EXCEL DATASETS AND STATISTICAL ANALYSIS PACKAGES

While Excel provides some analysis tools, researchers often discover that they need to read Excel worksheets into more complete analysis packages. This FAQ offers advice on how to set up worksheets in Excel so they can be read easily by the analysis tools commonly used at Stanford.

SETTING UP YOUR SPREADSHEETS

1) Organize the data so that the experimental units (i.e., people/animals) are listed as the rows.

2) Use a column heading in the first row of the worksheet. The naming conventions used by different analysis packages are not consistent. The following guidelines should allow your column headings to be read by any of them:
   a. Do not use any symbols other than letters or numbers → no $, , . , ! ( ) > _ - + = ' " '
   b. Do not include spaces in the names.
   c. Start the names with a letter of the alphabet.
   d. Be aware that some popular analysis tools pay attention to capitalization so you will have to remember the capitalization patterns.
   e. Remove unnecessary words like prepositions. “Date birth” is as informative as “date of birth”.
   f. If you are tempted to use several words in the title, type in “camel case” instead of uppercase or lowercase. That is, the first letter of each word is capitalized (like the humps in the middle of a camel). For example type date of birth as dateBirth.
   g. Use short (easy to type) meaningful descriptors:
      Bad: “S “is not meaningful.
      Good: “Sex”, “gender” or “isMale”

3) Make a column for each distinct type of information.
   a. You typically want to have a single column in the spreadsheet corresponding to each answer on a questionnaire.
   b. However, if a question can take on multiple choice values, one of which is “other” and there is a fill in the blank for it (other), the spreadsheet should have two columns. One holds the multiple choice answer and the other holds the text for the other category.
   c. If you have a compound answer to a single question, like a systolic and a diastolic blood pressure, the two answers should be stored in two columns.

4) Put only one type of unit in each column. For example, if you have a column for children’s ages, do not type the ages in days for one child, weeks for another and years for yet another.
5) Do not include extra type in the body of the sheet. For example, do not add a prefix like “sys:” for the systolic blood pressure values.

OTHER USEFUL TIPS FOR IMPORTING

1) At the time of writing (Aug 2008), no major statistical package can read in Excel 2007 or Excel 2008 files. So, save your files as Excel 2003 or Excel 2004 file formats.

2) Be extremely cautious of putting mixtures of letters and numbers in the same column. Analysis packages and databases distinguish between columns holding letters and numbers.

   If you have a column that begins with many rows containing numbers, and somewhere down the column you have a character of the alphabet (typically a typo), or a special symbol like > or <, the value will be set to missing and you will not be warned.

   By default, analysis packages only check the first 8 records to figure out if a column has letters or numbers, resulting in missing values following import of the dataset for those values that are of a different type than those in the first 8 records. This major problem can be avoided by tweaking your windows registry or by using a dummy record (see below).

3) Use a dummy record as the first row below the column headings. Within the dummy record, type a number in those columns that hold numbers. If the column can contain letters, type a placeholder with as many letters as the longest word (or phrase) that can go in the column. For example, if you have a column that holds the diagnosis and the longest diagnostic phrase is 10 letters wide, type x234567890 as the value on the dummy record. That will assure that the analysis package will set aside enough room.
IMPORTING EXCEL FILES INTO SAS WITH CODE

Set your column types to be of type “mixed”. You can use this to spot cases where you have a record with characters in the middle of column that should hold numbers.

```
proc import out = work.stuff datafile = "C:\projects\blah.xls";
mixed = yes;
sheet = "Sheet1";
run;
```

IMPORTING EXCEL FILES INTO R

1. Install PERL on your computer then use the read.xls function in the gdata package.

2. run library(gdata)

3. run stuff = read.xls("C:\\Projects\\blah.xls")

IMPORTING WITH R ON WINDOWS

If you work in Wndows, the point and click “R Commander” package has a menu choice for importing Excel

Type library(Rcmdr) then look at the Data menu for Import Data… item.