STATISTICAL METHODS FOR ASSESSING AGREEMENT FOR ORDINAL DATA

Ulf Jakobsson, PhD, Senior lecturer

Albert Westergren, PhD, Senior lecturer

Department of Health Sciences, Faculty of Medicine, Lund University

P.O. Box 157, SE-221 00 Lund, Sweden

Address for correspondence;

Ulf Jakobsson, Department of Health Sciences, Faculty of Medicine, Lund University.

P.O. BOX 157, SE-221 00 Lund, Sweden.

E-mail: Ulf.Jakobsson@med.lu.se

Telephone: +46 46 222 19 24, Fax: +46 46 222 19 34

ABSTRACT

Evaluation of various methods in clinical practice is often based on interpretations by two or more observers. Such data need to be analysed with correct statistics, or the results and conclusions may be misleading. In this article the use of measures of agreement for ordinal data in five international nursing journals is reviewed and various methods for measuring agreement are presented and discussed. Analyses of agreement did not seem to be very common in nursing research, but a great variation was found regarding the choice of method for analysing agreement. Both acceptable and unacceptable methods were found in the articles reviewed. When choosing among various methods for agreement the *weighted* Kappa coefficient is probably the most useful for ordinal data, but several issues of concern arise and need to be taken into consideration when using these types of analyses.

Keywords: Ordinal data, Agreement, Statistics, Kappa, Kappa statistics, Kappa coefficient, Nursing, Caring Sciences

INTRODUCTION

In clinical research, agreement between observers is often analysed when evaluating various methods. Agreement between observers (inter-rater agreement) can be measured in different ways, and some methods may be regarded as more accurate than other. Depending on which method one uses, one can obtain quite different values (1). The Kappa coefficient (2) has traditionally been used to evaluate inter-rater reliability between observers of the same phenomenon, and was originally proposed to measure agreement by classifying subjects in nominal scales, but it has since been extended to the classification of ordinal data as well. Other measures such as percentage agreement (also called exact agreement) and weighted Kappa coefficient are also used in various studies.

Ordinal data in general are often not presented or analysed appropriately in research studies, as has been shown in previous reviews (3, 4, 5). Avram (3) and colleagues reviewed two American anaesthesia journals from 1981 and 1983; Lavalley and Felson (4) reviewed three rheumatology journals from 1999; and Jakobsson (5) reviewed three nursing journals from 2003 for their presentation and analysis of ordinal data. The reviews found appropriate presentation in about 39–49% of the articles and appropriate analysis in 57–63%. However, these reviews had a general focus and hence more studies focusing on specific analyses such as inter-rater reliability are needed.

AIM

To review the literature regarding the use of statistical methods for measuring agreement for ordinal data, and show various examples of and discuss how to most appropriately measure agreement for this type of data.

THE REVIEW OF THE LITERATURE

A review of the literature (international peer-reviewed nursing journals) regarding the use of measures for agreement on ordinal scales was performed. The review comprised all the 2004 issues of *Applied Nursing Research, Nursing & Health Sciences, Nursing Research, Pain Management Nursing* and *Scandinavian Journal of Caring Sciences*. Only full-length research articles were reviewed. Ordinal data in the articles were identified according to the criteria of Siegel and Castellan (6).

INSERT TABLE 1

A total of 183 articles were reviewed and 9 (4.9%) of the articles were found to analyse agreement for ordinal data (Table 1). Ninety-six (60%) articles were quantitative studies, 44 (24%) were qualitative studies, while 14 (8%) had a combination of a quantitative and a qualitative design. The rest of the articles were categorised as "other articles" (e.g. review articles, methodological articles). The most common analyses for agreement of ordinal scale were (unweighted) Kappa analysis (n=3) and weighted Kappa (n=4). Other methods were percent of agreement (n=3), percentage of "acceptable agreement" (± 1 point difference) (n=1), Spearman rank-order correlation (n=1), paired Student's t-test (difference in total score) (n=1), some kind of unknown "inter-rater reliability" analysis (n=1) and another unknown "testing for content validity using a 5-point Likert scale" analysis (n=1).

The number of articles that handled analyses of agreement for ordinal data was found to be rather small, one reason for which might be that the review covered only one year. If the review had been broadened to include more years it would probably not have given any different result. A review (5) of issues from 2003 in three international nursing journals confirms this assumption, hence we do not see any reason to think that the sample is skewed.

MEASURES OF AGREEMENT

Inter-rater agreement can be calculated as *percentage agreement*, *Cohen's Kappa coefficient (K)* and *weighted Kappa coefficient (K_W)*. The different methods will be illustrated and discussed below using an example based on fictitious data (Table 2). The easiest way of calculating agreement is percentage of agreement, that is, the sum of the diagonal of the matrix (25+9+12+21=67) divided by the sum of the observations (N=85). The percentage of agreement (i.e. exact agreement) will then be, based on the example in table 2, 67/85=0.788, i.e. 79% agreement between the grading of the two observers (Table 3). However, the use of only percentage agreement is insufficient because it does not account for agreement expected by chance (e.g. if one or both observers were just guessing and/or the agreement happened by chance).

INSERT TABLE 2

Correlation is sometimes also used as a measure of agreement. However, correlation, like the chi-square test, is a measure of association and does not satisfactorily measure agreement (1, 7). Association can be defined as two variables that are not independent, while agreement is a special case of association where the data in the diagonal (perfect agreement) are of most interest. It should be noticed that perfect association does not automatically mean perfect agreement because a perfect correlation (r=1.0) can be obtained even if the intercept is not zero and the slope is not 1.0. An example of this is when one of the observers constantly grades the scores a little higher than the other observer. This will give a high association but low agreement. Thus, correlation does not account for systematic biases. Furthermore, the correlation coefficient tends to be higher than the “true” reliability (8). This can be seen in our example when Spearman’s rank order correlation is compared with the other three methods (Table 3).

INSERT TABLE 3 & 4

When two or more observers are asked to grade an item on some criterion, Cohen’s Kappa coefficient is an adequate method to measure agreement. The advantage of the Kappa statistic (K) is that it does account for both percentage agreement and the percentage of agreement expected by chance. The interpretation of Cohen’s Kappa coefficient is (theoretically) 1.0 for perfect agreement while chance agreement would equate to zero. Values are often interpreted as follows: below 0.20 regarded as poor, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as good and >0.80 as very good agreement (7). Landis and Koch (9) had a classification that had more categories; 0.00 was regarded as poor, 0.00–0.20 as slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect. Unlike Spearman’s rank order correlation, the Kappa coefficient accounts for systematic biases. Kappa is calculated as the ratio of the observed and expected frequencies (Table 4):

$$K= \frac{p_o - p_e}{1 - p_e}$$

Observed proportional agreement: $p_o = \sum \frac{\mathrm{f}_{\mathrm{ii}}}{n} = \frac{a+d}{n}$

Expected proportional agreement: $p_e = \sum \frac{\mathrm{r}_{\mathrm{i}}\ \mathrm{c}_{\mathrm{i}}}{n^2} = \frac{\mathrm{n}_1 \times \mathrm{n}_3}{n^2} + \frac{n_2 \times n_4}{n^2}$

where n is the total number of observations and f_{ii} is the number of agreements in the diagonal, r_i and c_i are the row and columns totals respectively for the *i*th category.

INSERT TABLE 4 & 5

Cohen's Kappa coefficient, as stated earlier, considers both percentage agreement and the percentage of agreement expected by chance. However, the limitation of the Kappa statistics is that it does not take any account of the degree of disagreement. In fact, all disagreement is treated equally (i.e. the unweighted Kappa gives zero weight to all disagreement cells). This is an important issue when calculating agreement of ordinal data and the magnitude of disagreement becomes significant. To "solve" this problem a weighted Kappa coefficient can be calculated, which is a generalisation of the unweighted Kappa coefficient (10). Different weights are given according to the magnitude of the disagreement. The determination of weights is a rather subjective issue and some different ways have been proposed for giving the disagreements their weights (cf. 10, 11, 12). To be noticed, the coefficient of the weighted Kappa can vary due to the weighting method (i.e. choice of weights) employed (13). In this article we illustrate a rather common and easy way to calculate weighted Kappa, known as "absolute error weights".

If *i* denotes the cells in rows and *j* denotes the columns, the weight w_{ij} is calculated as:

$$w_{ij} = 1 - \frac{(|i-j|)}{(g-1)}$$

where g is the number of categories and ($1 \leq i,j \leq g$). Thus, all cells on the diagonal give a weight of 1. The other weights will then be as presented in table 5. The value of weighted

Kappa in table 3 is calculated with “absolute error weights”. Another method for weighting is the “square error weights” which are calculated as:

$$w_{ij} = 1 - \frac{(i-j)^2}{(g-1)^2}$$

This method is similar to the one previously described, but the weights and the Kappa value will be a little bit different.

Regardless of what kind of weighting method that is used the weighted observed (p_{ow}) and expected (p_{ew}) agreement are then obtained as:

$$p_{ow} = \frac{\sum\sum w_{ij} f_{ij}}{n}$$

$$p_{ew} = \frac{\sum\sum w_{ij} r_i c_j}{n^2}$$

and, the weighted Kappa are calculated as:

$$K_w = \frac{p_{ow} - p_{ew}}{1 - p_{ew}}$$

The interpretation of the weighted Kappa coefficient is the same as for the unweighted one, where 1.0 means perfect agreement while zero signifies chance agreement. Most of the statistical software also present some kind of hypothesis testing where a p-value is obtained. The null hypothesis is often set to H_0: K=0, which means that a significant p-value only tells us that the Kappa value is not zero (with some certainty). Thus, the test of significance may seem irrelevant to the questions of agreement, especially when comparing two methods that are developed to measure the same phenomenon. A 95% confidence interval (CI) will then give us more information about the estimation of the Kappa value. However, this will imply a not too large sample because too large a sample will give a very narrow confidence interval, hence will make the interpretation difficult. Firstly a standard error, *se*(K), is calculated, which in turn is used to calculate the confidence interval. An approximate standard error can be calculated as follows:

$$se(\mathrm{K}) = \sqrt{\frac{\mathrm{p_o}(1-\mathrm{p_o})}{(\mathrm{n}(1-\mathrm{p_e})^2}}$$

and the 95% confidence interval will then be:

$\mathrm{K} \pm 1.96 \times se(\mathrm{K})$

In the example (Table 2 & 3) the *se*(K) will be 0.060 and the 95% CI will be ranging from 0.591 to 0.827 for the unweighted Kappa value. For the weighted Kappa in table 3 the *se*(K) will be 0.028 and the 95% CI will range from 0.72512 to 0.8349.

Is then the weighted Kappa really the best way to analyse agreement of ordinal data, with a minimum of errors? The answer to this is not clear, and some studies have discussed this issue. One well-known problem with Kappa statistics is that the Kappa value depends on the prevalence in each category, which leads to difficulties comparing the Kappa values from different studies with different prevalence in the categories (cf. 14). The number of categories in the variable assessed also affects the Kappa coefficient. With an increasing number of categories the Kappa value tends to be lower. In our example K=0.709 for four categories (Table 2), but collapsing the table into a 2x2 table where [0, 1]=0 and [2, 3] = 1 gives us K=0.857. To reduce this effect of changes in the number of categories, the intra-class correlation has been argued to be a better alternative because it is less sensitive to these changes (15). However, the intra-class correlation instead tends to increase with an increasing number of categories, so one may wonder whether this is a better method to be use for ordinal data. The intra-class correlation is probably best suited for continuous data.

Graham and Jackson (13) argued that the (squared error) weighted Kappa in some respects can be regarded more as a measure of association than of agreement. This is because the weighted Kappa coefficient is calculated on the marginal frequencies and is not sensitive for changes in the matrix i.e. when the figures in the diagonal change but the marginal frequencies are unaffected. So, even if the exact agreement changes, the Kappa value is unchanged if the row and column marginal are unchanged (13).

Weighted Kappa, as previously stated, is a measure of the agreement (reliability) of ordinal data. In the same way, the intra-class correlation is often used as a measure of reliability in quantitative scales. These two measures are known to be used on the respective scale types and should not be used interchangeably. However, some authors have presented some exceptions to this “rule”. Cohen (2) has previously shown that for a 2×2 table where the marginal distributions are identical, the same Kappa coefficient (weighted as well as unweighted) may be interpreted as the phi coefficient. Furthermore, for a general $m \times m$ table with identical marginal distributions and weights, weighted Kappa is equal to the product-moment correlation (10). However, such comparison is only valid for ordinal scales where the category is scored 1 for the first category, 2 for the second, and so on. Fleiss and Cohen (16) argued in their article that weighted Kappa coefficient is asymptotically equivalent to the intra-class correlation. This equivalence is only true when using square error weights and when systematic variability between observers is included as a component of total variation (16). Thus, there are some similarities (under certain circumstances) between correlation and agreement (Kappa coefficient), but it seems much easier use the Kappa coefficient when investigating agreement for ordinal scales.

CONCLUSIONS

Even if analyses of agreement were not very common in nursing research, great variation was found regarding the choice of method for analysing agreement. Both acceptable and unacceptable methods were found in the articles reviewed regarding analyses of agreement for ordinal data as well as the handling of ordinal data in general. Several issues of concern appear when measuring agreement in ordinal data. The following list gives some key points regarding analysis of agreement discussed in this article:

- Chi-square test and correlation (both Pearson's and Spearman's) are measures of association and not agreement, hence are not be used to measure agreement in ordinal data.
- The *unweighted* Kappa cannot detect differences that are not in the diagonal of the matrix and therefore cannot provide a complete description of the agreement in ordinal data.
- The *weighted* Kappa coefficient is probably the most useful measure for agreement in ordinal data.
- However, it is important to remember that the Kappa coefficient depends upon the prevalence in the cells as well as the number of categories in the variable, which makes it difficult to compare results from different studies.
- Furthermore, because various weighting method are used in different studies, comparison of the studies is difficult.

ACKNOWLEDGEMENTS

The authors wish to thank Alan Crozier for revising the English. This study was supported by grants from the Vårdal Foundation and Department of Health Sciences, Faculty of Medicine, Lund University.

REFERENCES

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 327: 307-310.

2. Cohen JA. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37-46.

3. Avram MJ, Shanks CA, Dykes MHM, Ronai AK, Stiers WM. Statistical methods in anaesthesia articles: an evaluation of two American journals during two six-month periods. *Anesth Analg* 1985; 64: 607-611.

4. Lavalley MP, Felson DT. Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. *Arthritis Rheum (Arthritis Care Res)* 2002; 47: 255-259.

5. Jakobsson U. Statistical presentation and analysis of ordinal data in nursing research. *Scand J Caring Sci* 2004; 18: 437-440.

6. Siegel S, Castellan NJ Jr. *Nonparametric statistics for behavioral sciences*. 1988, McGraw-Hill, London.

7. Altman DG. *Practical statistics for medical research.* 1991, Chapman & Hall, London.

8. Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use.* (2nd ed.) 1995, Oxford University Press, Oxford.

9. Landis JR, Kock GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-174.

10. Cohen JA. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; 70: 213-220.

11. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin* 1971; 76: 365-377.

12. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; 76: 378-382.

13. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol* 1993; 46: 1055-1062.

14. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; 43: 543-549.

15. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; 126: 161-169.

16. Fleiss JL, Cohen JA. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973; 33: 613-619.

Table 1. The result from the literature review.

Journal	Number of articles (total)	Number of articles using ordinal scales and analysing agreement
Applied Nursing Research	34	3
Nursing & Health Sciences	36	0
Nursing Research	45	2
Pain Management Nursing	16	0
Scandinavian Journal of Caring Sciences	52	4

Table 2. An example of the outcome of the grading for two observers.

		Observer 1				
		1	**2**	**3**	**4**	**Total**
Observer 2	**1**	25	7	1	0	**33**
	2	3	9	1	0	**13**
	3	2	2	12	2	**18**
	4	0	0	0	21	**21**
	Total	**30**	**18**	**14**	**23**	**85**

Table 3. Agreement between the two observers' ratings of the items in Table 1. Comparison of Cohen's Kappa coefficient (weighted; K_w) and ("unweighted"; K), Spearman's rank order correlation (r_s) and percentage agreement.

Inter-rater agreement (n=85)			
K_w	K	r_s	*Percentage agreement*
0.780	0.709	0.877	0.788

Table 4. Matrix for calculating Kappa statistics

		Observer 1		
		Yes	**No**	**Total**
Observer 2	**Yes**	a	b	n_3
	No	c	d	n_4
	Total	n_1	n_2	n

Table 5. Weights given to each cell when calculating weighted Kappa

		Observer 1				
		1	**2**	**3**	**4**	**Total**
Observer 2	**1**	1.00	2/3	1/3	0	**33**
	2	2/3	1.00	2/3	1/3	**13**
	3	1/3	2/3	1.00	2/3	**18**
	4	0	1/3	2/3	1.00	**21**
	Total	**30**	**18**	**14**	**23**	**85**