J Rehab Med 2001; 33: 47–48

GUIDELINES TO STATISTICAL EVALUATION OF DATA FROM RATING SCALES AND QUESTIONNAIRES

Elisabeth Svensson

From the Department of Statistics, Örebro University, Örebro, Sweden

Correspondence address: Elisabeth Svensson, Department of Statistics, Örebro University, SE-701 85 Örebro, Sweden.

(Accepted August 15, 2000)

INTRODUCTION

Questionnaires and rating scales are commonly used to measure qualitative variables, such as feelings, attitudes and many other behavioural and health-related variables. There are different types of instruments ranging from single scales to multi-dimensional, multi-item questionnaires. The scaling of the responses can vary from the dichotomous alternatives "yes" and "no" to a mark on a line, as in the visual analogue scale (VAS). Numerical labels are commonly used for the recordings. Nevertheless, irrespective of the type of scaling, the item responses indicate only an ordered structure and not a numerical value in a mathematical sense. Such data are often called ordered categorical or ordinal (1–4). Statistical methods for data from rating scales must take account of the rank-invariant properties of ordinal data, which means that the methods must be unaffected by a relabelling of the scale categories. Hence, statistical methods applicable to data from rating scales differ completely from the traditional methods for quantitative variables, since calculations based on adding or subtracting ordinal data are not appropriate. Sum scores of multi-item assessments, the mean value, standard deviation and calculation of differences for description of change in score do not have an interpretable meaning and must be avoided in the statistical evaluation of data from rating scales and questionnaires (4–6).

Traditionally, in applied research, there is a temptation to treat data from rating scales as numerical on an interval level (4, 5). It should be emphasized, however, that data on an interval level are quantitative, which means that such data have the mathematical properties of well-defined size and equidistance, but the same variable does not have the same ratio when it is measured in different units (1). Hence, qualitative data could never gain the properties required for being treated as interval data. Statistical methods for quantitative data are valid only when data have the mathematical properties of well-defined size and distance, and conclusions drawn from such analyses are solely interpretable and reliable for quantitative data. However, quantitative data such as blood pressure could be treated as ordinal when categorized as "low", "normal" and "high". Such categorization changes the choice of appropriate statistical methods of analysis.

GUIDELINES FOR STUDIES INCLUDING RATING SCALES AND/OR QUESTIONNAIRES

The measurement process

In the absence of standard instruments there is a considerable variety in the types of instruments and scales that are available to assess the same qualitative variable. The authors should therefore motivate the choice of measurement instrument and, in the case of a known instrument, refer to the main source. Important considerations for the choice are the operationalization process, which includes the theoretical framework and the operational definitions of the variable, which means identification of measurable indicators of the qualitative variable. The study purpose, the properties of study groups and whether the assessments are self- or observer-reported are also important factors in the choice of an instrument (3–7). The structure of the instrument should be described, for example the dimensions of the variable, the number of items and the types of item responses (6, 8). The author should motivate the use of sum scores, if present, by referring to the manual, but also be aware of the risk of invalidating the result of the measurements (4, 6, 8).

Describing data

Sums and differences of data from rating scales are inappropriate. The median level and the quartiles, or in the case of small samples, minimum and maximum (range) are appropriate measures for describing the distribution of ordinal data. Bar charts, point plots of VAS assessments, and box and whisker plots are recommended for the graphical display of the distribution of ordinal data (9). The results from assessments on multi-item scales could preferably be presented as median profiles on item levels (8). The joint frequency distribution of paired assessments could be presented in contingency tables or, in the case of VAS assessments, in scatter plots (6, 7, 10).

Agreement and association of paired data

Reliability studies concern the level of agreement in paired assessments. The percentage agreement (PA) in categories between two assessments on the same scale is a basic measure. Cohen's kappa, weighted for ordinal data, is commonly used, but the lack of comparability between studies and scales should be noted (6, 9, 11). For a comprehensive evaluation of the quality in paired ordinal assessments the Goodman–Kruskal gamma (11), the measure of monotonic of agreement (7, 12) and/or the measures of systematic (bias) and occasional disagreements proposed by Svensson (6, 10) could be consid-

 ISSN 1650–1977

ered. All of these measures take into account the non-metric properties of data from rating scales.

There is a widespread misuse of correlation in reliability studies (9). It should be emphasized that measures of correlation are not appropriate in agreement studies. The correlation coefficient measures the degree of association between two variables and reflects the strength of predictable relationship between pairs of variables; it does not measure the level of agreement or interchangeability between two assessments. A strong correlation does not indicate that two assessments produce equivalent results. Appropriate measures of association, when at least one set of data is ordinal, are the Spearman rank-order correlation coefficient (r_s) and the Kendall tau-b. Both measures should be adjusted for tied observations (9, 11).

There is also a misuse of measures that are based on the parametric correlation coefficient (r), which requires quantitative normally distributed data. The calculations of Cronbach's alpha and various reliability coefficients are based on the assumption of normality, which is not achievable in data from rating scales.

Comparisons between independent groups

In the case of comparing the categorical distributions or the median values between independent groups of data, a large number of different non-parametric tests is available, such as the Wilcoxon–Mann–Whitney test, the chi-square and the *U*-test. It should be noted that these tests have different criteria for their use (9, 11).

Analysis of change in ordinal assessments

The evaluation of change in qualitative variables is preferably performed in paired studies, in which each individual is its own control, or in matched-pair studies. The sign test and McNemar's test are appropriate for analysis of change in ordinal data (9, 11, 13). It should be stressed that the Wilcoxon signed-rank test is a non-parametric test, but nevertheless not appropriate for analysis of change in data from rating scales. This test is based on ranks of the differences between paired measurements, and since calculation of difference between ordinal ratings is not appropriate, this is an example of a non-parametric test that does not take account of the rank-invariant properties of ordinal data. For a comprehensive evaluation of change in qualitative variables, the approach by Svensson is suggested (13, 14).

Other statistical methods

The choice of more complex statistical methods of analysis or statistical modelling should be clearly motivated and the assumptions for their application must be considered. Computer-intensive statistical methods are commonly based on assumptions that are unrealistic in practice. For example, the use of factor analysis requires multivariate normally distributed variables, and the lack of uniqueness in performance implies that different approaches to the same set of data would achieve different results (15).

Ethical considerations

Altman (9) pointed out the ethical implications of inappropriate choice of statistical methods of analysis. These ethical considerations must be taken into account in the choice of rating scales and questionnaires and in the choice of statistical methods for evaluation of scale assessments. The decision-making process during diagnosis, treatment and rehabilitation is often influenced by the results of subjective assessments on scales, and so should not be undermined by an inappropriate choice of statistical methods of analysis.

REFERENCES

1. Stevens SS. On the theory of scales of measurement. Science 1946; 103: 677–680.
2. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989; 70: 308–312.
3. Hand DJ. Statistics and the theory of measurement. J R Statist Soc A. 1996; 159: 445–492.
4. Kind P. The development of health indices. In: Teeling Smith G, ed. Measuring health: a practical approach. Chichester: John Wiley & Sons 1988: pp. 23–43.
5. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Statist Med 1995; 14: 331–345.
6. Svensson E. Analysis of systematic and random differences between paired ordinal categorical data. Stockholm: Almquist & Wiksell; 1993.
7. Svensson E. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometr J 2000; 42: 417–434.
8. Svensson E. Construction of a single global scale for multi-item assessments of the same variable. Statist Med 2000: (in press).
9. Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991.
10. Svensson E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. J Epidemiol Biostatist 1998; 3: 403–409.
11. Siegel S, Castellan NJ. Non parametric statistics for the behavioural sciences. 2nd edn. Singapore: McGraw-Hill; 1988.
12. Gosman-Hedström G, Svensson E. Parallel reliability of the functional Independence Measure and the Barthel ADL index. Disabil Rehabil 2000: (in press).
13. Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Statist Med 1998; 17: 2923–2936.
14. Sonn U, Svensson E. Measures of individual and group changes in ordered categorical data: application to the ADL Staircase. Scand J Rehabil Med 1997; 29: 233–242.
15. Chatfield C, Collins AJ. Introduction to multivariate analysis. London: Chapman and Hall; 1989: pp. 82–89.