Basic statistics

Menu:	QCExpert	Basic statistics

The basic statistics module is useful for a preliminary data analysis, as well as for a more detailed look at the data. Various tools from this module can also be used to test whether the data are consistent with assumptions needed for a successful application of other statistical methods. Some of the basic and common assumptions about data are: normality, independence and homogeneity. Therefore, no outliers and gross errors should occur in the data.

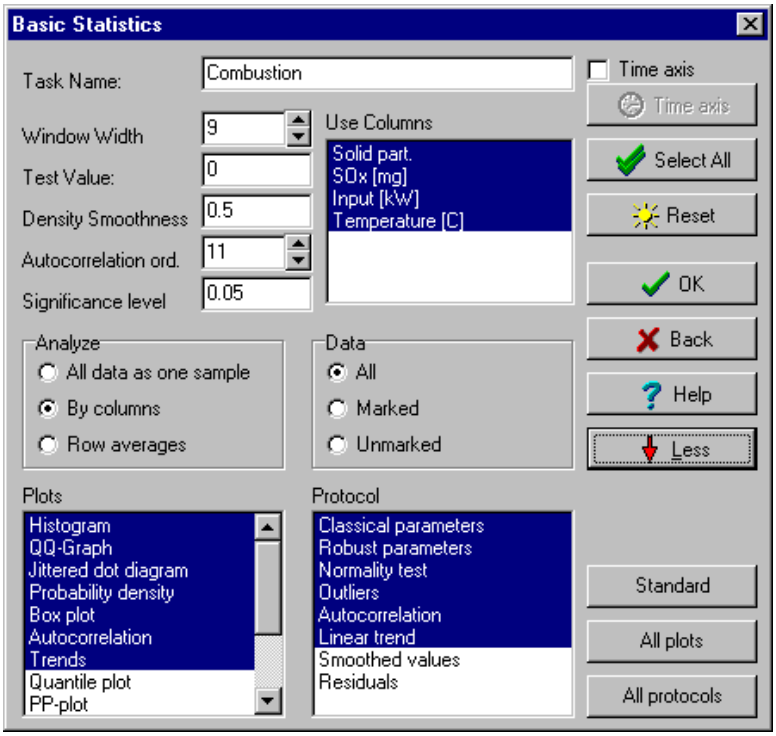


Fig. 1 Basic data analysis dialog panel

Data and parameters

Data are organized into columns (variables). The first row always contains column names. When "All data" or "Columns" choices are selected in the Basic data analysis dialog panel, columns of different lengths are allowed. When computations are requested for "Subgroup means", length of all columns should be the same. Minimum number of columns is 1. Minimum number of data points is 3. Columns to be analyzed can be selected in the *Columns* window inside the *Basic data analysis* panel, see Figure 17. Various other parameters can be set there as well.

Trend order determines how many consecutive data points will be used to compute moving averages and moving medians. The value should be smaller than half of the sample size.

Test the mean value The value μ_0 for a t-test is entered here. The program tests whether the true mean of the data is different from μ_0 at a specified significance level.

Density smoother The kernel width of kernel smoother is entered here. The smoother is used to estimate probability density function. High values of the parameter result in smoother probability density estimate and vice versa. The parameter has to be positive, a value about 0.5 is recommended.

Autocorrelation order gives the maximum lag for which the autocorrelation coefficient is computed. The value has to be smaller than the sample size minus 2.

Significance level gives significance level for statistical tests and confidence level for confidence intervals. It has to be positive and smaller than 0.5. The parameter multiplied by 100 gives the value in percent. A commonly used value is 0.05 (i.e. 5%).

Computations done for:
All data Data from all selected columns will be used for computations as if they came from a single column.
Columns Computations will be applied for each of the selected columns separately.
Subgroup means Row means for selected columns are computed. If columns differ in length or they have missing values, the computation is performed for complete rows only. This computation is most useful for data diagnostics and X bar control charts.
Time axis window is checked when there is a column containing time values among the data. The time column is identified by pressing the *Time axis* button.
Select all selects all columns in the active sheet for further computations.
Default values This button can be used when one is not sure whether previously entered parameter values are correct. When the button is pressed, parameters in the dialog panel are set to default values.
More/Less button specifies amount of output requested. Requested amount of both printed and graphical output is specified here. Pressing the *Standard* button produces the usual (reduced) amount of output containing the most important information only. *All graphs* and *All protocols* buttons are used to request the complete output.

Note: the size of objects produced when *Smoothed values* and *Residuals* items are selected depends on the sample size. They can fill the output sheet completely for large data sets.

Protocol

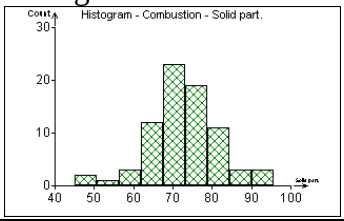
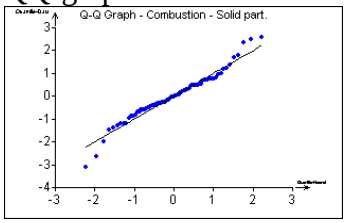
Column	Column name.
Row number	Total number of rows in the analyzed dataset.
Number of valid data points	Total number of valid data points in the dataset.
Number of missing values	Number of empty cells in the dataset.
Classical parameters	
Arithmetic mean	Mean value estimate for normal data.
Lower limit	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on arithmetic mean.
Upper limit	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on arithmetic mean.
Variance	Variance estimate.
Standard deviation	Square root of the variance estimate.
Skewness	Estimate of the third moment, skewness.
Difference from 0	Skewness for normal as well as other symmetrical distributions is zero. When the skewness is significantly different from 0, the data cannot be assumed to come from a symmetrical distribution. The test for normality (see later) is more useful.
Kurtosis	Estimate of the fourth moment, kurtosis.
Difference from 3	Kurtosis for normal distribution is 3. When the kurtosis is significantly different from 3, the data should be assumed to come from a non normal distribution. The test for normality (see later) is more useful
Half-sum	Half sum estimate, i.e. half of (maximum plus minimum)
Modus	Estimate of the modus, i.e. location of the probability density

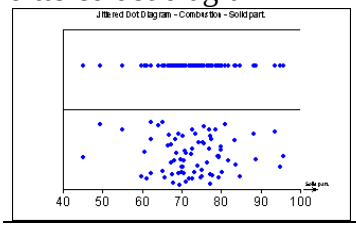
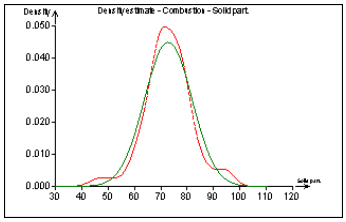
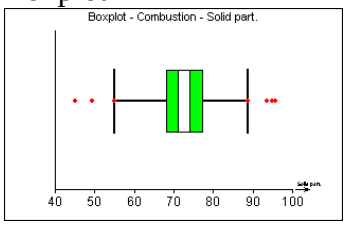
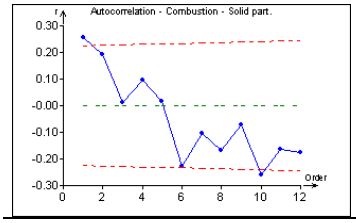
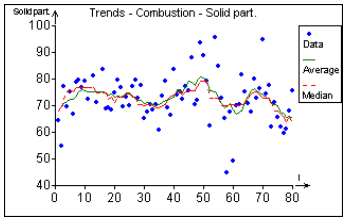
	function maximum.
t-test	
Hypothesized value	The value, entered to the "Test the mean value" field from the "Basic data analysis" panel.
Difference from the hypothesized value.	A comment, describing in words whether mean value is significantly different from the hypothesized value at specified significance level. The hypothesized value can be entered in the Basic data analysis panel, see Figure 17.
Calculated	Calculated test criterion.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the equality of true mean to the hypothesized value is rejected when using the observed data. When the p-value is smaller than a selected significance level, the true and hypothesized mean are significantly different.
Robust parameters	
Median	Estimate of the median, i.e. 50^{th} percentile. It might be a more useful estimate of the mean value than the arithmetic mean, when normality does not hold or when outliers are present in the analyzed data.
CI lower	Lower limit of the confidence interval for the median, computed for a specified confidence level.
CI upper	Upper limit of the confidence interval for the median, computed for a specified confidence level.
Standard deviation	Median based standard deviation.
Variance	Median based variance.
10% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 5% smallest and 5% largest values. This robust estimate of the mean is recommended when outliers are expected.
CI lower	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Variance estimate based on median.
Standard deviation	Standard deviation estimate based on median.
20% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 10% smallest and 10% largest values. This robust estimate of the mean is recommended when suspecting outliers presence.
CI lower	Lower limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Median based variance.
Standard deviation	Median based standard deviation.

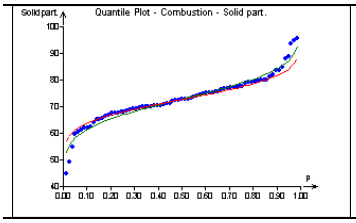
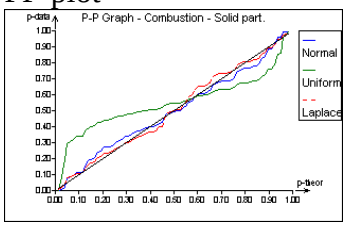
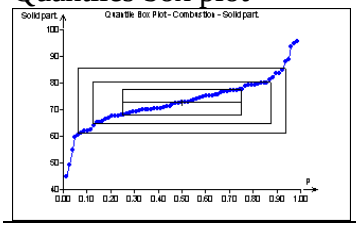
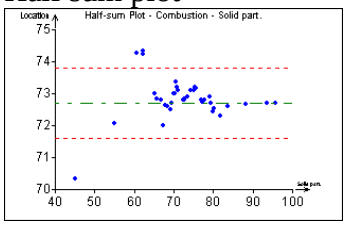
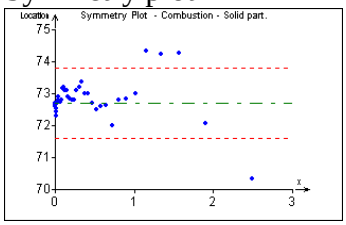
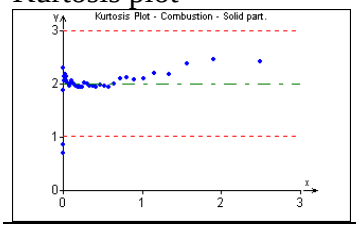
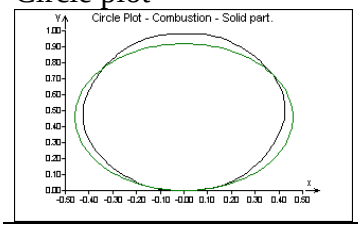
40% trimmed mean	The arithmetic mean computed from symmetrically trimmed data, i.e. after omitting 20% smallest and 20% largest values. This robust estimate of the mean is recommended when suspecting outliers presence.
CI lower	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
CI upper	Upper limit of the confidence interval for the mean value. It is computed for a specified confidence level and based on the trimmed mean.
Variance	Median based variance estimate.
Standard deviation	Median based standard deviation.
Runs test	Test for randomness in the sequence of values larger and smaller than the mean. When the larger/smaller pattern is too regular or when the differences from the mean form long runs with many consecutive values of the same sign, the data are suspect for the lack of independence. Such data are considered to be dependent.
Small sample analysis	Mean value estimate and confidence interval computed by the Horn's quantile based method. This estimator is recommended for small samples of 3 and more data points. The method usually produces more correct values than the arithmetic mean for such small sample sizes. The method should not be used for N>20.
N	Sample size.
Mean value	Estimate of the mean value.
Lower limit	Lower limit of the confidence interval.
Upper limit	Lower limit of the confidence interval.
Test for normality	A test for checking normality, based on both skewness and kurtosis. The output contains values of classical statistical characteristics
Normality	Conclusion of the test performed at a specified significance level, described in words.
Calculated	Calculated test criterion.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the normality is rejected using the observed data.
Outliers	A robust, quantile based test procedure to check whether outliers are present.
Homogeneity	Conclusion of the test for outlier presence, commented in words.
Number of outliers found	The number of data points, which can be considered as outliers, based on the previous test.
Lower limit	The data below this limit are considered to be outliers.
Upper limit	The data exceeding this limit are considered to be outliers.
Autocorrelation	Autocorrelation estimates and related tests performed at a selected significance level.
Autocorrelation order	Autocorrelation order.
Coefficient	Autocorrelation coefficient estimate. It is the correlation

	coefficient for a pair of variables, discussed in the paragraph 5.3., computed for a special choice of the two variables.
p-value	The smallest significance level for which the zero autocorrelation is rejected using the observed data. When the p-value is smaller than a specified significance level, the autocorrelation is significant.
R0Crit	Critical value for the autocorrelation test. Autocorrelation estimates larger than this value are significant.
Result	Conclusion of the autocorrelation test, described in words.
Test for a linear trend	Test for linear trend in the data. Even when the linear trend is not significant, a trend of nonlinear character might still be present in the data, e.g. periodic oscillations. Some nonlinear trends might be detected by other tests, e.g. by the runs test.
Slope	Slope of the fitted line.
Significance	Conclusion of the linear trend test in terms of slope significance, commented in words.
Theoretical	Appropriate t-distribution quantile.
p-value	The smallest significance level for which the hypothesis of no linear trend is rejected using the observed data. When the p-value is smaller than a specified significance level (usually 0.05), the trend is considered to be present (significant).
Smoothed values	Smoothed (detrended) values obtained by running means or running medians. With an appropriate choice of the running mean/median length a trend can be estimated by the smoother. Detrended data can be obtained by subtracting the trend estimate from the original data. When showing a strong trend, data should be appropriately detrended before they are used in control charts.
Residuals	Residuals, i.e. the differences between observed and smoothed data.

Graphs

Histogram 	Frequency histogram with a constant bin width; optimum number of bins is selected automatically, with respect to the sample size. Clicking the right mouse button on a bar in the dynamical graph shows frequency and limits of the class selected.
QQ-graph 	A graphical tool for checking normality and outliers presence; for normal data without outliers the points should lie close to the line; for normal data with outliers, points in the central parts should lie close to the line the endpoints further away from the line; for data coming from a positively skewed distribution (e.g. lognormal, exponential) the shape should be nonlinear, convex ⌊; for data coming from a negatively skewed distribution the shape should be nonlinear, concave ⌈; for data coming from a distribution with kurtosis higher than normal, i.e. those showing high concentration around the mean (e.g. Laplace), the shape should be concave-convex ⌊⌈; for data coming from a distribution with kurtosis smaller than normal, i.e. those with small concentration around the mean

	(e.g. uniform), the shape should be convex-concave ⌒. One advantage of the QQ-graph, compared to the statistics describing skewness, kurtosis etc. is that one can visually check whether the lack of normal appearance (nonlinearity) is caused by just a few points or whether it is a general tendency shared by all data.
Jittered dot diagram 	Data are plotted along the x-axis. y-axis has no physical meaning. The y-coordinates of the plotted points are random. This technique allows for a better recognition of individual data points which might coincide if plotted just along the x-line when their observed values are similar or the same.
Probability density function 	Comparison of the normal probability density curve (solid green line) with a kernel density estimate, computed from the data (dashed red line). The kernel estimate uses the Gaussian kernel. Smoothness of the estimate is given be the kernel width, entered as "Density smoother" parameter in the dialog panel shown in Figure 18. Smaller parameter values cause more rugged shape of the estimate (following more details of the data). When the data are not homogeneous and show a clustering tendency, several local maxima of the density estimate can occur. For normal data, both curves should be close to each other. On the other hand, one should realize that with small enough smoothing parameter, local maxima occur for any data.
Boxplot 	This is a standard diagnostic tool. The large box contains 50% of the data, its upper edge corresponds to 75th percentile, its lower edge to the 25th percentile. Median is located in the middle of the white rectangle inside the green box. Width of the white rectangle inside the green box corresponds to the width of the confidence interval for the median. Two black lines correspond to the inner fence. The data points outside the inner fence are marked red. They might be considered as outliers.
Autocorrelation 	Plot of autocorrelation coefficients against lag up to the maximum lag, specified in the dialog panel, see the Figure 17. Red lines show the limits beyond which the coefficient is considered to be significant at a previously selected significance level. When some of the coefficients exceed these limits, the data should be considered to be dependent.
Trends 	Plot of the smooth trend in the data, estimated by the running mean smoother (solid line) and the running median smoother (dashed line). The "Trend order" parameter is inputted in the dialog panel shown in Figure 17. Higher values of the parameter result in smoother curves, less sensitive to local behavior of the data, showing a global trend. Small parameter values lead to curves sensitive to a local data behavior. The running median smoother is less sensitive to errors and isolated outliers in the data (it is robust), so that it is recommended in cases when such problems are expected. When the linear trend test yields a significant result, the regression line is plotted as well
Quantile plot	Shows empirical quantiles and the inverse cumulative distribution function (the quantile function, QF) of the fitted normal distribution. The green curve corresponds to the normal QF with the classical estimates of the parameters (non-robust), the red curve corresponds to the median based estimates of the parameters (robust). Depending on which of the curves fits the data better, either mean or median might be chosen as an estimate

	of the mean value.
PP-plot 	Compares data distribution with several theoretical models, using the empirical cumulative distribution function and cumulative distribution function of normal (solid blue curve), Laplace, and uniform distributions. A model which fits data well should plot approximately as the $y = x$ line. The plot can be used to distinguish among symmetrical distributions according to their kurtosis. Apparent similarity to the uniform distribution suggests that the data were truncated (both small and large values excluded).
Quantiles box plot 	Points are plotted in the same way and have the same meaning as for the quantile graph. Relative position of the plotted rectangles show symmetry resp. asymmetry of the data distribution. The horizontal line inside the smallest rectangle corresponds to median, the vertical edge of the smallest rectangle corresponds to the confidence interval for median.
Half sum plot 	A sensitive indicator for distributional asymmetry. Ideally, the points should lie on a horizontal line. Green horizontal line corresponds to median and dashed red lines to its confidence limits. When the data distribution is asymmetric, the plot shows a clear trend (increasing for a negative skewness and decreasing for a positive skewness), going far beyond the dashed lines. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.
Symmetry plot 	It has a similar use as the Half sum plot from previous paragraph. Slope of a trend is proportional to skewness. When the data distribution is asymmetric, the plot shows a clear trend (increasing for a negative skewness and decreasing for a positive skewness), going far beyond the dashed lines. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.
Kurtosis plot 	The meaning is analogous to the previous two plots. Slope of its trend is proportional to the difference (kurtosis-3). When the kurtosis is very different from normal, the plot shows a clear trend. Pairs of data points (first-last, second-second largest, etc.) are used when constructing the plot, so when selecting a point on the plot, two data points are marked in the data table.
Circle plot 	It is used for a complex visual assessment of normality, considering skewness and kurtosis simultaneously. Green circle (ellipse) is an ideal (for a normal distribution), black “circle” is constructed from data. Both curves should be close to each other for normal data.