Beginning Steps in SPSS

Keith E. Emmert

Soad Emmert

Department of Mathematics, Tarleton State University
E-mail address: emmert@tarleton.edu

Department of Mathematics, Tarleton State University
E-mail address: semmert@tarleton.edu

Contents

CHAPTER 1

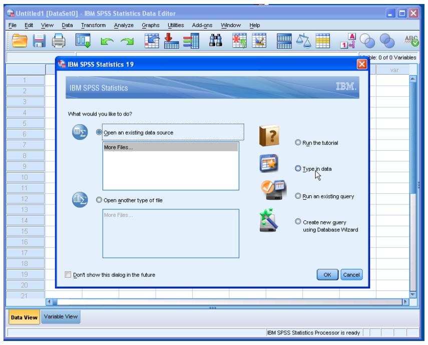
First Steps with SPSS

1.1. Introduction

SPSS stands for Statistical Package for the Social Sciences. It was developed in 1968, by Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent. Norman, frustrated with mainframe software which was inadequate for his needs, needed a program to quickly analyze volumes of social science data gathered through various methods of research. They wrote SPSS, the first of its kind, and targeted the PC. They began selling to other professors at other universities and by 1974 was making over $200,000 per year – without marketing!

1.2. Entering Data

When you first open SPSS, you will see a window asking you what you would like to do in SPSS, much like that in Figure 1. To start creating your own dataset, select the "Type in data" option.


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
IBM SPSS Statistics 19
IBM SPSS Statistics
IBM.
More Files...
More Files...
OK
Cancel
Data View
Variable View

What would you like to do?

● Open an existing data source

○ Open another type of file

○ Run the tutorial

○ Type in data

○ Run an existing query

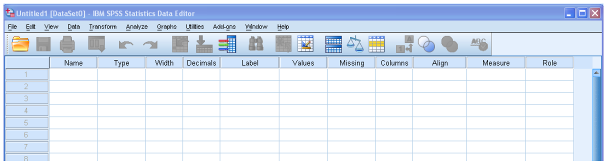
○ Create new query using Database Wizard

□ Don't show this dialog in the future

IBM SPSS Statistics Processor is ready

FIGURE 1. Splash Screen for SPSS

Once you click the "OK" button, you will be given a blank data table. Now, click on the **Variable View** tab located at the bottom left corner of the window. Here is where you will declare your variables. You will notice that the column headers have changed. The column headers should look like the headers in Figure 2

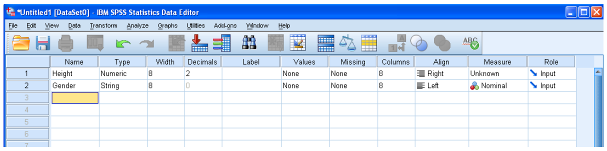

Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1											
2											
3											
4											
5											
6											
7											
8											

FIGURE 2. Variables Tab for SPSS

For the purposes of this tutorial, we will create a simple dataset with two variables. The dataset will contain a list of people's height and gender.

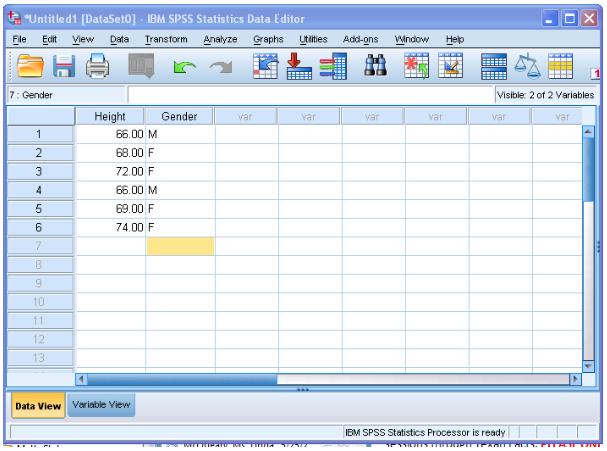
To start entering your variables, type the name of the variable into the "Name" box. The default values are loaded into all the other boxes. To change the variable type, click the grey dotted area in the "Type" box. The window shown in Figure 3 should open. This is where you can select the variable type. For our example we leave "Height" as a numeric and change the default value of "Gender" to be a string.


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Height	Numeric	8	2		None	None	8	Right	Unknown	Input
2	Gender	String	8	0		None	None	8	Left	Nominal	Input
3											
4											
5											
6											

FIGURE 3. Entering Variables in SPSS

Now that we have declared our variables, we can start entering the data. To add the data, click on the **Data View** tab located at the bottom left corner. The following window will appear, where you can add the data. The window will look as shown in Figure 4


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
7 : Gender
Visible: 2 of 2 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

	Height	Gender	var	var	var	var	var	var
1	66.00	M						
2	68.00	F						
3	72.00	F						
4	66.00	M						
5	69.00	F						
6	74.00	F						
7								
8								
9								
10								
11								
12								
13								

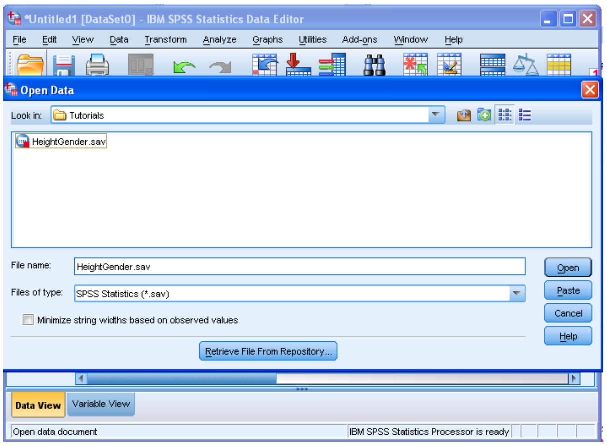
FIGURE 4. Entering Data in SPSS using the Data View

To save the given dataset, click on the **File** menu and choose **Save**. The following window will appear. Here you can choose the location of where you want to save the dataset. By default your file is saved in a ".sav" format.

1.3. Opening an Existing Dataset

To open a dataset, open SPSS and the following screen will appear as shown in Figure 5. Choose "Open an existing data source."

The textbook contains many datasets which are located on the CD. By default, the only datasets that will be visible will be datasets that have the file extension ".sav." However, many of the older datasets will have the ".sys" extension. To view these, just change the "Files of types:" drop-down menu setting to "SYS/PC+" and you will be able to see the .sys files.



Look in: Tutorials

HeightGender.sav

File name: HeightGender.sav

Files of type: SPSS Statistics (*.sav)

☐ Minimize string widths based on observed values

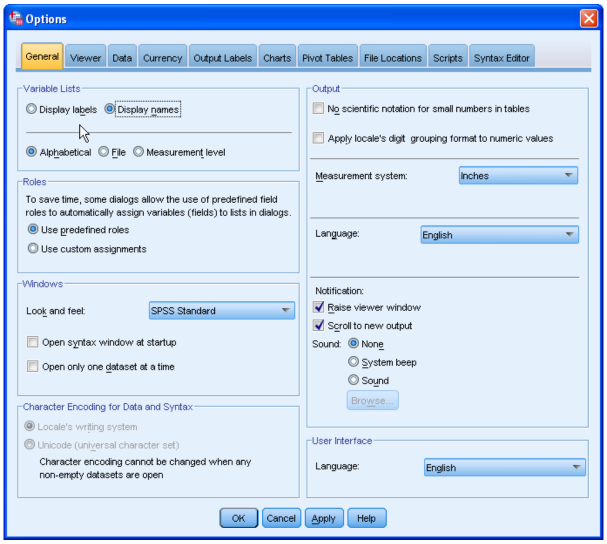
Open data document

IBM SPSS Statistics Processor is ready

FIGURE 5. Open a Saved Data Set in SPSS

1.4. Changing How Variables are Listed

When you start to run analyses, you will find that the variables will be listed by their label in the order they appear on the dataset. Because some datasets can contain over 100 variables, finding a specific variable can be difficult. You can change the way the variables are listed by using the **Options** window. This can be reached by choosing **Options** under the **Edit** menu. Here we can change the display to list the data sets in alphabetical order (instead of file) and to display the "Names" instead of the "Labels." After choosing your options, click "OK," as can be ssen in Figure 6. A warning about the changing of options in the **Variable List** group will reset all dialog box setting will be displayed. Just click "OK." Now, when you run an analysis, the variables will be listed by their name in alphabetical order. This greatly speeds up the process of finding variables.



Variable Lists

○ Display labels ● Display names

● Alphabetical ○ File ○ Measurement level

Roles

To save time, some dialogs allow the use of predefined field roles to automatically assign variables (fields) to lists in dialogs.

● Use predefined roles

○ Use custom assignments

Windows

Look and feel: SPSS Standard

☐ Open syntax window at startup

☐ Open only one dataset at a time

Character Encoding for Data and Syntax

● Locale's writing system

○ Unicode (universal character set)

Character encoding cannot be changed when any non-empty datasets are open

Output

☐ No scientific notation for small numbers in tables

☐ Apply locale's digit grouping format to numeric values

Measurement system: Inches

Language: English

Notification:

☑ Raise viewer window

☑ Scroll to new output

Sound: ● None

○ System beep

○ Sound

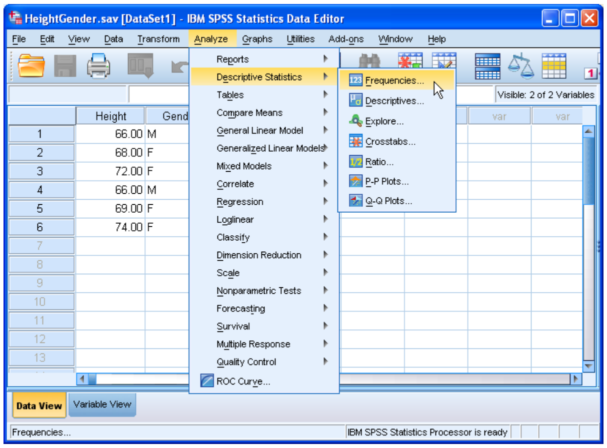
User Interface

Language: English

FIGURE 6. Options Window in SPSS. Change to alphabetical listing of variables and display names.

1.5. Analyzing a Dataset

This is where we begin to perform statistics. Click on the **Analyze** menu. We will be using the Frequencies analysis, so go to the **Descriptive Statistics** sub-menu and select **Frequencies...**, as shown in Figure 7.



HeightGender.sav [DataSet1] - IBM SPSS Statistics Data Editor

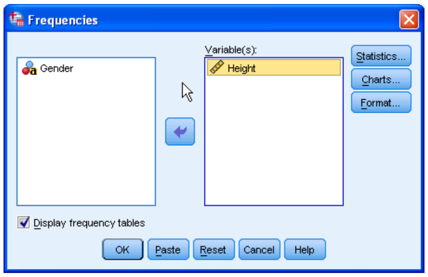
	Height	Gend
1	66.00	M
2	68.00	F
3	72.00	F
4	66.00	M
5	69.00	F
6	74.00	F

Frequencies... IBM SPSS Statistics Processor is ready

FIGURE 7. Select to perform descriptive statistics using the **Frequencies...** option.

The **Frequencies...** option allows you to analyze variables individually as opposed to analyzing them in relation with another variable. With this type of analysis, you can measure a variable's individual properties such as mean, median, mode, etc.

To start the analysis, in the **Frequencies** window, move the variables you wish to analyze from the left list to the right list. This is done by selecting the variables from the left list and clicking the right-arrow button located between the two lists. If you wish to remove a variable, simply reverse the process and move it back to the left list. See Figure 8.



Frequencies

Gender

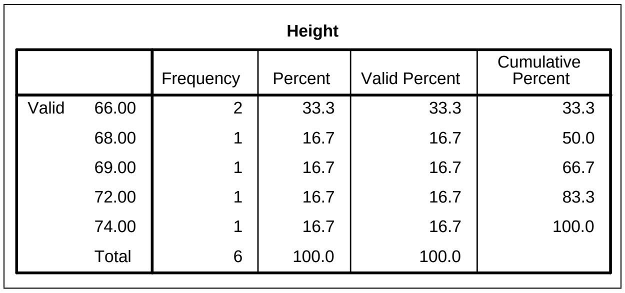
Variable(s): Height

☑ Display frequency tables

FIGURE 8. The **Frequencies...** dialog window.

Once you have chosen the required variables and set your display options, click the "OK" button. Another window will open up with the outputs displayed in the following format as shown in Figure 9.

You can save this output window by clicking on the "Save" button. You can close this window once the output has been saved. You can also delete the output, by clicking on the "Output" option in the left window, and then hitting the "Delete" button. If you leave this window open, and run another analysis, then the output for that run will be stacked under the existing output. If you print this output, then



Height

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	66.00	2	33.3	33.3	33.3
	68.00	1	16.7	16.7	50.0
	69.00	1	16.7	16.7	66.7
	72.00	1	16.7	16.7	83.3
	74.00	1	16.7	16.7	100.0
	Total	6	100.0	100.0	

FIGURE 9. The **Frequencies...** results.

all the analysis outputs will be printed, so it is recommended that you delete the output analysis from the output window once you have saved them.

For the rest of this section, you will need to change how the variables are listed in the analysis windows. If you need help, please refer to **Section Four: Changing how variables are listed**.

If you wish to see some basic descriptive statistics, then choose **Analyze** followed by **Descriptive Statistics** followed by **Frequencies...**. Again, you select your variable (Height) and choose the **Statistics...** button as shown in Figure 10.

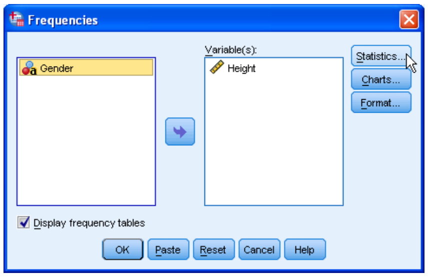


FIGURE 10. The **Frequencies...** dialog - accessing the Statistics...options..

This will open a new window, the **Frequencies: Statistics** dialog box. Check **Mean**, **Median**, **Mode**, **Variance**, **Standard Deviation**, **Range**, and **Quartiles**. Press the "Continue" button.

Now choose "OK." SPSS should stack the results below your frequencies as seen in Figure 12. Notice that SPSS calculates the five number summary, plus $\bar{x}$, s, s^2, and range.

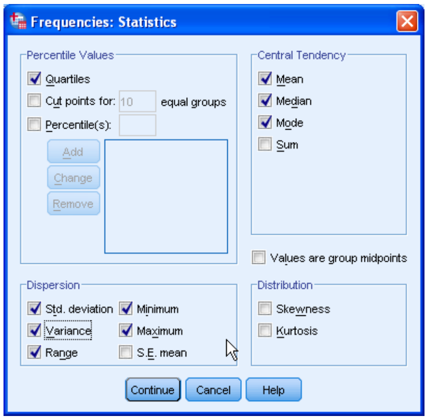
1.6. Creating a Box plot

Let's open up a larger data set. SPSS provides us with "cars.sav." There are several variables in the "cars.sav" data set: mpg, engine, horse, weight, accel (0 - 60 mph), year (1970 - 1982), origin, cylinder, and filter_$.

First, let' create a boxplot for mpg. Choose **Analyze - Descriptive Statistics - Explore...** to open the **Explore** dialog box as shown in Figure 13. Go ahead and add "mpg" to the **Dependent List** and select "Plots" i the **Display** group.

Next, select the **Plots...** button which opens the **Explore:Plots** dialog box, as can be seen in Figure 14. Select "Factor levels together" (it should already be selected) in the "Boxplots" group and uncheck "Stem-and-leaf" in the "Descriptive" group.

Select **Continue** and then **OK**. The results will appear in an output window and should look like the image in Figure 15. Notice that one outlier (observation #330) is plotted as a circle.



Percentile Values
☑ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

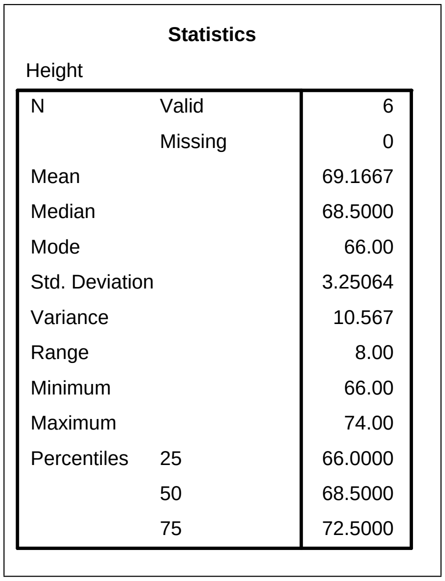
Central Tendency
☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

Dispersion
☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☑ Range ☐ S.E. mean

Distribution
☐ Skewness
☐ Kurtosis

FIGURE 11. The **Frequencies: Statistics** dialog - accessing the additional descriptive statistics.

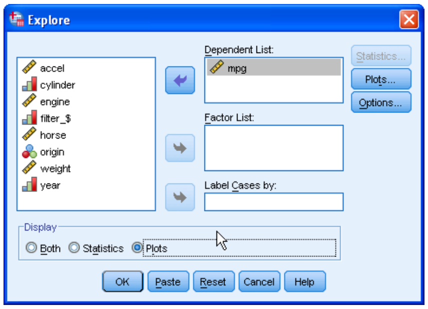


Statistics

Height

N	Valid	6
	Missing	0
Mean		69.1667
Median		68.5000
Mode		66.00
Std. Deviation		3.25064
Variance		10.567
Range		8.00
Minimum		66.00
Maximum		74.00
Percentiles	25	66.0000
	50	68.5000
	75	72.5000

FIGURE 12. The **Descriptives...** results: Five number summary, plus $\bar{x}$, s, s^2, and range.



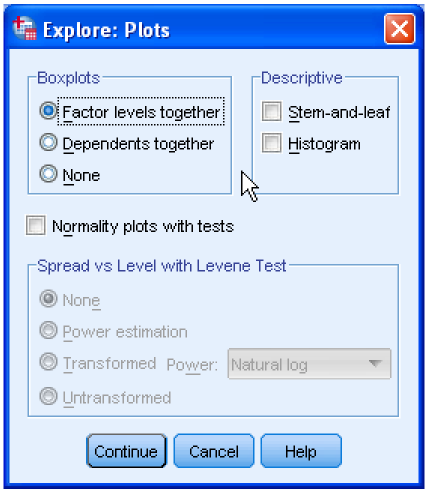
Dependent List: mpg

Factor List:

Label Cases by:

Display
○ Both ○ Statistics ● Plots

FIGURE 13. The **Explore...** dialog for exploring data sets.



Explore: Plots

Boxplots
- ● Factor levels together
- ○ Dependents together
- ○ None

Descriptive
- □ Stem-and-leaf
- □ Histogram

□ Normality plots with tests

Spread vs Level with Levene Test
- ● None
- ○ Power estimation
- ○ Transformed Power: Natural log
- ○ Untransformed

FIGURE 14. The **Explore:Plots...** dialog for exploring data sets.

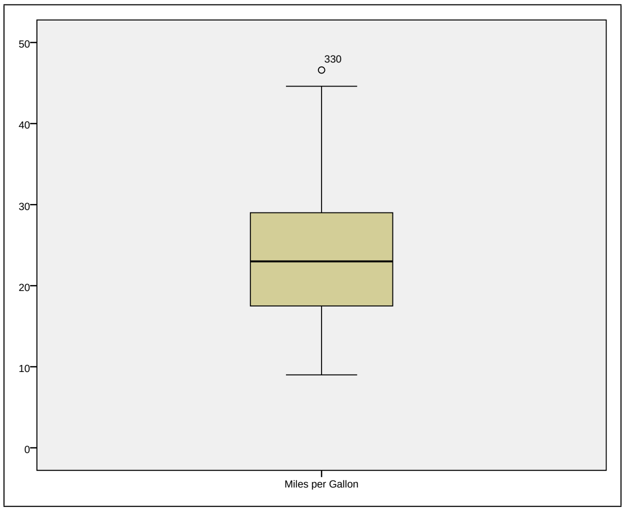


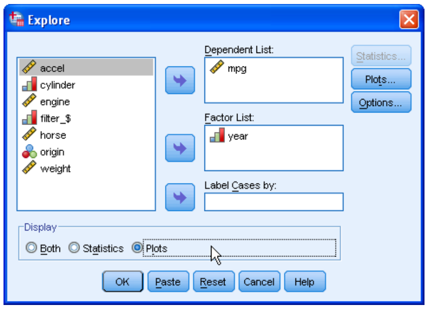
FIGURE 15. The box plot for mpg.

Of course, this is a bit misleading since it combines cars over a span of 13 years. Perhaps we should break them up into smaller pieces.

To make box plots for the mpg based upon a given year, choose **Analyze - Descriptive Statistics - Explore...** to open the **Explore** dialog box as shown in Figure 16. Notice that I have already added "mpg" to the **Dependent List**, "year" to the **Factor List**, and selected "Plots" in the **Display** group.

Click the **Plots...** button to open the **Explore: Plots** dialog box as shown in Figure 17. Choose the bullet for "Factor levels together" to make a separate Boxplot for each of the variables in the "Dependent List" in the Explore dialog box. If there are several variables in the Dependent List, choose the bullet for Dependents together to obtain side-by-side Boxplots.

Select **Continue** and then **OK**. The resulting side-by-side box plots are shown in Figure 18.



Explore

accel
cylinder
engine
filter_$
horse
origin
weight

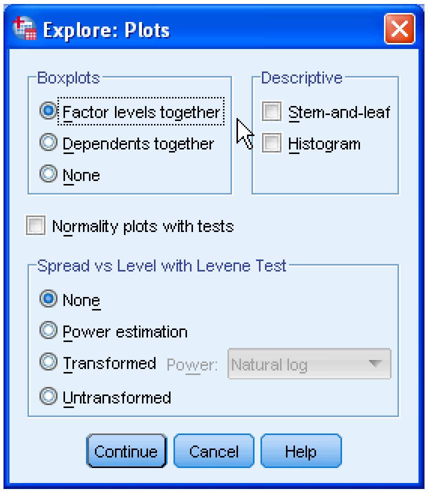
Dependent List: mpg

Factor List: year

Label Cases by:

Display: ○ Both ○ Statistics ● Plots

FIGURE 16. The **Explore...** dialog for exploring data sets.



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
□ Stem-and-leaf
□ Histogram

□ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

FIGURE 17. The **Explore:Plots...** dialog for exploring data sets.

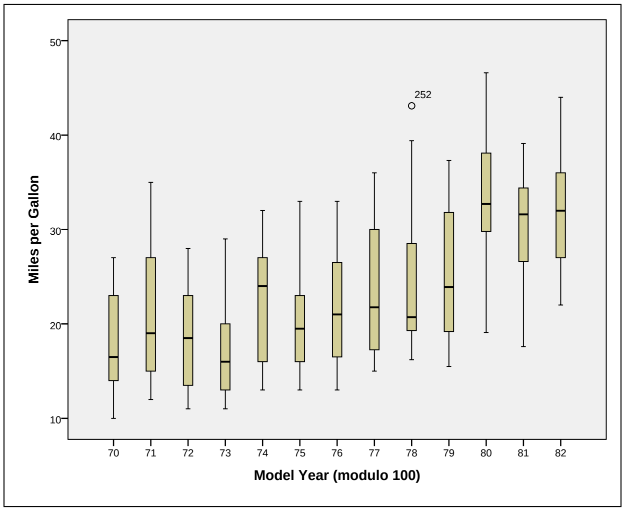


FIGURE 18. The side-by-side box plots of mpg based upon a given year.

Suppose that you wish to view side-by-side box plots of mpg and acceleration based upon year. The procedure is virtually the same. Just add the “Accel” variable to the “Dependent List” and click on the **Plots...** button to open the **Explore:Plots** dialog. In the **Explore:Plots** dialog, make sure you select “Dependents together” in the “Boxplots” section. The resulting side-by-side boxplots are shown in Figure 19.

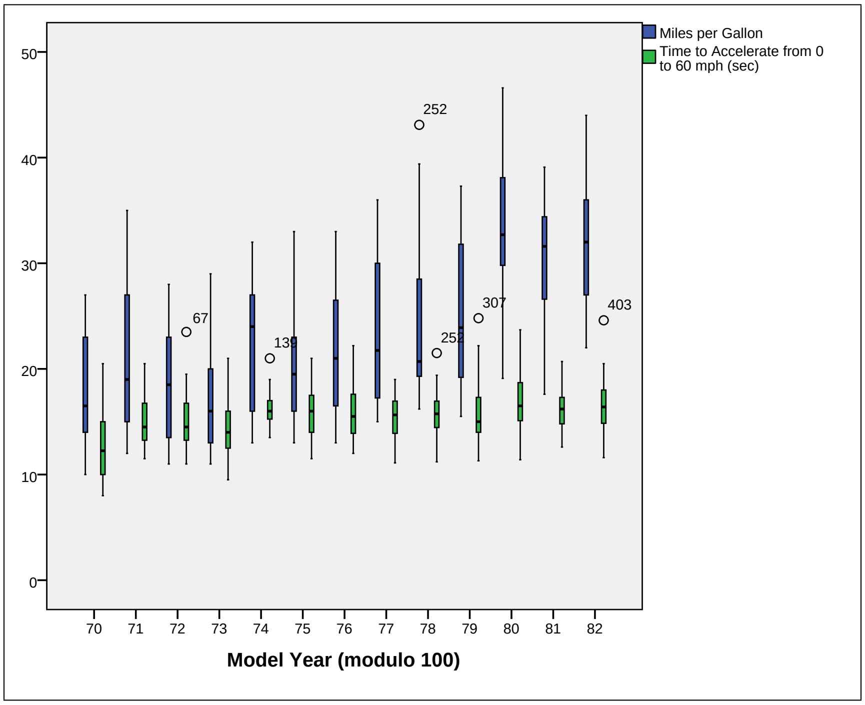


FIGURE 19. The side-by-side box plots of mpg and acceleration based upon a given year.

It is interesting to notice that during certain years, ’72, ’74, ’78, ’79, and ’82, there are instances of cars that have extremely slow acceleration. In ’78, there is a car that received exceptional gas mileage. Notice that this was not the slowest car that year, either (but if you look at the data set, it was close).

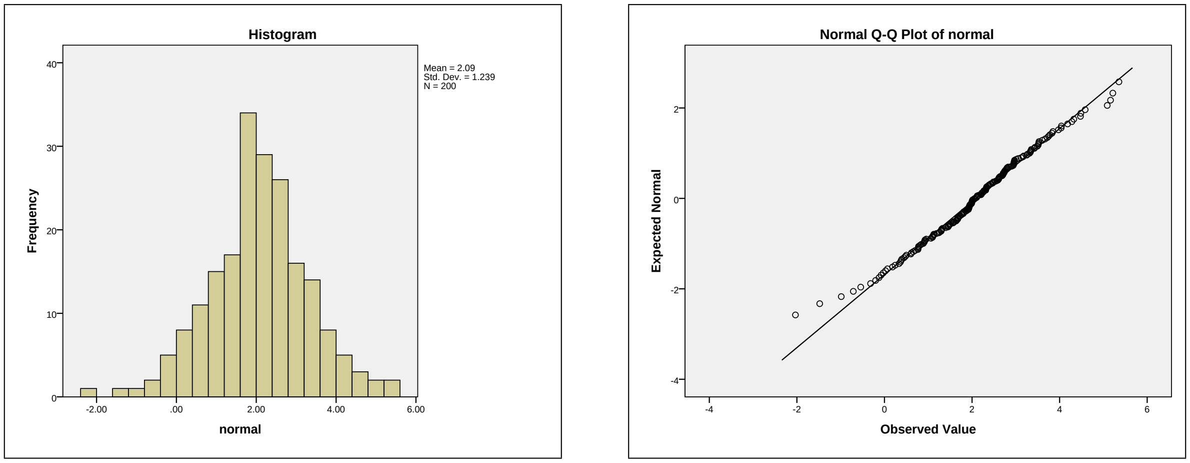
1.7. Creating Histograms and Q-Q Plots

Let’s first see what how a histogram and Q-Q plot from a normal distribution looks like. The plots in Figure 20 were generated from 200 samples of a random variable which follows $N((2, 1.2649^2)$.

Notice that in Figure 20a, we see a histogram. It is unimodal (mound shaped) and symmetric, both of which are characteristics of a normal distribution. Now, in Figure 20b, we see a Q-Q Plot. Most of the sample are grouped in the middle (again indicating symmetry). More importantly, most of the data (i.e. in the middle) lies on a straight line. The straight line is a “perfect” normal distribution (using the sample mean and standard deviation, that is $N(2.09, 1.239^2)$), so the closer we come to this line, the more likely we have stumbled across a normal distribution.

Now, let’s compare some non-normal histograms and Q-Q plots. We’ll consider samples from a binomial random variable, binomial$(10, 0.2)$, log-normal with parameters 2 and 0.5, and finally a uniform with parameters -1.8 and 5.8.

First, the binomial random variable. Figures 21a and 21b are the histogram and Q-Q Plot for this binomial random variable. Notice that the mean is $\mu = 2$ and the standard is $\sigma = 1.2649$, which is the

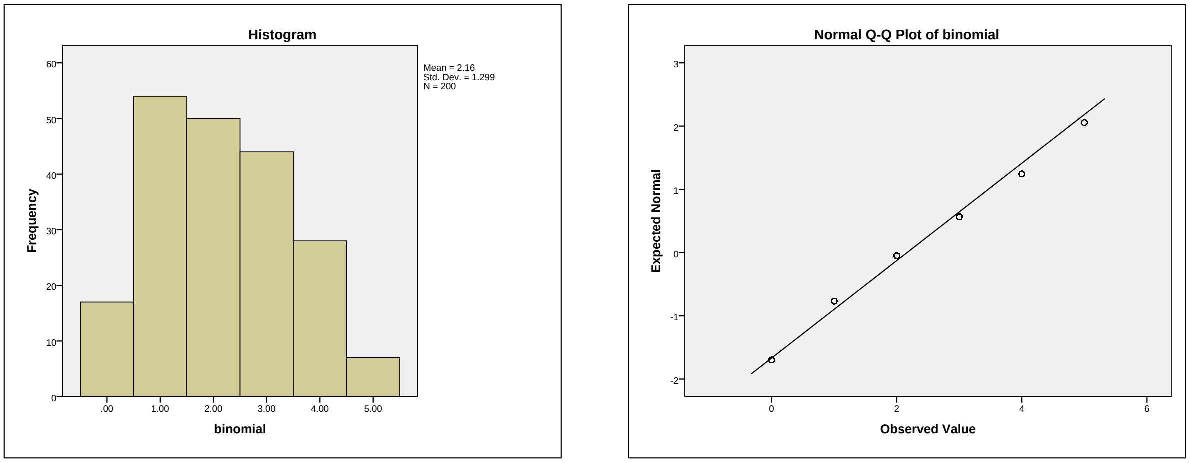


(A) Histogram for a Normal Random Variable

(B) Q-Q Plot for a Normal Random Variable

Figure 20. Histogram and QQ-Plot for a Normal Random Variable.

same as the mean and standard deviation of the normal used above. Notice that the histogram shows that the binomial is a bit right skewed and certainly not very symmetric. The Q-Q Plot only lists points on integer values (duh!, it's a binomial! It counts successes). Clearly, this is not a normal distribution.

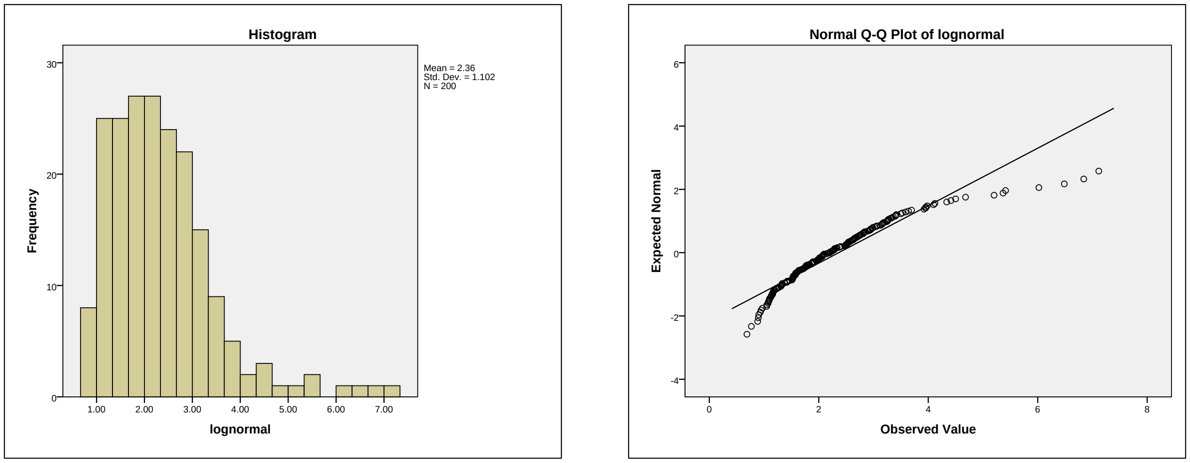


(A) Histogram for a Binomial Random Variable

(B) Q-Q Plot for a Binomial Random Variable

Figure 21. Histogram and QQ-Plot for a Binomial Random Variable.

Next, the log-normal random variable, which is a continuous random variable. Figures 22a and 22b are the histogram and Q-Q Plot for this log-normal random variable. Notice that the histogram shows that the log-normal is a definitely right skewed and not symmetric. In the Q-Q PLot, data is grouped closely together, just not the middle 50% portion of the data. Also, on the ends of the Q-Q Plot, the data moves significantly away from the line. This indicates potential outliers and is not usual for a normal random variable. Again, this is not a normal distribution.

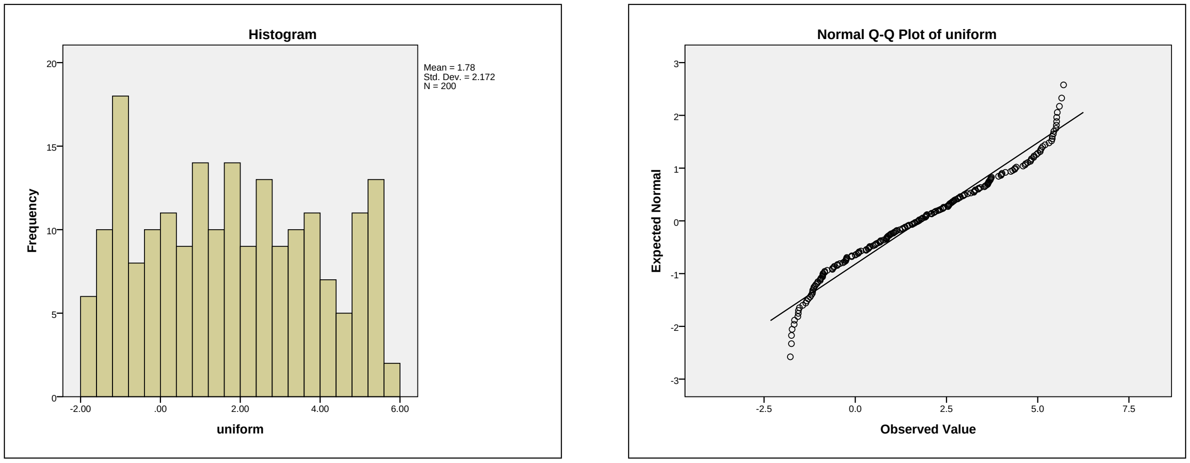


(A) Histogram for a Log-Normal Random Variable

(B) Q-Q Plot for a Log-Normal Random Variable

FIGURE 22. Histogram and QQ-Plot for a Log-Normal Random Variable.

Finally, we consider a uniform random variable, which is a continuous random variable. Figures 23a and 23b are the histogram and Q-Q Plot for this uniform random variable. Notice that the histogram shows that the uniform is a definitely right skewed and not symmetric. In the Q-Q PLot, data is grouped closely together, just not the middle 50% portion of the data. Also, on the ends of the Q-Q Plot, the data moves significantly away from the line. This indicates potential outliers and is not usual for a normal random variable. Again, this is not a normal distribution.



(A) Histogram for a Uniform Random Variable

(B) Q-Q Plot for a Uniform Random Variable

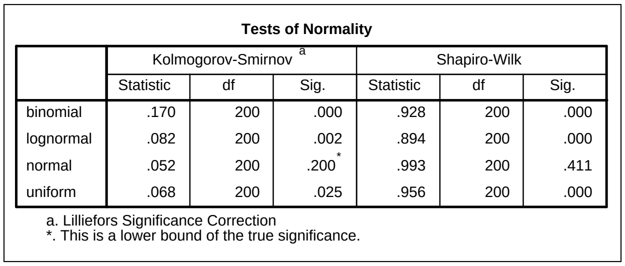
FIGURE 23. Histogram and QQ-Plot for a Uniform Random Variable.

One final check is using one of the tests of normality. SPSS uses Kolmogorov-Smirnov and Shapiro-Wilk. Note that Shapiro-Wilk only applies to sample sizes up to 2,000. There are other tests that are more useful in other situations. It is quite dangerous to use these tests on small sample sizes, especially ≤ 10. The Kolmogorov-Smirnov test may have some problems with large sample sizes, say $\geq 1,000$. In all cases, the particular hypotheses being tested are

$$H_0 : \text{The data is from a normal distribution.}$$
$$H_a : \text{The data is NOT from a normal distribution.}$$

A failure to reject indicates that the sample appears to come from a normal distribution.

See Figure 24. For the sample coming from a normal distribution, the Sig. is 0.200 for Kolmogorov-Smirnov and 0.411 for Shapiro-WIlk. Hence, we fail to reject H_0. It appears that the sample (based upon this test) comes from a normal distribution. Compare this to binomial, lognormal, and uniform. All have small p-values, certainly less that 0.05. Hence, the conclusion is that the sample does not appear to be from a normal distribution.



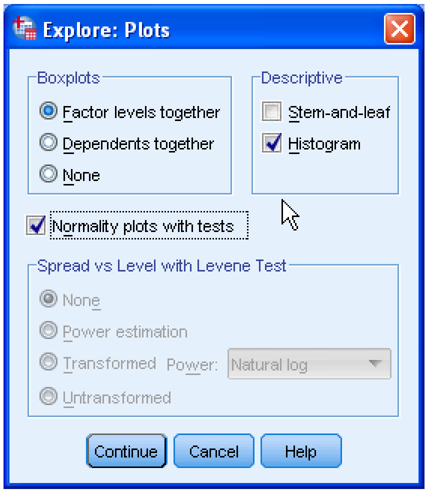
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
binomial	.170	200	.000	.928	200	.000
lognormal	.082	200	.002	.894	200	.000
normal	.052	200	.200*	.993	200	.411
uniform	.068	200	.025	.956	200	.000

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

FIGURE 24. Summary of Tests of Normality

First, you need to open a data file to analyze. We will analyze "EmployeeData.sav." Select Analyze→Descriptive Statistics→Explore. Select the variable "Salary" and add it to the "Dependent List." Click on "Plots..." and select "Histogram" and "Normality plots with tests." Deselect the "Stem-and-leaf" box. See Figure 25. Click "Continue" and "OK" to run the procedure. You should obtain



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
□ Stem-and-leaf
☑ Histogram

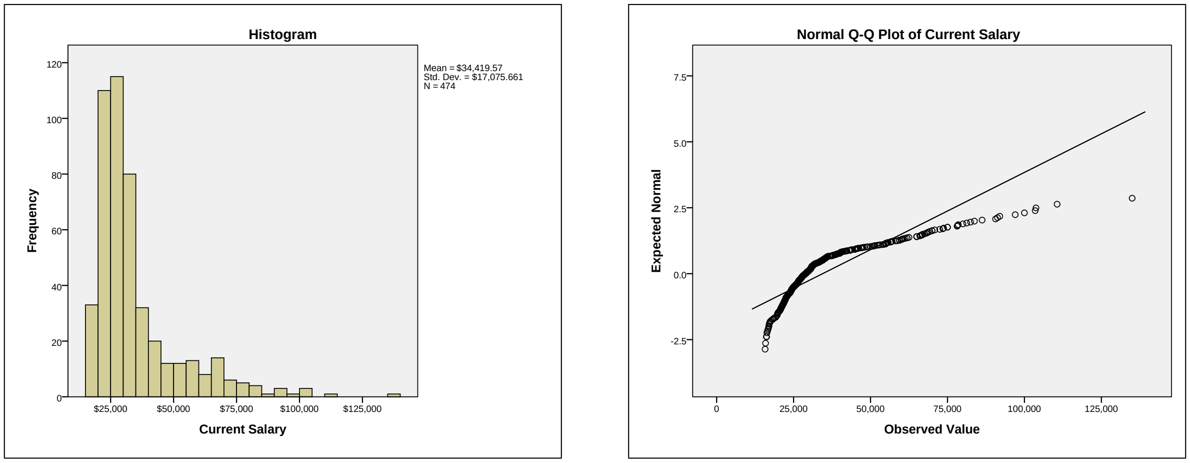
☑ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

FIGURE 25. Explore:Plots, used to create Histograms and Q-Q Plots

output as seen in Figure 26.

The histogram and Q-Q Plot are shown in Figures 26a and 26b. Notice that most of the data is concentrated to the left. This is clearly seen in the histogram. The Q-Q Plot also has a distinct bend in it. These characteristics suggest that the data is not normally distributed.



(A) Histogram for Salary

(B) Q-Q Plot for Salary

FIGURE 26. Histogram and QQ-Plot for Salary.

Another simple visual check is to overlay a normal curve using the calculated sample mean and sample standard deviation on top of your histogram. If the data is normally distributed, then they should (mostly) agree. In the histogram shown in Figure 28a, a normal curve in also plotted with the same mean and standard deviation as the data. Notice that this does not match up well with the histogram. Compare this to the data which was taken from a normal distribution (i.e. $N(2.1, 1.2649^2)$ at the beginning of this section; see also Figure 20a, the histogram without the normal curve). The histogram with a superimposed normal curve shown in Figure 28b has a much closer fit.

In order to superimpose a normal curve onto a histogram, open your data set, such as “EmployeeData.sav.” Select Analyze→Descriptive Statistics→Frequencies. Then, select the variable, such as “Salary” and click the “Charts” button. Select the “Histograms” radio button and then click the “Show normal curve on histogram” box. See Figure 27.

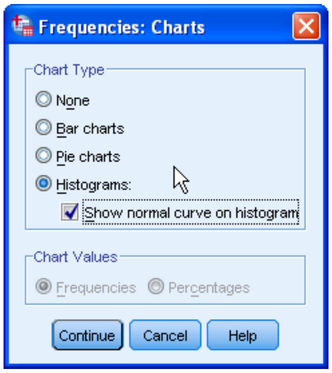
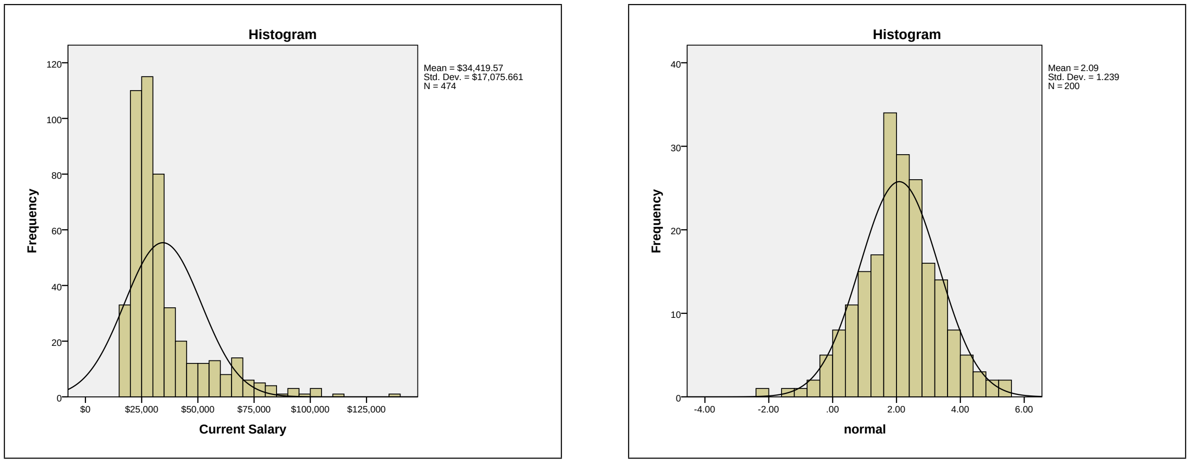

Frequencies: Charts

Chart Type
- ○ None
- ○ Bar charts
- ○ Pie charts
- ● Histograms:
  - ☑ Show normal curve on histogram

Chart Values
- ● Frequencies ○ Percentages

FIGURE 27. Frequences: Charts - Creating Histograms with Normal Curves

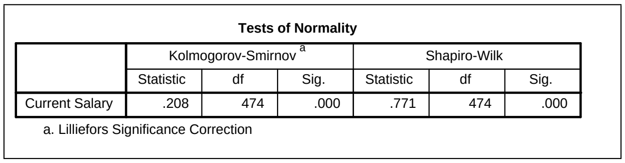


(A) Histogram with Normal Curve for Salary

(B) Histogram with Normal Curve using Normal Data

Figure 28. Histogram and Normal Curves.

Finally, in Figure 29, we see the statistical tests for normality. The extremely low significance level of 0.0000 indicates a very strong rejection of H_0. That is, we are fairly confident that the data does not come from a normally distributed population.



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Current Salary	.208	474	.000	.771	474	.000

a. Lilliefors Significance Correction

Figure 29. Summary of Tests of Normality for Salary.

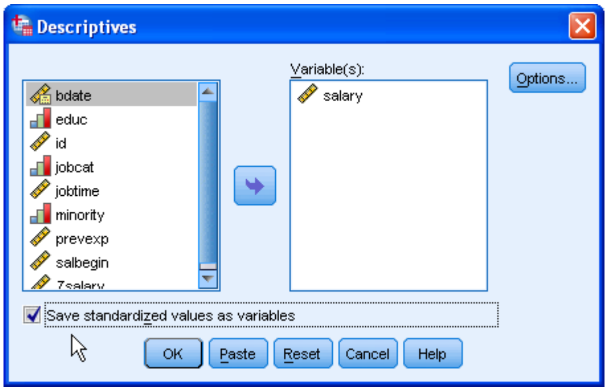
1.8. Standardized Scores z-Scores

Creating standardized scores is quick and painless using SPSS. Open any data file, such as "EmployeeData.sav." Select Analyze→Descriptive Statistics→Descriptives. Select a variable of interest, such as "Salary" and click the "Save Standardized Values as Variables" box, as shown in Figure 30. A new column of data will be created for each selected variable.

1.9. Scatter Plots and Regression

One of the first things you should do with a new data set is to look at it pictorially. If you have response and explanatory variables, then one simple graph is a scatter plot. Open the data set "customer_subset.sav". Using SPSS, select Graphs→Legacy Dialogs→Scatter/Dot. Select the "Simple Scatter" option in the Scatter/Dot dialog box and click "Define". For the y-axis, use the variable "carvalue" and for the x-axis use the variable "income". Click "OK“. The results can be found in Figure 31.

Notice that there does appear to be a linear trend in the data. So, perhaps we should perform linear regression. Linear regression is quickly performed by selecting Analyze→Regression→Linear. This opens



Variable(s): salary

☑ Save standardized values as variables

Figure 30. Saving Selected Variables as z-Scores.

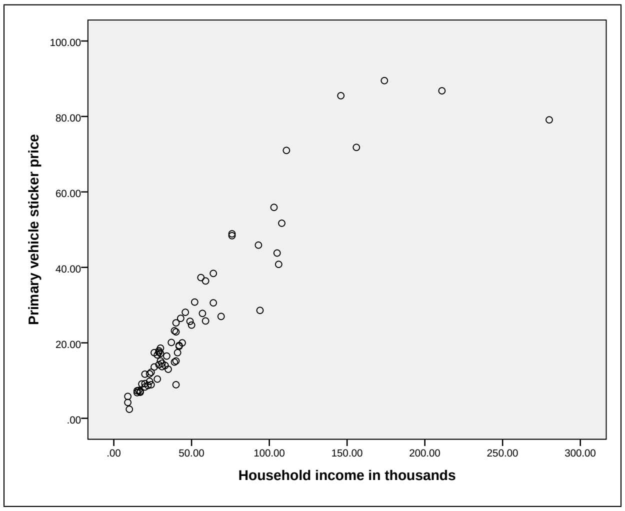


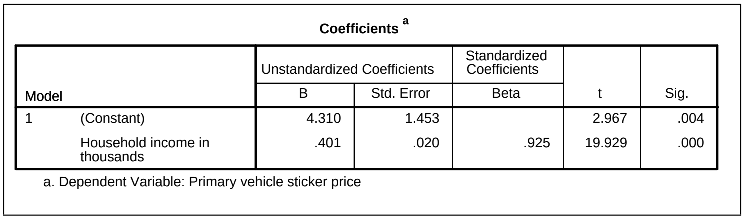
Figure 31. Scatter Plot of Primary Vehicle Sticker Price vs Household Income in Dollars.

the Linear Regression dialog box. Select "carvalue" for the Dependent (y-variable) and "income" for the Independent (x-variable). Clicking "OK" performs the requested regression. A lot of output is generated, but the more interesting part is shown in Figure 32. Notice that the significance level for the slope is 0.000 (which really means that it is really small...i.e. < 0.001 but probably not zero). Thus we reject the null hypothesis that the slope of the regression line is zero and conclude that linear regression does appear to be appropriate for our situation. The regression line is given by

$$\hat{y} = 0.401x + 4.310$$

so an increase of \$1,000 dollars in household income increases the sticker price by \$0.401 thousands, that is \$401.

Another useful bit of information is the R value, Pearson's Correlation Coefficient. Based on the information shown in Figure 33, we see that $R = 0.925$, which indicates a strong, positive, linear association between the two variables.

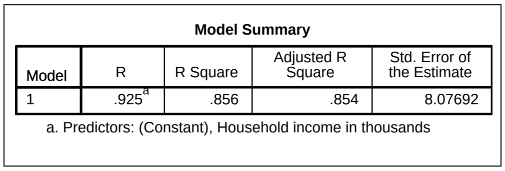


Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.310	1.453		2.967	.004
	Household income in thousands	.401	.020	.925	19.929	.000

a. Dependent Variable: Primary vehicle sticker price

Figure 32. Linear Regression of Primary Vehicle Sticker Price vs Household Income in Dollars.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925[a]	.856	.854	8.07692

a. Predictors: (Constant), Household income in thousands

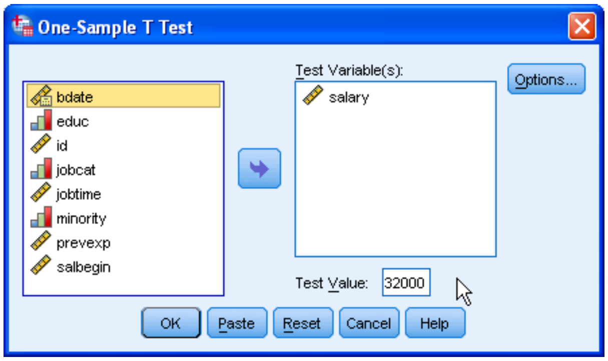
Figure 33. R from Linear Regression of Primary Vehicle Sticker Price vs Household Income in Dollars.

1.10. Testing Hypotheses about Population Means Using the t-Distribution

Open any data file, such as “EmployeeData.sav.” At a significance level of $\alpha = 5\%$, we wish to test the following hypothesis

$$H_0: \quad \text{The mean salary is \$32,000.}$$
$$H_1: \quad \text{The mean salary is greater than \$32,000.}$$

To perform the appropriate test, select Analyze → Compare Means → One-Sample t-test. Please enter the test value of $32,000 in the dialog box. Select a variable of interest, such as “Salary.” See 34. After clicking “OK” SPSS generates the output shown in Figure 36a.



One-Sample T Test

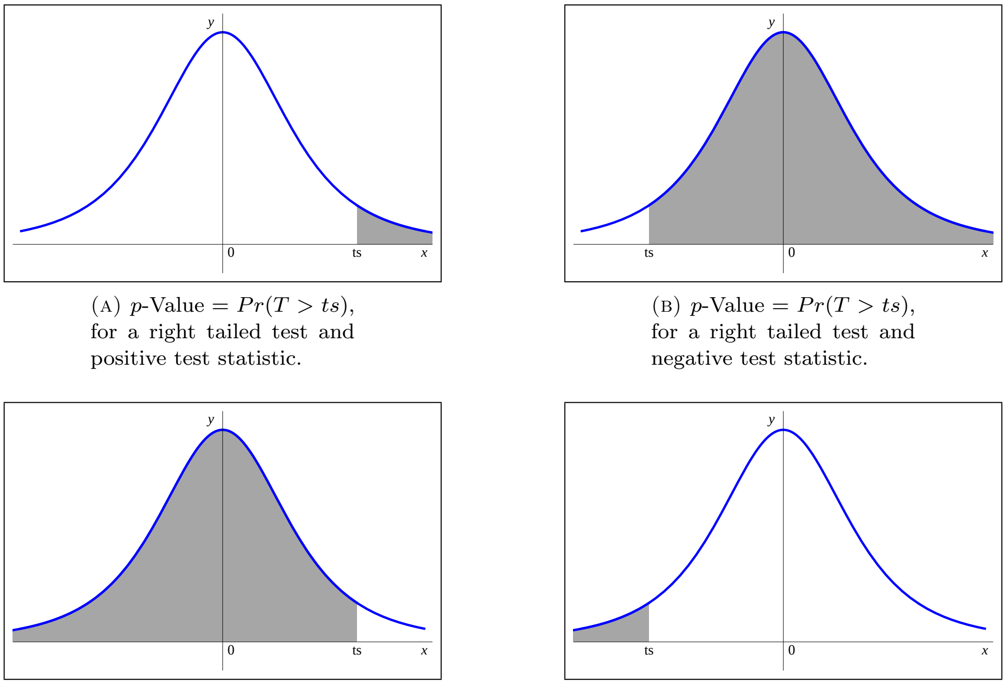
Test Variable(s): salary

Test Value: 32000

Figure 34. Dialog Box for the One Sample t-Test.

One unfortunate thing about SPSS is that it only performs 2-sided tests, and this is a right-tailed test. So, the significance level should be divided by two, as long as your test statistic is in the predicted direction, that is, as long as it is Positive, see Figure 35a. This indicates that your test statistic is in the same direction as the rejection region. If your test statistic is negative for a right-tailed test, then your p-value should be greater than 0.5, a very obvious failure to reject. In fact, it will be $1 - \frac{\text{Reported } p\text{-Value}}{2}$! See Figure 35b.

Of course, if this were a left tailed test, you could divide the p-value in half as long as your test statistic is Negative, as in Figure 35d. Again, for a left tailed test, a positive test statistic indicates your p-value is at least 0.5, and you would fail to reject. Once more, the actual p-Value will be $1 - \frac{\text{Reported } p\text{-Value}}{2}$! See Figure 35c.



(A) p-Value $= Pr(T > ts)$, for a right tailed test and positive test statistic.

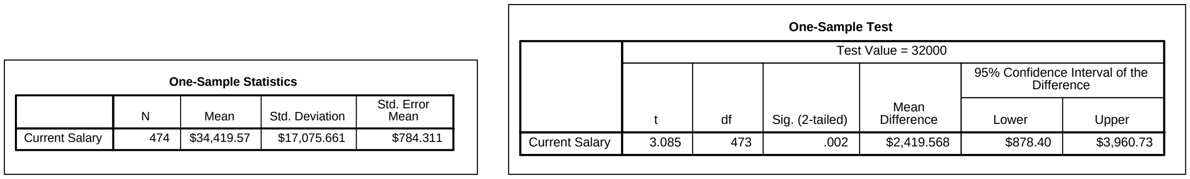
(B) p-Value $= Pr(T > ts)$, for a right tailed test and negative test statistic.

(C) p-Value $= Pr(T < ts)$, for a left tailed test and positive test statistic.

(D) p-Value $= Pr(T < ts)$, for a left tailed test and negative test statistic.

Figure 35. p-Value computation for left and right tailed tests using positive and negative test statistics.

Notice that in Figure 36a, we have some basic summary statistics, the sample size N, the sample mean and standard deviation, etc. In Figure 36b, we have the test statistic $t = 0.535$ and degrees of freedom, $df = 473$. The two-tailed significance level is 3.085. We have a positive test statistic of $ts = 3.085$, which lies in the direction of extreme, so the one-sided p-value is $\frac{0.002}{2} = 0.001$, which indicates we should reject the null hypothesis. We can conclude that the mean salary appears to be greater than \$32,000.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	474	$34,419.57	$17,075.661	$784.311

(A) One Sample Statistics

One-Sample Test

	Test Value = 32000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Current Salary	3.085	473	.002	$2,419.568	$878.40	$3,960.73

(B) One Sample Test

Figure 36. The one sample statistics and test.

1.11. χ^2 Goodness of Fit

Recall that the χ^2 Goodness of Fit test is a categorical variable test. So, we need a nice categorical experiment. Refer to the *Teaching Sociology* (July 2006) study of the fieldwork methods used by qualitative sociologists. Fieldwork methods can be categorized as follows

Table 1. Data for the Goodness of Fit

Fieldwork Method	Number of Papers
Interview	5,079
Observation + Participation	1,042
Observation Only	848
Grounded Theory	537

Suppose a sociologist claims that 70%, 15%, 10%, and 5% of the fieldwork methods involve interview, observation plus participation, observation only, and grounded theory, respectively. Does the data refute the claim with a significance level of 5%?

Notice that the hypotheses for this test are

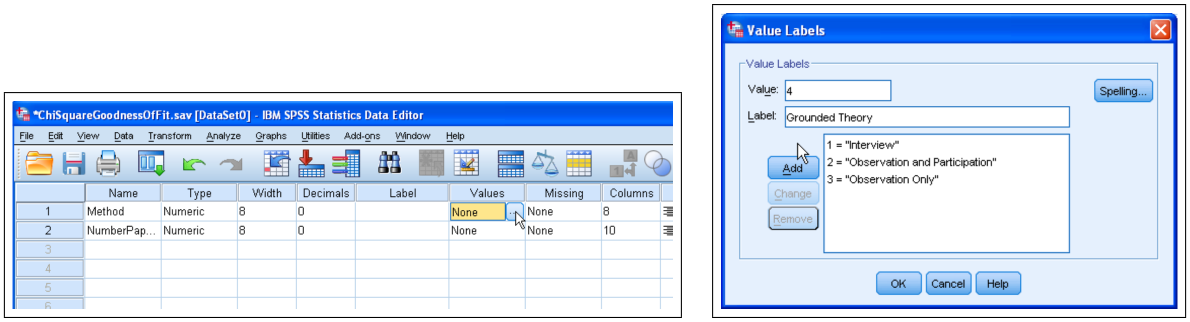
$$\begin{aligned} H_0: &\ \text{Interview } 70\%, \\ &\ \text{Observation and Participation } 15\%, \\ &\ \text{Observation Only } 10\%, \text{ and} \\ &\ \text{Grounded Theory } 5\% \\ H_a: &\ \text{The percentages are different.} \end{aligned}$$

For SPSS to successfully analyze this problem, enter the two variables "Method" and "NumberPapers" as integer variables (no decimals). In the "Method" row (we're still in the "Variable View" tab), click the cell in the "Values" column, as shown in Figure 37a. This opens the "Value Labels" dialog box, as seen in Figure 37b. We need to assign categorical names to the numeric values. Use the following values

Value	Label
1	Interview
2	Observation and Participation
3	Observation Only
4	Grounded Theory

As each pair is entered, use the "Add" button to record the assignment. When all four assignments are recorded, press "OK."

You should now enter the data. Swap to the "Data View" tab. Enter 1, 2, 3, and 4 in the Method column and for NumberPapers us the data as shown in Table 1.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	Method	Numeric	8	0		None	None	8
2	NumberPap...	Numeric	8	0		None	None	10
3								
4								
5								
6								

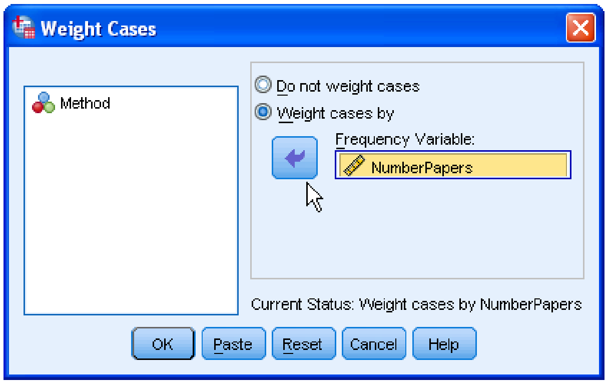
Value: 4
Label: Grounded Theory
1 = "Interview"
2 = "Observation and Participation"
3 = "Observation Only"

(A) Values Label Button

(B) Values Label Dialog Box.

FIGURE 37. Assigning labels to certain values in a numeric variable.

Since our data is a frequency table, we must tell SPSS how to weight the different cases; that is we should tell SPSS that NumberPapers records the frequency in each category listed by the variable Method. Use Data → Weight Cases to open the Weight Cases dialog. Make the changes as shown in Figure 38.



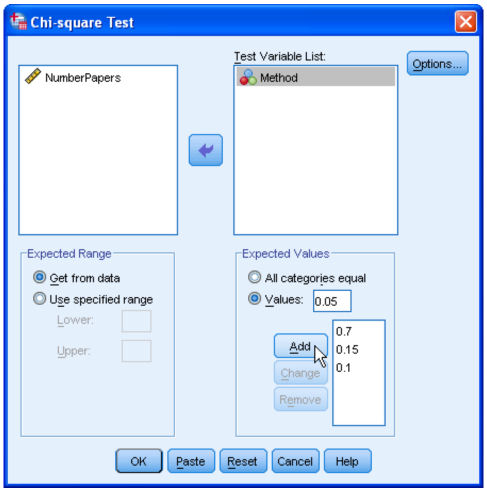
Method
○ Do not weight cases
● Weight cases by
Frequency Variable:
NumberPapers
Current Status: Weight cases by NumberPapers

FIGURE 38. Dialog Box when weighting cases.

Finally, let's perform the goodness of fit test. Select Analyze → Nonparametric Tests → Legacy Dialogs → Chi Square. This opens the "Chi-square Test" dialog box. Move the variable Method into the "Test Variable List." Under "Expected Values" select "Values." This allows you to enter the expected values: 0.70, 0.15, 0.10, and 0.05 (of course these numbers are from the percentages found in the null hypothesis!). Click "Add" after each entry. See Figure 39.

Pressing "OK" generates the following output shown in Figure 40. Notice that SPSS reports the observed, expected, and residual (Observed - Expected) and totals, as shown in Figure 40a. In the Test Statistics box, Figure 40b, the test statistic is reported, $\chi^2_{ts} = 94.02$, the degrees of freedom is $df = 3$, and the p-Value $= 0.000$ (my calculator reported a p-Value of 2.4821×10^{-20}, an extremely small number). Finally, in the Tests Statistics box, SPSS is kind enough to remind you that you should always have at least 5 of each category when performing this test.

Our conclusion should be reject the null hypothesis. In other words, the researcher is mistaken and the percentages are different.



Chi-square Test

NumberPapers

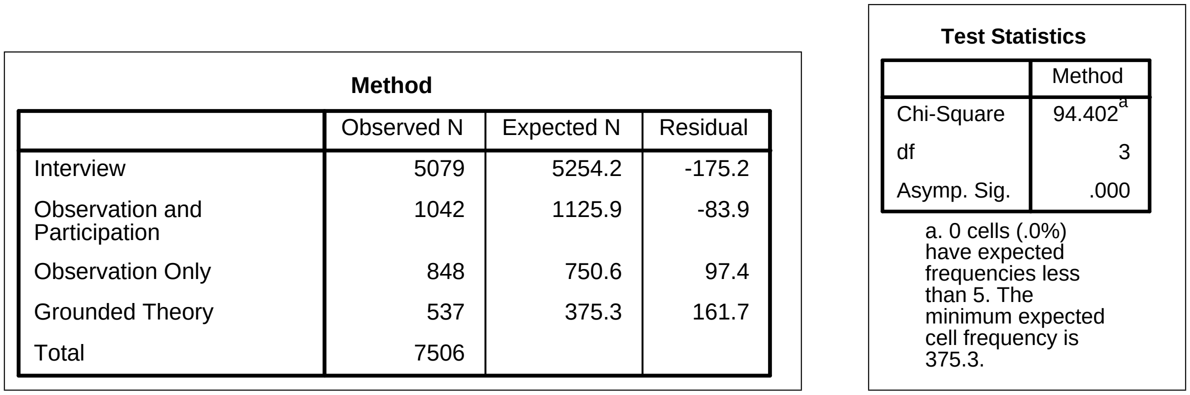
Test Variable List:
Method

Expected Range
● Get from data
○ Use specified range
Lower:
Upper:

Expected Values
○ All categories equal
● Values: 0.05

0.7
0.15
0.1

FIGURE 39. Dialog Box when setting up the χ^2 test for goodness of fit.



Method

	Observed N	Expected N	Residual
Interview	5079	5254.2	-175.2
Observation and Participation	1042	1125.9	-83.9
Observation Only	848	750.6	97.4
Grounded Theory	537	375.3	161.7
Total	7506		

(A) Observed and Expected Frequencies

Test Statistics

	Method
Chi-Square	94.402[a]
df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 375.3.

(B) Test Statistics

FIGURE 40. The results of a χ^2 goodness of fit test.

1.12. Contingency Tables or Cross Tabulations – Testing for Independence

A contingency table helps us look at whether the value of one variable is associated with, or "contingent" upon, that of another. It is most useful when each variable contains only a few categories. The hypotheses tested are

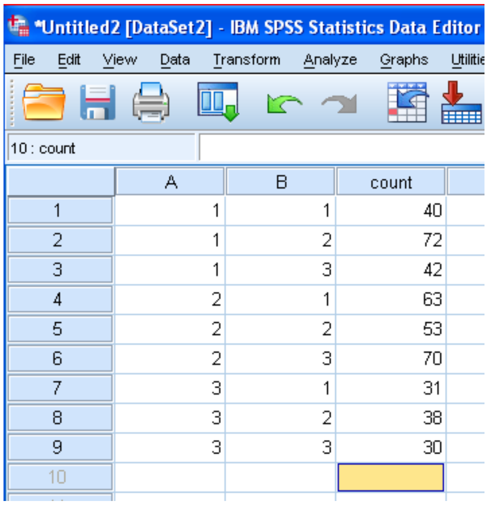
$$H_0: \quad \text{The variables are independent - i.e. have no relationship.}$$
$$H_a: \quad \text{The variables are dependent - i.e. have a relationship.}$$

Let's consider the following problem. Suppose you wish to test the null hypothesis of independence of the two classifications A and B of the 3×3 contingency table shown here. Test whether or not A and B are independent at $\alpha = 0.05$. Consider Table 2. To understand this frequency table, the level A_1 (i.e. row 1) and B_2 (i.e. column 2) has a frequency of 72. To enter this into SPSS, you need to have three variables, an A and a B for the different levels, and the number of entries, say count. Thus, to represent

TABLE 2. 3 × 3 Contingency Table

		B		
		B_1	B_2	B_3
A	A_1	40	72	42
	A_2	63	53	70
	A_3	31	38	30

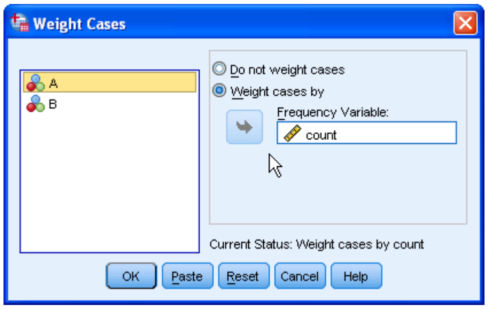
the entry of the first row and second column, enter 1 (in the A column), 2 (in the B column), and 72 (in the count column). Create these three integers in the "Variable View" and then swap to "Data View" to enter the data. See Figure 41 for the complete data set entered into SPSS.


*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilitie
10 : count

	A	B	count
1	1	1	40
2	1	2	72
3	1	3	42
4	2	1	63
5	2	2	53
6	2	3	70
7	3	1	31
8	3	2	38
9	3	3	30
10			

FIGURE 41. The frequency table as a data set in SPSS.

As before, you need to weight the data based upon the variable count. You do this via Data → Weight Cases. Select "Weight cases by" radio button and use the variable count for the "Frequency Variable." See Figure 42.

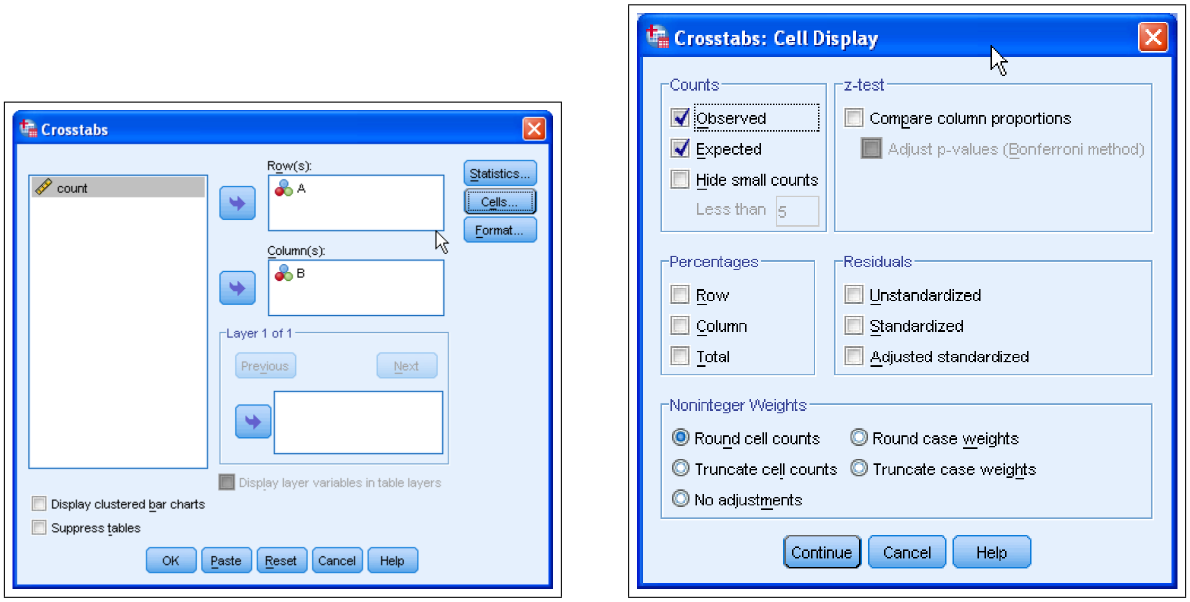


○ Do not weight cases
● Weight cases by
Frequency Variable: count

Current Status: Weight cases by count

FIGURE 42. The Weight Cases dialog box.

Next, we need to open the contingency table dialog box using Analyze → Descriptive Statistics → Crosstabs. Use A for your rows and B for your columns. Before clicking "OK", click the "Cells" button and make sure that "Observed" and "Expected" are both checked. Click "Continue" and "OK." See Figure 43 for these two dialog boxes.



Crosstabs

count

Row(s): A

Column(s): B

Layer 1 of 1

□ Display layer variables in table layers

□ Display clustered bar charts

□ Suppress tables

Crosstabs: Cell Display

Counts

☑ Observed

☑ Expected

□ Hide small counts

Less than 5

z-test

□ Compare column proportions

□ Adjust p-values (Bonferroni method)

Percentages

□ Row

□ Column

□ Total

Residuals

□ Unstandardized

□ Standardized

□ Adjusted standardized

Noninteger Weights

● Round cell counts ○ Round case weights

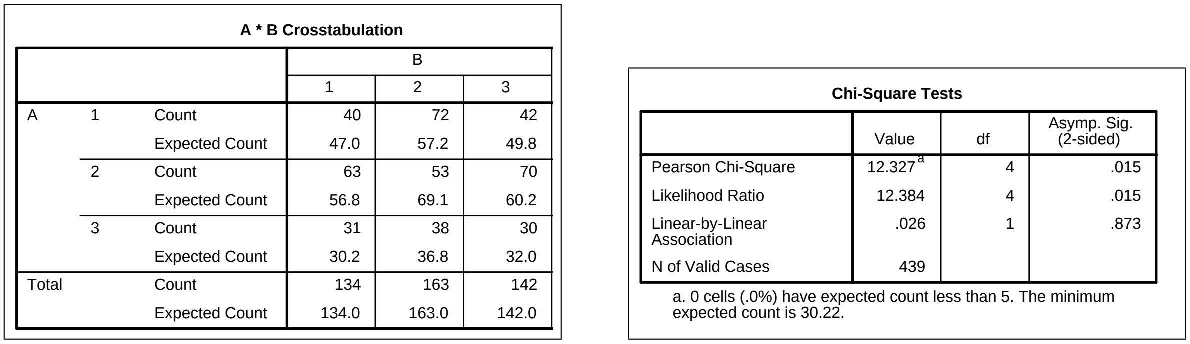
○ Truncate cell counts ○ Truncate case weights

○ No adjustments

(A) Basic Crosstabs Dialog Box (B) Crosstabs: Cell Display

FIGURE 43. Crosstabs Dialog Boxes.

This generates several tables. In Figure 44a, the count and expected count are shown. In Figure 44b several tests are listed. The one we're interested in is the Pearson χ^2-square test. Note that there are 4 degrees of freedom (remember $df = (\text{row} - 1)(\text{column} - 1)$), the test statistic is 12.327, and the p-value is 0.015. Thus, at $\alpha = 0.05$, we reject the null hypothesis and conclude that A and B are dependent.



A * B Crosstabulation

			B		
			1	2	3
A	1	Count	40	72	42
		Expected Count	47.0	57.2	49.8
	2	Count	63	53	70
		Expected Count	56.8	69.1	60.2
	3	Count	31	38	30
		Expected Count	30.2	36.8	32.0
Total		Count	134	163	142
		Expected Count	134.0	163.0	142.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12.327[a]	4	.015
Likelihood Ratio	12.384	4	.015
Linear-by-Linear Association	.026	1	.873
N of Valid Cases	439		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 30.22.

(A) Basic Crosstabs Dialog Box (B) Crosstabs: Cell Display

FIGURE 44. Crosstabs Dialog Boxes.

Table 3. Before and After Blood Pressure Levels

	SBP Level	
Id	No Drug	With Drug
1	115	128
2	112	115
3	107	106
4	119	128
5	115	122
6	138	145
7	126	132
8	105	109
9	104	102
10	115	117

1.13. Two-Sample Inference

1.13.1. The Paired t-Test. Recall that for the paired t-test, you need to have two matched or paired dependent populations in a "Before-After" situation. One important assumption is

- the data are chosen from a normal distribution.

Let μ_D be the mean difference between observations, then the hypotheses are

$$\begin{aligned} H_0 : &\quad \mu_D = 0 \\ H_a : &\quad \mu_D \neq 0 \end{aligned}$$

The test statistic is t with $n-1$ degrees of freedom, where n is the number of objects in the study. Since this is a two-tailed test, your p-value is

$$p - \text{Value} = Pr(t > |t_{ts}|),$$

where t_{ts} is the test statistic.

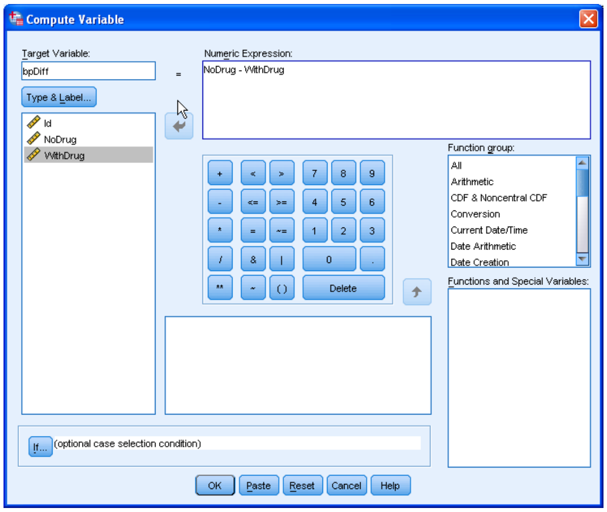
Suppose that you are studying blood-pressure levels (mm Hg) in 10 women while not using and while using a particular drug. The data are found in Table 3

Determine whether or not there is a difference in the men blood-pressure levels.

Enter the data into SPSS using variables "Id," "NoDrug," and "WithDrug." You can use integers since decimals are not present in any of the data.

First, we really should check to see if the differences in the data comes from a normal distribution. So, we should add a new variable to hold the differences. Instead of computing the differences by hand, we might as well use SPSS to do them for us. Click on Transform → Compute Variable. In the dialog box, type bpDiff in the "Target Variable" box. In the "Numeric Expression" box, type NoDrug - WithDrug. Notice that you can use the arrow to move defined variables into the "Numeric Expression" box...just supply the "$-$." See Figure 45

OK, now we have our new column. Test for normality. At a minimum, run a QQPlot and the Normality Tests. A Histogram is often nice to have as well, especially if you overlay a Normal Curve (double click the histogram to edit it). Select Analyze → Descriptives → Explore. Use "bpDiff" in the "Dependent List" box. Click the "Statistics" button and check "Histogram" and "Normality plots with tests." In Figures 46a and 46b are the histogram (with a superimposed normal curve) and the QQ Plot. Things appear a little disturbing towards the ends of the distribution, but not overly so. It does appear that we have a symmetric distribution. In Figure 47, we fail to reject the null hypothesis that the variable "bpDiff" follows a normal distribution. So, perhaps we can proceed.

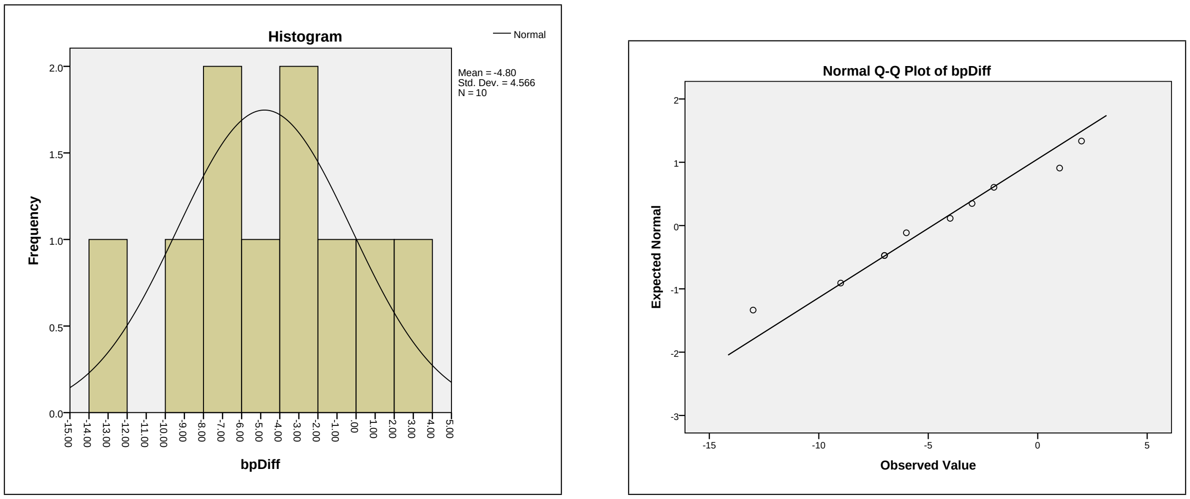


Compute Variable

Target Variable: bpDiff

Numeric Expression: NoDrug - WithDrug

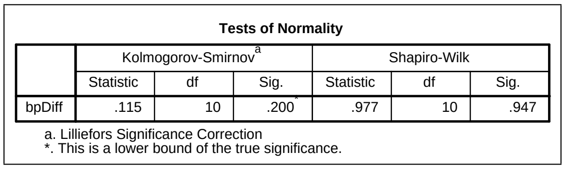
FIGURE 45. The Compute Variable dialog box.



(A) Histogram

(B) QQ Plot

FIGURE 46. A QQ Plot and a histogram.



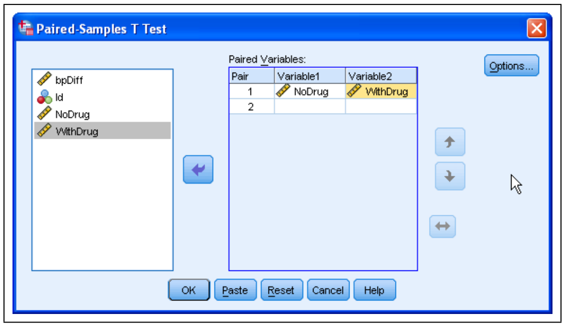
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
bpDiff	.115	10	.200*	.977	10	.947

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

FIGURE 47. Normality Tests.

Select Analyze → Compare Means → Paired-Samples T Test to open the "Paired-Samples T Test" dialog box. Select the variables "NoDrug" and "WithDrug" as can be seen in Fiugre 48.

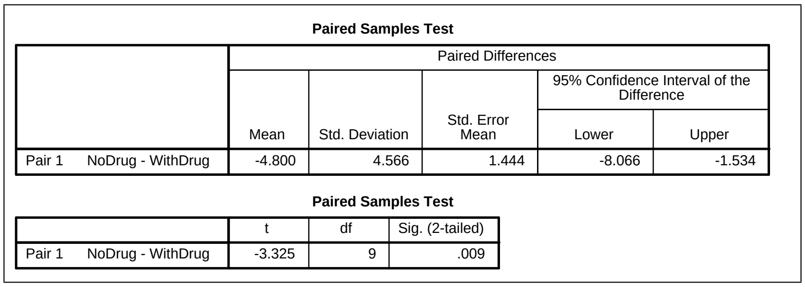


Paired Variables:

Pair	Variable1	Variable2
1	NoDrug	WithDrug
2		

FIGURE 48. Paired-Samples T Test Dialog Box.

The results from this are shown in Figure 49. The reported (two-tailed) p-Value is 0.009 with 9 degrees of freedom and a test statistic of -3.325. We also have some summary statistics such as the mean and standard deviation. Hence, based upon the p-value, we reject the null hypothesis and conclude that the medication appears to effect blood pressure.



Paired Samples Test

		Paired Differences				
					95% Confidence Interval of the Difference	
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper
Pair 1	NoDrug - WithDrug	-4.800	4.566	1.444	-8.066	-1.534

Paired Samples Test

		t	df	Sig. (2-tailed)
Pair 1	NoDrug - WithDrug	-3.325	9	.009

FIGURE 49. Results of the Paired T Test.

1.13.2. Independent Samples. Here we have two groups of data and wish to compare the means. The hypotheses are

Let μ_D be the mean difference between observations, then the hypotheses are

$$H_0: \quad \mu_D = 0$$
$$H_a: \quad \mu_D \neq 0$$

The test statistic is t. Since this is a two-tailed test, your p-value is

$$p\text{-Value} = Pr(t > |t_{ts}|),$$

where t_{ts} is the test statistic. Things are a bit complicated since the test changes, depending upon whether or not the variances are equal (or appear to be equal!). When the variances are equal, then it is a pooled test, when non equal, it is a non-pooled test. SPSS is kind enough to run Levene's test for you and then reports both pooled and non-pooled results. It is up to you to choose the correct results.

The main assumption (apart from independent variables) is

- both variables are normally distributed.

Table 4. Results of Vitamin C Study

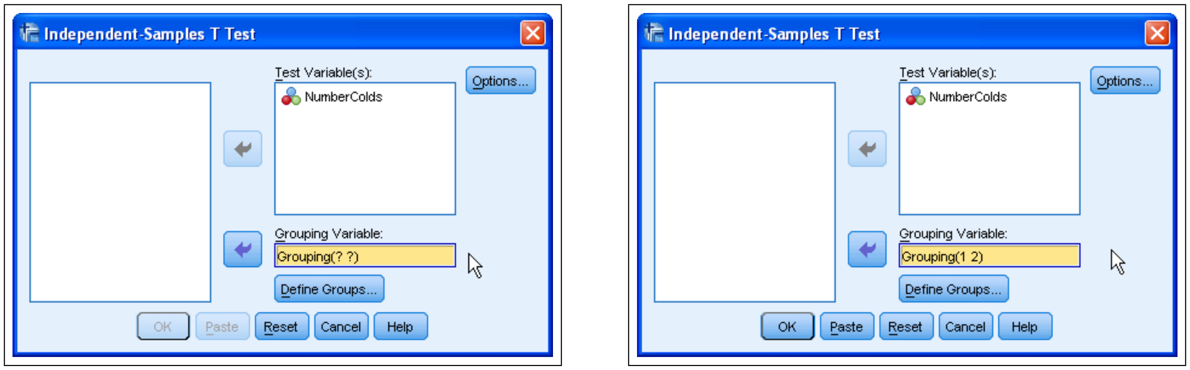
Vitamin C	4	0	3	4	4	3	4	3	2	6
hline Placebo	7	8	4	6	6	4	6	4	6	6

As an example, a study was conducted on the ability of Vitamin C to prevent colds. A total of twenty individuals were used in the study. Ten individuals were randomly selected to receive vitamin C and another ten were randomly selected to receive a placebo. The number of colds over a 12-month period were tracked and appear in Table 4. Test the hypothesis that vitamin C prevents the common cold.

Well, enter the data, using interesting names for variables such as “NumberColds” and a grouping variable, such as “Grouping.” Go ahead and label the variable “Grouping” according to the following key:

Vitamin C	1
Placebo	2

Remember, to do this, while in “Variable View” click o the “Values” cell for the varialbe “Grouping.” You can then enter the required labels.



Test Variable(s): NumberColds

Grouping Variable: Grouping(? ?)

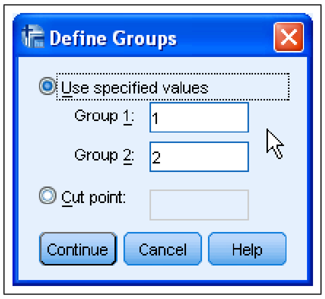
Test Variable(s): NumberColds

Grouping Variable: Grouping(1 2)

(A) Independent-Samples T Test before Grouping

(B) Independent-Samples T Test after Grouping

Figure 50. Independent-Samples T Test Dialog Box.



● Use specified values

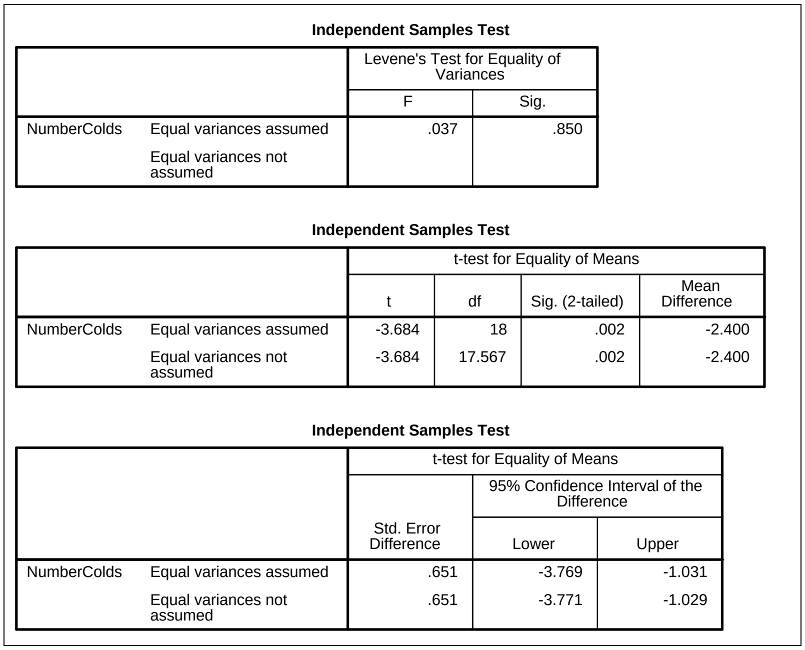
Group 1: 1

Group 2: 2

○ Cut point:

Figure 51. Define Groups Dialog Box

Select Analyze → Compare Means → Independent-Samples T Test. Move the variable "NumberColds" into the "Test Variable(s)" box and the variable "Grouping" into the "Grouping Variable" box. Notice, as in Figure 50a, SPSS lists "Grouping(? ?)" in the "Grouping Variable" box. This is because it is unsure how to group items. Click the "Define Groups" button and enter the information as shown in Figure 51. Now, notice that the "Independent-Samples T Test" now has the grouping variable listed as "Grouping(1 2)" as Figure 50b shows. Click OK to run the test.



Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
NumberColds	Equal variances assumed	.037	.850
	Equal variances not assumed		

Independent Samples Test

		t-test for Equality of Means			
		t	df	Sig. (2-tailed)	Mean Difference
NumberColds	Equal variances assumed	-3.684	18	.002	-2.400
	Equal variances not assumed	-3.684	17.567	.002	-2.400

Independent Samples Test

		t-test for Equality of Means		
			95% Confidence Interval of the Difference	
		Std. Error Difference	Lower	Upper
NumberColds	Equal variances assumed	.651	-3.769	-1.031
	Equal variances not assumed	.651	-3.771	-1.029

Figure 52. Results of the Independent T Test.

The results of the independent T Test is shown in Figure 52. First, Levene's Test for Equality of Variances indicates that the variances appear to be equal (The reported p-value of 0.850 is rather large!). Hence, we report information from the "Equal Variances Assumed" row of the table. The test statistic is -3.684, with 18 degree of freedom, and a significance (2-tailed p-value) of 0.002. Thus, we reject the null hypothesis and conclude that vitamin C appears to ward off the common cold.

1.14. One-Way ANOVA

One-way ANalysis Of VAriance (ANOVA) is used to determine whether there are significant differences between three or more population means. The hypotheses are

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \quad \text{At least two means are different.} \end{aligned}$$

Important assumptions include

(1) The dependent variable is normally distributed in each group. When this fails, either
  - transform the data so that the normality condition passes
  - use a non-parametric test such as the Kruskal-Wallis Test.

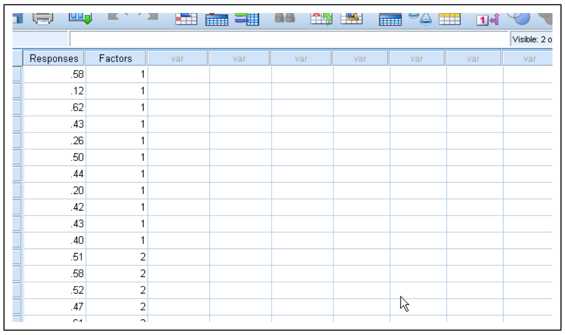
(2) There is homogeneity of variances, that the population variances are equal. When this fails, the Welch (or maybe the Brown-Forsythe) test is most likely the way to go.

TABLE 5. Can you sober up?

AR	AC	A	P
0.51	0.50	0.16	0.58
0.58	0.30	0.10	0.12
0.52	0.47	0.20	0.62
0.47	0.36	0.29	0.43
0.61	0.39	-0.14	0.26
0.00	0.22	0.18	0.50
0.32	0.20	-0.35	0.44
0.53	0.21	0.31	0.20
0.50	0.15	0.16	0.42
0.46	0.10	0.04	0.43
0.34	0.02	-0.25	0.40

(3) Independent variables (i.e. sample independently or randomly assign the experimental units to the treatments). When this fails, go talk to a statistician. You've got problems.

As an example, *Experimental and Clinical Psychopharmacology*, (Feb 2005) studied 44 male college students. The study concerned how stimulants (such as coffee or police about to ticket you) help you sober up after drinking. They were asked to memorize 40 words (20 on a green list and 20 on a red list). Students were randomly assigned to four groups, and students in three of the groups were given two alcoholic beverages to drink prior to performing a word completion task. Student in Group A received only the alcoholic drinks, student in group AC had caffeine powder dissolved in their drinks, and Group AR received two alcoholic drinks and a monetary award for correct responses. Group P (the placebo group) were told that they would receive alcohol, but received two drinks with carbonated beverage (no doubt they felt cheated). They all consumed their drinks and rested for 25 minutes before performing the word completion task. The reported score represents the difference between the proportion of correct responses on the green list and the proportion of incorrect responses on the red list. See Table 5. Test for equality of means under a significance level of 5%.



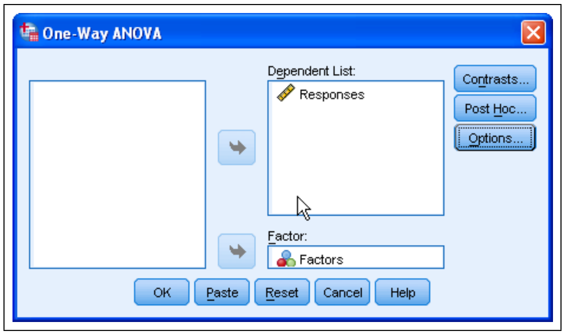
Responses	Factors	var	var	var	var	var	var	var
.58	1							
.12	1							
.62	1							
.43	1							
.26	1							
.50	1							
.44	1							
.20	1							
.42	1							
.43	1							
.40	1							
.51	2							
.58	2							
.52	2							
.47	2							

FIGURE 53. Entering Data for an ANOVA.

So, we have two variables, the categorical variable with values P, AR, AC, and A (which we will code using numbers 1, 2, 3, and 4) and the actual numeric results of the tests. Place the overall responses in a variable labeled "Responses." Create a new variable, "Factors," which is an integer and corresponds to the following table of categorical values.

1	2	3	4
P	AR	AC	A

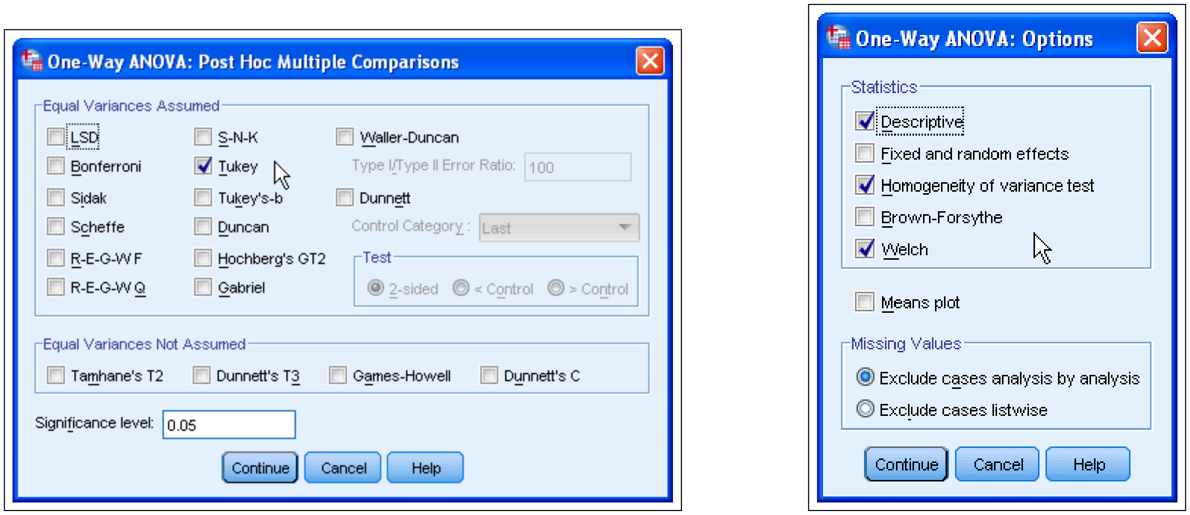
Remember, that you can create labels using the “Values” column (when in “Variable View”) which opens the “Value Labels” dialog box (enter 1 for a Value and P for a Label, etc.). So, enter one column of data, say P, in the “Responses” column and place a 1 in the “Factors” column. When you enter the next group of numbers in the “Responses” column, say AR, use a 2 in the “Factors” column. See Figure 53.



Dependent List: Responses

Factor: Factors

FIGURE 54. Entering the Factor and Dependent Variables for an ANOVA.



□ LSD
□ Bonferroni
□ Sidak
□ Scheffe
□ R-E-G-W F
□ R-E-G-W Q
□ S-N-K
☑ Tukey
□ Tukey's-b
□ Duncan
□ Hochberg's GT2
□ Gabriel
□ Waller-Duncan
Type I/Type II Error Ratio: 100
□ Dunnett
● 2-sided

□ Tamhane's T2
□ Dunnett's T3
□ Games-Howell
□ Dunnett's C

Significance level: 0.05

☑ Descriptive
□ Fixed and random effects
☑ Homogeneity of variance test
□ Brown-Forsythe
☑ Welch

□ Means plot

● Exclude cases analysis by analysis
○ Exclude cases listwise

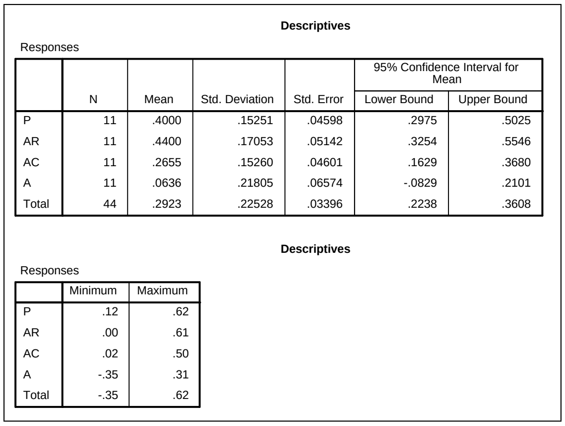
(A) One-Way ANOVA: Post Hoc Tests

(B) One-Way ANOVA: Options

FIGURE 55. Options for an ANOVA.

Now that the data is entered, select Analyze → Compare Means → One Way ANOVA. Enter “Factors” in the “Factor” box, and “Responses” into the “Dependent List.” Click the “Post Hoc” button and select the “Tukey” checkbox. Click “Continue” and then the “Options” button. Select the following check boxes: “Descriptive,” “Homogeneity of variance test,” and “Welch.” Click the “Continue” button and then the “OK” button. See Figures 54, 55a, and 55b, respectively.

First, let’s talk about the One-Way ANOVA options selected.



Descriptives

Responses

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
P	11	.4000	.15251	.04598	.2975	.5025
AR	11	.4400	.17053	.05142	.3254	.5546
AC	11	.2655	.15260	.04601	.1629	.3680
A	11	.0636	.21805	.06574	-.0829	.2101
Total	44	.2923	.22528	.03396	.2238	.3608

Descriptives

Responses

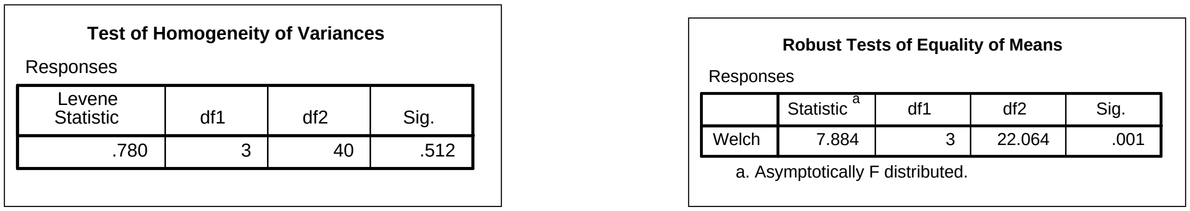
	Minimum	Maximum
P	.12	.62
AR	.00	.61
AC	.02	.50
A	-.35	.31
Total	-.35	.62

FIGURE 56. Descriptive Summaries from an ANOVA.

- Descriptive. This gives summary statistics such as Mean, Standard Deviation, etc. See Figure 56
- Homogeneity of Variances. This runs Levene's Test. The null hypothesis is

  $$\begin{aligned} H_0 :& \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_a :& \quad \text{At least two are different.} \end{aligned}$$

  We really would prefer not to reject the null hypothesis in this test. See Figure 57a for the results of this test. If, however, we do, then one of the assumptions for ANOVA fails and we should consider Welch's Test, shown in the Robust Tests of Equality of Means Table rather than the ANOVA Table.
- Welch. This runs a different type of test for equality of means and is reported in the Robust Tests of Equality of Means Table shown in Figure 57b Only consider this if the variances appear to be different, i.e. you reject H_0 in the Homogeneity of variances Test.



Test of Homogeneity of Variances

Responses

Levene Statistic	df1	df2	Sig.
.780	3	40	.512

(A) Homogeneity of Variances Table from an ANOVA.

Robust Tests of Equality of Means

Responses

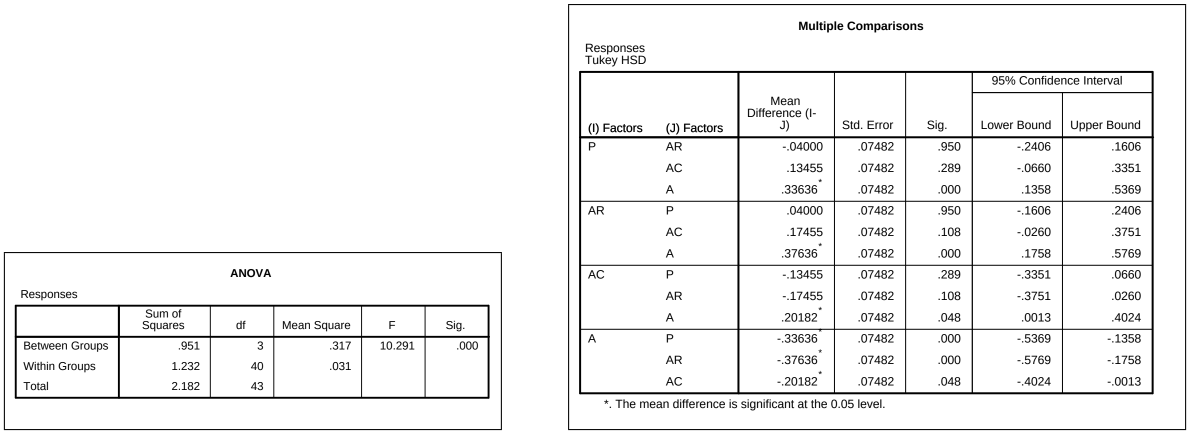
	Statistic[a]	df1	df2	Sig.
Welch	7.884	3	22.064	.001

a. Asymptotically F distributed.

(B) Robust Tests of Equality of Means Table from an ANOVA.

FIGURE 57. Testing for Homogeneity of Variances and the Robust Tests of Equality of Means.

Notice that we fail to reject the null hypothesis that the variances are different. Hence, Welch's Test (Robust Tests of Equality of Means) is not required. So, we skip directly to the ANOVA Table, as shown in Figure 58a. This indicates there is some sort of difference between the means, we just don't know which one is different.



ANOVA

Responses

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.951	3	.317	10.291	.000
Within Groups	1.232	40	.031		
Total	2.182	43			

(A) The ANOVA Table.

Multiple Comparisons

Responses
Tukey HSD

					95% Confidence Interval	
(I) Factors	(J) Factors	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
P	AR	-.04000	.07482	.950	-.2406	.1606
	AC	.13455	.07482	.289	-.0660	.3351
	A	.33636*	.07482	.000	.1358	.5369
AR	P	.04000	.07482	.950	-.1606	.2406
	AC	.17455	.07482	.108	-.0260	.3751
	A	.37636*	.07482	.000	.1758	.5769
AC	P	-.13455	.07482	.289	-.3351	.0660
	AR	-.17455	.07482	.108	-.3751	.0260
	A	.20182*	.07482	.048	.0013	.4024
A	P	-.33636*	.07482	.000	-.5369	-.1358
	AR	-.37636*	.07482	.000	-.5769	-.1758
	AC	-.20182*	.07482	.048	-.4024	-.0013

*. The mean difference is significant at the 0.05 level.

(B) Post Hoc Table from an ANOVA

FIGURE 58. Results After an ANOVA.

Post-hoc tests or *posteriori tests* are only considered if there is a difference detected between the means. These tests attempt to control the experiment wise error rate. If your data meet the assumption of homogeneity of variances, then either use Tukey's honestly significant difference (HSD) or Scheffe's Test. Note that Scheffe's test is more conservative (less likely to detect differences) than Tukey's HSD. If your data fail the assumption of homogeneity of variances, then try Games Howell or Dunnett's C. Again, from Figure 57a, we see that it appears that the homogeneity of variances is satisfied, so use Tukey's HSD. The resulting Post Hoc Tests Table in Figure 58b indicates that A is different from P, AR, and AC. However, none of the other combinations are significantly different.