PC Lab 2 – Introduction to SPSS

**Introduction to SPSS**

The purpose of this tutorial is for you to familiarise yourself with the IBM SPSS software which you will use to complete your Assessment Tasks throughout the semester. This is your chance to become familiar with the IBM SPSS statistics software before you have to use it in assessments in future exercises so take full advantage of this opportunity.

**Some concepts that you see today have not yet been covered in lectures.**

**Please keep a copy of this document on hand so you can come back to it later in the semester.**

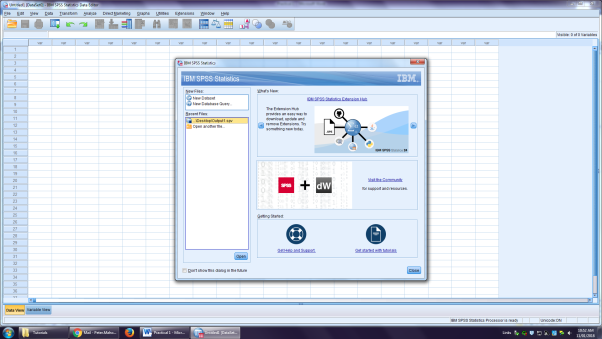
Work through these instructions to gain an understanding of setting up and running hypothesis tests in SPSS. This tutorial covers basic procedures and tests that you will use regularly in the unit.

**If you are struggling, ask your tutor for HELP. This is important!**

**Part 1: Entering data into SPSS**

There are several methods for inputting data into SPSS. However, in this unit we will concentrate on entering data directly, and copying/pasting/importing data from Microsoft Excel into SPSS. The correct arrangement of data depends on what you want to accomplish, and this will be discussed as each new type of hypothesis test is presented through the semester.

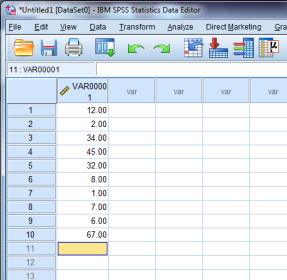
**Open IBM SPSS on your computer**. You initially see a database-like page with a pop-up window hovering above it in the centre of the screen (similar to Figure 1). You can simply click **CLOSE** the pop-up window to get rid of it.



**Figure 1.** The initial screen of IBM SPSS

Once the pop-up dialog window is closed, you will be left with a screen that looks very much like an Excel worksheet. Although similar to Excel in some ways, this screen has some key differences.

**Select the first cell in the first column** and enter a number, any number you like. Use the arrow key to move down a cell and enter another number. Repeat this process until you have ten cells with numbers entered (Figure 2).



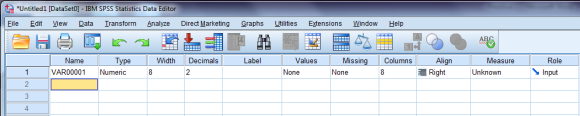
**Figure 2.** Data entered into SPSS cells

**Part 2: Data view and variable view**

Notice that the numbers are displayed to two decimal places and that the column heading is automatically set to VAR00001 (for variable 1). Before you can do very much with your data, you need to tell SPSS what type of data it is dealing with.

To achieve this, look at the bottom-left of the screen where there are two tabs: one tab labelled “Data View” shows you your raw data; and the second tab labelled “Variable View” shows meta-data about your raw data.

Click on the **VARIABLE VIEW** tab and you will see a summary of different ways that you can describe your data to SPSS (Figure 3).



**Figure 3.** SPSS Variable View showing a summary of your data

There are several columns in “Variable View”, each with information about your data. You do not need to worry about all of these columns at this stage, but several are quite important.

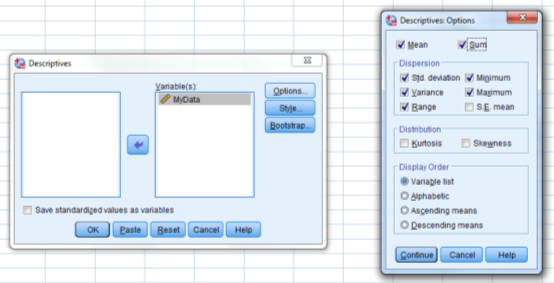
* **Name:** Here you can enter a name for your variable. We strongly recommend that you name variables with sensible descriptions so that you can keep track of them (in future weeks, you may have many variables on the screen at any one time). One caveat here is that SPSS will not allow spaces or unusual characters in variable names. **Try naming your variable.**
* **Type:** If you click on the type descriptor for your variable, a dialogue box appears with many options. Choose the option that is most appropriate for your variable. In this instance, keep the type as “Numeric” because our data are numbers. If your data consisted of words, you would select the option labelled “String”.
* **Width:** This option sets the maximum number of characters allowed in any cell for the variable. By default, it is set at 8 but you can change this as necessary. If you do not want anyone to be able to enter a 6 digit number into the data cells for a variable, you could set this field to “5”. Similarly, if your data are strings of characters, you can use this field to limit the maximum number of characters in the strings.
* **Decimals:** This mandates the number of decimal places to which numbers are displayed. By default, this is set to 2 (notice that the data you entered are all displayed at 2 decimal places) but you can easily change this. **Change your data to show no decimal places.**
* **Align:** controls the alignment of data in cells – this should be pretty familiar to anyone who has used Excel.
* **Measure:** This is, perhaps, the most important field of all. This tells SPSS what **scale of measurement** is to be used with your data (see Lecture 2 for a refresher on this). The options in SPSS are **nominal, ordinal and scale**. The first two should be familiar to you. The third option is used with data that are measured on interval or ratio scales (SPSS treats these both the same way). If you do not set this correctly, your hypothesis tests may not work properly. **Set the scale of measurement to ‘Scale’.**

**Part 3: Descriptive statistics**

One obvious use of SPSS is to calculate descriptive statistics about your data.

Switch back to Data View in SPSS and, from the menu at the top of the screen, select **ANALYZE → DESCRIPTIVE STATISTICS → DESCRIPTIVES (You can also choose FREQUENCIES to do a frequency analysis)**. This will launch a pop-up window. In the pop-up window, you should see your variable highlighted on the left.

Use the arrow to move your variable to the “Variable(s)” column and then click on **OPTIONS**. In the Options pop-up, select all of the descriptive statistics options shown in Figure 4 and then click **CONTINUE**. Finally click **OK**.



**Figure 4.** Descriptive statistics options

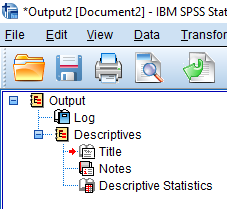
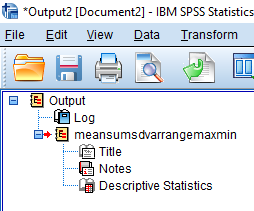
SPSS automatically calculates all chosen descriptive statistics and presents the output in a new window. SPSS presents all output in a window separate to the data window. This means that you need to change windows to get back to your data. During the semester, you will be doing some descriptive statistics by hand and some using Excel. SPSS does all of this work for you…so you should learn to appreciate it!

Before presenting the output table, SPSS generates some code such as below:

DESCRIPTIVES VARIABLES=testvariable

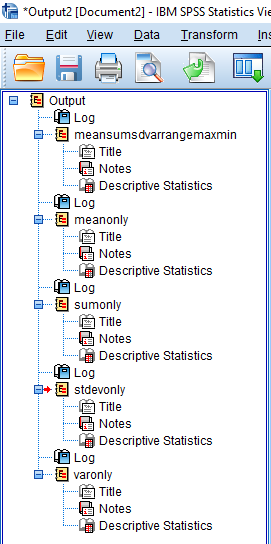
/STATISTICS=MEAN SUM STDDEV VARIANCE RANGE MIN MAX.

This is known as syntax and is useful for pinpointing specific tests that you have ran. Over the course of this unit, you will be conducting many statistical tests that will show up on the same output window. A quick way to find a test is to use the navigation menu and renaming “Descriptives” to another name of your choice. An example of this is shown below:

 to 

**Figure 5.** Navigation menu to pinpoint specific tests. Here “Descriptives” has been renamed to “meansumsdvarrangemaxmin” to indicate what parameters were selected.

When you click on the  icon, it will highlight the output of the test. Give this a try yourself, run the descriptives test again another four times but this time with mean only, sum only, std. deviation only and variance only respectively. Your navigation menu should look like Figure 6 below after appropriate renaming:



**Figure 6.** Navigation menu including all five tests.

Save your SPSS output file by selecting **File 🡪 Save As …**

In the dialogue box, navigate to your **ACU OneDrive** (labelled with your Student ID) or **USB storage device** and save a copy for future use. You may wish to save a second copy to your own personal device or email it to yourself. The file name should be in the form:

*Lab\_2\_StudentID*\_*FirstName*\_*LastName*.spv

Now try copying and pasting your first test (meansumsdvarrangemaxmin) data below. You can do this by right clicking the output table that SPSS has generated and selecting “copy”. Then right click in the space below and choose under “paste options”, “keep source formatting (K)”. Note if you go back to SPSS and right click on the table again and select “Copy as” and “image”, you can copy over your SPSS output as an image file.

Paste your descriptive statistics SPSS output in the space below:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

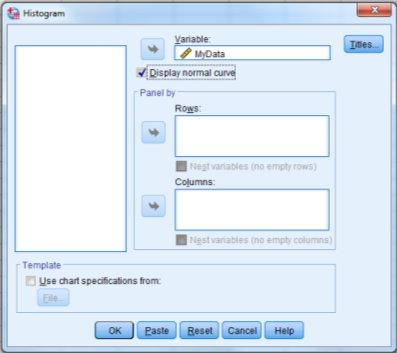
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Part 4: Graphing data in SPSS**

SPSS is capable of creating a range of graphs. These are generally not presented in an attractive fashion and may need editing. However, SPSS is highly suited to creating many of the styles of graph used in this unit, in particular histograms and boxplots, which we may use in the coming weeks.

*Histograms:*

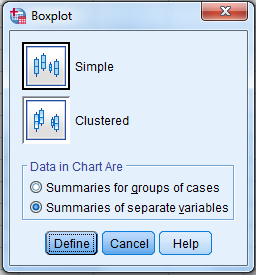
From the menu at the top of the screen, select **GRAPHS → LEGACY DIALOGS → HISTOGRAM**. In the pop-up window that launches, move your variable into the “Variable” box at the top-right. If you select the check-box labelled “Display normal curve”, this will superimpose a normal distribution curve over your histogram to act as a reference guide (Figure 7). **Create a histogram of the data for your variable by following the above steps and clicking OK.**



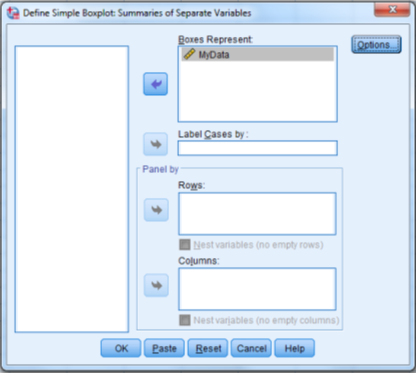
**Figure 7.** Histogram set-up dialog

*Boxplots:*

Select **GRAPHS → LEGACY DIALOGS → BOXPLOT** and a pop-up window will launch. Select “Simple” and “Summaries of separate variables”(Figure 8). Click **DEFINE**. Your variable will be highlighted on the left of a new pop-up window. Move your variable to the “Boxes represent” area on the right and then click **OK to create a boxplot of your data** (Figure 9). Box plots show much information about your data. They summarise the median, inter-quartile range, limits and any outliers (Figure 10).



**Figure 8.** Choosing boxplot design parameters in SPSS Legacy Dialogs



**Figure 9.** Defining your box plot

A diagram of a chart

Description automatically generated

**Inter-quartile range**

**Median**

**Range of data (excluding outliers)**

**Outlier**

**Figure 10.** Boxplot example

Note that the boxplot also indicates how the data are distributed. If the data are symmetrically distributed then the median will be in the centre of the range. In the example in Figure 8, the median is towards the bottom end of the box, indicating that the data are right-skewed. **If your boxplot does not show any outliers, try adding another data point to your variable that is very far from the existing data points and then creating a new boxplot.**

Look back at your left-hand side menu. You should now see two new entries for the histogram and boxplot that you have created. Again, this is a useful way to pinpoint a specific test or graph you have created in your output window.

**Part 5: Basic hypothesis testing in SPSS**

We are going to look at two common hypothesis tests in SPSS - the one-sample *t*-test and the Kolmogorov-Smirnov test. **T-tests and the Kolmogorov-Smirnov test will be covered in more detail later in the semester so don’t worry if you don’t understand what they mean now...we are just using them as examples.**

*One-sample t-test:*

A one-sample *t*-test assesses the null hypothesis (H0) that the sample (our Test Variable) mean IS NOT different to a population with a specific mean. SPSS produces an output table that shows the calculated test statistic (*t*), degrees of freedom (df) and *P*-value (Sig.) for this test (Table 1).  **You will learn these terms as we progress through the semester. Here are brief descriptions:**

**Test statistic** –a calculated variable that helps us determine if we should accept or reject the null hypothesis (H0) based on the data that we have. It is used to calculate the *P*-value.

**Degrees of freedom** – Often a tricky concept to grasp and you do not have to understand it to perform a statistical analysis but this website might give you a bit more of an idea: <http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-degrees-of-freedom-in-statistics>

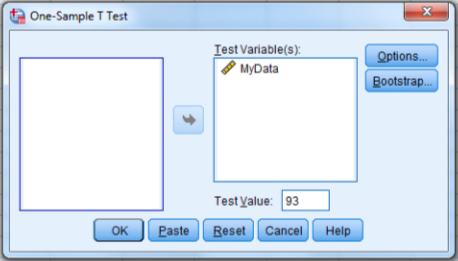
**P-value** – is the probability of getting an observed or extreme result, if the null hypothesis is true.

Your **alternate hypothesis (H1)** is that there is a difference between your sample (your test variable in SPSS) mean and a population which has a specific mean value of 93.

**Write the corresponding null hypothesis (H0):**

Test your null hypothesis using the **one-sample t-test** in SPSS as follows:

Select **ANALYZE → COMPARE MEANS → ONE-SAMPLE T Test** and a pop-up dialog window will launch. Move your variable into the “Test Variable(s)” area and enter the value you wish to test (93 for the above H1 and H0) against that distribution in the “Test Value” box, then click **OK** (Figure 11).



**Figure 11.** One-sample *t*-test setup dialog box

**Table 1.** SPSS one-sample *t*-test output

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **One-Sample Test** | | | | | | |
|  | Test Value = 93 | | | | | |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| MyData | -15.559 | 9 | .000 | -78.000 | -89.34 | -66.66 |

Note that SPSS labels P-values as “Sig.” and has shown the *P*-value as 0.000 – however, this does not mean that the probability of getting this result is actually zero. Some versions of SPSS are limited to three decimal places and a *P*-value of 0.000 really just means that the real value was too small to show in three decimal places. When reporting such values in your work, you should round this up and report that *P* < 0.001.

**Run one-sample t-tests on your data set with several different test values. Perhaps try the mean values from some of your fellow students to see if their samples are likely to come from the same population.** Record the corresponding alternate and null hypotheses.

*Kolmogorov-Smirnov test:*

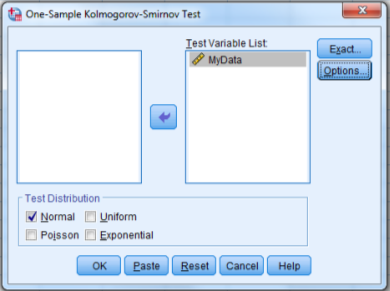
You have now run your first hypothesis test in SPSS. As we progress through the semester, you will learn that hypothesis tests have **assumptions.** These should be checked before using the test. You will use the Kolmogorov-Smirnov test throughout the semester.

One-sample *t*-tests have an assumption that the data are normally distributed. You did not check that assumption before conducting the *t*-test(s). Given the way you selected the data, they may very well NOT be normally distributed, in which case the results of the *t*-test are not reliable.

You need to test the data for normality. You could do this by looking at the histogram you created in Part 4, but SPSS has a better way. The Kolmogorov-Smirnov test (often just called a K-S test, for obvious reasons) tests data to see whether they are significantly different from a chosen type of distribution. We will use this test A LOT in this unit to assess whether data have a normal distribution.

Our null hypothesis for this type of test will, therefore, be: **the data are NOT significantly different to a normal distribution**.

Select **ANALYZE → NON PARAMETRIC TESTS → LEGACY DIALOGS → 1-SAMPLE K-S** to launch the set-up dialog for this test. Move your variable to the “Test Variable List” and select “Normal” as the type of distribution you want to test your data against, then click **OK** (Figure 12).



**Figure 12.** Kolmogorov-Smirnov set-up dialog box

SPSS provides an output table identifying the calculated test statistic and the *P-*value (labelled as asymp. Sig). You need to decide whether or not to reject the null hypothesis of the test based on the *P*-value. Do not be misled by the notation at the bottom that says “Test distribution is normal” – this is NOT saying that your data are normal, it is a reminder that you tested the data against a normal distribution and not one of the other options in the set-up dialog (Table 2).

**Table 2.** SPSS Kolmogorov-Smirnov test output

|  |  |  |
| --- | --- | --- |
| **One-Sample Kolmogorov-Smirnov Test** | | |
|  | | MyData |
| N | | 10 |
| Normal Parametersa,b | Mean | 15.00 |
| Std. Deviation | 15.853 |
| Most Extreme Differences | Absolute | .275 |
| Positive | .275 |
| Negative | -.189 |
| Test Statistic | | .275 |
| Asymp. Sig. (2-tailed) | | .031c |
| a. Test distribution is Normal. | | |
| b. Calculated from data. | | |
| c. Lilliefors Significance Correction. | | |

If the *P*-value for your test is below 0.05 then you would reject your null hypothesis and conclude that your data ARE significantly different to a normal distribution (i.e. not normal). In that situation the results of your earlier *t*-test would be unreliable.

**Part 6: Transforming data in SPSS**

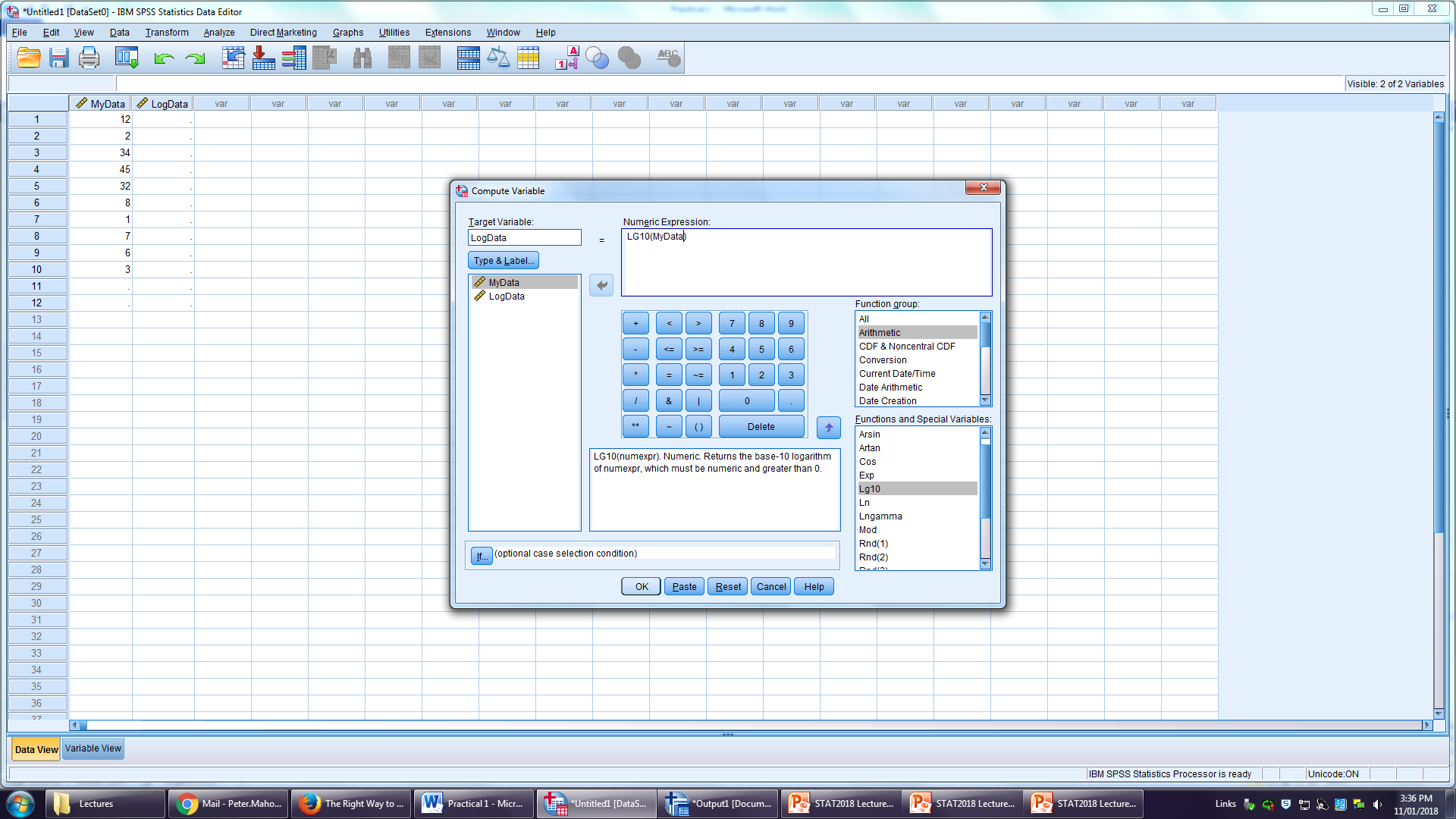
If you needed to conduct that one-sample *t*-test, but your data were not normal, you would need to try **transforming** your data. Data transformations are important and today you will learn the process of carrying out calculations for transformations. SPSS has built-in procedures to make data transformations quick and easy.

SPSS will insert the transformed data into any column that you designate. To that end, it is best to create a new column for the transformed data before starting. **Switch to Variable View** and create a new variable name for your transformed data, then set all of the rest of the properties of the variable appropriately.

Switch back to **Data View** and select **TRANSFORM → COMPUTE VARIABLE** to launch the data transformation dialog window. In the “Target Variable” box at the top-right, type the name you chose to call your transformed variable.

In the “Function Group” menu, select “Arithmetic” and then double-click **Lg10** from the “Functions and Special Variables” menu to choose a log 10 transformation.

In the “Type and Label” menu, highlight the variable you wish to transform and use the arrow button to move it into the “Numeric Expression” box, then simply click **OK** (Figure 13).



**Figure 13.** Data transformation set-up dialog box

SPSS will automatically take you to the output page but, for once, there is nothing interesting to see there. If you return to the data page you will see that the new column you set up has now been populated with transformed data.

Now **run a Kolmogorov-Smirnov test on the transformed data** to check whether they are normally distributed and, if they are, run your one-sample *t*-test again – this time with reliable results.

If your transformed data are not normally distributed, you can try other types of transformation. Simply follow the steps used to transform data above and select different transformation types from the “Functions and Special Variables” menu as desired…

* Ln will produce a Log base e (natural log) transformation
* SQRT will produce a square root transformation
* Arsin will produce an arcsine transformation

Just create a new column with an appropriate name for each transformation and be very sure to transform only the original data, not some already transformed data.

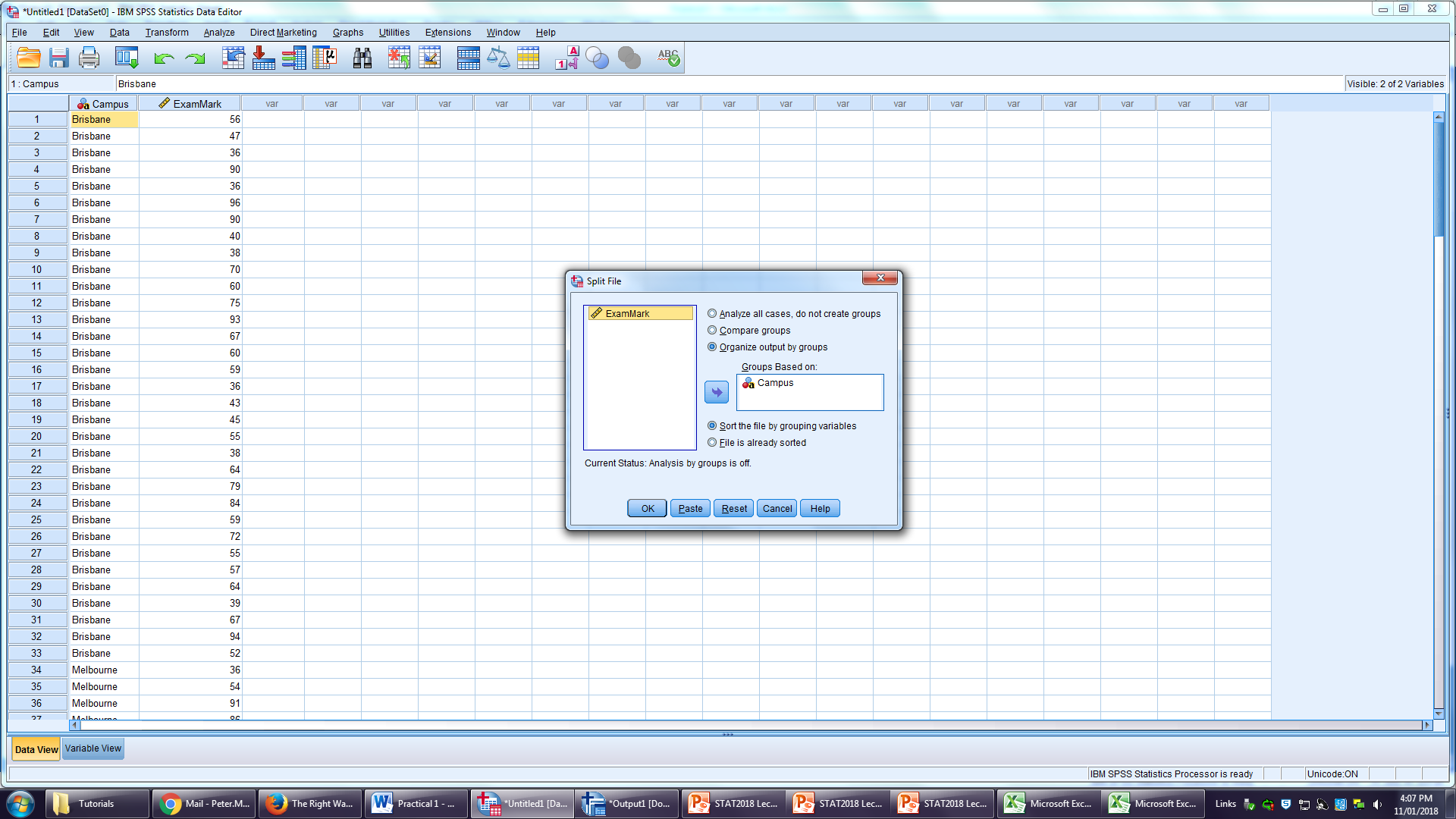
**Part 7: A couple of extra tricks to save you time and effort**

As mentioned at the beginning of this document. You can cut and paste data directly from Excel into SPSS.

Open the Excel file in LEO called **<Exam Marks>** and have a go at copying/importing the data directly into SPSS. You will need to do this many times in the coming weeks so get used to doing it now. You will still need to organize the variable view in SPSS before you can work with the data.

You will notice that the data consists of two variables. One is a list of the ACU campuses attended by sampled students and the other is the exam mark that each student achieved in the statistics exam in 2018.

If you needed to calculate descriptive statistics for each group, you could copy parts of the data into different columns and then carry out three separate procedures in SPSS. However, there is a quicker way. Select **DATA → SPLIT FILE** and the Split File pop-up dialog box will launch. Select “Organize output by groups” and drag the “Campus” variable into the “Groups Based On” area, then click **OK** (Figure 14).



**Figure 14.** The Split File dialog box

Now try calculating descriptive statistics (as you did in Part 3 of this prac) and see what happens. Then try creating a histogram of the data (as you did in Part 4 of this prac) and see what happens.

This trick can save you a lot of time and effort through the semester but you have to be very careful to **unsplit** the file when you are done by opening the Split File dialog box, selecting “Analyze all cases, do not create groups” and clicking **OK**.

Another neat trick is being able to recode your variables. Have a look at how ACU range their grades below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Final Result Grade** | **Notation** | **Range Guide (%)** | **GPA Value** |
| High Distinction | HD | 85-100 | 7 |
| Distinction | DI | 75-84 | 6 |
| Credit | CR | 65-74 | 5 |
| Pass | PA | 50-64 | 4 |
| Fail | NN | 0-49 | 1 |
|  |  |  |  |

Can you recode the “Exam Marks” variable to determine how many students got each of the grades from all three campuses? Here’s a hint: look for an option in SPSS called “**Recode into Different Variables…**”. By using recode and split file together, you can even see the frequency of grades on each campus. Good luck!

Now you have seen the basics of SPSS. Keep this file handy in future classes as it will help you navigate your way through SPSS, minimizing the time you take to finish and maximizing your marks!

Getting familiar with SPSS

**Descriptive Statistics using Excel and SPSS**

Open the Excel file <**Patient Biometrics**>.

You have been given the biometric data of patients from seven different cities in Australia.

1. Using Excel, calculate the 95% CI for each of the height, forearm length, lower leg length and hand span variables.

Import the <**Patient Biometrics**> Excel file into SPSS.

2. Using SPSS, calculate the mean, median, mode, standard deviation, variance, range, minimum and maximum of the variable ‘Age’.

3. Using SPSS, can you calculate the 95% confidence interval (for the 4 continuous variables in Q1) for just Melbourne patients alone?

4. Create a frequency distribution for the variable ‘Sex at birth’. How many males and females are there?

5. Create a histogram for the variable ‘Age’. What shape distribution does it appear to have?

6. Create a boxplot for the variables ‘Height’, ‘Forearm length’, ‘Lower leg length’ and ‘Handspan’. Do you see any outliers in any of these variables? If so, which one(s)?

7. Create a scatterplot of the variables ‘Height’ and ‘Lower leg length’. Can you find the option to put a line of best fit and r-squared value on your graph? What does the r-squared value mean? What value did you get for it?

8. Using the Chart Builder in SPSS, create a bar chart of the variables ‘Height’ and ‘Australian City’. Which variable should be on the x-axis? How does a bar chart differ from a histogram? Is there one city that has an average height shorter than the other cities? If so, which one? Can you determine this value using what you’ve learnt already in SPSS?

**Checking for Normality using SPSS**

There are a number of ways to test whether a sample of data follows a normal distribution. One such way is to use a Kolmogorov-Smirnov test (K-S test).

1. Write down the null and alternate hypotheses for a K-S test.

2. Explain with reference to the p-value and your answers to Q1 above, how you would determine if a dataset is normal.

3. Run a K-S test on each of the four continuous variables above and determine if any of them are normal. Then split the file according to their Australian cities and run the K-S test again. Do any of the continuous variables follow a normal distribution in individual Australian cities?

Two other ways to check for normality is to use the Shapiro-Wilk’s test (S-W test) or a QQ plot. Both of these can be done together using SPSS. It is important to note that the Kolmogorov-Smirnov and Shapiro-Wilk’s test are only useful for a sample size that is less than 30. Sample sizes larger than this typically use the QQ plot to check for normality.

Can you find the function in SPSS that gives you both the SW test and QQ plot? HINT: Look for “Explore” under the ‘Analyze’ menu. Check the normality of your four continuous variables again. Paste your QQ plots below. How do you know if they are normal or not?

**SPSS Value Labels Exercise**

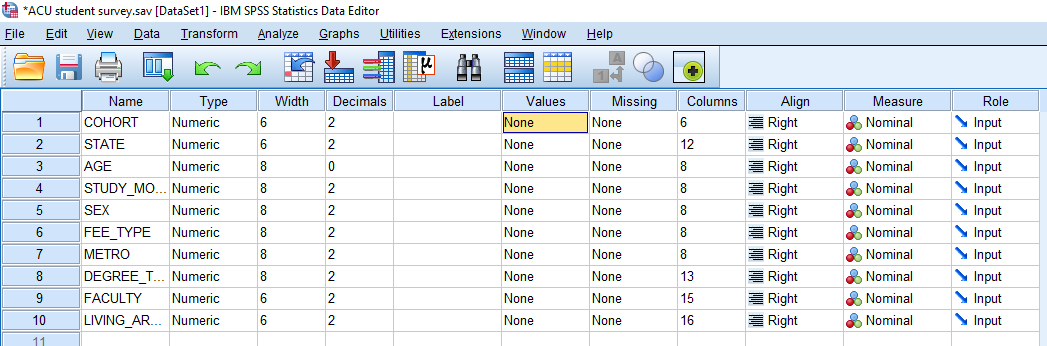
Open the SPSS file <**ACU student survey**>. The ACU student survey is a fictional dataset of students enrolling into their courses from 2005 to 2012. There are a number of demographic variables in which the students were asked to complete. This dataset is incomplete because it hasn’t been properly labelled. Run a frequency analysis (descriptive statistics) on all the variables to see this.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **cohort** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 1.00 | 3259 | 8.4 | 8.4 | 8.4 |
| 2.00 | 3615 | 9.3 | 9.3 | 17.8 |
| 3.00 | 3944 | 10.2 | 10.2 | 28.0 |
| 4.00 | 4086 | 10.6 | 10.6 | 38.5 |
| 5.00 | 5010 | 13.0 | 13.0 | 51.5 |
| 6.00 | 5687 | 14.7 | 14.7 | 66.2 |
| 7.00 | 6383 | 16.5 | 16.5 | 82.7 |
| 8.00 | 6697 | 17.3 | 17.3 | 100.0 |
| Total | 38681 | 100.0 | 100.0 |  |

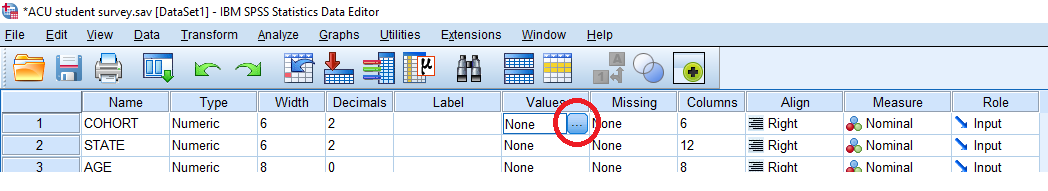
See how the numbers 1.00 to 8.00 show up in the first column? This doesn’t make a lot of sense to the reader and needs to be properly labelled. There are two ways to do this:

The first is to just simply edit the table in the SPSS output file (.spv). This can be done by right-clicking the table and choosing “Edit Content” 🡪 “In Viewer”. Now if you double click the “1.00” you can change it to “2005”. You can subsequently do this for all tables you present in SPSS.

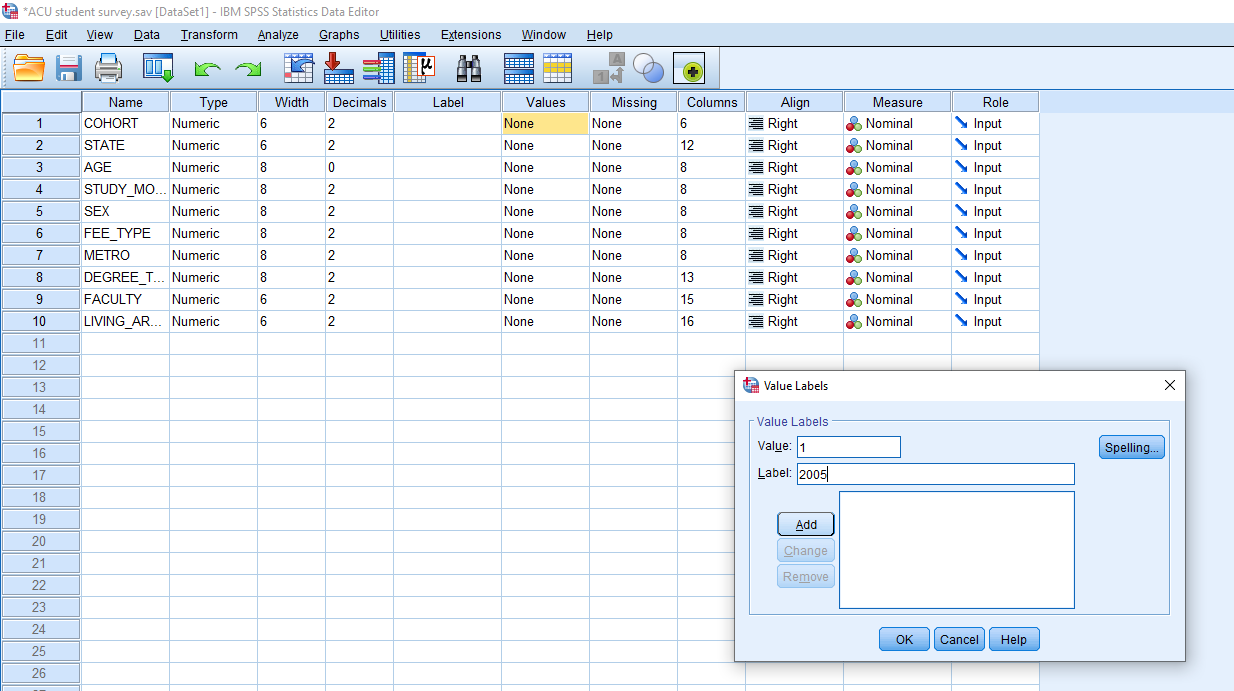
The second approach requires you to go into the Variable View, which currently looks like this:



For each variable you select in the “Values” column, you will see a box with “…” as shown below:



If you click this box, it will open up a “Value Labels” box where you can enter in your labels. Use the information below to enter in all the labels for each of the demographic variables. An example is given below…don’t forget to click “Add”. Note: you do not need to do a label for “AGE”.



**COHORT**

1 = 2005

2 = 2006

3 = 2007

4 = 2008

5 = 2009

6 = 2010

7 = 2011

8 = 2012

**STATE**

1 = NSW

2 = Victoria

3 = Queensland

4 = ACT

**STUDY\_MODE**

1 = FT

2 = PT

**SEX**

1 = Male

2 = Female

**FEE\_TYPE**

1 = Domestic

2 = International

**METRO**

1 = Metro

2 = Non-metro

**DEGREE\_TYPE**

1 = Single

2 = Double

**FACULTY**

1 = Arts and Sciences

2 = Education

3 = Health Sciences

4 = Theology and Philosophy

5 = Business

**LIVING\_ARRANGEMENT**

1 = At home

2 = College/student accom

3 = Independently