
This document is the first of a series of four modules intended for beginning SPSS users, providing an overview of SPSS for Windows. This first module introduces readers to the SPSS for Windows environment, and discusses how to create or import a dataset, transform variables, manipulate data, and perform descriptive statistics. The second module describes some commonly used inferential statistics, the third module discusses graphical display of output, and the fourth module covers other advanced topics. Throughout these modules, a single dataset, *Employee data.sav*, is used for all examples. This example dataset that is provided with recent versions of SPSS.

Section 1: Overview

Introduction to SPSS

SPSS is a software package used for conducting statistical analyses, manipulating data, and generating tables and graphs that summarize data. Statistical analyses range from basic descriptive statistics, such as averages and frequencies, to advanced inferential statistics, such as regression models, analysis of variance, and factor analysis. SPSS also contains several tools for manipulating data, including functions for recoding data and computing new variables as well as merging and aggregating datasets. SPSS also has a number of ways to summarize and display data in the form of tables and graphs.

Overview of SPSS for Windows

SPSS for Windows consists of five different windows, each of which is associated with a particular SPSS file type. This document discusses the two windows most frequently used in analyzing data in SPSS, the *Data Editor* and the *Output Viewer* windows. In addition, the *Syntax Editor* and the use of SPSS command syntax is discussed briefly. The Data Editor is the window that is open at start-up and is used to enter and store data in a spreadsheet format. The Output Viewer opens automatically when you execute an analysis or create a graph using a dialog box or command syntax to execute a procedure. The Output Viewer contains the results of all statistical analyses and graphical displays of data. The Syntax Editor is a text editor where you compose SPSS commands and submit them to the SPSS processor. All output from these commands will appear in the Output Viewer. This document focuses on the methods necessary for inputting, defining, and organizing data in SPSS.

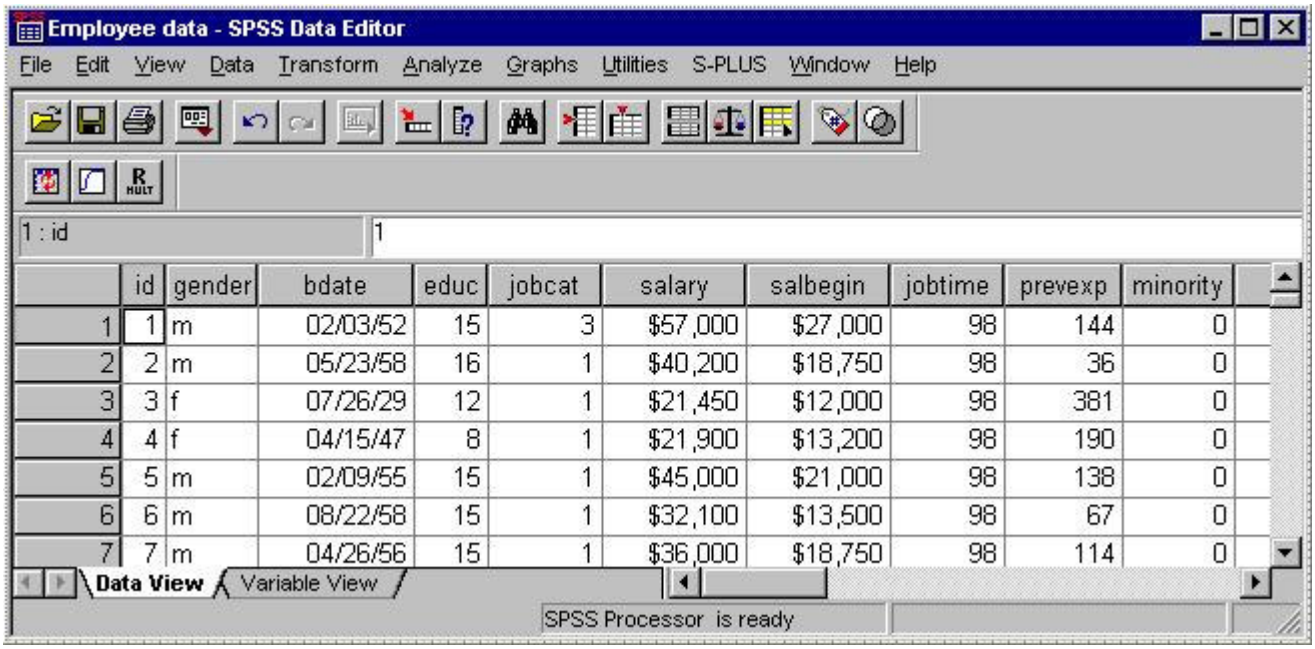
Section 2: Entering Data in SPSS

Starting SPSS

To start SPSS, go to the *Start* icon under Windows 95, Windows 98, Windows 2000, and Windows NT. You should find an SPSS icon under the *Programs* menu item. You can also start SPSS by double-clicking on an SPSS file.

The Data Editor

The Data Editor window displays the contents of the working dataset. It is arranged in a spreadsheet format that contains variables in columns and cases in rows. There are two sheets in the window. The *Data View* is the sheet that is visible when you first open the Data Editor and contains the data. You can access the second sheet by clicking on the tab labeled *Variable View* and while the second sheet is similar in appearance to the first, it is does not actually contain data. Instead, this second sheet contains information about the dataset that is stored with the dataset. Unlike most spreadsheets, the Data Editor can only have one dataset open at a time. However, beginning with version 10.0, you can open multiple Data Editors at one time, each of which contains a separate dataset. Datasets that are currently open are called *working datasets* and all data manipulations, statistical functions, and other SPSS procedures operate on these datasets. The Data Editor contains several menu items that are useful for performing various operations on your data. Here is the Data Editor, containing the *Employee data.sav* dataset:



Data can be directly entered in SPSS, or a file containing data can be opened in the Data Editor. From the menu in the Data Editor window, choose the following menu options (or type Alt+F+O+A):

File
Open...

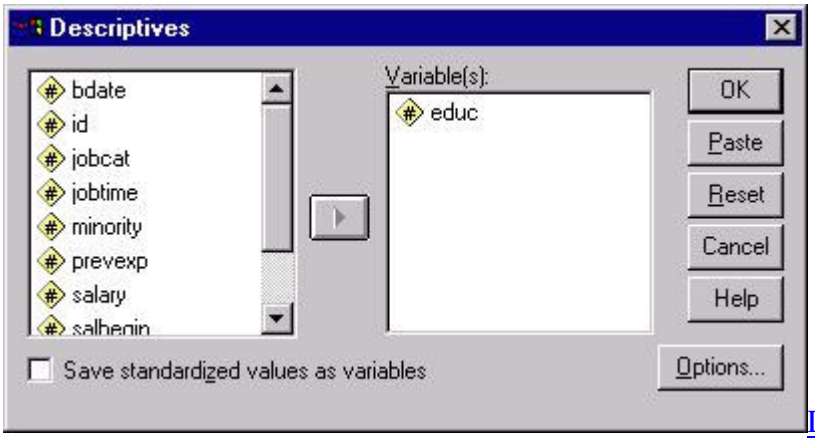
If the file you want to open is not an SPSS data file, you can often use the *Open* menu item to import that file directly into the Data Editor. If a data file is not in a format that SPSS recognizes, then try using the software package in which the file was originally created to translate it into a format that can be imported into SPSS (e.g., tab-delimited data).

The Syntax Editor

Another important window in the SPSS environment is the Syntax Editor. In earlier versions of SPSS, all of the procedures performed by SPSS were submitted through the use of syntax which instructed SPSS on how to process your data. More recent versions contain pull-down menus with dialog boxes that allow you to submit commands to SPSS without ever writing syntax. These SPSS for Windows tutorials focus on the use of the dialog boxes to execute procedures; however, there are a couple of important reasons why you should be aware of SPSS syntax even if you plan to primarily use the dialog boxes. First, not all procedures are available through the dialog boxes. Therefore, you may occasionally have to submit commands from the Syntax Editor. Second, you should be aware of the Syntax Editor so that you can save procedures as syntax to be rerun at a latter date. The dialog boxes available through the pull-down menus have a button labeled **Paste** which will print the syntax for the procedure you are running in the dialog box environment to the Syntax Editor.

Thus, you can easily generate SPSS syntax without typing in the Syntax Editor. This process is illustrated below.

The following dialog box is used to generate descriptive statistics. Here, only the **Paste** button in the dialog box is relevant. The process used for generating descriptive statistics is described later.



D

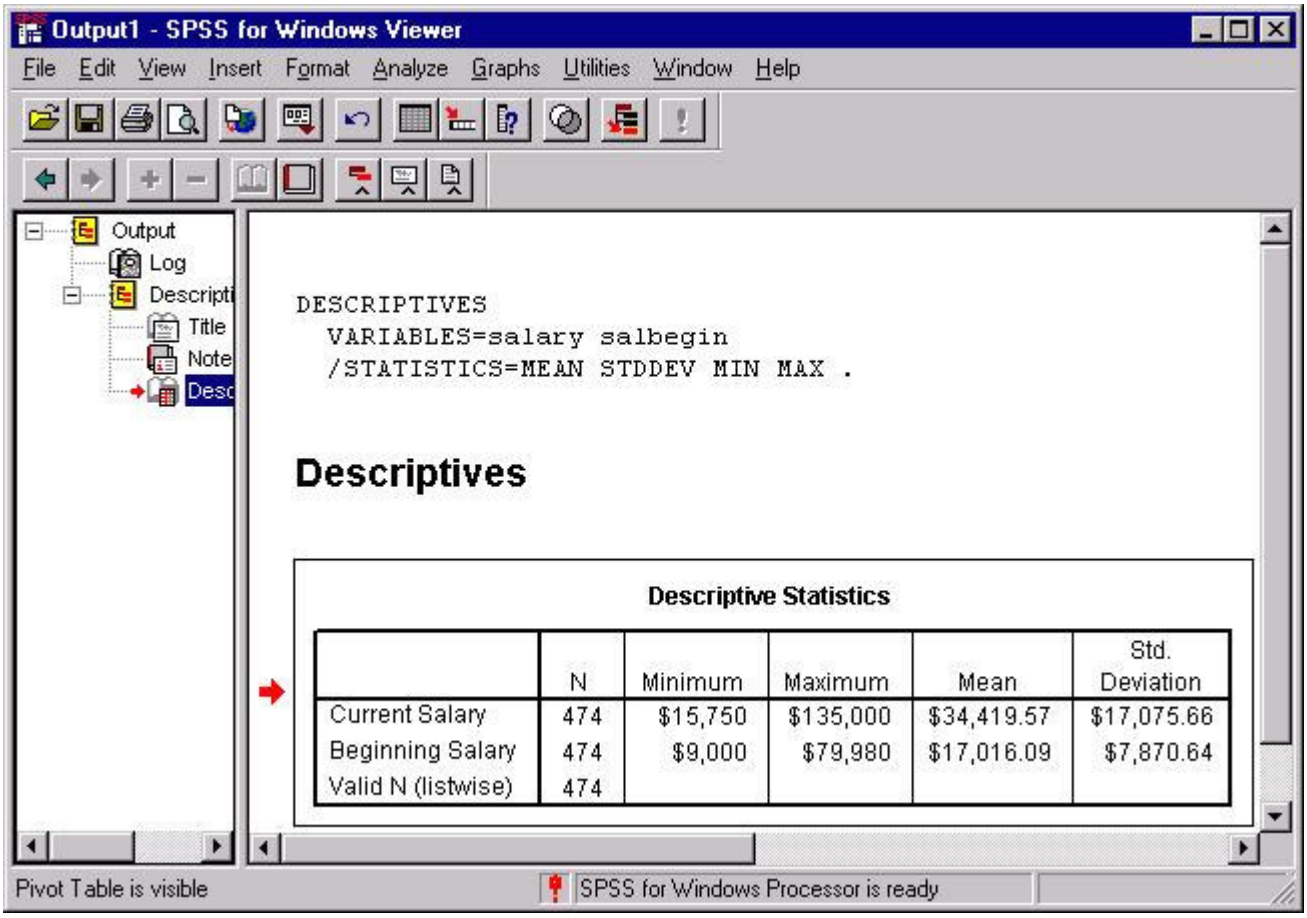
By clicking on the **Paste** button, the procedure that the above dialog box is prepared to run will be written in the form of SPSS syntax to the Syntax Editor. Thus, clicking the **Paste** (or Alt + P) button in the above example would produce the following syntax:

```
DESCRIPTIVES
  VARIABLES=educ
  /STATISTICS=MEAN STDDEV MIN MAX .
```

This syntax will produce exactly the same output as would be generated by clicking the **OK** button in the above dialog box. The syntax that is printed to the Syntax Editor can then be saved and run at a later time as long as the same dataset, or at least a dataset continuing the variables with the same names, is active in the Data Editor window. Saving syntax is useful if you think you may want to rerun your analysis after you add more data or if you want to run the same analysis on another dataset that contains the same variables.

The Output Viewer

All output from statistical analyses is printed to the Output Viewer window as well as other useful information. When you execute a command for a statistical analysis, regardless of whether you used syntax or dialog boxes, the output will be printed in the Output Viewer. Some other output that you may want to have printed to the Output Viewer are command syntax, titles, and error messages. The Output Viewer is shown below:



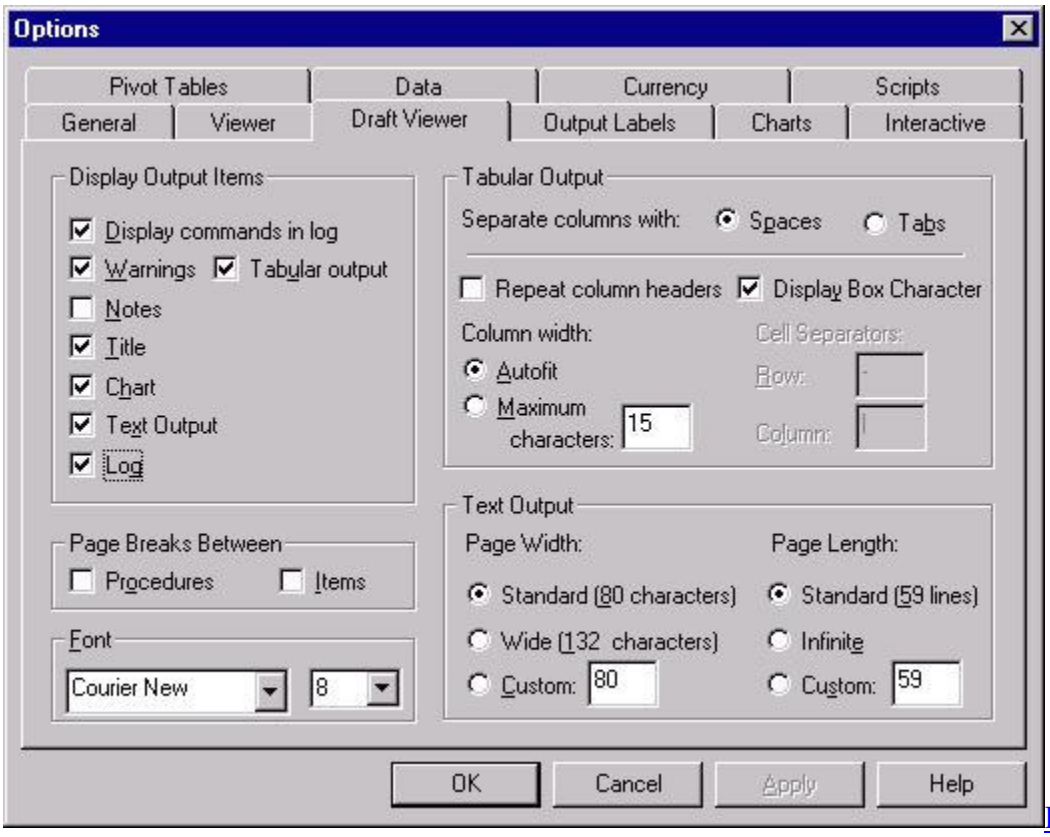
D

The left frame of the Output Viewer contains an outline of the objects contained in the window. For example, the icon labeled *Log* represents the command syntax shown at the top of the figure. Everything under *Descriptives* in the outline refers to objects associated with the descriptive statistics. The *Title* object refers to the bold title *Descriptives* in the output while the highlighted icon labeled *Descriptive Statistics* refers to the table containing descriptive statistics. The *Notes* icon has no referent in the above example, but it would refer to any notes that appeared between the title and the table. This outline can be useful for navigating in your Output Viewer when you have large amounts of output. By clicking on an icon, you can move to the location of the output represented by that icon in the Output Viewer. You can also copy, paste, or delete objects by first highlighting them in the outline and then performing the operation you want.

You can control what is displayed in your output by using the *Options* menu item (or Alt+E+N) on the *Edit* menu:

Edit
Options...

Selecting this option will produce the following dialog box:



D

This figure shows the *Options* dialog box with the *Draft Viewer* tab selected, to choose which options you want to appear in the Output Viewer. Most commands are selected by default. Here, the *Display commands in log* option, normally unselected, was selected so that the command syntax will be written to the log in the Output Viewer. This can be useful for keeping track of which procedures you have executed.

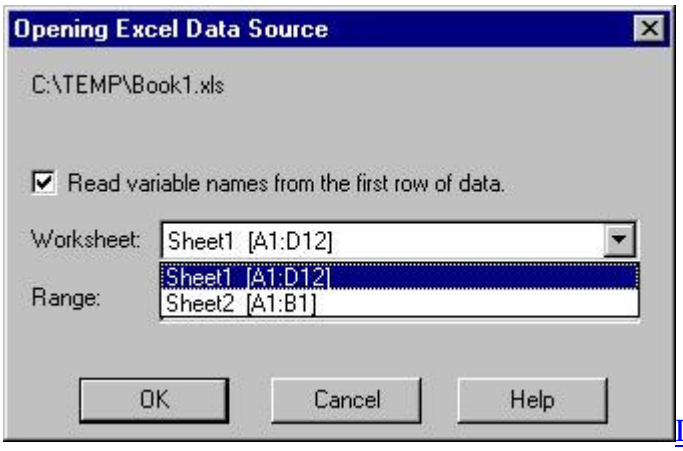
Importing Data From Excel Files

Data can be imported into SPSS from Microsoft Excel and several other applications with relative ease. This document describes a method for importing an Excel spreadsheet into SPSS. If you are working with a spreadsheet in another software package, you may want to save your data as an Excel file, then import it into SPSS. If you have a spreadsheet that is arranged in a database format (e.g., you have several tables in your Workbook that are related through identification fields), there is another method for importing Excel file that you might consider that will merge tables within your database as part of the import procedure. It is described in the fourth module of this tutorial series, Data Manipulation and Advanced Topics, in the Database Capture section.

To open an Excel file, select the following menu options from the menu in the Data Editor window in SPSS (or Alt+F+O+A):

File
Open...

First, select the desired location on disk using the *Look in* option. Next, select Excel from the *Files of type* drop-down menu. The file you saved should now appear in the main box in the *Open File* dialog box. You can open it by double-clicking on it. You will be presented with one more dialog box:



D

This dialog box allows you to select a spreadsheet from within the Excel Workbook. The drop-down menu in the example shown above offers two sheets from which to choose. As SPSS only operates on one spreadsheet at a time, you can only select one sheet from this menu. This box also gives you the option of reading variable names from the Excel Workbook directly into SPSS. Click on the *Read variable names* box to read in the first row of your spreadsheet as the variable names. If the first row of your spreadsheet does indeed contain the names of your variables and you want to import them into SPSS, these variables names should conform to SPSS variable naming conventions (eight characters or fewer, not beginning with any special characters). If you do not have variable names, use the procedure described below in Creating and Defining Data to add variable names to your dataset after you have imported your data into SPSS. You should now see data in the Data Editor window. Check to make sure that all variables and cases were read correctly. Next, save your dataset in SPSS format by choosing the *Save* option in the *File* menu.

If you are using a version of SPSS that was released prior to SPSS 10.0, there are a few additional steps that are necessary for opening an Excel spreadsheet directly into SPSS. You will need to save the file as an Excel version 4.0 or lower *Excel Worksheet* which is a file containing a single spreadsheet. More recent versions of Excel use the *Excel Workbook* format, which contains several spreadsheets. Because SPSS only allows one dataset to be active at any given time, it can read Excel spreadsheets which are a single spreadsheet but not Excel Workbooks which are several spreadsheets. To save an Excel file as a Excel Worksheet, choose the following from the menus (or Alt+F+A):

File
Save As...

After you assign the file a new name in the *File name* box and choose a location on disk with the *Save in* box, make sure you select the *Microsoft Excel 4.0 Worksheet (*.xls)* option from the *Save as type* pull-down menu. You will receive the following warning: "The selected file type does not support workbooks that contain multiple sheets." This warning is letting you know that only the visible worksheet in the Excel Workbook will be saved --not all of the sheets in the workbook. Click the **OK** button here if the spreadsheet you want to import into SPSS is currently the visible sheet in your Workbook. After saving the file in this format, be sure to close the file in Excel because SPSS cannot open a file that is currently open in Excel. At this point, the file is ready to be opened in SPSS and can be opened using the procedures for SPSS 10.0 described above with the exception that you will not be offered the option to choose from available sheets as there is only a single sheet in the Worksheet.

Importing data from ASCII files

Data are often stored in an ASCII file format, alternatively known as a text or flat file format. Typically, columns of data in an ASCII file are separated by a space, tab, comma, or some other character. SPSS 9.0 has a *Text Import Wizard* that will help you import data in an ASCII file format. The Text Import Wizard will open automatically when a ASCII file (a file with a *.txt* or *.dat* extension) is opened using the *Open* option in the *File* menu. If the data file you want to open does not have a *.txt* or *.dat* extension but you know that it is an ASCII file, then you can open the data file by opening the Data Import Wizard from the *File* menu (or Alt+F+R):

File
Read Text Data

The Text Import Wizard will first prompt you to select a file to import. After you have selected a file, you will go through a series of dialog boxes that will provide you with several options for importing data. Once you have imported your data and checked it for accuracy, be sure to save a copy of the dataset in SPSS format by selecting the *Save* or *Save As* options from the *File* menu:

File
Save
Save As...

(or Alt+F+S or Alt+F+A)

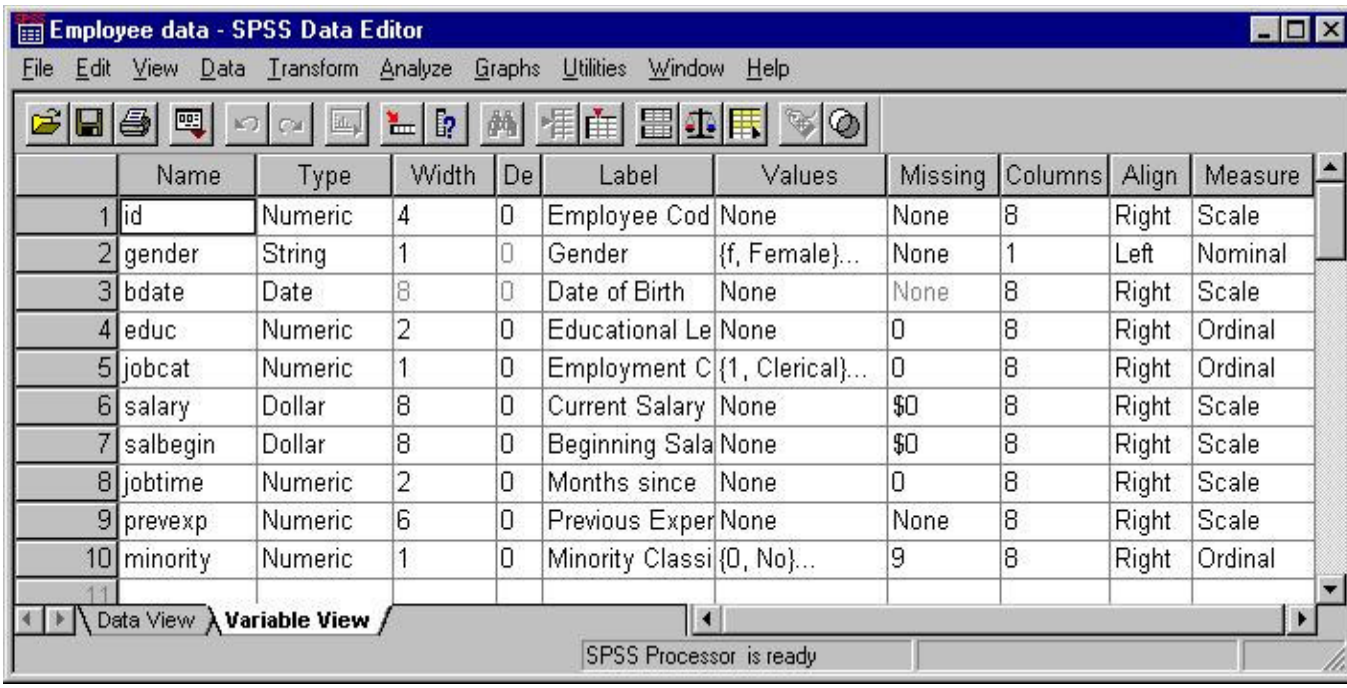
Section 3: Creating and Modifying Data in SPSS

Creating and Defining Variables

After data are in the Data Editor window, there are several things that you may want to do to describe your data. Before describing the process for defining variables, an important distinction should be made between two often confused terms: *variable* and *value*. A variable is a measure or classification scheme that can have several values. Values are the numbers or categorical classification representing individual instances of the variable being measured. For example, a variable could be created for job classification status. Each individual in the dataset would be assigned a value representing their job classification. For instance, we could assign custodians the value 1, clerks the value 2, and managers the value 3.

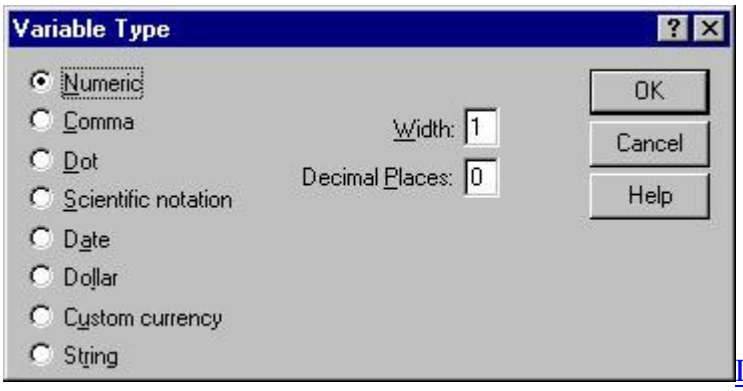
One reason to define information about your variables is to help you interpret the output. For example, if you have a variable representing employment categories that is coded as either 1, 2, or 3 for various employment categories, say, clerical custodial and managerial, it may be unwieldy to read the output if you are constantly trying to remember which number represents the which categories. One advantage of defining variables is that these values can be assigned labels that will appear in your output, thus making it much easier to interpret. Another aspect of defining variable information is to provide SPSS with information about the type of data in your dataset, which is often critical for SPSS to correctly process analyses.

You can define information about your variables by clicking the *Variable Information* tab. Doing so will bring the *Variable Information* sheet to the foreground. You can also access this sheet by double-clicking one of the gray boxes at the top of the columns in the Data Editor. The advantage of the second method is that it takes you to the row for the variable whose column head you clicked. Finally, you can also use the keystroke Ctrl+T to toggle between the windows. Regardless of the method you use, you will see a spreadsheet organized as the one below:



D

Many of the cells in the spreadsheet contain hidden dialog boxes that can be activated by clicking on a cell. If you see a gray box appear on the right side of the cell when you first click on the cell, this indicates that there is a hidden dialog box which can be accessed by clicking on that box. For example, clicking on the box in the cell for the *Type* column for the variable jobcat produces the following dialog box:

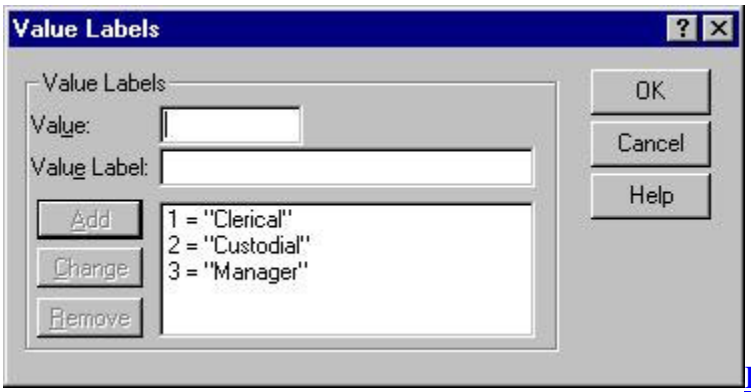


D

This box allows you to define the type of data for variables. For example, you will be presented with *Numeric*, *String*, and *Date* options among others. Thus, if you want to define the variable *jobcat*, a variable representing employment category as a string variable rather than the default variable type, numeric, you would click on the the cell in the *jobcat* row and the *Type* column, then click the gray box to produce the *Variable Type* dialog box. Here, you would choose the *String* option.

The *Missing Values* column allows you to define which values of a variable should be treated as missing data. The *Label* column is used to define labels for variables. The

Values column is used to assign labels to the particular values of a variable. For example, the following dialog box shows a variable that has been assigned the values 1, 2, and 3 for the labels *Clerical*, *Custodial*, and *Manager*.



D

To define variables as shown above, you should first enter the value (e.g., 1) in the box labeled *Value*, then enter the label associated with that value (e.g., *Clerical*), and click on the **Add** button. Repeat this process for each value you want to label.

Inserting and Deleting Cases and Variables

You may want to add new variables or cases to an existing dataset. The Data Editor provides menu options that allow you to do that. For example, you may want to add data about participants' ages to an existing dataset. To insert a new variable, click on the variable name to select the column in which the variable is to be inserted. To insert a case, select the row in which the case is to be added by clicking on the row's number. Clicking on either the row's number or the column's name will result in that row or column being highlighted. Next, use the insert options available in the *Data* menu in the Data Editor:

Data
Insert Variable
Insert Case

(or Alt+D+V for Variables and Alt+D+I for Cases)

If a row has been selected, choose *Insert Case* from the *Data* menu; if a column has been selected, choose, *Insert Variable*. This will produce an empty row or column in the highlighted area of the Data Editor. Existing cases and variables will be shifted down or to the right, respectively.

You may want to delete cases or variables from a dataset. To do that, select a row or column by highlighting as described above. Next, use the **Delete** key to delete the highlighted area. Or you can use the *Delete* option in the *Edit* menu to do it.

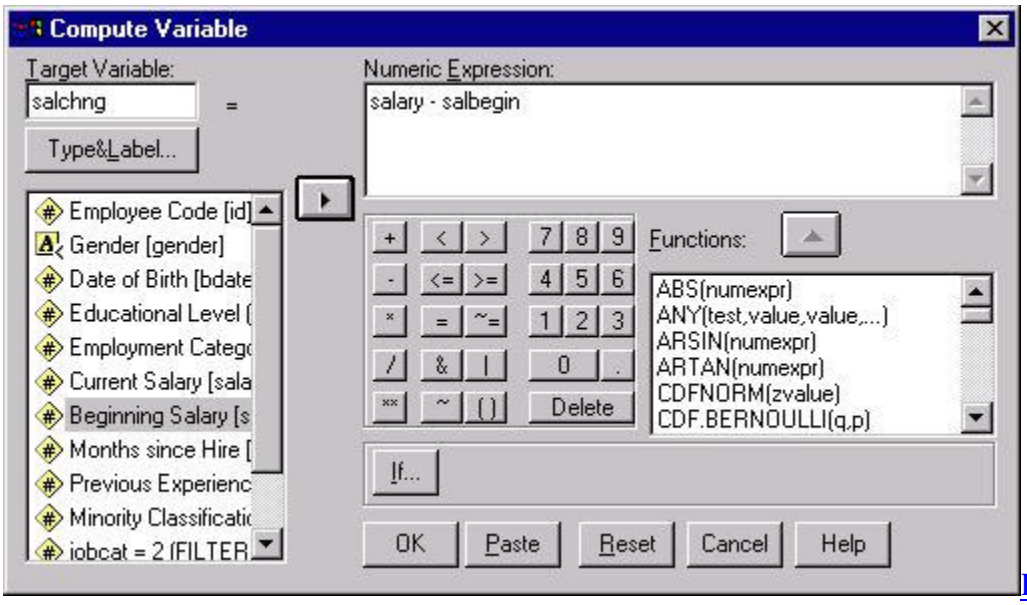
Computing New Variables

You may want to modify the values of the variables in your datasets. For example, if a dataset contained employees' salaries in terms of their beginning and current salaries but you wanted the difference between starting salary and present salary, a new variable could be computed by subtracting the starting salary from the present salary. In other situations, you may also want to transform an existing variable. For example, if data were entered as months of experience and you wanted to analyze data in terms of years on the job, then you could recompute that variable to represent experience on the job in numbers of years by dividing number of months on the job by 12.

Both variables that are created as a numeric expression of existing variables and variables whose values are modified by an operation can be computed using the *Compute* option available from the menu in the Data Editor (or Alt+T+C):

Transform
Compute...

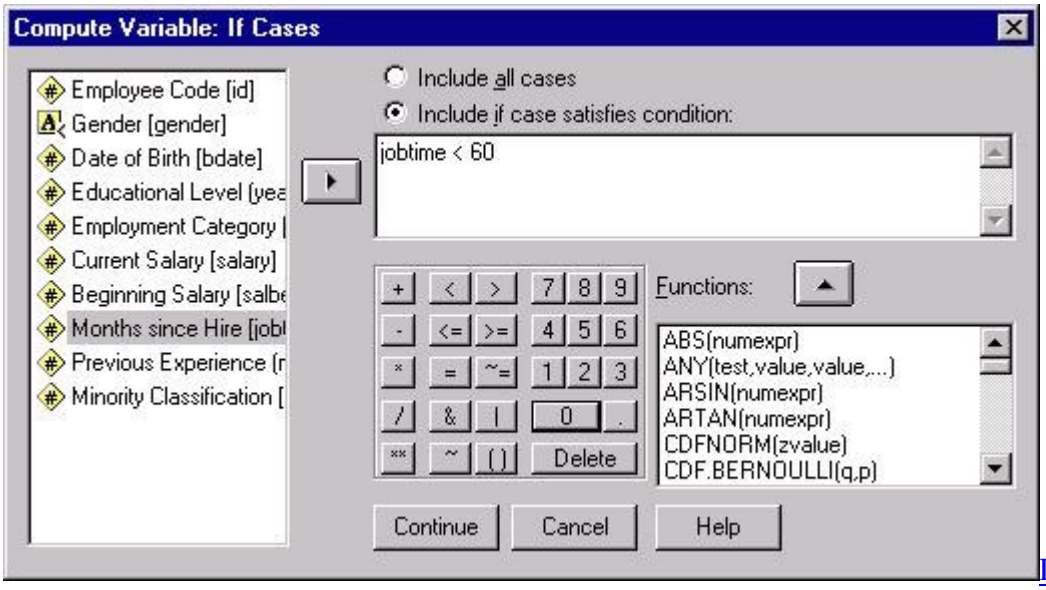
This will result in the following dialog box:



D

To create a new variable, type its name in the box labeled *Target Variable*. Alternatively, you may want to modify the value of an existing variable, in which case you would type its name in the box labeled *Target Variable*. In both cases, the expression defining the variable being computed will appear in the box labeled *Numeric Expression*. This expression can either be typed into the box directly, or you can use the buttons located below the *Numeric Expression* box to input values or operators. The example shown above demonstrates the computation of a new variable. This new variable, *salchng*, will be the difference between an employee's current salary and the employee's beginning salary. The new variable will appear in the rightmost column of the working dataset.

Variables can also be computed conditionally. For instance, if, in the above example, you were only interested in the change in salaries for people who began working for the company within the last five years, you could create a condition that would compute a new variable only if an employee had begun employment within the last five years. To do this, first click on the button labeled **If**, which will produce the following dialog box:



D

First, click on the button labeled, *Include if case satisfies condition* to activate the gray areas of the dialog box. Then, specify the condition for computing a new variable in the input box at the top right of this dialog box. You can either type in the condition or click on variables in the variable list on the left side of the dialog box and use the buttons on the bottom middle of the dialog box. Variables can be moved to the conditional box by selecting by clicking on the variable's name, then clicking the arrow button between the two boxes. Clicking on the buttons on the bottom left of the dialog box will cause the character on the button to be displayed at the location of the cursor in the input box.

The above example illustrates the definition of a condition that requires cases to have less that five years (60 months) experience in order to be included in the computation of the new variable: the variable *jobtime* represents the number of months since an employee has been hired. Thus, only cases which have fewer than sixty months, or five years, since they were hired will be included. Click the **Continue** button to return to the previous dialog box.

Recoding Variables

You can also modify the values of existing variables in your dataset. For example, if a dataset contains a variable that classifies an employee's status in three categories, but for a particular analysis you want to combine two of these classifications into a single

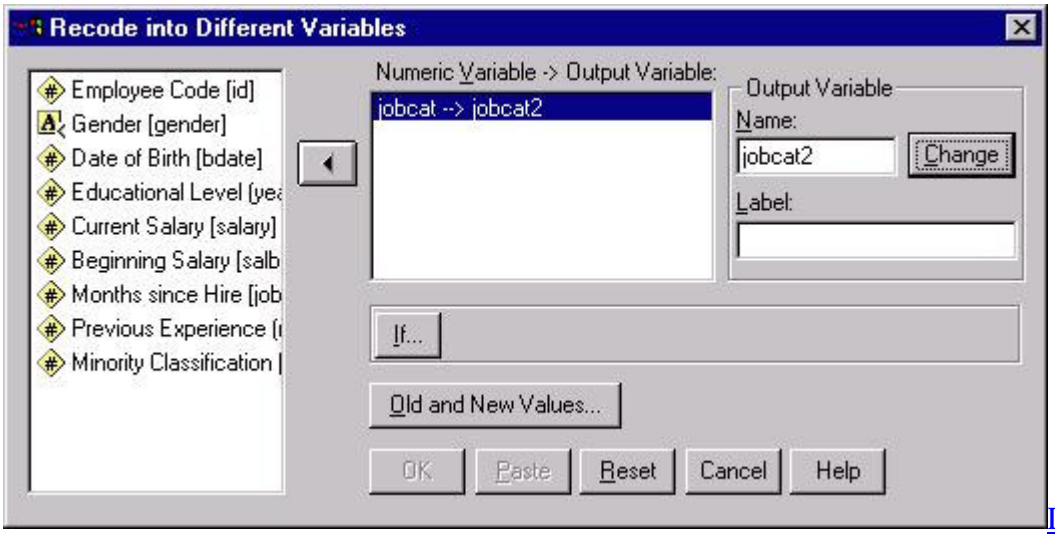
category, then two of the values would need to be recoded into a single value so that there are two total groups.

The *Recode* option (or Alt+T+R) is available from the menu in the Data Editor:

Transform
Recode

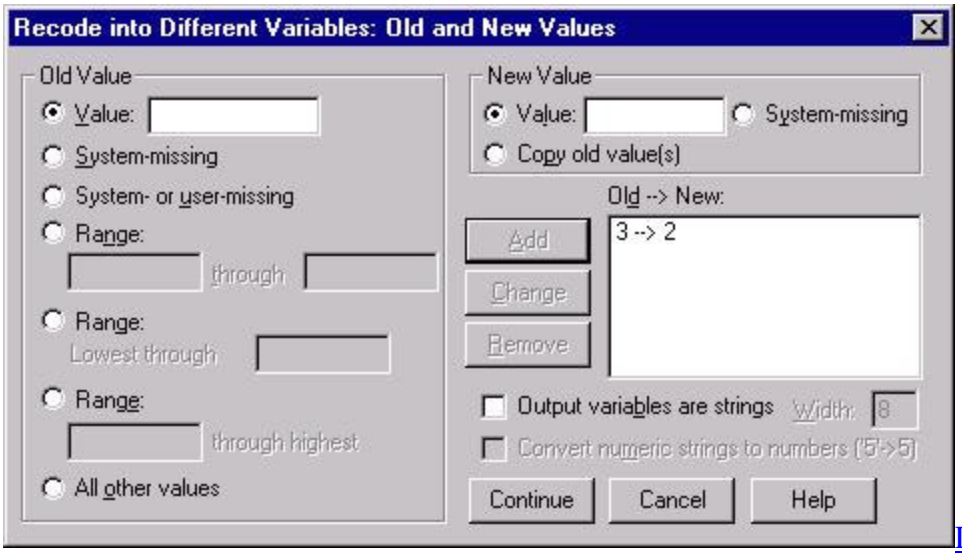
Additionally, there are two options for recoding variables in the *Recode* submenu. The *Into Same Variables* (Alt+T+R+S) option changes the values of the existing variables, whereas the *Into Different Variables* (Alt+T+R+D) option is used to create a new variable with the recoded values. Both options are essentially the same, except that recoding into a different variable requires you to supply a new variable name. You should use the *Into Different Variables* option, because you may change your mind about your recoding scheme at a later date. Thus, if your do change you mind, you still have the original values.

The following example illustrates the use of the *Recode* option to recode values into a new variable. When that option is selected from the menu, the following dialog box will appear:



D

First, a variable from the existing dataset should be selected by clicking on that variable, then clicking the arrow button in the middle of the dialog box. This will result in the selected variable being displayed in the box labeled, *Numeric Variable -> Output Variable*. Next, you must supply the name of the new variable, and optionally you can supply a label for the new variable. After a new variable name has been supplied, click on the button labeled *Old and new Values.* This will result in the following dialog box:



D

The above dialog box is the same regardless of whether you are recoding values into the same variable or creating a new variable. The original value of the variable being recoded is entered in the box labeled *Old Value*, and the new value is entered in the box labeled *New Value*. After values are entered in these boxes, click on the button labeled **Add** to complete the recode process.

Continuing with the above example, a variable with three values, such as *jobcat*, could be recoded into a variable with two values by recoding one of the values. In the example dataset, *jobcat* has three values: 1, 2, and 3. If the goal were to combine cases with the values 2 and 3, this could be accomplished by recoding cases with the value 3 into 2's. For example, by entering 3 in the box labeled *Old Value* and entering 2 in the box labeled *New Value* then clicking **Add**, all of the cases labeled 3 would take on the value 2. This can be repeated for as many of the values as necessary.

Values can also be recoded conditionally. The process for recoding values on the basis of a condition is essentially identical to the process for conditionally computing new variables discussed in the previous section: when you click on the **If** button in the main *Recode* dialog box, the same dialog box that was obtained from clicking **If** in the the *Compute* dialog box will appear with the same options.

Sorting Cases

Sorting cases allows you to organize rows of data in ascending or descending order on the basis of one or more variable. For example, the data could be sorted by job category so that all of the cases coded as job category 1 appear first in the dataset, followed by all of the cases that are labeled 2 and 3 respectively. The data could also be sorted by more than one variable. For example, within job category, cases could be listed in order of their salary. The *Sort Cases* (or Alt+ D+O) option is available under the *Data* menu item in the Data Editor:

Data
Sort Cases...

The dialog box that results from selecting *Sort Cases* presents only a few options:

D

To choose whether the data are sorted in ascending or descending order, select the appropriate button. You must also specify on which variables the data are to be sorted. The hierarchy of such a sorting is determined by the order in which variables are entered in the *Sort by* box. Variables are sorted by the first variable entered, then the next variable is sorted within that first variable. For example, if *jobcat* was the first variable entered, followed by *salary*, the data would first be sorted by *jobcat*, then, within each of the job categories, data would be sorted by *salary*.

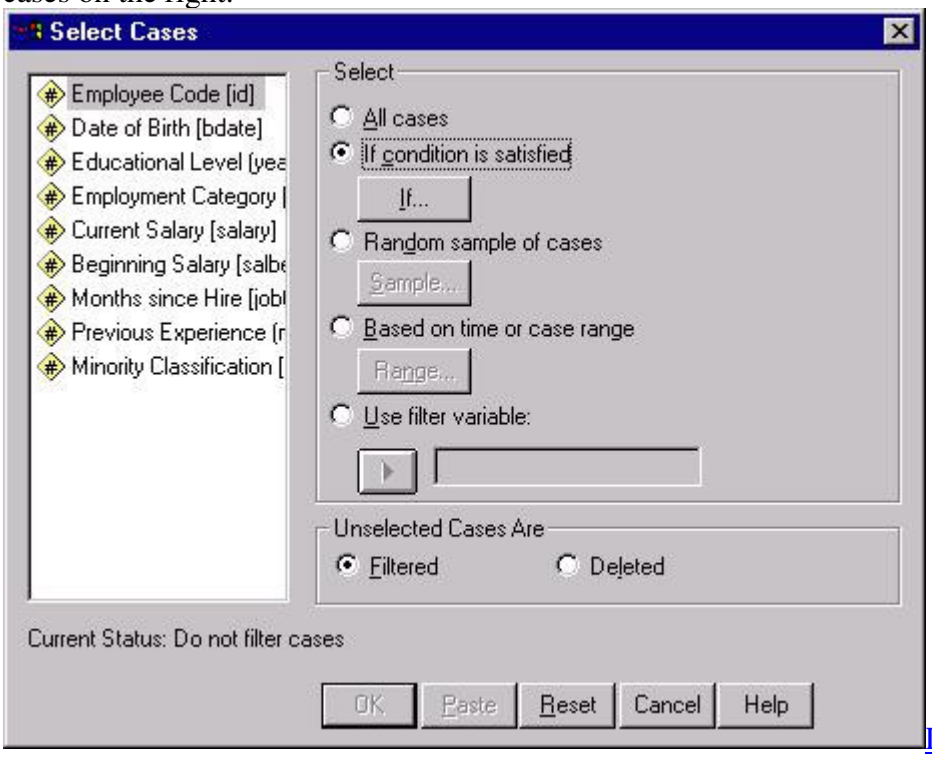
Selecting Cases

You can analyze a specific subset of your data by selecting only certain cases in which you are interested. For example, you may want to do a particular analysis on employees only if the employees have been with the company for greater than six years. This can done by using the *Select Cases* menu option, which will either temporarily or permanently remove cases you didn't want from the dataset. The *Select Cases* option (or Alt+D+C) is available under the *Data* menu item:

Data
Select Cases...

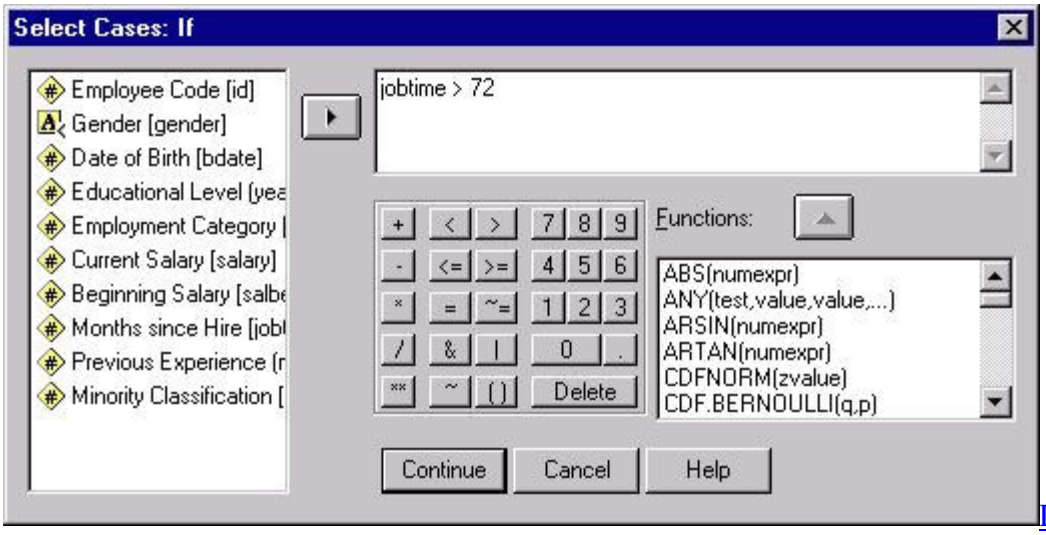
Selecting this menu item will produce the following dialog box. This box contains a list of the variables in the active data file on the left and several options for selecting

cases on the right.



D

Selecting one of these options will produce a second dialog box that prompts you for the particular specifications in which you are interested. For example, selecting the *If condition is satisfied* option and clicking on the **If** button (as was done in the example) results in a second dialog box, as shown below. The portion of the dialog box labeled *Unselected Cases Are* gives you the option of temporarily or permanently removing data from the dataset. The *Filtered* option will remove data from subsequent analyses until the *All Cases* option is reset, at which time all cases will again be active and used in further analyses. If the *Deleted* option is selected, the unselected cases will be removed from the working dataset. If the dataset is subsequently saved, these cases will be permanently deleted.

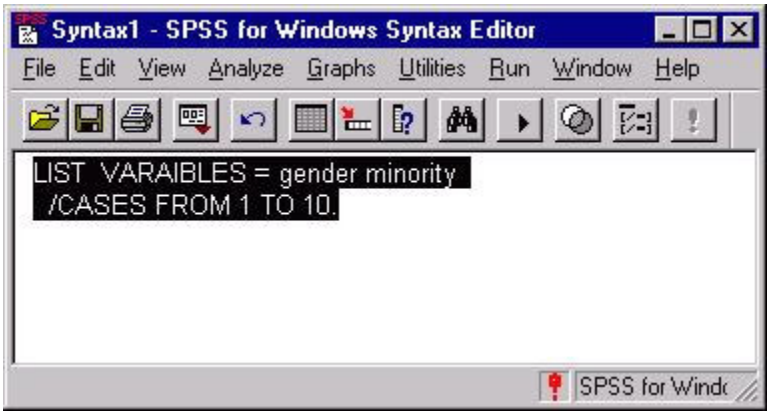


D

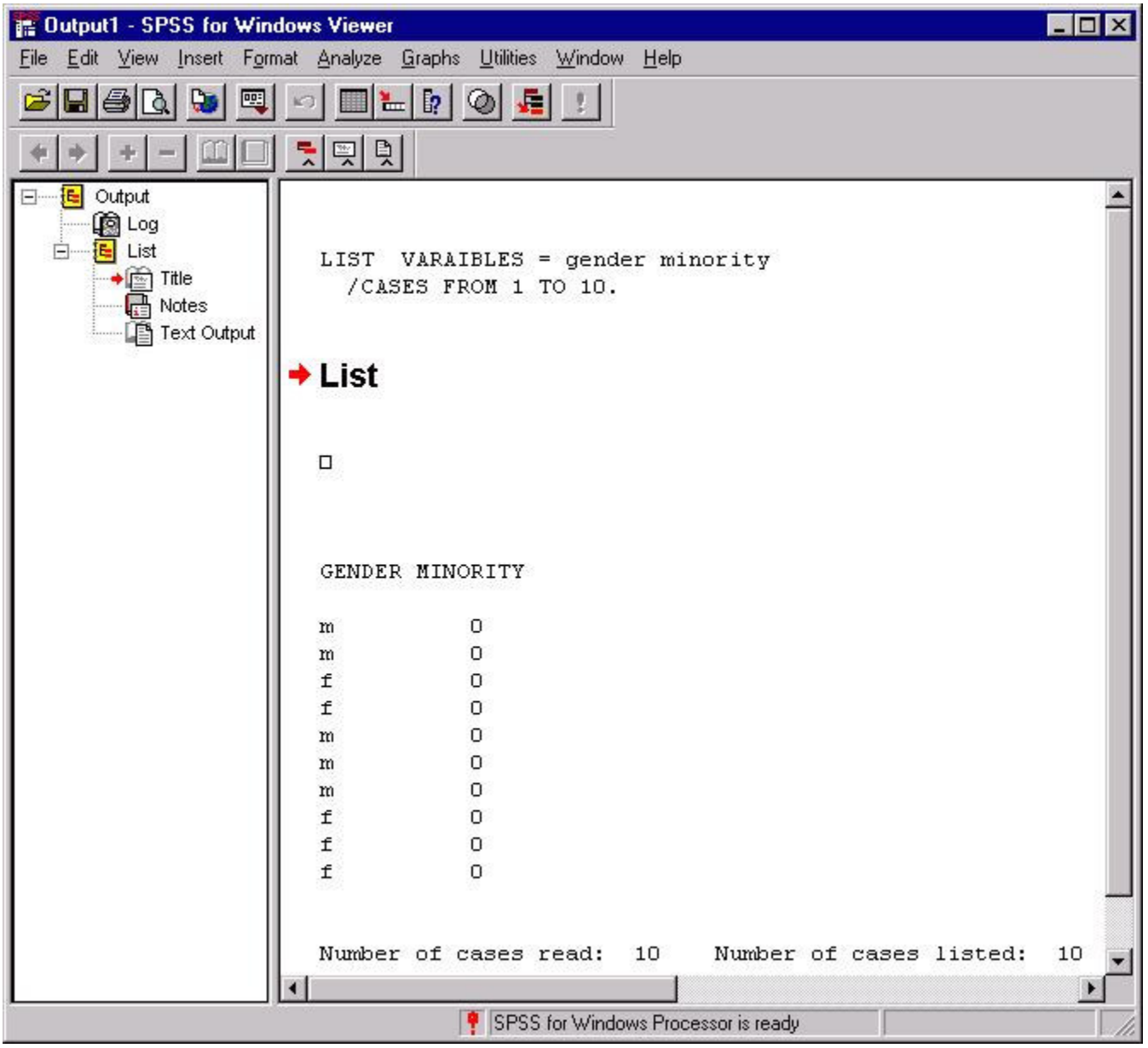
The above example selects all of the cases in the dataset that meet a specific criterion: employees that have worked at the company for greater than six years (72 months) will be selected. After this selection has been made, subsequent analyses will use only this subset of the data. If you have chosen the *Filter* option in the previous dialog box, SPSS will indicate the inactive cases in the Data Editor by placing a slash over the row number. To select the entire dataset again, return to the *Select Cases* dialog box and select the *All Cases* option.

Listing Cases

You may sometime want to print a list of your cases and the values of variables associated with each case, or perhaps a list of only some of the cases and variables. For example, if you have two variables that you want to examine visually, but this cannot be done because they are at very different places in your dataset, you could generate a list of only these variables in the Output Viewer. The procedure for doing this cannot be performed using dialog boxes and is available only through command syntax. The syntax for generating a list of cases is shown in the Syntax Editor window below. The variable names shown in lower case below instruct SPSS which variables to list in the output. Or, you can type in the command **ALL** in place of variables names, which will produce a listing of all of the variables in the file. The subcommand **/CASES FROM 1 TO 10**, is an instruction to SPSS to print only the first ten cases. If this instruction were omitted, all cases would be listed in the output.



To execute this command, first highlight the selection by pressing on your mouse button while dragging the arrow across the command or commands that you want to execute. Next, click on the icon with the black, right-facing arrow on it. Or, you can choose a selection from the *Run* menu. Executing the command will print the list of variables, *gender* and *minority* in the above example, to the Output Viewer. The *Output Viewer* is the third window with which you should be familiar. It is the window in which all output will be printed. The Output Viewer is shown below, containing the text that would be generated from the above syntax.



D

This document is the second module of a four module tutorial series. This document describes the use of SPSS to obtain descriptive and inferential statistics. In this module, you will be introduced to procedures used to obtain several descriptive statistics, frequency tables, and crosstabulations in the first section. In the second section, the Chi-square test of independence, independent and paired sample *t* tests, bivariate and partial correlations, regression, and the general linear model will be covered. If you are not familiar with SPSS or need more information about how to get SPSS to read your data, consult the first module of this four part tutorial, SPSS for Windows: Getting Started. This set of documents uses a sample dataset, *Employee data.sav*, that SPSS provides. It can be found in the root SPSS directory. If you installed SPSS in the default location, then this file will be located in the following location: C:\Program Files\SPSS\Employee Data.sav.

Some users prefer to use keystrokes to navigate through SPSS. Information on common keystrokes are available in our SPSS 10 for Windows Keystoke Manual.

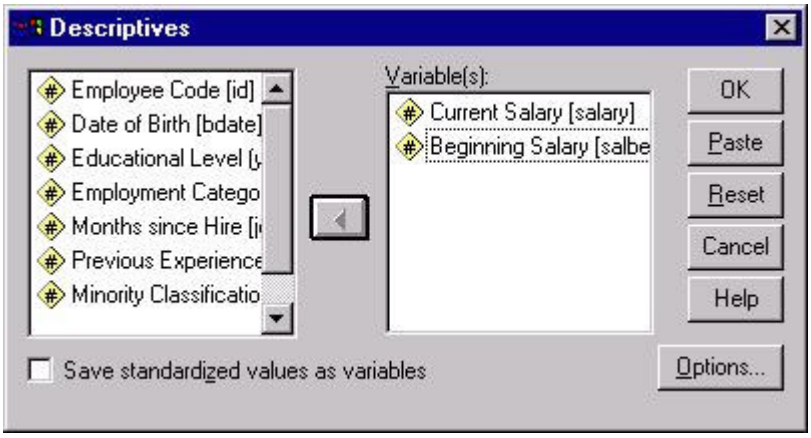
Section 4: Summarizing Data

Descriptive Statistics

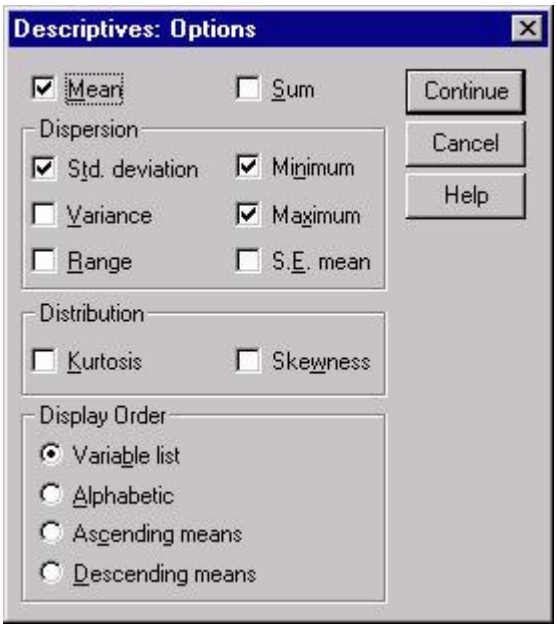
A common first step in data analysis is to summarize information about variables in your dataset, such as the averages and variances of variables. Several summary or descriptive statistics are available under the *Descriptives* option available from the *Analyze* and *Descriptive Statistics* menus:

Analyze
  Descriptive Statistics
    Descriptives...

After selecting the *Descriptives* option, the following dialog box will appear:



This dialog box allows you to select the variables for which descriptive statistics are desired. To select variables, first click on a variable name in the box on the left side of the dialog box, then click on the arrow button that will move those variables to the *Variable(s)* box. For example, the variables *salbegin* and *salary* have been selected in this manner in the above example. To view the available descriptive statistics, click on the button labeled **Options**. This will produce the following dialog box:



Clicking on the boxes next to the statistics' names will result in these statistics being displayed in the output for this procedure. In the above example, only the default statistics have been selected (mean, standard deviation, minimum, and maximum), however, there are several others that could be selected. After selecting all of the statistics you desire, output can be generated by first clicking on the **Continue** button in the *Options* dialog box, then clicking on the **OK** button in the *Descriptives* dialog box. The statistics that you selected will be printed in the Output Viewer. For example, the selections from the preceding example would produce the following output:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	474	$15,750	$135,000	$34,419.57	$17,075.66
Beginning Salary	474	$9,000	$79,980	$17,016.09	$7,870.64
Valid N (listwise)	474				

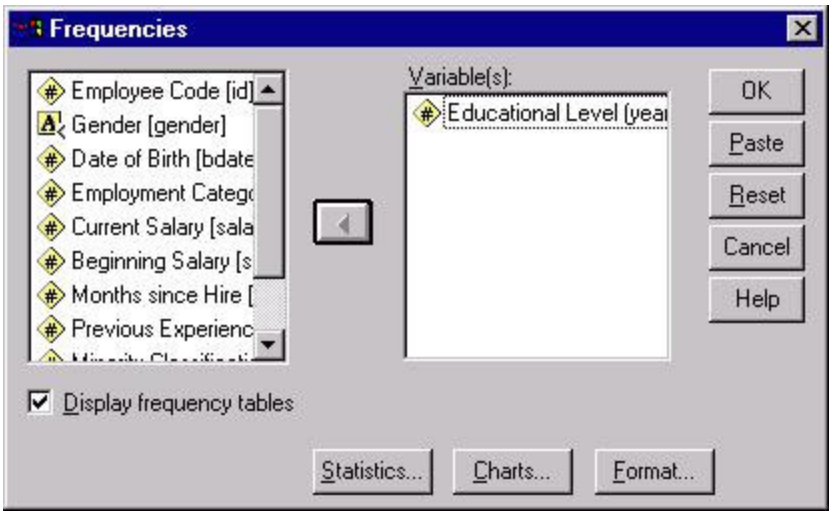
This output contains several pieces of information that can be useful to you in understanding the descriptive qualities of your data. The number of cases in the dataset is recorded under the column labeled *N*. Information about the range of variables is contained in the *Minimum* and *Maximum* columns. For example, beginning salaries ranged from $9000 to $79,980 whereas current salaries range from $15,750 to $135,000. The average salary is contained in the *Mean* column. Variability can be assessed by examining the values in the *Std.* column. The standard deviation measures the amount of variability in the distribution of a variable. Thus, the more that the individual data points differ from each other, the larger the standard deviation will be. Conversely, if there is a great deal of similarity between data points, the standard deviation will be quite small. The standard deviation describes the standard amount variables differ from the mean. For example, a starting salary with the value of $24,886.73 is one standard deviation above the mean in the above example in which the variable, *salary* has a mean of $17,016.09 and a standard deviation of $7,870.64. Examining differences in variability could be useful for anticipating further analyses: in the above example, it is clear that there is much greater variability in the current salaries than beginning salaries. Because equal variances is an assumption of many inferential statistics, this information is important to a data analyst.

Frequencies

While the descriptive statistics procedure described above is useful for summarizing data with an underlying continuous distribution, the *Descriptives* procedure will not prove helpful for interpreting categorical data. Instead, it is more useful to investigate the numbers of cases that fall into various categories. The *Frequencies* option allows you to obtain the number of people within each education level in the dataset. The *Frequencies* procedure is found under the *Analyze* menu:

Analyze
Descriptives Statistics
Frequencies...

Selecting this menu item produces the following dialog box:

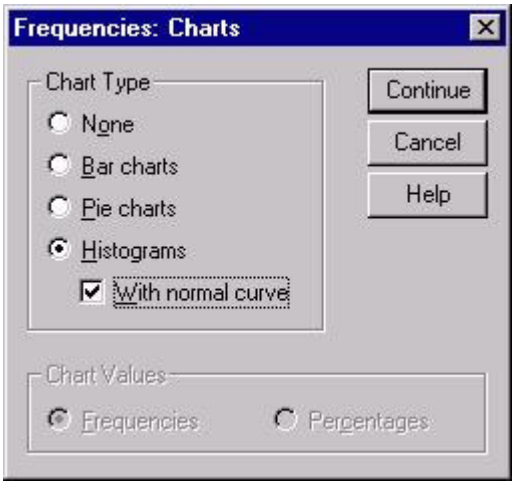


Select variables by clicking on them in the left box, then clicking the arrow in between the two boxes. Frequencies will be obtained for all of the variables in the box labeled *Variable(s)*. This is the only step necessary for obtaining frequency tables; however, there are several other descriptive statistics available, many of which are described in the preceding section. The example in the above dialog box would produce the following output:

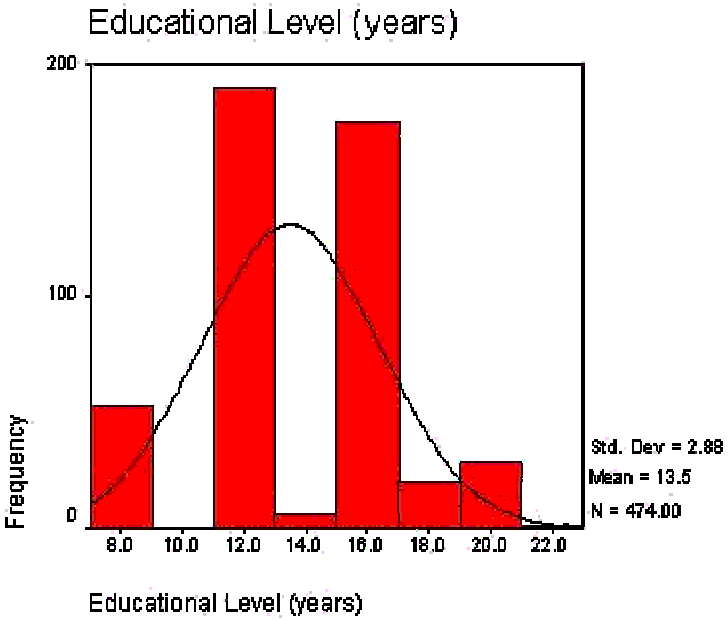
Educational Level (years)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	53	11.2	11.2	11.2
	12	190	40.1	40.1	51.3
	14	6	1.3	1.3	52.5
	15	116	24.5	24.5	77.0
	16	59	12.4	12.4	89.5
	17	11	2.3	2.3	91.8
	18	9	1.9	1.9	93.7
	19	27	5.7	5.7	99.4
	20	2	.4	.4	99.8
	21	1	.2	.2	100.0
	Total	474	100.0	100.0	

Clicking on the **Statistics** button produces a dialog box with several additional descriptive statistics. Clicking on the **Charts** button produces the following box which allows you to graphically examine their data in several different formats:



Each of the available options provides a visual display of the data. For example, clicking on the *Histograms* button with its suboption, *With normal curve*, will provide you with a chart similar to that shown below. This will allow you to assess whether your data are normally distributed, which is an assumption of several inferential statistics. You can also use the *Explore* procedure, available from the *Descriptives* menu, to obtain the *Kolmogorov-Smirnov test*, which is a hypothesis test to determine if your data are normally distributed.



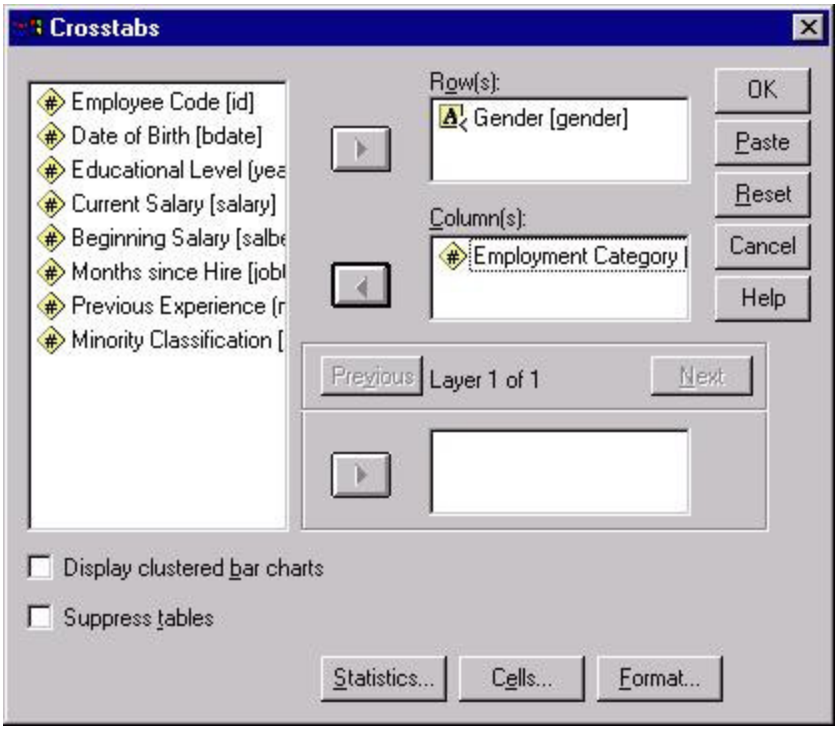
Crosstabulation

While frequencies show the numbers of cases in each level of a categorical variable, they do not give information about the relationship between categorical variables. For example, frequencies can give you the number of men and women in a company AND the number of people in each employment category, but not the number of men

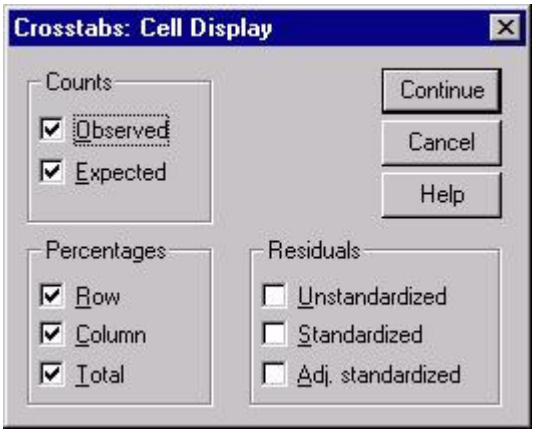
and women IN each employment category. The *Crosstabs* procedure is useful for investigating this type of information because it can provide information about the intersection of two variables. The number of men and women in each of three employment categories is one example of information that can be crosstabulated. The *Crosstabs* procedure is found in the *Analyze* menu in the Data Editor window:

Analyze
Descriptive Statistics
Crosstabs…

After selecting *Crosstabs* from the menu, the dialog box shown above will appear on your monitor. The box on the left side of the dialog box contains a list of all of the variables in the working dataset. Variables from this list can be selected for rows, columns, or layers in a crosstabulation. For example, selecting the variable *gender* for the rows of the table and *jobcat* for the columns would produce a crosstabulation of gender by job category.



The options available by selecting the **Statistics** and **Cells** buttons provide you with several additional output features. Selecting the **Cells** button will produce a menu that allows you to add additional values to your table. For example, the dialog box shown below illustrates an example in which *Expected* option in the *Counts* box and the *Row*, *Column*, and *Total* options in the *Percentages* box have been selected.



The combination of the two dialog boxes shown above will produce the following output table:

Gender * Employment Category Crosstabulation

			Employment Category			
			Clerical	Custodial	Manager	Total
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
		% within Gender	95.4%	.0%	4.6%	100.0%
		% within Employment Category	56.7%	.0%	11.9%	45.6%
		% of Total	43.5%	.0%	2.1%	45.6%
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
		% within Gender	60.9%	10.5%	28.7%	100.0%
		% within Employment Category	43.3%	100.0%	88.1%	54.4%
		% of Total	33.1%	5.7%	15.6%	54.4%
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0
		% within Gender	76.6%	5.7%	17.7%	100.0%
		% within Employment Category	100.0%	100.0%	100.0%	100.0%
		% of Total	76.6%	5.7%	17.7%	100.0%

The crosstabulation statistics provide several interesting observations about the data. In the above table, there appears to be an association between gender and employment category as the expected values, which are the values expected by chance, and the actual counts are different from each other. The following section will discuss how to further examine this relationship with inferential statistics.

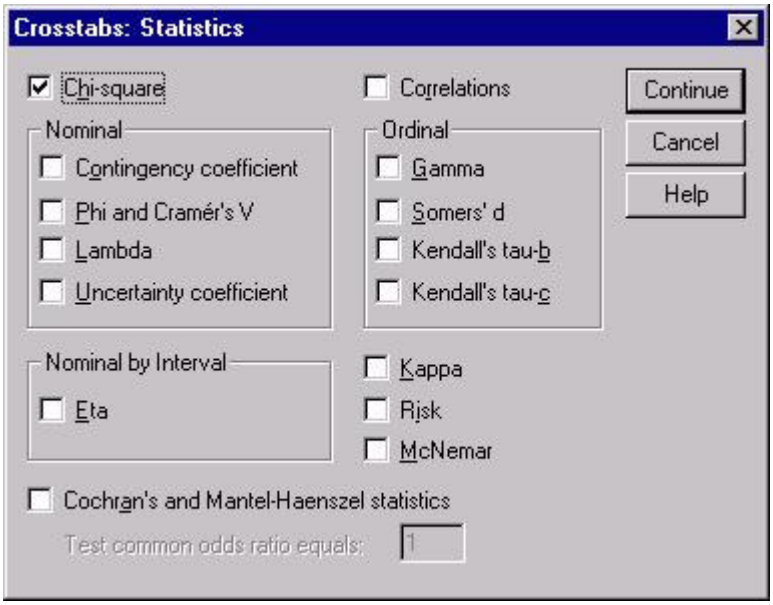
Section 5: Inferential Statistics

Chi-Square Test

The Chi-square test for independence is used in situations where you have two *categorical variables*. A categorical variable is a qualitative variable in which cases are classified in one and only one of the possible levels. A classic example is gender, in which cases are classified in one of two possible levels. The example in the above section, in which *Gender* and *Employment Category* are crosstabulated using the SPSS *Crosstabs* procedure, is an example of data with which you could conduct a *Chi-square test of independence* testing the null hypothesis that there is no relationship between the two variables.

For instance, you could conduct a test of the hypothesis that there is no relationship between *Gender* and *Employment Category*. If this hypothesis were true, you would expect that the proportion of men and women would be the same within each level of *Employment Category*. In other words, there should be little difference between *observed* and *expected* values, where the expected values represent the numbers that would be in each cell when the variables are independent of each other. The difference between observed and expected values is the basis of the Chi-square statistic: it evaluates the likelihood that the differences between the observed and expected values would occur under the null hypothesis that there is no difference between these values. The expected values can be obtained by clicking on the **Cells** box in the *Crosstabs* dialog box, as described in the preceding section. Examining the table above, it appears that it is indeed the case that gender and employment category are not independent of each other. It appears that there are more women in clerical positions than would be expected by chance, whereas there are more men in custodial and managerial positions than would be expected by chance. Conducting a Chi-square test of independence would tell us if the observed pattern is statistically different from the pattern expected due to chance.

The Chi-square test of independence can be obtained through the *Crosstabs* dialog boxes that were used above to get a crosstabulation of the data. After opening the *Crosstabs* dialog box as described in the preceding section, click the **Statistics** button to get the following dialog box:



By clicking on the box labeled *Chi-Square*, you will obtain the Chi-square test of independence for the variables you have crosstabulated. This will produce the following table in the Output Viewer:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.30.

Inspecting the table in the previous section, it appears that the the two variables, gender and employment category, are related to each other in some way. This finding is implicated by the substantial differences in the observed and expected counts: these differences represent the difference between values expected if gender and employment classification were independent of each other (expected counts) and the actual numbers of cases in each cell (observed counts). For example, if gender and employment classification were unrelated, then it is expected that 38.3 women would be in the manager classification as opposed to the observed number, 10. In this example, the expected value of 38.3 represents the fact that 45.6% of the cases in this dataset are women, so it is expected that 45.6% of the 84 managers in the dataset would also be women if gender and employment classification were independent of each other. The output above provides a statistical hypothesis test for the hypothesis that gender and employment category are independent of each other. The large Chi-Square statistic (79.28) and its small significance level ($p < .000$) indicates that it is

very unlikely that these variables are independent of each other. Thus, you can conclude that there is a relationship between a person's gender and their employment classification.

T tests

The *t* test is a useful technique for comparing mean values of two sets of numbers. The comparison will provide you with a statistic for evaluating whether the difference between two means is statistically significant. *T* tests can be used either to compare two independent groups (independent-samples *t* test) or to compare observations from two measurement occasions for the same group (paired-samples *t* test). To conduct a *t* test, your data should be a sample drawn from a continuous underlying distribution. If you are using the *t* test to compare two groups, the groups should be randomly drawn from normally distributed and independent populations. For example, if you were comparing clerical and managerial salaries, the *independent populations* are clerks and managers, which are two nonoverlapping groups. If you have more than two groups or more than two variables in a single group that you want to compare, you should use one of the General Linear Model procedures in SPSS, which are described below.

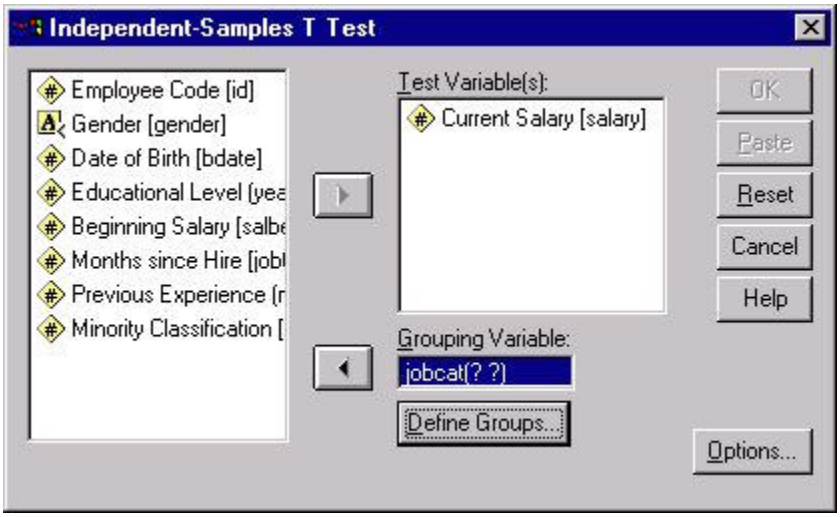
There are three types of *t* tests; the options are all located under the *Analyze* menu item:

Analyze
Compare Means
One-Sample *T* test...
Independent-Samples *T* test...
Paired-Samples *T* test...

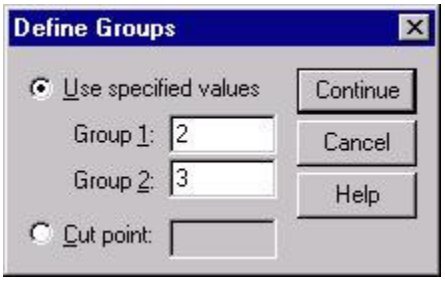
While each of these *t* tests compares mean values of two sets of numbers, they are designed for distinctly different situations:

- The *one-sample t test* is used compare a single sample with a population value. For example, a test could be conducted to compare the average salary of managers within a company with a value that was known to represent the national average for managers.
- The *independent-sample t test* is used to compare two groups' scores on the same variable. For example, it could be used to compare the salaries of clerks and managers to evaluate whether there is a difference in their salaries.
- The *paired-sample t test* is used to compare the means of two variables within a single group. For example, it could be used to see if there is a statistically significant difference between starting salaries and current salaries among the custodial staff in an organization.

To conduct an independent sample *t* test, first select the menu option shown above, to produce the following dialog box:



To select variables for the analysis, first highlight them by clicking on them in the box on the left. Then move them into the appropriate box on the right by clicking on the arrow button in the center of the box. Your independent variable should go in the *Grouping Variable* box, which is a variable that defines which groups are being compared. For example, because employment categories are being compared in this analysis, the *jobcat* variable is selected. However, because *jobcat* has more than two levels, you will need to click on **Define Groups** to specify the two levels of *jobcat* that you want to compare. This will produce another dialog box as is shown below:



Here, the groups to be compared are limited to the groups with the values 2 and 3, which represent the clerical and managerial groups. After selecting the groups to be compared, click the **Continue** button, and then click the **OK** button in the main dialog box. The above choices will produce the following output:

Group Statistics

	Employment Category	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Custodial	27	$30,938.89	$2,114.62	$406.96
	Manager	84	$63,977.80	$18,244.78	$1,990.67

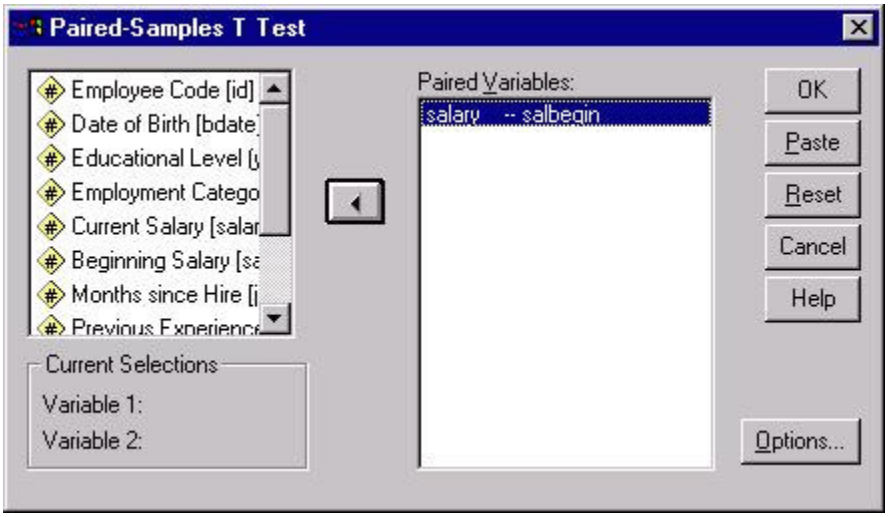
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	29.2	.000	-9.36	109	.000	-$33,038	$3,530	-$40,034	-$26,044
	Equal variances not assumed			-16.3	89.6	.000	-$33,038	$2,032	-$37,076	-$29,002

The first output table, labeled *Group Statistics*, displays descriptive statistics. The second output table, labeled *Independent Samples Test*, contains the statistics that are critical to evaluating the current research question. This table contains two sets of analyses: the first assumes equal variances and the second does not. To assess whether you should use the statistics for equal or unequal variances, use the significance level associated with the value under the heading, *Levene's Test for Equality of Variances*. It tests the hypothesis that the variances of the two groups are equal. A small value in the column labeled *Sig*. indicates that this hypothesis is false and that the groups do indeed have unequal variances. In the above case, the small value in that column indicates that the variance of the two groups, clerks and managers, is not equal. Thus, you should use the statistics in the row labeled *Equal variances not assumed*.

The SPSS output reports a *t statistic* and *degrees of freedom* for all *t* test procedures. Every unique value of the *t* statistic and its associated degrees of freedom have a significance value. In the above example in which the hypothesis that clerks and managers do not differ in their salaries, the *t* statistic under the assumption of unequal variances has a value of -16.3, and the degrees of freedom has a value of 89.6 with an associated significance level of .000. The significance level tells us that the probability that there is no difference between clerical and managerial salaries is very small: specifically, less than one time in a thousand would we obtain a mean difference of $33,038 or larger between these groups if there were really no differences in their salaries.

To obtain a paired-samples *t* test, select the menu items described above and the following dialog box will appear:



The above example illustrates a *t* test between the variables *salbegin* and *salary* which represent employees' beginning salary and their current salary. To set up a paired-samples *t* test as in the above example, click on the two variables that you want to compare. The variable names will appear in the section of the box labeled *Current Selections*. When these variable names appear there, click the arrow in the middle of the dialog box and they will appear in the *Paired Variables* box. Clicking the **OK** button with the above variables selected will produce output for the paired-samples *t* test. The following output is an example of the statistics you would obtain from the above example.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Current Salary - Beginning Salary	$17,403.48	$10,814.62	$496.73	$16,427.41	$18,379.56	35.04	473	.000

As with the independent samples *t* test, there is a *t* statistic and degrees of freedom that has a significance level associated with it. The *t* test in this example tests the hypothesis that there is no difference in clerks' beginning and current salaries. The *t* statistic, (35.04), and its associated significance level ($p < .000$) indicate that this in not the case. In fact, the observed mean difference of $17,403.48 between beginning and current salaries would occur fewer than once in a thousand times if there really were no difference between clerks' beginning and current salaries.

Correlation

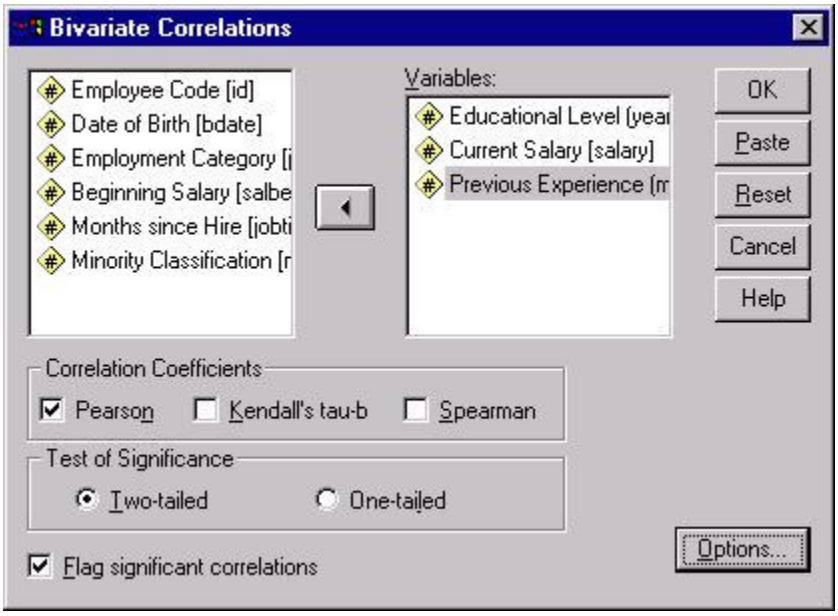
Correlation is one of the most common forms of data analysis both because it can provide an analysis that stands on its own, and also because it underlies many other analyses, and can can be a good way to support conclusions after primary analyses have been completed. *Correlations* are a measure of the linear relationship between two variables. A correlation coefficient has a value ranging from -1 to 1. Values that are closer to the absolute value of 1 indicate that there is a strong relationship between the variables being correlated whereas values closer to 0 indicate that there is little or no linear relationship. The sign of a correlation coefficient describes the type of relationship between the variables being correlated. A positive correlation coefficient indicates that there is a positive linear relationship between the variables: as one variable increases in value, so does the other. An example of two variables that are likely to be positively correlated are the number of days a student attended class and test grades because, as the number of classes attended increases in value, so do test grades. A negative value indicates a negative linear relationship between variables: as one variable increases in value, the other variable decreases in value. The number of days students miss class and their test scores are likely to be negatively correlated because as the number of days of missed classed increases, test scores typically decrease.

To obtain a correlation in SPSS, start at the *Analyze* menu. Select the *Correlate* option from this menu. By selecting this menu item, you will see that there are three options for correlating variables: (1) *Bivariate*, (2) *Partial*, and (3) *Distances.* This document will cover the first two types of correlations. The *bivariate correlation* is for situations where you are interested only in the relationship between two variables. *Partial correlations* should be used when you are measuring the association between two variables but want to factor out the effect of one or more other variables.

To obtain a bivariate correlation, choose the following menu option:

Analyze
Correlate
Bivariate...

This will produce the following dialog box:



To obtain correlations, first click on the variable names in the variable list on the left side of the dialog box. Next, click on the arrow between the two white boxes which will move the selected variables into the *Variables* box. Each variable listed in the *Variables* box will be correlated with every other variable in the box. For example, with the above selections, we would obtain correlations between *Education Level* and *Current Salary*, between *Education Level* and *Previous Experience*, and between *Current Salary* and *Previous Experience*. We will maintain the default options shown in the above dialog box in this example. The first option to consider is the type of correlation coefficient. Pearson's is appropriate for continuous data as noted in the above example, whereas the other two correlation coefficients, Kendall's tau-b and Spearman's, are designed for ranked data. The choice between a one and two-tailed significance test in the *Test of Significance* box should be determined by whether the hypothesis you are testing is making a prediction about the direction of effect between the two variables: if you are making a prediction that there is a negative or positive relationship between the variables, then the one-tailed test is appropriate; if you are not making a directional prediction, you should use the two-tailed test if there is not a specific prediction about the direction of the relationship between the variables you are correlating. The selections in the above dialog box will produce the following output:

Correlations

		Educational Level (years)	Current Salary	Previous Experience (months)
Educational Level (years)	Pearson Correlation	1.000	.661**	-.252**
	Sig. (2-tailed)	.	.000	.000
	N	474	474	474
Current Salary	Pearson Correlation	.661**	1.000	-.097*
	Sig. (2-tailed)	.000	.	.034
	N	474	474	474
Previous Experience (months)	Pearson Correlation	-.252**	-.097*	1.000
	Sig. (2-tailed)	.000	.034	.
	N	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

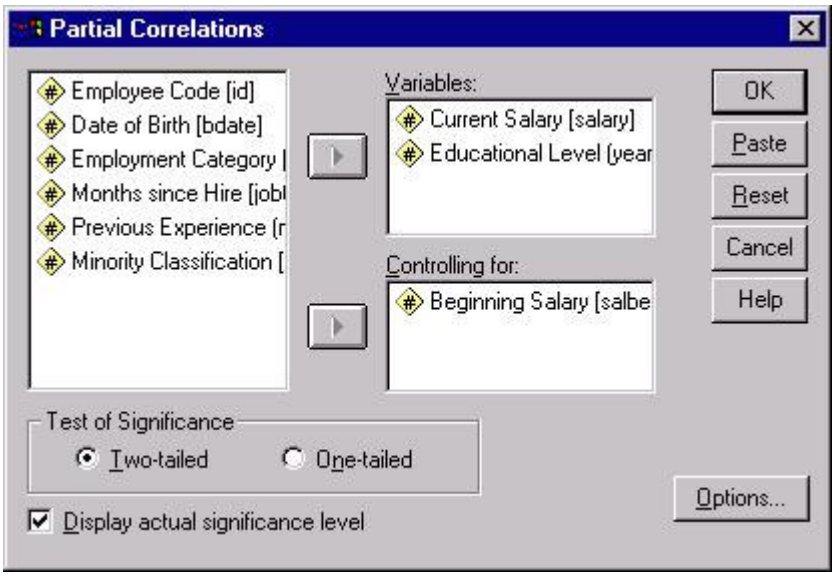
*. Correlation is significant at the 0.05 level (2-tailed).

This output gives us a correlation matrix for the three correlations requested in the above dialog box. Note that despite there being nine cells in the above matrix, there are only three correlation coefficients of interest: (1) the correlation between current salary and educational level, the correlation between previous experience and educational level, and the correlation between current salary and previous experience. The reason only three of the nine correlations are of interest is because the diagonal consists of correlations of each variable with itself, always resulting in a value of 1.00 and the values on each side of the diagonal replicate the values on the opposite side of the diagonal. For example, the three unique correlation coefficients show there is a positive correlation between employees' number of years of education and their current salary. This positive correlation coefficient (.661) indicates that there is a statistically significant ($p < .001$) linear relationship between these two variables such that the more education a person has, the larger that person's salary is. Also observe that there is a statistically significant ($p < .001$) negative correlation coefficient (-.252) for the association between education level and previous experience, indicating that the linear relationship between these two variables is one in which the values of one variable decrease as the other increases. The third correlation coefficient (-.097) also indicates a negative association between employee's current salaries and their previous work experience, although this correlation is fairly weak.

The second type of correlation listed under the *Correlate* menu item is the partial correlation, which measures an association between two variables with the effects of one or more other variables factored out. To obtain a partial correlation, select the following menu item:

Analyze
Correlate
Partial...

This will produce the following dialog box:



Here, we have selected the variables we want to correlate as well as the variable for which we want to control by first clicking on variable names to highlight them on the left side of the box, then moving them to the boxes on the right by clicking on the arrow immediately to the left of either the *Variables* box or the *Controlling for* box. In this example, we are correlating current salaries with years of education while controlling for beginning salaries. Thus, we will have a measure of the association between current salaries and years of education, while removing the association between beginning salaries and the two variables we are correlating. The above example will produce the following output:

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T
S  - - -

Controlling for..    SALBEGIN

               SALARY        EDUC

SALARY         1.0000        .2810
               (    0)     (  471)
               P= .        P= .000

EDUC            .2810      1.0000
               (  471)     (    0)
               P= .000     P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Notice that the correlation coefficient is considerably smaller in the output above than in the bivariate correlation example: the correlation between these variables was .661, whereas it is only .281 in the partial correlation. Nevertheless, a statistically significant association ($p < .001$) exists between these variables. While it may be obvious in the above example that starting and current salaries will be highly correlated, the example illustrates how partial correlations can be used to assess the extent to which variables can be used to explain unique variance by removing the effects of other variables that may be highly correlated with the relationship of interest.

Partial correlations can be especially useful in situations where it is not obvious whether variables possess a unique relationship or whether several variables overlap with each other. For example, if you were attempting to correlate anxiety with job performance and stress with job performance, it would be useful to conduct partial correlations. You could correlate anxiety and a job performance measure while controlling for stress to determine if there were a unique relationship between anxiety and job performance or whether perhaps stress is highly correlated with anxiety--which would result in little remaining variance that could be uniquely attributed to the association between anxiety and job performance.

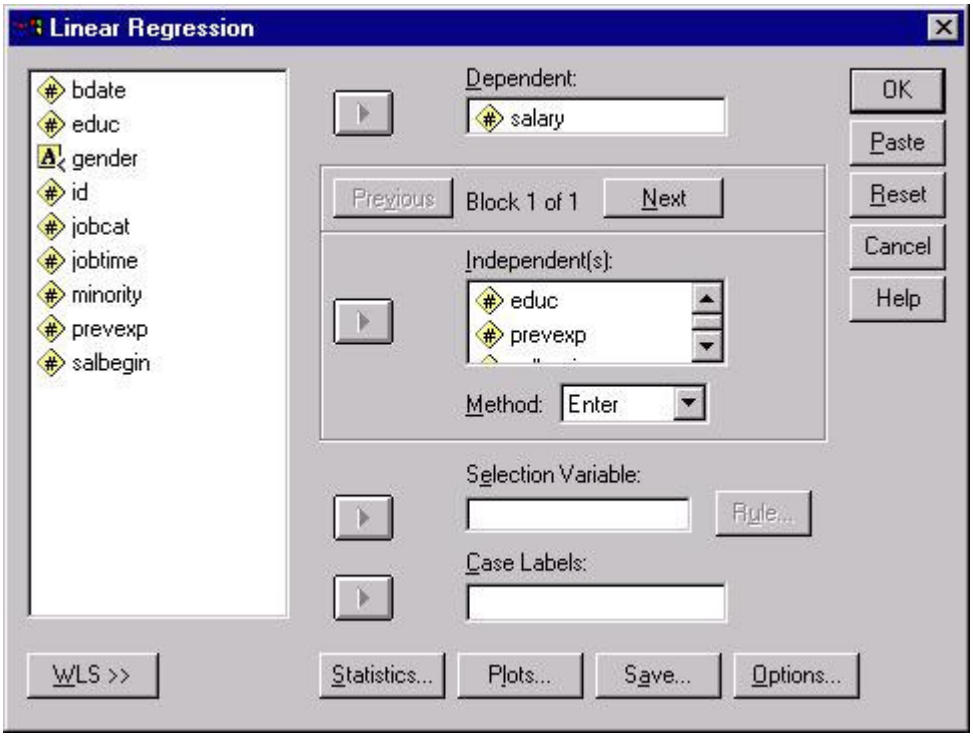
Regression

Regression is a technique that can be used to investigate the effect of one or more predictor variables on an outcome variable. Regression allows you to make statements about how well one or more independent variables will predict the value of a dependent variable. For example, if you were interested in investigating which variables in the employee database were good predictors of employees' current salaries, you could create a regression equation that would use several of the variables in the dataset to predict employees' salaries. By doing this you will be able to make statements about whether knowing something about variables such as employees' number of years of education, their starting salary, or their number of months on the job are good predictors of their current salaries.

To conduct a regression analysis, select the following from the *Analyze* menu:

Analyze
Regression
Linear...

This will produce the following dialog box:



This dialog box illustrates an example regression equation. As with other analyses, you select variables from the box on the left by clicking on them, then moving them to the boxes on the right by clicking the arrow next to the box where you want to enter a particular variable. Here, employees' current salary has been entered as the dependent variable. In the *Independent(s)* box, several predictor variables have been entered, including education level, beginning salary, months since hire, and previous experience.

NOTE: Before you run a regression model, you should consider the method that you use for selecting or rejecting variables in that model. The box labeled *Method* allows you to select from one of five methods: *Enter*, *Remove*,*Forward*, *Backward*, and *Stepwise*. Unfortunately, we cannot offer a comprehensive discussion of the characteristics of each of these methods here, but you have several options regarding the method you use to remove and retain predictor variables in your regression equation. In this example, we will use the SPSS default method, *Enter*, which is a standard approach in regression models. If you have questions about which method is most appropriate for your data analysis, consult a regression text book, the SPSS help facilities, or contact a consultant.

The following output assumes that only the default options have been requested. If you have selected options from the *Statistics*, *Plots*, or *Options* boxes, then you will have more output than is shown below and some of your tables may contain additional statistics not shown here.

The first table in the output, shown below, includes information about the quantity of variance that is explained by your predictor variables. The first statistic, *R*, is the

multiple correlation coefficient between all of the predictor variables and the dependent variable. In this model, the value is .90, which indicates that there is a great deal of variance shared by the independent variables and the dependent variables. The next value, *R* Square, is simply the squared value of *R*. This is frequently used to describe the goodness-of-fit or the amount of variance explained by a given set of predictor variables. In this example, the value is .81, which indicates that 81% of the variance in the dependent variable is explained by the independent variables in the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.900[a]	.810	.809	$7,465.14

a. Predictors: (Constant), Months since Hire, Previous Experience (months), Beginning Salary, Educational Level (years)

The second table in the output is an ANOVA table that describes the overall variance accounted for in the model. The *F* statistic represents a test of the null hypothesis that the expected values of the regression coefficients are equal to each other and that they equal zero. Put another way, this *F* statistic tests whether the R square proportion of variance in the dependent variable accounted for by the predictors is zero. If the null hypothesis were true, then that would indicate that there is not a regression relationship between the dependent variable and the predictor variables. But, instead, it appears that the four predictor variables in the present example are not all equal to each other and could be used to predict the dependent variable, current salary, as is indicated by a large *F* value and a small significance level.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111779919524.3	4	27944979881.07	501.450	.000[a]
	Residual	26136575912.07	469	55728306.849		
	Total	137916495436.3	473			

a. Predictors: (Constant), Educational Level (years), Months since Hire, Previous Experience (months), Beginning Salary

b. Dependent Variable: Current Salary

The third table in standard regression output provides information about the effects of individual predictor variables. Generally, there are two types of information in the *Coefficients* table: coefficients and significance tests. The coefficients indicate the increase in the value of the dependent variable for each unit increase in the predictor variable. For example, the unstandardized coefficient for *Educational Level* in the example is 669.91, which indicates to us that for each year of education, a person's

predicted salary will increase by $669.91. A well known problem with the interpretation of unstandardized coefficients is that their values are dependent on the scale of the variable for which they were calculated, which makes it difficult to assess the relative influence of independent variables through a comparison of unstandardized coefficients. For example, comparing the unstandardized coefficient of *Education Level*, 669.91, with the unstandardized coefficient of the variable *Beginning Salary*, 1.77, it could appear that Educational Level is a greater predictor of a person's current salary than is *Beginning Salary*. We can see that this is deceiving, however, if we examine the standardized coefficients, or *Beta coefficients*. Beta coefficients are based on data expressed in standardized, or *z* score form. Thus, all variables have a mean of zero and a standard deviation of one and are thus expressed in the same units of measurement. Examining the Beta coefficients for *Education Level* and *Beginning Salary*, we can see that when these two variables are expressed in the same scale, *Beginning Salary* is more obviously the better predictor of *Current Salary*.

In addition to the coefficients, the table also provides a significance test for each of the independent variables in the model. The significance test evaluates the null hypothesis that the unstandardized regression coefficient for the predictor is zero when all other predictors' coefficients are fixed to zero. This test is presented as a *t* statistic. For example, examining the *t* statistic for the variable, *Months Since Hire*, you can see that it is associated with a significance value of .000, indicating that the null hypothesis, that states that this variable's regression coefficient is zero when all other predictor coefficients are fixed to zero, can be rejected.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-16149.7	3255.470		-4.961	.000
	Months since Hire	161.486	34.246	.095	4.715	.000
	Previous Experience (months)	-17.303	3.528	-.106	-4.904	.000
	Beginning Salary	1.768	.059	.815	30.111	.000
	Educational Level (years)	669.914	165.596	.113	4.045	.000

a. Dependent Variable: Current Salary

General Linear Model

The majority of procedures used for conducting analysis of variance (ANOVA) in SPSS can be found under the *General Linear Model* (GLM) menu item in the *Analyze* menu. Analysis of variance can be used in many situations to determine whether there are differences between groups on the basis of one or more outcome variables or if a continuous variable is a good predictor of one or more dependent variables. There are three varieties of of the general linear model available in SPSS: univariate,

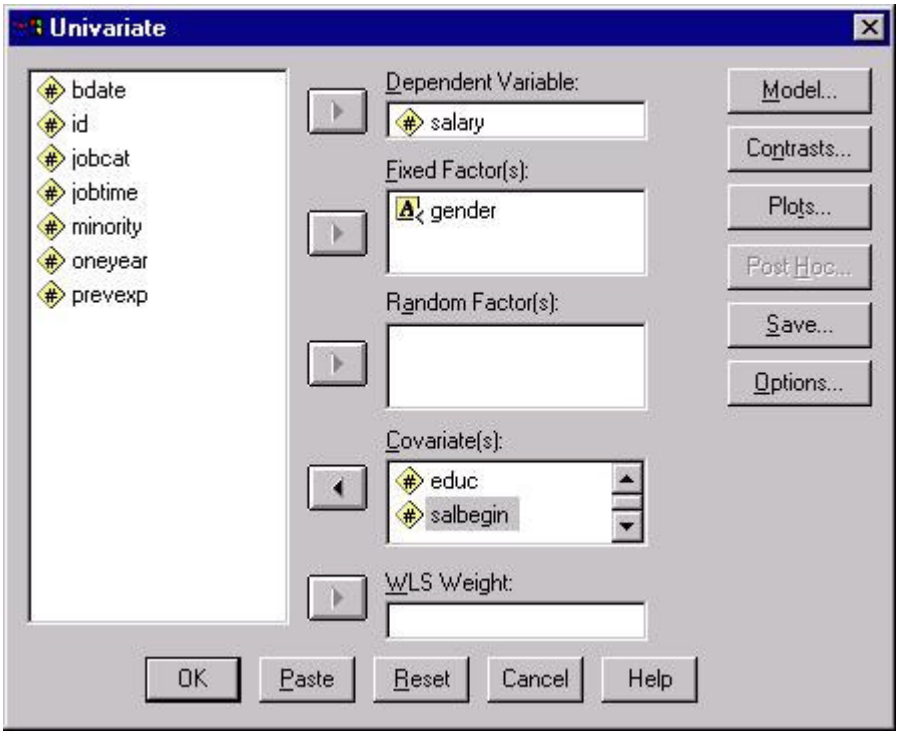
multivariate, and repeated measures. The *univariate general linear model* is used in situations where you only have a single dependent variable, but may have several independent variables that can be fixed between-subjects factors, random between-subjects factors, or covariates. The *multivariate general linear model* is used in situations where there is more than one dependent variable and independent variables are either fixed between-subjects factors or covariates. The *repeated measures general linear model is* used in situations where you have more than one measurement occasion for a dependent variable and have fixed between-subjects factors or covariates as independent variables. Because it is beyond the scope of this document to cover all three varieties of the general linear model in detail, we will focus on the univariate version of the general linear model with some attention given to topics that are unique to the repeated measures general linear model. Several features of the univariate general linear model are useful for understanding other varieties of the model that are provided in SPSS: understanding the univariate model will prove useful for understanding other GLM options.

The univariate general linear model is used to compare differences between group means and estimating the effect of covariates on a single dependent variable. For example, you may want to see if there are differences between men and women's salaries in a sample of employee data. To do this, you would want to demonstrate that the average salary is significantly different between men and women. However, in doing such an analysis, you are likely aware that there are other factors that could affect a person's salary that need to be controlled for in such an analysis. For example, educational background and starting salary are some such variables. By including these variables in our analysis, you will be able to evaluate the differences between men and women's salaries while controlling for the influence of these other variables.

To specify a univariate general linear model in SPSS, go to the analyze menu and select univariate from the general linear model menu:

Analyze
General Linear Model
Univariate...

This will produce the following dialog box:



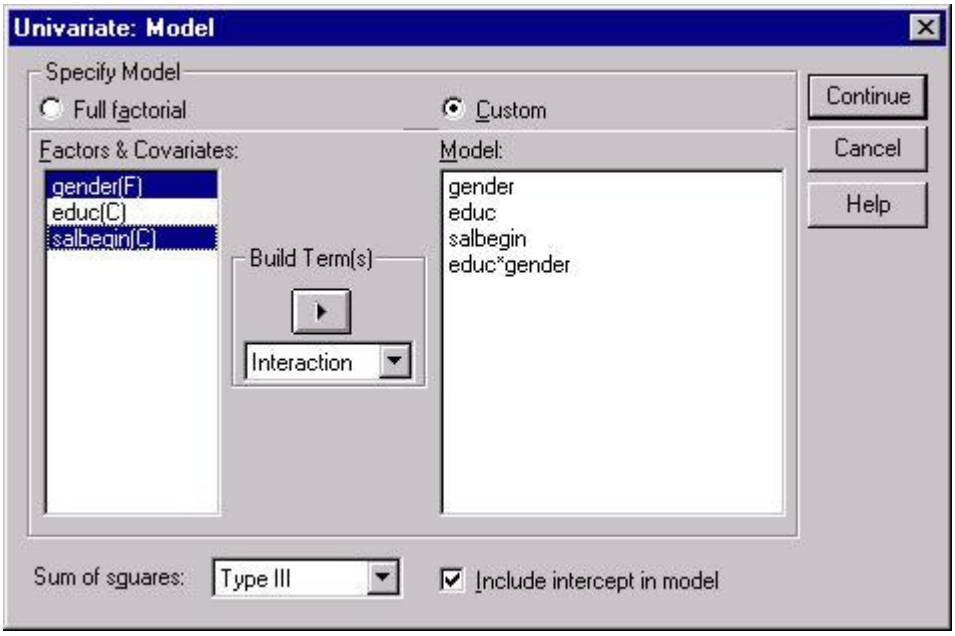
The above box demonstrates a model with multiple types of independent variables. The variable, gender, has been designated as a *fixed factor* because it contains all of the levels of interest.

In contrast, *random variables* are variables that represent a random sample of the possible levels that could be sampled. There are not any true random variables in our dataset; therefore, this input box has been left blank here. However, you could imagine a situation similar to the above example where you sampled data from multiple corporations for our employee database. In that case, you would have introduced a random variable into the model--the corporation to which an employee belongs. Corporation is a random factor because you would only be sampling a few of the many possible corporations to which you would want to generalize your results.

The next input box contains the covariates in your model. A *covariate* is a quantitative independent variable. Covariates are often entered in models to reduce error variance: by removing the effects of the relationship between the covariate and the dependent variable, you can often get a better estimate of the amount of variance that is being accounted for by the factors in the model. Covariates can also be used to measure the linear association between the covariate and a dependent variable, as is done in regression models. In this situation, a linear relationship indicates that the dependent variable increases or decreases in value as the covariate increases or decreases in value.

The box labeled *WLS Weight* can contain a variable that is used to weight other variables in a weighted least-squares analysis. This procedure is infrequently used however, and is not discussed in any detail here.

The default model for the SPSS univariate GLM will include main effects for all independent variables and will provide interaction terms for all possible combinations of fixed and random factors. You may not want this default model, or you may want to create interaction terms between your covariates and some of the factors. In fact, if you intend to conduct an analysis of covariance, you should test for interactions between covariates and factors. Doing so will determine whether you have met the *homogeneity of regression slopes* assumption, which states that the regression slopes for all groups in your analysis are equal. This assumption is important because the means for each group are adjusted by averaging the slopes for each group so that group differences in the covariate are removed from the dependent variable. Thus, it is assumed that the relationship between the covariate and the dependent variable is the same at all levels of the independent variables. To make changes in the default model, click on the **Model** button which will produce the following dialog box:



The first step for modifying the default model is to click on the button labeled *Custom,* to activate the grayed out areas of the dialog box. At this point, you can begin to move variables in the *Factors & Covariates* box into the *Model* box. First, move all of the main effects into the *Model* box. The quickest way to do that is to double-click on their names in the *Factors & Covariates* box. After entering all of the main effects, you can begin building interaction terms. To build the interactions, click on the arrow facing downwards in the *Build Term(s)* section and select interaction, as shown in the figure above. After you have selected the interaction, you can click on the names of the variables with which you would like to build an interaction, then click on the arrow facing right under the *Build Term(s)* heading. In the above

example, the *educ*gender* term has already been created. The *salbegin*gender* and *salbegin*educ* terms can be created by highlighting two terms at a time as shown above, then clicking on the right-facing arrow. Some of the other options in the *Build Terms* list that you may find useful are the *All n-way* options. For example if you highlighted all three variables in the Factors & Covariates box, you could create all of the three possible 2-way interactions by selecting the *All 2-way* option from the *Build Terms(s)* drop-down menu, then clicking the right-facing arrow.

If you are testing the homogeneity of regression slopes assumption, you should examine your group by covariate interactions, as well as any covariate by covariate interactions. In order to meet the ANCOVA assumption, these interactions should not be significant. Examining the output from the example above, we expect to see nonsignificant effects for the *gender*educ* and the *gender*salbegin* interaction effects:

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	109713955893[a]	6	18285659315.6	302.788	.000
Intercept	1153099.423	1	1153099.423	.019	.890
GENDER	1035163.021	1	1035163.021	.017	.896
SALBEGIN	435965029.184	1	435965029.184	7.219	.007
EDUC	80521627.420	1	80521627.420	1.333	.249
GENDER * SALBEGIN	39870280.811	1	39870280.811	.660	.417
GENDER * EDUC	48780222.556	1	48780222.556	.808	.369
SALBEGIN * EDUC	90179603.501	1	90179603.501	1.493	.222
Error	28202539543.0	467	60390876.966		
Total	699467436925	474			
Corrected Total	137916495436	473			

a. R Squared = .796 (Adjusted R Squared = .793)

Examining the group by covariate effects, you can see that both were nonsignificant. The *gender*salbegin* effect has a small *F* statistic (.660) and a large significance value (.417), the *educ*salbegin* effect also has a small *F* statistic (1.808) and large significance value (.369), and the *salbegin*educ* effect also has a small *F* statistic (1.493) and large significance level (.222). Because all of these significance levels are greater than .05, the homogeneity of regression assumption has been met and you can proceed with the ANCOVA.

Knowing that the model does not violate the homogeneity of regression slopes assumption, you can remove the interaction terms from the model by returning to the *GLM Univariate* dialog box, clicking the **Model** button, and selecting *Full Factorial*. This will return the model to its default form in which there are no interactions with

covariates. After you have done this, click **OK** in the *GLM Univariate* dialog box to produce the following output:

Tests of Between-Subjects Effects

Dependent Variable: Current Salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	109520439289[a]	3	36506813096.4	604.246	.000
Intercept	842156178.459	1	842156178.459	13.939	.000
EDUC	2287985765.92	1	2287985765.92	37.870	.000
SALBEGIN	42057263689.0	1	42057263689.0	696.115	.000
GENDER	234049991.723	1	234049991.723	3.874	.050
Error	28396056147.3	470	60417140.739		
Total	699467436925	474			
Corrected Total	137916495436	473			

a. R Squared = .794 (Adjusted R Squared = .793)

The default output for the univariate general linear model contains all main effects and interactions between fixed factors. The above output contains no interactions because gender is the only fixed factor. Each factor, covariate, or other source of variance is listed in the left column. For each source of variance, there are several test statistics. To evaluate the influence of each independent variable, look at the *F* statistic and its associated significance level. Examining the first covariate, education level, the *F* statistic (37.87) and its associated significance level (.000) indicate that it has a significant linear relationship with the dependent variable. The second covariate, *salbegin*, also has a significant *F* statistic (696.12) as can be seen from its associated significance level (.000). In both cases, this indicates that the the values of the dependent variable, *salary*, increases as the values of education level and beginning salaries increase. The next source of variance, gender, provides us with a test of the null hypothesis that there are no differences between gender groups, or more specifically, there are not differences between men and women's salaries. This test provides a small *F* statistic (3.87) and a significance level that is not statistically significant ($p = .05$). In the above model containing education level and beginning salaries as covariates, we are not able to say that there is a statistically significant difference between men and women's salaries. That is, when the model takes into account the variance accounted for by education level and beginning salaries, the variance that can be uniquely attributed to gender is not significantly different from a model in which gender explains no variance.

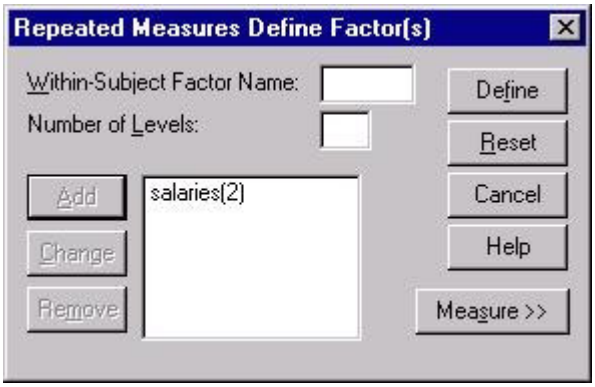
The repeated measures version of the general linear model has many similarities to the univariate model described above. However, the key difference between the models is that there are multiple measurement occasions of the dependent variable in repeated measures models, whereas the univariate model only permits a single dependent variable. You could conduct a similar model with repeated measurements

by using beginning salaries and current salaries as the repeated measurement occasions.

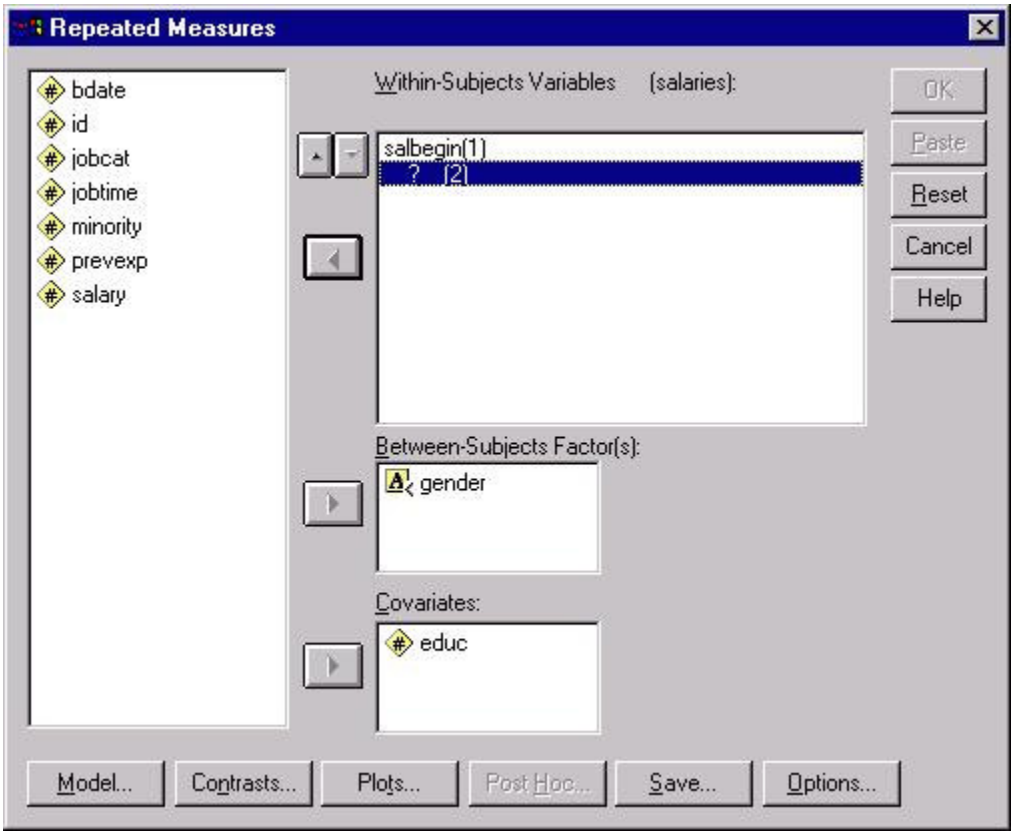
To conduct this analysis, you should select the *Repeated Measures* option from the *General Linear Model* submenu of the *Analyze* menu:

Analyze
General Linear Model
Repeated Measures...

Selecting this option will produce the following dialog box:



This dialog box is used for defining the repeated measures, or within-subjects, dependent variables. You first give the within-subject factor a name in the box labeled *Within-Subject Factor Name*. This name should be something that describes the dependent variables you are grouping together. For example, in this dialog box, salaries are being analyzed, so the within-subject factor was given the name *salaries*. Next, specify the number of levels, or number of measurement occasions, in the box labeled *Number of Levels*. This is the number of times the dependent variable was measured. Thus, in the present example, there are two measurement occasions for salary because you are measuring beginning salaries and current salaries. After you have filled in the *Within-Subject Factor Name* and the *Number of Levels* input boxes, click the **Add** button which will transfer the information in the input boxes into the box below. Repeat this process until you have specified all of your within-subject factors. Then, click on the **Define** button, and the following dialog box will appear:



When this box initially appears, you will see a slot for each level of the within-subject factor variables that you specified in the previous dialog box. These slots are labeled numerically for each level of the within-subject factor but do not contain variable names. You still need to specify which variable fills each slot of the within-subject factors. To do this, click the variable's name in the variable list on the left side of the dialog box. Next, click on the arrow pointing towards the *Within-Subject Variables* dialog box to move the variable name from the list to the top slot in the within-subjects box. This process has been completed for *salbegin*, the first level of the *salaries* within-subject factor. The same process should be repeated for *salary*, the variable representing an employee's current salary.

After you have completed the specifications for the within-subjects factors, you can define your independent variables. Between-subject factors, or fixed factors should be moved into the box labeled *Between-Subjects Factors(s)* by first clicking on the variable name in the variable list, then clicking on the arrow to the left of the *Between-Subjects Factor(s)* box. In this example, gender has been selected as a between-subjects factor. Covariates, or continuous predictor variables, can be moved into the *Covariates* box in the same manner as were the between-subjects factors. Above, *educ*, the variable representing employee's number of years of education, has been specified as a covariate.

This will produce several output tables, but we will focus here on the tables describing between-subject and within-subject effects. However, these tables for

univariate analysis of variance may not always be the appropriate. The univariate tests have an additional assumption: the assumption of sphericity. If this assumption is violated, you should use the multivariate output or adjust your results using one of the correction factors in the SPSS output. For a more detailed discussion of this topic, see the usage note, *Repeated Measures ANOVA Using SPSS MANOVA* in the section, "Within-Subjects Tests: The Univariate versus the Multivariate Approach." This usage note can be found at http://www.utexas.edu/cc/rack/stat.html.

The following output contains the statistics for the effects in the model specified in the above dialog boxes:

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
SALARIES	Sphericity Assumed	697223511.2	1	697223511	18.915	.000
	Greenhouse-Geisser	697223511.2	1.0	697223511	18.915	.000
	Huynh-Feldt	697223511.2	1.0	697223511	18.915	.000
	Lower-bound	697223511.2	1.0	697223511	18.915	.000
SALARIES * EDUC	Sphericity Assumed	6345201343	1	6345201343	172.136	.000
	Greenhouse-Geisser	6345201343	1.0	6345201343	172.136	.000
	Huynh-Feldt	6345201343	1.0	6345201343	172.136	.000
	Lower-bound	6345201343	1.0	6345201343	172.136	.000
SALARIES * GENDER	Sphericity Assumed	924057792.9	1	924057793	25.068	.000
	Greenhouse-Geisser	924057792.9	1.0	924057793	25.068	.000
	Huynh-Feldt	924057792.9	1.0	924057793	25.068	.000
	Lower-bound	924057792.9	1.0	924057793	25.068	.000
Error(SALARIES)	Sphericity Assumed	17361784863	471	36861539.0		
	Greenhouse-Geisser	17361784863	471	36861539.0		
	Huynh-Feldt	17361784863	471	36861539.0		
	Lower-bound	17361784863	471	36861539.0		

This table contains information about the within-subject factor, *salaries*, and its interactions with the independent variables. The main effect for salaries is a test of the null hypothesis that all levels of within-subjects factor are equal, or, more specifically, it is a test of the hypothesis that beginning and current salaries are equal. The *F* statistic (18.91) and its associated significance level ($p < .000$) indicate that you can reject this hypothesis as false. In other words, it appears that there is a statistically significant difference between beginning salaries and salaries after one year of employment. After you have tested this hypothesis, you can then investigate whether the increase in salaries is the same across all values or levels of the other independent variables that are included in the model. The first interaction term in the table tests the hypothesis that the increase in salaries is constant, regardless of educational background. The *F* statistic (172.10) and its associated significance level ($p < .000$) allow us to reject this hypothesis as well. The knowledge that this

interaction is significant indicates that it is worthwhile to examine characteristics of the interaction. Here, the interaction reflects the fact that employees with higher levels of education received greater pay raises than those with lower levels of education. However, you should always investigate the properties of your interactions through graphical displays, mean comparisons, or statistical tests because a significant interaction can take on many forms. Finally, the third interaction tests the hypothesis that the increase in salaries in the first year of employment differs by gender. The *F* statistic (25.07) and its significance level ($p < .000$) indicate that the increase in salaries does vary by gender.

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1992569475.0	1	1992569475.0	13.631	.000
EDUC	40631029552	1	40631029552	277.955	.000
GENDER	8155460237.7	1	8155460237.7	55.791	.000
Error	68850070836	471	146178494.34		

The output for the repeated measures general linear model also provides statistics for between-subject effects. In this example, the model contains two between-subjects factors: employees' education level and their gender. Education level was entered as a covariate in the model, and therefore the statistics associated with it are a measure of the linear relationship between education level and salaries. In contrast, the statistics for the between-subjects factor, gender, represents a comparison between groups across all levels of the within-subjects factors. Specifically, it is a comparison between males and females on differences between their beginning and current salaries. In the above example, both education level and gender are statistically significant. The *F* statistic (277.96) and significance level ($p < .000$) associated with education level allows us to reject the null hypothesis that there is not a linear relationship between education and salaries. By rejecting the null hypothesis, you can conclude that there is a positive linear relationship between the two variables indicating that as number of years of education increases, salaries do as well. The *F* statistic (55.79) for gender and its associated significance level ($p < .000$) represent a test of the null hypothesis that there are no group differences in salaries. The significant *F* statistic indicates that you can reject this null hypothesis and conclude that there is a statistically significant difference between men and women's salaries.

This document is the third module of a four module tutorial series. This document describes the use of SPSS to create and modify tables which can be exported to other applications. Graphical displays of data are also discussed, including bar graphs and scatterplots as well as a discussions on how to modify graphs using the SPSS Chart Editor and Interactive Graphs. If you are not familiar with SPSS or need more information about how to get SPSS to read your data, consult the first module, SPSS for Windows: Getting Started, of this four part tutorial.

Some users prefer to use keystrokes to navigate through SPSS. Information on common keystrokes are available in our SPSS 10 for Windows Keystoke Manual. This set of documents uses a sample dataset, *Employee data.sav*, that SPSS provides. It can be found in the root SPSS directory. If you installed SPSS in the default location, then this file will be located in the following location: C:\Program Files\SPSS\Employee Data.sav.

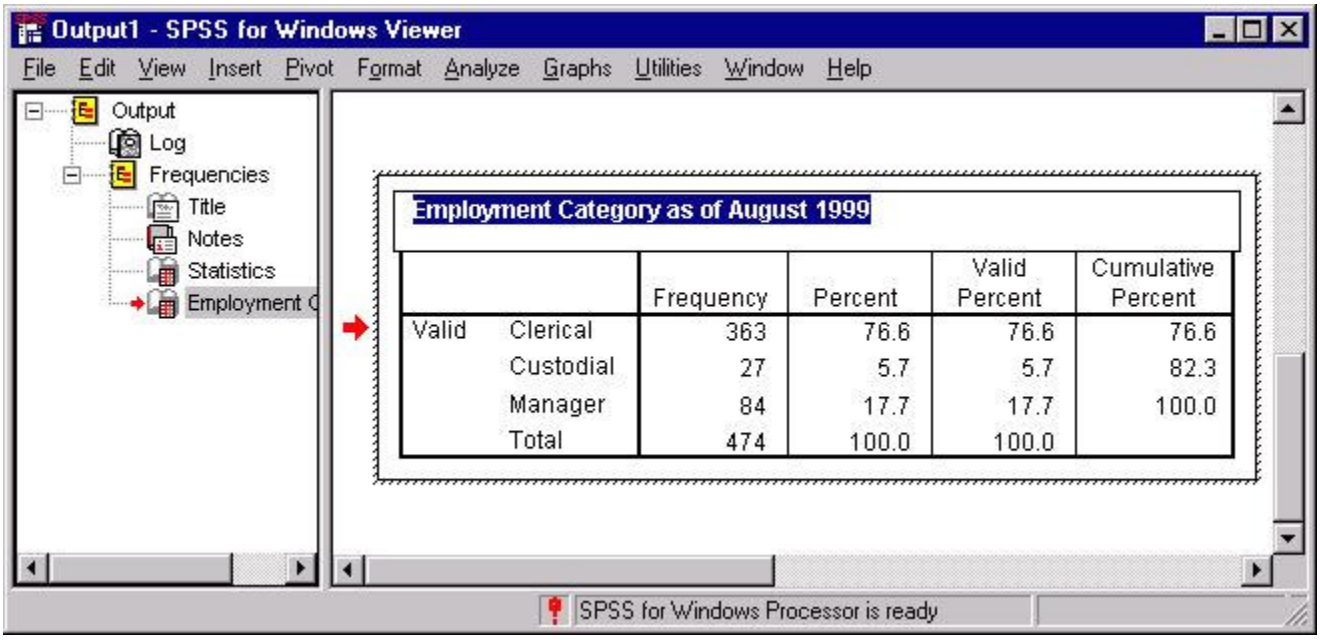
Section 6: Displaying Data

Tables

Much of the output in SPSS is displayed in a pivot table format. While these pivot tables are professional quality in their appearance, you may wish to alter their appearance or export them to another application. There are several ways in which you can modify tables. In this section, we will discuss how you can alter text, modify a table's appearance, and export information in tables to other applications.

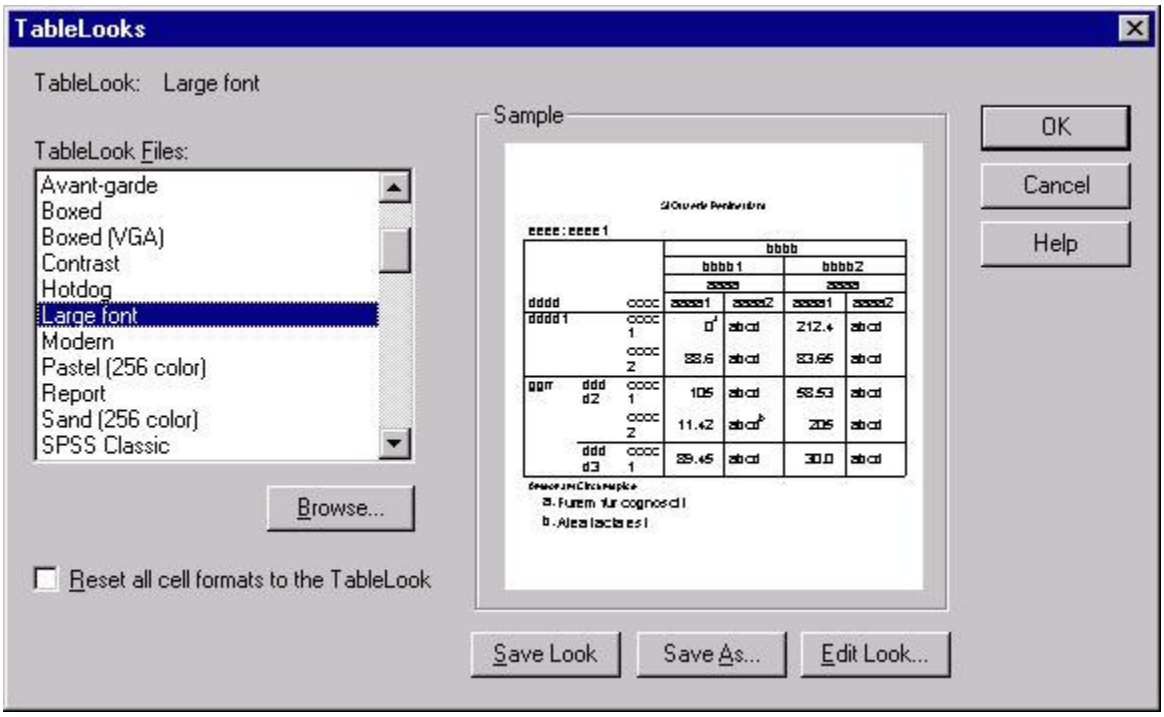
To edit the text in any SPSS output table, you should first double-click that table. This will outline dashed lines, as shown in the figure below, indicating that it is ready to be edited. Some of the most commonly used editing techniques are changing the width of rows and columns, altering text, and moving text. Each of these topics are discussed below:

- **Changing column width and altering text**. To change column widths, move the mouse arrow above the lines defining the columns until the arrow changes to a double-headed arrow faceing left and right. When you see this new arrow, press down on your left mouse button, then drag the line until the column is the width you want, then release your mouse button.
- **Editing text**. First double-click on the cell you wish to edit, then place your cursor on that cell and modify or replace the existing text. For example, in the frequency table shown below, the table was double-clicked to activate it, then the pivot table's title was double-clicked to activate the title. The original title, "Employment Category," was modified by adding the additional text, "as of August 1999."
- **Using basic editing commands, such as *cut*, *copy*, *delete*, and *paste***. When you cut and copy rows, columns, or a combination of rows and columns by using the *Edit* menu's options, the cell structure is preserved and these values can easily be pasted into a spreadsheet or table in another application.



D

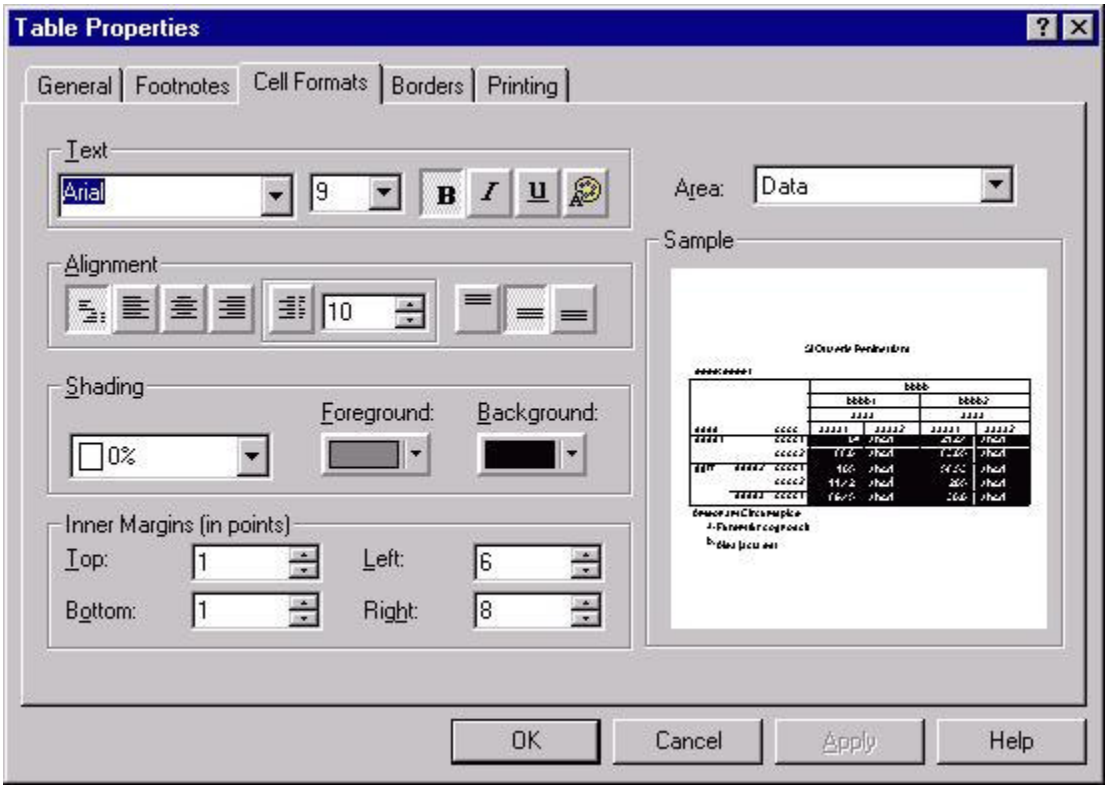
Aside from changing the text in a table, you may also wish to change the appearance of the table itself. But first, it is best to have an understanding of the SPSS *TableLook* concept. A TableLook is a file that contains all of the information about the formatting and appearance of a table, including fonts, the width and height of rows and columns, coloring, etc. There are several predefined TableLooks that can be viewed by first right-clicking on an active table, then selecting the *TableLooks* menu item. Doing so will produce the following dialog box:



D

You can browse the available TableLooks by clicking on the file names in the *TableLook Files* box, as shown above. This will show you a preview of the TableLook in the *Sample* box.

While the *TableLooks* dialog box provides an easy way to change the look of your table, you may wish to have more control of the look or create your own TableLook. To modify an existing table, right-click on an active pivot table, then select the *Table Properties* menu item. This will produce the following dialog box:



D

The above figure shows the *Table Properties* dialog box with the **Cell Formats** tab selected. You can alternate between tabs (e.g., *General*, *Footnotes*, etc.) by clicking on the tab at the upper left of the dialog box. While a complete description of the options available in the *Table Properties* dialog box is beyond the scope of this document, there are a few key concepts that are worth mentioning. Note the *Area* box at the upper right of the dialog box. This refers to the portion of the box that is being modified by the options on the left side of the box. For example, the color in the *Background* of the *Data* portion of the table was changed to black and the color of the text was changed to white by first choosing *Data* from the *Area* box, then selecting black from the *Background* drop-down menu and selecting white for the text by clicking on the color palette icon in the *Text* area on the left side of the dialog box.

The *Printing* tab also has some useful options. For example, the default option for three-dimensional tables containing several layers is that only the visible layer will be printed. One of the options under the *Printing* tab allows you to request that all layers be printed as individual tables. Another useful *Printing* option is the *Rescale wide/long tables to fit page*, which will shrink a table that is larger than a page so that it will fit on a single page.

Any modifications to a specific table can be saved as a TableLook. By saving a TableLook, you will be saving all of the layout properties of that table and can thus apply that look to other tables in the future. To save a TableLook, click on the **General** tab in the *Table Properties* dialog box. There are three buttons on the bottom

right of this box. Use the **Save Look** button to save a TableLook. That button will produce a standard *Save As* dialog box with which you can save the TableLook you created.

Exporting Tables in SPSS

In addition to modifying a table's appearance, you may also wish to export that table. There are three primary ways to export tables in SPSS. To get a menu that contains the available options for exporting tables, right-click on the table you wish to export. The three options for exporting tables are: *Copy*, *Copy object*, and *Export*.

The *Copy* option copies the text and preserves the rows and columns of your table but does not copy formatting, such as colors and borders. This is a good option if you want to modify the table in another application. When you select this option, the table will be copied into your system clipboard. Then, to paste the table, select the *Paste* command from the *Edit* menu in the application to which you are importing the table. The *Copy* option is useful if you plan to format your table in the new application; the disadvantage of this method is that only the text and table formatting remains and you will therefore lose much of the formatting that you observe in the Output Viewer.

The *Copy object* method will copy the table exactly as it appears in the SPSS Output Viewer. When you select this option, the table will be copied into your clipboard and can be imported into another application by selecting the *Paste* option from the *Edit* menu of that application. When you paste the table using this option, it will appear exactly as it is in the Output Viewer. The disadvantage of this method is that it can be more difficult to change the appearance of the table once it has been imported.

The third method, *Export*, allows you to save the table as an HTML or an ASCII file. The result is similar to the the *Copy* command: you will have a table that retains the text and cell layout of the table you exported, but it will retain little formatting. This method for exporting tables to other applications is different from the above two methods in that it creates a file containing the table rather than placing a copy in the system clipboard. When you select this method, you will immediately be presented with a dialog box allowing you to choose the format of the file you are saving and its location on disk. The primary advantage of this method is that you can immediately create an HTML file that can be viewed in a Web browser.

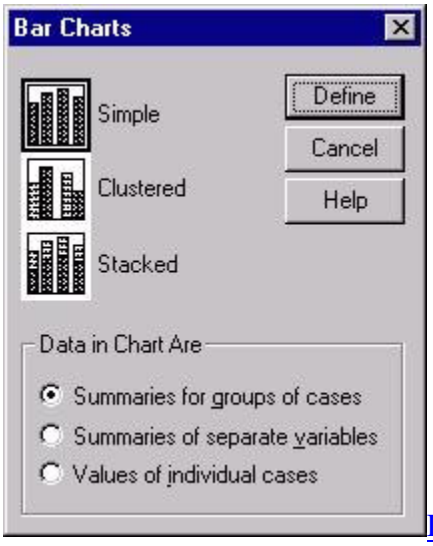
Bar Graphs

Bar graphs are a common way to graphically display the data that represent the frequency of each level of a variable. They can be used to illustrate the relative frequency of levels of a variable by graphically displaying the number of cases at each level of the variable. For example, you could compare the distribution of employees among employment categories by plotting the frequency of each job category next to the other job categories. If you want to develop a similar display for a continuous variable, examine the *Histogram* option on the *Graphs* menu. To access

the SPSS facilities for visually displaying data, including the bar graph procedure, select the *Bar* option from the *Graphs* menu:

Graphs
Bar...

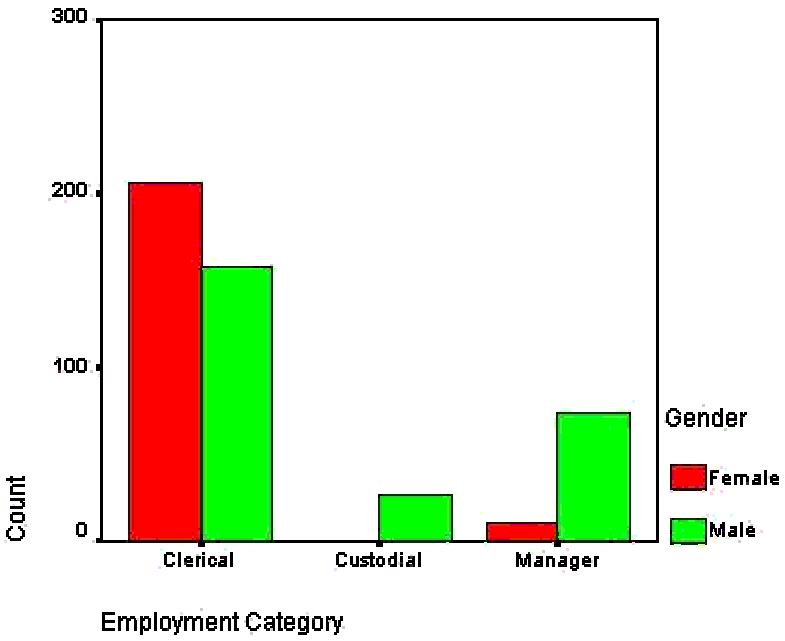
This will produce the following dialog box:



D

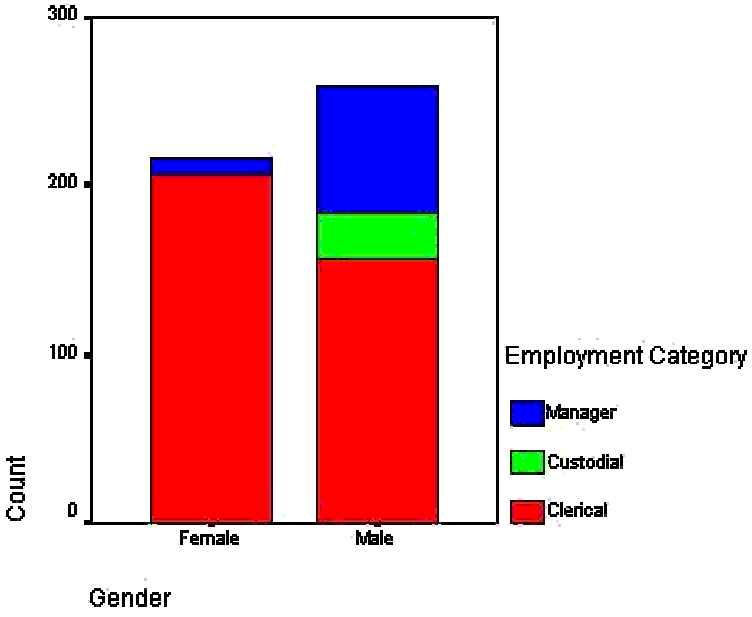
In this dialog box, you must choose between the three types of bar graphs offered by SPSS. The *Simple* bar graph is the most common one and is used to graph frequencies of levels within a variable. In the *Simple* bar graph, the category axis is the variable and each bar represents a level of the variable.

The other two types of bar graphs are used in situations where you want to graph frequencies for more than one variable. For example, if you want to graph job category by gender of employee, then you would choose either the *Clustered* or *Stacked* options in the above dialog box. The *Clustered* option groups together bars representing levels of a category. For example, in graphing gender by job category, you could produce a bar graph in which male and female, the levels of the gender variable, are grouped next to each other for each of the three levels of the job-category variable:



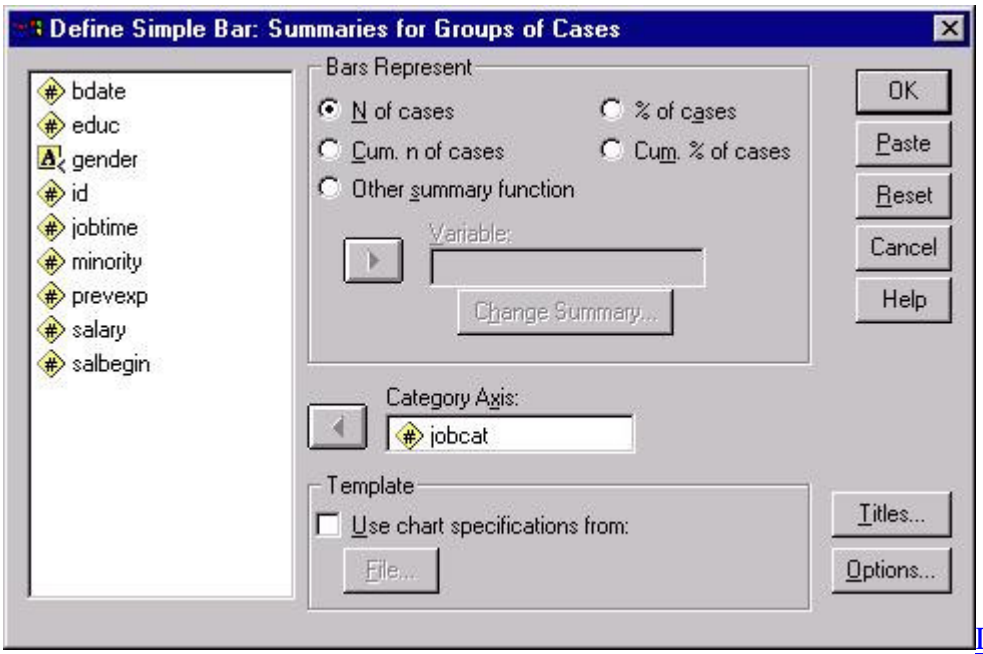
D

The other bar graph option for two variables is the *Stacked* option, which will have a bar for each level of one of the variables, while the levels of the other variable are placed on top of each other within each bar. The frequency of each level within a variable in a stacked bar graph is represented by a different color. For example, you could create a graph in which there was a bar for the levels of gender, male and female, and the numbers of males in females within each employment category are represented by a different color within a bar:



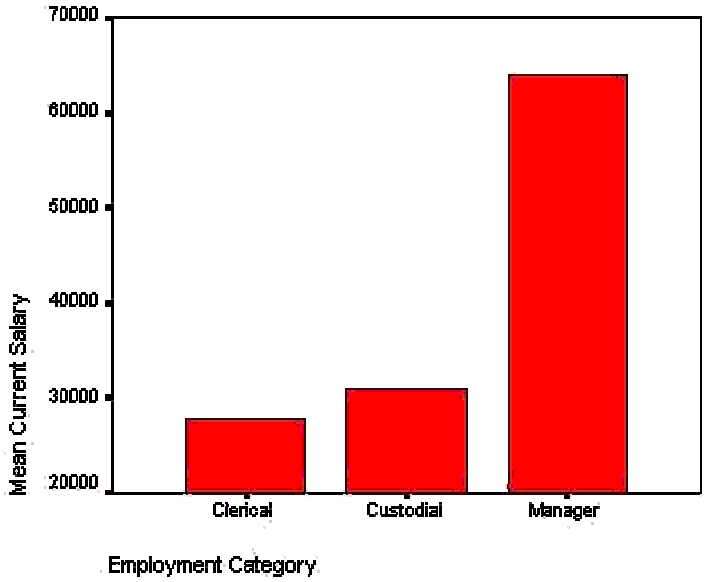
D

This discussion will focus on the the *Simple* graph. The options for all three are similar, so an understanding of the *Simple* option will suffice for either the *Clustered* or *Stacked* options. To get started with the bar graph, click on the icon representing the type of graph that you want, then click on the **Define** button to produce the following dialog box:



In this dialog box, the *Category Axis* is the only box that you must fill in. Click on a variable name in the list of variables at left and move it to the *Category Axis* box by clicking on the arrow next to that box. You should also select the summary information that you want in the *Bars Represent* section of the box. The *N of cases* option is the default and is the most common way of summarizing data in a bar graph. There is little difference between using the *N of cases* and the *% of cases* options, other than the unit of measurement in the vertical axis. Using either the *Cum. n of cases* or the *Cum. % of cases* options will produce a cumulative graph in which the first bar represents the level of a variable with the fewest number of cases, then then next bar represents the level with the second fewest number of cases in addition to the level that has previously been graphed, and so on.

In addition to these four options, you can also use the *Other summary function* option to specify another summary statistic, such as a variable's mean, sum, or variance. As in the above example, begin by selecting a *Category Axis* variable. Next, select the *Other summary function*, then place another variable in the box labels *Variable*. For example, you could select the variable *salary* to obtain the mean salary for each level of the variable *jobcat*. Doing so would produce the following table:



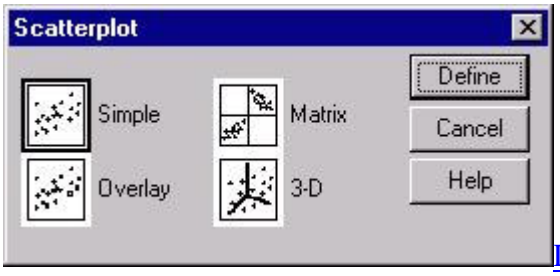
D

Scatterplots

Scatterplots give you a tool for visualizing the relationship between two or more variables. Scatterplots are especially useful when you are examining the relationship between continuous varaibles using statistical techniques such as correlation or regression. Scatterplots are also often used to evaluate the bivariate relationships in regression analyses. To obtain a scatterplot in SPSS, go to the *Graphs* menu and select *Scatter*:

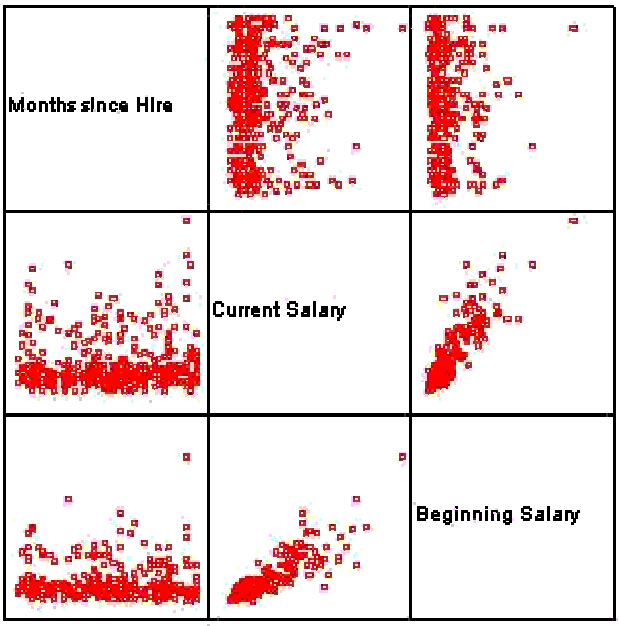
Graphs
Scatter...

This will produce the following dialog box:



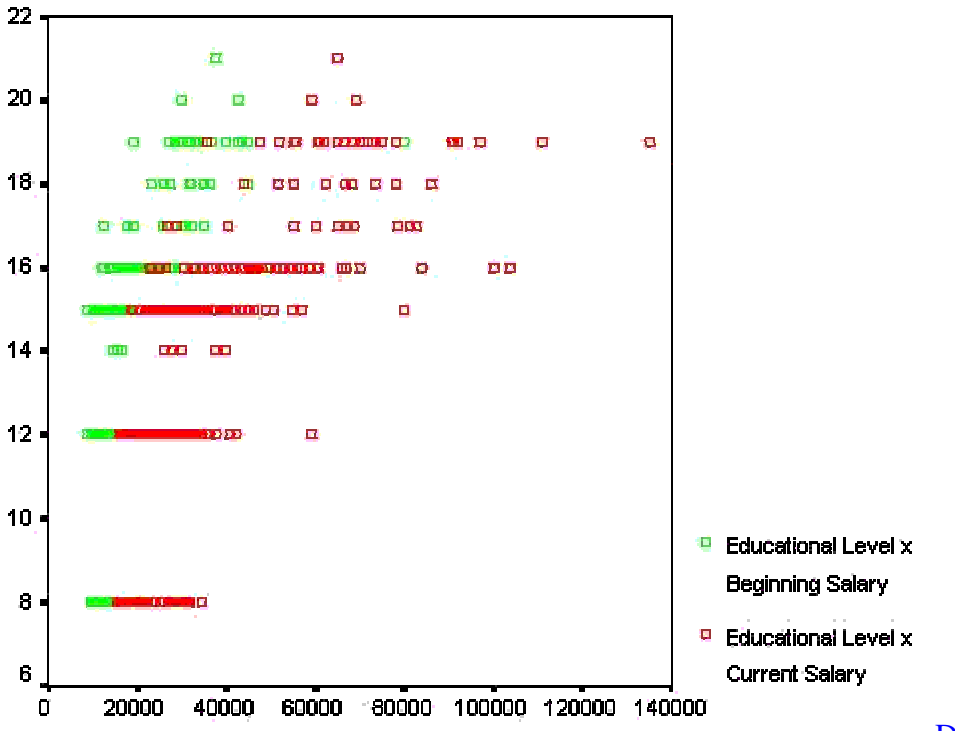
D

Note that there are four scatterplot options. The *Simple* option graphs the relationship between two variables. The *Matrix* option is for two or more variables that you want graphed in every combination: variable is plotted with every other variable. Every combination is plotted twice so that each variable appears on both the X and Y axis. For example, if you specified a *Matrix* scatterplot with three variables, *salary*, *salbegin*, and *jobtime*, you would receive the following scatterplot matrix:



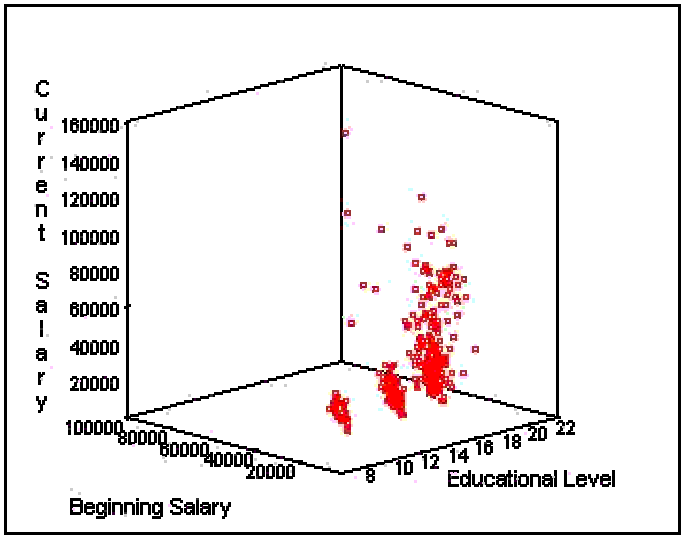
D

The third scatterplot option is the *Overlay* option. It allows you to plot two scatterplots on top of each other. The plots are distinguished by color on the overlaid plot. For example, you could plot education by beginning and current salaries by pairing the variables *educ* with *salbegin* and *educ* with *salary* using our example dataset. Doing so would produce the graph shown below. In that graph, the green points represent the *educ* x *salbegin* plot whereas the red points represent the *educ* x *salary* plot.



D

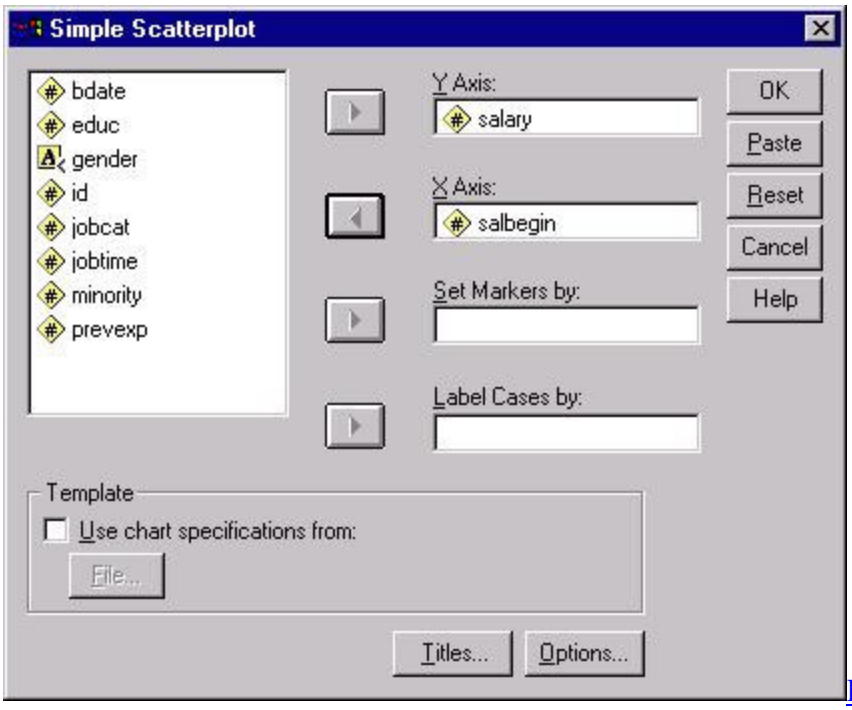
The fourth option for scatterplots is the 3-D scatterplot. This is used to plot three variables in three dimensional space. Here is an example of the *3-D* option, containing the variables, *educ*, *salary*, and *salbegin*:



D

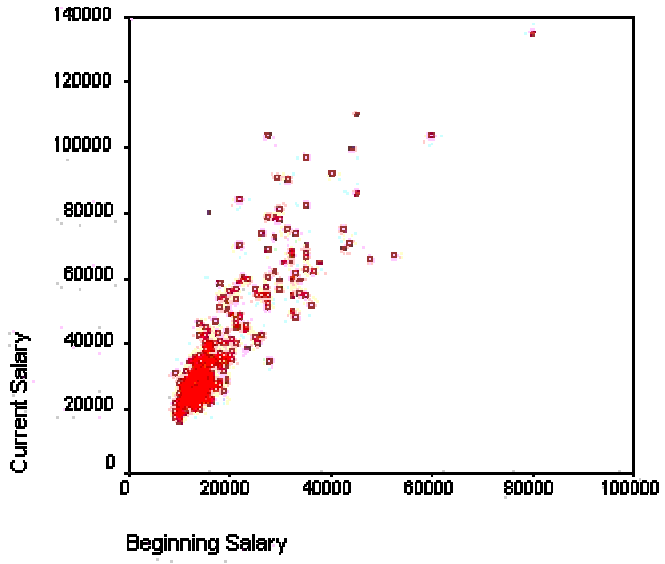
Returning to the *Simple* scatterplot option, you can examine some of the the options that are commonly used with a basic scatterplot by plotting the values of two

variables. When you select the *Simple* option from the initial dialog box, you will get the following dialog box:



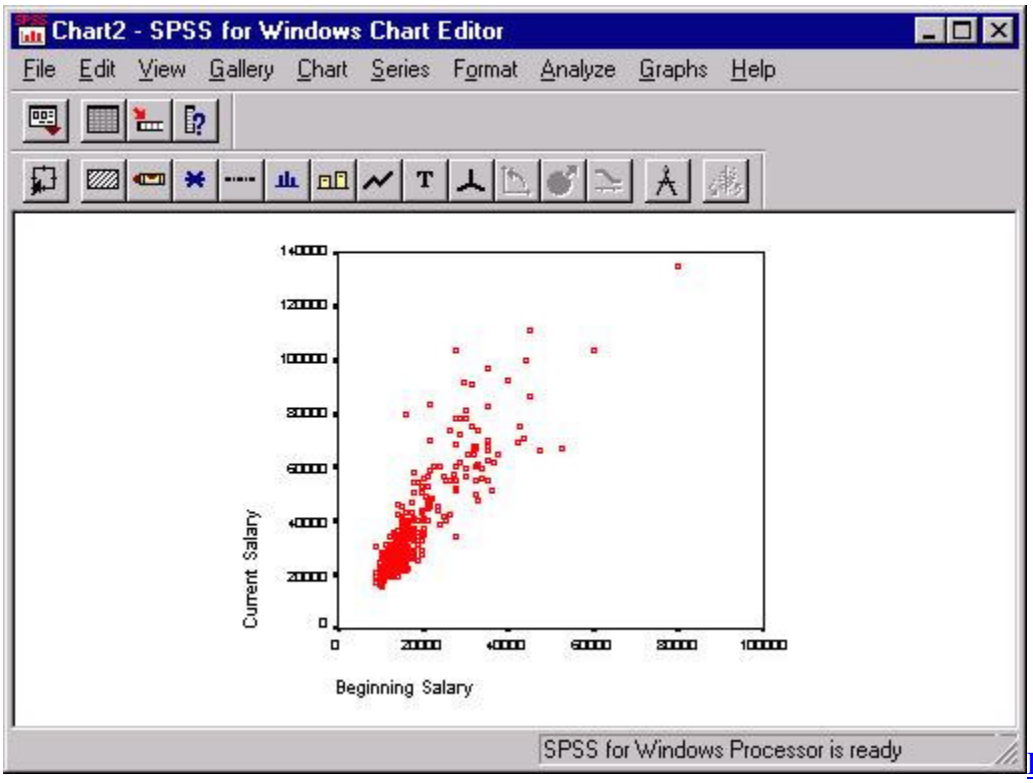
D

Select the two varaibles that you want to plot from the list on the left, placing one in the *Y axis* box and the other in the *X axis* box. If you have multiple groups (e.g., males and females) in your dataset, you can have SPSS draw different colored markers for each group by entering a group variable in the *Set Markers by* box. You can add titles to your chart by clicking on the **Titles** button. Clicking **OK** will produce the following scatterplot:



D

Some of the most useful options for modifying your scatterplot are only available after you have the initial scatterplot created. To use these options, double-click on the chart. This will open the chart in a new window, as shown below:

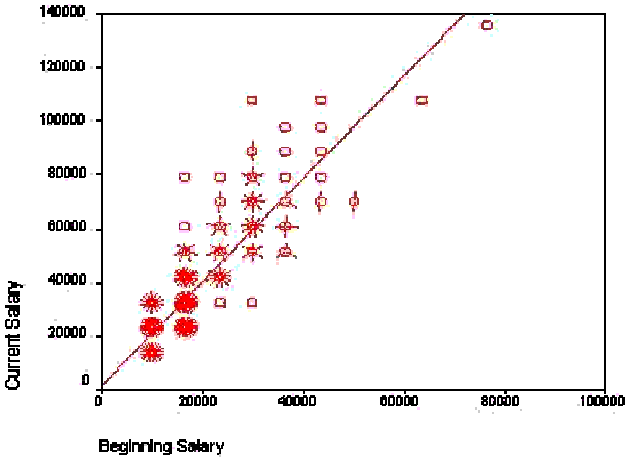


D

From this window, you have several options for modifying your chart. We will only deal with scatterplot-specific options here, as many of the general options will be covered in the next section. To get the scatterplot options, select *Options* from the *Chart* menu:

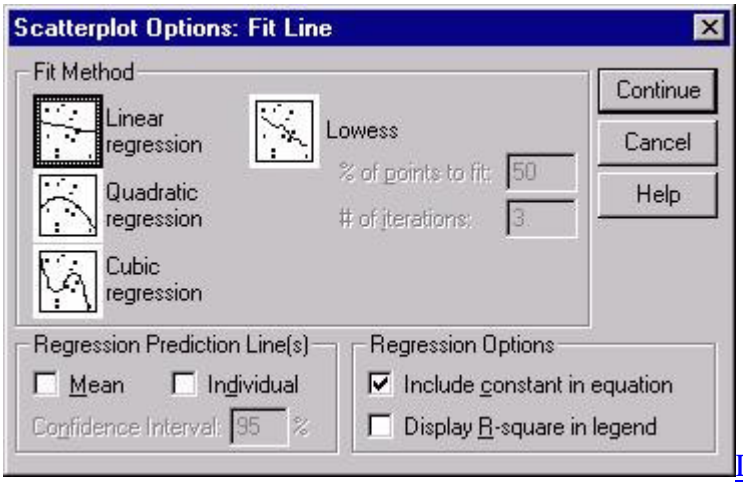
Chart
Options...

To get the following dialog box:



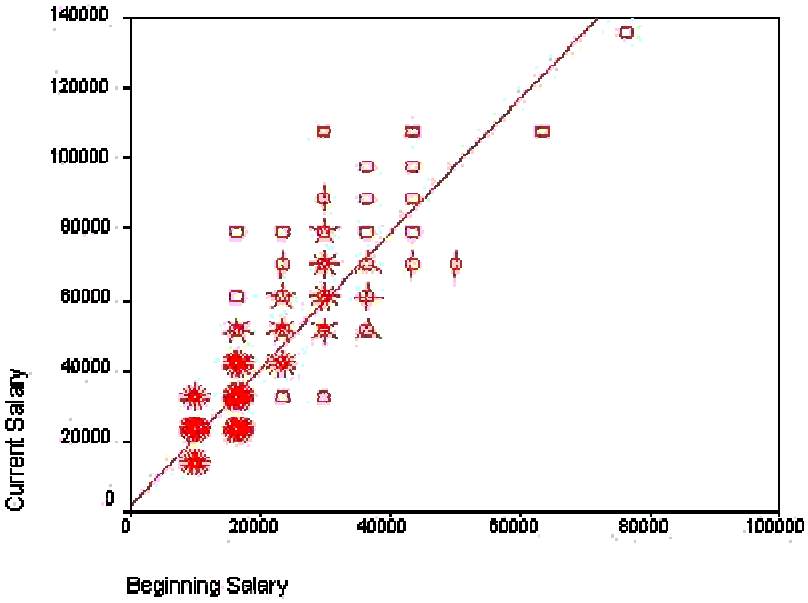
D

Some of the most useful options that will add information to your scatterplot are the *Sunflowers* and the *Fit Line* options. The *Show sunflowers* option draws spikes that resemble petals for each case where there are overlapping cases in the scatterplot. Clicking on the **Sunflower Options** button will then give you several options for defining how fine or coarse you want the spikes to appear and how many cases are represented by each spike in your scatterplot. See the chart at the end of this section for an example of sunflowers on a scatterplot. The *Fit Line* option will allow you to plot a regression line over your scatter plot. Click on the **Fit Options** button to get this dialog box:



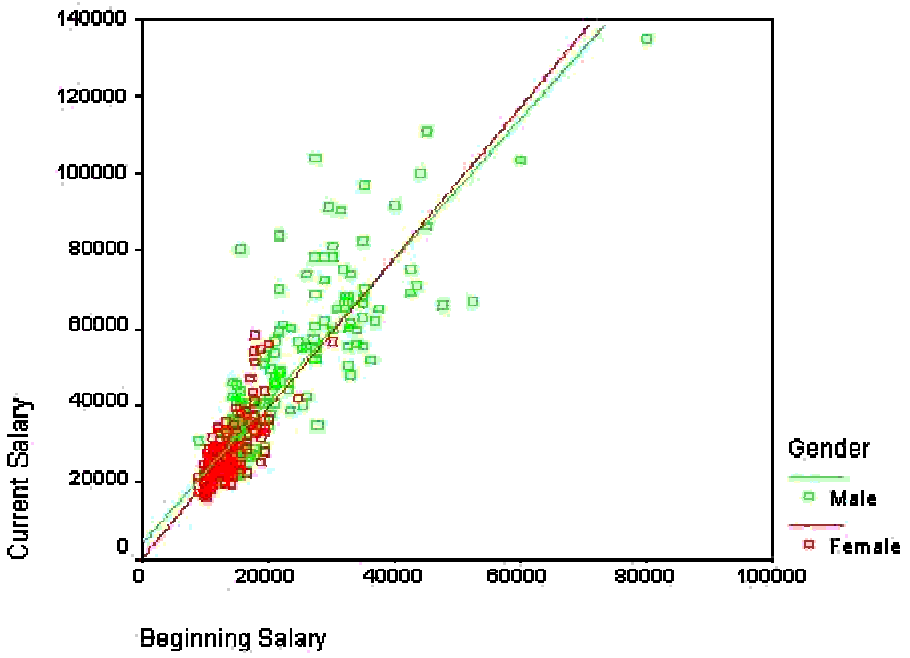
D

Next, click on the type of regression line you want in the *Fit Method* box. The *Linear regression* line is the most commonly plotted regression line, and it represents the linear increase in the Y variable as a function of increases in the X variable. The graph below illustrates a linear regression line plotted on a sunflower scatterplot. If you have a quadratic or cubic term in your model, you will likely be interested in plotting one of those fit lines.



D

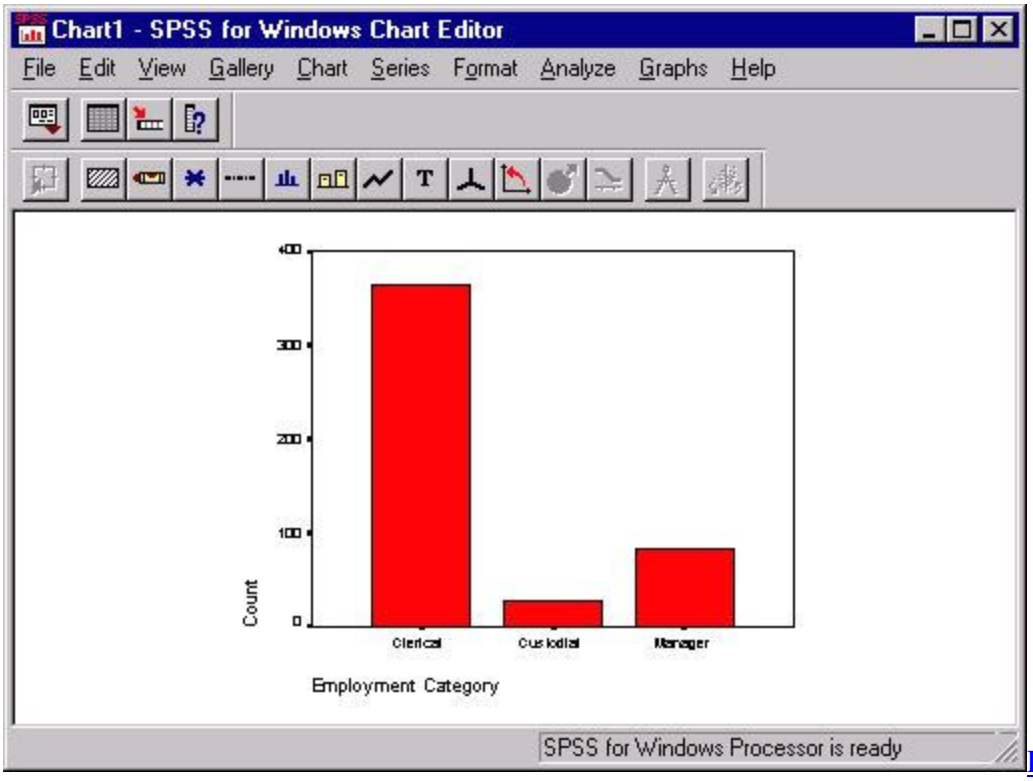
One common extension of the above example is to plot separate regression lines for subgroups. For example, you could plot separate regression lines for men and women in the above example to visually examine whether the relationship between beginning salaries and current salaries is the same for both men and women. To do this, you will first need to select a categorical variable in the main scatterplot dialog box (this is the dialog box that is labeled *Simple Scatterplot*). Place a categorical variable in the box labeled *Set Markers by*. After doing that, follow the rest of the above steps until you reach the dialog box labeled *Scatterplot Options*. There, you can define separate regression lines for your groups by selecting the *Subgroups* option in the *Fit Line* section of the *Scatterplots* dialog box. Doing so produces the graph shown below:



D

Modifying and Exporting Graphs

The primary tool for modifying charts in SPSS is the *Chart Editor*. The Chart Editor will open in a new window, displaying a chart from your Output Viewer. The Chart Editor has several tools for changing the appearance of your charts or even the type of chart that you are using. To open the Chart Editor, double-click on an existing chart and the Chart Editor window will open automatically. The Chart Editor shown below contains a bar graph of employment categories:



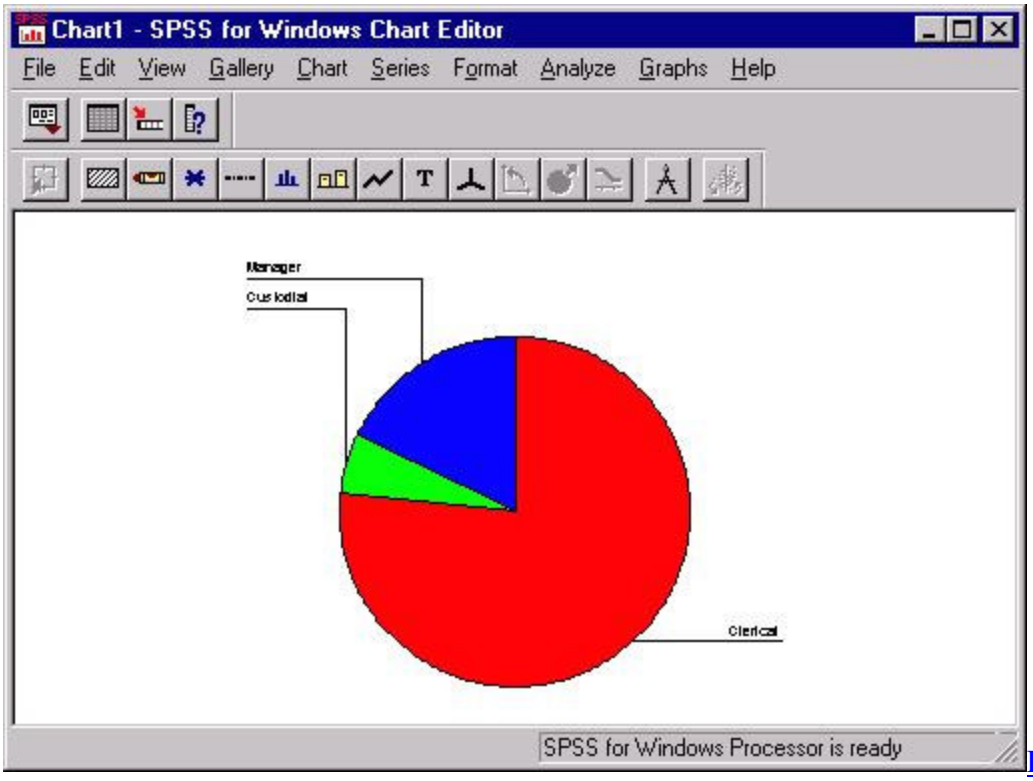
D

While there are many useful features in the Chart Editor, we will concentrate on the three of them: changing the type of chart, modifying text in the chart, and modifying the graphs.

You can change the type of chart that you are using to display your data using the Chart Editor. For example, if you want to compare how your data would look when displayed as a bar graph and as a pie chart, you can do this from the *Gallery* menu:

Gallery
Pie...

Selecting this option will change the above bar graph into the following pie chart:



D

Once you have selected your graphical look, you can start modifying the appearance of your graph. One aspect of the chart that you may want to alter is the text, including the titles, footnotes, and value labels. Many of these options are available from the *Chart* menu. For example, the *Title* option could be selected from the *Chart* menu to alter the charts title:

Chart
Title...

Selecting this menu item will produce the following dialog box:

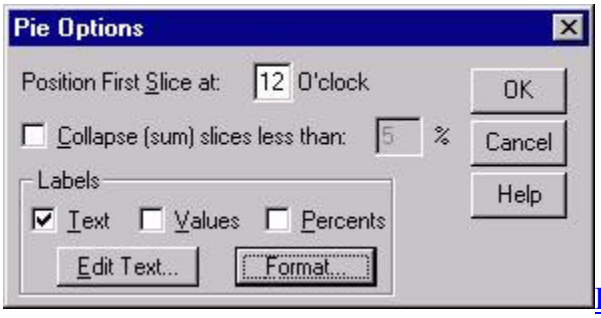
D

The title "Employment Categories" was entered in the box above and the default justification was changed from left to center in the *Title Justification* box. Clicking **OK** here would cause this title to appear at the top center of the above pie chart. Other text in the chart, such as footnotes, legends, and annotations, can be altered

similarly. The labels for the individual slices of the pies can also be modified, although it may not be obvious from the menu items. To alter the labels for areas of the pie, choose the *Options* item from the *Chart* menu.

Chart
Options...

This will produce the following dialog box:



D

In addition to providing some general options for displaying the slices, the *Labels* section enables you to alter the text labeling slices of the pie chart as well as format that text. You can click the **Edit Text** button to change the text for the labels. Doing so will produce the following dialog box:

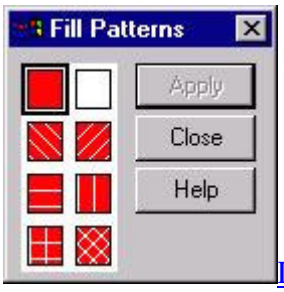
D

To edit individual labels, first click on the current label, which will be displayed below the *Label* box, then alter the text in the *Label* box. When you finish, click the **Continue** button to return to the *Pie Options* dialog box. You can make changes to the format of your labels by clicking the **Format** button here. If you do not want to to change formatting, click on **OK** to return to the Chart Editor.

In addition to altering the text in your chart, you may also want to change the appearance of the graph with which you are working. Options for changing the appearance of graphs can be accessed from the *Format* menu. Many options available from this menu are specific to a particular type of graph. There are some general options that are worth discussing here. One such is *Fill Pattern* option, which changes the pattern of the graph. It can be obtained by selecting the *Fill Pattern* option from the *Format* menu:

Format
Fill Pattern...

This will produce the following dialog box:



D

First, click on the portion of the graph where you want to change the pattern, then select the pattern you want by clicking on the pattern sample on the left side of the dialog box. Then, click the **Apply** button to change the appearance of your graph.

One other formatting option that is generally useful is the the ability to change the colors of your graphs. To do that, select the *Color* option from the *Format* menu:

Format
Color...

To produce the following dialog box:

D

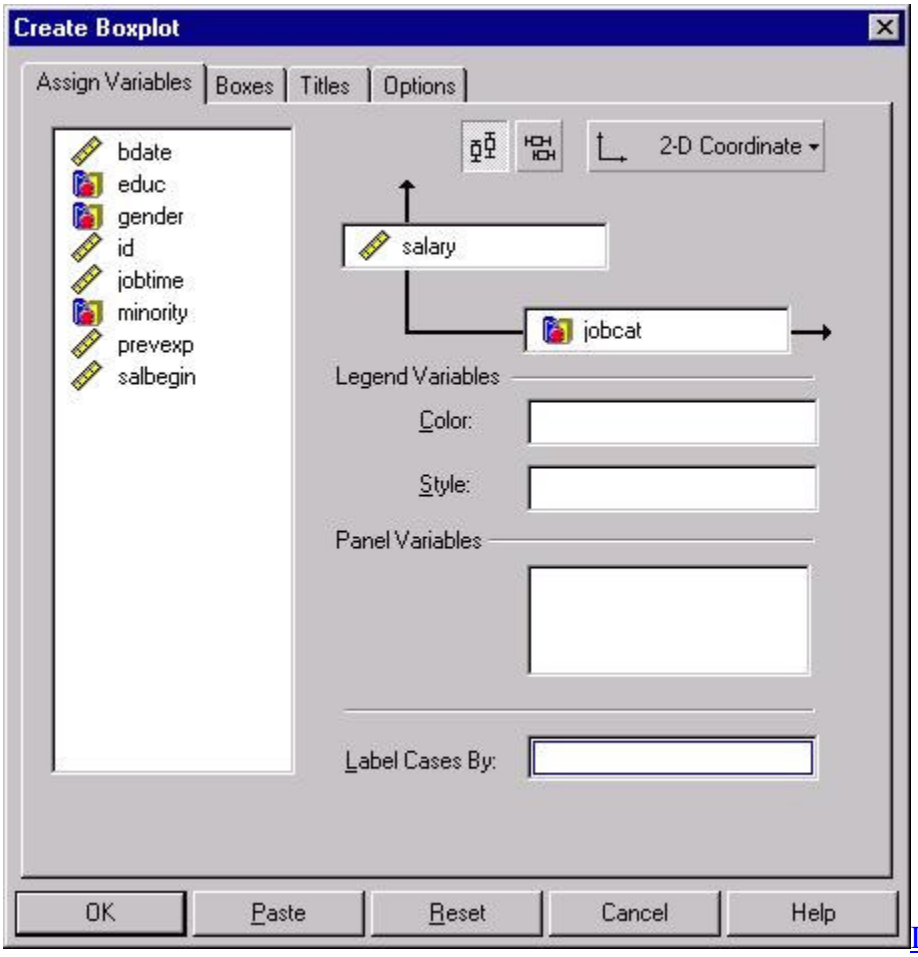
This will allow you to change the color of a portion of a graph and its border. First, select the portion of the graph for which you would like to change its color, then select the *Fill* option if you want to change the color of a portion of the graph and select the *Border* option if you want to change the color of the border for a portion of the graph. Next, click on the color that you want and click **Apply**. Repeat this process for each area or border in the graph that you want to change.

Interactive Charts

Many of the standard graphs available through SPSS are also available as *interactive charts*. Interactive charts offer more flexibility than standard SPSS graphics: you can add variables to an existing chart, add features to the charts, and change the summary statistics used in the chart. To obtain a list of the available interactive charts, select the *Interactive* option from the *Graphs* menu:

Graphs
Interactive

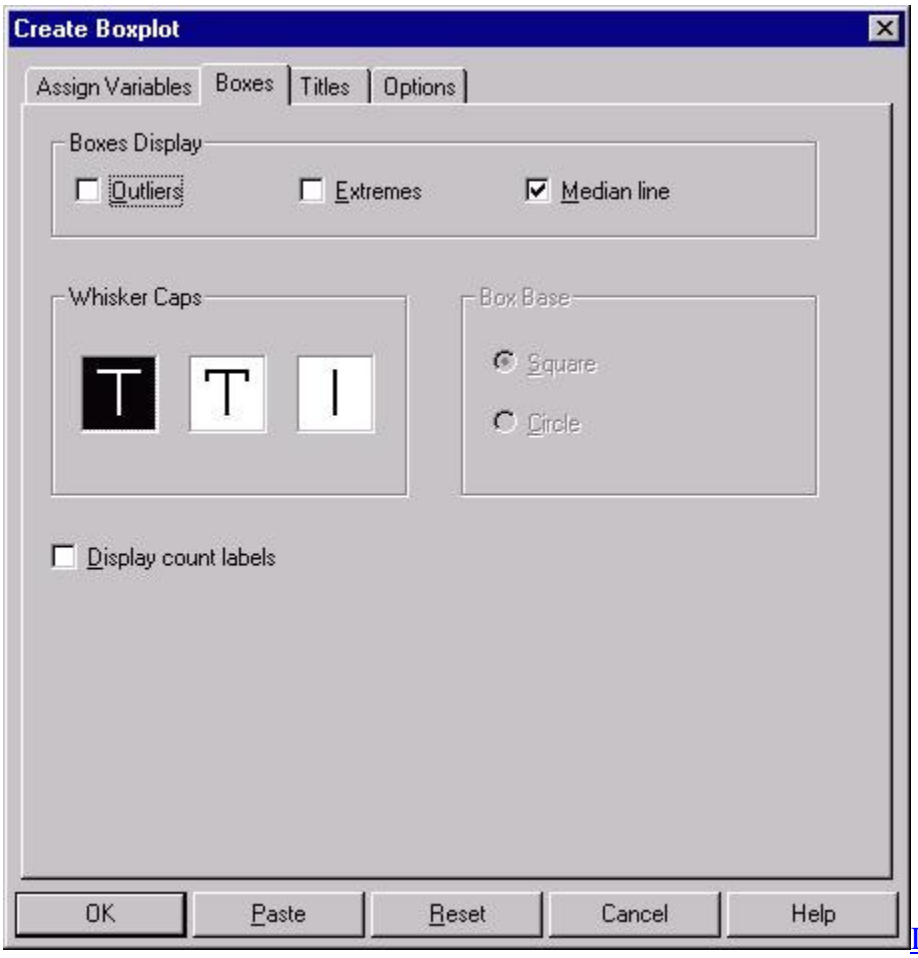
Selecting one of the available options will produce a dialog for designing an interactive graph. For example, if you selected the *Boxplot* option from the menu, you would get this dialog box:



D

Dialog boxes for interactive charts have many of the same features as other SPSS dialog boxes. For example, in the above dialog box, the variable type is represented by icons: scale variables, such as the variable *bdate*, are represented by the icon that resembles a ruler, while categorical variables, such as the variable *educ*, are represented by the icon that resembles a set of blocks. Variables in the variable list on the left of the dialog box can be moved into the boxes on the right side of the screen by dragging them with your mouse, in contrast to using the arrow button used in other

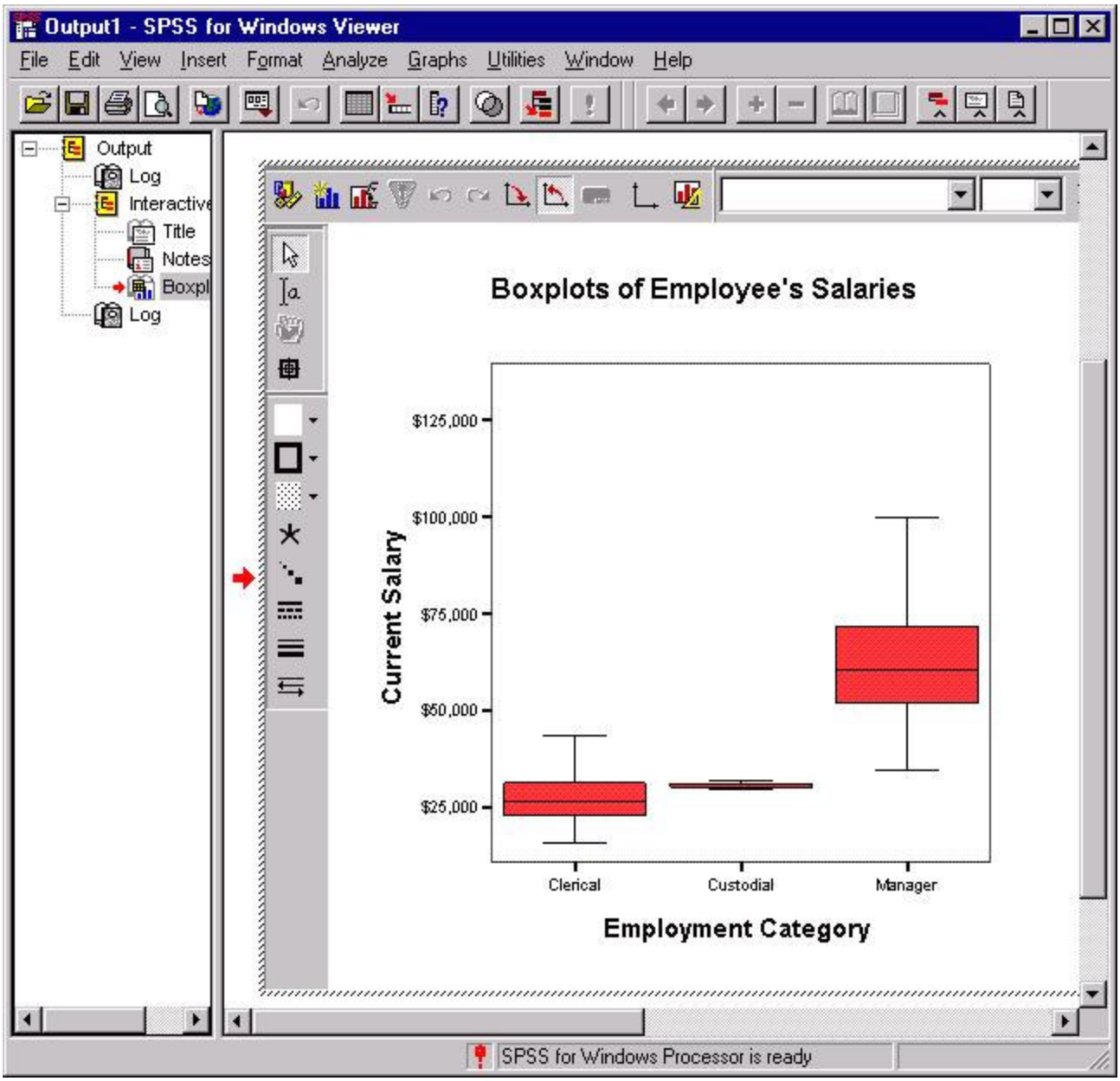
SPSS dialog boxes. Options in interactive graphs can be accessed by clicking on the tabs. For example, clicking on the *Boxes* tab produces the following dialog box:



D

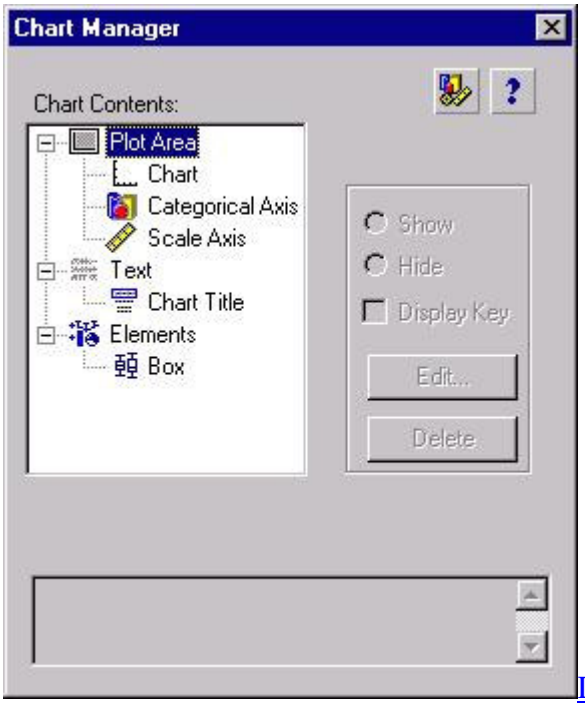
Here, you have several choices about the look of your boxplot. The choice to display the median line is selected here, but the options to indicate outliers and extremes are not selected. The *Titles* and *Options* tabs offer several other choices for altering the look of your table as well, although a thorough discussion of these is beyond the scope of this document. When you have finished the specifications for a graph, click the **OK** button to produce the graph you have specified in the Output Viewer.

Interactive graphs offer several choices for altering the look of the chart after you have a draft in the Output Viewer. To get the menus for doing that, double-click on the interactive graph that you want to alter. For example, double-clicking on the boxplot obtained through the above dialog box will produce the following menus:



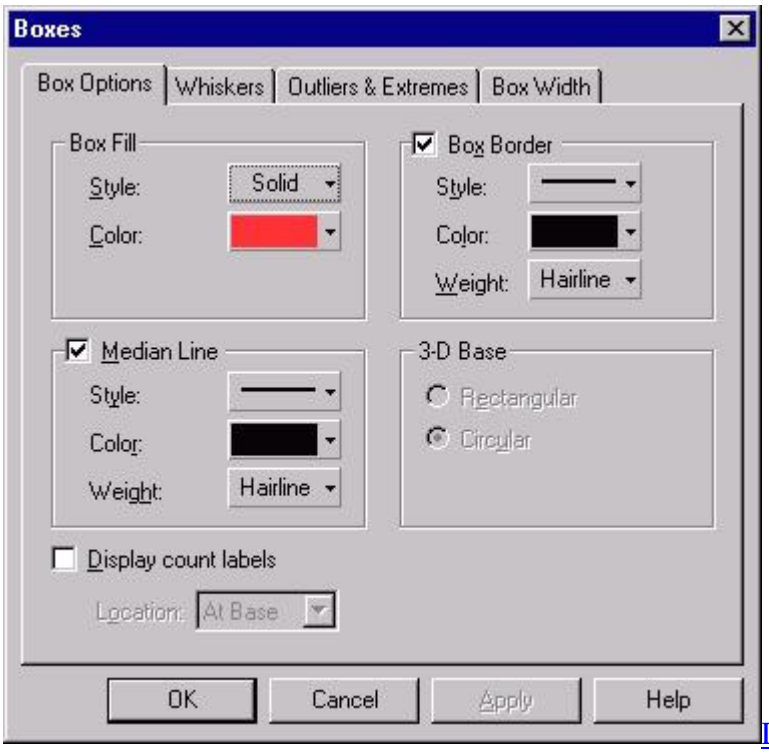
D

The icons immediately surrounding the graph provide you with several possibilities for altering the look of your graph. The three leftmost items in the horizontal menu are worthy of mention. The leftmost icon produces a dialog box that resembles the original *Interactive Graphs* dialog box and contains many of the same options. For example, you could change the variables that you are graphing using this dialog box. The next icon, the small bar graph, lets you add additional graphical information. For example. you could overlay a line that graphed the means of the three groups in the above graph by choosing the *Dot-Line* option from the menu, or you could add circles representing individuals salaries within each group by choosing the *Cloud* option. The third icon provides several options for changing the look of your chart. Selecting that icon will produce the following dialog box:



D

Each icon in this dialog box can be double-clicked to produce a dialog box that contains the properties of the component of the chart represented by that icon. For example, you could obtain the properties of the boxes in the interactive graph above by double-clicking on the icon labeled *Box*. Doing so would produce this dialog box:



D

Changing the properties in this or any other dialog box that controls the properties of any portion of the chart will change the look of the graph in the Output Viewer. For example, you could change the colors of the boxes and their outlines by selecting a different color from the above drop-down menus.

This document is the fourth module of a four module tutorial series. This module describes the use of SPSS to do advanced data manipulation such as splitting files for analyses, merging two files, aggregating datasets, and combining multiple tables in a database into an SPSS dataset in the first section. Several advanced topics are included in the second section, including the use of SPSS syntax, the SPSS Visual Basic editor, and SPSS Macros.

Some users prefer to use keystrokes to navigate through SPSS. Information on common keystrokes are available in our SPSS 10 for Windows Keystoke Manual. If you are not familiar with SPSS or need more information about how to get SPSS to read your data, consult the first module, SPSS for Windows: Getting Started, of this four part tutorial. This set of documents uses a sample dataset, *Employee data.sav*, that SPSS provides. It can be found in the root SPSS directory. If you installed SPSS in the default location, then this file will be located in the following location: C:\Program Files\SPSS\Employee Data.sav.

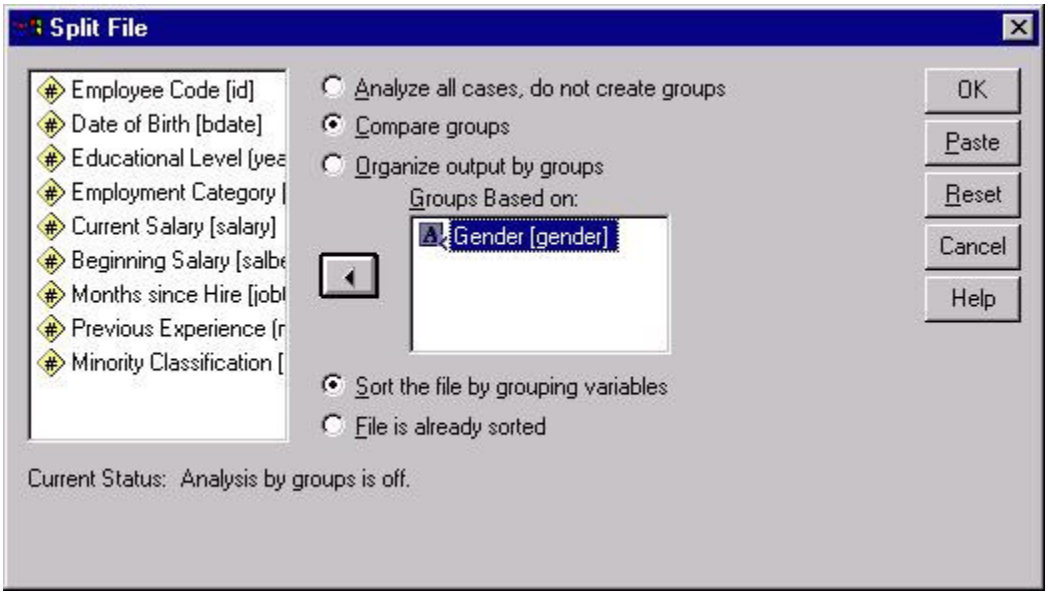
Section 7: Data Manipulation

Splitting Files

In some situations, you may want to perform the same analysis on different groups within the same dataset. For example, if you are comparing gender differences in salaries for men and women, you may want to first obtain separate sets of descriptive statistics such as the mean, standard deviation, etc. for men and women. Analyses such as these can be conducted by first selecting the *Split File* function from the *Data* menu in the Data Editor:

Data
  Split File...

This will produce the following dialog box:



You have three possible choices for the organization of your analyses. The default choice is the *Analyze all cases, do not create separate groups*. With this option, all analyses are conducted on the entire dataset. Because the *split file* command remains in effect indefinitely, you should reset this option when you no longer want a split file analysis.

The next two choices, *Compare groups* and *Organize output by groups* result in the same values in the output, regardless of the analysis being performed, but they differ in the way in which the output is presented. These two options require you to select one or more variables from the variable list on the left side of the dialog box by first clicking on variables by which the file is to be split, then clicking on the arrow to the right of the variable list. This will display the variables by which the file is to be split in the *Groups Based on* box. Clicking the **OK** button is the final step to activate the split file. If the *Compare groups* option was specified, as shown above, then all groups will be analyzed separately, but information on all groups will be contained in the same table in the output. Here is an example of some descriptive statistics with this option active:

Descriptive Statistics

Gender		N	Minimum	Maximum	Mean	Std. Deviation
Female	Current Salary	216	$15,750	$58,125	$26,031.92	$7,558.02
	Valid N (listwise)	216				
Male	Current Salary	258	$19,650	$135,000	$41,441.78	$19,499.21
	Valid N (listwise)	258				

The *Organize output by groups* option will produce separate tables for each group. The example below shows the result of selecting the identical options for descriptive statistics, as above, but with the *Organize output by groups* option active.

Gender = Female

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	216	$15,750	$58,125	$26,031.92	$7,558.02
Valid N (listwise)	216				

a. Gender = Female

Gender = Male

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	258	$19,650	$135,000	$41,441.78	$19,499.21
Valid N (listwise)	258				

a. Gender = Male

Merging Files

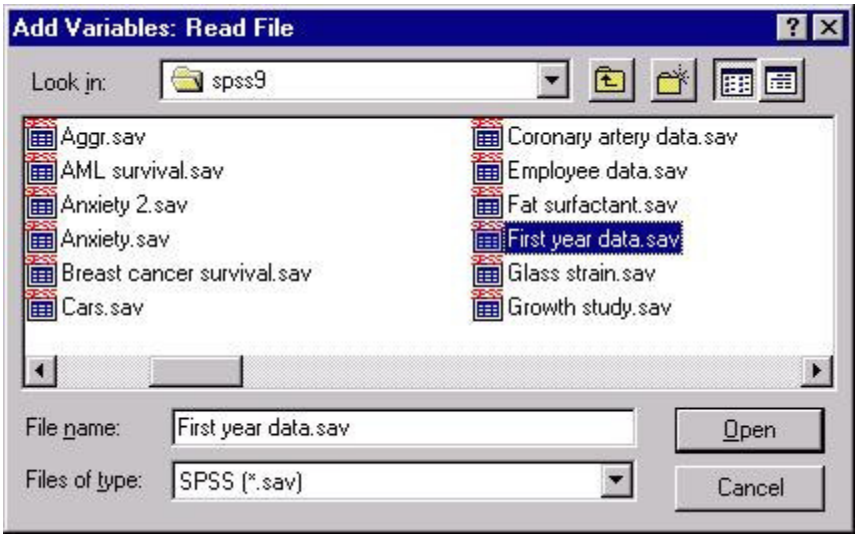
At times, you may want to conduct an analysis on data that is stored in more than one data file. Because SPSS only has one working data file at a time, you must first merge data into a single file for the purpose of analysis. For example, if there were another dataset that contained employee's salaries after one year and you wanted to analyze that information with the data in the current dataset, then you would first need to merge these two files.

But before merging files, keep these three considerations in mind: (1) There should be at least on identification variable in both files that contains the same values so that the files can be matched. (2) Variables with the same names in both datasets that are not being used to match cases will be need to be recode or they will be excluded. Therefore, if both datasets contain variables that have the same name but different values, one of these variables will be excluded. (3) You must sort your data in ascending order before performing a file merge.

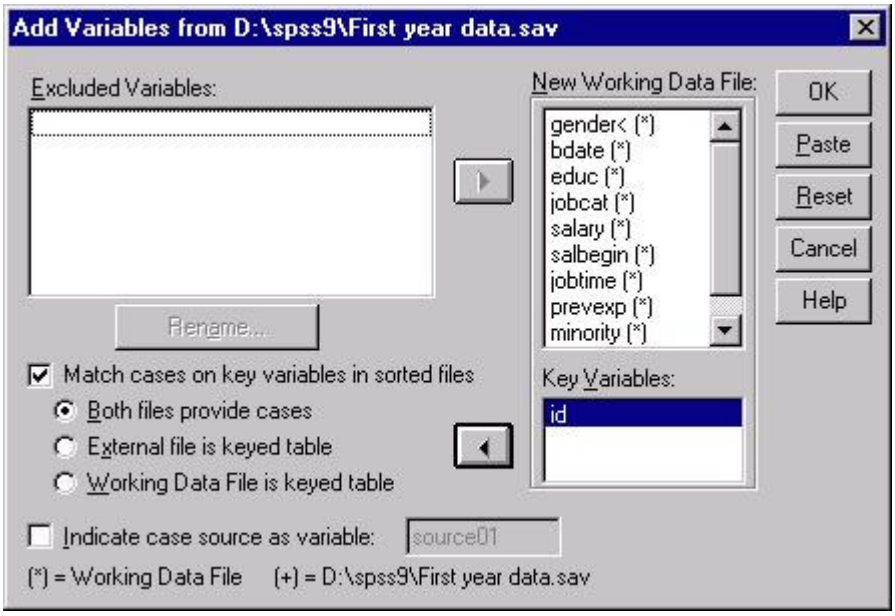
Data files can be merged using dialog boxes available under the *Data* menu item in the Data Editor:

Data
Merge Files

Selecting *Merge Files* will give you two options: *Add Cases* and *Add Variables*. The *Add Cases* option combines two files with different cases that have the same variables, whereas the *Add Variables* adds new variables on the basis of variables that are common to both files. The following example illustrates the use of the *Add Variables* option to combine two files that share a common variable. Here, the *Add Variables* option will be used to combine the working data file, *Employee data.sav* with an external file (*First year data.sav*) that contains two variables, employee ID (*id*) and salary after one year (*oneyear*). The variable *id* has identical values to the variable with the same name in the *Employee data.sav*, dataset whereas the variable *oneyear* is unique to the external data file. After selecting the *Add Variables* option, the following dialog box will appear:

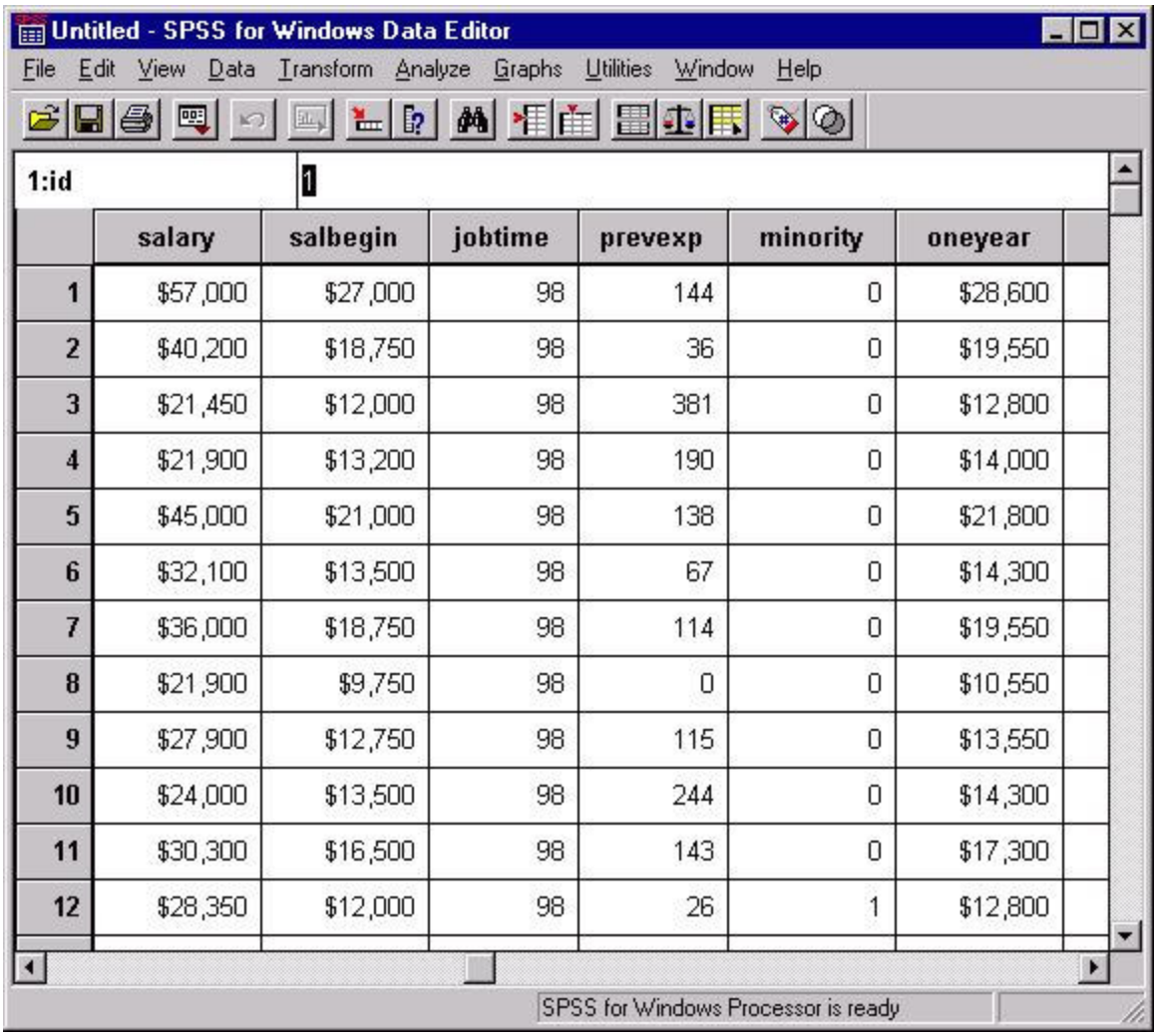


This is the first of two dialog boxes, in which you can select a file to merge with the working data file. Here, the dataset, *First year data.sav* has been selected to merge with the working dataset. This is the dataset containing the data on employees' salaries after one year. After selecting a file for the ensuing merge, click the **Open** button and the following dialog box will appear:



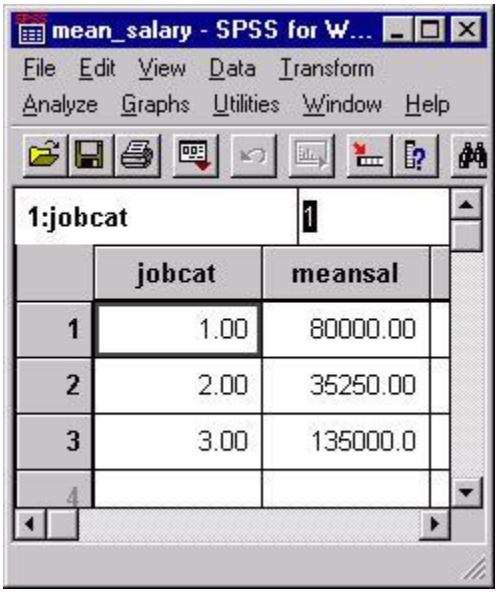
When this box first appears, variable names that appear in both datasets will be in the box labeled *Excluded Variables*. In this example, *id* is the only variable that fits this description. At least one variable must be common to both files in order to perform a merge. All variables that are unique to one of the data files will appear in the box labeled *New Working Data File*, indicating that they will appear in the file that is the merged product of the two files. The file in which these variables originated is indicated by the symbol following the variable name: variables from the working dataset are assigned the '*'; variables that originate from the external dataset are assigned a '+'. The default setting merges all of the nonduplicated variables. Any variables that you do not want in the merged file can be highlighted in the box labeled *New Working Data File* and moved to the *Excluded Variables* box by clicking on the arrow in between these two boxes.

To match files on the basis of one or more variables, click on the box labeled *Match cases on key variables in sorted files*. This will activate the options underneath this heading. Select the option *Both files provide cases* for a standard file merge. The arrow to the left of the box labeled *Keep Variables* will then become active, allowing you to select a variable from the box labeled *Excluded Variables*. In the above example, only the variable *id* appeared in this box because it was the only variable that is in both data files. To move it from the *Excluded Variables* box to the *Key Variables* box, as has already been done above, click on the arrow button to the left of the *Key Variables* box. You are now ready to run the procedure by clicking on the **OK** button. Doing so will produce the following dataset:



	salary	salbegin	jobtime	prevexp	minority	oneyear
1	$57,000	$27,000	98	144	0	$28,600
2	$40,200	$18,750	98	36	0	$19,550
3	$21,450	$12,000	98	381	0	$12,800
4	$21,900	$13,200	98	190	0	$14,000
5	$45,000	$21,000	98	138	0	$21,800
6	$32,100	$13,500	98	67	0	$14,300
7	$36,000	$18,750	98	114	0	$19,550
8	$21,900	$9,750	98	0	0	$10,550
9	$27,900	$12,750	98	115	0	$13,550
10	$24,000	$13,500	98	244	0	$14,300
11	$30,300	$16,500	98	143	0	$17,300
12	$28,350	$12,000	98	26	1	$12,800

In addition to matching files where there is a one-to-one match between cases in the files, you may also want to match two files on the basis of a particular variable. For example, you may want to add to your data file a data column containing the average salary for a person's job category. In this situation, you could match the employee dataset containing 474 cases with a dataset that contained only three rows. Each row in this dataset will represent one of the three possible job categories. The dataset for this example, *mean_salary.sav*, is shown below, containing a variable for job category, *jobcat*, and a variable representing the average salary for that job category, *meansal*.



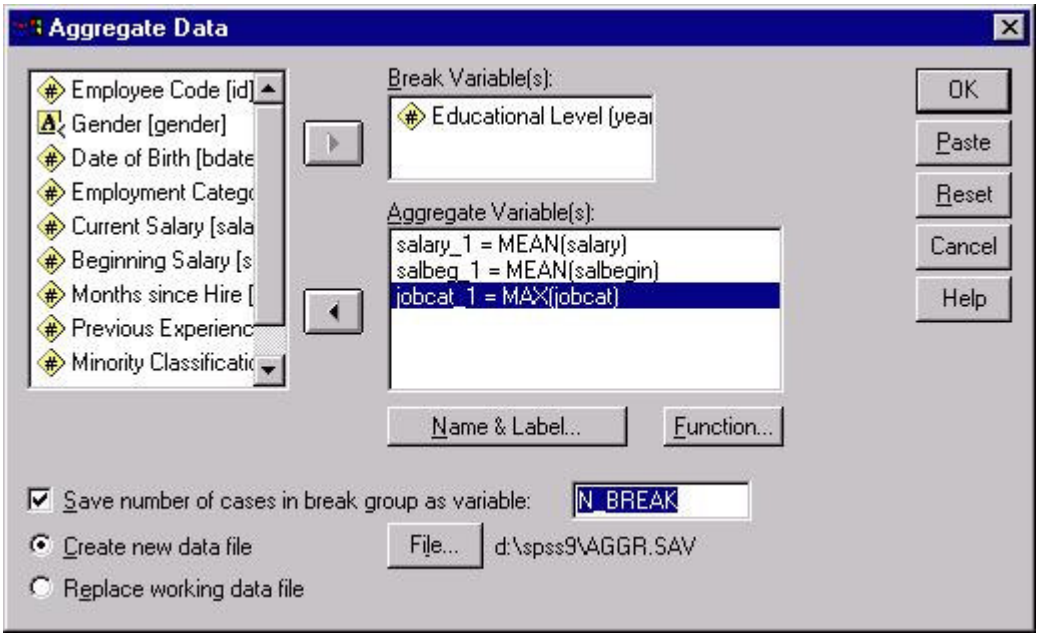
To do the type of file merge described above, select the *External file is keyed table* option in the *Merge* dialog box (this assumes that the *Employee data.sav* file is open in the Data Editor as opposed to the *mean_salary.sav* dataset--in which case you would have selected the *Working data file is keyed table* option). This option will require that you specify a variable on which the two files can be matched. Here, you should match on the basis of *jobcat*, as it is the one variable that the two files have in common. To designate a key variable, highlight the variable you want to use for your match in the *Excluded Variables* box and move it into the *Key Variables* box by clicking on the arrow next to that box. Clicking **OK** at this point will execute the match. You will receive a dialog box reminding you that both files need to be sorted on the key variable. If the files were not sorted on the key variable, *jobcat* in the above example, the file merge will fail. The file resulting from this process will contain the variable *meansal* for all cases, and the value will be the average salary for each person's job category.

Aggregating Files

Aggregating files is another frequently used data manipulation procedure. The *Aggregate* procedure allows you to condense a dataset by collapsing the data on the basis of one or more variables. For example, to investigate the characteristics of people in the company on the basis of the amount of their education, you could collapse all of the variables you want to analyze into rows defined by the number of years of education. To access the dialog boxes for aggregating data, choose the menu item:

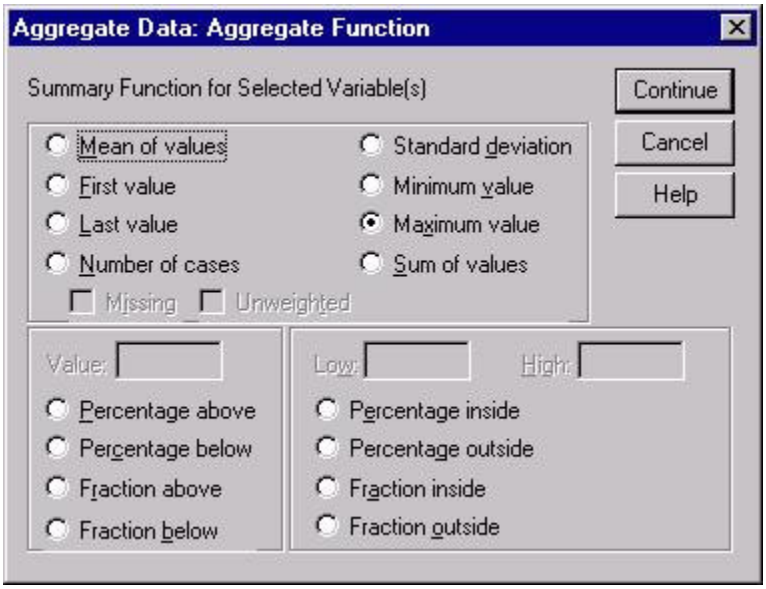
Data
Aggregate...

This will produce the following dialog box:

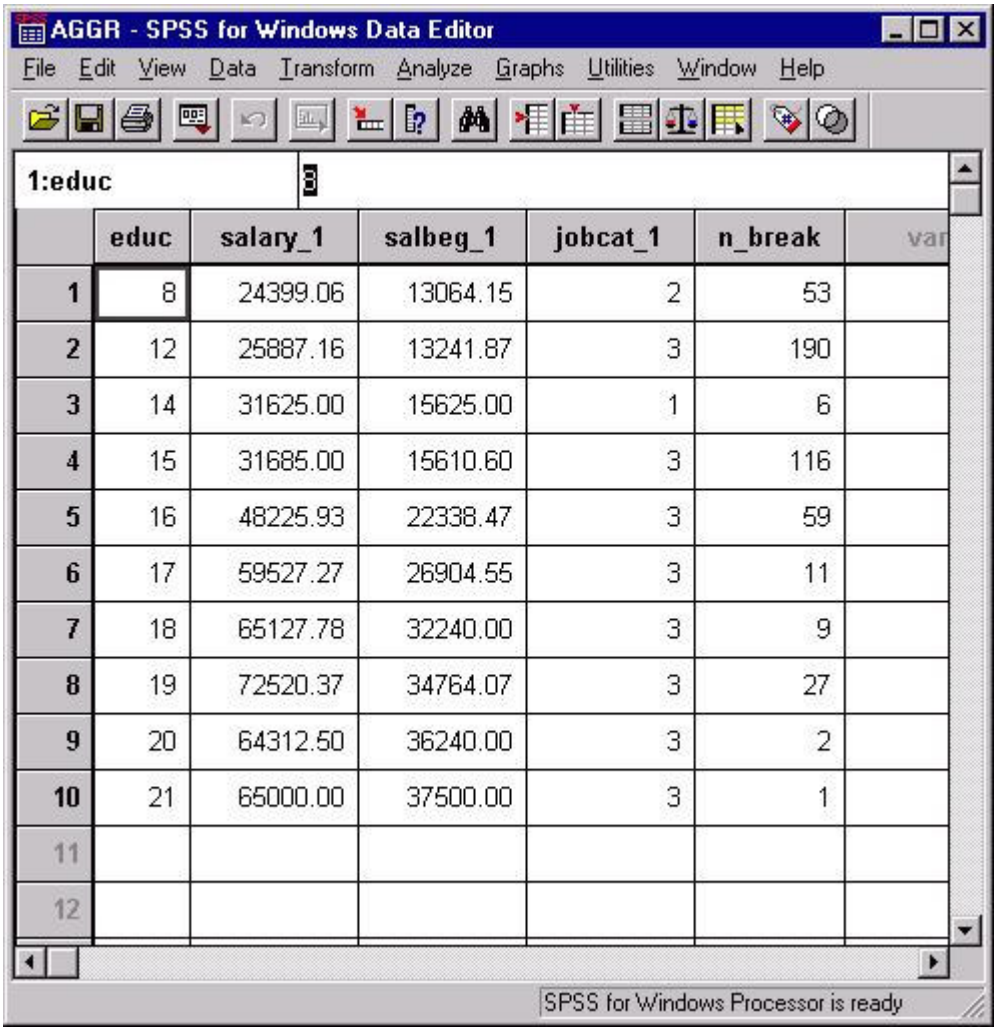


This dialog box shows an example aggregation. The top box, labeled *Break Variable(s)*, contains the variable within which other variables are summarized. Here, the variable contains a value representing a person's years of education. The box below, labeled *Aggregate Variable(s)*, contains the variables that will be collapsed. The variables *salary*, *salbegin*, and *jobcat* will all be summarized for each level of the variable *educ*. Two options available in the bottom left corner of the dialog box should be considered each time you conduct an aggregate function. The first allows you to save the number of cases that were collapsed at each level of the break variable or variables as a new variable. The second option determines whether the aggregated working data file is saved as a new file or replaces the working dataset. If you click on the button labeled *Create new data file*, you can then select the location and name of the file by clicking on **File**. If you click on the button labeled *Replace working data file*, then the resulting data file will be written over the original data file. These changes will be permanent if you save the ensuing aggregated dataset.

There are several options for summarizing variables. In the above example, the aggregated file will contain the means for *salary* and *salbegin* and the maximum value for *jobcat* within each level of *educ*. The summary function is indicated in the *Aggregate Variable(s)* dialog box in front of the original variable name. To the left of this expression is a new variable name that is assigned automatically. To change the summary function, first click on the name of the variable whose function you want to change, then click on the button labeled *Function*, which will produce the following dialog box:



This dialog box allows you to change the summary function for each variable individually. For example, here the function for the variable *jobcat* was changed from the default function, *Mean of values*, to the *Maximum value* function, which returns the maximum value of the variable within each level of the break variable. To select the specific function that you want to use for summarizing a variable, click on the radio button next to it. If you select one of the percentage or fraction functions in the bottom half of the box, you will also have to supply a value or two. When you finish defining your function using those dialog boxes, you can execute the aggregation by clicking on the **OK** button in the main aggregate dialog box. Doing so will write the data shown in the following figure to the file *D:\spss9\AGGR.SAV*.



	educ	salary_1	salbeg_1	jobcat_1	n_break	var
1	8	24399.06	13064.15	2	53	
2	12	25887.16	13241.87	3	190	
3	14	31625.00	15625.00	1	6	
4	15	31685.00	15610.60	3	116	
5	16	48225.93	22338.47	3	59	
6	17	59527.27	26904.55	3	11	
7	18	65127.78	32240.00	3	9	
8	19	72520.37	34764.07	3	27	
9	20	64312.50	36240.00	3	2	
10	21	65000.00	37500.00	3	1	
11						
12						

The first column, *educ*, is the break variable that represents the number of years of education an employee has had. The next two columns represent the mean current salary and mean beginning salary of employees within each educational level. These values were obtained by using the mean function to summarize these variables. The *Maximum value* option was used to produce the value in the fourth column, labeled *jobcat_1*. In this column, the value 3 represents a managerial position. The last column, labeled *n_break*, contains the number of cases that were collapsed into each level of *educ*.

One common use of aggregation is to collapse a dataset into a single row that contains a summary statistic for every variable in the dataset. To do this, you will need a break variable that is identical for every case in the dataset. Thus, when you collapse across this break variable that is the same for all cases, the resulting aggregated dataset will be reduced to a single row. To create this break variable, use the compute procedure. The use of dialog boxes to obtain a new variable is described in Section 3 of the first module of this SPSS tutorial series, "Getting Started." If you use the dialog boxes to obtain a new break variable, you would simply type in the name of a new variable in the *Target Variable* box and type 1 in the *Numeric*

Expression box. Or, you could use the following SPSS syntax to compute a new break variable:

```
COMPUTE breakvar = 1.
EXECUTE.
```

This will produce a new variable, named *breakvar*, containing the value 1 for all cases in the dataset. This variable would then be used as the break variable for the *Aggregate* procedure.

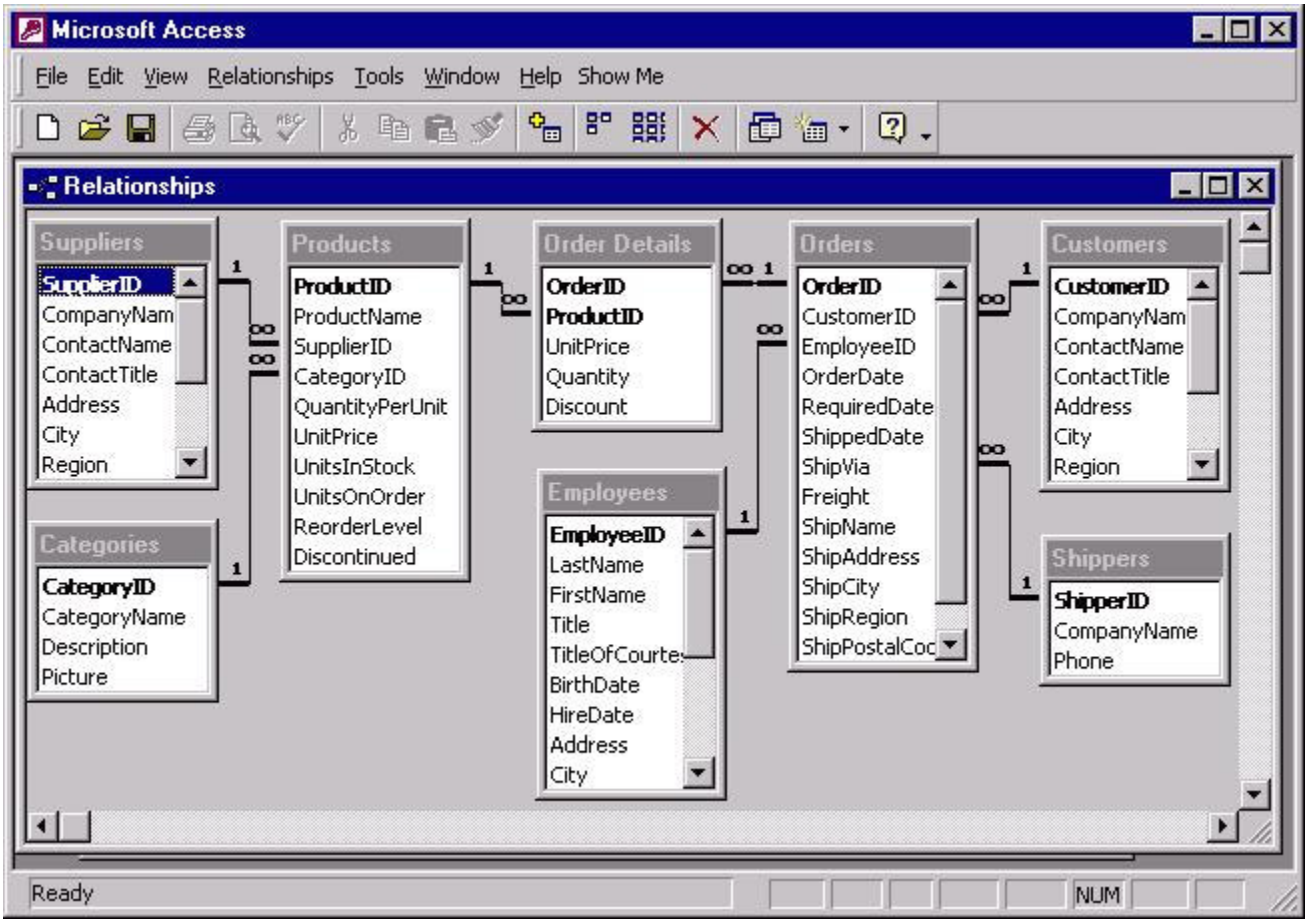
Database Capture

SPSS can be used to import data stored in databases into the Data Editor. The SPSS *Data Capture Wizard* can import variables that are stored in different tables in a database. To illustrate the use of the Data Capture Wizard, we will leave the *Employee data.sav* dataset used so far and introduce a database that is a sample database included with Microsoft Access: *Northwind.mdb*. It includes variables containing information about a company's inventory, products, suppliers, etc.

Before learning the process for importing a database into SPSS, it is important to understand some database terms and their corresponding terms in SPSS datasets:

- A *relational database* consists of several individual datasets called *tables.* Relational databases are called that because they consist of several related databases.
- Tables are roughly equivalent to individual datasets: they contain several variables, organized in columns.
- Variables are called *fields* in relational databases.

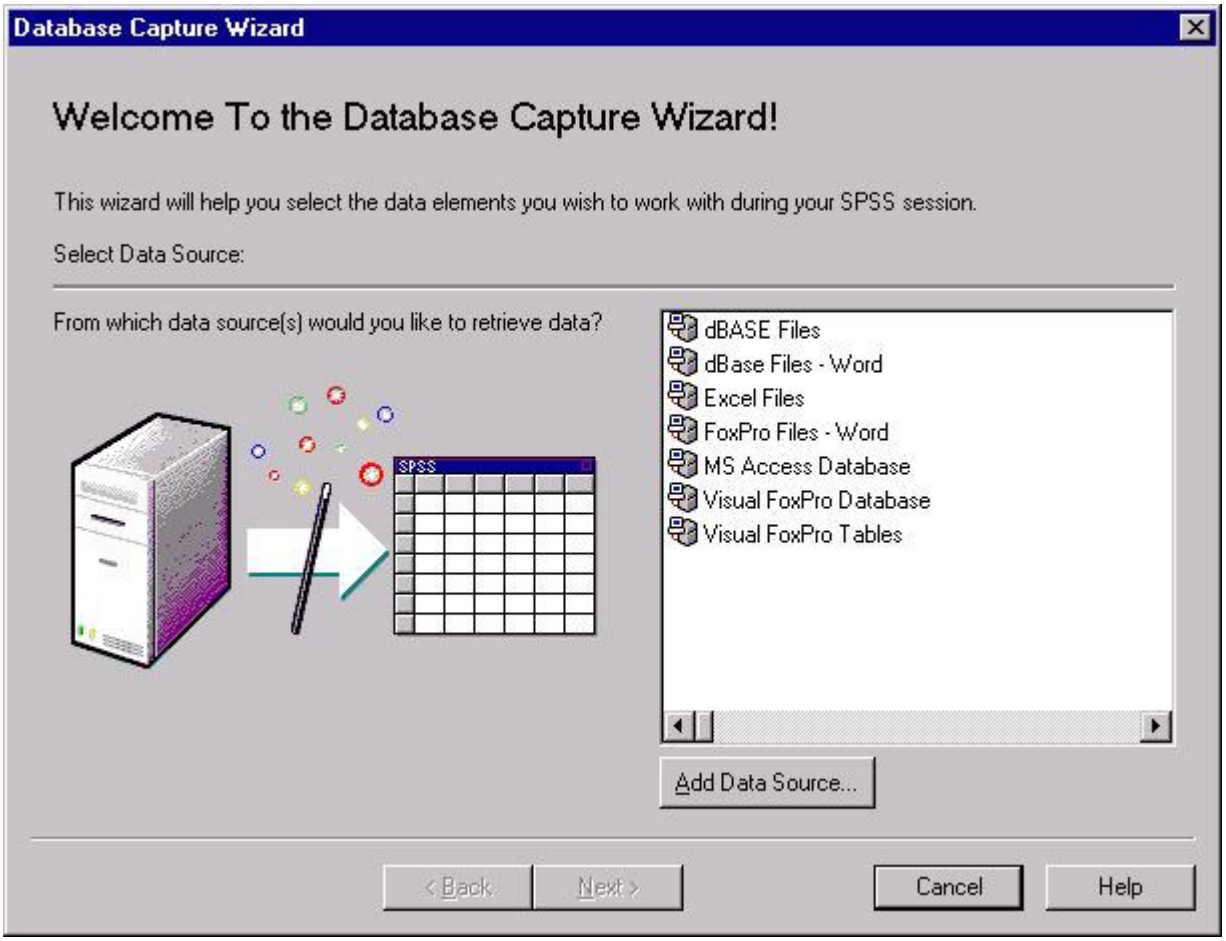
Below is a figure containing a graphical display of the Microsoft Access relational database, *Northwind.mdb*, that illustrates the above terms. Each box represents an individual table in the database. Within each of the tables is a list of the fields contained within each table. The relations in the database below are illustrated by the lines connecting fields having the same name in different tables.



The Database Capture Wizard can be accessed from the SPSS *File* menu:

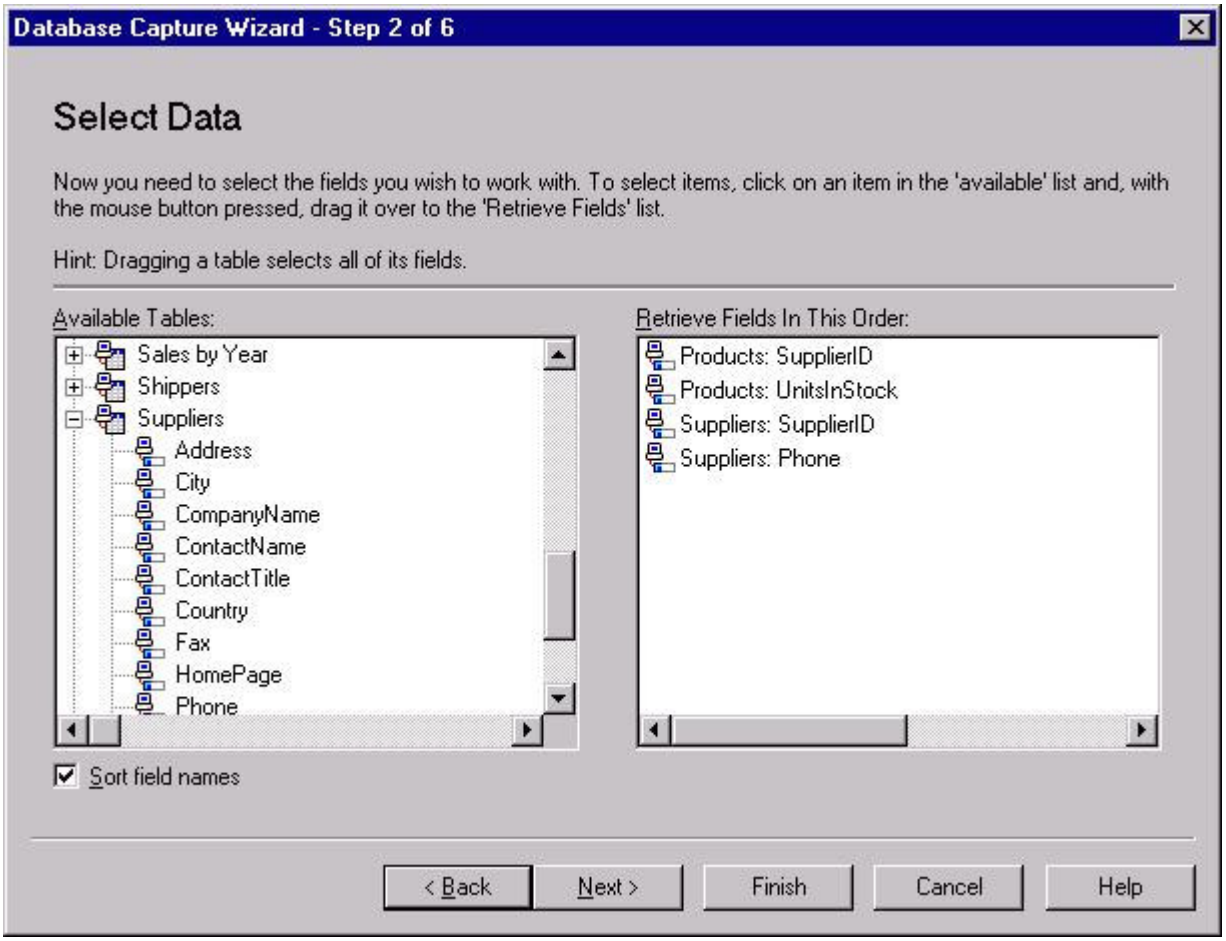
File
Database Capture
New Query...

Selecting this menu item will produce the following dialog box:

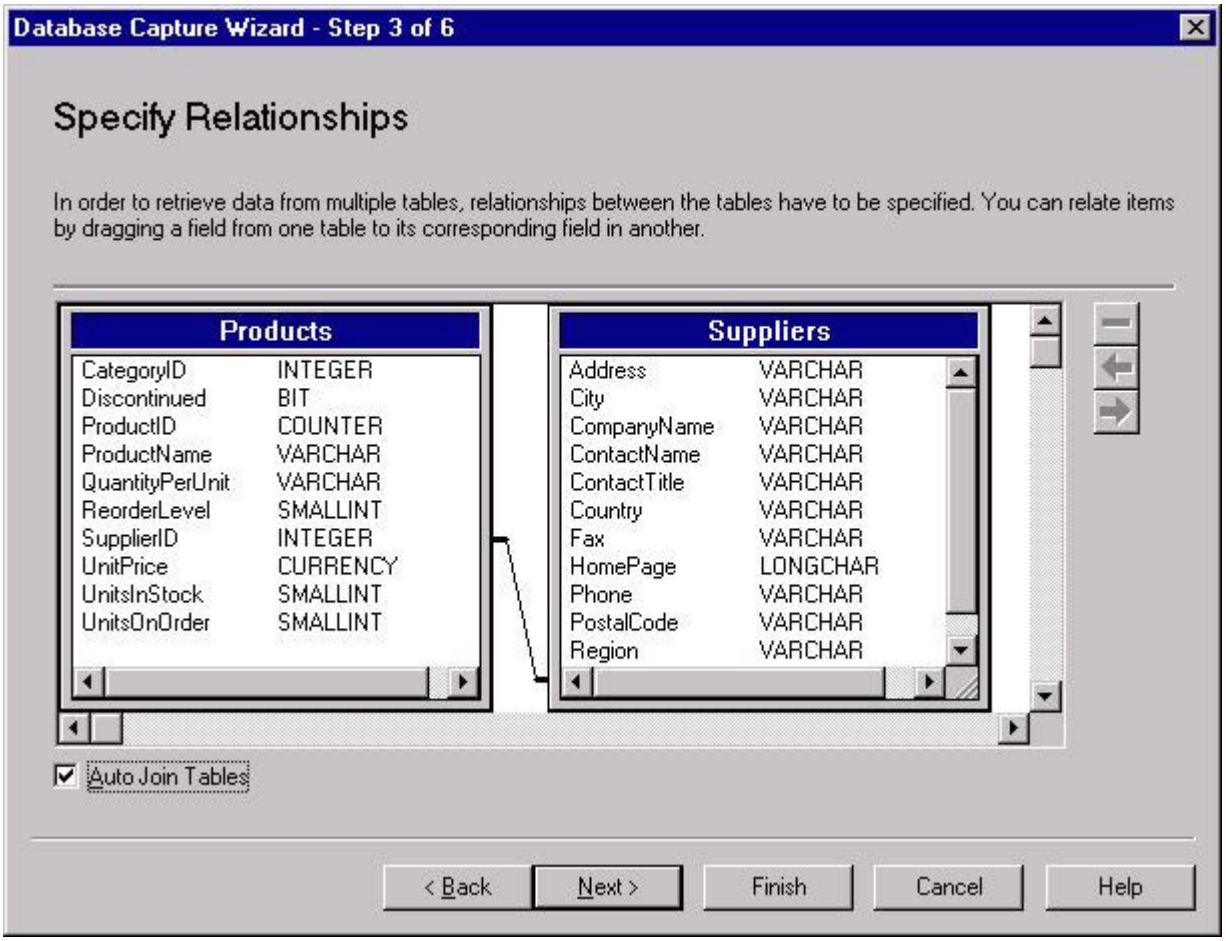


The Wizard consists of six dialog boxes, each of which performs a single function. While there are numerous features of the Database Capture Wizard, only the basics of importing data into a SPSS dataset are discussed here. Some useful features of the Wizard not covered here include: selection of a subset of variables that contain particular values, renaming variables, and saving syntax for the import process.

The first dialog box, shown above, will prompt you to select the type of file that you want to import into SPSS. Here, select the *MS Access database* option. After selecting the type of database and clicking **Next**, you will be provided with a dialog box from which you can select a database located on disk. The second dialog box in the series, shown below, will appear automatically once you have selected the database you want to import.



This dialog box is critical for selecting the information you want to include in the SPSS dataset you are creating. The dialog box will display a list of all of the tables contained within a relational database. You can click the plus sign next to the name of an individual table to expand it so that you have a list of the fields contained within that table. To create a new dataset, you can select either entire tables or you can get a list of individual fields in a table by clicking on the plus sign next to the table name and then selecting fields from within tables. To select items, double-click on them or use your mouse to drag them from the left window to the right window. In the example above, the plus sign next to the *Suppliers* table has been selected. Four fields have been selected, two of which are from the *Products* table and two from the *Suppliers* table. These include the *SupplierID* and *UnitsInStock* fields from the *Products* table and *SupplierID* and *Phone* fields from the *Suppliers* table. The SPSS dataset being created in this example will consist of these four fields. Clicking **Next** advances you to the next dialog box, shown below:



This dialog box illustrates the relationship between tables from which you are importing data. The type of line that connects the tables defines their relationship. There are three possibilities: (1) rows in the tables are included only if a common value appears in both fields (e.g., rows are included only if a particular *SupplierID* exists in both tables), (2) All rows from the table on the left are included and rows on the right table are included only if the value of an identifier variable exists in the left table (e.g., all rows in the *Products* table are included, but only rows from the *Suppliers* table that contain a *SupplierID* value that exists in the *Products* table are included in the SPSS dataset), and (3) All rows from the table on the right are included, and rows on the left table are included only if the value of a identifier variable exists in the left table (e.g., all rows in the *Suppliers* table are included, but only rows from the *Products* table that contain a *SupplierID* value that exists in the *Suppliers* table are included in the SPSS dataset). Double-clicking on the line connecting the tables will produce a dialog box from which you can select one of the above relationships between each table in your database. Clicking **Finish** here or at any other point will complete the process and import the data as you've defined it to the SPSS Data Editor.

The final three steps of the Database Capture Wizard are optional and are beyond the scope of this tutorial. The fourth step allows you to select only records with certain

values. The fifth step lets you rename your variables. The sixth step allows you to print the SPSS syntax to the Syntax Editor, which will replicate the process which you accomplished with the Database Capture Wizard.

Section 8: Advanced Topics

Syntax

SPSS syntax was first mentioned in the first module of this tutorial series, "Getting Started," in the discussion of the Syntax Editor in Section II. This section will introduce some generalizations about the use of SPSS syntax and resources for creating SPSS syntax, including options for executing syntax, important syntactic characters, and two modes for running syntax in SPSS: batch and interactive. With these core concepts, you should be able to navigate SPSS syntax and generate your own syntax.

The SPSS Syntax Editor section in the "Getting Started" module illustrates the use of the **Paste** button, which is available in most dialog boxes. The **Paste** button can generate syntax which can then be used to send commands to the SPSS processor.

The example from the "Getting Started document illustrates some features of SPSS syntax that generalize to virtually all syntax. The example syntax can be used to generate descriptive statistics:

```
DESCRIPTIVES
  VARIABLES=educ
  /STATISTICS=MEAN STDDEV MIN MAX .
```

First, know the difference between commands and subcommands. The command name is always on the first line. If it is a procedure that generates statistical output, then the names of variables to be included in the analysis will immediately follow. Here, the command name is **DESCRIPTIVES** followed by the subcommand, **VARIABLES**, on the following line. After the **VARIABLES** subcommand, the variable name *educ* appears, which tells the SPSS processor the variable on which to calculate descriptive statistics. This could also be a list of variables if you want to generate statistics on more than one variable at a time. The **VARIABLES** subcommand is a typical subcommand for specifying variables; however, many procedures have special syntax that is used to specify different types of variables, such as fixed factors and covariates.

The next line contains the subcommand, **/STATISTICS**. It shows the slash character that is common to all subcommands other than those which specify variable information. The **/STATISTICS** subcommand also illustrates the use of options in a subcommand. Following the equals sign, there is a list of four statistics, **MEAN**

STDDEV MIN MAX, which are four of several possible descriptive statistics that can be specified in this command. Notice the period following the last line of the above code. It indicates the end of a command and is always placed after the last subcommand for a particular procedure.

Another important distinction is that between *batch* and *interactive* modes. In interactive mode, you can maintain interaction with the SPSS processor while commands are being processed. In batch mode, the SPSS processor cannot be interrupted once commands have been submitted. If it *is* interrupted by a user or even by an error in the syntax, then it stops processing commands. Interactive mode is more frequently used and is the mode in which commands executed from dialog boxes or syntax windows are processed. Batch mode is the default mode when using the SPSS Production Facility and can be called from the Syntax Editor by using the **INCLUDE** command.

There are a few differences in the way commands are processed in batch and interactive modes, and there are some critical differences in their syntax. Batch mode has some features that are not in interactive mode. Interactive mode will print the line numbers of code in the output and will print the level of control when you are using **DO** loops or other redundant processes. Key differences in syntax are covered here; for details about other differences, see a consultant.

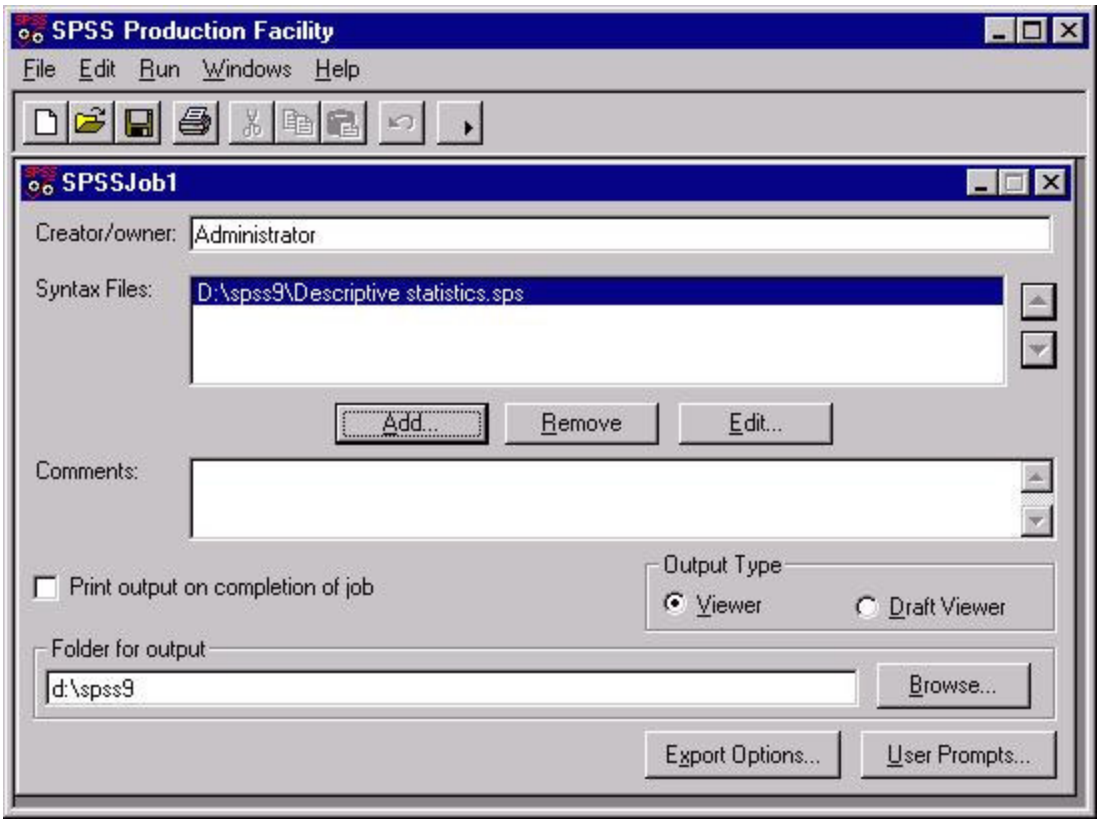
Generally, batch mode has stricter conventions about the format of syntax than does interactive mode. There are three key difference between the modes:

- Commands must start in the first column in batch mode.
- Continuation lines must be indented at least one column in batch mode.
- Periods at the ends of commands are optional in batch mode.

Our above syntax example conforms to the conventions of batch mode, as does all syntax generated using the **Paste** button: the command name, **DESCRIPTIVES**, begins in the first column, and all continuation lines are indented at least one space. The period at the end of the commands is not necessary but will not interrupt processing. In interactive mode, you could put all of the commands on a single line and submit them to get the same results.

The Production Facility

The *Production Facility* is an application that stands apart from the standard SPSS application. It can be accessed from the Window's *Start* menu. The Production Facility's primary purpose is to process syntax files in batch mode. You can process as many syntax files as you like by using the **Add** button to add syntax files to the list of files in the *Syntax Files* box. Clicking on the **Add** button, will present you with a dialog box from which you can select a syntax file. This process has been done for the file *Descriptive statistics.sps* in the example below. Repeat this process until you have added all of the files that you want to include in the production job.



The production job can be saved to include all of the options and a record of the syntax files included in that job. Selecting *Save* from the *File* menu will save the entire dialog box as it appears above. Doing so will include options such as the *Print output on completion of job* which causes your entire output file to be printed when the job has been completed. Saving a production job is useful for situations where you want to repeat an analysis on a regular basis. If you need to make changes in a syntax file in a saved production job, click on the **Edit** button, which will open the presently highlighted syntax file in a text editor.

Scripts using Visual Basic

The SPSS scripting facility enables you to add numerous custom features to SPSS. It uses the Visual Basic programming language to operate on most any feature of the SPSS environment. You can use any commands available in Visual Basic in addition to the Visual Basic commands that are unique to SPSS's scripting facility. While you may not be proficient in Visual Basic or interested in learning it, there are a number of scripts available that you may find useful. For example, one sample script will change the text, "Total," in a pivot table so that it is blue and in a bold font. This script contains comments that will indicate how to change the script so that it could find text other than "Total" and change it to colors other than blue.

There a few key concepts with which you should be familiar if you intend to use scripts in SPSS. Visual Basic operates on *objects*, which can refer to virtually any

definable entity in the SPSS environment. For example, everything from the SPSS application to a cell in a pivot table are objects in the SPSS environment. For a useful overview of such objects, select the *Objects* menu item from the *Help* menu:

Help
Objects

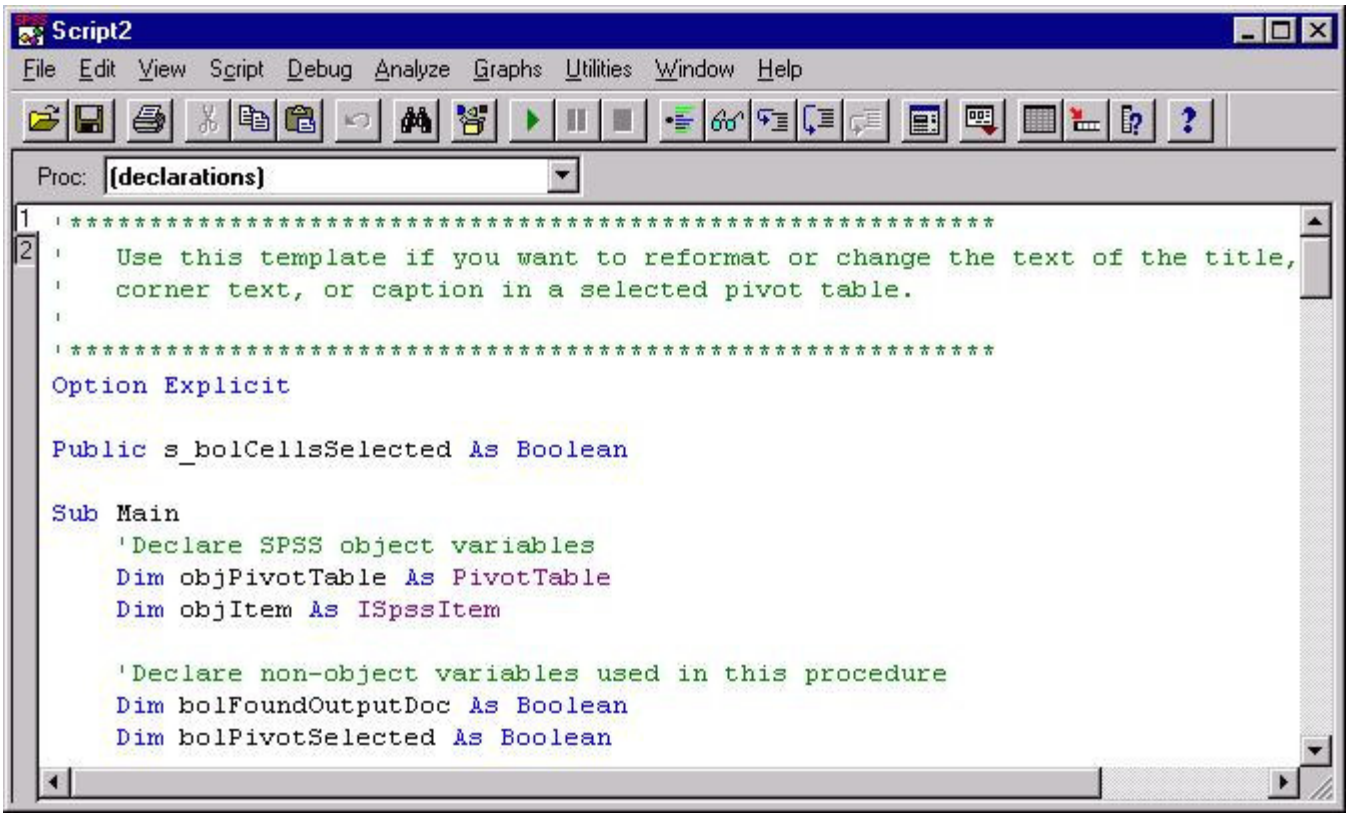
Two other key concepts in the Visual Basic environment are *methods* and *properties.* Methods refer to operations that can be performed on an object. For example, the **ResizeColumn** method will change the width of a column in a pivot table. A property refers to features of an object. For example, cells in a pivot table have properties such as height, width, and color, which can be changed with various methods. One such property is **TextFontAt**, which will return the type of font in a user-defined cell.

One final concept that is critical to using scripts in SPSS is the concept of *active* objects. At any given time, only one object is active in the SPSS environment. For a method or property command to be performed on an object, that object must be active. There are two ways to activate an object. The first is to use the Visual Basic language. This topic is not covered here. The second way is by clicking on that object. Thus, when you move your Output Viewer to the foreground, you have activated that object. Or when you highlight a single cell in a pivot table, that object is activated. This feature can be especially useful if you only want to run a brief program on a specific object.

To run a script, open the SPSS scripting facility from the *File* menu:

File
New
Script

This will produce a dialog box containing several script files that are included with SPSS. To use one of these, select it from the dialog box. To create your own script, select the **Cancel** button to move into the SPSS scripting facility. Here is a dialog box containing the script, *Reformat misc pivot.sbs*:



If you are using one of the scripts provided with SPSS, you will often have several choices for easily modifying the programs. Usually, these involve removing a comment from the text. Comments are lines with green text that begin with the ' character. By removing the ' character, the line becomes a line of code rather than a comment. One example of a commented line of code in the above program is a line that allows you to change the title of the active pivot table. If you scroll down in the script shown above, you will eventually see the following commented lines:

'Remove the ' before the next line to change the Title of the selected PivotTable
'objPivotTable.TitleText = "PUT YOUR TEXT HERE"

The first line is a comment that instructs you to remove the comment character from the following line if you want to alter a pivot table's title. The second line will actually alter the title. Thus, to replace the default title for a descriptive statistics table from *Descriptive Statistics* to *New Title*, alter the above code to change the title by removing the ' character and replacing the text, "PUT YOUR TEXT HERE", with "New Title". This altered code would look like this:

'Remove the ' before the next line to change the Title of the selected PivotTable
objPivotTable.TitleText = "New Title"

To run this code, first click the pivot table that you want to alter to make sure that it is active (there will be a box around it when it is active). Next, click on the icon on your toolbar with the green arrow on it to run the program. These steps will alter the title of your pivot table.

One other useful source of SPSS script code are the examples available through the object diagram. Open the object diagram from the *Help* menu as described above. Next, click on the object on which you want to perform an operation. This will produce a text box that contains a description of the object. Click the **Methods** or **Properties** buttons to get a list of the methods and properties that can be performed on the selected object. Then select one of the methods or properties from the list by clicking on it, and click the **Display** button to get a description of the command you have selected. This will provide you with a syntax diagram. To get example syntax, click the **Example** button.

Macros

SPSS provides the capability to create macros that contain text and SPSS command syntax. Macros can perform the same command repeatedly, combine several procedures, or calculate user-defined statistics. Any functions you can perform using SPSS syntax can be incorporated into a macro. When a macro is called, it is expanded so that any text including commands will be used in its place.

There are two commands used to create a macro: **DEFINE** and **!ENDDEFINE**. To begin the definition of a macro, type the **DEFINE** command in the Syntax Editor followed immediately by the user-defined macro name. Be careful in choosing this name as it will activate the macro whenever it is used. It should be a name that does not appear elsewhere in your syntax. Following the user-defined macro name are parentheses. These required parentheses may or may not contain parameters. Next, you can type syntax or any other text that will be called each time your macro is called. After you have entered all of the text or syntax that you want to include in your macro, type **!ENDDEFINE** to end the definition. The following syntax illustrates the use of a macro that contains a list of variables names. The macro is then used in the **DESCRIPTIVES** procedure.

```
DEFINE varlist ()
     educ jobcat salary salbegin jobtime
!ENDDEFINE.

DESCRIPTIVES
  VARIABLES=varlist .
```

This syntax defines a macro named *varlist*. The text to be called each time *varlist* appears later in the syntax is a list of variable names: *educ jobcat salary salbegin jobtime*. For example, *varlist* appears in the **VARIABLES** subcommand of the **DESCRIPTIVES** command in the above syntax, at which time it is replaced with the

text contained in the macro definition and descriptive statistics are calculated for all five of those variables. SPSS macros can also include SPSS syntax, as illustrated by the following macro:

```
    DEFINE varlist ()

educ jobcat salary salbegin jobtime
    !ENDDEFINE.

    DEFINE descrip ()
    DESCRIPTIVES
        VARIABLES=varlist
    !ENDDEFINE.

    DESCRIP.
```

The above macro is identical to the prior example except that the **DESCRIPTIVES** command has been included in a new macro called *descrip*, which is then called after the macro definition. The following processes take place in this case: First, the *varlist* macro is defined so that each time *varlist* appears, it is replaced by a variable list. Second, the *descrip* macro is defined to execute the **DESCRIPTIVES** command for the variables contained in the *varlist* macro. Last, the *descrip* macro is called in the line containing the macro's name, which then executes the **DESCRIPTIVES** command for the variables defined in the *varlist* command.

Thus far, none of our macro examples have contained *arguments*, which are additional terms such as variable names that you may want to call with a macro. The example below illustrates the use of a macro in which one macro is a variable list that is inserted in two statistical procedures.

```
    DEFINE varlist ()
        educ jobcat salary salbegin jobtime
    !ENDDEFINE.

    DEFINE stats (arg = !CHAREND ('/')).
    DESCRIPTIVES
        VARIABLES=!arg .
    FREQUENCIES
     nbsp;VARIABLES=!arg
      /STATISTICS=MIN MAX.
    !ENDDEFINE.

    STATS arg = varlist /.
```

Here, the first step again begins with the varlist macro which defines variables. The second step defines a second macro with the name *stats*. It contains a user-defined

argument, *arg,* which will be replaced with variable names when the macro is called. The argument is followed by the SPSS keyword, **!CHAREND**, which defines the character that signifies the end of the argument list. This is one of several possible keywords (See the *Syntax Guide* in the *Help* menu for more options). The third step is to execute a series of statistics: **DESCRIPTIVES** and **FREQUENCIES**. Notice in the **VARIABLES** subcommand that *!arg* is in the place where a variable list normally is. After the **!ENDDEFINE** command, the *stats* macro is called with the argument varlist followed by the '/.' The argument varlist replaces the arg in the stats macro with the variable list which it defines and the '/' signifies the end of the arguments. Thus, the output for the above example will include descriptive statistics and frequency tables for the variables, *educ*, *jobcat*, *salary*, *salbegin*, and *jobtime.*
