Maximizing Statistical Interactions Part II: Database Issues

Provided by: The Biostatistics Collaboration Center (BCC) at Northwestern University

While your data tables or spreadsheets may look good to you, it is important to note that computers and people read data differently. This handout is a summary of some general principles and specific hints that may be helpful if you are collecting and/or organizing data and preparing to share it with a statistician for analysis.

1. Deciding what to measure

   An early part of the investigation is deciding exactly what you want to know about each person (or other unit). It is often helpful to organize your variables into logical groups, such as demographic, medical history, current clinical, current laboratory, and so on. Maybe you are collecting responses to a standard set of questions. Data collected by different people or at different times should be separated. If you are asking someone to help you organize your data, this kind of information will be very helpful.

2. Deciding how to measure it.

   You may remember from your first statistics course that there are several different types of variables (nominal, ordinal, interval, ratio…). The good news is that the computer recognizes fewer "types" of data:

   - string = words – what you see is exactly what you get. These are usually discouraged, as any deviations (e.g., capital letters vs. lowercase letters, extra spaces, etc.) are seen as a "new" possible outcome.
   - Numeric = numbers (doesn't matter if this is "code" or actual values)
   - Specific formats. Certain data, like dates have formats that accompany them. Excel has many more "potential" formats (for instance, % or $). In general, we discourage the use of any specific format other than dates, which should be consistent (e.g., mm/dd/yy).

3. Identifying people

   A. We recommend using some unique subject identification number to identify people. Numeric IDs are preferable to names or initials for many reasons (apart from Health Insurance Portability and Accountability Act (HIPAA) Privacy rule requirements).

   - Names are stored as string. They are often misspelled and sometimes upper and lower cases are used. Middle initials come and go, and sometimes names change. Names are not necessarily unique.
   - Numbers are easier to use sorting. For large datasets, check digits can be built into IDs so that many typographical errors can be caught.
   - If names or medical record numbers (MRN) are known (and if MRN is not used as the numeric ID), then a master list matching names or MRNs with the numeric ID should be kept separately from the data. **Avoid sending the BCC MRNs or other identifying information.**
   - If the data is collected via paper records (or data collection sheets), whether or not it is identifiable – the numeric ID should be written on the collection sheet as well as included in

the dataset. This will be useful if any data needs to be “checked” (i.e., a data value that is not valid).

B. Please do not use the Excel row number to identify patients. Data can be sorted by different variables, so there needs to be a consistent “link” to identify individuals with records, or other potentially data sources. A unique, consistent ID is the best way to accomplish this.

4. Entering your data into a worksheet (or multiple worksheets).
Whether you are entering your data into Excel or some statistical package (like SPSS or Stata), please do not enter your data into a table in word. Word tables do NOT transition into statistical packages very well at all. Whether your data is structured vertically or horizontally (more on this later), there are basic rules to which we would like you to adhere in setting up your data.

A. In general, statistical packages like **numeric data** rather than strings. Data that uses strings need to be “cleaned” and the result would eventually be numeric coding. For example:

- Sex can be coded numerically as *0* or *1*, or as a string *female* or *male*. Most packages will distinguish between *Male* and *male* and *M* and *m*.
- Grade or stage of disease, frequently these (for medical purposes) are regarded as 1, 2, 2+, 3, 3+… and so on. Computers see the “+” and will consider these text. Numeric coding can solve this problem; you can code your data as 1, 2, 2.5, 3, 3.5,… to keep the ordinal nature intact.
- For variables that have mutually exclusive and exhaustive categories (like disease stage – where a study subject has one and only one value), have one variable with the codes. Do not include several variables that are Yes/No indicating which stage the subject is.
- If you have variables that can assume multiple values at the same time (like symptoms, see D. below), then it is reasonable to have several variables.

INCORRECT!						
ID	sex	stage	stage1	stage2	stage3	Stage+
John	m	2	No	yes	No	No
Mary	f	1	Yes	no	No	No
Pete	Male	2+	No	yes	No	Yes
Dave	male	1	Yes	no	No	No
Sue	Female	3	No	No	yes	No

CORRECT!		
ID	sex	Stage
1	1	2
2	0	1
3	1	2.5
4	1	1
5	0	3

Codebook:
Sex: 1 = male, 0 = female
Stage: 1 = stage1, 1.5 = 1+, 2 = stage 2, 2.5 = 2+, etc through stage 4+, -9 = unknown

B. (Naturally following A) Consider **a data “codebook”** - or data dictionary that is <u>separate from your data</u> – it may be in a separate sheet of a workbook, or it may be a separate file (like word or even a completely separate excel file). Here you will include each variable name, and any longer name that may better describe the variable, units in which the variable is measured (if appropriate) and the corresponding “codes” for any nominal (named, like gender) or ordinal data that you have coded numerically. Additionally, you can include valid ranges for your data (i.e., what numbers are reasonable and what would be considered "out of range"), or how a variable was calculated if it was calculated from your other data (e.g., BMI).

Please resist including any of that information in the same sheet as your data (say, on lines 1-3, or at the end) – again, the computer will not be able to understand. Your data will require "cleaning."

		COLUMNS ↓ A	↓ B	↓ C
ROWS	1 ->	(cell A1)	(cell B1)	
	2 ->	(cell A2)		
	3 ->			

C. In an Excel spreadsheet, rows are horizontal and are indexed by numbers (1,2,3,…). Columns are vertical and are indexed by letters (A, B, C,…). Cells are indexed by a letter-number combination (e.g., A2). The first row of your spreadsheet should have your **VARIABLE NAMES**. Previously there were many restrictions on what variable names could be, but over time, statistical packages have relaxed many of them.
   - They do not need to completely make sense, but in some way, consider them like typical usernames for websites.
   - They need to start with a letter.
   - No spaces should be in them.
   - Minimize the number of characters (try for length ≤8 characters).
   - You may include numbers, but avoid symbols (e.g., don't use: #, &, %, or -).
   - Different variables should have different names. If weight was collected twice, don't call both variables *weight*, consider *weight0* and *weight1*.
   - Examples: id, sex, height0, weight0, weight1 weight2, group. Your codebook would include a more detailed description: id = unique patient identifier, sex: 1 = male, 0 = female, height0 = height at baseline (inches), weight0 = weight at baseline (lbs), weight1 = weight (lbs) at one month, weight2 = weight (lbs) at 6 months, group: 1 = diet, 2 = diet and exercise, 3 = exercise, 4 = monitor.

D. If your data include blood pressure or other variables that are similar (various side effects), it is best to consider using **separate variables** for the "parts" – for instance:
   - rather than enter 140/90 for BP, have SBP and DBP as separate variables. Again, the computer will "see" the "/" and consider the variable a string.
   - Variables with several non-mutually exclusive values, like side effects. If you actually had a variable SE (side effects) and coded it 0 = none, 1=headache, 2=abdominal pain, 3=rash… and then for a subject who reported a headache and rash, if you type in 1,3 – the computer will not be able to understand that. Rather have 3 variables (or more) named: headache, abdpain, and rash; code those variables 0 = no, and 1=yes.

E. Multiple possible values: For variables as listed at the end of part D (side effects), it may be helpful to collect a smaller **preliminary dataset**, where you have set Side effects listed, but also have other variables: “other” and **“othertext”** where you can list “0 = no and 1=yes” but also collect what other side effects were common in the population you want to study. After several subjects are collected, examining the text variable might indicate that you may need to include another “common side effect” that you did not originally consider.

F. **Dates**. Be sure that the format you use for dates is consistent (e.g., MM/DD/YYYY), and that you don’t have inaccurate dates listed (e.g., 11/31/2009 – note, there are only 30 days in November). If there are inaccurate dates, when some packages read those, the computer will see the entire set of dates as text, or they will skip over those observations. Both will cause issues when trying to analyze data.

G. **Missing data.** You may also want to consider “coding” any missing data or reasons why data are missing. Usually you can pick values that are out of the range of possible values and code missing data appropriately. For example, if a person refuses to allow a BP measurement, that SBP and DBP measure might be coded as -99, whereas if they cannot have BP measured because an adequate cuff size is not available, that might be coded as -88. Be sure to include these “codes” in your data dictionary! You can leave the cell empty, but please do not type in NA or other text (see 1A on mixing string and numeric data).

H. **Calculations.** Although it is tempting to do (since Excel has made it easy!), please use caution when calculating a variable from data that you have entered. For example if you have height0, weight0 and weight1, calculating BMI at baseline, or change in weight will be superfluous. Although Excel is “getting better” – there used to be issues when calculating and then manipulating the database that would “recalculate” what was done… and sorting messed everything up… - most statistical packages are able to calculate these for you, and can be “re done” easily if a data error is found in the original variables. Syntax can also be saved that will make calculating composite scores from raw data easier (e.g,. calculating the Physical or Mental composite scores, or any sub-scores from the SF12 or SF36).

I. **Colors and fonts:** although it is tempting and somewhat visually appealing, color-coding your data is not helpful to statistical packages. If you are tempted to do so, please make sure that any reason for color coding is also included as a separate variable.

J. **Blank Lines:** Although the dataset might not “look” great, and admittedly, it is easier to “see” groups if there are blank lines in between (that is true only if you are not a computer), many statistical packages will stop reading in data once a blank line is “seen.” Please do not include blank rows.

5. Organizing your data. There are two main ways to organize data (Horizontal or Vertical)

A. Usually (for cross-sectional data), it makes sense to organize data in a **horizontal** fashion. That is, every row will correspond to a unique study subject.

B. However, if you have paired measures, or a study with multiple visits that obtain similar information throughout the study per person, it *may* make sense to have **multiple rows per person**, indexed by pair number or visit number. Be sure to include the unique ID on all lines pertaining to the same person or group (do not leave ID blank even if it is the same as the row above). This is frequently referred to as a “vertical” (or long or tall) arrangement of data.

C. It is quite okay to have both data structures used for a single study (hence at least <u>two files</u>

comprising study data). One example might be if baseline information is extensive, and follow- ups are more focused: information such as medical history and demographics are collected once and presented in a horizontal manner, but any information that might change or that is collected at subsequent follow-ups is collected in a vertical manner). As long as a unique ID is included, these files can be merged using more advanced statistical packages.

Horizontal (wide) data format				
Id	Date1	Weight1	Date2	Weight2
1	1/20/2008	120	2/20/2008	125
2	1/30/2008	160	2/30/2008	150
3	2/6/2008	155	3/6/2008	150

Vertical (long) data format			
id	visit	date	Weight
1	1	1/20/2008	120
1	2	2/20/2008	125
2	1	1/30/2008	160
2	2	2/30/2008	150
3	1	2/6/2008	155
3	2	3/6/2008	150

6. Updating or selecting a subset of data.

   A. Please try to have the dataset as complete as possible before submitting it to a statistician for analysis. Invariably, there will be some data-cleaning on our part (reformatting, etc) before we are able to “upload it” into a statistical package before analysis. If an updated spreadsheet, or additional new or changed data is provided after that point, all data cleaning and analysis would need to be done yet again. Please if at all possible, submit a “frozen” or “locked” dataset so that any work that needs to be done on it will be done once.

   B. If additional data are needed, or added (we know this does happen), please talk to your statistician about the best way to *amend* (rather than replace!) the files that have already been analyzed or cleaned. Likely, if you have new variables that need to be added, the new variables (only) with appropriate study ID is all that needs to be provided (we can then merge this data with the old dataset by the study ID). If it is strictly appended data (new IDs), be sure all of the variable names and types match, so that we can merge by variable names.

   C. Lastly, if we provide an analysis, please do not decide “post hoc” that you want the analysis done on a subset of the data (or “exclude the following cases…”). This is part of the hypothesis testing theoretical framework that statisticians do not like shaking. Once you know one set of results, any further questions you have pertaining to a particular overall question needs to have that type I error adjusted for fear that any “sub-results” are not simply a matter of chance. Note this does not mean that we will not do any additional analyses, but rather that we wish to ensure that all analyses have statistical integrity.