# **Chapter 1**

# **Introduction**

[PSPP](https://www.gnu.org/software/pspp/) is a statistical analysis program made available at no charge to users by the *Free Software Foundation*. There are two versions: the “syntax” version and the less comprehensive but more user-friendly GUI (Graphic User Interface) version. This tutorial, written primarily for beginning students, describes the GUI version.

PSPP is similar to [SPSS.](https://www.ibm.com/analytics/spss-statistics-software) SPSS (officially IBM SPSS Statistics) has, through many iterations and over many years, served as a widely used standard for analyzing quantitative data. The authors of this tutorial have used SPSS in their teaching and research, and continue to consider it extremely useful. It has become increasingly so for undergraduate instruction as, over time, it has become much more user-friendly, so that even students with no background in statistical analysis can master it as part of a single introductory research methods course. It has also added more and more features, notably including the ability to produce a wide range of graphs.

Despite its many advantages, one thing that SPSS is not is *free*. As of this writing, the cost of a base subscription to SPSS starts at $99 *per month*! This need not trouble you if you are a student at a college or university that has purchased a site license. Even if this is not the case, you can obtain a special version available only to faculty and students at a deep discount. But still not free. (Check Amazon.com or other vendors for details.)

In this tutorial, we’ll be using a subset of the General Social Survey (GSS) for use in conjunction with the software. The GSS is a biannual national survey conducted by the National Opinion Research Center, and used for teaching and research in a variety of disciplines since 1972. We’ve created a subset of the 2018 survey. A version that can be read by PSPP can be downloaded here.

# **The Rest of the Tutorial**

In Chapters 2 through 7, we’ll show you how to use PSPP. Chapter 8 will show you how to incorporate output from PSPP into reports. Here’s a more detailed breakdown of the remaining chapters.

* Chapter 2 – Creating, Saving, and Opening Data Files
  + Creating your own data file
  + Variable names
  + Variable labels
  + Value labels
  + Missing values
  + Entering the data definitions
  + Entering the data values and saving the file
  + Opening SPSS files in PSPP
  + Opening an Excel file in PSPP
  + Next chapter
* Chapter 3 – Transforming Data
  + Weighting cases
  + Recoding variables
  + Creating new variables using compute
  + Next chapter
* Chapter 4 – Univariate Analysis
  + Frequencies
  + Descriptives
  + Charts and graphs
  + Explore
  + Next chapter
* Chapter 5 – Comparing Means
  + Working with means
  + Introduction to hypothesis testing
  + One-sample t test
  + Independent-samples t test
  + Paired-samples t test
  + One-way analysis of variance (ANOVA)
  + Next chapter
* Chapter 6 – Crosstabulation
  + Two-variable contingency tables
  + Three-variable contingency tables
  + Next chapter
* Chapter 7 – Correlation and Regression
  + The correlation coefficient
  + Regression
  + Next chapter
* Chapter 8 – Writing Research Reports
  + An outline of your research report
  + Tables
  + Footnotes or endnotes?
  + Citing articles, papers, and other material
  + Plagiarism
  + Proofreading
  + Other guides to writing reports
* Appendix A – Codebook for data used in this tutorial
* Appendix B – Working with PSPP

# 

# **Other Resources**

For another tutorial on the Graphic User Interface (GUI) version of PSPP, see Gary Fisk, [*PSPP for Beginners*](https://garyfisk.com/pspp/)

GNU has produced a comprehensive [users’guide](https://www.gnu.org/software/pspp/manual/) for the command (syntax) version of PSPP. Available in various formats.

The GUI version of PSPP is perhaps most limited in its very minimal coverage of graphics, offering only pie charts, bar charts, histograms, and scatterplots, and these only with very few options. Fortunately, another package, also freely available, is [*Statistics Open for All*](http://sofastatistics.com/home.php) *(SOFA),* which includes much more extensive graphics capabilities.George Self has developed a comprehensive lab manual for this package. Designed for use by his own students, he has not published it on the Internet. He has, however, generously granted us permission to post it [here](https://ssric.org/files/2019-10/G_SELF_LabManual.pdf) on the site of the Cal State University Social Science Research and Instructional Center. He has copyrighted his manual under the Creative Commons “Attribution-ShareAlike 4.0 International license,” which is even more open than the “Attribution-NonCommercial-ShareAlike 3.0 license” which, except where noted, governs our own site.

# **About the Authors**

Edited by Ed Nelson (Sociology, California State University Fresno; [ednelson@csufresno.edu](mailto:dnelson@csufresno.edu)) and John Korey (Political Science, California State Polytechnic University, Pomona; [jlkorey@cpp.edu](mailto:jlkorey@cpp.edu)). Ed is the primary author of Chapters 2 through 5, and Chapter 8. John is the primary author of Chapters 6 and 7.

## **Next Chapter**

Chapter 2 will discuss how to create a PSPP data file and how to open PSPP data files and files in other formats such as Excel, SAS, and Stata.

**Chapter 2**

**Creating, Saving, and Opening Data Files**

If you don’t have PSPP installed on your computer, now is the time to do it. The easiest way to download PSPP is by clicking on this [link](http://pspp.awardspace.info/) and looking for the “Downloads” box. Then download the latest version in either 32-bit or 64-bit format. If you’re not sure which version to download, go to the control panel and click on “System” and look for your system type. Then follow the instructions to download and install the program.

**Creating Your Own Data File in PSPP**

Once you have PSPP installed and opened on your computer, the first thing you need to do is to either create your own data file or open an existing data file. Let’s start by assuming that you have data that you have collected and you want to analyze the data in PSPP. So, you need to create your own data file. As an example, let’s assume your data includes the following variables.

* Case identification number (i.e., a unique identification number for each case in your data) (*id*)
* Support or oppose same-sex marriage (*same\_sex\_marriage*)
  + 1 = Support same-sex marriage
  + 2 = Oppose same-sex marriage
  + 3 = Undecided
  + 9 = Don’t know or refuse to answer
* Political party preference (*partyid*)
  + 1 = Democrat
  + 2 = Republican
  + 3 = Independent
  + 4 = Other party
  + 9 = Don’t know or refuse to answer
* Political views (*polviews*)
  + 1 = Conservative
  + 2 = Middle-of-the-road
  + 3 = Liberal
  + 9 = Don’t know or refuse to answer
* Age in years (*age*)
  + 98 = 98 or older
  + 99 = Don’t know or refuse to answer
* Subjective social class (*class*)
  + 1 = Upper
  + 2 = Middle
  + 3 = Lower
  + 9 = Don’t know or refuse to answer
* Gender (*gender*)
  + 1 = Male
  + 2 = Female
  + 9 = Refuse to answer
* Education (*educ*)
  + 1 = Less than high school
  + 2 = High school degree
  + 3 = Some college
  + 4 = Bachelor’s degree
  + 5 = Some postgraduate work
  + 6 = Postgraduate degree
  + 9 = Don’t know or refuse to answer

Notice that all these variables are numeric. While PSPP can handle string variables, most of the time we use numeric variables. String variables may contain letters, numbers, and special characters while numeric variables contain only numbers. We’ll only consider numeric variables in this tutorial.

Notice also that we always allow for missing data. For various reasons, some information may not be available. In the case of a survey, this may be because respondents don’t know the answer or it may be that they don’t want to answer. We’ll have more to say later about how PSPP handles missing data, but for now keep in mind that we always have one or more codes to account for missing data.

Before we enter the data, we’re going to give each variable a name. Each variable has codes to account for the different ways respondents answer the questions. These are called values. For example, political views has four values – 1, 2, 3, and 9. Each value can be assigned a value label. Value 1 could be assigned the value label **conservative** and value 2 the label **middle-of-the-road**. Most variables will have one or more codes for missing data. For political views, 9 is our missing value code and could be assigned the label **don’t know or refuse to answer**.

**Variable Names**

Each variable must have a unique name. No two variables can have the same name. Names can be as long as 64 characters but it’s advisable to use relatively short names. Here are some simple rules to follow in naming your variables.

* Start with a letter.
* Use letters and/or numbers in the variable name.
* The underscore (\_) can be used to separate characters if desired.
* There can be no blank spaces in your variable names.
* Letters can be either upper or lower case.

Look back at the list of variables in the example on the first page of this chapter. Possible variable names are in parentheses. Note that the parentheses are not part of the variable name.

**Variable Labels**

Variable names are typically short and sometimes don’t supply much information about the variable. Sometimes users use variable names like *q1* or *var1*. To make the nature of the variable clearer, you can create a variable label that can be up to 256 characters. In our example, the variable named *partyid* could be given the label **political party preference**. Variable labels can contain letters (lower or upper case), numbers, special characters, and blank spaces. Variable labels are optional.

**Value Labels**

Values are the numbers that you use to represent different characteristics of the case. In our example, for the variable *partyid*, the values are 1, 2, 3, 4, and 9. For the variable *gender*, the values are 1, 2, and 9.

To tell the user what these values stand for we could give each value an extended value label. For the variable *partyid*, 1 could be given the label **Democrat** and 2 the label **Republican**. Value labels can contain letters (lower and upper case), numbers, special characters, and blank spaces. Value labels are optional.

**Missing Values**

Sometimes the information for a particular variable is unavailable. This can be for a number of reasons. If the cases are respondents to a survey, the respondent may not know how to respond to a particular question. If the question asks for the respondents’ yearly family income, they may not know their income. Another possibility is that they don’t want to tell us their income. If the cases are geographical areas such as counties or states, a particular piece of information might be unavailable. If the variable describes the violent crime rate of the area, the information might be unavailable for various reasons.

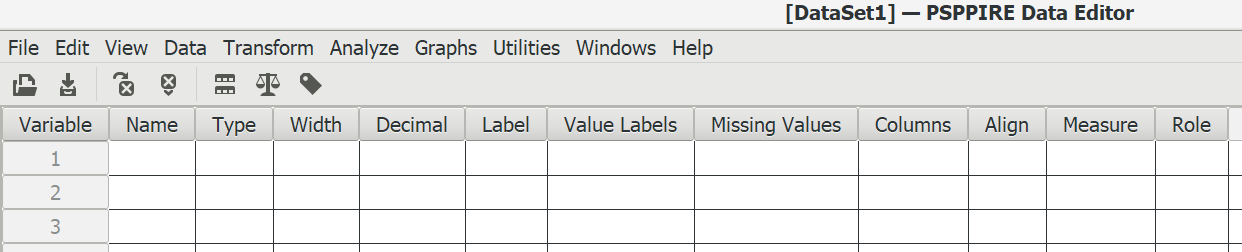
There are two types of missing values -- user-defined missing values and system missing values. In our example, 9 is the missing value for *partyid* and 99 is the missing value for *age*. There can also be more than one missing value. For example, we might want to use 8 for don’t know and 9 for refused. PSPP limits you to three missing value specifications.

Sometimes the user may use a blank space for missing information. PSPP automatically treats blank spaces as missing values. This is referred to as a system missing value. There are other examples of system missing values that we will discuss later.

**Entering the Data Definitions**

Think of the variable names, variable labels, value labels, and missing values as information that defines your data. How do we actually enter that information into PSPP?

PSPP opens in one of two views – data view or variable view. Data view displays the values or the value labels for the cases and the variables in your data set. It’s more convenient to display the values. If PSPP displays the VALUE LABELS, click on VIEW and uncheck the VALUE LABELS box and now it will display the values. Variable view displays the information that defines your data. Click on the variable view tab at the bottom of the screen. You should see the following.

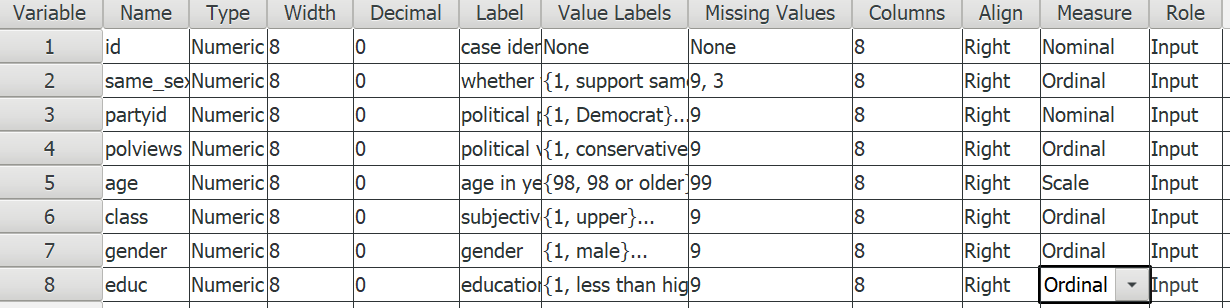


**Figure 2-1**

We can enter the various data definitions into this matrix.

* Enter the variable names for all eight variables in the example at the beginning of this chapter into the NAME box. Once you enter the variable name PSPP will enter the default values into some of the remaining cells. These defaults can be edited. After you have entered the variable name, press the Enter key to move down to the next variable.
* Enter the variable labels into the LABEL box and click ENTER to go to the next variable label.
* Enter the value labels into the VALUE LABELS box. Click in the far right-hand part of the box and a dialog box will pop up. Enter the value in the VALUE box and then enter the value label in the VALUE LABEL box. If you want to make a change, click on the label you want to change, make your change, and then click on APPLY. If you want to delete the label, click on REMOVE. When you are done, click on OK.
* Enter the missing values into the MISSING VALUES box. Click in the far right-hand part of the box and a dialog box will pop up. You can enter up to three different values (e.g., 9, 99) or you can enter one range of values and one value. Notice that the default is no missing values. When you are done, click on OK.
* There’s one other box that deserves our attention – the MEASURE[[1]](#footnote-1) box. Click anywhere in that cell. Now click on the drop-down arrow and you will have three choices – scale, ordinal, and nominal. Notice that scale is the default.
  + Scale refers to a continuous variable such as *age*. In a continuous variable, the values have the properties of real numbers. They can be added, subtracted, divided, and multiplied like real numbers.
  + Ordinal refers to categories that have an inherent order to them. Some categories are higher or lower than other categories. But you can’t treat them like real numbers. All you can say is that some categories are higher and others are lower. For example, think of social class. Upper class is higher that middle class and middle class is higher than lower class. We can use numbers to represent these different categories. But we can’t carry out mathematical operations such as addition and subtraction with them.
  + Nominal refers to categories that have no inherent order to them. For example, political party preference has four categories – Democrat, Republican, Independent, other. We can’t say that one category is higher or lower than another category. All we can do is to say they are different.
  + Treat dichotomies as ordinal.
  + Enter the type of measure for each variable.
* You can probably use the default values for the other columns in Figure 2-1. You might want to change the decimal value. The default value is two for all variables. If your values are integers, you might want to change the decimal value to zero.

Once you have filled in all the cells in Figure 2-1, your matrix should look like Figure 2-2.



**Figure 2-2**

**Entering the Data Values and Saving the File**

Now that you have defined all your variables, it’s time to start entering the data values for each case. One way to do this is to enter them in PSPP. Click on the DATA VIEW tab at the bottom of your screen. Notice that PSPP has filled in the variable names at the top. The variables are in the columns of the matrix and the cases are in the rows. Since this is a hypothetical data set, make up values for four cases and enter them into the matrix. Make sure that you are entering values that are within the ranges that you specified. Include some missing values as well.

All that is left now is to save the data file. Click on FILE and then on SAVE AS. Browse to where you want to save the file on your computer and enter the file name at the top of the screen. Press enter and PSPP has saved your data file. Open your file manager and make sure you saved it where you want to store it. Now close PSPP. PSPP will save your file as a .sav file.

Here’s how you can open the data file in PSPP that you just created.

* Open PSPP by clicking on the PSPP icon.
* Click on FILE in the menu bar at the top.
* Click on OPEN in the drop-down menu.
* Browse to the location on your computer where the .sav file was saved[[2]](#footnote-2) and double-click on it.

Some users prefer to enter their data in Microsoft Excel instead of PSPP. To do this, open Excel on your computer. Use the first row on your spreadsheet for the variable names. Then starting with row 2, enter the values for each case. Once you have entered the values for all the cases, save your Excel file as either a comma delimited file (.csv) or a tab delimited file (.txt) to wherever you want to store it on your computer. Later in this chapter we’ll show you how to open an Excel file in PSPP. The variable names will be entered in the NAME column on the VARIABLE VIEW tab. You’ll have to enter the other data definitions yourself.

**Opening SPSS Files in PSPP**

Another possibility is that you already have a data file that has been saved as a SPSS data file.[[3]](#footnote-3) SPSS has two different types of data files -- .sav and .por. PSPP will open both types of files.[[4]](#footnote-4) Here’s how to open a SPSS data file.

* Open PSPP.
* Click on FILE in the menu bar at the top of your screen.
* Click on OPEN.
* Browse to where your .sav or .por file is located.
* Double click on the file name.

**Opening an Excel File in PSPP**

Still another possibility is that you have a data file that was saved as an Excel file. It’s **essential** to keep in mind that PSPP can only **open** Excel data files that were saved as either a comma delimited file (.csv) or a tab delimited file (.txt). Here’s how you would do that.

* Open PSPP by clicking on the PSPP icon.
* Click on FILE in the menu bar at the top of the screen.
* Click on IMPORT DATA in the drop-down menu.
* Browse to where the .csv or .txt file is located on your computer and highlight the file name. Now click NEXT at the bottom-right of your screen.
* Select the lines of data that you want to open. Normally this would be all cases. Then click NEXT.
* Select the first line of the file that contains data and then click NEXT. Normally this would be the second line of your data file since the first line will contain the variable names.
* The next screen asks for the characters that separate your variables. This should be filled in for you. For example, in a comma separated file (.csv), this would be the comma and in a tab delimited file, it would be the tab. Click NEXT.
* The next screen allows you to adjust the variable formats (which you shouldn’t have to do). Click APPLY.
* Your file should now open in PSPP. You can save it, so you don’t have to go through this the next time you open the file.

**Next Chapter**

Chapter 3 will discuss various ways that you can transform your data. These include WEIGHT to weight your cases, RECODE to combine categories of your variables, and COMPUTE to create new variables out of existing variables.

# **Chapter 3**

# **Transforming Data**

This chapter explains how to change, or transform, the values associated with your variables. PSPP can transform the values in several ways by:

* weighting cases so that some cases count more heavily than others,
* combining values of a variable into several categories, and
* creating new variables out of old variables.

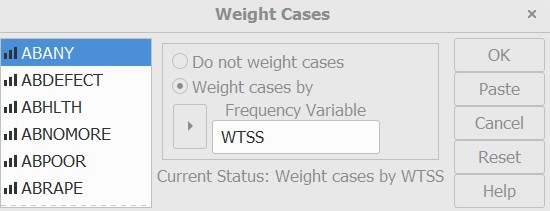
These procedures are called WEIGHT, RECODE, and COMPUTE.

We’re going to use the GSS18A data file in this chapter. See Chapter 2 if you need a refresher on how to do this.

## **Weighting Cases**

Sometimes, especially when analyzing survey data, you need to weight the cases in your data file so some cases count more heavily than others. If you get your data from an archive such as the Inter-university Consortium for Political and Social Research at the University of Michigan or the Roper Center for Public Opinion Research at Cornell University, most of the time there will be a weight variable. Weighting is typically used to make the data more closely represent the population from which the sample was selected.[[5]](#footnote-5) Be sure to read the codebook and user notes to see which weight variable to use (sometimes there will be more than one).

Once you have determined which weight variable you should use, it’s easy to weight the data. Click on DATA in the menu bar at the top of the screen, then click on WEIGHT. Choose WEIGHT CASES BY and then select the variable you want to weight the cases by and move it to the box below (see Figure 3-1). In the GSS18A data file you will be using in this chapter we have already weighted by the variable *wtss*.



**Figure 3.1**

You’ll notice as you work through the examples in this chapter that the frequencies are typically decimals. So, for example, the frequencies might be something like 17.23. Don’t worry about this. It’s a result of the weighting procedure.

## **Recoding variables**

Recoding Is a way of combining the values of a variable into fewer categories. Let’s say you have conducted a survey and one of the questions in your survey was the age of the respondent. Entering the actual age in years would be the simplest way to record the information in your data file. But let’s say that you want to compare respondents of different age categories such as young, middle young, middle old, and older. Using PSPP you could recode the data into these four 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 6), you probably want to limit the number of categories to three or four. The second thing you need to know is which of the old values are going to be combined into the new categories. For example, you might do something like this.

|  |  |
| --- | --- |
| The actual age of the respondent as originally recorded in the data file | The new, collapsed, categories |
| 18 through 29 years | Young |
| 30 through 49 years | Middle young |
| 50 years through 69 years | Middle old |
| 70 years or over | Older |

Or perhaps you asked respondents how often they attend church services, and the original categories were more than once a week, once a week, several times a month, once a month, several times a year, once a year, and never. With recoding we can combine these answers into four categories that we might call often, sometimes, infrequently, and never.

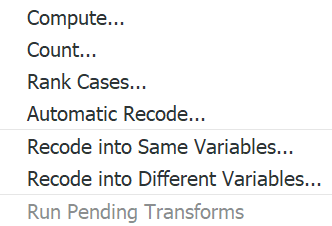
|  |  |
| --- | --- |
| The original response categories | The new, collapsed, categories |
| More than once a week, once a week | Often |
| Several times a month, once a month | Sometimes |
| Several times a year, once a year | Infrequently |
| Never | Never |

Recoding can be tricky sometimes. There are several things to keep in mind.

* You want to avoid having so few categories that you lose important information. For example, you probably wouldn’t want to recode age into only two categories – under 50 and 50 and over.
* You also want to avoid having so many categories that it gets confusing. Recoding age into two-year categories (e.g., 18-19, 20-21, etc.) probably isn’t a good idea either.
* Another problem to avoid is have too few cases in a category. For example, having an age category such as over 85 would result in too few cases for analysis.

Recoding is the process in PSPP that will carry out these examples. We’ll start by recoding the variable *age*.

Click on TRANSFORM. The drop-down menu will look like Figure 3.2.

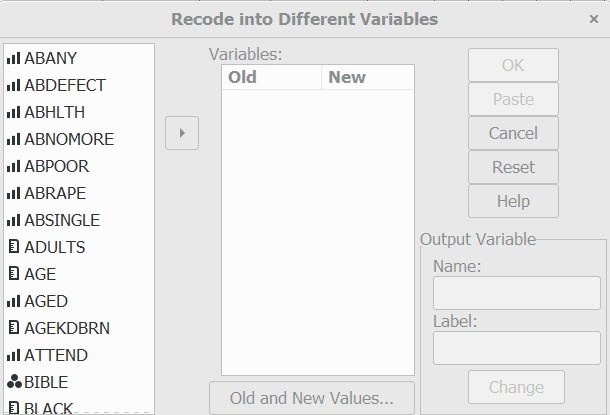


**Figure 3-2**

Now we have two recode options: Recode Into Different VARIABLES and Recode Into Same Variables. We strongly suggest that you only use the Recoding Into Different Variables option. One reason is that If you make an error, your original variable is still in the file and you can try again. Another reason is that you might want to recode into several sets of categories and then see which set works best.

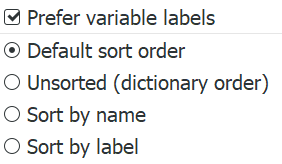
### **Recoding into Different Variables**

Recoding into a different variable starts with giving the new variable a name.[[6]](#footnote-6) For example, if we want to recode *age* into a different variable, we could call this new variable *age1* and then recode *age* into a different set of categories called *age2*. You should already have opened the GSS18B data file in PSPP. Click on TRANSFORM and then click on Recode Into Different Variables. Your screen should like Figure 3-3.



**Figure 3-3**

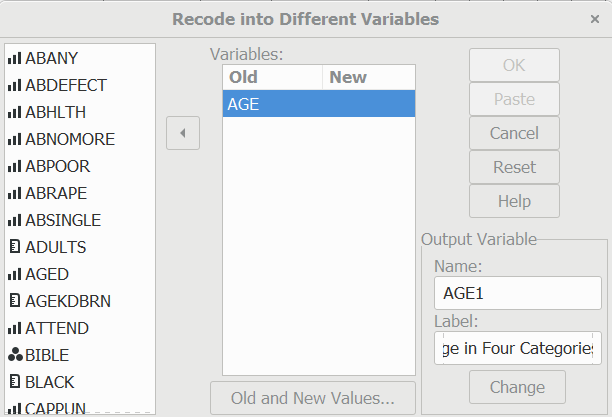
The variables are listed in the left-hand pane. If PSPP displays the variable labels instead of the variable names, right-click on the list and you should see Figure 3-4.



**Figure 3-4**

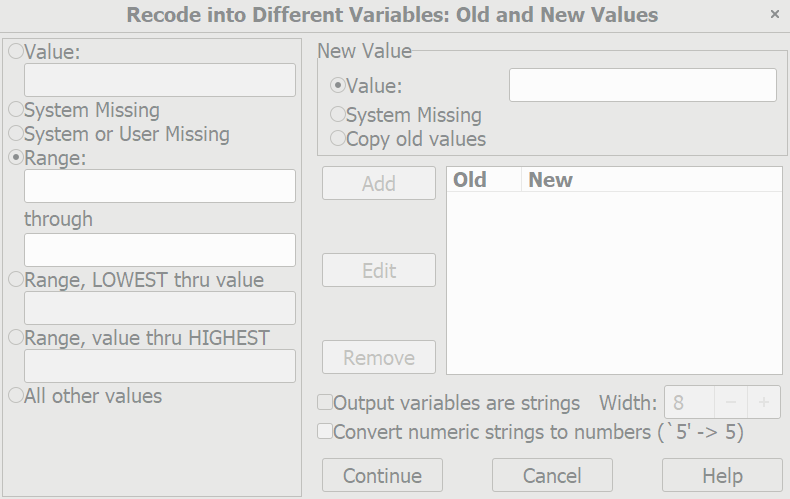
Uncheck the PREFER VARIABLE LABELS box by clicking on it and you should see Figure 3-3. Find *age* in the list of variables, click on it to highlight it, and then click on the arrow just to the left of the VARIABLES box in the middle of the dialog box. This will move *age* into the list of variables to recode.[[7]](#footnote-7)

You want to give a name to this new variable. Click on *age* in the VARIABLES box (i.e., that’s the big box in the middle, not the list on the left) and enter the new name,*age1*, in the NAME box (e.g., located under OUTPUT VARIABLE in the lower-right). You can also enter a variable label for this new variable in the box just below the NAME box. Try typing **Age in Four Categories** as your label. Your screen should look like Figure 3-5.



**Figure 3-5**

Click on the CHANGE button in the lower-right of the dialog box. Notice that the new variable name, *age1*, now appears in the VARIABLES box under NEW. Now we have to tell PSPP how to create these four categories. Click on the OLD AND NEW VALUES button at the bottom of the window. Your screen will look like Figure 3-6.



**Figure 3-6**

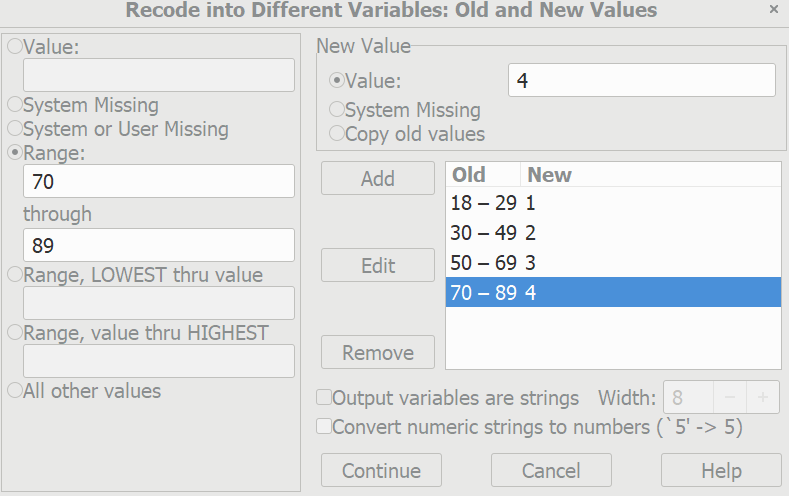
There are several options. You can change a particular value into a new value by entering the value to be changed in the VALUE box and the new value into the NEW VALUE box and then clicking on ADD. For example, if you wanted to change 1 to 2, you would enter 1 in the VALUE box and 2 in the NEW VALUE box.

In our example, we want to change a range of values to a new value. Click on the fourth bubble from the top labelled RANGE. Notice how this marks your choice by filling in the bubble. Then type **18** (the youngest age in the data file) in the box below RANGE and **29** in the box below THROUGH. Enter **1** in the NEW VALUE box to the right. This will tell PSPP to combine all ages from 18 through 29 into a single category and give it the value of 1. Now click on ADD.[[8]](#footnote-8)

Repeat this process for the other categories. To enter a new recode, you will have to delete what you entered previously and then enter the new recode. Enter the following recodes.

* 30 through 49 should be changed to 2
* 50 through 69 becomes 3
* 70 through 89 (the oldest age in the data file)[[9]](#footnote-9) becomes 4.

Your screen should look like Figure 3-7.



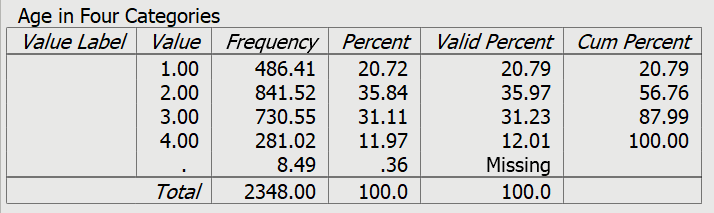
**Figure 3-7**

To change one of your categories, highlight that category in the OLD AND NEW VALUESbox, make the changes, and then click on EDIT. The revised category should appear in the box on the right. To remove a category, highlight it, and click on REMOVE.

Now we want to tell PSPP to carry out the recoding. Click on CONTINUE at the bottom of the screen. This will take you back to the RECORD INTO DIFFERENT VARIABLES box. Click on OK and PSPP will carry out your recode commands. PSPP will open an Output window and show you the commands it just executed.

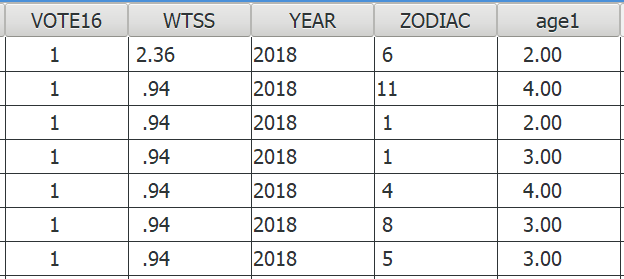
PSPP will open several windows in the course of a session. Click on WINDOWS in the menu bar at the top of the screen. There should be two windows open – the Output window and the data file window. Whichever window is open at the moment will have a check in the small box on the left. Check the box for the data window and you should now see the data window.

Click on ANALYZE in the menu bar at the top of your screen, point your mouse at DESCRIPTIVE STATISTICS, and then click on FREQUENCIES. (Frequency tables are discussed in more detail in Chapter 4.) If PSPP lists the variables by variable label, right-click on the list and uncheck the box that says PREFER VARIABLE LABELS. Note that *age1* has appeared in the list of variables on the left. Click on it to highlight it and then click on the arrow to move it to the VARIABLE box. Now click on OK. To look at your output, click on WINDOWS at the top of the screen and then click on OUTPUT. The output window will open, and your screen should look like Figure 3-8.[[10]](#footnote-10)



**Figure 3-8**

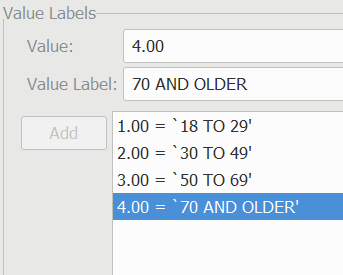
Let’s take a look at the data file. Click on WINDOW in the menu bar at the top of the screen and then on the data window. Use the scroll bar in the lower-right of the window to scroll to the right until you see a column titled *age1*. (It will be the last column in the matrix.) This is the new variable you just created. Your screen should look like Figure 3-9.



**Figure 3-9**

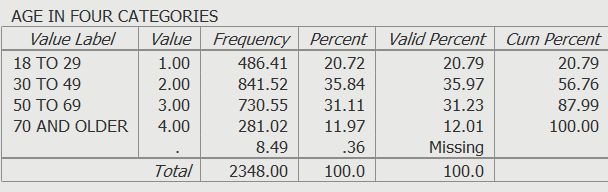
Look back at Figure 3-8 and you will see that there are no value labels for categories 1 through 4 for the new variable *age1*. If you want to add value labels, point your mouse at the variable name at the top of the column (*age1*) in the data window and double click. This will open the VARIABLE VIEW tab. Now you’re going to enter labels for the values in the recoded variable using what you learned in Chapter 2.

Double-click in the VALUES LABELS box for *age1* and PSPP will open a box where you can enter your value labels. At the top of this box you will see the VALUE and VALUE LABEL boxes. Click in the VALUE box and enter the value 1. Then click in VALUE LABEL box and enter the label for the first category, **18 to 29**. Click on ADD and the new label will appear in the box below. Now enter the labels for the other values. If you made a mistake you can click on the label and make the necessary changes, and then click on APPLY. If you want to delete a label, click on it and then click on REMOVE. Your screen should look like Figure 3-10.



**Figure 3-10**

Click on OK. Now click on ANALYZE, point your mouse at DESCRIPTIVE STATISTICS, and then click on FREQUENCIES and rerun the frequencies for *age1*. This time it should include the value labels you just entered. Your screen should look like Figure 3-11.



**Figure 3-11**

We said that recoding into different variables allowed you to recode a variable in more than one way. Recode *AGE* again, but this time recode into three categories – 18 through 34, 35 to 59, and 60 and over. We’ll call this new variable *age2*. Click on RESET in the RECODE INTO DIFFERENT VARIABLES box to get rid of the recoding instructions for *age1*. When you are done, run a frequency distribution for *age2*.

Let’s look at one more example before we close this section. The variable *race* is coded into three categories – white (value 1), black (value 2) and other (value 3). If we combined values 2 and 3 and left value 1 unrecoded we would have two categories – white (value 1) and non-white (values 2 and 3). We would have converted race which is a nominal variable[[11]](#footnote-11) into a dichotomy. Go ahead and recode race in this manner and call your new variable *race1*. Assign a variable label and value labels to your recoded variable. Then run a frequency distribution and make sure you did the recoding correctly.

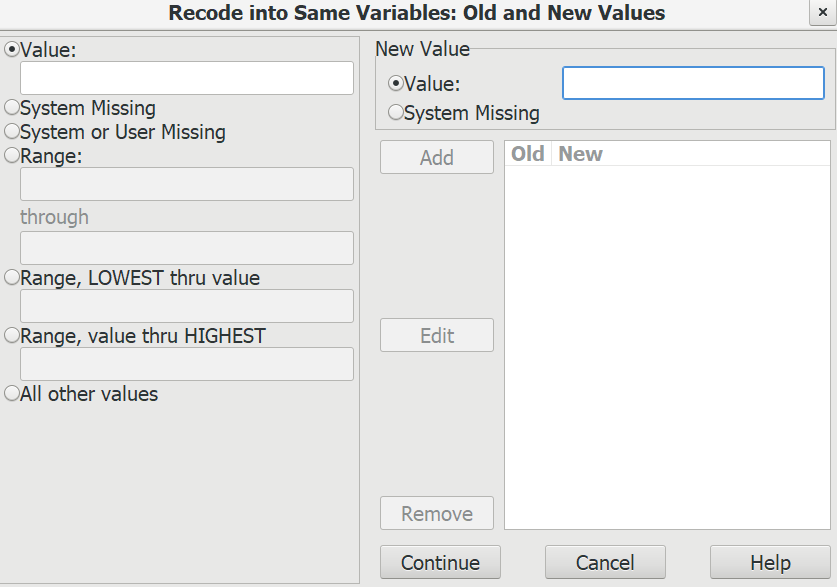
There are two more important points to discuss. Look back at Figure 3‑6. It shows the RECODE INTO DIFFERENT VARIABLES: OLD AND NEW VALUES box. There are three options in the VALUE box that we haven't discussed. Two are different ways of entering ranges. You can enter the lowest value of the variable 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, since 99 is the missing value for age, 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? Any 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 VALUE box and click on COPY OLD VALUES in the NEW VALUE box and then click on ADD.

### **Recoding Into the Same Variable**

Earlier we suggested that you always recode into a different variable but in case you want to recode into the same variable, we’re going to show you how to do it. This means that the recoded variable replaces the original variable, but remember the warning given you earlier in this chapter.

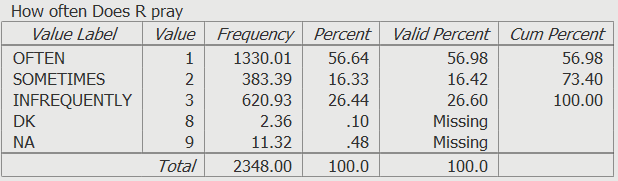
Click on TRANSFORM and then click on RECODE INTO SAME VARIABLES. Let's recode the variable called *prey*. Find *pray* on the list of variables on the left, click on it to highlight it, and then click on the arrow to the left of the VARIABLE box. This will move the variable *pray* into the big box in the middle of the window. Click on the OLD AND NEW VALUES button. This will open the RECODE INTO SAME VARIABLES: OLD AND NEW VALUES box. Your screen should look like Figure 3‑12.



**Figure 3-12**

This looks very much like the box you just used (see Figure 3‑6). Combine the values 1 and 2 by clicking on the fourth circle from the top under OLD VALUE and entering a **1** in the box below RANGE and a **2** in the box below THROUGH and then entering a **1** in the NEW VALUE box and then clicking on ADD. Now combine values 3 and 4 into a category called **2** and combine values 5 and 6 into a third category called **3**. Click on CONTINUE and then on OK. 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 window by scrolling to that variable. Point your mouse at the variable name (*pray*) and double click. This will open the VARIABLE VIEW window. Double click in the VALUES LABELS box and change the labels to whatever label you want to assign to the values of the recoded variable. You could call them something like often, sometimes, and infrequently. After you have entered the change you want make to the label, click on APPLY to actually make the change. To remove a label, click on the REMOVE button. Follow the instructions in the previous section on recoding into different variables. When you finish, click on OK. To see what your recoded variable looks like, click on ANALYZE, then point your mouse at DESCRIPTIVE STATISTICS, click on FREQUENCIES and move *prayer* over to the VARIABLES box and click on OK. Your screen should look like Figure 3‑13.



**Figure 3-13**

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.

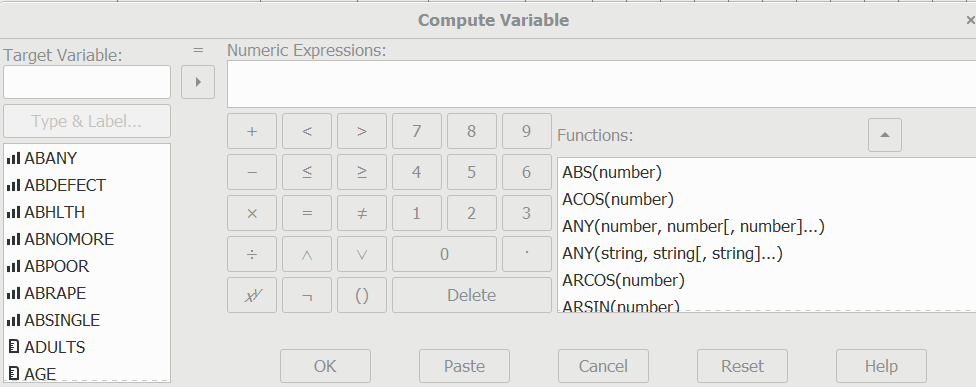
Recoding is a very useful procedure and one that you will probably use a lot. It's worth spending time practicing how to recode so you will be able to do it with ease when the time comes.

## **Creating New Variables Using Compute**

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

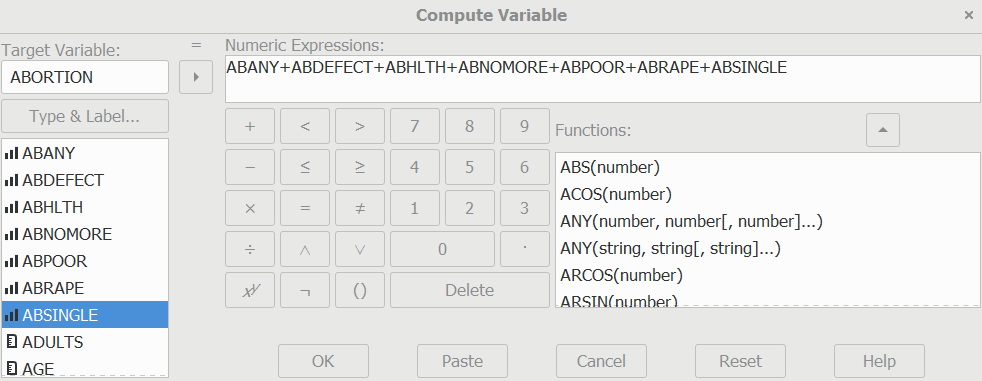
COMPUTE will allow us to combine these seven variables by adding the values, creating a new variable that we will call *abortion*. If a person said yes to all seven questions, the new variable would equal 7 and if the respondent said no to all seven questions, the new variable would equal 14. But what 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, click on TRANSFORM and then click on COMPUTE. Your screen should look like Figure 3-14.



**Figure 3-14**

Type the name of the new variable, *abortion*, in the TARGET VARIABLE box. Then enter the formula for this new variable in the NUMERIC EXPRESSIONS 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 EXPRESSIONS 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 EXPRESSIONS box and use the arrow keys and the delete and backspace keys to make corrections. 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. Your screen should look like Figure 3‑15.

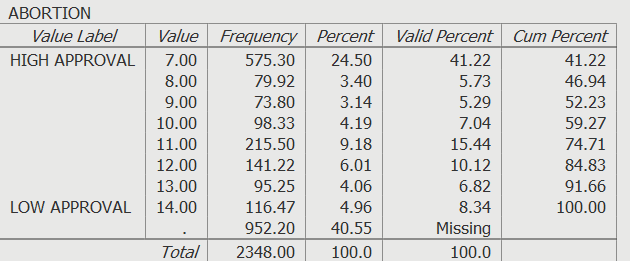


**Figure 3-15**

Click on OK to indicate that you want PSPP to create this new variable. You can use the scroll bar to scroll to the far right of the data file and view the variable you just created.

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 file and double clicking. This will open the VARIABLE VIEW window. You can enter the variable and value labels the way you were taught earlier in this chapter and in Chapter 2.

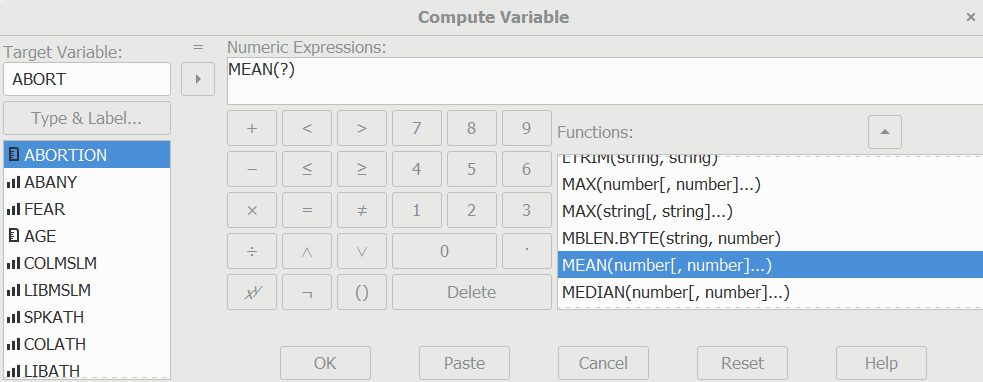
Enter the variable label **Sum of Seven Abortion Variables** andthe value label **High Approval** for the value 7 and **Low Approval** for the value 14. (Remember that seven means they approved of abortion in all seven scenarios and fourteen means they disapproved all seven times.) Click on OK. You should check your new variable to see that it was calculated correctly. Go to ANALYZE, then DESCRIPTIVE STATISTICS, and then FREQUENCIES. Click on RESET to get rid of what is already in the box. Find the variable *abortion*, highlight it and click on the arrow to the left of the VARIABLES box. Then click on OK. Your screen should look like Figure 3‑16. The lowest number should be 7 and the highest number should be 14. The dot (.) below 14.00 in the VALUE column indicates a system missing value.



**Figure 3-16**

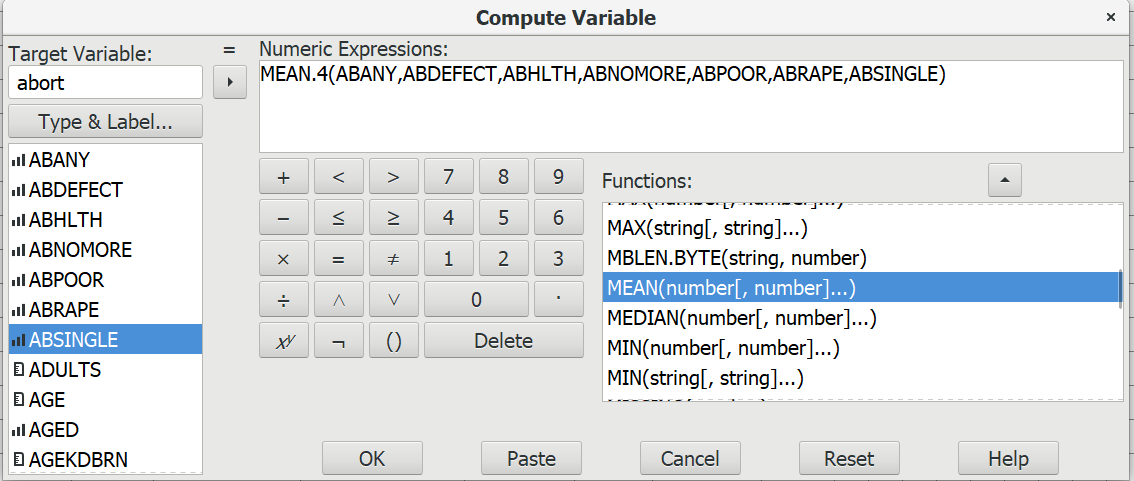
One of the problems with this approach is that the new variable (*abortion*) will be assigned a system missing value if even one of the original variables has a missing value. Notice that there 952 cases with system missing values in Figure 3-16. We can avoid this problem by summing the values of the original variable and dividing by the number of variables with valid values.[[12]](#footnote-12) For example, if six of the seven original variables had valid values, then we would divide the sum by six. We can also tell PSPP to create this new variable only if at least four (or whatever number we choose) of the original variables have valid values. If fewer than four of the original variables have valid values, PSPP 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 FUNCTIONS box, scroll down and double-click on MEAN. Your screen should look like Figure 3-17.



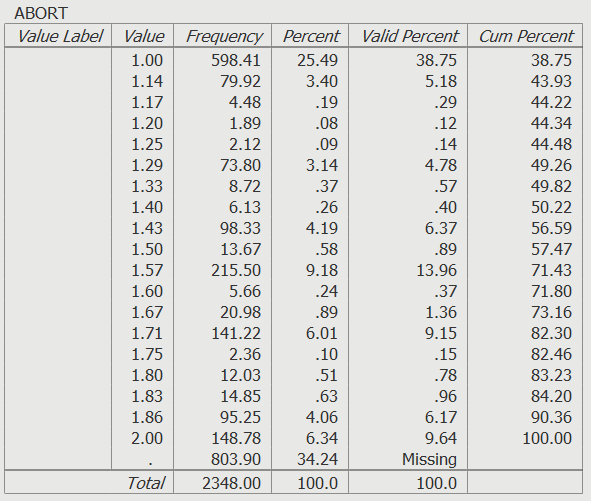
**Figure 3-17**

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*). Be sure to separate the variable names with commas. All that is left is to tell PSPP 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 3-18.



**Figure 3-18**

Click on OK and run a frequency distribution to see what your new variable looks like. You screen should look like Figure 3-19. Notice that there are only 804 cases with system missing values now.



**Figure 3-19**

Try creating another variable. There are six variables that measure tolerance for letting someone speak in your community who may have different views than your own (*spkath, spkcom*, *sphomo*, *spkmil*, *spkslm*, and *spkrac*). For each of these variables, 1 means that they would allow such a person to speak and 2 meansthat they would not allow it. Create a new variable (call it *speak)*, which is the sum of these six variables. This new variable would have a range from 6 (would allow a person to speak in each of the six scenarios) to 12 (would not allow a person to speak in any of the six scenarios). Do a frequency distribution for this new variable to see what it looks like.

Now that you have created *speak*, create another variable (call it *spk)* which would be the sum of these six variables divided by the number of variables with valid values. Let’s also tell PSPP 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, PSPP will assign it a system missing value.

You have already seen that PSPP uses + for addition. It also uses ‑ for subtraction, \* for multiplication, / for division, 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.

There are other ways you can transform variables but WEIGHT, RECODE, and COMPUTE are the most commonly used commands.

## **Next Chapter**

In the next chapter we’ll discuss how to analyze variables one at a time also known as univariate analysis. We’ll look at three different PSPP commands – FREQUENCIES, DESCRIPTIVES, and EXPLORE.

# **Chapter 4**

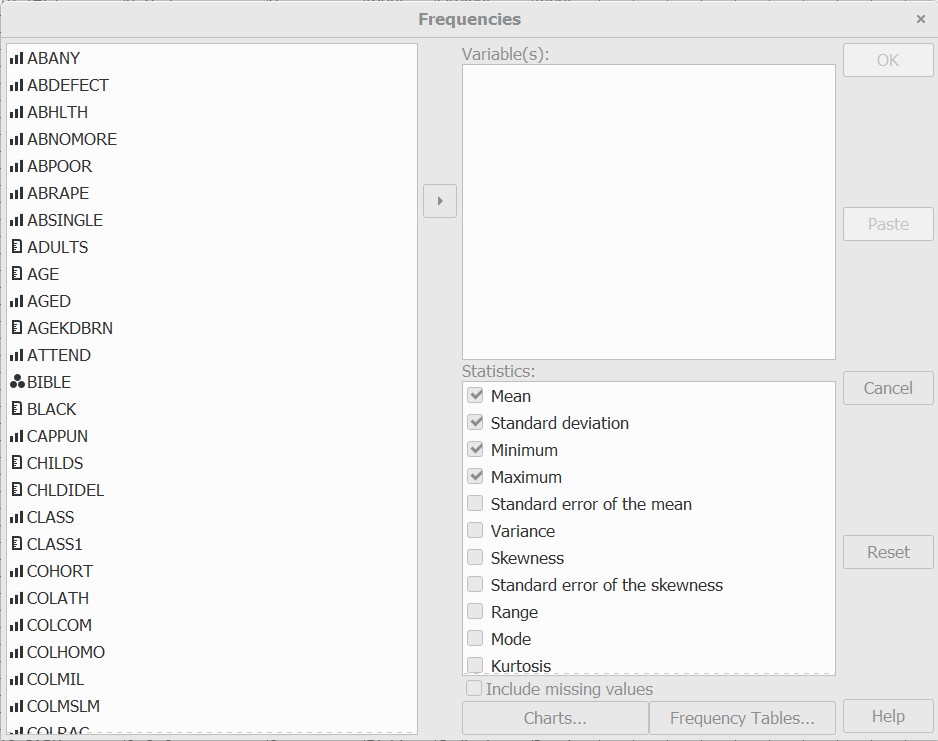
# **Univariate Analysis**

This chapter explains how to analyze variables one at a time. We’ll look at three different PSPP commands.

* FREQUENCIES
* DESCRIPTIVES
* EXPLORE

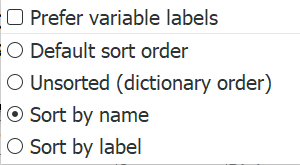
## **Frequencies**

Frequency distributions show you the number of cases for each category of your variable. They also convert these frequencies to percents and tell you how many cases had missing information. Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and finally on FREQUENCIES and you should see Figure 4-1. The list of variables will be on the far left. If PSPP displays the variable labels instead of the variable names, right-click on the list and deselect PREFER VARIABLE LABELS by clicking on it and now you should see Figure 4-1.



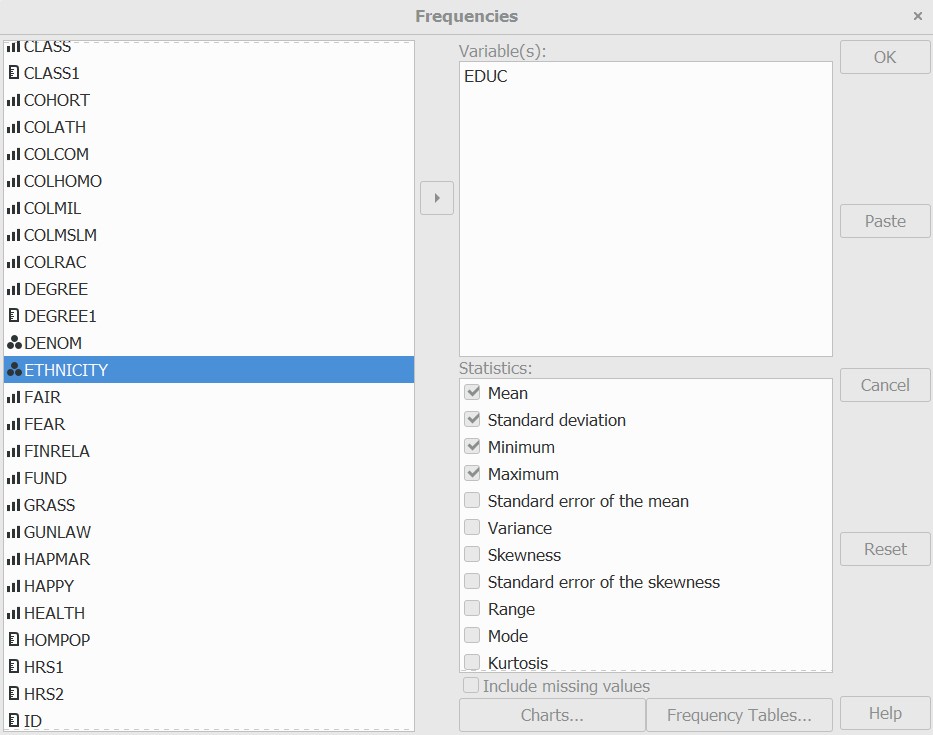
**Figure 4.1**

Notice that the variables in Figure 4-1 are in alphabetical order. What if they aren’t? Right click on the list of variables and select SORT BY NAME. Your screen should look like Figure 4-2.



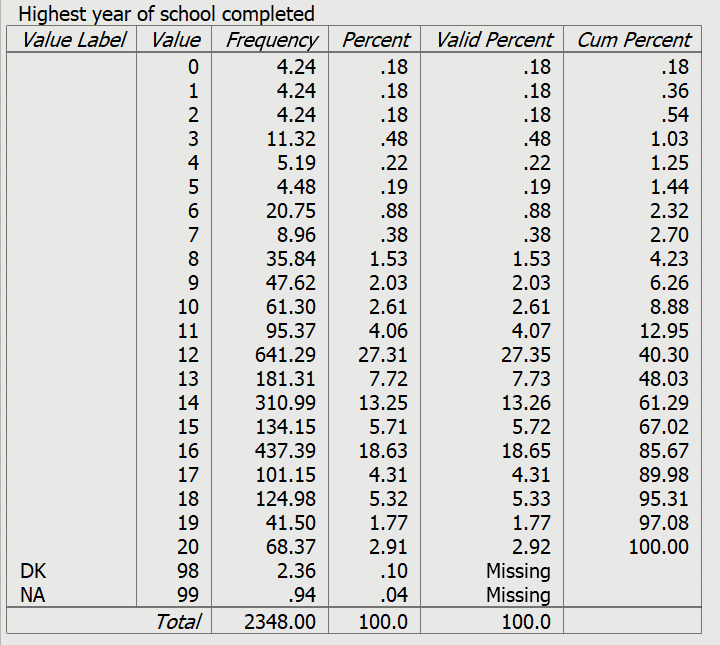
**Figure 4-2**

Now that you have adjusted the list of variables to make it easier to find a particular variable, select the variable(s) for which you want to get a frequency distribution by clicking on them and moving them to the VARIABLES box by clicking on the arrow that should be pointing to the right. We’re going to use *educ* in this example so move *educ* over to the VARIABLES box and you should see Figure 4.3.



**Figure 4-3**

To get the frequency distribution click on OK and you should see Figure 4-4. You’ll notice that PSPP also gives you some additional information such as the name of the data file you are using and various statistics which we’ll discuss next. It also shows you the PSPP commands that it just ran in syntax form. There are two ways to run PSPP. You can use the interface by clicking on the commands you want to run and then filling in the dialog boxes or by writing out the command in syntax form.[[13]](#footnote-13) We’re only going to discuss the interface in this tutorial.



**Figure 4-4**

### **Statistics**

In the lower right-hand corner of Figure 4-3, you’ll see a list of statistics that PSPP will compute. This list is longer that what fits in the box, so you’ll have to scroll down to see the full list. Notice that the first four statistics are already checked (i.e., mean, standard deviation, minimum value, maximum value). These are the default statistics which you will automatically get when you run a frequency distribution. You can choose not to get these statistics by unchecking them and you can also choose to get other statistics by checking them.

The statistics you choose to run are partially dictated by the level of measurement of the variables. Levels are often classified as nominal, ordinal, interval, and ratio.[[14]](#footnote-14)

* A **nominal measure** is one in which respondents are sorted into a set of categories which are qualitatively different from each other. The categories in a nominal level measure have no inherent order to them. This means that it wouldn’t matter how we ordered the categories. They could be arranged in any number of different ways. In our data file, *marital* is an example of a nominal measure.
* An **ordinal measure** is a nominal measure in which the categories are ordered from low to high or from high to low. In our data file, *class* is an example of an ordinal measure. But notice that while the categories are ordered they lack an equal unit of measurement.  That means that the differences between categories are not necessarily equal.  For example, the difference in class between values 1 and 2 may not be the same as the difference between values 3 and 4.
* An **interval measure** is an ordinal measure with equal units of measurement. Temperature measured in degrees Fahrenheit would be an example of an interval measure. The difference between 20 degrees and 40 degrees is the same as the difference between 70 degrees and 90 degrees. Now these numbers have the properties of real numbers and we can add them and subtract them. But notice one thing about the Fahrenheit scale. There is no absolute zero point. There can be both positive and negative temperatures. That means that we can’t compare values by taking their ratios. For example, we can’t divide 80 degrees Fahrenheit by 40 degrees and conclude that 80 is twice as hot at 40. To do this we would need a measure with an absolute zero.
* A **ratio measure** is an interval measure with an absolute zero point. The variable *educ* is an example of a ratio measure. Notice that it has an absolute zero point; you can’t have less than zero years of school.

Since *educ* is a ratio variable, we could use the mean, median, and mode as our measures of central tendency and the standard deviation and variance as our measures of variability. If our variable was *CLASS* (i.e., ordinal), then we couldn’t use the mean but could use the median and the mode as our measures of central tendency; the standard deviation and variance wouldn’t be appropriate measures of variability. If our variable was *marital* (i.e., nominal), then we could only use the mode as our measure of central tendency.

So, for *educ* we’re going to ask for the mean, median, mode, minimum value, maximum value, and standard deviation. Notice that in the output the median is listed as the 50th percentile. Figure 4-5 shows the PSPP output for these choices.



**Figure 4-5**

There are some other options in Figure 4-35 which we will briefly note.

* You can choose to include the missing values in the computation of the various statistics. However, you would rarely want to do that.
* You can select CHARTS. We’ll discuss charts and graphs next.
* You can select FREQUENCY TABLES and change the way the frequency distributions are displayed. We’ll be using the defaults here, so you won’t need to worry about this.

## **Charts and Graphs**

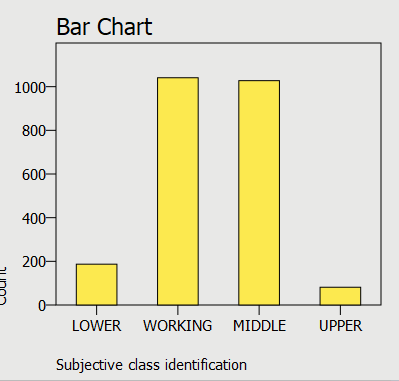
As part of the FREQUENCIES procedure, PSPP will construct pie charts, bar charts, and histograms. (Bar charts and histograms can also be produced by choosing GRAPHS in the menu bar.)

A pie chart is a chart that shows the frequencies or percents of a variable with a small number of categories.  It is presented as a circle divided into a series of slices.  The area of each slice is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables but can be used with interval or ratio variables which have a small number of categories. Figure 4-6 is a pie chart for *class*.

## Pie chart for CLASS

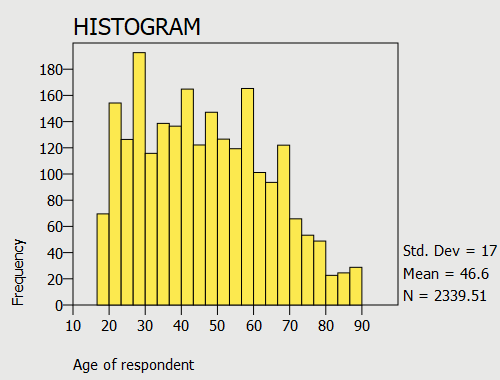
**Figure 4-6**

A bar chart is a chart that shows the frequencies or percents of a variable and is presented as a series of vertical bars that do not touch each other.  The height of each bar is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables. Figure 4-7 is a bar chart of this same variable (*class*)*.*



**Figure 4-7**

A histogram is a graph that shows the frequencies or percents of a variable with a larger number of categories. It is presented as a series of vertical bars that touch each other. The height of each bar is proportional to the number of cases or the percent of cases in each category. It is used with interval or ratio variables. Figure 4-8 is a histogram of *AGE*.

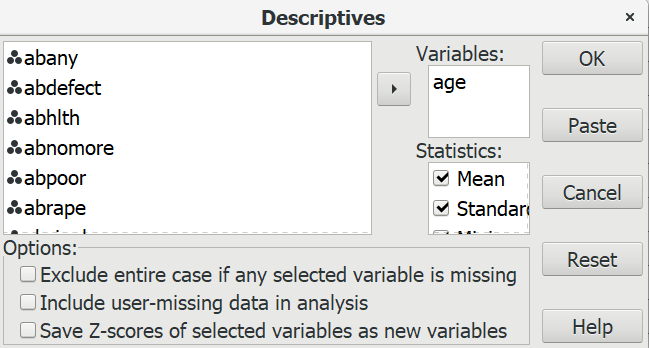


**Figure 4-8**

To get a chart from PSPP, click on the CHARTS button and check the box for the type of chart you want. If you don’t want to get the frequency distribution, click on FREQUENCIES TABLES and select NEVER under DISPLAY FREQUENCIES TABLES.

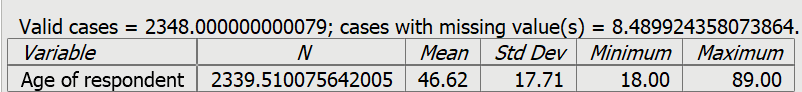
## **Descriptives**

The DESCRIPTIVES procedure is similar to FREQUENCIES except that it does not produce frequency distributions. It should be used when you only want the statistics. Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and finally on DESCRIPTIVES. Move the variable (*age* in our example) you want to use into the VARIABLES box and select the statistics you want to use. You should see Figure 4-9.



**Figure 4-9**

Now click on OK. This time we’re going to use the default statistics (i.e., mean, standard deviation, minimum, maximum). Your output should look like Figure 4-10.



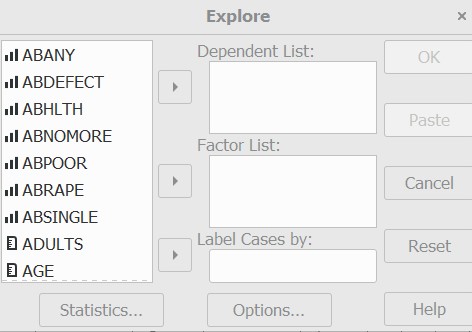
**Figure 4-10**

## **Explore**

Explore can be used to look at your data in various ways.

* It shows how much missing data there is.
* It displays the extreme values.
* It shows the central tendency, variability, skewness and kurtosis of your variables.
* It also shows the percentiles for your variables.

Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and finally on EXPLORE. Your screen should look like Figure 4-11.

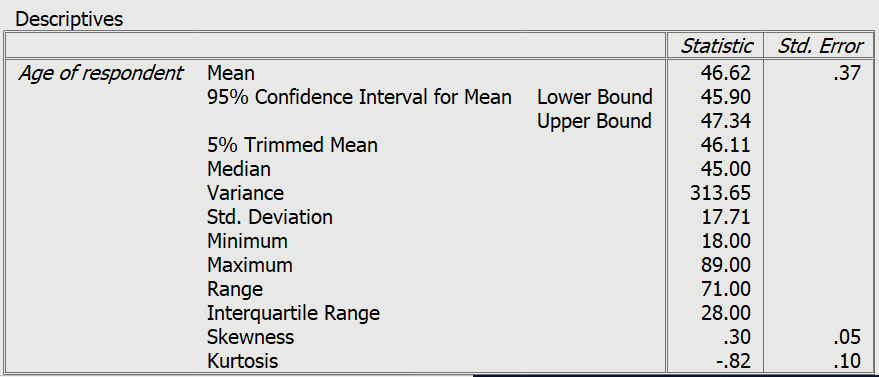


**Figure 4-11**

Move the variables that you want to describe or explore into the DEPENDENT LIST box. For this chapter, ignore why it calls them dependent variables. Let’s focus on age so move the variable *age* into the DEPENDENT LIST box.

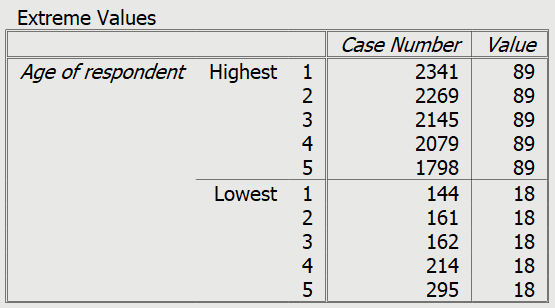
PSPP computes three sets of statistics to describe the variables you choose – DESCRIPTIVES, EXTREMES, PERCENTILES. Click the boxes for all three sets. Click on CONTINUE and then on OK.

* In Figure 4-12 DESCRIPTIVES computes a wide array of different ways of describing central tendency, variability, skewness, and kurtosis.



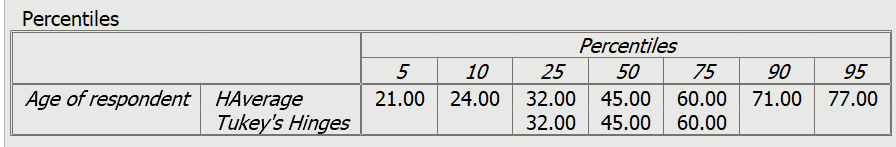
**Figure 4-12**

* EXTREMES (see Figure 4-13) shows you the five highest and five lowest values in your variables.



**Figure 4-13**

* PERCENTILES displays various percentiles in Figure 4-14.



**Figure 4-14**

### **Factors**

Let’s say you want to explore the distributions separately for men and for women. In this case, you would enter the variable *sex* in the FACTORS LIST box so go ahead and move *sex* into it and click on OK. Now PSPP will display your output three times – once for everyone, a second time for males, and then for females.

We’re going to skip over the LABEL CASES BY box.

## **NEXT CHAPTER**

In the next chapter we’re going to start looking at bivariate analysis which involves focusing on the relationship between pairs of variables. One way to do that for interval and ratio variables is to compare means. PSPP offers several different ways of comparing means.

# **Chapter 5**

# **Comparing Means**

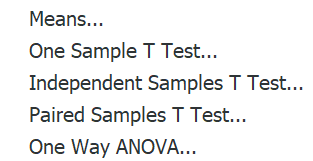
PSPP will compute means for variables that you specify and will also compute means for combinations of these variables. We’ll explain this in the next section of this chapter. But PSPP will do more than just compute means; it will also carry out various forms of hypothesis testing including the following.

* One sample t test
* Independent samples t test
* Paired samples t test
* One-way analysis of variance, often called ANOVA

## **Working with Means**

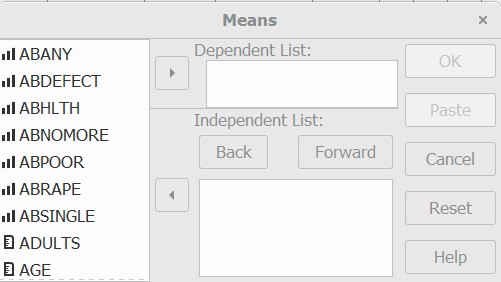
Keep in mind that computing means only makes sense for variables that are interval or ratio level measurement. See Chapter 4 for a discussion of levels of measurement.

Click on ANALYZE and then on COMPARE MEANS. You screen should look like Figure 5-1.



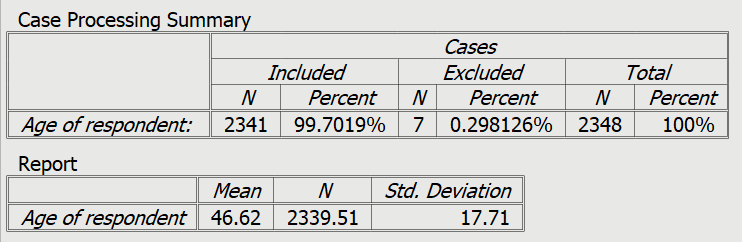
**Figure 5-1**

These are the five options you have for comparing means. Now click on MEANS and you should see Figure 5-2.



**Figure 5-2**

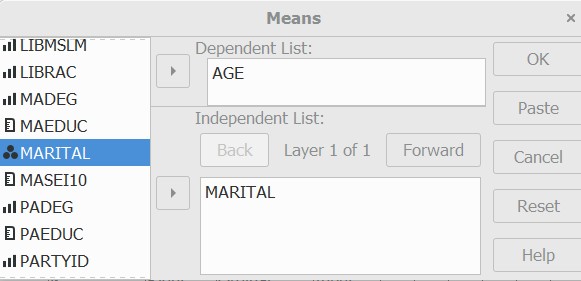
The DEPENDENT LIST box is where you put the variables for which you want to compute means. Remember that these must be interval or ratio level variables. For this example, move *age* over into the DEPENDENT LIST box. Now click on OK. You should see Figure 5-3.



**Figure 5-3**

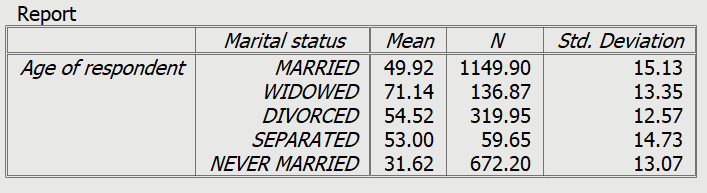
This shows the mean, standard deviation, and number of cases for the variable *age*.

What if you wanted to get these statistics broken down by marital status? To do that we would enter the variable *MARITAL* in the INDEPENDENT LIST box. Your screen will look like Figure 5-4.



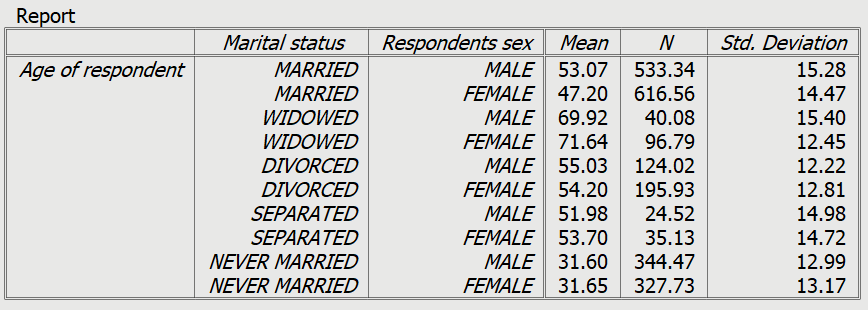
**Figure 5-4**

Now click on OK and you should see Figure 5-5.



**Figure 5-5**

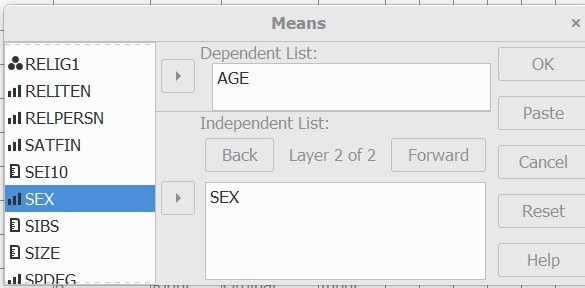
As you would expect, widowed respondents are the oldest and those who have never married are the youngest. You could break these five marital categories down further by the variable *sex*. Since *marital* has five categories and *sex* has two categories, this breakdown will have 5 x 2 or 10 categories. To do this add the variable *sex* to the INDEPENDENT LIST box and click on OK. Now you should see Figure 5-6.



**Figure 5-6**

This table show the means, standard deviations, and number of cases for each of these 10 categories. You can see that mean age does not vary much by sex but does vary considerably by marital status.

Look back at Figure 5-4 and you’ll see LAYER 1 OF 1 just below INDEPENDENT LIST. Remove *sex* from the INDEPENDENT LIST box. Now click on the FORWARD button and add the variable *sex* below LAYER 2 OF 2. You screen should look like Figure 5-7.



**Figure 5-7**

If you click on the BACK button you should see *marital* in LAYER 1 OF 1. Now click on the FORWARD button and you’ll see *sex* in LAYER 2 OF 2. Once you have made sure that you have it right, click on OK and you should see Figure 5-8. Now PSPP will break *age* down by *marital* and by *sex* separately.

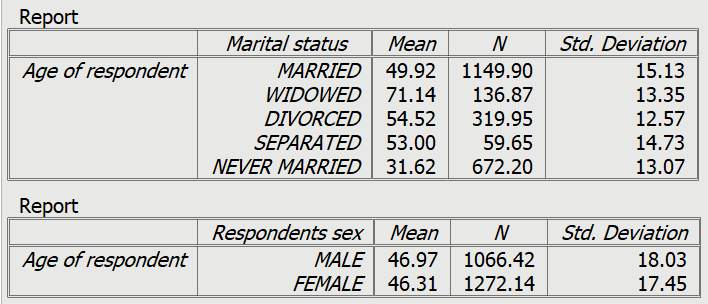


Figure 5-8

Now try it on your own. Break *tvhours* (i.e., the number of hours that respondents watch television per day) down by *race* and *sex* separately and then jointly.

Notice in this section how we have moved from looking at variables one at a time (i.e., univariate analysis) to looking at the relationship between two variables (i.e., bivariate analysis) and then to looking at the relationships among three variables (i.e., multivariate analysis). In later chapters we’ll look further at relationships among sets of three or more variables.

Before we can proceed to the other options in COMPARE MEANS (see Figure 5-1), we need to introduce the idea of hypothesis testing.

## **Introduction to Hypothesis Testing**

It’s important to understand the difference between a population and a sample. A population is the complete set of objects in which you are interested. For the General Social Survey that we’re using in these exercises, it’s all non-institutionalized adults (18 years of age and older) living in the continental United States. Populations are often very large and it’s virtually impossible to collect data on every object in the population. The solution is to select a subset of the population which we will use to make inferences about the population. That subset is called a sample. This process of drawing an inference from sample data to the population is referred to as statistical inference. When we make statistical inferences, it’s assumed that we are using a probability sample which is a sample in which the probability of selection is known for all cases in the population.

No sample is ever a perfect representation of the population. There’s always some amount of error which is referred to as sampling error. We can’t eliminate sampling error, but we can try to minimize it. Fortunately, there are certain things we can do to reduce sampling error.

* Increasing the sample size will reduce sampling error. Of courses, there is a practical limit to how big a sample we can select. The data file that we are using in this chapter is a little over 2,300 respondents (i.e., 2,348 to be exact).
* There are ways to select a sample that will also reduce sampling error. Stratification is one way to do that, but stratification is beyond the scope of this tutorial.

To summarize, we use data from our sample to make inferences about the population. Using what you learned in the previous section, compute the mean years of school competed (*educ*) of our sample. It turns out to be 13.73 years of school, a little more than 1.5 years after high school. Our sample mean is an estimate of our population mean (i.e., the mean age for all adults in the U.S.[[15]](#footnote-15)). We know that because of sampling error 13.73 isn’t actually the mean age of the population. But can we conclude that education in the population is higher than 12 years?

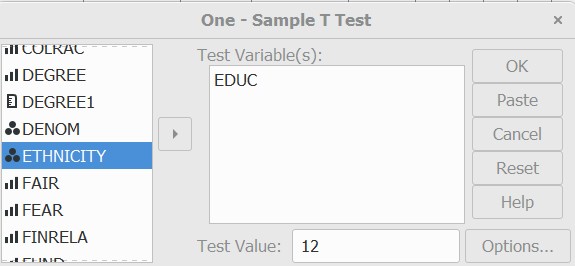
How are we going to answer this question? We would like to be able to prove that the population mean is greater than 12 years but there’s no way we can actually do that. What we can do is set up another hypothesis that says that the population mean is actually equal to 12. We have names for these hypotheses – the research hypothesis and the null hypothesis.

* Research hypothesis – the population mean is greater than 12
* Null hypothesis – the population mean is equal to 12

The research hypothesis is what we think is true and the null hypothesis says that our research hypothesis is not true. We’re going to test the null hypothesis. If we can reject the null hypothesis then we have evidence to support our research hypothesis. If we can’t reject the null hypothesis then we don’t have any evidence in support of the research hypothesis. In the next section we’ll show how to carry out this test using PSPP.

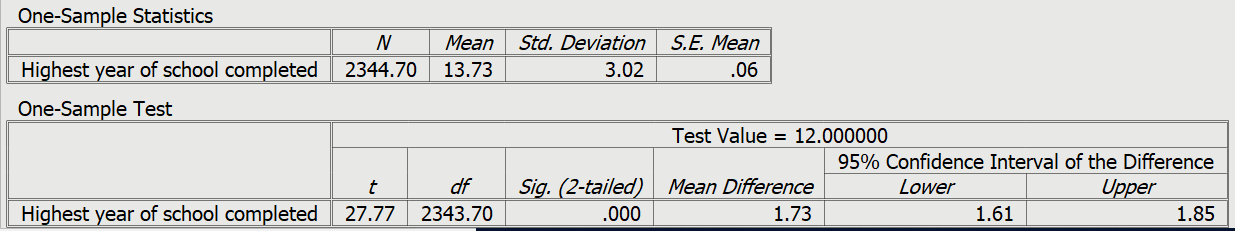
## **One-Sample t Test**

Click on ANALYZE and then on COMPARE MEANS which will open another drop-down menu and then click on ONE-SAMPLE T TEST. Move the variable, *educ*, over to the TEST VARIABLE box on the right. Below the box on the right you will see a box called TEST VALUE. This is where we enter the value specified in the null hypothesis which in our case is **12[[16]](#footnote-16)**. Your screen should look like Figure 5-9.



**Figure 5-9**

Click on OK and your output should look like Figure 5-10.



**Figure 5-10**

The first box will have four values in it.

* N is the number of cases for which we have valid information (i.e., the number of respondents who answered the question). In this problem, N equals 2,345.
* Mean is the mean years of school completed by the respondents in the sample who answered the question. In this problem, the sample mean equals 13.73.
* Standard Deviation is a measure of dispersion. In this problem, the standard deviation equals 3.02.
* Standard Error of the Mean is an estimate of how much sampling error there is. In this problem, the standard error equals .06.

The second box will have five values in it.

* t is the value of the t test
* df is the number of degrees of freedom
* Significance (2-tailed) value
* Mean Difference
* 95% Confidence Interval of the Difference

The significance value is a probability. It’s the probability that you would be wrong if you rejected the null hypothesis. It’s .000 which you would think is telling you that there is no chance of being wrong if you rejected the null hypothesis. But this is actually a rounded value and it means that the probability is less than .0005 or less than five in ten thousand. So, there is a chance of being wrong but it’s really, really small.

Now all we have to do is figure out how to use the one-sample t test to decide whether to reject or not reject the null hypothesis. Look again at the significance value, which is less than .0005. This tells you that the probability of being wrong if you rejected the null hypothesis is less than five out of ten thousand. With odds like that we’re going to reject the null hypothesis. A common rule is to reject the null hypothesis if the significance value is less than .05 or less than five out of one hundred.

But wait a minute. The PSPP output said that this was a two-tailed significance value. What does that mean? Look back at the research hypothesis which was that the population mean was greater than 12. We’re actually predicting the direction of the difference. We’re predicting that the population mean will be greater than 12. That’s called a one-tailed test and we have to use a one-tailed significance value. If the two-tailed significance value is less than .0005 then the one-tailed significance value is half that or less than .00025. We still reject the null hypothesis which means that we have evidence to support our research hypothesis. We haven’t proven the research hypothesis to be true, but we have evidence to support it.

Now try your hand at this. There is another variable in the GSS called *hrs1* which is the number of hours that respondents worked last week if they were employed. Many people have suggested that Americans are working longer hours than they used to. Since the traditional work week is 40 hours, if it’s true that we’re working more hours our research hypothesis would be that the mean number of hours worked last week would be greater than 40. Do a one-sample t test to test your hypothesis. Then decide whether you should reject or not reject the null hypothesis and what this tells you about your research hypothesis.

## **Independent Samples t Test**

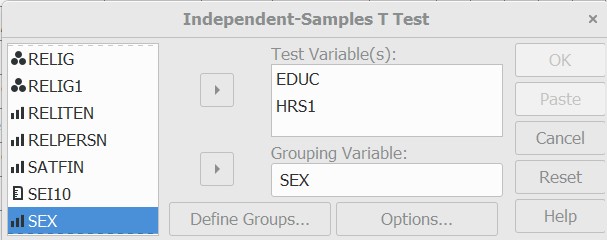
What do we mean by independent samples? Independent samples are samples in which the composition of one sample does not influence the composition of the other sample. In this chapter we’re using the GSS which is a sample of adults in the United States. If we divide this sample into men and women, we would have a sample of men and a sample of women and they would be independent samples. The individuals in one of the samples would not influence who is in the other sample.

In this section we’re moving from looking at one sample to comparing two samples by taking the difference between their means. We’ll start by asking two questions.

* Do men and women differ in the number of years of school they completed (*educ*)?
* Do men and women differ in the number of hours they worked in the last week (*hrs1*)?

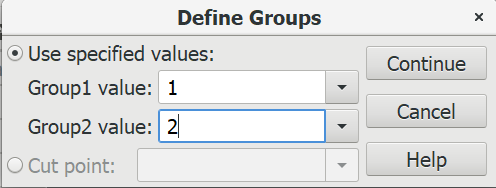
A t test is used when you want to compare **two** groups. The grouping variable defines these two groups. The variable, *sex* is a dichotomy. It has only two categories – male (value 1) and female (value 2). But any variable can be made into a dichotomy by establishing a cut point or by recoding. For example, the variable *satfin*(satisfaction with financial situation) has three categories – satisfied (value 1), more or less satisfied (value 2), and not at all satisfied (value 3). The cut point is the value that makes this into a dichotomy. All values less than the cut point are in one category and all values equal to or larger than the cut point are in the other category. If your cut point is 3, then values 1 and 2 are in one category and value 3 is the other category.

Click on ANALYZE and then on COMPARE MEANS and finally on INDEPENDENT-SAMPLES T TEST. Move the two variables listed above into the TEST VARIABLE box. These are the variables for which you want to compare the mean scores. Right below the TEST VARIABLE box is the GROUPING VARIABLE box. This is where you indicate which variable defines the groups you want to compare. In this problem the grouping variable is *sex*. Your screen should look like Figure 5-11.



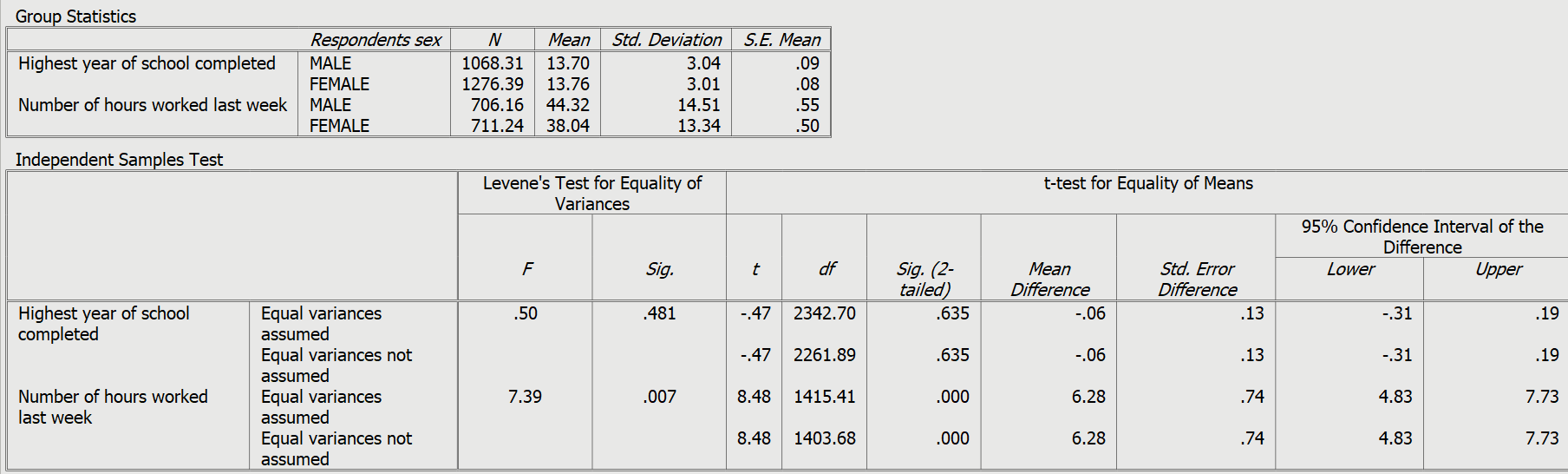
**Figure 5-11**

Once you have entered the grouping variable, then click on DEFINE GROUPS and enter either the values of the two groups or the cut point. In our case, you would enter 1 for male into GROUP 1 and 2 for female into GROUP 2. It wouldn’t matter which was group 1 and which was group 2. Your screen should look like Figure 5-12.



**Figure 5-12**

Click on CONTINUE and then on OK and you should see Figure 5-13.



**Figure 5-13**

The first box gives you four pieces of information.

* N which is the number of males and females on which the t test is based. This includes only those cases with valid information. In other words, cases with missing information (e.g., don’t know, no answer) are excluded.
* Means for males and females.
* Standard deviations for males and females.
* Standard error of the mean for males and females which is an estimate of the amount of sampling error for the two samples.

The second box has more information in it. The first thing you notice is that there are two t tests for each variable. One assumes that the two populations (i.e., all males and all females) have equal population variances and the other doesn’t make this assumption. In our two examples, both t tests give about the same results. We’ll come back to this in a little bit. The rest of the second box has the following information. Let’s look at the t test for *educ*.

* t is the value of the t tests.
* Degrees of freedom of the t tests.
* The significance (two-tailed) value.
* The mean difference is the mean for the first group (males) minus the mean for the second group (females).
* The standard error of the difference which is an estimate of the amount of sampling error for the difference score.
* 95% confidence interval of the difference which tells us that there’s a 95% probability that the population difference is within this range.

Notice how we are going about this. We have a sample of adults in the United States. We calculate the mean years of school completed by men and women in the sample who answered the question. But we want to test the hypothesis that the mean years of school completed by men and women in the population are different. We’re going to use our sample data to test a hypothesis about the population.

The hypothesis we want to test is that the mean years of school completed by men in the population is different than the mean years of school completed by women in the population. We’ll call this our research hypothesis. It’s what we expect to be true. But there is no way to prove the research hypothesis directly. So, we’re going to set up another hypothesis that says that the research hypothesis is not true and call this the null hypothesis. If we can’t reject the null hypothesis then we don’t have any evidence in support of the research hypothesis. If we can reject the null hypothesis then we have evidence that supports the research hypothesis. We haven’t proven the research hypothesis, but we have support for this hypothesis.

Here are our two hypotheses.

* research hypothesis – the population mean for men minus the population mean for women does not equal 0. In other words, they are different from each other.
* null hypothesis – the population mean for men minus the population mean for women equals 0. In other words, they are not different from each other.

It’s the null hypothesis that we are going to test.

Now we’re going to use the t test to decide whether to reject or not reject the null hypothesis. Look again at the significance value which is .635 for both t tests. That tells you that the probability of being wrong if you rejected the null hypothesis is just about .64 or 64 times out of one hundred. With odds like that we’re not going to reject the null hypothesis. A common rule is to reject the null hypothesis if the significance value is less than .05 or less than five out of one hundred.

The PSPP output said this was a two-tailed significance value. What does that mean? Look back at the research hypothesis which was that the population mean for men minus the population mean for women does not equal 0. We’re not predicting that one population mean will be larger or smaller than the other. That’s called a two-tailed test and we have to use a two-tailed significance value. If we had predicted that one population mean would be larger than the other that would be a one-tailed test.

We still haven’t explained why there are two t tests. As we said earlier, one assumes that the two populations (i.e., all males and all females) have equal population variances and the other doesn’t make this assumption. To decide which t test we should use we need to look at the Levene’s test for the equality of variances which is in the second box in your PSPP output. For this test, the null hypothesis is that the two population variances are equal. The appropriate test would be the F test. The significance value for the variable *EDUC* is 0.481 which is not less than .05 so we do not reject the null hypothesis that the population variances are equal. This means that we would use the t test that assumes equal population variances.

Now work through the output for the t tests comparing the number of hours worked last week by men and women. Would you use the t test assuming equal population variances or the t test making no such assumption? Would you reject the null hypotheses for this t test? What does that tell you?

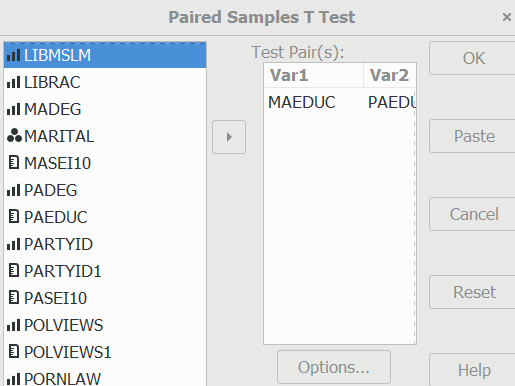
In the next section of this chapter we’ll discuss paired or dependent samples.

## **Paired Samples t Test**

Paired or dependent samples are samples in which the composition of one sample influences the composition of the other sample. For example, if we have a sample of married couples and divide that sample into two samples of men and women, then the men in one of the samples determines who are the women in the other sample, and vice-versa. The composition of the samples are interdependent.

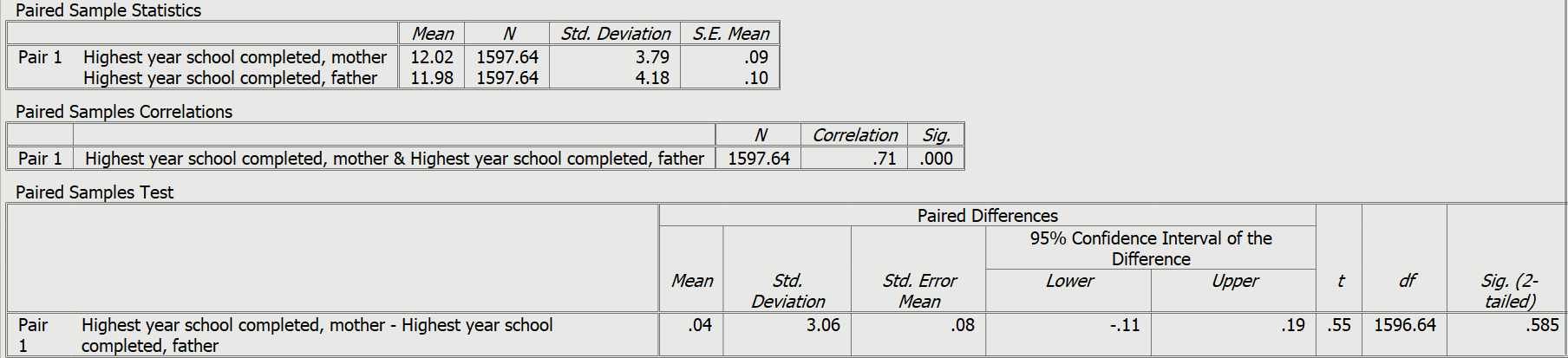
Let’s try a different example. The GSS includes questions about the years of school completed by the respondent’s parents – *maeduc* and *paeduc*.  Let’s assume that we think that respondents’ mothers have more education than responden’s fathers.  We would compare the mean years of school completed by mothers with the mean years of school completed by fathers.  Since our samples are paired or dependent samples, we’ll need to use the Paired Samples t Test in PSPP.

Click on ANALYZE in the menu bar and then on COMPARE MEANS and finally on PAIRED SAMPLES T TEST.  Select the variable *maeduc* and click on arrow to move it into the TEST PAIRS box under VAR1. Then in the TEST PAIRS box, scroll over to the right until you see VAR2 and click right below it. Now move *paeduc* below VAR2,[[17]](#footnote-17)(You may need to scroll to the right to see VAR2.) These are the variables for which you are going to compute means. Your screen will look like Figure 5-14.



**Figure 5-14**

Click on OK and your output should look like Figure 5-15.



**Figure 5-15**

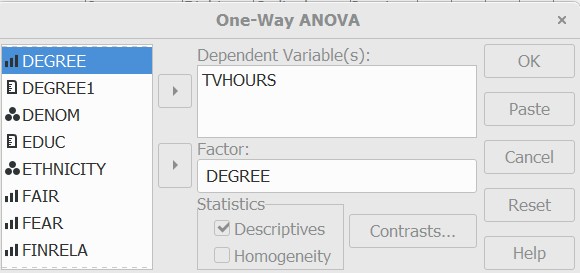
The first box shows the sample means and standard deviations for *maeduc* and *paeduc*. The second box shows the correlation between the two variables which we’ll cover in Chapter 7. The third box shows the mean difference for *maeduc* and *paeduc*, the t test and the 2-tailed significance value (among other things). Our research hypothesis is that the mean years of school completed by mothers is more than the mean years of school for fathers. Our null hypothesis is that the two means are the same. Remember that our hypotheses always refer to the population means. We’re using the sample means to make an inference about the population means. The two-tailed significance value is .585. Since we are predicting direction, we want the one-tailed significance value. Divide .585 in half to get the one-tailed value (.2925). Since this is not less than .05, we do not reject the null hypothesis and have no evidence to support our research hypothesis.

Now it’s your turn. We also have data on the respondent’s years of school completed (*educ*) and their spouse’s education if married (*speduc*). Carry out the paired samples t test and discuss what this test tells you.

## **One-Way Analysis of Variance (ANOVA)**

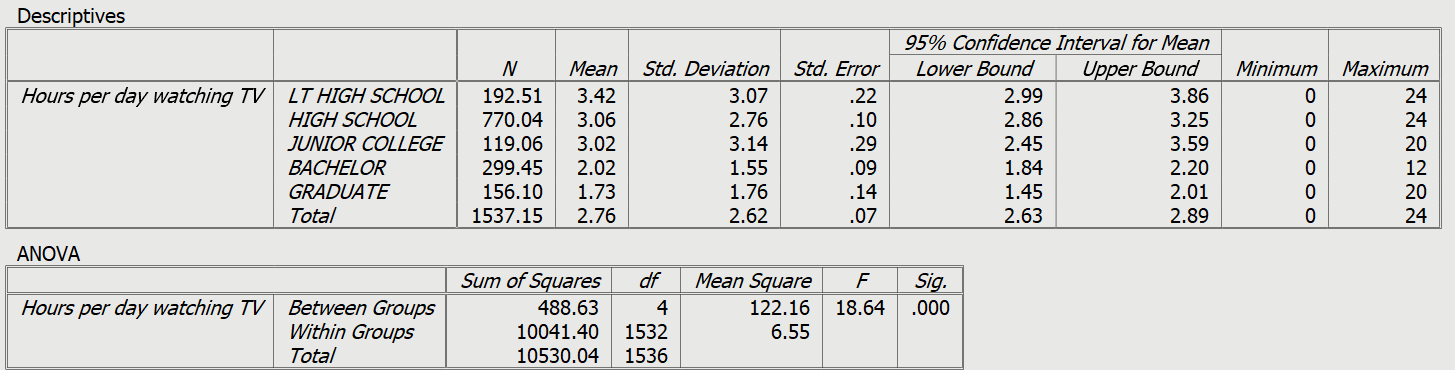
In this section we’re going to see if the mean hours of television watched per day varies for different levels of education and if these differences are statistically significant. That’s a job for one-way analysis of variance (ANOVA).

Click on ANALYZE in the menu bar and then on COMPARE MEANS and finally on ONE-WAY ANOVA.  Select the variable *tvhours* and move it to the DEPENDENT VARIABLES box.  Then select the variable *degre* and move it to the FACTOR box.  Click on DESCRIPTIVES in the STATISTICS box. Your screen should look like Figure 5-16.



**Figure 5-16**

Click on OK and your output should look like Figure 5-17.



**Figure 5-17**

The output gives you results of the one-way analysis of variance.  We’re not going to go into detail on these statistics. But here’s the brief version.

Our research hypothesis is that the mean number of hours watching television is higher for some levels of education than for other levels in the population. This is what we expect to be true.  But there is no way to prove the research hypothesis directly.  So, as before, we set up another hypothesis that says that the mean number of hours watching television is the same for all levels of education in the population and call this the null hypothesis. If we can reject the null hypothesis, we have evidence that there is at least one pair of educational categories for which there is a statistically significant difference. If we can’t reject the null hypothesis then we don’t have any evidence in support of our research hypothesis.

Here are our two hypotheses.

* research hypothesis – the mean number of hours watching television for at least one level of education is different from at least one other population mean.
* null hypothesis – the mean number of hours watching television is the same for all five levels of education in the population.

It’s the null hypothesis that we are going to test.

We can use the F test to decide whether to reject or not reject the null hypothesis.  Look at the significance value which is 0.000.  Since this is a rounded value it tells you that the probability of being wrong if you rejected the null hypothesis is less than .0005 or less than 5 times out of ten thousand.  With odds like that we’re going to reject the null hypothesis.

So, what have we learned?  We learned that the mean number of hours watching television for at least one educational level is different from at least one other educational level.[[18]](#footnote-18)  But which ones?  There are statistical tests for answering this question but we’re not going to cover them in this tutorial.

## **Next Chapter**

In the next chapter we’re going to discuss crosstabulation which can be used to compare two or more variables that are nominal or ordinal level measurement.

# **Chapter 6**

# **Crosstabulation**

This chapter explains how to display as a table the relationship between variables when at least one is measured at the nominal or ordinal level. We’ll do this first by considering just two variables. We’ll then show you how to introduce a third (control) variable into the analysis. Throughout the chapter, we’ll be looking at variables that might help explain people’s choice of political party.

## **Two-variable Contingency Tables**

A contingency table (also called a crosstabulation, or crosstab for short) displays the relationship between two or more “categorical variables” (a variable in which cases are coded as having a relatively small number of categories).  It is called a “contingency table” because it allows us to examine a hypothesis that the values of one variable are contingent (dependent) upon those of another.

When one or both variables are nominal, the statistics used will be different from what we would use if both are ordinal. We’ll look at one example of each situation.

Example 1: To what extent do people identify with a political party based on their ethnicity?

Our first example examines the relationship between political party identification, an ordinal variable, and ethnicity, a nominal variable.

To simplify matters, we’ve recoded *partyid* into a new variable called *partyid1* with three categories, Democrat, Independent, and Republican. For ethnicityrespondents will be classified as White, Black, and Hispanic. We’ll need to delete a fourth category (Other). There are two reasons for this: 1) there are too few cases in this category to permit reliable analysis, and 2) the cases within the category vary from one another too much for it to make sense to group them all together. We can delete these cases by changing to variable view, finding the variable *ethnicity*, clicking in the missing values box, and adding the value 4 to one of the empty boxes for discrete missing values. This is the approach we’ve chosen. If you do the same, be sure to undo this change when you are finished so that these cases won’t be eliminated in later analyses. An alternative approach would be to use the RECODE INTO DIFFERENT VARIABLES procedure to create another variable that doesn’t include “Other” as a valid value.

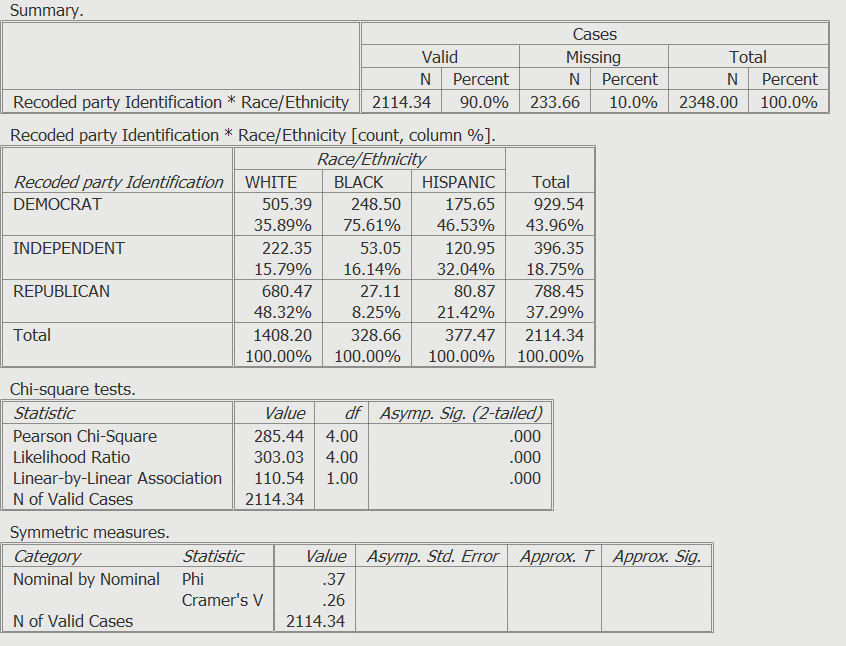
A table consists of a matrix of rows and columns.  Both variables happen to have three categories. The table is referred to as a “three by three” table.  (Note that, if they differ, the number of rows is always given first.)  The values of one of the variables are placed in the rows and those of the other in the columns.   Usually, as in this example, the dependent variable (*partyid1*) is the row variable and the independent variable (*ethnicity*) is the column variable.  There is, however, no requirement that this be the case.  If it fits the available space better to put the dependent variable in the columns, go ahead and do so.

Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and on CROSSTABS. Select *partyid1*, and then on the arrow to the left of ROWS. Select *ethnicity*, and then on the arrow to the left of COLUMNS.

By default, PSPP calculates the number of cases in each cell, along with the row, column, and total percentages. In using contingency tables to test hypotheses, always percentage in the direction of the independent variable.  This is necessary because we are testing the idea that different categories of the independent variable will tend to have different values for the dependent variable. Comparisons are much easier when each category of the independent variable totals to 100%. In other words, you want the percents to sum to 100% within each category of the independent variable. Click on CELLS. In the dialog box that pops up, deselect ROW and TOTAL. Click on CONTINUE.

To see how strong and statistically significant the relationship is, click on STATISTICS. Chi-square is a widely accepted measure of statistical significance. It is selected by default, and can be employed even with nominal level measures. We’ll be using the “Pearson Chi-square,” the most common version, though PSPP also provides a couple of alternatives.

To test for the strength of a relationship between two variables when one or both is nominal, you can use Cramer’s V, a measure that standardizes chi-square so that it ranges from 0 to a number approaching 1. Select “PHI.,” Phi is a special case of Cramer’s V for two by two tables, but choosing phi will also produce Cramer’s V for tables of any size. Click on CONTINUE and on “OK.” PSPP will return the table shown in Figure 6-1.



**Figure 6-1**

The “Summary” simply shows the number and percent of cases in the table (that is, those for which there are valid values for both variables), the number and percent of cases with missing values, and the total of both. The second table shows the actual crosstab, while the third provides chi-square tests for statistical significance, and the fourth measures the strength of the relationship.

The column percentages displayed in the contingency table show major differences among the ethnicity categories. White respondents tend to favor Republicans over Democrats by a margin of about 12.5 percentage points, whereas Blacks prefer Democrats over Republicans by a huge margin (75.61% Democrats versus 8.25% Republicans, or a ratio of about nine to one). Hispanics are in between: the margin is 46.53% Democrats versus 21.42% Republicans, a ratio of about two to one. Another way to do this is to compare the percents straight across. Compare the 35.89 to the 75.61 to the 46.53 for Democrats which shows that Blacks are much more likely to be Democrats. Then do this for each row.

In our example the Chi-square measure is highly significant statistically: a Chi-square as high as 285.44 would occur by chance less than one time in a thousand. The Cramer’s V of .26 may not seem too impressive, but with weak (nominal) level data, this is to be expected.

Example 2: To what extent do people identify with a political party based on their ideology (that is, their views on political issues, typically expressed as liberal, moderate, and conservative)?

For this example, we will be testing the hypothesis that a person’s recoded party identification (which we’ve used in exercise 1) tends to depend on where that person stands on political issues (recoded into three categories as *polviews1[[19]](#footnote-19)*). Click on ANALYZE, then on DESCRIPTIVE STATISTICS, and on CROSSTABS. Select *partyid1*, and then on the arrow to the left of ROWS. Select *polviews1*, and then on the arrow to the left of COLUMNS.As in exercise 1, select COUNT and COLUMNS under CELLS. We’ll also need to measure the strength and significance of the relationship. Since we have two ordinal variables, we can employ more powerful tests than we could if one or both were nominal. A good measure of strength of association between two ordinal variables is Somer’s D. Its value ranges from 0 (no relationship) to plus or minus[[20]](#footnote-20) 1 (a perfect relationship, either positive or negative). Since PSPP doesn’t “know” which variable we are treating as dependent, it calculates 3 versions of D, one assuming that *PARTYID1* is the dependent variable, another treating *polviews1* as dependent, and a third which makes no assumption either way. For each, PSPP also generates a test of its statistical significance called a “t” test. Click on STATISTICS, then unselect CHISQ (which has been selected by default), and select D. Unselect any other statistics that are checked. Click on CONTINUE, then on OK. The result should look like Figure 6.2.

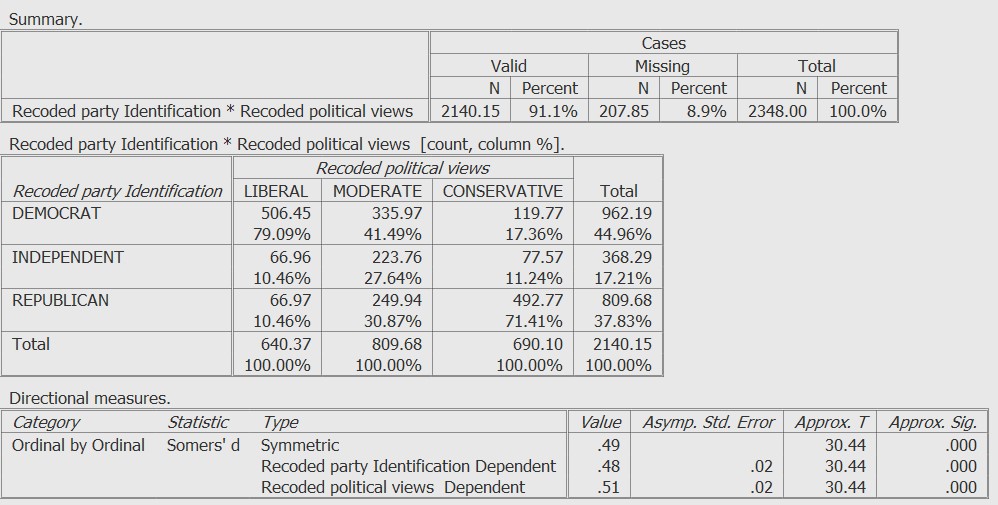


Figure 6-2

As expected, Figure 6-2 shows that a very large percentage of liberals are Democrats and that, to a lesser degree, conservatives identify as Republicans. Moderates are somewhat, though not overwhelmingly, more likely to favor Democrats over Republicans.

Since we are treating *PARTYID1* as the dependent variable, the value for Somer’s D is .48, a moderately strong association.

Along with Somer’s *D*[[21]](#footnote-21), PSPP also generates a test of its statistical significance called a “t” test. The "Approx. Sig." for this test is the significance of t, that is, the probability (p) that the relationship could have occurred by chance. In this instance, the “approximate significance” of t is .000 (that is, p<.0005). This is for a two-tailed test (used if we were making no assumption about the direction of the relationship). For a one-tailed test, it would be half that <.00025, since PSPP doesn’t measure any probability less than .001).

## **Three-variable Contingency Tables**

Example 3: Party identification by ethnicity, controlling for social class.

We’ve found (see example 1) that Blacks identify with the Democratic Party far more than do Whites, with Hispanics somewhere in between. We might suspect that these differences in party identification simply reflect differences in perceived social class. Party differences could simply be the result of Blacks being more likely than Whites to self-identify as working or lower class, while Whites are more likely to self-identify as middle or upper class, and Hispanics again somewhere in between. If so, party identification could really be a function of class rather than ethnicity. To test this, we can break the table shown in Figure 6-1 into sub-tables defined by social class: working class (including lower class) and middle class (including upper class). (One reason why we’ve grouped “lower” with “working” class and “upper” with “middle” class is that relatively few Americans consider themselves either lower or upper class.)

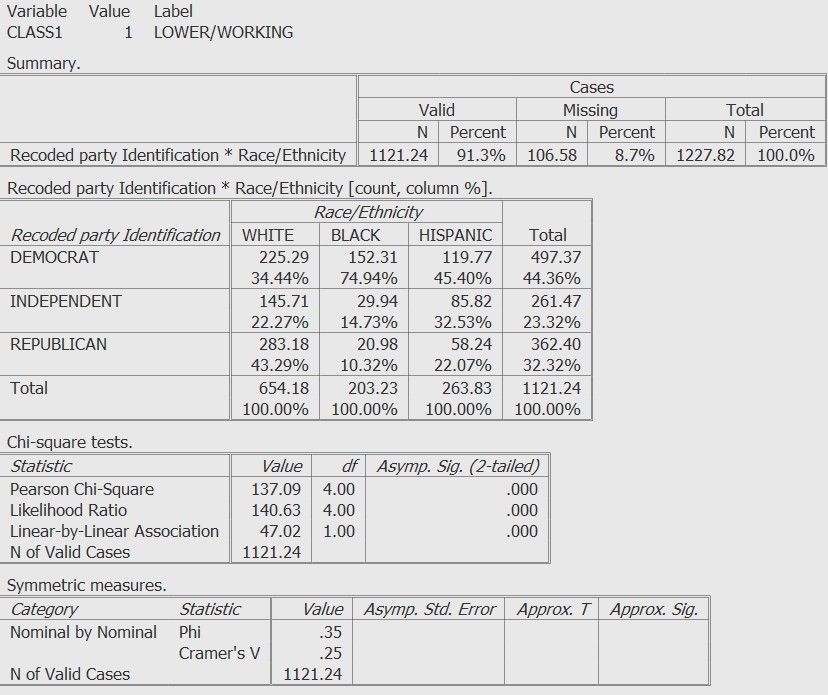
To break a crosstab into different categories of a control variable:

1. If necessary, recode the control variable into a new variable that has the categories you want to use in your three-variable crosstab. So that you won’t have to, we’ve recoded *class* into a new variable we’ve called *class1*, with two values (1) lower or working and (2) middle or upper.

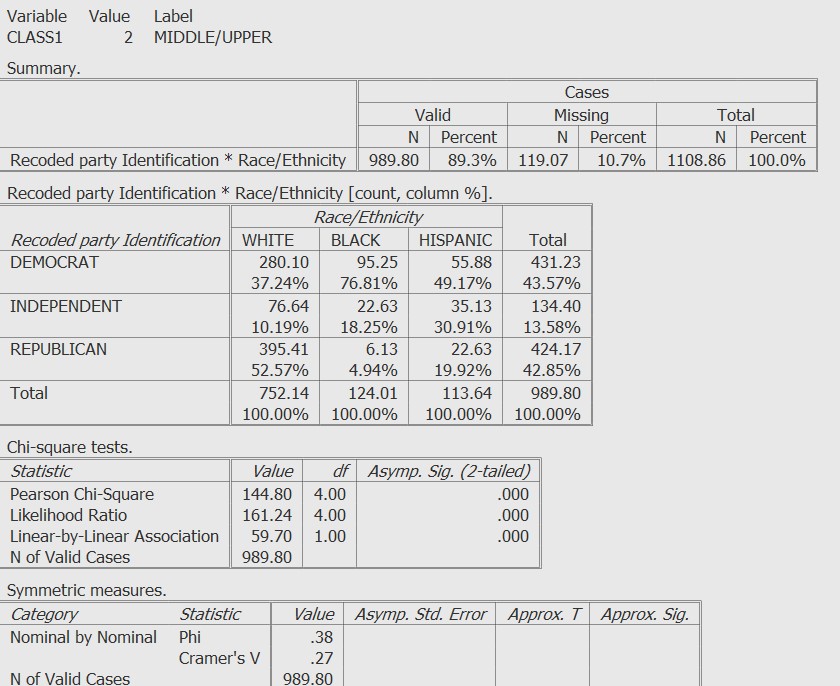
2. Use SELECT CASES to filter **out** cases for which *class* has a missing value. If you omit this step, you will get a separate sub table for each value of *class1*, including a meaningless table for the handful of cases for which the respondent’s social class is not reported. To filter out these cases, click on DATA, then on SELECT CASES. Select *class1* from the list of variables on the left, then click on the right arrow to move it under USE FILTER VARIABLE. Warning: be sure that UNSELECTED CASES are FILTERED. If they are DELETED, they will be dropped from your data file. Click on OK. (If you then wish to run another procedure without a filter, first go back to SELECT CASES and click on SELECT… ALL CASES.)

3. Split the file by *CLASS1*. Click on DATA, then on SPLIT FILE. Click on ORGANIZE OUTPUT BY GROUPS, then select CLASS*1* from the list of variables on the left and, using the right arrow, move it over to the window on the right under GROUPS BASED ON. Click on OK. (If you then wish to run another procedure without splitting the file, go back to SPLIT FILE and click on ANALYZE ALL CASES. DO NOT SELECT GROUPS.)

4. Run the crosstab with *partyid1* in the rows and *ethnicity* in the columns, asking for COUNT and COLUMN percentages in each cell, and the CHI-SQUARE and PHI statistics. The results should look like Figures 6-3 and Figure 6-4.



**Figure 6.3**



**Figure 6-4**

By comparing these tables, we can see that inter-ethnic differences are about the same regardless of class. In both tables, the values of Cramer’s V are almost the same and, in both cases, the relationship is highly unlikely to be due to chance.

Example 4: Party identification by ideology, controlling for level of education.

For a variety of reasons, college graduates are probably more likely than non-grads to think of politics in ideological terms like “liberal” (or “progressive”) versus “conservative.” (For example, college grads have likely been more heavily exposed in lectures and readings to the use of such terms to describe differences among politicians and parties.) If this is true, we would anticipate that the party identification of college graduates would bear a closer relationship to ideology than would be true for non-grads. We can split our file into these two categories, and run a contingency table for each. If our hypothesis is correct, the relationship between party id and ideology should be stronger for college grads.

1. Recode *degree* into a new variable (which we’ve called *degree1* with just two categories: those with less than a baccalaureate degree (codes 0 through 2), and those with a college degree (codes 3 and 4). By the way, when introducing a control variable into contingency table analysis, it doesn’t matter what the level of measurement of the control variable is. See Chapter 3 to review recode procedures.

2. As discussed above with example 3, when introducing a control variable into a crosstab, we will usually want to exclude any cases that have a missing value for that control variable. We can use SELECT CASES to do that. Since, as it happens, there are no missing values for *degree1*, we can skip this step here.

3. Split the file as we did in example 3, but substitute *degree1* instead of *class1*.

4. Set up and run a crosstabs table identical to the one you produced for the two-variable analysis we ran above in exercise 2. Figures 6-5 and 6-6 display, respectively, the results when respondents are divided between those who don’t have college degrees and those who do.

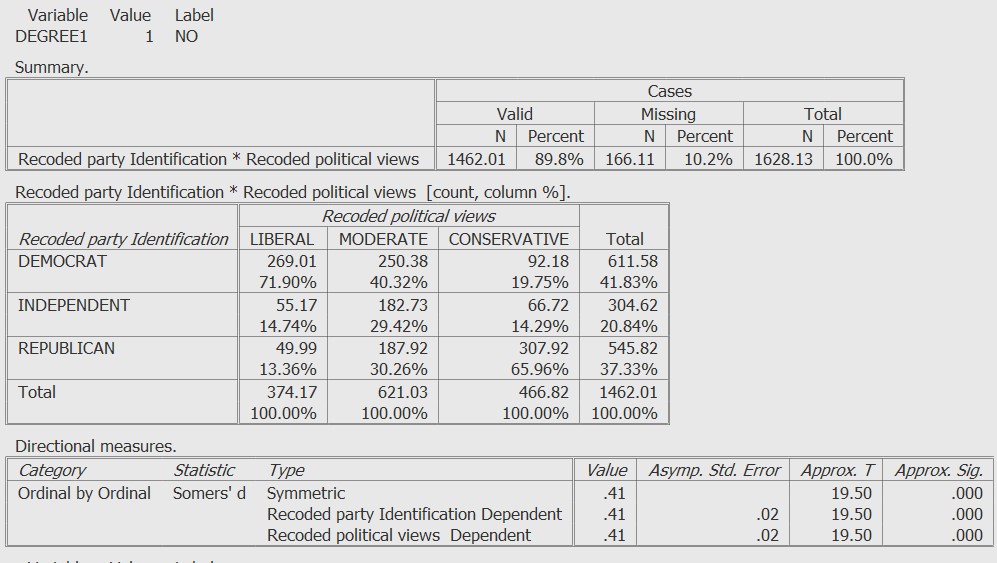
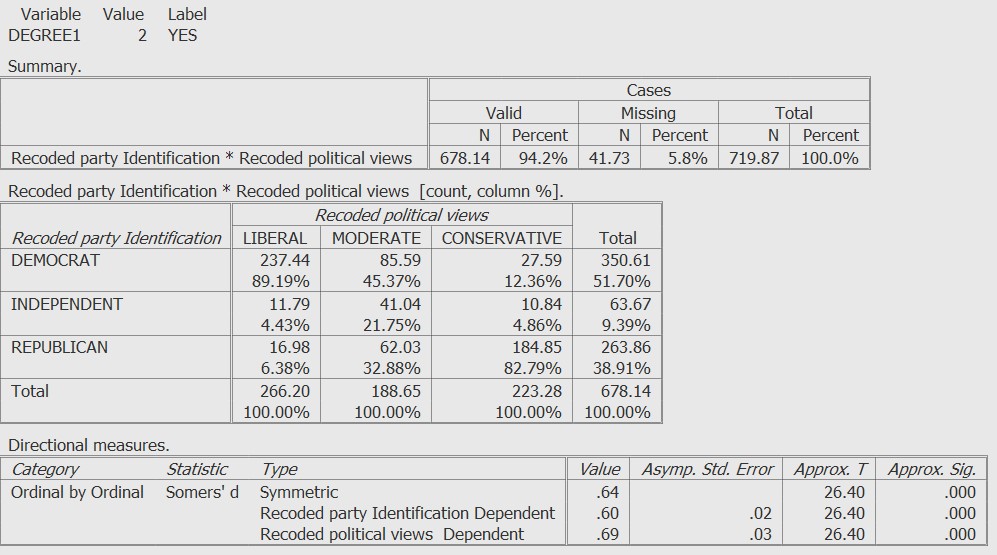


Figure 6-5



**Figure 6-6**

In both figures, the chi-squares are highly significant statistically (p<.001); in each case, the odds are less than one in a thousand that the results are due to chance. However, the value of Somer’s D in Figure 6-6 is higher than in Figure 6-5. Comparing the two contingency tables, we see that party identification is more closely associated with ideology among college grads than among non-grads.

## **Next Chapter**

In the next chapter we’re going to discuss correlation and regression. These are very powerful techniques, but they generally require variables that are interval or ratio level.

**Chapter 7**

**Correlation and Regression**

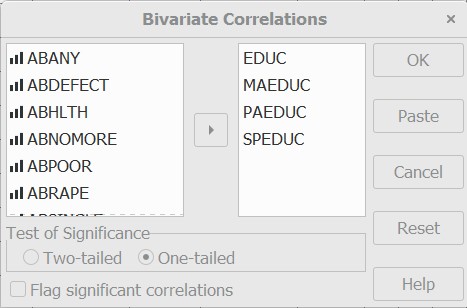
In this chapter, we’ll show how PSPP can be used to 1) calculate the degree to which two variables co-vary (correlation), and 2) develop models in which several independent variables can be used to predict the value of one dependent variable (regression). Our discussion will be limited to techniques used to analyze variables measured at the interval or ratio level.

**The Correlation Coefficient**

The values of two variables may co-vary because they are measuring the same or similar underlying concepts, because they are causally related, or because they are both related to one or more additional variables. To measure the degree to which two variables co-vary we can calculate a correlation coefficient. Unless otherwise specified, a correlation coefficient refers to the “Pearson product-moment correlation coefficient,” represented by the symbol “r.” This coefficient requires interval or ratio data. Pearson’s r is useful if you wish to know the extent to which, as one variable increases in value, the value of another increases or decreases in a linear manner. The value of r will range from 0.00 (no relationship) to ±1.00 (a perfect positive or negative relationship). A positive value indicates that as one variable increases, the other variable tends to increase as well. A negative value indicates that as one variable increases, the other variable tends to decrease. The relationship can also be represented graphically, as shown below.

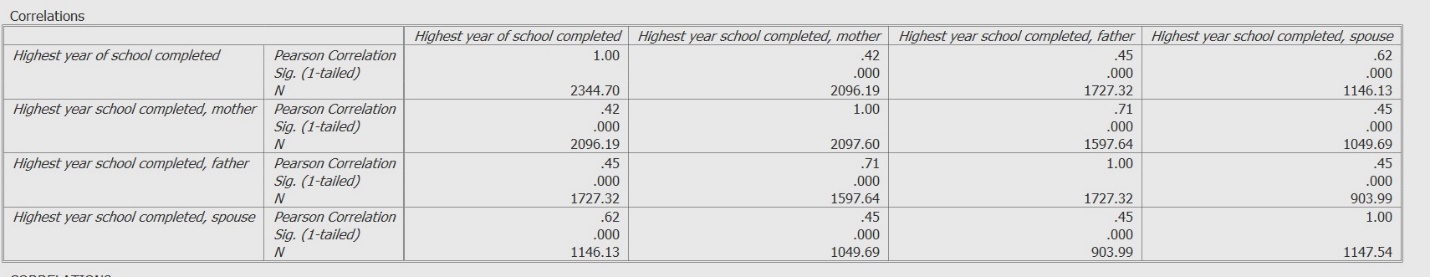
The General Social Survey includes several questions about respondents’ education. Respondents were asked how many years of education they had (*educ*). They were also asked the same question about each of their parents (*maeduc* and *paeduc*). Married respondents were also asked about their spouses’ education (*speduc*). We would expect these relationships to be positive. The education of one’s parents would likely be a causal factor in the amount of education received by a respondent. The relationship between a mother’s and a father’s education, or between a respondent and a spouse, is more likely a case of “homogamy” or the tendency of people to marry others of similar backgrounds (also called “associative mating”).

To calculate a “matrix” of correlation coefficients, click on ANALYZE in the menu bar. From there click on BIVARIATE CORRELATION. From the menu that appears, select *educ* in the left window, and move it over to the right window. Do the same for *maeduc*, *paeduc*, and SPEDUC. Under TEST OF SIGNIFICANCE, select ONE-TAILED, because we are predicting the direction (positive) of all relationships. The result should look like Figure 7-1.



**Figure 7-1**

Now click on OK. The result should look like Figure 7-2.



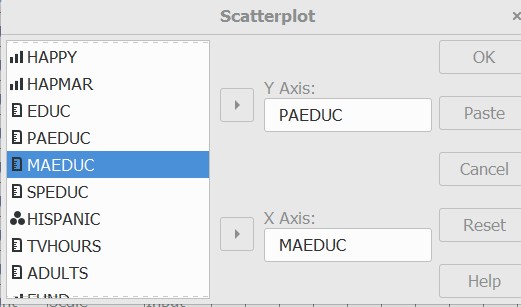
**Figure 7-2**

The first entry in each cell shows the value of r, the second the one-tailed significance level, and the third the number of cases with valid values for both variables. By the way, the square of r (not-surprisingly, r2) is known as the coefficient of determination. This is a very useful number: it tells us how much of the variation in one of the variables can be “explained by” (or associated with) the other variable. For example, if r2 is .40, then 40% of the variation in one variable can be explained by the other variable.

From Figure 7-2, we can see that, as expected, each variable has a moderate, positive correlation with each of the others, and that all relationships are statistically significant.

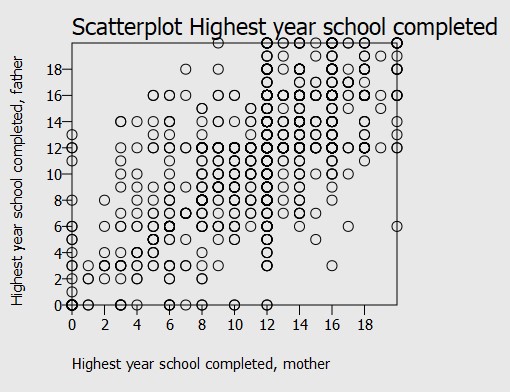
**Scatterplots**

We can also create a graph showing the relationship between any two variables that are at least interval. For example, we might wish to examine the extent to which the parents of our respondents came from similar educational backgrounds. From Figure 7-2 we can see that the correlation between the number of years of education of respondents’ fathers and the number of years of respondents’ mothers is .71. We can see what this “looks” like as follows. Click on GRAPHS in the menu bar. In the dialog box that appears, select either *paeduc* or *maeduc* in the left window, and move it over to the right window under “Y-Axis.” More the other variable under “X-axis” so that the dialog box looks like Figure 7-3.



**Figure 7-3**

Since we aren’t hypothesizing that either one of these variables “causes” the other, it doesn’t matter which variable we put in which axis. Normally, it matters a great deal: by convention, the dependent variable **always** goes on the X axis, and the dependent variable on the Y axis. This convention is very well established. If you violate it, you will only cause confusion. Click on OK. The result will look like Figure 7-4.

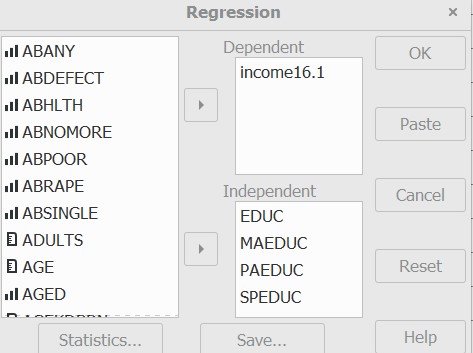


**Figure 7-4**

From Figure 7-4 we can see that, in general, the more years of education one parent has had (the higher the value X is), the more the other will have had as well (the higher the value of Y will be).

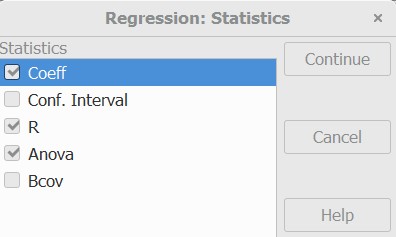
**Regression**

Suppose we want to find out how much education affects income. The GSS subset includes a measure of total family income in 2016 (*income16*) which as coded, is ordinal, but which we have recoded to make it approximately ratio, and called it *income16.1*[[22]](#footnote-22). Our independent variables will be *educ*, *maeduc,* *paeduc*, and *speduc*. To build a regression model, click on ANALYZE in the menu bar. From there click on REGRESSION, and LINEAR. From the menu that then appears, select *income16.1* in the left window, and move it over to the upper right window under DEPENDENT. Move *educ*, *maeduc,* *paeduc*, and *speduc* over to the lower right-hand window under INDEPENDENT. Your screen should look like Figure 7-5.



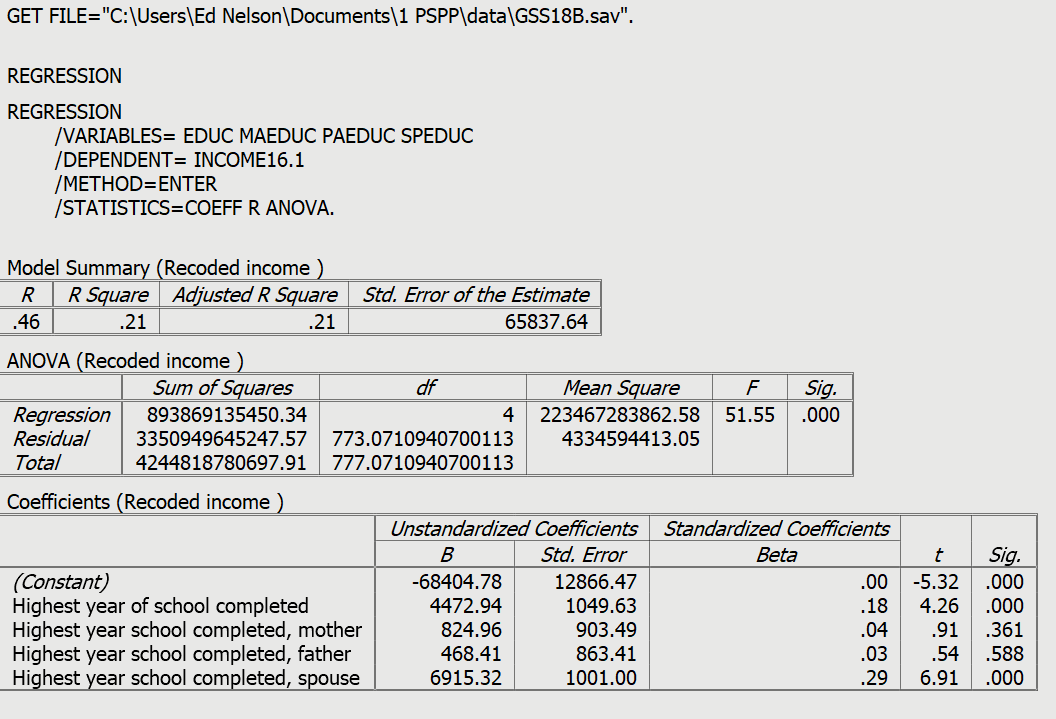
**Figure 7-5**

Click on STATISTICS. In the window that appears, you’ll see (Figure 7-6) that COEFF, R, and ANOVA are the defaults, so just click on CONTINUE. The results should look like Figure 7-6.



**Figure 7-6**

Now click on OK. The results should look like Figure 7-7.



**Figure 7-7**

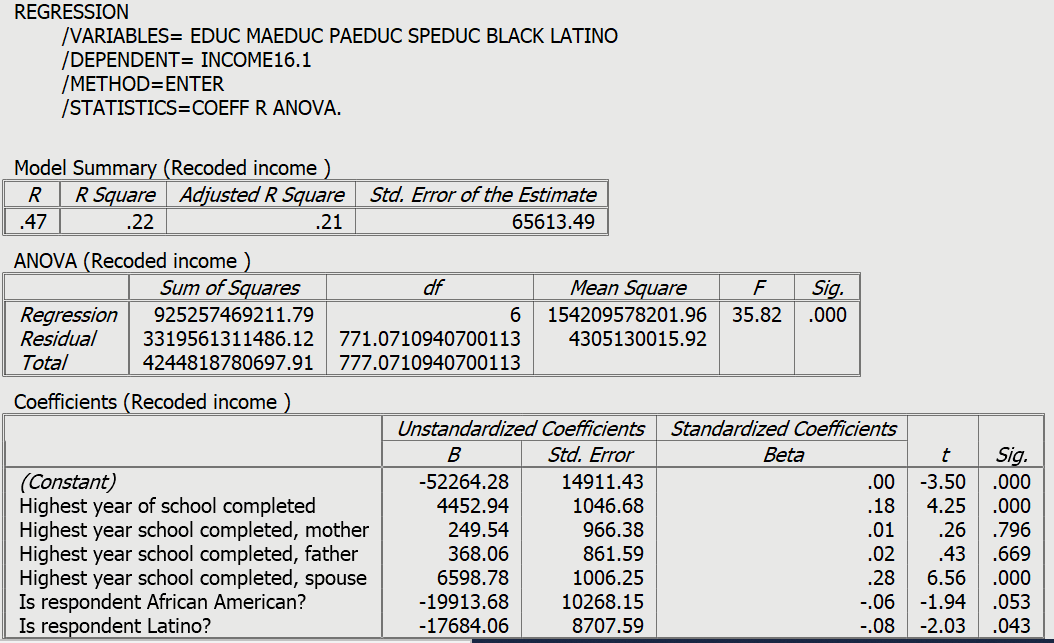
We’ve generated quite a bit of output. We’ll just look at the key numbers.

* From the first (model summary) table, we obtain the values of R, R2, and the adjusted R2 (which takes into account the number of variables in the equation.) The adjusted R2 is .21. This means that all four independent variables taken together explain about a fifth of the variation in total family income. Note that, in multiple regression (that is, with more than one independent variable), correlation coefficients are represented by upper case symbols.
* From the second (ANOVA) table, we see that the model as a whole is statistically significant. An R2 this high would occur by chance fewer than one time in five thousand.
* In the third table, we find the following.
  + The B coefficients show the change in the dependent variable (in dollars) for a change of one unit in any independent variable which, in our model, are all expressed in years of education. Because these coefficients are expressed in whatever units are used to code the variables in question, they are called “unstandardized” regression coefficients. For example, the coefficient for spouses’ years of education means that, holding all other variables in the equation constant, an increase of one unit (i.e., year) of spouse’s education is associated with an additional annual household income of $6,915.32.
  + The Beta (β) coefficients express all the variables as standardized units, with each variable having a mean of zero and a standard deviation of one. These are called “standardized” regression coefficients. The standardized coefficient of .29 for spouse’s education means that, again controlling for all the other variables in the equation, for every increase of 1 standard deviation in a spouse’s years of education, a respondent will have on average a household income that is .29 standardized deviations higher, which is more than for any other variable in the equation. Since all four independent variables are measured in similar units of analysis (years of education), it doesn’t matter much whether we use standardized or unstandardized coefficients. This won’t always be the case, as we’ll see below.
  + The t test is a test of significance that determines whether the coefficients are significantly greater than zero.
  + The level of significance for each t value. This tells us if the relationship between the dependent variable and each independent variable is statistically significant, again controlling for all the other independent variables.

When an independent variable is a dichotomy (that is, has just two values), it can be entered into a regression equation like any other variable.  Called ***dummy variables*** in this context, dichotomies are usually coded “0” and “1.”   To illustrate this, let’s look at the effect, if any, that ethnicity has on the relationship between education and family income. The relevant variable in our dataset, *ethnicity*, has four categories: WHITE, BLACK, HISPANIC, and OTHER. Since there are few cases in the “OTHER” category, and since it includes respondents from a variety of ethnicities, these cases won’t be included in the analysis. OTHER has the value of 4 in our data set so tell PSPP to treat 4s as missing values. (See the instructions for doing this earlier.)

* Now that we have eliminated the OTHER category from our analysis, use the recode procedure to create two new variables One of these variables we’ll call *latiino* and it and will equal 1 if the *ethnicity* variable equals 3 (i.e., Hispanic) and 0 otherwise. The other variable we’ll call *black* which will equal 1 if the *ethnicity* variable equals 2 (i.e., Black) and 0 otherwise. We should not add a third variable for White respondents since, if we specify the value of a case for two categories, we have in effect already specified its value for the third. In more general terms, we use N-1 dummy variables in the regression equation for a variable having N categories.
* Click on ANALYZE in the menu bar. From there click on REGRESSION, and LINEAR. From the menu that then appears, add *black* and *latino* to the list of independent variables (see figure 7-5, above), Click on OK. The results should look like Figure 7-8. Both standardized and unstandardized coefficients are useful in interpreting results. In both cases, however, the relationship between the dependent variable and each independent variable is calculated holding constant (or “controlling for”) all the other independent variables.

.



**Figure 7-8**

Here we can see more clearly the different way in which unstandardized and standardized regression coefficients are interpreted. The unstandardized (*B*) coefficients for the two dummy variables compare *black* and *latino* with *white*, which was the category we omitted when we created our dummy variables. The omitted category always becomes the comparison group. The coefficient for *black* is -19913.68. This means that Blacks have an average income that is $19,913.68 less than Whites controlling for all the other variables in the equation. Latinos have an average income that is $17,684.06 less than Whites controlling for all the other variables.

While the unstandardized coefficients for the dummy variables are very high, the standardized coefficients are low. A frequency distribution of *ethnicity* (not shown) would reveal that, although they are represented in rough proportion to their share of the U.S. population, Black and Latino respondents are a relatively small share of the sample, and so could not possibly contribute much to an overall explanation of the variation in income. It’s worth noting, however, that the regression coefficient for Latinos is statistically significant, and the coefficient for Blacks falls just short of statistical significance.

**Next Chapter**

In Chapter 8, we’ll show you how to prepare a research report using output generated with PSPP.

**Chapter 8**

**Writing Research Reports**

This chapter will focus on how to write research reports including:

* how to organize your report,
* creating tables,
* whether to use footnotes or endnotes,
* citing articles, papers and other materials,
* avoiding plagiarism,
* proofreading, and
* other guides to writing reports.

**An Outline of your Research Report**

In the previous chapters we discussed how to use PSPP to analyze your data. We talked about using PSPP to describe your data, analyzing the relationship between pairs of variables, and extending our analysis to include sets of three or more variables. Now we need to think about how to write a research report so that others may read it and learn from our analysis.  This report might be for a class you are taking, or it might be a report that you are submitting to a research conference.  If you are going to submit your report to a journal for possible publication, you need to look carefully at the instructions that all journals provide on preparing a manuscript for publication.

Here's an outline for your report.  Don't think that this is the only way you can organize your report, but this is one way to do it.

* Title page including your name, date, and class or institutional affiliation.
* Abstract – An abstract is a short summary of what you did in the paper and the major findings of your analysis.  Abstracts are really short, so you need to be succinct. It should be less than 200 words or even shorter depending on the requirements of your professor or the research conference to which you are submitting your paper.
* Table of contents (optional).
* Body of the paper.
  + An introduction to the paper which explains why you wrote the report and provides an introduction to the topic of the paper.
  + Your review of the literature that summarizes what others discovered about this topic.  Almost everything you might do has been written about by others.  You should review the relevant literature and summarize what others have found.  You don't want to simply list the relevant literature and consider the articles and books one by one.  Rather you want to summarize what others have done and look for themes around which you can organize your literature review.  If you are having trouble finding relevant literature, go to the library at your university or a nearby university and talk with a reference librarian.  They are trained in searching for relevant literature and will be able to help you.
  + The methodology of your study.
    - If you collected your own data, discuss how you chose your sample, how you measured the concepts, and how you collected your data.
    - If you used an existing data set, discuss the sampling (if using sample data), measurement, and data collection used in that study.  Studies that are part of data archives such as the Inter-university for Political and Social Research at the University of Michigan and the Roper Center for Public Opinion Research at Cornell University provide good summaries for all data sets that are housed at their archive.
  + Theory and Hypotheses – If you are using a theoretical perspective and/or testing hypotheses, describe the theory and state the hypotheses you plan to test.  Be sure to cite supporting literature that form the basis for your theory and hypotheses.
  + Empirical findings and interpretation – What are the empirical findings that came out of your data analysis and what did they tell you?  If you are testing hypotheses, did your analysis support your hypotheses?  Remember that you are telling a story.  Start simple and build up.  That means starting with looking at variables one at a time (i.e., univariate analysis), then proceeding to relationships between pairs of variables (i.e., bivariate analysis), and then looking at sets of three or more variables (i.e., multivariate analysis) to consider such things as spuriousness.
  + Conclusions and summary. This is a little like your abstract but not as short.  What did you do, what did you find in your study and what does it mean?
* Tables.  You may choose to put your tables in the body of your paper, or you may decide to put them all at the end of your paper.
* References.  For every article or book that you cite, you need to provide a full bibliographic reference at the end of the report.

**Tables**

There are advantages and disadvantages to putting your tables in the body of the report or at the end of the report.  Putting them in the body of the report keeps them front and center for the reader but they often are bulky and get in the way of reading the report.  Putting them at the end of the report gets them out of the way and allows the reader to spread them out and look at them as he or she is reading the paper.  Your instructor or the research conference will usually tell you where to put your tables.

If they are placed at the end of the paper, put a note in the body of the report that says something like "Table 1 about here."  That will let the reader know where the table fits into your report.

Constructing a good table is important.  Sometimes your instructor will tell you to copy tables from the program you are using for statistical analysis (e.g., PSPP, SPSS, and others) into your paper.  Other times you will construct the tables yourself.  A good reference on creating tables is *The Chicago Guide to Writing About Numbers* by Jane E. Miller.[[23]](#footnote-23)  Your word processing program (e.g., Word in Microsoft Office) will provide you with templates that you can choose for your tables.

**Footnotes or Endnotes**

Often you want the reader to be aware of something, but you don't want to put it in the body of the paper.  It may be a technical issue such as how you recoded a variable or why you chose a particular statistic.  Or you may want to tell the reader that you will discuss something later in the paper.  You can put comments like these in either a footnote or an endnote.  A footnote goes at the bottom of the page and an endnote goes at the end of the paper.  Your word processing program will allow you to enter either footnotes or endnotes in your paper.  Which you use is up to you unless your instructor or the research conference tells you that one or the other is required.

**Citing Articles, Papers, and Other Materials**

There are many styles such as American Psychological Association (APA) or Modern Language Association (MLA) that you could use to cite materials that you refer to in your paper.  Remember that anytime you refer to someone else's work, you must acknowledge the source.  Your instructor or research conference will often specify which style you should use.

**Avoiding Plagiarism**

Plagiarism is using someone else's words or ideas without acknowledging the source.  If you are quoting from a document, you must cite the source.  Even if you are paraphrasing, you must acknowledge the source.  If you are using someone else's ideas, you must also acknowledge the source.  There is a good review of plagiarism written by Earl Babbie that can be found on the Internet by clicking [here](http://www1.chapman.edu/~babbie/plag00.html).  Click on the red arrows at the top to go forward or backward in this review of plagiarism.

**Proofreading**

Be sure to proofread your paper several times before submitting it. Use the spell and grammar checker in your word processing program. You could also ask a friend to read it and tell you about any errors or parts that are confusing.

**Other Guides to Writing Reports**

There are many other guides to writing research reports.  One that is commonly used in Sociology is the *Guide to Writing Sociology Papers*.[[24]](#footnote-24)  You can find others on the internet by entering "writing research reports" in the search box.

# **Appendix A**

# **Codebook for the Subset of the 2018 General Social Survey**

The General Social Survey (GSS) is a large, national probability sample of adults in the United States. It began in 1972 and continued on an almost yearly basis until 1996. In 1996, the GSS became a biannual survey and the sample size increased. Many questions are asked on each survey, while other questions are rotated from survey to survey. This subset from the 2018 GSS includes all the cases (2,348) and 97 variables. This data set has already been weighted using the weight variable supplied by the GSS (*wtss*). Some of the original GSS variables were recoded and a few new variables created. Some of the new variables have names similar to those in the original GSS data set. The data set was created to accompany this PSPP tutorial and is named GSS18A.SAV.

**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

BLACK Is respondent African American?

CAPPUN Favor or oppose death penalty for murder

CHILDS Number of children

CHLDIDEL Ideal number of children

CLASS Subjective class identification

CLASS1 Recoded from subjective class identification [CLASS]

COHORT Cohort

COLATH Allow anti‑religionist to teach

COLCOM Allow communist to teach

COLHOMO Allow homosexual to teach

COLMIL Allow militarist to teach

COLMSLM Allow anti-American Muslim Clergyman to teach in college

COLRAC Allow racist to teach

DEGREE Respondent's highest degree

DEGREE1 Recoded from R’s highest degree [DEGREE] – does respondent have a four-year college or postgraduate degree

DENOM Specific Protestant denomination

EDUC Highest year of school completed

ETHNICITY Respondent’s race/ethnicity[[25]](#footnote-25)

FAIR People fair or try to take advantage

FEAR Afraid to walk at night in neighborhood

FINRELA Opinion of family income

FUND Fundamentalism of respondent’s religion

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 respondent worked last week

HRS2 Number of hours respondent usually works a week

ID Respondent’s identification (id) number

INCOME16 Total family income (2017)

INCOME16.1 Recoded from total family income (2017) [INCOME16]

LATINO Is respondent Latino?

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

LIBMSLM Allow anti-American Muslim clergyman'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

MASEI10 Mother’s socioeconomic status using scale developed in 2010

PADEG Father's highest degree

PAEDUC Highest year school completed, father

PARTYID Political Party Affiliation

PARTYID1 Recoded from Political Party Identification [PARTYID]

PASEI10 Father’s socioeconomic status using scale developed in 2010

POLVIEWS Think of self as liberal or conservative

POLVIEWS1 Recoded from Think of self as liberal or conservative [POLVIEWS]

PORNLAW Feelings about pornography laws

POSTLIFE Belief in life after death

PRAY How often does respondent pray?

PRAYER Support Supreme Court Decision on prayer in public schools

PRES12 Vote for Romney or Obama in 2012

PRES16 Vote for Clinton or Trump in 2016

RACE Race of respondent

REGION Region of interview

RELIG Respondent's religious preference

RELIG1 More detailed breakdown of religious preference

RELITEN Strength of religious affiliation

RELPERSN Respondent considers self a religious person

SATFIN Satisfaction with financial situation

SEI10 Respondent’s socioeconomic status using scale developed in 2010

SEX Respondent's sex

SIBS Number of brothers and sisters

SIZE Size of place respondent lives in thousands

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

SPKMSLM Allow anti-American Muslim clergyman to speak

SPKRAC Allow racist to speak

SPSEI10 Spouse’s socioeconomic status using scale developed in 2010

TRUST Can people be trusted?

TVHOURS Hours per day watching television

VOTE12 Did respondent vote in 2012?

VOTE16 Did respondent vote in 2016?

WTSS Weight variable for GSS18 (data subset already weighted by the variable WTSS)

YEAR Year of survey (2018 for all respondents)

ZODIAC Respondent's astrological sign

# **Appendix B**

# **Working with PSPP**

SPSS is the most widely used statistical package in the California State University System and is the package that the authors are most familiar with. However, some colleges, including many community colleges, do not have a site license for SPSS. The cost of SPSS is often prohibitive, so we’re writing this PSPP tutorial for beginning users. For more information on PSPP, click [here](http://www.gnu.org/software/pspp/). Their website says that “GNU PSPP is a program for statistical analysis of sampled data. It is a Free replacement for the proprietary program SPSS and appears very similar to it with a few exceptions.”

The easiest way to download PSPP is click [here](http://pspp.awardspace.info/) and look for the “Downloads” box. Then download the latest version in either 32-bit or 64-bit format. If you’re not sure which version to download, go to the control panel and click on “System” and look for your system type. Then follow the instructions to download.

You can open SPSS data files (both .sav and .por) in PSPP. You can also open Excel files (both .csv and .txt). And, of course, you can create your own data file in PSPP. Instructions are in Chapter 2.

You can run PSPP from either the interface or in syntax mode. In this tutorial, we’re going to focus exclusively on using the interface by pointing and clicking at appropriate points.

Some PSPP commands do not have the full capabilities of their corresponding SPSS commands. Other PSPP commands work somewhat differently from their SPSS counterparts. For example, running a three-variable table in PSPP is different from how you do it in SPSS. You can use SELECT CASES in PSPP, but it works differently than in SPSS.

You can also run SPSS syntax (.sps) files in PSPP but we don’t recommend this for beginning students. Most students are not familiar with writing syntax in SPSS and you wouldn’t want to try teaching syntax to most undergraduates, and not a few graduate students.

PSPP will list the variables in your data file and you can select those variables you want to use. It’s easier to find the variables if they are listed by variable names. You can change the way PSPP lists the variables by right clicking anywhere on the list of variables and checking or unchecking the box for PREFER VARIABLE LABELS. If you want your list of variables to be in alphabetical order, right click on the list and select SORT BY NAME. Note: Even if the SORT BY NAME button is already selected, check to be sure that variable names are, in fact, in alphabetical order. If this is not the case, select DEFAULT SORT ORDER, then select SORT BY NAME.

PSPP is very limited in terms of what you can do with charts and graphs. There is another free software package called SOFA ([Statistics Open for All](http://sofastatistics.com/home.php)) that you can use for charts and graphs.

We’re using various conventions in this tutorial. These include the following.

* All variable names in the text are lower case with italics.
* All PSPP commands and dialog choices are in upper case.
* Everything that the user enters is bolded.

Since most students are probably not familiar with SPSS, we won’t mention differences between PSPP and SPSS in this tutorial. Rather, we will note a few major ones here in this appendix.

* Analyze/Compare Means/Means works differently in PSPP than in SPSS. For example, in PSPP if you want to compute the mean number of hours that respondents watch television daily (*tvhours*) by marital status (*marital*) and sex (*sex*) you would put *marital* and *sex* in the INDEPENDENT LIST box. This will compute the mean *tvhours* for each of the ten categories (i.e., male-married, female-married, male-widowed, female-widowed, etc.). If you wanted to compute the mean *tvhours* first by *marital* and then by *sex,* you would put *marital* in LAYER 1 and *sex* in LAYER 2. This would give you the mean hours first for the five marital categories and then separately for the two sex categories.
* Select Cases. Use SELECT CASES to filter cases with missing values **out** of the analysis. If, for example, we use *cappun* as a filter variable, indicate that unselected cases are filtered (not deleted), and then run *FREQUENCIES* for *cappun.* This means that you will be left with only cases with non-missing values. For a step by step example, see Chapter 6.
* Select If. In SPSS, if we wanted to do an analysis of only same cases in the file (let’s say, Millennials only), we could easily do this with SELECT IF. In PSPP, there is no such procedure in the GUI version, but there is a work-around. Suppose we wanted to do an analysis that would include only Millennials. We could:
  + recode cases into a new variable with just two categories, Millennials (coded 1) and those older or younger (coded 2), with 2 treated as a missing value.
  + SELECT CASES, thereby filtering out non-Millennials.
  + Run the analysis.
* Split File. To carry out a three-variable crosstab in PSPP, use SPLIT FILE to divide the data into two or more categories, and then run a two-variable crosstab for each category. For a step by step example, see Chapter 6.

1. Usually referred to as “level of measurement,” which is discussed in greater detail in Chapter 4. For now, note that what is usually called “interval” and what is usually called “ratio” are both called “scale” in PSPP. [↑](#footnote-ref-1)
2. There are two types of data files in SPSS -- .sav and .por files. PSPP will open both types of files. [↑](#footnote-ref-2)
3. There are many data archives where you will find data files in SPSS format. There are membership consortiums such as the Inter-university Consortium for Political and Social Research and the Roper Center for Public Opinion Research. There are other data archives that you don’t have to join such as the Pew Research Center and the Public Policy Institute of California. For an extensive list of data archives, click [here](https://libguides.princeton.edu/politics/opinion). [↑](#footnote-ref-3)
4. There are many other formats that data might be saved in such as SAS and Stata. To open these types of files you will need to resave the files in a format that PSPP can open. There is a commercial program called Stat/Transfer that will do this for you. However, it is moderately expensive. There is, however, an academic discount for Stat/Transfer. To find out more about Stat/Transfer, click on this [link](https://stattransfer.com/). [↑](#footnote-ref-4)
5. For a more detailed explanation of weighting and for examples, click [here](https://www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Weighting.aspx). [↑](#footnote-ref-5)
6. See Chapter 2 for instructions on how to name a variable. [↑](#footnote-ref-6)
7. Notice that the arrow changes direction. When it points to the right, it moves whatever is highlighted to the right. When it points left, it moves it to the left. [↑](#footnote-ref-7)
8. There are also other options. For example, you can work with what are called “system-missing” values. All blanks will automatically be changed to system-missing values. You can recode 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. [↑](#footnote-ref-8)
9. All respondents who are 89 or older are coded 89. [↑](#footnote-ref-9)
10. The output window is cumulative. As you request different analyzes from PSPP, it adds them to the bottom of the output file. So, you will have to scroll down to see what you most recently requested. [↑](#footnote-ref-10)
11. A nominal variable consists of a set of unordered categories. [↑](#footnote-ref-11)
12. Valid responses refer to all values except those that indicate missing values. [↑](#footnote-ref-12)
13. PSPP will run SPSS commands in syntax form. You don’t have to modify the SPSS commands to run them. This can be particularly helpful for those who are familiar with SPSS. [↑](#footnote-ref-13)
14. See also Dan Osherson and David M. Lane, Levels of Measurement, <http://onlinestatbook.com/2/introduction/levels_of_measurement.html> [↑](#footnote-ref-14)
15. Remember that our population is actually all non-institutionalized adults (18+) adults in the continental United States. That takes a long to time to say. We’ll just refer to this as all adults in the U.S. for brevity. [↑](#footnote-ref-15)
16. We’re using 12 because that’s represents a high school education. [↑](#footnote-ref-16)
17. It doesn’t matter which is VAR1 and which is VAR2. [↑](#footnote-ref-17)
18. Remember that our hypotheses always refer to population values. [↑](#footnote-ref-18)
19. The fact that both variables have three categories is coincidental. [↑](#footnote-ref-19)
20. The sign can be misleading, since it depends on the way the values are coded. Here 1 is liberal, 2 is moderate, and 3 is conservative. They could be coded as 1 for conservative, 2 for moderate, and 3 for liberal. This would flip the sign of D. [↑](#footnote-ref-20)
21. PSPP provides several other measures of association for ordinal variables, but only gives the statistical significance level for D. For fuller discussions of significance and association in contingency table analysis, see Edward Nelson, *General Social Survey (2014): Statistics (PSPP version)*, exercises 9-11 (2016). [*https://ssric.org/node/559*](https://ssric.org/node/559) (Accessed February 5, 2019) and/or John Korey, ”Contingency Table Analysis,” in *Introduction to Research Methods in Political Science* (2009). [*https://www.cpp.edu/~jlkorey/POWERMUTT/Topics/contingency\_tables.html*](https://www.cpp.edu/~jlkorey/POWERMUTT/Topics/contingency_tables.html) (Accessed February 5, 2019). [↑](#footnote-ref-21)
22. The original variable groups income into ranges (e.g., $35,000 to $39,999). For each category except for the highest ($170,000 and over) we have assigned the middle of the range. For the highest (open ended) category, we have arbitrarily assigned the value of $250,000. [↑](#footnote-ref-22)
23. Jane E. Miller. *The Chicago Guide to Writing About Numbers*. 2015 (2nd edition). Chicago: University of Chicago Press. [↑](#footnote-ref-23)
24. Sociology Writing Group. *A Guide to Writing Sociology Papers*. 2013 (7th edition). New York: Worth Publishers. [↑](#footnote-ref-24)
25. This variable was created by combining responses to a question asking the respondent’s race (coded as White, Black, and Other), and another question asking whether the respondent is Hispanic. Any respondent identifying as Hispanic was so classified, regardless of race. [↑](#footnote-ref-25)