UCDAVIS HEALTH | Clinical and Translational Science Center

Guidance for Database Development for Efficient Import to Statistical Software

The CTSC Biostatistics Core primarily uses R and SAS to for statistical analyses. Researchers in some fields (e.g., psychology) may use other software packages, such as SPSS. Most commonly, data is imported into these programs as a comma delimited text file, but REDCap can to some extent export data directly to data formats used in SAS and SPSS. These packages have different albeit some common requirements and restrictions with respect to data structures. Here we provide guidelines to facilitate easy and accurate importation of databases into statistical software.

Variable Names

- Make variable names (usually in the 1st row) as one full word without any space. Fields names are limited to 32 characters or less in SAS.
- Do not use punctuation such as /, -, $, ?, etc, in a variable name. The only permissible punctuation for use in variable names is an underscore (e.g., last_visit).
- Do not start with a number – for example, A1 is a valid variable name but not 1A.

Data Entries

- Each column needs to have either all numeric data or all character data.
- Do NOT use ’ ‘ or ” or # or, in text fields or as codes.
- When possible, enter dichotomous variables as numeric with 1=yes and 0=no.
- Variable values should be one word or string of numbers with no spaces or punctuation.
- Categorical variables can be represented by numeric values (1=yes, 0=no), character values (Y = yes, N=no) or text (yes, no). If using text, use one word (e.g., “white” versus “white or Caucasian) and no punctuation.
- For numeric data, it is preferable to retain the data as it was collected and not round or truncate values.
- For character data, make sure there are no leading or trailing spaces to entries.

Subject ID

- Indicate each subject with a unique identification number and use it consistently across data tables and for all observations/measurements of the same subject. Numeric ID is preferred, e.g., 11343 instead of ID11343.
- DO NOT include any personal information in data exported for statistical analysis. HIPPA regulation requires statisticians to not know any of this information. Personal identifiers include but are not limited to name, address, phone number, SSN, medical record number, autopsy number.

September 19, 2014

Missing Data

- Use consistent and unambiguous entries to indicate missing values.

For numerical data

- Missing data can be indicated by . (i.e., dot) or something like 9999. DO NOT use 0 or some other plausible value, and do not just leave a field blank. Be consistent for all numeric fields.

For character data

- Missing data can be indicated by any reasonable word or character string, such as NA or Missing. Be consistent for all character fields.
- Use upper case and lower case consistently. Some statistical packages do not distinguish by case, but others do.