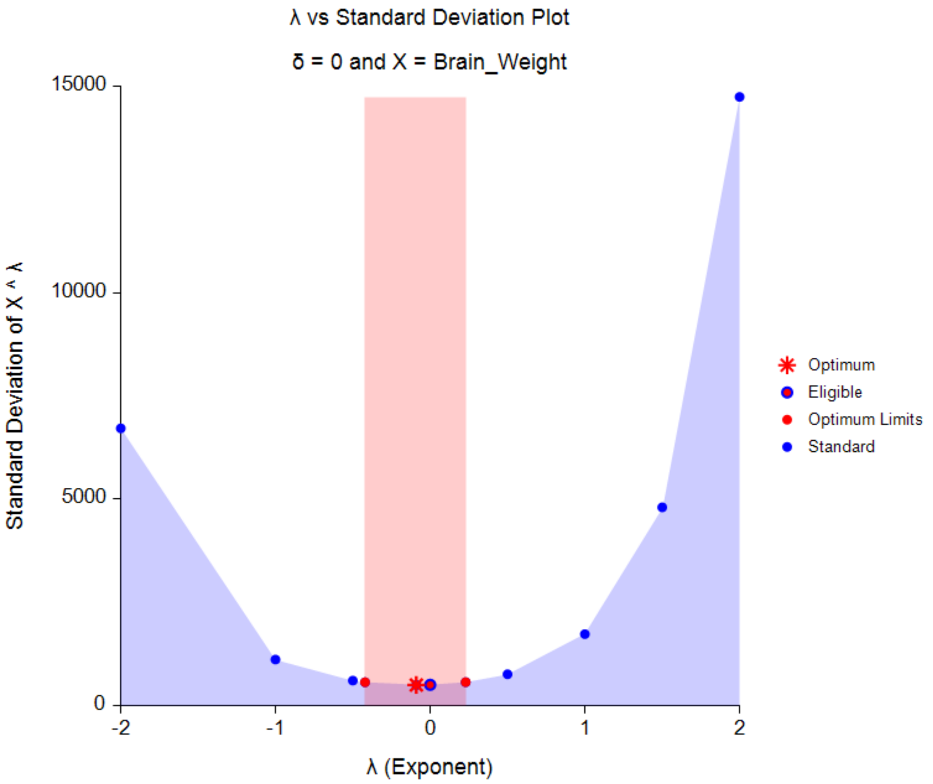
Chapter 190

Box-Cox Transformation

Introduction

This procedure finds the appropriate Box-Cox power transformation (1964) for a single batch of data. It is used to modify the distributional shape of a set of data to be more normally distributed so that tests and confidence limits that require normality can be appropriately used. It cannot correct every data ill. For example, data that contains outliers may not be properly normalized by this technique.

Example of the Box-Cox λ Plot



The Box-Cox transformation has the following mathematical form

$$Y = (X + \delta)^{\lambda}$$

where λ is the exponent (power) and δ is a shift amount that is added when X is zero or negative. When λ is zero, the above definition is replaced by

$$Y = \ln(X + \delta)$$

Usually, the standard λ values of -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, and 2 are investigated to determine which, if any, is most suitable. The program will also solve for the optimum value of λ using maximum likelihood estimation. The program also calculates confidence limits about the optimum value. The usual procedure is to adopt the most convenient standard value between the confidence limits. For example, if the confidence limits were 0.4 to 1.1, *λ* would be set to the standard value of '1' (no transformation) since this is the most convenient. Care must be used when using the confidence limits, because they are heavily dependent on the sample size.

Box-Cox Algorithm

Suppose you have a sample of *n* response values $X_1, X_2, \ldots, X_n$. Further suppose you visually determine a value of *δ* that will keep all $X + \delta > 0$. Calculate a set of Z_i's corresponding to the X_i's using

$$Z = \begin{cases} \left[(X+\delta)^{\lambda} - 1\right] / \left[\lambda H^{\lambda-1}\right] & \lambda \neq 0 \\ H \ln(X+\delta) & \lambda = 0 \end{cases}$$

where *H* is the geometric mean of $X + \delta$. That is,

$$H = \sqrt[n]{\prod_{i=1}^{n}(X+\delta)}$$

Scaling by *H* is intended to keep the standard deviation of the *Z*'s approximately the same as the standard deviation of the *X*'s so that the standard deviations can be compared at various values of *λ*.

Maximum Likelihood Estimation of λ

In this case, the likelihood for a given *λ* is inversely proportional to the standard deviation of the corresponding *Z*'s. The likelihood function is maximized when the standard deviation is minimized. A bracketing search algorithm is conducted that continues to tighten the boundaries until a specified precision (bracket width) is reached.

Approximate Confidence Interval for λ

An approximate confidence interval for *λ* is based on likelihood function which in turn is proportional to the standard deviation (SD) of *Z*. The confidence limits correspond to the two values of *λ* at which

$$SD_{\lambda}^{2} = SD_{\hat{\lambda}}^{2} \exp\left(\frac{\chi_1^2(1-\alpha)}{n}\right)$$

where $\hat{\lambda}$ is the maximum likelihood estimate of *λ* and $\chi_1^2(1-\alpha)$ is the percentage point of the chi-squared distribution with one degree of freedom.

Data Structure

The data may be entered as a single variable or as several variables that are joined together into one variable during the analysis.

Example 1 – Box-Cox Transformation of Brain Weights

This section presents an example of how to run a Box-Cox transformation analysis of the brain weights of 15 types of mammals. The data used are found in the Mammals dataset.

Setup

To run this example, complete the following steps:

1 Open the Mammals example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Mammals** and click **OK**.

2 Specify the Box-Cox Transformation procedure options

- Find and open the **Box-Cox Transformation** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Variable(s) ....**Brain_Weight**

Report Options (*in the Toolbar*)

Variable Labels...**Column Names**

3 Run the procedure

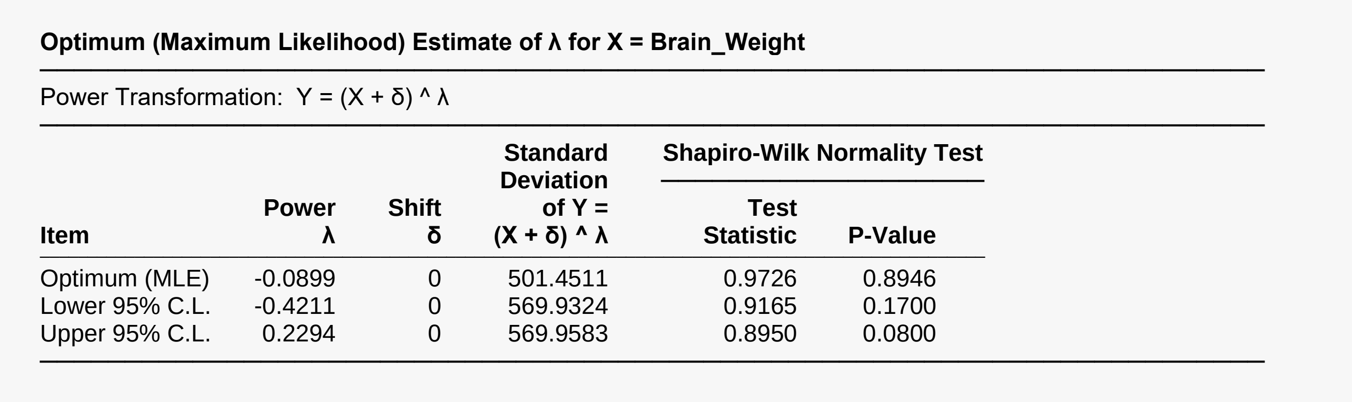
- Click the **Run** button to perform the calculations and generate the output.

Run Summary

```
Run Summary for Brain_Weight
─────────────────────────────────────────────────────────────────────────────────
Parameter                Value          Parameter                 Value
δ (Shift)                0              Rows Processed            15
Optimum λ (Power)        -0.0899        Rows Used                 15
Minimum λ Searched       -5
Maximum λ Searched       5              Geometric Mean (with δ)   326.7287
Target Search Width of λ 0.0001         Minimum                   26
MLE Iterations Used      26             Maximum                   5712
Max MLE Iterations       50
─────────────────────────────────────────────────────────────────────────────────
```

This report summarizes the run by showing main results as well as the input settings that were used. You should pay particular attention to the *Rows* lines to make sure that they are as you expect. Also, if the number of MLE Iterations is equal to the Max MLE Iterations, the search algorithm may not have converged properly.

Optimum (Maximum Likelihood) Estimate of λ



```
Optimum (Maximum Likelihood) Estimate of λ for X = Brain_Weight
─────────────────────────────────────────────────────────────────────────────
Power Transformation:  Y = (X + δ) ^ λ
─────────────────────────────────────────────────────────────────────────────
                                      Standard    Shapiro-Wilk Normality Test
                                     Deviation    ───────────────────────────
                    Power   Shift       of Y =          Test
Item                    λ       δ  (X + δ) ^ λ     Statistic     P-Value
─────────────────────────────────────────────────────────────────────────
Optimum (MLE)     -0.0899       0     501.4511        0.9726      0.8946
Lower 95% C.L.    -0.4211       0     569.9324        0.9165      0.1700
Upper 95% C.L.     0.2294       0     569.9583        0.8950      0.0800
─────────────────────────────────────────────────────────────────────────────
```

This report gives the results for the maximum likelihood estimation portion of the analysis.

Item

The name of item being reported on this line of the report.

Power (λ)

The value of λ for this item. This is the transformation exponent.

Shift (δ)

The value of δ, the shift value.

Standard Deviation of Y = (X + δ) ^ λ

This is the standard deviation of the transformed data values. Actually, the data have not only been shifted and raised to the indicated power, but they have also been scaled by the geometric mean so that the standard deviations are directly comparable. Note the geometric mean is not used when using the λ that has been found by this algorithm.

Shapiro-Wilk Normality Test Statistic

The value of the Shapiro-Wilk normality test statistic calculated on transformed data.

Shapiro-Wilk Normality Test P-Value

The p-value of the Shapiro-Wilk normality test. Since the desire is to transform the data to be more normally distributed, you are looking for large (non-significant) values. Remember that this value is not only influenced by the normality of the data, but also by the sample size.

Standard λ's

```
Standard λ's for X = Brain_Weight
─────────────────────────────────────────────────────────────────────────────
Power Transformation:  Y = (X + δ) ^ λ
─────────────────────────────────────────────────────────────────────────────
                                  Standard     Shapiro-Wilk Normality Test
                                 Deviation     ───────────────────────────
              Power     Shift       of Y =          Test
Item              λ         δ  (X + δ) ^ λ     Statistic      P-Value
─────────────────────────────────────────────────────────────────
1              -2.0         0    6717.1155        0.4763       0.0000
2              -1.0         0    1117.7134        0.7051       0.0003
3              -0.5         0     607.4627        0.8896       0.0661
4               0.0         0     506.5852        0.9640       0.7615
5               0.5         0     761.9966        0.7694       0.0015
6               1.0         0    1733.7314        0.5836       0.0000
7               1.5         0    4804.2570        0.4961       0.0000
8               2.0         0   14744.4980        0.4583       0.0000
─────────────────────────────────────────────────────────────────────────────
λ's between the maximum likelihood confidence limits are bolded.
```

This report displays the results for each of the standard λ's.

Item

The number of items being reported on this line of the report.

Power (λ)

The value of λ for this item. This is the transformation exponent.

Shift (δ)

The value of δ, the shift value.

Standard Deviation of Y = (X + δ) ^ λ

This is the standard deviation of the transformed data values. Actually, the data have not only been shifted and raised to the indicated power, but they have also been scaled by the geometric mean so that the standard deviations are directly comparable. Note the geometric mean is not used when using the λ that has been found by this algorithm.

Shapiro-Wilk Normality Test Statistic

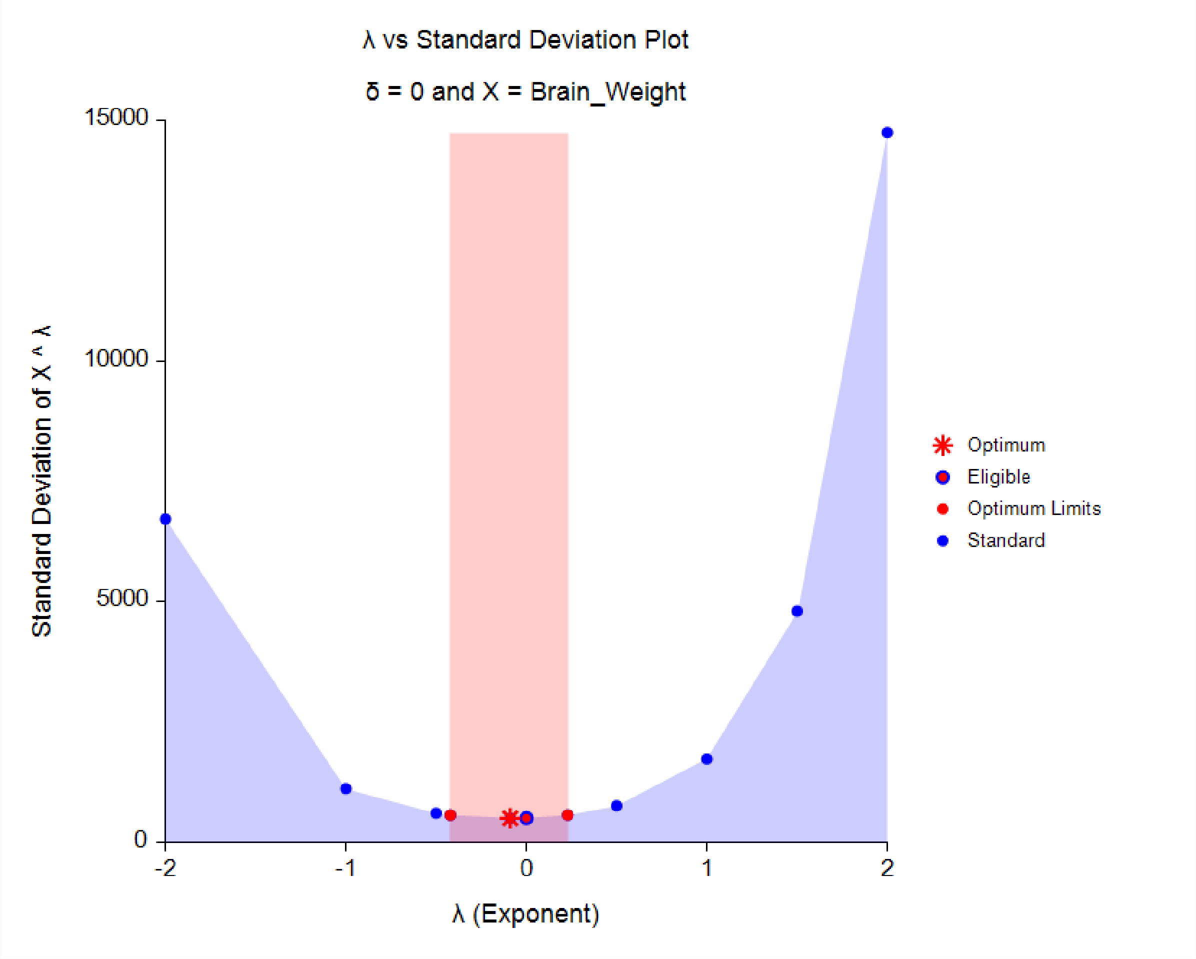
The value of the Shapiro-Wilk normality test statistic calculated on transformed data.

Shapiro-Wilk Normality Test P-Value

The p-value of the Shapiro-Wilk normality test. Since the desire is to transform the data to be more normally distributed, you are looking for large (non-significant) values. Remember that this value is not only influenced by the normality of the data, but also by the sample size.

Plots

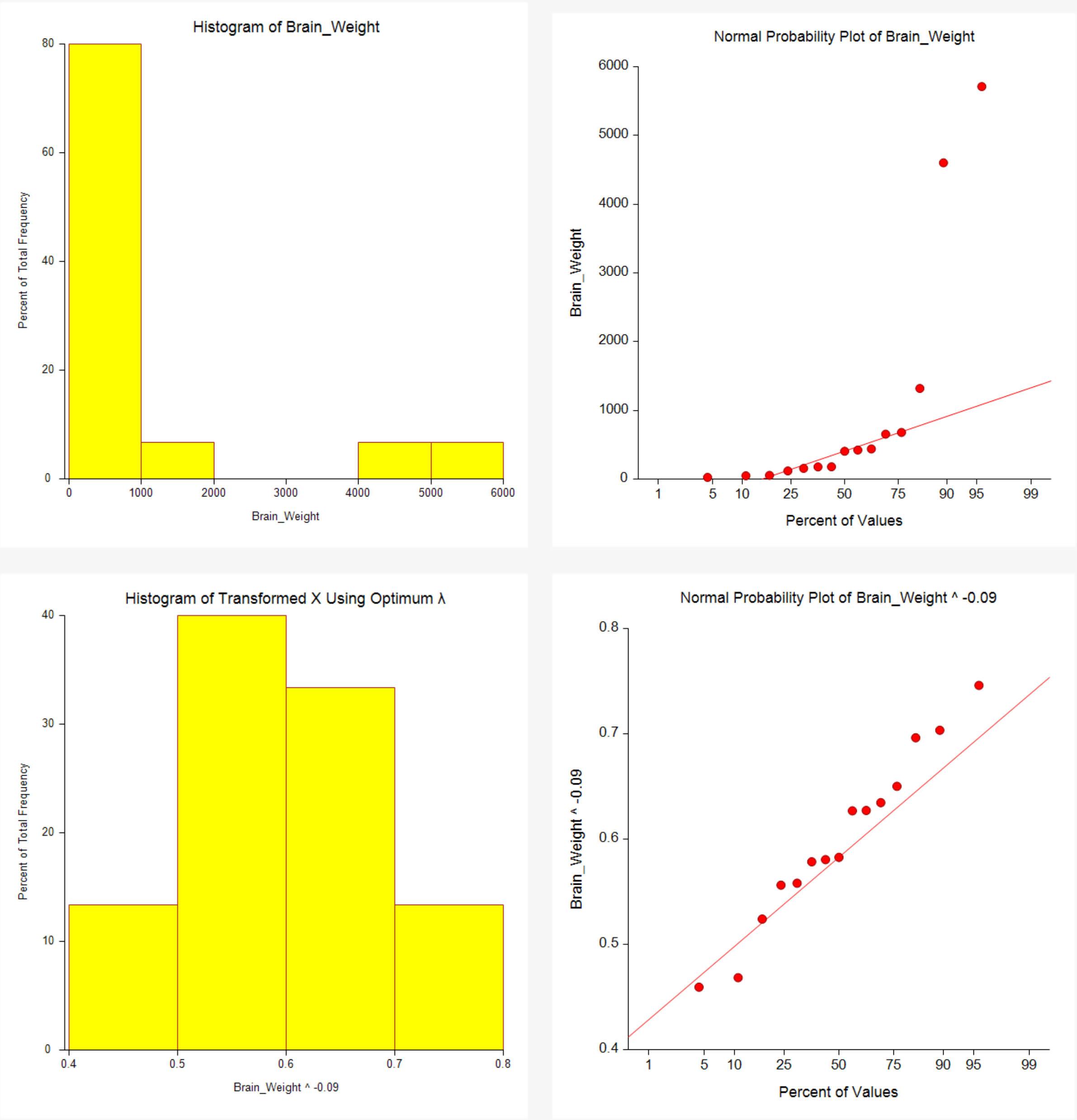
Plot for Choosing λ



This plot gives a visual representation that will help you select the value of λ that you want to use. The optimum value found by maximum likelihood is plotted with a large, red asterisk. This value is usually inconvenient to use, so a convenient (standard) value is sought for that is close to the optimum value. These convenient values are plotted using a blue circle with a red center. In this example, it is obvious that λ = 0 is certainly a reasonable choice. The large, shaded area in the middle of the plot highlights the values of λ that are within the confidence interval for the optimum.

Note that this plot was created using the Scatter Plot procedure. The shading effects and different plot symbols were made by making several groups of data.

Plots for Assessing Normality at Various λ's



These plots let you see the improvement towards normality achieved by the power transformation. The top row shows the histogram and probability plot of the original data. The lack of normality is evident in the two plots. The bottom row shows the same two plots applied to the data that has been transformed by the optimum λ. They are now much closer to being normally distributed.