Teaching data documentation with Stata

Svend Juul

I have seen quite a few accidents like:

- Not being able to reconstruct what modifications were made to a dataset
- Not being able to reproduce an analysis
- Not discovering errors due to lack of error-checking
- Mistakes about what a numerical code represents

Such experiences led to a two-day course in data documentation; first time in 1998, using SPSS; since 2002 using Stata. The main target group is Ph.D. students in the health sciences. Most of these students are not very sophisticated concerning statistics and computing; they have other things in their minds. However, most of their projects involve collection of own data, and safe handling of these is of crucial importance.

A key concept is the *audit trail*: As a bookkeeper you must be able to go back from the final balance sheet to the individual vouchers. This is necessary to identify and correct your own errors, and it is a request for audit. The same principle applies when working with research data.

The main aim of the course is to help the student researcher handle his/her own data in a consistent and safe way, thus preventing errors, mistakes, and loss of data. It starts with advice on methods and safeguards when entering data and ends with the student archiving data and documentation after analysis of a moderately complex dataset. The quality of the archive is assessed and a feedback is given.

In the course I use the booklet *Take good care of your data*. You may download it and the data used for the course from http://www.folkesundhed.au.dk/uddannelse/software. The concepts also had a strong influence on the contents and structure of *Introduction to Stata for Health Researchers* (S. Juul, Stata Press, 2006).

The course gets good evaluations (mostly); frequent comments are: "I should have taken this course a year ago" and "This course should be compulsory". About 80% pass the 'driving test' (the assessment of the quality of the students' archives); about 20% fail. There are no sanctions against those who fail, but I tell them that they are a hazard to their own data. Often they redo the exercise and ask for another assessment – and most pass the second time.

The presentation will give more detail about the teaching experience. Also, useful Stata tools will be presented. My main reason for giving the presentation is the feeling that these are important, but frequently neglected issues.

Svend Juul
Tel. +45 8942 6090
Email: sj@soci.au.dk

Department of Epidemiology
Institute of Public Health
Vennelyst Boulevard 6
DK-8000 Aarhus C, Denmark