Ascension Wisconsin Research Integrity and Protection

Preparing Research Data in a Spreadsheet

Accurate data entry is critical for the success of the study. Using a spreadsheet (Excel, Google sheets, etc.) is a common method used to collect and store research data. It is easy to learn and use, and researchers can use simple statistical and plotting functions to help gain insight into their data.

However, most projects require more extensive statistical techniques that require the use of additional statistical software packages such as SAS or SPSS. Researchers must appropriately set up their data collection tools so that data can be easily and accurately imported into a statistical software package.

The guidance below outlines tips and examples for data entry into spreadsheets.

Variable Names

Variable names should be compatible with the statistical program. In general, variable names should:

- Be entered in the Column heading in Row 1 (do not extend to multiple rows)
- Begin with a letter
- Contain allowable characters including any letter, digit, a period, or the symbols @, #, _, or $
- Do not contain blanks or other special characters (for example, !, ?, ', and *)
- Cannot end with a period or an underscore
- Be easily recognizable and short (<=16 characters is best, but cannot exceed 64 characters)
- Be unique, no duplication allowed
- Not use reserved keywords. SPSS keywords include: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH
- Consider case-sensitivity (SPSS is not case-sensitive but other programs may be). Capital letters can be used for variable names followed by numeric.

Entering Data

Below are tips for setting up your spreadsheet and entering your data.

- Each row should represent one (1) subject or observation.
- Avoid blank rows.
- Use a 4-digit year format (i.e. mm/dd/yyyy). Be sure the date and time format used is recognized by the statistical program used.
- Mathematical functions (i.e. means, sums) can be performed later in SPSS, don't include as a column.
- Use a separate column for each piece of information.
  For example:

DO THIS

SystolicBP	DiastolicBP	Option1	Option2	Option3
120	80	1	2	3

NOT THIS

BloodPressure	Option
120/80	1,2,3

- Avoid mixing numeric (only numbers are suitable for numeric calculations) and string information (letters, numbers and other characters that you can't use in calculations) in the same data column. If the information is necessary, include a separate column to record that information.

- Use numbers, not words, to record data.
  For example:

DO THIS

Treated
2
1
1

NOT THIS

Treated
control
treated
treated

- Don't use formatting or features like hidden columns, colors, italics, bold, etc. or add a graph to the spreadsheet that will be used for data analysis (you can use these for yourself or during data collection).
- If you have missing data, do not leave that cell blank. Define a missing value code for numbers and letters and place that code in any cell that contains missing data; be sure it cannot be confused with a "real" data value. Computer programs handle missing data more effectively through the use of commands rather than precoding.
- Using numeric averaging, deletion, or fill methods should be consistent with review of study methodology and data Real Statistics Resource Packs.

General Tips and Best Practice

- Be consistent with your data entry. This especially important when entering information such as dates and/or time which will need to be formatted in the statistical program.
- If you are working with a statistician, data analyst, or programmer, make sure to ask them about the set up before beginning.
- Develop your research/study question(s) plan using the study variables to outline your proposed statistical tests. This promotes a fluid way of learning the process of analysis and define a method to express the results from descriptive to outcome findings.
- Document your database with a data dictionary and/or codebook (Manual of Operations) that outlines what your variables are and what they mean. This will help you and other users (i.e. statistician, programmer, data manager) better understand your data and database.
  The data dictionary should include the variable names, data type that corresponds to the variable, a label or longer name that describes the variable including the units it is measured in, the codes for any categorical variables, and any notes for the variable. This can be a separate worksheet or document file.
  For example:

DATA DICTIONARY

Variable Name	Data Type	Description	Value Codes	Missing Code
ID	String	Patient ID	na	00000
Age	Numeric	Age (years)	na	-9
Sex	Numeric	Gender	1=Male 2=Female	-9
Height	Numeric	Height (cm)	na	-9
Weight	Numeric	Weight (kg)	na	-9
Treated	Numeric	Treatment Group	1=control 2=treated	-9
TestDate	Date	Test Date	na	None