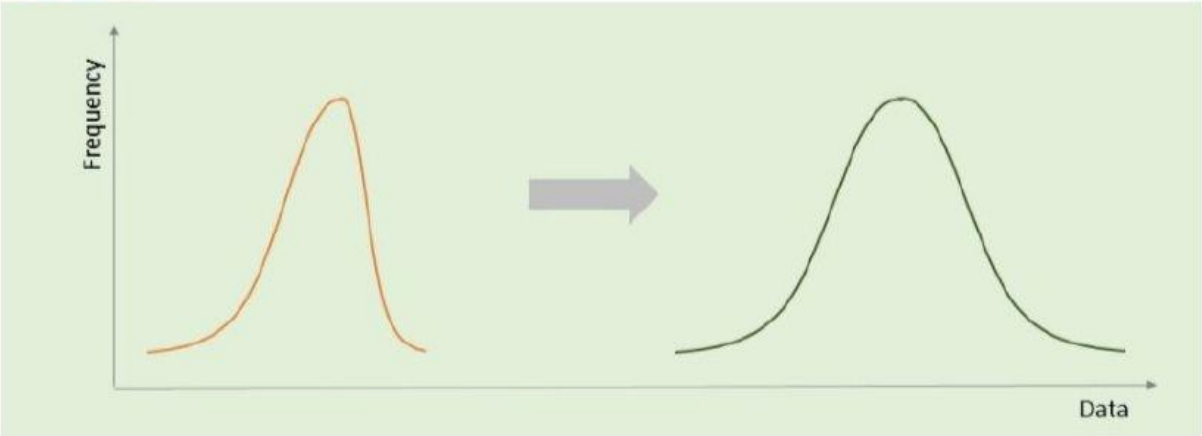
How do I transform my data to a normal distribution?

View All Blogs



Check for Outliers

The first thing we need to do check if the data is not normal because of any outliers. A normal data does not have any outliers – hence, if there are outliers in your data, then that may be the reason that the data is not normally distributed. First we need to check if the outliers in the data are because of any data entry errors. If so, we can correct the data and then check if the data is normally distributed. If there are no data entry errors, the next question to ask is if the outliers are because of some special causes which are not going to recur in the future. If so, it may be okay to note the reasons and then delete these outliers. However, if these outliers have a chance of recurring in the future then it would not be appropriate to just blindly delete them from analysis. We need to look for other ways of handling this data.

Box-Cox Transformation

The second approach is to transform the data such that the transformed data is normally distributed. There are some transformations that have been found to make the transformed data normal. For example, if you square the data values, the squared values may be normal. Or, in some cases, the square root of the data or the reciprocal of the data may be normally distributed. In other cases, the logarithm of the data may be normally distributed. Such simple transformations of the data to make the data normal can be grouped together under a transformation called the Box-Cox transformation. The Box-Cox transformation is given by the following general formula: $y = x^{\lambda}\ for\ \lambda \neq 0\ and\ y = \ln(x)\ for\ \lambda = 0$ **Where, x is the raw data and y is the transformed data and lambda is the transformation constant. If lambda =**

1, then there is no transformation. If lambda = 2, then it is the square transformation and so on. The following table provides the names of some standard transformations:

Lambda	Standard Transformation
-3	Inverse Cube
-2	Inverse Square
-1	Inverse
-0.5	Inverse Square Root
0	Logarithmic
0.5	Square Root
1	No Transformation
2	Square
3	Cube

How to Fit the Box-Cox Transformation

There are several approaches to determine the value of lambda for the Box-Cox transformation. The most commonly used approach is to use the Most Likely Estimate (MLE) approach. Getting into the details about this approach is out of scope of this article. A simple approach to determine the value of lambda is to vary the value of lambda from -5 to +5 and then determine which value of lambda produces a distribution that is as close to a normal distribution as possible. The value of lambda is selected that provides the smallest value of the standard deviation of the variation between the transformed data and a normally distributed data. Of course, lambda can take any value (say 1.63), but it may be hard to explain what that means to others. Some users like to choose the Box-Cox transformation to the values of lambda shown in the above table so that the transformation can be easily understood by the users. Note that there is no guarantee that a Box-Cox transformation will always result in a normal distribution. It is possible that none of the values of lambda can result in a normally distributed data..

Johnson Transformation

A third approach to transform the data to a normal distribution is to use another type of more complex transformation called the Johnson family of transformations. There are three different families of Johnson distributions:

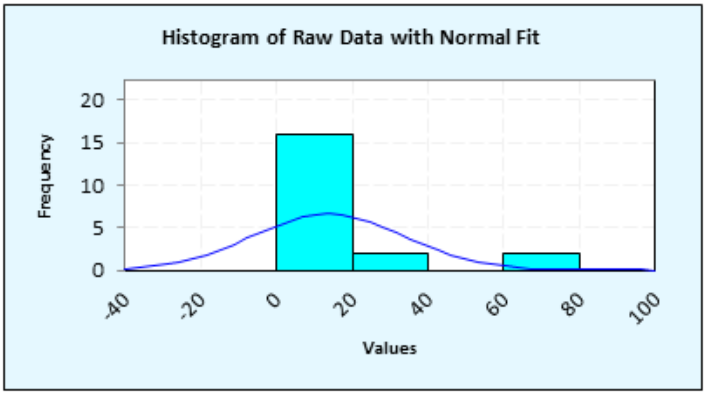
Johnson Distribution	Formula
SU	$Y = \gamma + \eta sinh^{-1}\left(\frac{x-\epsilon}{\lambda}\right)$
SB	$Y = \gamma + \eta log\left(\frac{x-\epsilon}{\lambda+\epsilon-x}\right)$
SL	$Y = \gamma + \eta \log\left(\frac{x-\epsilon}{\lambda}\right)$

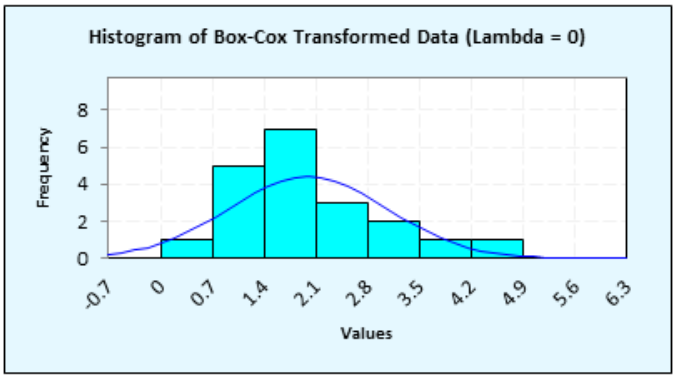
:

Where, Y is the transformed data, X is the raw data, and eta, epsilon, and lambda are the Johnson parameters. Decision rules have been formulated for the selection of the appropriate Johnson family of distributions SU, SB, and SL. There are several algorithms available to fit the Johnson parameters for a given data set. However, due to complex nature of these algorithms, the solutions are not very straightforward and require the use of appropriate software to estimate these parameters. Similar to a Box-Cox transformation, a computer can run through several combinations of these Johnson parameters to determine which set of parameters makes the transformed data as close to normal as possible. Since there are several parameters to fit the Johnson transformation, we usually find that a Johnson transformation does a better job of transforming the data to a normal distribution compared to a Box-Cox transformation. Similar to the Box-Cox transformation, there is no guarantee that a Johnson transformation will be successful in transforming a data to the normal transformation. It should be pointed out that when you transform the raw data using one of these transformations, the specification limits also need to be transformed if you need to calculate the process capability.

Example

Let's say that we need to determine if a delivery process is in control. The primary metric of interest is the time to deliver an order from placement in the system till the order it signed and received by the customer. One of the ways to do this is to develop a control chart. In this exercise, since the primary metric is continuous and data is collected individually (subgroup size of 1), the appropriate chart to use is the I-MR chart. I-MR chart assumes that the data needs to be normally distributed. The delivery time data for this example are: 22, 76, 10, 2, 3, 6, 7, 5, 6, 2, 8, 1, 3, 5, 63, 4, 22, 13, 4, 12. We first draw the histogram of the data and check for the normality of the data. Looking at the histogram and reviewing the P-value, it is clear that the data is not normal. The next step is to check for any data entry errors or the reason for the outliers – let's say we have investigated them and did not find any issues with this data. The next step is to transform the data. Let's try to use a Box-Cox transformation. The results of the Box-Cox transformation are shown in the following figure. The P value of the raw data was <0.001 (not-normal) and after the transformation, the P value is 0.381 (normal)





A Johnson transformation is also shown in the figure below. From the transformed data, it is clear that the data is transformed into a normally distributed data. The P value of the transformed data is 0.99 (normal). We can see that the Johnson transformation did an excellent job of transforming the data to a normal distribution. We can now plot the I-MR chart for this transformed data which shows that the process is in control (this topic is out of scope of this article).