Two-sample comparison

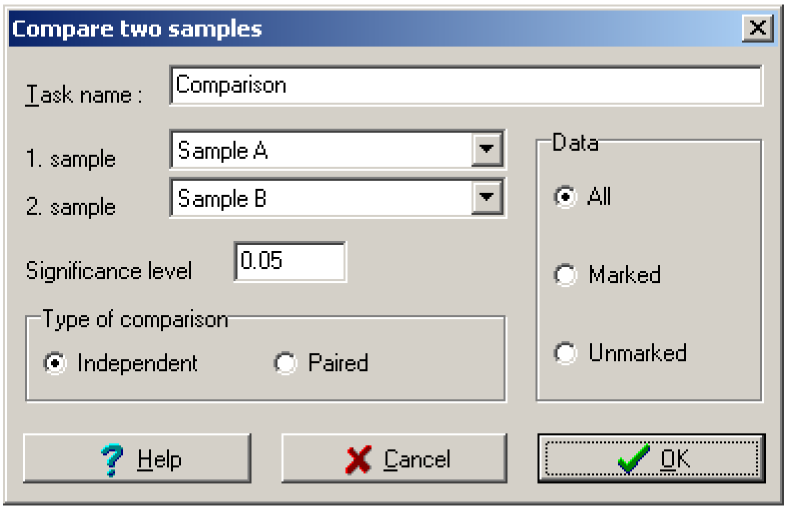
Menu:	QCExpert	Two-sample comparison

This module is intended for a detailed analysis of two datasets (two samples). The module offers two analyses: independent samples comparison, and paired samples comparison.

Independent samples *x*, *y* feature no mutual relationship. They can have different sample sizes, in general. Ordering of the elements of both samples is arbitrary and can be changed without any information loss. Main point of this analysis is to decide, whether the expected values E(*x*) and E(*y*) of the two samples are different. Weight of peanuts from two different locations can serve as an example of two independent samples. On each location, a few dozens of the peanuts are selected at random and weighted individually.

On the contrary, the paired test focuses on comparison of two related datasets, for instance on two sets of measurements, taken on the same units, under different circumstances. Measurements of each unit come in *x*, *y* pairs. The paired test can be performed to decide whether the different conditions influence measurements on the same unit. Technically, the paired comparison goes through the test of whether the expected value of the ***difference*** between first and second variable, E($x - y$) is significantly different from zero. For example, consider comparison of blood cholesterol levels for a group of patients, measured before and after a particular medical treatment. There have to be the same number of pre- and post-treatment measurements (patients who might dropped from the study during the treatment are omitted). Relative ordering of the pre- and post-treatment measurements is important: both measurements of the same patient have to appear on the same line.

Data and parameters



Compare two samples

Task name : Comparison

1. sample: Sample A

2. sample: Sample B

Significance level: 0.05

Type of comparison: ● Independent ○ Paired

Data: ● All ○ Marked ○ Unmarked

Fig. 1 Dialog panel Two-sample comparison

Names of the columns holding values of the first and second variable have to be entered in the dialog panel. In the *Comparison type* part, one has to specify whether the test for *Independent samples* or *Paired samples* is requested. Although the *Significance level* is set to the 0.05 (5%) by default, it can be edited. Similarly as with all other modules, analysis can be requested either for *All* data, or *Marked* data, or *Unmarked* data.

Independent samples
Data are in two columns, whose lengths can be different. Empty cells will be omitted.
Paired samples
Data are in two columns, whose lengths should be the same. If any of the two values in the same row is missing, whole row is omitted.

Protocol

Protocol content is different when independent samples and when paired samples were tested. The same is true for graphical outputs. Both output versions are described below.

Independent samples

Task name	Project name taken from the dialog panel.
Significance level	Required significance level α.
Columns to compare	Names of the columns containing samples to compare.
Sample size	The sample size of first dataset (n_1) and second dataset (n_2).
Average	Arithmetic averages of the first and second column, x_1, x_2.
Standard deviation	Standard deviations of the first and second sample, s_1 a s_2.
Variance	Variance of the first and second sample, s_1^2 and s_2^2.
Correl. coeff. R(x,y)	This entry, together with the warning „Significant correlation!“ will appear only in the case that correlation between the two columns is significant (significantly different from zero) at the significance level α. In such a case, there might be a serious problem with the data and/or their collection procedures, or paired comparison might be called for. If this row is not included in the Protocol, correlation coefficient is not significantly different from zero.
Variance equivalence test	Also called Variance homogeneity test. Tests whether the two sample variances are different. The test is based on approximate normality. Specifically, the data should not contain any outliers. If that is not the case, robust variance estimates should be used instead (see below).
Variance ratio	Test statistic, $\max(\sigma_1^2/\sigma_2^2, \sigma_2^2/\sigma_1^2)$
Degrees of freedom	Degrees of freedom that are used to look up the critical value, i.e. the value of the quantile of the F-distribution with n_1-1 and n_2-1 degrees of freedom
Critical value	F-distribution quantile, $F(\alpha, n_1-1, n_2-1)$
Conclusion	Variance homogeneity test conclusion in words: „Variances are not different“, or „Variances are different“.
p-value	p-value corresponds to the smallest significance level on which the null hypothesis about variance homogeneity were rejected for the given data.
Robust variance test	Alternative variance homogeneity test for two samples. It is intended for non-normal data, mainly those coming from distributions differing from the normal distribution by skewness. The test should not be used for normal data (due to a lower power).
Variance ratio	Test statistic, $\max(\sigma_1^2/\sigma_2^2, \sigma_2^2/\sigma_1^2)$.
Corrected degrees of freedom	Degrees of freedom corrected for the departure from normality.
Critical value	F-distribution quantile.
Conclusion	Variance homogeneity test conclusion in words: „Variances are not different“, or „Variances are different“.
p-value	p-value corresponds to the smallest significance level on which the null hypothesis about variance homogeneity were rejected for the given data.

Mean equivalence test for Equivalent variances	Test of the null hypothesis of equal means in the case of equal variances. When the variances are significantly different, *unequal* variances version of the test needs to be used, see below.
t-statistic	Test statistic.
Degrees of freedom	Degrees of freedom for the t-test.
Critical value	t-distribution quantile.
Conclusion	Test conclusion in words.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about equal means would be rejected for given data.
Mean equivalence test for Different variances	Test of the null hypothesis of equal means in the case of unequal variances. When the variances are not significantly different, *equal* variances version of the test needs to be used, see above.
t-statistic	Test statistic.
Degrees of freedom	t-test degrees of freedom.
Critical value	t-distribution quantile.
Conclusion	Test conclusion in words.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about equal means is rejected for given data.
Goodness of fit test	
Two sample K-S test	Kolmogorov-Smirnov test, comparing distributions generating the two independent samples. It is based on maximum difference between empirical distribution functions (computed from the two samples). Note that it is possible that both means and variances are not significantly different, while the KS test shows significant difference between the distributions. Typically, this is connected to a substantial difference of at least one of the distributions from normality (usually asymmetry or bimodality). Data are not suitable for the simple t-test, then.
Difference DF	Maximal empirical distribution functions difference. It is the test statistic for the KS test.
Critical value	KS-distribution critical value.
Conclusion	Test conclusion in words: „Distributions are significantly different“ or „Distributions are not significantly different“

Paired samples

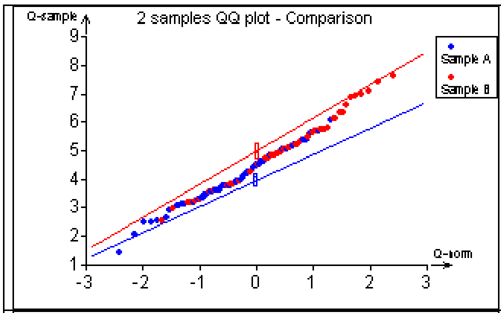
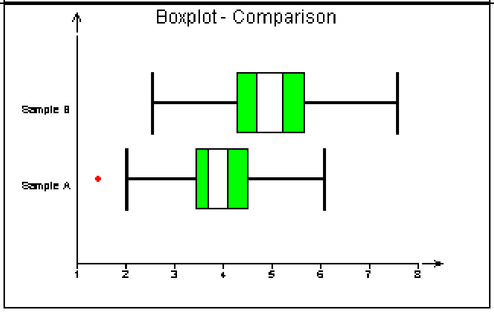
Task name	Project name taken from the dialog panel.
Significance level	Required significance level α.
Columns to compare	Names of the columns containing samples to compare.
Analysis of	

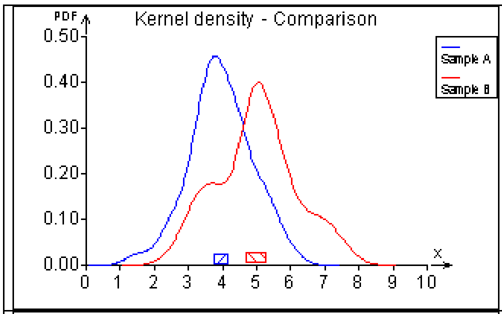
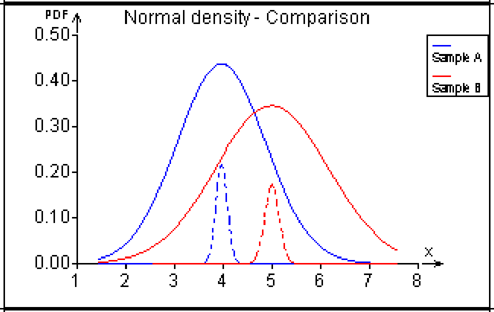
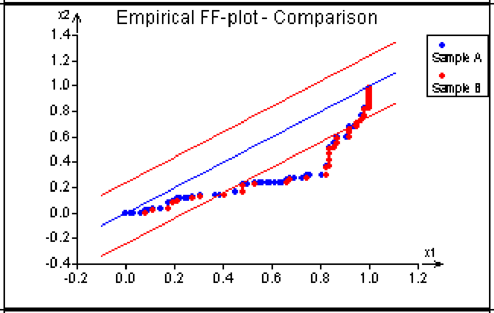
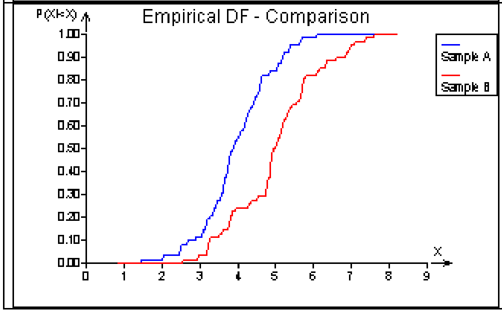
differences	
Sample size	Number of data pairs, *n*.
Average difference	Arithmetic mean of the difference between the first and second variable $x_1 – x_2$ (first of a pair – second of a pair), x_d
Confidence interval	(1-α)% confidence interval for arithmetic mean of differences.
Standard deviation	Standard deviation of the differences, s_d.
Variance	Variance of the differences, s_d^2.
Correlation coefficient R(x,y)	Sample correlation coefficient *r*. It estimates correlation between the first and second data column. When the correlation is not significant, red warning will appear. Paired comparison choice is somewhat suspicious. There might be some problem with the dataset. For instance, relative ordering of the first and second columns might be distorted. Or, the x_1, x_2 pairs come from a box that is too narrow/low.
Test of difference	The test of difference between the first and second pair members.
t-statistic	Test statistic, $x_d \cdot \sqrt{n} / s_d$.
Degrees of freedom	Number of degrees of freedom, $n–1$.
Critical value	
Conclusion	Test conclusion in words. The differences are either „NOT SIGNIFICANTLY different from zero“, or „SIGNIFICANTLY different from zero“.
p-value	*p*-value corresponds to the smallest significance level on which the null hypothesis about mean difference being equal to zero is rejected for given data.

Graphs

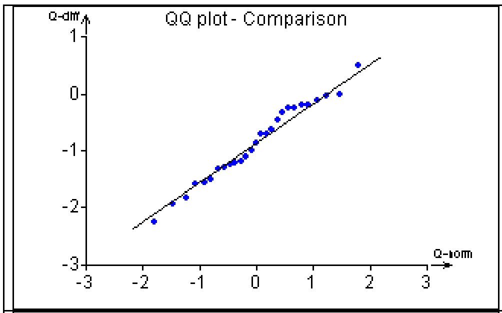
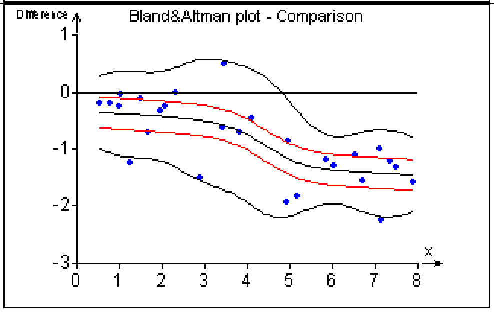
Graphical output is different, according to whether paired or independent samples comparison was choice was selected (similar to the Protocol differences).

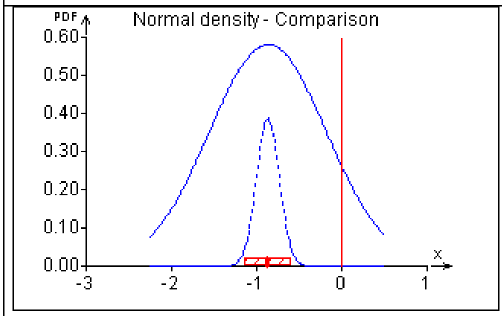
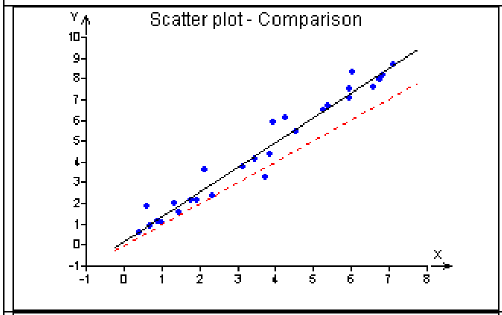
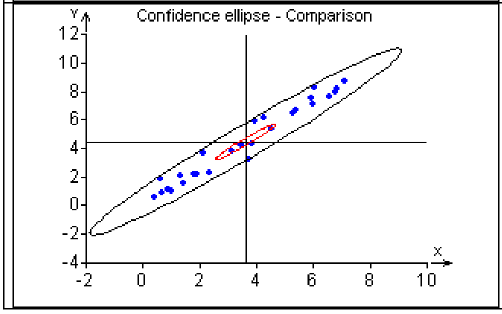
Independent samples

	Q-Q plot for all data. All data are plotted as one sample. First or second sample data are plotted in different colors (see the legend). The two sample means are marked and their confidence intervals are plotted as hatched boxes. Plotted lines' slopes correspond to standard deviations of the two samples. Hence, the steeper line corresponds to sample with a larger standard deviation.
	Boxplots help to compare the samples visually. Larger box contains inner 50% of the data. Right border of the green box corresponds to the 75th percentile. The left border of the green box corresponds to the 25th percentile. Center of the white band corresponds to the median. White band corresponds to the confidence interval for median. Two black whiskers correspond to the so-called inner fences. All data beyond the inner fences are plotted individually, as red points. They are suspicious and can be considered as outliers. Asymmetric placement of the white band in the green box shows data distribution asymmetry.

	Kernel density estimates computed for the two samples separately. Blue curve corresponds to the first sample, while the red curve corresponds to the second sample. Confidence intervals for means are plotted as hatched boxes. When these boxes do not overlap, the means are statistically different on the significance level selected.
	Gauss‘ density curves corresponding to the two samples‘ means and variances. The colors assignment is the same as on the previous plot. For comparison purposes, there are densities for the two arithmetic averages plotted as well (the y-coordinate is shrunken down).
	Joint empirical F-F plot for testing distributional differences between two independent samples. Empirical distribution function values for the first and second samples are plotted as *x* and *y* coordinates. (The empirical distribution functions are shown on the next plot.) If the two distributions are not significantly different, the points are close to the central (blue) line. If the any of the points falls beyond one of the two red lines, then the distributions differ significantly.
	Empirical distribution plot for the first and second sample. The *Y* coordinate corresponds to the distribution function value (i.e. to the probability that there is a measurement smaller than or equal to the value of the X coordinate).

Paired samples

	Q-Q plot for checking normality of differences between first and second member of the pair graphically. If the points are placed close to the line, the data do not look non-normal. When this is not the case, substantial departure from normality is suggested. Information content of the tests reported in the Protocol can be seriously impaired then.
	Bland and Altman plot. Average of a particular data pair is plotted on the X-axis, while the difference for the same pair is plotted on the Y-axis. This plot helps to detect possible dependence between variability (estimated by the pair members difference) and value attained. For a better orientation, horizontal zero line is added to the plot. Smoothed average of the differences is plotted as a function of average (black curve). Corresponding confidence interval of this estimate is plotted, with the two red curves. Smoothed value

	±2σ, where σ is estimated nonparametrically and it is plotted as two black curves. Ideally, (when the difference is uniformly zero for any value of average), the red curves should contain horizontal zero line, and the ±2σ band should be approximately linear, parallel to the horizontal zero line.
	Gaussian density curve. The parameters are estimated from the differences between pair members under the normality assumption. The inner curve corresponds to the approximate density of the arithmetic mean of the difference (*Y* coordinate is shrunken for better readability). Vertical red line corresponds to zero difference. The hatched box corresponds to the mean difference confidence interval. If zero is contained in the interval, then the mean difference between first and second sample is not significantly different from zero.
	This plot is useful when judging degree of interdependence between the first and second sample. *y*=*x* line is plotted in red (dashed). It corresponds to the zero difference. The second line corresponds to the best *y*-depends-on-*x* line, fitted to the data.
	The same points as on the previous plot are plotted here. The data points here are viewed as a sample from a bivariate normal distribution, however. Black ellipse corresponds to the region containing approximately 100.(1-α)% of the data (under bivariate normality). The red ellipse corresponds to the border of the 100.(1– α)% confidence region for the vector of two means.