Data and Analytics in SPSS: A Step-By-Step Guide to Analysis and Interpretation

Henry Njagi

2021-10-05

Contents

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Introduction

2.1 Preliminary

The use of SPSS accommodates all people including the novice computer user and those with no previous experience using SPSS. The study chapters are divided into sub-sections that explains the statistics to be used, with underlying assumptions, and methodology for the results interpretation and express then in a research report.

The genesis of the book explains how to start the SPSS, variable definitions, data entry and saving. All the necessary and fundamental statistics techniques are explained. These statistics included the descriptive statistics, data graphing, association and prediction, inferential statistics for parametric and nonparametric, and statistics for test construction.

The learners are expected to follow up with the sample screenshots.

New to this edition 1:

Data and real output are now available for all Phrasing Results sections – eliminating hypothetical output or hypothetical data

Henry Njagi is a lecturer and a Statistician with advanced knowledge in data and Analytics (MSc. in Applied Statistics, Jomo Kenyatta University of Agriculture and Technology).

Chapter 3

Getting Started

3.1 Starting SPSS

We want to appreciate different users are using different operating systems. The procedure to startup the SPSS may differ slightly. This part uses screenshots from the window version of SPSS. Other window versions such as Unix and/or MacOS will have the same functionality, but could appear differently than what is being depiced herein.

When you start the SPSS, you will be presented with the above dialog box. This depends with the system administrator options that is working with your version program. On the dialog box, click Type in data and OK. Or just (Close). This will present a blank data window.

The provided blank window have basic interface for SPSS.

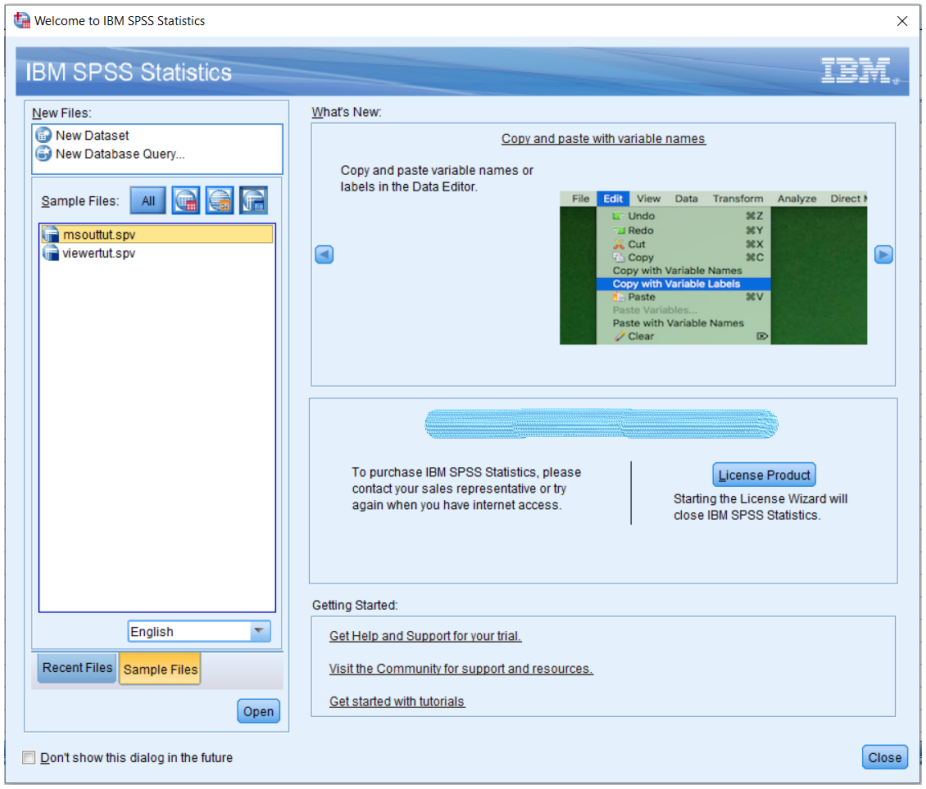
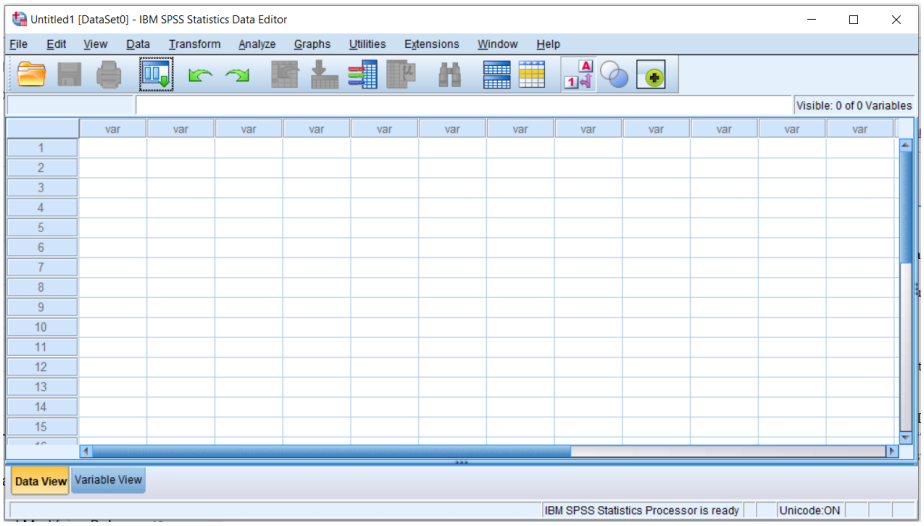


Figure 3.1: SPSS Version 25.



Entering Data

Before you get to know the operability of SPSS, you must understnd the key to success with SPSS. You should understand how SPSS stores data and uses the data. In order to illustrate the hasics data entry with SPSS, we will use an example 2.1.1

Example 2.1.1

A survey was given to different students from a college who were studying different courses (Enginnering, Economics, Computer Science, and Agriculture). The students were asked whether or not they revise extensively in the morning or late in the evening. This survey also asked for their first semester grade in the class (100% being the highest grade possible for all the courses). The response sheets fro the first 5 students are presented below

StudetID	Course	Revion Time	Grade %
1	Engineering	Morning	98
2	Agriculture	Evening	72
3	Computer Science	Morning	90
4	Economics	Evening	77
5	Agriculture	Morning	89

The main goal is to enter the data from the five college students into SPSS for the analysis. We first need to determine the variables that will be entered. A variable is any informtion that will keep on varying among participants. The variables we need from example 2.1.1 are:

StudetID:

Course:

Revion Time:

Grade %

Recall the blank SPSS window, there is two button (Data View and Variable View) on the bottom-left of the window. In the data view, we have columns and rows, the columns represents variables, and rows represent participants. Since the variables we have defined from example 2.1.1 have only four variables, therefore, we will create a data file with four columns, and five rows, each representing an entry from individual respondent/student.

3.2 Defining Variables

Before we can enter any data, we must first enter some basic information about each variable into SPSS. For instance, variables must first be given names that

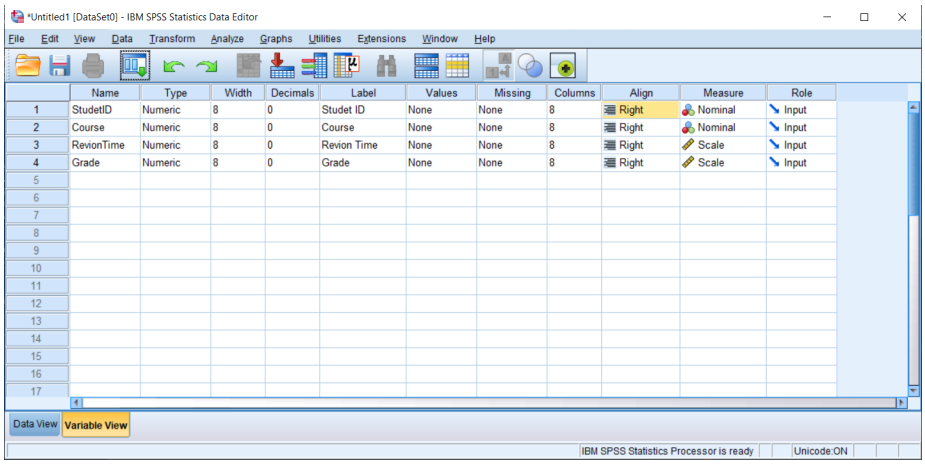
– begin with a letter, and – do not contain a space.

Thus, the variable name "Q7" is acceptable, while the variable name "7Q" is not. Similarly, the variable name "PRE_TEST" is acceptable, but the variable name "PRE TEST" is not. Capitalization does not matter, but variable names are capitalized in this text to make it clear when we are referring to a variable name, even if the variable name is not necessarily capitalized in screenshots.

To define a variable, click on the Variable View tab at the bottom of the main screen (refer to the SPSS Blank Window). This will show you the Variable View window. To return to the Data View window, click on the Data View tab.

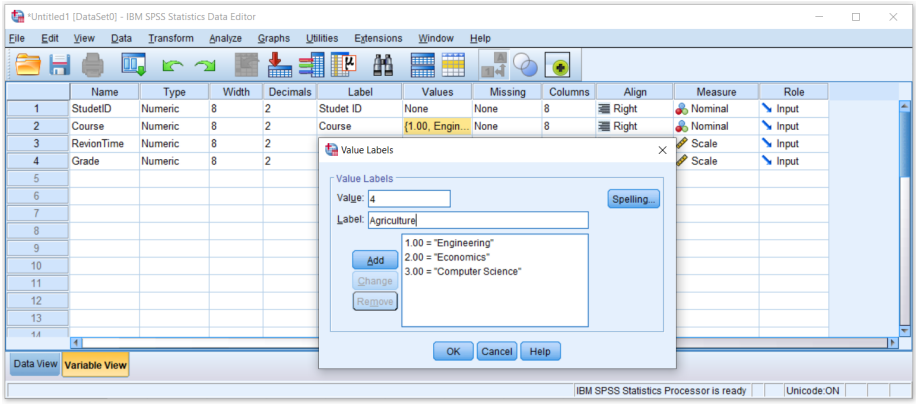
From the Variable View screen, SPSS allows you to create and edit all of the variables in your data file. Each column represents some property of a variable, and each row represents a variable. All variables must be given a name. To do that, click on the first empty cell in the Name column and type a valid SPSS variable name. The program will then fill in default values for most of the other properties.

One useful function of SPSS is the ability to define variable and value labels. Variable labels allow you to associate a description with each variable.



Value labels allow you to associate a description with each value of a variable. For instance, for most procedures, SPSS requires numerical values. Thus, for data such as the course (i.e., Engineering, Economic, Computer Science, and Agriculture), we need to first code the values as numbers. We can assign the number 1 to Engineering, 2 to Economic, 3 to Computer Science, and 4 to Agriculture. To help us keep track of the numbers we have assigned to the values, we use value labels.

To assign value labels, click in the cell you want to assign values to in the Values column (in this case, for Variable 2 i.e., the course). This will bring up a small gray button (shown below). Click on that button to bring up the Value Labels dialog box. The revision time should also be coded as 1 = Morning, and 2 = Evening.



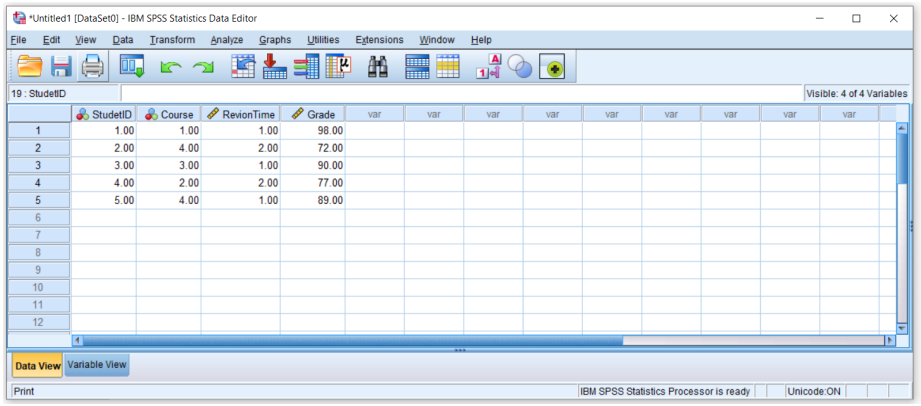
When you enter a value label, you must click Add after each entry. This will move the value and its associated label into the bottom section of the window. When all labels have been added, click OK to return to the Variable View window.

In addition to naming and labeling the variable, you have the option of defining

the variable type. To do so, simply click on the Type, Width, or Decimals columns in the Variable View window. The default value is a numeric field that is eight digits wide with two decimal places displayed. If your data are more than eight digits to the left of the decimal place, they will be displayed in scientific notation (e.g., the number 2,000,000,000 will be displayed as 2.00E+09). SPSS maintains accuracy beyond two decimal places, but all output will be rounded to two decimal places unless otherwise indicated in the Decimals column.

There are several other options available in this screen, which are beyond the scope of this text. In our example, we will be using numeric variables with all the default values.

After you have coded the book, click on the Data View tab to open the data-entry screen. Enter data horizontally, beginning with the first student's ID number. Enter the code for each variable in the appropriate column. To enter the GRADE variable value, enter the student's class/semester grade.



The previous data window can be changed to look like the screenshot on the next page by clicking on the Value Labels icon (see below). In this case, the cells display value labels rather than the corresponding codes. If data are entered in this mode, it is not necessary to enter codes, as clicking the button that appears in each cell as the cell is selected will present a drop-down list of the predefined labels. You may use whichever method you prefer.

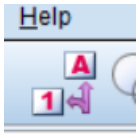


Figure 3.2: Data View label1.

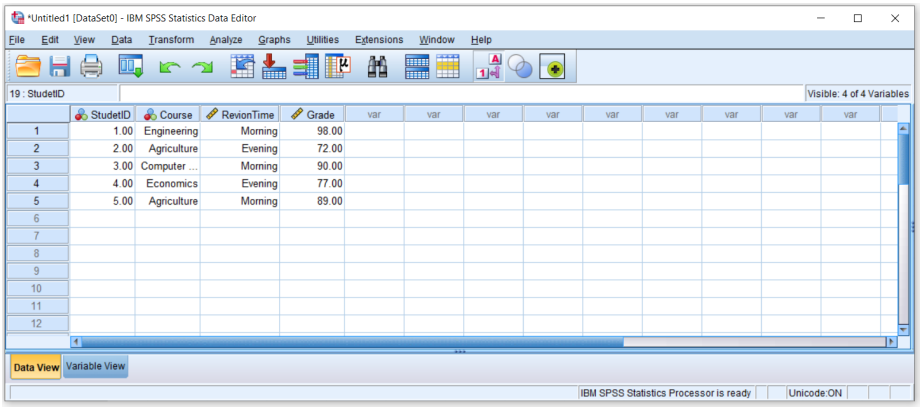


Figure 3.3: Data View label2.

3.3 Loading and Saving Data Files

Once you have entered your data, you will need to save it with a unique name so that you can retrieve it when necessary for later use.

Loading and saving SPSS data files works in the same way as most Windows-based software. Under the File menu, there are Open, Save, and Save As commands. SPSS data files have a ".sav" extension, which is added by default to the end of the filename (that is, do not type ".sav" after the filename; SPSS will add it automatically). This tells Windows that the file is an SPSS data file. Other SPSS extensions include ".spv" for saved output files and ".sps" for saved syntax files.

When you save your data file (by clicking File, then clicking Save or Save As to specify a unique name), pay special attention to where you save it. You will probably want to save your data on a removable USB drive so that you can take the file with you.

3.4 Load Your Data

When you load your data (by clicking File, then clicking Open, then Data, or by clicking the open file folder icon), you get a similar window. This window lists all files with the ".sav" extension. If you have trouble locating your saved file, make sure you are looking in the right directory.

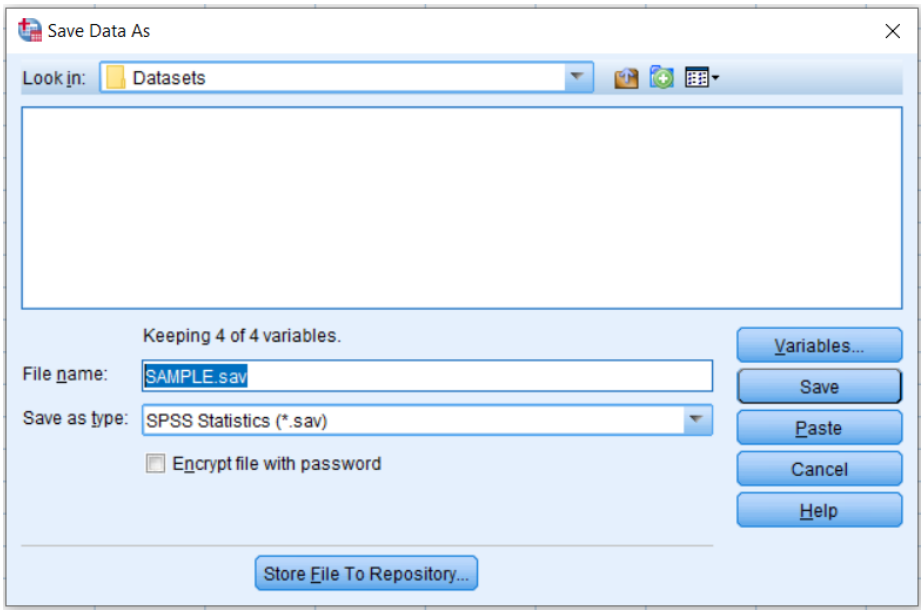


Figure 3.4: Saving File.

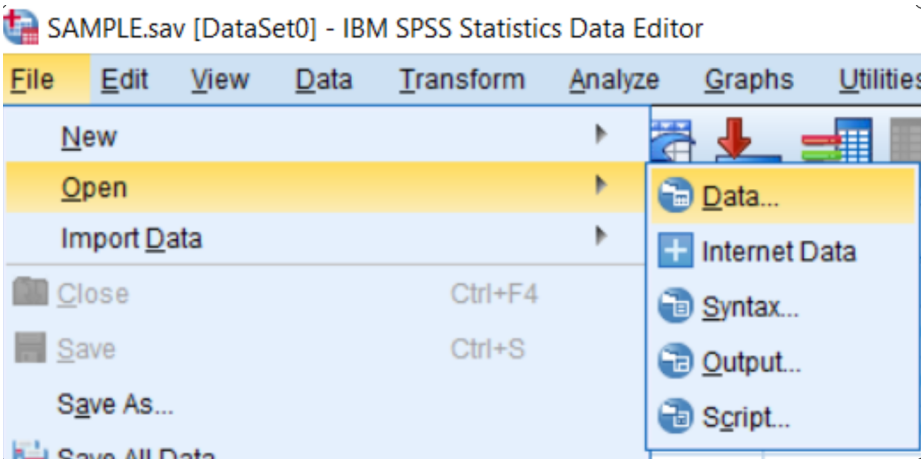


Figure 3.5: Saving File.

3.5 Running Your First Analysis

Any time you open a data window, you can run any of the analyses available. To get started, we will calculate the students' average grade. (With only five students, you can easily check your answer by hand, but imagine a data file with 50,000 student records.)

The majority of the available statistical tests are under the Analyze menu. This menu displays all the options available for your version of the SPSS program (the menus in this book were created with SPSS Statistics Version 25). Other versions may have slightly different sets of options.

To calculate a mean (average), we are asking the computer to summarize our dataset. Therefore, we run the command by clicking Analyze, then Descriptive Statistics, then Descriptives.

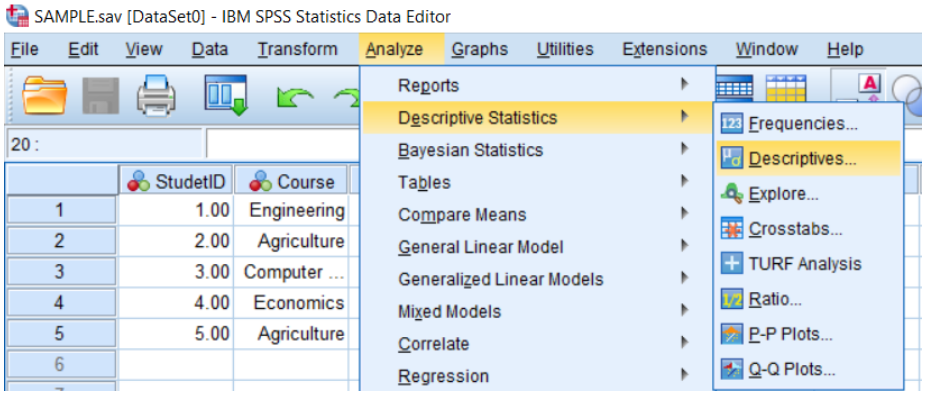


Figure 3.6: Analysis

This brings up the Descriptives dialog box. Note that the left side of the box contains a list of all the variables in our data file. On the right is an area labeled Variable(s), where we can specify the variables we would like to use in this particular analysis.

We want to compute the mean for the variable called GRADE. Thus, we need to select the variable name in the left window (by clicking on it). To transfer it to the right window, click on the right arrow between the two windows. The arrow always points to the window opposite the highlighted item and can be used to transfer selected variables in either direction. Note that double-clicking on the variable name will also transfer the variable to the opposite window. Standard Windows conventions of "Shift" clicking or "Ctrl" clicking to select multiple variables can be used as well. Note: Some configurations of SPSS show the variable names, and others show the variable labels (if any). This can be changed under Edit → Options → General.

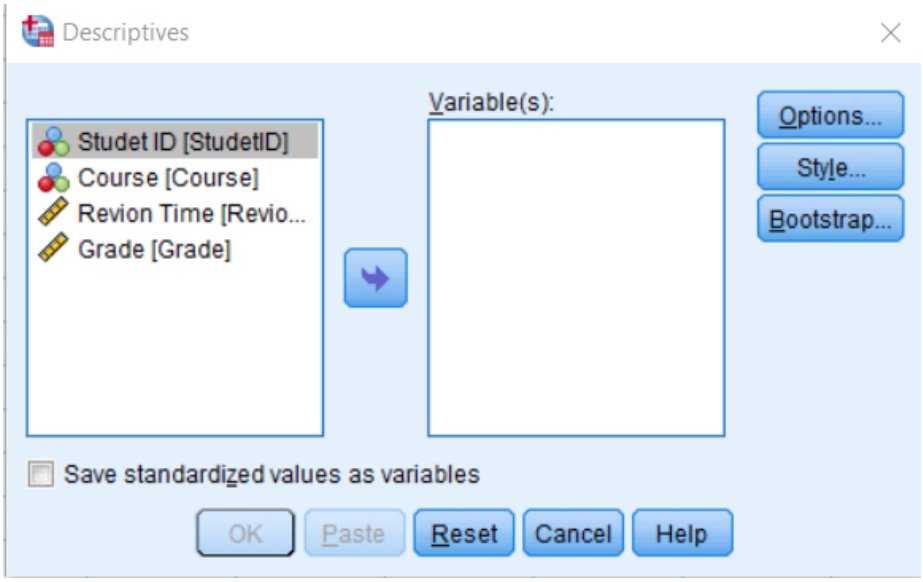


Figure 3.7: Analysis

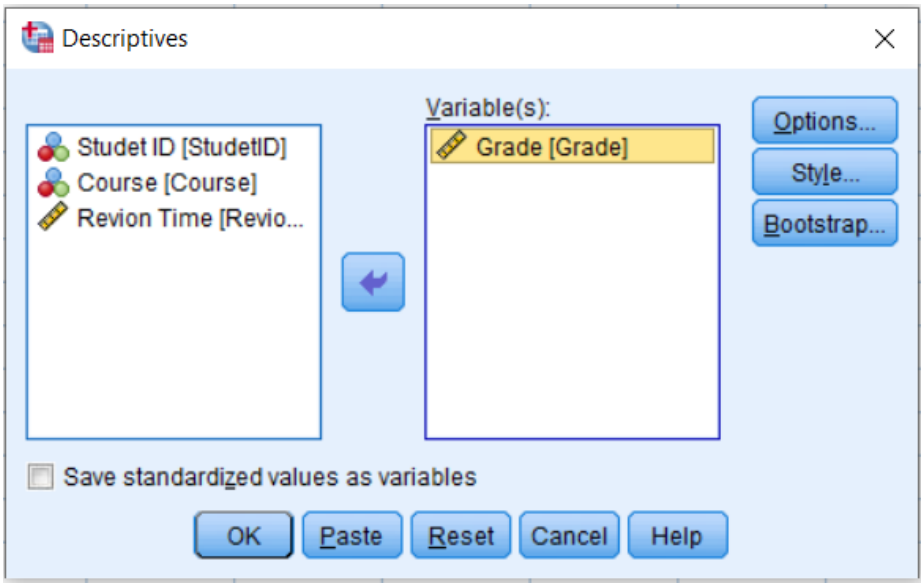


Figure 3.8: Analysis

When we click on the OK button, the analysis will be conducted, and we will be ready to examine our output.

3.6 Examining and Printing Output Files

After an analysis is performed, the output is placed in the output window, and the output window becomes the active window. If this is the first analysis you have conducted since starting SPSS, then a new output window will be created. If you have run previous analyses and saved them, your output is added to the end of your previous output.

To switch back and forth between the data window and the output window, select the desired window from the Window menu bar. Alternately, you can select the window using the taskbar at the bottom of the screen.

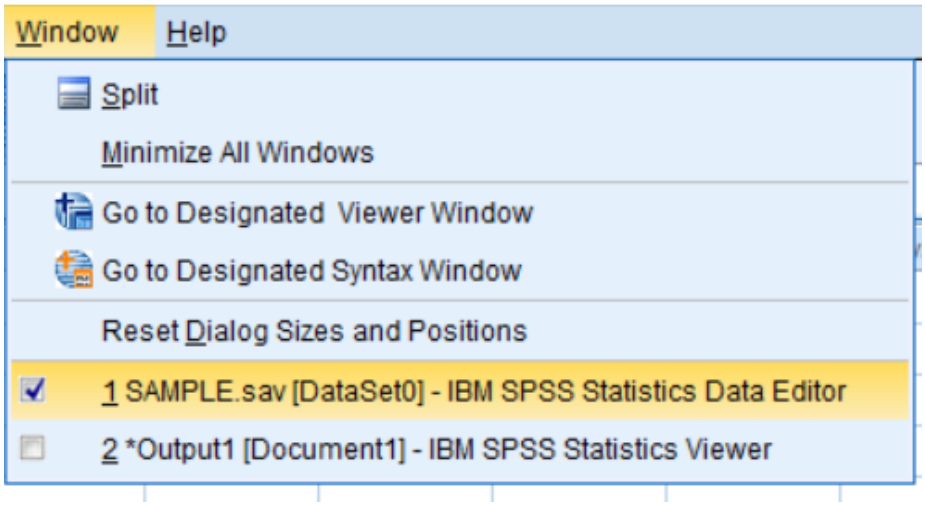


Figure 3.9: Analysis

The output window is split into two sections. The left section is an outline of the output (SPSS refers to this as the outline view). The right section is the output itself.

The section on the left of the output window provides an outline of the entire output window. All of the analyses are listed in the order in which they were conducted. Note that this outline can be used to quickly locate a section of the output. Simply click on the section you would like to see, and the right window will jump to the appropriate place.

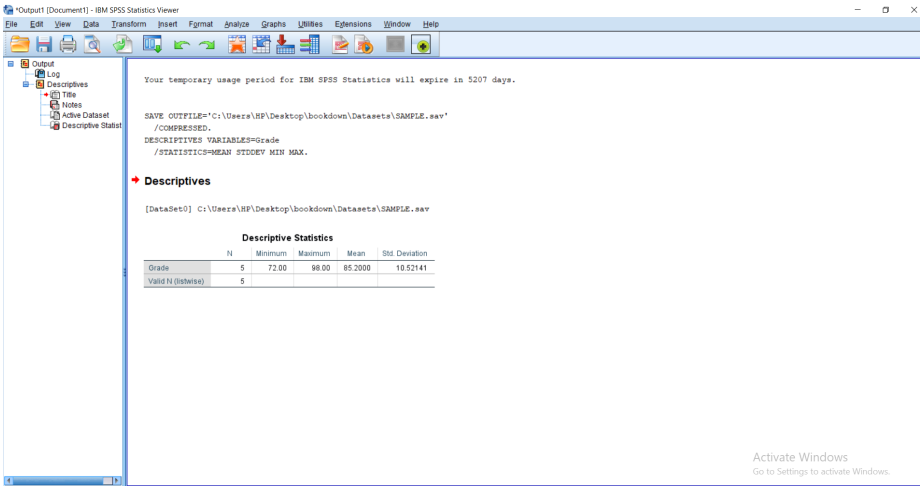


Figure 3.10: Analysis

Clicking on a statistical procedure also selects all of the output for that command. By pressing the Delete key, that output can be deleted from the output window. This is a quick way to be sure that the output window contains only the desired output. Output can also be selected and pasted into a word processor or spreadsheet by clicking Edit, then Copy to copy the output. You can then switch to your word processor and click Edit, then Paste.

To print your output, simply click File, then Print, or click on the printer icon on the toolbar. You will have the option of printing all of your output or just the currently selected section. Be careful when printing! Each time you run a command, the output is added to the end of your previous output. Thus, you could be printing a very large output file containing information you may not want or need.

One way to ensure that your output window contains only the results of the current command is to create a new output window just before running the command. To do this, click File, then New, then Output. All your subsequent commands will go into your new output window.

You can also save your output files as SPSS format files (.spv extension). Note that SPSS saves whatever window you have open. If you are on a data window you will save your data. If you are on an output window it will save your output.

Practice Exercise

Load the sample data file you created earlier (SAMPLE.sav). Run the Descriptives command for the variable GRADE, and print the output. Next, select the data window and print it.

3.7 Modifying Data Files

Once you have created a data file, it is really quite simple to add additional cases (rows/participants) or additional variables (columns).

StudetID	Course	RevionTime	Grade
1.00	Engineering	Morning	98.00
2.00	Agriculture	Evening	72.00
3.00	Computer ...	Morning	90.00
4.00	Economics	Evening	77.00
5.00	Agriculture	Morning	89.00
6.00	Engineering	Evening	68.00
7.00	Engineering	Evening	65.00

Figure 3.11: Data added

To add the data, simply place two additional rows in the Data View window (after loading your sample data). Notice that as new participants are added, the row numbers become bold. When done, the screen should look like the screenshot above.

New variables can also be added. For example, if the first three participants were given special training on time management, and the four new participants were not, the data file can be changed to reflect this additional information. The new variable could be called TRAINING (whether or not the participant received training), and it would be coded so that 0 = No and 1 = Yes. Thus, the first three participants would be assigned a "1" and the last four participants a "0." To do this, switch to the Variable View window, then add the TRAINING variable to the bottom of the list. Then switch back to the Data View window to update the data.

Adding data and variables are logical extensions of the procedures we used to originally create the data file. Save this new data file. We will be using it again later in this book.

Practice Exercise

Follow the previous example (where TRAINING is the new variable). Make the modifications to your SAMPLE.sav data file and save it. Good!

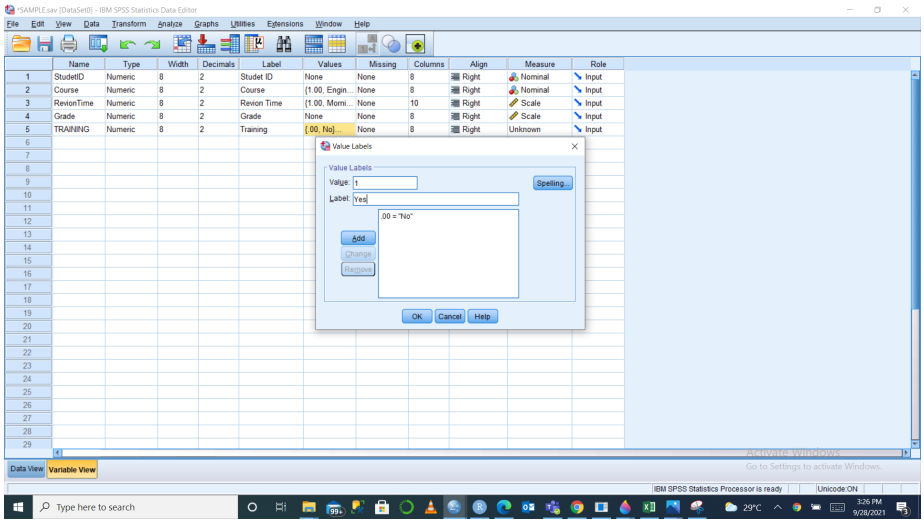


Figure 3.12: Variable added

Chapter 4

Entering and Modifying Data

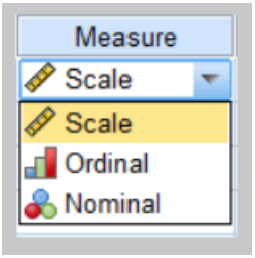
4.1 Variables and Data Representation

In SPSS, variables are represented as columns in the data file. Participants are represented as rows. Thus, if we collect four pieces of information from 100 participants, we will have a data file with four columns and 100 rows.

4.1.1 Measurement Scales

There are four types of measurement scales: nominal, ordinal, interval, and ratio. While the measurement scale will determine which statistical technique is appropriate for a given set of data, SPSS generally does not discriminate. Thus, we start this section with this warning: If you ask it to, SPSS may conduct an analysis that is not appropriate for your data.

Newer versions of SPSS allow you to indicate which types of data you have when you define your variable. You do this using the Measure column. You can indicate Scale, Ordinal, or Nominal (SPSS does not distinguish between interval and ratio scales).



Look at the SAMPLE.sav data file we created in Chapter

3. We calculated a mean for the variable GRADE. GRADE was measured on a ratio scale, and the mean is an acceptable summary statistic (assuming that the distribution is normal).

We could have had SPSS calculate a mean for the variable TIME instead of GRADE. If we did, we would get the output presented on the next page.

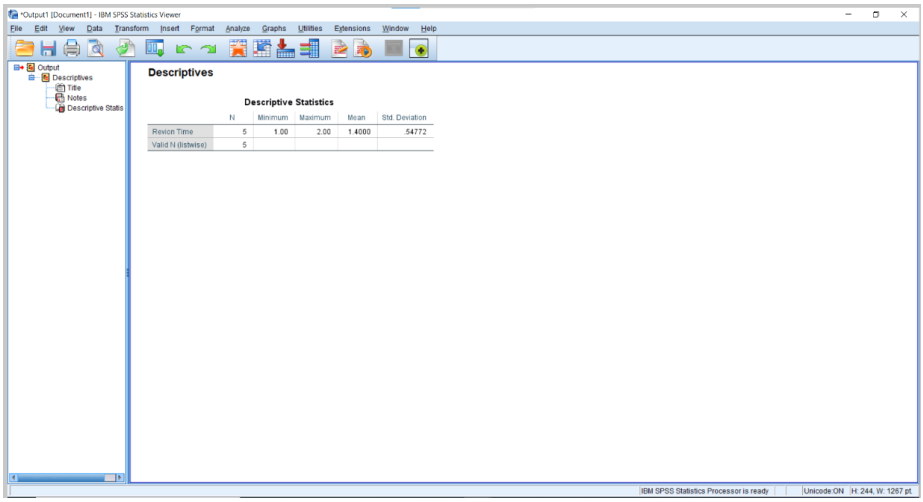


Figure 4.1: Measure.

The output indicates that the average TIME was 1.4. Remember that TIME was coded as an ordinal variable (1 = morning class, 2 = afternoon class). Though the mean is not an appropriate statistic for an ordinal scale, SPSS calculated it anyway. The importance of considering the type of data cannot be overemphasized. Just because SPSS will compute a statistic for you does not mean that you should use it. Later in the text, when specific statistical procedures are discussed, the conditions under which they are appropriate will be addressed. Please note that there are some procedures (e.g., graphs and nonparametric tests) where SPSS limits what you can do based on the measurement scale. However, more often than not, it is up to the user to make that decision.

4.1.2 Missing Data

Often, participants do not provide complete data. For example, for some students, you may have a pretest score but not a posttest score. Perhaps one student left one question blank on a survey, or perhaps she did not state her age. Missing data can weaken any analysis. Often, a single missing answer can eliminate a subject from all analyses.

	Q1	Q2	Total
1	3.00	3.00	6.00
2	2.00	1.00	3.00
3	1.00	2.00	3.00
4	2.00	.	.
5	4.00	1.00	5.00

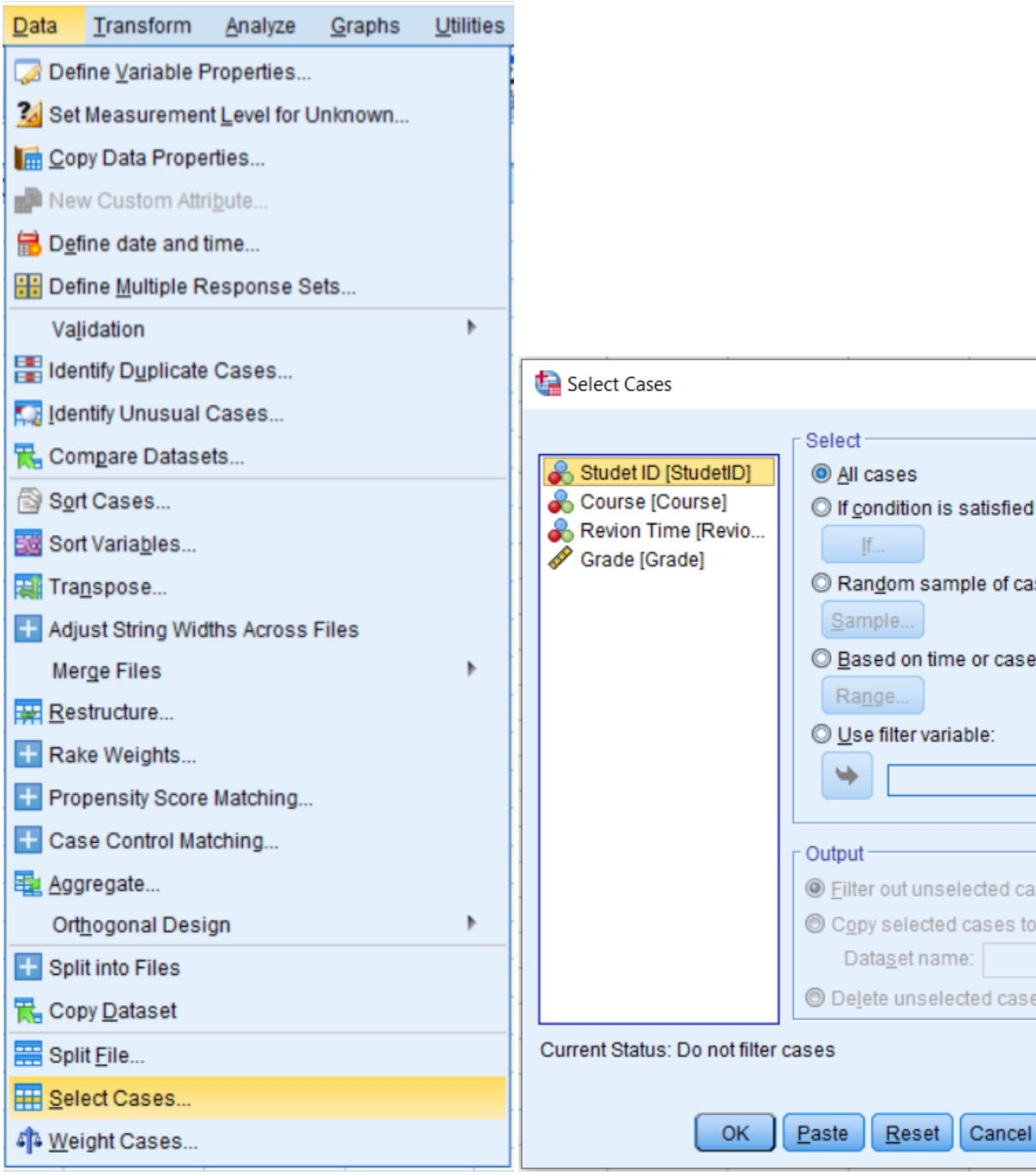
If you have missing data in your dataset, leave that cell blank. In the example shown above, the fourth subject did not complete Question 2 (q2). Note that the total score (which is calculated from both questions) is also blank because of the missing data for Question 2. SPSS represents missing data in the data window with a period (although you should not enter a period—just leave it blank). It is NOT good practice to create a filler value (e.g., "999" or "0") to represent blank scores, because SPSS will see it as a value with meaning, whereas it will treat truly blank values as missing.

4.2 Selection and Transformation of Data

We often have more data in a data file than we want to include in a specific analysis. For instance, our sample data file contains data from four participants, two of whom received special training and two of whom did not. If we wanted to conduct an analysis using only the two participants who did not receive the training, we would need to specify the appropriate subset.

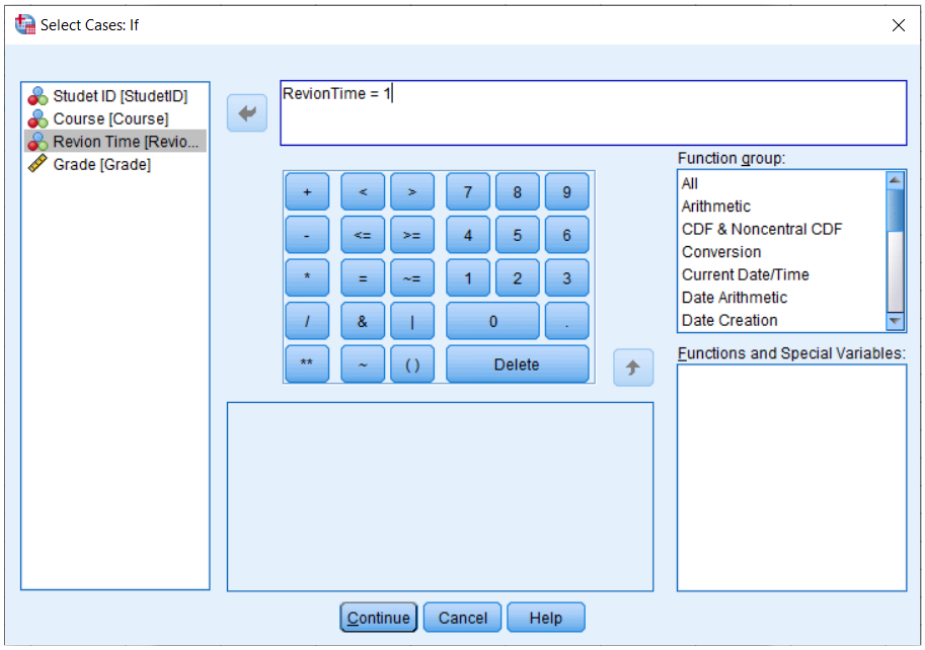
4.2.1 Selecting a Subset

We can use the Select Cases command to specify a subset of our data. The Select Cases command is located under the Data menu. When you select this command, the dialog box below will appear. (Note the icons next to the variable names that indicate that all variables were defined as being measured on a nominal scale except grade, which was defined as scale.)



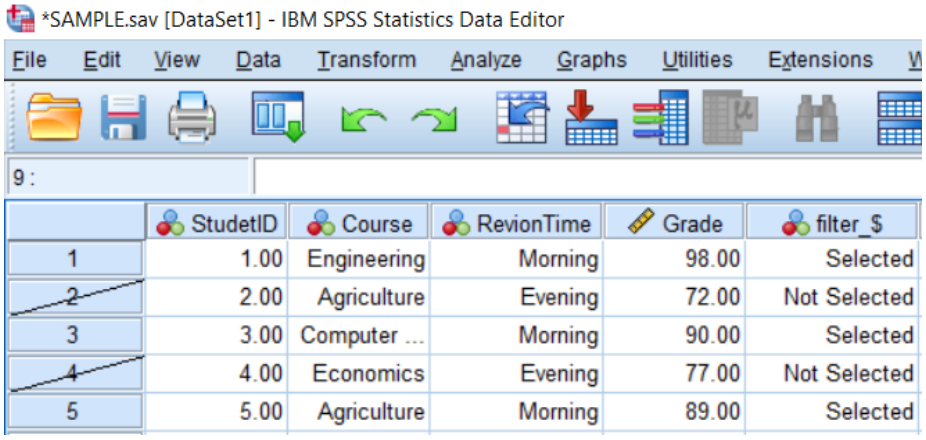
You can specify which cases (participants) you want to select by using the selection criteria, which appear on the right side of the Select Cases dialog box.

By default, All cases will be selected. The most common way to select a subset is to click If condition is satisfied, then click on the button labeled If. This will bring up a new dialog box that allows you to indicate which cases you would like to use.



You can enter the logic used to select the subset in the upper section. If the logical statement is true for a given case, then that case will be selected. If the logical statement is false that case will not be selected. For instance, you can select all cases that were coded as Morning/Evening by entering the formula RevisionTime = 1 in the upper-left part of the window. If RevisionTime is 1, then the statement will be true, and SPSS will select the case. If RevisionTime is anything other than 1, the statement will be false, and the case will not be selected. Once you have entered the logical statement, click Continue to return to the Select Cases dialog box. Then, click OK to return to the data window.

After you have selected the cases, the data window will slightly change. The cases that were not selected will be marked with a diagonal line through the case number. For instance, for our sample data, the second and fourth cases are not selected. Only the First and Third and Fifth cases are selected for this subset.



	StudetID	Course	RevionTime	Grade	filter_$
1	1.00	Engineering	Morning	98.00	Selected
2	2.00	Agriculture	Evening	72.00	Not Selected
3	3.00	Computer ...	Morning	90.00	Selected
4	4.00	Economics	Evening	77.00	Not Selected
5	5.00	Agriculture	Morning	89.00	Selected

An additional variable will also be created in your data file. The new variable is called FILTER_$ and indicates whether a case was selected or not.

If we calculate a mean GRADE using the subset we just selected, we will receive the output here. Notice that we now have a mean of 1.00 with a sample size (N) of 3 instead of 5.

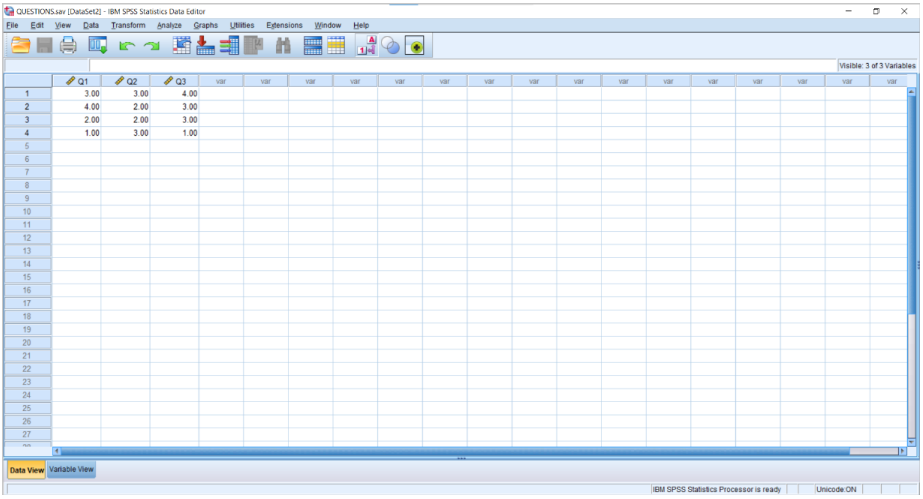
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RevionTime=1 (FILTER)	3	1	1	1.00	.000
Valid N (listwise)	3				

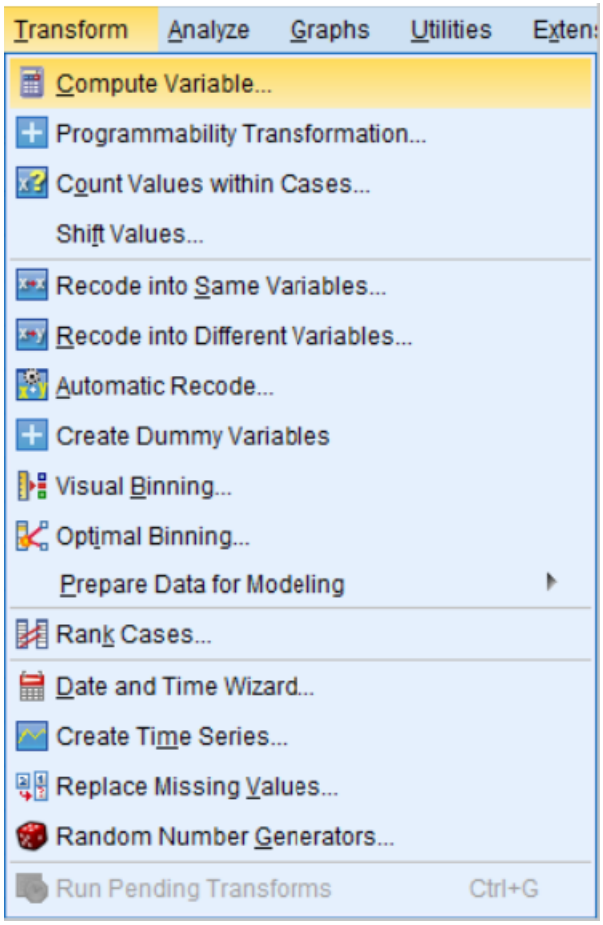
Be careful when you select subsets. The subset remains in effect until you run the command again and select all cases. You can tell if you have a subset selected because the bottom of the data window will indicate that a filter is on. In addition, when you examine your output, N will be less than the total number of records in your dataset if a subset is selected. The diagonal lines through some cases will also be evident when a subset is selected. Be careful not to save your data file with a subset selected, as this can cause considerable confusion later.

4.2.2 Computing a New Variable

SPSS can also be used to compute a new variable or manipulate your existing variables. To illustrate this, we will create a new data file. This file will contain data for four participants and three variables (Q1, Q2, and Q3). The variables represent the number of points each participant received on three different questions. Now enter the data shown on the screen below. When done, save this data file as "QUESTIONS.sav." We will be using it again in later chapters.



Now you will calculate the total score for each subject. We could do this manually, but if the data file were large, or if there were a lot of questions, this would take a long time. It is more efficient (and more accurate) to have SPSS compute the totals for you. To do this, click Transform, and then click Compute Variable.



After clicking the Compute Variable command, we get the dialog box shown below.

The blank field marked Target Variable is where we enter the name of the new variable we want to create. In this example, we are creating a variable called TOTAL, so type the word total.

Notice that there is an equals sign between the Target Variable blank and the Numeric Expression blank. These two blank areas are the two sides of an equation that SPSS will calculate. For instance, total = Q1 + Q2 + Q3 is the equation that is entered in the sample presented here (screenshot shown above). Note that it is possible to create any equation here simply by using the number and operational keypad at the bottom of the dialog box. When we click OK, SPSS will create a new variable called TOTAL and make it equal to the sum of the three questions.

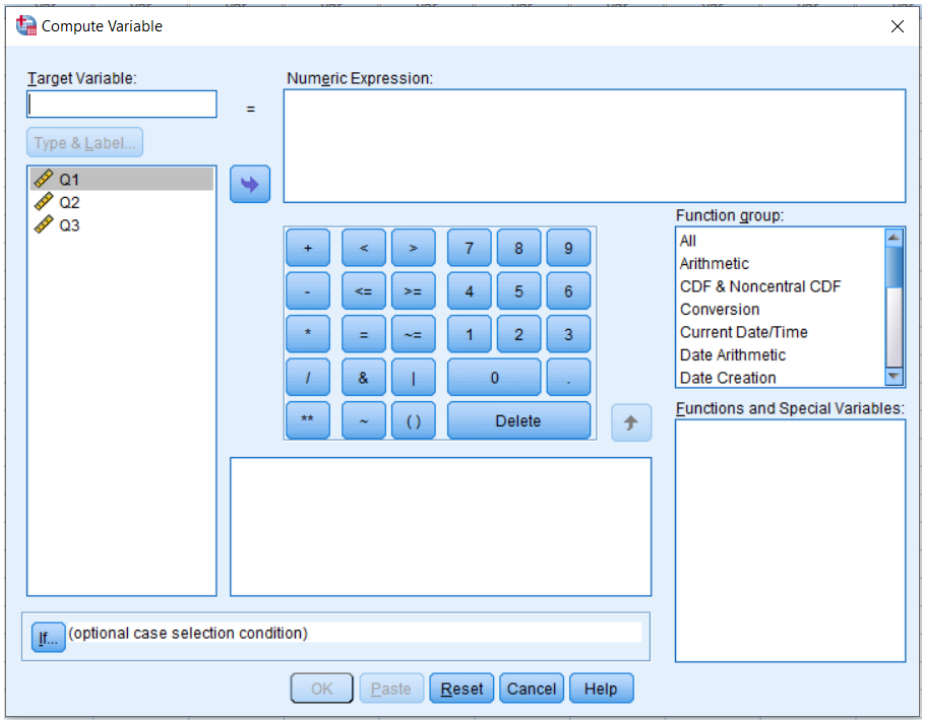
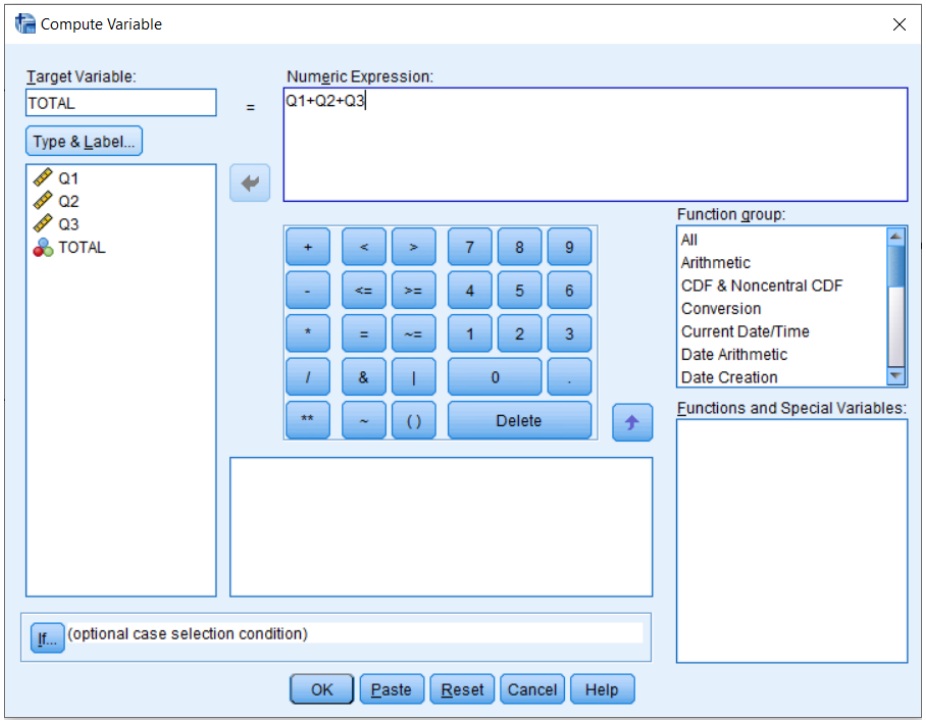


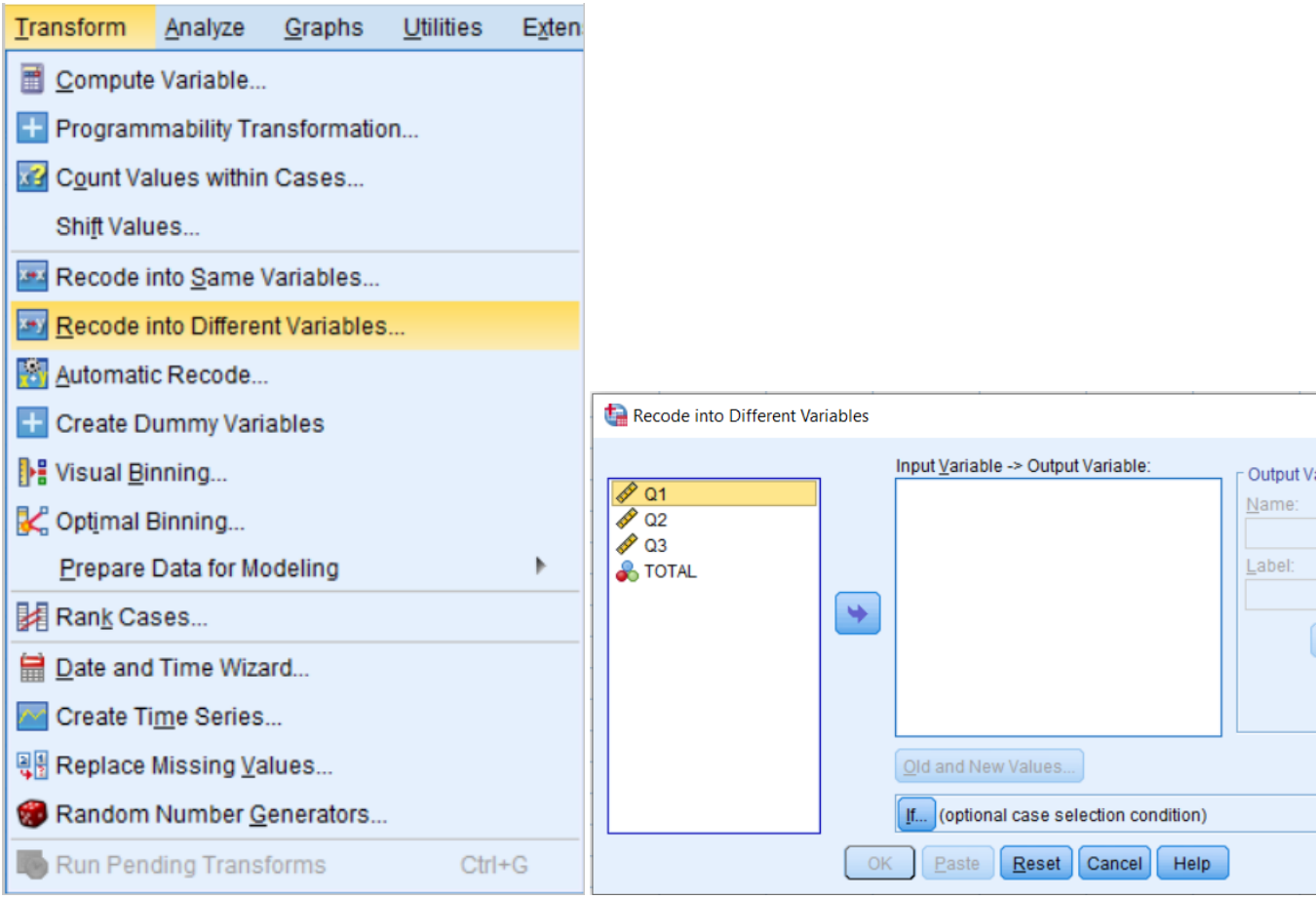
Figure 4.2: Tranformation



Save your data file again so that the new variable will be available for future sessions.

4.2.3 Recoding a Variable—Different Variable

SPSS can create a new variable based upon data from another variable. Say we want to split our participants on the basis of their total score. We want to create a variable called GROUP, which is coded 1 if the total score is low (less than or equal to 8) or 2 if the total score is high (9 or larger). To do this, we click Transform, then Recode into Different Variables.



This will bring up the Recode into Different Variables dialog box shown above. Transfer the variable TOTAL to the middle blank. Type group in the Name field under Output Variable. Click Change, and the middle blank will show that TOTAL is becoming GROUP, as shown below.

Click Old and New Values. This will bring up the Recode dialog box below.

In the example shown here, we have entered a 9 in the Range, value through HIGHEST field, and a 2 in the Value field under New Value. When we click Add, the blank on the right displays the recoding formula. We next entered an 8 on the left in the Range, LOWEST through value blank, and a 1 in the Value field under New Value. Click Add, then Continue.

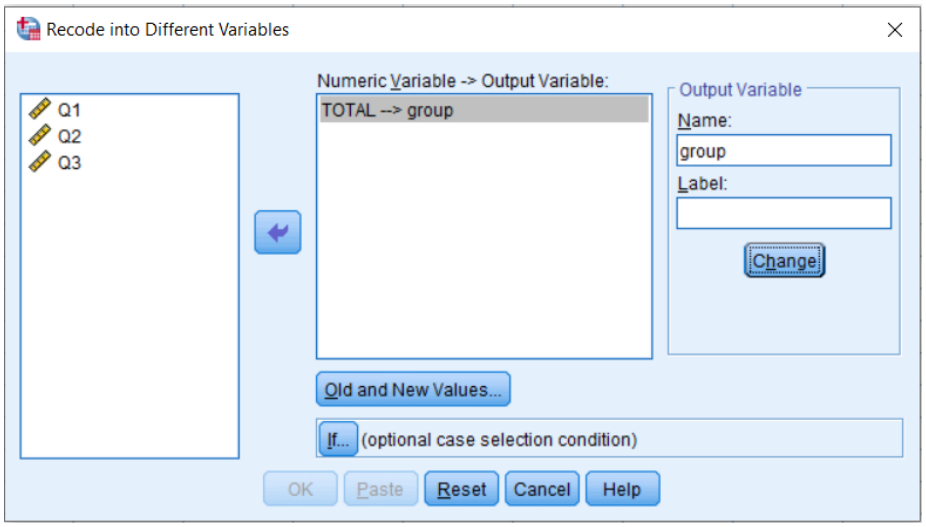
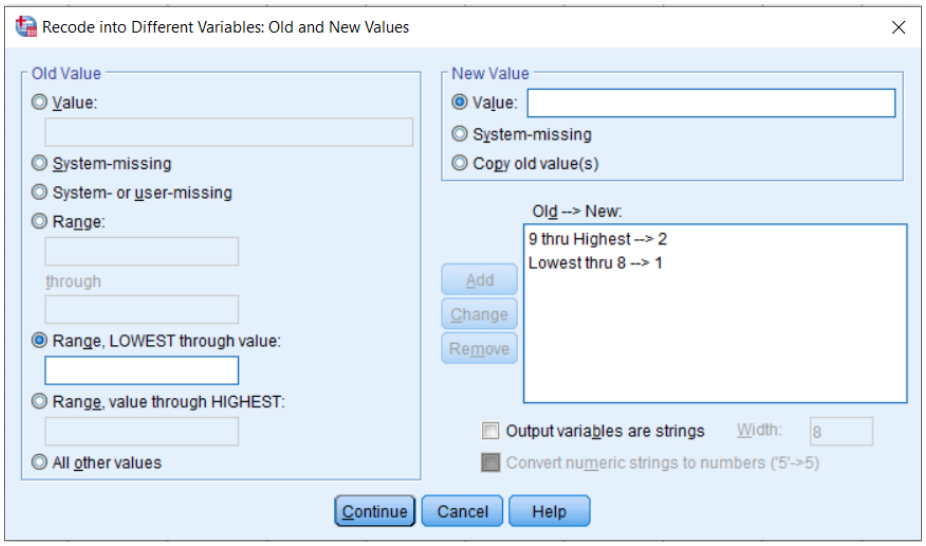


Figure 4.3: Recoding



Click OK. You will be redirected to the data window shown below. A new variable (GROUP) will have been added and coded as 1 or 2, based on TOTAL.

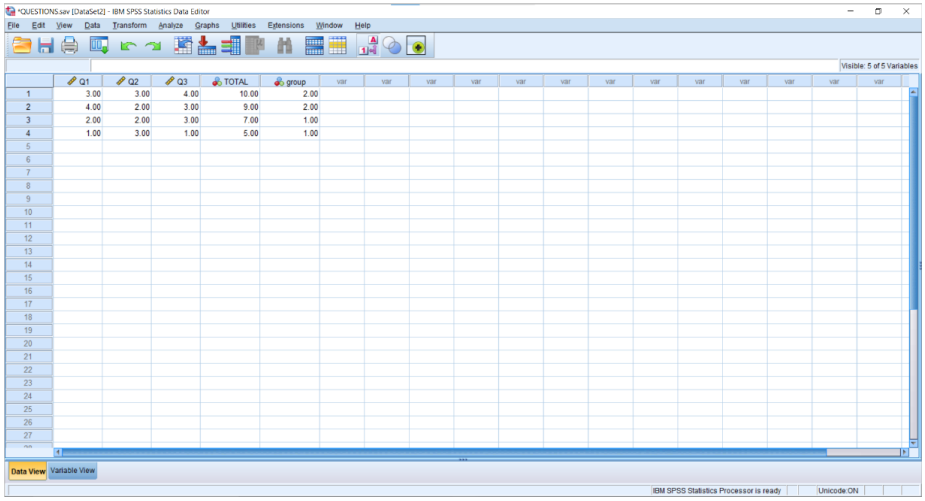


Figure 4.4: Recoded data

Chapter 5

Descriptive Statistics

5.1 Data

Chapter 6

Graphing Data

Chapter 7

Copyright

Visit the Author Gmail: hnjagimgmil.com