

Quantifying association in ordinal data

Pamela Warner

Background

Spearman's non-parametric correlation coefficient has been used to explore association in the paper by Kulczycki *et al.*¹ in this issue of the Journal. These notes are intended to provide a supplementary explanation of this method. [See Box 1 for a glossary of terms used in this article.]

What is non-parametric correlation?

A non-parametric (or ordinal) correlation coefficient for two variables quantifies the extent to which there is, in the study sample, a consistent relationship between the values for the two variables. The correlation coefficient indexes the strength of the relationship found, from zero (no discernible relationship) to a maximum value of 1 (denoting a perfect relationship). There are two main methods for obtaining non-parametric or rank correlation coefficients: Spearman rho (ρ) and Kendall tau (τ). The general principles are the same (both are based on calculations using ranked data) and for most purposes the interpretation is very similar.² This Noteworthy Statistics (NS) explanation will focus on Spearman rho, as used in the Kulczycki *et al.* paper.¹

An ordinal association might be manifested in two ways: direct or inverse. The most straightforward association to envisage is a **direct** relationship, where an individual who gives a 'higher' response (relative to group) on the one variable, tends also to give a 'higher' response on the other variable (or in the case of other individuals, mid-responses with mid, or lower with lower). An example might be the association between parity and 'wish for permanent contraception'. Alternatively, the true situation might be an **inverse** relationship, where higher values on the one variable tend to be found in association with *lower* values on the other variable (and vice versa) (e.g. between 'degree of concern about hormone use' and 'acceptability of hormonal implant contraception'). Direct and inverse relationships will be distinguishable from analysis output because, under conventional numerical coding schemes for responses, direct relationships result in positive correlation coefficients (>0 to 1), and indirect relationships in negative correlation coefficients (<0 to -1). This is why inverse relationships are also commonly termed **negative associations/correlations**. [In Table 2 in the Kulczycki *et al.* paper, all the correlations with 'experience in diaphragm fitting' are negative (i.e. the more experience, the less the perceived barrier).¹]

When/why is it useful?

When the aspect of interest is the strength of ordinal association between two variables, and both variables are at least ordinal (see Glossary), a non-parametric correlation coefficient is a succinct and informative method of summarising this. In the case of categorical ordinal

variables, Chi-square could be used to test the null hypothesis (NH) of 'no association' between them.³ However, there are two compelling advantages to non-parametric correlation:

1. Non-parametric correlation provides a single summary statistic (rho) that quantifies the strength of the association between a pair of variables. This is clearly useful when there are a number of factors under study (such as the seven potential barriers and four provider characteristics in Kulczycki *et al.*'s study)¹ and there is an interest in ascertaining/reporting the extent to which the provider characteristics are associated with barrier responses (28 associations in all). Furthermore, if there is a wish to draw inferences about the true strength of a correlation (rho), in the population from which the study sample was drawn, then the population estimate for rho can be reported with a confidence interval (CI), in accordance with good statistical practice. In contrast, Chi-square does *not* provide such a summary statistic (nor CI); it merely ascertains the significance *p*-value for the test of 'no association'.

2. Non-parametric correlation provides a more powerful test of ordinal association than Chi-square. Rho can be tested against the NH of no correlation ($\rho = 0$); and since rho encapsulates all the information in the data about the ordinal association between the two variables, the test also utilises this complete information. The Chi-square test, by contrast, is suitable for use with nominal data (i.e. categorical variables with no inherent ordering, such as blood group). Although it can be, and often is, used with ordinal variables, it takes *no account* of the ordering in the data. This means that it misses important information about the association that is available in the data, and hence it is an underpowered test of ordinal association.

What precautions are needed?

The only requirement for non-parametric correlation is that the two variables are at least ordinal. No assumptions are needed regarding the distributions of the responses for the two variables being correlated, so this method is suitable also for continuous variables, regardless of distribution.

Example

Table 1 shows some hypothetical data for two variables from a survey: 'extent of professional experience' and respondent's perception of the 'impact of a specific side effect on discontinuation of contraceptive method'. If Chi-square is applied to this table to test the NH of no association, it is found that Chi-square = 9.1 (6 df) and $p = 0.17$. Therefore, on the evidence of this survey the NH is entirely plausible. However, both variables are

Table 1 Side effect A impact responses by extent of professional experience ($n = 173$)^a

Professional experience Impact response	Low ($n = 49$)	Mid ($n = 69$)	High ($n = 55$)
No impact	10	13	7
Moderate	15	14	8
Strong	16	25	19
Very strong	8	17	21

^aData are fictitious, created for the purpose of illustration only.

J Fam Plann Reprod Health Care 2010; **36**(2): 83–85

Centre for Population Health Sciences, University of Edinburgh Medical School, Edinburgh, UK

Pamela Warner, BSc, PhD, Reader in Medical Statistics and Associate Editor, *Journal of Family Planning and Reproductive Health Care*

Correspondence to: Dr Pamela Warner, Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: p.warner@ed.ac.uk

conceptually ordinal. If Spearman's non-parametric correlation rho is calculated instead, it is found that $\rho = 0.21$, and if rho is tested against the NH of zero correlation, the p -value obtained is 0.006. So, applying the more appropriate ordinal analysis to these data, it is found that there is strong evidence against the NH. Therefore, the NH can be rejected, and we conclude there is an ordinal association. Figure 1 graphs these data, plotting the 'impact' responses as percentages for each experience subgroup (and stacking them up in the subgroup column), while the 'no impact' responses, although not plotted, would make each column up to 100%. The ordinal

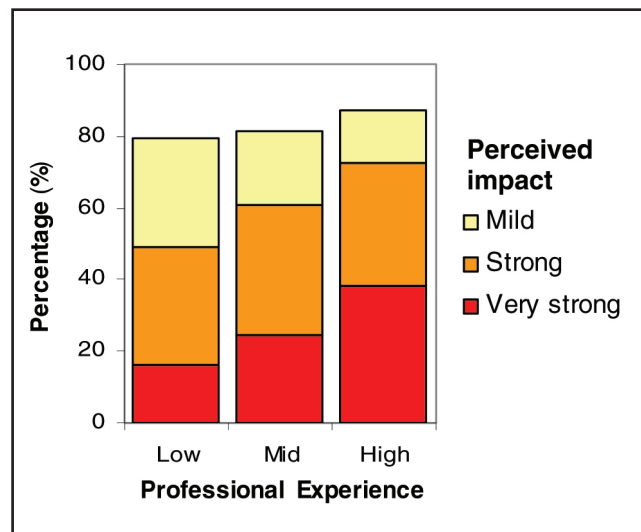


Figure 1 Perception of impact of side effect A by professional experience of respondents

association found above is confirmed by the graph, which shows that the greater the professional experience of the respondent, the more frequent the perception of impact at all (increasing heights of columns) and, in addition, the greater the experience, the more frequent the perception of 'very strong' impact, rather than 'mild'.

If the survey had been twice as large ($n = 346$), but obtained exactly the same pattern of combinations of responses on the experience and impact variables, so that Figure 1 (in terms of percentages) still represents the (doubled) data graphically, then the non-parametric correlation would be unchanged, at 0.21, but the p -value for the test against the NH would be much smaller, at 0.0001. This makes sense – the quantification of the strength of association is unchanged, because the association itself is unchanged, but the evidence against the NH is much stronger (p -value much smaller), because the calculated correlation is based on a larger amount of data, and hence is much more dependable as representing the truth.

Figure 2 presents results for analyses of perceived impact for other (hypothetical) side effects: B which appears to show virtually no association between experience and perceived impact; C which shows a stronger correlation than side effect A; and D, which shows the situation where there is an inverse correlation, such that the greater the experience of the respondent, the less frequently perceived was the impact of the side effect D, and where impact was perceived, the milder it was judged to be. Table 2 summarises the non-parametric correlation statistics for all the examples discussed, and also presents Chi-square test results. This confirms that whenever there is an ordinal association, the non-parametric correlation provides a more powerful test than

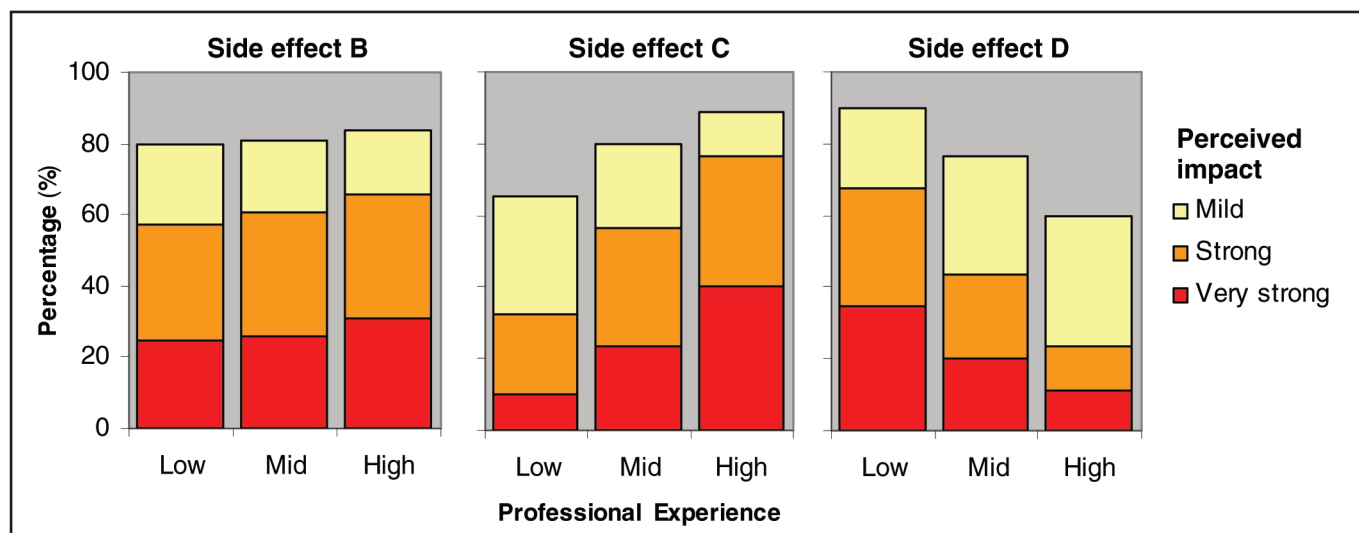


Figure 2 Impact responses by experience, for three further side effects

Table 2 Non-parametric correlations by side effect, and Chi-square tests^a

Side effect	<i>n</i>	Spearman correlation		Tests of rho <i>p</i> -value	Chi-square tests	
		rho	95% CI		Chi-square	<i>p</i> -value
A	173	0.21	0.07 to 0.35	0.006	9.1	0.17
A (<i>n</i> doubled)	346	0.21	0.11 to 0.31	0.0001	18.3	0.006
B	173	0.07	-0.08 to 0.22	0.35	0.96	0.99
C	173	0.35	0.22 to 0.48	0.000002	21.7	0.0009
D	173	-0.35	-0.48 to -0.22	0.000002	21.7	0.0009

^aData analysed are fictitious, created for the purpose of illustration only. CI, confidence interval.

Chi-square (i.e. gives smaller p -values). In the case of Survey D, it can be seen that apart from the sign, the rho and 95% CI are numerically identical with Survey C, meaning that this happens to be an equally strong correlation as for C, but in the reverse direction. Similarly, the p -value is identical, because there is exactly equal strength of the evidence against the NH, in terms of both how far the sample correlation rho is from the null (0.35), and the equal study size ($n = 173$). The results for the two A analyses show that while Spearman rho is a stable representation of the strength of association across different study sizes, it is more precisely estimated in the larger study (narrower CI), and that the Chi-square statistic is not stable, since it increases with n even when the correlation is unchanged. This is because it does not simply reflect the strength of the association, but also reflects the amount of evidence (i.e. the study size).

Overview

“Rank correlation should be used more often”.² Spearman rho provides a useful and succinct summary of the ordinal association between two variables. It can be used to test the NH of no ordinal association much more powerfully than the standard Chi-square test, and CIs can be calculated for estimates.

Statements on funding and competing interests

Funding None identified.

Competing interests None identified.

References

- 1 Kulczycki A, Qu H, Bosarge PM, Shewchuk RM. Barriers to diaphragm use: the views of advanced practice nurses. *J Fam Plann Reprod Health Care* 2010; **36**: 79–82.
- 2 Altman DG. *Practical Statistics for Medical Research*. London, UK: Chapman & Hall/CRC, 1999.
- 3 Warner P. Testing and quantifying association in binary data. *J Fam Plann Reprod Health Care* 2009; **35**: 26–27.

Box 1: Glossary of statistical terms used in this article

Association	Relationship between two variables. For categorical variables, the data can be envisaged as a cross-tabulation of counts of respondents with each combination of values for the two variables. ‘An association’ means the occurrence for an individual of a particular value of one variable is associated with (more likely to be in conjunction with) a particular value of the other variable. ‘No association’ means the distribution of values will be approximately the same in each column or, to put it a different but equivalent way, in each row. See also <i>Ordinal association</i> .
Categorical variable	This has a set of distinct values that can be nominal (i.e. simply descriptive, such as blood group) or ordered (such as degree of impact, duration of professional experience).
Chi-square test	Test applied to cross-tabulated data for two categorical variables, to assess association between them. It is designed for nominal data (no inherent ordering) so for any table of counts (e.g. Table 1), the same test result would be obtained whatever the order of rows and/or columns.
Confidence interval (CI) (95%)	This defines a range of values within which we are 95% confident the true population effect (in this article, rho) lies.
Correlation coefficient	See <i>Spearman rho</i> .
Degrees of freedom (df)	Can be thought of as the information capacity (or independent elements of information) in the data set. In the case of Chi-square, $df = (\text{the number of rows less } 1) \times (\text{the number of columns less } 1)$.
Direct association	See <i>Ordinal association</i> .
Inverse association	See <i>Ordinal association</i> .
Nominal variable	Has a set of distinct values, such as gender, recruitment setting.
Non-parametric	General label given to statistical methods that do not depend on distributional assumptions for the data values, and which in the main utilise ranked data. However, it is rho only that is truly non-parametric – the test of rho against NH, and calculation of CI, utilise asymptotic properties of rho (i.e. the way rho behaves for relatively large n), and hence depends on parametric distributions representing these.
Null hypothesis (NH)	A statement, prior to testing, of no effect (in this case, no association, or zero correlation). See also <i>Significance probability</i> .
Ordinal association	This is association between two ordinal variables: direct or inverse. A direct association applies when an individual who gives an ordinal ‘higher’ response (relative to group) on one of the two variables tends also to give a ‘higher’ response on the other variable (and by extension across the study group, mid-responses occur with mid, and lower with lower). Conversely, in an inverse association, higher values on the one variable tend to be found in association with lower values on the other variable (and vice versa). See also <i>Spearman rho</i> .
Ordinal variable	Variable where the values are conceptually ordered, such as categorical variables (e.g. degree of pain – none, mild, moderate, etc.), count variables (e.g. parity – 0, 1, 2, etc.) and continuous variables (e.g. weight – 62, 74, 91, etc.). For analysis of categorical variables it is conventional to assign numeric codes to the ‘values’ (e.g. 1, 2, 3 for moderate, severe, etc.). [Note that, if a variable is not ordinal conceptually, such as blood group, then such assigned numeric codes, despite looking like an ‘ordering’, do <i>not</i> make it ordinal!] It is important that codes assigned to ordinal variable ‘values’ reflect the implicit ordinal trend (i.e. increase or decrease as appropriate). So, for impact (none, moderate, strong, very strong) codes 3, 5, 9, 14 would do (provided increasing, the actual numbers do not matter). Coding as 9, 0, 17, 12 would lose the ordinality altogether, whereas 4, 3, 2, 1 would reverse it. Thus the meaningfulness of correlation coefficients depends on the numeric coding assigned, as does the interpretation as to whether indicating negative or positive associations.
Negative correlation	See <i>Ordinal association</i> – inverse.
p-value	See <i>Significance probability</i> .
Positive correlation	See <i>Ordinal association</i> – direct.
Power	This term is used here, loosely, as the probability of detecting from study data what is in fact the real situation.
Rank correlation	See <i>Spearman rho</i> .
Significance probability (p-value)	The probability, if the NH is true, of obtaining the observed data (combinations of responses on the two variables). The smaller p is, then the less likely these data would be under the NH, and so the greater our doubts that the NH is indeed true.
Spearman rho (ρ) (non-parametric correlation coefficient)	Spearman rho is an index of the strength of ordinal association between values for two variables measured on the same individuals. Rho takes values between -1 and 1 , with zero indicating no correlation, a positive value indicating a direct or positive correlation, and a negative value an inverse or negative correlation. Values 1 and -1 indicate perfect (direct or inverse) correlation. See also <i>Ordinal association</i> .