University of Idaho

COLLEGE OF AGRICULTURAL AND
LIFE SCIENCES

Statistical Programs
875 Perimeter Drive MS 2337
Moscow ID 83844-2337

jpiaskowski@uidaho.edu
uidaho.edu/cals/statistical-programs

February 29, 2024

Minimum Standards for Data Sets Submitted to Statistical Programs

1. Data should be free of gross errors.
2. Missing observation should be identified with a unique symbol; NA works for R and blank cells work in SAS.
3. Each observation (row) should be uniquely identified; this means a plot or subject ID.
4. File and sheet names should be concise, consistent, and short and begin with a letter.
5. Column (variable) names should be concise, consistent, and short and begin with a letter.
6. Group (treatment) names should be concise, consistent, and short and begin with a letter.
7. Identify questionable observations.
8. Create a DICTIONARY for your experiment, where each variable name used is documented. Units of measurement should be part of this dictionary and not of the variable name.
9. Use a consistent format for all dates, preferably with the standard format YYYY-MM-DD, for example, 2024-01-01.
10. Be careful about extra spaces within cells. A blank cell is different than a cell that contains a single space. E.g. “male” is different from “ male ”.

EXPLANATION AND EXAMPLES:

1. All data should be free from gross errors, e.g., typos, values that are impossible. Calculating min and max values for each column of data will identify those extreme values. Before submitting data for analysis, please double check your data before submitting it by looking at summary of each variable in the dataset. (e.g. use `summary(data)` in R or PROC UNIVARIATE and PROC FREQ).
2. Use a unique identifier to indicate missing observations. Code true zeroes as zero, not missing. R generally treats NA as a missing observation. Don’t use NA for data to be analyzed in SAS; you can either leave the cell empty (do not use a period) or enter a number that is clear out of range and biologically impossible. “-9” is a popular choice. The latter is preferable when converting files to a flat file format such as txt, prin, or csv.
3. Each observation (row) should be identified by a plot number or subject ID. Many experiments are of repeated measures nature and require a unique subject ID to

identify the experimental unit on which repeated observations are taken. Plot numbers are much easier to handle than subject IDs concatenated from Loc, Block, and Treatment information. Example: Plot = 1101, where digit 1 = location, 2 = block and 3&4 = plot is much easier than the concatenated Subject_ID = 'Everglade REC Block by the transformer next to the Canal Nrate =100".

4. Consistent and short file or sheet names makes reading data into the stats program of your choice much easier. Example: Let's assume you are taking data every month for two years (2017 and 2018). If you named the sheet as 3-letter month plus year as in Jan_17, Feb_17, Nov_18, Dec_18 you would have a concise and consistent naming system.
5. If your data consists of multiple sheets but similar or identical variables, ensure that the variable (column) names are consistent. For example, you can use any of these designations for replication unit: 'Rep', 'rep', 'REPLICATION', 'Replication'. But, please only use one of those.
6. All group (e.g., treatment) designations should be consistent. Example: Treatment "abcd" is not the same as treatment "AbCd" or "ABCD" due to differences in case. Use the filter function in EXCEL to check for consistency.
7. For questionable observations, do not highlight the row/observation, instead add a column with an indicator for questionable data. Highlighted cells and colors can only be seen in spreadsheet software and generally cannot be translated to statistical software.
8. Please see our example file "Data_sample.xlsx" for how to structure a data set and accompanying data dictionary.
9. You can use the format function in Excel to check and set specific data formats.
10. All computer programs deal with whitespace (leading, trailing or between characters) differently. It is best to avoid any white space in variable levels.