

SPSS for Windows

Version 19.0: A Basic Tutorial

Linda Fiddler, *California State University, Bakersfield*
Laura Hecht, *California State University, Bakersfield*
Edward E. Nelson, *California State University, Fresno*
Elizabeth Ness Nelson, *California State University, Fresno*
James Ross, *California State University, Bakersfield*

Preface

This book is intended for those who want to learn the basics of IBM SPSS (Version 19). This book can be used as a text in a class or by those working independently. It was written as a tutorial. Each chapter has instructions that guide you through a series of problems, as well as graphics showing you what your screen should look like at various steps in the process. The data set that accompanies the text, GSS10A, can be downloaded from the web by going to <http://www.ssrhc.org> and clicking on ONLINE TEXTBOOKS under Teaching Resources. Click on download GSS10A to download the file. The data have been weighted by the variable (wtss) that is provided by the National Opinion Research Center at the University of Chicago. There are also exercises at the end of each chapter for further practice and more exercises and teaching modules at this same web site.

In order to use the software that you will need for this book, you will either have to purchase IBM SPSS or find a computer laboratory in which IBM SPSS has been installed. Most colleges and universities have labs where you can use IBM SPSS. If you purchase IBM SPSS, we suggest you buy the Graduate Pack.

The authors would like to thank the faculty and students who have tested each version of these materials (6.0, 7.5, 9.0, 11.0, 13.0, 16.0 and now 19.0). We also want to thank the Social Science Research and Instructional Council of the California State University for their support and encouragement. Finally, we would like to thank our editor, James A. Doepke of McGraw-Hill Learning Solutions, for his help and support. If you would like to make suggestions, correct errors, or give us feedback, please email any of us at the following addresses:

Linda Fiddler (fiddlerlinda@yahoo.com)

Laura Hecht (lhecht@csu.edu)

Edward Nelson (ednelson@csufresno.edu)

Elizabeth Ness Nelson (elizn@csufresno.edu)

James Ross (jross1938@yahoo.com)

Table of Contents

GSS10A, the data set for this text can be downloaded from the web by going to <http://www.ssrlic.org> and clicking on ONLINE TEXTBOOKS under Teaching Resources.

CHAPTER ONE: GETTING STARTED WITH IBM SPSS FOR WINDOWS	1
USING WINDOWS.....	1
STARTING IBM SPSS FOR WINDOWS	2
LEAVING IBM SPSS.....	3
LOOKING AT DATA	3
A BRIEF TOUR OF IBM SPSS	5
OVERVIEW OF CHAPTERS	9
CHAPTER TWO: CREATING A DATA FILE.....	10
BASIC STEPS IN CREATING A DATA FILE	10
GETTING STARTED IN IBM SPSS	12
STUDENT SURVEY QUESTIONNAIRE	15
CHAPTER TWO EXERCISES	18
CHAPTER THREE: TRANSFORMING DATA	21
RECODING VARIABLES	21
RECODING INTO DIFFERENT VARIABLES	22
RECODING INTO THE SAME VARIABLE	26
CREATING NEW VARIABLES USING COMPUTE	27
CREATING NEW VARIABLES USING IF	33
USING SELECT CASES	34
CONCLUSION	36
CHAPTER THREE EXERCISES	37
CHAPTER FOUR: UNIVARIATE STATISTICS	39
FREQUENCIES	39
VARIABLE NAMES, VARIABLE LABELS, VALUES, VALUE LABELS, OH MY!	43
DESCRIPTIVES.....	45
EXPLORE	46
CONCLUSION	51
UNIVARIATE ANALYSIS AS YOUR FIRST STEP IN ANALYSIS	51
CHAPTER FOUR EXERCISES	53
CHAPTER FIVE: CROSS TABULATIONS.....	55
CHAPTER FIVE EXERCISES.....	59
CHAPTER SIX: COMPARING MEANS.....	60
INDEPENDENT-SAMPLES T TEST.....	60
PAIRED-SAMPLES T TEST	62
ONE-WAY ANALYSIS OF VARIANCE.....	63
CONCLUSION	65
CHAPTER SIX EXERCISES.....	66

CHAPTER SEVEN: CORRELATION AND REGRESSION	67
CORRELATION	67
REGRESSION.....	69
CHAPTER SEVEN EXERCISES	75
CHAPTER EIGHT: MULTIVARIATE ANALYSIS	76
CROSSTABS REVISITED	76
MULTIPLE REGRESSION.....	79
CHAPTER EIGHT EXERCISES	83
CHAPTER NINE: PRESENTING YOUR DATA	84
CHARTS.....	84
TABLES	86
COPYING AND PASTING CHARTS AND TABLES TO A DOCUMENT.....	87
APPENDIX A:	90
CODEBOOK FOR THE SUBSET OF THE 2010 GENERAL SOCIAL SURVEY	90

Chapter One: Getting Started With IBM SPSS for Windows

Using Windows

The Windows start-up screen should look something like Figure 1-1. Several standard desktop icons will always appear on start up. Note the Taskbar along the bottom, with the **Start** icon at the far left. If you own your own computer, you can do quite a bit of customizing of your desktop by choosing your favorite colors and scenes, rearranging and adding icons, moving the Taskbar to a different location, hiding it from view, and so on.

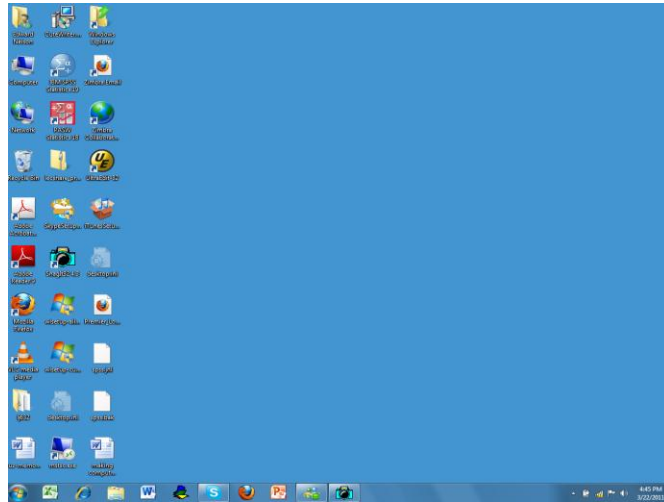


Figure 1-1

Although Windows makes use of the right-button key on the mouse, you will only use the left button for now. When we say to click on the mouse button, it will always mean the left one. A single click will often take you where you want to go, but if one click doesn't do anything, try a double click. (Double clicking means to press the left mouse button twice in rapid succession. If nothing seems to happen, you probably need to double click more rapidly.) Now move your mouse so the pointer touches the **Start** icon (don't click anything yet, just let the pointer rest somewhere on the icon). Notice that a label appears in a little rectangle telling you indicating this is the **Start** icon.

These floating labels will usually appear whenever the mouse cursor moves over an icon. They will then disappear after a few seconds so pay attention! Now that you have found the **Start** icon, click on it once and the **Start** menu will appear. Your screen should look something like Figure 1-2. This is the basic **Start** menu. It can also be customized by adding your most often used programs to it. Without clicking anything yet, move the mouse pointer up and down the **Start** menu. As you encounter each item, it will become highlighted. Some of the items

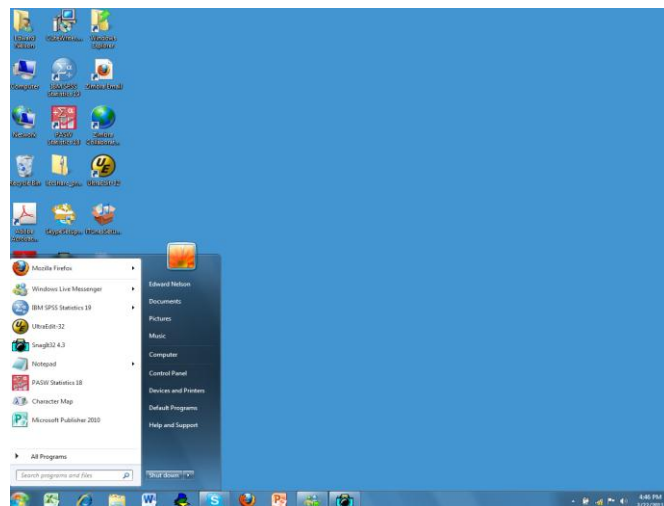


Figure 1-2

on the **Start** menu have little triangles on the right. If one is present, that means there are more menus for that item. If you have used programs that saved files, the last ones you saved will show up on this menu. This is a quick way to get back to whatever you were working on recently, since one click will open that file and the applications program it belongs to.

Move the mouse pointer around on those icons with a little triangle at the right (don't click anything yet) and watch what happens. When the pointer lands on an icon, it becomes highlighted and its sub-menu appears.

Starting IBM SPSS for Windows

The IBM SPSS 19 icon should be on the **Start** menu. If you are using a computer in a lab, it is common for the icon to be placed in a folder. If you customize your computer, all you have to do to start IBM SPSS is to point to the IBM SPSS 19 icon on the desktop and double click. Then wait while IBM SPSS loads.

After IBM SPSS loads, you may, depending on how IBM SPSS is set up, get a menu that asks, “What would you like to do?” For this tutorial, click **Cancel** to get rid of this. Now the screen should look like Figure 1-3.

The Taskbar is probably visible at the bottom of the screen. It shows an IBM SPSS icon. Whatever programs you have open will show on the Taskbar and the one you are currently using will be highlighted.

Next, observe the three small squares in the uppermost right-hand corner of the main IBM SPSS window. The one furthest to the right, with an

X in it, is used when you want to close any program you are using. Don't worry if you click on it accidentally, a dialog box will pop up asking if you want to save anything that changed since the last time you saved your work, before it actually lets you exit the program.

The middle of the three small squares allows you to have the window you are working in fill up the whole screen, or to shrink it down to a smaller size. If the middle square shows two cascading rectangles in it, the window is already as big as it can get—clicking on this square will reduce the window in size. Try this now. In this shrunken window,

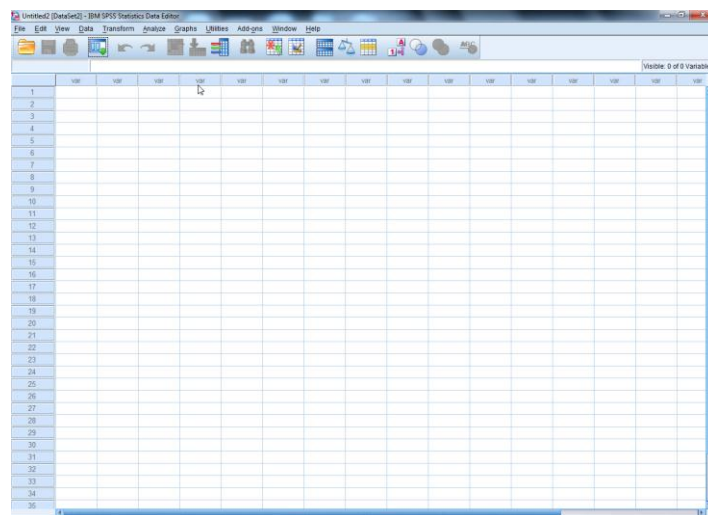


Figure 1-3

the middle button now shows only one rectangle. Click on it to get back to the full screen view.

The last square, to the left of the other two, has what looks like a minus sign on it. Click this and watch what happens. Look at your Taskbar. The icon for IBM SPSS is still there but it is no longer highlighted. Click on it and see what happens. You have just learned how to minimize a window.

Leaving IBM SPSS

We're not ready to actually use IBM SPSS yet, so let's close it. There are at least four ways to do this. Move your mouse until the arrow is pointing at the word **File** in the upper-left hand corner of the screen and press the left mouse button once. A menu will appear. Move the arrow so it is pointing at the word **Exit** and press the left mouse button. This should close IBM SPSS.

There is a second technique that can do the same thing. (Computers usually have more than one way to do everything.) Go back into IBM SPSS and move your mouse until the arrow is pointing at the word **File** in the upper-left corner of the screen and press the left mouse button once, but this time, don't release the mouse button. Hold the mouse button down and move the mouse down until the word **Exit** is highlighted. Now release the mouse button and IBM SPSS should close. This is called click and drag and is another way to use your mouse.

The third way to close IBM SPSS is to point your mouse at the IBM SPSS icon in the upper-left corner of the screen. The icon will be just to the left of the words **Untitled: IBM SPSS Statistics Data Editor**. Move your mouse to the icon and double click on it. This has the same effect as the first two procedures; it closes IBM SPSS.

The final and fourth way to close IBM SPSS is to point your mouse at the X in the upper-right corner of the screen and click. IBM SPSS will close.

Now you know how to move your mouse around and how to start and close IBM SPSS. We'll show you more about Windows, but not much more. If you want to learn more about Windows, there are a lot of books available. The nice thing about Windows is that you don't have to know much about it to use it.

Looking at Data

There is a data file to be used with this tutorial called GSS10A. This is a subset of the 2010 General Social Survey (Release 1). The General Social Survey is a large, national probability sample done every two years. You can download this data file from the web by going to the following address on the web, <http://www.ssric.org> and clicking on ONLINE TEXTBOOKS under Teaching Resources. Click on download GSS10A to download the file. Also, if you are in a computer lab, someone may have copied the data files onto your hard drive.

Let's begin by starting IBM SPSS just as you did above. Your screen should look like Figure 1-3. (If you see a box asking, "What do you want to do?," click on Cancel to close this box.) At the very top of the screen, you will see the words **IBM SPSS Statistics Data Editor**. Just below that line will be the menu bar with the following options: File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Add-ons, Window, and Help. Point your mouse at **File** and press the left mouse button.

A box will open that is the **File** menu. Point your mouse at **Open** and then at **Data** and click. (Also, you could have gotten to this point by clicking on the **Open File** icon just below **File** on the menu bar.)

This opens a larger box called the **Open Data** box. (You're beginning to get the idea of how Windows works.) Here you need to tell IBM SPSS where to find the data file to open. The file you are looking for is an SPSS "portable" file, meaning that it will run on any platform (such as Macintosh). You'll need to indicate this by clicking on the down arrow next to **File of type:** and choosing **Portable (*.por)**. In the upper part of the box, you'll see **Look in:**. Find the folder that contains the data file and click on the file name, **GSS10A**, to highlight it and then click on **Open**. In a few seconds, your data matrix will appear.

A data matrix is a very important concept. The rows contain the cases and the columns contain the variables. (If you're familiar with spreadsheets, that's what this is.) Row 1 is case 1; row 2 is case 2, and so on. The top of each column contains the variable name. In this data set, the variable names are abbreviations like *abany* and *abdefect*.

Unfortunately, the abbreviations for the variable names do not tell you very much. We need some way to find out what these variables are. So try this. On the menu bar at the top of your screen, you'll see the word **Utilities**. Point your mouse at **Utilities** and click the left button. This will open the **Utilities** menu. Point your mouse at **Variables** and click again. Your screen should look like Figure 1-4.

You'll see a list of all the variables in your data on the left side of the little window. (Also, see Appendix A for a list of variables.) Point your mouse at any of these variables and click. To the right of the **Variable List** you'll see a short description of this variable.

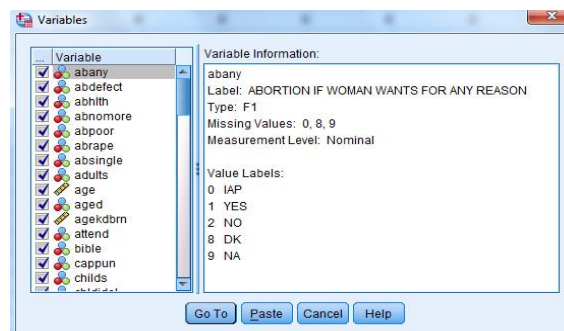


Figure 1-4

For example, point your mouse at the variable *abany* and press the left mouse button. This question asked if respondents thought that obtaining a legal abortion should be possible for a woman if she wants it for any reason. The possible answers are YES (value 1), NO (value 2), DK or don't know (value 8), NA or no answer (value 9), and IAP or not applicable (value 0). (Not applicable includes people who were

not asked the question.) As you will see in Chapter 3, these values are very important!

Now you know how to open a preexisting data file in IBM SPSS and how to find out what the variables are in the file. We will tell you more about this later, but here we just want to give you a brief introduction to IBM SPSS.

A Brief Tour of IBM SPSS

Now that you have the file opened, let's look at some things you can do with IBM SPSS. You're already familiar with the variable *abany*. Let's find out what percent of people surveyed thought it ought to be legal for a woman to have an abortion for any reason. (If you have the Variables window open showing the variable labels and values, point your mouse at the close button and click it.) On the menu bar you will see **Analyze**. Point your mouse at **Analyze** and click it. A box opens that looks like Figure 1-5.

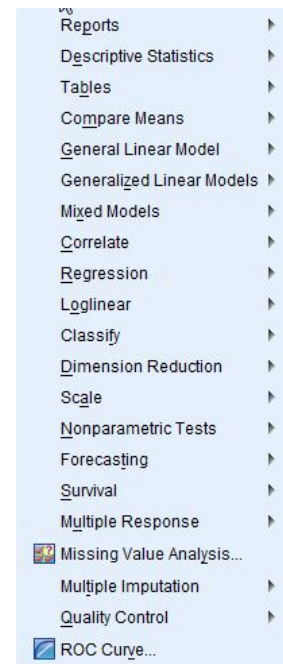


Figure 1-5

This lists the statistical procedures in IBM SPSS. We want to use **Descriptive Statistics** so point your mouse at **Descriptive Statistics**. This opens another box listing the statistical procedures you can use to summarize your data.

Point your mouse at **Frequencies** and click it. This opens the Frequencies box. Since *abany* is the first variable in the data, it's already highlighted.¹ Point your mouse at the right arrow next to the list of variables and click it. The label *abany* will move to the box called Variable(s). This is how you select variables. Point your mouse at **OK** and click it. In a few seconds, a new screen should appear that looks like Figure 1-6. We are now in a different part of IBM SPSS called the Output Window. This is where the results, or output, are displayed.

The Output Window is divided into two vertical frames or panes. The left-hand pane contains the outline of the output or information that IBM SPSS gives you. This information is in outline form and can be used to select what you want to view. Simply click on the information you want to look at and that information will appear in the right-hand pane. You can also collapse and expand the outline by clicking on the plus and minus signs in the left-hand pane. The plus sign indicates that the information is collapsed (or hidden) and the minus sign indicates that it is expanded (or shown). You can use the scroll bars on each pane to scroll through the Output Window.

¹ Instead of seeing a list of variables, you may see a list of variable labels. You can change this so IBM SPSS displays the list of variables. To do this, click on **Edit** in the menu bar, then click on **Options** and on the General tab. Look for Variable Lists in the General tab and click on Display names. You can put the variable names in alphabetical order by clicking on Alphabetical just below Display names. Finally, click on **OK**.

On the right side, the frequency distribution for *abany* is divided into four parts: (1) the title, (2) notes on the table (there aren't any for this table), (3) the active data set, (4) statistics (a summary of the number of missing and valid observations), and (5) the actual table showing the frequency distribution. Click on **Statistics** in the left-hand pane and you will see that there were 1,234 valid and 810 missing cases. Click on **abortion—if woman wants for any reason**

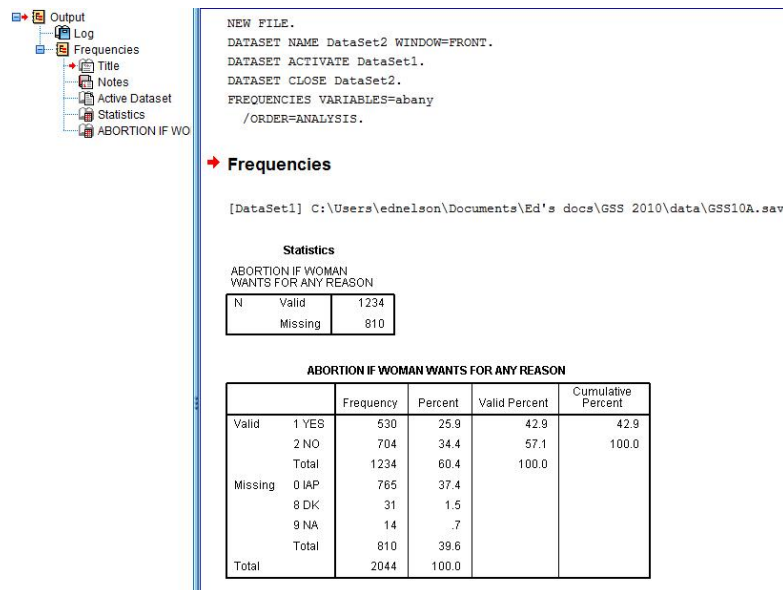


Figure 1-6

and you will see the frequency distribution. In Figure 1-6, you can see that 530 people said yes, 704 said no, 31 said they didn't know, 14 didn't answer the question, and 765 were coded not applicable. (These 765 respondents were not asked this question. In survey research it's very common to ask some, but not all, of the respondents a particular question. In this case, only 1,279 of the approximately 2,044 respondents were asked this question.) Of those who had an opinion, we want to know what percent of the respondents said yes or no, so we should look at the Valid Percents in the table. About 43% of the respondents who had an opinion thought it should be legal, while 57% thought it should be not be legal.

It would be interesting to know if men or women were more likely to favor allowing a legal abortion when the woman wants it for any reason. We're going to use a crosstab to determine this. Point your mouse at **Analyze** and press the left mouse button. Then point your mouse at **Descriptive Statistics** and finally, point your mouse at **Crosstabs** and press the mouse button. Your screen should look like Figure 1-7.

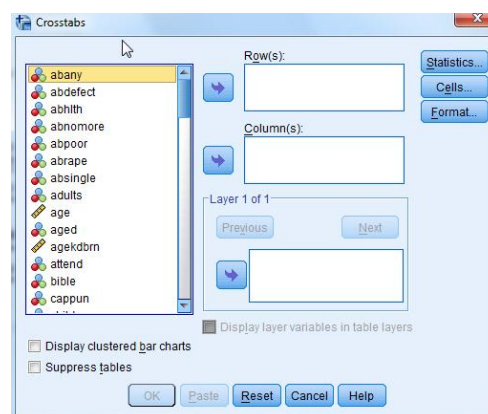


Figure 1-7

The list of variables in your data set is on the left of the screen. We want to move the variable *abany* into the box next to the list of variables where it says **Rows**. Click on the variable *abany*, which will highlight it.

Now click on the arrow pointing to the right, which is next to the Rows box. Notice that this moves *abany* into the Rows box. We also need to move the variable *sex* into the Columns box. You will have to use the scroll bar in the box containing the list of variables to find this variable. (You can also click anywhere in this box and then type the letter **S** to move to the first variable starting with the letter S.) Point your mouse at the down arrow next to the list of variables and click. If you keep pressing the mouse button, the list of variables will move down and eventually you will see the variable *sex*. Highlight it and click on the arrow pointing to the right, which is next to the Columns box. This moves *sex* into the Columns box. Now your screen should look like Figure 1-8.

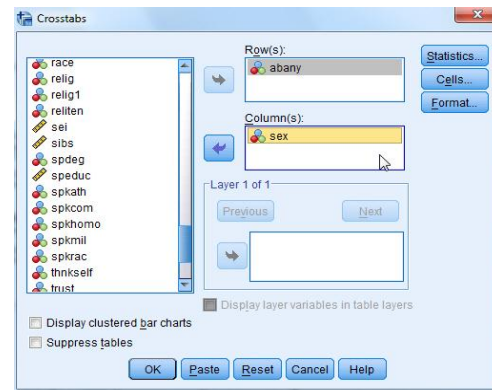


Figure 1-8

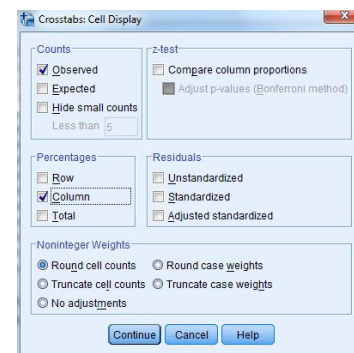


Figure 1-9

Frequencies by themselves are seldom useful. You'll need to compute percents. To get IBM SPSS to compute percentages, point your mouse at the button labeled **Cells** at the right of the screen and click on it. This will open the Crosstabs: Cell Display box. Find the box called Column Percentages and click on this box. This will place a check mark in this box and your screen should look like Figure 1-9.

Crosstabs

[DataSet1] C:\Users\ednelson\Documents\Ed's docs\GSS 2010\data\GSS10A.sav

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ABORTION IF WOMAN WANTS FOR ANY REASON * RESPONDENTS SEX	1234	60.4%	810.000	39.6%	2044.000	100.0%

ABORTION IF WOMAN WANTS FOR ANY REASON * RESPONDENTS SEX Crosstabulation						
			RESPONDENTS SEX		Total	
			1 MALE	2 FEMALE		
ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	260	270	530	
		% within RESPONDENTS SEX	45.3%	40.9%	42.9%	
	2 NO	Count	314	390	704	
		% within RESPONDENTS SEX	54.7%	59.1%	57.1%	
Total		Count	574	660	1234	
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%	

Figure 1-10

Now click on **Continue** and you will be back to the Crosstabs box. To tell IBM SPSS to run the Crosstabs procedure, click on **OK**. In a second your screen should look like Figure 1-10. Use the scroll bar to look at all the information that IBM SPSS gives you in the Output Window.

Figure 1-10 shows the results, or “output.” It shows, for example, that males and females differ very little in their opinions about a woman obtaining a legal abortion for any reason. Forty-five percent of the males and 41 percent of the females approve of a woman obtaining an abortion for any reason.

You can also examine other items in the survey to compare men and women. Who has more education? Is the average age at birth of first child younger for women than for men? Comparing means will answer these questions. Click on **Analyze**, point your mouse to **Compare Means**, and then click on **Means**. Your screen should look like Figure 1-11.

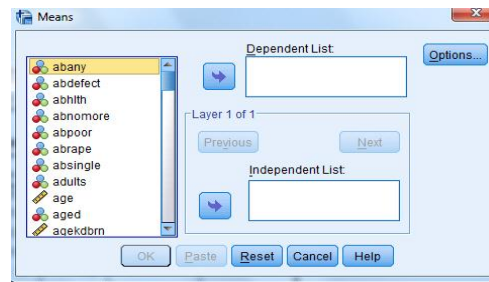


Figure 1-11

Now put age at birth of first child (*agekdbn*) and years of school completed (*educ*) in the Dependent List box and *sex* in the Independent List box. By now you have a good idea how to do this. Highlight *agekdbn* in the list of variables on the left of the screen by pointing your mouse at it and clicking. Then click on the arrow next to the Dependent List box. Do the same for *educ*. Now highlight *sex* and click on the arrow next to the Independent List box. This should move *agekdbn* and *educ* into the Dependent List box and *sex* into the Independent List box and your screen should look like Figure 1-12. Then click on **OK** and the output should look like Figure 1-13.

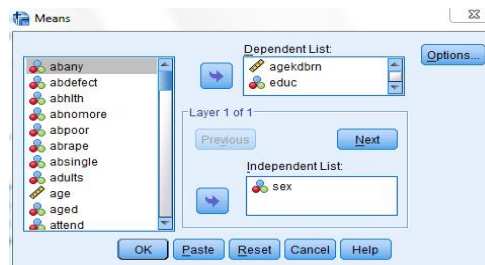


Figure 1-12

Report

RESPONDENTS SEX		R'S AGE WHEN 1ST CHILD BORN	HIGHEST YEAR OF SCHOOL COMPLETED
1 MALE	Mean	25.60	13.39
	N	628	921
	Std. Deviation	6.516	3.172
2 FEMALE	Mean	22.96	13.42
	N	855	1119
	Std. Deviation	5.413	3.054
Total	Mean	24.08	13.41
	N	1483	2039
	Std. Deviation	6.045	3.107

Figure 1-13

Women had their first child at an average age of 22.96 years, while the average for men is 25.60 years, a difference of less than three years. Now look at the mean years of school completed for men and women. There's virtually no difference (about 0.03 of a year) between men and women.

Another way of examining relationships is to look at Pearson Correlation Coefficients. One could hypothesize that respondents' education is correlated with the educational achievements of their parents. The Pearson Correlation Coefficient will tell us the strength of the linear relationship between father's education, mother's education, and the respondent's

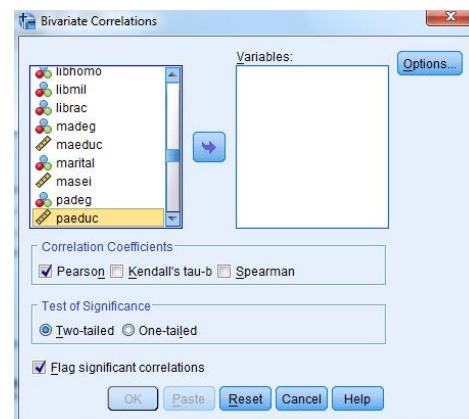


Figure 1-14

education. The closer the correlation is to 1, the stronger the relationship, and the closer it is to 0, the weaker the relationship.

Point your mouse at **Analyze** and press the mouse button. Now point your mouse at **Correlate** and then click on **Bivariate**. Your screen should look like Figure 1-14.

Now move the following three variables into the Variables box: *educ*, *maeduc*, and *paeduc*. These variables refer to the number of years of school completed by the respondent and the respondent's parents. Highlight each of these variables and press the arrow next to the Variables box. The screen should look like Figure 1-15. Click on **OK** and the correlations will appear in your Output box.

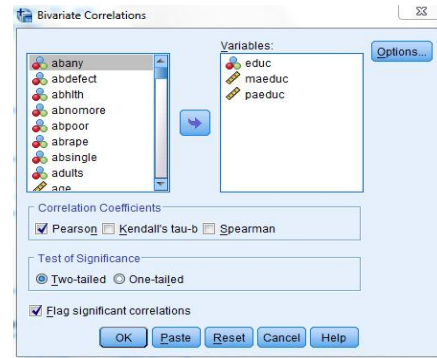


Figure 1-15

The strongest correlation is between father's and mother's education. As we predicted, there is also a fairly strong correlation between respondent's education and parent's education.

Overview of Chapters

Chapter 2 will acquaint you with how to enter new data into IBM SPSS using the **Data Editor**. Chapter 3 explains how to take your data, or preexisting data, and modify it in a way that makes it easier to understand. Chapter 4 starts the sections where you really get to see the results of your work. In Chapter 4 you will learn how to look at each variable, one at a time. We call this univariate analysis. Chapters 5 through 7 will show you how to look at two variables at a time, or what we call bivariate analysis. Chapter 5 will show you how to create a cross tabulation. Chapter 6 shows you alternative ways of comparing two variables and Chapter 7 will show you how to do this using linear regression techniques. Chapter 8 shows you how to explore relationships among sets of variables using multivariate cross tabulation and multiple regression. Finally, Chapter 9 will show you how to present your data effectively and will include charts and graphs.

Chapter Two: Creating a Data File

This chapter explains how to set up a file with new data. After finishing this chapter, you should be able to create an IBM SPSS data file that will include the data and some labeling that gives more detail about the data. To illustrate this process, we will use a shortened version of the questionnaire used by the General Social Survey conducted by the National Opinion Research Center. For this example, our students wanted to see if their opinions on social issues were similar to those of the national sample.

The students knew they were not a representative sample, even of college students, but this questionnaire is an interesting way to learn how to create a new data file. They decided to use the following questions²:

- What is your age?
- Are you male or female?
- What is your religious preference?
- Generally speaking, in politics, do you consider yourself as conservative, liberal, or middle of the road?
- What kind of marriage do you think is the more satisfying way of life: one where the husband provides for the family and the wife takes care of the house and children or one where both the husband and wife have jobs and both take care of the house and children?
- Do you think it should be possible for a pregnant woman to obtain a legal abortion:
 - If there is a strong chance of a serious defect in the baby?
 - If she is married and does not want any more children?
 - If the woman's own health is seriously endangered by pregnancy?
 - If the family has a very low income and cannot afford any more children?
 - If she became pregnant as a result of rape?
 - If she is not married and does not want to marry the man?
 - If the woman wants it for any reason?

Basic Steps in Creating a Data File

It is best to start a data file with some careful planning.

1. First we will assign each respondent an identification number. This is not so we can identify individuals, but so we can keep track of each case when we go back to check the accuracy of the data entering. Each question is a variable in our data set. It needs a variable name that is simple but expresses something important about the data.

² A copy of this questionnaire is included as Appendix 2-A at the end of this chapter.

(IBM SPSS limits variable names to 64 characters or fewer. They may be numbers or letters but not spaces and very few special characters, so don't use any odd symbols.) *Age* and *sex* would be good variable names for the first two questions.³ For the questions on abortion, we decided to use the first three characters of the variable names used by the General Social Survey. We used *mg* for the preferred type of marriage and called political orientation *conlib*. Each variable name can be given an extended variable label that gives more detail. (Extended variable labels can use spaces or special characters.) For example, *conlib* could have a variable label that said Conservative-Liberal.

2. After we have given each variable a name and label, we give each possible response to the question a code called a value label that is often the number corresponding to the order of the answers. (We could use another system, but this is the easiest because IBM SPSS works best with numeric codes to represent the data.) For example, *sex* could use 1 for male and 2 for female; *conlib* could use 1 for conservative, 2 for liberal, and 3 for middle of the road. Values would then be given value labels such as Male, Female, Conservative, Liberal, and Middle of the Road.
3. Sometimes respondents do not answer a question, give more than one answer, or do something else that makes their answers unusable. In our example, respondent #2 marked both yes and no on the last question, respondent #3 wrote in none on question 4, and respondent #13 didn't answer the marriage question. We assign these missing value codes so they don't distort the analysis. Often 9 is used to indicate missing data or 99 if it is a two-digit value.

Everything must be planned carefully before entering the data into IBM SPSS. It is useful to put the data in a matrix like Table 2.1 before entering it into the IBM SPSS Data Editor. For this exercise, we will use only the first four questions and five respondents. (The complete matrix is Appendix 2.B at the end of this chapter.)

Table 2.1 Matrix for Data-entry Exercise

<i>id</i>	<i>age</i>	<i>Sex</i>	<i>rel</i>	<i>conlib</i>
01	20	1	4	2
02	24	2	5	2
03	21	2	2	9
04	24	2	5	3
05	26	2	4	2

³ For this exercise, we used lower-case italics for the variable names.

Getting Started in IBM SPSS

To create the data file in IBM SPSS, open IBM SPSS (probably by clicking on the IBM SPSS icon on the desktop). (See Figure 2-1.)

This opens a matrix similar to a spreadsheet such as Excel or the matrix we just worked on. The rows will be the cases (the respondents) and the columns will be the variables (answers to the questions). So, the upper-left cell will contain an identification number for the first case and the cells to the right will be data about that case. The IBM SPSS Data Editor has tabs in the lower-left corner that let you work with your data in two ways. **Variable View** is used to set up the data—names, variable labels, value labels, etc. The other tab, **Data View**, is used to actually enter the data. IBM SPSS probably opened in the **Data View** mode, if not, click the **Data View** tab at the bottom left of the IBM SPSS screen now. (See Figure 2-2.)

Entering Variable and Value Names and Labels:

In **Data View**, we will use the first column for the respondents' ID numbers, so type **001** into the first cell and press Enter and 1.00 will appear. (See Figure 2-3.)⁴

We will use the **Variable View** tab to assign variable names and longer variable labels as well as value names and labels that will make it easier to use the data for tables and charts. Click **Variable View** now and click the VAR0001 in the top left column. Type in id. (Press Enter and VAR0001 changes to our variable name, id.) Go back to **Data View** and notice that the first column is now titled id. (See Figure 2-4.)

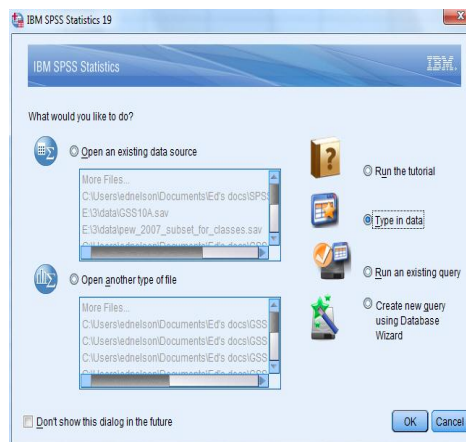


Figure 2-1

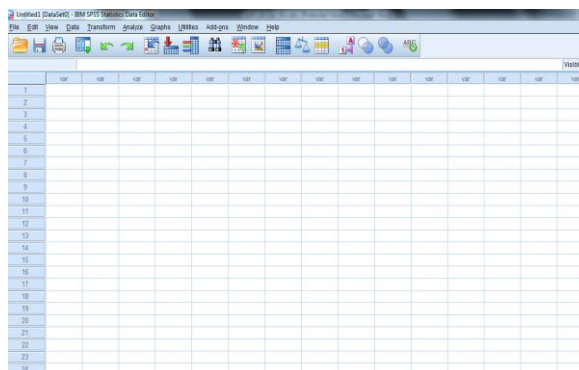


Figure 2-2

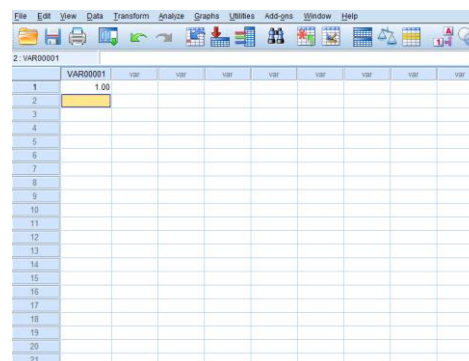


Figure 2-3

⁴ It is wise to save your computer work early and often. You might want to save this file and call it something like **Data Entry Exercise 1**. Notice that IBM SPSS saves it as a .sav file. This means it contains the data in the format for IBM SPSS analysis.

The second variable will be the student's age, so change back with the Variable View and type **age** under name in the second row. Remember to press Enter after you type in each entry.

IBM SPSS makes some assumptions about data that might not be appropriate. In the fourth column, notice that it plans to use two decimal points even when the values for age are integers. To avoid these inappropriate decimal points, select the **decimals** column by clicking the heading and use the blue box to go down to 0. (Remember to do this whenever a numeral doesn't really refer to a numerical value.)

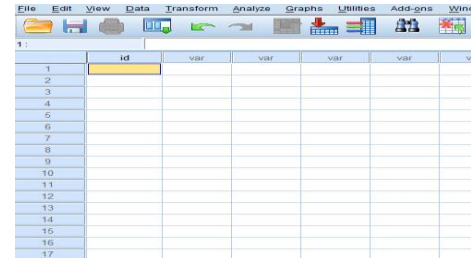


Figure 2-4

Since the short variable name usually doesn't give enough information about the variable, we want a longer or clearer variable label for our analysis. This one would be simple. To add a variable label to *age*, just tab over to the **label** column and type in **Age**. (See Figure 2-5.) Although, it may not seem necessary to have a variable label for age, for most variables a longer variable label is very useful.

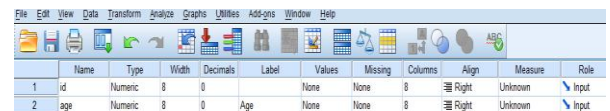


Figure 2-5

Sometimes respondents don't answer a question, give two answers, or do something else so the data can't be used in the analysis. To have accurate results, missing or invalid data need to be indicated. Still in Variable View, tab over to **Missing** and click the blue box. This dialog box lets you specify up to three missing values. For our data, click **Discrete** and type **99** in the first text box. Leave the other boxes empty. Then Click **OK**. Now if someone doesn't answer a question, it will be marked as missing. (See Figure 2-6.)

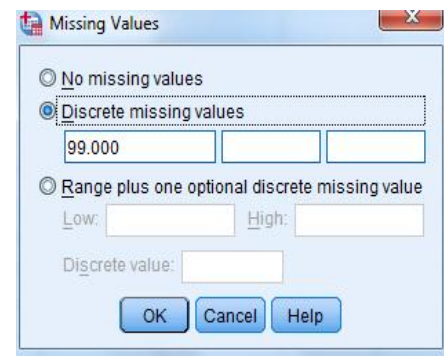


Figure 2-6

The third variable will be the sex of the respondent, so type **sex** in the third row under **Name** and **Sex** as the variable label. Since we're going to use the code 1 for males and 2 for females, we're going to need value names in words for each category. Tab over to the cell under **Values** and click the little blue box to get the value labels menu. Type a **1** in the Value box and then **Male** in the Value Label box. Click **Add** and it shows that Value 1 will be Male. Type a **2** in the Value box, and type **Female** in the Value

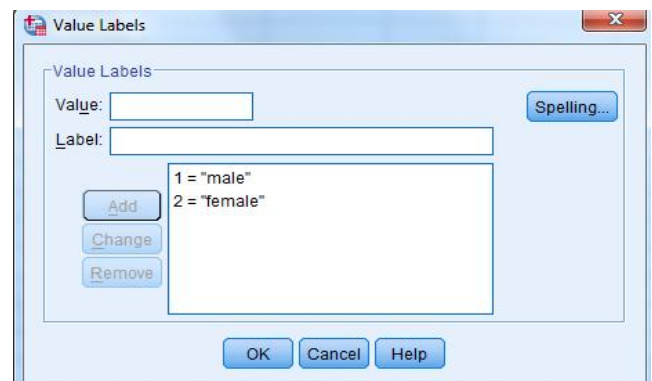


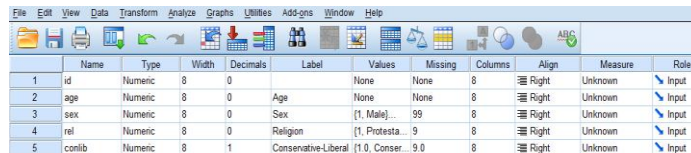
Figure 2-7

label space. Click **Add** and then click **OK** to save these. Now, IBM SPSS knows that 1 and 2 in **Sex** are really male and female respectively. (See Figure 2-7.)

For this exercise, we are also using religion and conservative-liberal as variables. Add those variables in rows 4 and 5. Give each variable a name and label—*rel* gets **Religion** and *conlib* gets something like **Conservative-Liberal** as variable name and label. Then add value names and labels.

Notice that *rel* has five possibilities—Protestant, Catholic, Jewish, other, and no religion. Go ahead and work out the value names and value labels.

Make arrangements for missing values just as you did before. (You can refer to Appendix 2-A, Codebook for Student Questionnaire at the end of this chapter.) Remember to type variable labels, value names, and value labels exactly the way you would want them in a table when you do the analysis—often this is with the first letter of each important word capitalized. (Your Variable View might look like Figure 2-8.)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	8	0		None	None	8	Right	Unknown	Input
2	age	Numeric	8	0	Age	None	None	8	Right	Unknown	Input
3	sex	Numeric	8	0	Sex	[1, Male]... 99		8	Right	Unknown	Input
4	rel	Numeric	8	0	Religion	[1, Protestan... 9		8	Right	Unknown	Input
5	conlib	Numeric	8	1	Conservative-Liberal	[1.0, Conser... 9.0		8	Right	Unknown	Input

Figure 2-8

Entering the Data:

Enter the codes for each variable using **Data View**⁵. Then check the accuracy of your data entry by scanning down each column looking for codes that would be impossible. For example, sex can have only three possibilities since male is 1, female is 2, and missing information is 9, so a 5 or 6 would be a mistake. Then check everything carefully. The best check is to have one person read the codes while another checks the entries on **Data View**.

⁵ Some people, especially those who are used to working with spreadsheets, like to enter all the data in **Data View** before they set up the variable names, etc. In this example, we set up the variable names, etc., before we enter any data. (You'll have to figure out what works best for you.) You can also enter data from a spreadsheet like Excel.

Student Survey Questionnaire

- (1) What is your age? _____
- (2) Are you _____ male or _____ female?
- (3) What is your religious preference?
_____ Protestant _____ Catholic _____ Jewish _____ Some other religion _____ No religion
- (4) Generally speaking, in politics, do you consider yourself as
_____ conservative _____ liberal _____ or middle of the road
- (5) What kind of marriage do you think is the more satisfying way of life?
_____ One where the husband provides for the family and the wife takes care of the house and children
_____ One where both the husband and wife have jobs and both take care of the house and children

Do you think it should be possible for a pregnant woman to obtain a legal abortion?

- (6) If there is a strong chance of serious defect in the baby?
_____ Yes _____ No _____ Don't Know
- (7) If she is married and does not want any more children?
_____ Yes _____ No _____ Don't Know
- (8) If the woman's own health is seriously endangered by pregnancy?
_____ Yes _____ No _____ Don't Know
- (9) If the family has a very low income and cannot afford any more children?
_____ Yes _____ No _____ Don't Know
- (10) If she became pregnant as a result of rape?
_____ Yes _____ No _____ Don't Know
- (11) If she is not married and does not want to marry the man?
_____ Yes _____ No _____ Don't Know
- (12) If the woman wants it for any reason
_____ Yes _____ No _____ Don't Know

Appendix 2A: Codebook for Student Questionnaire

Missing Values	9 or 99
Age	Age at last birthday
Sex	1 = male, 2 = female
Religious Preference	1 = Protestant, 2 = Catholic, 3 = Jewish, 4 = Other, 5 = None
Political Orientation	1 = Conservative, 2 = Liberal, 3 = Middle of the road
Preferred Marriage	1 = Traditional, 2 = Shared
Abortion if Birth Defect	1= Yes, 2 = No, 3 = Don't Know
Abortion if No More Children	1= Yes, 2 = No, 3 = Don't Know
Abortion if Health Risk	1= Yes, 2 = No, 3 = Don't Know
Abortion if Poor	1= Yes, 2 = No, 3 = Don't Know
Abortion if Rape	1= Yes, 2 = No, 3 = Don't Know
Abortion if Not Married	1= Yes, 2 = No, 3 = Don't Know
Abortion For Any Reason	1= Yes, 2 = No, 3 = Don't Know

Appendix 2B: Planning Matrix for Data-entry Exercise

	<i>age</i>	<i>sex</i>	<i>rel</i>	<i>c-l</i>	<i>mg</i>	<i>abd</i>	<i>abn</i>	<i>abh</i>	<i>abp</i>	<i>abr</i>	<i>abs</i>	<i>aba</i>
01	20	1	4	2	2	2	2	1	3	1	2	2
02	24	2	5	2	2	1	1	1	1	1	1	9
03	21	2	2	9	2	2	2	2	2	2	2	2
04	24	2	5	3	2	1	1	1	1	1	1	1
05	26	2	4	2	2	1	1	1	1	1	1	1
06	28	2	2	2	2	2	2	1	2	1	2	2
07	23	1	1	2	2	1	2	1	1	1	2	2
08	22	2	4	3	1	1	1	1	1	1	1	1
09	22	1	5	2	2	1	1	1	1	1	1	1
10	22	2	4	4	2	1	1	1	1	1	1	1
11	23	1	2	2	1	2	2	1	2	1	2	3
12	24	2	2	3	2	1	1	1	1	1	1	2
13	51	2	1	2	9	1	1	1	1	1	1	1
14	22	2	2	3	2	1	1	1	1	1	1	1
15	21	2	4	3	2	1	1	1	1	1	1	1
16	37	1	1	3	2	1	2	1	2	1	2	2
17	22	2	4	2	2	1	1	1	1	1	2	2
18	22	2	3	3	2	1	2	1	2	1	2	2
19	22	2	4	3	2	3	2	1	2	1	1	1
20	30	2	5	2	2	1	1	1	1	1	1	1
21	25	2	5	2	2	1	1	1	1	1	1	1
22	23	1	2	2	2	1	1	1	1	1	1	1
23	21	1	1	2	1	1	1	2	1	2	1	1

Chapter Two Exercises

Exercise 2-1. Clients of Friendly Visitor Service

At California State University, Fresno, the Friendly Visitors Service hires college students to do in-home care for elderly people so they can remain independent and stay in their homes as long as possible. The students do cleaning, yard work, shopping, etc. The staff begins by interviewing clients in their homes and assessing their need for services. The following information is used to match the seniors with the students who want employment:

- Age: Age at Last Birthday
- Sex: Male or Female
- Lives alone: Yes or No
- Low income: Yes = Eligible for Supplemental Security Income (SSI)
- Need for assistance with the activities of daily living (ADL): Bathing, dressing, toileting, transferring in/out of bed, eating
- Total number of ADLs needing help
- Need for assistance with the instrumental activities of daily living (IADL): Using telephone, shopping, preparing food, light housework, heavy housework, finances
- Total Number IADLs needing help

To keep track of the needs of potential clients, the program could create a data file and use it in IBM SPSS. (Data from one month's new applications are provided below. For this example, we'll just use the count of the number of activities for which the seniors need help, but note that they could include the yes/no responses for each of the activities of daily living.)

Sample Data Set: Friendly Visitor Service Clients

<i>id</i>	<i>age</i>	<i>sex</i>	<i>alone</i>	<i>income</i>	<i>#adl</i>	<i>#iadl</i>
001	74	M	N	N	0	4
002	66	M	N	N	4	6
003	81	M	N	N	2	5
004	76	F	N	N	0	4
005	74	M	N	N	1	5
006	69	F	N	Y	0	4
007	79	F	Y	N	0	4
008	80	M	N	Y	3	6
009	89	M	N	N	3	5
010	60	F	Y	N	2	6
011	88	F	Y	N	0	3
012	82	F	Y	N	2	4
013	79	F	Y	N	1	4
014	77	M	N	N	3	6
015	62	M	Y	N	1	4
016	83	M	N	N	4	6
017	80	F	Y	N	0	2
018	85	F	N	N	1	4
019	66	F	Y	N	1	3
020	84	M	N	N	4	6
021	74	F	N	N	4	4
022	74	M	N	N	0	2
023	74	F	Y	N	0	5
024	92	M	N	N	3	6
025	66	F	N	N	2	6

Exercise 2-2. Age at Death from Newspaper Obituaries

An interesting source of data for student practice with data analysis using IBM SPSS is the death notices in local newspapers. Although big city newspapers publish obituaries only for the rich and famous, many local newspapers provide information on almost everyone who dies in the community. (For example, see *The Fresno Bee* at www.fresnobee.com/obituaries, which publishes information provided by funeral homes on most deaths in the community.) Click on the date to get the alphabetized list of people for that day. From these death notices, you could set up a data file with the age and sex of each person who died at a particular time (for example, the first month in the term). The age or birthday is usually given, and you can infer sex from names or pronouns. (The longer, more-detailed obituaries provided by some families would not be a suitable sample for a statistical analysis.) We could use this with IBM SPSS for an analysis of age and sex at death, for example, obtaining frequency and percent distributions, various charts, and descriptive statistics in Chapter 4; cross tabulations in Chapter 5, and/or comparison of means in Chapter 6.

Chapter Three: Transforming Data




This chapter explains how to change, or transform, the values associated with your variables, like the values entered in the data entry process shown in Chapter 2. IBM SPSS can transform the values in several ways. IBM SPSS can:

- Combine values of a variable into several categories
- Create new variables out of old variables
- Select particular cases and analyze only these cases
- Weight cases so that some cases count more heavily than others

IBM SPSS provides many ways to transform data. Covered in this chapter are **recode**, **compute**, and **if**.

Recoding Variables

Recoding is a way of combining the values of a variable into fewer categories. Let me give a hypothetical example. Let's say you have conducted a survey and one of your demographic questions was the age of the respondent. Entering the actual age in years would be the simplest way of working with the data. But let's also say that you want to compare people of different age groupings. In other words, your data would be more useful if it was organized into collapsed categories, like "Young," "Middle age," and "Older." Using IBM SPSS you could reorganize the data so that you had these three groupings. There are two things you need to know before you recode the values. First, you need to decide the number of categories you want to end up with. Generally this will be determined by the way you plan to use the information. If you are going to analyze the data using a table where you crosstabulate two variables (see Chapter 5), you probably want to limit the number of new categories to three or four. The second thing you need to know is which of the old values are going to be combined into a new category. For example, you might do something like this.

The actual age of the respondent as originally recorded in the data file.	The new, collapsed, category.
 18 years through 35 years	Young
 36 years through 55 years	Middle age
 56 years of age or over	Older

Another example might be if respondents were asked how often they prayed, and the original responses were several times a day, once a day, several times a week, once a

week, less than once a week, once a month, once a year, or never. With recode we can combine the people who said “several times a day” with the people who said “once a day” and put all these respondents into a new category which we could call “a lot.” Similarly, we could combine the people who said “several times a week” with those who said “once a week” and call this category “sometimes” and combine those who said “less than once a week” and “once a month” and call this category “infrequently.” Those who said “once a year” or “never” could be combined into a fourth category called “hardly ever.” Recoding is the process in IBM SPSS that will do the above examples.

Starting IBM SPSS the way you were taught in Chapter 1, bring in the GSS10A file, as you did in Chapter 1. Our task is going to be to recode the variable called *age*, which is, of course, the respondent's age.

Click on Transform. Your screen will look like Figure 3-1.



Figure 3-1

Now we have two options: **Recode Into Different Variables** and **Recode Into Same Variables**. It is strongly suggested that the beginning student only use the **Recode Into Different Variables** option. If you make an error, your original variable is still in the file and you can try again. If you make an error using **Recode Into the Same Variables**, you have changed the original variable. If you also saved the file after doing this, and you did not have another copy of the file, you have just eliminated any chance of correcting your error.

Recoding Into Different Variables

The recoding into a different variable starts with giving the new variable a variable name. For example, if we recode into different variables we could combine ages into one set of categories and call this new variable *age1* and then recode ages into a different set of categories called *age2*. To do that, click on **Recode Into Different Variables**. Your screen will look like Figure 3-2.

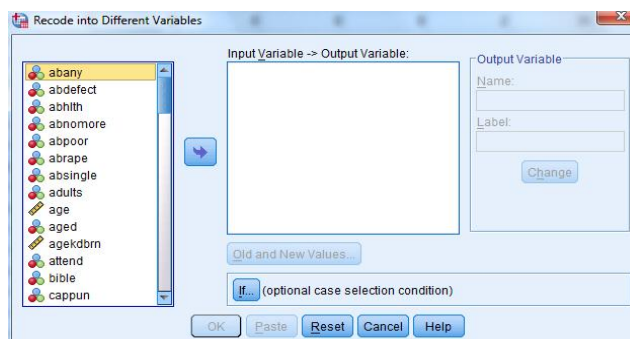


Figure 3-2

Find *age* in the list of variables on the left and click on it to highlight it, and then click on the arrow just to the left of the big box in the middle of the window. This will move *age* into the list of variables to recode.

You want to give a name to this new variable so click in the Name box under Output Variable and type the name **age1** in this box. You can even type a variable label for this new variable in the Label box just below the Name box. Try typing **Age in Four Categories** as your label. Click on the **Change** button to tell IBM SPSS to make these changes. Your screen will look like Figure 3-3.

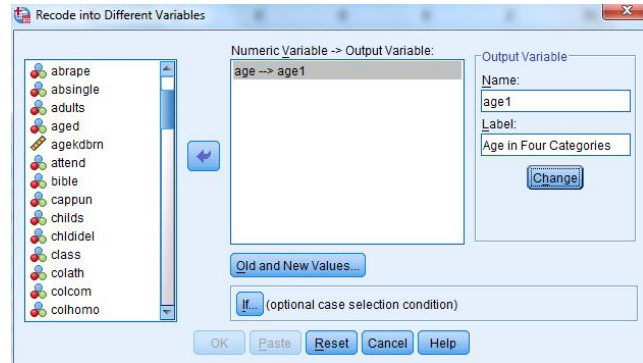


Figure 3-3

Now we have to tell IBM SPSS how to create these categories referred to as values. Click on the **Old and New Values** button at the bottom of the window. The screen will look like Figure 3-4.

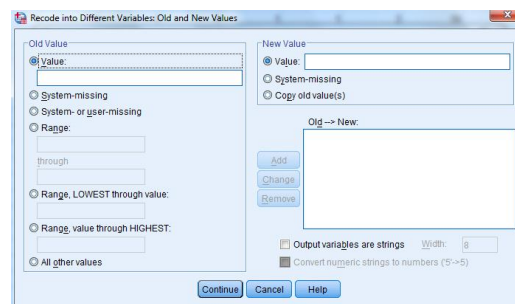


Figure 3-4

There are several options. You can change a particular value into a new value by entering the value to be changed into the Old Value box and the new value into the New Value box and then clicking on **Add**. You will usually change one “real” value to another “real” value. For example, change 18 thru 35 into value 1. (The next paragraph tells you how to do this.) There are also other options⁶.

As you can tell from the previous example, you can also change a range of values into a new value. Click on the fourth bubble from the top labeled **Range**. Notice how this marks this choice by filling in the bubble. Then type **18** (the youngest age in the data set) in the box above **Through**, click on the box below **Through**, and type **29** in that box. Then click on **Value** just below **New Value** and type **1** in that box. This will have IBM SPSS combine all ages from 18 through 29 into a single category and give it the value of 1. Then click on **Add**.

Repeat this process for the other categories. Click on the box under **Range** and type **30** in the box above **Through**, click on the box below **Through**, and type **49** in that box. Click on **Value** just below **New Value** and type **2** in that box and click on **Add**. Do the same thing for the category 50 to 69 (give this a new value of **3**) and the category 70 to 89 (the largest age in the data set). Give this last category a new value of **4**. Your screen should look like Figure 3-5.

⁶ For example, you can work with what SPSS calls “system-missing” values. All blanks will automatically be changed to system-missing values. You can change these system-missing values into another value, or you can change both the system-missing values and the missing values that you define into another value.

To change one of your categories, highlight that category in the Old->New box, make the changes in the Old or New Value box, and then click on **Change**. The new category should appear in the Old->New box. To remove a category, highlight it and click on **Remove**.

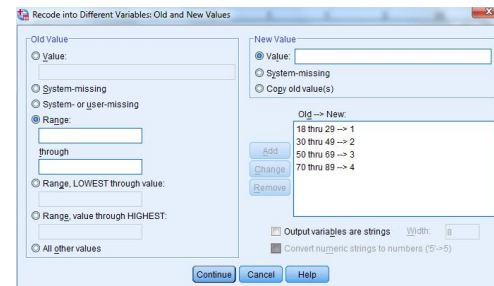


Figure 3-5

Now we want IBM SPSS to carry out the recoding. Click on **Continue** at the bottom of the window. This will take you back to the Record into Different Variables box. Click on **OK** and IBM SPSS will carry out your commands. IBM SPSS will show you the command it just executed. When it says that the IBM SPSS Processor is **Ready** at the bottom of the window you know that IBM SPSS has finished with the recoding.

Statistics

Age in Four Categories

N	Valid	2041
	Missing	3

Age in Four Categories

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	407	19.9	19.9
	2.00	750	36.7	56.7
	3.00	668	32.7	89.5
	4.00	215	10.5	100.0
	Total	2041	99.8	100.0
Missing	System	3	.2	
Total		2044	100.0	

Figure 3-6

Click on **Analyze**, then point your mouse at **Descriptive Statistics**, and then click on **Frequencies**. Notice that *age1*⁷ has appeared in the list of variables on the left. Click on it to highlight it and click on the arrow to move it to the Variables box. Then click on **OK**. An output window will open. Your screen will look like Figure 3-6.

	spkml	spkrac	thnkself	trust	thhours	vote04	vote08	zodiac	wtss	year	age1	var
1	0	0	1	1	0	3	1	9	46	2010	1.00	
2	0	0	2	3	6	1	1	9	93	2010	3.00	
3	2	2	2	3	-1	2	2	12	93	2010	1.00	
4	2	2	0	0	5	1	1	8	46	2010	4.00	
5	0	0	4	3	5	1	1	10	46	2010	4.00	
6	0	0	1	3	0	1	1	3	46	2010	2.00	
7	0	0	1	1	0	1	2	6	46	2010	2.00	
8	1	2	0	0	5	1	1	3	2.53	2010	3.00	
9	0	0	1	1	0	3	3	5	93	2010	2.00	
10	1	2	5	3	-1	2	1	8	93	2010	1.00	
11	2	2	0	0	3	8	1	3	46	2010	2.00	
12	0	0	1	1	0	1	1	2	93	2010	3.00	
13	1	1	1	1	-1	1	1	5	1.68	2010	3.00	
14	2	2	1	1	-1	3	3	5	93	2010	2.00	
15	2	2	0	0	3	1	1	3	84	2010	4.00	
16	2	1	2	1	-1	1	1	1	93	2010	4.00	
17	2	2	0	0	3	1	1	2	46	2010	3.00	
18	0	0	2	1	2	2	1	5	93	2010	2.00	
19	1	1	1	1	-1	1	1	11	46	2010	4.00	
20	1	1	0	0	2	1	1	3	93	2010	3.00	
21	0	0	1	3	5	1	1	7	46	2010	4.00	
22	1	2	0	0	2	1	1	5	93	2010	3.00	
23	1	1	1	1	-1	1	1	10	93	2010	1.00	
24	2	2	0	0	2	3	2	6	1.68	2010	1.00	
25	0	0	9	1	3	1	1	99	93	2010	3.00	
26	2	2	1	3	-1	2	1	11	93	2010	2.00	
27	0	0	1	3	1	1	1	7	46	2010	2.00	
28	0	0	1	1	4	1	1	12	93	2010	3.00	
29	0	0	1	1	2	1	1	6	93	2010	2.00	
30	1	1	1	1	-1	3	2	7	93	2010	1.00	
31	0	0	4	1	2	1	1	8	93	2010	3.00	
32	0	0	3	1	4	1	1	3	93	2010	4.00	
33	1	2	0	0	2	2	1	8	46	2010	1.00	
34	0	0	2	3	3	2	2	4	93	2010	2.00	

Figure 3-7

Let's take a look at the data matrix. Click on **Window** in the menu bar. In the box that is opened you will see a list of all the windows you have opened. One of these windows will be called **GSS10A - IBM SPSS Data Editor**. Click on that line and the data matrix window will be moved to the foreground and you will see it on your screen. Use the scroll bar in the lower-right part of the window to scroll to the right until you see a column titled *age1*. (It will be the last column in the matrix.) This is the new variable you just created. Your screen should look like Figure 3-7.

⁷ If your list shows labels, you can change the display. Check Footnote 1 in Chapter 1.

Use the other scroll bar to scroll down and see the values in this variable. Look back at Figure 3-6 and you will see that there are no value labels for categories 1 through 4 for the new variable *age1*.

If you want the output to give you more information about what each category means, you need to insert value labels. To do this, point your mouse at the variable name at the top of the column (*age1*) and double click. This will open the Variable View tab in the Data Editor. Now you're going to enter labels for the values in the recoded variable using what you learned in Chapter 2.

Click in the Values box and you will see a small gray button in the right-hand side of the box. Point your mouse at this box and click. This will open the Value Labels box. You will see two more boxes, Value and Label. Click in the Value box and type the value **1**.

Then click in the Label box and type the label for the first category, **under 30**. Then click on **Add** and the new label will appear in another box just to the right of the **Add** button. Then click in the Value box and type the value **2** and type the label for the second category, **30 to 49**, and click on **Add**. Do this for values **3** and **4**.

If you make a mistake you can use the **Change** and **Remove** buttons, which work the same way we just described. Your screen should look like Figure 3-8.

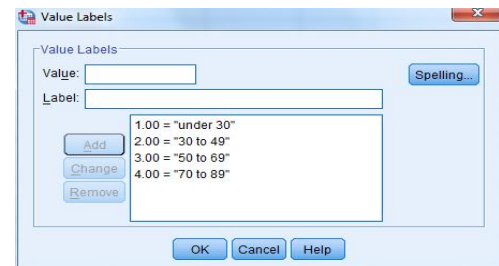


Figure 3-8

Click on **OK**. Now click on **Analyze**, point your mouse at **Descriptive Statistics**, and then click on **Frequencies** and rerun the frequencies distribution for *age1*. This time it should have the value labels you just entered on the output. Your screen should look like Figure 3-9.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	under 30	407	19.9	19.9	19.9
	30 to 49	750	36.7	36.8	56.7
	50 to 69	668	32.7	32.8	89.5
	70 to 89	215	10.5	10.5	100.0
	Total	2041	99.8	100.0	
Missing	System	3	.2		
Total		2044	100.0		

Figure 3-9

We said that recoding into different variables allowed you to recode a variable in more than one way. Let's recode *age* again, but this time let's recode age into three categories—18 through 34, 35 to 59, and 60 and over. Let's call this new variable *age2*. Retracing the steps you used to create *age1*, recode *age* into *age2*.

Be sure to click on **Reset** in the Recode into Different Variables box to get rid of the recoding instructions for *age1*. When you are done, do a frequency distribution for *age2*.

There are two more important points to discuss. Look back at Figure 3-4. It shows the Recode into Different Variables: Old and New Values box. There are three options in the Old Value box that we haven't discussed. Two are different ways of entering ranges.

You can enter the lowest value of the variable through some particular value and you can enter some particular value through the highest value of the variable. Make sure that you do not include your missing values in these ranges or your missing values will become part of that category. For example, if 99 is the missing value for *age*, then recoding 70 through highest would include the missing values with the oldest age category. This is probably not what you want to do. So be careful.

Here is another important point. What happens if you don't recode a particular value? If it is a missing value, it retains its status as a missing value in the new variable. But what if it isn't a missing value? Any value (other than a missing value) that is not recoded is changed into a system-missing value. If you want to leave a value in its original form, then click on All Other Values in the Old Value box and click on Copy Old Value in the New Value box and then click on Add.

Recoding into the Same Variable

Now we are going to recode and have the recoded variable replace the old variable. This means that we will not create a new variable. We will replace the old variable with the recoded variable, but remember the warning given you earlier in this chapter. Click on **Transform** and click on **Recode Into Same Variables**. Let's recode the variable called *pray*. Find *pray* on the list of variables on the left, click on it to highlight it, and then click on the arrow to the left of the Variable box. This will move the variable *pray* into the big box in the middle of the window. Click on the **Old and New Values** button. This will open the Recode into Same Variables: Old and New Values box. Your screen should look like Figure 3-10.

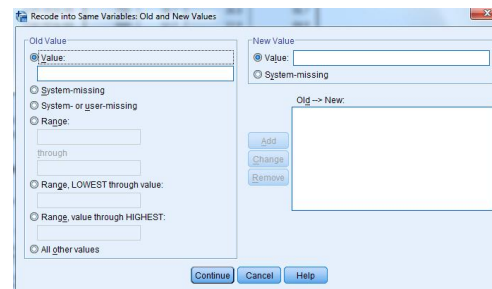


Figure 3-10

This looks very much like the box you just used (see Figure 3-4). Combine the values 1 and 2 by clicking on the fourth circle from the top under **Old Value** and entering a **1** in the box above **through** and a **2** in the box below **through** and then entering a **1** in the New Value box and then clicking on **Add**. Now combine values 3 and 4 into a category called **2**. Then combine values 5 and 6 into a third category called **3**. Since this is not a new variable, it will still be called *pray*.

You will want to change the value labels. Find the variable *pray* in the **Data View** by scrolling to that variable. Point your mouse at the variable name (*pray*) and double click. This will open the **Variable View** tab in the **Data Editor**. Click in the **Values** box and then click on the small gray box and change the labels to an appropriate name for the new variable. You will have to use the **Change** and **Remove** buttons to do this. Follow the instructions we just went through for recoding into different variables.

When you finish, click on **Analyze**, then point your mouse at **Descriptive Statistics**, then click on **Frequencies** and move *pray* over to the Variables box and click on **OK**. Your screen should look like Figure 3-11.

HOW OFTEN DOES R PRAY					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	one or more times a day	1158	56.7	57.2	57.2
	several times a week	381	18.7	18.8	76.1
	less than once a week	484	23.7	23.9	100.0
	Total	2023	99.0	100.0	
Missing	DK	12	.6		
	NA	8	.4		
	Total	21	1.0		
Total		2044	100.0		

Figure 3-11

When you recode into the same variable, a value that is not recoded stays the same as it was in the original variable. If we had decided to keep “never” (value 6) as a separate category, we could have left it alone and it would have stayed a 6. Or we could have changed it to another value such as 4. This is an important difference between recoding into the same and different variables.

Recoding is a very useful procedure and one that you will probably use a lot. It's worth spending time practicing how to recode so you will be able to do it with ease when the time comes.

Creating New Variables Using COMPUTE

You can also create new variables out of old variables using **compute**. There are seven variables in the data set we have been using that ask respondents if they think a woman ought to be able to obtain a legal abortion under various scenarios. These are the variables *abany* (woman wants abortion for any reason), *abdefect* (possibility of serious birth defect in baby), *abhlth* (woman's health is seriously threatened), *abnomore* (woman is married and doesn't want any more children), *abpoor* (woman is poor and can't afford more children), *abrape* (pregnant as result of rape), and *absingle* (woman is not married). Each variable is coded 1 if the respondent says yes (ought to be able to obtain a legal abortion) and 2 if the person says no. The missing values are 0 (not applicable, question wasn't asked), 8 (don't know), and 9 (no answer).

Compute will allow us to combine these seven variables, creating a new variable that we will call *abortion*. If a person said yes to all seven questions, the new variable would equal 7 and if he or she said no to all seven questions, the new variable would equal 14. But what about missing values? If any of the seven variables have a missing value, then the new variable would be assigned a system-missing value.

To use **Compute** to do this, click on **Transform** and then click on **Compute**. Your screen should look like Figure 3-12.

Type the name of the new variable, *abortion*, in the Target Variable box. Then enter the formula for this new variable in the Numeric Expression box. There are two ways to do this. One method is to click on the first of the seven variables, *abany*, in the list of variables on the left, then click on the arrow to the right of this list. This will move *abany* into the Numeric Expression box. Now click on the plus sign and the plus sign moves into the box.

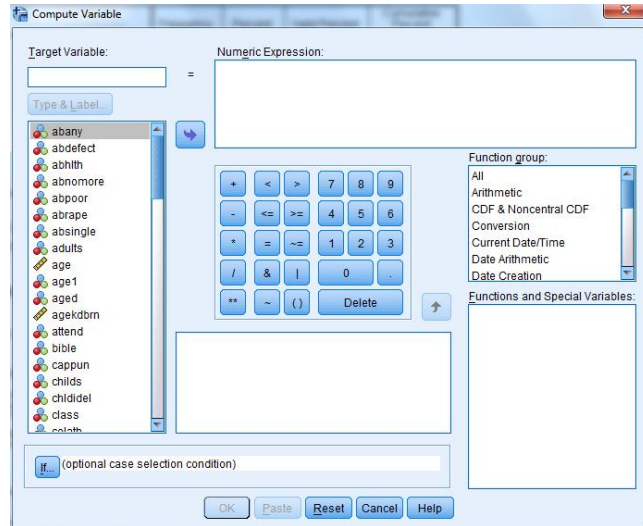


Figure 3-12

Continue doing this until the box contains the following formula: *abany* + *abdefect* + *abhlth* + *abnomore* + *abpoor* + *abrape* + *absingle*. (Don't type the period after *absingle*.) If you make a mistake, just click in the Numeric Expression box and use the arrow keys and the delete and backspace keys to make corrections. Your screen should look like Figure 3-13.

Click on **OK** to indicate that you want IBM SPSS to create this new variable. You can use the scroll bar to scroll to the far right of the matrix and view the variable you just created. A second way to enter the formula in the Numeric Expression box is to click in the box and type the formula directly into the box using the keyboard.

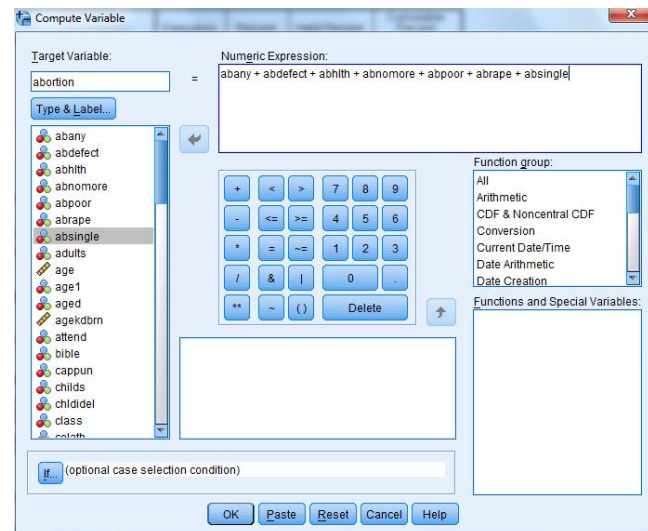


Figure 3-13

You can add variable and value labels to this variable by pointing your mouse at the variable name (*abortion*) at the top of the column in the data matrix and double clicking. This will open the **Variable View** tab in the **Data Editor**. You can enter the variable and value labels the way you were taught earlier in this chapter and in Chapter 2.

Enter the variable label **Sum of Seven Abortion Variables**. Enter the value label **High Approval** for the value 7 and **Low Approval** for the value 14. (Remember that seven means they approved of abortion in all seven scenarios and fourteen means they disapproved all seven times.) Click on **OK**.

You should check your new variable to see that it was calculated correctly. Go to Analyze, then Descriptive Statistics, and then Frequencies. Click on Reset to get rid of what is already in the box. Find the variable *abortion*, highlight it and click on the arrow to the left of the Variables box. Then click on OK. Your

		abortion			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	high approval	409	20.0	36.9	36.9
	8.00	66	3.2	6.0	42.8
	9.00	64	3.1	5.7	48.5
	10.00	79	3.8	7.1	55.6
	11.00	180	8.8	16.2	71.8
	12.00	122	6.0	11.0	82.9
	13.00	86	4.2	7.8	90.6
	low approval	104	5.1	9.4	100.0
	Total	1109	54.2	100.0	
	Missing System	935	45.8		
Total		2044	100.0		

Figure 3-14

screen should look like Figure 3-14. The lowest number should be 7 and the highest number should be 14. Do you remember why?

One of the problems with this approach is that the new variable (*abortion*) will be assigned a system missing value if one or more of the original variables have a missing value. We can avoid this problem by summing the values of the original variable and dividing by the number of variables with valid values. For example, if six of the seven original variables had valid values, then we would divide the sum by six. We can also tell IBM SPSS to create this new variable only if at least four of the original variables have valid values. If fewer than four of the original variables have valid values, IBM SPSS will assign it a system missing value.

We can do this by clicking on Transform and then on Compute and entering the new variable name in the Target Variable box. Let's call this variable *abort*. In the Function Group box, scroll down and click on Statistical. This will list the statistical functions in the Functions and Special Variables box. Double-click on Mean. Your screen should look like Figure 3-15.

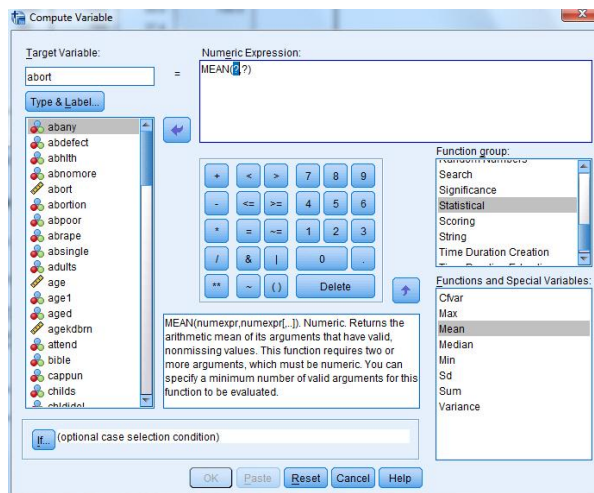


Figure 3-15

Notice that *Mean(?,?)* has been inserted in the Numeric Expression box. What you want to do is to replace the *(?,?)* with the list of the seven original variables. It should now read *(abany, abdefect, abhlth, abnomore, abpoor, abrape, absingle)*. All that is left is to tell IBM SPSS that you want to create this new variable only if at least four of the

original variables have valid values. Do this by entering **.4** following Mean so the expression reads **MEAN .4 (abany, abdefect, abhlth, abnomore, abpoor, abrape, absingle)**. Your screen should look like Figure 3-16.

Click on OK and run a frequency distribution to see what your new variable looks like. Your screen should look like Figure 3-17.

Try creating another variable. Two of the variables in the data set are the number of years of education of the respondent's father (*paeduc*) and of the respondent's mother (*maeduc*). If we divide *paeduc* by *maeduc* we will get the ratio of the father's education to the mother's education. Any value greater than one will mean that the father has more education than the mother and any value less than 1 means the mother has more education than the father. Any value close to 1 means that the father and mother have about the same education.

We have a small problem though. If the mother's education is zero, then we will be dividing by zero, which is mathematically undefined. Let's recode any value of zero for *maeduc* so it becomes a one. This will avoid dividing by zero and still give us a useful ratio of father's to mother's education. Click on Transform, then click on Recode into same variables. (You may need to click on Reset to get rid of the recoding instructions used earlier.) Move *maeduc* into the Variables box by highlighting it in the list of variables on the left and clicking on the arrow to the right of this list. Click on Old and New Values and then type **0** into the Value box under Old Value and click in the Value box under New Value. Type **1** in this box and click on Add. Your screen should look like Figure 3-18.

Click on Continue and then on OK in the Recode Variables box. Now we have changed each 0 for *maeduc* into a 1.

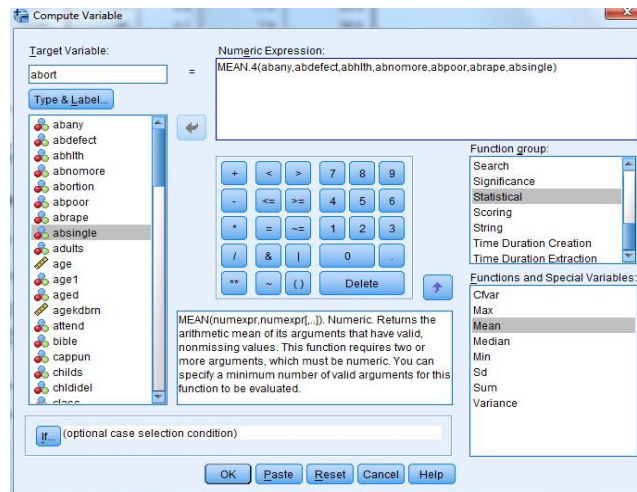


Figure 3-16

abort					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	428	21.0	34.1	34.1
	1.14	66	3.2	5.2	39.3
	1.17	6	.3	.5	39.8
	1.20	1	.1	.1	39.9
	1.25	5	.2	.4	40.3
	1.29	64	3.1	5.1	45.3
	1.33	8	.4	.7	46.0
	1.40	12	.6	.9	46.9
	1.43	79	3.8	6.2	53.2
	1.50	19	.9	1.5	54.7
	1.57	180	8.8	14.3	68.9
	1.60	5	.2	.4	69.3
	1.67	17	.8	1.4	70.7
	1.71	122	6.0	9.7	80.4
	1.75	1	.0	.1	80.5
	1.80	6	.3	.4	80.9
	1.83	19	.9	1.5	82.5
	1.86	86	4.2	6.8	89.3
	2.00	135	6.6	10.7	100.0
	Total	1258	61.6	100.0	
Missing	System	786	38.4		
Total		2044	100.0		

Figure 3-17

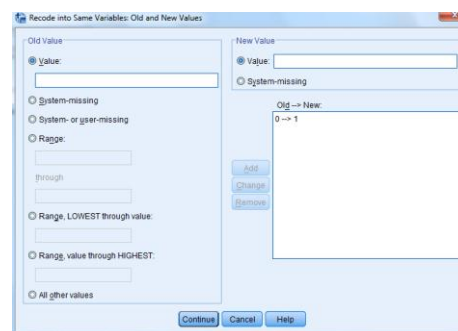


Figure 3-18

To create our new variable, click on **Transform** and then on **Compute**. (If necessary, click on **Reset** to get rid of the formula for the *abort* variable you just created.) Call this new variable *ratio*. So type **ratio** in the Target Variable box. Now we want to write the formula in the Numeric Expression box. Click in the list of variables on the left and scroll down until you see *paeduc*. Click on it to highlight it and click on the arrow to the right of the list to move it into the Numeric Expression box.

IBM SPSS uses the slash (/) to indicate division, so click on the / in the box in the center of the window. Click on the list of variables again and scroll up until you see *maeduc* and click on it to highlight it. Move it to the Numeric Expression box by clicking on the arrow. Your screen should look like Figure 3-19.

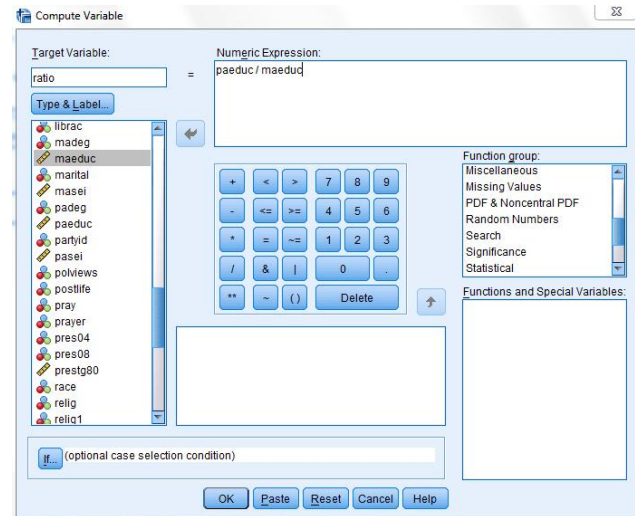


Figure 3-19

Click on **OK** and IBM SPSS will create your new variable. Use the scroll bar to scroll to the right in the data matrix until you can see the new variable you called *ratio*. Scroll up and down so you can see what the values of this variable look like. You may want to do a frequencies distribution as a check to make sure the new variable was created correctly.

After looking at the frequencies distribution, it is obvious that it would be easier to understand if we grouped some of the scores together, so create a new variable by recoding it into a Different Variable. Click on **Transform** and then click on **Recode Into Different Variables**. Find the variable *ratio* in the list of variables on the left and click on it to highlight it. (Again, you may have to click **Reset** if there is old information still in the boxes.) Click on the arrow to the right of this list to move it into the box in the middle of the window. Type **ratio1** in the Name box under Output Variable and type **Recoded Ratio** in the label box. Then click on **Change**.

Click on **Old and New Values** to open the Recode Into Different Variables: Old and New Values box. Click on the fifth bubble from the top under **Old Value** and then type **0.89** in the box to indicate that you want to recode the lowest value through 0.89. Click on the **Value** box under **New Value** and type **1** in that box, and then click on **Add**. Click on the fourth bubble from the top under **Old Value** and type **0.90** in the

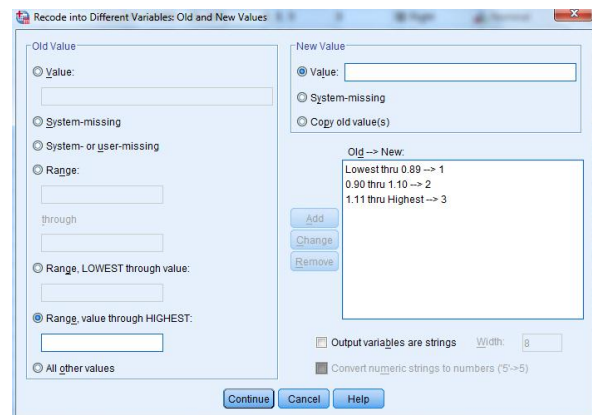


Figure 3-20

box above through and **1.10** in the box below. Then type **2** in the Value box under New Value and click on Add. Finally, click on the sixth bubble from the top under Old Value and type **1.11** in the box. Type **3** in the Value box under New Value and click on Add. Your screen should look like Figure 3-20. Click on Continue and then on OK in the Recode Into Different Variables box.

Let's add value labels to the new values. Find the variable *ratio1* in the data matrix and double click on the variable name, *ratio1*. This will open the Variable View tab in the Data Editor. Click the Values box and then click in the small gray box and enter the labels. Type **1** in the Value box and **under 0.90** in the Value Label box and then click on Add.

Do this twice more to add the label **0.90 through 1.10** to the value 2 and **over 1.10** to the value 3. Your screen should look like Figure 3-21. (This should seem familiar to you now.)

Click on OK in the Value Labels box. Run a frequencies distribution on the new variable to double-check your work. Your screen should look like Figure 3-22.



Figure 3-21

The first category (under 0.90) means that Father's Education was less than 90% of Mother's Education. The second category (0.90 through 1.10) means that Father's and Mother's Education were about the same, while the third category (over 1.10) means that Father's Education was more than 110% of Mother's Education. You can see that about 46% of the respondents have fathers and mothers with similar education, while about 27% have fathers with substantially less education than the mother and another 27% have fathers with substantially more education than the mother.

		ratio1			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	under 0.90	386	18.9	27.1	27.1
	0.90 through 1.10	662	32.4	46.5	73.5
	over 1.10	377	18.4	26.5	100.0
	Total	1425	69.7	100.0	
Missing	System	619	30.3		
	Total	2044	100.0		

Figure 3-22

You have already seen that IBM SPSS uses + for addition and / for division. It also uses - for subtraction, * for multiplication, and ** for exponentiation. There are other arithmetic operators and a large number of functions (e.g., square root) that can be used in compute statements.

Creating New Variables Using IF

The IF command is another way to create new variables out of old variables. Perhaps we want to compare the level of education of each respondent's father to that of his or her mother. Now, however, we're not interested in the precise ratio, but just want to know if the father had more education than the mother, the same amount, or less. We'll create a new variable that will have the value 1 when the father has more education than the mother, 2 when both have the same amount of education and 3 when the mother has more education.

Click on Transform and then click on Compute. (You may need to click on Reset to get rid of the instructions for creating *ratio*.) Type the name of the new variable, *compeduc*, in the Target Variable box. Then click on the Numeric Expression box and enter 1. So far, this is just like what you did in the previous section. Your screen should look like Figure 3-23. This time, however, click on IF.

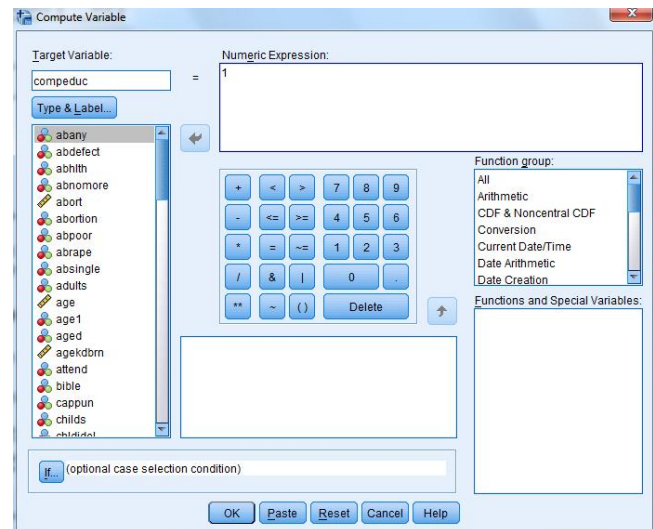


Figure 3-23

Click on Include if case satisfies the condition. Find *paeduc* in the list of variables on the left and click on it to highlight it. Then click on the arrow to the right of this list. This will move *paeduc* into the box to the right of the arrow. Now click on > (greater than). Find *maeduc* in the list of variables on the left, click on it, and click on the arrow to add *maeduc* to the formula. (Alternatively, you could click on the box to the right of the arrow and directly enter the formula, *paeduc* > *maeduc*.) Your screen should look like Figure 3-24. Now click on Continue.

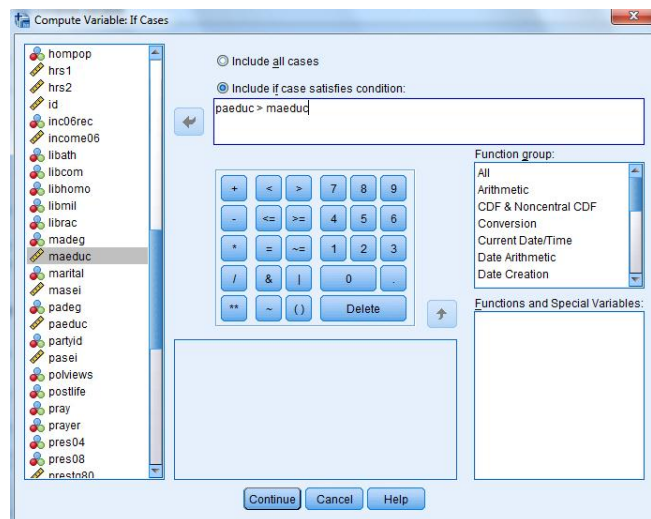


Figure 3-24

Click on OK. Now repeat the same procedures as above, but this time setting the value of *compeduc* to 2 (instead of 1) and the formula to *paeduc* = *maeduc*. When you are asked if you want to Change existing

variable, click on OK. Now repeat the procedures a third time, but change the value of *compeduc* to 3 and the formula to *paeduc < maeduc*.

You can add variable and value labels to this variable, just as you did earlier in this chapter and in Chapter 2. To do this, point your mouse at the variable name at the top of the column (*compeduc*) and double click. This will open the Variable View tab in the Data Editor. Click in the Values box and then in the small gray button in the right-hand side of the box. Point your mouse at this box and click. This will open the Value Labels box. Click in the box next to Value and type 1. Click on the box next to value label (or press the Tab key) and type **Dad More**. Now click on Add. Repeat this procedure for values 2 and 3, labeling them **Same** and **Mom More** respectively. Click on Continue, then on OK. Now run frequencies on your new variable to double-check your work.

Using Select Cases

IBM SPSS can also select subsets of cases for further analysis. One of the variables in the data set is the respondent's religious preference (*relig*). The categories include Protestant (value 1), Catholic (2), Jewish (3), none (4), as well as other categories. The missing values are 98 (don't know) and 99 (no answer). We might want to select only those respondents who have a religious preference for analysis. We can do this by using the Select Cases option in IBM SPSS.

Click on Data and then on Select Cases. This will open the Select Cases box. Your screen should look like Figure 3-25. Notice that All Cases is currently selected. (The circle to the left of All Cases is filled in to indicate that it is selected.) We want to select a subset of these cases so click on the circle to the left of If condition is satisfied to select it. At the bottom of the window it says that unselected cases are filtered out. This means that the cases you do not select can be used later if you click on All cases. If you had selected Delete unselected

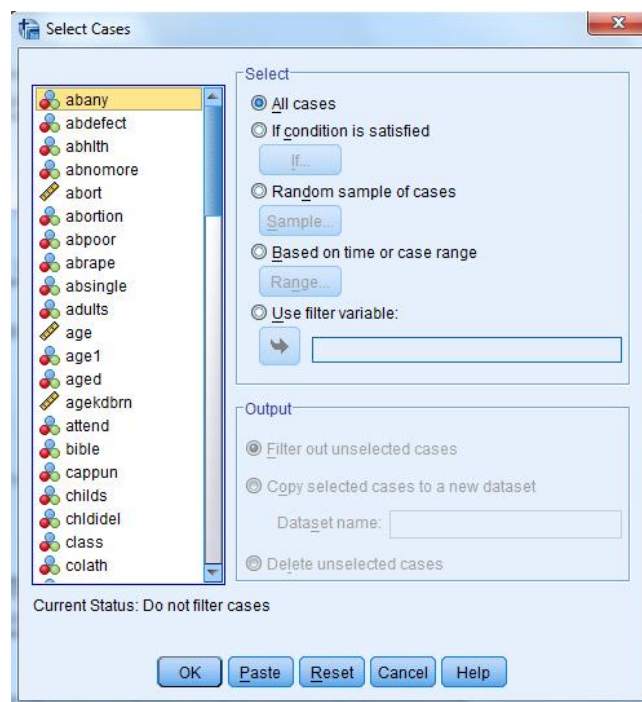


Figure 3-25

cases, these unselected cases could not be used later. You should be very careful about saving a file after you have deleted cases because they are gone forever in that file. (You could, of course, get another copy of the data file by clicking on File and on Open.)

Click on IF (below the button that says If condition is satisfied) and this will open the Select Cases: IF box. Scroll down the list of variables on the left until you come to *relig* and then click on it to highlight it. Click on the arrow to the right of this list to move *relig* into the box in the middle of the window. We want to select all cases that are not equal to 4 so click on the \neq sign. This symbol means “not equal to.” Now click on 4 and the expression in the box will read *relig* \neq 4 which means that the variable *relig* does not equal 4 (the code for no religious preference). Your screen should look like Figure 3-26. Click on Continue and then on OK in the Select Cases box.

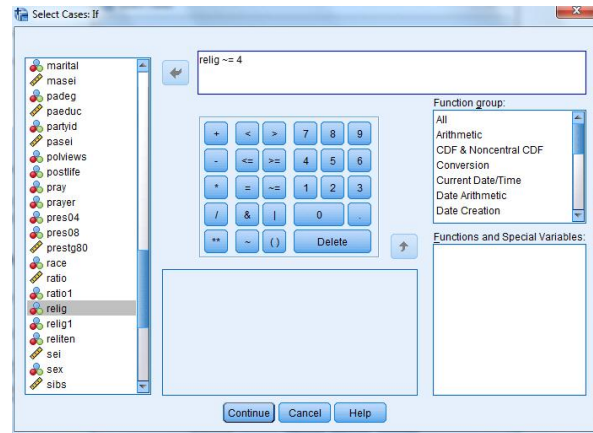


Figure 3-26

Run a frequencies distribution and check that your new variable gives you a reasonable range of values. Your screen should look like Figure 3-27.

RS RELIGIOUS PREFERENCE					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	PROTESTANT	1036	62.1	62.1	62.1
	CATHOLIC	513	30.8	30.8	92.9
	JEWISH	33	2.0	2.0	94.9
	OTHER	85	5.1	5.1	100.0
	Total	1667	100.0	100.0	

Figure 3-27

There are no respondents without a religious preference (value 4) in this table because you selected only those cases with values not equal to four.

What if we wanted to analyze only Protestants and Catholics? Click on Data and then on Select Cases. Click on Reset to eliminate what you had entered previously. Click on IF condition is satisfied and then on IF. Scroll down the list of variables and click on *relig* and then click on the arrow to the right of the list to move it into the box. Click on = and then on 1 so the expression in the box reads *relig* = 1. IBM SPSS uses the symbol & for AND and the symbol | for OR. We want all cases for which *relig* is 1 or 2.

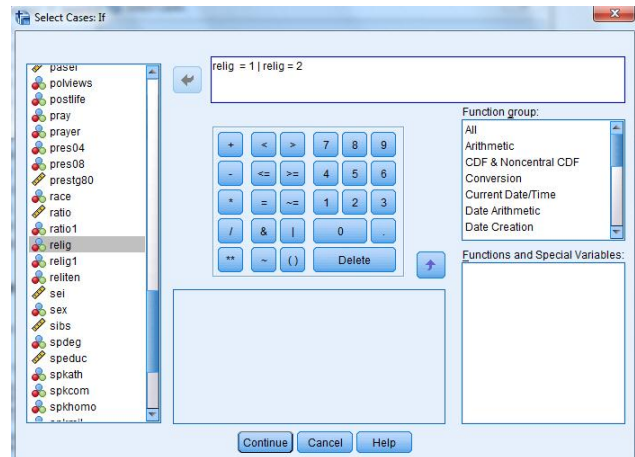


Figure 3-28

Now click on |. Click on *relig* in the list of variables again and on the arrow to move it into the box. Then click on = and then on 2 so the expression in the box reads *relig* = 1 | *relig* = 2 which means that *relig* will equal 1 or 2. Your screen should look like Figure 3-28. Click on Continue and on OK in the Select Cases box.

Run a frequencies distribution on the new variable to see what it looks like. Your screen should look like Figure 3-29. You will only have Protestants (1) and Catholics (2) in your table because you selected only those cases with values one and two on *relig*.

RS RELIGIOUS PREFERENCE					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	PROTESTANT	1036	66.9	66.9	66.9
	CATHOLIC	513	33.1	33.1	100.0
	Total	1549	100.0	100.0	

Figure 3-29

After you have selected cases for analysis, you will probably want to continue your analysis with all the cases. To do this, click on **Data**, then on **Select Cases**, and then click on the circle to the left of **All Cases**. Click on **OK** and IBM SPSS will select all the cases in the data file. This is very important. If you don't do this, you will continue to work with just the cases you have selected. This will work only if you selected filter out unselected cases in the Select Cases box when you began using select cases. If you selected delete unselected cases, you will have to get another copy of the data file by clicking on **File** and then on **Open**.

Conclusion

In this part of the book you have learned how to recode, create new variables using compute and if, and select particular cases for analysis. You can do more complicated things with these commands than we have shown you but these are the basics. In the rest of this book, we will focus on some of the statistical procedures that IBM SPSS can do for you.

Chapter Three Exercises

Use the data set 2010 General Social Survey for all these exercises.

RECODE Exercises

There are two variables that refer to the highest year of school completed by the respondent's mother and father (*maeduc* and *paeduc*). Do a frequency distribution for each of these variables. Now recode each of them (into a different variable) into three categories: under 12 years of school, 12 years, and over 12 years. Create new value labels for the recoded categories. Do a frequency distribution again to make sure that you recoded correctly.

Income06 is the total family income. Do a frequency distribution to see what the variable looks like before recoding. Recode (into a different variable) into eight categories: under \$10,000, \$10,000 to \$19,999, \$20,000 to \$29,999, \$30,000 to \$39,999, \$40,000 to \$49,999, \$50,000 to \$59,999, \$60,000 to \$74,999, and \$75,000 and over. Be very careful that you recode the values, not the labels associated with the values. Call the new variable *income1*. Create new value labels for the recoded categories. Do another frequency distribution to make sure you recoded correctly.

Now recode *income06* again (into a different variable). This time use only four categories: under \$20,000, \$20,000 to \$39,999, \$40,000 to \$59,999, and \$60,000 and over. Call the new variable *income2*. Create new value labels for the recoded categories. Do another frequency distribution to make sure you recoded correctly.

COMPUTE Exercises

In this chapter we created a new variable called *abortion*, which was the sum of the seven abortion variables in the data set. Create a new variable called *ab1*, which is the sum of *abdefect*, *abhlth*, and *abrape*. Do a frequency distribution for this new variable to see what it looks like. How is this distribution different from the distribution for the *abortion* variable based on all seven variables?

There are five variables that measure tolerance for letting someone speak in your community who may have very different views than your own (*spkath*, *spkcom*, *spkhomo*, *spkmil*, and *spkrac*). For each of these variables, 1 means that they would allow such a person to speak and 2 means that they would not allow it. Create a new variable (call it *speak*), which is the sum of these five variables. This new variable would have a range from 5 (would allow a person to speak in each of the five scenarios) to 10 (would not allow a person to speak in any of the five scenarios). Do a frequency distribution for this new variable to see what it looks like.

IF Exercises

There are two variables that describe the highest educational degree of the respondent's father and mother (*pdeg* and *mdeg*). Create a new variable (call it *mapaeduc*) that indicates if the father and mother have a college education. This variable should equal 1 if both parents have a college education, 2 if only the father has a college education, 3 if only the mother has a college education, and 4 if neither parent has a college education. Create new value labels for the recoded categories. Do a frequency distribution for this new variable to see what it looks like.

One variable indicates how often the respondent prays (*pray*) and another variable indicates if the respondent approves or disapproves of the Supreme Court's decision regarding prayer in the public schools (*prayer*). Create a new variable (call it *pry*) that is a combination of these two variables. This variable should equal 1 if the respondent prays a lot (once a day or several times a day) and approves of the Supreme Court's decision, 2 if the respondent prays a lot (once a day or several times a day) and disapproves of the Supreme Court's decision, 3 if the respondent doesn't pray a lot and approves of the Supreme Court's decision, and 4 if the respondent doesn't pray a lot and disapproves of the Supreme Court's decision. Do a frequency distribution for this new variable to see what it looks like.

SELECT IF Exercises

Select all males (1 on the variable *sex*) and do a frequency distribution for the variable *fear* (afraid to walk alone at night in the neighborhood). Then select all females (2 on the variable *sex*) and do a frequency distribution for *fear*. Are males or females more fearful of walking alone at night?

Select all whites (1 on the variable *race*) and do a frequency distribution for the variable *pres04*. Did they vote for Bush, Kerry, or someone else in 2004? Then select all blacks (2 on the variable *race*) and do a frequency distribution for *pres04*. Were whites or blacks more likely to vote for Kerry?

Chapter Four: Univariate Statistics

Univariate analysis, looking at single variables, is typically the first procedure one does when examining data for the first time. There are a number of reasons why it is the first procedure, and most of the reasons we will cover at the end of this chapter, but for now let us just say we are interested in the “basic” results. If we are examining a survey, we are interested in how many people said, “Yes” or “No,” or how many people “Agreed” or “Disagreed” with a statement. We aren't really testing a traditional hypothesis with an independent and dependent variable; we are just looking at the distribution of responses.

The IBM SPSS tools for looking at single variables include the following procedures: Frequencies, Descriptives and Explore, all located under the Analyze menu.

This chapter will use the GSS10A file used in earlier chapters, so start IBM SPSS and bring the file into the Data Editor. (See Chapter 1 if you need to refresh your memory on how to start IBM SPSS.) To begin the process, start IBM SPSS, choose “Open an existing data source” then find and open the GSS10A data file. Under the Analyze menu, choose Descriptive Statistics and the procedure desired: Frequencies, Descriptives, Explore, Crosstabs ...

Frequencies

Generally a frequency is used for looking at detailed information for nominal (category) data that describes the results. Categorical data is for variables such as gender, i.e., males are coded as “1” and females are coded as “2.” Frequencies options include a table showing counts and percentages, statistics including percentile values, central tendency, dispersion and distribution, and charts including bar charts and histograms. The steps for using the frequencies procedure is to click the Analyze menu choose Descriptive Statistics then from the sub menu choose Frequencies and select your variables for analysis. You can then choose statistics options, choose chart options, choose format options, and have IBM SPSS calculate your request.

For this example we are going to check out attitudes on the abortion issue. The 2010 General Social Survey, GSS10A, has the variable *abany* with the label ABORTION-FOR ANY REASON. We will look at this variable for our initial investigation.

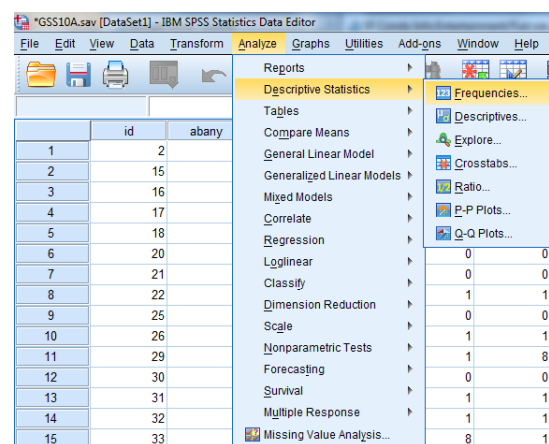


Figure 4-1

Choosing Frequencies Procedure:

From the Analyze menu, highlight Descriptive Statistics, Figure 4-1, then move your mouse across to the sub menu and click on Frequencies.

A Dialog box, Figure 4-2, will appear providing a scrollable list of the variables on the left, a Variable(s) choice box, and buttons for **Statistics**, **Charts** and **Format** options.⁸

Selecting Variables for Analysis:

First select your variable from the main **Frequencies** Dialog box, Figure 4-2, by clicking the variable name on the left side. (Use the scroll bar if you do not see the variable you want.) In this case *abany* is the first variable and will be selected (i.e., highlighted). Thus, you need not click on it.

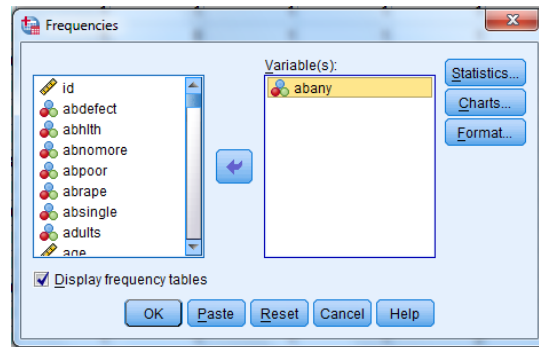


Figure 4-2

Click the arrow on the right of the Variable List box, Figure 4-2, to move *abany* into the Variable(s) box. All variables selected for this box will be included in any procedures you decide to run. We could click OK to obtain a frequency and percentage distribution of the variables. In most cases we would continue and choose one or more statistics.

Choosing Statistics for Variables:

Click the **Statistics** button, right top of Figure 4-2, and a Dialog box of statistical choices will appear, Figure 4-3.

This variable, *abany*, is a nominal (category) variable so click only the Mode box within the central Tendency choices. See Figure 4-3.

After clicking the Mode box, click the **Continue** button, bottom left, and we return to the main **Frequencies** dialog box, Figure 4-2.

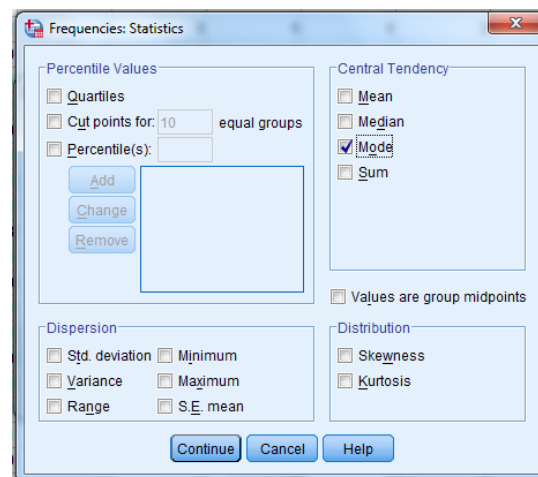


Figure 4-3

We could now click OK and IBM SPSS would calculate and present the frequency and percent distribution (click OK if you want) but, in the more typical manner, we will continue and include choices for charts and check out the **Options** possibilities. If you clicked OK, just press the **Analysis** menu then choose **Descriptive Statistics** and then **Frequencies** from the sub menu and you will be back to this point with your variable and statistics chosen.

⁸ If you want to change the display to labels or know more information about a variable, the label, codes, etc., place the mouse pointer on the variable name in the Variable List, right click the mouse button.

Choosing Charts for Variables:

On the main frequencies window, click the **Charts** button, Figure 4-2, and a Dialog box of chart choices, Figure 4-4, will appear.

Click **Bar Chart**, as I have done, since this is a categorical variable, then click **Continue** to return to the main Frequencies window box. If you have a continuous variable, choose **Histograms** and the **With Normal Curve** option would be available. Choose the **With Normal Curve** option to have a normal curve drawn over the distribution so that you can visually see how close the distribution is to normal. Note: **Frequencies** is automatically chosen for chart values but if desired you could change that to **Percentages**, bottom Figure 4-4.

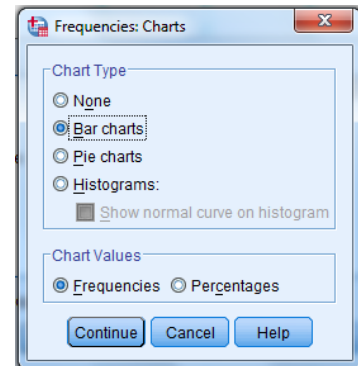


Figure 4-4

Now click **OK** on the main frequencies dialog box and IBM SPSS will calculate and present a frequency and percent distribution with our chosen format, statistics, and chart. (Note: We could look to see if additional choices should be made by clicking the **Format** button. In this case we don't need to do this because all the **Format** defaults are appropriate since we are looking at one variable.)

Looking at Output from Frequencies:

We will now take a brief look at our output from the IBM SPSS frequencies procedure. (Patience, processing time for IBM SPSS to perform the analysis in the steps above will depend on the size of the data set, the amount of work you are asking IBM SPSS to do and the CPU speed of your computer.) The IBM SPSS Output Navigator, left side, and the output, right side, will appear when IBM SPSS has completed its computations. Either scroll down to the chart in the right window, or click the **Bar Chart** icon in the outline pane to the left of the output in Figure 4-5.

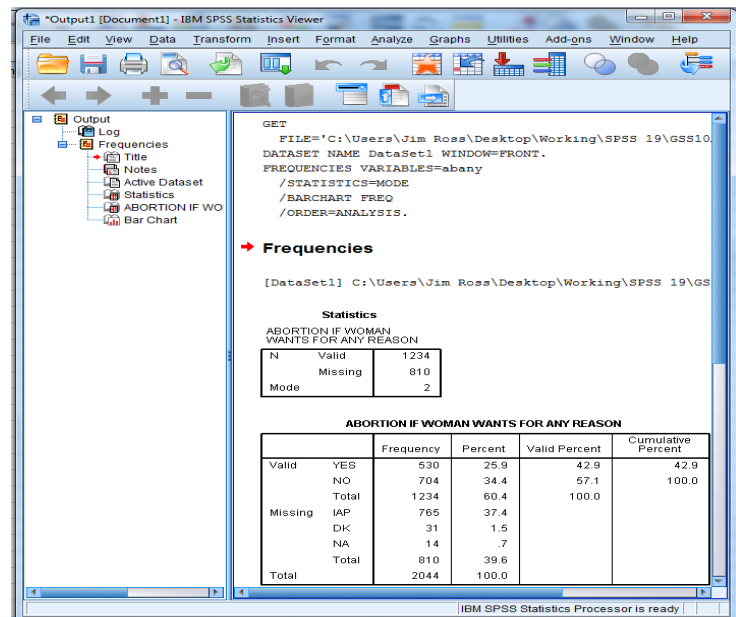


Figure 4-5

Interpreting the Chart:

We now see the chart, Figure 4-6. The graphic is a bar chart with the categories at the bottom, the X axis, and the frequency scale at the left, the Y axis. The variable label **ABORTION IF WOMAN WANTS FOR ANY REASON** is displayed at the top of the chart. We see from the frequency distribution that there are more “no,” 34.4%, answers

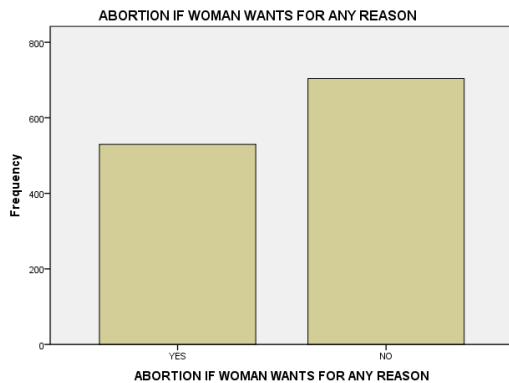


Figure 4-6

ABORTION IF WOMAN WANTS FOR ANY REASON					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	530	25.9	42.9	42.9
	NO	704	34.4	57.1	100.0
	Total	1234	60.4	100.0	
Missing	IAP	765	37.4		
	DK	31	1.5		
	NA	14	.7		
	Total	810	39.6		
Total		2044	100.0		

Figure 4-7

than “yes,” 34.4% answers (see Figure 4-7), when respondents were asked if a woman should be able to get an abortion for any reason. A much smaller number, which does not appear on this chart, 1.5% (see Figure 4-7), selected “don't know,” “DK.” If a chart were the only data presented for this variable in a report, you should look at the frequency output and report the total responses and/or percentages of YES, NO and DK answers. Not asked this question, NA, was 14% of the sample. You should also label the chart with frequencies and/or percentages. There are a lot of possibilities for enhancing this chart within IBM SPSS (Chapter 9 will discuss presentation).

If we choose to copy our chart to a word processor program for a report, first select the chart by clicking the mouse on the bar chart. A box with handles will appear around the chart. Select **Copy** from the **Edit** menu. Start your word processing document, click the mouse where you want the chart to appear then choose **Paste Special** from the down arrow on **Edit**. Choose an option in the paste special dialog box that appears and click **OK** to paste the chart into your document.

Interpreting Frequency Output:

To view the frequency distribution, move the scroll bar on the right of our output window to maximize the size of the table. Another way is to click the **Frequencies** icon in the Outline box to the left of the output window. To view a large table you may want to click on the Maximize Arrow in the upper right corner of the IBM SPSS Output Navigator window to enlarge the output window. Use the scroll bars to display different parts of a large table. The most relevant part of the frequency distribution for *abany* is in Figure 4-7.

We can now see some of the specifics of the IBM SPSS frequencies output for the variable *abany*. At the top is the variable label ABORTION IF WOMEN WANTS FOR ANY REASON. The major part of the display shows the value labels (YES, NO, Total), and the missing categories, IAP (Inapplicable), DK (Don't Know), and NA (Not Answered), Total and the Frequency, Percent, Valid Percent, Cumulative Percent (the cumulative % for values as they increase in size), for each classification of the variable. The "Total" frequency and percent is listed at the bottom of the table. When asked if a woman should be able to have an abortion for any reason, 34.4 % responded no. DK, don't know, was chosen by 1.5 % and .7% were NA [Not Answered]. The 37.4 % "IAP [Inapplicable], was that portion of the sample that was not asked this question. In a written paper, you should state that the "Valid Percent" excludes the "missing" answers.

Variable Names, Variable Labels, Values, Value Labels, Oh My!

Options in Displaying Variables and Values:

It is important to use these concepts correctly so a review at this point is appropriate. A Variable name is the short name you gave to each variable, or question in a survey. The table below is designed to help you keep these separate.

Variable Name	Variable Label	Value	Value Label
<i>SEX</i>	Respondent's gender.	1 or 2	(1) Male, (2) Female
<i>AGE</i>	Respondent's age at last birthday.	18, 19, 20, 21... 89, 98, 99	None needed
<i>AGED</i>	Should aged live with their children.	1, 2, 3, 0, 8, 9	(1) A good idea, (2) Depends, (3) A bad idea (0) IAP [Inapplicable], (8) DK [Don't Know], (9) NA [Not Answered]
<i>BIBLE</i>	Feelings about the bible	1, 2, 3, 4, 0, 8, 9	(1) Word of God, (2) Inspired Word, (3) Book of Fables, (4) Other, (0) IAP, (8) DK, (9) NA

Understanding these allows you to intelligently customize IBM SPSS for Windows so that it is easier for you to use. You can set IBM SPSS so that you can see the variable names when you scroll through a listing of variables, or so that you can see the variable labels as you scroll through the listing. You can set IBM SPSS so that you get only the values, only the labels, or both in the output. Below are two examples of a Frequencies Dialog box.

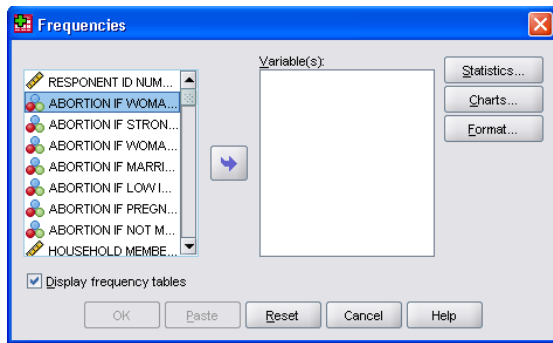


Figure 4-8

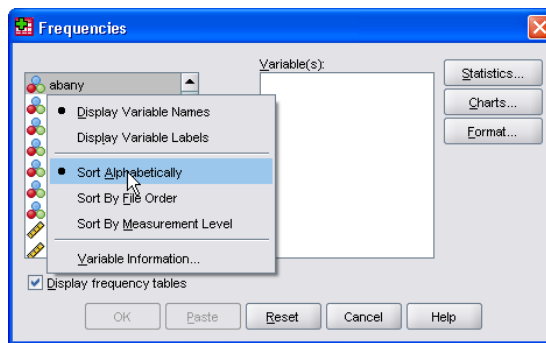


Figure 4-9

Figure 4-8 shows the listing as variable labels. This is the default setting when IBM SPSS for Windows is installed. This example has the cursor on the variable label ABORTION IF WOMAN WANTS FOR ANY REASON (is displayed). You can change the listing however, so that you see only variable names, *abany*, as in Figure 4-9. Changing this is a matter of personal taste. This chapter uses variable names, Figure 4-9.

You can change the display listing when running a procedure by right clicking on the list in the left box of a procedure and choosing a display format, Figure 4-9. For this chapter we choose Display Names and Alphabetical so that variable names will be displayed alphabetically as in Figure 4-9.

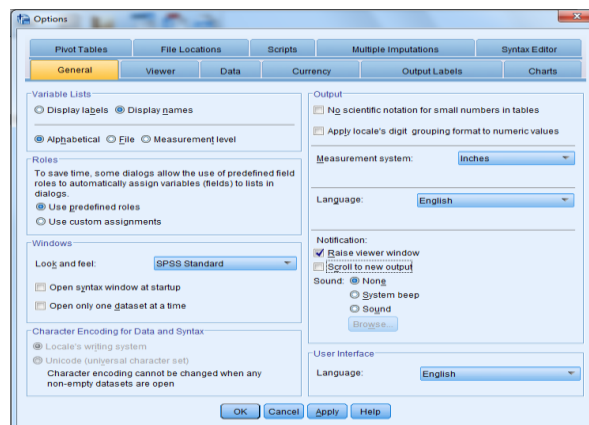


Figure 4-10

Changing the display option for the Variable Selection dialog box, as well as other display formats, can be done for all dialog choices *before* running a procedure. After starting IBM SPSS, to set the display option, click Edit then choose Options. The General tab on the Options dialog box will appear, Figure 4-10. Under Variable Lists section, top right quadrant, click your choices, again we choose Display Names and Alphabetical, then click OK.

Displaying Values, Value Labels or Both in Your Output:

One other option you might want to make is in the table format for your IBM SPSS output. You can choose to have displayed variable labels, values (e.g., 1, 2, 3, etc.), value labels (YES, No, DK, etc.) or both values and labels (1 YES, 2 NO, 3 DK). To make these choices, click the Edit menu and choose Options, then click the Output labels tab, click your choices on the options dialog box. My choices are seen in Figure 4-11. The output resulting from my choices for a Frequencies procedure is Figure 4-12.

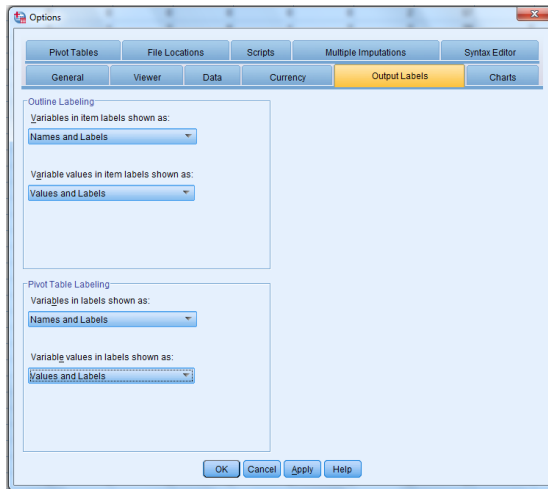


Figure 4-11

abpoor ABORTION IF LOW INCOME--CANT AFFORD MORE CHILDREN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	554	27.1	44.9	44.9
	2 NO	679	33.2	55.1	100.0
	Total	1233	60.3	100.0	
Missing	0 IAP	765	37.4		
	8 DK	34	1.7		
	9 NA	12	.6		
	Total	811	39.7		
Total		2044	100.0		

Figure 4-12

Descriptives

Descriptives (Analyze, Descriptive Statistics, Descriptives, Figure 4-13) is used to obtain summary information about the distribution, variability, and central tendency of continuous variables. Possibilities for Descriptives include mean, sum, standard deviation, variance, range, minimum, maximum, S.E. mean, kurtosis and skewness. For this example we are going to look at the distribution of age and education for the General Social Survey sample. Since both these variables were measured at interval/ratio level, different statistics from our previous example will be used.

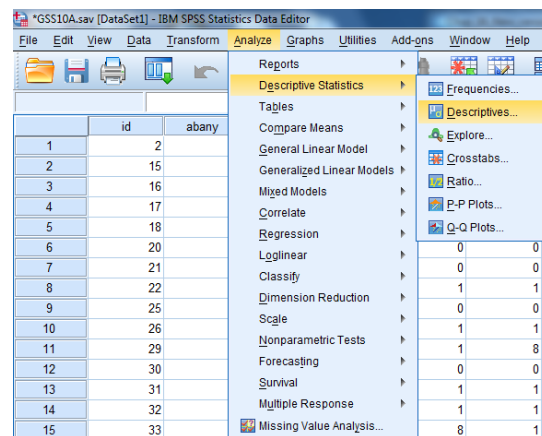


Figure 4-13

Choosing Descriptive Procedure:

First click the Analyze menu and select Descriptive Statistics, then move across to the sub menu and select Descriptives (see Figure 4-13). The Variable Choice dialog box will appear (see Figure 4-14).

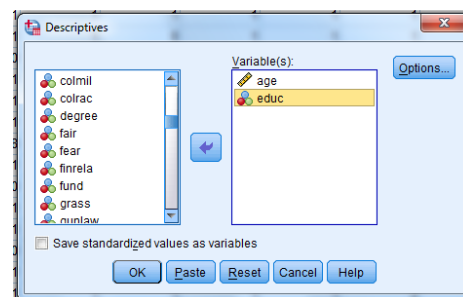


Figure 4-14

Selecting Variables for Analysis:

First click on *age*, the variable name for AGE OF RESPONDENT. Click the select arrow in the middle and IBM SPSS will place *age* in the Variable(s) box. Follow the

same steps to choose *educ*, the variable name for HIGHEST YEAR OF SCHOOL COMPLETED. The dialog box should look like Figure 4-14.

We could click OK and obtain a frequency and percentage distribution, but we will click the Options button and decide on statistics for our output. The Options dialog box, Figure 4-15, will open.

Since these variables are interval/ratio measures, choose: Mean, Std. deviation, Minimum and Maximum. We will leave the defaults for the Distribution and Display Order.

Next, click the Continue button to return to the main Descriptives dialog box, (Figure 4-14). Click OK in the main Descriptives dialog box and IBM SPSS will calculate and display the output seen in Figure 4-16.

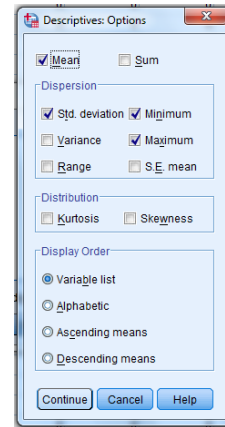


Figure 4-15

Interpretation of the Descriptives Output:

In the Interpretation of Figure 4-16, AGE OF RESPONDENT has a mean of 46.64 and a standard deviation of 17.367. The youngest respondent was 18 and the oldest was 89. Look at your IBM SPSS output for HIGHEST YEAR OF SCHOOL COMPLETED. It has a mean of 13.41 (a little more than 1 year beyond high school) and a standard deviation of 3.107. Some respondents indicated no “0” years of school completed. The most education reported was 20 years.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
age AGE OF RESPONDENT	2041	18	89	46.64	17.367
educ HIGHEST YEAR OF SCHOOL COMPLETED	2039	0	20	13.41	3.107
Valid N (listwise)	2036				

Figure 4-16

Explore

Explore is primarily used to visually examine the central tendency and distributional characteristics of continuous variables. Explore statistics include M-estimators, outliers, and percentiles. Grouped frequency tables and displays, as well as Stem-and-leaf and box-plots, are available. Explore will aid in checking assumptions with Normality plots and Spread vs. Level with the Levene test.

Choosing the Explore Procedure:

From the Analyze menu choose Descriptive Statistics, drag to the sub menu and select Explore.

Selecting Variables:

As in the other procedures, find and click the variable you want to explore, and then click the select arrow to include your variable in the Dependent List box. Choose the variable *educ* and move into the Dependent List box. The dialog box should look like Figure 4-17.

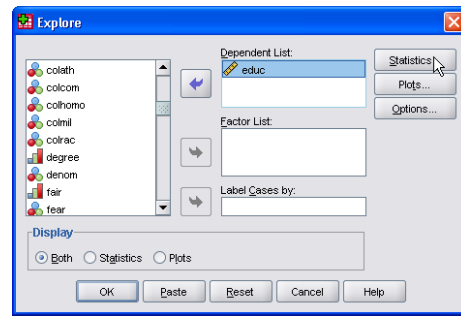


Figure 4-17

Selecting Displays:

In the Display box on the bottom left, you may choose either Both, Statistics, or Plots. We left the default selection, Both, to display statistics and plots.

Selecting Statistics:

Click the Statistics button, bottom middle of Figure 4-17, and the Explore: Statistics dialog box will open, Figure 4-18.

Leave checked the Default box for Confidence Interval for the Mean 95%, and click the Outliers box so we can look at the extreme observations for our variable. Click Continue to return to the main explore dialog window.

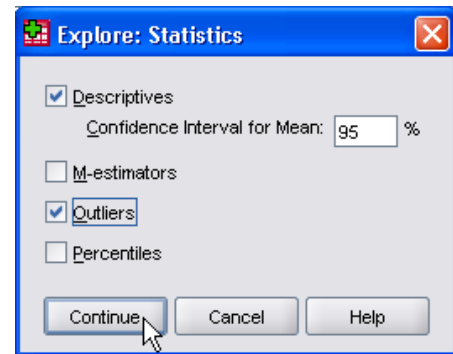


Figure 4-18

Selecting Plots:

Click the Plots button on the main Explore Dialog box, Figure 4-17, and the Explore: Plots dialog box, Figure 4-19, will open.

The default choices in the Boxplots box are good so click Stem-and-leaf and Histogram in the Descriptive box. Click on Normality Plots with Test so we can see how close the distribution of this variable is to normal. Leave the default for Spread vs Level with Levene Test. Click Continue to return to the main explore dialog box, Figure 4-17.

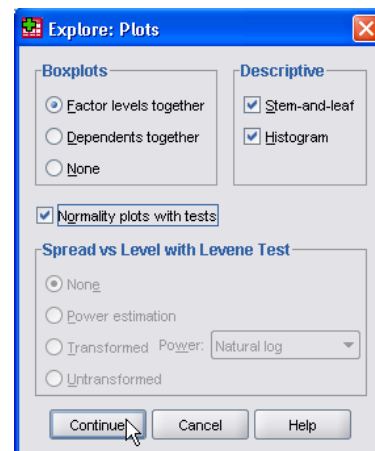


Figure 4-19

Selecting Options:

Click the Options button in the main explore dialog box, Figure 4-17, and the Explore: Options dialog box, Figure 4-20, will be displayed.

No changes are needed here since the default of **Exclude cases listwise** is appropriate. Now click **Continue** to return to the main Explore dialog box, Figure 4-17. Click **OK** in the main Explore dialog box and IBM SPSS will perform the chosen tasks and display the data in the IBM SPSS Output Navigator.

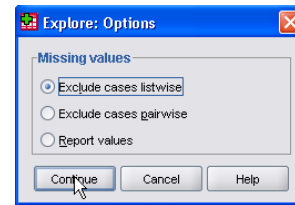


Figure 4-20

Interpretation of Explore Output:

Use the scroll bar to view any part of the output. The first part of the output is the **Case Processing Summary**, Figure 4-21.

	Case Processing Summary					
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
educ HIGHEST YEAR OF SCHOOL COMPLETED	2039	99.8%	5	.2%	2044	100.0%

Figure 4-21

We can see that 2039 (99.8%) of our respondents answered this question. The other ten, .2% of the sample, was **Missing**, not answering the question in this case. The GSS in recent years has had a split sample where not all respondents in the sample are asked the same questions. This is a question where all respondents were asked the question, so the total sample size was 2044 (100%).

The **Descriptives** statistics output should look like Figure 4-22.

We can see all the typical descriptive statistics on this output: mean (13.41), lower bound (13.27) and upper bound (13.54) for a 95% confidence of the mean (in polling terminology this says that we are 95% confident that the mean for the population is between 13.27 and 13.54).

Also shown, the median (13.00), variance (9.654), standard deviation (3.107), minimum (0), maximum (20), range (20), interquartile range (4.00), skewness (-.472), kurtosis (1.492). A narrative explaining the education of first marriage for the US population in 2010 would be somewhat like the following:

Descriptives			Statistic	Std. Error
educ HIGHEST YEAR OF SCHOOL COMPLETED	Mean		13.41	.069
	95% Confidence Interval for Mean	Lower Bound	13.27	
		Upper Bound	13.54	
	5% Trimmed Mean		13.49	
	Median		13.00	
	Variance		9.654	
	Std. Deviation		3.107	
	Minimum		0	
	Maximum		20	
	Range		20	
	Interquartile Range		4	
	Skewness		-.472	.054
	Kurtosis		1.492	.108

Figure 4-22

Our sample from the 2010 General Social Survey indicates that the average education for those over 18 in the US in 2002 was 13.41 years with a 95% confidence that the real

Interpretation of the Q-Q Plot of Age:

Continue scrolling down the IBM SPSS Output Navigator to the Normal Q-Q Plot of HIGHEST YEAR OF SCHOOL COMPLETED (see Figure 4-27).

A Q-Q plot charts observed values against a known distribution, in this case a normal distribution. If our distribution is normal, the plot would have observations distributed closely around the straight line. In Figure 4-27, the expected normal distribution is the straight line and the line of little boxes is the observed values from our data. Our plot shows the distribution deviates somewhat from normality at the low end.

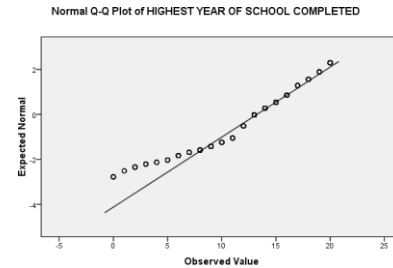


Figure 4-27

The high end of the distribution is pretty much normal.

The Detrended Normal Q-Q plot, shows the differences between the observed and expected values of a normal distribution. If the distribution is normal, the points should cluster in a horizontal band around zero with no pattern. Figure 4-28, of HIGHEST YEAR OF SCHOOL COMPLETED, indicates some deviation from normal especially at the lower end. Our overall conclusion is that this distribution is not normal. Most researchers would see this as close enough to treat as a normal distribution.

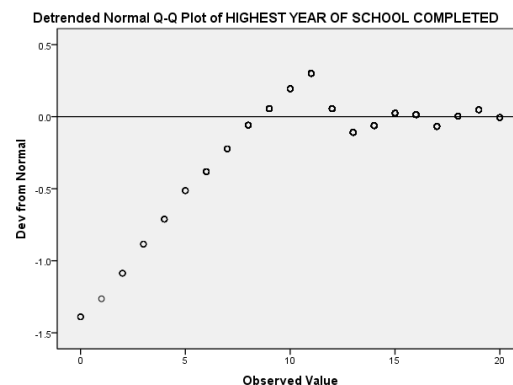


Figure 4-28

Interpretation of the Boxplot:

In the IBM SPSS Output Navigator, scroll to the boxplot of HIGHEST YEAR OF SCHOOL COMPLETED. The boxplot should look like Figure 4-29.

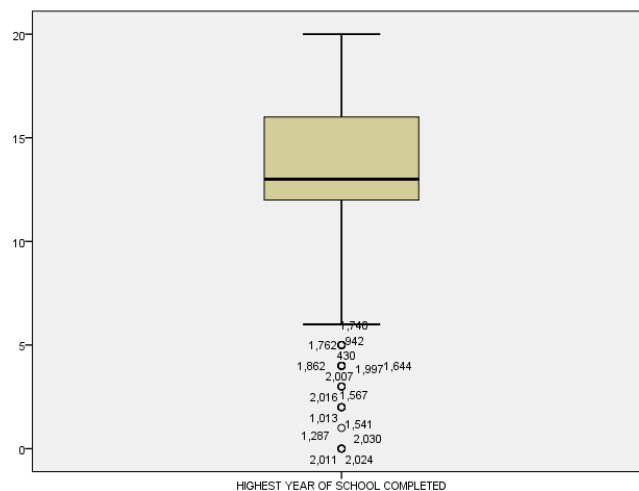


Figure 4-29

Once again the major part of our distribution deviates from normal, there are significant outliers, the cases beyond the lower line of our boxplot. Our outliers are at the lowest end of the distribution, people with little or no education. There are also more observations above than below the mode.

Conclusion

In performing univariate analysis, the level of measurement and the resulting distribution determine appropriate analysis as well as further multivariate analysis with the variables studied. The specific output from IBM SPSS one uses in a report is chosen to clearly display the distribution and central tendencies of the variables analyzed. Sometimes you report a particular output to enable comparison with other studies. In any case, choose the minimal output that best accomplishes this goal. Don't report every IBM SPSS output you obtained.

Univariate Analysis as Your First Step in Analysis

Why do univariate analysis as your first step in data analysis? There are five reasons:

1. As discussed at the beginning of this chapter, the frequency distribution may actually be all you are interested in. You may be doing research for people with little statistical background or they are really only interested in the percentage or count of people that said "Yes" or "No" to some question.
2. You can check for "dirty" data. Dirty data is incorrectly entered data. "Data cleaning" is correcting these errors. Remember, in Chapter 2 you were instructed to give each case an ID number. One primary reason for the ID number is to help us clean our data in case there are data entry or logically inconsistent errors. One way to do this is by determining when there are codes in the data outside the range of the question asked and determining which cases, the ID number, is in error. You can then check all the way back to the original questionnaire and correct the entry or, if that's not possible, change the erroneous code to the "Missing values" code.

An example might be if you had a question in a questionnaire where responses were coded in the following way:

Global warming is a scientific fact.

1. Strongly Agree
2. Agree
3. Neutral
4. Disagree
5. Strongly Disagree

But suppose you run a frequency distribution and find that two respondents have a code of "6." That wasn't one of the codes! What happened? Your data entry person, who may have been you, hit the 6 on the keyboard instead of some other number. We can correct this error. In fact, when we locate this error, we may find others because often errors occur in streaks. The data entry person gets something out of order, or they get their fingers on the wrong keys. These problems can happen to any of us. Our goal is to correct the errors as best possible.

You can have IBM SPSS for Windows select only those cases that have the code of “6” (see Chapter 3) for that variable, and then tell it to do a Frequencies on the variable ID. This will tell you the case numbers that have the error and you can correct it. Be sure to double check the codes, before and after, to make sure they are correct.

3. A third reason for running a **Frequencies** on your variables as your first step in analysis is that you can tell if you need to combine categories and, if so, what codes should be combined. You would know if there were too few respondents giving “Strongly Agree” or “Strongly Disagree” and for analysis they should be folded into either “Agree” or “Disagree.” Another common combination of categories is for age groups. For example, you would do this if you wanted to compare age groups born before and after a significant event (i.e., those born before Vietnam compared to those born after Vietnam).
4. You can also determine if everything that should be defined as “Missing” is actually defined as missing. For example, if you find that 8 “Don’t Know” is a response that has been left in your calculations, your analysis will include all of the eight’s. Even your mean statistics will have these “extra” eight’s included in the calculation. You need to go into the definition of the variable and make these codes “Missing values” or recode these so they are not included, say as a “System Missing” value (Chapter 3).
5. Finally, you may want to examine the distributions for your variables. This should help you determine characteristics of your sample, make some conclusions, and decide further steps in your analysis. You might find that in a 1-5 agree/disagree question, discussed in Step 2 above, almost everyone disagreed. You may discover you do not have a normal distribution and should not use statistics requiring normal distributions. You could also decide that you want to “fix” the distribution using various transformation techniques to convert the data into a normal distribution. These and related techniques are often referred to as “exploratory data analysis” and are beyond the scope of this text.

Chapter Four Exercises

These exercises are designed to familiarize you with the IBM SPSS univariate procedures. They are open-ended with no specific answers.

1. In this chapter we looked at *abany* (ABORTION—FOR ANY REASON), one of the variables in the GSS10A data measuring people’s attitudes about abortion. There are other variables measuring different aspects of the abortion issue. These are:

- *abdefect* (ABORTION--STRONG CHANCE OF SERIOUS DEFECT)
- *abhlth* (ABORTION--WOMAN'S HEALTH ENDANGERED)
- *abnomore*, (ABORTION--MARRIED, WANTS NO MORE CHILDREN),
- *abpoor* (ABORTION--LOW INCOME, CAN'T AFFORD MORE CHILDREN)
- *abrape* (ABORTION--PREGNANT AS RESULT OF RAPE)
- *absingle* (ABORTION--NOT MARRIED)

Pick one of these variables and perform the appropriate techniques discussed in this chapter for the variable. Write up a short narrative explaining what you found about this variable. (Looking back at what we did with *abany* should help you with this. Your write up should be designed to best explain what you found, so do not report all the IBM SPSS output, just that output necessary to clearly and accurately describe your findings.)

2. In this chapter we looked at *educ* (HIGHEST YEAR OF SCHOOL COMPLETED). There are similar variables measuring respondent’s parents’ education:

- *paeduc* (HIGHEST YEAR SCHOOL COMPLETED, FATHER)
- *maeduc* (HIGHEST YEAR SCHOOL COMPLETED, MOTHER)

Pick one of these variables and perform the appropriate techniques discussed in this chapter for describing the variable. Write up a short narrative explaining what you found about this variable. (You might want to look back at what we did with *educ*. Your write up should be designed to best explain what you found so do not report all the IBM SPSS output, just that output necessary to clearly and accurately describe your findings.)

3. The GSS10A file provides answers to a wide range of questions from a sample of respondents in the US in 2010 on their lifestyle and attitudes. Look over the attitude variables in the survey. You can do this by clicking the **Utilities** menu and choosing **Variables**. This will provide a dialog box, which can be used to examine the variable and value labels for our data file. There is also a codebook for this data set in **Appendix A** that lists all the variable information. Pick a couple of interesting attitude questions and use an appropriate IBM SPSS univariate procedure discussed in this chapter to describe the responses for these variables by this sample. Write a narrative description of your IBM SPSS output. (You might want to take another look at what we did in this chapter. Your write up should be designed to best explain

what you found so do not report all the IBM SPSS output, just that output necessary to clearly and accurately describe your findings.)

4. One way to evaluate how close a sample is to the population from which it was drawn is by a comparison of known variables of the population with the same variables in the sample. The 2010 General Social Survey has variables for which we pretty much know the US population distribution (age, race, gender, etc.) from the census. Pick a few of these and find their distribution in our GSS sample. Use the procedures we learned in this chapter. See how close the sample distribution for the variables you choose comes to matching the U.S. population distribution for the same variables. You can find US distributions by checking a library or Internet source for US census data (Statistical Abstracts is one source, <http://factfinder.census.gov/> is a web source). If there is a difference, try and speculate why. Write a short narrative, explaining the differences you found and why you think this difference occurred. Explaining the difference between the sample and the population may be a challenge. You might want to look at the web site for the General Social Survey to determine how the survey was conducted and who was chosen.) The General Social Survey web site is: <http://www.norc.umd.edu/projects/gensoc.asp>.

Chapter Five: Cross Tabulations

In this chapter, we'll look at how IBM SPSS can be used to create contingency tables, sometimes called cross tabulations (or crosstabs), bivariate, or two-variable tables. A contingency table helps us look at whether the value of one variable is associated with, or “contingent” upon, that of another. It is most useful when each variable contains only a few categories. Usually, though not always, such variables will be nominal or ordinal. Some techniques for examining relationships among interval or ratio variables are presented in later chapters.

To make it easier to follow the instructions in this chapter, we recommend that you set certain options in IBM SPSS in the same way that we have. First, click on **Edit** in the menu bar, then on **Options**, and **General**. Under **Variable Lists**, click on **Display names**, and **Alphabetical**. These choices will ensure that the variables in dialog boxes will look like they do in our examples (see Figure 5-1).

Now click on **Pivot Tables** in the tabs. IBM SPSS offers a number of different “looks” for contingency tables. You might want to experiment with the different choices. For now, however, click on **Academic** under **TableLook** (see Figure 5-2). Then click on **OK**.

To illustrate the Crosstabs technique, we'll use the General Social Survey subset (GSS10A). Open this file following the instructions in Chapter 1 under “Getting a Data File.”

Crosstabs are particularly useful for exploring the relationship between variables. We're going to use the question, “Let's see if there is any difference between men and women in their attitudes towards abortion.” To create a contingency table (crosstabs), from the menu, click on **Analyze**, **Descriptive Statistics**, and **Crosstabs**. This will open the dialog box shown in Figure 5-3.

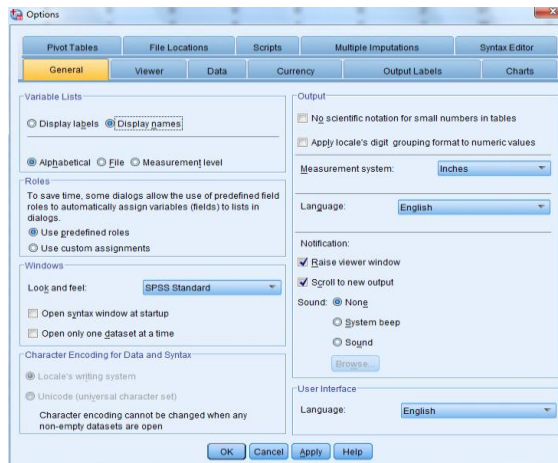


Figure 5-1

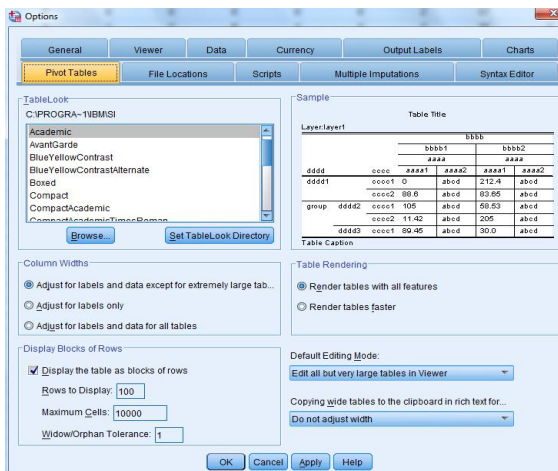


Figure 5-2

You then choose the row (usually the dependent) variable and column (usually the independent) variable.⁹ In Appendix A, you will see that there are seven variables that deal with opinions about abortion. Let's choose *abhlth* (abortion if the woman health is endangered) for our row variable and *sex* (respondent's sex) for the column variable. To do this, select the variable you want from the list and click on it to highlight it, then use the arrow keys to the right of the List box to move the variable into either the Row or the Column box (for now, ignore the Bottom box – more about it in Chapter 8). If you've done everything correctly, your screen will look like Figure 5-4, but don't click OK yet!

In the buttons within the Crosstabs dialog box, click on **Cells**. Here you have a number of choices for the information you would like to have in each cell of your table. The Observed box should already be selected—it shows the actual number of cases in each cell. You will also want to see percentages as well as raw numbers so that you can easily compare groupings of different sizes. You should always make sure that each category of the independent variable totals 100%; our general rule is to have the dependent variables be the rows and the independent variables the columns. So choose **Columns** for the percentages as in Figure 5-5.

Now click on **Continue** to get back to the Crosstabs dialog box. Once you are back there, click **OK**. IBM SPSS will now open the **Output** window, which will show you your table (see Figure 5-6).

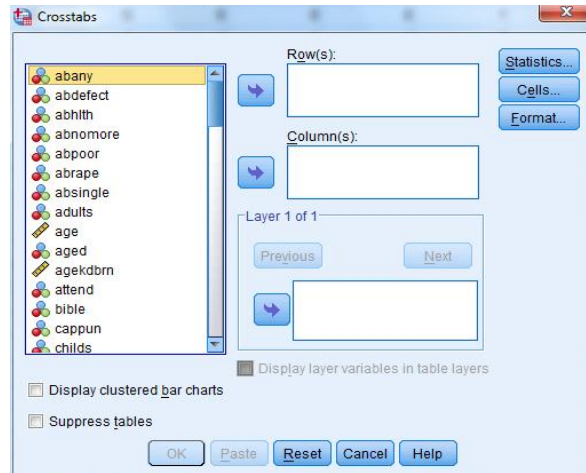


Figure 5-3

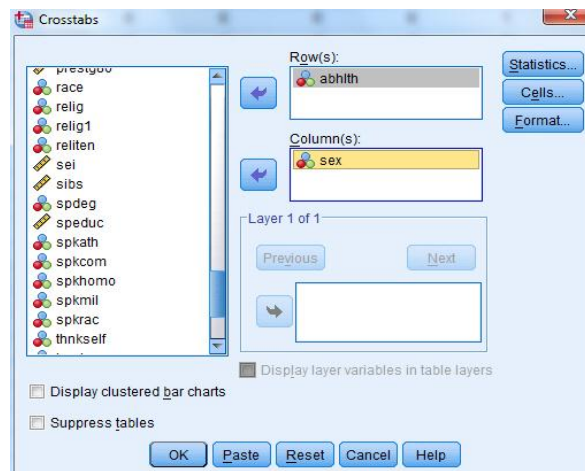


Figure 5-4

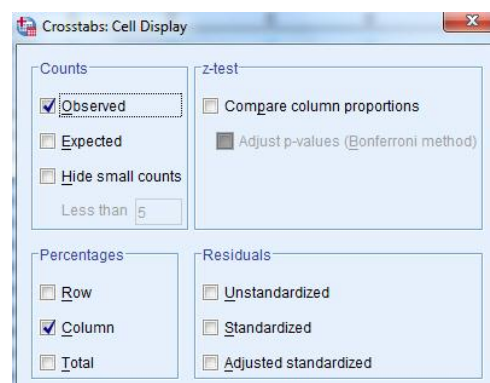


Figure 5-5

⁹ The independent variable is the causal variable.

The Case Processing Summary shows the Valid, Missing, and Total cases. The high percent of missing cases here reflects the people who were not asked this particular question in the survey. Only the valid cases are used in the table.

The Crosstabs shows the 1,229 valid cases arranged in a table that shows what percent of men and women said either Yes or No to the *abhlth* question. Note that 87.1% of the men and 85.6% of the women said Yes, a percentage point difference of only 1.5.

Your initial conclusion here might be that on abortion issues, there's virtually no difference between men and women in their responses. Is this correct or did you stop your analysis a little too soon? Let's look at a different abortion question. Repeat the steps above, but use *abnomore* as your dependent variable this time. Your results should look like Figure 5-7.

Now we see that 50.3% of the men and 45.3% of the women said Yes to "Abortion if a woman is married and wants no more children." When we compare Figure 5-6 with Figure 5-7, we see there is a large difference between total Yes answers (86% compared with 48%), which indicates that abortion as an issue needs to be broken down into specific conditions if you want to study it in depth. We also see that there is now a larger difference between men and women on this particular question. But is it a significant difference? To answer this we will need to do some statistical analysis.

For our next cross tabulation, again go to the menu and choose Analyze, Descriptive Statistics, and Crosstabs. In the Crosstabs dialog box place *abnomore* as the row variable and *sex* as the column variable. Now click on the Statistics button, then Chi-Square to obtain a measure of statistical significance, and on Phi and Cramer's V, which are measures of the strength of association between two variables when one or both are at the nominal level of measurement. Phi is appropriate for tables with two rows and two columns, while Cramer's V is appropriate otherwise. Your dialog box should look like Figure 5-8.

ABORTION IF WOMANS HEALTH SERIOUSLY ENDANGERED * RESPONDENTS SEX Crosstabulation					
			RESPONDENTS SEX		
			MALE	FEMALE	Total
ABORTION IF WOMANS HEALTH SERIOUSLY ENDANGERED	YES	Count	506	555	1061
		% within RESPONDENTS SEX	87.1%	85.6%	86.3%
	NO	Count	75	93	168
		% within RESPONDENTS SEX	12.9%	14.4%	13.7%
Total		Count	581	648	1229
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

Figure 5-6

ABORTION IF MARRIED--WANTS NO MORE CHILDREN * RESPONDENTS SEX Crosstabulation					
			RESPONDENTS SEX		
			MALE	FEMALE	Total
ABORTION IF MARRIED-- WANTS NO MORE CHILDREN	YES	Count	289	294	583
		% within RESPONDENTS SEX	50.3%	45.3%	47.7%
	NO	Count	285	355	640
		% within RESPONDENTS SEX	49.7%	54.7%	52.3%
Total		Count	574	649	1223
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

Figure 5-7

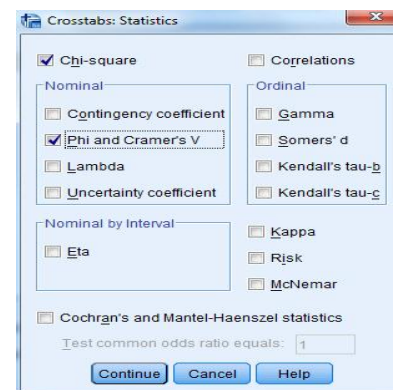


Figure 5-8

Click on **Continue**, then **OK**. The table in Figure 5-7 reappears, but with some additional information (you might have to scroll down to see it)—look for “Chi-Square Tests” (Figure 5-9).

The Pearson Chi Square test indicates that the relationship is not statistically significant. It would occur by chance approximately 78 times out of 1000. The Cramer’s V of .050 in Figure 5-10 (Symmetric Measures) indicates that there is virtually no relationship.

Let’s look at a somewhat different table. We’re going to consider the relationship between social class and political views. Click on **Analyze**, **Descriptive Statistics**, and **Crosstabs**. If the variables you used before are still there, click on the **Reset** button, then move *polviews* to the Row box and *inco6rec* to the Column box. Since both of these variables are ordinal, we’ll want to obtain different statistics to measure their relationship. Click on **Statistics** and then on **Chi-square** and **Kendall’s tau c**. (Tau c is a measure of association that is appropriate when both variables are ordinal and do not have the same number of categories.)

Now click on **Continue** and then on **Cells** and then on **Column percents**. Now click on **Continue** and then click on **OK**. What do the results show? While the Chi-square statistic is statistically significant, the value of Kendall’s tau c is quite low indicating that there is virtually no relationship between these two variables. The pattern to the percents shows the same lack of relationship.

Chi-Square Tests				
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	3.111 ^a	1	.078	
Continuity Correction ^b	2.912	1	.088	
Likelihood Ratio	3.112	1	.078	
Fisher's Exact Test				.085
Linear-by-Linear Association	3.109	1	.078	
N of Valid Cases	1223			.044

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 273.62.
b. Computed only for a 2x2 table

Figure 5-9

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	.050	.078
	Cramer's V	.050	.078
N of Valid Cases		1223	

Figure 5-10

Chapter Five Exercises

1. Suppose we measure class, not by income, but by what people perceive their social class to be (using the variable named *class*). How closely is this measure related to a person's self-identified political views (*polviews*)? Note: before running this crosstab, look at the frequency distribution for *class*. (See Chapter 4 on univariate statistics.) You may want to recode this variable before proceeding. (See Chapter 3 on transforming data.)
2. Consult the codebook in **Appendix A** describing this dataset. Other than income and self-perceived class, what background variables (such as age, marital status, religion, sex, race, or education) might help explain a person's political views? (Here as well, you may need to recode some variables before proceeding.)
3. Is trust related to race? Run crosstabs for *trust* (Can people be trusted?) with *race* and see what you find.
4. Is ideology a general characteristic or is it issue specific? That is, are people who are liberal (or conservative) on one issue (such as capital punishment) also liberal (or conservative) on other issues (such as gun control or legalizing marijuana)?

Chapter Six: Comparing Means

Cross tabulation is a useful way of exploring the relationship between variables that contain only a few categories. For example, we could compare how men and women feel about abortion. Here our dependent variable consists of only two categories—approve or disapprove. But what if we wanted to find out if the average age at birth of first child is younger for women than for men? Here our dependent variable is a continuous variable consisting of many values. We could recode it so that it only had a few categories (e.g., under 20, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 and older), but that would result in the loss of a lot of information. A better way to do this would be to compare the mean age at birth of first child for men and women.

Open the 2010 General Social Survey to answer this question. Click on **Analyze**, point your mouse at **Compare Means**, and then click on **Means**. We want to put age at birth of first child (*agekdbrn*) in the **Dependent List** and *sex* in the **Independent List**. Highlight *agekdbrn* in the list of variables on the left of your screen, and then click on the arrow next to the **Dependent List** box. Now click on the list of variables on the left and use the scroll bar to find the variable *sex*. Click on it to highlight it and then click on the arrow next to the **Independent List** box. Your screen should look like Figure 6-1. Click on **OK** and the Output Window should look like Figure 6-2. On the average, women are a little less than three years younger than men at the birth of first child.

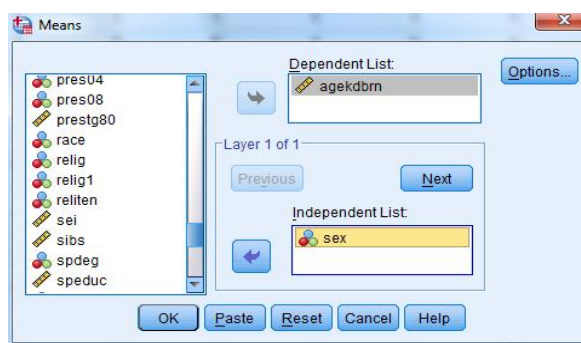


Figure 6-1

Report			
agekdbrn R'S AGE WHEN 1ST CHILD BORN			
sex RESPONDENTS SEX	Mean	N	Std. Deviation
1 MALE	25.60	628	6.516
2 FEMALE	22.96	855	5.413
Total	24.08	1483	6.045

Figure 6-2

Independent-Samples T Test

If women are, on average, a little less than three years younger than men at birth of first child, can we conclude that this is also true in our population? Can we make an inference about the population (all people) from our sample (about 1,500 people selected from the population)? To answer this question, we need to do a t test. This will test the hypothesis that men and women in the population do not differ in terms of their mean age at birth of first child. By the way, this is called a null hypothesis. The particular version of the t test that we will be using is called the independent-samples t test since our two samples are completely independent of each other. In other words, the selection of cases in one of the

samples does not influence the selection of cases in the other sample. We'll look later at a situation where this is not true.

We want to compare our sample of men with our sample of women and then use this information to make an inference about the population. Click on **Analyze**, then point your mouse at **Compare Means** and then click on **Independent-Samples T Test**. Find *agekdbrn* in the list of variables on the left and click on it to highlight it, then click on the arrow to the left of the Test Variable box. This is the variable we want to test so it will go in the Test Variable box.

Now click on the list of variables on the left and use the scroll bar to find the variable *sex*. Click on it to highlight it and then click on the arrow to the left of the Grouping Variable box. *Sex* defines the two groups we want to compare so it will go in the Grouping Variable box. Your screen should look like Figure 6-3.

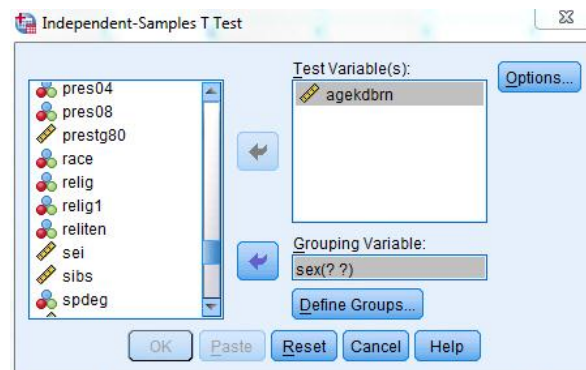


Figure 6-3

Now we want to define the groups so click on the **Define Groups** button. This will open the Define Groups box. Since males are coded 1 and females 2, type **1** in the Group 1 box and **2** in the Group 2 box. (You will have to click in each box before typing the value.) This tells IBM SPSS what the two groups are that we want to compare. (If you don't know how males and females are coded, click on **Utilities** in the **Data Editor** menu bar, then on **Variables** and scroll down until you find the variable *sex* and click on it. The box to the right will tell you the values for males and females. Be sure to close this box.) Now click on **Continue** and on **OK** in the Independent-Samples T Test box. Your screen should look like Figure 6-4.

Group Statistics					
	sex:RESPONDENTS SEX	N	Mean	Std. Deviation	Std. Error Mean
agekdbrn:RS AGE WHEN 1ST CHILD BORN	1 MALE	628	25.60	6.516	.260
	2 FEMALE	855	22.96	5.413	.185

Independent Samples Test									
Levene's Test for Equality of Variances					t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
agekdbrn:RS AGE WHEN 1ST CHILD BORN	Equal variances assumed	19.883	.000	8.486	1481	.000	2.634	.310	2.025 3.243
	Equal variances not assumed			8.251	1197.597	.000	2.634	.319	2.008 3.260

Figure 6-4

This table shows you the mean age at birth of first child for men (25.60) and women (22.96), which is a mean difference of 2.64. It also shows you the results of two t tests. Remember that this tests the null hypothesis that men and women have the same mean age at birth of first child in the population. There are two versions of this test. One assumes that the populations of men and women have equal variances (for *agekdbrn*), while the other doesn't make any assumption about the variances of the populations. The table also gives you the values for the degrees of freedom and the observed significance level. The significance value is .000 for both versions of the t test. Actually, this means less than .0005 since IBM SPSS rounds to the nearest third decimal place. This significance value is the probability that the t value would be this big or bigger simply by

chance if the null hypothesis was true. Since this probability is so small (less than five in 10,000), we will reject the null hypothesis and conclude that there probably is a difference between men and women in terms of average age at birth of first child in the population. Notice that this is a two-tailed significance value. If you wanted the one-tailed significance value, just divide the two-tailed value in half.

Let's work another example. This time we will compare males and females in terms of average years of school completed (*educ*). Click on **Analyze**, point your mouse at **Compare Means**, and click on **Independent-Samples T Test**. Click on **Reset** to get rid of the information you entered previously. Move *educ* into the Test Variable box and *sex* into the Grouping Variable box. Click on **Define Groups** and define males and females as you did before. Click on **Continue** and then on **OK** to get the output window. Your screen should look like Figures 6-5. There isn't much of a difference between men and women in terms of years of school completed. This time we do not reject the null hypothesis since the observed significance level is greater than .05.

Group Statistics				
sex RESPONDENTS SEX	N	Mean	Std. Deviation	Std. Error Mean
1 MALE	831	13.38	3.172	.105
2 FEMALE	1119	13.42	3.054	.091

Independent Samples Test									
Levene's Test for Equality of Variances					t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
educ HIGHEST YEAR OF SCHOOL COMPLETED	1.706	.192	-.252	2037	.801	-.035	.138	-.306	.236
Equal variances assumed									
Equal variances not assumed			-.252	1932.721	.801	-.035	.138	-.307	.237

Figure 6-5

Paired-Samples T Test

We said we would look at an example where the samples are not independent. (IBM SPSS calls these paired samples. Sometimes they are called matched samples.) Let's say we wanted to compare the educational level of the respondent's father and mother. *Paeduc* is the years of school completed by the father and *maeduc* is years of school for the mother. Clearly our samples of fathers and mothers are not independent of each other. If the respondent's father is in one sample, then his or her mother will be in the other sample. One sample determines the other sample. Another example of paired samples is before and after measurements. We might have a person's weight before they started to exercise and their weight after exercising for two months. Since both measures are for the same person, we clearly do not have independent samples. This requires a different type of t test for paired samples.

Click on **Analyze**, then point your mouse at **Compare Means**, and then click on **Paired-Samples T Test**. Scroll down to *maeduc* in the list of variables on the left and click on it and click on the arrow to the left of the paired Variables box to move it to variable 1 in the paired Variables box. Now click on *paeduc* in the list of variables on the left and click on it and click on the arrow to the left of the paired Variables box to move it to variable 2 in the paired Variables box.

Your screen should look like Figure 6-6. Click on OK and your screen should look like Figure 6-7. This table shows the mean years of school completed by mothers (11.71) and by fathers (11.69), as well as the standard deviations. The t-value for the paired-samples t test is 0.238 and the 2-tailed significance value is 0.812. (You may have to scroll down to see these values.) This is the probability of getting a t-value this large or larger just by chance if the null hypothesis is true. Since this probability is greater than .05, we do not reject the null hypothesis.

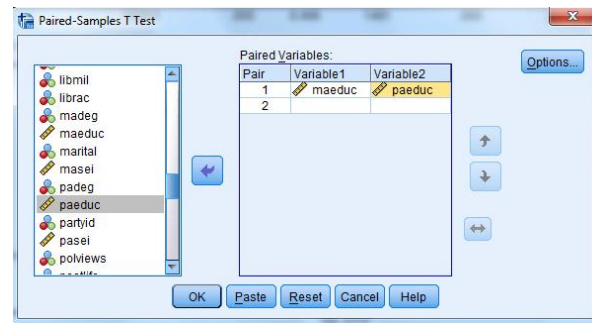


Figure 6-6

Paired Samples Statistics				
	Mean	N	Std. Deviation	Std. Error Mean
Pair 1: maeduc: HIGHEST YEAR SCHOOL COMPLETED, MOTHER	11.71	1423	3.745	.099
paeduc: HIGHEST YEAR SCHOOL COMPLETED, FATHER	11.69	1423	4.175	.111

Paired Samples Test								
Paired Differences								
95% Confidence Interval of the Difference								
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1: maeduc: HIGHEST YEAR SCHOOL COMPLETED, MOTHER - paeduc: HIGHEST YEAR SCHOOL COMPLETED, FATHER	.019	2.970	.079	-.136	.173	.238	1422	.812

Figure 6-7

There is no statistical basis for saying that the respondent's fathers and mothers probably have different educational levels. Notice that if we were using a one-tailed test, then we would divide the two-tailed significance value of .812 by 2 which would be .406. For a one-tailed test, we would also not reject the null hypothesis since the one-tailed significance value is greater than .05.

One-Way Analysis of Variance

In this chapter we have compared two groups (males and females). What if we wanted to compare more than two groups? For example, we might want to see if age at birth of first child (*agekdbrn*) varies by educational level. This time let's use the respondent's highest degree (*degree*) as our measure of education. To do this we will use One-Way Analysis of Variance (often abbreviated ANOVA). Click on Analyze, then point your mouse at Compare Means, and then click on Means. Click on Reset to get rid of what is already in the box. Click on *agekdbrn* to highlight it and then move it to the Dependent List box by clicking on the arrow to the left of the box. Then

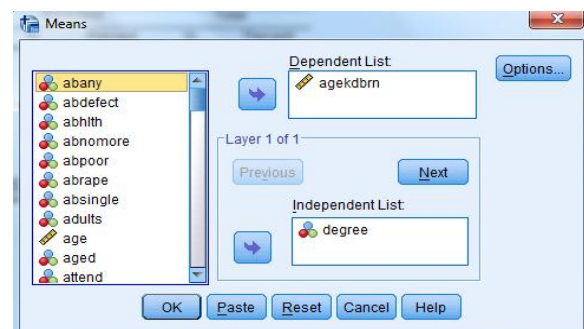


Figure 6-8

scroll down the list of variables on the left and find *degree*. Click on it to highlight it and move it to the Independent List box by clicking on the arrow to the left of this box. Your screen should look like Figure 6-8. Click on the Options button and this will open the Means: Options box. Click in the box labeled Anova table and eta. This should put a check mark in this box indicating that you want IBM SPSS to do a One-Way Analysis of

Variance. Your screen should look like Figure 6-9. Click on **Continue** and then on **OK** in the Means box and your screen should look like Figure 6-10.

In this example, the independent variable has five categories: less than high school, high school, junior college, bachelor, and graduate. Figure 6-10 shows the mean age at birth of first child for each of these groups and their standard deviations, as well as the Analysis of Variance table including the sum of squares, degrees of freedom, mean squares, the F-value and the observed significance value. (You will have to scroll down to see the Analysis of Variance table.) The significance value for this example is the probability of getting a F-value of 79.317 or higher if the null hypothesis is true. Here the null hypothesis is that the mean age at birth

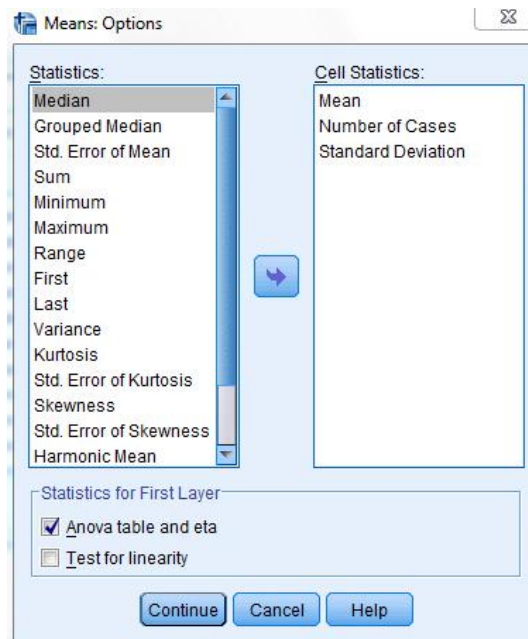


Figure 6-9

of first child is the same for all five-population groups. In other words, that the mean age at birth of first child for all people with less than a high school degree is equal to the mean age for all with a high school degree and all those with a junior college degree and all those with a bachelor's degree and all those with a graduate degree.

Since this probability is so low ($<.0005$ or less than 5 out of 10,000), we would reject the null hypothesis and conclude that these population means are probably not all the same.

Report				
agekdbm R'S AGE WHEN 1ST CHILD BORN				
degree RS HIGHE	Mean	N	Std. Deviation	
0 LT HIGH SCHOOL	21.40	262	5.531	
1 HIGH SCHOOL	22.91	731	5.214	
2 JUNIOR COLLEGE	24.15	112	5.292	
3 BACHELOR	27.38	227	5.787	
4 GRADUATE	29.38	150	6.374	
Total	24.08	1483	6.045	

ANOVA Table						
		Sum of Squares	df	Mean Square	F	Sig.
agekdbm R'S AGE WHEN 1ST CHILD BORN *	Between Groups	9571.358	4	2392.839	79.317	.000
degree RS HIGHEST DEGREE	Within Groups	44575.798	1478	30.168		
	Total	54147.156	1482			

Measures of Association		
	Eta	Eta Squared
agekdbm R'S AGE WHEN 1ST CHILD BORN *	.420	.177
degree RS HIGHEST DEGREE		

Figure 6-10

There is another procedure in IBM SPSS that does One-Way Analysis of Variance and this is called One-Way ANOVA. This procedure allows you to use several multiple comparison procedures that can be used to determine which groups have means that are significantly different.

Conclusion

This chapter has explored ways to compare the means of two or more groups and statistical tests to determine if these means differ significantly. These procedures would be useful if your dependent variable was continuous and your independent variable contained a few categories. The next chapter looks at ways to explore the relationship between pairs of variables that are both continuous.

Chapter Six Exercises

Use the GSS10A data set for all these exercises.

1. Compute the mean age (*age*) of respondents who voted for Obama, McCain, or someone else (*pres08*). Which group had the youngest mean age and which had the oldest mean age?
2. Use the independent-samples t test to compare the mean family income (*income06*) of men and women (*sex*). Which group had the highest mean income? Was the difference statistically significant (i.e., was the significance value less than .05)?
3. Use the independent-samples t test to compare the mean age (*age*) of respondents who believe and do not believe in life after death (*postlife*). Which group had the highest mean age? Was the difference statistically significant (i.e., was the significance value less than .05)?
4. Use the paired-samples t test to compare mother's socioeconomic status (*masei*) and father's socioeconomic status (*pasei*). Who has the highest mean socioeconomic status? Was the difference statistically significant (i.e., was the significance value less than .05)?
5. Use One-Way Analysis of Variance to compare the mean years of school completed (*educ*) respondents who voted for Obama, McCain, or someone else (*pres08*). Which group had the most education and which had the least education? Was the F-value statistically significant (i.e., was the significance value less than .05)?

Chapter Seven: Correlation and Regression

Correlation and regression analysis (also called “least squares” analysis) helps us examine relationships among interval or ratio variables. As you will see, results of these two tests tell us slightly different things about the relationship between two variables. In this chapter, we’ll explore techniques for doing correlation and bivariate regression. Chapter 8 will include a look at multiple regression and correlation.

Correlation

How does education influence the types of occupations that people enter? One way to think about occupations is in terms of “occupational prestige.” Your data set includes a variable, *prestg80*, in which a prestige score was assigned to respondents’ occupations. Higher numbers indicate greater prestige

Let’s hypothesize that as education increases, the level of prestige of one’s occupation also increases. To test this hypothesis, click on **Analyze, Correlate, and Bivariate**. The dialog box shown in Figure 7-1 will appear on your screen. Click on *educ* and then click the arrow to move it into the box. Do the same with *prestg80*.

The most widely used bivariate test is the Pearson correlation. It is intended to be used when both variables are measured at either the interval or ratio level and each variable is normally distributed. However, sometimes we do violate these assumptions. If you do a histogram of both *educ* and *prestg80*, you will notice that neither is actually normally distributed. Furthermore, if you noted that *prestg80* is really an ordinal measure, not an interval, you would be correct. Nevertheless, most analysts would use the Pearson correlation because the variables are close to being normally distributed, the ordinal variable has many ranks, and because the Pearson correlation is the most common. IBM SPSS includes another correlation test, Spearman’s rho, that is designed to

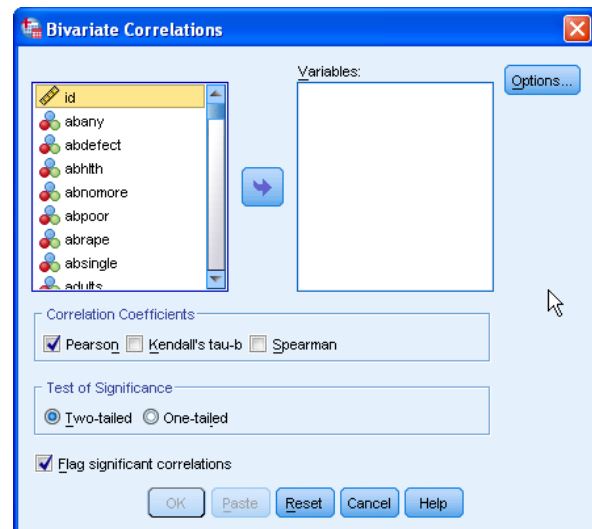


Figure 7-1

analyze variables that are not normally distributed, or are ranked, as is *prestg80*. We will conduct both tests to see if our hypothesis is supported, and also to see how much the results differ depending on the test used—in other words, whether those who use the Pearson correlation on these types of variables are seriously off base.

In the dialog box, the box next to Pearson is already checked, as this is the default. Click in the box next to Spearman. Your dialog box should now look like the one in Figure 7-2. Click OK to run the tests.

Your output screen will show two tables: one for the Pearson correlation and one for the Spearman's rho. The results of the Pearson's correlation, which is called a correlation matrix, should look like the one in Figure 7-3:

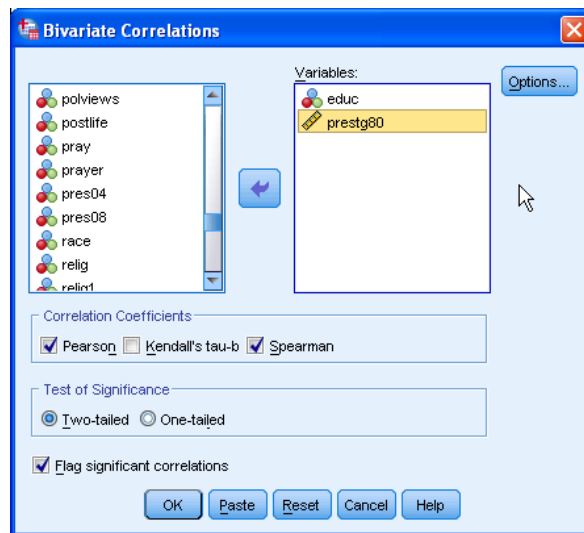


Figure 7-2

The correlation coefficient may range from -1 to 1, where -1 or 1 indicates a “perfect” relationship. The further the coefficient is from 0, regardless of whether it is positive or negative, the stronger the relationship between the two variables. Thus, a coefficient of .453 is exactly as strong as a coefficient of -.453. Positive coefficients tell us there is a direct relationship: when one variable increases, the other increases. Negative coefficients tell us that there is an inverse relationship: when one variable increases, the other one decreases. Notice that the Pearson coefficient for the relationship between education and occupational prestige is .511 and it is positive. This tells us that, just as we predicted, as education increases, occupational prestige increases. But should we consider the relationship strong? At .511 the coefficient is only about half as large as is possible. It should not surprise us, however, that the relationship is not “perfect” (a coefficient of 1).

Education appears to be an important predictor of occupational prestige, but no doubt you can think of other reasons why people might enter a particular occupation. For example, someone with a college degree may decide that they really wanted to be a cheese-maker, which has an occupational prestige score of only 29, while a high-school dropout may one day become an owner of a bowling alley, which has a prestige score of 44. Given the variety of factors that may influence one's occupational choice, a coefficient of .513 suggests that the relationship between education and occupational prestige is actually quite strong.

Correlations				
			educ HIGHEST YEAR OF SCHOOL COMPLETED	prestg80 RS OCCUPATIO NAL PRESTIGE SCORE (1980)
Spearman's rho	educ HIGHEST YEAR OF SCHOOL COMPLETED	Correlation Coefficient	1.000	.511**
		Sig. (2-tailed)	.	.000
		N	1887	1743
	prestg80 RS OCCUPATIONAL PRESTIGE SCORE (1980)	Correlation Coefficient	.511**	1.000
		Sig. (2-tailed)	.000	.
		N	1743	1745

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

Figure 7-3

The correlation matrix also gives the probability of being wrong if we assume that the relationship we find in our sample accurately reflects the relationship between education and occupational prestige that exists in the total population from which the sample was drawn (labeled as Sig. [2-tailed]). The probability value is .000 (remember that the value is rounded to three digits), which is well below the conventional threshold of $p \leq .05$. Thus, our hypothesis is supported. There is a relationship (the coefficient is not 0), it is in the predicted direction (positive), and we can generalize the results to the population ($p \leq .05$).

Recall that we had some concerns about using the Pearson coefficient, given that *prestg80* is measured as an ordinal variable. Figure 7-4 shows the results using Spearman's rho. Notice that the coefficient is .511! What do you conclude?

Correlations				
			educ HIGHEST YEAR OF SCHOOL COMPLETED	prestg80 RS OCCUPATIO NAL PRESTIGE SCORE (1980)
Spearman's rho	educ HIGHEST YEAR OF SCHOOL COMPLETED	Correlation Coefficient	1.000	.511**
		Sig. (2-tailed)	.	.000
		N	1887	1743
	prestg80 RS OCCUPATIONAL PRESTIGE SCORE (1980)	Correlation Coefficient	.511**	1.000
		Sig. (2-tailed)	.000	.
		N	1743	1745

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

Figure 7-4

Regression

We can also analyze the relationship between education and occupational prestige using regression analysis. But first, let's look at the relationship graphically by creating a scatterplot. Click on Graphs, Chart Builder. This will open up the dialog box shown in Figure 7-5. You will have to do three things to make the box on your screen look like the one in Figure 7-5. First, in the "Choose from," list at the lower left, click on Scatter/Dot. Then, shift your attention to the sample graph patterns, and click on the first one (upper left; see arrow). Holding down the mouse button, drag the sample to the large chart preview window.

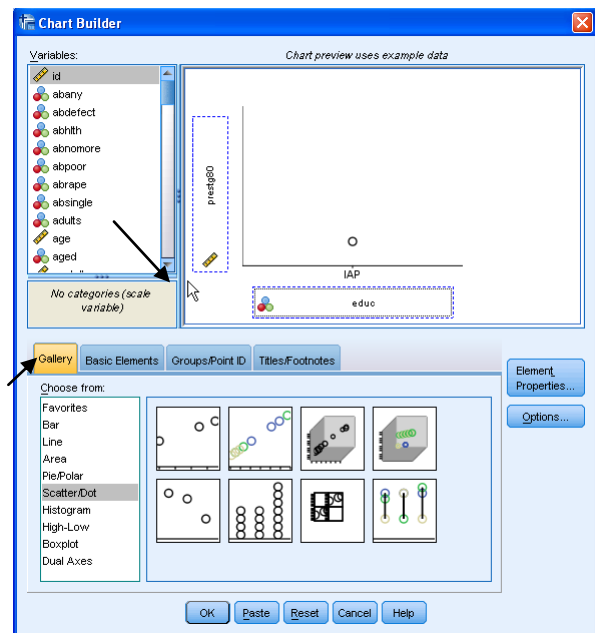


Figure 7-5

Finally, add the variables to the chart preview window. From the list of variables, click on *educ* and drag it to the box located on the horizontal axis (the X Axis, because it is the independent variable in our hypothesis). Next, click on *prestg80* and drag it into the box located on the vertical axis (the Y Axis, because it is the dependent variable). Your dialog box should now look like the one in Figure 7-5. Then, click OK.

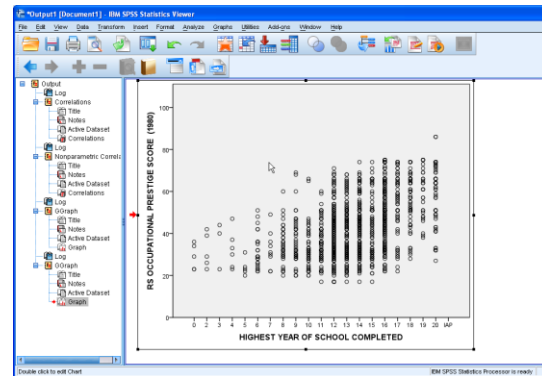


Figure 7-6

What you see is a plot of the number of years of education by the occupational prestige score for persons in the data set who have a job. Your scatterplot should look like the one in Figure 7-6. You can edit your graph to make it easier to interpret. First, double-click anywhere in the graph. This will cause the graph to open in its own window. Double-click on the Y in the menu bar. A dialog box will open. Click the **Scale** tab at the top of the box. Then, click the check box next to **Major Increment**. You will see that the box to the right, in the Custom column, will no longer be shaded.

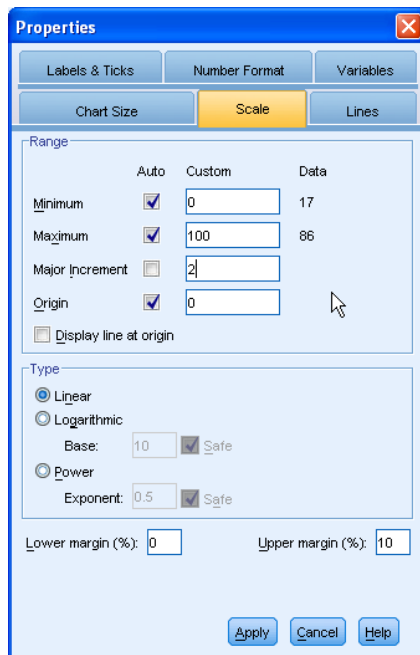


Figure 7-7

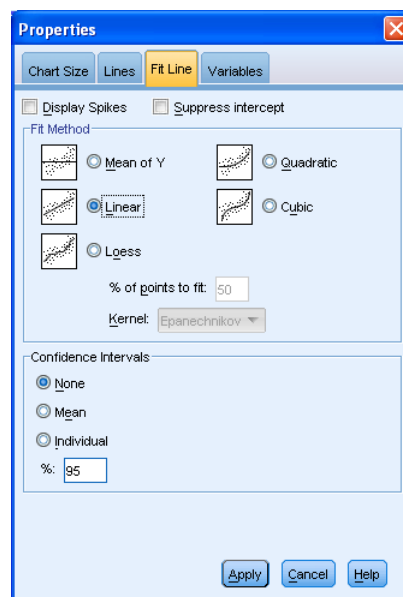


Figure 7-8

Enter the number **2**. When you finish, it should look like the one in Figure 7-7. Then, click on **Apply** and then **Close**.

Now, on the menu bar, click on **Elements**, then **Fit Line at Total**. You will get a dialog box that looks like the one in Figure 7-8. In the **Fit Method** section, click on **Linear** (it is the default) and then click on **Apply** and close the box.

Your graph now looks like the one in Figure 7-9. Notice the Fit Line that is now drawn on the graph. Regression (and correlation) analyze linear relationships between variables, finding the line that “best fits” the data (i.e., it keeps the errors, the distances of each point from the line, to a minimum). The Fit Line shows you the line that describes the linear relationship. Also notice the R-square statistic listed to the right of the graph. Multiplied by 100, this statistic tells us the percentage of the variation in the dependent variable (*prestg80*, on the Y-axis) that is explained by the scores on the independent variable (*educ*, on the X-axis). Thus, years of education predicts 26.1% of the variation in occupational prestige in our sample. Recall that the Pearson coefficient was .511. If you square the Pearson coefficient (.511 x .511), you get .261—the same as the R-square.

Thus, by knowing the correlation coefficient, you can also know the amount of variance in one variable (dependent) that is explained by the other variable (independent) in a bivariate analysis.

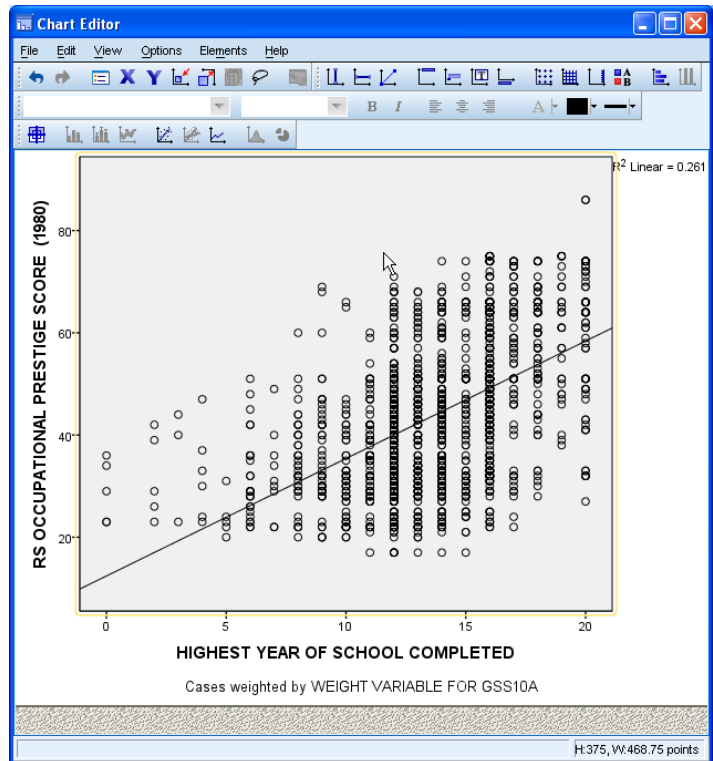


Figure 7-9

Doing a regression analysis can help us to understand the Fit Line in more detail. Close the IBM SPSS Chart Editor. Click on Analyze, Regression, and Linear. This opens up the dialog box shown in Figure 7-10. Move *prestg80* to the Dependent box, and *educ* to the Independent(s) box. Click OK. The results should look like those shown in Figure 7-11.

The first table just shows the variables that have been included in the analysis. The second table, “Model Summary,” shows the R-square statistic, which is .261. Where have you seen this before? What does it mean?

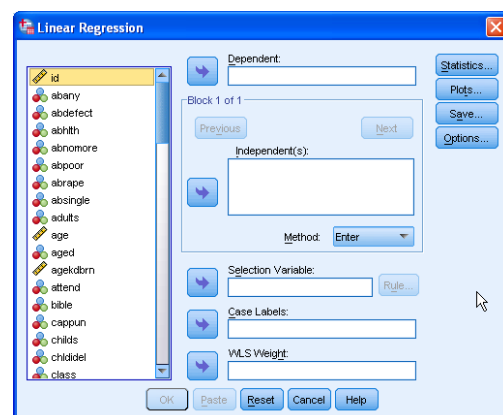


Figure 7-10

The third table, ANOVA, gives you information about the model as a whole. ANOVA is discussed briefly in Chapter 6. The final table, Coefficients, gives results of the

regression analysis that are not available using only correlation techniques. Look at the “Unstandardized Coefficients” column. Two statistics are reported: B, which is the regression coefficient, and the standard error. Notice that there are two statistics reported under B: one labeled as (Constant), the other labeled as EDUC. The statistic labeled as EDUC is the regression coefficient, which is the slope of the line that you saw on the scatterplot (note that in scholarly reports, it is conventional to refer to the regression coefficient using the lower case, b). The one labeled as (Constant) is not actually a regression coefficient, but is the Y-intercept (IBM SPSS reports it in this column for convenience only).

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	educ HIGHEST YEAR OF SCHOOL COMPLETED	.	Enter

a. All requested variables entered.

b. Dependent Variable: prestg80 RS
OCCUPATIONAL PRESTIGE SCORE (1980)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.511 ^a	.261	.261	11.834

a. Predictors: (Constant), educ HIGHEST YEAR OF SCHOOL COMPLETED

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	92961.277	1	92961.277	663.840	.000 ^a
	Residual	263044.327	1878	140.036		
	Total	356005.604	1879			

a. Predictors: (Constant), educ HIGHEST YEAR OF SCHOOL COMPLETED

b. Dependent Variable: prestg80 RS OCCUPATIONAL PRESTIGE SCORE (1980)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	12.423	1.236		10.053	.000
	educ HIGHEST YEAR OF SCHOOL COMPLETED	2.298	.089	.511	25.765	.000

a. Dependent Variable: prestg80 RS OCCUPATIONAL PRESTIGE SCORE (1980)

Figure 7-11

What do these numbers mean? You may recall from your statistics course that the formula for a line is:

$$Y = a + bX$$

Y refers to the value of the dependent variable for a given case, a is the Y-intercept (the point where the line crosses the Y-axis, listed as Constant on your output), b is the slope of the line which describes the relationship between the independent and dependent variables (B for *educ*), and X is the value of the independent variable for a given case.

We know that the linear relationship between X and Y (*educ* and *prestg80*) is not perfect. The correlation coefficient was not 1 (or -1), and the scatterplot showed plenty of cases that did not fall directly on the line. Thus, it is clear to us that knowing someone's education will not tell us without fail what their occupational prestige is, and furthermore, we are only analyzing a sample of cases and not the whole population to which we want to generalize our findings. It is clear that there is some error built into our findings. This is the reason that the Fit Line is usually called the "Best Fit Line." For these reasons, it is conventional to write the formula for the line as

$$\hat{Y} = a + bX + e, \text{ where } e \text{ refers to error.}$$

What can we do with this formula? One thing we can do is make predictions about particular values of the independent variable, using just a little arithmetic. All we have to do is plug the values from our output into the formula for a line (for our purposes, we will ignore the "e"):

$$\hat{Y} = 12.423 + 2.298X$$

12.423, the Y-intercept (or Constant), is interpreted as the average occupational prestige score (our dependent, or Y variable), holding constant the effects of education (our independent, or X variable). 2.298 is the slope of the line. That is, if you refer back to the scatterplot, if pick any point on the fit line, then you move one unit to the right on the X-axis, then move 2.298 units upward, you will intersect with the regression line (fit line). (It is possible to have a negative coefficient. In that case, to intersect with the line, you would move one unit to the right, and then B units downward.)

What occupational prestige score would our results predict for a person who completed high school, but no higher education? Since X refers to the value of the independent variable, and *educ* is our independent variable, all we have to do is enter 12 (as in twelve years of education) into our equation as follows:

$$\begin{aligned}\hat{Y} &= 12.423 + 2.298(12) \\ \hat{Y} &= 39.999\end{aligned}$$

We find that having 12 years of education is associated with an occupational prestige score of 39.999. But what about the error? We know that not every high school graduate

has this exact prestige score. We acknowledge this when we discuss results by stating that, *on average*, those with 12 years of education will have occupations with prestige scores of 39.999. This language points out to our readers that it is likely that some of those respondents scored higher and some lower, but that 39.999 represents a central point. In sum, the error tells us about the distance from actual values of Y (the answers that the GSS survey respondents gave) and predicted values of Y (the ones you calculate based on the GSS respondent's information in the "X" variable). Thus, the error is the difference between a predicted value of Y for a given case and the actual value of Y for a given case ($\hat{Y} - Y$).

More generally, though, when we discuss regression results, we rarely compute predicted scores for particular values of the independent variable. Instead, in scholarly reports, we usually point out the general process at work. In our case, we would say that "each additional year of education is associated with a 2.298 increase on the occupational prestige scale." Note that we refer to "an additional year of education" because our independent variable was measured as years of school completed. Thus, the "unit" of measurement is "years." We say there was a 2.298 increase in prestige with a unit increase in education, because that is the distance we have to move to intersect with the Y-axis, which represents occupational prestige.

Chapter Seven Exercises

1. It is likely that people of different ages watch different amounts of television. How do you think these may be related? Write a hypothesis that predicts the direction of the relationship between *age* and *tvhours*.
2. Do a Pearson correlation to test your hypothesis. Was your hypothesis supported? Explain. Remember that whether or not your hypothesis is supported depends on three things: whether or not the coefficient was 0, whether your prediction of the hypothesized direction of the relationship (+ or -) was correct, and the significance (the probability that you will be wrong if you generalize your finding to the population from which the sample was drawn). Be sure to discuss all three in your explanation.
3. Discuss the strength of the relationship between *age* and *tvhours*. Then, speculate about a second factor that might also influence the amount of television that people watch.
4. How much of the variance in *tvhours* is explained by *age*? Tell how you found out.
5. Do a regression analysis of the relationship between *age* and *tvhours*. Be sure to place your variables into their proper boxes (in other words, correctly identify the independent and dependent variable). If you were writing a scholarly report, how would you describe the relationship between *age* and *tvhours* based on your results? (Hint: If it is small, IBM SPSS may have expressed your regression coefficient in scientific notation in order to save space. If you see something like 2.035E-2 on your IBM SPSS output, that is scientific notation. The E-2 is telling you to move the decimal point two places to the left. Thus, 2.035E-2 becomes .02035. If you don't want to move the decimal yourself, click rapidly several times on the coefficient in the output screen and IBM SPSS will show you the actual value of the coefficient.)
6. Do the results of the regression analysis suggest that your hypothesis is supported? Be sure to discuss the magnitude of the regression coefficient, the direction (+ or -), and the probability.
7. How many hours of television does your model predict that people aged 21 tend to watch each day? People aged 42? Show how you calculated these predicted scores

Chapter Eight: Multivariate Analysis

Up until now, we have covered univariate (“one variable”) analysis and bivariate (“two variables”) analysis. We can also measure the simultaneous effects of two or more independent variables on a dependent variable. This allows us to estimate the effects of each independent variable on the dependent variable, while controlling for the effects of one or more other independent variables. This is called multivariate (“multiple variables”) analysis. In this Chapter we review two ways to do that by using techniques that you have already used: crosstabs and regression analysis.

Crosstabs Revisited

Recall that the crosstabs procedure is used when variables are nominal (or ordinal). Simple crosstabs, which examine the influence of one variable on another, should be only the first step in the analysis of social science data (refer to Chapter 5). One might hypothesize that people who identify as Republicans are more likely to oppose abortion than those who identify as Democrats, then run the crosstabs, and then conclude you were right. However, this one-step method of hypothesis testing is very limited. What if most of the Republicans in your sample are strongly religious and most of the Democrats are not strongly religious? Is it the political party that best explains your findings or is it religious orientation?

Let’s explore this issue. First, we will test whether political conservatism is related to attitudes toward abortion. Before we run the Cross tabulation, we will recode the independent variable (*partyid*) to make our interpretation easier. Refer back to Chapter 3 to review how to recode into a new variable. Naming your new variable *partyidr* (for *partyid* recode), recode *partyid* into three categories as follows:

<u>Values for <i>partyid</i></u>	<u>Values for <i>partyidr</i></u>	<u>Value Label for <i>partyidr</i></u>
0-2	1	Democrat
3	2	Independent
4-6	3	Republican
All other values	System-missing	

Then, run Crosstabs using *partyidr* as the independent variable (in the column position), and *abany* as the dependent variable (in the row position). You may want to review Chapter 5 for direction on how to do this. You should get results that look like the table shown in Figure 8-1.

As the results show, 42.6% of all respondents think that abortion should be allowed for any reason. However, political party identification does appear to be related to attitudes toward abortion. Just over half (51.9%) of respondents who identify as Democrats are in favor of allowing abortion for any reason, compared to less than a third of those identifying as Republicans (31.6%). So, our hypothesis appears to be supported. But we are still left with the question of whether the strength of religious beliefs actually accounts for this relationship.

We will do a multivariate cross tabulation to test this idea, also called an elaboration analysis. If the relationship between party identification and attitudes toward abortion persists no matter what the strength of religious beliefs, then we would conclude that party identification is the better predictor of these attitudes. If the relationship between party identification and attitudes disappears when the strength of religious beliefs is controlled (that is, nearly the same percentage of Democrats and Republicans approve of abortion within each level of religiosity), then we would conclude that the strength of religious beliefs is the better predictor. But what if the relationship between political party identification is different, depending on the strength of religious beliefs? If that is the case, then we would say that there is an interaction between the two independent variables; they work together in shaping attitudes toward abortion.

abany ABORTION IF WOMAN WANTS FOR ANY REASON * partyidr Crosstabulation						
			partyidr			Total
			1.00 Democrat	2.00 Independent	3.00 Republican	
abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	279	97	131	507
		% within partyidr	51.9%	40.6%	31.6%	42.6%
	2 NO	Count	259	142	283	684
		% within partyidr	48.1%	59.4%	68.4%	57.4%
Total	Count	538	239	414	1191	
	% within partyidr	100.0%	100.0%	100.0%	100.0%	

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	39.593 ^a	2	.000
Likelihood Ratio	39.998	2	.000
Linear-by-Linear Association	39.454	1	.000
N of Valid Cases	1191		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 101.74.

Figure 8-1

Recall that your original crosstabs procedure produces one contingency table, with as many rows as there are categories (or values) of the dependent variable, and as many columns as there are categories of the independent variable. When you start using control (sometimes called test) variables, you will get as many separate tables as there are categories of the control variable. There are four categories in the *reliten* variable, the measure of strength of religious beliefs. Thus, we should expect to get four contingency tables, each one showing the relationship between *partyidr* and *abany* for a different category of *reliten*.

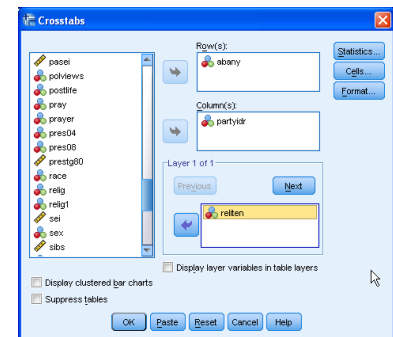


Figure 8-2

To do the crosstabs, click **Analyze, Descriptive Statistics, Crosstabs**, just as you did for testing the bivariate relationship between *partyidr* and *abany*. This time, click on *reliten* and move it into the box called Layer 1 of 1. The dialog box should look like the one shown in Figure 8-2. Click **OK**. Your results should look like the table shown in Figure 8-3.

Notice that the relationship between *partyidr* and *abany* is different depending on the different categories of *reliten*. What does that tell you?

Try other variables as a control (i.e., in place of *reliten*) to see what happens. As a general rule, here is how to interpret what you find from this elaboration analysis:

- If the partial tables, such as those shown in Figure 8-3 are similar to the zero-order table (this is the original bivariate table shown in Figure 8-1), you have *replicated* your original findings, which means that in spite of the introduction of a particular control variable, the original relationship persists. The only way to convince us that this is indeed a strong, or even causal, relationship is if you control for all the other logical independent variables you can think of, and still find essentially no differences between the zero-order tables and their partials.

abany ABORTION IF WOMAN WANTS FOR ANY REASON * partyidr * reliten STRENGTH OF AFFILIATION Crosstabulation

reliten STRENGTH OF AFFILIATION				partyidr			Total
				1.00 Democrat	2.00 Independent	3.00 Republican	
1 STRONG	abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	63	11	31	105
			% within partyidr	38.4%	18.3%	16.1%	25.2%
		2 NO	Count	101	49	162	312
			% within partyidr	61.6%	81.7%	83.9%	74.8%
	Total		Count	164	60	193	417
			% within partyidr	100.0%	100.0%	100.0%	100.0%
2 SOMEWHAT STRONG	abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	6	5	7	18
			% within partyidr	20.0%	19.2%	31.8%	23.1%
		2 NO	Count	24	21	15	60
			% within partyidr	80.0%	80.8%	68.2%	76.9%
	Total		Count	30	26	22	78
			% within partyidr	100.0%	100.0%	100.0%	100.0%
3 NOT VERY STRONG	abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	118	36	75	229
			% within partyidr	57.8%	50.0%	45.2%	51.8%
		2 NO	Count	86	36	91	213
			% within partyidr	42.2%	50.0%	54.8%	48.2%
	Total		Count	204	72	166	442
			% within partyidr	100.0%	100.0%	100.0%	100.0%
4 NO RELIGION	abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	70	39	15	124
			% within partyidr	63.6%	56.5%	55.6%	60.2%
		2 NO	Count	40	30	12	82
			% within partyidr	36.4%	43.5%	44.4%	39.8%
	Total		Count	110	69	27	206
			% within partyidr	100.0%	100.0%	100.0%	100.0%
Total	abany ABORTION IF WOMAN WANTS FOR ANY REASON	1 YES	Count	257	91	128	476
			% within partyidr	50.6%	40.1%	31.4%	41.6%
		2 NO	Count	251	136	280	667
			% within partyidr	49.4%	59.9%	68.6%	58.4%
	Total		Count	508	227	408	1143
			% within partyidr	100.0%	100.0%	100.0%	100.0%

Figure 8-3

- If the difference shown in all the partial tables (the separate tables for each category of the control variable) are significantly smaller than those found in the original AND IF your control variable is antecedent (occurs prior in time) to both the other variables, you have found a *spurious* relationship and explained away the original. In other words, the original relationship was due to the influence of that control variable, not the one you first hypothesized.
- If the differences you see in the partial tables are less than you saw in the original table AND IF your control variable is intervening (that is, the control variable occurs in time after the original independent variable), you have *interpreted* the relationship. If the time sequence between the independent and control variable is

not determinable (or otherwise unclear), then you don't know whether you have explanation or interpretation, but you do know that the control variable is important.

- If one or more of the differences shown in the partial tables is stronger than in the original and one or more is weaker, you have discovered the conditions under which the original relationship is strongest. This is referred to as *specification* or the interaction effect.
- If the zero order table showed weak association between the variables, you might still find strong associations in the partials (which is a good argument for keeping on with your initial analysis of the data even if you didn't “find” anything with bivariate analysis). The addition of your control variable showed it to have been acting as a *suppressor* in the original table.
- Last, if a zero order table shows only a weak or moderate association, the partials might show the opposite relationship, due to the presence of a *distorter* variable.

Look at the patterns in each of the four partial tables and compare them to the original (bivariate) table. What do you conclude about the relationships? Try some of your own three-way (or higher) tables using some of the variables in the GSS10A data set. Recall that for this procedure, there should be few categories for each variable, particularly your control variables (so you might need to recode), and you are limited to variables measured at, or recoded to, nominal or ordinal levels.

Multiple Regression

Another statistical technique estimating the effects of two or more independent variables on a dependent variable is multiple regression analysis. This technique is appropriate when your variables are measured at the interval or ratio level, although independent variables with only two categories can be treated as interval level (these are generally called “dummy variables” in regression analysis). Although technically designed for interval/ratio variables, it is common for researchers to use multiple regression with ordinal variables as well.

You may remember from a statistics course that when we use regression analysis, we make several assumptions about our data: that there is a linear relationship between each independent variable and the dependent variable, that the distribution of values in your variables follows a normal distribution, and the variables are measured at interval or ratio levels.

The variable *hrsrelax* indicates the amount of leisure time respondents have. Since it is measured in the number of hours per day, it is an

sexr				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid .00 Male	923	45.2	45.2	45.2
1.00 Female	1121	54.8	54.8	100.0
Total	2044	100.0	100.0	

Figure 8-4

interval level variable. There is much existing research that finds that women have less leisure time than men do. Let's see if this pattern holds for our data as well. Since we are only working with two variables at this stage, we will first do a bivariate regression analysis, just as we did in Chapter 7. But first, we must recode the sex variable. The mathematics on which regression analysis is based requires that variables having two categories be coded as 0 and 1. In the GSS10A data, sex is coded as 1 (male) and 2 (female). So, you must first recode sex. If necessary, review the directions for recoding into a new variable in Chapter 3. Name the new variable *sexr* (sex recoded), and code male as 0 and female as 1. Do a frequency distribution of the new variable to make sure you did it correctly. It should look like the one shown in Figure 8-4. Don't forget to add value labels to the new variable.

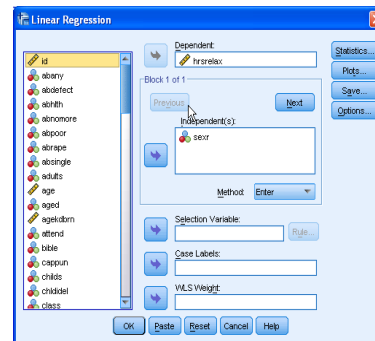


Figure 8-5

Now we are ready to analyze the data. We will start first by examining the relationship between sex and leisure time. On the menu, click **Analyze**, **Regression**, **Linear**. Click on *hrsrelax* and move it into the Dependent box at the top of the dialog box. Click on *sexr* and move it into the Independent(s) box. It should look like the one shown in Figure 8-5. Click OK.

Your results should look like those shown in Figure 8-6. Looking first at the Model Summary table, you will see that the R-squared value is .018. As you recall from Chapter 7, this means that 1.8% of the variation in the dependent variable (hours of relaxation time) is explained by knowing the sex of the respondent. This doesn't seem like very much. However, our goal was not to explain all of the factors that influence how much time people have to relax. Our goal was to find out whether women report having less relaxation time than men do. For this, we need to look at the Coefficients table. If you look at the B coefficient for *sexr*, you will see that it is -.490. If there was absolutely no relationship between *sexr* and *hrsrelax*, the coefficient would be zero. How do we interpret this coefficient? Recall the discussion in Chapter 7: a one unit change in the independent variable (*sexr*) is associated with a change in the dependent variable (*hrsrelax*) equal to the value of B. So, if we increase the value of *sexr* by 1, on average, we get a change of -.490 in *hrsrelax*. Since *sexr* is a dichotomous variable, increasing the value by 1 simply means going from "male" to "female." Here is what one might say in a report: Women

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	sexr ^a	.	Enter

a. All requested variables entered.
b. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.093 ^a	.009	.008	2.619

a. Predictors: (Constant), sexr

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	69.087	1	69.087	10.069	.002 ^a
	Residual	7903.923	1152	6.861		
	Total	7973.010	1153			

a. Predictors: (Constant), sexr
b. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	3.939		35.077	.000
	sexr	-.490	.154	-.093	.002

a. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Figure 8-6

report .490 fewer hours than men in their average amount of relaxation time per day. (It is fewer, because the coefficient is negative; so, when X, the independent variable, goes up, from 0 to 1, Y, the dependent variable, goes down). Notice that the t is statistically significant, with a value of .002. This means that since the sample was randomly selected, the pattern you see in the data is very likely an accurate reflection of the average experiences in the population from which the sample was drawn.

Now, that analysis tells us that women report fewer hours of relaxation per week, but it doesn't tell us why. Two variables that might be relevant are the number of children that people have, and their income level. One reason might be because most women spend more time doing childcare and housework than most men and therefore have fewer hours of relaxation time. So, we would expect to see a negative relationship between the number of children and the amount of leisure time. Another might be income; the higher the family income, the more money there is available to hire others to do some of the work, like house cleaning, babysitting, gardening, etc., making more time for leisure. If that is the case, then we would expect to find a positive relationship between income and leisure time. Finally, if these two variables explain why women have less leisure time than men do, then controlling for them should result in a smaller B coefficient for *sexr*. Lets test this out using multiple regression analysis.

Once again, click **Analyze**, **Regression**, **Linear**. Click on *hrsrelax* and move it into the **Dependent** box at the top of the dialog box. Click on *sexr* and move it into the **Independent(s)** box. Then, move *childs* and *income06* into **Independent(s)** box as well. It should look like the one shown in Figure 8-7. Click **OK**.

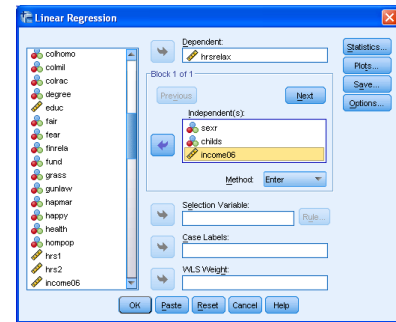


Figure 8-7

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	income06 TOTAL FAMILY INCOME, childs NUMBER OF CHILDREN, sexr	.	Enter

a. All requested variables entered.
b. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.124 ^a	.015	.013	2.607

a. Predictors: (Constant), income06 TOTAL FAMILY INCOME, childs NUMBER OF CHILDREN, sexr

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	112.932	3	37.644	5.539	.001 ^a
	Residual	7238.129	1065	6.796		
	Total	7351.061	1068			

a. Predictors: (Constant), income06 TOTAL FAMILY INCOME, childs NUMBER OF CHILDREN, sexr
b. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.772	.310		15.386	.000
	sexr	-.472	.161	-.090	-2.940	.003
	childs NUMBER OF CHILDREN	-.020	.052	-.012	-.392	.695
	income06 TOTAL FAMILY INCOME	-.046	.015	-.095	-3.096	.002

a. Dependent Variable: hrsrelax HOURS PER DAY R HAVE TO RELAX

Figure 8-8

Look at the B values in the Coefficients in the table (your results should look like those in Figure 8-8).

Notice that none of the B coefficients are 0, meaning that each of the three independent variables have some relationship with the number of hours of relaxation per week, when the effects of the others are controlled. Furthermore, for two of the variable, the t statistic is significant ($p \leq .05$), so that you can generalize the patterns you see in the data to the population from which the sample was drawn. In contrast, the t statistics is not statistically significant for the remaining variable (*chlds*). Thus, even though the results are that all three independent variables have some relationship to the dependent variable (the B is not 0), we can generalize these results to the population for only two – sex and total family income. How do we interpret these patterns? First, we would say that the number of children does not appear to be related to the number of hours of relaxation time, controlling for sex and family income. Family income is related to relaxation time, but not in the expected direction. The effect of income is also negative—those who earn more have fewer hours of relaxation, a result which was contrary to our hypothesis. Thus, even though the results show that a relationship exists, our hypothesis is not supported (because we predicted in the wrong direction). So, neither of our hypotheses were supported by the data. Interestingly, the decrease in the value of B was very small from the bivariate analysis (-.490) to the multivariate analysis (-.472). This suggests that the number of children does not appear to account for the pattern in which women report working fewer hours per week than men. What do you think accounts for this difference?

Chapter Eight Exercises

1. How would you hypothesize the relationship between *fear* (Afraid to walk at night in neighborhood) and *sex*?
 - a. Write out your hypothesis.
 - b. Run a crosstabs to test your hypothesis and report your results.
 - c. Now, do a second crosstabs, this time controlling for *class*. Report your results.
 - d. Now run *fear* and *sex* but control for *trust*. Report your results.
2. Choose three independent variables from the dataset that you think influence the number of hours people watch television (*tvhours*, the dependent variable).
 - a. Write up your hypotheses (how and why each independent variable is associated with the dependent variable).
 - b. Run a multivariate regression to test your hypotheses and report your results.

Chapter Nine: Presenting Your Data

This chapter discusses methods for presenting your data and findings in your reports. Most of this chapter is devoted to introducing you to methods for creating and editing charts. Then, we review ways to edit the tabular output from the various statistical procedures so that you convey just the information you need. Finally, we show you how to copy your work from the IBM SPSS output screen into a word-processing document (i.e., Microsoft Word).

Charts

Deciding on a Chart Type:

Charting a Single Variable: DEGREE

First, let's decide how we might display education as measured by *degree*. Click on **Graphs**, then **Chart Builder**. The Chart Builder box is shown in Figure 9-1. You will be creating your chart within the window shown in this box.

Notice that the text contained in the chart window indicates that there are two ways to build a chart (by dragging a gallery chart or by clicking on the Basic Elements tab). We will be using the first method, dragging a Gallery chart to use as a starting point.

First, you might want to familiarize yourself with the items contained in the **Chart Builder**. Making sure that the **Gallery** tab is active (this is the default), click on the various choices of chart types (i.e., Bar, Line, Area, etc.) to review the forms that each type takes. Also, notice the other tabs located to the right of the **Gallery** tab. We will return to these tabs later to add a title to the chart.

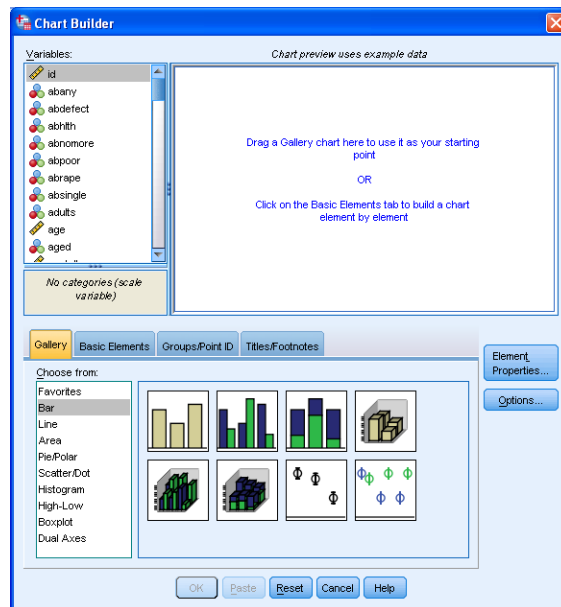


Figure 9-1

Returning to the Gallery tab, it looks like a Bar Chart would be a good way to display *degree*. Click on **Bar** from the list of chart types. Click on the first one, the simple bar chart (you can see the names of each chart as you hold your mouse over it), and drag it up into the Chart Builder window. In addition to seeing the simple chart show up in the Chart Builder window, you will see that a second dialog box, called **Element Properties**, has opened. Figure 9-2 shows what you should be seeing on your screen. For the moment, ignore the Element Properties box.

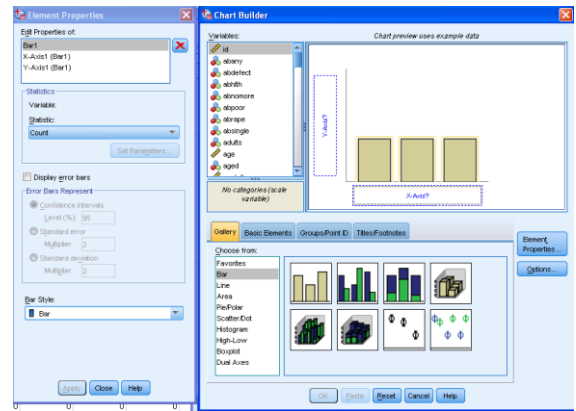


Figure 9-2

We are creating a chart to summarize only one variable. So, we will only be using the X-axis. Locate *degree* in the Variable List, click on it, and drag it to the box labeled “X-axis?”. When you do that, your screen should look like the one shown in Figure 9-3.

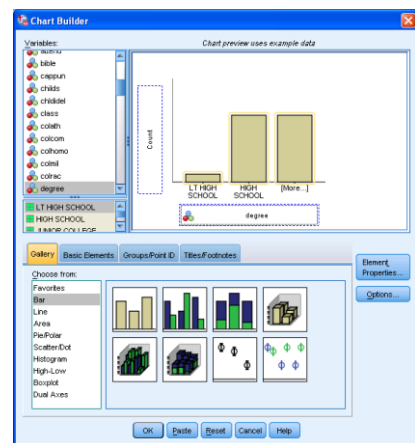


Figure 9-3

Notice that the box that previously was labeled “Y-axis?” is now labeled **Count**. What does this mean?

What if you would rather display the percentage of respondents who have various degrees rather than the number of respondents? You can do this by using the **Element Properties** dialog box. Click on **Bar 1**, then click on the drop-down menu under

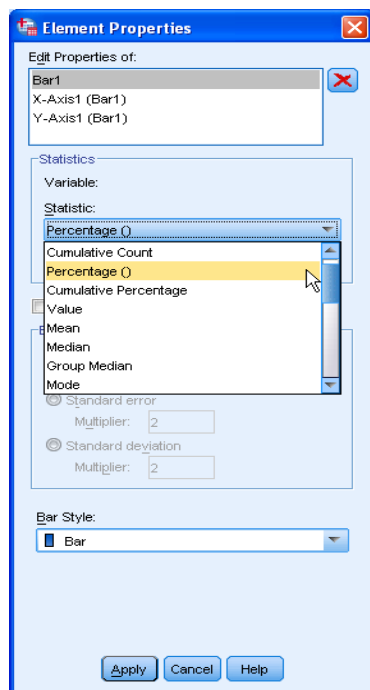


Figure 9-4

Statistic, click **Percentage (?)**, then **Apply**. Your screen should look like Figure 9-4. Notice that at the bottom of the **Element Properties** box, you may change your **Bar Style**. Choose a new one if you wish, but be sure to click **Apply** afterward.

The final step is to give the chart a title. Click on the **Titles/Footnotes** tab, then click the box next to **Title 1**. You should notice that the Elements box now looks like the one shown in Figure 9-5.

In the Content window, type in your title (for example, **Figure 1: Highest Degree Earned**). Then, click **Apply**.

We are now finished defining what our chart should look like. Moving back to the Chart Builder window, click **OK**. Your finished chart should look like Figure 9-6.

If you wish, you may continue to edit your chart from the Output screen. To do this, double-click anywhere in the chart, and it will open in the Chart Editor. Explore the menus in the Chart Editor to experiment with what you can do. Try this: click on one of the bars in your chart. Then, click **Edit**, then **Properties**. Choose a new color for the bars, then click **OK**. Explore some of the other menus in the Chart Editor to find out what they do.

Tables

Let's create a cross tabulation of *sex* and *fear*. Click on **Analyze**, then **Descriptive Statistics**, then

Crosstabs. Put *fear* in the Row box and *sex* in the Column box (recall that in cross tabulations, the independent variable always goes in the column position). Now click on **Cells** and select **Column** in the Percentages box, and then click on **Continue**, then **OK**.

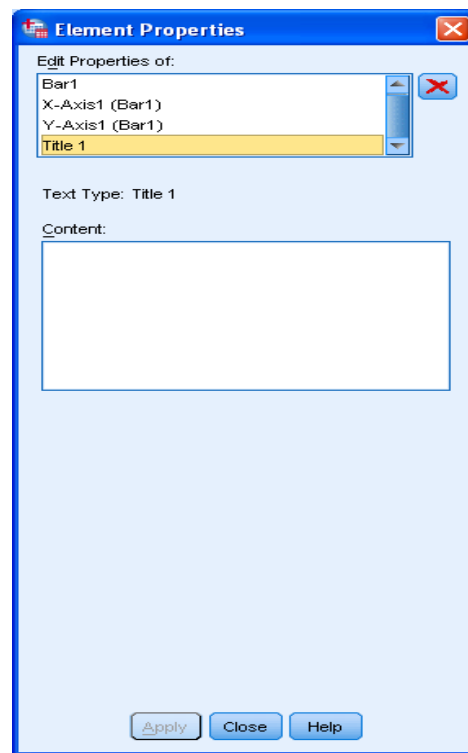


Figure 9-5

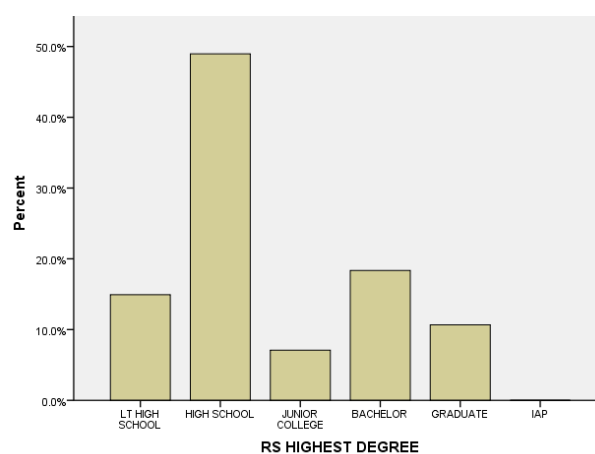


Figure 9-6

The Output Window will appear, and your screen should look like Figure 9-7.

The left-hand frame will contain an outline of the tables in the right-hand frame. Click on **FEAR AFRAID TO WALK AT NIGHT IN NEIGHBORHOOD * SEX RESPONDENTS SEX** Cross tabulation. This will select that table in the right-hand frame. The red arrow to the left of the table indicates that it has been selected. Double click anywhere inside the selected table. The border of the table will change from a solid line to a dotted line. You can now edit the table. Let's say you want to change the title. Point your mouse at the title and click once. This highlights the title (see Figure 9-8).

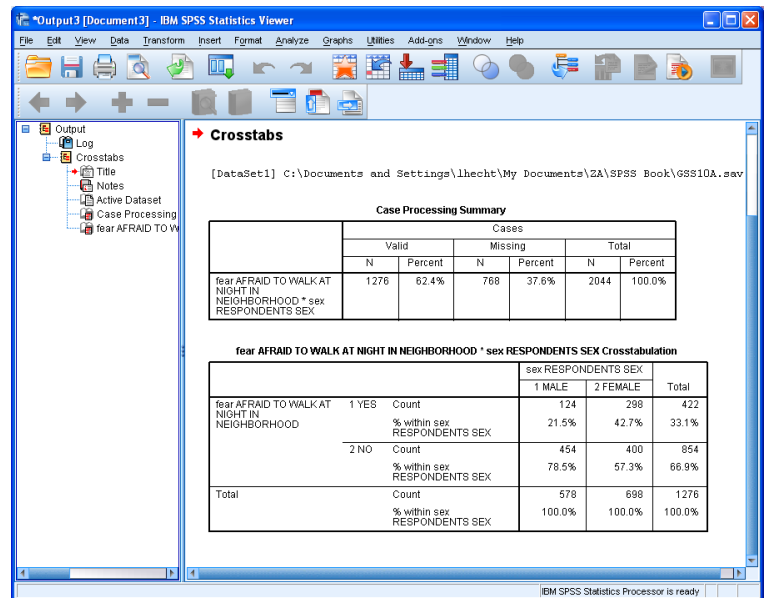


Figure 9-7

You can now edit the title. Delete the old title, and type in: **Table 1: Crosstabulation of Fear of Walking at Night, by Gender**, then click anywhere in the table (outside of the title). You can click in any area of the table, and make changes. By clicking in the various cells and deleting and/or editing contents, you can simplify the table so that it looks like the one shown in Figure 9-8. You can also use the toolbar that appeared when you double clicked the table to change fonts, etc. If you do not want to make changes using the toolbar, simply close that box.

Copying and Pasting Charts and Tables to a Document

Since you will probably be using a word processing program to prepare the report of your results, it will be useful to copy your charts and tables from IBM SPSS into your word-processing document. Let's start with the table we just created. There are two ways to do this. The simplest way is to click on the table using the right mouse button. A small menu will appear; click on **Copy**. Then, go to your word-

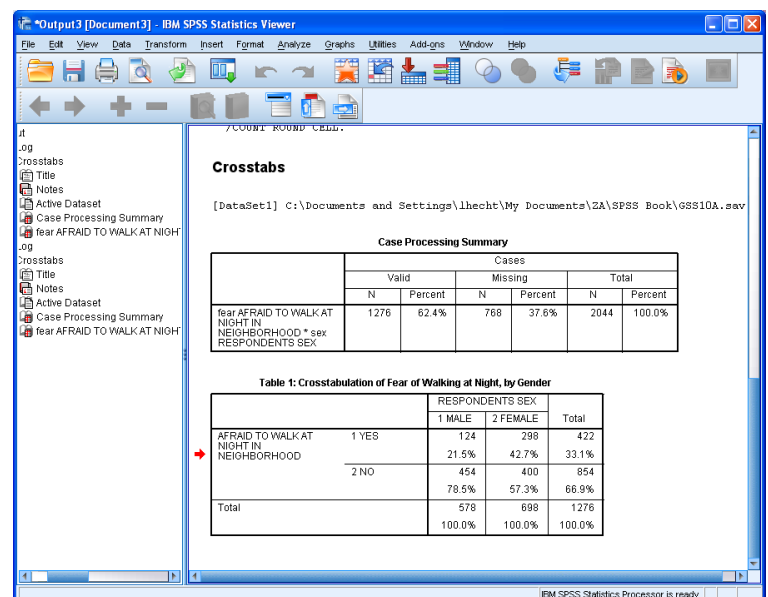


Figure 9-8

processing document, and right-click where you want the table to appear. The small menu will appear again; click **Paste**.

The second way to copy the table is by using the menu commands. Make sure the table you want is selected (you will see the red arrow pointing to it, and the table will have an outline around it). Click on **Edit** on the menu bar, then click on **Copy**. Switch over to your word-processing document. Click the mouse where you want to paste your table. Click on **Edit** on the menu bar, then click on **Paste**. You might want to paste your graph into a Text box. This will make your graph easier to move. You could also click on **Paste Special** instead of **Paste**. This would give you choices about the format for your table. For example, using **Paste Special**, you can paste is as a picture rather than as text. Note: The method for copying and pasting charts is exactly the same as the method as for copying and pasting tables.

Chapter Nine Exercises

1. Make a bar chart of *trust*. Then, edit the chart by giving it a proper title. Copy and paste the chart into a word processing file. Write a few sentences that describe the pattern shown in the chart.
2. Do a cross-tabulation of *hapmar* and *trust*. Since *hapmar* is the independent variable, place it in the column location, and show column percentages (see Chapter 5 for a review). Be sure that your table is properly titled. Copy and paste the table into a word processing file. Write a few sentences that discuss the relationship of the information shown in the table to the information shown in the chart you created for Question 2.

Appendix A:

Codebook for the Subset of the 2010 General Social Survey

The General Social Survey (GSS) is a large, national probability sample of adults in the United States. It began in 1972 and continued on an almost yearly basis until 1996. In 1996, the GSS became a biannual survey and the sample size increased. Many questions are asked on each survey, while other questions are rotated from survey to survey. This subset from the 2010 GSS includes all the cases (2,044) and 80 variables.

Variable	Description of Variable
ABANY	Abortion if woman wants for any reason
ABDEFECT	Abortion if strong chance of serious defect
ABHLTH	Abortion if woman's health seriously endangered
ABNOMORE	Abortion if married and wants no more children
ABPOOR	Abortion if low income and can't afford more children
ABRAPE	Abortion if pregnant as result of rape
ABSINGLE	Abortion if not married
ADULTS	Household members 18 years and older
AGE	Age of respondent
AGED	Should aged live with their children
AGEKDBRN	Respondent's age when first child born
ATTEND	How often respondent attends religious services
BIBLE	Feelings about the bible
CAPPUN	Favor or oppose death penalty for murder
CHILDS	Number of children
CHLDIDEL	Ideal number of children
CLASS	Subjective class identification
COLATH	Allow anti-religionist to teach
COLCOM	Should communist teacher be fired?
COLHOMO	Allow homosexual to teach
COLMIL	Allow militarist to teach
COLRAC	Allow racist to teach
DEGREE	Respondent's highest degree
EDUC	Highest year of school completed
FAIR	People fair or try to take advantage
FEAR	Afraid to walk at night in neighborhood
FINRELA	Opinion of family income
FUND	How fundamentalist R is
GRASS	Should marijuana be made legal
GUNLAW	Favor or oppose gun permits
HAPMAR	Happiness of marriage
HAPPY	General happiness
HEALTH	Condition of health
HOMPOP	Number of persons in household
HRS1	Number of hours R worked last week

HRS2	Number of hours R usually works a week
HRSRELAX	Number of hours R has to relax per day
INCOME06	Total family income in 2009
INC06REC	Recoded total family income in 2009
LIBATH	Allow anti-religious book in library
LIBCOM	Allow communist's book in library
LIBHOMO	Allow homosexual's book in library
LIBMIL	Allow militarist's book in library
LIBRAC	Allow racist's book in library
MADEG	Mother's highest degree
MAEDUC	Highest year school completed, mother
MARITAL	Marital status
MASEI	R's mother's socioeconomic index
PADEG	Father's highest degree
PAEDUC	Highest year school completed, father
PARTYID	Political Party Affiliation
PASEI	R's father's socioeconomic index
POLVIEWS	Think of self as liberal or conservative
POSTLIFE	Belief in life after death
PRAY	How often does respondent pray
PRAYER	Support Supreme Court Decision on prayer in public schools
PRES04	Vote for Kerry, Bush, Nader in 2004
PRES08	Vote for Obama, McCain in 2008
PRESTG80	Respondent's occupational prestige score (1980)
RACE	Race of respondent
RELIG	Respondent's religious preference
RELIG1	More detailed breakdown of religious preference
RELITEN	Strength of affiliation
SEI	Respondent's socioeconomic index
SEX	Respondent's sex
SIBS	Number of brothers and sisters
SPDEG	Spouse's highest degree
SPEDUC	Highest year school completed, spouse
SPKATH	Allow anti-religionist to speak
SPKCOM	Allow communist to speak
SPKHOMO	Allow homosexual to speak
SPKMIL	Allow militarist to speak
SPKRAC	Allow racist to speak
THNKSELF	To think for ones self
TRUST	Can people be trusted
TVHOURS	Hours per day watching television
VOTE04	Did R vote in 2004
VOTE08	Did R vote in 2008
WTSS	Weight variable
YEAR	Year (2010)
ZODIAC	Respondent's astrological sign