Power Transformations

Summary

The **Power Transformations** procedure is designed to determine a normalizing transformation for a column of numeric observations that do not come from a normal distribution. In such cases, it is often possible to find a power transformation that will make the data approximately normal. Given such a transformation, statistical procedures that assume normality can then be applied to the transformed data.

The procedure uses the method proposed by Box and Cox (1964).

Sample StatFolio: *powertransforms.sgp*

Sample Data:

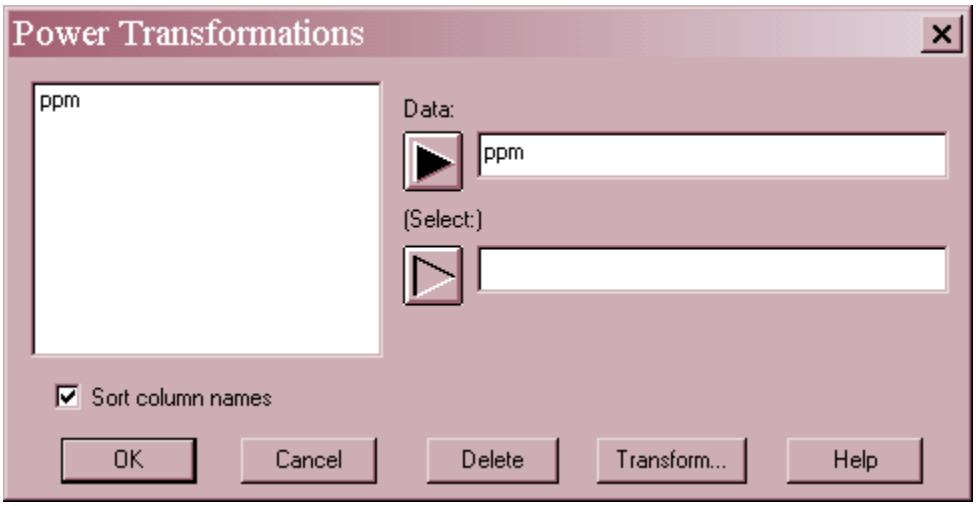
The file *groundwater.sgd* contains $n = 47$ measurements of the concentration of uranium in groundwater samples taken from a location in northwest Texas. The table below shows a partial list of the data from that file:

ppm
8.25
2.82
4.16
18.66
12.72
8.75
2.29
7.22
9.76
7.72
27.38
5.14

The concentration is measured in parts per million.

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 analyzed.
- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows the transformation derived for the data.

Power Transformations
Data variable: ppm
Number of observations = 47

Box-Cox Transformation
Power (lambda1): 0.204
Shift (lambda2): 0.0
(optimized)

Geometric mean = 9.01355

Approximate 95% confidence interval for power: -0.077 to 0.505

The procedure automatically determines the best power transformation by finding the value of λ_1 that minimizes the standard deviation of the observations when transformed according to the Box-Cox transformation:

$$Y = 1 + \frac{(X + \lambda_2)^{\lambda_1} - 1}{\lambda_1 g^{\lambda_1 - 1}} \qquad \text{if} \qquad \lambda_1 \neq 0 \qquad (1)$$

$$Y = 1 + g \ln(X + \lambda_2) \qquad \text{if} \qquad \lambda_1 = 0 \qquad (2)$$

where g is the geometric mean of the observations after adding λ_2:

$$g = \left(\prod_{i=1}^{n} (X_i + \lambda_2) \right)^{1/n} \quad (3)$$

The parameter λ_2 is set to 0 unless the analyst specifies a non-zero value on the *Analysis Options* dialog box.

At the heart of the above transformation is the power to which the data is raised, λ_1. Often, a power between –2 and +2 will make the data approximately normal. This includes many common transformations:

Power λ_1	*Transformation*
-2.0	reciprocal square
-1.0	reciprocal
-0.5	reciprocal square root
-0.3333	reciprocal cube root
0	logarithm
0.3333	cube root
0.5	square root
1.0	none
2.0	square

In general, the further λ_1 is from 1.0, the stronger the transformation. Powers less than 1.0 are required to normalize positively skewed data, while power greater than 1.0 are required for negatively skewed data.

Important information included in the output is:

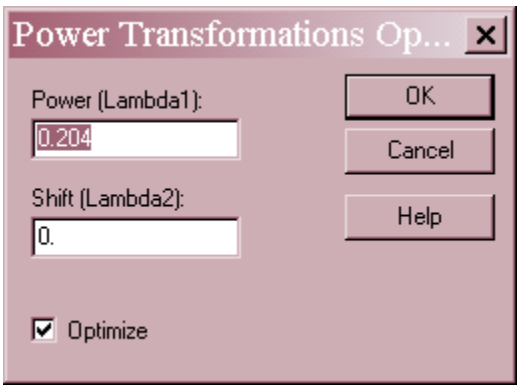
1. **Power (λ_1):** the optimal power for the data. For the sample data, it appears that $ppm^{0.204}$ is the optimal transformation to achieve normality.

2. **Shift (λ_2):** a user-specified constant to add to each observation before raising it to a power. In some cases, shifting the data prior to performing the power transformation improves the fit.

3. **Geometric mean (g):** the geometric mean of the observations after adding the shift parameter.

4. **Approximate confidence interval for power**: an approximate confidence interval for the power parameter λ_1. Since the derived power transformation is based on a sample of data, it is only a point estimate of the best power for the population from which the data were taken. The confidence interval shows the estimated margin of error. In this case, any power between –0.077 and 0.505 could be a reasonable value for λ_1. This includes a logarithmic transform and a square root.

Analysis Options



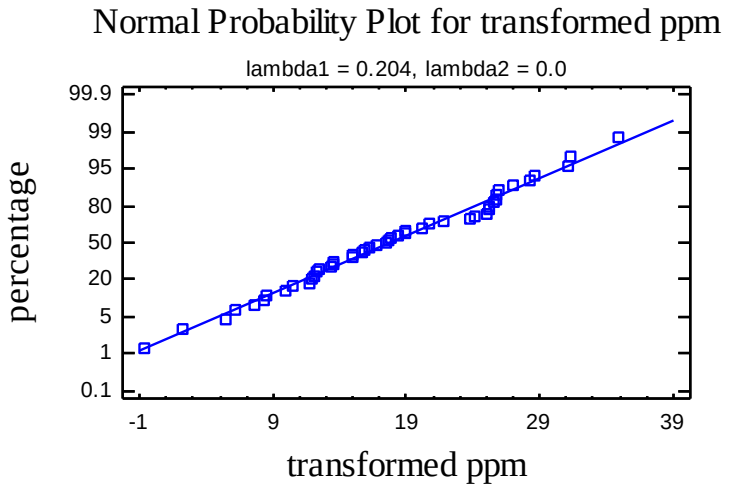
- **Power (lambda1):** the power parameter λ_1. If *Optimize* is checked, this value will be automatically determined by the procedure.

- **Shift (lambda2):** the shift parameter λ_2. This value is added to the observations before the power transformation is performed.

- **Optimize:** check this button to have the procedure determine an optimal value for λ_1 using the Box-Cox method.

Normal Probability Plot

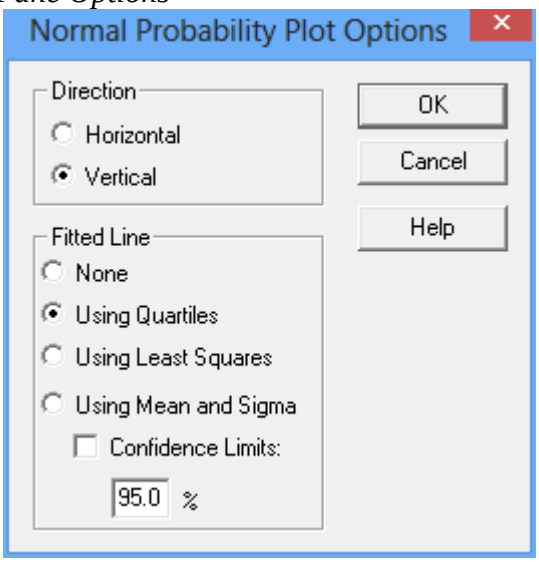
This pane shows a normal probability plot for the transformed values.



If the transformation was successful in normalizing the data, it should fall approximately along a straight line. For the sample data, the transformation was very effective.

Details regarding normal probability plots may be found in the *Normal Probability Plot* documentation.

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.

- **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.

MSE Comparison Table

The Box-Cox procedure finds the value λ_1 that minimizes the mean squared error

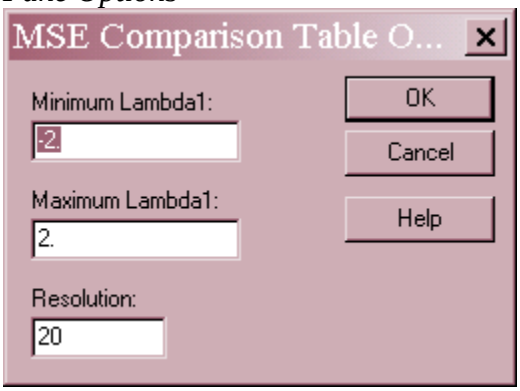
$$\text{MSE} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n} \qquad (4)$$

This table shows the MSE for various values of λ_1.

MSE Comparison Table
Shift (lambda2): 0.0

lambda1	*MSE*
-2.0	*10323.0*
-1.8	4809.42
-1.6	2311.36
-1.4	1154.69
-1.2	605.55
-1.0	337.376
-0.8	202.379
-0.6	132.407
-0.4	95.3914
-0.2	75.9491
0.0	66.6195
0.2	63.853
0.4	66.1757
0.6	73.3865
0.8	86.2915
1.0	106.778
1.2	138.193
1.4	186.112
1.6	259.687
1.8	373.974
2.0	553.922

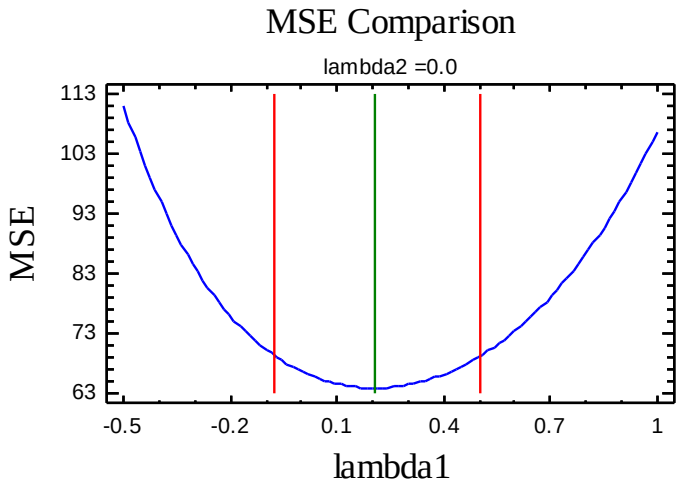
Pane Options



- **Minimum Lambda1**: minimum value of λ_1 to display in the table.
- **Maximum Lambda1**: maximum value of λ_1 to display in the table.
- **Resolution**: the number of increments between the minimum and maximum values.

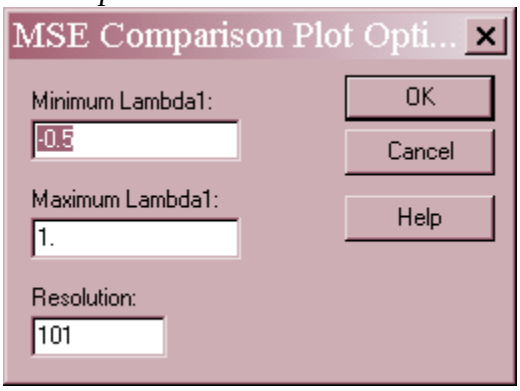
MSE Comparison Plot

This plot shows the MSE as a function of λ_1.



Vertical lines are drawn at the derived λ_1 and its confidence limits.

Pane Options



- **Minimum Lambda1**: minimum value of λ_1 to display on the plot.
- **Maximum Lambda1**: maximum value of λ_1 to display on the plot.
- **Resolution**: the number of values of λ_1 at which to plot the MSE.

Tests for Normality

This pane shows the results of several tests to determine whether a normal distribution adequately models the transformed data.

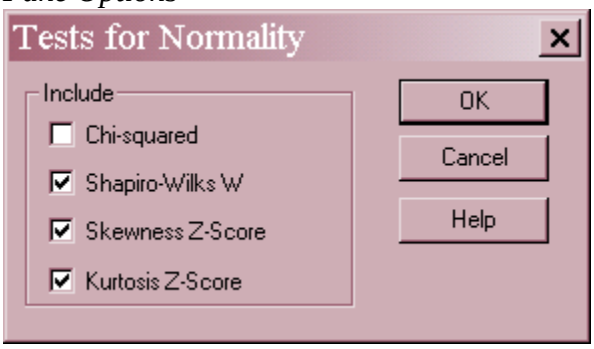
Tests for Normality
Power (lambda1): 0.204
Shift (lambda2): 0.0

Test	*Statistic*	*P-Value*
Shapiro-Wilks W	0.981806	0.8057
Skewness Z-score	-0.0732864	0.941573
Kurtosis Z-score	-0.697379	0.485563

Small P-Values for any test (less than 0.05 is operating at the 5% significance level) lead to a rejection of the hypothesis that the transformed data follows a normal distribution. For the sample data, the transformation seems to have adequately normalized the data.

For more details about tests for normality, refer to the documentation on *Distribution Fitting (Uncensored Data)*.

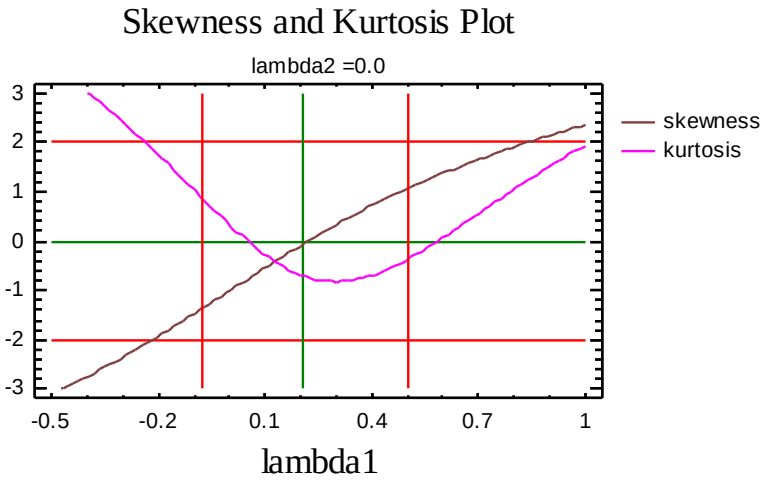
Pane Options



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

Skewness and Kurtosis Plot

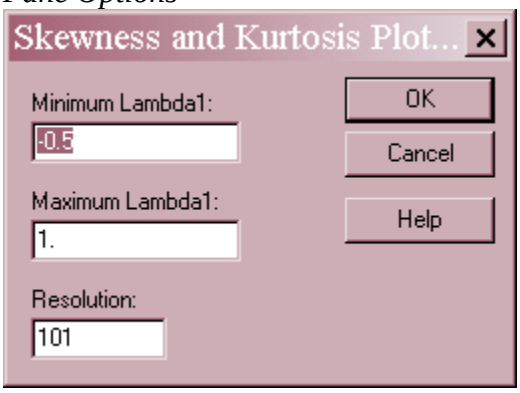
This plot shows the values of the standardized skewness and standardized kurtosis as a function of the power parameter λ_1.



The standardized skewness and standardized kurtosis should both be between –2 and +2 for a transformation that adequately normalizes the data. The plot shows horizontal lines at –2 and +2, with the vertical lines indicating the optimal value of λ_1 and its confidence limits.

Clearly, there is a wide range of values for λ_1 that would create a reasonable transformation of the data.

Pane Options



- **Minimum Lambda1**: minimum value of λ_1 to display on the plot.
- **Maximum Lambda1**: maximum value of λ_1 to display on the plot.
- **Resolution**: the number of values of λ_1 at which to plot the statistics.

Save Results

You may save the *Transformed Data* values Y to a column of the datasheet.

Calculations

Standardized Skewness & Standardized Kurtosis

Calculated using the method described under *Tests for Normality* in the documentation for *Distribution Fitting (Uncensored Data)*.

