Some properties of measures of disagreement and disorder in paired ordinal data

To my beloved wife
Monica

Örebro Studies in Statistics 4

HANS HÖGBERG

Some properties of measures of disagreement and disorder in paired ordinal data

Title: Some properties of measures of disagreement and disorder in paired ordinal data.

Publisher: Örebro University 2010
www.publications.oru.se
trycksaker@oru.se

Print: Intellecta Infolog, Kållered 11/2010

ISSN 1651-8608
ISBN 978-91-7668-769-7

Abstract

Hans Högberg (2010): Some properties of measures of disagreement and disorder in paired ordinal data. Örebro Studies in Statistics 4, 38 pp.

The measures studied in this thesis were a measure of disorder, D, and a measure of the individual part of the disagreement, the measure of relative rank variance, RV, proposed by Svensson in 1993. The measure of disorder is a useful measure of order consistency in paired assessments of scales with a different number of possible values. The measure of relative rank variance is a useful measure in evaluating reliability and for evaluating change in qualitative outcome variables.

In Paper I an overview of methods used in the analysis of dependent ordinal data and a comparison of the methods regarding the assumptions, specifications, applicability, and implications for use were made. In Paper II an application, and a comparison of the results of some standard models, tests, and measures to two different research problems were made. The sampling distribution of the measure of disorder was studied both analytically and by a simulation experiment in Paper III. The asymptotic normal distribution was shown by the theory of U-statistics and the simulation experiments for finite sample sizes and various amount of disorder showed that the sampling distribution was approximately normal for sample sizes of about 40 to 60 for moderate sizes of D and for smaller sample sizes for substantial sizes of D. The sampling distribution of the relative rank variance was studied in a simulation experiment in Paper IV. The simulation experiment showed that the sampling distribution was approximately normal for sample sizes of 60-100 for moderate size of RV, and for smaller sample sizes for substantial size of RV. In Paper V a procedure for inference regarding relative rank variances from two or more samples was proposed. Pair-wise comparison by jackknife technique for variance estimation and the use of normal distribution as approximation in inference for parameters in independent samples based on the results in Paper IV were demonstrated. Moreover, an application of Kruskal-Wallis test for independent samples and Friedman's test for dependent samples were conducted.

Keywords: agreement, augmented ranks, categorical data, disorder, jackknife, paired ordinal data, rating scales, sample size, sampling distribution, simulation, U-statistics, variance

Hans Högberg, Handelshögskolan
Örebro University, SE-701 82 Örebro, Sweden

List of papers

This thesis consists of an introductory part and the following five papers:

I. Högberg, H. and Svensson, E. An overview of methods in the analysis of dependent ordered categorical data: assumptions and implications. Working Papers, Swedish Business School, Örebro University, No. 2008:7
II. Högberg, H. and Svensson, E. Comparison of methods in the analysis of dependent ordered categorical data. Working Papers, Swedish Business School, Örebro Univerity, No. 2008:6
III. Högberg, H. Statistical properties of a nonparametric measure of discordance in paired ordinal data. Manuscript.
IV. Högberg, H. Rank-based methods for analysis of individual variations in paired ordinal data. Manuscript.
V. Högberg, H. Statistical aspects on multiple comparisons of relative rank variance in paired ordinal data. Manuscript.

Contents

1 Introduction

Rating scales are commonly used in clinical research for assessing qualitative outcomes [1-6]. A characteristic of rating scales are that they produce ordinal data [7-10]. According to Stevens [11] ordinal data have an ordered structure but lack information about size and distance. The ordered categories may be assigned numbers, letters or other labels. Ordered labels assigned to the different categories of the variable should not alter the results of statistical analysis [12-14]. This rank-invariant property should be reflected in the statistical methods used in the analysis. Rank-based statistical methods are thus appropriate to use [10, 13, 15-20].

When studying the quality of data from rating scales, agreeable data from repeated observations are desirable. Thus, agreement is an important concept to evaluate and paired ordinal data typically arises. Especially when analyzing paired ordinal data the rank-invariant property has an important consequence since it is not appropriate to taking differences of the paired observations [10, 12-14].

Standard statistical methods often address a specific type of disagreement. There are several tests of marginal homogeneity and marginal models try to model the relations between the repeatedly observed variables, given assumptions of the variations within the observed pairs. On the other hand, specific tests and measures are constructed to measure the extent of agreement in the observed pairs. Some of the methods are parametric methods and dichotomize the scale categories while other measures of agreement are conditional on marginal homogeneity.

Disagreement may occur from unclear definition of the categories in a rating scale, that the rating task is not clearly stated, or that the cut-off points between adjacent categories are not clearly agreed upon [13, 20]. These exemplify reasons for systematic disagreement, and refer to a property of the rating scale or the rating situation. Systematic disagreement may be corrected once the reasons have been identified. Disagreement may also occur from occasional, haphazard events on individual basis in the rating situation of the subjects. Such disagreement is harder to correct. Thus, it is important to quantify and separate the systematic and individual disagreement. The reasons for systematic disagreement may require one kind of action while the remaining individual disagreement may require another kind of action [14, 21-23].

Svensson [13, 20] has developed an approach to simultaneously assess both systematic and individual disagreement based on an augmented ranking method. The approach and its measures are free from any distributional restriction on the data.

This thesis deals with two measures of individual disagreement; the measure of disorder and the measure of relative rank variance. Both measures have been used in many different studies regarding development of questionnaires, evaluations of validity, reliability and change in various application disciplines.

What are the main differences between these measures and other commonly used measures and methods in empirical studies? What are the characteristic properties and assumptions of the measures, tests and models? What are their implications for applicability? To what questions do they apply? What are the results from the different methods and do they agree? These questions comprise the starting point for this thesis and determined the aims of the first two papers.

Inferences about the corresponding population parameters have been based on estimates of variances either by an empirical counterpart to the theoretical variance, by jackknife estimates or bootstrap technique [13, 20]. In order to further study valid inferences, the distributional properties were investigated in the third and fourth papers in this thesis. It was also considered important to develop methods for testing the difference of intra-rater agreement in different items in a multi-item questionnaire. This was done in the fifth paper.

2 Aims of the thesis

The overall aim of the present thesis was to further investigate the properties of the two measures of individual variability, developed by Svensson, such as the possibility of asymptotic normality, and to suggest approaches for interval estimation and tests. Moreover, an aim was to show the importance of considering the properties of dependent ordered categorical data in choosing methods for statistical analysis. This overall aim was formulated as the following five specific aims and presented in separate papers:

I. To overview and discuss the relative merits of standard methods and measures for analysis of dependent ordered categorical data and the measures in Svensson's approach. The focus was at the assumptions of the models and data, the usefulness of the methods for descriptions and inferences, and their implications for use.

II. To apply various measures, models, and tests, including Svensson's non-parametric approach to two different data sets of paired ordinal data and to compare and interpret the results and show their conditions for use.

III. To derive the asymptotic distributional properties of the empirical measures of disorder and monotonic agreement. Another aim was to investigate the distributional properties of these measures for sample sizes encountered in practice and to study how well the variance estimators compare to the sampling variance. A final aim was to apply the measures and two classical measures of concordance to an empirical data set.

IV. To study the distributional properties of the empirical measure of the relative rank variance RV for sample sizes met in practice. Another aim was to illustrate the methods for inference regarding the variance of the relative rank difference in an empirical data set.

V. To discuss and develop statistical methods for inference when comparing the individual disagreement measured by the measure of the relative rank variance RV between different items in a multi-item questionnaire. Another aim was to illustrate the methods in an empirical data set.

3 A theoretical background

The Svensson approach to analysis of paired ordinal data makes it possible to evaluate the systematic part of an observed disagreement in paired assessments on a rating scale separately from the individual variability of assessments [13, 20]. The observed agreement/disagreement pattern is described by the distribution of the pairs of data in a contingency table when the assessments are made on a rating scale having a discrete number of ordered categories or by a scatter plot when assessments are made on a visual analogue scale with 101 possible values. Figure 1 shows a contingency table with the notation used by Svensson and in the present thesis [13].

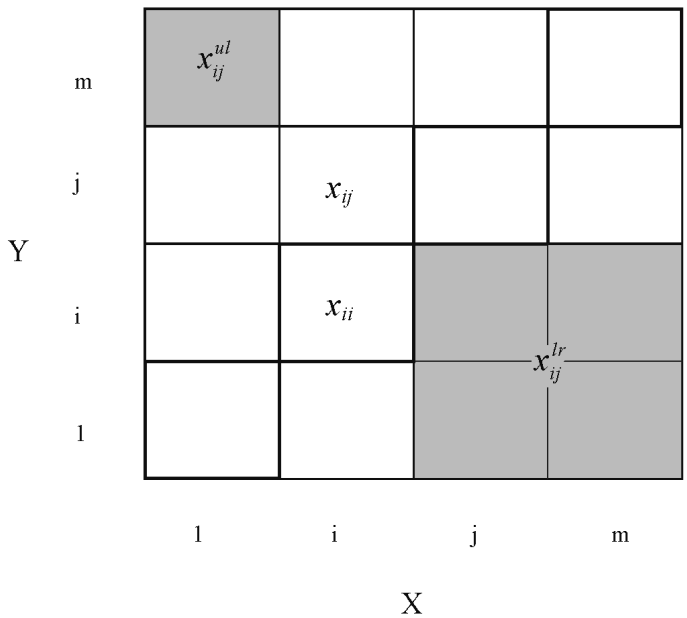

Figure 1. Schematic illustration of the basic notations in a contingency table with m categories used in formulas [13].

When the two sets of frequency distributions of assessments - also called marginal distributions - differ, systematic disagreement is present. Svensson [13, 20] defined two measures of systematic disagreement. The measure of Relative Position, is a measure of a systematic shift in the use of the scale categories between the two assessments, which means a case when one frequency distribution is stochastically larger than the other

$$\gamma = P(X < Y) - P(Y < X) \tag{1}$$

Another reason for a systematic disagreement could be a systematic disagreement in how the assessments are concentrated on the scale categories between the two assessments. The measure is called the Relative Concentration,

$$\delta = P(X_l < Y_k < X_m) - P(Y_l < X_k < Y_m) \tag{2}$$

These measures are thoroughly described in [13, 20, 24] and are not treated further in this thesis.

Svensson has shown that it is always possible to construct one unique distribution of pairs of data to each set of marginal distributions, which is a rank-transformable pattern of agreement [13, 20]. This pattern illustrates the distribution of paired data that is expected when all disagreement is explained by systematic disagreement only, given the observed marginal distributions.

In the rank-transformable pattern of agreement, each pair of assessments will have the same rank ordering when ranking the assessments X, and the assessments Y, respectively. Svensson [13, 14, 20] proposed an augmented ranking approach by which the pairs of rank values given to the observations are tied to the pairs and not to each marginal distribution. Then the mutual relationship between the paired assessments on the individual by the two raters is utilized.

Empirical data sets commonly have individual variations in repeated assessments on scales. Then the observed distribution of pairs of data differs from the rank-transformable pattern of agreement, and so do the two set of ranks allocated to the pairs of data. Besides the measures of systematic disagreement Svensson has proposed two measures of such individual variability in an observed disagreement pattern; the *measure of disorder* and the *measure of relative rank variance* [13, 25, 26].

The augmented mean ranks for assessments X are calculated by:

$$\overline{R}_{ij}^{(X)} = \sum_{k=1}^{i-1}\sum_{l=1}^{m} x_{kl} + \sum_{l=1}^{j-1} x_{il} + \frac{1}{2}(1 + x_{ij}) \tag{3}$$

where x_{ij} is the ij:th cell frequency. The augmented mean ranks for Y are defined correspondingly as:

$$\overline{R}_{ij}^{(Y)} = \sum_{k=1}^{m}\sum_{l=1}^{j-1} x_{kl} + \sum_{k=1}^{i-1} x_{kj} + \frac{1}{2}(1 + x_{ij}) \tag{4}$$

Differences in augmented mean ranks indicate dispersed observations from the rank transformable pattern of agreement and define the empirical measure of individual variability in disagreement, the relative rank variance (*RV*) [13, 20]. This is a normed estimate of the parameter of the variance of the relative rank differences. The relative rank variance, *RV*, is defined as

$$RV = \frac{6}{n^3}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}(\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)})^2 = \frac{6}{n^3}\sum_{\upsilon=1}^{n}(R_{\upsilon}^{(X)} - R_{\upsilon}^{(Y)})^2 \qquad (5)$$

where υ is the *υ:th* of n subjects and which estimates the parameter

$$6\sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij}(q_{ij}^{ul} - q_{ij}^{lr})^2 = \psi \qquad (6)$$

where p_{ij} is the *ij*:th cell probability, q_{ij}^{ul} is the upper-left region probability, and q_{ij}^{lr} is the lower-left region probability. The relative rank variance is the recommended measure of individual variability in test-retest assessments on the same scale, which is the common design for evaluation of reliability of assessments and for evaluation of change in qualitative outcome variables [20, 27, 28].

In validity studies, the consistency in assessments on different scales of the same variable will be evaluated. The two comparing scales can have different number of possible values, and then the measure of disorder is a useful measure of order consistency in the paired assessments.

The measure of disorder, proposed by Svensson [25, 26], defines the level of discordance relative to total agreement in ordering irrespective of the scale levels and marginal distributions. This is in contrast to traditional measures like Kendall's tau-b [29, 30], Stuart's tau-c [31] and Goodman-Kruskal's gamma [32]. Svensson [25] and others, e.g. [32, 33] have demonstrated how the different approaches adjust for ties in different ways and how the measures depending on scaling and marginal distributions limit the possibility to attain the limit of unity.

The parameter is defined

$$\Pi_D = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} p_{ij}(q_{ij}^{ul} + q_{ij}^{lr}) \qquad (7)$$

Svensson [13] showed the expression for the variance. The parameter Π_D is the parameter of reversed order classification and equals the parameter of disordered observation of which the measure *D* is an estimator, except for the correction

factor for tied observation in the denominator. The empirical measure of the parameter Π_D can be written as

$$T = \frac{\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} x_{ij}(x_{ij}^{ul} + x_{ij}^{lr})}{n(n-1)} \tag{8}$$

and the variance

$$V(T) = \frac{2}{n(n-1)}\Pi_D(1-\Pi_D) + 4\frac{1}{n}\frac{n-2}{n-1}(\Pi_{DD} - \Pi_D^2) \tag{9}$$

where Π_{DD} is the probability that out of three pairs, the second and the third pairs are disordered to the first pair. This variance is then estimated by substituting the empirical relative frequencies for the unknown probabilities. In particular

$$\hat{\Pi}_{DD} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \hat{p}_{ij}(\hat{q}_{ij}^{ul} + \hat{q}_{ij}^{lr})^2 \tag{10}$$

provided the existence of two such pairs disordered the pair in the *ij*:th cell.

The variance may also be estimated by jackknife or bootstrap techniques. The explicit asymptotic variance for the empirical measure of disorder, D, has not been shown yet and the sampling distributions of the empirical measures for sample sizes met in practice have not been demonstrated.

The variance of *RV* is a complicated expression and even an asymptotic approximation is cumbersome to utilize for an empirical variance estimator [13]. Jackknife or bootstrap techniques may be used for variance estimation and inferences. The sampling distribution for *RV* is not known so the estimates of the variance of *RV* cannot be used directly for inference.

Bootstrap tests and confidence intervals are thus a possible strategy but to use the measures of disorder and relative rank variance thoroughly it is important to know more about the distributional properties.

4 Summary of the papers

4.1 Paper I: An overview of methods in the analysis of dependent ordered categorical data: Assumptions and implications

The aim of the first paper was to give an overview of the methods used for the analysis of dependent ordered categorical data and to compare standard methods with Svensson's measures. The exposition was focused on the assumptions, specifications, applicability and implications for the appropriate areas of application. The approach was to call attention to the different problems in analysing dependent ordered categorical data and put together and sum up the results. The overview gives a picture of standard methods as well as state-of-the-art methods and serves as an inventory of problems and a rationale to the development of Svensson's measures.

At first some fundamental asymmetric models for categorical data are described followed by a description of how these fundamental models were elaborated to dependent ordinal data. These fundamental models are basic multinomial generalized linear models (GLM), such as the cumulative logit model, the adjacent categories logit model, and the continuation ratio model, which are used in marginal models or conditional models. In marginal models focus is on population-average effects such as marginal homogeneity and in conditional models focus is on cluster-specific effects as well [34, 35]. The cluster was here a general concept that may describe subjects that are rated repeatedly over time or subjects rated by several raters. Special cases are paired ratings at two time points or paired rating by two raters.

A second major group of models that is described is log linear models. In contrast to asymmetric models which distinguish between response and explanatory variables log linear models are symmetric and are useful for modelling association [36]. In certain cases there are correspondences between logit models for dependent data and log linear models. The log linear models are described, giving special attention to those models developed for dependent ordinal data with applications to agreement studies. Many models in this class of log linear models are hierarchically structured and through an analysis of goodness-of-fit statistics from these hierarchical models conclusions about agreement patterns may be drawn [21, 37].

Further, summary measures for describing order consistency and agreement are described, such as Kendall's tau-b, Goodman-Kruskal's gamma, and Cohen's kappa. Although specific, easy to understand, and easily accessible, they are often used inadequately. For example, assessment of association is not generally

the same as assessment of agreement. It was important to point out the original objective of the measures and their shortcomings.

The overview concludes with a description of the augmented ranking approach and Svensson's measures. In parallel to the development of these models and measures, Svensson brought up the lack of methods that were rank-invariant, and were able to quantify different important aspects of agreement and disagreement and, at the same time, could be used for analysis of change in ordinal outcome variables [13, 20, 27].

In the discussion it was concluded that using models is often considered to be superior to tests and summary measures due to the models' elaborated and more facetted information, but models may also become more and more complicated to parameterize and interpret. The risk of misspecification and that the fundamental assumptions behind the models are violated are obvious in using such models in applied research. It is, on the other hand, difficult to capture the many aspects of change, association or agreement by one single measure. Many models use some scoring system and the models are then not invariant to any transformation of the scores or to merging or splitting categories. Some link functions in the models are not even palindromic invariant. Models and measures for ordered categorical data should be rank invariant.

Many of the existing measures may be regarded as attractive as they are easy to understand and to apply. But some of these are not adequately used, e.g. correlation for measuring agreement, and some have serious drawbacks, e.g. the coefficient of kappa. Svensson's approach is rank-invariant, non-parametric and uses the paired ordered information in the ranking procedure. By the complementary use of a few measures it is possible to evaluate both systematic and individual disagreement. The measures are equally apt to be used in designs for evaluation of change in response to some treatment as they are for assessment of reliability or validity. In contrast to other measures of concordance or agreement, the limits of their range for the measures of Svensson are attainable, irrespective of the number of possible response categories and the type of scaling and the category distributions.

4.2 Paper II: Comparison of methods in the analysis of dependent ordered categorical data

As a second step in the treatment of non-parametric methods for analysis of dependent ordered categorical data, a comparison of some standard measures, models and tests with Svensson's measure using two empirical data sets was made. The novel approach was to bring together results from applying a variety

of different measures, models, and tests on two very typical examples of research problems, and to compare those with the results from the measures of Svensson.

The empirical data sets represent two types of studies commonly encountered in clinical research. The first empirical data set was from a study concerning agreement in judging biopsy slides for carcinoma of the uterine cervix [38]. The purpose of that study was to investigate the variability in classification and the degree of agreement in ratings among pathologists. The original data has since been published, frequently served as an illustration in methodological papers [21, 39, 40]. The second empirical data set was from a study of individual and group changes in the patients' social outcome after aneurysmal subarachnoid haemorrhage between two occasions [28]. The intension of the study was to publish the results, but it was also used to illustrate some aspects of Svensson's measures. The measures, models and tests were determined by the different aims of original studies. These were various agreement and association measures, and models of agreement based on log linear models with parameters describing symmetry, quasi-symmetry and marginal homogeneity. Furthermore, tests of marginal homogeneity and symmetry were applied, and in the case of testing change between two time points the sign test was applied. Svensson's measures of systematic and individual disagreement could be applied to both study purposes [20]. The various standard measures, models and test were outlined but Svensson's measures were presented more thoroughly.

To study reliability, as in the first application example, it was not sufficient to use one of the standard agreement or association measures, such as Cohen's kappa, Goodman-Kruskal's gamma, or Kendall's tau-b. It was necessary to supplement the measures by one or more log linear models and models of marginal homogeneity. Most important though, was to use models and measures constructed for paired ordinal data and relevant to the question of reliability.

To study change in ordinal response variables certain tests are in common use. Such an example was the sign test in the second application. Change on group level may also be tested by marginal models as was demonstrated. Log linear models could be an option as they models cell frequencies. The independence model is often used. This model may be expanded depending on what patterns in the cell frequencies are relevant to study. But even if the choice of log linear models for analyses of change patterns is not a common option, they were used in paper two to contrast them with Svensson's measures. Svensson's measures are rank-invariant and utilize the fact that the data consists of pairs. Furthermore, the measures gave more comprehensive information about systematic and individual variations.

The study showed that the standard measures, models and tests gave diverging conclusions, which thus implies difficulties in the interpretation of such findings. Further, the study showed that the measure RC of Svensson revealed systematic

differences in concentration in the biopsy slides assessment study, which could not be detected by the traditional methods. This is an indication of that the assumption of stochastic ordering, crucial to many models, was not fulfilled. Moreover, the study showed that the measure RV of Svensson indicated individual occasional causes of change in the social outcome study, which could not be detected explicitly by the traditional methods. One conclusion of the study was that it was not as easy to detect aspects of the kind of systematic and individual reasons for variations by the various standard measures, models or tests as it was by the measures of Svensson. As an example, the model of agreement plus uniform association gave information of the kind of association and agreement, but no information about the systematic disagreement in concentration using the scale. A remaining question was whether the assumptions for the model was fulfilled. An implication of this was that the researcher has to be very aware of using the models for its intended purpose and check for their adequacy.

4.3 Paper III: Statistical properties of a nonparametric measure of discordance in paired ordinal data

One purpose of the study was to derive the large sample properties of the measure of disorder, D. Another purpose was to investigate the distributional properties of the measure and to compare variance estimators in sample sizes encountered in practice. The measure and two classical measures of concordance were applied to an empirical data set and compared.

Measures based on indicators of disordered and ordered observations are traditionally built up as the excess of concordant pairs over discordant pairs adjusted for the number of tied observations. The measures differ in the way they consider tied observations. The measure of disorder, *D*, was defined by Svensson [25, 26] as:

$$D = \frac{\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} x_{ij}\,(x_{ij}^{ul} + x_{ij}^{lr})}{n(n-1) - t} \qquad (11)$$

where the number of individuals classified to the *i:*th and *j*:th category respectively is denoted x_{ij} and x_{ij}^{ul} and x_{ij}^{lr} is the number of observations in the upper-left region and lower-right region relative the *ij*:th cell, respectively, and where *t* is the correction factor for tied observations

$$t = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} x_{ij}(x_{ij} - 1) \tag{12}$$

This way of defining tied observations imply that pairs of observations may be either discordant, concordant or else tied if the pairs of observations are identical, which means that the number of concordant pairs is differently defined than in classical measures. When there is total agreement in ordering, no pairs of observations are in the upper-left or lower-right regions relative to the cells, which mean that $x_{ij}^{ul} = x_{ij}^{lr} = 0$ and D=0. The maximum value of D=1 indicates total inconsistency in ordering. A measure of monotonic agreement (MA) was also defined as [25, 26]

$$MA = 1 - 2D \tag{13}$$

The asymptotic distribution of D is not known. The large sample properties were shown by the theory of U-statistics [41] and the finite sample properties was investigated by a simulation experiment. The theory of U-statistics was well suited for non-parametric theoretical study of large sample distributional properties provided the existence of the second moment of the kernel function of the U-statistic. Application of a theorem in Hoeffding [41] regarding a function of U-statistics leads to the general formula for the variance in the limiting normal distribution for a ratio of two U-statistics

$$AsVar\left[\frac{U}{W}\right] = \sum_{\gamma=1}^{2}\sum_{\delta=1}^{2} m(\gamma)m(\delta)\left(\frac{\partial g(y)}{\partial y^{(\gamma)}}\right)_{y=\theta}\left(\frac{\partial g(y)}{\partial y^{(\delta)}}\right)_{y=\theta}\zeta_1^{(\gamma,\delta)} \tag{14}$$

where m is the degree of the U-statistic and

$$\zeta_1^{(\gamma,\delta)} = E\left\{\Psi_1^{(\gamma)}(X_1)\Psi_1^{(\delta)}(X_1)\right\} \tag{15}$$

is the first order variance and covariance term in the decomposition of the kernel functions in conditional expectations of the U-statistics.

The simulation experiment was designed to study how fast the theoretical results of asymptotical normality works in practice and what the empirical sampling distributions looked like in sample sizes encountered in practice. Furthermore, the asymptotic variance and the approximate variance estimator derived in Svensson [13] were to be compared with the empirical variance in the simulation experiment. The sample sizes varied from 20 to 1000. Various appearance of contingency tables were chosen to serve as populations, based on various amount

of disorder, various sizes of the tables, various probabilities of tied observations, and various appearance of observation along the agreement diagonal. Tests of normality were performed and normal Q-Q plots were used for complementary evaluation of normality. The measures and the asymptotic results were then applied to a data set from an investigation of order consistency. A comparison with Goodman-Kruskal's gamma [32] and with Kendall's tau-b [29, 30] was also made.

The measures D and MA were shown to be asymptotic normally distributed with a specified variance by the theory of U-statistics. The variance for D in the asymptotic normal distribution was

$$AsVar[D] = AsVar\left[\frac{U}{W}\right] = \sum_{\gamma=1}^{2}\sum_{\delta=1}^{2} m(\gamma)m(\delta)\left(\frac{\partial g(y)}{\partial y^{(\gamma)}}\right)_{y=\theta}\left(\frac{\partial g(y)}{\partial y^{(\delta)}}\right)_{y=\theta} \zeta_1^{(\gamma,\delta)} =$$

$$= \frac{4}{n(1-\Pi_t^2)^2}\left[(\Pi_{DD} - \Pi_D^2) + \frac{2\Pi_D}{1-\Pi_t^2}(\Pi_{Dt} - \Pi_D\Pi_t^2) + \frac{\Pi_D^2}{(1-\Pi_t^2)^2}\Pi_t^2(1-\Pi_t^2)\right] \tag{16}$$

where Π_{Dt} is the probability that out of three pairs of observations two of them are tied and the third is disordered and Π_t^2 is the probability that two pairs of observations are tied. The analogy of this result to the asymptotic results for Kendall's tau was pointed out. For various experimental settings the sampling distribution of the measure D approached the normal distribution at different rates of convergence. Very small values of D required large sample sizes, at least 100, to be able to use the normal distribution. For moderately large D (near 0.1) the sampling distribution had an approximate normal distribution for sample sizes of about 60, and for large D (above 0.15) the sampling distribution was close to normal for sample sizes of about 20 to 40. Besides the values of D, the appearances of some of the contingency tables in the experiment afflicted the sampling distribution. Tables with a high probability of ties and a high probability of observations in cells that influenced the value of D resulted in excessively skewed and jagged sampling distributions. The square root of the asymptotic variance was close to the sampling standard error in most cases. In summary, the normal distribution could be used as an approximate sampling distribution for tables showing moderate to large disorder and in large sample sizes even for small amounts of disorder. The application illustrated the performance and comparative advantage relative the two classical measures as well as results of confidence intervals using the asymptotic results.

4.4 Paper IV: Rank-based methods for analysis of individual variations in paired ordinal data

One aim in the fourth paper was to study the sampling distribution of the measure RV for possible use in the construction of approximate test statistics and confidence intervals as an alternative to bootstrap methods. Another aim was to illustrate the findings by empirical data.

The sampling distribution of RV was studied by means of a simulation experiment. The simulation experiment was designed to study how the empirical sampling distributions looked like in sample sizes encountered in practice for various amount of RV. The sample sizes varied from 20 to 1000. Various data sets were chosen to serve as populations, based on various amounts of individual disagreement, different numbers of categories, various probabilities of tied observations, and various appearance of observation along the agreement diagonal. The Shapiro-Francia W' test, tests of skewness and kurtosis, and a joint skewness and kurtosis test were used for test of normality [42]. Normal Q-Q plots were used for complementary evaluation of normality. These plots are useful for evaluating discrepancies from the normal distribution in the tail regions [43]. The jackknife technique for variance estimation in inference regarding the parameter estimated by RV was suggested. The findings were applied to data from a study of test-retest stability by intra-individual agreement in a multi-item questionnaire [2]. The items in the questionnaire were tested for evidences of lack of stability and for items prone to be influenced by individual disturbing factors.

In some experimental setting the sampling distributions were skewed even for a large sample size. Sample sizes had to be in the order of at least 100 when RV was small, i.e. smaller than 0.1. Very small values of RV are desirable in reliability studies. But in important situations when there were substantial amount of individual disagreement the sampling distributions converged to an approximately normal distribution for moderate sample sizes. For RV values above 0.12 a sample size in the range 60 to 100 was generally sufficient. The approximation to the normal distribution seemed primarily to depend on the size of RV, but also on how those observations that contributed to the RV were located in the contingency table. In the application, the test based on the normal approximation resulted in rejection of the null hypothesis, and the confidence interval illustrated the uncertainty of the point estimate. Despite the small sample size, the amount of individual variation still made it possible to use the normal approximation in the inference.

4.5 Paper V: Statistical aspects on multiple comparisons of relative rank variance in paired ordinal data

The aim was to discuss and develop statistical methods for inference when comparing the individual change measured by RV between two or more different items in a multi-item questionnaire or between two or more independent groups. The methods were illustrated by an empirical data set.

The methods discussed were divided in methods for independent samples and methods for dependent samples. In the first method for independent samples the simulation results in paper four were utilized to use the normal distribution in inference. The jackknife estimates of variances were used in the pooling of variances. Pooling the jackknife estimates of variances for the different items served as an estmate of the common variance under the null hypothesis, which then was used as estimate of the sampling variance of the difference in the test statistic for independent samples. The estimator of the variance of RV equals the jackknife estimator of the variance except for a constant close to one for large sample sizes

$$\hat{V}ar(RV) = \left(\frac{n-1}{n}\right)^4 \hat{\sigma}^2_{jack}(RV) \tag{17}$$

and the jackknife estimator is convenient to use in practice [13]. In calculation of the variance for the difference of two RV the jackknife estimates of each RV is pooled by

$$\hat{V}ar_{pooled}(RV) = \frac{\left(\frac{n_1}{n_1-1}\right)^5 \hat{V}ar(RV_1) + \left(\frac{n_2}{n_2-1}\right)^5 \hat{V}ar(RV_2)}{n_1+n_2-2} \tag{18}$$

and then

$$\hat{V}ar_p(RV_1 - RV_2) = \hat{V}ar_{pooled}(RV)\left[\frac{(n_1-1)^6}{n_1^5} + \frac{(n_2-1)^6}{n_2^5}\right] \tag{19}$$

This variance estimator is then used in the test statistic

$$z = \frac{RV_1 - RV_2}{\sqrt{\hat{V}ar_p(RV_1 - RV_2)}} \tag{20}$$

and critical values are determined from the standard normal distribution.

The second method for independent samples expanded the application of the distribution-free Kruskal-Wallis one-way ANOVA test [44]. The ranks used in the test statistic were the ranks of the squares of the augmented mean rank differences in the construction of the measure RV. The test statistic, corrected for ties, is defined as

$$H^* = \frac{\frac{12}{N(N+1)}\sum_{k=1}^{s}\frac{R_k^2}{n_k} - 3(N+1)}{1 - \frac{\sum_{t=1}^{g}(d_t^3 - d_t)}{N^3 - N}} \tag{21}$$

where R_k is the sum of ranks in the *k:*th sample and the ranks are the ordered values of $(R_\upsilon^{(X)} - R_\upsilon^{(Y)})^2$ for each sample. A follow-up test procedure, originating in Dunn [45], which controls the overall significance level and considers the dependencies between the mean rank differences, was also used. This procedure amounts to calculate

$$Z_{kl}^* = \frac{\left|\overline{R}_k - \overline{R}_l\right|}{\sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{t=1}^{g}(d_t^3 - d_t)}{12(N-1)}\right]\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}} \tag{22}$$

for $1 \leq k < l \leq s$ and compare to critical values

$$z' = z_{\alpha/[s(s-1)]} \tag{23}$$

The overall significance level α is recommended to be set to about α=0.2 [45, 46].

For dependent samples, the Friedman two way analysis of variance by ranks [47, 48], was applied to the ranks of the squares of the augmented mean rank differences. The Friedman test statistic, corrected for ties, is

$$Q = \frac{12(s-1)\sum_{k=1}^{s}\left(R_k - \frac{n(s+1)}{2}\right)^2}{ns(s^2-1) - \sum_{t=1}^{g}(d_t^3 - d_t)} \tag{24}$$

The ranks in the rank sum R_k for sample k was, like in the test statistic H^* based on the values $(R_\upsilon^{(X)} - R_\upsilon^{(Y)})^2$. Both H^* and Q are asymptotically χ^2_{s-1}.

Two follow-up test procedures were applied; one based on the Dunn's procedure for the Kruskal-Wallis test and the other was pair-wise application of the Friedman test. The follow-up test statistic

$$Z_{kl} = \frac{|R_k - R_l|}{\sqrt{\frac{ns(s+1)}{6}}} > z' \tag{25}$$

is compared to the critical value z'.

The different inferential procedures were applied on a data set of patients' assessment of four different items in the dimension Vitality in the questionnaire Short-Form-36 (SF-36) [6]. The items were regarded as dependent so the correct test to use was Friedman's test. The data was nevertheless used for independent samples with the purpose of working through the test procedures and formulas for illustration.

The RV values of the four items were all high. The Friedman test could not reject the hypothesis of equal individual change for the four items. The follow-up tests were not necessary, but were carried out for illustration. The follow-up test procedures showed some inconsistencies. The Kruskal-Wallis test also resulted in a non-significant result. In the pairwise testing, the large sample normal theory based test and Wilcoxon-Mann-Whitney test were used. The results of the pairwise testing gave the same conclusion but showed varied p-values. The sample sizes were between 94 and 98, so the large sample normal approximation would be appropriate in this case.

For small sample sizes in comparisons of two or more RV values, it is possible to use the Friedman test and the Kruskal-Wallis test and, if significant, the follow-up procedure by Dunn. If sample sizes are large, Friedman's test and Kruskal-Wallis test could still be used, but the follow-up tests could be performed by the z-test in the case of independent samples.

5 Discussion and conclusion

The goal of this thesis was to further develop the two measures of individual variability, developed by Svensson, to increase their applicability in research. To justify the need of developing the measures, an overview of current models, tests, and measures in analysis of dependent ordinal data was considered important. A review was made with respect to assumptions, specifications, applicability, and implications. Several of the models, tests, and measures are sometimes used mechanically for various types of research questions and results in varying conclusions. Through the review and comparison of methods, Svensson's measures were placed in a methodological context and some advantages could be highlighted.

The choice of the measures of disorder (D) and individual disagreement (RV) was based on the presumption that study of individual variations is an important aspect of total variation, and that there is a scarcity of competitive alternatives. An important domain of potential development for the measures to consider was methods for inference. The methods used so far are bootstrap for interval estimation and jackknife technique for variance estimation. Inference based on normal approximation for the measures D and RV has not been studied thoroughly up to now; the present thesis constitutes a beginning of such development. The measure D is a somewhat simpler measure in construction and interpretation than RV. For evaluating order consistency and interchangeability, and comparability between scales, the measure D is appropriate. For evaluating reliability in terms of agreement and also for evaluating change, the measure RV is appropriate.

Within many research areas there is a great need of comparing the size of disorder or disagreement of different items or between several points of time and make some inferential conclusions. With the results of the study of the statistical properties of single D and RV, the development of inferential methods for comparison of several parameters has commenced.

The overview in Paper I was performed by reviewing the statistical, scientific literature. No general search engines or databases covering most of the statistical literature were available so the search for relevant literature started with some selected articles and the references included therein. Specific databases, e.g. MathSciNet and JSTOR or publishing companies were also searched. Search engines and databases are in rapid progress so systematic review has become easier to perform.

The review of the methods in this thesis had its departure from a broad and general perspective, arriving at more specific methods for specific purposes. Asymmetric models are used for modelling effects of, for example, treatments over time and covariates.

The overview and the comparison of methods in the first two papers showed that standard models sometimes have complicated specifications and strong assumptions which are often not fulfilled when analysing paired ordinal data. Users must have comprehensive and thorough knowledge of the models to evaluate them and choose the appropriate ones for their specific research question. The results of the comparison showed that different results may occur amongst the various methods that are commonly used for the same objective. Some of the standard measures have drawbacks or limitations and are used for research questions which they are not constructed for. The choice of measures and models in the papers fell upon measures and models commonly used in analysis of paired ordinal data. One purpose was to show the results by using them in an analysis and show that the results could diverge if the researcher is not careful in choosing the adequate model or measure for the problem at hand. An example of this is using a standard marginal homogeneity test (Stuart's test) for paired ordinal data, even though this test does not consider ordinality. Goodman-Kruskal's gamma and Kendall's tau are measures of association. Certain log linear models are used for analysis of agreement and some of them may be used for tests of marginal homogeneity. Furthermore, the marginal model approach by generalized estimating equation (GEE) and the random effects cluster-specific model approach may be used for analysis of marginal homogeneity. In general, marginal models and conditional models may be used in studies of change with different foci on the effects. The marginal models focus on population effects and the conditional models focus on cluster effects.

One of the most fundamental issues is the question of assigning scores in models for ordinal data. In most models scores are assigned to the ordinal categories. There are several log linear models adequate for modelling aspects of reliability in squared tables such as agreement, association and category distinguishability [21, 22, 37, 49]. A critically assumption then concerns the assignment of scores. Most often equidistant scores are used and of these, integer scores are most common. The most general of log linear models permit arbitrary scores but they must be fixed. Scores as parameters lead to non log linear models. Cumulative logit models do not require scores for the dependent variable but for the independent variable. Agresti in [50] p. 138 says: "An obvious disadvantage of the ordinal loglinear models… is the necessity of assigning scores to the categories of ordinal variables. For many variables no obvious choice of scores exists. Yet parameter estimates and the goodness of fit of the models depends of that choice." In Svensson's measures no scores are used. In the measure of disagreement and the coefficient of monotonic agreement indices of discordance and concordance are used and in the measure RV the number of discordant pairs expresses a non-metric distance from agreement in ordering, not a metric function of distances of scores.

Asymptotic properties of estimators may be studied theoretically or by simulations. Sometimes the asymptotic properties are known theoretically, but the properties in finite samples do not necessary behave as if in infinite sequences. In other cases the asymptotic properties are not known in theory. In both situations simulation may be useful. In Paper III the asymptotic properties of the measure of disagreement D was derived and some of the behaviour of the sampling distribution for finite samples was studied by simulations. In the Paper IV the asymptotic properties of the measure of relative rank variance RV were not studied theoretically, but the sampling distribution was studied by means of simulations. The greater complexity of the expression of RV made it harder to study theoretically. The simulations generated ideas about the asymptotic properties as well as illustrations of the behaviour of the sampling distribution in finite samples. In both papers representative and valid examples of cases of data were chosen. The sampling distribution was tested for normality and visually evaluated regarding similarity to the normal distribution. By the use of normal Q-Q plots, the deviances of the sampling distribution from the normal distribution could be indicated. Furthermore, through the study of the sampling distributions for different types of data sets, the dependencies of the observations contributing to the measures D and RV could be studied. Since the measures studied are non-parametric it was considered adequate to analyse the asymptotic properties by non-parametric based theory such as the theory of U-statistics. The theory of U-statistics offers a convenient way to show asymptotic normality without any assumptions except that of the existence of the second moment of the kernel function of the U-statistic [41, 51, 52]. The theory also applies for discrete populations [41]. The theory of U-statistics has been used in derivation of asymptotic results for measures of concordance and association [53].

The result in Paper IV was used to continue the development of the applicability of the measures to inference for individual disagreement in two or more samples. Employing the jackknife technique for variance estimation and pooling the estimates according to the assumed equality of the parameters and using the simulation results, makes it possible to use the normal approximation in inference for the difference of two parameter values in independent samples. For small sample sizes, application of the non-parametric Kruskal-Wallis test and Friedman's test for independent and dependent samples, respectively, was discussed. The novel application of the rank-based test statistics were to use the squares of the augmented rank differences in the formula for RV as observed values and then rank these observations.

The choice of measures to develop was primary based on the urgent need to use these measures in inferential applied research. Furthermore, a modified version of the measure of relative position, RP, has been studied in Wahlström [24]

and both RP and the measure of relative concentration RC have been studied in Yang [54].

Another approach to study the asymptotic properties of D and RV would have been to exploit the fact that the measures are functions of relative frequencies and the corresponding parameters are functions of probabilities. By assuming multinomial distribution and using the multivariate delta method, asymptotic normality and variance in the asymptotic distribution could have been shown. The theory of U-statistics makes it possible to take the study of the asymptotic properties in a unified way even further. The only assumption needed to be verified for the theory of U-statistic to apply is the existence of the second moment of the kernel. For asymptotic normality one has to check for the existence of the first order variance for the U-statistic. Otherwise, it is possible to verify asymptotic chi-square distribution [41]. Furthermore, the rate of convergence and results of the behaviour of the sequences to the limit may be studied within the framework of the theory of U-statistics [52, 55]. This theoretical work has not been done in the present thesis. The simulation experiments indicate skewed sampling distributions for small to moderate samples when the target value of RV is close to zero, so it would be of interest to theoretically examine the asymptotic properties further.

In the study of asymptotic properties and sampling distributions no restrictions have been made to hypothesis tests and the assumptions under the null hypothesis. Under such assumptions, evaluating test statistics in finite samples by permutation methods could have been an option. The purpose of the simulation experiments could have been to simulate the behaviour of some test statistic under the null hypothesis, focusing on the actual type I error rates for different nominal type I errors, different sample sizes, and different amount of disorder or disagreement etc. In Wahlström [24] such a type of simulation experiment was conducted where tests of equivalence of actual type I error rates to nominal type I error rates were performed. The properties of the confidence interval could also have been studied by simulation experiments. This has not been done in the present thesis.

Comparing the results from papers three and four, one conclusion may be that the normal approximation works better for the measure D than it does for the measure RV. The normal distribution may be used for tests and confidence intervals for D in smaller sample sizes than it could for RV. Moreover, for the measure D an asymptotic variance was derived, whereas for the measure RV an approximation of the theoretical variance or a jackknife approximation may be used in practice.

By means of the theoretical study and the simulations of D some conclusions in accordance with those relating to the gamma coefficient of Goodman and Kruskal in [56] for the asymptotic normal distribution may be drawn. The first is

that if D=0 the estimate of the variance in the asymptotic distribution is zero. As long as the parameter $\Theta_D > 0$ the probability for D=0 decreases as $n \to \infty$ while if $\Theta_D = 0$, D equals 0. If Θ_D is close to zero, the frequency of D=0 may be high so that the asymptotic variance does not exists if the sample size is not large enough. The sample size is critical for the asymptotic results to apply and depends on Θ_D when close to zero. This is also in accordance with the theory of U-statistics for the asymptotic distribution. If the second moment does not exist, the asymptotic normal distribution will not apply. This shows that in practice, the asymptotic results not only concern the sample size, but also about when D is close to zero and Θ_D equals or is close to zero, which was shown in the simulations. A discussion of the procedure to verify the first order variance for the modified version of RP can be found in Wahlström [24].

It has been shown in the present thesis that the sample sizes required to use the normal distribution in inference, especially in using RV, may be large. But for substantial individual disagreement, which is important to statistically verify, sample sizes of about 60 was required. Depending on the distribution of observations contributing to a large RV, sample sizes low as about 20 to 40 may suffice. For a large amount of disorder, a sample size in the range of 20 to 40 was enough for using the normal distribution as an approximation. This may be compared to the simulation results for the difference of two treatments using the modified version of the measure of relative position RP, studied in Wahlström [24]. The simulation experiment showed that the type I error rates were within equivalence limits for sample sizes as small as 10 in each of the two groups.

The research in the field of statistical methods for analysis of dependent ordinal data in recent years has concentrated on development of models rather than on specific measures. Examples concerning models for agreement are log linear models such as uniform association models, non-uniform association models, and agreement plus uniform association models [49, 57-59]. Research regarding models for agreement in continuous variables has progressed in recent years [60] and some research has been conducted on developing methods of inference for correlation and concordance coefficients [61, 62].

Methods, originally for continuous data, are frequently used in analysis of ordinal data [63]. In their paper, Liu and Agresti [63] invited scholars to come with ideas of what can be done for disseminating knowledge and increasing the application of proper methods in the analysis of ordinal data. The response to the invitation showed the need was considered urgent and many of the suggestions were about demonstrating the merits and limitations of various methods and to make adequate methods accessible.

Many topics remain for further research. The theoretical analysis of asymptotic properties of RV by applying the theory of U-statistics may give answers to the convergence rates and alternative asymptotic distribution of RV. This would

also give an asymptotic variance to be estimated and compared with the jackknife variance estimator. The statistical properties of D and RV could be studied further in simulation experiments. Tentative answers about actual type I error rates using the normal distribution compared to the nominal type I error rates of α=0.05 and α=0.01 for different sample sizes and types of data should further increase the credibility in using the normal distribution. The properties of confidence intervals, such as coverage, length, and location could also be studied by simulations.

There is a great necessity to use appropriate methods for assessment of agreement and analysis of change when data consist of paired ordinal data. The results in this thesis have shown which methods are in use and of assumptions, implications, and of possible shortcomings and deficiencies in these methods. Furthermore, the results have shown some of the advantages with Svensson's approach and expanded the knowledge of the statistical properties for some of the measures. Svensson's approach and the measures of disorder and individual disagreement have been used in many research areas in recent years. Common usage has been in evaluation of construct validity by measuring order consistency, interchangeability and comparability between scales. Measures, like D, which can be used in situation where two scales have different number of categories, are valuable. Another important field of application has been analysis of reliability. For example, comparison of test-retest stability for a number of items in a multi-item questionnaire is of great value when evaluating intra-individual agreement, or lack of agreement, to be able to improve or eliminate the items with poor agreement. This has been accomplished by the measure RV. With a measuring instrument such as a rating scale possessing verified good validity and reliability, the analysis of change in treatment studies may proceed. The present thesis has shown that the studied measures have advantages compared to other models, tests, and measures. Above all, the thesis has shown that in certain circumstances, the normal distribution can be used in inference. Interval estimation by confidence interval and test of hypothesis regarding single or multiple parameters by use of the normal approximation may now be added to the means of statistical analysis.

Acknowledgments

There are many people who have helped and supported me in completing this thesis. I want to express my deepest gratitude to all of you.

Elisabeth Svensson, professor emerita and former professor in Statistics at the Swedish Business School at Orebro University, my supervisor and friend, for being so encouraging and engaged in me and my work. Your passion for science and inspiring lectures convinced me to start the long journey towards completing doctoral studies in statistics. Your support, constructive criticism and always being available for answering questions and for discussions have been of immense value for me.

Marieann Högman, professor and head of the Centre for Research and Development Uppsala University and County Council of Gävleborg (CFUG), for being encouraging and supportive in times of despair and for letting me work on my thesis within my employment. Your firm and generous offers to help made me feel comfortable.

Lennart Fredriksson, research supervisor and colleague at CFUG, for giving me support in technical matters regarding the preparation of ready-to-print manuscripts of the thesis, but above all, for encouragement, interest in my work, and stimulating discussions of research and science. You were there when I needed you most.

Inga-Lill Stenlund, for always being so helpful – a real doer – and for being so engaged in making me feel comfortable and happy and for your cheerful mood. The bell-ringing twice a day reminded me of reality.

Dag Rissén, research supervisor and colleague at CFUG, for support, sharing experiences, and helping me ease my workload.

Magnus Lindberg, former colleague and fellow doctoral student, for good cooperation and stimulating discussions about the many aspects of research and being a doctoral student. Your kindness has been greatly appreciated.

Vivi-Anne Rahm, former head of CFUG and FoU-forum, for letting me begin my doctoral studies within my employment and for giving me support and encouragement with your great enthusiasm.

Lawrence Teeland, former colleague, for considerable cooperation, mentorship and for valuable discussions of science and arts.

Jan Teeland, for excellent linguistic advice and for being so service-minded.

My present and former fellow doctoral students and senior researchers at CFUG for the many interesting discussions at the coffee table. These gatherings have been most inspiring and have opened one's mind in many directions. With the risk of forgetting some of you, I wish to name and especially thank: Johan Ahlgren, Catrine Björn, Kristina Bröms, Sevek Engström, Ingalill Feldmann, Per Fessé, Anders Holmlund, Benny Holmström, Ulla Johansson, Maria Lindberg, Anna-Greta Mamhidir, Arne Mordenfeld, Bernice Skytt, Eva Sving, Gösta Ullmark, Tomas Weitoft, and Kristina Vroland Nordstrand.

Emina Hadžibajramovic, fellow doctoral student, for valuable cooperation, comments on drafts of my manuscripts, and for taking notes at the final seminar.

Last, by not the least, to the most important person in my life, my beloved wife Monica for always giving me strong support, encouragement, endless love and creating happiness in my life. From now on we will have time for our common passions.

This work was financed by the Centre for Research and Development Uppsala University and the County Council of Gävleborg.

References

1. Bowling, A., *Measuring health: a review of quality of life measurement scales*. 3 ed. 2005, Buckingham: Open University Pr.
2. Dahlin-Ivanoff, S., U. Sonn, and E. Svensson, *Development of an ADL instrument targeting elderly persons with age-related macular degeneration.* Disability and rehabilitation, 2001. **23**(2): p. 69-79.
3. Fayers, P.M. and D. Machin, *Quality of life: the assessment, analysis, and interpretation of patient-reported outcomes*. 2 ed. 2007, Chichester: Wiley.
4. Lund, I., et al., *Evaluation of variations in sensory and pain threshold assessments by electrocutaneous stimulation.* Physiotherapy Theory and Practice, 2005. **21**(2): p. 81-92.
5. Streiner, D.L. and G.R. Norman, *Health Measurement Scales. A Practical Guide to Their Development and Use*. 3 ed. 2003, Oxford: Oxford University Press.
6. Svensson, E., et al., *The Balanced Inventory for Spinal Disorders. The Validity of a Disease Specific Questionnaire for Evaluation of Outcomes in Patients With Various Spinal Disorders.* SPINE, 2009. **34**(18): p. 1976-1983.
7. Cox, D.R., et al., *Quality-of-life Assessment: Can We Keep It Simple?* Journal of the royal statistical society. Series A (General), 1992. **155**(3): p. 353-393.
8. Hand, D.J., *Statistics and Theory of Measurements.* Journal of the Royal Statistical Society, Series A, 1996. **159**: p. 445-492.
9. Kampen, J. and M. Swyngedouw, *The Ordinal Controversy Revisited.* Quality and Quantity, 2000. **34**: p. 87-102.
10. Merbitz, C., J. Morris, and J.C. Grip, *Ordinal scales and foundations of misinference.* Archive of Physical Medicine and Rehabilitation, 1989. **70**(4): p. 308-312.
11. Stevens, S.S., *On the theory of scales of measurements.* Science, 1946. **103**: p. 677-680.
12. McCullagh, P., *Regression Models for Ordinal Data.* Journal of the Royal Statistical Society, Series B, 1980. **42**(2): p. 109-142.
13. Svensson, E., *Analysis of systematic and random differences between paired ordinal categorical data*. 1993, Stockholm: Almqvist & Wiksell International.
14. Svensson, E., *A Coefficient of Agreement Adjusted for Bias in Paired Ordered Categorical Data.* Biometrical Journal, 1997. **39**(6): p. 643-657.

15. Coste, J., J. Fermanian, and A. Venot, *Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals.* Statistics in Medicine, 1995. **14**(4): p. 331-345.
16. Kendall, M. and J.D. Gibbons, *Rank Correlation Methods*. Fifth ed. 1990, London: Hodder and Stoughton Limited.
17. Lehmann, E.L. and H.J.M. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day series in probability and statistics. 1975: New York ; San Francisco : Holden-Day.
18. Siegel, S. and J.N. Castellan, *Nonparametric statistics for the behavioral sciences*. 2 ed. 1988, New York: McGraw-Hill.
19. Svensson, E., *Guidelines to statistical evaluation of data from rating scales and questionnaires.* Journal of Rehabilitation Medicine, 2001. **33**(1): p. 47-48.
20. Svensson, E. and S. Holm, *Separation of systematic and random differences in ordinal rating scales.* Statistics in Medicine, 1994. **13**(23-24): p. 2437-2453.
21. Agresti, A., *A model for agreement between ratings on an ordinal scale.* Biometrics, 1988. **44**(2): p. 539-548.
22. Agresti, A., *Modelling patterns of agreement and disagreement.* Statistical Methods in Medical Research, 1992. **1**(2): p. 201-218.
23. Brennan, P. and A. Silman, *Statistical methods for assessing observer variability in clinical measures.* British Medical Journal, BMJ, 1992. **304**: p. 1491-1494.
24. Wahlström, H., *Nonparametric Tests for Comparing Two Treatments by Using Ordinal Data*. Örebro Studies in Statistics. Vol. 2. 2004, Örebro: Örebro University, University Library.
25. Svensson, E., *Concordance between ratings using different scales for the same variable.* Statistics in Medicine, 2000. **19**(24): p. 3483-3496.
26. Svensson, E., *Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating Scales.* Biometrical Journal, 2000. **42**(4): p. 417-434.
27. Svensson, E., *Ordinal invariant measures for individual and group changes in ordered categorical data.* Statistics in Medicine, 1998. **17**(24): p. 2923-2936.
28. Svensson, E. and J.-E. Starmark, *Evaluation of Individual and group changes in social outcome after aneurysmal subarachnoid haemorrage: A long-term follow-up study.* Journal of Rehabilitation Medicine, 2002. **34**: p. 251-259.
29. Kendall, M.G., *A new measure of rank correlation.* Biometrika, 1938. **30**(1/2): p. 81-93.
30. Kendall, M.G., *The treatment of ties in ranking problems.* Biometrika, 1945. **33**(3): p. 239-251.

31. Stuart, A., *The estimation and comparison of strengths of association in contingency tables.* Biometrika, 1953. **40**(1/2): p. 105-110.
32. Goodman, L.A. and W.H. Kruskal, *Measures of Association for Cross Classification.* Journal of the american statistical association, 1954. **49**(268): p. 732-764.
33. Somers, R.H., *A New Asymmetric Measure of Association for Ordinal Variables.* American Sociological Review, 1962. **27**(6): p. 799-811.
34. Agresti, A., et al., *Random-Effects Modeling of Categorical Response Data.* Sociological Methodology, 2000. **30**: p. 27-80.
35. Agresti, A. and R. Natarajan, *Modeling clustered ordered categorical data: A survey.* International Statistical Review, 2001. **69**(3): p. 345-371.
36. Goodman, L.A., *Simple Models for the Analysis of Association in Cross-Classifications having Ordered Categories.* Journal of the American Statistical Association, 1979. **74**(367): p. 537-552.
37. Schuster, C. and A. von Eye, *Model for Ordinal Agreement Data.* Biometrical Journal, 2001. **43**(7): p. 795-808.
38. Holmquist, N.S., C.A. McMahon, and O.D. Williams, *Variability in Classification of Carcinoma in situ of the Uterine Cervix.* Archives of Pathology, 1967. **84**: p. 334-345.
39. Landis, R.J. and G.G. Koch, *An application of hierarchial kappa-type statistic in the assessment of majority agreement among multiple observers.* Biometrics, 1977. **33**(2): p. 363-374.
40. Svensson, E., *Application of a rank-invariant method to evaluate reliability of ordered categorical assessments.* Journal of Epidemiology and Biostatistics, 1998. **3**(4): p. 403-409.
41. Hoeffding, W., *A class of statistics with asymptotically normal distribution.* The Annals of Mathematical Statistics, 1948. **19**(3): p. 293-325.
42. Thode, H.C., Jr., *Testing for Normality*. Statistics: Textbooks and Monographs. Vol. 164. 2002, New York: Marcel Decker, Inc.
43. Gan, F.F., K.J. Koehler, and J.C. Thompson, *Probability Plots and Distribution Curves for Assessing the Fit of Probability Models.* The American Statistician, 1991. **45**(1): p. 14-21.
44. Kruskal, W.H. and W.A. Wallis, *Use of Ranks in One-Criterion Variance Analysis.* Journal of the american statistical association, 1952. **47**(260): p. 583-621.
45. Dunn, O.J., *Multiple comparisons using rank sums.* Technometrics, 1964. **6**(3): p. 241-252.
46. Gibbons, J.D. and S. Chakraborti, *Nonparametric Statistical Inference*. 5 ed. 2010, Boca Raton: Chapman & Hall/CRC.
47. Friedman, M., *The Use of Ranks to Avoid the Assumtion of Normality Implicit in the Analysis of Variance.* Journal of the american statistical association, 1937. **32**(200): p. 675-701.

48. Friedman, M., *A Comparison of Alternative Tests of Significance for the Problem of m Rankings.* The Annals of Mathematical Statistics, 1940. **11**(1): p. 86-92.
49. Bagheban, A.A. and F. Zayeri, *A Generalization of the Uniform Association Model for Assessing Rater Agreement in Ordinal Scales.* Journal of Applied Statistics, 2010. **37**(8): p. 1265-1273.
50. Agresti, A., *Analysis of ordinal categorical data.* 1984, New York: John Wiley & Sons.
51. Lehmann, E.L., *Elements of large sample theory*. 1 ed. Spinger texts in statistics. 1998, New York: Springer-Verlag.
52. Serfling, R.J., *Approximation theorems of mathematical statistics*. 1980, New York: Wiley.
53. Carr, G.J., K.B. Hafner, and G.G. Koch, *Analysis of rank measures of association for ordinal data from longitudinal studies.* Journal of the American Statistical Association, 1989. **84**(407): p. 797-804.
54. Yang, Y., *Comparison of Change Between Groups with Data Having Rank-Invariant Properties Only*. 2009, Swedish Business School, Örebro University.
55. Lee, A.J., *U-statistics: theory and practice*. Statistics: Textbooks and Monographs. Vol. 110. 1990, New York: Marcel Dekker.
56. Goodman, L.A. and W.H. Kruskal, *Measures of association for cross classifications III: Approximate sampling theory.* Journal of the American Statistical Association, 1963. **58**(302): p. 310-364.
57. Aktaş, S. and S. Tülay, *Estimation of symmetric disagreement using a uniform association model for ordinal agreement data.* Advances in Statistical Analysis, 2009. **93**(3): p. 335-343.
58. Valet, F., et al., *Quality assessment of ordinal scale reproducibility: log-linear models provided useful information on scale structure.* Journal of Clinical Epidemiology, 2008. **61**(10): p. 983-990.
59. Valet, F., C. Guinot, and J.Y. Mary, *Log-Linear Non-Uniform Association Models for Agreement between two Ratings on an Ordinal Scale.* Statistics in Medicine, 2007. **26**(3): p. 647-662.
60. Hutson, A.D., *A multi-rater nonparametric test of agreement and corresponding agreement plot.* Computational statistics and data analysis, 2010. **54**(1): p. 109-119.
61. Woods, C.M., *Confidence Intervals for Gamma-Family Measures of Ordinal Association.* Psychological Methods, 2007. **12**(2): p. 185-204.
62. Woods, C.M., *Consistent Small-Sample Variances for Six Gamma-Family Measures of Ordinal Association.* Multivariate Behavioral Research, 2009. **44**(4): p. 525-551.
63. Liu, I. and A. Agresti, *The Analysis of Ordered Categorical Data: An Overview and a Survey of Recent Developments.* Sociedad de Estadistica e Investigacion Operativa. Test, 2005. **14**(1): p. 1-73.

Publications *in the series*
Örebro Studies in Statistics

1. Werner, Peter. *(2003) On the Cost-Efficiency of Mixed Mode Surveys Using the Web.*
2. Wahlström, Helen. *(2004) Nonparametric Tests for Comparing Two Treatments by Using Ordinal Data.*
3. Westling, Sara. *(2008) Cost efficiency of nonresponse rate reduction efforts – an evaluation approach.*
4. Högberg, Hans. (2010) *Some properties of measures of disagreement and disorder in paired ordinal data.*