Testing for Normality using SPSS

?

Introduction

An assessment of the normality of data is a prerequisite for many statistical tests as normal data is an underlying assumption in parametric testing. There are two main methods of assessing normality - graphically and numerically.

This guide will help you to determine whether your data is normal and, therefore, that this assumption is met in your data for statistical tests. The approaches can be divided into two main themes - relying on statistical tests or visual inspection. Statistical tests have the advantage of making an objective judgement of normality but are disadvantaged by sometimes not being sensitive enough at low sample sizes or overly sensitive to large sample sizes. As such, some statisticians prefer to use their experience to make a subjective judgement about the data from plots/graphs. Graphical interpretation has the advantage of allowing good judgement to assess normality in situations when numerical tests might be over or under sensitive but graphical methods do lack objectivity. If you do not have a great deal of experience interpreting normality graphically then it is probably best to rely on the numerical methods.

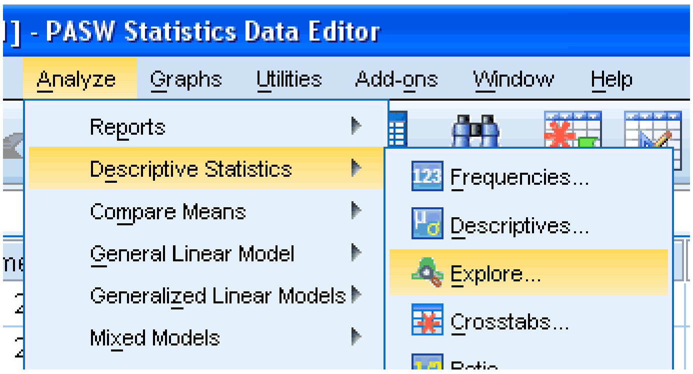
Methods of assessing normality

SPSS allows you to test all of these procedures within **Explore...** command. The**Explore...** command can be used in isolation if you are testing normality in one group or splitting your dataset into one or more groups. For example, if you have a group of participants and you need to know if their height is normally distributed then everything can be done within the **Explore...** command. If you split your group into males and females (i.e. you have a categorical independent variable) then you can test for normality of height within both the male group and the female group using just the **Explore...**command. This applies even if you have more than two groups. However, if you have 2 or more categorical, independent variables then the **Explore...** command on its own is not enough and you will have to use the **Split File...** command also.

Procedure for none or one grouping variable

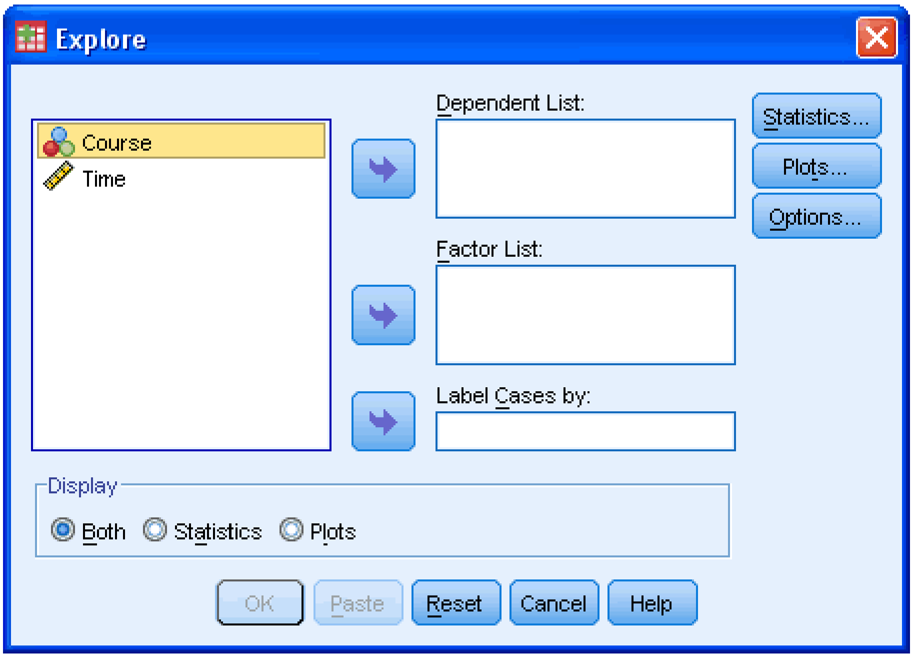
The following example comes from our guide on how to perform a one-way ANOVA in SPSS.

1. Click **Analyze > Descriptive Statistics > Explore...** on the top menu as shown below:



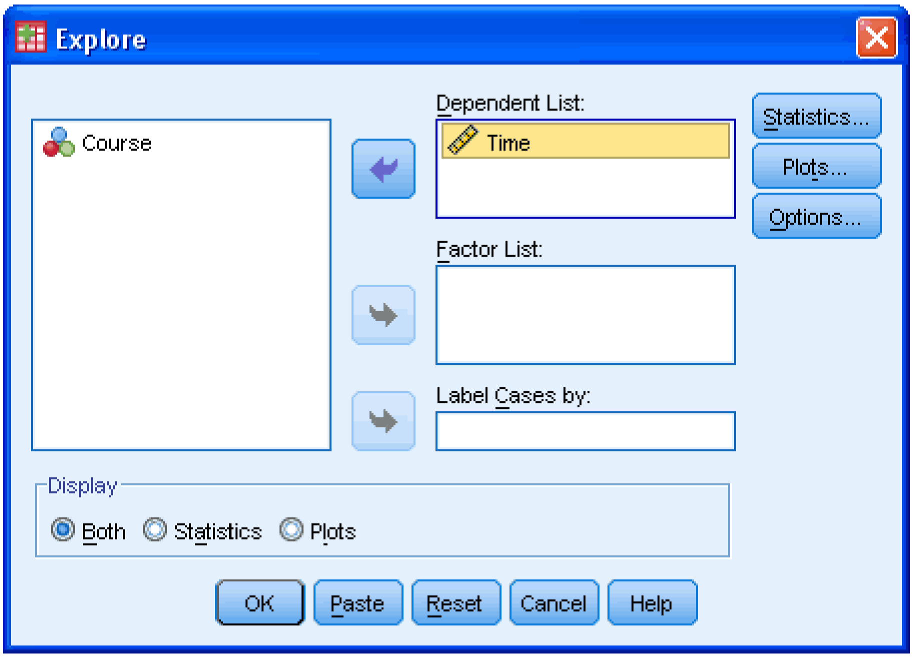
Published with written permission from SPSS Inc, an IBM Company.

2. You will be presented with the following screen:



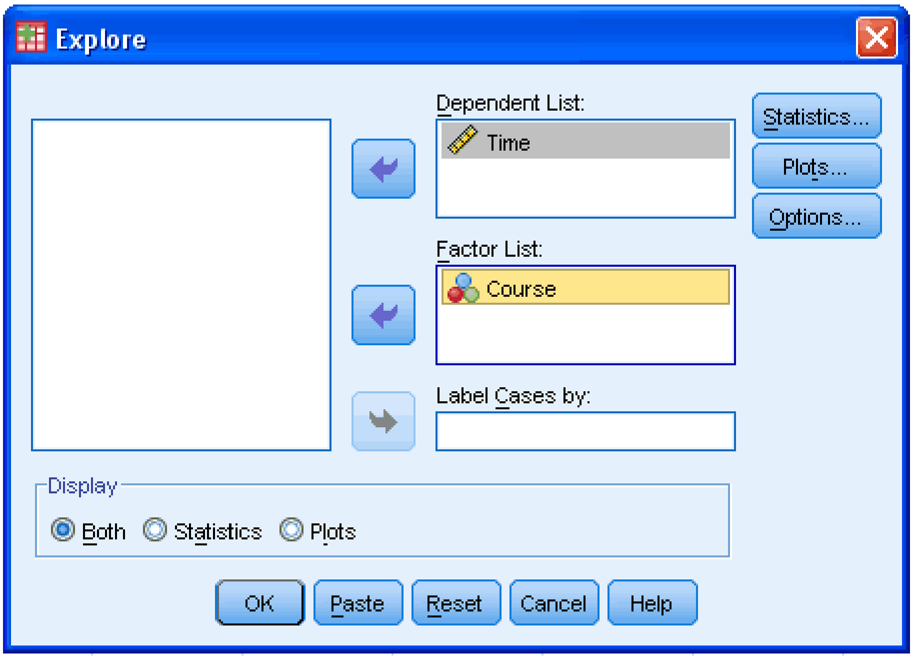
Published with written permission from SPSS Inc, an IBM Company.

3. Transfer the variable that needs to be tested for normality into the "Dependent List:" box by either drag-and-dropping or using the  button. In this example, we transfer the "Time" variable into the "Dependent List:" box. You will then be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

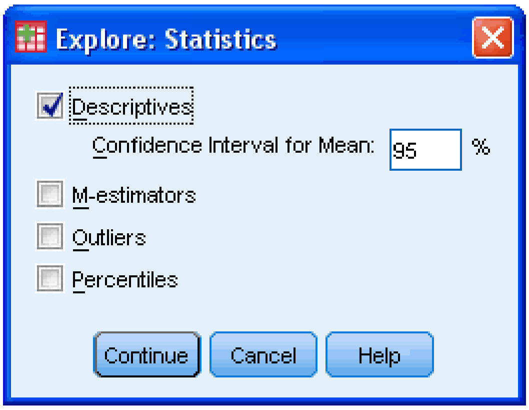
4. [Optional] If you need to establish if your variable is normally distributed for each level of your independent variable then you need to add your independent variable to the "Factor List:" box by either drag-and-dropping or using the  button. In this example, we transfer the "Course" variable into the "Factor List:" box. You will be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

5. Click the 

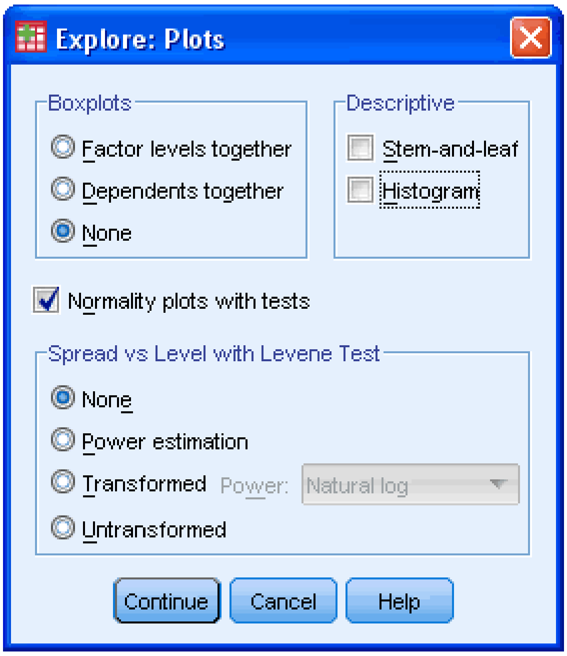
button. You will be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

Leave the above options unchanged and click the Continue button.

6. Click the Plots... button. Change the options so that you are presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

Click the Continue button.

7. Click the OK button.

Output

SPSS outputs many table and graphs with this procedure. One of the reasons for this is that the **Explore...** command is not used solely for the testing of normality but in describing data in many different ways. When testing for normality, we are mainly interested in the **Tests of Normality** table and the **Normal Q-Q Plots**, our numerical and graphical methods to test for the normality of data, respectively.

Shapiro-Wilk Test of Normality

Tests of Normality

	Course	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Time	Beginner	.177	10	.200*	.964	10	.827
	Intermediate	.166	10	.200*	.969	10	.882
	Advanced	.151	10	.200*	.965	10	.837

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

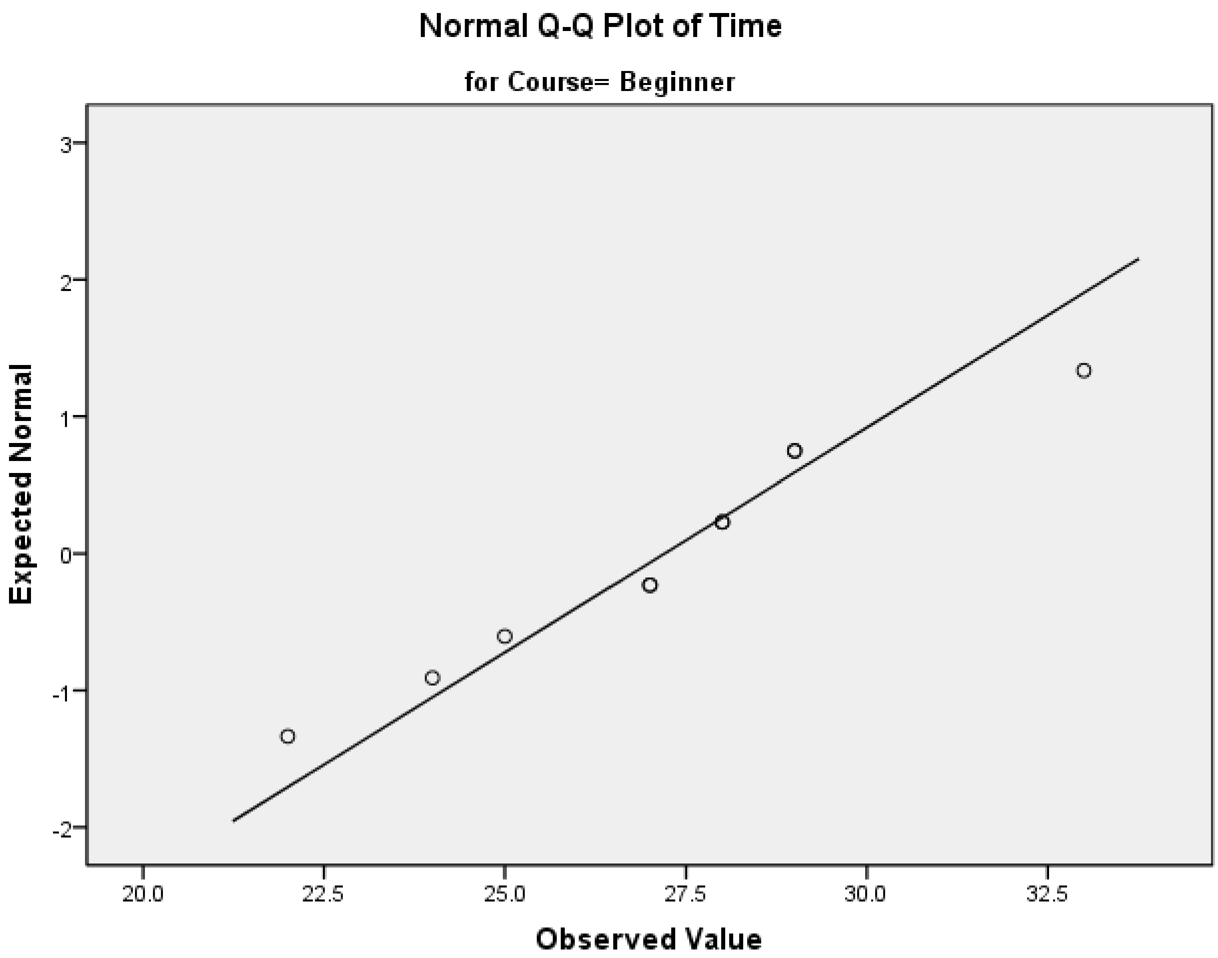
The above table presents the results from two well-known tests of normality, namely the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test. We Shapiro-Wilk Test is more appropriate for small sample sizes (< 50 samples) but can also handle sample sizes as large as 2000. For this reason, we will use the Shapiro-Wilk test as our numerical means of assessing normality.

We can see from the above table that for the "Beginner", "Intermediate" and "Advanced" Course Group the dependent variable, "Time", was normally distributed. How do we know this? If the **Sig.** value of the Shapiro-Wilk Test is greater the 0.05 then the data is normal. If it is below 0.05 then the data significantly deviate from a normal distribution.

If you need to use skewness and kurtosis values to determine normality, rather the Shapiro-Wilk test, you will find these in our upgraded Premium SPSS guide. Check out our low prices here.

Normal Q-Q Plot

In order to determine normality graphically we can use the output of a normal Q-Q Plot. If the data are normally distributed then the data points will be close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion then the data are not normally distributed. As we can see from the normal Q-Q plot below the data is normally distributed. If you at all unsure of being able to correctly interpret the graph then rely on the numerical methods instead as it can take a fair bit of experience to correctly judge the normality of data based on plots.



If you need to know what Normal Q-Q Plots look like when distributions are not normal (e.g. negatively skewed), you will find these in our upgraded Premium SPSS guide. Check out our low prices here.

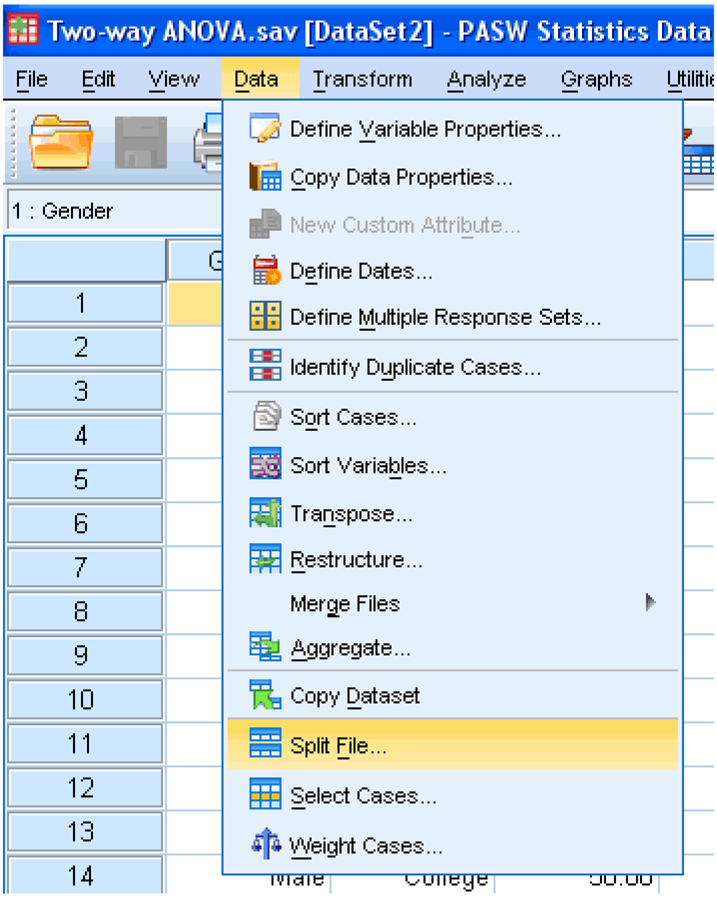
Testing for Normality using SPSS (cont...)

12

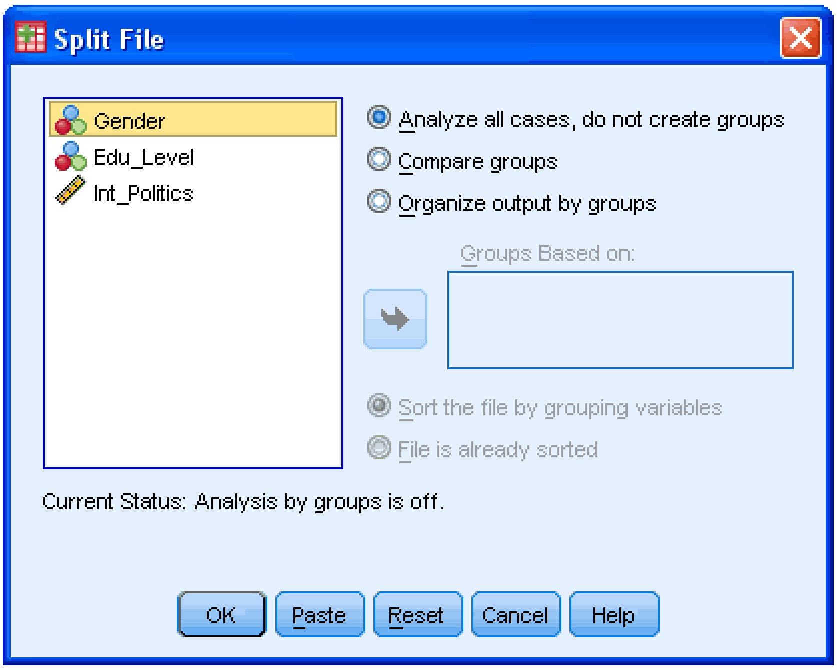
Procedure when there are two or more independent variables

The **Explore...** command on its own cannot separate the dependent variable into groups based on not one but two or more independent variables. However, we can perform this feat by using the **Split File...** command.

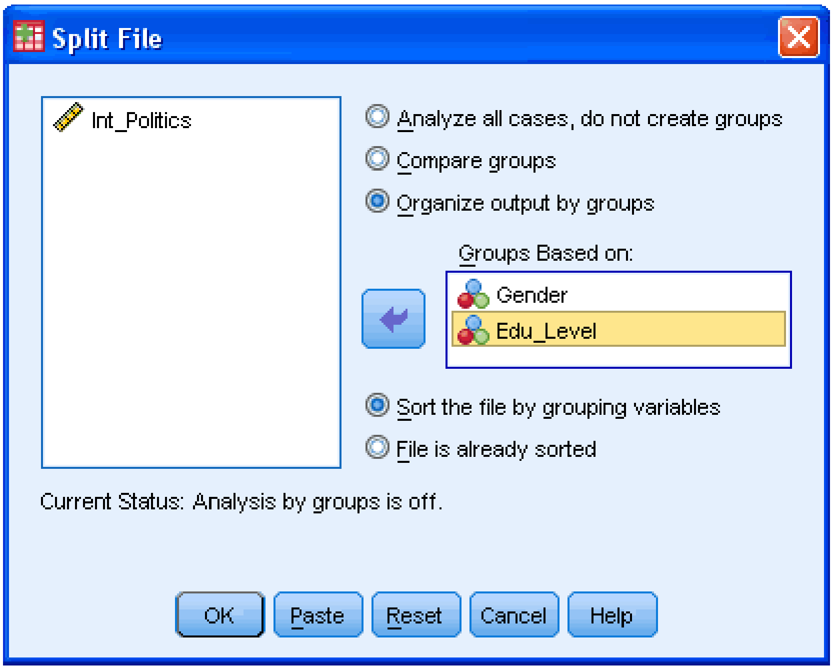
1. Click **Data > Split File...** on the top menu as shown below:



2. You will be presented with the following screen:



3. Click the radio option, "Organize output by groups". Transfer the independent variables you wish to categorize the dependent variable on into the "Groups Based on:". In this example, we want to know whether interest in politics (Int_Politics) is normally distributed when grouped/categorized by Gender AND Edu_Level (education level). You will be presented with the following screen:

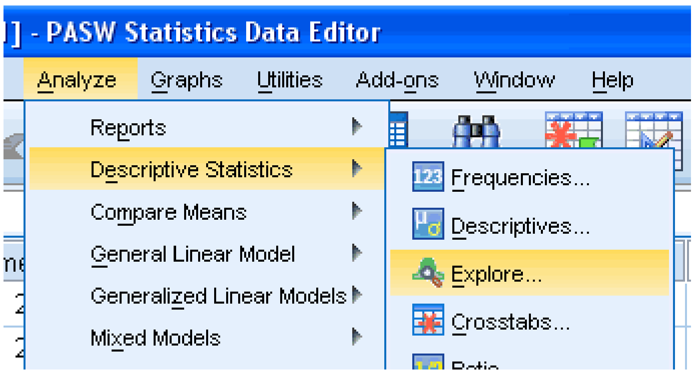


Published with written permission from SPSS Inc, an IBM Company.

Click the  button.

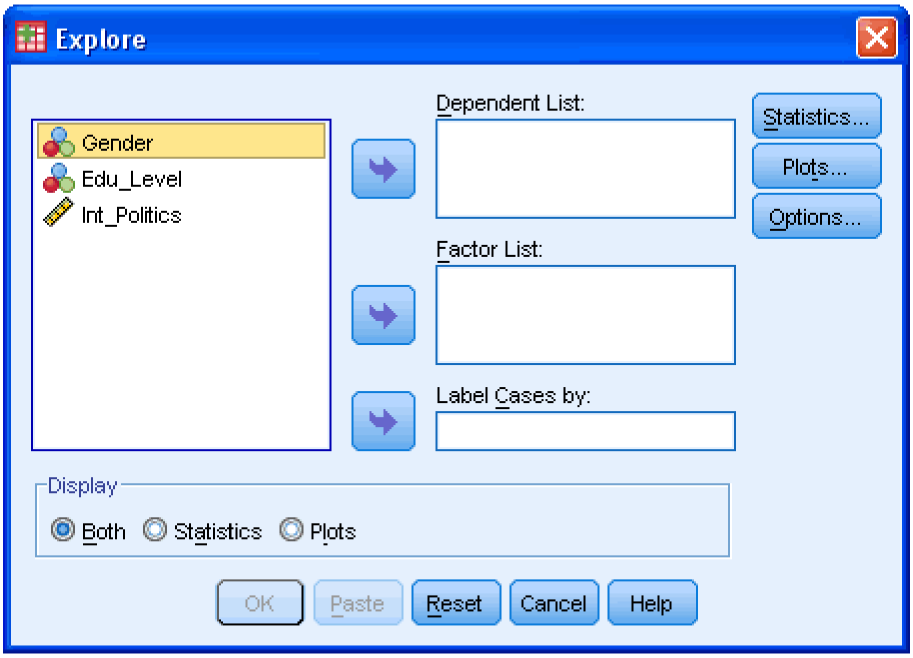
[Your file is now split and the output from any tests will be organized into the groups you have selected.]

4. Click **Analyze > Descriptive Statistics > Explore...** on the top menu as shown below:



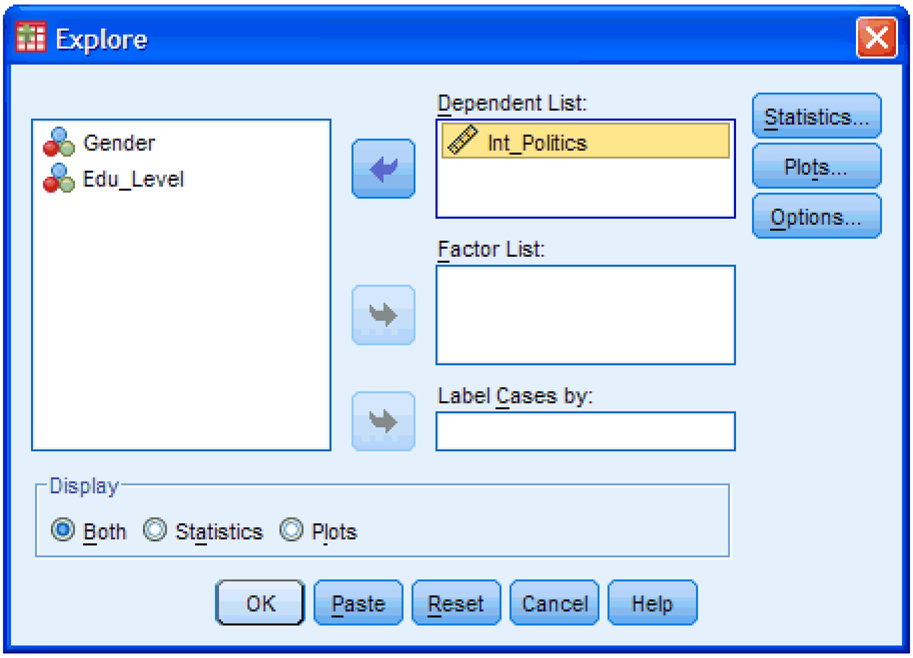
Published with written permission from SPSS Inc, an IBM Company.

5. You will be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

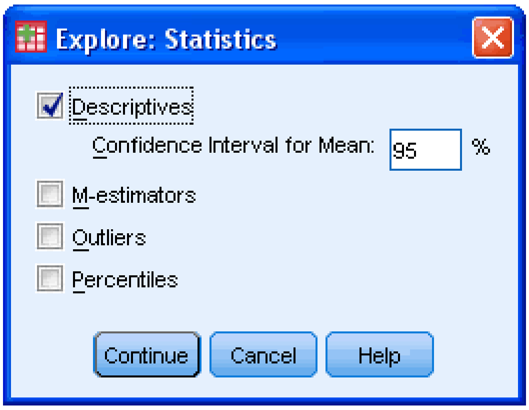
6. Transfer the variable that needs to be tested for normality into the "Dependent List:" box by either drag-and-dropping or using the button. In this example, we transfer the "Int_Politics" variable into the "Dependent List:" box. You will then be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

[*There is no need to transfer the independent variables "Gender" and "Edu_Level" into the "Factor List:" box as this has been accomplished with the* ***Split File...****command. Why not simply transfer these two independent variables into the "Factor List:" box? Because this will not achieve the desired result. It will first* *analyse "Int_Politics" for normality with respect to "Gender" and then with respect to "Edu_Level". It does NOT analyse "Int_Politics" for normality by grouping individuals into both "Gender" and "Edu_Level" AT THE SAME TIME.*]

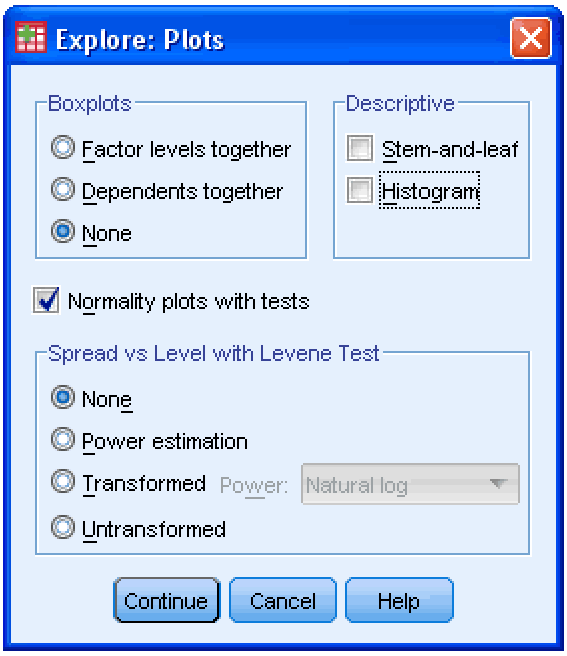
7. Click the Statistics... button. You will be presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

Leave the above options unchanged and click the Continue button.

8. Click the Plots... button. Change the options so that you are presented with the following screen:



Published with written permission from SPSS Inc, an IBM Company.

Click the Continue button.

9. Click the OK button.

Output

You will now see that the output has been split into separate sections based on the combination of groups of the two independent variables. As an example we show the tests of normality when the dependent variable, "**Int_Politics**", is categorized into the first "**Gender**" group (male) and first "**Edu_Level**" group (School). All other possible combinations are also presented in the full output but we will not shown them here for clarity.

Gender = Male, Edu_Level = School

Under this above category you are presented with the **Tests of Normality** table as shown below:

Tests of Normality[b]

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Int_Politics	.178	10	.200*	.944	10	.599

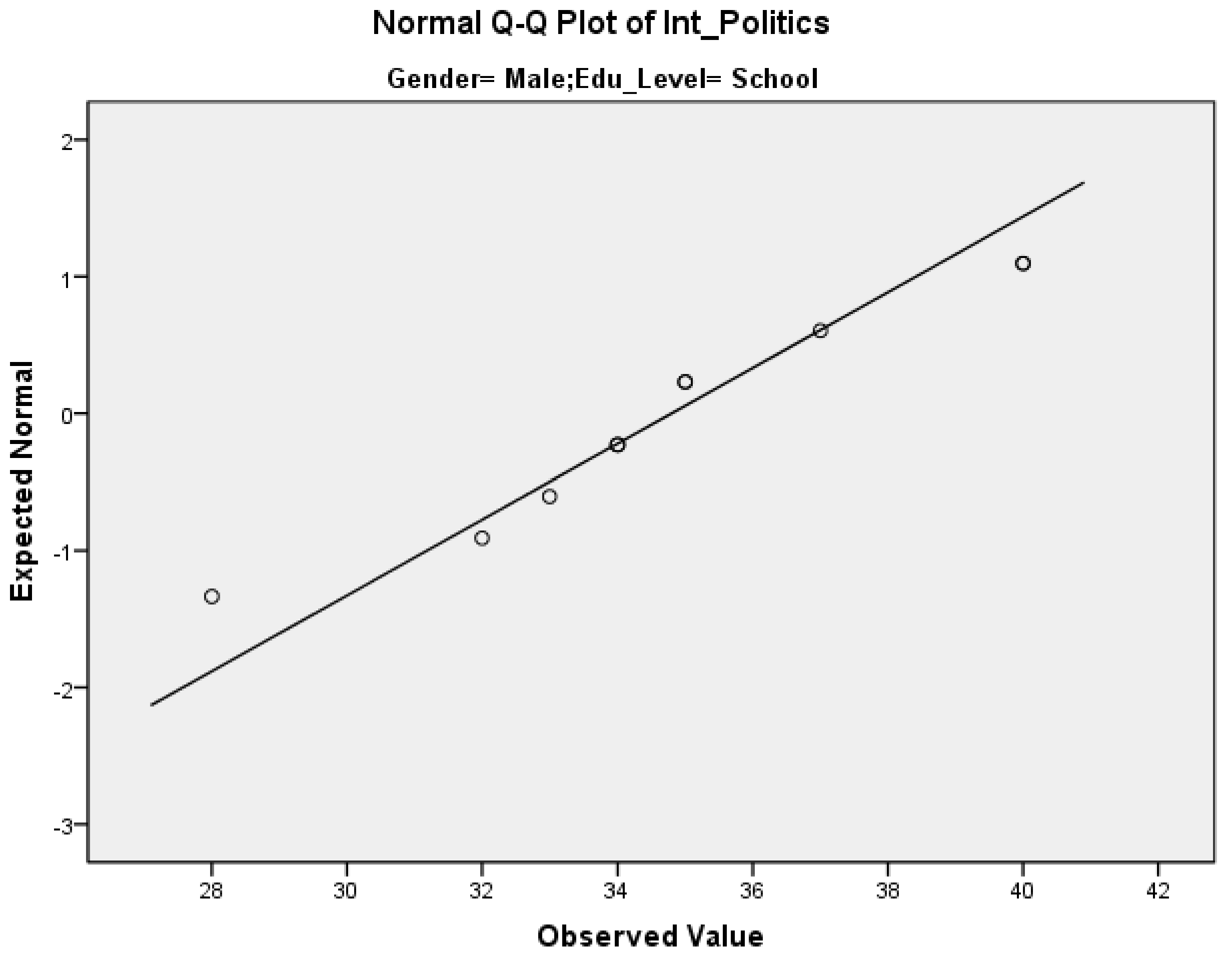
a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

b. Gender = Male, Edu_Level = School

The Shapiro-Wilk test is now analyzing the normality of "Int_Politics" on the data of those individuals that are classified as both "male" in the independent variable "Gender" and "school" in the independent variable "Edu_Level". As the **Sig.** value under the Shapiro-Wilk column is greater than 0.05 we can conclude that "Int_Politics" for this particular subset of individuals is normally distributed.

The same data from the same individuals are now also being analyzed to produce a Normal Q-Q Plot as below. From this graph we can conclude that the data appears to be normally distributed as it follows the diagonal line closely and does not appear to have a non-linear pattern.


Normal Q-Q Plot of Int_Politics
Gender= Male;Edu_Level= School
Expected Normal
Observed Value
2
1
0
-1
-2
-3
28
30
32
34
36
38
40
42

One-way ANOVA using SPSS

91

Objectives

The one-way analysis of variance (ANOVA) is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups. This guide will provide a brief introduction to the one-way ANOVA including the assumptions of the test and when you should use interpret the output. This guide will then go through the procedure for running this test in SPSS using an appropriate example, which options to choose and how to interpret the output. Should you wish to learn more about this test before doing the procedure in SPSS, please click here.

What does this test do?

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are significantly different from each other. Specifically, it tests the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a significant result then we accept the alternative hypothesis (H_A), which is that there are at least 2 group means that are significantly different from each other.

At this point, it is important to realise that the one-way ANOVA is an *omnibus* test statistic and cannot tell you which specific groups were significantly different from each other, only that at least two groups were. To determine which specific groups differed from each other you need to use a *post-hoc* test. *Post-hoc* tests are described later in this guide.

Assumptions

- Independent variable consists of **two or more categorical independent groups**.
- Dependent variable is either **interval or ratio (continuous)** (see our guide onTypes of Variable).
- Dependent variable is **approximately normally distributed for each category of the independent variable** (see our guide on Testing for Normality).
- Equality of variances between the independent groups (**homogeneity of variances**).
- Independence of cases.

Example

A manager wants to raise the productivity at his company by increasing the speed at which his employees can use a particular spreadsheet program. As he does not have the skills in-house, he employs an external agency which provides training in this spreadsheet program. They offer 3 packages - a beginner, intermediate and advanced course. He is unsure which course is needed for the type of work they do at his company so he sends 10 employees on the beginner course, 10 on the intermediate and 10 on the advanced course. When they all return from the training he gives them a problem to solve using the spreadsheet program and times how long it takes them to complete the problem. He wishes to then compare the three courses (beginner, intermediate, advanced) to see if there are any differences in the average time it took to complete the problem.

Setup in SPSS

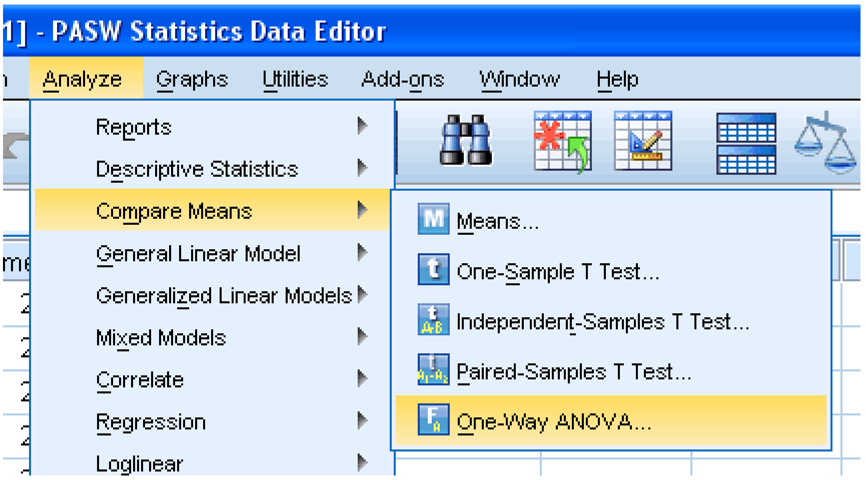
In SPSS we separated the groups for analysis by creating a grouping variable called "**Course**" and gave the beginners course a value of "**1**", the intermediate course a value of "**2**" and the advanced course a value of "**3**". Time to complete the set problem was entered under the variable name "**Time**". To know how to correctly enter your data into SPSS in order to run a repeated measures ANOVA please read our Entering Data in SPSStutorial.

Testing assumptions

See how to test the normality assumption for this test in our Testing for Normality guide.

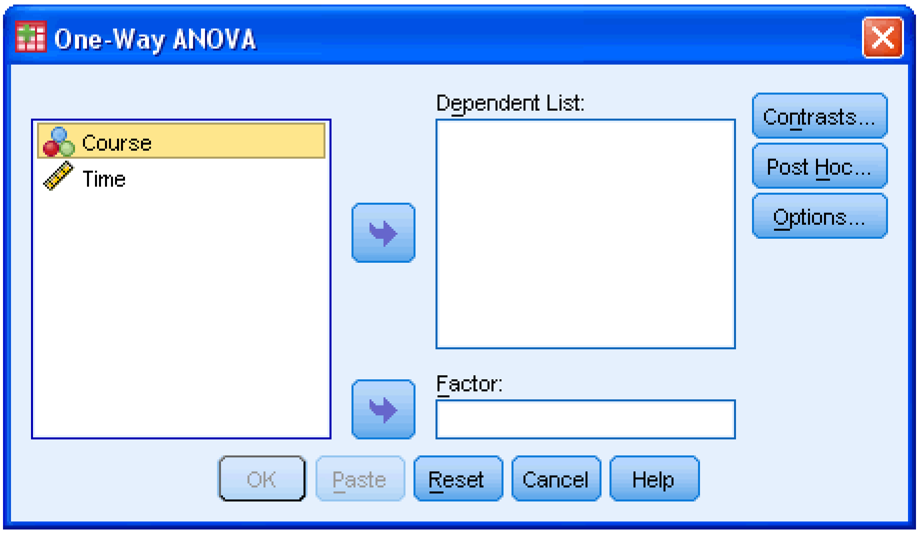
Test Procedure in SPSS

1. Click **Analyze > Compare Means > One-Way ANOVA...** on the top menu as shown below.



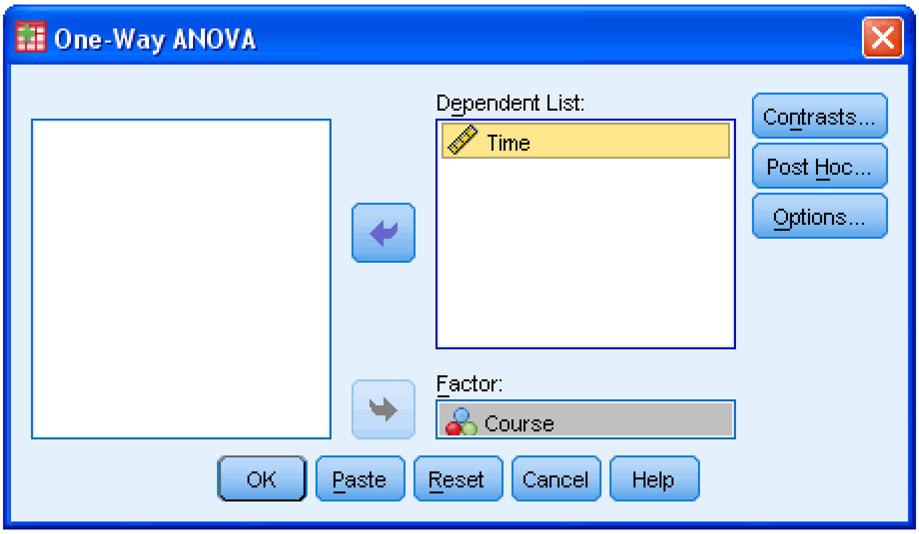
Published with written permission from SPSS Inc, an IBM Company.

2. You will be presented with the following screen:



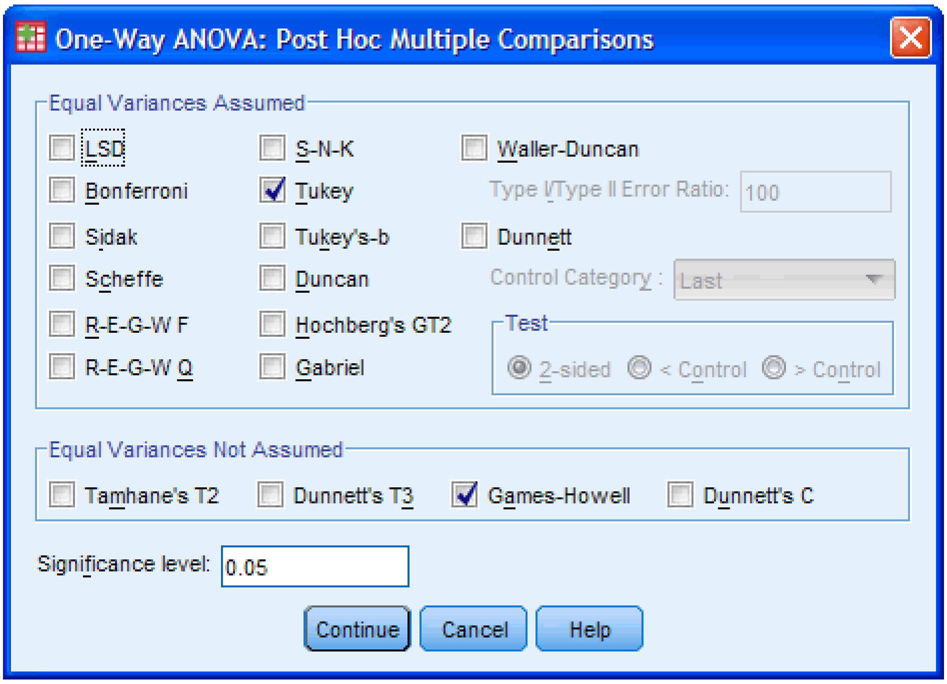
Published with written permission from SPSS Inc, an IBM Company.

3. Drag-and-drop (or use the  buttons) to transfer the dependent variable (Time) into the *Dependent List:* box and the independent variable (**Course**) into the*Factor:* box as indicted in the diagram below:



Published with written permission from SPSS Inc, an IBM Company.

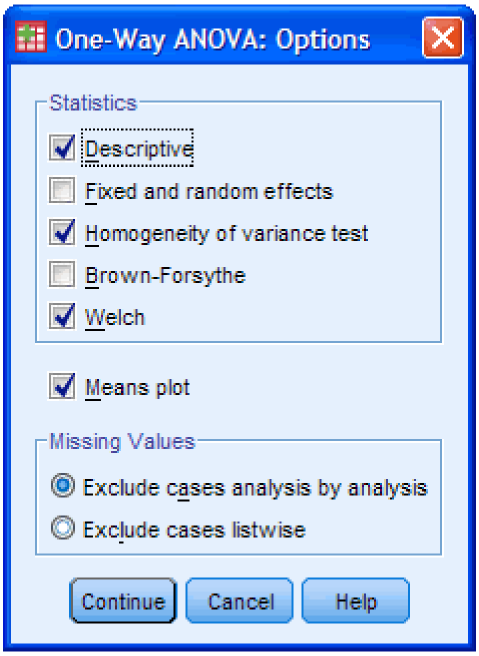
4. Click the Post Hoc... button. Tick the "Tukey" checkbox as shown below:



Published with written permission from SPSS Inc, an IBM Company.

Click the Continue button.

5. Click the Options... button. Tick the "Descriptive", "Homogeneity of variance test", "Brown-Forsythe", and "Welch" checkboxes in the *Statistics* area as shown below:



Published with written permission from SPSS Inc, an IBM Company.

Click the Continue button.

6. Click the OK button.

Go to the next page for the SPSS output and an explanation of the output.

SPSS Output of the one-way ANOVA

SPSS generates quite a few tables in its one-way ANOVA analysis. We will go through each table in turn.

Descriptives Table

The descriptives table (see below) provides some very useful descriptive statistics including the mean, standard deviation and 95% confidence intervals for the dependent variable (Time) for each separate group (Beginners, Intermediate & Advanced) as well as when all groups are combined (Total). These figures are useful when you need to describe your data.

Descriptives

Time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Beginner	10	27.2000	3.04777	.96379	25.0198	29.3802	22.00	33.00
Intermediate	10	23.6000	3.30656	1.04563	21.2346	25.9654	18.00	29.00
Advanced	10	23.4000	3.23866	1.02415	21.0832	25.7168	18.00	29.00
Total	30	24.7333	3.56161	.65026	23.4034	26.0633	18.00	33.00

Homogeneity of Variances Table

One of the assumptions of the one-way ANOVA is that the variances of the groups you are comparing are similar. The table Test of Homogeneity of Variances (see below) shows the result of Levene's Test of Homogeneity of Variance, which tests for similiar variances. If the significance value is greater than 0.05 (found in the **Sig.** column) then you have homogeneity of variances. We can see from this example that Levene's *F* Statistic has a significance value of 0.901 and, therefore, the assumption of homogeneity of variance is met. What if the Levene's *F* statistic was significant? This would mean that you do not have similar variances and you will need to refer to the Robust Tests of Equality of Means Table instead of the ANOVA Table.

Test of Homogeneity of Variances

Time

Levene Statistic	df1	df2	Sig.
.105	2	27	.901

ANOVA Table

This is the table that shows the output of the ANOVA analysis and whether we have a statistically significant difference between our group means. We can see that in this example the significance level is 0.021 (P = .021), which is below 0.05 and, therefore, there is a statistically significant difference in the mean length of time to complete the spreadsheet problem between the different courses taken. This is great to know but we do not know which of the specific groups differed. Luckily, we can find this out in the**Multiple Comparisons Table** which contains the results of *post-hoc* tests.

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

Robust Tests of Equality of Means Table

We discussed earlier that even if there was a violation of the assumption of homogeneity of variances we could still determine whether there were significant differences between the groups by not using the traditional ANOVA but using the Welch test. Like the ANOVA test, if the significance value is less than 0.05 then there are statistically significant differences between groups. As we did have similar variances we do not need to consult this table for our example.

Robust Tests of Equality of Means

Time

	Statistic[a]	df1	df2	Sig.
Welch	4.522	2	17.978	.026

a. Asymptotically F distributed.

Multiple Comparisons Table

From the results so far we know that there are significant differences between the groups as a whole. The table below, Multiple Comparisons, shows which groups differed from each other. The Tukey *post-hoc* test is generally the preferred test for conducting post-hoc tests on a one-way ANOVA but there are many others. We can see from the table below that there is a significant difference in time to complete the problem between the group that took the beginner course and the intermediate course (P = 0.046) as well as between the beginner course and advanced course (P = 0.034). However, there were no differences between the groups that took the intermediate and advanced course (P = 0.989).

Multiple Comparisons

Dependent Variable:Time

	(I) Course	(J) Course	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Beginner	Intermediate	3.60000*	1.43088	.046	.0523	7.1477
		Advanced	3.80000*	1.43088	.034	.2523	7.3477
	Intermediate	Beginner	-3.60000*	1.43088	.046	-7.1477	-.0523
		Advanced	.20000	1.43088	.989	-3.3477	3.7477
	Advanced	Beginner	-3.80000*	1.43088	.034	-7.3477	-.2523
		Intermediate	-.20000	1.43088	.989	-3.7477	3.3477
Games-Howell	Beginner	Intermediate	3.60000	1.42205	.052	-.0314	7.2314
		Advanced	3.80000*	1.40633	.037	.2096	7.3904
	Intermediate	Beginner	-3.60000	1.42205	.052	-7.2314	.0314
		Advanced	.20000	1.46363	.990	-3.5356	3.9356
	Advanced	Beginner	-3.80000*	1.40633	.037	-7.3904	-.2096
		Intermediate	-.20000	1.46363	.990	-3.9356	3.5356

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

Reporting the Output of the one-way ANOVA

There was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,27) = 4.467$, $p = .021$). A Tukey post-hoc test revealed that the time to complete the problem was statistically significantly lower after taking the intermediate (23.6 ± 3.3 min, $P = .046$) and advanced (23.4 ± 3.2 min, $P = .034$) course compared to the beginners course (27.2 ± 3.0 min). There were no statistically significant differences between the intermediate and advanced groups ($P = .989$).

If you are interested in calculating an effect size for a one-way ANOVA, we explain how to do this in our Premium articles. Find out more here.