

# SPSS for Windows Version 13.0: 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 SPSS (Version 13). Additional information can be found in the *SPSS Base 13.0 User'-s Guide* (SPSS, Inc., 2004). 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, GSS02A, can be downloaded from [here](#), and follow the instructions for downloading the file and accessing other text information. 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 SPSS or to find a computer laboratory in which SPSS has been installed. Most colleges and universities have labs where you can use SPSS. If you purchase 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 and now 13.0). 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, Julie Kehrwald of McGraw-Hill Custom Publishing, for her 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](mailto:fiddlerlinda@yahoo.com))

Laura Hecht ([lhecht@csub.edu](mailto:lhecht@csub.edu))

Edward Nelson ([ednelson@csufresno.edu](mailto:ednelson@csufresno.edu))

Elizabeth Nelson ([elizn@csufresno.edu](mailto:elizn@csufresno.edu))

James Ross ([jross@csub.edu](mailto:jross@csub.edu))

## Table of Contents

*Preface*

*Chapter One: Getting Started with SPSS for Windows*

*Using Windows*

*Starting SPSS for Windows*

*Leaving SPSS*

*Looking at Data*  
*A Brief Tour of SPSS*  
*Overview of Chapters*

*Chapter Two: Creating a Data File*

*Basic Steps in Creating a Data File*  
*Getting Started in SPSS*  
*Student Survey Questionnaire*  
*Codebook for Student Questionnaire*  
*Planning Matrix for Data-entry Exercise*  
*Chapter Two Exercises*  
*Sample Data Set: Friendly Visitor Service Clients*

*Chapter Three: Transforming Data*

*Recoding Variables*  
*Recoding Into Different Variables*  
*Recoding into the Same Variable*  
*Creating New Variables Using COMPUTE*  
*Creating New Variables Using IF*  
*Using Select Cases*  
*Weighting Cases*  
*Summary*  
*Chapter Three Exercises*

*Chapter Four: Univariate Statistics*

*Frequencies*  
*Variable Names, Variable Labels, Values, Value Labels*  
*Descriptives*  
*Explore*  
*Conclusion*  
*Univariate Analysis as Your First Step in Analysis*  
*Chapter Four Exercises*

*Chapter Five: Cross Tabulations*

*Exercises for Chapter Five*

*Chapter Six: Comparing Means*

*Independent-Samples T Test*  
*Paired-Samples T Test*  
*One-Way Analysis of Variance*  
*Summary*  
*Chapter Six Exercises*

*Chapter Seven: Correlation and Regression*

*Correlation*  
*Regression*  
*Chapter Seven Exercises*

*Chapter Eight: Multivariate Analysis*

*Cross Tabulation Revisited*

*Multiple Regression*

*Chapter Nine: Presenting Your Data*

*Charts*

*Tables.*

*Copying and Pasting Charts and Tables to a Document*

*Chapter Nine Exercises*

*Appendix A: Codebook for the Subset of the 2002 General Social Survey*

# Chapter One: Getting Started With SPSS for Windows

## Using Windows

Your Windows opening screen should look something like [Figure 1-1](#). Several standard desktop icons, such as **My Computer** and **Recycle** will always appear on start up. Note the Taskbar along the bottom, with the Start button 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.

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** button (don't click anything yet, just let the pointer rest somewhere on the button). Notice that a label appears in a little rectangle, just above the **Start** button, showing **Click here to begin**.

These floating labels will usually appear in your active window whenever the mouse cursor moves over a button icon. They will then disappear after a few seconds, so pay attention! Now that you have found the Start button, click on it once and the Start Menu will appear, and your screen should look something like [Figure 1•2](#).

This is the basic Start Menu, and 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. Nothing other than that will happen with the Help, Run, or Turn Off icons, but letting the pointer hover over Programs, Documents, Settings, or Search will cause other menus to slide out across your screen. The little triangles at the right side of the Start Menu are your clue here. If one is present, that means there are more menus for that item. If this is the first time you are using Windows, the Documents menu will say "empty". 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 (such as Word, SPSS, Excel, etc.).

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 a program icon, it becomes highlighted, and its sub•menu appears.

## Starting SPSS for Windows

The SPSS 13 for Windows 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. On my computer, all you have to do to start SPSS is to point to the SPSS 13 icon on the desktop and double click. Then wait while SPSS loads.

After SPSS loads you may, depending on how 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 **SPSS** button. 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 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, 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 button for **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 SPSS

We’re not ready to actually use 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 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 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 SPSS should close. This is called click and drag and is another way to use your mouse.

The third way to close SPSS is to point your mouse at the SPSS icon in the upper-left corner of the screen. The icon will be just to the left of the words **Untitled: SPSS 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 SPSS.

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

Now you know how to move your mouse around and how to start and close 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 to use it.

## Looking at Data

There is a data file to be used with this tutorial called GSS02A. This is a subset of the 2002 General Social Survey. 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.csub.edu/~jross/projects/SPSS/>. Click on **SPSS Text** and follow the instructions on downloading the data 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 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'll see the words **SPSS 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 which 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 File box. (You're beginning to get the idea of how Windows works.) Here you need to tell SPSS where to find the data file to open. 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, GSS02A, 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. 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 NAP 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 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 SPSS for Windows.

## A Brief Tour of SPSS

Now that you have the file opened, let's look at some things you can do with 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](#).

This lists the statistical procedures in 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 [\[1\]](#) the **Frequencies** box. Since **ABANY** is the first variable in the data, it's already highlighted. 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 SPSS for Windows, 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 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**.

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) statistics (a summary of the number of missing and valid observations), and (4) the actual table showing the frequency distribution. Click on **Statistics** in the left-hand pane and you will see that there were 900 valid and 1865 missing cases. Click on **abortion--for any reason** and you will see the frequency distribution. In [Figure 1•6](#), you can see that 387 people said yes, 513 said no, 23 said they didn't know, 1 didn't answer the question, and 1841 were coded not applicable. (These 1841 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 924 of the approximately 2800 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](#).

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](#).

Frequencies by themselves are seldom useful. You'll need to compute percents. To get SPSS to compute percentages, point your mouse at the button labeled **Cells** at the bottom 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](#).

Now click on **Continue** and you will be back to the **Crosstabs** box. To tell SPSS to run the **Crosstabs** procedure, click on **OK**. After a few seconds your screen should look like [Figure 1•10](#). Use the scroll bar to look at all the information that 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-four 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](#).

Now put age at birth of first child (**AGEKDBRN**) 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 **AGEKDBRN** 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 **AGEKDBRN** 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](#).

Women had their first child at an average age of 22.74 years, while the average for men is 25.39 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 the 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 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. The output should look like [Figure 1•16](#).

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.

We can also look at a scatterplot showing the relationship between father's education and the respondent's education. Click on **Graphs** in the menu bar and then click on **Scatter/Dot**. This will open the Scatter/Dot box. Click on **Simple Scatter** and then on **Define**. This will open the Simple Scatterplot box. Scroll down the list of variables on the left until you see **EDUC** and click on it to highlight it. Then click on the arrow to the left of the Y Axis box to move **EDUC** into this box. Scroll down this same list until you find **PAEDUC** and click on it. Then click on the arrow to the left of the X Axis box to move **PAEDUC** into it. Your screen should look like

[Figure 1•17.](#)

Now click on **OK** and an output box should open. Your screen should look like [Figure 1•18.](#)

[2]

Each dot in the scatterplot represents a case in your data set. In general, the higher the education of the father, the higher the education of the child. However, it is far from a perfect relationship. Many fathers with more education have children with less education and many fathers with less education have children with more education.

## Overview of Chapters

Chapter 2 will acquaint you with how to enter new data into SPSS for Windows 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 teach 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 more than one variable at a time, and Chapter 7 will teach 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.

---

[1]

Instead of seeing a list of variables, you may see a list of variable labels. You can change this so 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 and to the left of **Display names**. Finally, click on **OK**. This will open another dialog box that asks about changing options in the variable list. Click on **OK**.

[2]

Depending on the number of cases in the scatterplot, a dot may represent more than one case.

## 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 a 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 [\[1\]](#) learn how to create a new data file. They decided to use the following questions :

- 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, 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:

[\[2\]](#)

If there is a strong chance of a serious defect in the baby? [ABDEFECT ]

If she is married and does not want any more children? [ABNOMORE]

If the woman's own health is seriously endangered by pregnancy? [ABHLTH]

If the family has a very low income and cannot afford any more children? [ABPOOR]

If she became pregnant as a result of rape? [ABRAPE]

If she is not married and does not want to marry the man? [ABSINGLE]

If the woman wants it for any reason [ABANY]

### 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 about the data. (SPSS limits variable names 64 characters or fewer). They can use numbers or letters but not spaces and very few special characters, so don't use any odd

[\[3\]](#)

symbols.) AGE and SEX would be good variable names for the first two questions. 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 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, Middle of the Road.

3. Sometimes respondents do not answer a question, give more than one answer, or do something else that would make 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 can 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 SPSS. It is useful to put the data in a matrix like Table 2.1 before entering it into the 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

ID	AGE	SEX	REL	CONLIB
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

## Getting Started in SPSS

To create the data file in SPSS, open SPSS (probably by clicking on the SPSS icon on the desktop. When it says, What would you like to do?, choose Type in data and click OK. (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 the identification number for the first case and the cells to the right will be data about that case. The SPSS Data Editor has tabs in the lower-left 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. SPSS probably opened in the **Data View** mode, if not, click the **Data View** tab at the bottom left of the 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. (See [Figure 2-3](#).) [4]
- We will use the **Variable View** tab to assign variable names and longer variable labels plus value names and labels that will make it easier to use the data in tables and charts. Click **Variable View** now and click the VAR00001 in the top left column. Type in **ID**. (We use all capital letters for variable names to differentiate them from other terms.) Press **Enter** and VAR00001 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](#).)
- 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. SPSS makes some assumptions about data that might not be appropriate. Notice that it uses two decimal points even if the values are integers. To hide these inappropriate decimals, select the **decimals** column by clicking the heading and type **0** instead of **2**. Remember to do this whenever a numeral doesn't really refer to a numerical value. 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, but for most variables a longer variable label is very useful.
- Sometimes respondents don't answer a question or 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 gray box. This dialog box lets you specify up to three distinct missing values. For our data, click **discrete** and type **99** in the first text box and leave the other two empty. Then Click **OK**. Now if someone doesn't answer a question, it will be marked as missing. Click **Add** to save this. (See [Figure 2-6](#).)

5. 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 gray box to get the **Value Labels** menu. Type a **1** in the **value** box and then **Male** in the **value label** box and click **Add**. Then, click the gray box again, type a **2** in the **value** space, and type **Female** in the **value label** space. Click **Add** and then click **OK** to save these. Now, SPSS knows that 1 and 2 in **SEX** are really male and female respectively. (See [Figure 2-7](#).)

6. For this exercise, we are also using religion and conservative-liberal as variables. Add those variables in rows 4 and 5. Give each a variable label—REL gets **Religion** and CONLIB gets something like **Conservative-Liberal**. 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 variable labels, as well as value names and value labels. Make arrangements for missing values just as you did in #4 above. (You can refer to the 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 data file might look like [Figure 2-8](#).)

## Entering the Data

[5]

7. Enter the codes for each variable using **Data View**. 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**.

## 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, \_\_\_ middle of the road, or

(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

**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

**Planning Matrix for Data-entry Exercise**

	AGE	SEX	REL	C-L	MG	ABD	ABN	ABH	ABP	ABR	ABS	ABA
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 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

ID	AGE	SEX	ALONE	INCOME	#ADL	#IADL
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 SPSS is the death notices in local newspapers. Although big city newspapers publish obituaries only on the rich and famous, many local newspapers provide information on almost everyone who dies in the community. (See, for example, see *The Fresno Bee*--[www.fresnobee.com/obituaries](http://www.fresnobee.com/obituaries)--which publishes information provided by funeral homes for most deaths in the community as well as more detailed obituaries provided by some families). 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. This could be used for analysis with SPSS, for example, 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.

[1]

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

[2]

This, ABDEFECT, is the variable name of this question in the General Social Survey.

[3]

For this book, we use all caps for variable names.

[4]

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

[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'll 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.

## 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. SPSS for Windows can transform the values in several ways. SPSS for Windows can:

1. combine values of a variable into several categories,
2. create new variables out of old variables,
3. select particular cases and analyze only these cases,
4. weight cases so that some cases count more heavily than others.

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

### 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 SPSS for Windows 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.
<input type="text"/> 18 years through 35 years	Young
<input type="text"/> 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 SPSS that will do the above examples.

Starting SPSS for Windows the way you were taught in Chapter 1, bring in the GSS02A 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 and then point your mouse at Recode. Your screen will look like [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 **Recording 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 **Into Different Variables**. Your screen will look like [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 SPSS to make these changes. Your screen will look like [Figure 3•3](#).

Now we have to tell 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](#).

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 [\[1\]](#) also other options .

As you can tell from the previous example, you can also change a range of values into a new value and that is what we are going to do. 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 to the left of through, click on the box to the right of through, and type **29** in that box. Then click on **VALUE** just below NEW VALUE and type **1** in that box. This will have 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 to the left of through, click on the box to the right of 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](#).

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**.

Now we want SPSS to carry out the recoding. Click on **CONTINUE** at the bottom of the window. This will take you back to the **RECODE INTO DIFFERENT VARIABLES** box. Click on **OK** and SPSS will take a few seconds to carry out your commands. The data matrix should appear on the screen. When it says that the SPSS Processor is **READY** at the bottom of the window you know that SPSS has finished with the recoding.

Click on **ANALYZE**, then point your mouse at **DESCRIPTIVE STATISTICS**, and then click on [\[2\]](#)  
**FREQUENCIES**. Notice that **AGE1** 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](#).

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 **GSS02A • 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](#).

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 Two.

In the Values box 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 Value Label. Click in the Value box and type the value **1**. Then click in the Value 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](#).

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.

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. Your screen should look like [Figure 3•9](#).

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 then point your mouse at **Recode**. This time 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 **NUMERIC 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](#).

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 to the left of **through** and a 2 to the right of **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 the left. Point your mouse at the variable name (PRAY) and double click. This will open the Variable View tab in the Data Editor. Click on the small gray box in the Values 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](#).

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 women 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.

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 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.

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 seven and **Low Approval** for the value fourteen. (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 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 has 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 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, 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](#).

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 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 one means the mother has more education than the father. Any value close to one 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 point your mouse at **Recode**, and finally click on **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.

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.

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](#).

Click on **OK** and 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 point your mouse at **Recode** and then click on **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 box to the left of through and **1.10** in the box to the right. 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 to the left of through. 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 on the small gray box in the Values 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 Define Variable box. Run a frequencies distribution on the new variable to double-check your work. Your screen should look like [Figure 3•22](#).

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 44% of the respondents have fathers and mothers with similar education, while about 31% have fathers with substantially less education than the mother and another 26% have fathers with substantially more education than the mother.

You have already seen that 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**.

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**.

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. In the **Values** box 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.

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

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 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. This means that the cases you do not select can be used later if you click on **All Cases**. If you had selected **Deleted**, 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**.)

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 **~=** sign. This symbol means "not equal to." Now click on **4** and the expression in the box will read **RELIG ~= 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. 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](#).

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. 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". SPSS uses the symbol **&** for **AND** and the symbol **|** for **OR**. We want all cases for which **RELIG** is 1 or 2. 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**.

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 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 **Unfiltered** in the Select Cases box when you began using select cases. If you selected **Deleted**, then you will have to get another copy of the data file by clicking on **FILE** and then on **OPEN**.

## Weighting Cases

Sometimes you may want to weight some cases in your data more heavily than others. Each household represented in the General Social Survey (i.e., the data set you have been using in Chapter Three) had an equal probability of selection. If there was more than one person eligible in the household (18 years of age or older), then one of these individuals was randomly selected. If there was one eligible person in the household, then that person had a 1 out of 1 chance of being selected. If there were two eligible people, then each person had a 1 out of 2 chance. If there were three eligibles, then each person had a 1 out of 3 chance and so on.

In other words, the more eligible people in the household, the smaller the chance of selection for any one of them. We can correct for this by weighting each case by the number of eligible people

in their household. There is a variable called ADULTS which is the number of people 18 years of age or older in the household and this is, of course, also the number of eligible people in the household.

The number of adults in the household varied from one to five. The following table shows what this distribution looks like.

Weighting Cases by Number of Eligible Adults in Household (ADULTS)		
Number of Eligible Adults	Number of Cases	Weighted Number of Cases
1	1047	1047
2	1364	2728
3	258	774
4	75	300
5	21	105
Total	2765	4954

The weighted number of cases is just the number of eligible adults multiplied by the number of cases. This means that each case with two eligible adults has a weight twice that of each case with one eligible adult, each case with three eligible adults has a weight three times that of each case with one eligible, and so on.

The problem with this is that we started with 2,765 cases and ended up with 4,954 cases. This artificially inflates the size of the sample which we really don't want to do. There is an easy way to fix this. If we divide 4,954 (the weighted sum of cases) by 2,765 (the actual number of cases) we get 1.792. We can divide each weight by 1.792 to get an adjusted weight. This would produce the following weighted data.

Weighting Cases Using Adjusted Weights			

Number of Eligible Adults	Adjusted Weight	Number of Cases	Weighted Number of Cases
1	$1/1.792=0.558$	1047	584.23
2	$2/1.792=1.116$	1364	1522.22
3	$3/1.792=1.674$	258	431.89
4	$4/1.792=2.232$	75	167.40
5	$5/1.792=2.790$	21	58.59
Total		2765	2764.33

Notice that when using the adjusted weights, the weighted number of cases equals the number of cases (except for a small amount of rounding error). Let's use compute to create our new adjusted weight variable. We'll call this variable WADULTS for weighted adults. Click on **TRANSFORM** and then on **COMPUTE**. Click on **RESET** to get rid of what you entered previously. Type **WADULTS** in the Target Variable box. Find the variable ADULTS 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 to move it into the Numerical Expression box. Now click on / (for division) and then enter the value 1.792 by clicking on the one, then the decimal, then seven, then nine, and finally on two. The formula in the box should read "ADULTS/1.792" and your screen should look like [Figure 3•30](#). Click on **OK** and SPSS will create the new variable called WADULTS.

Now we want to weight the data using this variable we just created. Click on **Data** and then on **Weight Cases**. Click on the circle to the left of Weight cases by. Notice that this fills the circle in to indicate that it has been selected. Scroll down the list of variables on the left and find the variable WADULTS. Click on it to highlight it and then click on the arrow to the right of the list to move this variable into the Frequency Variable box. Your screen should look like [Figure 3•31](#). Click on **OK** and SPSS will weight the data appropriately.

Get a frequency distribution for the variable ADULTS using the weighted data. Click on **ANALYZE**, then point your mouse at **DESCRIPTIVE STATISTICS**, and then click on **FREQUENCIES**. Move the variable ADULTS into the Variables box and click on **OK**. The weighted frequency distribution should look like [Figure 3•32](#).

Notice that the frequencies are equal to the weighted number of cases produced by using the

adjusted weights we computed above. (Any differences would be due to rounding error.)

If you want to go back to the unweighted data, you will have to click on **Data** and then on **Weight cases**. Click on the circle to the left of **Do not weight cases** and then on **OK**. Now you are using the unweighted cases again.

## Summary

In this part of the book you have learned how to recode, create new variables using compute and if, select particular cases for analysis, and weight cases. You can do more complicated things with these commands than we have shown you, but these are the basics. You can use the *SPSS Base 13.0 User's Guide* (SPSS, Inc., 2004) to learn what else you can do with these commands. In the rest of this book, we will focus on some of the statistical procedures that SPSS can do for you.

## Chapter Three Exercises

Use the GSS02A data set for all these exercises.

### RECODE Exercises

1. 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.

2. INCOME98 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 a new variable INCOME1. Create new value labels for the recoded categories. Do another frequency distribution to make sure you recoded correctly.

Now recode INCOME98 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

3. 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?
  
4. 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

5. There are two variables that describe the highest educational degree of the respondent's father and mother (PADEG and MADEG). 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.
  
6. 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

7. 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?
  
8. Select all whites (1 on the variable RACE) and do a frequency distribution for the variable PRES00. Did they vote for Bush, Gore, or Nader in 2000? Then select all blacks (2 on the variable RACE) and do a frequency distribution for PRES00. Were whites or blacks more likely to vote for Gore?

---

[1]

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.

[2]

If your list shows labels, you can change the display. Check footnote 1 on page 5.

## Chapter Four: Univariate Statistics

Univariate analysis, looking at single variables, is typically the first procedure one does when examining first time data. 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 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 GSS02A file used in earlier chapters, so start SPSS and bring the file into the Data Editor. ( See Chapter 1 to refresh your memory on how to start SPSS). To begin the process start SPSS, then open the 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 in a nominal (category) data set 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 submenu choose **Frequencies** and select your variables for analysis. You can then choose statistics options, choose chart options, choose format options, and have SPSS calculate your request.

For this example we are going to check out attitudes on the abortion issue. The 2002 General Social Survey, GSS02A, has the variable **ABANY** with the label "ABORTION--FOR ANY REASON." We will look at this variable for our initial investigation.

### Choosing Frequencies Procedure:

From the **Analyze** menu, highlight **Descriptive Statistics**, [Figure 4•1](#), then move 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. [\[1\]](#)

### Selecting Variables for Analysis:

First select your variable from the main frequencies dialog box, [Figure 4•2](#), by clicking the variable name once. (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.

Click the arrow to the left of the Variable(s): box, [Figure 4•2](#), to move ABANY into the 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, bottom 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, top right, and we return to the main Frequencies dialog box, [Figure 4•2](#).

We could now click OK and 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.

### 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.

Now click OK on the main frequencies dialog box and 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 SPSS frequencies procedure. (Processing time for 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 SPSS to do and the CPU speed of your computer). The SPSS Output Navigator, left side, and the output, right side, will appear when 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 as we did in [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. To display the chart, drag the scroll bar on the right of your table. The variable label ABORTION•• FOR ANY REASON is displayed at the top of the chart. We see from the frequency distribution that there are more "no", 18.6%, answers than "yes", 14.0% 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, .8% (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, were 66.6% 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 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 **Edit** menu. Choose **Picture** in the paste special dialog box that appears and click **OK** to paste the chart into your document.

### Interpreting Frequency Output:

To enlarge 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 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 SPSS frequencies output for the variable ABANY. At the top is the variable label ABORTION•• FOR ANY REASON. The major part of the display shows the value labels (YES, NO, Total), and the missing categories NAP [Not Appropriate], DK [Don't Know], 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 18.6 % responded no. DK, don't know was chosen by .8 % and .0% were NA [Not Answered]. The 66.6 % "NAP" [Not Appropriate] was that portion of the sample that were 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
SEX	Respondent's gender.	1 or 2	Male, Female
AGE	Respondent's age at last birthday.	18, 19, 20, 21... 89, 98, 99	None needed
AGED	Should aged live with their children.	1, 2, 3, 0, 8, 9	A good idea, Depends, A bad idea NAP [Not Appropriate], DK [Don't Know], NA [Not Answered]
BIBLE	Feelings about the bible	1, 2, 3, 4, 0, 8, 9	Word of God, Inspired Word, Book of Fables, NAP, DK, NA

Understanding these allows you to intelligently customize SPSS for Windows so that it is easier for you to use. You can set 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 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](#) shows the listing as Variable Labels. This is the default setting when SPSS for Windows is installed. This example has the cursor on the variable ABPOOR so ABORTION-LOW INCOME-CAN'T AFFORD MORE CHILDREN (is displayed). You can change the listing however, so that you see only variable names as in [Figure 4-9](#). Changing this is a matter of personal taste. This chapter uses variable names, [Figure 4-9](#).

Changing the display option for the variable selection dialog box must be done *before* the data file is opened. If you have SPSS open with a data file click **File, New and Data** and the data editor will be cleared.

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 then **OK**. For this chapter we choose **Display Names** and **Alphabetical** so that variable names will be displayed alphabetically as in [Figure 4-9](#).

### Displaying Values, Value Labels or Both in Your Output

One other option you might want to make is in the table format for your 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, click the options dialog box and make your choices. My choices are seen in [Figure 4-11](#). The output resulting from my choices is [Figure 4-12](#).

## Descriptives

Descriptives (Analysis, 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.

### 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](#)).

### Selecting Variables for Analysis:

First click on AGE, the variable name for AGE OF RESPONDENT. Click the select arrow in the middle and 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. Click Options and the Descriptives: 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 SPSS will calculate and display the output seen in [Figure 4-16](#).

### Interpretation of the Descriptives Output

In the Interpretation of [Figure 4-16](#), AGE OF RESPONDENT has a mean of 46.28 and a standard deviation of 17.370. The youngest respondent was 18 and the oldest was 89. On your SPSS output HIGHEST YEAR OF SCHOOL COMPLETED, has a mean of 13.36 (a little more than 1 year beyond high school) and a standard deviation of 2.9. Some respondents indicated no "0" years of school completed. The most education reported was 20 years.

## 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 The dialog box should look like [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.

### 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.

### 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 SPSS will perform the chosen tasks and display the data in the SPSS Output Navigator.

### **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](#).

We can see that 2753 (99.6%) of our respondents answered this question. The other 12, .4% 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 2765 (100%).

The Descriptives statistics output should look like [Figure 4-22](#).

We can see all the typical descriptive statistics on this output: mean (13.26), lower bound (13.16) and upper bound (13.37) 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 13.14 and 13.36), median (13.00), variance (8.23), standard deviation (2.89), minimum (0), maximum (20), range (20), inter quartile range (4.00), skewness (.134), kurtosis (.781). A narrative explaining the education of first marriage for the US population in 2002 would be somewhat like the following:

Our sample from the General Social Survey of 2002, indicates that the average education for those over 18 in the US in 2002 was 13.26 years with a 95% confidence that the real average would fall between 13.17 and 13.37 years. The least years of education reported was found to be 0 and the most was 20. The exact middle point of the population with 50% falling below and 50% above, the median was 13.00.

The Extreme Values can be seen in [Figure 4-23](#). This Figure shows the five highest and the five lowest values for our variable. More than five respondents listed their years of education as 20. On the low end there was one with 0 education, two with 2 years of education and more than five people listed their education as 3 years in our sample. The Test of Normality is shown next (see [Figure 4-24](#)). This shows that this distribution is significantly different from the expected normal distribution. This is a pretty stringent test, most researchers would not require the distribution to be this close to normality.

The histogram, [Figure 4-25](#), shows a rough bell shaped distribution. SPSS divided our distribution into nine groups with a width of 2.5 years of education for each group.

The largest group has a little more than 1100 cases, a visual estimate. The smallest group has very few cases (we know there are a number of respondents who reported 3 years of education from our Extreme Values and the 2.5 bar). The statistics on the histogram tell us that the standard deviation is 2.87 with a mean of 13.3 for a total N of 2808.

The Stem•and•Leaf is next. [Figure 4-26](#), once again, shows a close but not quite normal distribution with significant outliers on the end of the distribution and a high number of observations above the mode. We saw this in our earlier output.

## Interpretation of the Q•Q Plot of Age:

Continue scrolling down the 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. 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. Many researchers would see this as close enough to treat as a normal distribution.

## Interpretation of the Boxplot:

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

Once again the major part of our distribution is not normal and 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 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 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 and/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 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:

- 1 is the code used for "Strongly Agree"
- 2 is the code used for "Agree"
- 3 is the code used for "Neutral"
- 4 is the code used for "Disagree"
- 5 is the code used for "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. The trick is to correct the errors as best possible.

You can have 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. Related to number 3 is that you can find 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 may 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 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 GSS02A 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 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 (FATHERS HIGHEST DEGREE)
- 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 SPSS output, just that output necessary to clearly and accurately describe your findings.)

3. The GSS02A file provides answers to a wide range of questions from a sample of respondents in the US in 2002 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 SPSS univariate procedure discussed in this chapter to describe the responses for these variables by this sample. Write a narrative description of your 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 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 2002 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.org/projects/gensoc.asp>.)

---

[1]

If you want to know more about a variable, the label, codes, etc. with the mouse pointer on the variable name in the variable list, right click the mouse button and choose **Variable Information**.

## Chapter Five: Cross Tabulations

In this chapter, we'll look at how SPSS for Windows 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 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. 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.tlo** under the **TableLook**, choices on the left side of the dialog box. Then, click on **Labels and Data** under **Adjust Column Widths** (see [Figure 5-2](#)). Then click on **OK**.

To illustrate the Crosstabs technique, we'll use the General Social Survey subset (GSS02A.sav). Open this file following the instructions in chapter 1 under "Getting a Data File."

Crosstabs are particularly useful in hypothesis testing such as 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](#).

You then choose the row (usually the dependent) variable and column (usually the independent) [\[1\]](#)

variable. 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 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**. SPSS will now open the Output Viewer window which will show you your table (see [Figure 5-6](#)).

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. The Valid N (number of cases) is used in the table.

The Crosstabs shows the 901 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 91.9% of the men and 91.2% of the women said Yes, a percentage point difference of only 0.7.

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 46.5% of the men and 42.9% 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 (92% compared with 45%), 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 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](#).

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 28 times out of 100. The Cramer's V of .036 in [Figure 5-10](#) (Symmetric Measures) indicates that there is virtually no relationship.

Let's look at a somewhat different table. For many years, scholars have observed that in the U.S., compared to other industrialized countries, social class has relatively little impact on political attitudes and behavior. To find this out, 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 INCOME98 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 **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.) Click on **Continue**, then on **OK**. What do the results show?

## 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 region of country, 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)?

---

[1]

The independent variable is the causal variable such as "gender" in the hypothesis that gender determines income.

## Chapter 6: 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 (abortion) 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 2002 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.

### 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 2,800 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, and 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](#). 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 SPSS what the two groups are 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](#).

This table shows you the mean age at birth of first child for men (25.39) and women (22.74) which is a mean difference of 2.641. 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 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.

## Paired-Samples T Test

We said we would look at an example where the samples are not independent. (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. It will automatically move it to the Current Selections box as Variable 1. Now click on PAEDUC to automatically move it to the Current Selections box as Variable 2. Click on the arrow to the left of the Paired Variables box to move this pair of variables into the box in the middle of the window. 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 (11.58) and by fathers (11.42), as well as the standard deviations. The t-value for the paired-samples t test is 2.219 and the 2-tailed significance value is 0.027. (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 less than .05, we reject the null hypothesis. There is a 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 .027 by 2 which would be .0135. For a one-tailed test, we would also reject the null hypothesis since the one-tailed significance value is less 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 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 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 73.643 or higher if the null hypothesis is true. Here the null hypothesis is that the mean age at birth 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.

There is another procedure in 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. If you want to use these procedures, consult the *SPSS Base 13.0 User's Guide* (SPSS, Inc., 2004).

## Summary

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 GSS02A data set for all these exercises.

1. Compute the mean age (AGE) of respondents who voted for Bush, Gore, and Nader (PRES00). 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 (INCOME98) 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) of respondents who voted for Bush, Gore, and Nader (PRES00). 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. (To get more information about how the occupational prestige scale was constructed, you can go to <http://www.csub.edu/~jgross/projects/spss/>)

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 one they are used to. SPSS includes another correlation test, Spearman's rho, that is designed to 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](#):

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  $.496$ , 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  $.496$ , 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  $.496$  suggests that the relationship between education and occupational prestige is actually quite strong.

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,  $.513$ , is nearly identical to coefficient obtained using the Pearson correlation. What do you conclude?

## 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, Scatter/Dot** and **Define** (we will use the default format, Simple). This will

open up the dialog box shown in [Figure 7•5](#). In the box on the left, click on EDUC then on the arrow key that is pointing toward the box labeled **X Axis** (because it is the independent variable in our hypothesis). Next, click on PRESTG80 and move it into the box labeled **Y Axis** (because it is the dependent variable). Your dialog box should look like the one in [Figure 7-5](#). Then, click **OK**.

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 **X** 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. Enter the number 2. When you finish, it should look like the one in [Figure 7-7](#). Then, click on **Apply** and then close the box.

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, distances of points 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 24.6% of the variation in occupational prestige in our sample. Recall that the Pearson coefficient was .496. If you square the Pearson coefficient ( $.496 \times .496$ ), you get .246 – 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.

Doing a regression analysis can help us to understand the Fit Line in more detail. Close the 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 .246. Where have you seen this before? What does it mean?

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 (SPSS reports it in this column for convenience only).

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 (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

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

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”):

$$\bullet = 12.557 + 2.334X$$

12.557, 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). 2.334 is the slope of the line. That is, if you refer back to the scatterplot, if you move one unit to the right on the X-axis, then move 2.565 units upward, you will intersect with the regression 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:

- =  $12.557 + 2.334(12)$
- = 40.565

We find that having 12 years of education is associated with an occupational prestige score of 40.565. But what of 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 40.565. This language points out to our readers that it is likely that some of those respondents scored higher and some lower, but that 40.62 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 one's 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 ( $\bullet - Y$ ).

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 2.334 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 2.334 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, 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 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 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 Five). 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 Crosstabulation, 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:

Values for PARTYID	Values for PARTYIDR	Value Label for PARTYIDR
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, 43.4% 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. Half (50.1%) of respondents who identify as Democrats are in favor of allowing abortion for any reason, compared to only about a third of those identifying as Republicans. 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 crosstabulation 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.

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.

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 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.
- 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 partials are less 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.

Try some of your own three-way (or higher) tables using some of the variables in the GSS02A 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, and 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 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 GSS02A 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.

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 HRSRELXS 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 .016. As you recall from Chapter 7, this means that 1.6% 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  $-.722$ . 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  $-.722$  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 report  $.722$  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  $.000$ . 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. Because most women spend more time doing child care and housework than most

men do, that might be one reason that they 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 INCOME98 into **Independent(s)** box as well. It should look like the one shown in [Figure 8a-7](#). Click **OK**.

Look at the B values in the Coefficients in the table [Figure 8b-7](#). None of our additional hypotheses are supported by the data. Looking at the coefficient for CHILDS, we see that the value is  $-.034$ . It is not 0, so there appears to be some relationship between number of children and hours of relaxation time in the data, but you see that the t statistic is not statistically significant, which means that we cannot generalize from the sample to the population from which it is drawn. Thus, although there is a non-zero coefficient, we cannot draw any inferences based on this result. In contrast, the coefficient for family income is statistically significant, however we had hypothesized that as family income increases, so does leisure time. But the B coefficient is negative. The coefficient tells us that with each additional movement of one unit upward in income (review a frequency distribution of INCOME98 to see what a unit increase means), leisure time goes down by  $.052$  hours. Finally, we hypothesized that controlling for the number of children and family income would “explain” some of the gender difference in the number of hours of relaxation time, which would result in a smaller B coefficient. Instead, notice that the coefficient is actually larger! How might we interpret this finding?

## Chapter Eight Exercises

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.

Choose three independent variables for the GSS02A data 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.

### **New Exercise - Not in the text**

Does education (educ), religiosity (measured by church attendance), and sex influence the number of children that someone has? Write a hypothesis for the relationship between each IV and the DV (you should have three hypotheses). Is each hypothesis supported? How does education influence the age of childbearing? How much change in age of childbearing occurs for a one year increase in education? Do people with higher degrees of religiosity have kids at younger or older ages? How do you know? Do women have children at a younger age than men? (Refer to a statistics text book for help with these questions) [Question contributed by Pat Jennings]

## Chapter 9: 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 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 **Gallery**. The Chart Gallery is a guide that SPSS 13 provides to help you to decide which type of chart fits your needs, and to show you the steps for obtaining the chart you choose. The Chart Gallery box is shown in [Figure 9-1](#). Point your mouse over some of the icons to see the names of the different chart types. It looks like a Bar Chart would be a good way to display DEGREE. Click on **Bar** in the Chart Gallery (the first icon on the left side). You will see a window that looks like the one in [Figure 9-2](#).

Click on the icon labeled **Simple** to see some examples of bar charts (see [Figure 9-3](#)). Since we are displaying only one variable, it looks like Chart A, "Summaries for Groups of Cases," is the best choice. It shows the number of cases which fall into each category (i.e. DEGREE value).

Scroll down to the descriptions of each chart type, and click the **How To** link under the title "A. Summaries for Groups of Cases." Here, you will see the steps to follow to create this type of chart (see [Figure 9-4](#)).

#### Creating Your Chart

Lets go through the steps in [Figure 9-4](#). On the SPSS Data Editor menu bar, click on **Graphs**, then **Bar**. You will see the box shown in [Figure 9-5](#).

You can see that **Simple** is already selected, because it has a black box around it. "Summaries for groups of cases" is also already selected. So, click on **Define** so that you can tell SPSS which variable you want to show in your chart. Find DEGREE in the list of variables on the left, then click the arrow below **Category Axis** to enter degree into the box. See [Figure 9-6](#) to see how it should look. (Note: In this exercise, you are going to create a chart that shows the number

of respondents in each category. If you preferred to show the percentage of respondents, you would click on “% of cases” in the **Bars Represent** section.)

Click **OK** to create the chart. Your chart should look like the one in [Figure 9-7](#).

You may make changes to the way your chart looks by opening it in its own window. Double-click anywhere in the chart, and it opens in the Chart Editor. Explore the menus in the Chart Editor to experiment with what you can do. Try this: click on **Options**, then **Title**. Two things will happen. The Properties dialog box will appear [see [Figure 9-8](#)], and you will also see the word “Title” contained in an outlined box at the top of your chart (see [Figure 9-9](#)). Simply type in the title you want for your chart. If you would like to change the font or the text color used for your chart title, do that in the Properties box. If you choose something other than the default, click **Apply**, then close the box by clicking on the **X** in the upper-right corner. If you are happy with the default, just close the box.

We could have chosen other chart types to display the data in the DEGREE variable. Experiment with some of the other chart types. You may want to explore the definitions of other charts in the Chart Gallery.

## Charting Several Variables: Abortion Attitudes

When you are working with several related variables, you may want to show them in one chart, rather than having separate charts for each. Let's show how responses to the questions about abortion attitudes vary with the condition under which abortion would be allowed. Let's also make it more interesting by showing the similarities and differences in men's and women's attitudes.

To obtain this chart, click on **Graphs**, then **Bar**. When the dialog box opens, click on **Clustered**. Doing this will “cluster” the data by respondents' gender. In the lower section, click on **Summaries of Separate Variables**. This will allow you to show the results for all of the ‘abortion attitude’ variables. The dialog box should look like the one in [Figure 9-10](#). Click **Define**.

A dialog box for entering your variables opens once you click Define. Move each of the ‘abortion attitude’ variables into the larger box (the one labeled “Bars Represent.” The order you add them into the box will determine the order they will print out in the chart. You may want to enter them in the following order ABHLTH, ABRAPE, ABDEFECT, ABNOMORE, ABPOOR, ABSINGLE, ABANY. Then, click on SEX, and move it into the space labeled Category Axis. At this point, your dialog box should look like the one shown in [Figure 9-11](#).

Notice that in the dialog box, the attitude variables are preceded with **Mean**. This tells you that the default is for the bars to represent the mean of each variable. If we were charting interval or ratio variables, such as education or income, this would make sense. However, the possible meaningful responses to the abortion questions are “yes” and “no.” It seems that reporting the percentage who answered “yes” would convey the information much more meaningfully. Do this by using the **Change Statistic** button. First, click on all of the abortion variables to highlight them (you will need to hold down the Shift key while you do this). Then, click the **Change Statistic** button.

The dialog box that opens gives you several choices about how you want your data to be summarized and displayed. Recall that for the abortion variables, a “Yes” response was coded as 1, and “No” was coded as 2. (If you aren’t sure how variable categories were coded, just look at the labels in the “Variable View.”) We can tell SPSS to display the percent who said “Yes” by clicking on Percent below, and entering the number 2 in the Value box. This way, only those responses coded with a value lower than 2 will be charted, which in this case, are the “Yes” responses. When you finish, your dialog box should look like the one in [Figure 9-12](#). Click **Continue**, then **OK**. Your chart should look like the one shown in [Figure 9-13](#).

## Tables

Let's create a crosstabulation 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 crosstabulations, 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**. The Output Window will appear, and your screen should look like [Figure 9•14](#).

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 Crosstabulation**. 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 and it will have a series of cross-hatches on it. 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•15](#)).

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 9-16](#). 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 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 object. Then, go to your word-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 Objects**. Switch over to your word-processing document. Click the mouse where you want to paste your table. Click on **Paste Special** and then **OK**.

The method for copying and pasting charts is exactly the same as the method as for copying and pasting tables. Note that in both cases, what you are pasting into your word processing document is an object: a graphic (or picture). This means that you can resize it to make it bigger or smaller, but you cannot edit it.

## 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. Are happily married people more trusting? Create a “clustered” bar chart which shows levels of TRUST grouped by HAPMAR (hint: the HAPMAR categories should appear along the x-axis, and the TRUST categories should appear in the legend). This time, have the bars show the percentages of respondents in each category, rather than the number (so, click on “% of cases” in the “Bars Represent” section of the dialog box). Show the % represented by each bar, and make sure the chart is properly titled and labeled. Copy and paste the chart into a word processing file. Write a few sentences that describe the patterns

shown in the chart.

3. 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 discusses 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 2002 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 doubled from about 1,500 to nearly 3,000. Many questions are asked on each survey, while other questions are rotated from survey to survey. This subset from the 2002 GSS includes all the cases (2,765) and 74 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
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
HRSRELAX	Hours per day R has to relax
INCOME98	Total family income
INC98REC	Recoded total family income
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
PRES00	Vote for Gore, Bush, Nader in 2000
PRES96	Vote for Clinton, Dole, Perot in 1996
PRESTG80	Respondent's occupational prestige score (1980)
RACE	Race of respondent
RELIG	Respondent's 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
ZODIAC	Respondent's astrological sign