Statistical Notes VII: Managing Data

Adrian Cook, Aziz Sheikh

Introduction

In this, the last in a series of seven papers on the statistical analysis of medical data, we describe the steps needed in order to transform data that has been collected into a form that is suitable for analysis. We consider issues of how best to organise data, the usefulness of database software and the choice of statistical software. We have also included practical tips that we hope will ease this task since it need not be an onerous one

Organising Data

Data is usually arranged in a grid, where each row represents an individual and each column a variable. Having reached the analysis stage, it is seldom necessary to know an individual patient's name or address so these need not be entered, and indeed constitute a confidentiality risk if they are. Instead, a unique identifier should be used in computer files containing data, and elsewhere a list that maps identifiers to individuals should be stored securely.

The unique identifier may simply be the row number or it may contain more information. For example, in a study where several practices contribute up to 10 patients each, a three digit number might be used where the first digit indicates the practice and the next two indicate a patient from that practice. For large studies, data is more manageable if stored in a number of smaller files instead of one large one. The unique identifier can then be used to link the different files together. Statistical packages readily merge data from separate files, although records in each file usually need to be sorted in order of the identifier

Data Entry

Data can either be entered using a database package or entered directly into a statistical package. For large studies the use of a dedicated database package such as Microsoft ACCESS is advisable, although data must then be transferred to a statistical package for analysis. One advantage of using database software is reduced errors, since erasing or altering data is made deliberately difficult. A further benefit is that data entry can be simplified by making the computer screen look just like the paper questionnaire, in ACCESS this is done in the 'Forms' window. For smaller studies data may be entered directly into the statistics package, in SPSS (originally called Statistical Package for Social Scientists) has well-developed facilities for doing so. Before entering any values the variables should be defined, this definition telling the computer what to expect and further reducing the chance of error. In ACCESS this is done in the 'Design View' window

and in SPSS by right clicking at the top of a column and choosing 'Define Variable'.

Defining variable

Questionnaire information falls into two main categories, numeric and textual (string). While either type of data may be entered onto a computer, numeric variables allow easier analysis and more reliable transfer between packages. Wherever possible textual data should therefore be coded numerically, for example, frequency of asthma exacerbations reported as daily/weekly/monthly/annually could be coded as 1/2/3/4.

Label

It is important to label coded data, to avoid confusion at a later stage. Two types of label are used, variable labels and value labels. Variable labels are simply the names given to variables, so for example 'Variable 01' becomes 'Asthma exacerbations'. Value labels relate to the specific values which the variable may take, so for asthma exacerbations 1=daily, 2=weekly and so on. Labeling data should be done in the statistics package and not the database, if used

Missing value

Missing data causes problems when transferring data if the number of items on each row of the database varies, since values will be read into the wrong column and cause major problems if undetected. The solution is to enter something in every box, using a plausible value where data is missing. For example if age is missing enter 999. The computer must then be told not to include these values: in SPSS the missing value can be included in the definition, in package without this facility the value should be recoded as missing immediately before analysis.

Choosing a statistical package

Two statistical packages are widely used in medical research: SPSS and Stata (Box 1). SPSS tends to be favoured by the less statistically-minded because it is relatively easy to use, the type of analysis is usually selected from a drop-down menu and then options are specified precisely in a box. By contrast, Stata is command-driven and thus requires the user to learn at least the basic set of commands. However, for anyone planning to do much analysis the effort to learn Stata will pay dividends. Once the basics have been mastered, it is easy and pleasant to use and gives a feeling of being in control, whereas SPSS is by contrast more of a 'black box' approach.

Adrian Cook
Statistician

Aziz Sheikh
NHS R&D National
Primary Care Training
Fellow

Correspondence to:
Adrian Cook
Department of Primary
Health Care and General
Practice
Faculty of Medicine
Imperial College School
& Science, Medicine
& Technology
Charing Cross Campus
St Dunstan's Road
London
W6 8R

a.d.cook@ic.ac.uk

Date submitted: 08/08/0
Date accepted: 13/09/0

Prim Care Respir 2001
10(3) 17-22

Alternatives to SPSS and Stata include:

- Epi-Info, a basic package but available free from the internet;
- SAS, which is very difficult to use but capable of handling extremely large datasets;
- Microsoft EXCEL, actually a spreadsheet but with a good data analysis module and worth considering because of its availability

Transferring data

Data transfer is much easier than it used to be, and it is now frequently possible to save data from one package directly into the format of another package. Where this is not possible a good intermediate file format is 'comma separated variables' (csv), files which can be created and read by most available packages.

It is still, however, easy for data to become corrupted during transfer because of missing values. The problem of blank fields has already been described. A further problem is that the definition of a missing value might be lost, so a value such as 999 may be included in the analysis erroneously. A good check on data once it has been transferred is to calculate a few summary statistics such as mean age, age range or the range of peak expiratory flow, if anything has gone wrong it will usually be evident from this.

Repeatable Result

It is good practice to be able to recreate the results of an analysis, enabling the work to be readily checked if questioned. SPSS is often run interactively by clicking on the drop-down menus and windows, but it is hard to go back and remember exactly what was done. The answer is to use the syntax window, where the actual commands used by SPSS are visible. To use this facility one selects an analysis in the usual way, but before clicking 'OK' click 'Paste', the syntax appears in a separate window and can be saved to a file. On returning to the analysis this file can be opened and run to repeat the analysis exactly, and with a little experience one can learn to modify the commands if required.

Box 1: Statistical software information available on the world wide web

- SPSS, SPSS Inc. Illinois USA. <http://www.spss.co>
- Stata statistical software, Stata Corporation, Texas, USA. <http://www.stata.co>
- SAS, SAS Institute Inc, North Carolina, USA. <http://www.sas.co>
- Epi-Info, EPI Info 2000. <http://www.cdc.gov/epiinfo>

Being able to repeat analyses in Stata is a little easier

The commands used simply need to be stored in a text file with a '.do' extension. On returning to Stata the command 'do filename' will use the file to rerun the commands as before, or to make changes the file can be edited and then rerun. Statistical analysis of data is frequently an iterative process, with small changes made after discussion with others, and such changes become easy when analyses are programmed to be repeatable.

Conclusion

Statistical analysis can only be done on a computer after data has been entered, and possibly transferred between packages. For the analysis to be correct these steps must be performed without error. Basic data management skills are therefore essential for anyone intending to analyse data using a computer, but these are fortunately neither difficult to learn nor time consuming to use.

Acknowledgement

We thank Dr Mark Levy for his support throughout the preparation of this series. Aziz Sheikh is supported by a NHS R&D National Primary Care Award.