Data Management in SPSS

Good data management is like changing the oil in a car: Your advisor probably pays someone else to do it. ☺

Hm. Let's try that again. Good data management is like changing the oil in a car: It might not seem worth the bother right now, but if you don't do it, you're risking a lot of trouble later. Keep in mind that most good journals require that you maintain your "raw data" for at least five years after the date of publication; this includes being able to explain to anyone who asks what values are in which columns of your data file.

Speadsheets: Rows vs. Columns

Within SPSS (as well as Excel and almost all other systems), data are usually entered and stored using what has become the *de facto* standard. Each subject gets a row in the speadsheet and all of the data from a given subject appear on one row. There are, of course, exceptions to this, including single-subject designs and the old-fashioned, random-factors format -- under which each observation gets a row -- but these are rare and will not be covered in detail in this course.

✂ Technically, what-ever is going to be the "target of generalization" defines the rows. Thus, if your goal is to generalize to the population of <u>people</u> from which your subjects were sampled, then each <u>person</u> gets a row; this is the case about 90% of the time. One specific example of a situation that does not use subjects as the target of generalization (and, therefore, does not give each subject a row in the data file) occurs in language research, where many analyses are replicated using item (i.e., word) to define the rows. This is done because these folks also want to be able to generalize from the specific words used in the experiment to all words in the language.

Each column in a spreadsheet holds one of two things: (1) the values of a *control variable*: i.e., numbers or nominal codes [that can also be numbers] that specify the value of an IV or SV, or (2) the values of a *data variable*: i.e., numbers or nominal codes that specify the value of a DV.

⊕ Note: In SPSS, you do not have to "waste" a column specifying the subject number. This is implied by row. You only need to have a column for subject number if you are using the old-fashioned, random-factors format for the file.

Given that SVs are often treated as if they were IVs (as under quasi-experimental designs), you might be wondering why I have named the two types of columns *control* and *data*, as opposed to *IV* and *DV*. The answer is that not all IVs show up as nominal codes or numbers in a single column; sometimes the levels of an IV show up as a set of related columns.

Here are the rules on this:

- If the IV corresponds to a between-subjects factor, then the level of the IV will be specified using a control variable.
- If the IV corresponds to a within-subjects factor, then there won't be a control variable for the level of this factor; instead, there will be a separate data column for each level of the factor.

Hopefully this will become more clear after a few examples.

The other reason for referring to columns that hold IVs or SVs as “control variables” is that SVs are often used as covariates, in which case they are being used to “control for” differences between the subjects.

Example #1 - one between-subjects IV (defining two groups) and one DV

Assume that you are interested in whether anxiety can be reduced by consuming chocolate. One group of five subjects is fed a Hersey’s Kiss™ and then asked to fill out a form that is condensed to provide a single, numerical value for anxiety. Let’s call this group ***E***, for experimental. The other group of five subjects is not fed any chocolate, but still fills out the anxiety questionnaire. This is group ***C***, for control. The data would be tabled using two columns. The first would be a control variable and each entry would be either an *E* or a *C*. The second would be a data variable and each entry would be a number (i.e., the subject’s anxiety score). There would be 10 rows; one for each subject.

Example #2 - one within-subjects IV (defining two conditions) and one DV

Assume that you’re still interested in the miraculous powers of chocolate, but have decided to use a within-subjects design, instead. In this case, each of 10 subjects fills out the anxiety questionnaire twice (and, therefore, has two different anxiety scores): one of these measures is taken after the subject has eaten a Kiss™; the other is taken after no chocolate. The data will again be tabled using two columns, but this time they are both data variables. One of the columns holds the with-chocolate anxiety scores; the other column holds the without-chocolate scores. To be clear: even though this experiment has an IV, the spreadsheet contains **no** control variables. This can happen in any case where all of the IVs are manipulated within-subjects.

Stop here for a moment and think about that second example some more. Is there any information missing from the spreadsheet as described?

Yes! When you use a within-subjects design, in most cases the data from the various conditions are not collected at the same time. Thus, the conditions have an order. And even if you take the “counter-balance order and then forget about it” approach (under which you don’t include order in the analysis) you really should include order information in the spreadsheet (because you might need to know this later for some reason). So there ought to be a control variable in the second spreadsheet after all: a control variable that specifies whether the with-chocolate condition was run before or after the without-chocolate condition.

Naming columns

It is strongly suggested that you get in the habit of consistently naming the columns in your SPSS data files. (This is done by double-clicking on the head of the column.) In many versions of SPSS, you are limited to 8 characters with no spaces, although you can use underscores to separate characters.

⊕ Warning: many versions of SPSS are case-insensitive. In particular, most versions convert all upper-case letters to lower-case at the moment you hit ENTER; I don’t think that there is anything you can do about this. In general, avoid using case as a way to define values. (For example, don't use **C** for *chocolate* and **c** for *control* because the critical information could be lost.)

In general, you can take one of two approaches to naming columns: ***direct labels*** and ***codes***. An example of direct labels would be "with_c" (for *with chocolate*) and "wo_c" (*without chocolate*) for the Example #2, above. An example of codes would be "data1" and "data2" for the same set of data (on the grounds that both of these are data variables). Clearly, in the long run, direct labels would be better, as you'd be much less likely to forget what each of the column contains. However, under more complicated designs (that will be analyzed using multi-way ANOVA), it is often better to use codes. This is true because you'll want to keep these columns together (even after SPSS sorts them) so that they can be moved from one zone to another as a group.

My general suggestion with regard to naming data variables goes like this: work down through the following list, stopping at the **first** set of conditions that describes your situation:

- If there is going to be only one data variable (because each subject contributes exactly one observation), then use a direct label for the data column. The label should refer to what was measured (e.g., "anxiety" for anxiety score, or "depressn" for depression score).

- If there is one within-subjects IV and one DV, then use direct labels for the columns. The labels should refer to the condition under which the measurements were taken (e.g., "with_c" for *with chocolate* or "wo_c" for *without*).

- If there is one within-subjects IV and more than one DV, then use direct labels for the columns with a prefix to identify the DV (e.g., "rt_ w_c" for *response time with chocolate*, "rt_wo_c" for *response time without chocolate*, "ac_w_c" for *accuracy with chocolate*, and "ac_wo_c" for *accuracy without*). The reason for using a prefix, as opposed to a suffix, is to keep the related columns together after SPSS sorts by alphabetical order.

- Use codes. First work out a coding scheme that assigns numbers or letters to each level of each within-subjects factor. Then set up the columns in terms of the DV and factors. For example, for a two-by-two, within-subjects design with two DVs, use "dv1_1_1" for the first DV when Factor 1 = 1 and Factor 2 = 1, and "dv1_1_2" for the first DV when Factor 1 = 1 and Factor 2 = 2, etc. Do not lose your coding scheme! (Better yet: Read the SPSS help files and learn how to use the **Labels** option when you double-click the column header.)

With regard to control variables, these should always be given a direct label. For example, a control variable that specifies sex should be named "sex" and a control variable that specifies whether the subject was given chocolate should be named "chocolat" (or something like that). The one exception to using very simple labels is when a bunch of IVs or SVs are going to be used together (as a "set") in an MRC analysis. In this case, you might want to preface each name with the same one, two, or three letters (so that SPSS doesn't list them separately when it alphabetizes). For example, if socio-economic status is operationally defined as a combination of annual income and value of home, then use "ses_incm" and "ses_home" for these two columns.

Other Attributes of a Column

It is a good habit to specify at least two other attributes of each column in the data file (by double-clicking on the header of the column). First, set the **Measurement** to nominal, ordinal, or scale (for qualitative, ordinal, or quantitative, respectively). Second, set the **Type** to numeric for ordinals or quantitatives, or to string for qualitatives. These actions will prevent you from ever treating a set of nominal codes that happen to be numbers as if the numbers should be taken seriously as numbers. This is important because you can get some completely crazy (and incorrect) results if you make this mistake.

Coding Nominal Scales

In Example #1 above, I suggested that you use ***E*** to code for "experimental group" and ***C*** to code for "control group." This may have caused you to pause while reading. *Why not use **1** for "experimental group" and **0** for "control"?* (You might even have been taught to do this before.)

This is a tricky issue. To understand my reason for using letters, we have to say a few things about the difference between ANOVA (which includes *t*-tests) and MRC (which includes bivariate correlation); we also have to introduce a few more technical terms.

First, one of the two differences between ANOVA and MRC is that ANOVA treats all IVs as if they were nominal, while MRC treats all predictors as if they were quantitative. (FYI: the only other difference is that ANOVA tests for all possible interactions between IVs by default, while MRC does not.) In other words, even if you use numbers as the values of a control variable, when that variable is used as an IV in an ANOVA, the values in the column are not taken seriously as numbers. In contrast, you cannot use a variable that contains non-numbers as a predictor in an MRC, because MRC demands that all predictors be quantitative.

Second, in technical terms, a nominal variable with *k* levels is said to have *k-1* ***aspects***. The number of aspects is the minimum number of independent differences that exist within the variable. For example, if a nominal variable takes on just two values (e.g., ***E*** and ***C***), then it has only one aspect: the difference between the two values. If a nominal variable takes on three values (e.g., ***A***, ***B***, & ***C***), then it has two aspects: the difference between ***A*** and ***B***, plus the difference between ***B*** and ***C***. (The difference between ***A*** and ***C*** is implied by the other two aspects; in other words, all aspects are *transitive*.)

Because ANOVA treats all IVs as if they were nominal, it does not matter if the control variable (for a given between-subjects factor) contains letters, letter-strings, or numbers. SPSS will automatically search through the column and identify all of the different values and then group the data by value. However, because MRC treats all predictors as if they were quantitative, you cannot (in general) use numbers to directly represent the values of a nominal scale. Instead, you must use as many columns as there are aspects and let the pattern of values across the *k-1* columns stand for the nominal value.

We will not cover coding nominals for MRC until much later, but please take my word on this: It would be better to get in the habit of avoiding using numbers for groups when all of the groups are in one column. You will save yourself from having to unlearn this later.

☞ With all that said, there is one exception to the “don’t use numbers to code nominal IVs” rule: when there are only two values on the nominal scale. In this case, both ANOVA and MRC require that you use exactly one column (that contains two different values). Although it does not make a difference, it is traditional to use the values of ***0*** and ***1*** to code the two levels/values; it is also traditional to use ***0*** for the control condition, if one exists. Thus, to return to Example #1, it would have been OK to use ***0*** and ***1***, as opposed to ***C*** and ***E***, but I’d rather if you didn’t. I’d rather that you avoid using numbers until we get to MRC.

Arranging the Data

There are myriad ways to organize a spreadsheet and as long as you are consistent, you should be OK. My suggestion is that you use chronology. Working from left to right, put the SVs in first column(s) [since they have been true about the subject for a very long time], then the codes for between-subject factors (IVs) in the next column(s) [since they were assigned before the data was collected], then codes for the order of any within-subject factors, and then the data variables last. At a minimum, I suggest that you put all of the control variables to the left of all of the data variables. When multiple control variables are used to code for a nominal scale, put all of the columns for a given variable next to each other. As to multiple DVs, there are two lines of thought. Most people group all the columns by what was measured; for example, these people put all the response-time columns together and then all of accuracy columns together. The reason for doing this is to keep the columns that will be analyzed together near to each other.

Key: be consistent.