Making Data Normal Using Box-Cox Power Transformation

isixsigma.com/tools-templates/normality/making-data-normal-using-box-cox-power-transformation

February 26, 2010

Normally distributed data is needed to use a number of statistical analysis tools, such as individuals control charts, C_p/C_{pk} analysis, *t*-tests and analysis of variance (ANOVA). When data is not normally distributed, the cause for non-normality should be determined and appropriate remedial actions should be taken. (An introduction to remedial actions for non-normal data can be found in "Dealing with Non-normal Data: Strategies and Tools.")

Data transformation, and particularly the Box-Cox power transformation, is one of these remedial actions that may help to make data normal. By understanding both the concept of transformation and the Box-Cox method, practitioners will be better prepared to work with non-normal data.

What Are Transformations?

Transforming data means performing the same mathematical operation on each piece of original data. Some transformation examples from daily life are currency exchange rates (e.g., U.S. dollar into Euros) and converting degree Celsius into degree Fahrenheit.

These two transformations are called linear transformations because the original data is simply multiplied or divided by a specific coefficient or a constant is subtracted or added. But these linear transformations do not change the shape of the data distribution and, therefore, do not help to make data look more normal (Figure 1).

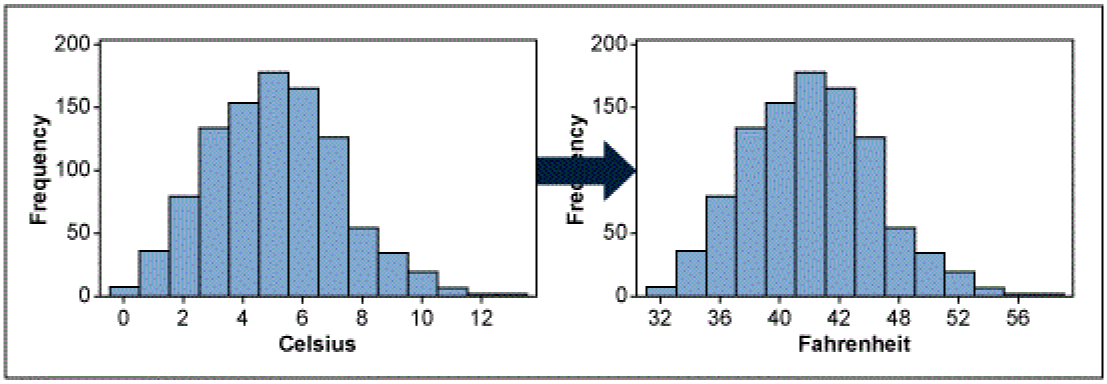

Figure 1: Linear Transformation of Degrees Celsius to Degrees Fahrenheit

What is the Box-Cox Power Transformation?

The statisticians George Box and David Cox developed a procedure to identify an appropriate exponent (Lambda = l) to use to transform data into a "normal shape." The Lambda value indicates the power to which all data should be raised. In order to do this, the Box-Cox power transformation searches from Lambda = -5 to Lamba = +5 until the best value is found. Table 1 shows some common Box-Cox transformations, where *Y'* is the transformation of the original data *Y*. Note that for Lambda = 0, the transformation is NOT *Y* (because this would be 1 for every value) but instead the logarithm of *Y*.

Table 1: Common Box-Cox Transformations

l	Y'
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(Sqrt(Y))$
	log(Y)
0.5	$Y^{0.5} = Sqrt(Y)$
1	$Y^1 = Y$
2	Y^2

Handpicked Content: Tips for Recognizing and Transforming Non-normal Data

An example: Figure 2 shows non-normally distributed cycle time data. Using the Box-Cox power transformation in a statistical analysis software program provides an output that indicates the best Lambda values (Figure 3).

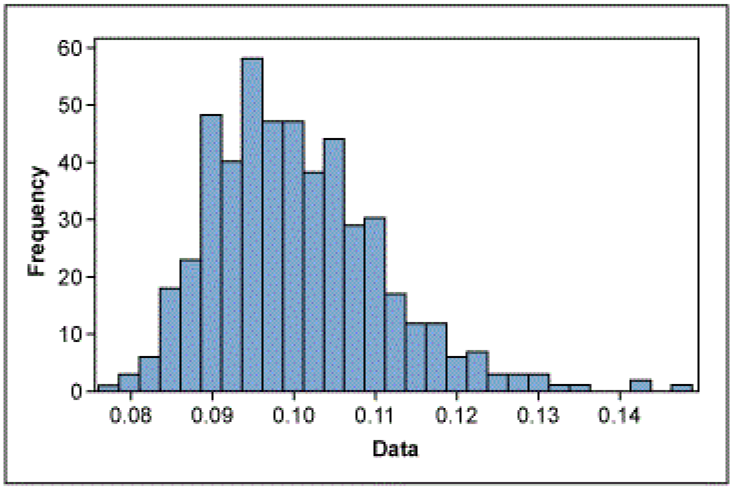

Figure 2: Example of Non-normally Distributed Cycle Time Data

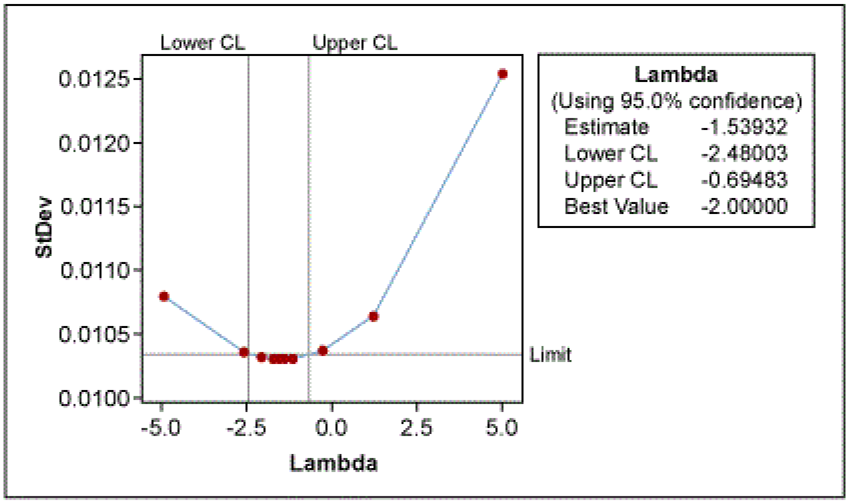

Figure 3: Example Box-Cox Plot of Data

The lower and upper confidence levels (CLs) show that the best results for normality were reached with Lambda values between -2.48 and -0.69. Although the best value is -1.54 (estimate in Figure 3), the process works better if this value is rounded to a whole number; this will make it easier to transform the data back and forth. The best whole-number values here are -1 and -2 (the inverse function of *Y* and Y^2, respectively). The histogram in Figure 4 shows the transformed data using Lambda = -1, now more normally distributed.

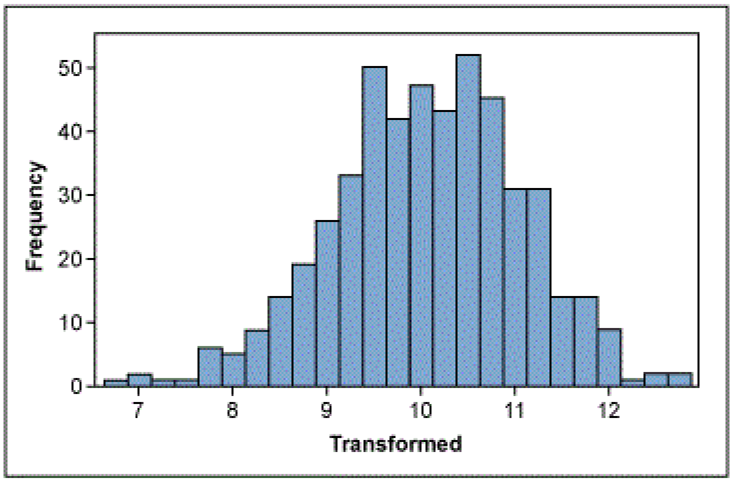

Figure 4: Data Transformed Using Lambda = -1

Does Box-Cox Always Work?

The Box-Cox power transformation is not a guarantee for normality. This is because it actually does not really check for normality; the method checks for the smallest standard deviation. The assumption is that among all transformations with Lambda values between -5 and +5, transformed data has the highest likelihood – but not a guarantee – to be normally distributed when standard deviation is the smallest. Therefore, it is absolutely necessary to always check the transformed data for normality using a probability plot.

Additionally, the Box-Cox Power transformation only works if all the data is positive and greater than 0. This, however, can usually be achieved easily by adding a constant (*c*) to all data such that it all becomes positive before it is transformed. The transformation equation is then:

Handpicked Content: Resource Page: A Primer on Non-normal Data

$Y' = (Y+C)^l$

Application Example

A project team collected cycle time data from a purchase order-generation process. One team member created a control chart of this data (Figure 5) and was about to ask what special cause had happened for data point 40 when the Green Belt remembered that using an individuals control chart requires normally distributed data. A look at the probability plot of the data (Figure 6) revealed non-normal distribution. Therefore, the control limits of the control chart were useless.

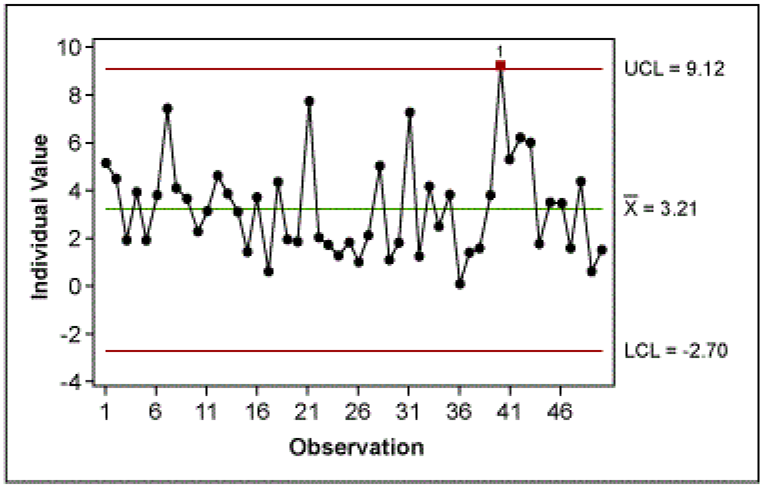

Figure 5: Control Chart of Original Cycle Time Data

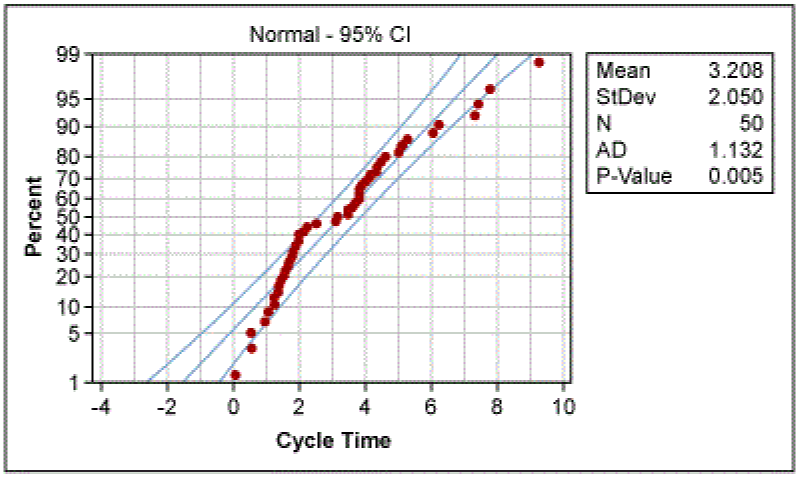

Figure 6: Probability Plot of Original Cycle Time Data

The Green Belt used the Box-Cox power transformation to determine whether the data could be transformed (Figure 7). Box-Cox suggested a best Lambda value of 0.5 for transformation (i.e., the square root of the original data). And the transformation really worked: The new probability plot confirms normality (Figure 8).

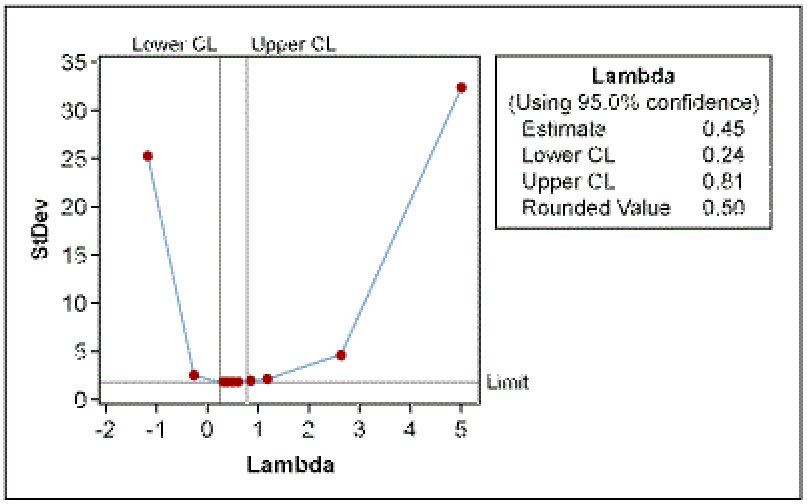

Figure 7: Box-Cox Plot of Cycle Time Data

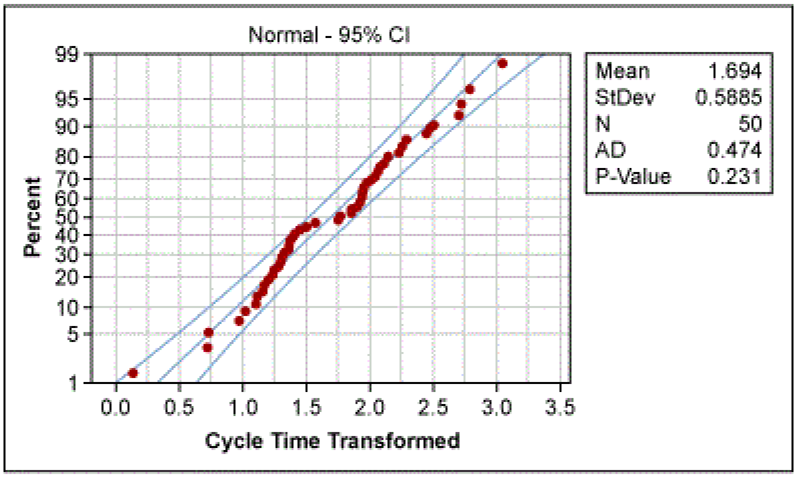

Figure 8: Probability Plot of Transformed Cycle Time Data

After the transformation, the Green Belt created a control chart of the transformed data and showed that the purchase order-generation process was actually quite stable, i.e., all variation was due to common causes (Figure 9).

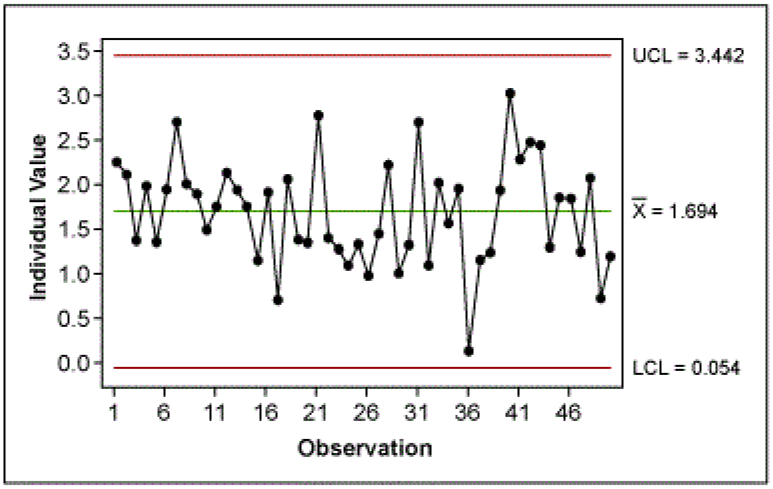

Figure 9: Control Chart of Transformed Cycle Time Data

Because the individual values of the transformed data have no practical meaning, the Green Belt had to re-create a control chart for the original data, but this time with the correct control limits (Figure 10). To do this, the Belt used the upper and lower CLs from the control chart of the transformed data and transformed them back into their original values. Because the transformation operation was taking the square root, the back-transformation involved squaring the transformed data:

Handpicked Content: Are You Sure Your Data Is Normal?

UCL = UCL'2 = 3.4422 = 11.847
LCL = LCL'2 = -0.0542= 0.003

For the mean, the Belt used the mean of the original data.

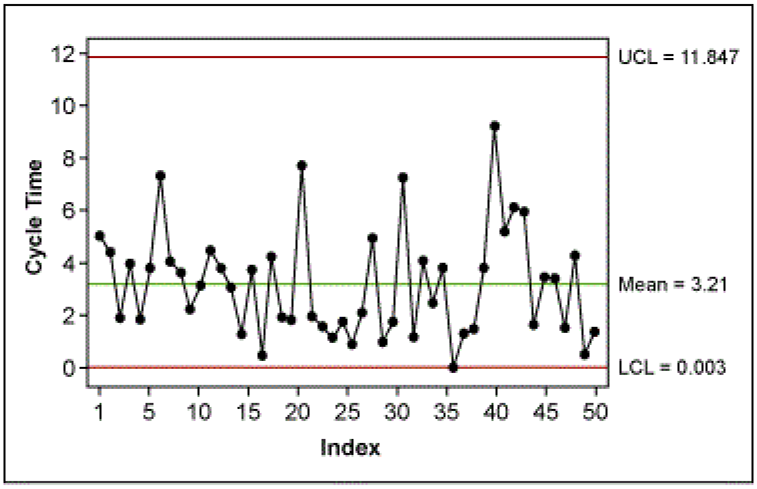

Figure 10: Control Chart of the Original Data with Correct Control Limits

This control chart could then be used for the ongoing monitoring of the purchase order-generation process.