CORRELATION 2

Contents

If viewing on screen you can click on a contents item above to jump to the page the item is on

1. A variety of correlation coefficients

Although the most useful correlation coefficient in the analyses of the results of psychological assessment is the Pearson Product Moment r, several other correlation coefficients exist which are occasionally useful, either in the interpretation or construction of tests.

The Table below summarises several of these and inicates the situations in which they are used.

Varieties of correlation coefficients	
Correlation coefficient	Situation in which used
Pearson product moment	When both variables are continuous
Biserial	One scale continuous, the other an artificial dichotomy of a continous scale
Point biserial	One scale continuous and the other a true dichotomy
Phi	Both scales are dichotomous
Spearman’s Rho	Both scales are rank orders
Kendall’s Tau	Both scales are rank orders

2. Biserial correlation coefficient

This correlation coefficient is is used when both of two continuous distributions have been dichotomised in some way.

Suppose we had a group of elderly people to whom we had given a memory for designs test. Suppose further that we decided to divide our sample into those aged 69 or less and those aged 70 or more.

Hypothetical data are shown in the table below.

Age	Score on test
65	10
67	7
67	8
68	3
69	9
71	6
72	7
74	4
75	5
79	2

We could of course simply calculate the usual correlation coefficient. If we do so we find its value equals - 0.72.

For the purposes of computing the biserial r we organise the data slightly differently

Memory test scores of older and younger groups	
Older (aged 70 or more)	Younger (aged 69 or less)
6	10
7	7
4	8
5	3
2	9
Mean score (M_1) = 4.8	Mean score (M_0) = 7.4
Proportion of older people = p = 0.5	Proportion of younger people = q = 0.5
Standard deviation of total group = 2.6	
Ordinate of normal distribution which corresponds to the division point between p and q = .3989 (from normal curve table)	

The formula for the biserial correlation is:

$$\frac{M_1 - M_0}{\sigma_t} \times \frac{pq}{y_o}$$

where:

M_1 = the mean continuous variable score (test score) of the higher group on the dichotomised variable (in this case the mean test score of the older group)
M_u = the mean continuous variable score of the lower group on the dichotomised variable
p = the proportion of all the cases who are in the higher group
q = the proportion of cases who are in the lower group.
y – the ordinate of the normal curve at the point which divides p from q
σ_t = the standard deviation of all the scores on the continuous variable

Applying this formula to the data in the table above we get

$$\frac{4.8-7.4}{2.6}\times\frac{.50\times.50}{.3989}=-.63$$

(Incidentally, if we take a cut-off memory score of 6 or less versus 7 or more, the Phi coefficient for these data is .60. Phi is discussed in a later section)

3. The point biserial coefficient

The point biserial coefficient is used when on of the variables is a dichotomy and the other is a continuous variable.

This coefficient has the formula:

$$r_{p.bis}=\frac{M_1-M_0}{\sigma_t}\times\sqrt{pq}$$

where:

M_1 = the mean score of those in one category of the dichotomised variable

M_0 = the mean score of those scoring in the other category

p = the proportion scoring in the first category

q = the proportion scoring in the other category.

σ_t = the standard deviation of all the scores on the continuous variable

The point biserial correlation is often used in item analysis, so lets apply the formula to a grossly oversimplified example.

The data in the following table show scores on each of the five items of a test of depression. These are the dichotomous variable, each being scored 0 or 1. We wish to correlate the items with total scores on the proposed test to see which items correlate most highly with total score.

People	Item 1	Item 2	Item 3	Item 4	Item 5	Total score
A	1	1	0	0	0	2
B	1	1	0	1	1	4
C	0	0	0	1	1	2
D	0	1	1	0	1	3
E	1	1	1	1	1	5
p	.60	.80	.40	.60	.80	

Let’s take Item 1

The mean total score of those who pass it is (2 + 4 + 5)/3 = 3.7

The mean score of those who fail it is (2+3)/2 = 2.5

The proportion who pass the item is so p = .6 and q = (1 - .6)

The standard deviation of the total scores (y) is 1.3

So for this example $r_{p.bis}$ will equal :

$$\frac{3.7-2.5}{1.3}\times\sqrt{(.6\times .4)} = 0.59$$

This is the correlation between Item 1 and total score

Suppose we now repeat the process for Item 3.

This time the mean score of those passing the item will be ((5+3)/2 = 4.0

The mean score of those who fail the item will be (2+4+2)/3 = 2.7

p this time is .4 so *q* will be .6, therefore r_{pbis} will be;

$$\frac{4.0-2.7}{1.3}\times\sqrt{(.4\times .6)} = 0.49$$

Thus Item 3 is a poorer predictor than Item 1

It must be emphasised that this is a very simplified example. In real life there would be many more subjects, and of course we would want to test the significance of the differences between the correlations of the individual items with total score.

4. The Phi Coefficient

The Phi Coefficient, symbolised as r_{phi} or as ϕ is used when both variables to be correlated are dichotomous.

As an example, suppose we are interested in the strength of the relationship between early morning waking and depression. Hypothetical data are shown below.

	Depressed	Not depressed	Totals
Early morning waking	50	10	60
Normal sleep	20	50	70
Totals	70	60	130

This is of course the sort of 2 x 2 table for which we often compute Chi Square.

Interestingly enough one of the formulas for Phi is as follows:

$$r_{phi} = \sqrt{\frac{\chi^2}{N}}$$

The more usual formula is derived from the 2 x 2 table as follows:

		Variable 2		
		yes	no	Total
Variable 1	yes	a	b	a + b
	no	c	d	c + d
	Total	a + c	b + d	a + b+ c +d

The formula is:

$$\frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

In the example above r_{phi} becomes;

$$\frac{(50 \times 50) - (10 \times 20)}{\sqrt{60 \times 70 \times 60 \times 70}} = 0.548$$

5. The Binomial effect Size Display

Rosenthal, Rosnow, and Rubin have, in a number of papers, argued for the use of the binomial effect size display in the interpretation of correlation coefficients.

For example the correlation between receiving psychotherapy and cure from a neurotic disorder was, at one time, thought to be 0.30.

How should we interpret this? If we use the coefficient of determination we would say that psychotherapy accounts for about 9 percent of the variance in cure rate.

Rosenthal and Rubin proposed using the binomial effect size display instead

With a Phi Coefficient of 0.3 the two by two table would look like this.

	Proportion cured	Proportion not cured
Psychotherapy	65	35
No psychotherapy	35	65

It turns out that the proportions occurring in the cells can be easily worked out as follows

	Cured	Not cured
Psychotherapy	A = 0.5 + r/2	B = 1 - A
No psychotherapy	C = 0.5 – r/2	D = 1 - C

Phi , it turns out, is, in this situation, equal to the difference in the success rates between the two groups.

This will be true in all situations where we have the subjects divided so that 50 percent are in the experimental (treated) group, and 50 percent in the control group, and where the success rate overall is 50percent.

So, suppose the correlation between psychotherapy and cure had been .4 instead of .3. What would have been the values in the various cells?

The table would now look like this.

	Cured	Not cured
Psychotherapy	70	30
No psychotherapy	30	70

Or, to return to our age and memory for designs example, where the Phi Coefficient was 0.6, What proportions would we expect in in each cell of a 2 x 2 table?

	Memory score 7 or higher	Memory score 6 or lower
Aged 69 or less	A = 0.5 + .3 = .8	B = 1 - .8 = .2
Aged 70 or more	C = 0.5 – .3 = .2	D = 1 - .2 = .8

The values we actually got were as follows

	Memory score 7 or higher	Memory score 6 or lower
Aged 69 or less	4	1
Aged 70 or more	1	4

Thus with 50-50 splits in the data the Binomial Effect Size display is a very convenient and easily understood interpretation of correlation coefficients.

Suppose however that we apply the psychotherapy results to a situation where, say, only 10 percent get psychotherapy and 90 percent do not, and we look at 1000 patients

The numbers in our 2 x 2 table will now look like this:

	Number cured	Number not cured
Psychotherapy	.65 x 100 = 65	.35 x 100 = 35
No psychotherapy	.35 x 900 = 315	.65 x 900 = 585

Phi is now only 0.185.

If we were to work out the proportions expected in the given cells, we would get:

	Cured	Not cured
Psychotherapy	A = 0.5 + r/2 = 0.5925	B = 1 – A = 0.4075
No psychotherapy	C = 0.5 – r/2 = 0.4075	D = 1 – C = 0.5925

So the value of Phi will change as the the numbers in the experimental and control samples depart from being equal.

The BESD will also change as the success rate changes. When there is an overall fifty percent success rate, Phi will equal the difference between proportions successful in the experimental and control groups. Let’s call this difference ‘*d*’.

Preece has shown that the formula for estimating the difference between experimental and control proportions successful is:

$$d = 2\phi\sqrt{s(1-s)}$$

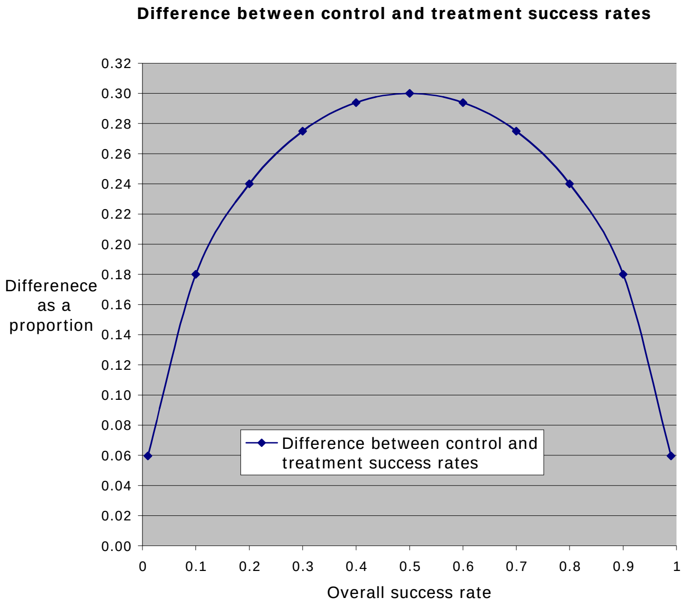
where:

d = difference between control and treatment success rates

s = the overall success rate.

When, and only when, the overall success rate is = .50 does d = Phi.

How would the difference between control and treatment success rates change if the correlation between psychotherapy and treatment remained aat .30, but the overall (control plus treated) percentage of patients cured varied? The graph below gives the answer.



As the cure rate departs from an overall 50 percent, so the difference in cure rates between treated and control groups lessens.

But within the overall cure rate range of 40 to 60 percent there is very little difference from the difference found with a cure rate of 50 percent.

So in general the BESD seems a reasonable way of illustrating/interpreting a correlation coefficient if the success/cure/hit rate lies between 40 and 60 percent.

Applying this to assessment situations

We can use the BESD to estimate or approximate the proportions of people falling into various categories just from the correlation between two tests.

For example, suppose we have a test of auditory and a test of visual memory, which correlate .7 with one another. What proportion of people of those who are below median on one test would also be below median on the other?

Using the BESD we could set up the following table

		Visual	
		High	Low
Auditory	High	A =.5 + (.7/2)	B = 1-A
	Low	C =.5 + (.7/2)	D = 1 - C

		Visual	
		High	Low
Auditory	High	A =.85	B = .15
	Low	C =.15	D = .85

As you can see, we would expect85 percent of those who are below median on the auditory memory test to also be below median on the visual memory test.

Test yourself
What would you expect the proportions to be if the correlation had been .5?

Answer

		Visual	
		High	Low
Auditory	High	A =.75	B = .25
	Low	C =.25	D = .75

If we know that the success/hit rate departs considerably from .5 we can use Preece's formula (given earlier). The difference in hit rates (d) will be:

$$d = 2\phi\sqrt{s(1-s)}$$

But in the case of diagnostic tests we usually have the data we need in the table on which we based our Phi coefficient anyway.

6. Estimating a correlation coefficient from a significance test result.

If we want to find the correlation between two variables and al that we have is a t-test value for the difference between them , or a Chi-square value with one degree of freedom we can still estimate the correlation between the dependent and the independent variables.

In the case of a t test we can estimate the point biserial r by using the formula:

$$r_{p.bis} = \sqrt{\frac{t^2}{t^2 + df}}$$

In the case of Chi-square, with 1 degree of freedom, the value of Phi coefficient (as mentioned earlier is :

$$r_{phi} = \sqrt{\frac{\chi^2}{N}}$$

these correlations can then be used in their usual ways.