Describe – Response Variable

This procedure calculates statistics for a single column of numeric values. It displays the data in several ways and tests hypotheses about the population from which the data were obtained.

The data for this analysis consist of *n* samples from a population. Let

x_i = *i*-th observation
n = sample size

Access

Highlight: a single *Response* column.

Select: *Describe* from the main menu.

Output Page 1: A histogram and summary statistics are displayed.

Output Page 2: A fitted distribution is plotted, and confidence intervals are displayed for the population mean and standard deviation. Hypothesis tests for the mean and variance are also performed.

Output Page 3: A quantile plot is displayed and selected percentiles are calculated.

Output Page 4: A Q-Q or quantile-quantile plot is displayed to help judge how well the data correspond to the fitted distribution.

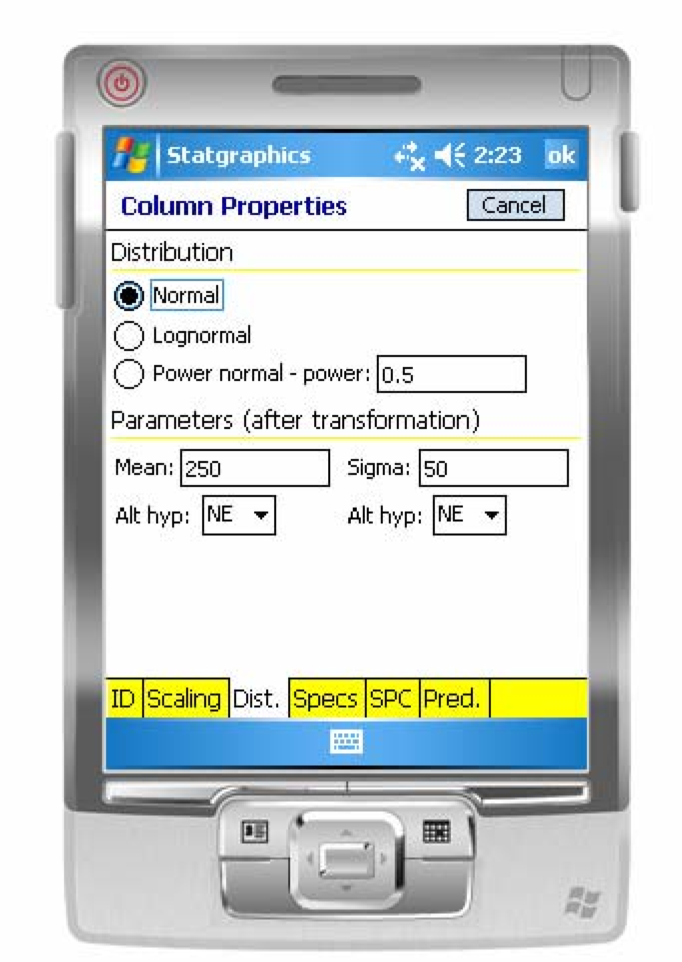
Output Page 5: An outlier plot is created with sigma limits. Grubbs' outlier test is also performed.

Options

The default assumption is that the data are random samples from a normal distribution. To select a different distribution:

1. Access the *Properties* dialog box for the *Response* variable by double-clicking on the column header.

2. On the *Dist.* tab, select the assumed distribution. The default selection assumes that the data follow a normal distribution. If you select *Lognormal*, the logarithms of the data will be assumed to follow a normal distribution. If you select *Power normal*, the data will be assumed to follow a normal distribution after raising them to the indicated *power*.



Column Properties

Distribution

● Normal

○ Lognormal

○ Power normal - power: 0.5

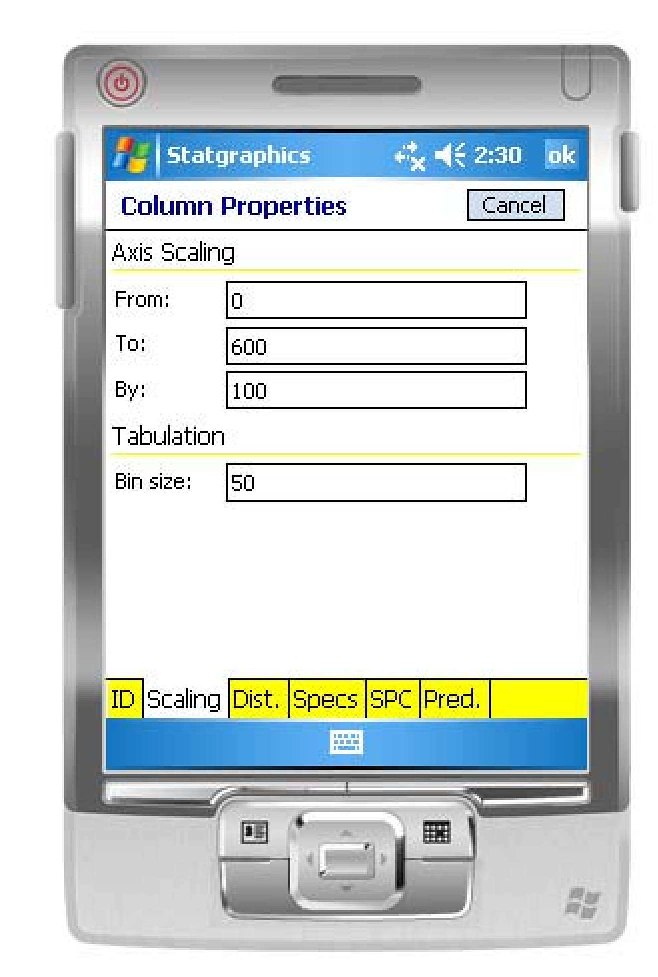
Parameters (after transformation)

Mean: 250 Sigma: 50

Alt hyp: NE Alt hyp: NE

If you intend to perform hypothesis tests concerning the mean and standard deviation of the population from which the data were sampled, specify the values to be used for the null hypothesis and the type of alternative hypothesis to be tested. If a non-normal distribution is selected, the mean and sigma should apply to the data after the indicated transformation has been applied.

The *Scaling* tab is also useful if you wish to specify the scale to be used when plotting the data:



Axis Scaling

From: 0

To: 600

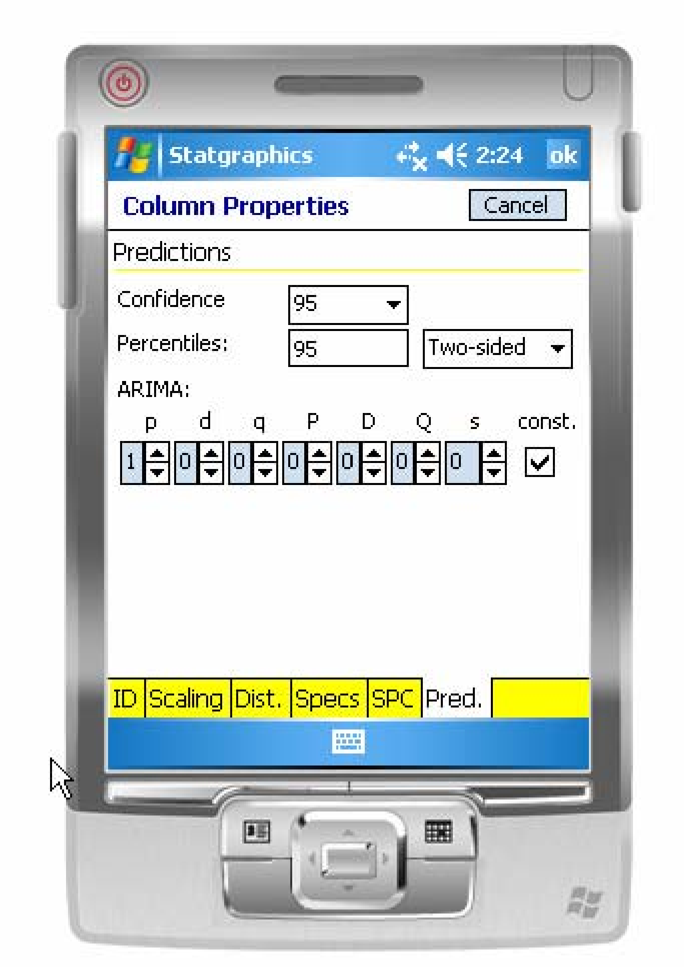
By: 100

Tabulation

Bin size: 50

In addition, you can set the width or *Bin size* of the intervals to be used when plotting a histogram.

The *Pred.* tab contains two additional fields that affect the output of this procedure:



Predictions

Confidence: 95

Percentiles: 95 Two-sided

ARIMA:

p	d	q	P	D	Q	s	const.
1	0	0	0	0	0	0	☑

The *Confidence* field sets the level of confidence used when displaying confidence intervals. The *Percentiles* field specifies the percentage at which percentiles will be calculated.

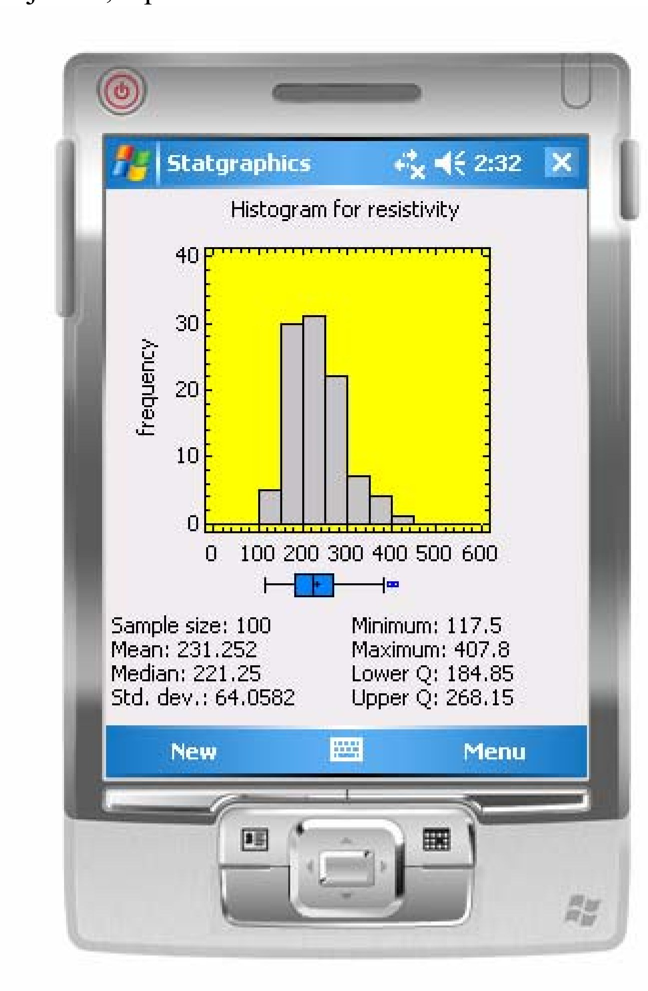
Sample Data

The file *resistivity.sgm* contains data on *n* = 100 samples, each containing the measured resistivity of a silicon wafer. The first several rows of the file are shown below:

Sample	Wafer 1
1	211.2
2	128.4
3	154.9
4	186.8
5	156.7
6	155.4
7	255.7
8	165.8
9	227.9
10	178.8

Histogram

The *Histogram* creates a chart that displays the number of observations falling within a set of adjacent, equal-width intervals.



```
Sample size: 100        Minimum: 117.5
Mean: 231.252           Maximum: 407.8
Median: 221.25          Lower Q: 184.85
Std. dev.: 64.0582      Upper Q: 268.15
```

Below the histogram is a 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).

- 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 that are not outside points.

The output also displays the values of the following statistics:

- **Count** - the sample size *n*, the number of non-missing entries in the column.

- **Average** or arithmetic **mean** - the center of mass of the data, given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

- **Median** - the middle value when the data are sorted from smallest to largest. If *n* is odd, the sample median equals $x_{(0.5+n/2)}$, where $x_{(i)}$ represents the *i-th* smallest observation If *n* is even, the sample median is the average of the two middle values:

$$\frac{x_{(n/2)} + x_{(1+n/2)}}{2} \tag{2}$$

- **Standard deviation** - the square root of the sample variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} \tag{3}$$

- **Minimum** - the smallest data value $x_{(1)}$.

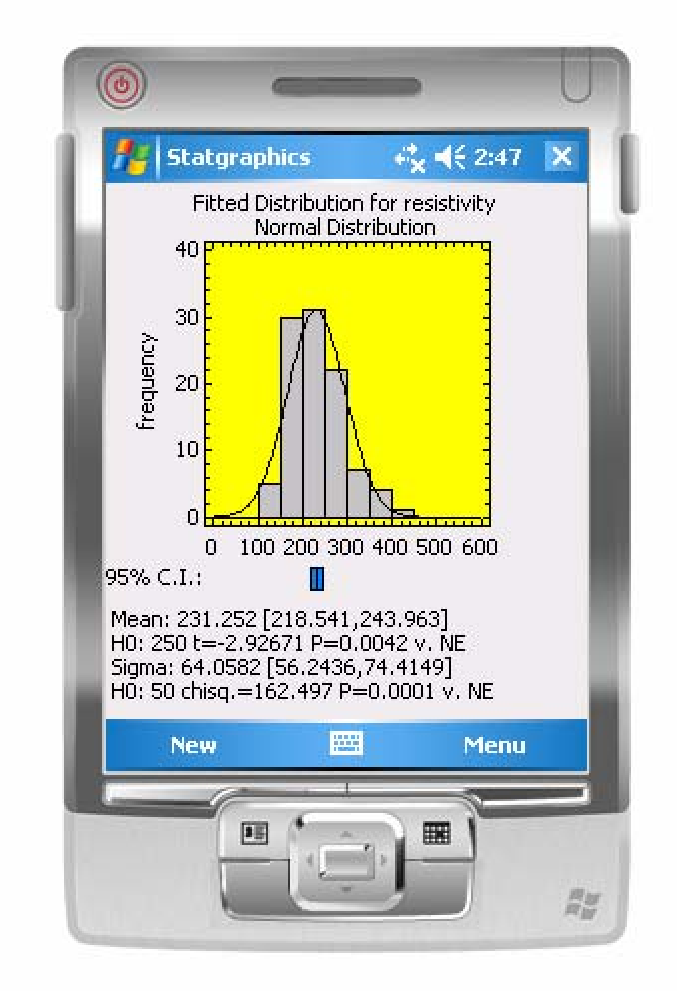
- **Maximum** - the largest data value $x_{(n)}$.

- **Lower quartile** - the 25-th percentile. Approximately 25% of the data values will lie below this value.

- **Upper quartile** - the 75-th percentile. Approximately 75% of the data values will lie below this value.

Distribution

This plot adds a plot of the fitted distribution to the output.



Fitted Distribution for resistivity
Normal Distribution

95% C.I.:

```
Mean: 231.252 [218.541,243.963]
H0: 250 t=-2.92671 P=0.0042 v. NE
Sigma: 64.0582 [56.2436,74.4149]
H0: 50 chisq.=162.497 P=0.0001 v. NE
```

Indicated below the plot are confidence intervals for the population mean and standard deviation. The interval for the mean is calculated using Student's t distribution with *n* – 1 degrees of freedom:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{4}$$

The interval for the standard deviation is calculated using the chi-square distribution with *n* – 1 degrees of freedom:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}},\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}\right] \tag{5}$$

If values for the mean and/or standard deviation have been specified on the *Dist*. tab of the *Column Properties* dialog box for the *Response* column, and if the data are assumed to come from a normal distribution, then the results of hypothesis tests are also included. Given a specified value μ_0 for the population mean, the program computes a t-statistic from

$$t = \frac{\bar{x} - \mu_o}{s / \sqrt{n}} \tag{6}$$

A P-value is also indicated. If $P < \alpha$, then the null hypothesis is rejected in favor of the alternative hypothesis at the 100α% significance level. Given a specified value σ_0 for the population standard deviation, the program computes a X^2-statistic from

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{7}$$

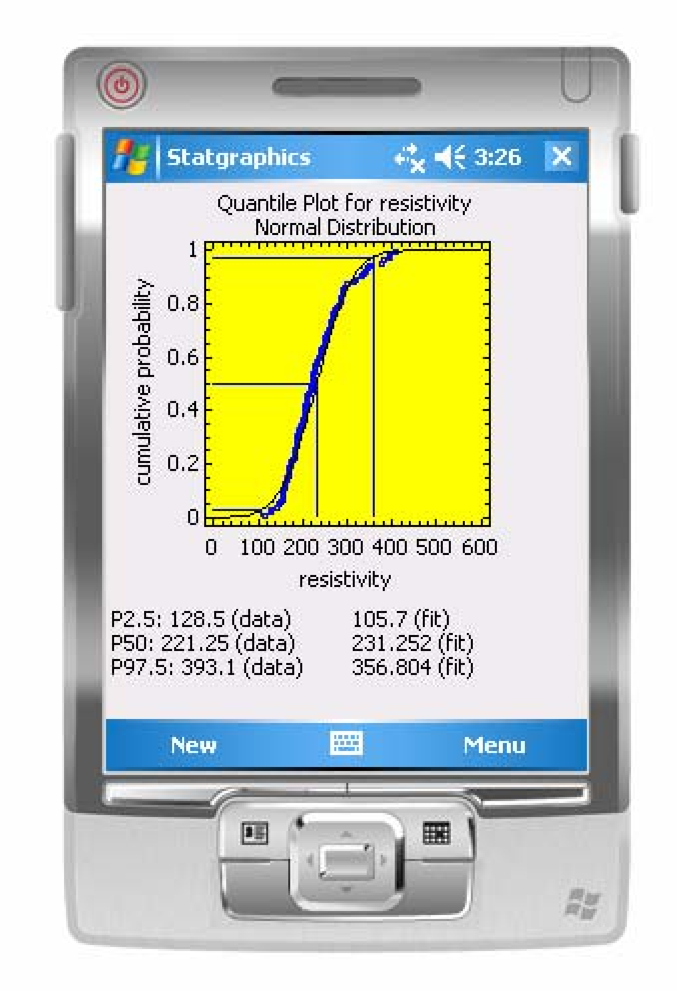
A P-value is also indicated.

In the above example, the hypothesis that $\mu = 250$ is rejected at the 1% significance level since the P-value is less than 0.01, as is the hypothesis that $\sigma = 50$.

If the data are assumed to come from a non-normal distribution, confidence and hypothesis tests are performed in the transformed metric.

Quantiles

This pane plots the quantiles (percentiles) of the data.



P2.5: 128.5 (data) 105.7 (fit)
P50: 221.25 (data) 231.252 (fit)
P97.5: 393.1 (data) 356.804 (fit)

In this plot, the data are sorted from smallest to largest and plotted at the coordinates

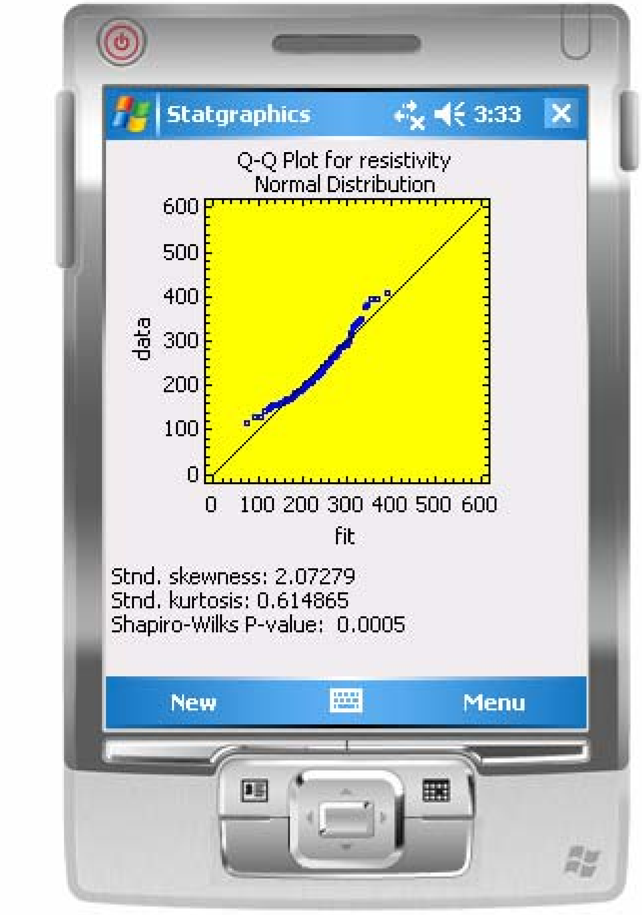
$$\left(x_{(j)}, \frac{j-0.5}{n} \right) \tag{8}$$

A solid line is drawn through the points showing the cumulative distribution function of the fitted distribution.

The bottom of the page lists the values of 3 percentiles, based both on the data and on the fitted distribution. In addition to P50 (50th percentile), percentiles are also displayed that bound the percentage indicated on the *Pred*. tab of the *Column properties* dialog box for the *Response* variable. Two-sided or one-sided bounds can be displayed. Lines are also drawn on the graph at the indicated percentiles.

Q-Q Plot

The *Quantile-Quantile Plot* shows the quantiles of the data plotted versus the equivalent percentiles of the fitted distribution.



Stnd. skewness: 2.07279
Stnd. kurtosis: 0.614865
Shapiro-Wilks P-value: 0.0005

If the data are well-modeled by the selected distribution, then the points should fall close to the diagonal line. Curvature such as that shown above is indicative of a discrepancy between the data and the selected distribution.

Also displayed are:

- **Standardized skewness** (measure of shape) – a measure of asymmetry calculated from

$$z_1 = \frac{g_1}{\sqrt{6/n}} \tag{9}$$

where g_1 is the coefficient of skewness

$$g_1 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} \tag{10}$$

For large samples, significant skewness can be asserted at the 5% significance level if z_1 falls outside the interval (-2, +2).

- **Standardized kurtosis** (measure of shape) - a measure of relative peakedness or flatness compared to a bell-shaped curve, calculated from

$$z_2 = \frac{g_2}{\sqrt{24/n}} \tag{11}$$

where g_2 is the coefficient of kurtosis

$$g_2 = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{12}$$

At the 5% significance level, significant kurtosis can be asserted if z_2 falls outside the interval (-2,+2).

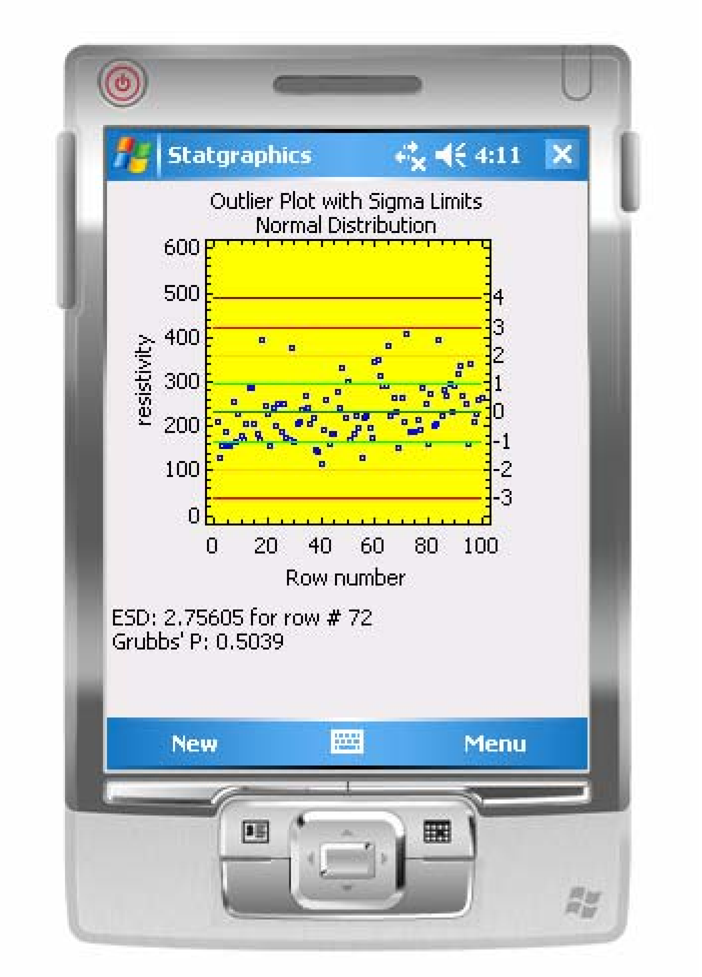
- **Shapiro-Wilks P-value** – The Shapiro-Wilks test, available when $2 \le n \le 2000$, uses a statistic derived by calculating how well the data fall along a straight line when plotted on a normal probability plot. Small P-values indicate that the data are not well modeled by a normal distribution. In computing the statistic and its P-value, Statgraphics uses Roysten's method.

For the sample data, the standardized skewness statistic exceeds 2.0 and thus indicates significant positive skewness in the data. The small P-value for the Shaipro-Wilks test also indicates that the data do not come from a normal distribution.

If a non-normal distribution is selected, all of the output applies to the data after making the appropriate transformation. The tests therefore indicate whether or not the selected transformation is appropriate for the data.

Outliers

This page of output is useful for identifying whether there are any significant outliers present in the data. It shows a plot of the data with horizontal lines at different sigma distances from the sample mean.



ESD: 2.75605 for row # 72
Grubbs' P: 0.5039

Provided n ≥ 3 , the page also displays the result of Grubbs' test, which is designed to detect outliers in a sample from a normal distribution. Also called the *Extreme Studentized Deviate (EDS)* test, it is based on the largest Studentized value t_{max}, which corresponds to largest absolute value of the Studentized deviates

$$t_i = \frac{x_i - \bar{x}}{s} \tag{13}$$

An approximate two-sided P-Value is obtained using Student's t-distribution with *n* - 2 degrees of freedom. A small P-value leads to the conclusion that the most extreme point is indeed an outlier. In the sample data, row #72 is farthest from the mean, approximately 2.76 standard deviations away. Since P is well above 0.05, this value is not a significant outlier.

If a non-normal distribution is assumed, all calculations are done on the transformed data.