Outlier Identification

Summary

The **Outlier Identification** procedure is designed to help determine whether or not a sample of *n* numeric observations contains outliers. By "outlier", we mean an observation that does not come from the same distribution as the rest of the sample. Both graphical methods and formal statistical tests are included. The procedure will also save a column back to the datasheet identifying the outliers in a form that can be used in the *Select* field on other data input dialog boxes.

Sample StatFolio: *outlier.sgp*

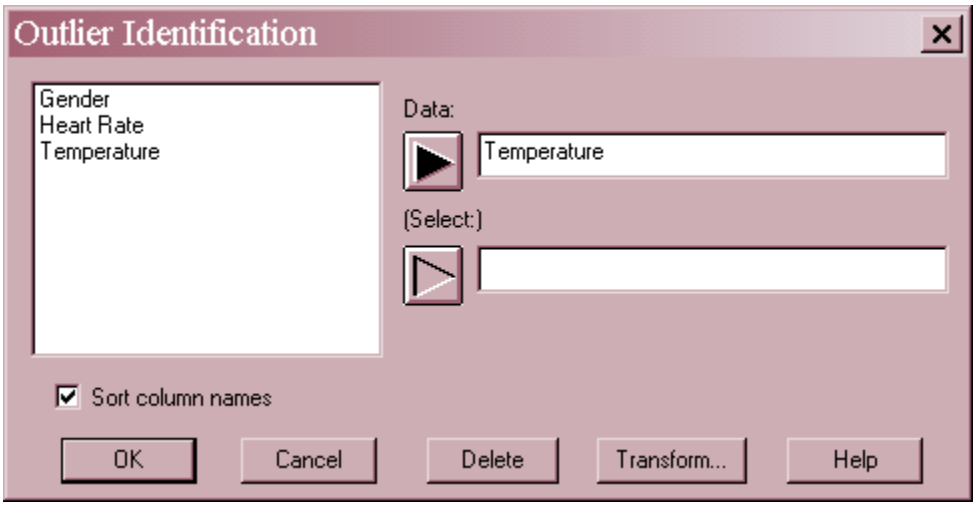
Sample Data:

The file *bodytemp.sgd* file contains data describing the body temperature of a sample of $n = 130$ people. It was obtained from the Journal of Statistical Education Data Archive (www.amstat.org/publications/jse/jse_data_archive.html) and originally appeared in the Journal of the American Medical Association. The first 20 rows of the file are shown below.

Temperature	*Gender*	*Heart Rate*
98.4	Male	84
98.4	Male	82
98.2	Female	65
97.8	Female	71
98	Male	78
97.9	Male	72
99	Female	79
98.5	Male	68
98.8	Female	64
98	Male	67
97.4	Male	78
98.8	Male	78
99.5	Male	75
98	Female	73
100.8	Female	77
97.1	Male	75
98	Male	71
98.7	Female	72
98.9	Male	80
99	Male	75

Data Input

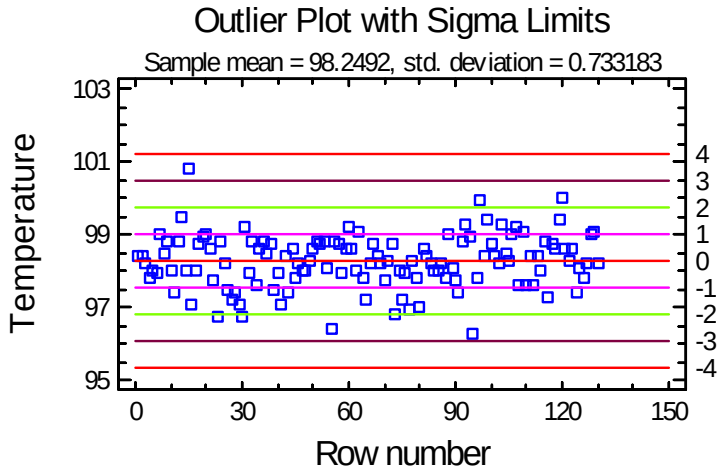
The data to be analyzed consist of a single numeric column containing $n = 2$ or more observations.



- **Data :** numeric column containing the data to be summarized.
- **Select:** subset selection.

Outlier Plot

A good place to begin when considering the possibility that a sample of n observations contains one or more outliers is the *Outlier Plot*.



This plot shows each data value together with horizontal lines at the sample mean plus and minus 1, 2, 3, and 4 standard deviations. Points beyond 3 sigma, of which there is one in the plot above, are usually deemed to be potential outliers and worthy of further investigation.

Analysis Summary

The *Analysis Summary* displays a number of statistics designed to be resistant to outliers, as well as the result of several formal outlier tests. The top section of the display is shown below:

Outlier Identification - Temperature
Data variable: Temperature
130 values ranging from 96.3 to 100.8
Number of values currently excluded: 0

Location estimates

Sample mean	98.2492
Sample median	98.3
Trimmed mean	98.2714
Winsorized mean	98.25

Trimming: 15.0%

Scale estimates

Sample std. deviation	0.733183
MAD/0.6745	0.74129
Sbi	0.714878
Winsorized sigma	0.708916

95.0% confidence intervals for the mean

	Lower Limit	*Upper Limit*
Standard	98.122	98.3765
Winsorized	98.1032	98.3968

Location estimates

Four statistics are provided that estimate the center or location of the population from which the data were sampled, including:

1. *Sample mean* – the arithmetic mean of the sample.

2. *Sample median* – the center or middle value of the sample.

3. *Trimmed mean* – the average value after dropping a specified percentage of the smallest and largest observations.

4. *Winsorized mean* – the average value after replacing a specified percentage of the smallest and largest observations with the most extreme values not within that percentage.

If the data come from a normal distribution, each of the four statistics estimates the population mean μ. However, the last 3 statistics are each less sensitive to the possible presence of outliers than the ordinary sample mean. In the current example, there is very little difference between the estimates. However, such is not always the case.

Scale Estimates

There are also four estimates of the dispersion of the data, each of which estimates the standard deviation σ provided the data come from a normal distribution:

1. *Sample standard deviation* – the usual standard deviation.

2. *MAD/0.6745* – an estimate based on the median absolute deviation (the median of the absolute differences between each data value and the sample median).

3. *Sbi* – an estimate based on a weighted sum of squares around the sample median, where the weights decrease with distance from the median.

4. *Winsorized sigma* – an estimate based on the squared deviations around the Winsorized mean.

The latter 3 estimates are designed to be resistant to outliers. For the current data, the estimates are very similar.

Confidence Intervals

Confidence intervals for the mean μ are displayed based on the usual sample mean and standard deviation and also using the Winsorized statistics. The fact that the intervals are so close implies that outliers are not a major problem in this data.

Extreme Values

The middle section of the table shows the 5 largest and 5 smallest observations in the data:

Sorted Values

		Studentized Values	*Studentized Values*	*Modified*
Row	*Value*	*Without Deletion*	*With Deletion*	*MAD Z-Score*
95	96.3	-2.65859	-2.74567	-2.698
55	96.4	-2.52219	-2.59723	-2.5631
23	96.7	-2.11302	-2.15912	-2.1584
30	96.7	-2.11302	-2.15912	-2.1584
73	96.8	-1.97663	-2.01521	-2.0235
...				
99	99.4	1.56955	1.59096	1.4839
13	99.5	1.70594	1.7323	1.6188
97	99.9	2.25151	2.30628	2.1584
120	100.0	2.3879	2.45231	2.2933
15	100.8	3.47903	3.67021	3.3725

The 3 rightmost columns show standardized values or *Z-Scores* that may be used to help identify outliers. Each statistic measures how many standard deviations the data values are from the center of the data.

Studentized values without deletion - using the sample mean and standard deviation, each data value is standardized by

$$t_i = \frac{x_i - \bar{x}}{s} \qquad (1)$$

These values measure the number of standard deviations each value lies from the sample mean and correspond to the right axis scale on the outlier plot. Grubbs' test, described below, is based on the most extreme Studentized value, which in this case equals 3.479.

Studentized values with deletion - each data value is removed from the sample one at a time and the mean $\bar{x}_{[i]}$ and standard deviation $s_{[i]}$ are calculated using the remaining n - 1 data values. Each data value is then standardized by

$$t_i = \frac{x_i - \bar{x}_{[i]}}{s_{[i]}} \qquad (2)$$

These values measure the number of standard deviations each value lies from the sample mean when that data value is not included in the sample. This is similar to the calculation of Studentized deleted residuals used in the regression procedures. The importance of deleting each observation prior to standardizing it is that a strong outlier, particularly in a small sample, can have such a big impact on the sample mean and standard deviation that it does not appear to be unusual.

Modified MAD Z-score - each data value is standardized by

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \qquad (3)$$

These values use the estimate of sigma based on the median absolute deviation (MAD). Iglewicz and Hoaglin (1993) suggest that any data value for which $|M_i|$ is greater than 3.5 be labeled an outlier, which is the rule used by the StatAdvisor in interpreting the results.

Grubbs' Test

The final section of the output shows the result of one or more formal outlier tests:

Grubbs' Test (assumes normality)
Test statistic = 3.47903
P-Value = 0.0484379

The first test is due to *Grubbs* and is calculated if $n \geq 3$. Also called the *Extreme Studentized Deviate Test* (ESD), it is based on the largest Studentized value (without deletion) t_{max}. The test statistic T is computed according to

$$T = \sqrt{\frac{n(n-2)t_{\max}^2}{(n-1)^2 - nt_{\max}^2}} \qquad (4)$$

An approximate two-sided P-Value is obtained by computing the probability of exceeding |T| based on Student's t-distribution with n - 2 degrees of freedom and multiplying the result by $2n$. A small P-value leads to the conclusion that the most extreme point is indeed an outlier. For small samples, one can refer instead to Iglewicz and Hoaglin (1993) who give 5% and 1% values for t_{max} in Appendix A of their monograph, as well as for a generalized test involving r > 1 potential outliers.

In the sample data, row 15 is the most extreme point, with a Studentized value equal to nearly 3.5. Since the P-Value is less than 0.05, that point may be declared to be a statistically significant outlier at the 5% significance level. This conclusion is made subject to the assumption of Grubb's test that all other data values come from a normal distribution.

Dixon's Test

For small samples with $4 \le n \le 30$, *Dixon's Test* is also performed. This test begins by ordering the data values from smallest to largest. Letting $x_{(j)}$ denote the j-th smallest data value, statistics are then computed to test for 5 potential situations:

Situation 1: **1 outlier on the right**. Compute:

$$r = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} \tag{5}$$

Situation 2: **1 outlier on the left**. Compute:

$$r = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \tag{6}$$

Situation 3: **2 outliers on the right**. Compute:

$$r = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} \tag{7}$$

Situation 4: **2 outliers on the left**. Compute:

$$r = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}} \tag{8}$$

Situation 5: **1 outlier on either side**. Compute:

$$r = \max\left[\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}\right] \tag{9}$$

The calculated statistic *r* is then compared to critical values in tables such as Appendix A.3 of Iglewicz and Hoaglin (1993). For each test, STATGRAPHICS indicates whether or not the result is statistically significant at the 5% level and at the 1% level. A significant result indicates the presence of the hypothesized situation.

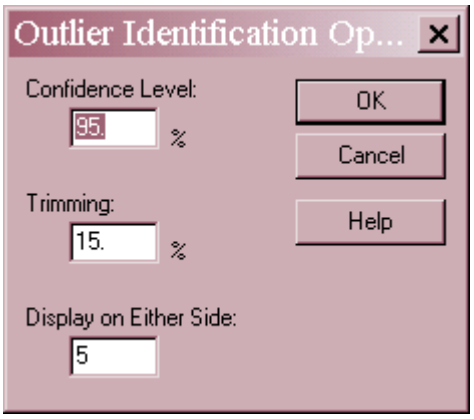
For example, arbitrarily selecting the first 30 rows of the data file, the following table is displayed:

Dixon's Test (assumes normality)

	Statistic	*5% Test*	*1% Test*
1 outlier on right	0.317073	Significant	Not sig.
1 outlier on left	0.0	Not sig.	Not sig.
2 outliers on right	0.439024	Significant	Significant
2 outliers on left	0.142857	Not sig.	Not sig.
1 outlier on either side	0.317073	Significant	Not sig.

Significant results are obtained at the 5% significance level for the hypothesis that 1 large outlier exists on the right, that 2 large outliers exist on the right, and that 1 large outlier exists on either side. When using this test, you should select the hypothesis of interest before looking at the test results.

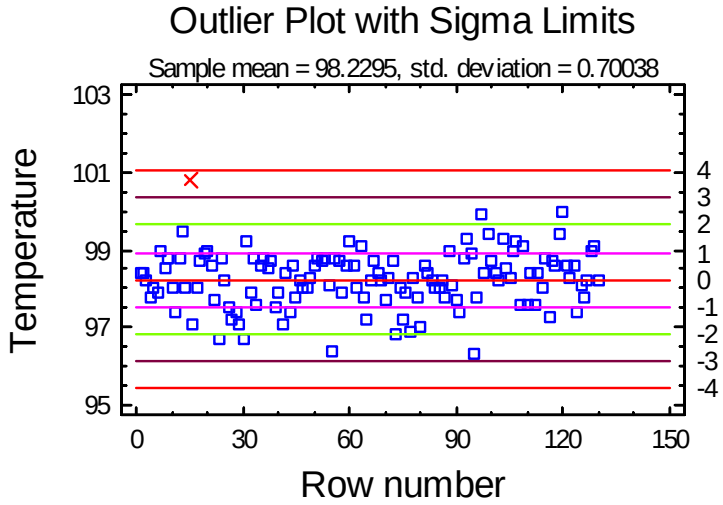
Analysis Options



- **Confidence Level:** level used to calculate the confidence intervals.
- **Trimming:** the percent of the data trimmed from each end when computing the trimmed mean and Winsorized statistics.
- **Display on Each Side:** the number of most extreme large and small values to include in the table.

Excluding Outliers

Data values that are determined to be outliers may be excluded graphically by clicking on the points in the *Outlier Plot* and then clicking on the *Exclude/Include* button on the analysis toolbar.



The excluded points will be marked by an X and all statistics throughout the procedure recalculated without that data. For example, *Grubbs' Test* now shows a very insignificant P-Value for the most extreme value in the remaining data:

Grubbs' Test (assumes normality)
Test statistic = 2.75487
P-Value = 0.676064

Summary Statistics

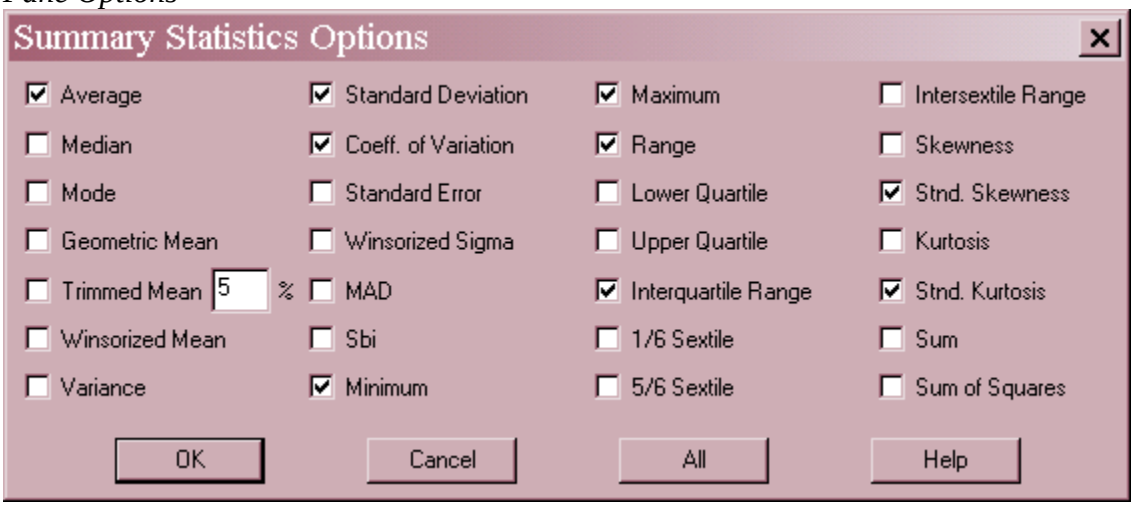
The *Summary Statistics* pane calculates a number of different statistics that are commonly used to summarize a sample of *n* observations:

Summary Statistics for Temperature

Count	130
Average	98.2492
Standard deviation	0.733183
Coeff. of variation	0.746248%
Minimum	96.3
Maximum	100.8
Range	4.5
Interquartile range	0.9
Stnd. skewness	-0.0205699
Stnd. kurtosis	1.81642

The statistics included in the table by default are controlled by the settings on the *Stats* pane of the *Preferences* dialog box. Within the procedure, the selection may be changed using *Pane Options*. Of particular interest here are the standardized skewness and standardized kurtosis. Both of these statistics should be between –2 and +2 if the data come from a normal distribution. Since this is an assumption of the outlier tests, you should check these values after excluding the outliers.

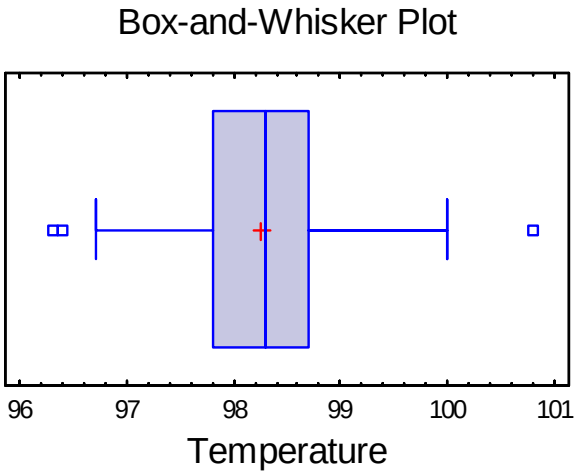
Pane Options



Select the statistics to be displayed. The meaning of each statistic is described in the documentation for the *One Variable Analysis* procedure.

Box-and-Whisker Plot

This pane displays the box-and-whisker plot.



The plot is constructed in the following manner:

- A box is drawn extending from the *lower quartile* of the sample to the *upper quartile*. This is the interval covered by the middle 50% of the data values when sorted from smallest to largest.

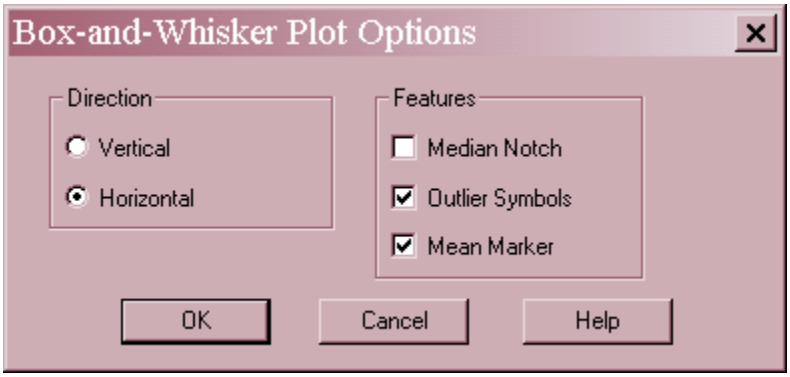
- A vertical line is drawn at the *median* (the middle value).

- If requested, a plus sign is placed at the location of the sample mean.

- Whiskers are drawn from the edges of the box to the largest and smallest data values, unless there are values unusually far away from the box (which Tukey calls *outside points*). Outside points, which are points more than 1.5 times the interquartile range (box width) above or below the box, are indicated by point symbols. Any points more than 3 times the interquartile range above or below the box are called *far outside points*, and are indicated by point symbols with plus signs superimposed on top of them. If outside points are present, the whiskers are drawn to the largest and smallest data values which are not outside points.

The above plot for the body temperature data is very symmetric. The plus sign for the mean lies very close to the line for the median, while the whiskers are of approximately equal length. There are 3 outside points. When sampling 130 observations from a normal distribution, outside points can be expected to occur just by chance about half the time, but usually only one or two. Far outside points, of which there is none, occur extremely rarely.

Pane Options



- **Direction**: the orientation of the plot, corresponding to the direction of the whiskers.
- **Median Notch**: if selected, a notch will be added to the plot showing an approximate 100(1-α)% confidence interval for the median at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu).
- **Outlier Symbols**: if selected, indicates the location of outside points.
- **Mean Marker**: if selected, shows the location of the sample mean as well as the median.

Tests for Normality

Several formal tests for normality are performed and the results displayed in the *Tests for Normality* pane.

Tests for Normality

Test	*Statistic*	*P-Value*
Chi-Squared	54.0154	0.000424234
Shapiro-Wilks W	0.986473	0.821435
Skewness Z-score	0.0151112	0.987938
Kurtosis Z-score	1.64492	0.0999861

Each of the tests is based on the following set of hypotheses:

H_0: data come from a normal distribution

H_A: data do not come from a normal distribution

Small P-Values (less than 0.05 if operating at the 5% significance level) lead to a rejection of the hypothesis of normality.

The four tests, details of which are given in the documentation on *Distribution Fitting (Uncensored Data)*, are the following:

- **Chi-Square Test** - divides the data into non-overlapping classes and calculates a statistic based on the differences between the observed frequencies in each class and

the expected frequencies if the data came from a normal distribution. This test should *not* be used if the data is heavily rounded, as in the current example, since the discrete nature of the data may easily distort the results.

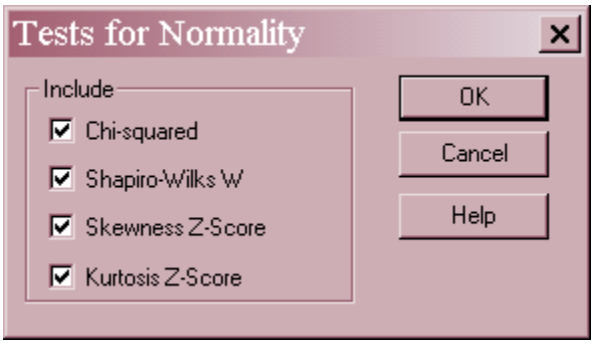
- **Shapiro-Wilks W** – available when $2 \le n \le 2000$, this test compares the fit of the least squares regression line to the data on the normal probability plot.

- **Z-score for skewness** – performs a test based on the estimated skewness in the data.

- **Z-score for kurtosis** – performs a test on the estimated kurtosis in the data.

Except for the chi-squared test, whose behavior can be explained by the fact that the data were rounded to the nearest tenth of a degree, there is no evidence to reject the hypothesis that the body temperatures follow a normal distribution.

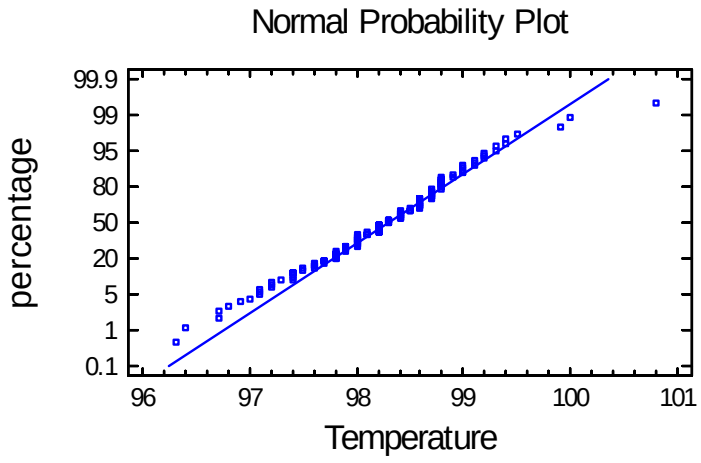
Pane Options



- **Include:** select one or more tests to perform.

Normal Probability Plot

The *Normal Probability Plot* displays the data from smallest to largest in a manner that makes it possible to judge whether or not the data come from a normal distribution.



The vertical axis is scaled in such a way that, if the data come from a normal distribution, the points should lie approximately along a straight line. In constructing the plot, the points are plotted at coordinates equal to

$$\left(x_{(j)}, \Phi^{-1}\left(\frac{j - 0.375}{n + 0.25} \right) \right) \tag{10}$$

where $\Phi^{-1}(u)$ represents the inverse standard normal distribution evaluated at u. The labels along the vertical axis equal 100u%, for values of u ranging between 0.001 and 0.999.

In order to help determine how closely the points correspond to a straight line, a reference line is superimposed on the plot corresponding to a normal distribution with mean μ and standard deviation σ. There are two options for fitting the line:

1. Using the median and the sample quartiles:

$$\hat{\mu} = \text{sample median} \tag{11}$$

$$\hat{\sigma} = \text{interquartile range / 1.35} \tag{12}$$

2. Fitting a least squares regression of the normal quantiles on the sorted data values.

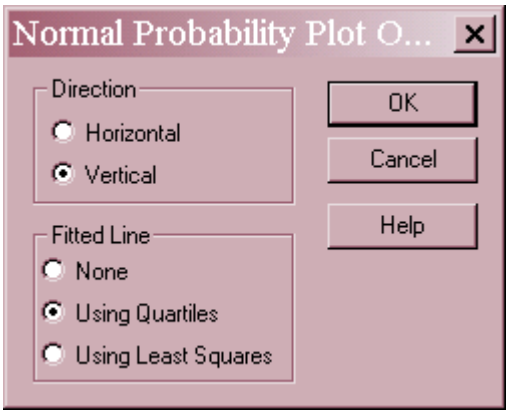
$$\hat{\mu} = \text{- intercept / slope} \tag{13}$$

$$\hat{\sigma} = \text{1 / slope} \tag{14}$$

The first method is more robust to deviations from normality in the tails of the distribution, since it essentially relies only on the middle half. Outliers or long tails will have a greater influence on the fit using the least squares method.

Note: set the default method for fitting lines on normal probability plots using the *EDA* pane on the *Preferences* dialog box, accessible from the *Edit* menu.

Pane Options



- **Direction**: the orientation of the plot. If vertical, the *Percentage* is displayed on the vertical axis. If *Horizontal*, *Percentage* is displayed on the horizontal axis.

- **Fitted Line**: the method used to fit the reference line to the data. If *Using Quartiles*, the line passes through the median when *Percentage* equals 50 with a slope determined from the interquartile range. If *Using Least Squares*, the line is fit by least squares regression of the normal quantiles on the observed order statistics. The former method based on quartiles puts more weight on the shape of the data near the center and is often enable to show deviations from normality in the tails that would not be evident using the least squares method.

Save Results

The *Save Results* button on the analysis toolbar allows the following values to be saved back to the datasheet:

1. **Winsorized data** – the data after Winsorization. The specified percentage of the largest and smallest values will have been replaced with the most extreme values not trimmed.

2. **Select flags** – a column containing a 0 for any value that you have manually excluded from the analysis using the *Exclude* feature on the *Outlier Plot*, and a 1 for all other values. In other procedures, enter the name of this column in the *Select* field to automatically exclude the same values from the analysis

3. .**Studentized values (no deletion)** – the standardized data values based on sample statistics for all observations.

4. **Studentized values (with deletion)** – the standardized data based on the mean and standard deviation calculated after deleting the observation.

5. **Modified Z-scores** – the standardized data based on the sample median and MAD estimate of sigma.

Calculations

Median Absolute Deviation

$$MAD = median_i \left\{ \left| x_i - \tilde{x} \right| \right\} \tag{15}$$

100α% Trimmed Mean

$$T(\alpha) = \frac{1}{n(1-2\alpha)} \left[k\left(x_{(r+1)} + x_{(n-r)} \right) + \sum_{i=r+2}^{n-r-1} x_{(i)} \right] \tag{16}$$

where $r = \lfloor \alpha n \rfloor$ and $k = 1 - (\alpha n - r)$.

100α% Winsorized Mean

$$T_W = \frac{1}{n} \left\{ \sum_{i=r+1}^{n-r} x_{(i)} + r\left[x_{(r+1)} + x_{(n-r)} \right] \right\} \tag{17}$$

Sbi

$$S_{bi} = \frac{\sqrt{n \sum_{i=1}^{n} \left(x_i - \tilde{x} \right)^2 \left(1 - u_i^2 \right)^4}}{\left| \sum_{i=1}^{n} \left(1 - u_i^2 \right)\left(1 - 5u_i^2 \right) \right|} \tag{18}$$

where

$$u_i = \frac{x_i - \tilde{x}}{9MAD} \tag{19}$$

Winsorized Sigma

$$S_W = \sqrt{\frac{n\left\{ \sum_{i=r+1}^{n-r} \left(x_{(i)} - T_W \right)^2 + r\left[\left(x_{(r+1)} - T_W \right)^2 + \left(x_{(n-r)} - T_W \right)^2 \right] \right\}}{(n-2r)(n-2r-1)}} \tag{20}$$

Winsorized Confidence Interval

$$T_W \pm t_{n-2r-1,\alpha/2} \frac{S_W}{\sqrt{n}} \tag{21}$$

