UNIVERSITY OF GÖTEBORG

Department of Statistics

RESEARCH REPORT 1991:3

ISSN 0349-8034

STATISTICAL RANK METHODS FOR ORDINAL CATEGORICAL DATA

by

Sture Holm and Elisabeth Svensson

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

STATISTICAL RANK METHODS FOR ORDINAL CATEGORICAL DATA

Sture Holm and Elisabeth Svensson
Department of Statistics
University of Göteborg
Viktoriagatan 13
S-411 25 GÖTEBORG
SWEDEN

STATISTICAL RANK METHODS FOR ORDINAL CATEGORICAL DATA

ABSTRACT

The aim of this paper is to present a new rank method for analysing ordinal scale problems, and to give some of its basic properties. The method is suitable for the assessment of validity and reliability of health measurement instruments. We will be able to separate systematic and random differences between judges or scales and also, in a suitable way, measure the size of these two types of differences.
Some methods for estimating systematic differences between raters will be given.
The model is illustrated in a worked example.

HOLM S, SVENSSON E.
STATISTICAL RANK METHODS FOR ORDINAL CATEGORICAL DATA,

CONTENTS

1. INTRODUCTION

Measurement

Measurements in clinical research have traditionally been based on mortality and objective criteria like laboratory measurement values in order to record symptoms as the basis of judgements focused upon the presence or absence of disease. Nowadays, in clinical and other sociomedical research as well, there is a development of measuring techniques aimed at qualitatively assessing the health status or the quality of life of an individual before and after medical or surgical treatment. Many health care researchers have developed instruments consisting of subjective judgements obtained from questionnaires or rating scales. An excellent guide written by McDowell and Newell describe 50 sociomedical measurement methods [1]. The difficulty in developing health measuring instruments lies in assessing their validity, reliability and reproducibility and Teeling Smith [2] points out that still much research is needed to produce universally accepted and validated instruments.

Traditionally, it has been assessed that **measurement** is the assignment of numerals to objects or events according to rules classifying measurement data into different kind of scales [3]. The lowest level of scaling is assignments restricted to distinguish between two or more different categories.The scale is called **nominal** or **categorical.** A classification of individuals into different diagnoses is an example of nominal measure. Labelling ordered categorical data creates an **ordinal scale.** The numerals in the ordinal scale remain invariant under all order-preserving transformations, which means that the numerical codes do not represent any mathematical

value except indicating the rank order between categories. The fact that one succession of numerals of an ordinal scale can be replaced by another and that there are unequal , unmeasurable, distances between adjacent categories, scale codes, restricts the application of common mathematical and statistical methods .
An example of an ordinal scaling is the measuring of satisfaction used in a Social Support Questionnaire: "very satisfied, fairly satisfied, a little satisfied, a little dissatisfied, fairly dissatisfied , very dissatisfied " [1]

Qualitative measures differ from the quantitative measures in the interval and ratio scales by the former´s unstandardized way of constructing the scale and by the means of translating the categorical variables into a numerical form suitable for statistical analysis.

Measuring instruments record abstract and subjective phenomena formulated as items, a common concept for measuring instruments, often consisting of questions to be answered in rating scales or statements with which the rater has to agree or disagree.

Different approaches have been used in developing health indices. Some efforts are made to create health measurement scales that can be treated as interval scales, which simplifies the statistical treatment. The most common type is however still the ordinal scale.[1,2,3]

There will usually be a lot of tests made, when developing a new scale based on qualitative measures in order to assess its validity and reliability. This is also necessary when a measuring scale, originally designed for another kind of population, is to be used.

Validity can be defined as the extent to which an instrument measures what it is intended to measure. Validity is based on an interpretation of the instrument on a special set of data, that means that it is valid for a particular purpose depending on the definitions of the variables and the population. In this context we will mention some of the many different concepts of validity. **Content validity** expresses to what extent the chosen items reflect the aim of the measurement. **Criterion** validity is traditionally defined as the association of a new scale to the true state. If the criterion is available simultaneous the **concurrent validity** is assessed. On the other hand if the agreement of the new scale is assessed to outcomes in the future the concept is **predictive validity** [1] The **construct validity** indicates how well an instrument correlate to other.

The usefulness of a measurement scale depends also on the degree to which it can be replicated. **Reliability** is concerned with the extent to which repeated measurements on the same subject yield similar results.[1,4,5]

The main difference between assessing validity and reliability is formulated in the following practical definitions of the concepts: **Validity** refers to agreement between the true state and a fallible rater. This agreement will reflect the conformity of the rater and the true state. It is however quite impossible to find a "true state" in a health measuring scale - the closest you can get is agreement with a "gold standard", for instance a very qualified rater.

The most common paired measuring situation is however the one with two equivalent judges. The concept **reliability** refers to the agreement or consistency between two fallible raters [6,7] judging the same individual.

Note that the definitions of validity and reliability are based on **agreement** and that there is a big difference between agreement and association. If two raters perfectly agree in judging individuals, all observations will lie in the diagonal of the matrix of the categories. The association between the paired observations is also complete and positive in this situation. As an index of association, the coefficient of correlation is sometimes used to determine the reliability. This is however not suitable, since an inter-observer association may be very strong despite a weak inter-observer agreement in the judgements [7,9].

Random and systematic error of a measurement

An observed score can be considered to consist of two main parts: an underlying true score and **error** components. The amount of systematic error is connected with the validity concept. The random error component depends on the ability of the instrument to measure in a reproducible and consistent way [8].

The concept of **reliability** is commonly used for measurements in the nominal and ordinal scale .The corresponding term for assessing the random variabilty of measurements in interval and ratio scales, for instance in the calibration of laboratory instruments, is **precision** or **reproducibility**. The standard deviation is a measure of the imprecision of a measurement instrument in the interval or ratio scale.Since this parameter has no real meaning for nominal and ordinal scales, other methods are used to assess the reliability of measuring instruments in these scales.

The aim of measurement is to get information about the true level of the variable of an individual. According to traditional definitions: **reliability** is the proportion of observed variation in scores that is due to the true subject-to-subject variation. The **unreliability** is the proportion of variation that is due to random error in measurements. [1,8]

There are two main sources of inconsistency with repeated measurements depending on how they are obtained; internally for an observer and externally between observers. The agreement of an observer with himself is usually called **intra-observer agreement** but also **test-retest-reliability**. This measurement assumes **stability**, that is no change over time in stable subjects. A serious problem with estimating test-retest-reliability is the fact that the observer will remember the previous judgement. The repeated measurements are not independent in this case.

The agreement of two or more observers judging the same individual using the same measurement scale is termed **inter-observer agreement** or **inter-rater reliability**[1,3,5].

Assessing agreement

There are different approaches to assessing the inter-observer agreement or concordance between the two measurements on the same individual. One measure of agreement is the proportion of agreement among the total number of judgements [3,9]. This index does not take into account the amount of agreement expected by chance and it also ignores partial agreement and disagreement. It is possible to improve it by defining weights to the judgements of partial agreement. But not even the weighted percentage agreement corrects for agreement expected by chance. [3,9]

Probably the most popular measure for summarizing degree of agreement between two raters is the **coefficient kappa** introduced by Cohen in 1960. There exist many reports on kappa statistics. An extensive and careful treatment is given in the thesis by Schouten [9] **Kappa** is the degree of agreement above the expected random agreement divided by the maximum possible excess agreement, i e $\kappa = (p_O - p_e)/(1-p_e)$.

The observed proportion agreement is here denoted p_O, while p_e is the chance expected proportion agreement. In the calculation of p_e it is supposed that the two observers are **independent.**

In spite of the popularity of kappa , several authors have pointed out some unsatisfactory features. For instance kappa is a summarized index of agreement, not distinguishing between systematic and random deviations. There are different approaches in calculating weighted kappa values [10,11,12,13] which may complicate the interpretation. Furthermore, kappa values from different samples are not comparable if the number of response categories is not the same or if the samples do not represent the same underlying population [9].

McCullagh [14] and Agresti [15] propose log -linear models for agreement analysis as well as for analysis of ordered categorical data for each judge. These are parametric models of a particular type.

The aim of the present paper is to present a new rank method for analysing ordinal scale problems, and to give some of its basic properties. We will be able to separate systematic and random differences between judges or scales and also in a suitable way measure the sizes of these two types of differences.

We will illustrate our method in a worked example on a material used in the thesis by Schouten 1985 [9].

2. A STATISTICAL RANK METHOD FOR ORDINAL SCALES

The model

Suppose there are two judgements in ordinal scales with m_1 and m_2 categories respectively. This first general discussion allows for different categorical scales with unequal number of categories. Later on the special case, $m_1 = m_2$ and the case of inter-rater reliability will be treated.

Suppose further that n individuals from the same population are used in the two judgements and that the judgements of different individuals are independent.

The probability of rating a randomly chosen individual to the i:th category by judgement 1 and to the j:th category by judgement 2 is denoted by p_{ij} . The numbers of judgements in cell (ij), x_{ij} i=1,.....,m_1 and j = 1,....., m_2 , have a multinomial distribution with parameters n and p_{ij}, i=1,...,m_1 and j=1,...,m_2.

A valid scale with reliable judgements will have high probability for scores close to the diagonal of agreement .

A particular ranking

We will now introduce a rank transformation, common in non parametric statistics and related to the ROC curve (Relative operating characteristic) used in some medical statistical problems [16]

The number of observations obtained in category i of rater 1 equals

$$x_{i.} = \sum_{j=1}^{m_2} x_{ij}$$

The ranking of judgement 1 gives the observations in category (i) the following ranks

$$\sum_{v=1}^{i-1} x_{v.} + 1, \ldots\ldots, \sum_{v=1}^{i} x_{v.}$$

In ranking the observations of judgement 1 we use the convention of making the internal ranking of category (i) according to the ranks of judgement 2. Thus the observations in cell (ij) get the following ranks from judgement 1:

$$\sum_{i_1=1}^{i-1} x_{i_1.} + \sum_{j_1=1}^{j-1} x_{i\,j_1} + 1 \quad \text{to} \quad \sum_{i_1=1}^{i-1} x_{i_1.} + \sum_{j_1=1}^{j} x_{i\,j_1}$$

and the mean rank

$$\bar{R}_{ij}^{(1)} = \sum_{i_1=1}^{i-1} x_{i_1.} + \sum_{j_1=1}^{j-1} x_{i\,j_1} + \frac{1}{2}(1+x_{ij}) \qquad (1)$$

In the same way we can define mean judgement 2 rank for cell (ij) as

$$\bar{R}_{ij}^{(2)} = \sum_{j_1=1}^{j-1} x_{\cdot j_1} + \sum_{i_1=1}^{i-1} x_{i_1 j} + \frac{1}{2}(1+x_{ij}) \qquad (2)$$

where $x_{\cdot j} = \sum_{i=1}^{m_1} x_{ij}$

Observe that $\bar{R}_{ij}^{(1)}$ and $\bar{R}_{ij}^{(2)}$ are defined only if $x_{ij} > 0$.

Definition:

Two sets of judgements of the same n individuals are called **rank transformable** if $\bar{R}_{ij}^{(1)} = \bar{R}_{ij}^{(2)}$ for all (ij) such that $x_{ij} \geq 1$.

Remark 1:

When two sets of judgements are rank transformable, there exists a common ranking for the two sets of ranking. Our convention to rank the individuals in cell (ij) in the same order in both ranking gives this common ranking if it exists.

Remark 2:

If there exists some very clear ordering among the individuals the judgements will be exactly or approximately rank transformable and the rates will essentially describe the individual interpretation of the measuring scale by the two raters.

If on the other hand, there is no definite ordering among the individuals, there will appear random differences between the two judgements, resulting in different mean ranks $\bar{R}_{ij}^{(1)}$ and $\bar{R}_{ij}^{(2)}$ for the cells (ij).

Assessing random differences

If two judgements are not rank transformable, the difference $\bar{R}_{ij}^{(1)} - \bar{R}_{ij}^{(2)}$ indicates locally a deviation from the rank transformable case, which means that there is a random error in the two judgements.

Our convention to rank the individuals within a cell (ij) in the same order for both judgements means that each observation in the cell has the same rank difference $\bar{R}_{ij}^{(1)} - \bar{R}_{ij}^{(2)}$

According to formulas (1) and (2) this difference can be written

$$\bar{R}_{ij}^{(1)} - \bar{R}_{ij}^{(2)} = \sum_{i_1<i} \sum_{j_1>j} x_{i_1 j_1} - \sum_{i_1>i} \sum_{j_1<j} x_{i_1 j_1} \qquad (3)$$

Given that one particular out of n observations appears in cell (ij) , its expected rank difference can be written

$$(n-1)\left[\sum_{i_1<i} \sum_{j_1>j} p_{i_1 j_1} - \sum_{i_1>i} \sum_{j_1<j} p_{i_1 j_1} \right] = (n-1)(q_{ij}^{(ul)} - q_{ij}^{(lr)}) \qquad (4)$$

where $q_{ij}^{(ul)}$ is the upper left probability $\sum_{i_1<i} \sum_{j_1>j} p_{i_1 j_1}$ of cell (ij)

and $q_{ij}^{(lr)}$ is the lower right probability $\sum_{i_1>i} \sum_{j_1<j} p_{i_1 j_1}$ of cell (ij), related to the diagonal of agreement.

Lemma 1:

The expected rank difference of a randomly chosen observation is

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (n-1)\,(q_{ij}^{(ul)} - q_{ij}^{(lr)})\, p_{ij} = 0$$

Proof:

For each $i_1 < i$ and $j_1 > j$ the product $p_{i_1 j_1}\, p_{ij}$ appear twice,
once with a plus sign and once with a minus sign in the calculation
QED

Given the event that one particular observation occurs in cell (ij), the **conditional variance of the rank difference** $(R_{ij}^{(1)} - R_{ij}^{(2)})$ associated with that single observation is

$$(n-1)\,[\, q_{ij}^{(ul)}\,(1 - q_{ij}^{(ul)}) + q_{ij}^{(lr)}\,(1 - q_{ij}^{(lr)}) + 2\, q_{ij}^{(ul)} q_{ij}^{(lr)}\,] =$$

$$(n-1)\,[\, q_{ij}^{(ul)} + q_{ij}^{(lr)} - (\, q_{ij}^{(ul)} - q_{ij}^{(lr)}\,)^2\,] \qquad (5)$$

The variance (V_R) of the rank difference associated with a randomly placed observation equals:

$$V_R = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij}\,(n-1)^2\,(q_{ij}^{(ul)} - q_{ij}^{(lr)})^2 +$$

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij}\,(n-1)\,[\, q_{ij}^{(ul)} + q_{ij}^{(lr)} - (\, q_{ij}^{(ul)} - q_{ij}^{(lr)}\,)^2\,] =$$

$$\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij}\,(n-1)\,(n-2)\,(\, q_{ij}^{(ul)} - q_{ij}^{(lr)}\,)^2 + \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij}\,(n-1)\,[q_{ij}^{(ul)} + q_{ij}^{(lr)}] \qquad (6)$$

This formula (6) is obtained by conditioning on the place of the random observation.

The variance of the rank difference is a measure of the random error in the two judgements of the same individuals. Its value increases with the number of observations .

A standardized variance, (= the variance of the relative rank difference) is obtained by dividing by $(n-1)^2$. An estimate of the standardized variance can be obtained by replacing the probabilities in V_R by the corresponding relative frequencies. The standardized variance can also be estimated by the mean of the squares of the obtained rank differences. Note that the expectation of the rank difference is 0 and so is also the mean of all obtained rank differences. The mean of the squares of rank differences equals

$$\frac{1}{n^3} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (\bar{R}_{ij}^{(1)} - \bar{R}_{ij}^{(2)})^2 \; x_{ij}$$

Assessing systematic differences

Consider next the systematic difference between two raters using the same scale, with m categories, independently judging n subjects. There is no systematic empirical disagreement between the two raters if

$$\sum_{i=1}^{\nu} \sum_{j=1}^{m} x_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{\nu} x_{ij} \quad \text{for } \nu = 1,2,....,m-1$$

A reasonable empirical measure of the systematic difference at the upper boundary of category v is the difference

$$\sum_{i=1}^{v}\sum_{j=1}^{m} x_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{v} x_{ij} = \sum_{i=1}^{v}\sum_{j=v+1}^{m} x_{ij} - \sum_{i=v+1}^{m}\sum_{j=1}^{v} x_{ij} \qquad (7)$$

which we denote by Y_v. The expectation of this difference equals

$$n\left(\sum_{i=1}^{v}\sum_{j=v+1}^{m} p_{ij} - \sum_{i=v+1}^{m}\sum_{j=1}^{v} p_{ij}\right) = n\left(q^{(ul)}_{v,v+1} - q^{(lr)}_{v+1,v}\right) \qquad (8)$$

By considering the multinomial distribution of the parts constituting this difference, we easily find its variance to be

$$n\left[q^{(ul)}_{v,v+1} + q^{(lr)}_{v+1,v} - \left(q^{(ul)}_{v,v+1} - q^{(lr)}_{v+1,v}\right)^2\right] \qquad (9)$$

The variance can be estimated by replacing the probabilities by the corresponding relative frequencies.

The random variables Y_v for different v:s are not independent.

If $v_1 < v_2$

$$\mathrm{Cov}\left(Y_{v_1}, Y_{v_2}\right) =$$

$$n\left[q^{(ul)}_{v_1,v_2+1} + q^{(lr)}_{v_2+1,v_1} - \left(q^{(ul)}_{v_1,v_1+1} - q^{(lr)}_{v_1+1,v_1}\right)\left(q^{(ul)}_{v_2,v_2+1} - q^{(lr)}_{v_2+1,v_2}\right)\right] \qquad (10)$$

The covariances can also easily be estimated by substituting probabilities by their corresponding relative frequencies.
Thus for the variables Y_v , $v = 1, \ldots, m-1$ we have now the whole covariance matrix as well as a basic estimate of that matrix.

The statistics Y_ν , ν =1, ..., m-1 can be used to estimate the detailed behaviour of a possible systematic difference between two judges, that is the systematic difference in the determination of the inter-rater reliability.

For increasing sample size n, the normalized statistics $Z_\nu = \frac{Y_\nu}{n}$ converge with probability 1 to the parameters ($q^{(ul)}_{\nu,\nu+1}$ - $q^{(lr)}_{\nu+1,\nu}$) which describe the detailed systematic inter-rater behaviour.

A simple parametric model for systematic differences

The systematic inter-rater difference is exhaustively described in a nonparametric way by all the category probabilities $p^{(\lambda)}_\nu$ ν=1,...., m for the two raters λ = 1,2 or the two corresponding sets of cumulative probabilities

$$q^{(\lambda)}_\nu = \sum_{i=1}^{\nu} p^{(\lambda)}_i \qquad \nu = 1,2,....,m-1$$

It might however be reasonable to use some simple model with a few parameters, e g two parameters describing a tendency for one rater to be shifted in some direction relative to the other or being more or less concentrated in the categories compared to the other.

Let

$g_1(q) = \frac{1}{2} - 2(\frac{1}{2} - q)^2 = 2(q - q^2)$

$g_2(q) = 4q\,(q-1)(\,q - \frac{1}{2}\,) = 4q^3 - 6q^2 + 2q$

and let $q^{(1)}$ and $q^{(2)}$ be two probabilities.

Then the equation

$$q^{(2)} - q^{(1)} = \Theta_1 \cdot g_1[\tfrac{1}{2}(q^{(1)} + q^{(2)})] + \Theta_2 \cdot g_2[\tfrac{1}{2}(q^{(1)} + q^{(2)})] \quad (11)$$

determines a curve in the rectangle [0,1] x [0,1].

The values of the parameters Θ_1, Θ_2 should satisfy $|\Theta_1| \leq 1$, $|\Theta_2| \leq 1$, $|\Theta_1 + \Theta_2| \leq 1$ and $|\Theta_1 - \Theta_2| \leq 1$.

The following figures show the two cases $\Theta_1 = \frac{1}{2}$, $\Theta_2 = 0$ and $\Theta_1 = 0$ and $\Theta_2 = \frac{1}{2}$.

In the first case, figure 1a, if $q^{(1)}$ and $q^{(2)}$ represent cumulative distribution functions in the same point, the distribution corresponding to $q^{(1)}$ is shifted to the right compared to the distribution of $q^{(2)}$. This means that there is a systematic difference in position of the two distributions.

Analogously in the second case, figure1b, There is a difference in concentration between the distribution functions of $q^{(1)}$ and $q^{(2)}$.

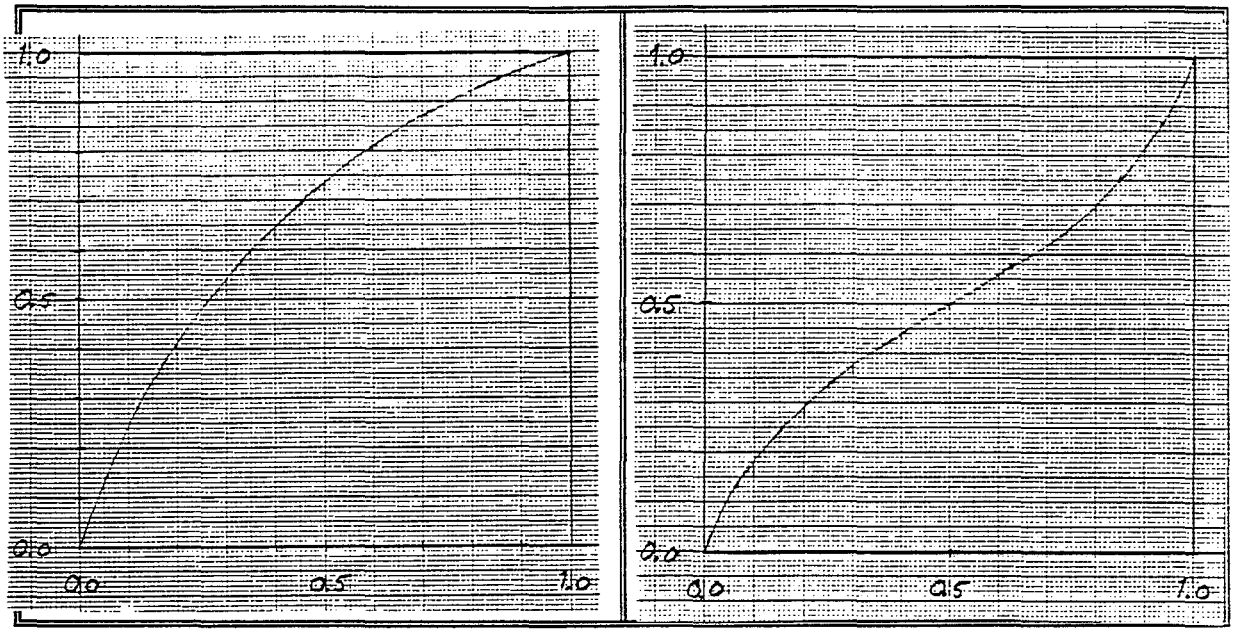

Figure 1a: An illustration of systematic difference in position between two distributions.
$\Theta_1 = \frac{1}{2}$, $\Theta_2 = 0$
$g_1(q) = 2(q - q^2)$

Figure 1b: An illustration of systematic difference in concentration between two distributions.
$\Theta_1 = 0$ and $\Theta_2 = \frac{1}{2}$.
$g_2(q) = 4q^3 - 6q^2 + 2q$

In the previous section we have also determined the variances and covariances for the differences of the ranks constituting the coordinate estimates of the curves.

Estimation

This estimated covariance matrix $\hat{C}$ could be used as a weight when estimating the parameters Θ_1 and Θ_2. We denote by Q the (m-1)x2 matrix with following elements

$Q_{\nu k} = g_k(\hat{q}_\nu)$ $\quad \nu = 1, \ldots, m-1 \quad k = 1,2$

where $\hat{q}_\nu$ is the estimate of the mean cumulative probability after category ν i.e.

$$\hat{q}_i = \frac{1}{2n}\left(\sum_{i=1}^{\nu}\sum_{j=1}^{m} x_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{\nu} x_{ij}\right)$$

Then a suitable estimate of $\Theta = \begin{bmatrix}\Theta_1 \\ \Theta_2\end{bmatrix}$ is obtained by minimizing $(Z - Q\Theta)' \hat{C}^{-1} (Z - Q\Theta)$, where Z is the (m-1)-dimensional observation vector with components Z_ν $\nu = 1,\ldots,m-1$.

The solution is $\hat{\Theta} = (Q' \hat{C}^{-1} Q)^{-1} Q' \hat{C}^{-1} Z$ (12)

Beside the computational force needed, there arises a practical problem in this context. In particular, for small sample sizes it might happen that

$$\sum_{i=1}^{\nu}\sum_{j=\nu+1}^{m} x_{ij} = \sum_{i=\nu+1}^{m}\sum_{j=1}^{\nu} x_{ij} = 0$$ i.e. the estimate of Var Y_ν equals zero.

In order to get a fully valid method, the above one needs some revision.

An attractive alternative in measuring relative position and concentration of two distributions of judgements would be the use of some kind of general parameter not related to any particular model.

Consider first two continuous distributions with cumulative distribution functions F(x) and G(x) respectively. Then a possible measure of their positions relative to each other is

$\int F(x)\, dG(x) - \int G(x)\, dF(x)$.

If X and Y are independent random variables with these distributions, the measure can be interpreted in terms of probabilities as

$P(X \leq Y) - P(Y \leq X)$.

Obviously the inequality $-1 \leq P(X \leq Y) - P(Y \leq X) \leq 1$ is applicable and both bounds can be attained.

An analogous measure of **the relative position for discrete distributions,** for example from inter-rater comparisons of the same scale with m categories, is obtained by

$$\sum_{v} P(X \leq v)\, P(Y = v) - \sum_{v} P(Y \leq v)\, P(X = v)$$

$$= \sum_{v} [\, P(X < v)\, P(Y = v) - P(Y < v)\, P(X = v)] \qquad (13)$$

A positive value of the relative position indicates that the distribution function of the X-values is shifted to the left relative the distribution function of Y.

For continuous cumulative distribution functions F and G the difference

$\int F(x)\,[1-F(x)]\,dG(x) - \int G(x)\,[1-G(x)]\,dF(x)$ can be interpreted probabilistically as $P(X_1 \le Y_1 < X_2) - P(Y_1 \le X_1 < Y_2)$ for independent random variables X_1, X_2 and Y_1, Y_2 with cumulative distribution functions F and G respectively. This difference measures a **difference of concentration** between the two distributions.

It can be shown that

$$-p(1-p) \le P(X_1 \le Y_1 < X_2) - P(Y_1 \le X_1 < Y_2) \le p(1-p)$$

where $p = P(X_1 \le Y_1)$.

Analogously for the **discrete case** with the same set of outcomes in X and Y:

$$P(X_1 < Y_1 < X_2) - P(Y_1 < X_1 < Y_2)$$

$$= \sum_{\nu} [P(Y=\nu)\,P(X<\nu)\,P(X>\nu) - P(X=\nu)\,P(Y<\nu)\,P(Y>\nu)]$$

where $\nu = 1,2,\ldots, m.$ (14)

Bounds of this difference are determined according to the following lemma:

Lemma 2:

$$-\min(p_0 - p_0^2,\ p_1 - p_1^2) \le P(X_1 < Y_1 < X_2) - P(Y_1 < X_1 < Y_2) \le \min(p_0 - p_0^2,\ p_1 - p_1^2),$$ where

$$p_0 = P(Y_1 \le X_1) = \sum_{\nu} P(Y_1 \le \nu)\,P(X=\nu) \text{ and}$$

$$p_1 = P(Y_1 < X_1) = \sum_{\nu} P(Y_1 < \nu)\,P(X=\nu) \quad (15)$$

Proof:

Consider the event $(Y_1 < X < Y_2)$ for independent X, Y_1, Y_2 with cumulative distribution functions F(X) and G(X) for X and Y respectively. Further, let $f(v)$ and $g(v)$ denote the corresponding probabilities for possible outcomes $v = 1,2,..., m$.

Then $P(Y_1 < X < Y_2)$

$$= \sum_v G(v-1)\, f(v)\, [1 - G(v)] \leq \sum_v G(v)\, f(v)\, [1 - G(v)]$$

Denoting $\sum_v G(v)\, f(v) = P(Y_1 \leq X) = p_0$, we get

$$P(Y_1 < X < Y_2) \leq p_0 - \sum_v G^2(v)\, f(v)$$

$$= p_0 - \sum_v (G(v) - p_0)^2 f(v) - p_0^2 \leq p_0 - p_0^2$$

In the same way

$$P(Y_1 < X < Y_2) = \sum_v G(v-1)\, f(v)\, [1 - G(v)]$$

$$\leq \sum_v G(v-1)\, f(v)\, [1 - G(v-1)] \leq p_1 - p_1^2$$

where $p_1 = \sum_v G(v-1)\, f(v) = P(Y_1 < X)$.

The corresponding proofs for $P(X_1 < Y_1 < X_2)$ are literally the same.

For a given p_0 or p_1 the bounds is attained when one distribution is completely concentrated between the parts p_0 and $(1- p_0)$ or p_1 and $(1-p_1)$ respectively of the other. QED

Remark 1:

The property in this lemma is related to the maximum variance of a Wilcoxon statistic for continuous distributions, obtained by Birnbaum and Klose,1957 [17] Our principle technique of proof would also apply to that problem and it is somewhat simpler than their technique.

Remark 2:

The boundaries of the difference in the lemma are given by the probability p_0 or p_1 depending on their distance from ½ . The probability value with the greatest absolute difference to ½ applies.

We can now make a suitably normalized measure of relative concentration for two judgers using the same scale. Denoting the upper bound M we will use the expression

$$\frac{1}{M}[P(X_1 < Y_1 < X_2) - P(Y_1 < X_1 < Y_2)] =$$
$$\frac{1}{M}[\sum_{v} F(v-1)\, g(v)\, (1 - F(v)) \; - \; \sum_{v} G(v-1)\, f(v)\, (1 - G(v)] \qquad (16)$$

which always has a value in the interval [-1, 1]. The two extreme values correspond to one distribution entirely concentrated in relation to the other.

The values of the relative position and the relative concentration of two judgements in a total agreement are both zero, while disagreement will result in nonzero values of one or both of the measures.

The case of equal distributions of X and Y is one example of zero relative concentration. Another example is obtained by having three possible outcomes, e g $v_1 < v_2 < v_3$ and

$g(v_1) = p_1$, $g(v_2) = 1-p_1$, $g(v_3) = 0$

$f(v_1) = 0$ $f(v_2) = 1-p_1$, $f(v_3) = p_1$ for some p_1, $0 < p_1 < 1$.

Intuitively these two distributions also have the same concentration relative to each other. Note that the values of v_1, v_2 and v_3 have no influence on the measure.

Suitable empirical measures are obtained by substituting relative frequencies for the corresponding probabilities in the theoretical measure.

3. EXAMPLE

Inter-rater agreement

In his dissertation 1985 Schouten [9] demonstrated the kappa statistics using results from a study designed to investigate the inter-rater reliability in a histological classification of carcinoma in situ. We will give a worked example of our rank model using parts of the same material and compare our measures with corresponding kappa value.

Two pathologists separately classified 118 biopsy slides into one of five ordered categories ranged from 1 = no signs of carcinoma to 5 = invasive carcinoma.

Figure 2 shows the result of the paired independent judgements of the 118 biopsy slides [9,pp6].

category by pathologist 1	1	2	3	4	5
5			3		3
4		1	14	7	
3		2	36		
2	5	7	14		
1	22	2	2		

category by pathologist 2

Figure 2. Observed frequencies x_{ij} of biopsy slides classified by two pathologist, the same material as in the thesis by Schouten [9].

The observed proportion of agreement between both pathologists is 64 percent since 75 of the 118 biopsy slides were equally classified by the two pathologists.

There are two main sources of disagreement of the pathologists; random misclassification on one hand and systematic error on the other.

Random differences

In order to assess the random differences of the paired classifications, the mean ranks were calculated and shown in figure 3. The observations in all cells but (1;1), (3;4) and (5;5) - provided observations - contribute to the random differences between the two classifications. The variance of the rank differences according to formula (6) is 36.299. The estimate of the standardized variance is 0.00265.

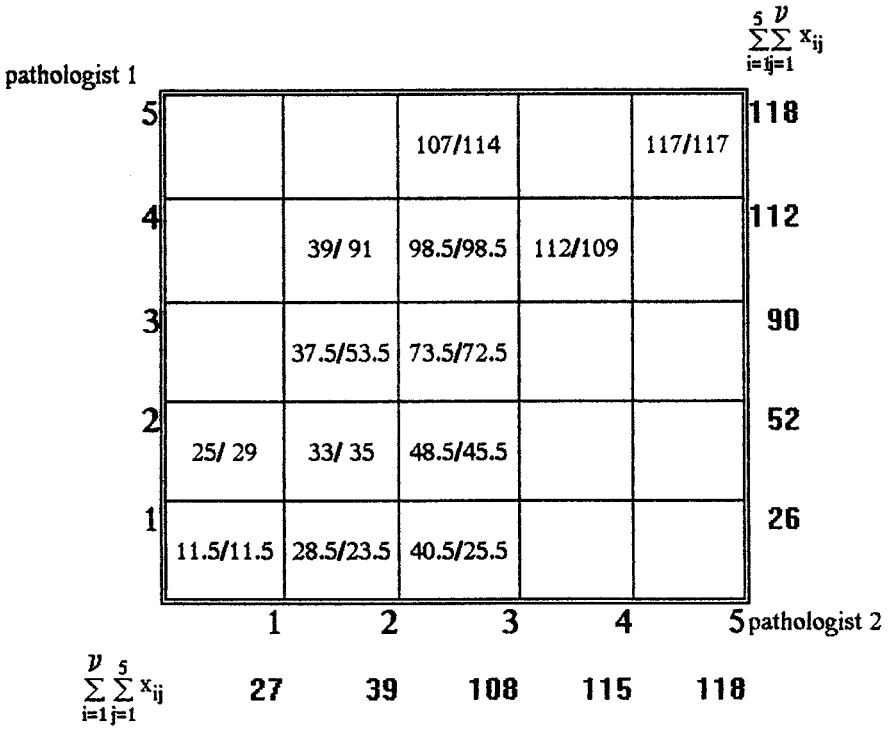

Figure 3: Mean ranks for the 118 biopsy slides independently classified to cell (ij) by pathologists (1) and (2), written in following way: $\overline{R}_{ij}^{(2)} / \overline{R}_{ij}^{(1)}$

Systematic differences

If the two pathologists do not agree on the item descriptions of the five categories there will be a systematic difference between the two judgements of the biopsy slides. Occurrence of systematic differences between two raters attenuates the validity of the measuring instrument. The two sets of cumulative frequencies for the five categories, also shown in figure 3, determine the relative lengths of the categories for the two judges. Consequently, these different lengths visualize the systematic differences as shown in figure 4.

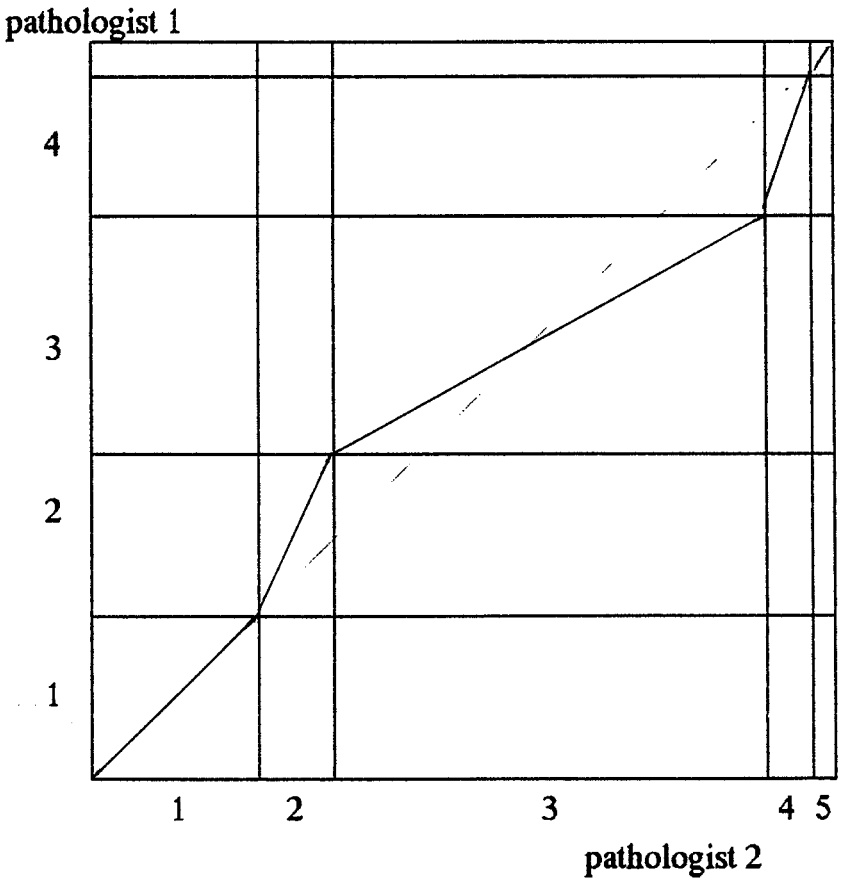

Figure 4: The systematic differences between the two pathologists appears as different lengths between categories, determined by Y_p .

The pathologists differ essentially in opinion about the categories 2,3 and 4. The items in the measuring instrument of the histological classification of carcinoma may be ambiguously described. This means that the biopsy slides have the risk of getting different classifications from the two pathologists.

The systematic differences between the two pathologists can be divided into the relative position and the relative concentration of the paired classifications.
Table 1 shows the distributions of the classifications made by the pathologists.

category v	1	2	3	4	5	**total**
pathologist 2 (X)	27	12	69	7	3	118
pathologist 1 (Y)	26	26	38	22	6	118

Table 1: The observed frequency distributions of the classifications into five categories made by two pathologists.

Thus the relative position according to formula (13) is 0,0276. The positive sign indicates that a greater part of the distribution function of the classifications made by the pathologist named 2 is shifted to the left relative the distribution function of the classifications made by pathologist named 1.

Further we have $p_0 = 0.573$ and $p_1 = 0.349$ giving following differences $|p_0 - 0.5| = 0.073$ and $|p_1 - 0.5| = 0.151$. Thus the norming constant in this example is: $M = 0.349(1-0.349) = 0.227$.

According to formulas (14) and (16) the estimated relative concentration is -0.0286 and the estimate of the normalized relative concentration of the paired classifications equals -0.126 .

Table 2 summarizes the different measures of agreement for the example.

measure of agreement	observed value	the value for total agreement
The random error: standardized variance of the rank differences	0.00265	0
The systematic error: * the value of relative position * the value of relative concentration	0.0276 -0.126	0 0
The coefficient kappa	0.499	+1
The weighted kappa (disagreement weights)	0.650	+1
The proportion of agreement	64%	100%
The weighted proportion of agreement	90%	100%

Table 2: Different agreement measures of the example.

4 DISCUSSION

Today health measurement scales are important complements to the measurements made by laboratory methods. The problem, though, with a measurement based on categorical data is to ensure its validity and reliability. A common approach is to use the correlation coefficient i order to assess the validity of a measuring instrument and coefficient kappa to assess its reliability. Those methods cannot separate the different sources of unreliability and unvalidity.

We suggest here non parametric measures which enables relevant descriptions of validity and reliability properties of ordinal scales. The worked example shows the behaviour of the method in practice. It is possible to visualize the systematic difference between two observers or two methods and directly point out those categories who have the greatest systematic difference. The systematic error between the two measurements will be calculated by using a relative position measure and a relative concentration measure of the two distribution functions.

Our metod is useful in developing instruments. It gives a possibility to validate the discriminant quality in different scale categories, which is important in the process of developing descriptions of items used in the scales.

In this paper we have not developed all relevant statistical properties for the suggested measures. Such a development is needed in order to get a full understanding of the meaning of the suggested measures. These supplementary properties of the method will be presented in a forthcoming paper.

5. REFERENCES

1. **McDowell I, Newell C.** Measuring health. a guide to rating scales and questionnaires.Oxford:Oxford University Press, 1987

2. **Teeling Smith G.** ed.Measuring health: a practical approach. Chichester: John Wiley & sons, 1988.

3. **Stevens SS.** On the theory of scales of measurement. Science 1946;103:677-80

4. **Streiner DL, Norman GR.**Health measurement scales -a practical guide to their development and use. Oxford:Oxford University Press, 1989

5. **Koran LM.** The reliability of clinical methods, data and judgements. The New England Journal of Medicine 1975;293:642-6 695-701.

6. **Spitznagel EL, Helzer JE.** A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985;42:725-8.

7. **Kramer MS, Feinstein AR.** Clinical biostatistics LII a primer on quantitative indexes of association. Clin Pharmacol Ther 1980;28:130-45

8. **Fleiss JL, Shrout PE.** Reliability considerations in planning diagnostic validity studies. In: Robins LN, Barrett JE. eds. The validity of psychiatric diagnosis. New York: Raven Press,1989:279-91

9. **Schouten HJA.** Statistical measurement of interobserver agreement. Rotterdam:Erasmus Universitet, Institute of Biostatistics, 1985. Thesis.

10.**Cicchetti DV.** Assessing inter-rater reliability for rating scales: resolving some basic issues. Brit J Psychiat 1976;129:452-6.

11.**Hall JH.** Inter-rater reliabilty of ward rating scales. Brit J Psychiat 1974;125:248-55.

12. **Kramer MS, Feinstein A.** Clinical biostatistics LIV The biostatistics of concordance. Clin Pharmacol Ther 1981;29:111-23.

13. **Maclure M, Willett WC.** Misinterpretation and misuse of the kappa statistic. Americal Journal of Epidemiology 1987;126:161-69

14. **McCullagh P.** Regression models for ordinal data. J Statist Soc B 1980;42(2):109-42.

15. **Agresti A.** A model for agreement between ratings on an ordinal scale. Biometrics 1988;44:539-48

16. **Swets JA.** Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. Psychological Bulletin 1986;99(1):100-17

17 **Birnbaum ZW, Klose O.** Bounds for the variance of the Mann-Whitney statistic. Ann Math Statist 1957;28:933-45.

1990:1	Holm, S.	Abstract bootstrap confidence intervals in linear models.
1990:2	Holm, S. & Dahlbom, U	On tests of equivalence
1991:1	Olofsson, Jonny	On some prediction methods for categorical data
1991:2	Jonsson, Robert	On the problem of optimal inference in the simple error component model for panel data
1991:3	Holm, S. & Svensson, E.	Statistical rank methods for ordinal categorical data