TEACHING: A WAY OF IMPLEMENTING NOVEL STATISTICAL METHODS FOR ORDINAL DATA TO RESEARCHERS

Elisabeth Svensson
Department of Statistics, Örebro University, Sweden
elisabeth.svensson@oru.se

The use of questionnaires, rating scales and other kinds of ordered classifications is unlimited and inter-disciplinary, so it can take long time before novel statistical methods presented in statistical journals reach researchers of applied sciences. Therefore, teaching is an effective way of introducing novel methods to researchers at an early stage. Assessments on scales produce ordinal data having rank-invariant properties only, which means that suitable statistical methods are non-parametric and often rank-based. These limited mathematical properties have been taken into account in my research regarding development of statistical methods for paired ordinal data. The aim is to present a statistical method for paired ordinal data that has been successfully implemented to researchers from various disciplines together with statisticians attending interactive problem solving courses of biostatistics.

INTRODUCTION

The use of questionnaires, rating scales and other kinds of ordered classifications are used for assessments of attitudes, perceptions and other qualitative variables and also for expert-rating of the severity of a diagnosis, grading of performance in health sciences. Assessments on rating scales produce ordinal data having rank-invariant properties only, meaning that the ordered categorical responses, also when numerically labeled, indicate only an ordering and not a mathematical value (Stevens, 1955; Hand, 1996). These limited mathematical properties have been taken into account in my research regarding development of statistical methods that takes account of the non-metric properties of data and that allows for comprehensive analysis of paired ordinal data (Svensson, 1993). In a review regarding statistical developments of methods for ordinal data analysis Liu and Agresti (2005) state that making the novel methods better known is the main challenge for future, in order to increase the quality of analyses and of scientific conclusions.

Teaching is one effective way of implementing good statistical practice and novel statistical methods to researchers. PhD students from all kinds of disciplines have the need for statistical learning and understanding in common. They are often open to new statistical methods, but are also tied to the supervisor's preference to statistical methods. My experience is that the preference for traditional methods is stronger than the willingness to use novel statistical methods, even when the traditional methods are inappropriate or wrong (see Svensson, 2001a, 2002).

Students of statistics need for experiencing real life problem solving before graduation. They often have a solid knowledge of statistical methods suitable for quantitative data, preferably parametric ones. During the years I have experienced the power of teaching good statistical practice and of introducing novel less known statistical methods for ordinal data to researchers and to statisticians by running inter-disciplinary problem solving courses of biostatistics (Svensson, 2002, 2009).

The aim is to present one of the most frequently demanded statistical methods for paired ordinal data that has been successfully implemented to researchers from various disciplines together with statisticians attending interactive problem solving courses of biostatistics.

THE TEACHING MODEL

The key feature of the problem solving course is that it is multi-disciplinary and open both for researchers and for second year students of statistics. The course is interactive in the sense that the statistical methods taught are applied to the participants' research problems, when applicable, in exercises and assignments. Communication skills and mutual understanding of the researchers' problems are naturally trained and the statisticians are dependent on the researchers, as they need problems to be solved. The measurement process and the measurement properties of data are thoroughly discussed, especially regarding data from rating scale assessments. The requirements for passing the assignments differ in the sense that the statisticians must add a statistical

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

perspective to the choice of methods for design and analysis.Detailed descriptions of the courses, the examination processes etc have been reported (see Svensson, 1998a, 1998b, 2001a, 2002, 2009).

MEASURES OF AGREEMENT AND DISAGREEMENT IN PAIRED ORDINAL DATA

Many of the participating researchers' studies have qualitative outcome variables, and data are collected by questionnaires and rating scales. Inter-and intra-rater reliability studies and studies of change after treatment or intervention have so far been the dominating types of studies, and these issues deal with agreement and disagreement between paired ordinal data out of different approaches. In reliability studies disagreement must be avoided, in change studies observed disagreement could be the evidence of a desired treatment effect.

The non-metric properties of ordinal data have obvious consequences for the choice of statistical methods for paired ordinal data. The statistical method developed by Svensson for paired ordinal data evaluates the reasons for observed disagreement between paired assessments, when present. The method makes it possible to identify and measure systematic disagreement separately from disagreement caused by occasional variations. These two sources of disagreement have different impacts on the interpretations of the analysis, since systematic disagreement is population based and occasional variations are related to individual pairs of assessments (Svensson, 2001b, 1998c).

The main steps of the evaluation of agreement and disagreement between pairs of assessments are demonstrated by a worked example. Figure 1 shows the frequency distribution of ordinal data from 40 paired assessments on a five-point scale, the ordered categories being A<B<C<D<E, and the two sets of assessments are denoted X and Y, respectively. The diagonal of identical pairs of categories is marked. Eight out of the 40 pairs are scored (E,E) and another two, (D,D), hence the *percentage agreement, PA*, is 25%.

		X				
		A	B	C	D	E	Total
Y	E			3	10	8	21
	D		2	5	2	2	11
	C	2	4				6
	B	2					2
	A						
	Total	4	6	8	12	10	40

Figure 1. The frequency distribution of 40 pairs of assessments, (X,Y), the categories being A<B<C<D<E. The agreement diagonal is marked.

The two marginal distributions show the frequency distributions of ordered categorical assessments according to X and Y, respectively. Different marginal distributions are sign of *systematic disagreement* in assessments. Two measures of systematic disagreement are defined**;** the relative position, RP, and the relative concentration, RC. The *empirical measure RP* estimates the difference between the probabilities of the marginal distribution Y being shifted toward higher categories than X and the opposite, P(X<Y)-P(Y<X). Possible values of RP range from (-1) to 1, and a positive value indicates that the data set Y has systematically more assessments on higher categories than has X. The *empirical measure RC* estimates the difference between the probabilities of differences in how the assessments are concentrated on the scale categories, in short P(X<Y<X)-

P(Y<X<Y). Possible values range from (-1) to 1, and a positive RC indicates that the assessments Y more likely are concentrated to central categories than the assessments X.

The systematic part of the observed disagreement is defined by the marginal heterogeneity, and by pairing off the two sets of marginal distributions, *the rank-transformable pattern of agreement (RTPA)* is constructed and shows the expected paired distribution in the case of systematic disagreement only. Figure 2 shows the RTPA defined by the marginal distributions X and Y, Figure 1. According to the RTPA one can expect that the distribution of Y assessments will systematically being one or two categories higher than those of X.

		X					
		A	B	C	D	E	Total
Y	E				11	10	21
	D		2	8	1		11
	C	2	4				6
	B	2					2
	A						
	Total	4	6	8	12	10	40

Figure 2. The rank-transformable pattern of agreement defined by the two sets of marginal distributions, X and Y, of Figure 1.

The observed distribution of pairs, Figure 1, differs slightly from the RTPA. This dispersion of pairs indicates additional presence of *occasional, individual-based disagreement*. The *relative rank variance, RV,* is a rank-based measure of the observed occasional variability and is defined by the sum squares of rank differences when the ranks are tied to the pairs of observations in the cells, so called augmented ranks. Non-zero RV indicates presence of occasional variations and the higher the value of RV, the more dispersed are the paired assessments in relation to the RTP having no occasional variation, i.e. RV=0.

The marginal heterogeneity of the set of paired data in Figure 1 indicates presence of systematic disagreement. The significant RP-value 0.37 (95% CI, 0.24 to 0.50) confirms the presence of a systematic disagreement in position of the scale categories, also evident by the RTPA, Figure 2. In this example the RC value is negligible (0.10, 95% CI, -0.15 to 0.35). The additional occasional variation is negligible, as RV is 0.042 (95% CI, 0 to 0.1).

What are the interpretations of these statistical measures when the data set is supposed to come from a reliability study and from a study of treatment effect, respectively?

APPLICATION TO RELIABILITY STUDIES:

Reliability expresses the extent to which repeated assessments yield the same result, which means agreement in paired data. A high level of reliability requires lack of systematic disagreement, RP=RC=0, a high percentage agreement, PA, and an RV-value close to zero. The two types of disagreements have different impacts on the quality of scale and assessments. An important consequence of systematic disagreement, bias, is that the PA will never reach 100%. Systematic disagreement is related to the group and can be reduced when identified. However, as the RV value is a measure of additional variability after adjustment for bias, the RV value could reach zero both in biased and unbiased repeated assessments. Therefore, a small value of RV in the presence of bias indicates that after eliminating the source of bias the reliability will be very high. On the other hand, a high RV indicates that the questions or the scale categories do not fit well to

the objects being classified and that the assessments are sensitive to disturbing factors of the test situation.

In *inter-rater reliability studies* the paired data are obtained by assessments made by two raters, X and Y, of the same objects. Referring to the paired distribution of Figure 1 the 75% disagreement is mainly explained by a systematic disagreement in how the raters interpret the scale categories. The rater Y systematically used higher categories than did X. According to the RP value (0.37) it was 37 percentage units more likely that the objects were scored higher by Y than by X rather than the opposite. This means, that the inter-rater reliability could be substantially improved by informing the raters about this bias and/or by training the raters. The individual variability was negligible. In *intra-rater reliability* studies paired data are obtained by test-retest assessments by the same rater, and systematic test-retest disagreement could occur in case of changes in conditions between the test-retest occasions. A high level of occasional variations, evident by non-zero RV is sign of low quality of scale and is hard to repair (Svensson & Holm, 1994).

There is a widespread *misuse of correlation* coefficient as a reliability measure. The correlation coefficient measures the *degree of association* between two variables and does not measure the level of agreement. In Figure 1 the PA is 25%, and the observed disagreement is mainly explained by a high level of systematic disagreement in position. The Spearman rank-order correlation coefficient of the pairs of data in Figure 1, r_s is 0.73, indicates a rather high association between the two assessments of the same objects, but this is not an evidence of reliability.

APPLICATION TO ANALYSIS OF CHANGE

Evaluation of change is often based on assessments made on two occasions, for example before and after treatment. The classical methods for analysis of change in ordinal data are the same as for dichotomous data, the sign test, which means that valuable information get lost when the individuals are categorized in two groups, those with higher and those with lower categorical levels on the second occasion than on the first.

A complete *agreement* in the assessments made on the two occasions imply *unchanged* outcome. Referring to the paired distribution of assessments, Figure 1, the two occasions are represented by X and Y, respectively. Ten of the 40 individuals have unchanged levels, and the observed *disagreement* is a sign of *change* in outcome between the occasions. A systematic disagreement is a sign *of a change in common for the group*, while occasional variability indicate *individual variation in changes,* not explained by the common group change. In the worked example of Figure 1, the significant RP value of 0.37 is a sign of a homogeneous group change which reflects the efficacy of a common treatment plan for the group. The negligible individual variability confirms the homogeneity of the group changes. In cases of high RV, the individual changes are sign of individual heterogeneity in changes, which in clinical studies indicates that individual interventions or treatments must be considered, (Svensson, 1998c).

DISCUSSION

The interactive inter-disciplinary courses not only improved the participants' learning and understanding, but also implemented good statistical practice, such as statistical methods for ordinal data and for non-normal quantitative data, and a side effect was that these methods were spread to research groups as a whole. Another important effect was the marketing of statistical methods by the researchers' publications in various journals and by the doctoral dissertations based on good statistical practice. The participating students of statistics became prepared to meet the complexity of real life problems, and some of them were offered positions as consultants or biostatisticians. The research questions presented in courses sometimes needed alternative solutions and further statistical methodological developments, which was the case in my research. My stat methods for ordinal data is an ongoing process especially in order to make the methods known and easily applied, free software and guidelines is now available (see www.oru.se/hh/elisabeth_svensson).

REFERENCES

Hand, D. J. (1996). Statistics and the theory of measurement. *J R Statistical Society A,* 159, 445-492.

Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent development. *TEST, 14*, 1-73.

Stevens, S. S. (1955). On the averaging of data. *Science, 121*, 113-116.

Svensson, E. (1993). *Analysis of systematic and random differences between paired ordinal categorical data* (dissertation). Göteborg: Göteborg University.

Svensson E., & Holm, S. (1994). Separation of systematic and random differences in ordinal rating scales. *Statistics in Medicine, 13*, 2437-2453.

Svensson, E. (1998a). Teaching biostatistics to clinical research groups. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statist*ics (pp. 289-294). Singapore: International Statistical Institute.

Svensson E. (1998b). Teaching the measurement process in biostatistics. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 1257- 1262). Singapore: International Statistical Institute.

Svensson, E. (1998c). Ordinal invariant measures for individual and group changes in ordered categorical data. *Statistics in Medicine*, *17*(24), 2923-2936.

Svensson, E. (2001a). Important considerations for optimal communication between statisticians and medical researchers in consulting, teaching and collaborative research–with a focus on the analysis of ordered categorical data. In C. Batanero (Ed.), Training researchers in the use of statistics (pp. 23-35). Granada: International Association for Statistical Education.

Svensson, E. (2001b). Guidelines to statistical evaluation of data from ratings scales and questionnaires. *Journal of Rehabilitation Medicine, 33*, 47-48.

Svensson, E. (2002). Teaching statisticians and applied researchers statistical methods for analysis of data from rating scales. Experiences from joint research courses in rating scale data analysis. In B. Phillips (Ed.), *Developing a statistically literate society. CD of the Proceedings of the Sixth International Conference on Teaching Statistics*, 7 - 12 July, 2002, Cape Town, South Africa. International Association for Statistical Education, International Statistical Institute.

Svensson, E. (2009). Experiencing the complexity of reality before graduation. *Proceedings of the IASE Satellite conference, Durban.*