Further short notes on data transformations

A question has been raised on the need for data transformation. This can be answered as below.

Many of the most common statistical procedures adopted are known as *parametric statistics*, meaning that they make certain assumptions about the distribution of the data population from which the sample has been drawn. For the detailed discussion on the types of assumption to be made, the previous blogs on "*DOE – Strategies for checking model assumptions – Parts 1 and 2*" can be referred.

If the sample data indicate that these assumptions have not been met, we have several options for analyzing the data. One is to use alternate, *nonparametric statistical* procedures which make fewer or no assumption of the data distribution.

Another possibility is to transform the data in some way so that the assumptions of the desired parametric statistical procedure are met. There are many ways to transform data, depending on the distribution involved and the assumptions violated.

For example, we can transform a data set in order to make it close to a normal distribution. The very first step in data transformation is to evaluate the data set and decide which, if any, transformation might be appropriate. In addition to the common way of graphing the data such as creating a histogram with a superimposed normal curve to allow a visual evaluation of the general shape of the data as well as the opportunity to identify outliers, we can also use another approach by computing one of the test statistics designed to test whether the data fits a particular distribution.

Two statistics commonly used for this purpose are the Anderson-Darling (A-D) and the Kolmogorov-Smirnov (K-S). We can find these test statistics in many statistical software packages and various statistical calculations are also available on the internet.

The Anderson-Darling statistic equations for data normality and independence are:

$$A^2 = -\frac{\sum_{i=1}^{n} (2i-1)\left[\ln(p_i) + \ln(1 - p_{n+1-i})\right]}{n} - n$$

$$A^{2*} = A^2(1+\frac{0.75}{n}+\frac{2.25}{n^2})$$

where:

A^2 is the A-D normal statistic estimate

A^2* is the corrected A-D normal statistic estimate

p_i is the normal probability value at data point *i*.

n is the number of data point

The variability of the data set is analyzed by using its sample standard deviation (s) as well as the sample standard deviation of its moving range (MR), defined as the difference between two successive data points, indicating the stability of these data. The evaluations of A_s^{2*} and A_{MR}^{2*} results are based on the following thumbs of rule:

a) If $A_s^{2*} < 1.0$ and $A_{MR}^{2*} < 1.0$, we accept that the data points are normal and independent;

b) If $A_s^{2*} > 1.0$ and $A_{MR}^{2*} > 1.0$, we conclude that the data are not normal and independent;

c) If $A_s^{2*} < 1.0$ and $A_{MR}^{2*} > 1.0$, it indicates that the data points are not fully independent on each other