APPENDIX C

HOW TO USE A STATISTICAL PACKAGE: SPSS

With the assistance of Lisa M. Gilman and with contributions by Joan Saxton Weber

Computers and statistical software such as the Statistical Package for the Social Sciences (SPSS) make complex statistical computations simple and fast. SPSS is one of the most popular comprehensive statistical software packages used in the social sciences. You can use it to calculate a great many statistics and to create charts and tables for presentations with just a few clicks of the mouse.

This appendix provides a basic introduction to SPSS. Even if you are unfamiliar with computers or apprehensive about statistics, you will find that SPSS for Windows is very user friendly. Please bear in mind that all of the examples use version 13 or 14 of SPSS, with data from the 2004 version of the General Social Survey (GSS); if you are using a different SPSS version or another year of the GSS, you may find some slight differences in procedures or answers.

BASIC PROCEDURES

To start SPSS for Windows, click or double-click on the SPSS icon using the left mouse button. If you are unable to locate the SPSS icon, click the Start button, then click on Programs, and then click on SPSS.

A new screen will open with a "What would you like to do?" box superimposed on the screen. For now, click on the Cancel button in the "What would you like to do?" box, to get it out of the way. You are now looking at the SPSS Data Editor window. This screen is where data to be analyzed are entered or data files that have already been created (datasets) are loaded.

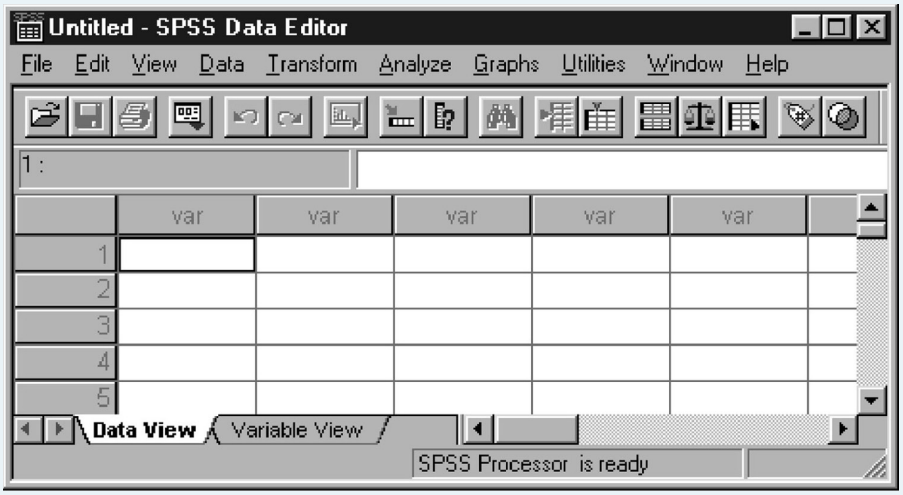
To access the data for this appendix, click:

File—Open—Data (or just click on the open folder icon).

Select (highlight) the GSS2004mini file. If the GSS2004mini file hasn't been transferred to your hard disk, you can find it on the CD-ROM that came with this text.

The Data View in the Data Editor consists of columns and rows, with each column representing a different variable—their names are at the top—and each row representing one case or "observation" (Exhibit C.1). If you are using the student version of SPSS, you are limited to no more than 50 variables and 1,500 observations. (The GSS2004mini dataset has 1,404 cases, or observations, and 48 variables. You can confirm this by moving the scroll bars on the bottom and right of the Data View). There are no variable/observation limits in the commercial version.

Exhibit C.1 SPSS Data Editor



Another screen you should be familiar with is the Variable View screen. To access the variable view screen, click on the Variable View tab at the bottom left of the Data Editor (not available in early versions of SPSS for Windows). The Variable View screen contains a list of all the variables included in the dataset and their characteristics. Each row corresponds to a single variable. Variable characteristics are indicated at the top of each column. When you are in the Variable View mode, you can create, edit, or view variable information. Now click on the Data View tab to return to the data view mode.

Looking at the Data

There are several ways to learn more about the variables and numbers that you see in the Data View screen. The names of the variables are listed across the top of the screen. You can also see the list of variables in the GSS2004mini file by clicking on the variable list icon at the top of the Data Editor screen. When you click on any variable in the resulting window, you can see the details that show how it was coded. But the easiest way to learn about the variables in the dataset is to switch from the Data View to the Variable View screen by clicking on the tab in the lower left.

You can tell what some variables are just by their SPSS variable names, like SEX. However, many variables are responses to specific questions and we can't tell just what they are on the basis of the shorthand variable name. Instead, we can inspect the variable label in the "Label" column corresponding to that variable in the Variable View screen (you may need to widen the Label column by dragging the separator line at the top of the column with your mouse).

SPSS understands numbers much better than text, so all the answers to each GSS question have been coded. This means that, for every variable, each response category has been assigned a numerical value. For some variables, such as AGE and EDUC, the number you see is simply the number of years. For other variables, such as SEX and RACE, each number corresponds to a particular response category. For example, go to the Variable View screen and click on the "Values" cell in the row corresponding to the variable SEX. Now click on the gray box in that cell and you can see the labels for specific codes, which show that men are coded "1" and women are coded "2."

Values that stand in for missing data are indicated in the "Missing" column. Special labels identifying the reason for missing data, such as DK for Don't Know or NA for No Answer, also appear in the Value Labels column. For ABRAPE, the variables dialog box shows that 8 corresponds with DK (Don't Know) and 9 corresponds with NA (Not Available). Therefore, for the variable ABRAPE, both 8 and 9 are missing values.

For most statistical calculations, SPSS ignores missing values and calculates the statistic based on the responses of the respondents who answered the question. In general, however, you should make sure missing values are not inadvertently included in your analysis. Common values used to identify missing data are –1, 0, 8, and 9. If the variable is two digits wide, such as AGE, the missing values are usually 97, 98, and/or 99.

What if you have data for new cases to enter, or you have new variables that you need to define? You can simply enter the data in the Data View and enter the labels and missing value codes in the Variable View. Click on File—Save after you are sure that everything is correct.

Before you proceed to the next step, make one minor adjustment in the display options. Click:

Edit—Options

The Options dialog box will open. You will see a series of "tabs" along the top of the window. You should be looking at the General options (if not, click on the General tab). On the upper left side of the General options screen, you will see the Variable Lists options.

Under Variable Lists, click the radio button next to Display Names (to mark it with a small dot). This will display the convenient short variable names rather than the descriptive variable label in some of the selection boxes.

UNIVARIATE STATISTICS

Now that you have an open data file and are somewhat familiar with the data itself, you can begin exploring the data statistically (statistics are discussed in Chapter 14).

Frequencies

A *frequency distribution* is a table that displays how many and what percentage of observations fall into given categories for a variable of interest. In other words, a frequency distribution tells you how many people said yes, how many people said no, and so on. The purpose of obtaining a frequency distribution is to summarize the data so that they are easy to understand. The variable ABRAPE (pregnant as a result of rape) will serve as a good example for a frequency distribution. Click:

Analyze—Descriptive Statistics—Frequencies

The frequencies dialog box will open, as shown in Exhibit C.2. Choose ABRAPE by clicking on it from the list on the left. Click the arrow in the center of the dialog box to move the ABRAPE to the "variable(s)" box on the right. Then click OK.

After SPSS has processed the command, the results appear in a new window titled "Output1–SPSS Viewer." The Output window has two panes. The left pane is referred to as the "output navigator." The output navigator contains an outline of everything you ask SPSS to do from the beginning of the session. It allows you to easily refer back to any given table or graph. To go to a specific table, you find it in the output navigator, click on the table you want, and it will be displayed in the right pane. As indicated, the right pane contains the actual output for the commands you gave SPSS. It is often referred to as the "output" or the "results."

In the example above, SPSS created a frequency distribution for the variable ABRAPE. SPSS produces two boxes (Exhibit C.3). The first box just displays the valid number of cases (*N*) and the total number of missing cases for the variable of interest. The second box contains

Exhibit C.2 The Frequencies Dialogue Box

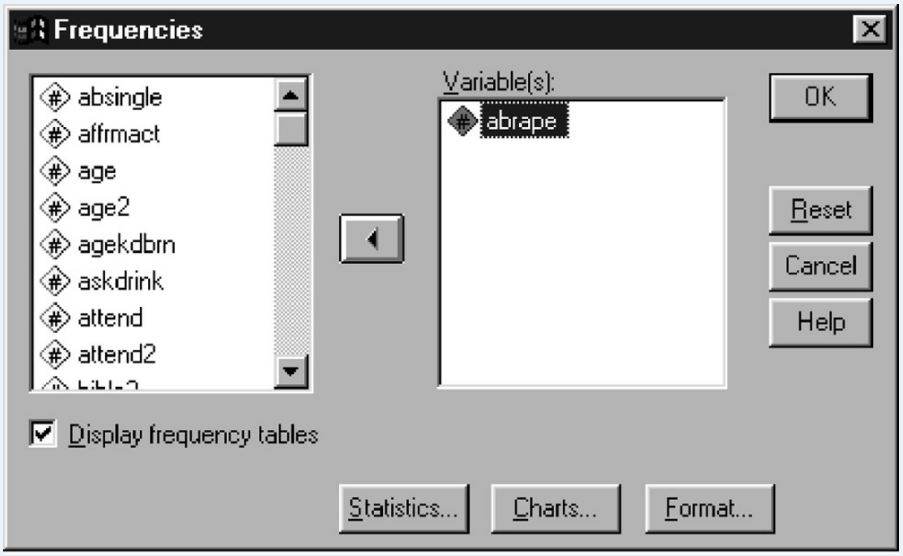
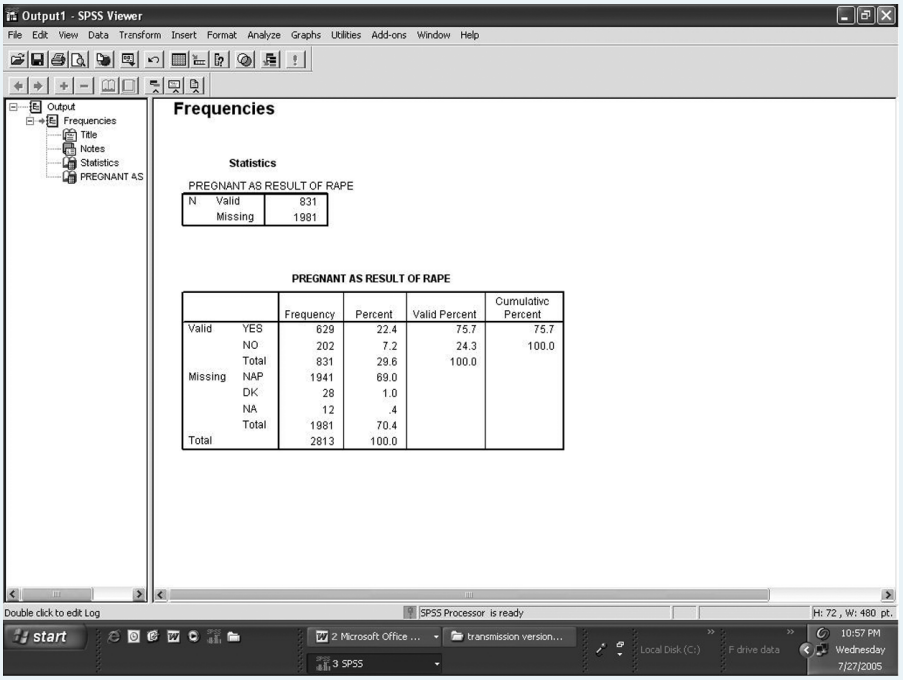


Exhibit C.3 SPSS Frequencies Output



the actual frequency distribution. If the entire table is not visible, use the scroll bar on the right to move down the output window. The distribution shows that about 76% of respondents believe abortion is permissible when a pregnancy is the result of rape, whereas 24% of respondents do not believe that abortion is permissible in this circumstance.

Descriptive Statistics

If you need to calculate the mean, median, or standard deviation of an interval or ratio level variable, you can do so simply with the Descriptive Statistics procedure. Click:

Analyze—Descriptive Statistics—Descriptives

and click the EDUC variable over into the Variables window before clicking OK.

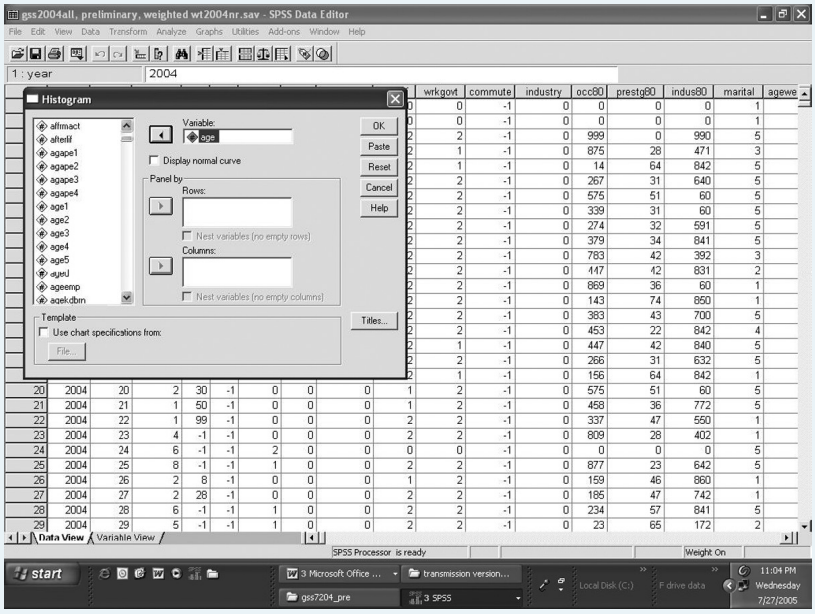
Graphing

Another method of describing the distribution of a single variable is through the use of a histogram. A histogram is useful for graphically displaying the distribution of a given variable. Histograms are useful when the variable(s) you are interested in are continuous in nature (interval/ratio) and have a large number of categories. For example, if you want to look at the number of respondents within certain age groups, it is much simpler to visually display the data than to look at counts. Let's look at how age is distributed in the GSS200mini data included with this book. Click:

Graphs—Histogram

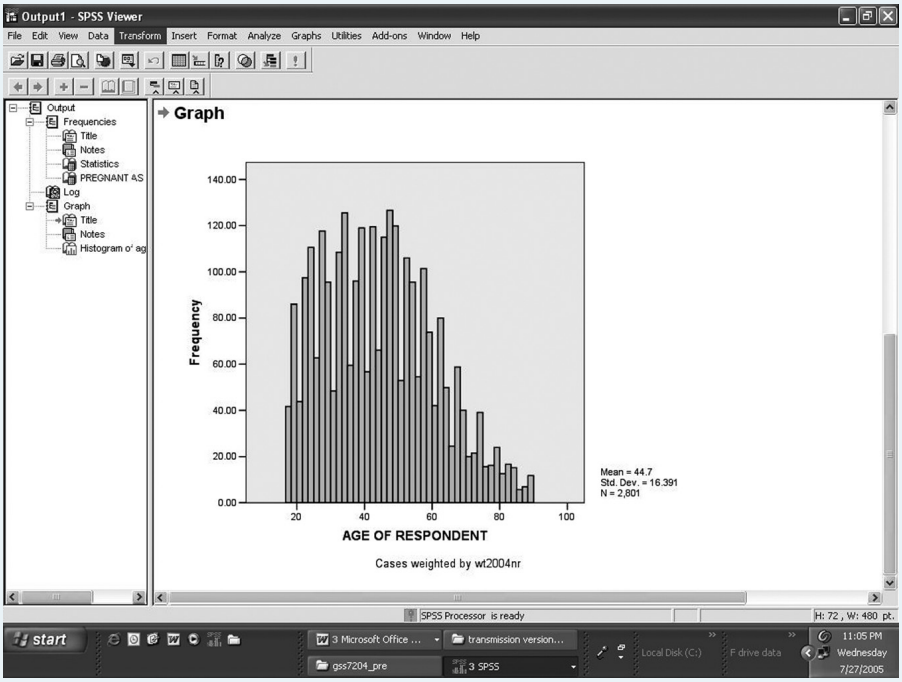
Click on the AGE variable in the variable list and move it to the Variable box and click OK (Exhibit C.4).

Exhibit C.4 Selecting a Variable for a Histogram



When you look at the histogram (Exhibit C.5), you can see that the age distribution has a slight positive skew. The majority of respondents are under the age of 60. The statistics to the right of the histogram indicate that the mean age of the respondents for this sample is approximately 45 years old with a standard deviation of about 16 years.

Exhibit C.5 Histogram of Age



A similar graphic device for nominal or ordinal variables is the bar chart. You can create a bar chart for Race by clicking:

Graphs—Bar

and then highlighting the options Simple and "Summaries for groups of cases" before clicking Define. Now scroll through the variable list to find race, click this into the Category Axis box, and select the Bars Represent . . . % of cases option. Now, click OK. And there you have it.

RECODING VARIABLES

In some instances, variables with many categories may be confusing to use and/or interpret. At other times, your research question may make it necessary to limit your analysis to certain categories. The variable EDUC (highest year of education completed) is a good example. You seldom want to know if people with 11 years of education are different from people who have 10 years of education. It might be better to examine differences of opinion among high school graduates versus college graduates.

To create these groups, it will be necessary to recode the EDUC variable. When recoding a variable, be sure to check the missing values and to look at the minimum and maximum values for valid responses with a frequency distribution.

Let's recode EDUC into the following four categories:

Those who have 0 to 11 years of education are put into category 1.

Those who have exactly 12 years of education are put into category 2.

Those who have 13 to 15 years of education are put into category 3.

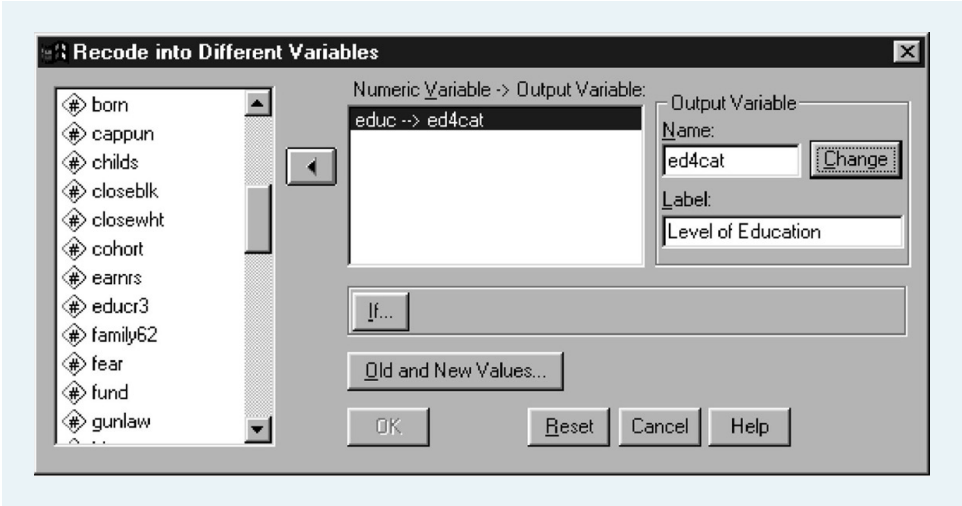
Those who have 16 or more years of education are put into category 4.

In this example, you will collapse the original 21 categories (years of education from 0 to 20) into the 4 categories indicated above. To do this, you will create a new variable called ED4CAT. (You should always recode variables into "different variables" to retain the original variable.) Click:

Transform Recode—Into Different Variables

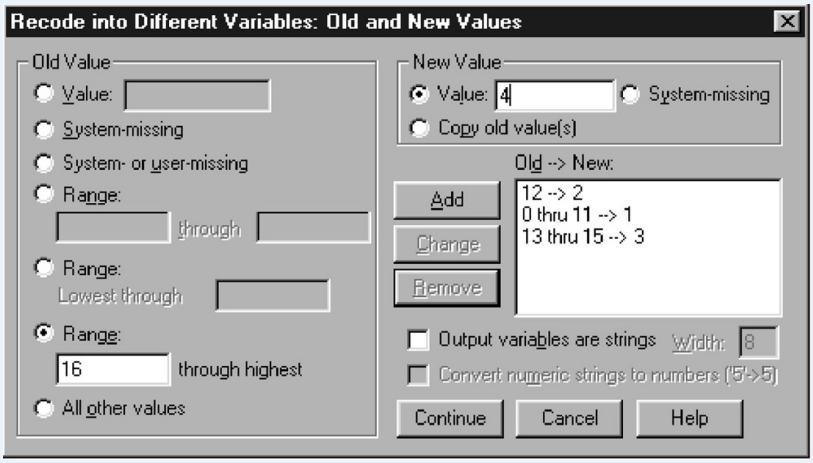
This takes you to the first recode dialog box, in which you tell SPSS the name of the old variable you are recoding from (EDUC) and the name of your new recoded variable (ED4CAT) (see Exhibit C.6). Using the scroll bar, find EDUC in the variable list on the left and move it to the "Numeric Variable—Output Variable" box. Under the "Output Variable" section, type the name of the new variable (ED4CAT) and a variable label, such as "Recoded Years of Education," and then click Change.

Exhibit C.6 Selecting Variable for Recoding



After doing this, click the "Old and New Values . . ." button. A new dialog box will open (Exhibit C.7). This second box is where you tell SPSS into what categories you want to recode the original data. To do this, you identify the original ("old") values of the variable on the left side of the box, and you put the values for your new variable on the right side of the box. When you recode a single value, use the "Value" box under the "Old Value" heading; when you recode a range of values to a single new value, specify the range under one of the "Range" options under "Old Value." Click the Add button to proceed with each conversion. Be sure

Exhibit C.7 Recoding Instructions

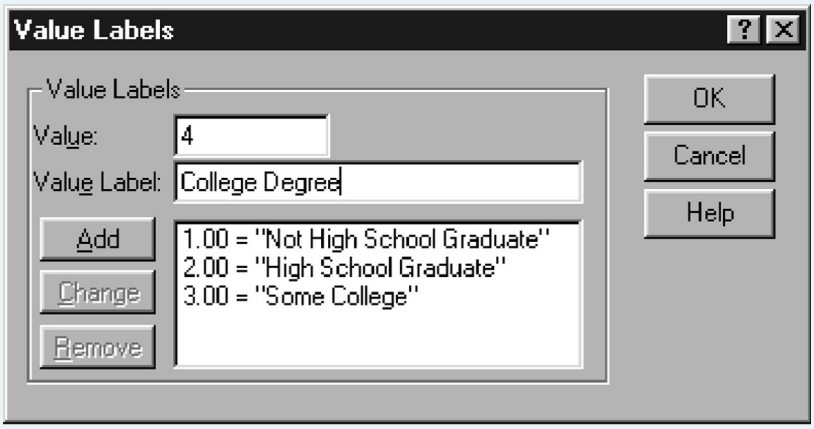


to recode any missing values (you could just recode "System- or user-missing" under "Old Value" to "System-missing" under "New Value").

After you have created the desired categories, click OK. You will be returned to the first recode box and will need to click OK again. The newly created variable (ED4CAT) will be located at the end of the data matrix. Obtain a frequency distribution for ED4CAT. You will notice that there are no value labels. Therefore, to make this output easier to read, it will be necessary to attach value labels to the numeric codes. Go back to the Data View window. Scroll to the far right of the data matrix. Double-click the name of the new variable ED4CAT. This will put you in Variable View mode, which will allow you to edit the characteristics of the variable.

Using the right arrow key on the keyboard, go to the "Values" cell. Click on the button located next to the word "None." A box entitled "Value Labels" will appear. Enter each category value and the corresponding value label where indicated (Exhibit C.8). After all value labels are entered, click OK.

Exhibit C.8 Defining Value Labels



Computing a New Variable

Recoding is very useful when you want to alter the format of a given variable. However, sometimes it may be necessary to combine multiple variables. For example, you may have a survey measuring aggressive behavior. To best measure levels of aggressiveness it would make sense to build an additive index from the data so that each respondent has an aggressiveness score (see Chapter 4). To do this, it is necessary to modify the data through the use of the Compute command. This command works well for combining two or more variables and for performing other types of mathematical transformations of the data.

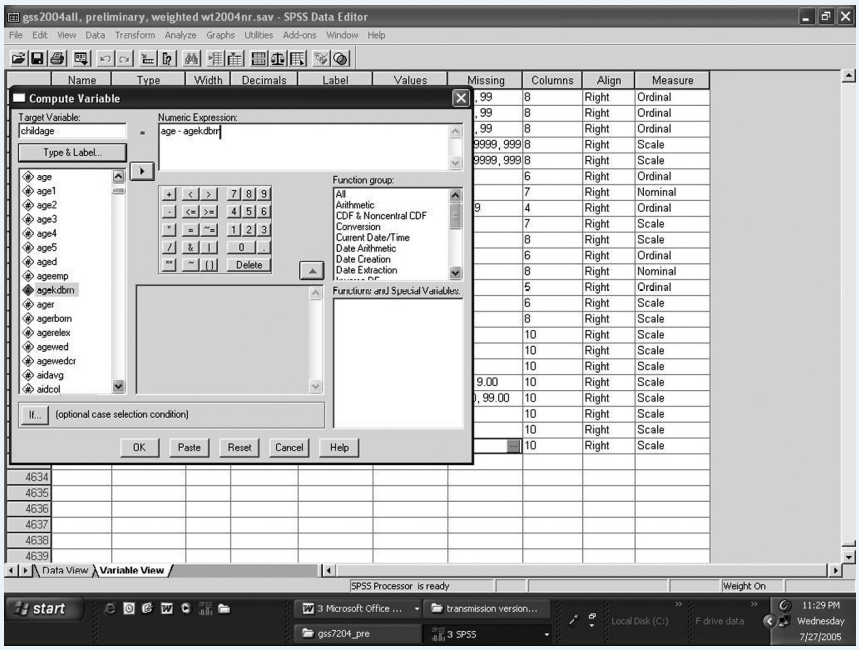
Using the Compute command, let's calculate the age of each respondent's eldest child. Click:

Transform—Compute

The Compute command will create a new variable based on the mathematical functions you select—the target variable. Name the target variable CHLDAGE (Exhibit C.9). To compute this, we need to calculate the difference between the respondents' age and the age of the respondent when his or her first child was born. Therefore, move the AGE variable to the "Numeric Expression" box. Using your mouse or keypad, click on the subtraction symbol. Now, move the AGEKDBRN variable into the "Numeric Expression" box, and click OK. You will find the new variable to the far right of the existing variables in the "Data View." If you generate a histogram for CHLDAGE, you will see that 24 is the mean age at which the first child was born.

If you scroll down the CHLDAGE column in the Data View, you will notice a number of cells with a "." in them. This is SPSS's default for "missing" data. The age of the eldest child could not be calculated for those individuals who did not have any children. Therefore, SPSS uses the "." to indicate that there were no data (or the data were missing) for those cases.

Exhibit C.9 Computing a New Variable



BIVARIATE AND MULTIVARIATE STATISTICS

Frequency distributions are good for describing the distribution of the variables in the dataset. However, to examine the relationship between two or more variables, a different set of statistical techniques are necessary. These are referred to as bivariate (two-variable) and multivariate (multiple-variable) statistics.

Crosstabulation

One way of exploring the relationship between two variables is through the use of a contingency table, more commonly referred to as a crosstab. Crosstabs are useful for exploring relationships between categorical variables. For example, we can use a crosstab to explore the relationship between level of education as measured by DEGREE and attitudes toward abortion for any reason (ABANY). Click:

Analyze—Descriptive Statistics—Crosstabs

For this analysis, DEGREE is the independent variable and ABANY is the dependent variable. Therefore, you will want DEGREE in the columns and ABANY in the rows. Locate these variables in the variable list (Exhibit C.10). Using the arrow buttons, place DEGREE and ABANY in their respective boxes.

SPSS automatically computes cell counts (the number of respondents that are "observed" within each category). However, you should also have SPSS calculate percentages for the independent variable to interpret differences. Therefore, click the Cells . . . button, choose the column percentages, and then click Continue to return to the first dialog box. If you have previous knowledge of statistics, you may want SPSS to also calculate a specific statistical test, such as chi-square. To do this, click the Statistics . . . button, make the appropriate choices, then click Continue. Click OK to run the crosstab command. Take a minute to inspect the resulting output (Exhibit C.11) and interpret the table (review Chapter 14 if necessary).

Exhibit C.10 Requesting Crosstabs

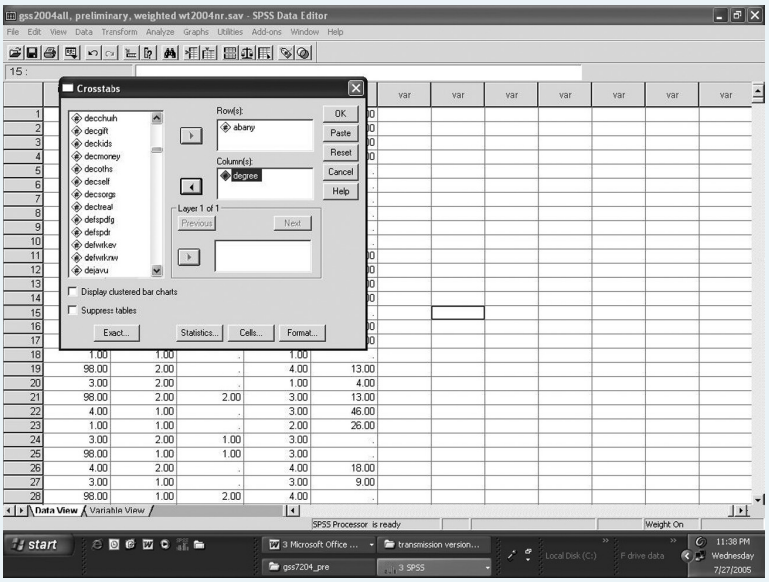
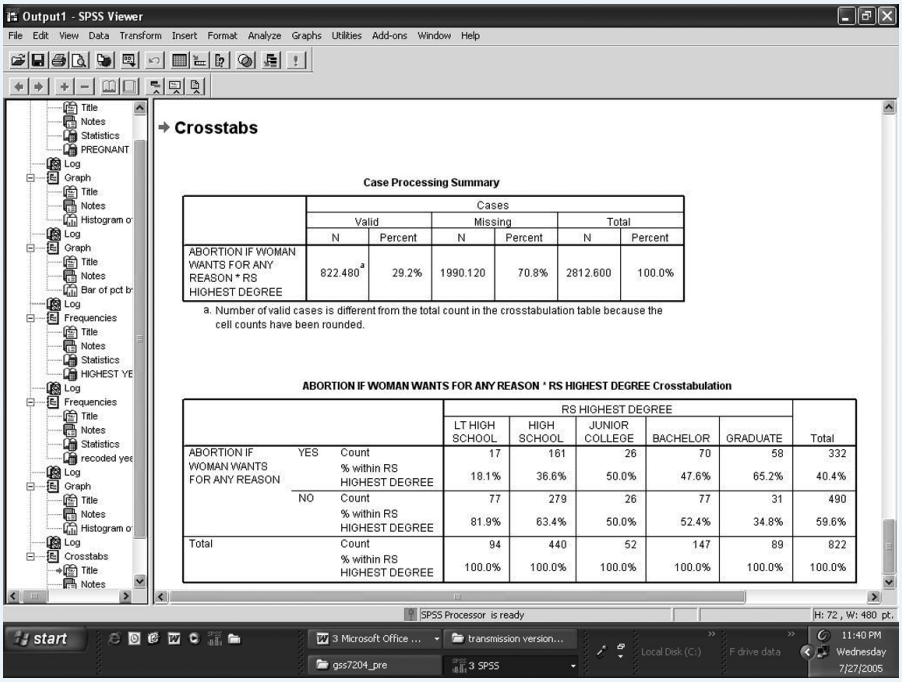


Exhibit C.11 Crosstabulation of ABANY by DEGREE



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ABORTION IF WOMAN WANTS FOR ANY REASON * RS HIGHEST DEGREE	822.480[a]	29.2%	1990.120	70.8%	2812.600	100.0%

a. Number of valid cases is different from the total count in the crosstabulation table because the cell counts have been rounded.

ABORTION IF WOMAN WANTS FOR ANY REASON * RS HIGHEST DEGREE Crosstabulation

			RS HIGHEST DEGREE					
			LT HIGH SCHOOL	HIGH SCHOOL	JUNIOR COLLEGE	BACHELOR	GRADUATE	Total
ABORTION IF WOMAN WANTS FOR ANY REASON	YES	Count	17	161	26	70	58	332
		% within RS HIGHEST DEGREE	18.1%	36.6%	50.0%	47.6%	65.2%	40.4%
	NO	Count	77	279	26	77	31	490
		% within RS HIGHEST DEGREE	81.9%	63.4%	50.0%	52.4%	34.8%	59.6%
Total		Count	94	440	52	147	89	822
		% within RS HIGHEST DEGREE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Now repeat this procedure, substituting INCOM98R for DEGREE. This will generate a crosstabulation of support for abortion by respondent's family income.

The results indicate that support for abortion if the woman wants one for any reason increases with education (rising from about 18% to over 65% across the five education categories) and also with family income (an increase of about 24% from the lowest to the highest recoded income categories).

Three-Variable Crosstabs

In the last example, it was determined that there was a relationship between income and support for abortion. Could this relationship be due to an extraneous factor? A three-variable crosstab can be created to assess whether controlling for another variable eliminates the apparent influence of income on support for abortion.

You have already seen that support for abortion increases with education (DEGREE). Because income also increases with degree, and we know that maximum educational attainment most often precedes employment as an adult, it is possible that support for abortion varies with income because both of these variables are, in turn, influenced directly by level of education. If this is the case, the relationship between income and support for abortion would be spurious—that is, it would be due to the extraneous factor of education.

We can control for degree in order to assess the effects of education on the relationship between income and support for abortion. To do this, click:

Analyze—Descriptive Statistics—Crosstabs

You should click ABANY into the rows box and INCOM98R into the columns box in the Crosstabs dialog box (unless they are still there after your last procedure). Select the variable

DEGREE from the variables list and move it to the empty box on the bottom titled "Layer 1 of 1." Select the column percentages and the chi-square statistic, if desired, then click OK.

This is a much bigger table to interpret, as it is really four subtables of ABANY by INCOM98R for the four different values of DEGREE. If you compare the percentages of YES responses from the lowest to the highest income level, first for the subtable representing respondents with less than a high school education and then for the other four subtables, you'll no longer find the clear pattern we found of increasing support for abortion as family income increases. Instead, the support for abortion increases a bit across some, but not all, of the income categories. The results indicate that education largely explains the effect of family income on support for abortion for any reason. You might take a minute to speculate about what might account for this pattern (and if you have had a course in statistics, and you check the chi-square statistic for these tables, you may realize that we cannot reject the possibility that the weak relationship between support for abortion and income is simply due to chance).

Comparing Means

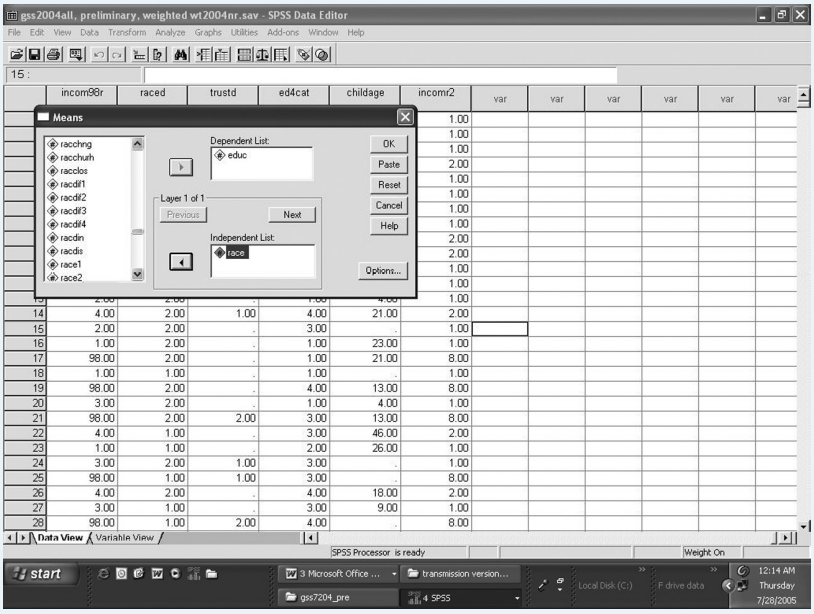
How would you compare mean years of education of whites and minorities? This type of question requires a comparison of means. A comparison of means is useful when you want to look at differences between two or more groups, such as by race or gender.

Click:

Analyze—Compare Means—Means

You must choose two variables: a dependent variable, which must be an interval/ratio level variable (or an ordinal variable that you are treating as interval), and an independent variable, which should be nominal or ordinal. For this example, EDUC is the dependent variable and RACE is the independent variable (Exhibit C.12).

Exhibit C.12 Requesting a Comparison of Means



After you identify your dependent and independent variables, click OK. SPSS will produce a table that shows the mean years of education for each category within the RACE variable.

The results indicate that the mean years of education for whites is about one year higher than the mean number of years of education for blacks, but it is about the same as for individuals in the "other" category.

SAVING AND RETRIEVING FILES IN SPSS

A number of the exercises in this book require you to recode variables. To eliminate the need to re-create variables you created via the Recode command, it is a good idea to save the dataset as a new file—with a new name—on either your hard drive or a floppy disk or a memory stick. In the Student Version of SPSS, the maximum number of variables allowed in any given dataset is 50. If you exceed the 50-variable limit, SPSS will save only the first 50 variables. You can delete variables you do not need by highlighting the corresponding column in Data View and pressing Delete on your keyboard. In any case, do *not* save the dataset with the same name. You might unintentionally make a permanent change in the values of variables in the dataset and later come to regret it.

Saving SPSS Data Files

To save the dataset as a new file, click:

File—Save As

The Save As dialog box will open. It should be similar to the File_Open dialog box shown at the beginning of this appendix.

If you are saving the file to your hard drive, you should save it in the SPSS folder or My Documents folder to find it easily later. After you type in the new name, click the Save As button.

Be forewarned: Output files can quickly take up a *lot* of space on your hard disk, or on a floppy. In most cases, you should just print the output instead of saving it in a file.

Opening Output Files

To open the output file(s) you saved, simply click:

File—Open—Output

The Open File dialog box will open. From the "Look in" drop-down text box, locate the folder where you saved the output file. Click on "ABANYFreq.spo" (or any other Viewer output file you want to open) to highlight it, and then click Open. The file will open in the Output window.

PRINTING YOUR OUTPUT

There are two ways to print from SPSS: from the menu bar and from the tool bar. First, let's look at how to print from the menu bar. Click:

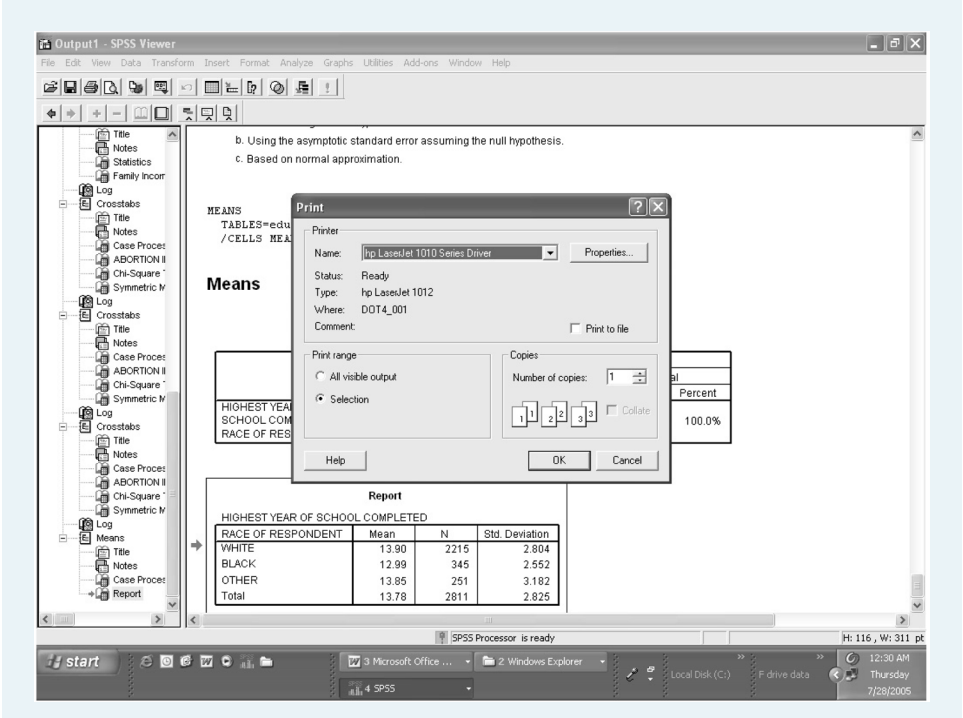
File—Print

From the print dialog box, you can choose to print all of your output or only selected items, the type of printer, and then number of copies to print. After you have made your selections, click OK. If you want to print only part of the output, you can select multiple portions by depressing the Ctrl key while clicking on each portion you would like to print (this must be done prior to selecting the print option). Then, from the print dialog box, click on the radio button for Selection in the "Print Range" section (to mark it with a small dot), and then click OK. Exhibit C.13 shows an illustration of some of these features in the print dialog box.

In addition to printing from the menu bar, you can also print by clicking the "Print" button on the tool bar (the printer icon). This brings up the same print dialog box.

Now that you have been introduced to a variety of different ways to communicate with SPSS, you may be feeling a bit overwhelmed. While you are still familiarizing yourself with SPSS, you may want to stick to the menu bar options as they guide you on how to proceed. When you become more comfortable using SPSS, you may like the convenience of the tool bar shortcuts. Keep in mind that you don't have to learn everything all at once.

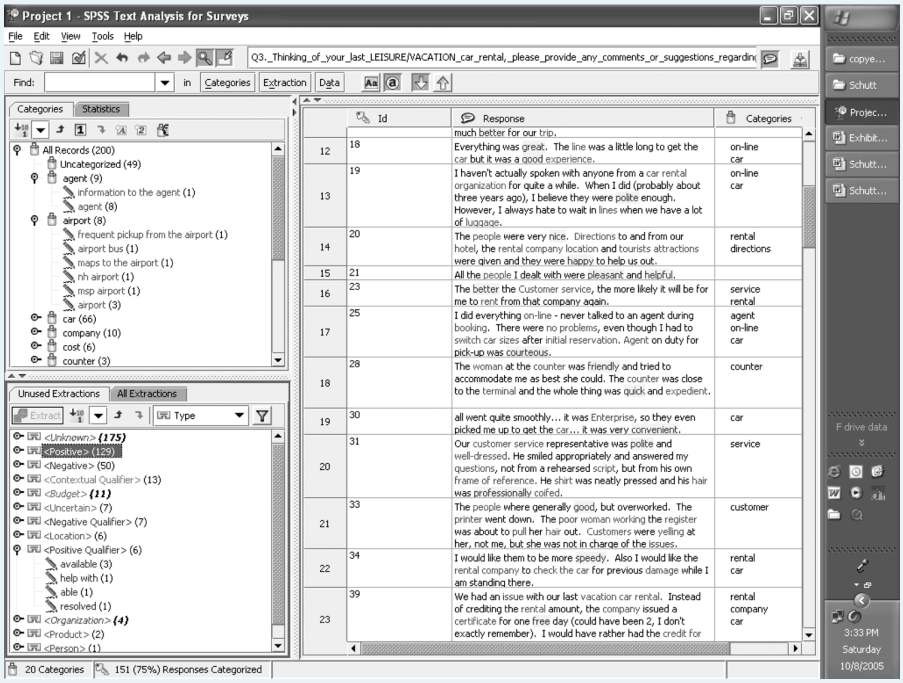
Exhibit C.13 Print Dialog Box



TEXT ANALYSIS

You can also use a supplementary SPSS module to code textual data systematically, if you or your university has purchased it. SPSS Text Analysis for Surveys combines a text search procedure with an automated linguistic technique that allows you to classify text according to key words and adjectives. Exhibit C.14 displays a Text Analysis screen that illustrates key features using an SPSS-supplied database. On the right, you can see some of the text that was entered into the case records in the original SPSS file and then imported into the Text Analysis program. In the upper left window, you can see the categories that were identified for the

Exhibit C.14 Data Categorized with SPSS Text Analysis Software



analysis, as well as the phrases that are grouped within the categories of "agent" and "airport." In the lower left window, you can see the text extracts that are used by the linguistic processing feature to identify statements about the categories as positive or negative. The words used in the categories and extracts are highlighted in the text, on the right. The categories also appear in a separate column on the right, for each case.

Development of the SPSS Text Analysis system reflects the way in which quantitative and qualitative data analysis procedures are converging in some areas, such as in the analysis of survey data that contains some textual data along with the quantitative responses. If you have access to this program, you will find many opportunities for using its capabilities.

CONCLUSION

At this point, you should be familiar with SPSS and able to complete all the exercises included in this book. Remember, the more you practice, the more comfortable you will be with using SPSS and statistics. If you have trouble, don't be afraid to refer to the SPSS Help file or Statistical Tutor, which are available from the Help drop-down menu, or ask your instructor or the computer lab consultant. Everyone runs into problems, but you can solve problems more quickly if you do the following:

- Write down the error message you are getting.
- Try to determine whether your problem is an SPSS problem. If so, look for help from the many sources of assistance out there: the Help drop-down menu in SPSS,

the SPSS manual, some campus computer lab consultants, your instructor, and quite possibly your classmates.

- Try to determine whether your problem is a Windows problem or a computer hardware problem. If so, your college probably offers computer lab technical support. Introductory books for Windows are an excellent resource because they typically describe problems common to Windows and to many software programs.

EXHIBIT C.1 Summary of SPSS Drop-Down Menus

The **File** menu enables you to open and save data files, import data created by other software, and print the contents of the data editor (or print the output, when you are in the SPSS Viewer window).

Use the **Edit** menu to cut, copy, and paste data; find data within the open data set; insert a new variable or case; and change options settings. You can also enter data directly in the "Data View" spreadsheet.

The **View** menu allows you to change how the data editor looks by changing fonts, turning toolbars on and off, turning grid lines on and off, and turning value labels on and off.

Use the **Data** menu to sort, select, or weight cases.

The **Transform** menu lets you make changes to selected variables, compute new variables, recode the values of existing ones, and replace missing values.

Use the **Analyze** menu to select the statistical procedure you want to use.

The **Graphs** menu allows you to produce a variety of two- and three-dimensional graphical displays of your data for purposes of data analysis and presentation.

Use the **Utilities** menu to obtain information about the variables in the open data set and select a list of variables to appear in dialog boxes.

The **Window** menu lets you switch between SPSS windows or minimize all open SPSS windows. Use the **Help** menu to access SPSS help topics and run the tutorial.