Introduction to SPSS

Written by: Emory Morrison

Center for Social Science Computation & Research

145 Savery Hall

University of Washington

Seattle WA 98195 U.S.A.

(206)543-8110

October 99

http://augustus.csscr.washington.edu/pdf/spss.pdf

INTRODUCTION TO SPSS

SPSS (Statistical Package for the Social Sciences) is a versatile software package which primarily assists users in performing complex statistical analyses of quantitative data sets. The software allows users to create, modify, and analyze data, as well as to produce graphics to display findings in reports or presentations. SPSS is comparable to other general statistical analysis software such as SAS, Stata, or S-Plus. Relative to these other software packages, SPSS is easier to learn and more simple to use. However, it is more restrictive than the other statistics packages; advanced users have a tougher time tailoring SPSS to meet their specialized analytical needs.

This document should help new users get started in SPSS. It is structured in question and answer format, and addresses in a logical sequence the questions that a new user might have. Once you are on your way, (or if you already have a basic knowledge of SPSS), you can use the set of manuals that SPSS provides, the help features within SPSS, or run the SPSS on-line tutorial for more comprehensive guidance.

Please note, at the time of the writing of this document, CSSCR is using version 9.0 of SPSS software. All information in this document will pertain to SPSS version 9.0.

How do I start using SPSS at CSSCR?

SPSS (version 9.0) is on all the PCs in CSSCR. The process of launching the software is the same as for any other frequently used package in the lab. At the initial window (once Windows has been launched) find the SPSS icon. Double click with your mouse, and voila.

This screen is the **data editor** window in SPSS. It is one of a seven basic windows that SPSS provides (other windows include: **the output viewer, syntax editor, chart editor, pivot table editor, text output editor, and script editor**). The data editor window is the most fundamental. With the exception of the more advanced syntactical procedures, you must enter data in the data editor window before SPSS can perform any operations.

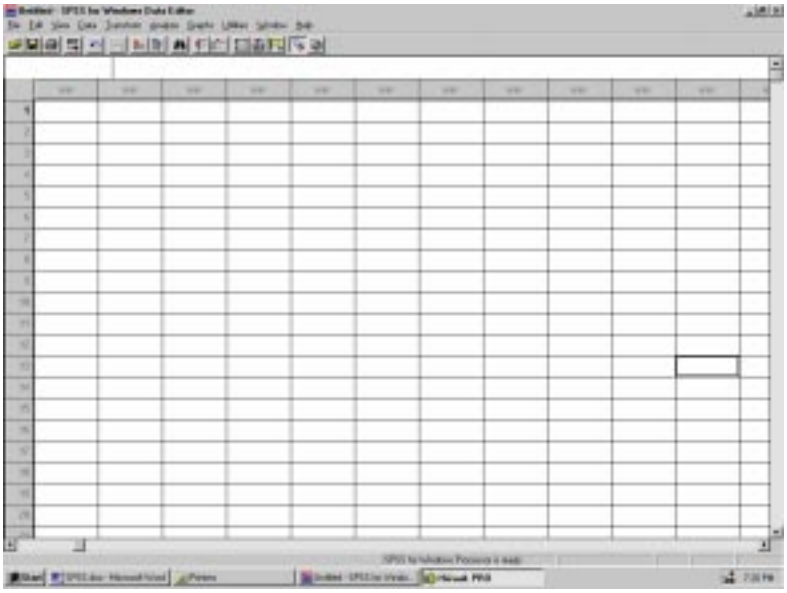


Figure 1

What Are Data?

SPSS is very restrictive in what it recognizes as data. For SPSS, data are members drawn from a similar set (**cases**), that are each described by the same set of features (**variables**). Cases are listed in **rows**, and variables are listed along **columns**. The block where a row and column intersect is known as a **cell**. Each cell contains the value on the variable (the **column**) for the case (the row). Thus, data must take the form of a **spreadsheet**. An illustration of data in the SPSS data editor window is presented below.

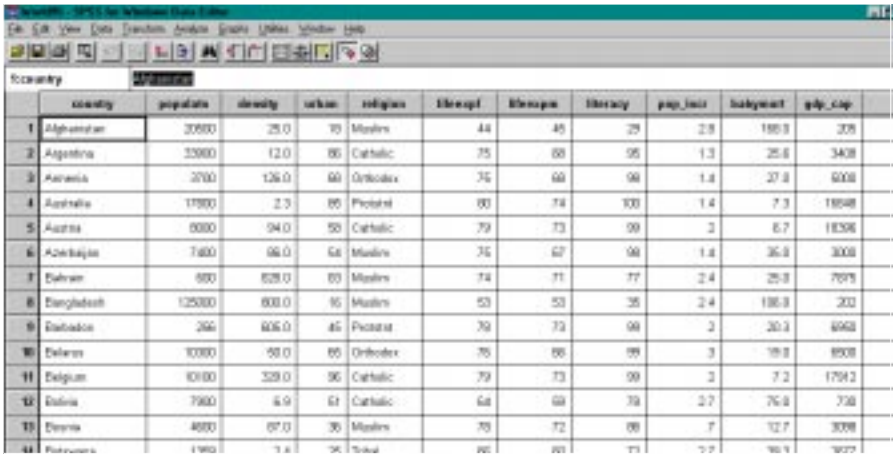


Figure 2

In Figure 2, the cases are countries of the world (thirteen of which appear in the figure) and the variables are characteristics of each of the countries (i.e. population, population density, percent of population residing in urban regions, etc.). Since SPSS only recognizes a variable title eight characters long, many of the variables have shortened names. For example, the column that contains information on population is titled populatn.

The value for the case Australia for the variable "populatn" is 17800 (which, since the units are in thousands, indicates that Australia has a population of 17.8 million people). The value for the case Afghanistan for the variable religion is Muslim. Values can either be numeric or in character (also known as string) format.

However, if you want to use a variable in an analysis, you will need to code it in numeric format (even if a character format makes more sense to you). So, if you want to analyze the relationship between population size and nations' principle religion, both variables will need to be coded in numeric format. Each religion should have a unique number, so that Muslim could be 1, Christianity could be 2, Buddhist could be 3, Hindu could be 4, etc.

What do I Do Once I Open SPSS?

Input the data that you want to analyze into the SPSS **data editor window**. SPSS can get data through a variety of methods. First, you can manually type the data directly into the data editor window. If you want to enter data this way, define your variables first. Start by moving your mouse to the gray area at the top of the left most column. Your cursor should form a white arrow. Double-click. Your screen should look like Figure 3.

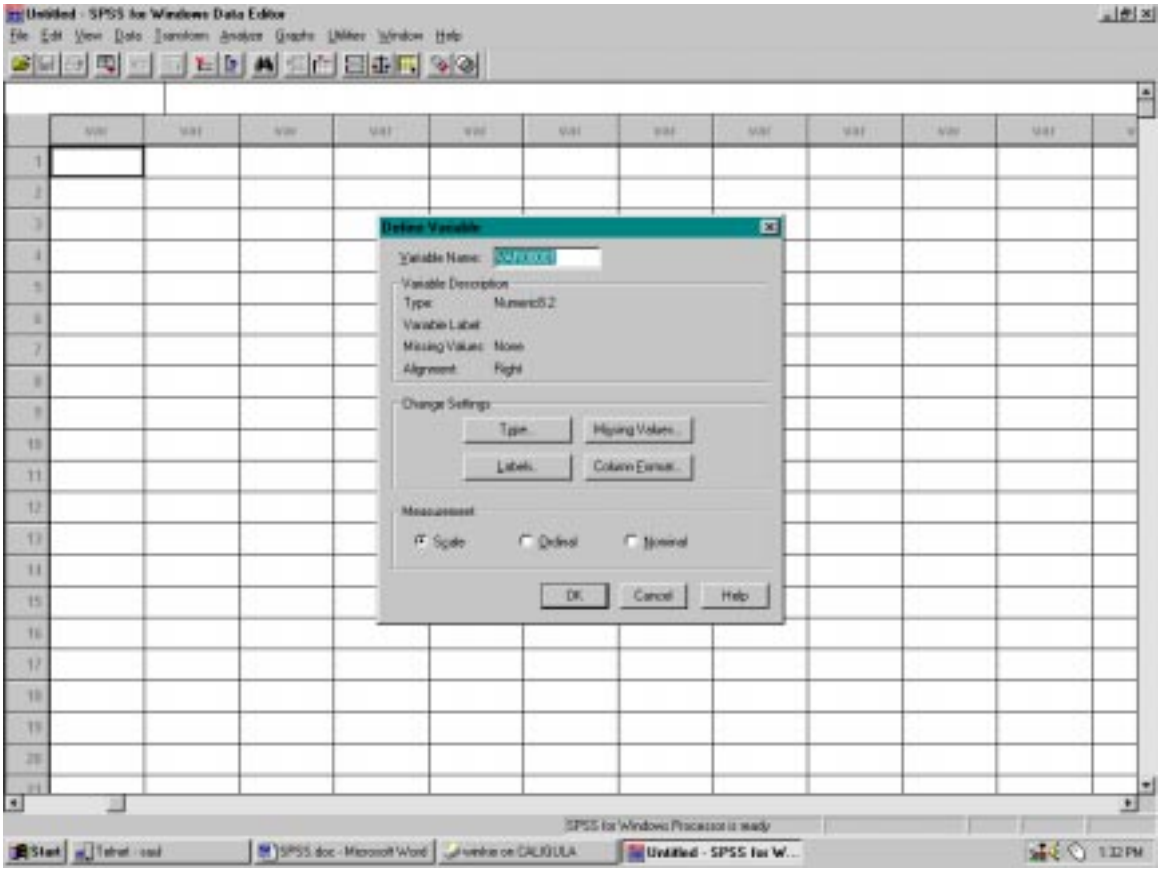


Figure 3

The gray box that appears in the middle of the spreadsheet is the **Define Variables dialogue box**. You should decide what to name your variable (eight characters or less) and type it into the white rectangle next to the prompt Variable Name.

Imagine that you have the data from Figure 3 on paper and you would like to type it into the computer. Type "Country" into the rectangle provided for a variable name. Since this variable is the case identification, and you want to define your cases using characters (rather than numbers), you'll need tell SPSS. Double click-on the Type button.

This will activate a new window which allows you to define the way the data editor interprets the data that you enter. Click the button to the left of the word "string." Next to where the prompt says "Characters," type the number of spaces you will need for your longest entry (perhaps 20 will suffice). Click Continue. This will return you to the Define Variable dialogue box.

Note that once you have told SPSS that you will be using string data for this variable, it automatically switches your Measurement setting to Nominal. Now, click on the OK button. You will be able to enter the data for the first column; strike the return key between entries to automatically put each new entry in a new row.

When you have entered the data for that column, double-click on the gray area above the second column. Type "Populatn" in the slot provided for Variable Name. Since the values for this variable are numeric, be sure that under "Type" the numeric button is marked. Next click the button marked Labels.

The top of this new dialogue box asks you for a **Variable Label.** In the box next to this prompt, type "population of the country in thousands." When you have a data set with a lot of variables and information, the variable labels will help you keep track of what each variable represents. Click Continue. Since population is a scalar variable, make sure that under the Measurement prompt, scale is marked. Click OK. Type in column 2 and continue in the same fashion until Column 5.

Once you get to the fifth column, the one titled Religion, you will have an interesting situation. The data clearly is in character format, but if you want to use it for analysis, it must be in numeric format. The solution to this problem is to assign each string value a numeric code, and then indicate how the coding works in the Define Variable window.

So, after you have double-clicked on the gray heading of the fifth column and have typed "Religion" next to the Variable Name, click on Labels. Next to the prompt Variable Label, type Principle religion of the country. Next to the prompt Value type the number 1. Hit the tab bar. You will now be positioned to tell SPSS what the number 1 indicates for this column. Type "Muslim" and click on the Add button. Repeat the process until you have entered the numeric code for each possible entry. When you are done, click Continue. Next, click on the Type button to ensure that the variable is numeric. Finally, since the measurement level is nominalist the Nominal button is activated. Click OK. You will now be able to type in the numeric codes for the religion variable.

Using .sav files

The manual method of entering data directly into SPSS is both cumbersome and time-consuming, especially when you want to analyze a large amount of data. A second method for introducing data to SPSS is to find a file that has already been prepared in SPSS format, and open it. Such files have a **.sav suffix** affixed to them. For example, the file in Figure 4 is titled "World95.sav." It is stored on each the PCs in CSSCR on the C drive, within the directory C:\program files\SPSS. In order to open it, click on the File menu. A list of options will appear under the menu bar; the second option is Open. Click on it.

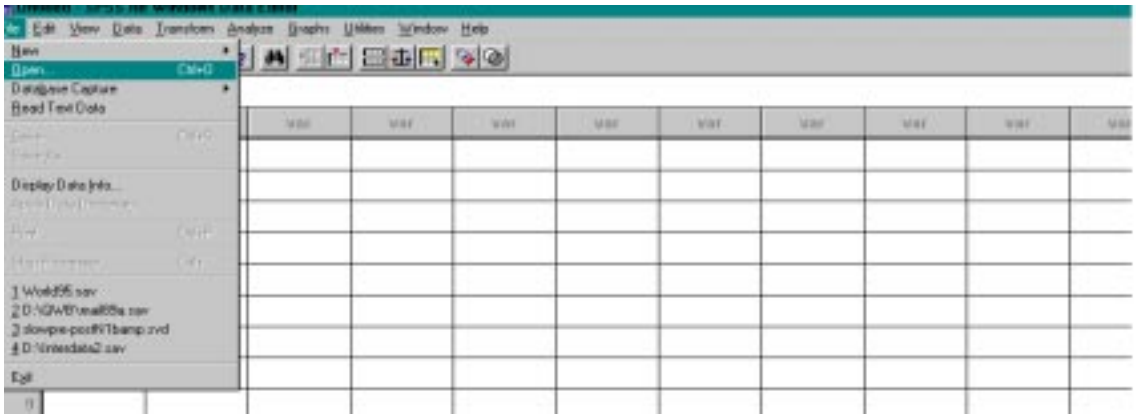


Figure 4

You will see the dialogue box below.

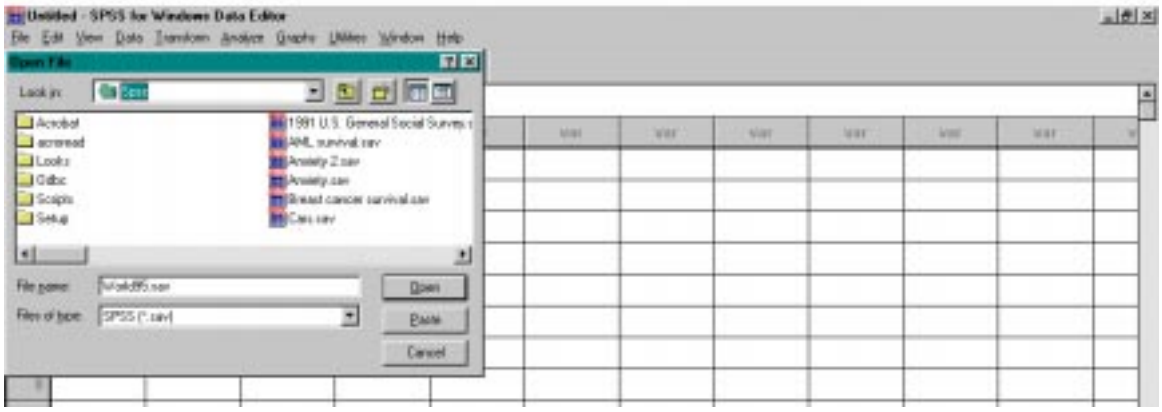


Figure 5

The top prompt, labeled "Look in:" provides a tree directory of your computer; find the drive and folders in which your file is stored. Next to the prompt File name type in the name. Next to Files of type check that the SPSS(*.sav) option is up. Click on the Open button. You will now be ready for analysis.

Aside from the above two procedures, other methods are possible to input data. One may take data that has been stored using other software packages (like excel or access). One may get data that has been stored in ASCII format. One can enter data though syntax. This document will not detail how to get data through these methods.

Okay, I've Got My Data, What Will SPSS Do With It?

Once you have your data entered, you can use SPSS to manipulate and analyze your data set. SPSS will allow you to add new variables to your data set, to pare down cases, to merge files together, etc. SPSS will also allow you to generate summary statistics for your variables, to run statistical tests like correlations on your variables, it will also allow you to run many types of regression analyses, as well as to run variants of the standard ANOVA models, and it will let you run all sorts of advanced statistical models and tests. SPSS will also let you create graphical representations of your data set (such as pie charts, bar graphs, scatter plots, etc.)

The new user will use items off the menu bar to direct SPSS to perform tasks. More experienced users may use syntax to direct SPSS. Since this document is directed to the new user, it will only address items that are on the menu bar. The menu bar contains ten headings: File, Edit, View, Data, Transform, Analyze, Graph, Utilities, Windows, and Help. Clicking on any of the ten headings will provide the user with a number of options. Most of the options under the File, Edit, Windows, and Help menus are not unique to SPSS; therefore, this document will address options under the other headings.

Click on View at the top of the menu bar. Figure 6 shows the View options.

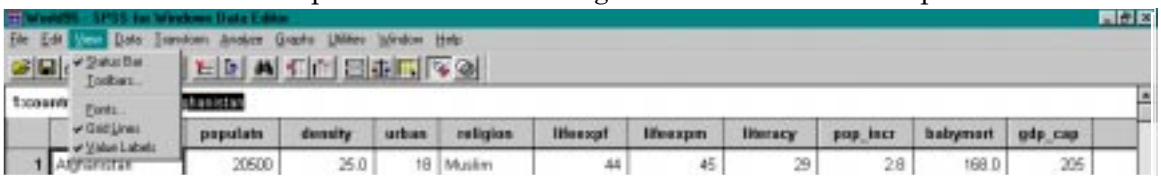


Figure 6

The only option worth noting here is the bottom one, **Value Labels.** Click on it to either activate it (check it), or deactivate it (uncheck it). When it is checked, the spread sheet will display in character form the values for variables. Thus, while it is checked, the column Religion displays the label Muslim. When it is unchecked, the spread sheet will display only the numeric value of the variable (so under religion the first entry would be 1).

Click on the Data pull-down menu in the menu bar. Figure 7 shows the commands under the Data menu.

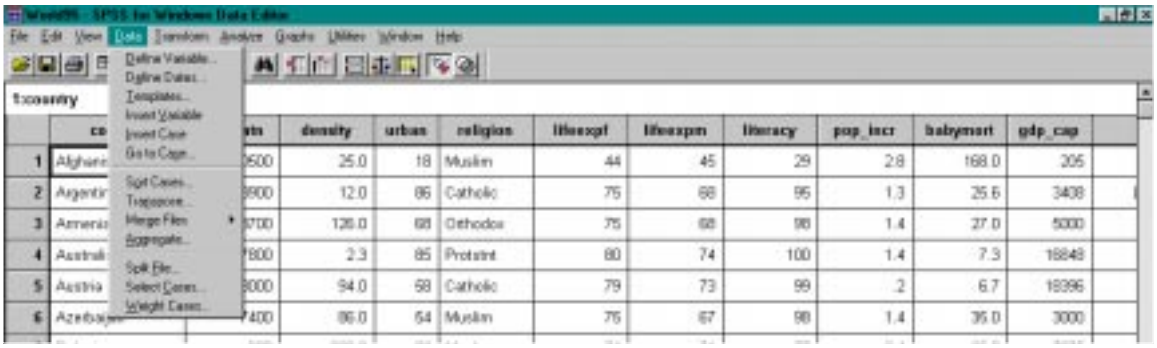


Figure 7

These options allow you to do things such as move within your data set **(Go to Case)**, edit your data set **(Insert Variable, Insert Case, Define Variable)**, expand your data set by joining two data sets together (**Merge Files**), contract your data set by selecting a subset of cases (**Select Cases**), Group your cases into categories for separate analyses (**Split File**), as well as perform other functions. For details on how these functions work, consult an SPSS manual.

Now click on the Transform pull-down menu. Here are the commands under the Transform menu.

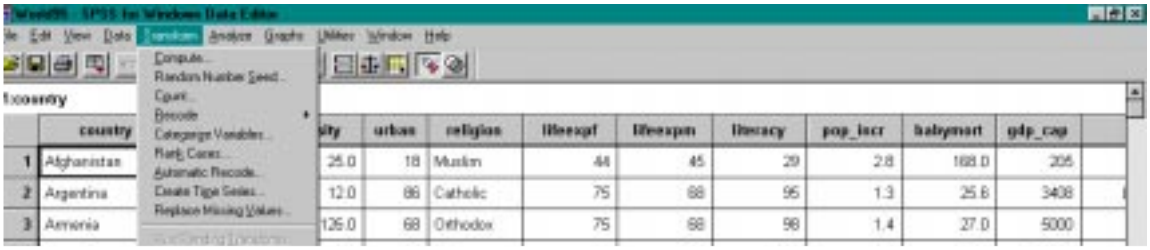


Figure 8

These options help you to create new variables, and to assign values to the new variables based on specific procedures. Two options that we will look more closely at are **Compute** and **Recode**, both of which allow you to use existing variables to create new ones. First, we shall examine the Compute function. Imagine that we wish to create a new variable which records the life expectancy differences between males and females in each country in the data set. The life expectancy for males already exists, lifeexpm and the life expectancy for females already exists, lifeexpf. In order to create our new variable (complete with values for all the cases) we need to click on the compute button. A new dialogue box will open up, which is displayed in Figure 9.

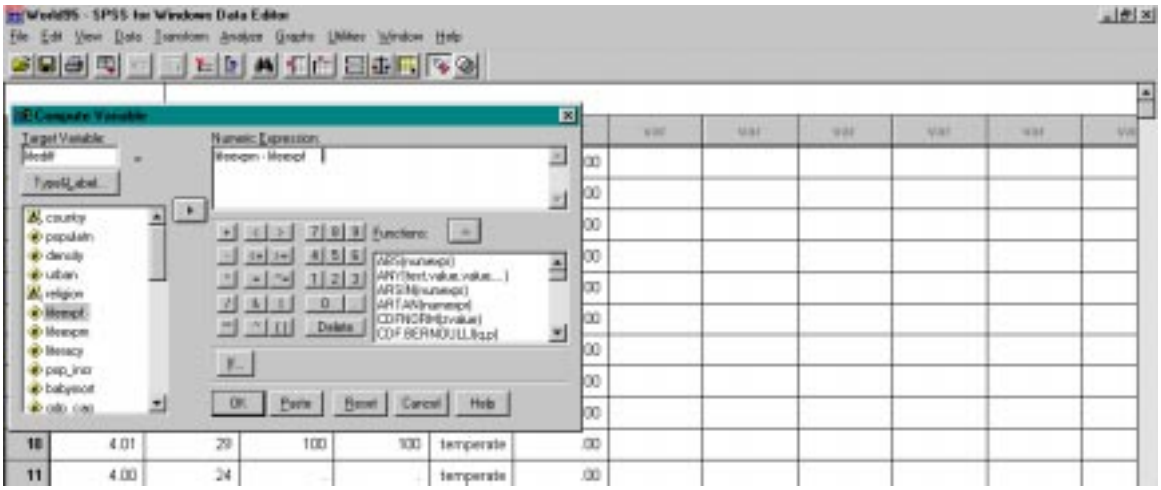
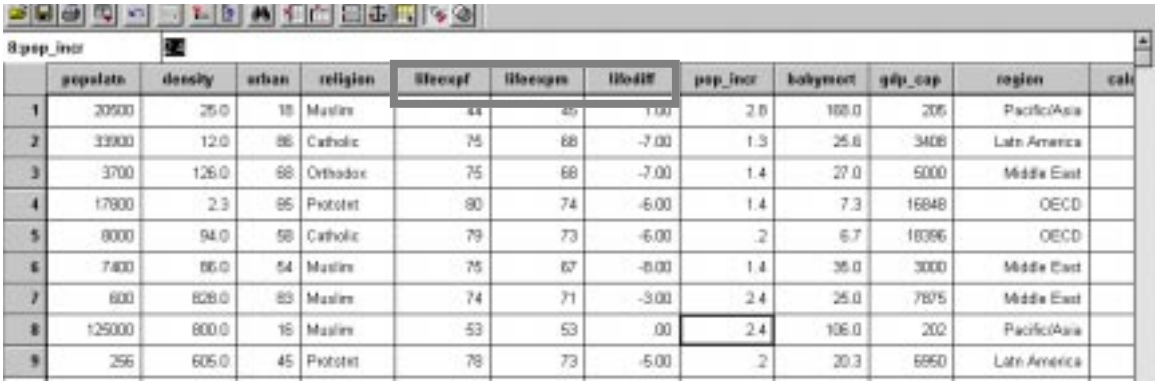


Figure 9

You need to pick a name for the new variable, and type it into the box labeled Target Variable. Next under the box Numeric Expression you need to type in a formula which defines how the new variable will be calculated. In the current case, the new variable, (the difference in life expectancy between men and women), will be named "lifediff." It will be "lifeexpm" minus "lifeexpf." Once this information is typed into the appropriate boxes, click the OK button. SPSS will have created the new variable and will have assigned values to it for all cases which had valid scores for both lifeexpm and lifeexpf.



	pop_incr	babymort	gdp_cap	region	lifediff
1	2.8	168.0	205	Pacific/Asia	
2	1.3	25.6	3408	Latin America	-7.00
3	1.4	27.0	5000	Middle East	-7.00
4	1.4	7.3	16848	OECD	-6.00
5	.2	6.7	18086	OECD	-6.00
6	1.4	35.0	3000	Middle East	-8.00
7	2.4	25.0	7875	Middle East	-3.00
8	2.4	106.0	202	Pacific/Asia	.00
9	2.0	20.3	6950	Latin America	-6.00

Figure 10

Notice in Figure 10 how SPSS formed a new column, lifediff. Note also how the values in lifediff are, in fact, the difference between the values in the columns lifeexpm and lifeexpf.

Recoding

In addition to calculating new variables, people frequently use SPSS to transform variables by categorizing them using the **Recode** option. From the Transform pull-down menu, choose Recode; click the Into Different Variable option. A Recode dialogue box will appear.

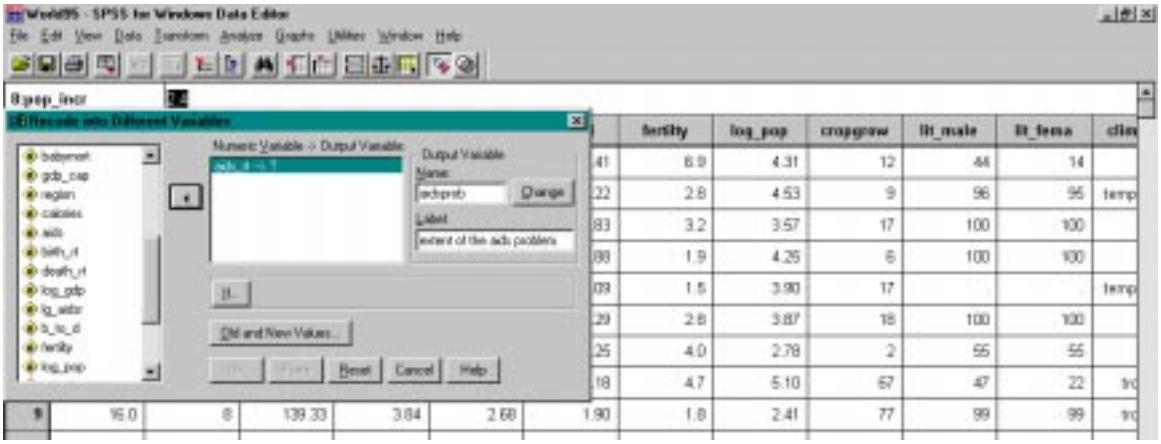


Figure 11

In the large white box in the middle, type in the name of the existing variable that you would like to recode. For example, imagine that you would like to recode the variable which reports the rate of aids cases that a country reports (per 100,000 population), aids_rt. This variable ranges from 0 to 327. You would like to identify four categories of countries based on the extent of the aids problem within the country. Some countries (perhaps where the aids rate is less than 1 per 100,000) you would like to label as controlled problem; other countries (perhaps where the aids rate is more than 1, but less than 10 per 100,000) you would like to label as problematic; still other countries (perhaps where the aids rate is between 10 and 100 per 100,000) you would like to label as in crisis; finally you would like to identify the countries (perhaps those with an aids rate greater than 100 per 100,000) as "severe crisis."

In order to perform this task, type "aids_rt" in the large white box. Type the name of your new variable in the Output Variable Name box. Once you type in the name of a new variable (and perhaps a label for it too), click on the Change button. Next click on Old and New Values... to tell SPSS how to change the values from one variable to another. After this step, you'll see a new dialogue box (Figure 12).

You will need to spend several minutes working with this new dialogue box. First,

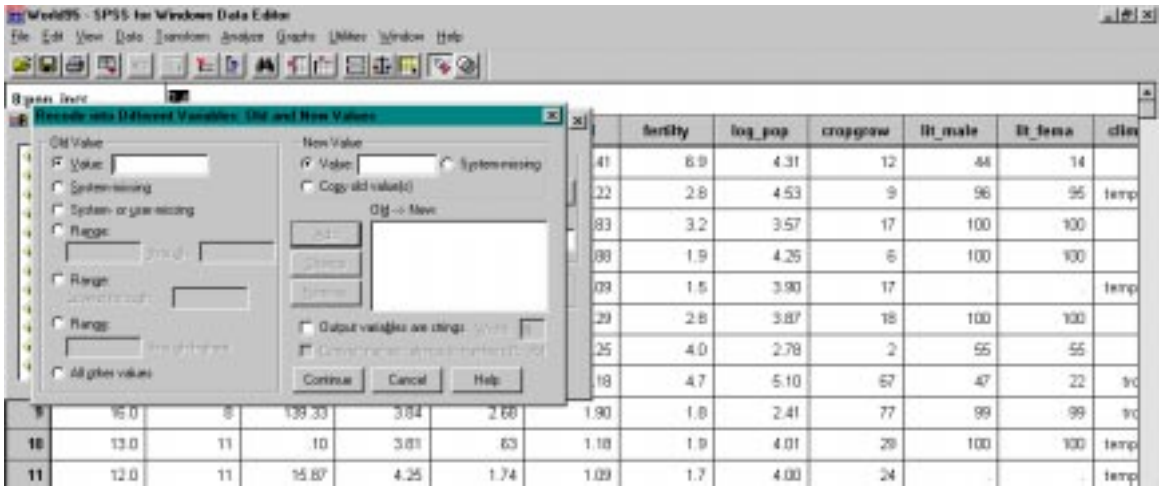


Figure 12

you need to direct SPSS to recode the first set of values. Click the Lowest through Range button. Type the number 1 in the box. Next, move to the New Value box; type a 1. Click on the Add button.

You will see Lowest thru 1 --> 1 appear in the box under Old --> New. Next, click the button next to the uppermost Range. Type 1 into the left rectangle and 10 into the right rectangle. Go back to New Value and type 2; click on the Add button.

Move back left again, click on the same Range button. In the left rectangle type 10; type 100 in the right one. Type a 3 into the New Value box. Click on the Add button.

Now, click on the through highest Range button. Type a 100 into the box. Under New Value type a 4. Click the Add button.

Click the Continue button; you will be returned to the first dialogue box. Click the OK button. You will have created the new variable that you wanted.

Now let us look at what types of analyses that we can do with SPSS. Figure 13 shows the options under the Analyze menu. Most of these analyses require that you have some knowledge of statistics so that you can interpret your results. However, you can find some easy-to-interpret information under the option Descriptive Statistics.

Imagine that you need to know about the distribution of values for the variable “fertility” (which reports the average number of births in a particular country). From the Analyze menu, go to Descriptive Statistics/Descriptives.

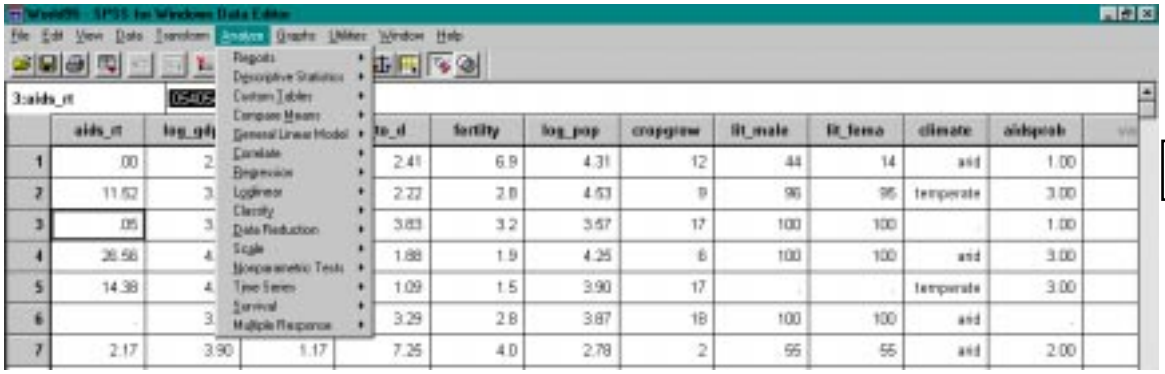


Figure 13

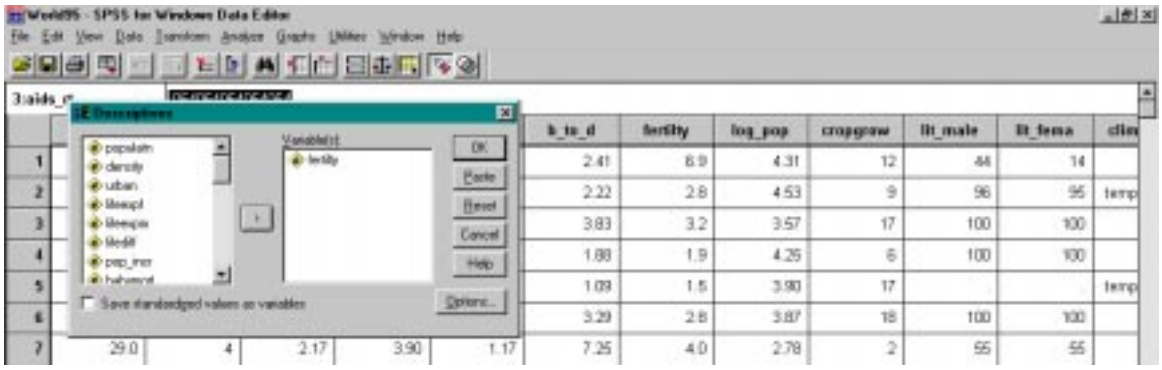


Figure 14

Drag the variable that you want information on from the left box (which lists all the variables) to the right box. In order to do this task: scroll down the list of variables on the left side; find the one you would like to analyze; double-click on it; voila -- it is done! Click OK. SPSS will show you the information in an **output file.** In this case the output will look like Figure 15.

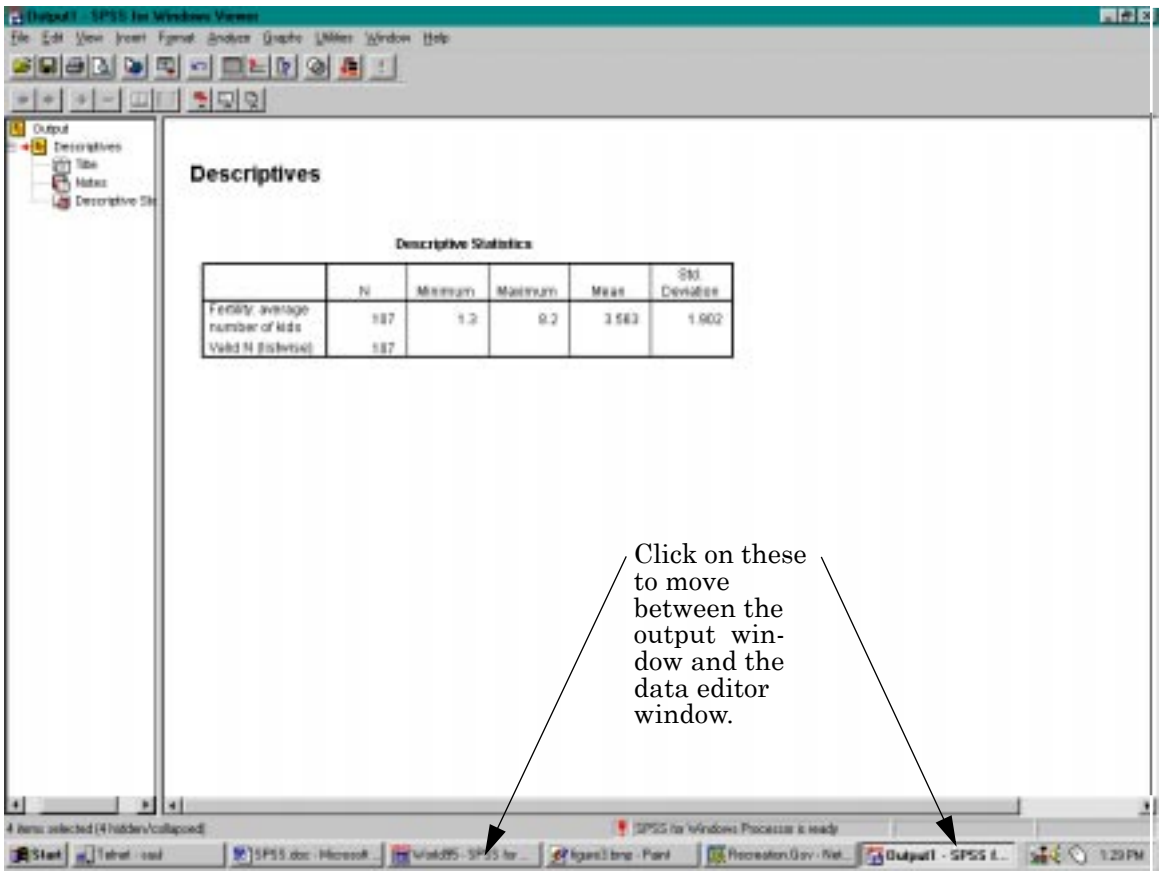


Figure 15

The window in which SPSS posts the results is separate from the data editor window. The output file window is also discrete from the data file. Therefore, if you want to print your output, work from your output window -- not your data editor window. Likewise, if you want to save your output, be sure that you are working from the correct window. SPSS assigns output files the suffix ".spo" and will only recognize files with this suffix as output files. Toggle between the open Program buttons on the task bar.

The output window displays the answers to the statistical questions you asked SPSS. In the case of Figure 15, SPSS had been asked to provide descriptive statistics for the distribution of scores for the fertility variable. The table in the main section of the output reports that 107 countries have valid (non-missing) scores for fertility; in the country with the rate is 1.3 children; in the country with the highest fertility the rate is 8.2 children. All countries measured average a 3.56 fertility rate with a standard deviation of 1.9.

To see the full distribution of scores for this variable, go to Analyze/Descriptive Statistics/Frequencies. The output that SPSS will generate is displayed in Figure 16.

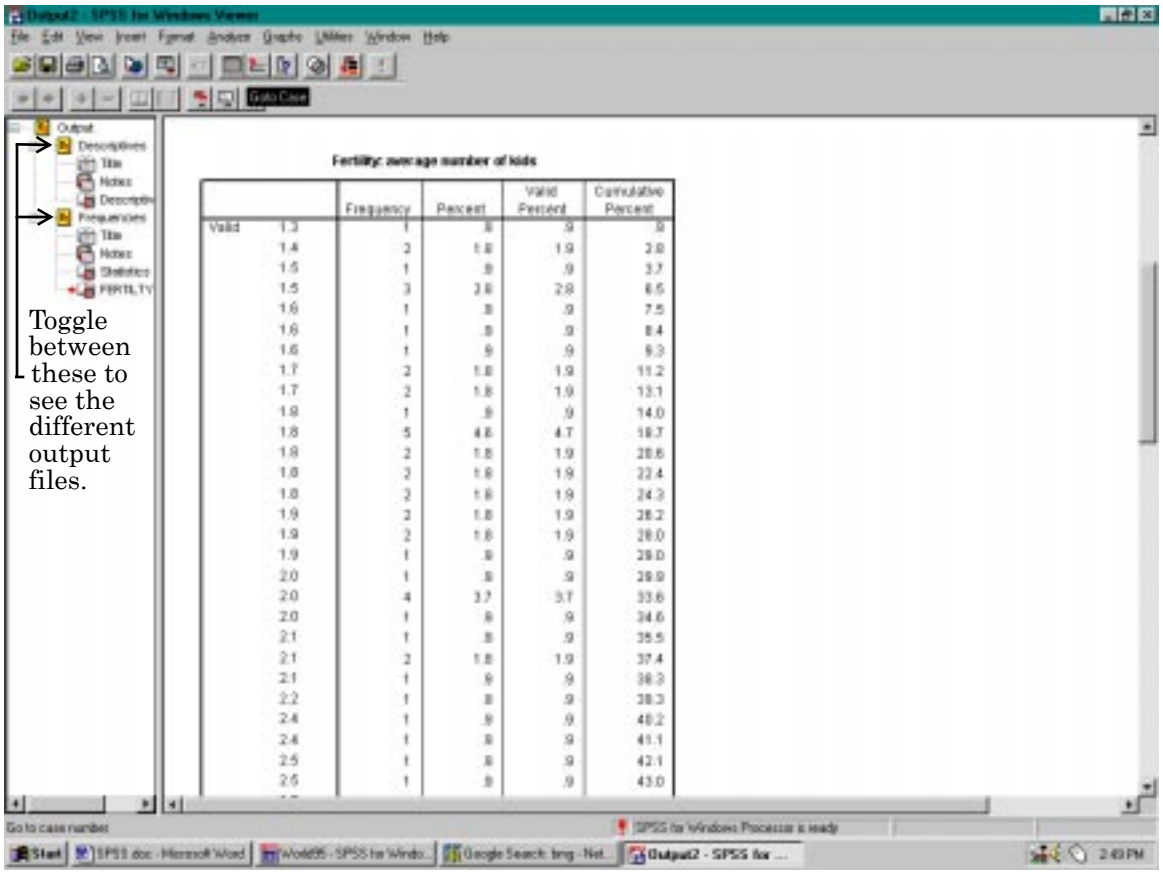


Figure 16

The table in Figure 16 displays five columns of information. The first column displays in ascending order the list of values that appear in the data set for fertility. Since this table has too many values to be displayed on one screen, you must scroll down the window to see more of the table. The second column shows how many cases in the data set resulted in the value to the column to the left. For example, one nation had an average of 1.3 kids per adult woman; two nations had on average 1.4 kids per woman. The third column lets you know what percent of the data set had the value in the first column. The fourth column calculates the same number but only considers the cases that do not have missing data. Finally, the fifth column reports the percentage of cases that have at least the value in the first column.

Note also, in the output window that is Figure 16, the left side has information that is distinct from the table. This information is the **navigator window.** With this window, you can see what findings are displayed in your output. The output file in Figure 16 contains two tables, one for "Descriptives" and one for "Frequencies." To see the output for the descriptives that were run earlier, double-click on the icon next to the word "Descriptives" in the navigator window.

Now, let's do some real analysis. Let's see how SPSS manages a simple ordinary least squares (OLS) regression. We're going to look at regressing fertility on women's literacy rates, on the log of the gross domestic product, and on the death rate of the nation. To perform this operation, go to Analyze/Regression/Linear. You will see the dialog box illustrated in Figure 17.

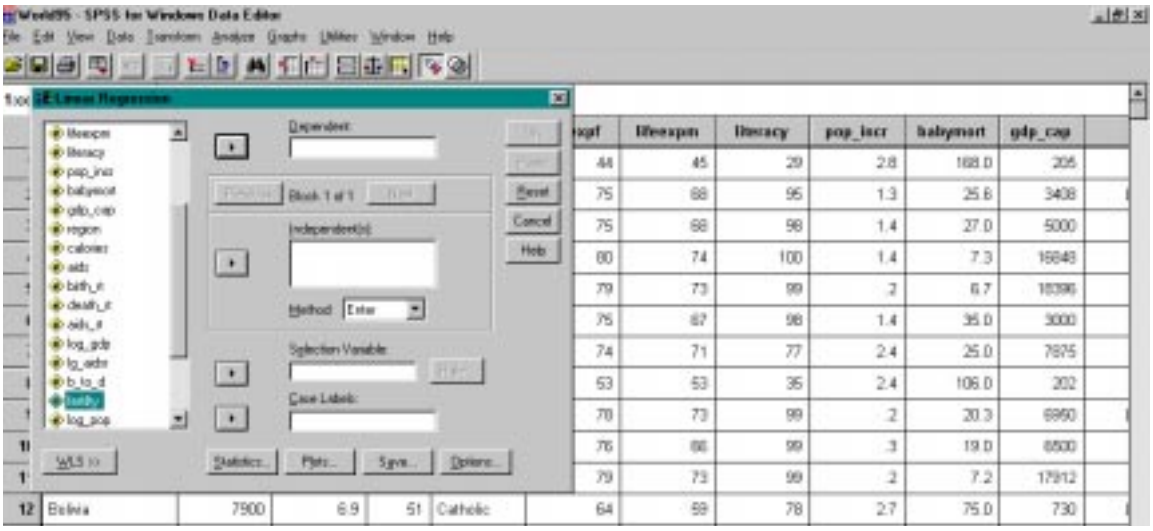


Figure 17

The left side of this dialogue box is a list of the variables contained in the data set. We must select the variables we want to include in the analysis from this list. First, we pick the "Dependent" variable, which in this case is "fertility." Find it in the list of variables and click on it. Now click on the little arrow beside the white rectangle under the heading Dependent. Observe how the variable "fertility" moves into the Dependent list. Next click on the variable "lit_fema;" click on the arrow next to the Independent(s) list. Notice that you've just moved this variable into the Independent(s) list. Repeat this routine with the variables "log_gdp" and "death_rt." SPSS can forthwith run the regression. Click on OK. The results are displayed in Figure 18.

For OLS regression, SPSS generates four tables of results -- as the navigator portion of the output window shows. In fact, SPSS generates too much information to display in one screen. Figure 18 displays three of the four tables of information. You should have a basic knowledge of statistics to interpret the information displayed in these tables.

The first table in Figure 18 (the second table that SPSS generates as part of this output) shows that the "R-squared" statistic indicates that the three independent variables explain 72 percent of the variance in fertility among nations. The final table suggests that two of the three variables in this model are significant predictors of the average fertility rate (at the 95 percent confidence level).

By comparing the "Beta" statistics to each other (also known as the standardized regression coefficient), we can tell that the most powerful predictor of fertility among the three is women's illiteracy. According to the "B" statistic associated with "females who read," for every one percent drop in women's literacy, we can anticipate a .04 increase in the fertility rate. This OLS regression analysis is only one of many that SPSS can perform. For descriptions of how to perform or interpret findings from other analyses, consult the SPSS manuals.

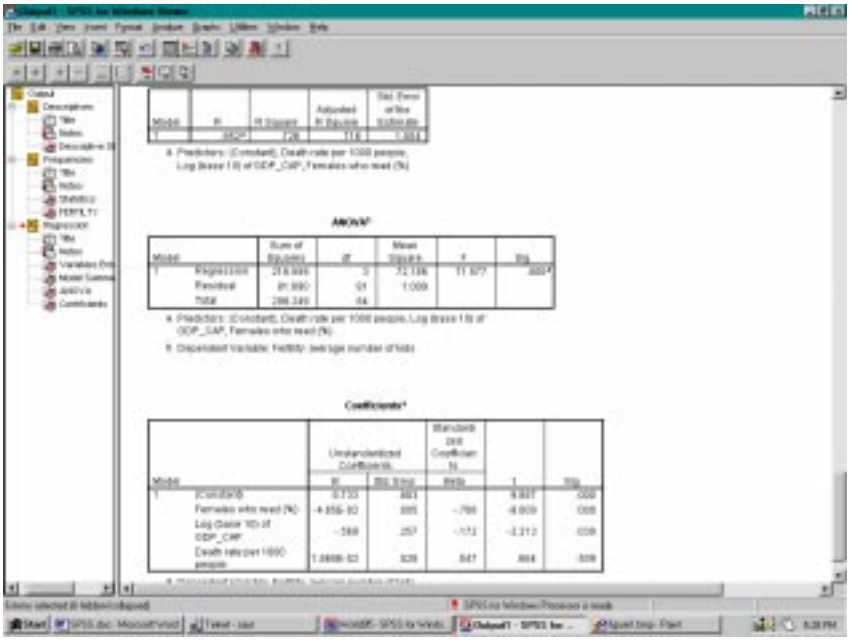


Figure 18

Graphics

SPSS also lets you create (and edit) graphics that allow you to display findings in ways that are visually more accessible than tabular format. The "Graphs" heading on the menu bar contains the options for creating these images. The options under Graphs are displayed in Figure 19.

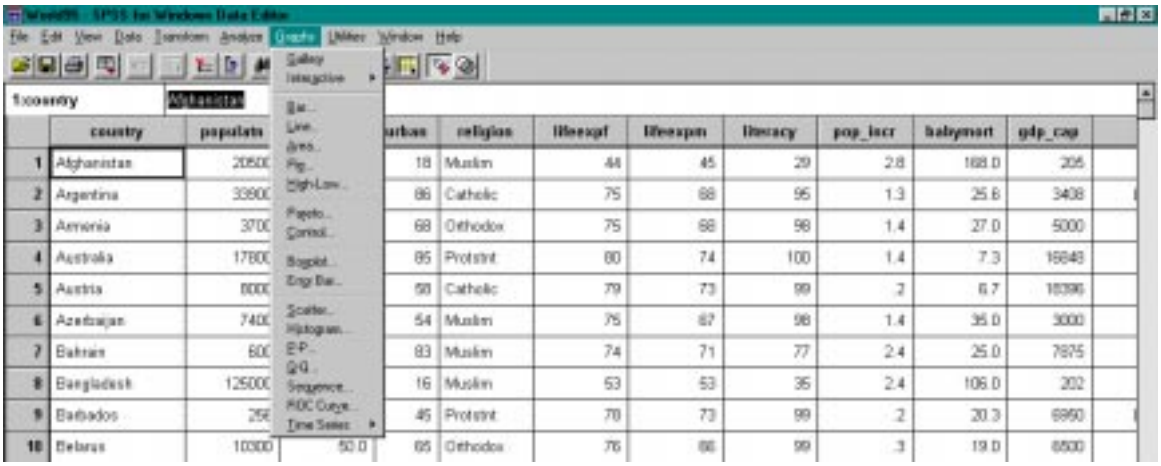


Figure 19

We can display the distribution of the data for the variable "fertility," in the same manner as our frequency analysis. Instead of generating a frequency table we will make a histogram (which is easier to look at). Go to Graphs/Histogram. The Histogram dialogue box should look like Figure 20.

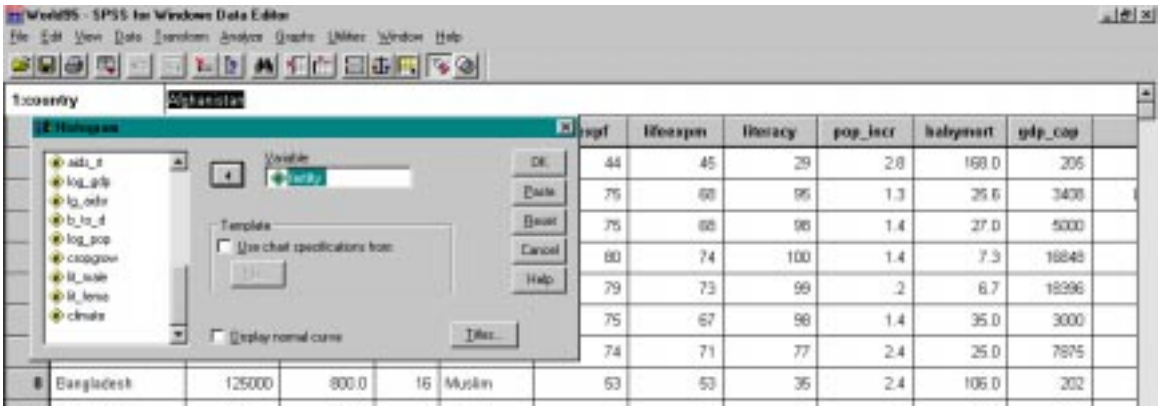


Figure 20

Move the variable "fertility" from the list of variables on the left side to the white rectangle under the heading Variable. Click OK. Figure 21 displays the histogram.

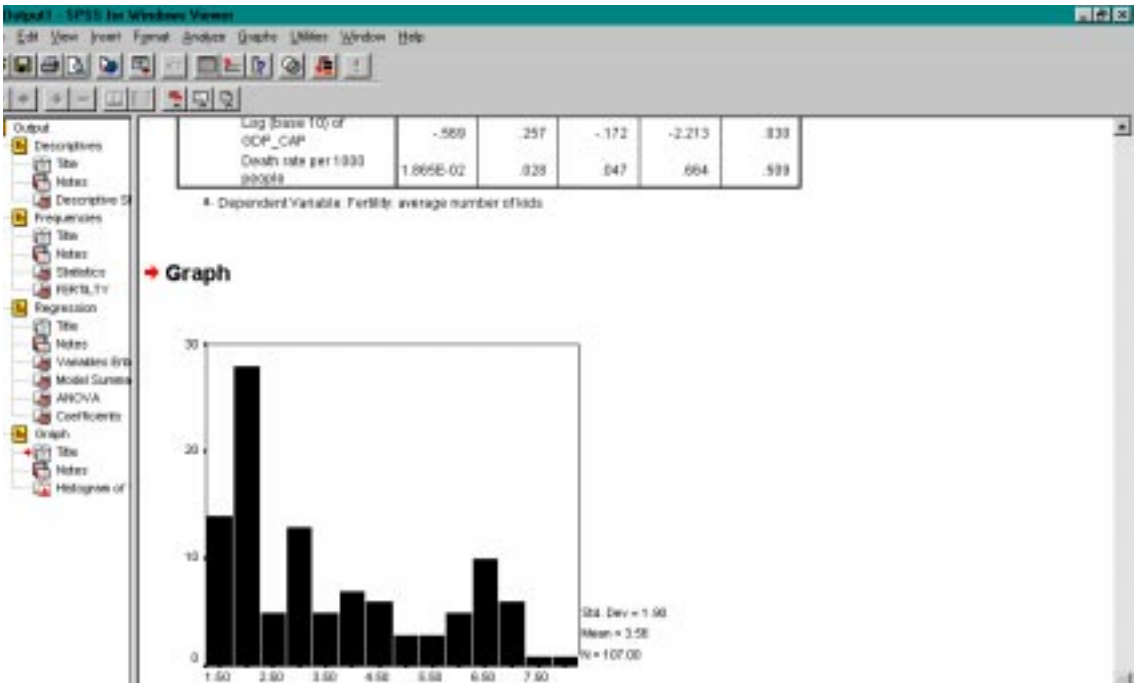


Figure 21

Notice that the histogram is displayed in the output window. You can edit this histogram if you wish. Double-click on it, and you will open up a **chart editor window** with your histogram inside.

From the chart editor window, you can do things like add text to the figure, change the scale of the axis, add lines or other graphic features, or even "fudge" the findings by changing bar heights. For information about how to use the chart editor window, consult the SPSS manuals.

You can also create graphics which use more than one variable. For example, you can make a scatter plot to show the relationship between female literacy and fertility (similar to the OLS analysis which we performed earlier). To do this, go to Graphs/Scatter/Simple/Define. You will see a dialogue box like Figure 22.

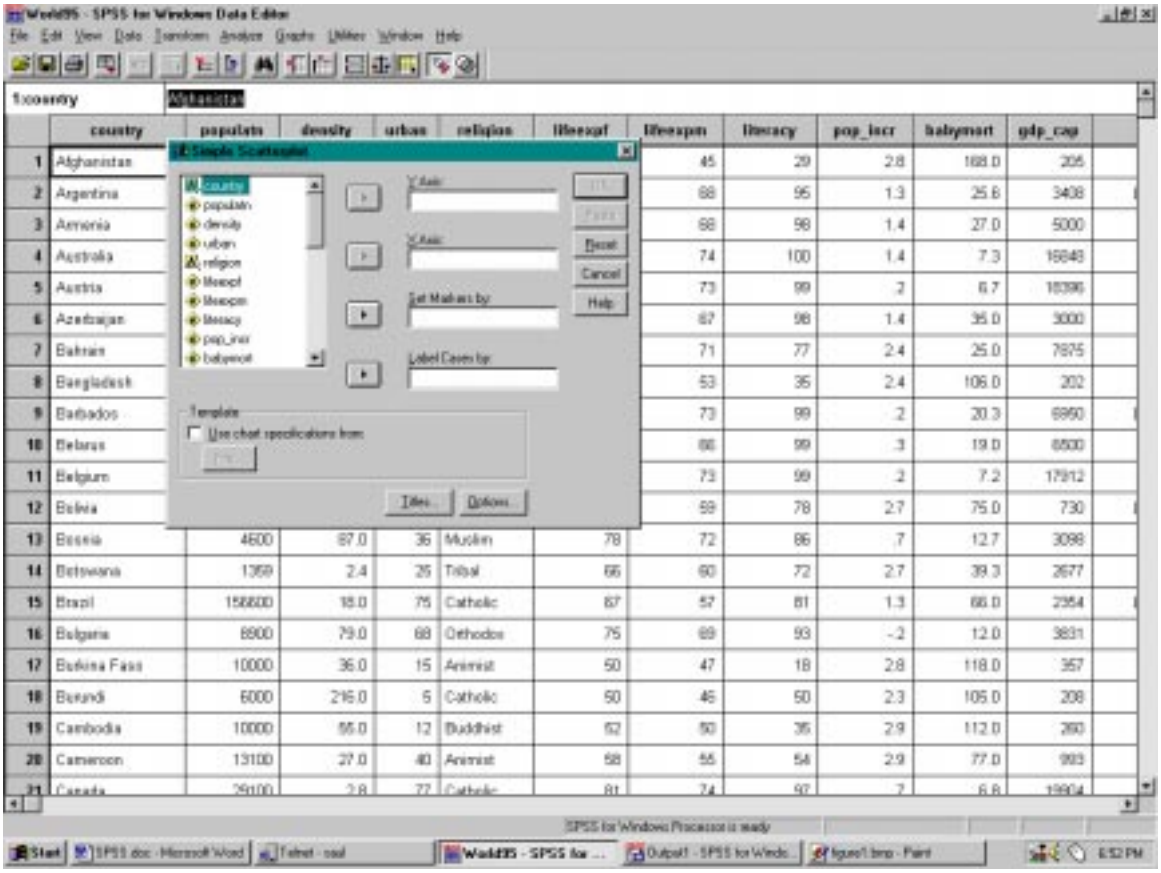


Figure 22

Place the "fertility" variable in the box under the heading "Y-axis"; place the "lit_fema" variable in the box under the heading "X-axis." Click OK. Figure 23 displays the results.

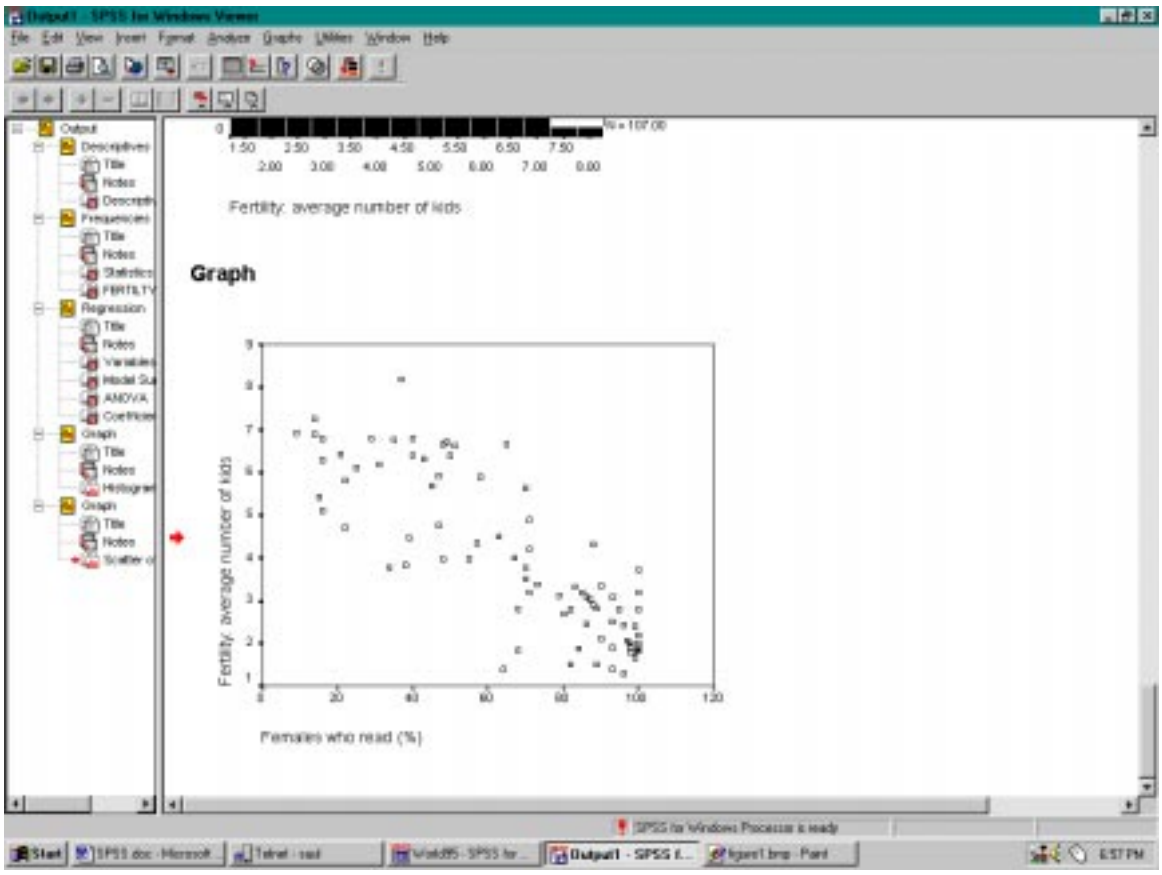


Figure 23

Figure 23 shows a nice linear relationship between the two variables: the more women read, the less children they have. As with the histogram, double-click on the graph to open up a chart editor window to manipulate it. You can easily add the best fit line, or display the regression equation.

Conclusion

If you have a basic knowledge of statistics, SPSS is among the easiest statistical analysis packages to learn and use. Learn it by experimenting with its features, using the **on-line tutorial** that is built into the Help menu, or by consulting the manuals that come with the package. I have three pieces of advice to keep in mind when you're using SPSS in CSSCR's labs.

1) If you are manipulating data files, save your original under one name, and save your updates under a new name in the "C:\temp" directory of your computer. Keep your original file intact in case you make mistakes in the changes you have made to your data set. You will also want to save frequently, because if you give SPSS a very demanding job that requires a lot of memory, you are in danger of crashing your machine and losing your work.

2) When you save or print your file, be very aware of which window you have open. For example, if you want to print your findings, you must have the output window open. If you have your data editor window open and hit the print button you will not print your output, but rather your data file, which could potentially be hundreds (or even thousands) of pages of nothing that you were interested in seeing on paper, tying up the printers, killing trees, costing you money.

3) If you have questions about how to perform a specific task, ask the CSSCR consultants. They should be able to help you.

Good luck with SPSS!