SPSS tutorial
Betalabb, 07/05/09

0. Introduction

SPSS (Statistical Package for the Social Sciences) is a program that is used for statistical analysis. In many ways, SPSS is comparable to Excel. It can be used to calculate, and to make diagrams. However, there are many more possibilities within SPSS than in Excel, and a number of procedures are much easier to do in SPSS. On the other hand, in some situations Excel may be the better option. The good news is that data can be exchanged between SPSS and Excel rather easily. SPSS exists for PC and for Macintosh. The licenses are rather expensive however. At SOL, SPSS is installed on the even-numbered computers in Beta-labbet, and on two Macintosh computers in Humlabbet.

This tutorial aims to give an introduction to SPSS. It is assumed that you are familiar with Excel.

1. Basics

Start the program (SPSS 14 for Windows), from the START menu. You get a spreadsheet that resembles an Excel worksheet. This is called the Data Window. Unlike Excel, a data window in SPSS can consist of one worksheet only. It is however possible to have multiple data windows open.

Each column in the worksheet represents a *variable*, each row represents a *case*. The names of the variables appear on top of each column. The numbers of the cases appear to the left of every row. This will become clear below.

We start with opening existing data. Data files are usually saved with the extension *.sav*. To open a data file, go up to the menu, to File > Open > Data. In the dialog window that comes up, find a folder on the desktop (*skrivbord*) named TUTORIAL plus the number of the computer that you are using (e.g., TUTORIAL02, TUTORIAL04, etc.). This will be your working directory during this tutorial, i.e., you use this directory for opening and saving files! In the working directory you will find a file called *namelist.sav*. Open it.

The file contains a list with five variables: persons (first column), where they live (second column), their age (third column), their gender (fourth column, *M* = man, *K* = Kvinna), and their income (fifth column). The list is currently sorted by the place names. Notice that the names of the five variables are shown on top of each column.

2. Make changes to the data

Changes can be made directly in a cell. This works more or less the same as in Excel: you simply select a cell and type over the text. Try this out by changing the first name in the first row from *Lars Möldener* to something else.

Notice also the case numbers to the left of each row. Scroll down to the bottom of the list to see how many cases there are in this data file.

If you want to add a new case to the data file, click once on an existing cell, and then go up to Edit > Insert Cases. This will add a new case (an empty row) to the data file. Fill this empty case with the following information:

Henrik Hansson Malmö 54 M 19456

You can navigate from one cell to the next using the mouse or the Tab-key. Note that the information is case sensitive, so the *M* has to be a capital *M*!

If you want to delete a case, then click on the case number so that an entire row is selected. Then go up to Edit > Clear (or press delete). The entire row will then disappear from the data file. Try this out by deleting a case from the data file.

3. Adding more info

Notice that there are two tabs on the bottom of the Data Window, one called Data View, the other Variable View. Click on the tab Variable View. This will change the layout of the Data Window, making it possible to enter additional information about the variables in the list, or to create new variables.

The first column (*Name*) shows the variable name. There are certain restrictions as to the names. For instance, they cannot contain spaces, or commas, and they cannot start with a number. So, *var 1* or *1var* are illegal names, but *var1* or *var.1* are OK.

The second column (*Type*) indicates whether the variable is a numeric variable (a number) or a string variable (text). In this data file, the variables Name, Place and Gender are of the string type, whereas age and income are of the numerical type.

The third column (*Width*) indicates how much space a variable takes up.

The fourth column (*Decimals*) indicates how many decimals should be displayed. This can only be specified for numerical variables. Try this out by increasing the number of decimals for income or age from zero to two. You can check the effect by switching back to *Data View*!

In the fifth column ("Label"), it is possible to give a longer name to a variable. This time it is OK to use commas or spaces and so on. The variable "Name" for

instance, could be labelled “Participant’s first and last name”. Try this out by giving a suitable label to one or more of the variables.

The sixth column (“Values”) can be very useful. In this data file, for example, the letters M and K are used to indicate male and female. This can be specified in the Value column. Click once in the Values column at line 4 (for Gender). You will see that a small grey square with three dots will appear. Click on this square with the dots. You will get a new window in which you can specify what the letters in the gender variable stand for. Type *K* (capital!) in the Value field, and *Kvinna* in the Label field, and then click on Add. Then type *M* in the Value field and *Man* in the Label field, and click on Add. Then click on OK to continue. Later on, the words *Kvinna* and *Man* will show up in the results, rather than the abbreviations *K* and *M*, which will make the output more readable.

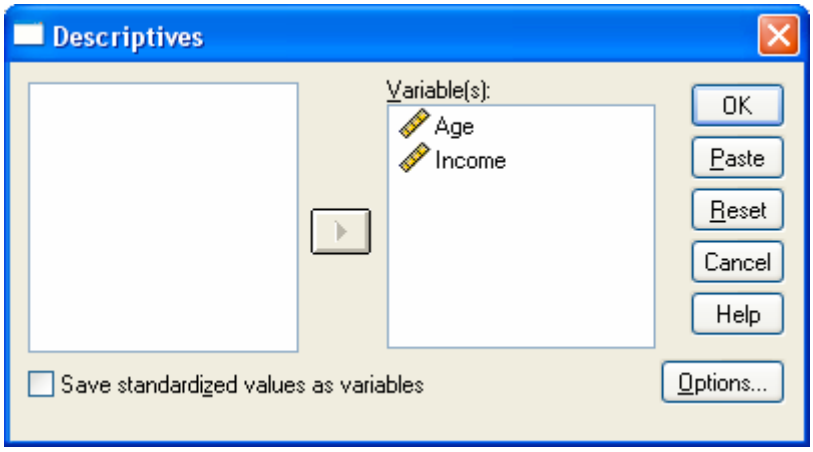
The seventh column (Missing) can be used to indicate that some values are missing. A missing value means that the value is unknown. These can be left open in the data file, or you can specify missing values yourself, for instance, ‘X’ means ‘Place unknown’, or ‘0’ means age unknown, etc.

The eighth and ninth column are used to change the layout (column width and alignment) in the Data View.

The final column indicates the measurement scale of the variable. Nominal means that the values indicate categories (for instance, 1 = red, 2 = green, 3 = blue). Ordinal means that the values indicate ranks (for instance 1 = very bad; 2 = bad; 3 = OK; 4 = good; 5 = very good). Scale means that the numbers are true numbers. Notice that string variables can only be nominal or ordinal. Numerical variables can be any kind.

4. Descriptive Statistics

We will start with a few descriptive statistics, e.g., what is the average age of the people in the list, what is the highest income, the lowest income, and so on. Go up to the menu to Analyze, and then to Descriptive Statistics > Descriptives. You get a window with the variables *age* and *income* on the left. The symbols in front of these variables indicate that these are numbers. The other three variables (*name*, *place* and *gender*) are not displayed in this window because they are string (text) variables, over which descriptive statistics cannot be calculated. Select *age* and *income*, and move them to the right, by clicking on the arrow in between the two window panes, as shown below.



Then click on **Options…**. You can now select which statistical measures you want. Make sure that at least **Mean**, **Minimum**, and **Maximum** are selected. Feel free to select other options as well. (**Standard deviation** and **Variance** are common measures for the amount of variation in the data. **Range** is simply the difference between the minimum and the maximum. **Skewness** is a measure for the symmetry of the data, and **kurtosis** is a measure for the shape of the distribution of the data.). Click on **Continue** when you are done, and then on **OK** to do the calculation.

Now, the output window should come to the foreground. Have a close look at the layout of this window. To the left-hand side of the main window is an overview part. The information on the left-hand side can help you navigate in the results. It shows that this part consists of "Descriptives" (indicated by the yellow book symbol) and the parts that it contains of. This can become handy when you need to navigate in an output window that has become very long.

The results of the analysis are shown in the main part of the window. Look at the table that is labelled "**Descriptive statistics**". The *N* in the second column indicates the number of cases over which the statistics were calculated. Otherwise, the table shows the options that you requested. Fill in the values in the following table:

	Age	Income
Mean		
Minimal		
Maximum		

5. Edit the output

It is possible to add or edit the text in the output. This can be good to keep the output organized. Double click on the Title ("Descriptives"). This will open a new window in which you can change the title text to something more meaningful.

It is also possible to delete unnecessary parts of the output. For instance, there is a field that says:

```
[DataSet1] C:\Documents and Settings\lingjwe\Skrivbord\TUTORIAL02\namelist.sav.sav
```

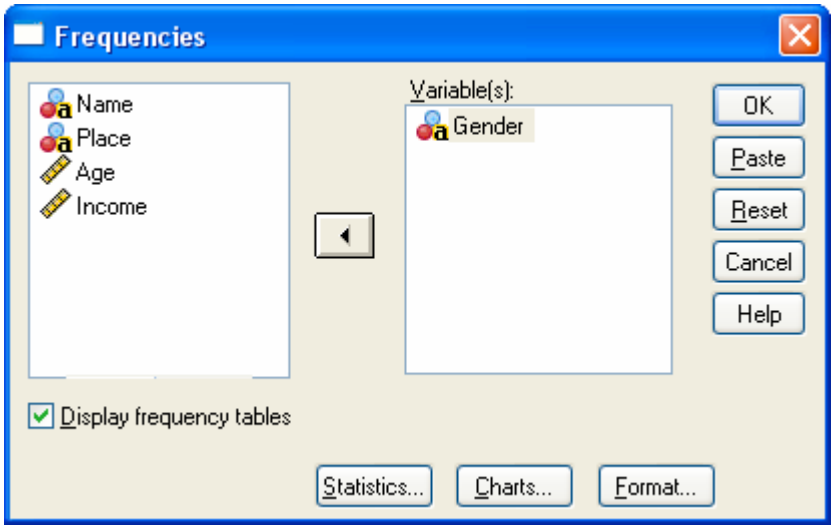
If you don't want this message in the output, then click on it once, and then press Delete to take it away.

6. Save the results

To save the results, go to the menu, File > Save. Give the results file a name, and save the output window in the working directory. By default, these output files are saved with the extension *.spo*, and they can only be opened in SPSS. It is also possible to save the output as a Word or HTML document. In that case, you can go up to the menu, File > Export…. Choose Word/RTF as the file type and save the file to the working directory (use the Browse button if necessary!). Try this out!

7. Frequencies

One of the things that is easier in SPSS than in Excel, is to do frequency counts. Suppose you want to know how many men and how many women there are in this list. In other words, you want to count the numbers of 'M' and 'K' in the fourth column. To do this, go up to the menu to Analyze > Descriptive Statistics > Frequencies. You will get a window with two panes. The five variables are on the left-hand side. Click on the gender variable, and then on the arrow to move it to the right-hand pane, as shown in the figure below:



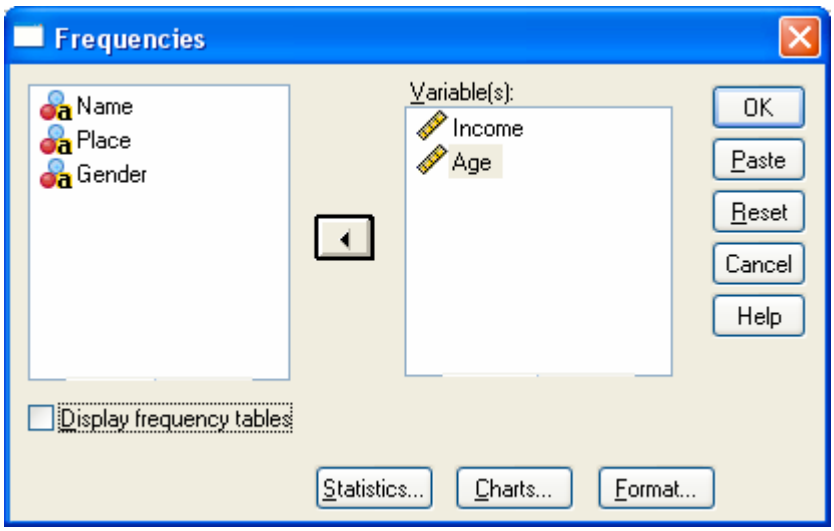
Then click on **Charts…**, and specify *Pie Chart* in the window that comes up. Click on Continue and then on OK, and wait for the result to come up in the output window!

The new results are appended to the previous results. There is a table labelled 'statistics' which shows the total number of cases that have been processed, and the number of missing data, i.e., cases where gender was not specified. The next table, labelled 'gender' shows the frequency counts for women and men. The table also shows these numbers as percentages. Fill in the values in the following table:

	Frequency	Percentage
Men		
Women		

The column valid percent becomes relevant when there are missing cases. The column cumulative percent shows the cumulative percentages, i.e., where percentages have been added (this is not so meaningful in the present example). Finally, you get a pie chart that shows the numbers as a graph.

Now gender is a so-called categorical variable, that is the values *M* and *K* denote categories, and in principle any two symbols could have been used to denote these two categories. Income and age, on the other hand, are so-called continuous (or scale) variables, that is, the values indicate real quantities. The frequencies option in SPSS is also useful for this type of variables. For instance, where is the limit to the highest 10% of all the incomes in the list? And, what is the limit if we were to divide the whole list into two equal groups of young and old people? In order to find this out, go up to Analyze > Descriptives > Frequencies once more. Move the variable Gender back to the left, and move the variables Age and Income to the right. Now first deselect the "Display frequency tables" option (which would result in very long and meaningless frequency tables). The window should look like the figure below:



Then click on Statistics. There are several ways of doing this, but the easiest way is to select the option "Cut points for 10 equal groups" (10 is the default). This will divide the two variables into ten subsequent groups (youngest 10%, next youngest 10-20%, etc.). Also select the Minimum and Maximum values, which will help with the interpretation of the output. Click on Continue. Then also click on Charts, and select Histograms and With Normal Curve. Click on Continue, and then on OK to run the analysis.

The output will give you a table with the cut points for the intervals. Fill in the upper and the lower boundaries in the following table:

	Age	Income
Lowest 10%		
Highest 10%		

Look once more at the SPSS output, and notice that the intervals get narrower towards the middle and widen towards the ends. This is because the data are so-called normally distributed, that is, there is tendency for the data to cluster around the midpoint, and to be less common further away from the midpoint. This can be seen in the two histograms, which show a bell-shaped curve which is the typical shape of a normal distribution. The ideal normal curve is indicated by the smooth black line drawn inside the two graphs.

A final comment: the 50% cut-off point is also called the *median*, that is the boundary at which 50% of all the data points are higher and 50% which are lower. Fill in the median and mean (calculated in the previous step) values for income age in the following table.

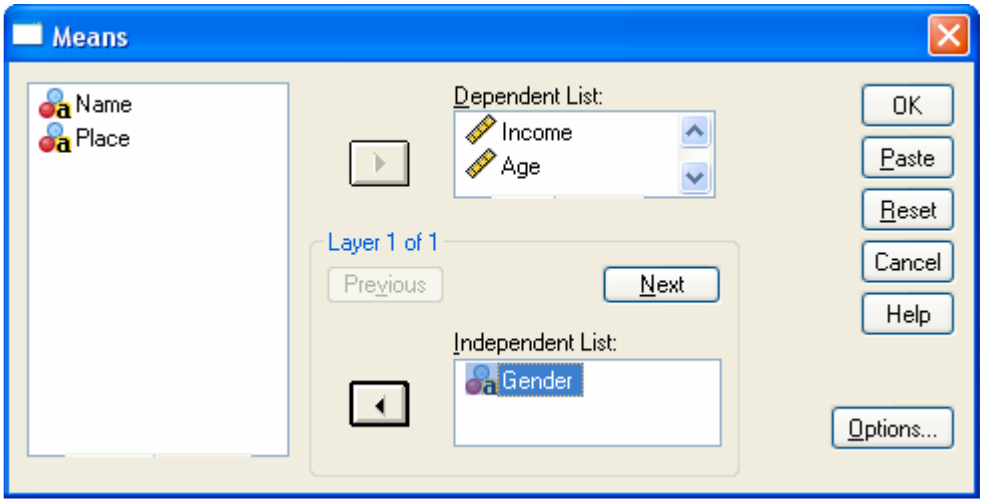
	Age	Income
Mean		
Median		

In this case, they are approximately equal, and that is because the data are normally distributed. This is not necessarily the case. When a distribution is asymmetrical (for instance when there are more higher than lower values), there will be a discrepancy between the mean and the median.

8. Compare Groups

Now we know how many men and women are in the list, we would like to know what their separate average age and their separate average income are. In order to do this, go up to Analyze in the menu bar, and then to Compare Means > Means.

You will get a window that has three variables on the left hand side, and two empty panes on the right hand side. Move the variables Income and Age to the top right, where it says "Dependent variable". Dependent variable means that it is a variable that varies due to the influence of another variable, the independent variable, in this example Gender. So, move Gender to the bottom left, that says Independent List. The window should look like the picture below:



Click on **Options…**. Here you can specify which statistical values you are interested in. In this example, the Mean value needs to be selected. Feel free to select additional statistics if you like. Then, click on **Continue**, and next on **OK** to run the analysis.

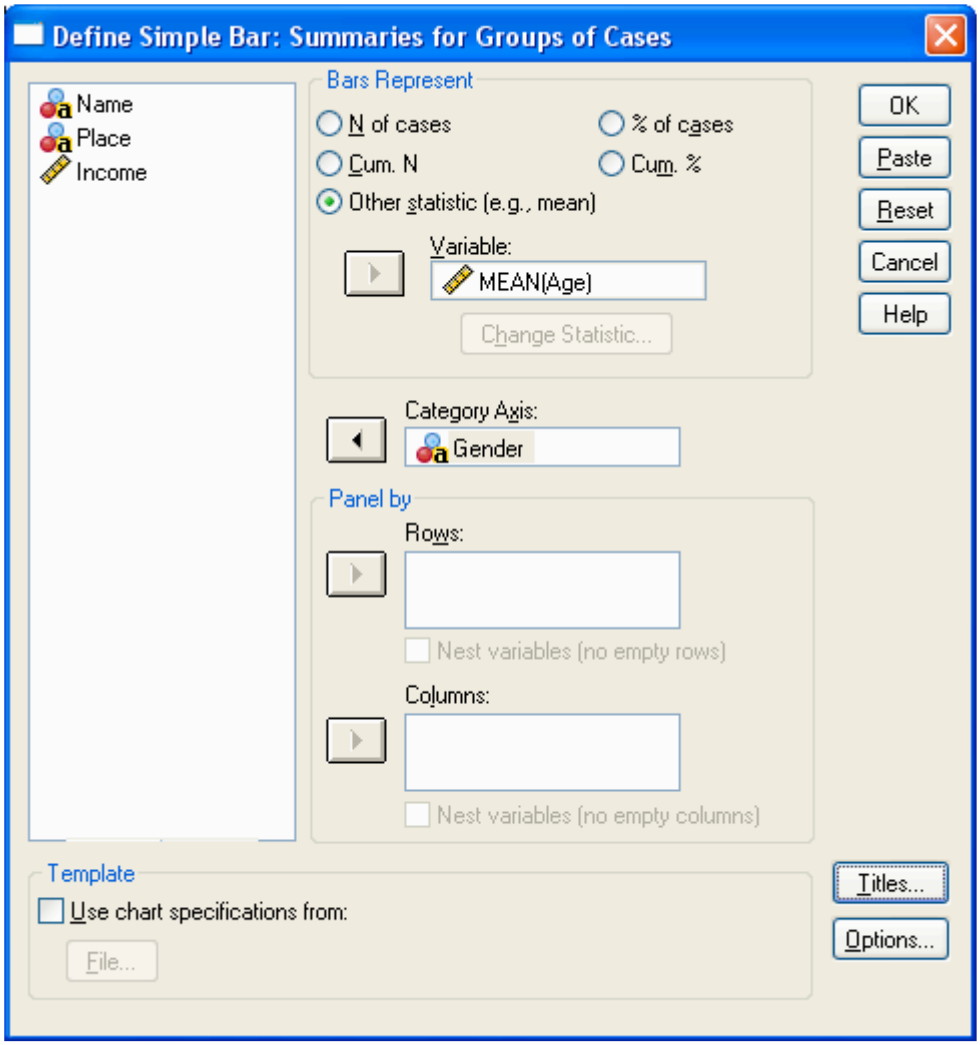
The new results are appended in the output window. This is also shown in the overview part to the left. The table that is of interest, is the table that is labelled "Report", which shows the average income and the average age of the men and the women in the list, including the averages for the two groups together. Fill in the values in the following table:

	Women	Men
Mean age		
Mean income		

9. Make a chart

A bar chart would be an appropriate way of visualizing the average values that you calculated in the previous step. In order to make a bar chart, go up to the menu to Graphs > Bar… Simple > Define. You will get a window in which the variables are on the left-hand side. We will make a bar chart of average age first. First select "Other Statistic", then click on the Age variable to the left. Move it to the right, by clicking on the arrow in front of the "Variable:" window.

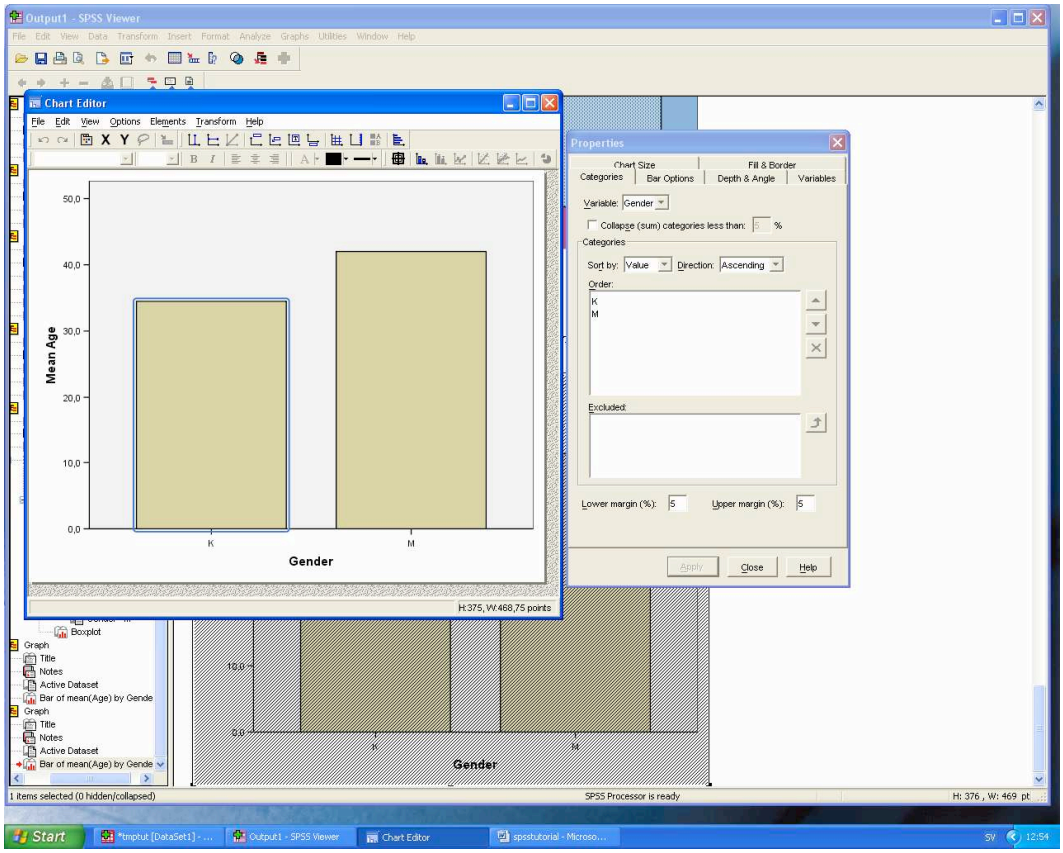
Furthermore, move the Gender variable to the right-hand side where it says Category Axis. If all is well, the window should look as shown below. If that is so, press **OK** to continue.



The resulting Bar Chart will appear in the output window. Now create another one with the average income for men and women!

10. Edit a graph

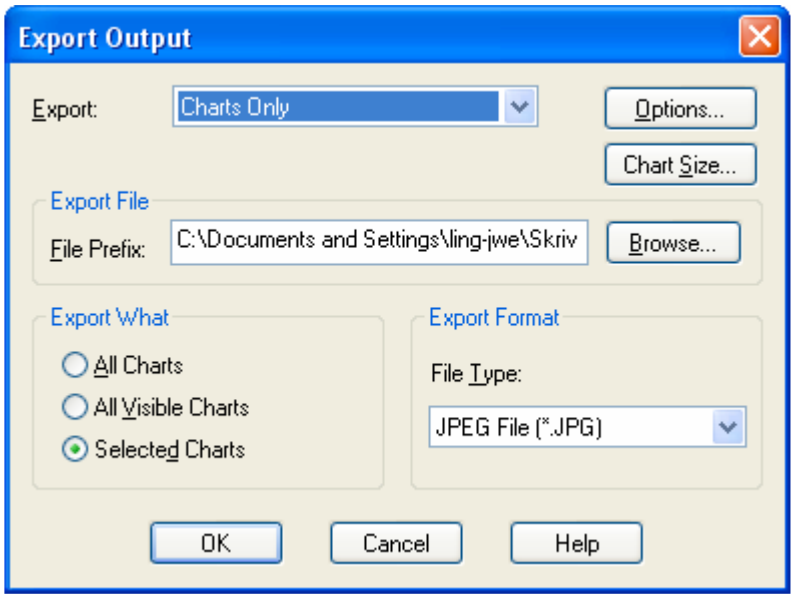
Double click on the bar chart. This will open the Chart Editor in which you can change the layout of a graph. Right-click on one of the two bars, and from the menu that comes up, choose Select > This bar. Then right-click once again, and choose Properties window (or type control-T). This opens the properties window. Arrange the Chart-editor window and the properties window next to each other on the screen, so that you can see them both as shown below.



Under the Fill & Border tab in the Properties Window, you can change the colour of parts of the graph. Use this to change the colour of the selected bar. Pick a colour from the alternatives, and press **Apply**. The colour in the chart should change to the one that you chose. Now click once on a different part of the chart (for instance, the other bar), and immediately go back to the Properties window, to change the colour of this part. When this works well, feel free to try out some of the other options. Under the Depth & Angle tab, you can make the graph three-dimensional. Under Bar Options, you can change the width of the bars. When you are done, close the properties window.

Now double-click on one of the numbers to the left of the Y-axis. This will open a different properties window in which you can change the scale and the layout of the axes. Go to the Number Format tab, type 0 (zero) in the Decimal places field, and press Apply. This will remove the decimal. Under the Scale tab, change the minimum number to 20, while keeping the maximum at 45. This will increase the difference between the two bars. Under the text style tab, you can change the font type and colour. Feel free to try out some of the possibilities.

When the chart is the way you want it, close the chart editor, which brings you back to the Output Window in SPSS. Now suppose you want to use this chart in another document. Go up to the Menu, File > Export. The option for Export should be “Chart Only”.



Now save the chart as a JPEG file in your working directory.

11. Sorting

The list is currently sorted by place name. Suppose you want the list to be sorted in some other way, e.g., by name, income, or age. To do this, go up to the menu to Data > Sort Cases. You will get a window where you can specify which variable or variables you want to sort by, and whether the direction of the sorting should be ascending or descending.

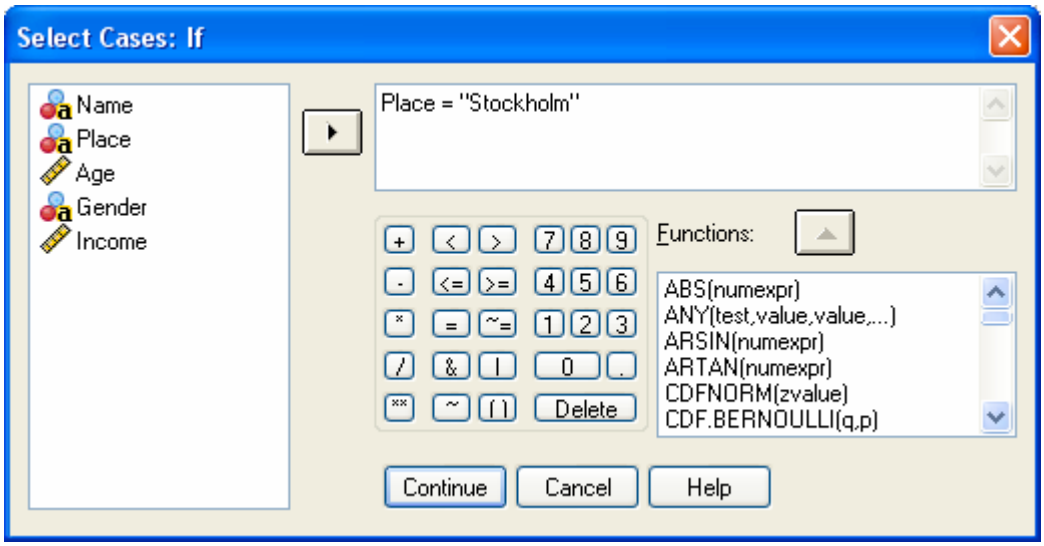
Now sort the data, once by age, and once by income. Then scroll up and down in the data window to fill in the values in the following table:

Oldest person	
Youngest person	
Richest person	
Poorest person	

12. Restricting the analysis to a subset of the data

Finally, suppose you want to select only a part of the list, for instance, only the people who live in Stockholm. This can be done as follows. Go up to the menu, to Data > select cases. In the window that comes up, click first on the button that says "If condition is satisfied", and then of the "If…" button. You get a new window with the five variables to the left. Click on Place, and then on the arrow, so that it moves to the right. Then after Place type **= "Stockholm"**. The quotes are

necessary, because this is a text variable! The window should look as shown below. If it does, then click first on Continue, and then on OK to terminate.



Now look at the data window, you will find that most of the record numbers have been crossed out. Besides, a new variable has been created with the name **filter_$**. The filter variable has the value 1 for all the selected cases, and zero for the unselected cases. You can verify this by scrolling up and down the data file. The values that are crossed out will not be included in a new analysis.

Now use SPSS to fill in the values in the following table:

Number of people from Stockholm:	
Average income from the people from Stockholm:	
Average age of the people from Stockholm:	
Number of women from Stockholm:	
Number of men from Stockholm:	

Note that if you want to undo the selection, then you go back to the Select Cases window, where you choose All cases.

13. Printing the output

The entire output or selected parts can be printed directly from SPSS. Suppose you want to print the frequencies analysis of gender, including the tables and the pie chart. The easiest way to do this, is to click once on the yellow-book symbol that indicates the Frequencies part of the output (in the left-hand side of the output window). This will select the parts that you want to print. Go up to File > Print Preview, to check that the result is the way you want it. Then choose Print…, and send it not to the PDF creator, but to the printer "\\student\solpr-betalabb".

14. Postscript

Here ends the tutorial, but this is only the beginning of SPSS. There are many more functions to explore. Please do not hesitate to contact me if you would like to use the program for something in specific. Just send an email to:

Joost van de Weijer
metodolog@sol.lu.se or vdweijer@ling.lu.se.