Normal Probability Plot

Summary

The *Normal Probability Plot* is used to help judge whether or not a sample of numeric data comes from a normal distribution. If it does not, you can often determine the type of departure from normality by examining the way in which the data deviate from the normal reference line.

Sample StatFolio: *probplot.sgp*

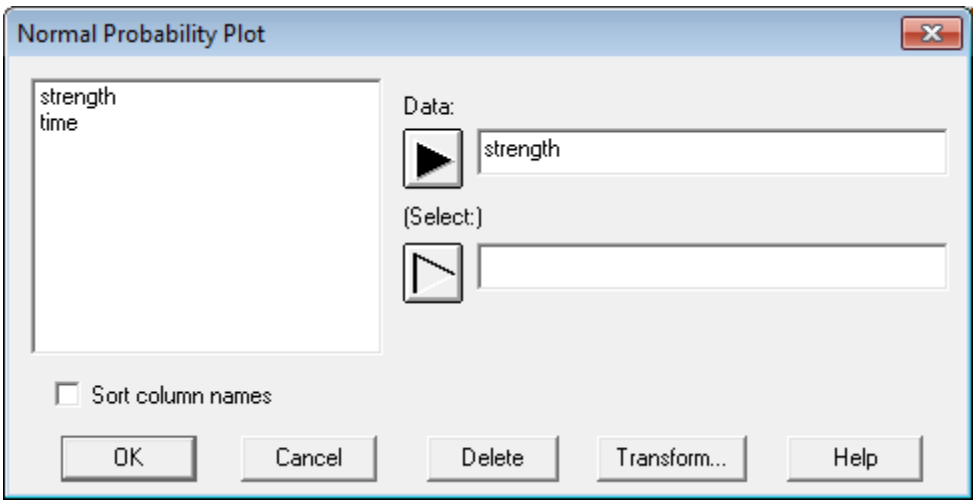
Sample Data

The file *bottles.sgd* contains the measured bursting strength of $n = 100$ glass bottles, similar to a dataset contained in Montgomery (2005). The table below shows a partial list of the data from that file:

strength
255
232
282
260
255
233
240
255
254
259
235
262

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.

Analysis Summary

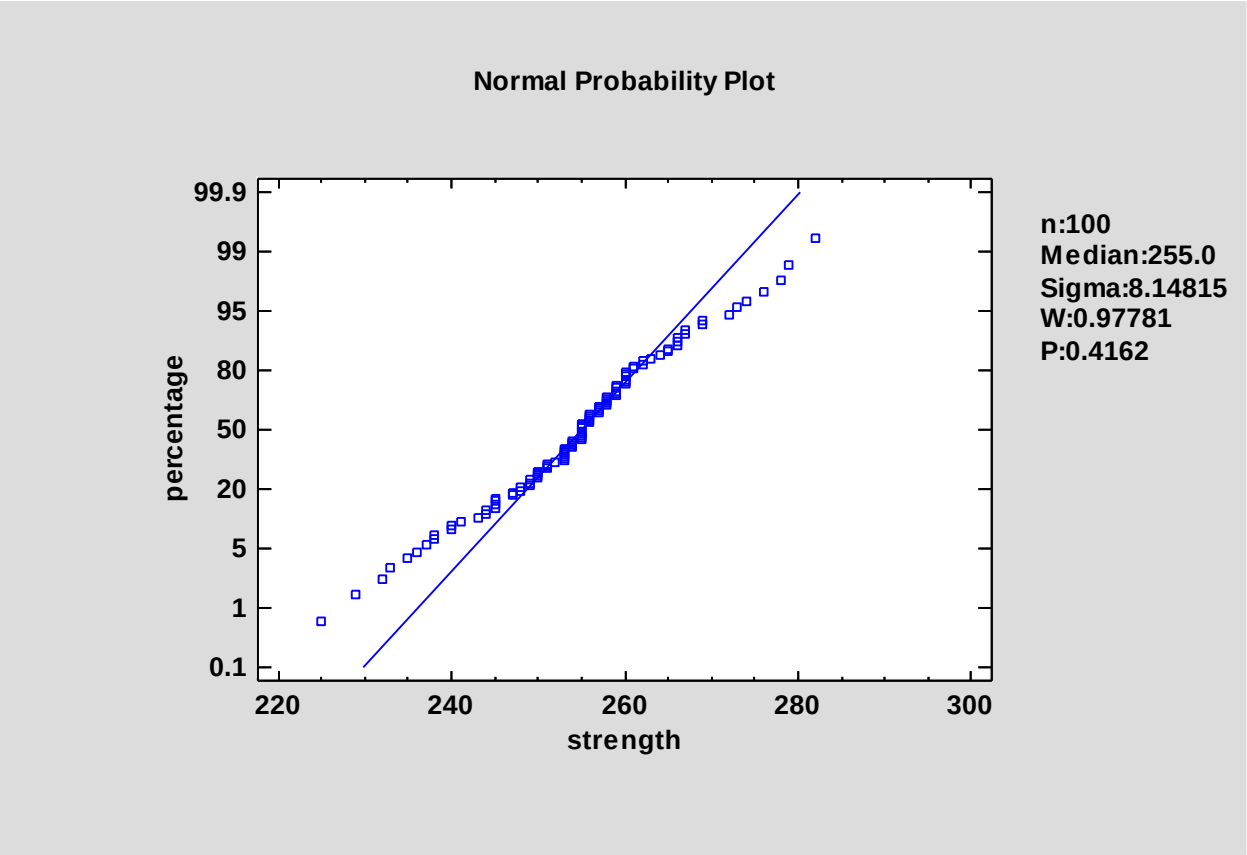
The *Analysis Summary* shows the number of observations in the data column.

Probability Plot - strength
Data variable: strength
100 values ranging from 225.0 to 282.0

Also displayed are the largest and smallest values.

Normal Probability Plot

This pane displays the probability plot.



The plot is constructed in the following manner:

- The data are sorted from smallest to largest and the order statistics are determined. By definition, the *j-th* order statistic is the *j-th* smallest observation in the sample, denoted by $x_{(j)}$.
- The data are then plotted at the positions

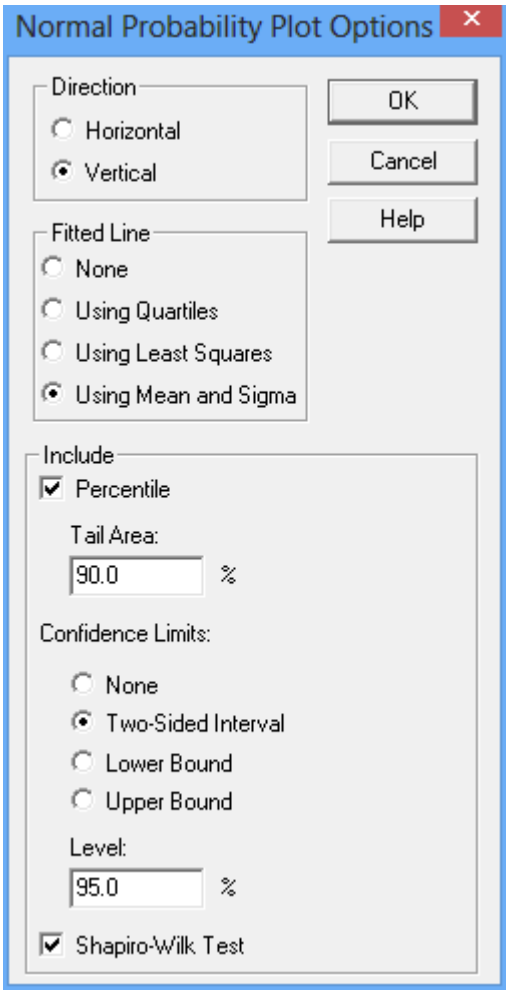
$$\left(x_{(j)}, \Phi^{-1}\left(\frac{j - 0.375}{n + 0.25} \right) \right) \qquad (1)$$

  where $\Phi^{-1}(u)$ indicates the inverse standard normal distribution evaluated at u.

- If desired, a straight line is fit to the data and added to the plot.

The normal probability plot is created in such a way that, if the data are random samples from a normal distribution, they should lie approximately along a straight line. In the above plot, the deviation of the values from the reference line at both ends indicates that the data may come from a distribution with relatively longer tails than a normal distribution.

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. If *Using Mean and Sigma*, the line is determined from the mean and standard deviation of the *n* observations. The method based on quartiles puts more weight on the shape of the data near the center and is often able to show deviations from normality in the tails that would not be evident using the other methods.

- **Include - Percentile:** displays a reference line at the specified percentile. The percentile is estimated using the fitted line.

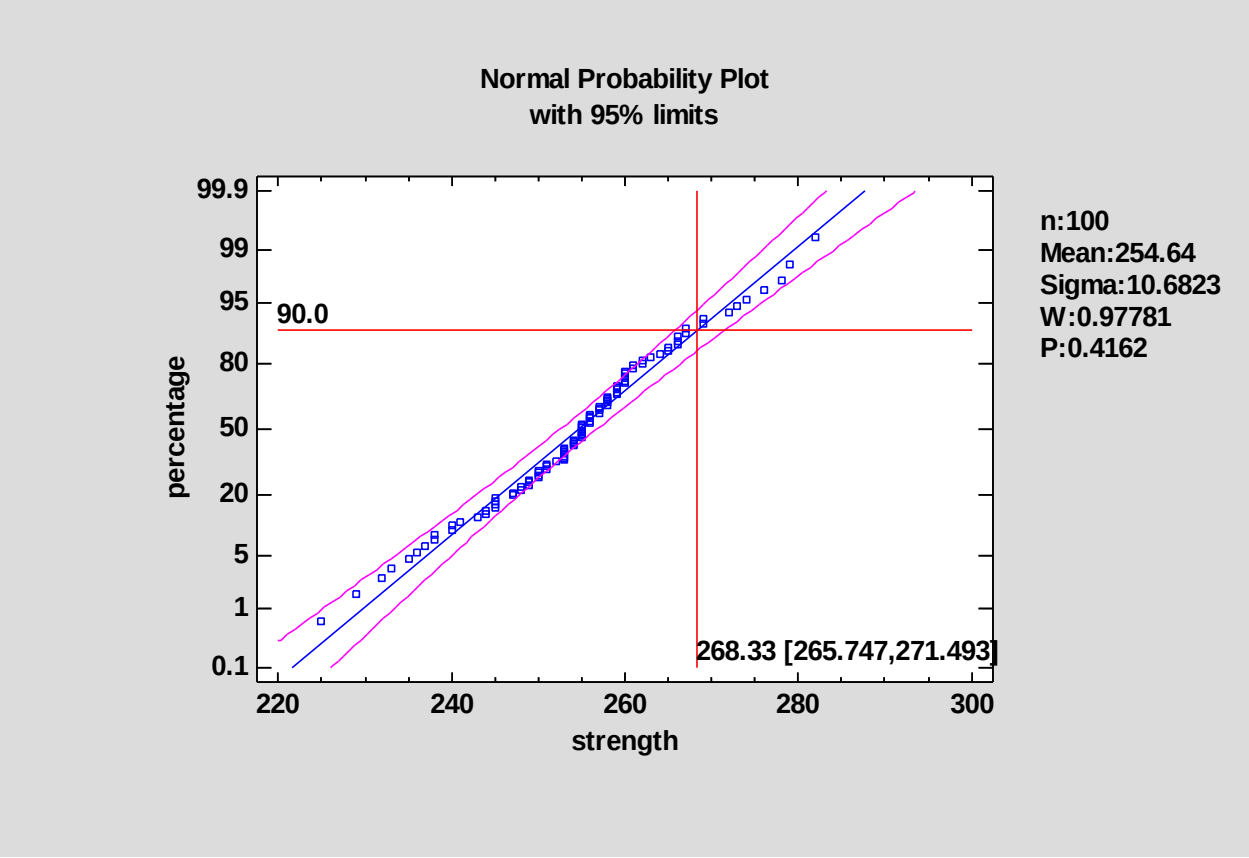
- **Include – Confidence Limits:** displays confidence limits around the fitted line (only available when estimating the line *Using mean and sigma*). The confidence level applies to each percentile separately.

- **Include – Shapiro-Wilk Test**: displays the calculated value of the Shapiro-Wilk W test and its associated P value. This test is described in the document titled *Distribution Fitting – Uncensored Data*. A small P-value indicates that the data are not well modeled by a normal distribution.

The *Direction* and *Fitted Line* defaults are determined from the settings on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu.

Example – Using Mean and Sigma with Confidence Limits and a Selected Percentile

The plot below displays the estimated line based on the sample mean and standard deviation, the estimated 90th percentile, and the 95% confidence limits.



Using the *Mean and Sigma* method on the sample data produces a more "normal" looking plot, since the line is forced to fit both the core and the tails.

