Preparing your data as a spreadsheet

so that it can be read by stats package.

Introduction

You can use a spreadsheet (such as Excel) to prepare your data for statistical analysis. Many packages can open Excel spreadsheets directly. All packages can open what are called 'tab-delimited files', and all spreadsheets can save them, so once your data are in a spreadsheet they are in a format compatible with any stats package.

How to do it

To make sure that your data can be read and analysed, you need to follow a couple of simple rules in Excel:

- Just one spreadsheet

Put all your data on one worksheet; do not use multiple worksheets.

- One row = one case

Each row of the spreadsheet should be all the data from one case (usually a person).

- One variable = one column

A variable contains one item of information only. Do not put things like '120/80' into a spreadsheet. 120/80 is two things, systolic and diastolic pressure. They go in separate columns. So does parity 2+1.

- Row 1 = variable names

The first row of the spreadsheet contains the names of the variables. You can give them long names. Just make sure that each variable has a unique and informative name. Keep the names of the variables in the first row only - do not use the second row as some sort of continuation. The second row of the spreadsheet should contain data on the first case.

- Cases and controls, or different groups of subjects

If you have, for example, cases and controls, put them ALL in the same spreadsheet. Use a column to record which are cases and which are controls. Do not put them in different places. If you have different groups of subjects, the same thing applies. Use one column to identify the group that a subject belongs to.

- ID numbers

Each case should have a unique identifier, so that errors in the spreadsheet can be traced and corrected. This can be a chart number, or you can make up a study number, but you should have some identifier.

- Pair numbers (matched data only)

If your study had matched pairs of cases and controls, each pair should have an identifying number. This is different from the unique identifier for each case. It tells the statistics package which case and control make up a matched pair. So if case 101 and control 201 are the first pair, then each would have the same number in this column.

- No text in columns that contain numbers

If a variable contains numbers, do not put extra text into it. For instance, do not write 'mm' to indicate millimetres.

- Missing data

If information is missing just leave the cell blank. Do not put in things like 'NA' or 'Unknown'. If you need to record why the data were missing, put this in a column called something like 'reason why missing'

- Writing numbers

Do not put commas into large numbers; write 2500, and not 2,500.
And make sure you don't use the letters O and l when you should have 0 and 1

- If you are recording the presence or absence of something, use 0 (zero) for absent and 1 for present. It makes doing statistics easier.

- Be careful and consistent with spelling. If you spell something in several different ways, it will emerge as different things. A computer thinks that 'Major depressive disorder' is different from 'Major Depressive disorder' and 'Major Depressive Disorder'.

Special types of data

- If some values are not known exactly (for instance because the lab test reported something like '< 1 part per million') you will need special statistics to analyse the data, and the data will have to be in a specific format suited to these statistics. The format is too fiddly to describe in general guidelines. If you have data that are not known exactly, ask me and I will send you specific instructions for formatting them.

- If you have data from a survival study, this needs to be in a specific format for analysis. Ask for specific instructions.

Which version of Excel?

It makes no difference.

What about SPSS?

You can use SPSS to enter your data. This allows you to specify missing values, attach labels to numeric values and several other useful features to speed up data entry. You can send me the SPSS file (.sav).

What about other stats packages?

I can read data from just about any statistical package, but check with me before you send it. It may be easier to save the data as tab-delimited text and send that instead of the datafile created by the stats package.

If all else fails -

Finally, if you are in doubt about how to put your data into a spreadsheet, ask. You can always put in a couple of cases and send them to me to check that everything makes sense.