Tips for Researchers: Setting Up Your Data File

Or, some data (and data file) situations that can cause problems and delays in statistical analysis….and how to correct them…..and facilitate stat analysis

UCLA School of Nursing Office of Research & Innovation
Statistical Support Core
11/21/23

Problematic data situation	**How to minimize problems and facilitate later analysis**
Variable names, labels, attributes	
1. a critical identifying variable given different names in data files within a study, e.g. subject identifier variable named *id* in file containing intervention group records and *id1* in file containing control group records	Be consistent in variable names across data files that will be merged for analysis
2. complex names for subject identifier variables (e.g., *studysubjectidentifier*)	Make ID variable numeric and simple (e.g.*id* or *subj_id*). If study has more than one time point and identifiers are complex, the probability of mistakenly changing IDs is high, making merging of different time point data problematic
3. no subject identifier variable	Subject (or case) identifier becomes extremely important if multiple data files need to be merged for analysis (e.g. baseline data, 1st followup, 2nd followup; or instrument 1, instrument 2, instrument 3)
4. complex/long variable names (e.g. *subject_hiv_aids_status* or *Beck_depression_scale_question_22* or indecipherable abbreviations such as *axtbd42mm27fqrs)*	Use relatively short variable names (*HIVstatus* or *HIV*; or for items, *HIV_q1, HIV_q2* etc.). Longer, more explicit labeling can be given as variable labels and in codebooks. Statisticians have to type these variable names when creating analysis programs—make it easier for them.
5. letters that can be confused with each other in variable names—i l l 1, O 0	Give variable names that are as clear as possible. OK to use these potentially confusing letters/numbers if used sparingly and consistently (e.g. using "i" consistently throughout items in an instrument as an abbreviation for "item" [*msas_i1*, *msas_ i2*, *msas_ i3*, etc.)
6. variable names that contain blanks or special characters or start with a number	Most statistical packages don't accept variable names that begin with a number, contain blanks, or contain certain special characters. Some tips: start each variable name with a letter; use an underscore "_" instead of a blank to separate parts of a variable name; don't use other special characters (e.g. $, %, @, etc)
7. different names and labels for the same variable at different time points	When separate files are prepared for each time point (and files merged later for analysis): if a question is named *A12* at baseline, make sure it is given the same name in all of the follow-up data files to avoid combining wrong variables together while merging data. Note that if all time points are entered in the same file ("wide" format), then keep the stem of

	the variable name the same across time points and add a time indicator into the variable name (e.g. *BDI3_t0, BDI3_t1, BDI3_t2* for item 3 on BDI at times 0, 1, and 2). See also tip 24 below.
8. multiple variables given the same label	Give each variable a unique label. If the same question is asked about different things, e.g. frequency of medication taken for a set of several different medications, label each one with the medication in question. Instead of labeling "how often do you take this medication" repeatedly, label them as, "how often do you take med1", "how often do you take med2" or "med1_how often?", "med2_how often?" etc.
9. overly complex variable attributes given (e.g. too many decimal places)	When creating datasets, make sure that variables with only integer codes (1, 2, 3 etc.) are not assigned decimal places. Statistical report/outputs look much cleaner and easy to read. (14 is much easier to read than 14.000)
10. a variable given different attributes in different data files that will eventually be merged, e.g. *subject_id* specified as numeric in the baseline file but character in the 3-month followup file	Be consistent with variable attributes across data files that will be merged for analysis
Coding, recoding, documentation	
11. use of special characters or letters in otherwise numeric coding	Most statistical analysis programs want to see numerically-coded variables for most analyses. When you enter special characters (e.g. "<5" or "M" or "N/A") as data responses along with numeric responses (e.g. 1, 2, 3), this automatically creates a character variable, which then would have to be translated to numeric for analysis; and the special codes would have to be recoded. OK to have open-ended questions as character variables—you will want to see these answers and you can categorize them later (into numeric coded categories) for analysis.
12. inconsistent coding within a variable	Be consistent in the entering the data. For example using "M" and "-9" to indicate missing values within the same variable or "Female" and "F" in the same variable will require recoding prior to analysis.
13. inconsistent coding across variables that have the same response categories, e.g. 0=no/1=yes for some variables and 0=yes/1=no for other variables	Be as consistent as possible in coding. Note, however, that sometimes inconsistency is unavoidable when you're using already-developed instruments that have their own coding—in this case, use the codes as given on the instrument and alert the statistician to the situation.
14. inconsistent coding of subject identifiers across various study data files	The identifier variable should have consistent codes or values for the same subject across files. E.g., IDs coded as *1* in the baseline file and *id01* in the post-intervention file and *s01* in the follow-up file (all for the same subject) will have to be recoded prior to merging them together.

15. inconsistent categories for categorical variables across observation points	The same categorical variables should have the same number of categories for each time point. If variable *b13* has 3 categories at baseline, it should not have 4 or 2 categories at any follow-up.
16. inconsistent entry (or coding) of dates, e.g. 1-26-2014 and Jan 26, 2014	Enter dates in a consistent format—in some data entry software, a column can be specified as a date variable in a certain format.
17. inadequate instructions for skip patterns to respondents and/or inadequate coding and documentation of skip patterns	Make sure skip patterns are appropriately labeled and highlighted to reduce mistakes. It is common to find some respondents who are not eligible (because of a previous answer) to answer a question responding to this question, and not following the skip pattern. If there are skip patterns, be sure to document clearly so that statistician will know, and specify what to do if respondents answer a question which he/she shouldn't have.
18. many skip patterns in the data collection instrument	Usually having too many skip patterns substantially increases the chance of the respondent missing questions or of having incorrectly administered questionnaires. Use skip patterns judiciously to reduce subject burden but not to make the questionnaire too complicated. Be sure to include special missing data codes for legitimately skipped items.
19. reversing the values of one or more variables to conform to scale, without proper documentation	No variable values should be changed in the main/original data set, even if reversing is required to create a scale or to conform to other requirements. Such actions should be left for later data sets created for analyzing data and clearly documented to avoid confusion and analysis error. In other words, the original data set should be just that – the exact responses of the respondents without any data manipulation.
20. calculating mean scores of scale variables as a single calculation (e.g., (V1+V2+V3) /3) without explicitly considering missing items	In such cases, the mean may be incorrect or the result may be a missing value if there is a missing value for any variables. Use statistical commands (e.g., Mean, Sum, etc.) to calculate these summary scores in a statistical package. Let statisticians know if you have scale variables, and whether they may have missing values. In many occasions, statisticians must double check these scale variables to see whether or not the composite scores/mean scores were calculated correctly. Also, check with scoring instructions to see if there is a minimum number of items that must be non-missing in order for a score to be calculated.
Other data file construction and data entry issues	
21. typing or coding mistakes	Typos and coding mistakes happen. Carefully check data before giving to statistician

22. duplicate records in database for same subject (or 2 not-quite-duplicate records for same subject, e.g. if data were corrected or if some measures were added to the data file at a different time from others)	Check for this before giving data file to statistician.
23. recording separate pieces of information as one variable (that is, entering multiple pieces of information in the same column), e.g. *130/70* for the 2 data elements systolic and diastolic blood pressure or *150/24.8* for the 2 data elements weight and bmi	If you have related but distinct measures of interest (e.g. systolic and diastolic blood pressure or weight and body mass index), create two separate variables (*systolicBP* and *diastolicBP* or *wt* and *bmi*) instead of entering them as a single variable.
Other helpful hints	
24. data file format	When same measures are collected at multiple time periods, it's usually easiest to enter data in "long" format with a record for each person for each time period. In this case, also create a variable *time* that records when a measurement was taken (e.g. *0, 2, 4 etc* for baseline, 2-months, 4-months).
25. data management process	Include your statistician when you are designing your data file format, coding, and data entry procedures. It will help reduce problematic data situations that may delay data analysis.

Links to other useful resources:

https://www.sussex.ac.uk/library/researchdatamanagement/organise/namingandorganisingfiles/
https://datamanagement.hms.harvard.edu/plan-design/file-naming-conventions
naming & organizing data files

https://www.data.cam.ac.uk/data-management-guide/organising-your-data
https://www.griffith.edu.au/__data/assets/pdf_file/0025/1233907/20210107-Guide-to-managing-research-data.pdf
managing data tiles

https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/tips-for-creating-an-excel-file-that-can-be-easily-moved-to-a-statistical-program-for-analysis/
formatting your excel data file for later statistical analysis

http://www.princeton.edu/~otorres/DataPrep101.pdf
data preparation

https://stats.oarc.ucla.edu/spss/
how to use SPSS—including setting up data files (click on "Data Management" on the SPSS page)