Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

SPSS I: Using SPSS for Basic Statistics

As noticed many times before in the theoretical discussions, statistics can become more and more difficult to perform by hand given that we often want to run analyses on very large datasets. The use of computers is common in the field of statistics. To get a general idea of what software is available, understand that there are two major types of statistical software or statistical packages. The first type are the *open source* packages, which are free to use and are often developed by a community of users. The second type are the *commercial* packages, which are developed by companies and are often expensive to use.

What are some open source packages? The most popular open source package is R. R is a programming language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. R

is the most specialized and powerful statistical software, which is why it is the most popular among *pure* statisticians. Increasingly, R is seeing uses in other fields, such as economics, finance, natural sciences, and other social sciences. However, R is not the most user-friendly software, and has a relatively steep learning curve. Among other open source packages are Python and Julia.

IBM SPSS Statistics is a commercial software package that is used for statistical analysis. It is available for Windows, Mac, and Linux, and stands for Statistical Packages for the Social Sciences.

SPSS is a very user-friendly software package that is used by many researchers. Among the other commercial packages are Stata and SAS. It is common to see that commercial software is used in disciplines which are not too statistics-heavy, such as psychology, sociology, among others. However, it is important to remember that while commercial software is often very easy to learn, it suffers from heavy costs and is often not as powerful as open source software.

In here, we will cover the basics of using SPSS for basic statistics, which include a very brief introduction to how to initialize the software, loading data and running analyses. We will also cover how to save results. There are tons of resources online to learn how to use SPSS, and this is just a very brief introduction. Please review SFU's SPSS tutorial for more information.

Overview

SPSS has several different versions. In these notes, I have been using SPSS 25, which is a relatively new version. Versions are often not too different from each other, so you should be able to follow along with any version of SPSS.

When you open SPSS, you will see a window that looks like this:

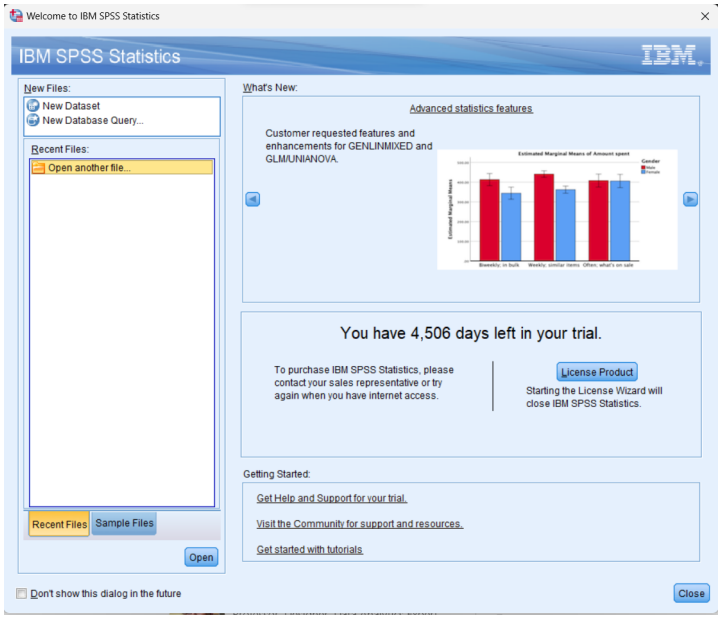


Figure 1: SPSS Welcome Screen

This screen allows you to open recently used datasets or SPSS output files, which contain the results from analysis you may have run. It also allows you to connect to sample or recent datasets, which is an important thing that we will do later

You can close the welcome window now. You will then notice that two other windows have been opened fro your. These are the program's two main windows: the *Data Editor* and the *Output Viewer*. The Data Editor is where you enter data and perform data manipulation. It has two tabs, one for Data View and one for Variable View. The Data View shows you all of the rows of your data, and the Variable View shows you all of the columns of your data. Remember that in statistics, we often refer to rows as *observations* and columns as *variables*. See an image below:

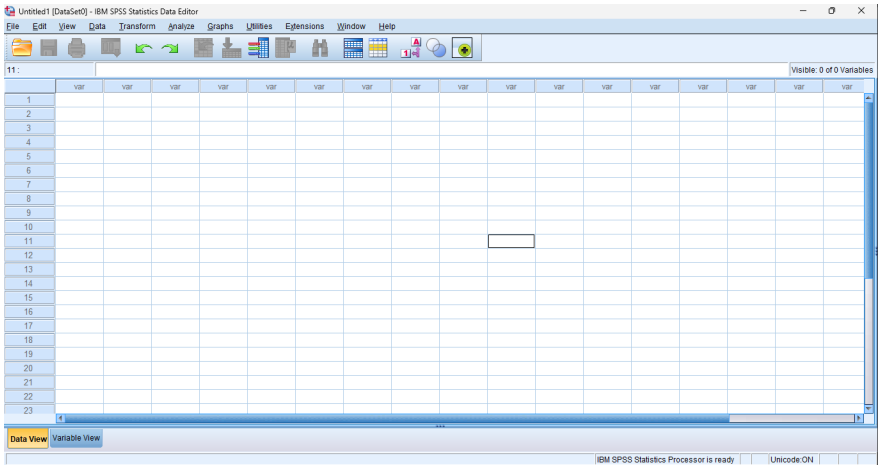


Figure 2: SPSS Data Editor

The Output or Statistics Viewer is where you see the results of your analyses, which can be seen in Figure 3 below.

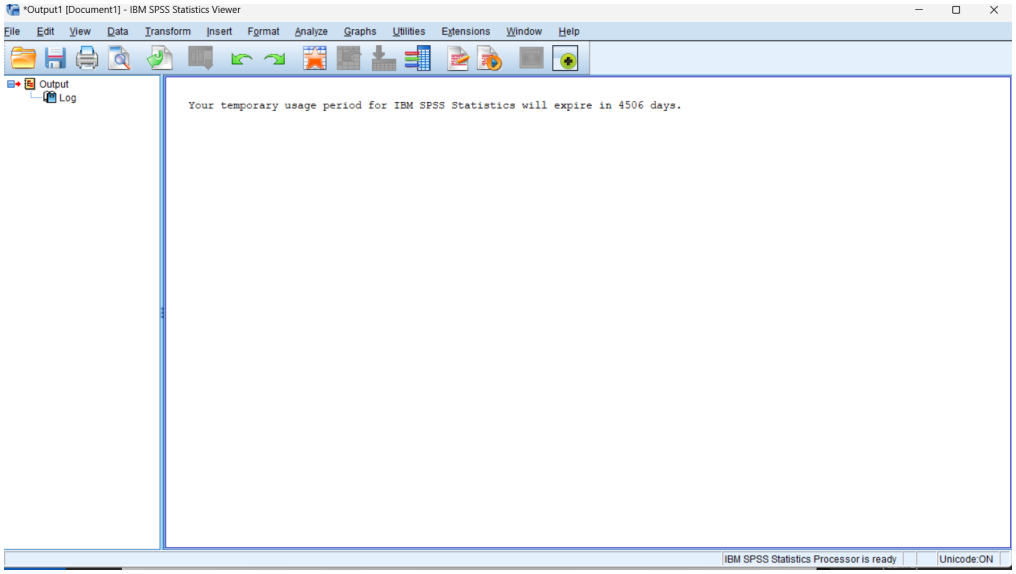


Figure 3: SPSS Output/Statistics Viewer

There are other windows, but we will refer to them only when we need to use them. For now, we will just call the Data Editor and the Output Viewer (shorthanded as Viewer) the *main windows*.

Using data

SPSS has a special type of data format, the `.sav` format. You will often find that datasets are provided in this format, specially in national surveys. However, you can also use other types of data formats, such as `.csv` or `.xlsx`.

One can also use the data editor to input data manually, however, for very large data collection efforts, there are specialized software that are used to collect data which export the resulting in the `.sav` format.

In the welcome screen, you can click on the *Open an existing data source* button to open a dataset. You can also click on the *New dataset* button to create a new dataset. We can also use the *File* menu to open or create a dataset.

The welcome screen also allows you to connect to sample datasets. This is a very useful feature to perform reproducible analysis. Use the *Sample* tab and connect to the `demo.sav` dataset.

The demo dataset

SPSS provides a description of every sample dataset available. The description of the `demo.sav` dataset is as follows: *This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.*

In the Data View of the Data Editor, you will see that every row is a customer, and every column is a variable. In the variable view, you can see every variable, and it gives us information about the type of variable that it is. A numeric variable is a variable that contains numbers, and a string variable is a variable that contains text.

Note, however, that a numeric variable doesn't necessary actually mean that the variable has a numerical meaning. The education variable, (`ed`) for instnace, is numerical, however, every number has a qualitative meaning. You can look at the *value labels* of a variable by clicking on the *Values* button in the Variable View for the corresponding variable. A view like this one will appear:

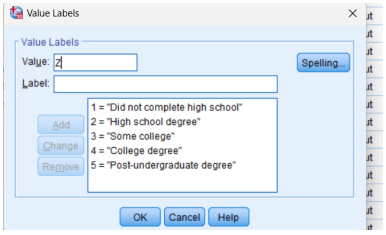


Figure 4: SPSS Values Viewer

So, whenever you see a person who has a value of 2, it means that they completed a high school education.

Another important thing to see in the Variable view is the "Label" column. Each variable in SPSS has a name which is often short and concise, and a label which is often longer and more descriptive. The label is what you will see in the output of your analyses, but you will need to use the name to refer to the variable in your analyses.

For instance, the `employ` variable has a label which says "Years with current employer". Good SPSS datasets have informative variable and value labels.

Running analyses

Now that we have a dataset, we can run analyses on it. We will start by running a simple analysis with descriptive statistics. Let us focus on the `age` variable, which is pretty self explanatory.

How to do this in SPSS? SPSS, like other languages, has a "point and click" interface. This means that we can go to the menus in the Data Editor and click buttons to run analyses. At some point we may need to quit this approach, but for now it is okay.

So, to get some descriptive statistics for the `age` variable, we can go to the *Analyze* menu, then to *Descriptive Statistics*, and then to *Descriptive*. A window like this one will appear:

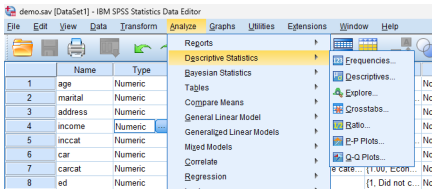


Figure 5: SPSS Menu

After selecting descriptives, another window will pop up, which asks us to select the variable that we want to get descriptive statistics for. We could even select more than one.

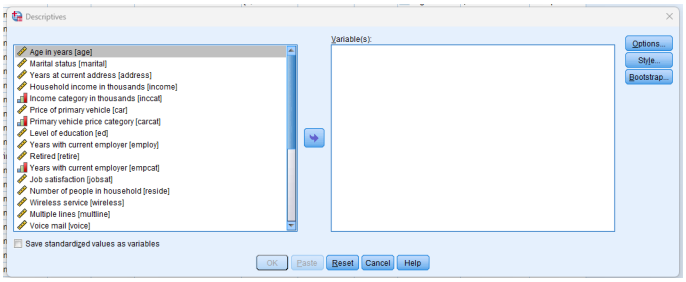
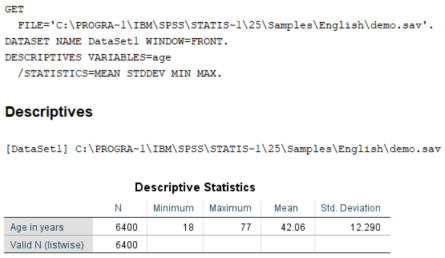


Figure 6: SPSS Descriptives

Afterwards, we will get results in the Viewer window. They will look like this:



```
GET
  FILE='C:\PROGRA~1\IBM\SPSS\STATIS~1\25\Samples\English\demo.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
DESCRIPTIVES VARIABLES=age
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

```
[DataSet1] C:\PROGRA~1\IBM\SPSS\STATIS~1\25\Samples\English\demo.sav
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age in years	6400	18	77	42.06	12.290
Valid N (listwise)	6400				

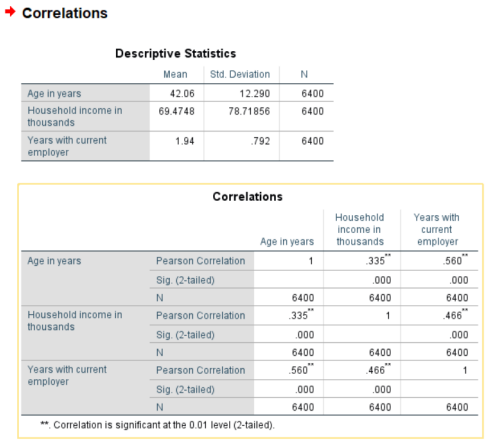
Figure 7: SPSS Descriptives Results

Many other analyses can be done in SPSS using this same approach. We click the Analyze menu, then we click the type of analysis that we want to run, and then we select the variables that we want to run the analysis on.

For instance, with categorical variables, such as the level of education, we run the same analysis, but we select the *Frequencies* option instead of the *Descriptives* option. We may customize and add additional output to the analysis by clicking the *Statistics* button in the Frequencies window. We can even include graphics which are automatic that come with the frequencies.

The crosstabs function permits to perform cross-tabulation, which is a very useful tool to understand the relationship between two categorical variables. Further, covariance and correlation can be calculated between two numerical variables by using the *Correlate* option in the *Analyze* menu. We choose the *Bivariate* option to calculate the correlation between two variables. The test comes with a two tailed hypothesis test. This gives the p-value that the correlation is not zero. If the p-value is less than 0.05, we can reject the null hypothesis that the correlation is zero. This will also allow the possibility to calculate a correlation matrix, which gives a correlation for many different variables.

An example result from a correlation matrix is shown below.



Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
Age in years	42.06	12.290	6400
Household income in thousands	69.4748	78.71856	6400
Years with current employer	1.94	.792	6400

Correlations

		Age in years	Household income in thousands	Years with current employer
Age in years	Pearson Correlation	1	.335**	.560**
	Sig. (2-tailed)		.000	.000
	N	6400	6400	6400
Household income in thousands	Pearson Correlation	.335**	1	.466**
	Sig. (2-tailed)	.000		.000
	N	6400	6400	6400
Years with current employer	Pearson Correlation	.560**	.466**	1
	Sig. (2-tailed)	.000	.000	
	N	6400	6400	6400

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

Figure 8: SPSS Correlation Matrix

How to read a correlation matrix? The correlation between two variables is a number between -1 and 1. A correlation of 1 means that the two variables are perfectly positively correlated, a correlation of -1 means that the two variables are perfectly negatively correlated, and a correlation of 0 means that the two variables are not correlated at all. We must carefully interpret the results of a correlation matrix by looking at it in a cuadricule fashion. The diagonal line which divides the matrix in two sandwich like parts is always of ones, because those are the correlations of a variable with itself. The others are the other correlations, which are interesting to us.

Notice that as you keep running analyses, the results will keep appearing in the Viewer window. You can click on the tabs to see the results of different analyses. This tree-like structure is useful to navigate all results.

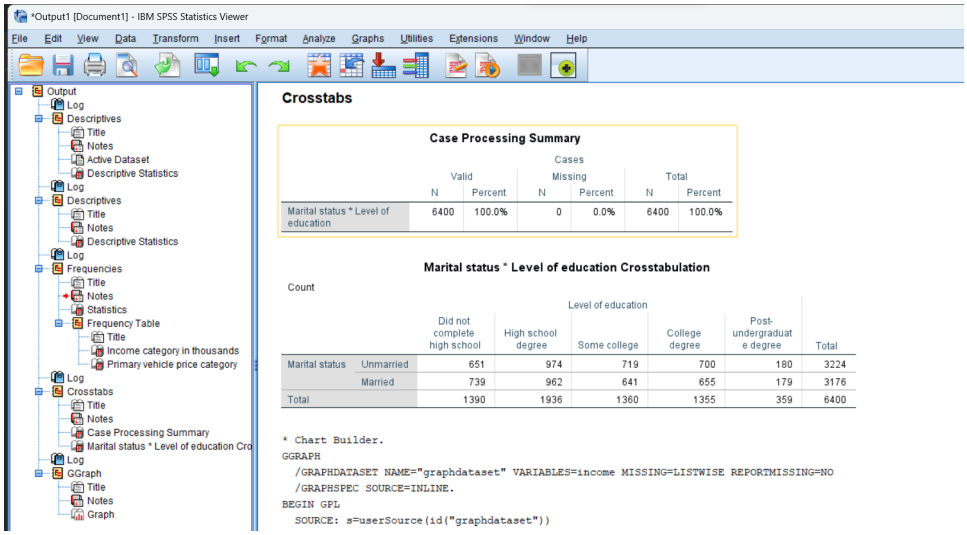


Figure 9: SPSS Output Viewer

Graphics

SPSS also allows us to create graphics. We can create histograms, scatterplots, bar charts, and many other types of graphics. We can do this by going to the *Graphs* menu, and then selecting the type of graph that we want to create.

The chart builder window will appear. In this window, we can select the type of graph that we want to create, and then we can drag and drop the variables that we want to use in the graph.

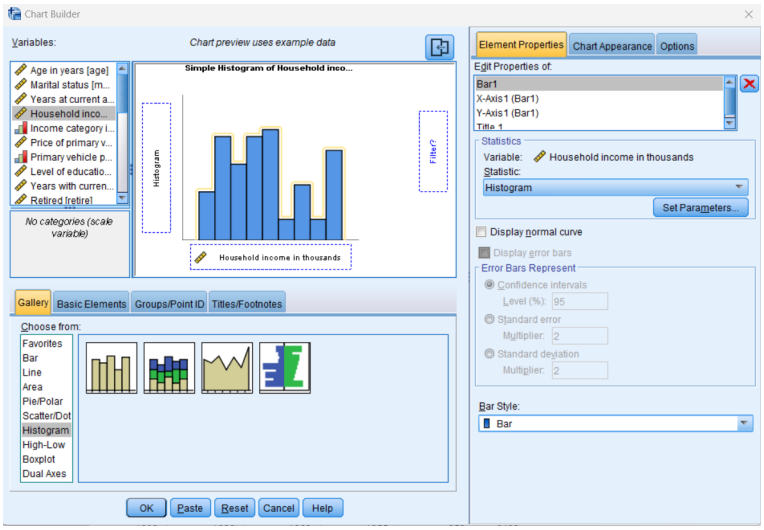


Figure 10: SPSS Chart Builder

Afterwards, we can click the *OK* button, and the graph will appear in the Viewer window.

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=income MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: income=col(source(s), name("income"))
  GUIDE: axis(dim(1), label("Household income in thousands"))
  GUIDE: axis(dim(2), label("Frequency"))
  GUIDE: text.title(label("Simple Histogram of Household income in thousands"))
  ELEMENT: interval(position(summary.count(bin.rect(income))), shape.interior(shape.square))
END GPL.
```

GGraph

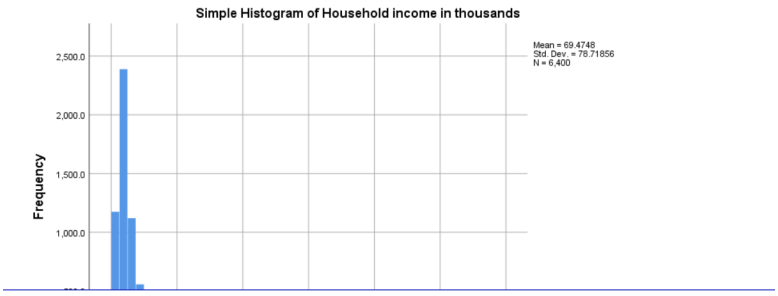


Figure 11: SPSS Chart Builder Results

By right clicking in the graphs, we can edit them. We can change the colors, the fonts, the labels, and many other things. We need to select the edit output option. Further, we can also export the graph to an image so it can be used elsewhere.

Preparing data

While, for our purposes, most of the time we will have data that was already prepared for analysis, in the real world data is very dirty and needs to be cleaned before it can be used. SPSS allows us to do this.

Using the *Transform* menu, we can perform many different types of data manipulation. We can create new variables, we can recode variables, we can compute new variables, we can select cases, we can split files, we can merge files, and we can sort cases.

To create new variables, we can use the *Compute Variable* option. This will allow us to create a new variable based on the values of other variables. For instance, we can create a new variable that is the sum of two other variables. We can also create a new variable that is the average of two other variables.

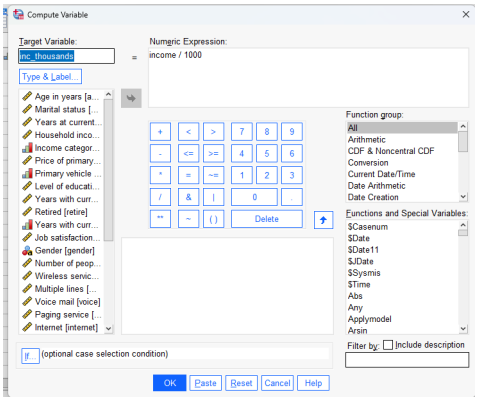


Figure 12: SPSS Transform

Specific analyses to be performed in SPSS

Group means

Calculating the average or mean value of a numeric variable is very common in statistics. We can do this in SPSS by using the *Analyze* menu, then the *Compare Means* option, and then the *Means* option. We then select the variable that we want to calculate the mean for as the *dependent* and the grouping variable as the *independent*.

As an example, we may calculate average household income by different levels of education. As always, after using the menus, we will get the results in the Viewer window.

Proportions

Remember that the proportion is a measure of relative frequency. For instance, we might want to know what are the sample proportions of people who have certain levels of education. The proportion is given in decimal form, so we need to multiply it by 100 to get the percentage.

We can calculate proportions in SPSS by using the *Analyze* menu, then the *Descriptive Statistics* option, and then the *Frequencies* option. We then select the variable that we want to calculate the proportion for.

There is another option in the *Compare Means* option called *One-Sample Proportion*. This option allows us to calculate the proportion of a categorical variable that has a certain value. For instance, we can calculate the proportion of people who have a high school education. This would yield the same result as using the *Frequencies* option, but it is a bit more convenient, because it presents automatically a hypothesis test. The hypothesis test is as follows:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Where p is the proportion of people who have a certain level of education, and p_0 is the value that we are testing. If the p-value is less than 0.05, we can reject the null hypothesis that the proportion is equal to p_0. The default for the hypothesized proportion is 0.5, but we may change it by clicking the *Tests* button in the *One-Sample Proportion* window.

This menu will only calculate one proportion, so we must select the variable that we want to calculate the proportion for and the level of such categorical variable that we want to calculate the proportion for. We will need to give the level in terms of how it actually appears in the dataset, so it might be useful to get familiar with the value labels of the variable in the Variable view in the Data Editor. Further, we can also calculate a composite proportion, which is the proportion of people who have a certain level of education or another. This is also known as the intersection of two events. We need only to select the variable and write down the two or more levels that we want, separated by a space, in the *Define Success* box of the menu.

Further, this option also gives the confidence interval of the proportion that we are calculating. The confidence interval is a range of values that we are confident that the true value of the proportion is in. The default confidence interval is 95%, but we can change it by clicking the *Confidence Interval* button in the *One-Sample Proportion* window.

T-tests

There are many different types of t-tests that we can run in SPSS. The most basic one is the one-sample t-test, which tests whether the mean of a variable is equal to a certain value. For instance, we may want to test whether the average income of people is equal to 50,000 dollars. This test would be as follows:

$$H_0 : \mu = 50,000$$

$$H_1 : \mu \neq 50,000$$

Where μ is the mean of the variable. If the p-value is less than 0.05, we can reject the null hypothesis that the mean is equal to 50,000.

We can run this test by using the *Analyze* menu, then the *Compare Means* option, and then the *One-Sample T Test* option. We then select the variable that we want to calculate the mean for and the value that we want to test.

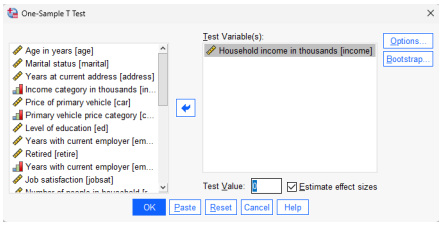


Figure 13: SPSS T-Test

How to modify the hypothesized value of the test, which in this case is $\mu_0 = 50,000$? We modify the *test value* in the *One-Sample T Test* window. The output will give us a p-value which we can use to reject or fail to reject the null hypothesis.

Another type of t-test is the independent samples t-test, which tests whether the means of two groups are equal. For instance, we may want to test whether the average income of people who have a high school education is equal to the average income of people who have a college education. This test would be as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where μ_1 is the mean of the variable for people who have a high school education, and μ_2 is the mean of the variable for people who have a college education. If the p-value is less than 0.05, we can reject the null hypothesis that the means are equal.

We can run this test by using the *Analyze* menu, then the *Compare Means* option, and then the *Independent Samples T Test* option. We then select the variable that we want to calculate the mean for and the grouping variable. The grouping variable should be some kind of categorical variable with only two levels, otherwise, we will need to specify two values for the grouping variable so that the test can be done.

The output should be interpreted in the row where it says *Equal variances not assumed*, and looking at the two sided p-value. We then use it to reject or fail to reject the null hypothesis.

Saving results

SPSS allows us to save the results of our analyses. We can save the results in the `.spv` format, which is the SPSS format for output files. We can also save the results in the `.pdf` format, which is a format that is easy to share with others.

To save, we simply go to the *File* menu, and then we click the *Save* option. We can also click the *Save As* option to save the results in a different format. Then, we can select the location where we want to save the file, and we can give it a name.

We may also export a manipulated dataset. To do this, we go to the *File* menu, and then we click the *Save As* option. Then, we can select the location where we want to save the file, and we can give it a name. We can also select the format that we want to save the file in. If you want to use it using SPSS as well, the `.sav` format is the best.

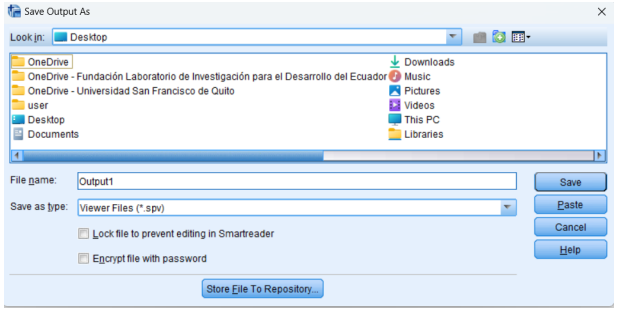


Figure 14: SPSS Save