Statistical consulting, a matter of breaking tradition in applied research

Elisabeth Svensson
Department of Biostatistics, ESI
Örebro University
SE 701 82 Örebro, Sweden
Elisabeth.svensson@esi.oru.se

1. Introduction

Statistical consulting can play different roles for scientists in applied research fields. A common task for a statistician is to analyse data and to deliver tables, curves and p-values for the paper to be written. A statistical consultant can also be involved in a project at an even later stage – to interpret what the computer has done. The accessibility and reliance on statistical software imply that advanced statistical analyses and modelling can be performed by non-statisticians without enough knowledge to judge the appropriateness or to interpret the results. As consultant I have experienced disagreement concerning the statistical approaches preferred between the supervisor, the doctoral students and others involved in the same research project. There is a strong tradition in medical research concerning the choice of statistical methods, and the argument is that it is necessary for being accepted. Hence there are many different sources of misuse of statistical methods, but there is also a resistance against novel or uncommon statistical methods (Altman 1998; Svensson 2001a). One way to overcome this problem is to run research courses in applied biostatistics (Svensson 1998), another is to take advantage of the ideal learning situation of statistical consultation. This paper will present some experiences from consulting in introducing novel statistical methods and breaking traditions in applied medical research.

The development of statistical methods for rating scales

Rating scales are very common in medical research for subjective assessments of qualitative variables. Scales can also be used by experts in the judgements of complex diagnostic or outcome variables. Assessments on scales produce ordered categorical data with a discrete or continuous number of possible values. Hence, such data have an ordered structure with no other arithmetic properties, which means that the statistical methods used for their analysis differ completely from the traditional methods for quantitative variables (Hand 1996; Svensson 1993). However, as the categories often are labelled with numerical scores, there is a temptation to treat such quantified data as numbers with the same arithmetic properties as quantitative data (Agresti 1990, Coste, Fermanian, Venot 1995).

With the popularity of questionnaires, there is an increased demand for statistical methods for ordinal data. My research concerns development of statistical methods that take into account the non-metric properties of the ordered categories in the evaluation of paired scale assessments. The basis of my statistical approach is an augmented ranking procedure that makes it possible to measure the systematic component of change in categories separately from individual variations (Svensson 1993). My method is useful in reliability and validity studies and for evaluation of change on both group and individual levels of explanation.

The demand for appropriate statistical methods for dependent ordinal data among applied scientists has also led to the early introduction of these methods in courses and in consultations. Therefore much experience concerning the consequences of introducing novel statistical methods to applied research is gained.

Breaking tradition in applied research

The examples of collaborative consultations concern inter- and intra-observer reliability, inter-scale comparisons when the scales have different numbers of response categories, analysis of change, global scaling, and development of instruments.

Quality assessments

The coefficient kappa is a very popular measure of inter-observer reliability in scale assessments (Cohen 1960). However, there are drawbacks with a single measure of reliability that depends on the number of categories and on the prevalence. Therefore, my method was used for analysis of the inter-observer reliability in neuroradiological diagnostic assessments on computerised tomography pictures of the brain. My statistical method was in fact improved thanks to the collaboration with the research group (Svensson, Holm 1994; Svensson, Starmark, Ekholm et al. 1996).

The statistical measures for systematic and random disagreement in paired ordered categorical assessments also turned out to be very useful for item reduction, test-retest reliability and responsiveness evaluations during development of an ADL instrument (Dahlin-Ivanoff, Sonn, Svensson 2001 and 2002).

Global scaling and sum scores

There is a common opinion in psychometric research that the precision in scale assessments increases with the number of items, and sum score of multi-items is therefore commonly used as a global index (Fayers, Machin 2000). However, each item produces ordered categorical data, and adding the categories to a sum score is not appropriate. A sum score transforms detailed qualitative information into one single value that is not univocally interpretable. My suggestion is to use item median profiles and well defined criteria for a global categorisation (Svensson 2001 b).

One example of breaking the tradition of calculating sum scores in an established 20-item instrument of assessment family function is demonstrated in a paper by Starke, Svensson 2001). By the use of within-dimensional medians and conditional paired criteria on the two main dimensions, an eight level global scale of family function was defined. However, a drawback of breaking the tradition in sociology was that the student was not allowed by the supervisor to use novel statistical methods in her doctoral thesis; so further research was performed without involvement of statisticians.

Median scores instead of sum scores were also used for assessment of coping strategies after liver transplantation and in the assessment of behaviour (Forsberg, Bäckman, Svensson 2002; Berg, Trollfors, Hugosson et al. 2002).

Development of scales

Development of diagnostic scales of well-defined criteria for the ordered categories broke the tradition of using Likert scales or visual analogue scales (VAS) in radiology, and increased the reliability and the validity of diagnostic performance (Svensson, Svensson, Hellström 2002; Svensson, Svensson, Lasson et al. 2002).

Evaluation of change

Traditionally, change in paired assessments is defined by the difference, and the non-parametric Wilcoxon sign rank test will be performed. However, calculating difference in ordered categories is not meaningful, so the sign test and the McNemar test for binary data are the appropriate tests for analysis of change in ordered categorical data as well. This means that the ordered categorical assessments are dichotomised in the number of individuals who changed towards higher and towards lower scale categories, respectively. In order to avoid loss of information, and to take account of the entire scale, a comprehensive analysis of group and individual changes was applied to several research projects. The researchers Sonn, Starmark, Lagging and Westin broke strong traditions at their departments, and they will have a strong impact on future research (Sonn, Svensson 1997; Lagging, Westin, Svensson et al. 2002;Svensson, Starmark 2002;; Westin, Lagging, Spak et al. 2002)

Inter-scale comparisons

The use of visual analogue scales (VAS) is very popular and VAS is often recommended for assessment of different variables, such as pain and quality of life. Studies have clearly shown the non-linearity of Vas and that VAS assessments are not consistent with discrete scale assessments of the same variable (Svensson 2000; Berntson, Svensson 2001)

In consulting, I have also experienced that more than one instrument could be used for assessment of the same variable. By means of my statistical approach the order consistency of different types of scales for the same variable have been evaluated (Gosman-Hedström, Svensson 2000;Claesson, Svensson 2001).

REFERENCES

Agresti, A. (1990). Categorical data analysis. New York: John Wiley & Sons.

Altman, D. G. (1998). Statistical reviewing for medical journals. Statistics in Medicine 17, 2661-2674.

Berg, S., Trollfors, B., Hugosson, S., Fernell, E. and Svensson, E. (2002). Long-term follow-up of children with bacterial meningitis with emphasis on behavioural characteristics. Eur J Pediatr 161,330-336

Berntson, L. and Svensson, E. (2001). Pain assessment in children with JCA; a matter of scale and rater. Acta Paediatrica 90,1131-1136

Claesson, L. and Svensson, E. (2001).Measures of order consistency between paired ordinal data: application to the functional independence measure and Sunnaas index of ADL. Journal of Rehab Med 33,1-8.

Cohen, J.(1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20,37-46

Coste, J., Fermanian, J., and Venot, A.(1995). Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. Statistics in Medicine, 14, 331-345.

.Dahlin-Ivanoff, S., Sonn, U. and Svensson, E (2001) Development of an ADL instrument targeting elderly persons with age-related macular degeneration. Disability and rehabilitation 23,69-79.

Dahlin-Ivanoff, S., Sonn, U. and Svensson, E.(2002). A health education program for elderly persons with visual impairment and its impact on perceived security in the performance of daily occupations. A randomised study. The American Journal of Occupational Therapy 56, 322-330.

Fayers, P.M. and Machin, D. (2000). Quality of life, assessment, analysis and interpretation. Wiley.Chichester.

Forsberg, A., Bäckman, L. and Svensson, E.(2002). Liver transplant recipients' ability to cope during the first 12 months after transplantation – a prospective study. Scandinavian J Caring Sciences 16, 345-352.

Gosman-Hedström, G., Svensson E.(2000). Parallel reliability of the Functional Independence Measure and the Barthel ADL index. Disability and Rehabilitation 22, 702-715.

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

Lagging, L.M., Westin, J., Svensson, E., Aires, N., Dhillon, A.P., Lindh, M.et al. (2002). Progression of fibrosis in untreated patients with Hepatitis C virus infection. Liver 22,136-144.

Sonn, U. and Svensson, E.(1997). Measures of individual and group changes in ordered categorical data: application to the ADL Staircase. Scand J Rehab Med 29,233-242.

Sonn, U., Törnquist, K. and Svensson, E.(1999). The ADL-Taxonomy - from individual categorical data to ordinal categorical data. Scand J of Occup. Ther 6,11-20

Starke, M. and Svensson, E.(2001). Construction of a global assessment scale of family function, using a questionnaire. Social Work in Health Care 34,131-142

Svensson, E.(1993). Analysis of systematic and random differences between paired ordinal categorical data. Almqvist & Wiksell International, Stockholm.

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

Svensson, E. (2000). Concordance between ratings using different scales for the same variable. Statistics in Medicine 19,3483-3496

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

Svensson, E.(2001b) Construction of a single global scale for multi-item assessments of the same variable. Statistics in Medicine 20,3831-3846.

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

Svensson, M.H., Svensson, E. and Hellström, M.(2002). Bowel wall visualisation at CT colonography. Acta Radiologica 43, 87-95.

Svensson. M.H., Svensson, E., Lasson, A. and Hellström, M (2002). Patient Acceptance of CT Colonography and Conventional Colonoscopy: Prospective Comparative Study in Patients with or Suspected of Having Colorectal Disease. Radiology 222, 337-345

Svensson, E., Starmark, JE., Ekholm, S., von Essen, C. and Johansson, A. (1996). Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans. Neurological Research 18, 487-494.

Svensson, E. and Starmark, JE.(2002). Evaluation of individual and group changes in social outcome after aneurysmal subarachnoid haemorrhage: a long-term follow-up study. J Rehab Med 34,251-259.

Westin, J., Lagging, L.M., Spak, F., Aires, N., Svensson, E., Lindh, M. et al. (2002). Moderate alcohol intake increases fibrosis progression in untreated patients with hepatitis C virus infection. J of Viral Hepatitis 9,235-241.