Al-Mustaqbal University
College of Health and Medical Technologies
Lecturer Mohammed Qasim Al Azzawi
mohammed.qasim.obayes@uomus.edu.iq

Data Handling

1. Reading (opening) the data set

Data can be obtained in several formats:
• SPSS files.
• Spreadsheet - Excel, Lotus.
• Database - dbase, paradox.
• Files from other statistical programs.
• ASCII text.
• Complex database formats - Oracle, Access.

Reading SPSS data

-In SPSS, go to FILE/OPEN.
-Click on the button “Files of Type.” As in Figure 1.
-Select the option “SPSS (*.sav).”
-Click on "Open.”

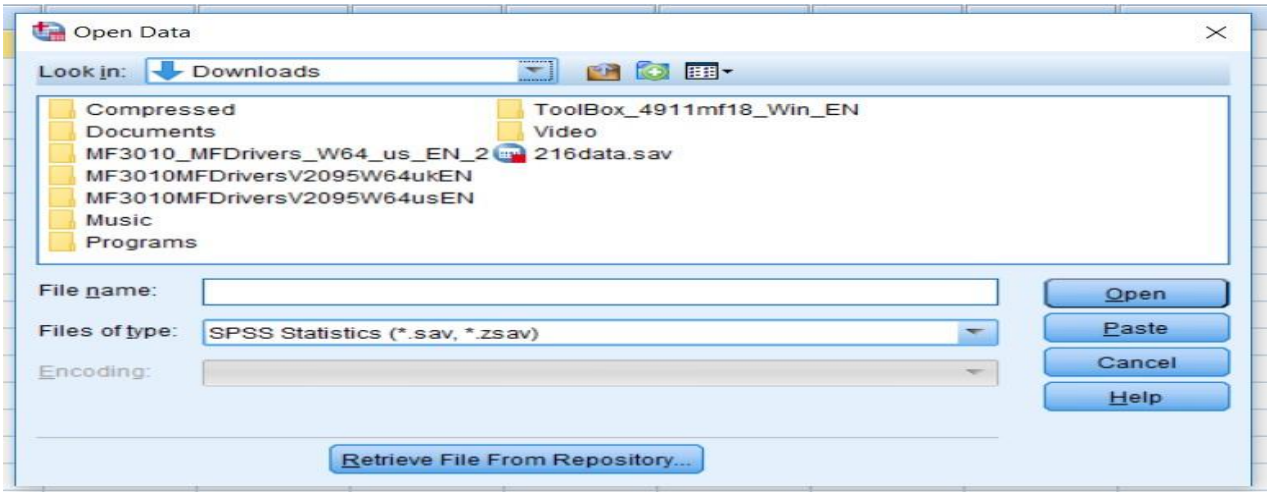

Figure (1): Opening data interface

Reading data from spreadsheet formats - Excel

In SPSS, go to FILE/OPEN. Click on the button “Files of Type.” Select the option “Excel (*.xlsx) .” Select the file, then click on “Open.” See Figure 2.

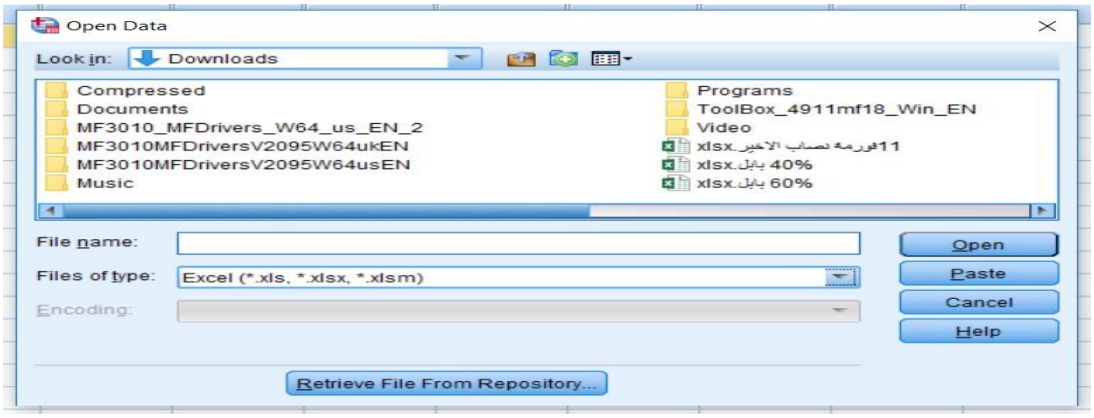

Figure (2): Opening excel data file

As shown in Figure 3. SPSS will request the range of the data in Excel and whether to read the variable names. Select to read the variable names and enter the range. Click on "OK."

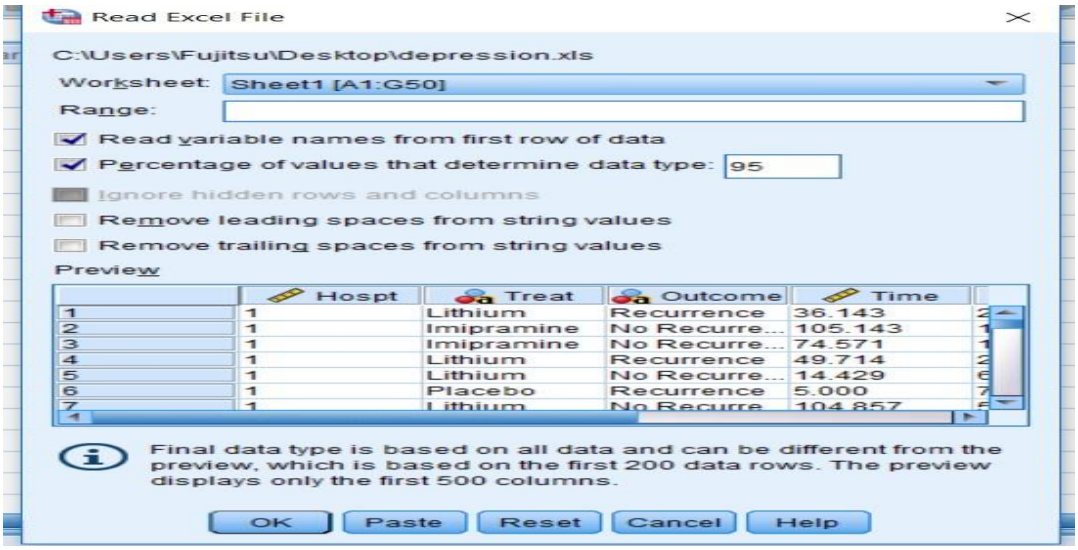

Figure (3): Reading excel file

The data within the defined range will be read. Save the opened file as a SPSS file by going to the menu option FILE/ SAVE AS and saving with the extension ".sav"

2. Sorting

Sorting defines the order in which data are arranged in the data file and displayed on your screen. When you sort by a variable, X, then you are arranging all observations

in the file by the values of X, either in increasing or decreasing values of X. If X is string variable, then the order is alphabetical. If it is numerical, then the order is by magnitude of the value.

Sorting a data set is a prerequisite for several procedures, including split file, replacing missing values, etc.

Go to Data/ Sort variables, then choose the variable and the type of sorting (ascending, or descending). See Figure 4.

Figure (4): Sorting variables of data

The order of selection is important - first the data file will be organized by *gender*. Then within each *gender* group, the data will be sorted by *education*. So, all males (*gender*=0) will be before any female (*gender*=1). Then, within the group of males, sorting will be done by*education* level. Let's assume you want to order *education* in reverse - highest to lowest. Click on *educ* in the box "Sort by" and then choose the option "Descending" in the area "Sort Order." Click on "OK." As shown in Figure 5.

gender	*educ*	*wage*	*age*	*work_ex*
0	21	34	24	2
0	15	20	23	2
0	8	21	25	5
0	0	6	35	20
1	12	17	45	25
1	8	14	43	27
1	6	11	46	25
1	3	7	22	2

Figure (5): example of sorting

3. Reducing sample size

Using random sampling

Let's assume you are dealing with 2 million observations. This creates a problem – whenever you run a procedure, it takes too much time, the computer crashes and/or runs out of disk space.

To avoid this problem, you may want to pick only 100,000 observations, chosen randomly, from the data set.

- Go to DATA/SELECT CASES.
- Select the option "Random Sample of Cases" by clicking on the round button to the left of it.
- Click on the button "Sample."
- Select the option "Approximately" by clicking on the round button to the left of it.
- Type in the size of the new sample relative to the size of the entire data set. As shown in Figure 6.

In this example the relative size is 5% of the entire data - SPSS will randomly select 100,000 cases from the original data set of 2 million. Click on "Continue."

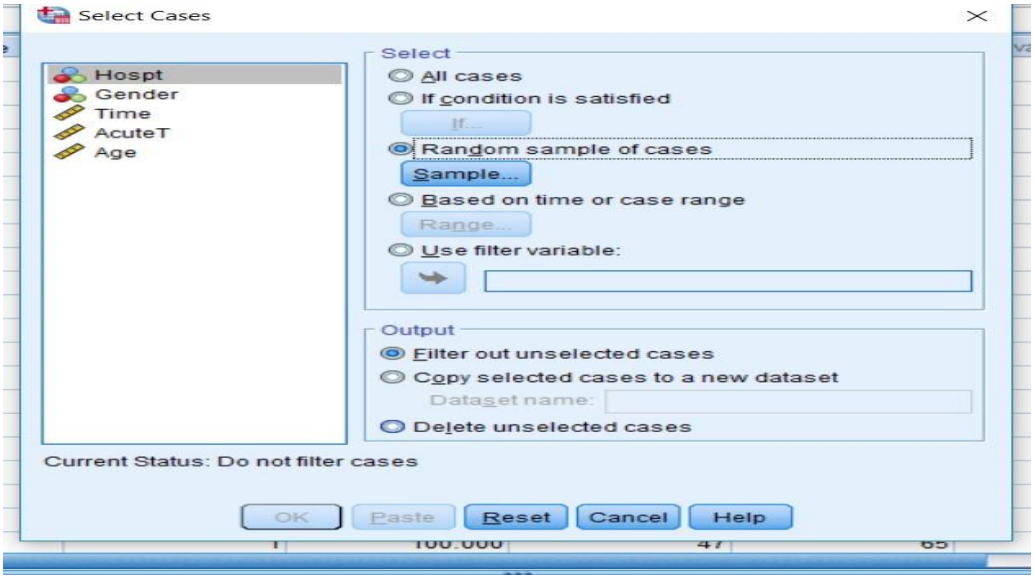

Figure (6): Filtering data

4. Filtering data

It will often be the case that you will want to select a Sub-set of the data according to certain criteria. For example, let's assume you want to run procedures on only

those cases in which *education* level is over 6. In effect, you want to temporarily "hide" cases in which *education* level is 6 or lower, run your analysis, then have those cases back in your data set. Such data manipulation allows a more pointed analysis in which sections of the sample (and thereby of the population they represent) can be studied while disregarding the other sections of the sample.
Similarly, you can study the statistical attributes of females only, adult females only, adult females with high school or greater *education* only, etc. If your analysis, experience, research or knowledge indicates the need to study such sub-set separately, then use DATA/ SELECT CASE to create such sub-sets.

A simple filter

Suppose you want to run an analysis on only those cases in which the respondent's *education* level is greater than 6. To do this, you must filter out the rest of the data. Go to DATA/ SELECT CASE When the dialog box opens, click on "If condition is satisfied." Click on the button "If."
The white boxed area "2" in the upper right quadrant of the box is the space where you will enter the criterion for selecting a Sub-set. Such a condition must have variable names. These can be moved from the box on the left (area "1"). Area "3" has some functions that can be used for creating complex conditions. Area "4" has two buttons you will use often in filtering: "&" and "|" (for "or"). As you read this section, the purpose and role of each of these areas will become apparent. See Figure 7.

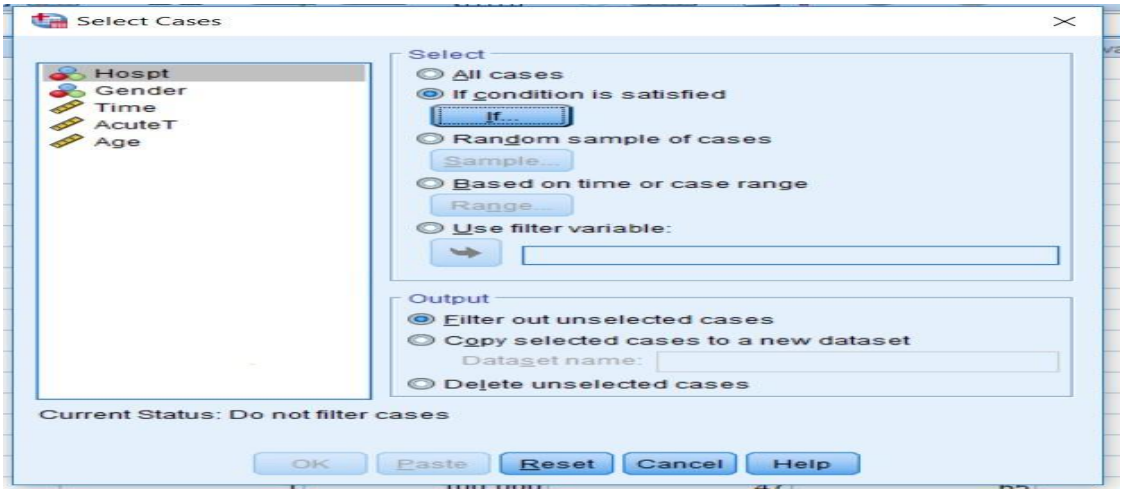

Figure (7): Selecting by condition