LAB 1 INSTRUCTIONS

T-TOOLS, DATA TRANSFORMATIONS, AND ALTERNATIVES

The instructions will assist you in learning how to describe and display univariate data with some statistical tools in SPSS. In particular, you will learn how to use the *Explore* procedure to obtain descriptive statistics, side-by-side boxplots and normality plots. Moreover, you will learn how to transform the original measurements when normality assumption is violated and how to use the T-Tools (independent samples T-Test and paired samples T-Test) and distribution–free methods to test for differences between two groups.

We will demonstrate the above statistical procedures in SPSS with a simple example.

Example: T-cells play a central role in the human body immune system. A new drug has just been developed that is expected to boost the immune function by increasing the number of T-cells in human body. In order to test the effectiveness of this drug, 20 healthy young male subjects were randomly selected and assigned randomly to two groups, control and treatment groups, each consisting of 10 subjects. The 10 subjects in the control group received placebo, the 10 subjects in the treatment group were given the new drug. The numbers of T-cells in each subject were measured twice, 10 minutes following the administration of the placebo or the new drug and 1 hour later. The study was double-blinded.

The data are available in the SPSS data file that can be downloaded by clicking the link below:

DOWNLOAD DATA

The following is a description of the variables in the data file:

Variable Name	Description of Variable
TCELLS1	Count of T-cells in the subject 10 minutes after administration of the placebo or the drug,
TCELLS2	Count of T-cells in the subject 60 minutes after administration of the placebo or the drug,
GROUP	c (control) or t (treatment).

Download the data to your workstation or home computer. Note that TCELLS1 and TCELLS2 are both numeric variables, GROUP is a string variable.

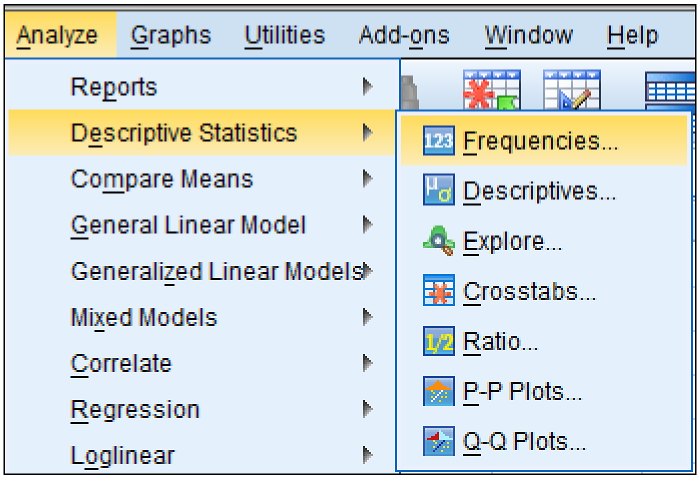
1. Displaying and Describing Data with SPSS

There are four basic steps to data analysis with SPSS:

(a) Bring your data into SPSS (open a previously saved SPSS data file, read a spreadsheet, or enter your data directly in the *Data Editor*),
(b) Select a procedure from the menus,
(c) Select the variables for the analysis (move the variables from the source list to the target list in a corresponding dialog box for the procedure),
(d) Run the procedure and examine the results.

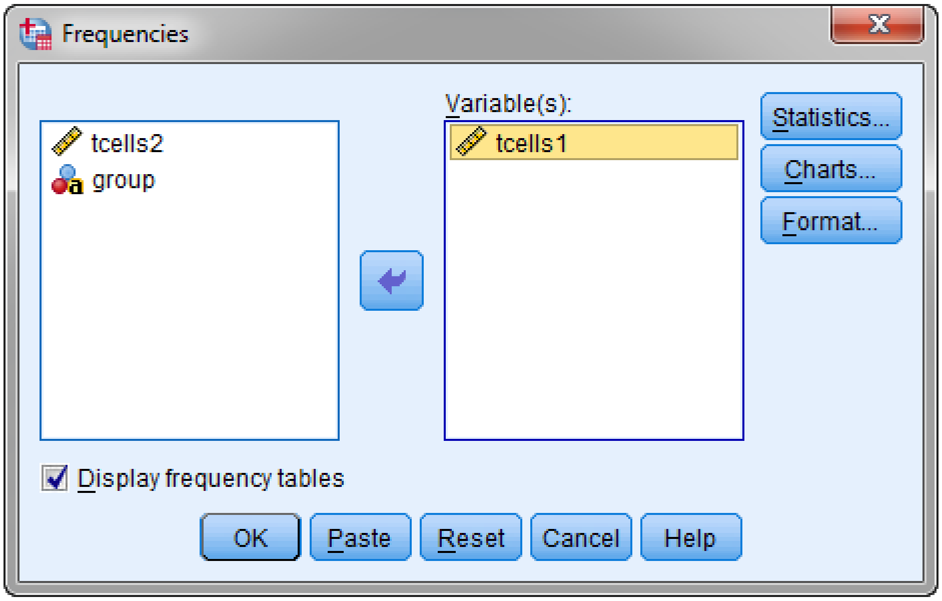
We will demonstrate the methodology by obtaining some charts and descriptive statistics for our sample data.

Frequencies, *Descriptives*, and *Explore* are the principal procedures for describing and exploring continuous data, the last one also allowing quantitative variables to be classified by categories of a qualitative variable (e.g. gender). In order to access them select *Analyze* from the main menu. You will obtain the following pull-down menu with submenus:



1.1 Frequencies

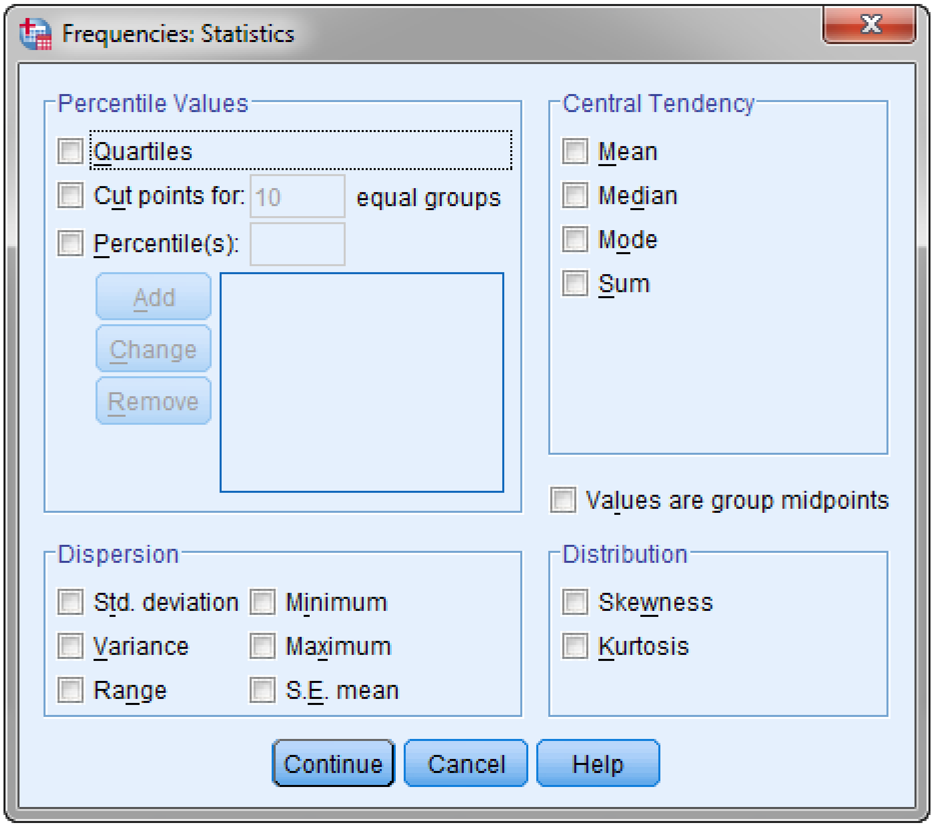
The first example is the use of *Frequencies* with the variable *tcells1*. Click *Frequencies* in the above pop-up menu and use the arrow button to move the variable *tcells1* into the *Variable(s)* box.



Variable(s): tcells1

☑ Display frequency tables

Notice that there are three tabs in the *Frequencies* dialog box: *Statistics…*, *Charts…*, and *Format…*. By clicking *Statistics* tab, you will obtain the following dialog box.



Frequencies: Statistics

Percentile Values

□ Quartiles
□ Cut points for: 10 equal groups
□ Percentile(s):

Central Tendency

□ Mean
□ Median
□ Mode
□ Sum

□ Values are group midpoints

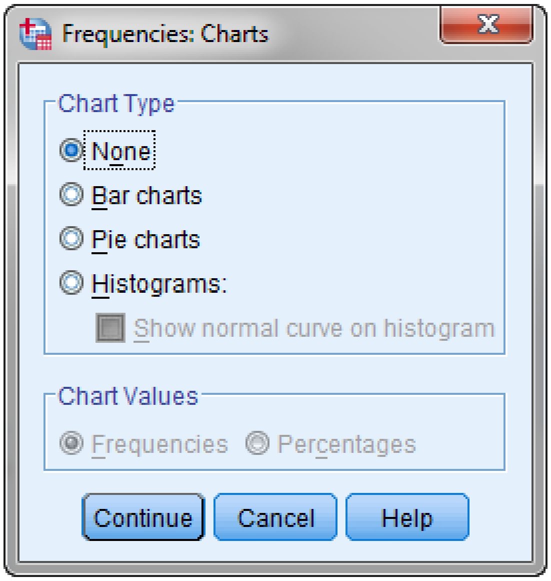
Dispersion

□ Std. deviation □ Minimum
□ Variance □ Maximum
□ Range □ S.E. mean

Distribution

□ Skewness
□ Kurtosis

Check on these measures of central tendency and dispersion you wish to obtain for the data. For example, you might choose the mean and the median as the requested measures of central tendency and standard deviation and quartiles as measures of dispersion.Then click *Chart* tab in the *Frequencies* dialog box to request a histogram of tcells1 count for the 20 subjects.



Frequencies: Charts

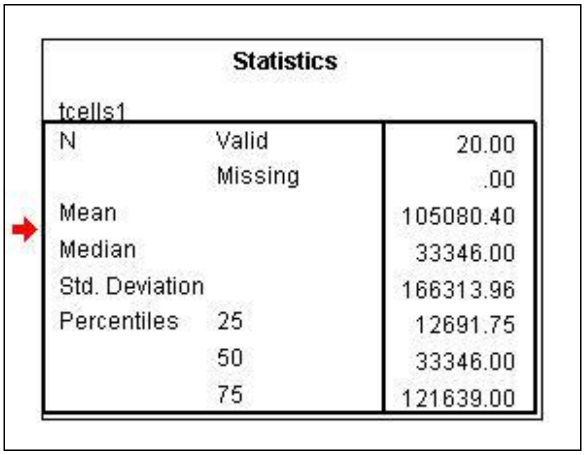
Chart Type

● None
○ Bar charts
○ Pie charts
○ Histograms:
□ Show normal curve on histogram

Chart Values

● Frequencies ○ Percentages

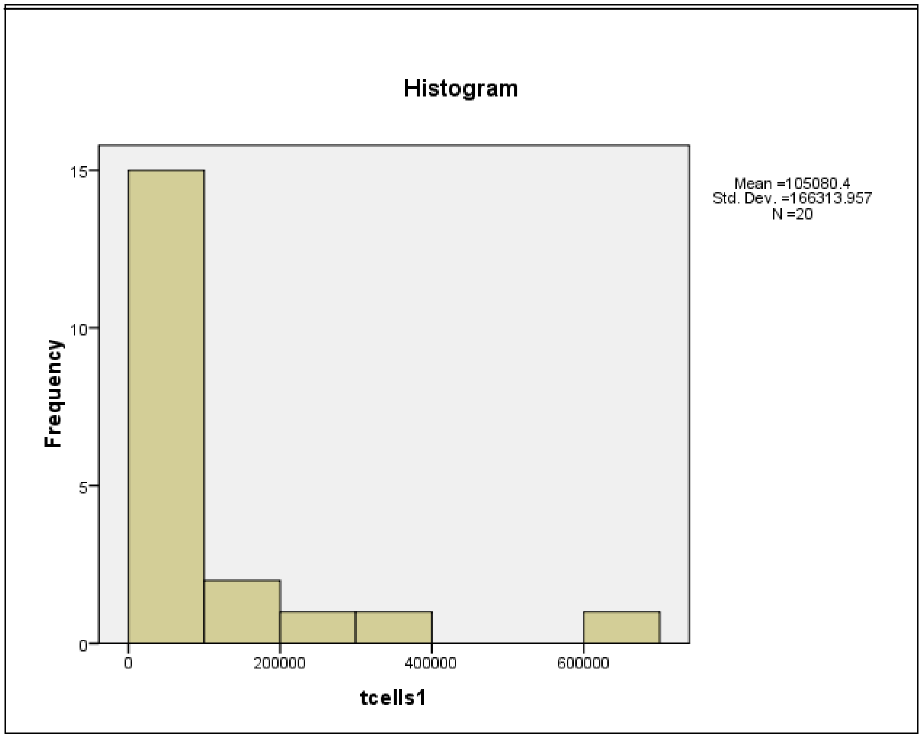
The summary statistics and the histogram will be displayed in the *Viewer* window. They are displayed on the next page.



Statistics

tcells1

N	Valid	20.00
	Missing	.00
Mean		105080.40
Median		33346.00
Std. Deviation		166313.96
Percentiles	25	12691.75
	50	33346.00
	75	121639.00



The output in this case has been obtained by running only one procedure *Frequencies* to the variable *tcells1*. It consists of the five items: *Title*, *Notes*, *Active Dataset*, *Statistics*, and *Histogram*.

You can use the scroll bars in the display pane to browse the results. Or you can click an item in the outline to go directly to the corresponding table or chart. To practice, click each item in the output to see its contents.

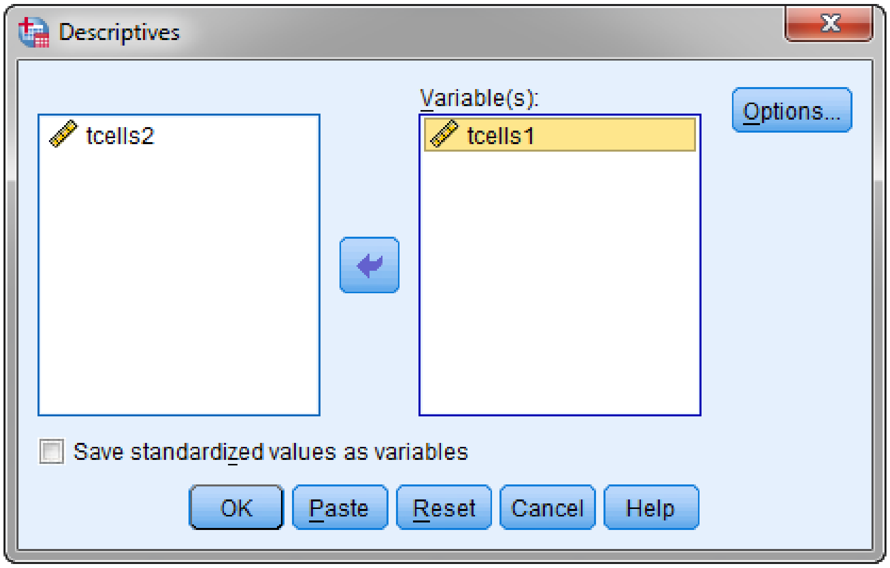
An arrow in front of an open book icon in the outline pane indicates that the corresponding item is currently visible in the display pane. To hide a table or chart in the display without deleting it, double-click its book icon. The open book changes to a closed book icon, indicating that the item is now hidden.

To change the position of tables or charts in the display, click the items in the outline pane, and drag them where you want to put them. Try it.

You can switch between the *Data Editor* and *Viewer* windows by clicking *Window* in either menu.

1.2 Descriptives

The *Descriptives* procedure in SPSS allows you to calculate basic univariate statistics for numeric variables. To open the Descriptives dialog box, choose Analyze from the main menu, then *Descriptive Statistics*, and finally *Descriptives*. Move the variable *tcells1* into the *Variables* box as shown below:

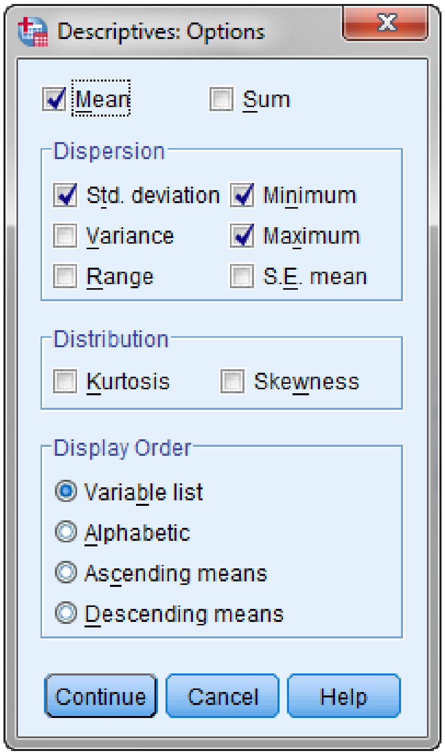


tcells2

Variable(s): tcells1

□ Save standardized values as variables

Clicking *Options* tab in the right panel opens the following dialog box:



☑ Mean □ Sum

Dispersion

☑ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S.E. mean

Distribution

□ Kurtosis □ Skewness

Display Order

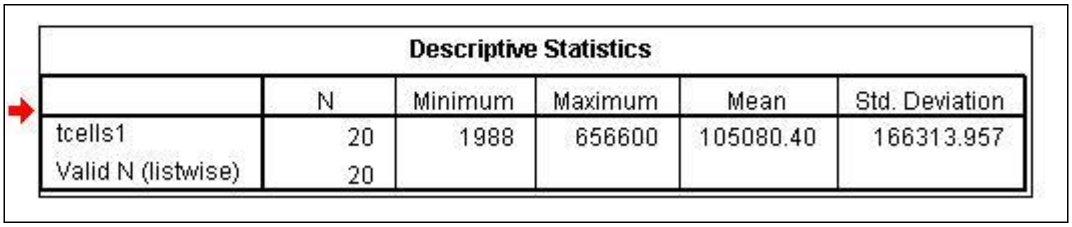
● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

You may choose the statistics you are interested in for the variable tcells1. Click *Continue* and *OK* in the *Descriptives* dialog box to obtain the following output:



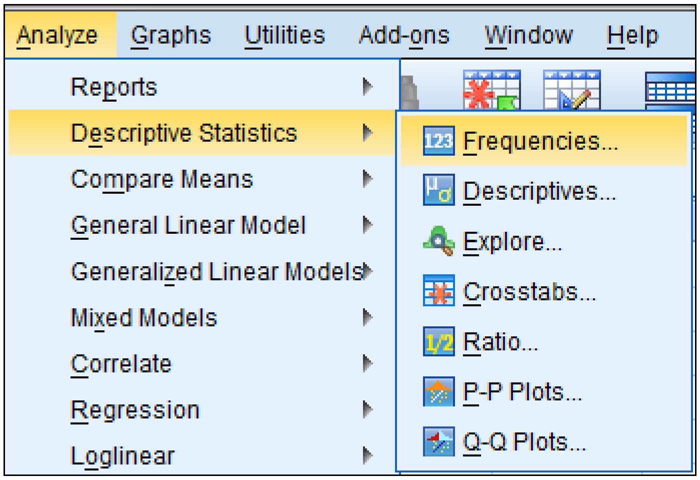
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
tcells1	20	1988	656600	105080.40	166313.957
Valid N (listwise)	20				

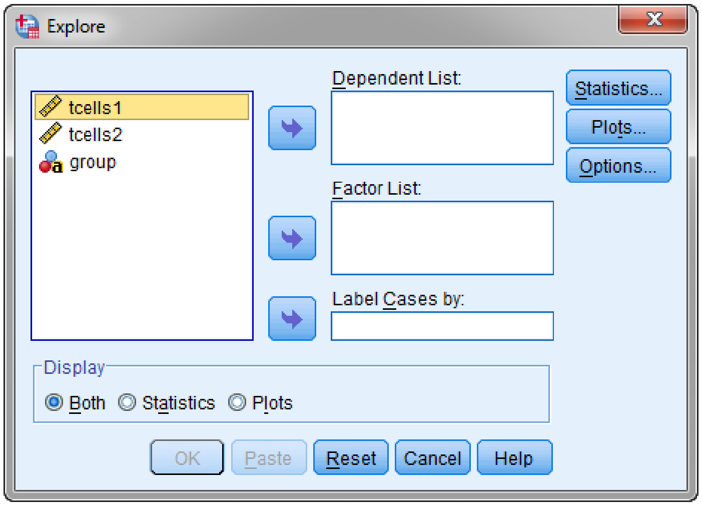
1.3 Explore Procedure

The Explore procedure in SPSS allows you to obtain descriptive statistics for your data, identify extreme observations, calculate the percentiles of the distribution, and display the data with a variety of plots.

In order to access the *Explore* procedure in SPSS, click *Analyze* in the menu bar, and then *Descriptive Statistics* from the pull-down menu

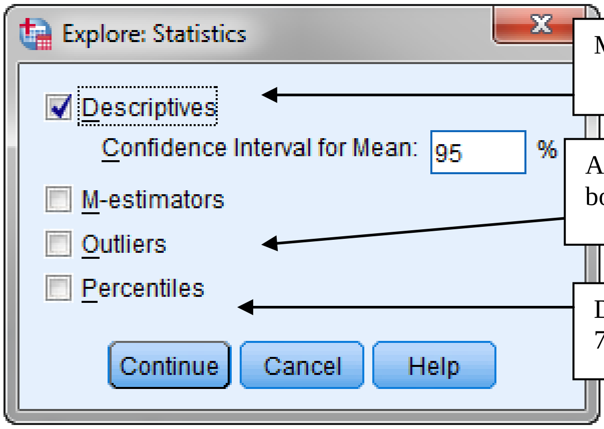


Clicking *Explore* tab opens the following dialog box:



In order to compare the number of T-cells 10 minutes after the treatment administration (placebo or drug) between the two treatment groups, select the *tcells 1* variable in the variables left panel of the Explore dialog box and move it into the *Dependent List* box by clicking the forward arrow button at the box. Make sure that the *group* variable is moved into the *Factor List:* entry box. Moreover, make sure that either *Statistics* or *Both* is selected in the Display group at the bottom left.

Clicking the *Statistics* tab opens the *Explore: Statistics* dialog box as shown below:



☑ Descriptives

Confidence Interval for Mean: 95 %

□ M-estimators

□ Outliers

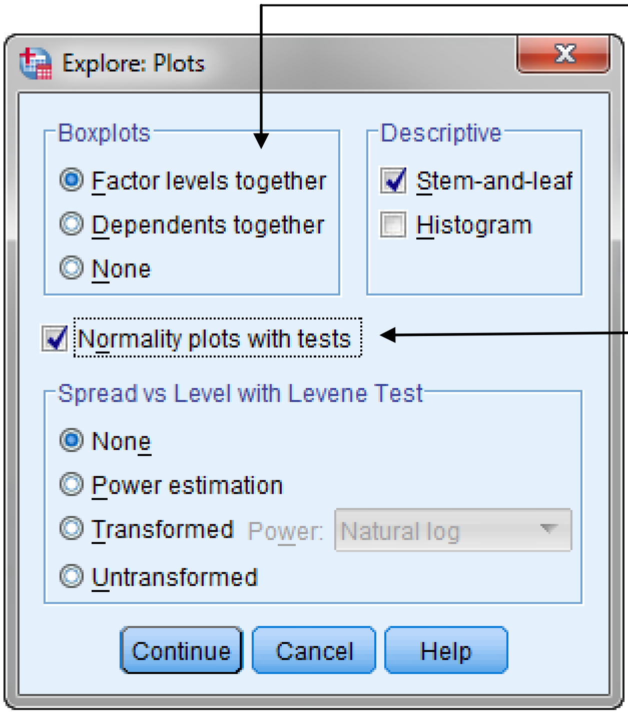
□ Percentiles

Measures of central tendency and dispersion

An outlier is defined as a value more than 1.5 box-lengths away from either end of the box

Displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles

The *Descriptives* check box is already checked (default). Click the *Continue* tab. Now click *Plots* button in the *Explore* dialog box. It opens the *Explore: Plots* dialog box. Select the *Normality plots with tests* check box to obtain the normality plot of tcells 1 variable for each group.



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

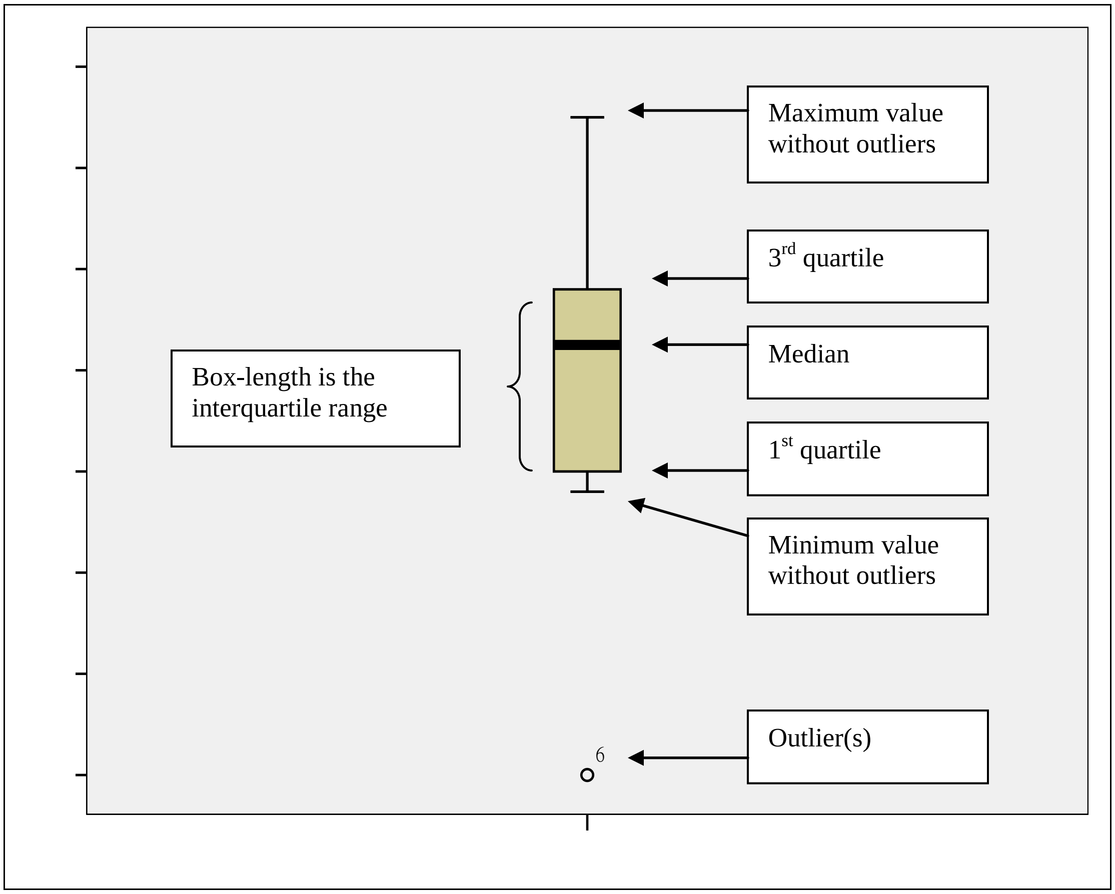
The default setting for the boxplots is a side-by-side plot for each group (placebo, treatment)

Select this check box to obtain the normality plot for each group and test whether the samples come from populations that have normal distribution

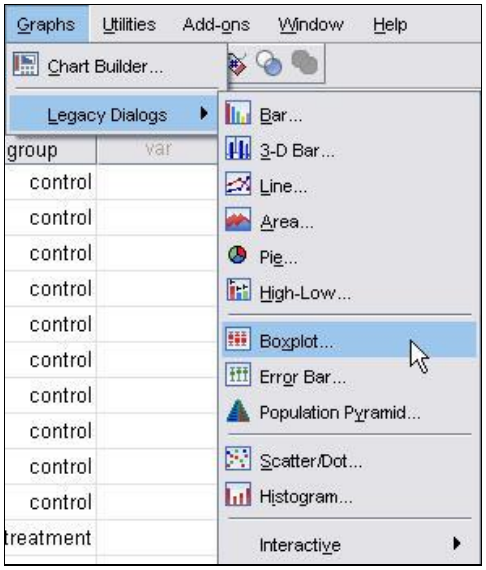
Click *Continue* and then *OK* to run the procedure. The output of the *Explore* procedure will be displayed in the *Viewer* window.

1.4 Boxplots

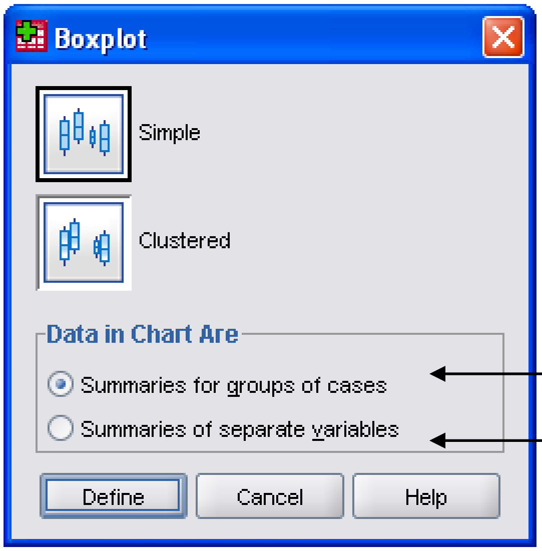
Boxplots are very helpful to better understand and visualize the distribution of a numerical variable. They simultaneously display the median, the quartiles, the interquartile range, and the smallest and largest values. The following picture summarizes the information provided by boxplots:


Maximum value without outliers
3rd quartile
Median
1st quartile
Minimum value without outliers
Box-length is the interquartile range
Outlier(s)
6

As we have seen in Section 1.3, boxplots can be obtained with the Explore procedure. They can also be requested directly by choosing the *Boxplot* item in the *Graphs* drop-down menu in the menu bar.



The following dialog box opens:

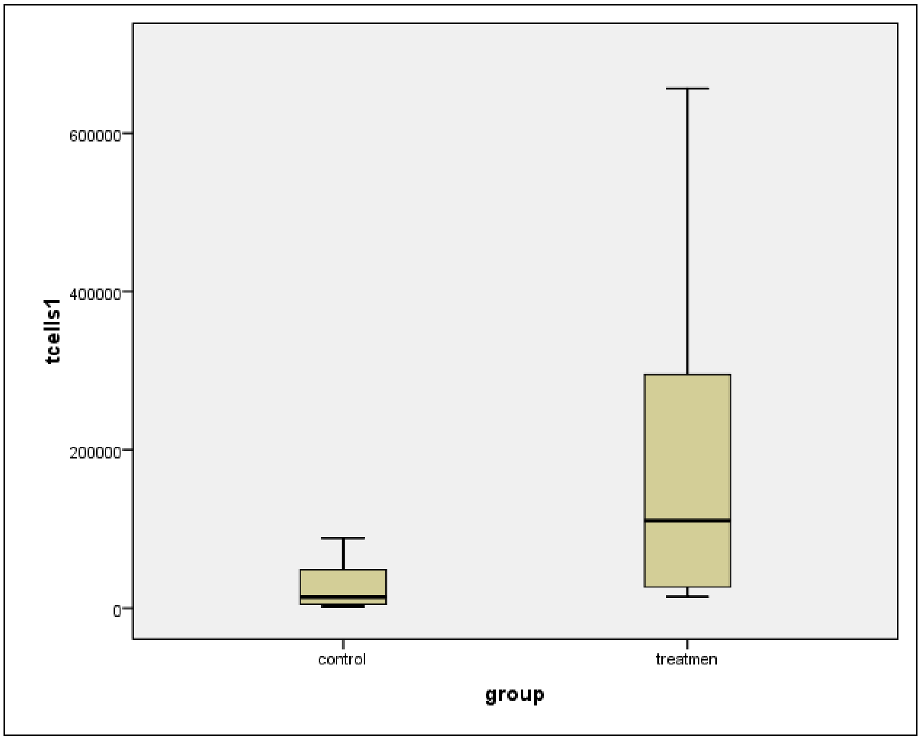

Boxplot
Simple
Clustered
Data in Chart Are
● Summaries for groups of cases
○ Summaries of separate variables
Define
Cancel
Help

Check the first radio button to obtain a side-by-side boxplot for each group (control, treatment)

Check the second radio button to obtain a separate boxplot for each group

In the *Boxplot* dialog box, click the *Simple* button. In the *Data in Chart Are* panel there are two radio buttons that refer to two possible options in the structure of your data: *Summaries for groups of cases* or *Summaries of separate variables* buttons. The first option creates a boxplot of a single quantitative variable within categories of another variable; the second option creates a boxplot of one or more quantitative variables.

In order to obtain the boxplot for the *tcells1* variable for each group, click *Summaries for groups of cases*. The side-by-side boxplot for the two groups is shown below:

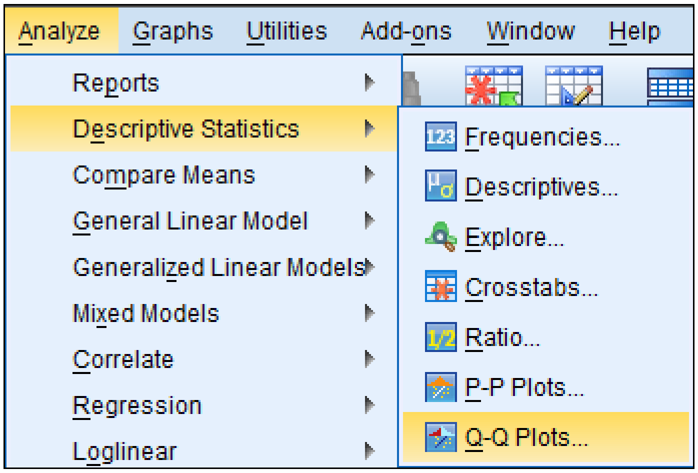


As you can see, the distribution of the number of T-cells for each group is extremely right skewed.

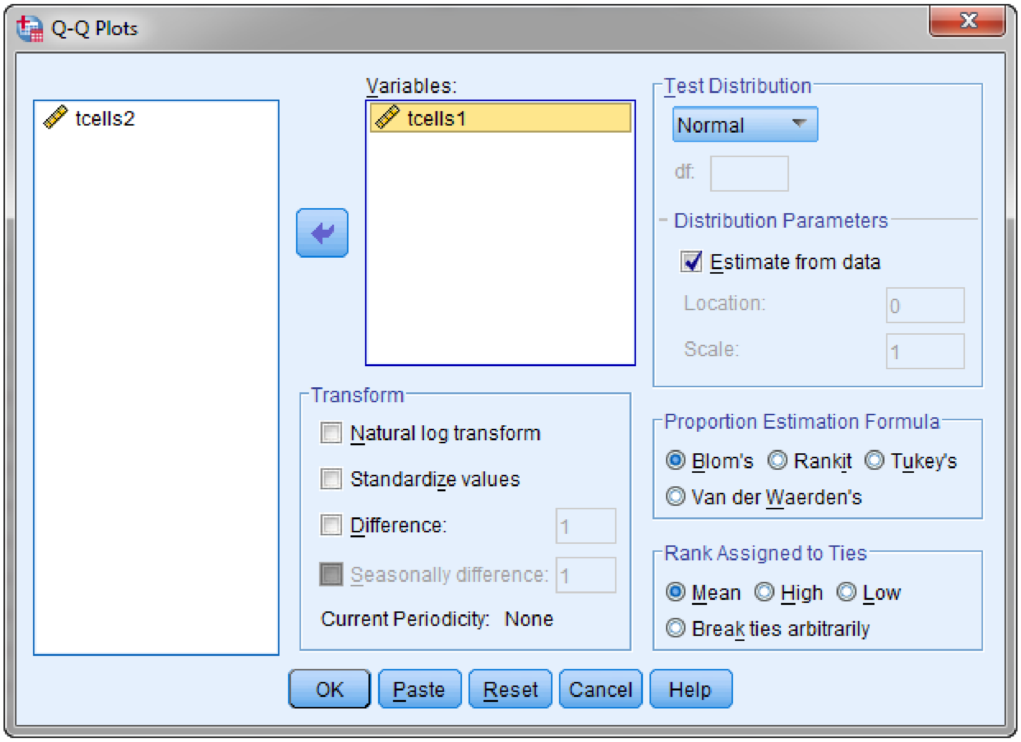
1.5 Normality Plots

Normality plots are used to evaluate the normality of the distribution of a variable, that is, whether and to what extent the distribution of the variable follows the normal distribution. The selected variable will be plotted in a scatterplot against the values "expected from the normal distribution."

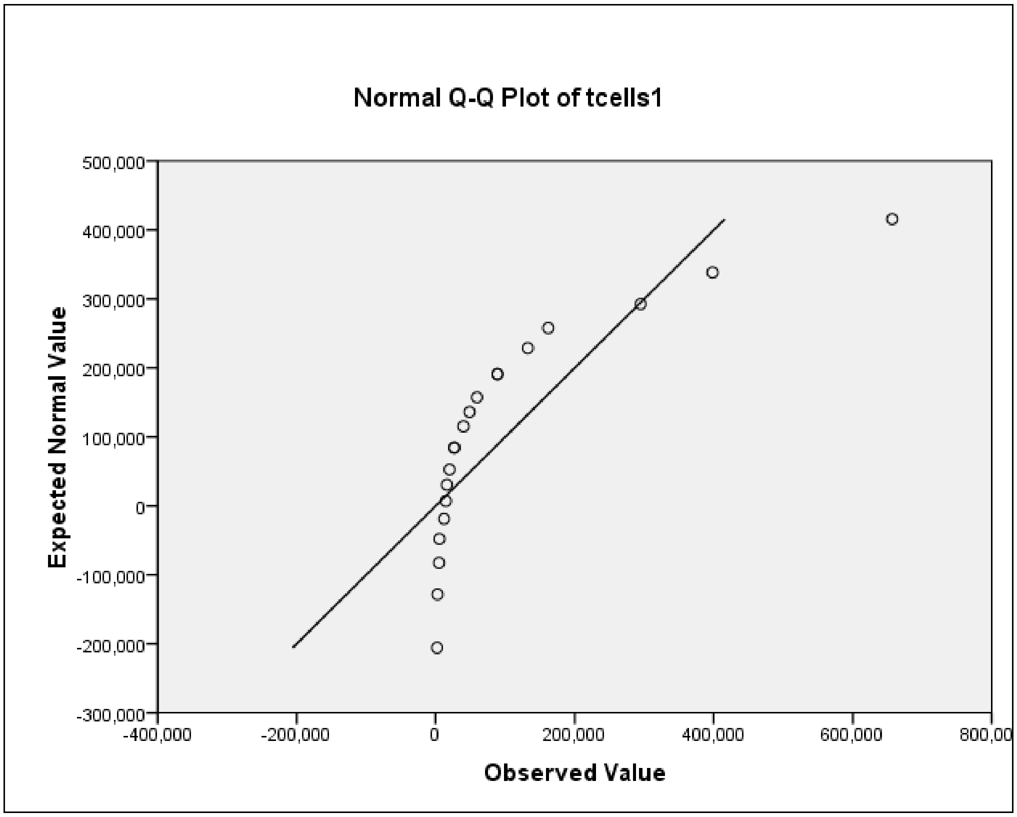
In order to obtain a normality plot, click *Analyze* in the menu and then select *Q-Q Plots*…option.



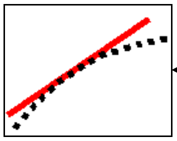
Now select and move the *tcells1* variable to the *Variables* box, and then click *OK* button.



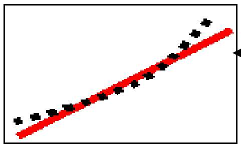
The QQ plot will be displayed in the *Viewer* window as shown below. If all points in the plot are reasonable close to a straight line, the data follow approximately a normal distribution.



Systematic deviations from a straight line indicate a non-normal distribution. Outliers appear as points that are far away from the overall pattern.



Right skewed distribution



Left skewed distribution

According to the above Q-Q plot, the counts of T-cells 10 minutes after treatment administration for the combined data (control and treatment) do not seem to come from a normal population; the data are clearly right skewed.

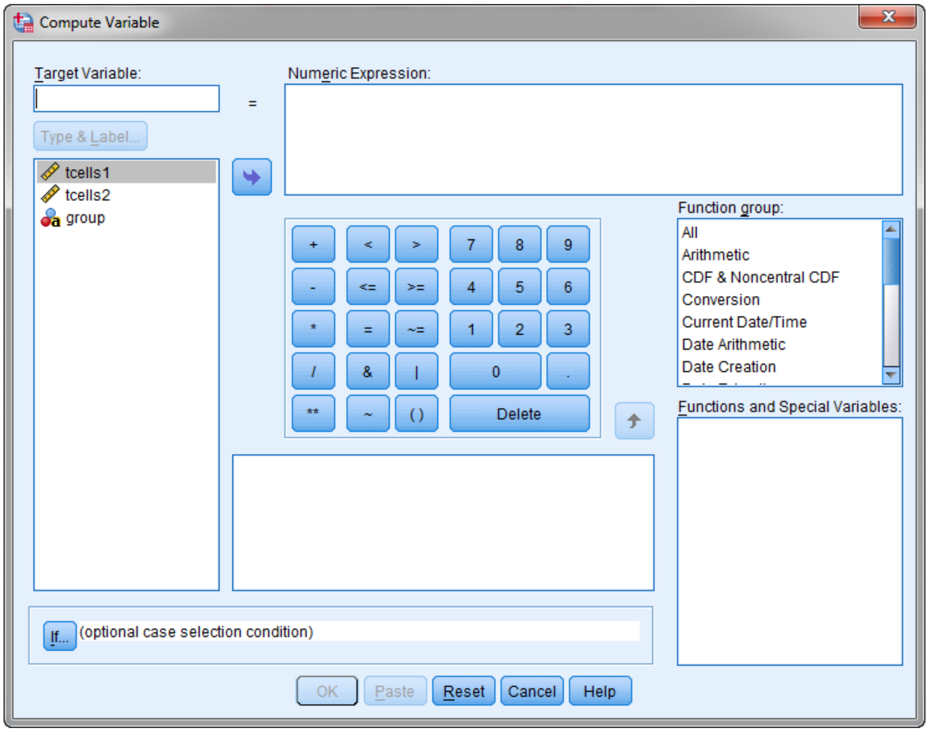
Notice that in order to obtain a separate Q-Q plot for each group (control and treatment) with the Q-Q Plots procedure, you have to split the data file by group first and then apply the Q-Q Plot procedure.

2. Data Transformations

Data transformations are a remedy for outliers, departures from the normality, linearity, or equal-variances assumptions. However, only those transformations are useful that allow us to interpret easily the transformed variables.

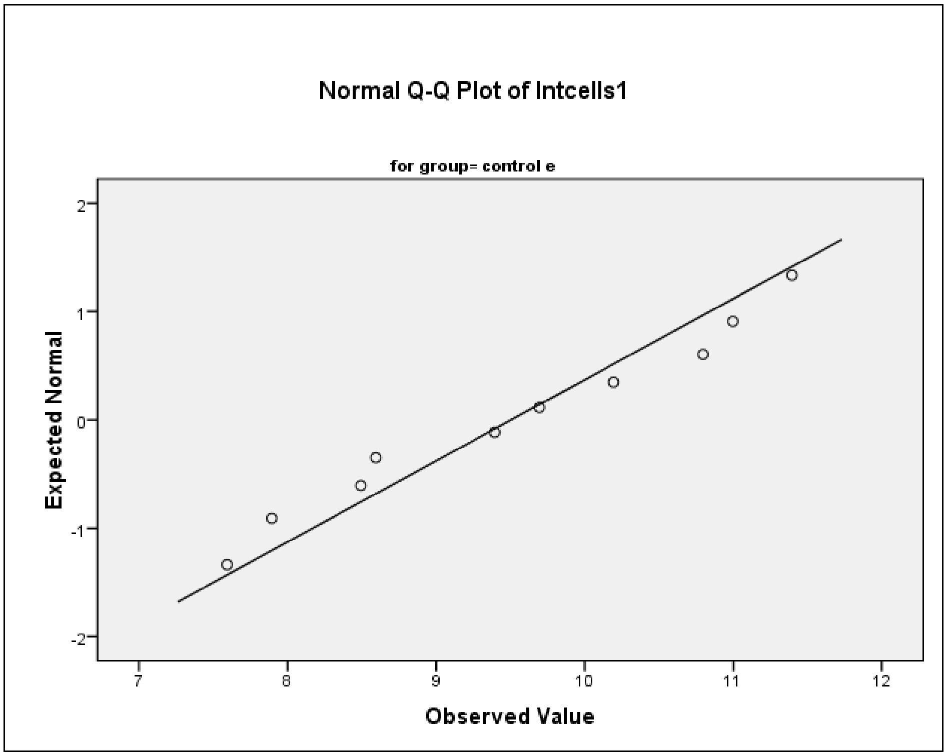
2.1 Natural Logarithm Transformation

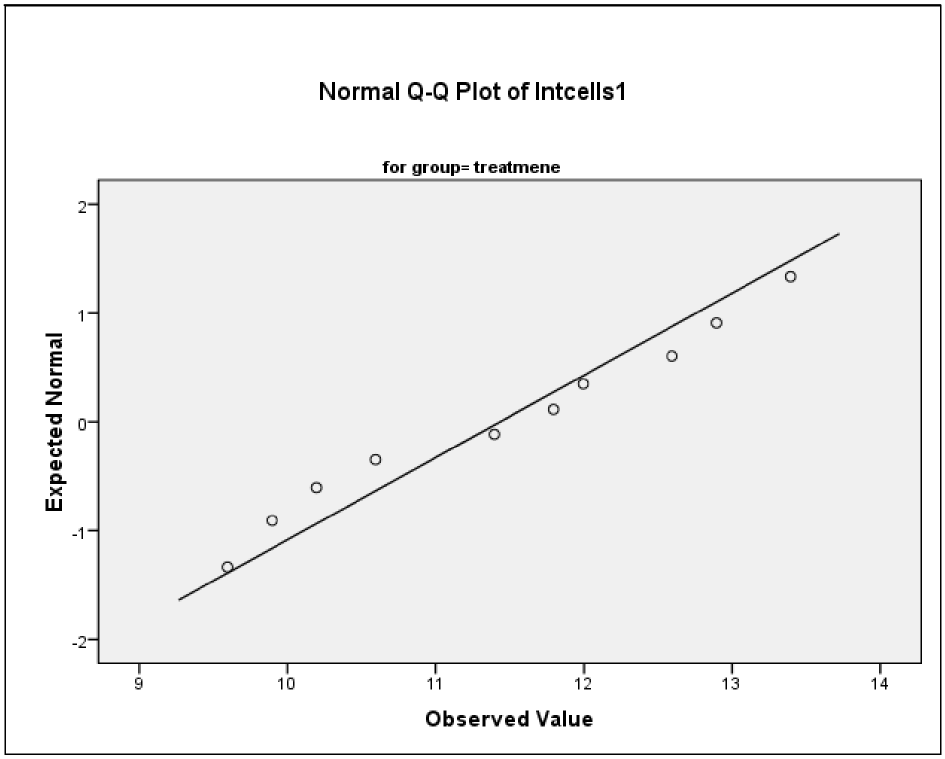
The log-transformation can be used to remove the observed skewness and outliers so that the assumption of normality is not violated. Click *Transform* and *Compute Variable*…The following dialog box is obtained:



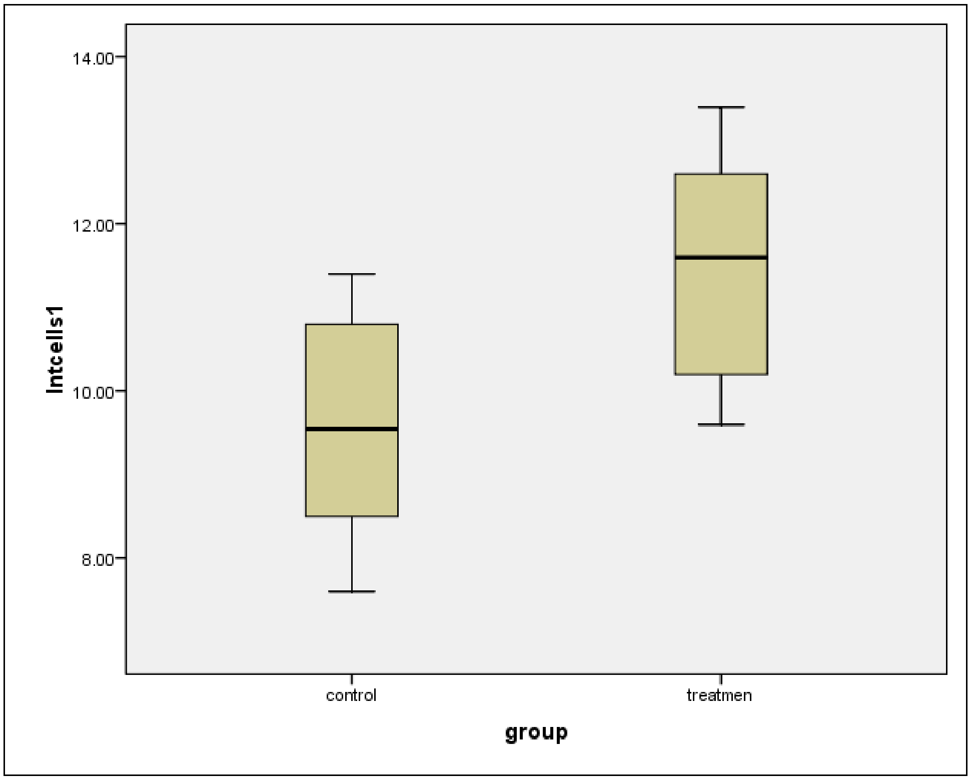
For example, let us define a new variable lntcells1 which is the natural logarithm transformation of *tcells1* for each group. Click arrow to move the Ln function to the *Numeric Expression:* box. Then select the *tcells1* variable in the left box and move it into the *Numeric Expression:* box. The target variable lntcells1 is defined as Ln(tcells1). Once you have completed the appropriate expression, click *OK* to close the *Compute Variable* dialog box. The new defined variable will be displayed in the *Data Editor* window.

The Q-Q plots for each group are shown below:



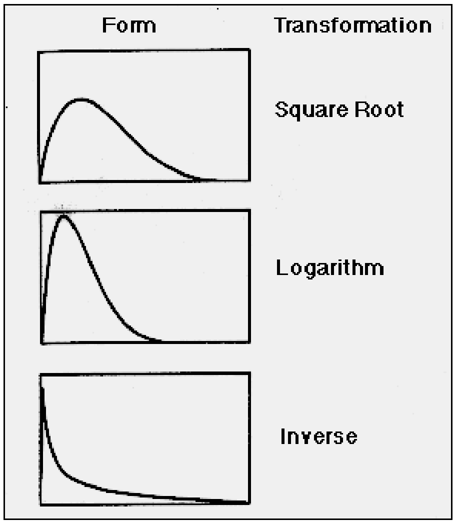


As you can see the natural logarithm transformation was successful to remove the right skewness in the two distributions. Moreover, it is easy to check that the two groups have approximately the same spread on the natural logarithm scale (see the side-by-side boxplot below):



2.2 Other Useful Data Transformations

The type of transformation used to remove skewness and/or outliers depend on the form of the distribution of the variable on the original scale of measurement. Some recommended data transformations for a given distribution are given in the diagram below:



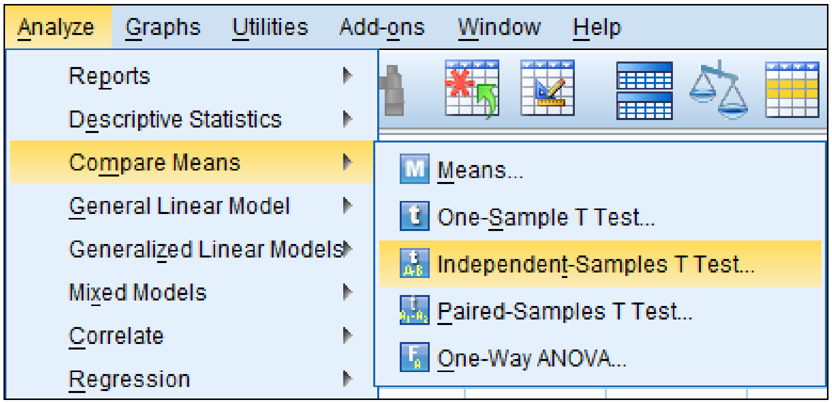
3. Tests for Differences between Two Independent Samples

In this section you will learn how to compare two populations or treatment groups with independent-samples t-test (under the assumption of normality) and Mann-Whitney test (no distributional assumptions required for this test).

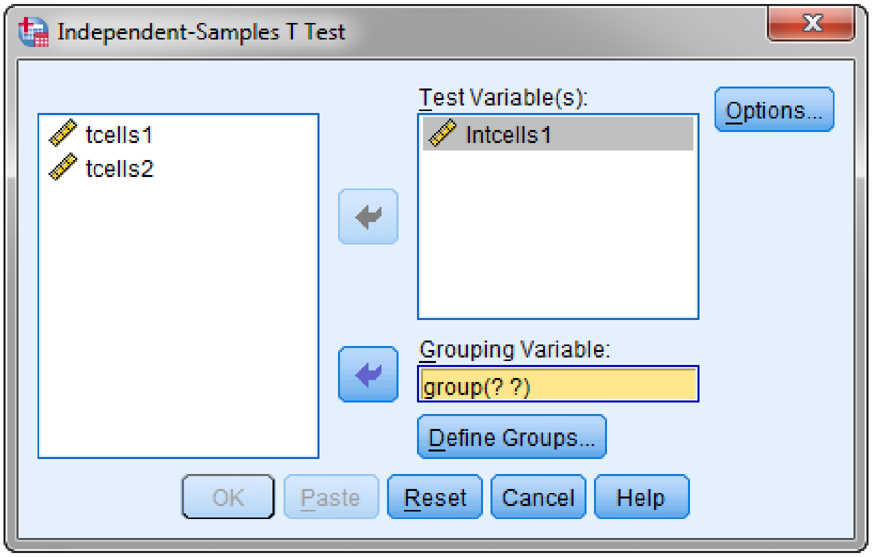
3.1 Independent-Samples T-Test

The Independent –Samples T Test procedure tests the null hypothesis that the population means for two groups are equal. It also calculates a confidence interval for the difference between the population means in the two groups. To run an independent-samples t test in SPSS, you must specify the variable(s) whose means you want to compare, and you must specify the two groups to be compared. The test is valid under the assumption that the data come from normal populations, though it is fairly robust to minor departures from normality.

To open the Independent-Samples T Test dialog box, choose *Analyze* from the main menu, *Compare Means*, and finally *Independent-Samples T Test*.



Suppose we wish to determine whether the new drug to increase the number of T-cells is effective. After the natural-log-transformation, the variable *lntcells1* is approximately normally distributed. Thus the independent two samples T-Test can be used to test for the differences between control and treatment groups:



Independent-Samples T Test

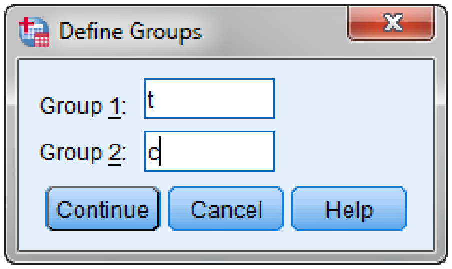
tcells1

tcells2

Test Variable(s): lntcells1

Grouping Variable: group(? ?)

Select and move the *lntcells1* variable to the *Test Variable(s)* box using the upper right arrow button. Then select and move the *group* variable to the *Grouping Variable* box using the lower right arrow button. Click *Define Groups* button to open the *Define Groups* dialog box.



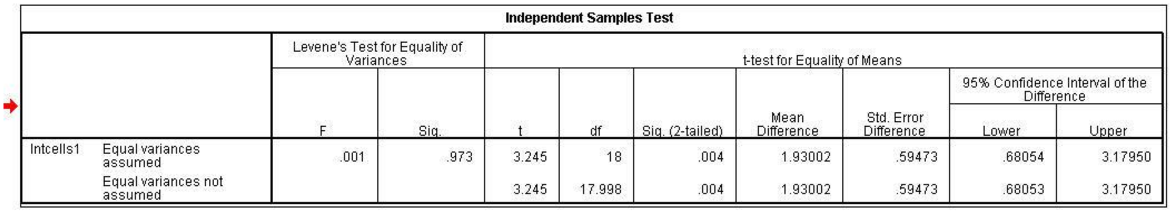
Define Groups

Group 1: t

Group 2: c

In our example, there are two groups, treatment (t) and control (c). Therefore, we enter these two values into the *Group 1* and *Group 2* boxes separately. Note that the values specified for Group 1 and Group 2 must be either short strings or numeric values (for example, 0 and 1).

Click *Continue* to close the *Define Groups* dialog box and then click *OK* to run the procedure. The output of the *Independent Samples Test* table is displayed below.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
lntcells1	Equal variances assumed	.001	.973	3.245	18	.004	1.93002	.59473	.68054	3.17950
	Equal variances not assumed			3.245	17.998	.004	1.93002	.59473	.68053	3.17950

Before we examine the above output, it is worth to notice that the difference in averages for the log transformed data is reported in the above output as $\overline{z}_t - \overline{z}_c = 1.93002$. The antilogarithm of the above difference is exp($\overline{z}_t - \overline{z}_c$) = 6.889648. Here the subscripts t and c stand for “treatment” and “control” groups, respectively.

For symmetric distributions, the mean and the median are approximately equal. As the log-transformed data have approximately symmetric distributions, the following relationships hold:

$$\overline{z}_t - \overline{z}_c \approx \text{Median}[\ln(y_t)] - \text{Median}[\ln(y_c)] = \ln[\text{Median}(y_t)] - \ln[\text{Median}(y_c)]$$

The last equality follows from the fact that log preserves ordering.

From the properties of logarithms, it follows that,

$$\overline{z}_t - \overline{z}_c \text{ estimates } \ln\left[\frac{\text{Median}(\text{y}_\text{t})}{\text{Median}(\text{y}_\text{c})}\right]$$

and, therefore

$$\exp(\overline{z}_t - \overline{z}_c) \text{ estimates } \left[\frac{\text{Median}(\text{y}_\text{t})}{\text{Median}(\text{y}_\text{c})}\right].$$

Therefore the ratio

$$\left[\frac{\text{Median}(\text{y}_\text{t})}{\text{Median}(\text{y}_\text{c})}\right] \text{can be estimated by 6.889648.}$$

In other words, the median T-cells count for the treatment group is approximately 6.89 times as large as the median T-cells count for the control group (10 minutes after administration of placebo or the treatment).

Now we examine the output of independent samples t test shown on the previous page.

The independent samples t-test is run either under the assumption that the population variances are equal (in this case sample variances are pooled to obtain a single estimate of the common variance) or without the assumption. The two-sample t-test with a pooled variance is slightly more powerful than the two-sample t-test without the equal variances assumption, but serious error can result if the variances are not equal. The two tests (under the assumption of equal variances and without the assumption) may produce slightly different values of the test statistics and consequently different p-values.

The assumption of equal variances is verified in SPSS with the Levene’s test. The null hypothesis for the Levene’s test is that the variances are homogeneous (identical). The table that reports the t test statistics also includes the Levene’s test of homogeneity of variance (see the output above).

There are two rows of statistics displayed in the above output. The statistics in the row labelled “*Equal variances assumed*” should be used whenever Levene’s test is not significant, that is, when the variances are homogeneous (more precisely, when the data do not provide sufficient evidence to reject the assumption of equal variances). The statistics in the row labelled “*Equal variance not assumed*” should be used whenever Levene’s test is significant, that is, when the variances are not homogeneous (more precisely, when the data provide strong evidence that the variances are not equal).

For this set of data the Levene’s test is not significant (the value of the test statistic F=0.001, P=0.973), indicating the null hypothesis cannot be rejected, that is, the data support the assumption of equal variances.

The value of the test statistic t in this case is 3.245 and the test statistic follows a t distribution with 18 degrees of freedom (under the null hypothesis that the means are equal). P-value for the test assuming equal variance is 0.004. We conclude that there is strong evidence of differences between the two means (on the natural logarithm scale)

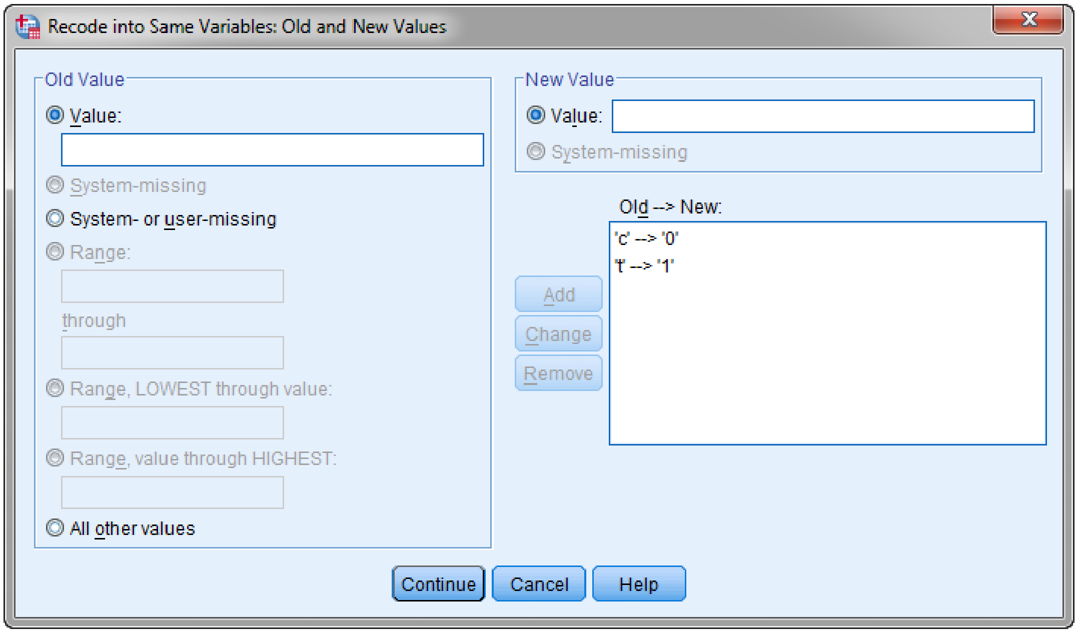
Note that the p-value produced by the t-test in the above output is for two-sided alternative hypothesis. The p-value of the one-sided test that the mean for the treatment group is larger than the mean for the control group (on the natural logarithm scale) is 0.004/2 = 0.002. The significant differences between the two means on the natural logarithm scale are equivalent to significant differences between the medians of the two groups on the original scale of measurement.

The output also includes a 95% confidence interval for the mean difference. That is, the mean difference between the counts of T-cells in control and treatment groups (control minus treatment) on the natural logarithm scale is between 0.68054 and 3.17950. To obtain the 95% confidence interval on the original scale we take the antilogarithm of the endpoints. Thus, the 95% confidence interval for difference in T-cells counts for the control and the treatment groups is (1.974944, 24.03473).

3.2 Nonparametric Tests: Mann-Whitney Test

In some situations the independent samples t-test cannot be applied, because the model assumptions of the t-test are seriously violated. For these situations, some other methods may be used. In particular, distribution-free methods or nonparametric tests are based on models that do not require any specific distribution assumptions. One of them, the Mann-Whitney test (rank-sum test) is the most commonly used alternative to the independent samples t test. The test is used for assessing whether two samples of observations come from the same distribution. It requires the two samples to be independent.

The test also requires the grouping variable *group* be numeric. Thus we have to recode the string variable *group* into the same variable with two possible values: "0" for the control group ("c" will be replaced by "0") and "1" for the treatment group ("t" will be replaced by "1").



Recode into Same Variables: Old and New Values

Old Value

● Value:

○ System-missing

○ System- or user-missing

○ Range:

through

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value:

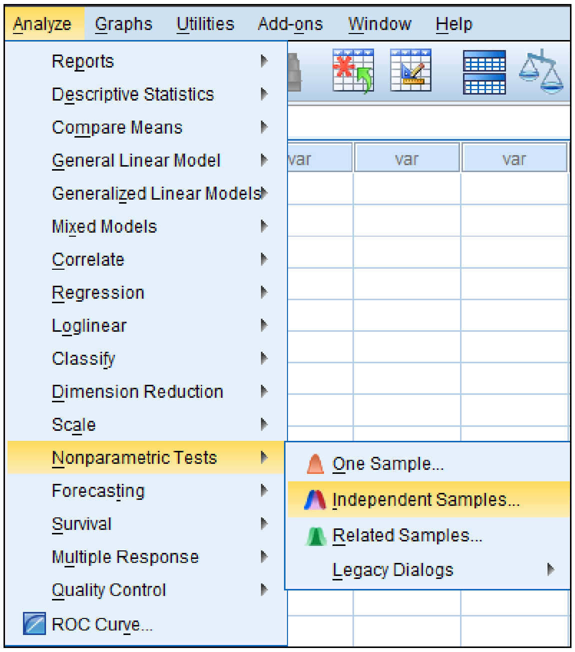
○ System-missing

Old --> New:

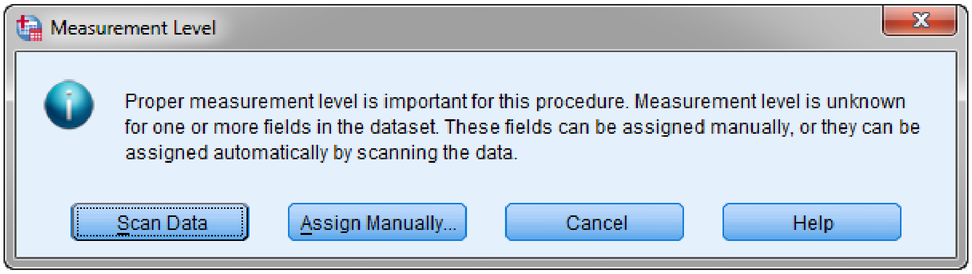
'c' --> '0'

't' --> '1'

In order to compare the variable tcells1 for the treatment and control groups, click *Analyze* in the menu bar, then *Nonparametric Tests* from the pull-down menu:

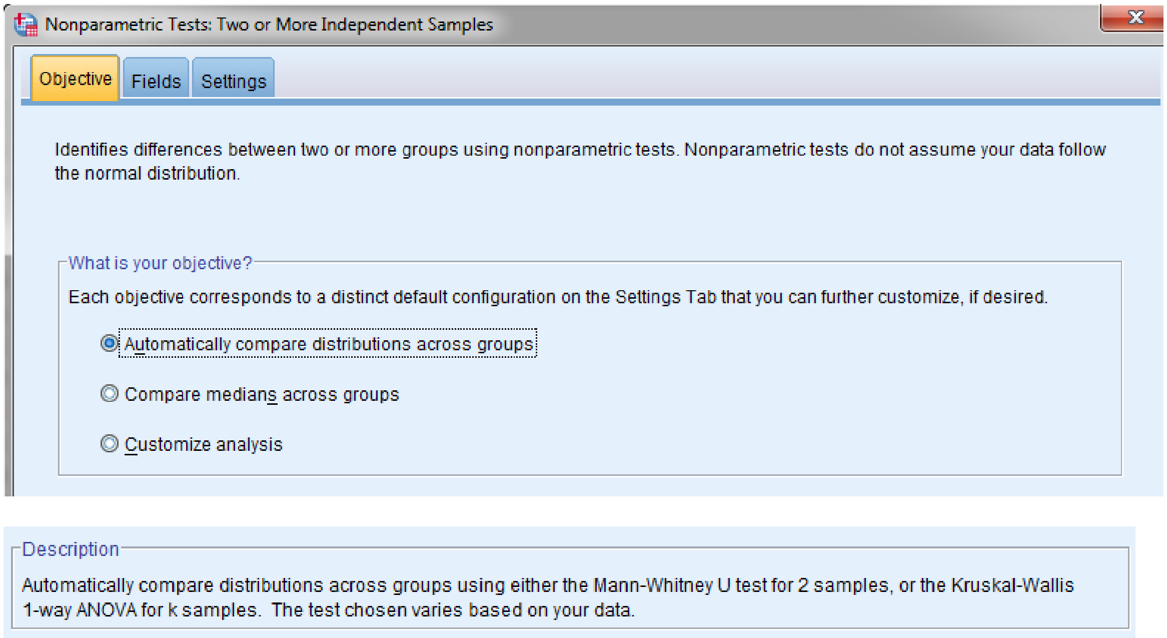


The following dialog box opens to allow the user or SPSS to specify the measurement level of the variables to be tested (nominal, ordinal or continuous).



Proper measurement level is important for this procedure. Measurement level is unknown for one or more fields in the dataset. These fields can be assigned manually, or they can be assigned automatically by scanning the data.

Click *Scan Data*. It opens the *Two-Independent-Samples Tests* dialog box as shown below:



Identifies differences between two or more groups using nonparametric tests. Nonparametric tests do not assume your data follow the normal distribution.

What is your objective?

Each objective corresponds to a distinct default configuration on the Settings Tab that you can further customize, if desired.

● Automatically compare distributions across groups

○ Compare medians across groups

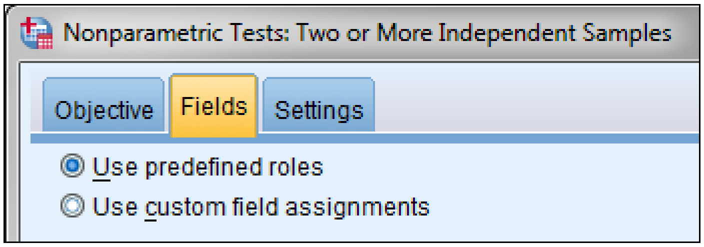
○ Customize analysis

Description

Automatically compare distributions across groups using either the Mann-Whitney U test for 2 samples, or the Kruskal-Wallis 1-way ANOVA for k samples. The test chosen varies based on your data.

As we wish to compare the distributions for the two groups (*tcells1* for the treatment and control groups), we should have the first option in the above dialog box checked. We may also choose to compare the medians of the two groups with Median Test by checking the second option (*Compare medians across groups*).

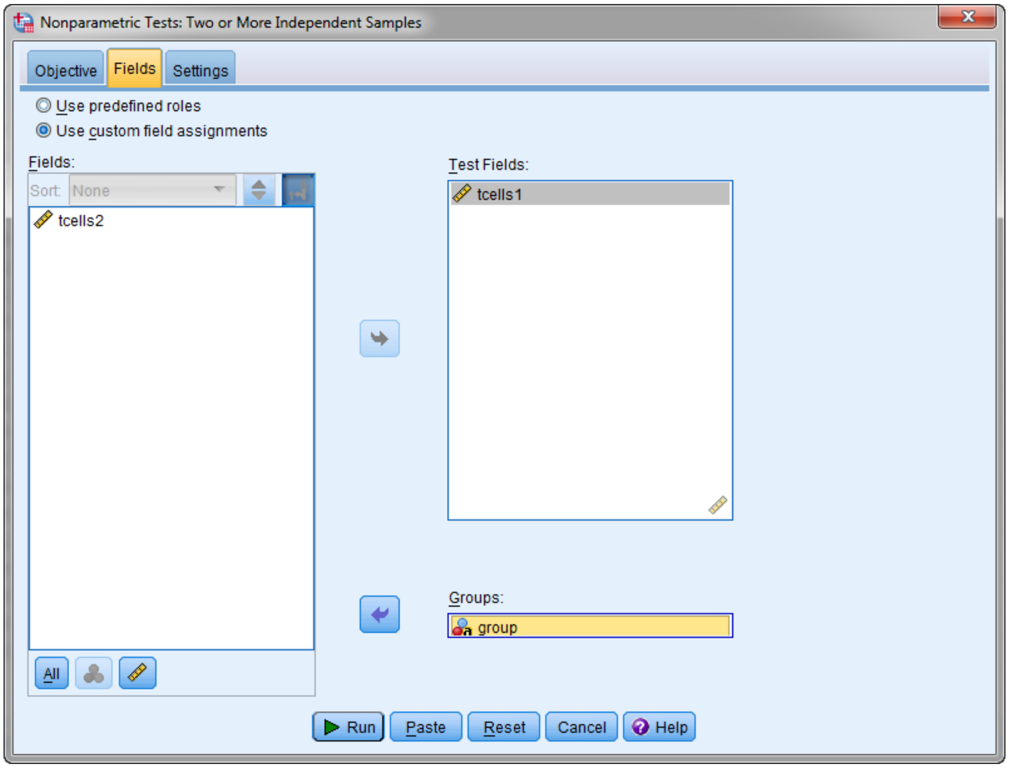
Fields tab specifies which fields should be tested and the field used to define groups. Two possible options are offered: *Use predefined roles* or *Use custom field assignments.*



● Use predefined roles
○ Use custom field assignments

Use predefined roles option uses existing variable information. Otherwise you must use custom field assignments and specify the variables to be tested. If there is a single categorical variable (i.e. *group*), it will be used as a grouping variable.

Select and move the *tcells1* variable to the *Test Fields:* box. Then select and move the numeric variable *group* to the *Groups* box. Click *Run* to run the procedure.



○ Use predefined roles
● Use custom field assignments

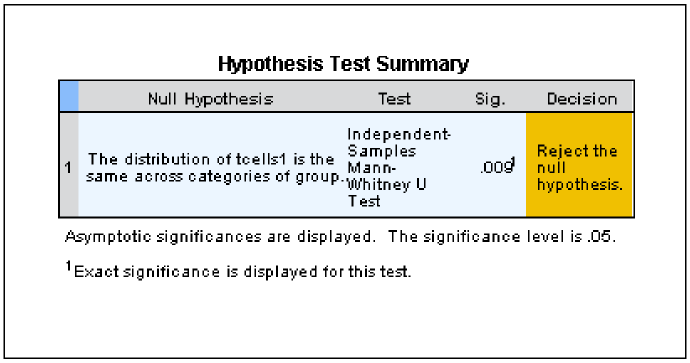
Fields: tcells2

Test Fields: tcells1

Groups: group

The *Settings* tab is optional and allows the user to specify the tests to be performed on the variables (fields) specified on the *Fields* tab. Otherwise, SPSS will automatically choose the tests based on the data. *Automatically choose the tests based on the data* setting applies the Mann-Whitney U test to data with

The *Mann-Whitney Test* output for our data is given below.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of tcells1 is the same across categories of group.	Independent-Samples Mann-Whitney U Test	.009[1]	Reject the null hypothesis.

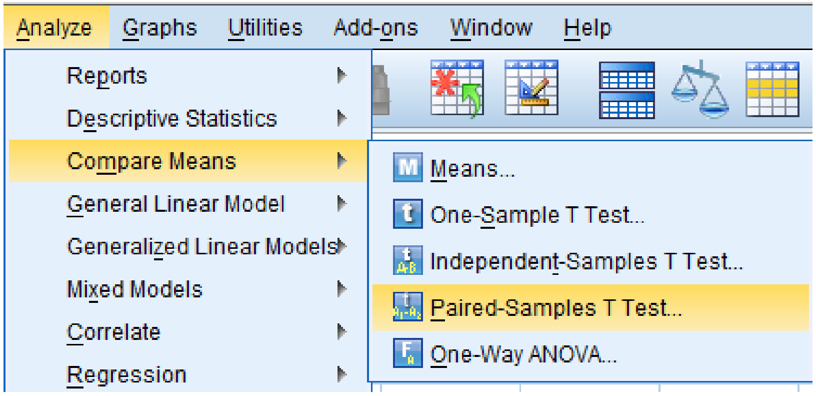
Asymptotic significances are displayed. The significance level is .05.

[1] Exact significance is displayed for this test.

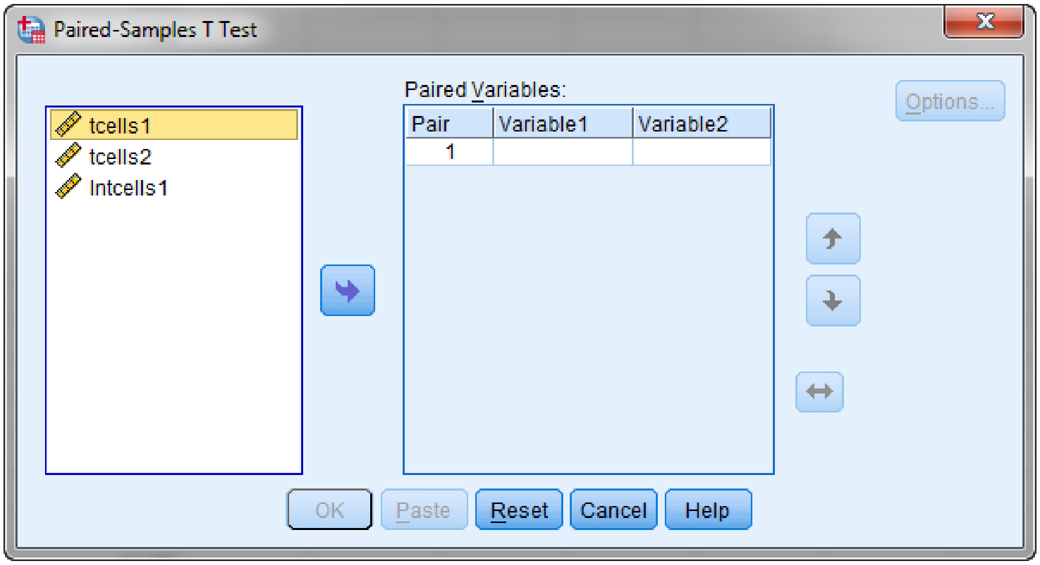
As you can see, the value of the test statistics is z=-2.572 and the corresponding p-value is 0.009. We conclude that there is strong evidence of a significant difference in T-cells counts for the treatment and control groups (10 minutes after the administration of placebo or the treatment).

3.3 The Paired-Samples T-Test

Suppose we wish to determine whether there is a significant difference in the mean number of T-cells for the treatment group 10 minutes after the administration of the treatment and 1 hour later. The comparison calls for the paired-data T-test applied to the *tcells1* and *tcells2* variables for the subjects in the treatment group only. Click *Analyze* in the menu bar, then *Compare Means* from the pull-down menu.

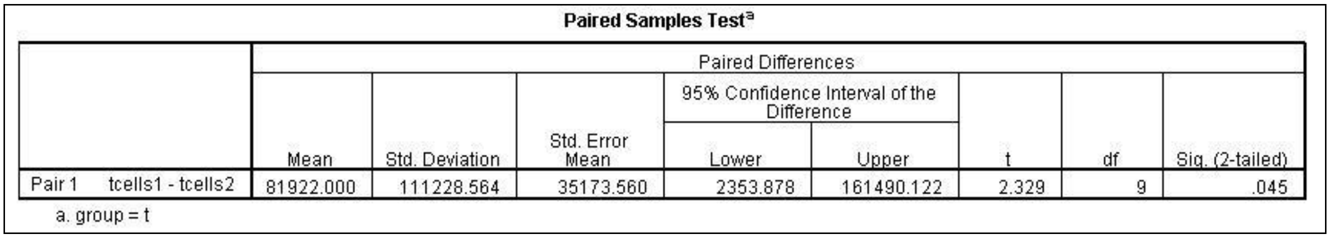


`Then select *Paired Samples T Test* and it opens the *Paired-Samples T Test* dialog box as shown below:



As the procedure should be applied to the subjects in the treatment group only, the file should be split first to allow for the comparison (there is no grouping variable option in the above dialog box). Use the *Split File* function in the *Data* menu to split the file by *group*.

Then open the *Paired-Samples T Test* dialog box and select the *tcells1* and *tcells2* variables and transfer them to the *Paired Variables* box. Then click *OK* to run the procedure. The *Paired Samples Test* table for the treatment group is given below.



Paired Samples Test[a]

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	tcells1 - tcells2	81922.000	111228.564	35173.560	2353.878	161490.122	2.329	9	.045

a. group = t

According to the above output, the value of the test statistics is 2.329, where t follows a t-distribution with 9 degrees of freedom. The two-sided p-value of the test is 0.045. Thus there is strong evidence that the mean T-cells count is different for the observations obtained 10 minutes and 60 minutes after the drug administration. The p-value of one-sided t test is 0.0225.