Identifying the distribution of data is key to analysis

There is a simple way to find the true distribution of your data so you can select the appropriate analysis.

Knowing the distribution of your data is essential to choosing the right statistical method. Suppose you need to assess the capability of your process. If you conduct an analysis that assumes the data follow a normal distribution but in fact the data are nonnormal, your results will be inaccurate.

To avoid this costly error, you must determine the distribution of your data.

So, how do you determine the distribution? Minitab's Individual Distribution Identification provides a simple way to find the distribution of your data so you can select the appropriate analysis. You can use this tool to:

- Verify that a distribution used historically is still valid for the current data.
- Choose the right distribution when you're not sure which to use.
- Transform your data to follow a normal distribution.

In many cases, your process knowledge helps you identify the distribution of your data. In these situations, you can use Individual Distribution Identification to confirm that this distribution fits the current data.

Suppose you want to perform a capability analysis to ensure that the weights of ice cream-filled containers from your production line are meeting specifications. In the past, these data have followed a normal distribution, but you want to confirm normality for the present fill weights. You can use Individual Distribution Identification to generate a probability plot and quickly assess the fit.

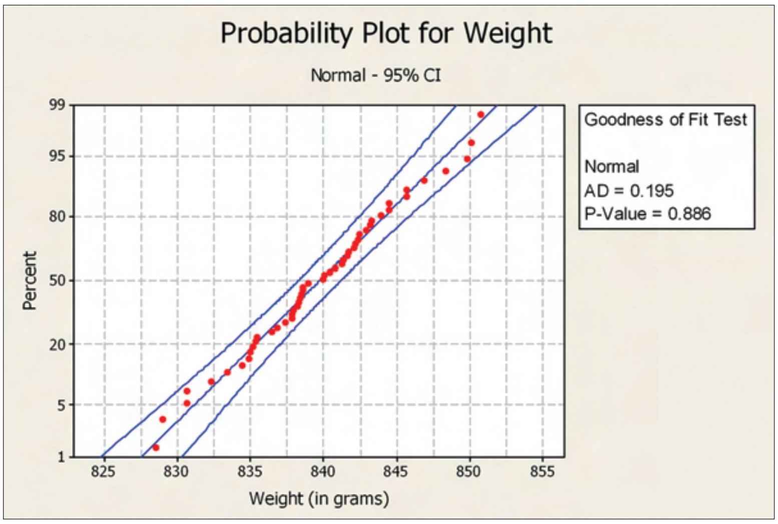

Fig. 1. The normal probability plot shows that the data follow a normal distribution.

A given distribution is a good fit if:

- The plotted points roughly follow a straight line.
- The goodness of fit test p-value is greater than 0.05 (or your chosen alpha level).

Based on these criteria, the ice cream fill weights data appear to follow a normal distribution. Therefore, you can justify the use of normal capability analysis.

Suppose you have successfully used more than one distribution in the past to model a particular measurement. You can use Individual Distribution Identification to help you decide which distribution best fits your current data. For example, you want to assess whether a particular weld strength is meeting customers' requirements. After measuring the weld strengths, you can use Individual Distribution Identification to choose the distribution that best fits your data.

In this case, the data are modelled using the lognormal, Weibull, smallest extreme value and logistic distributions. The probability plot shows that the lognormal distribution is a better fit than the other distributions because the plotted points on the lognormal probability plot roughly follow a straight line. In addition, the p-value for the lognormal distribution is the highest above 0.05.

You can evaluate up to 14 different distributions in Minitab, including 1-, 2-, and 3-parameter distributions. When you fit your data with both a distribution and its higher-parameter counterpart, the higher-parameter distribution often appears to be a better fit. For example, if you fit your data using both a 2-parameter and a 3-parameter Weibull distribution, the 3-parameter Weibull distribution may appear to provide a better fit. However, because the 3-parameter distribution is more restrictive, you would only want to use the 3-parameter Weibull distribution if it offers a significantly better fit. You can use the likelihood ratio test to assess the fit and choose between the two distributions.

While Minitab offers various options for working with nonnormal data, many practitioners simply prefer to use the broader palette of normal statistical techniques. The good news is that, in addition to finding the true distribution of your data, Minitab's Individual Distribution Identification can transform your

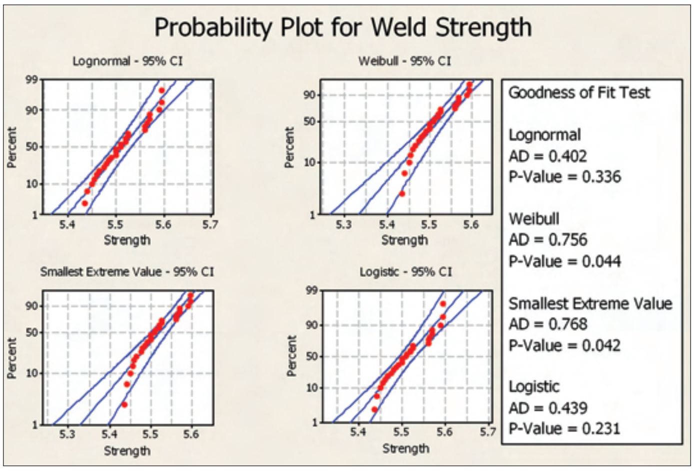

Fig. 2. Assess the fit of multiple distributions to see which distribution best fits your data.

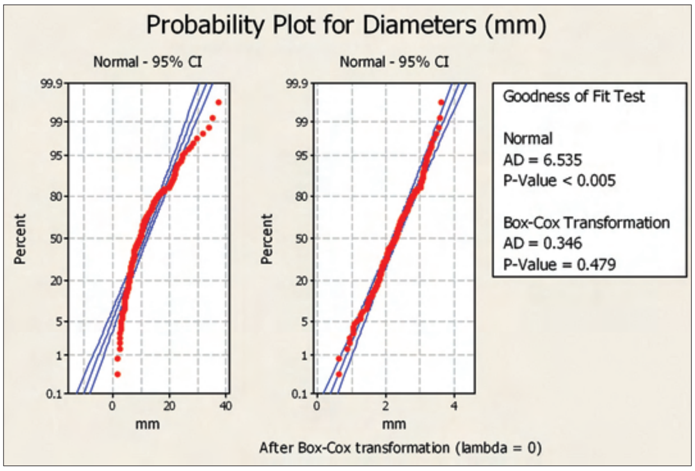

Fig.3. Use the Box-Cox transformation to produce data that follow a normal distribution.

nonnormal data to follow a normal distribution using the Box-Cox transformation or the Johnson transformation. You can then use the transformed data with any tool that assumes normality.

In this case, the probability plot and corresponding p-value suggest that the data are successfully transformed to follow a normal distribution when using the Box-Cox transformation. You can now use the transformed data for future analysis.

Transforming data does not always result in normal data. You must check the probability plot and p-value to assess whether the normal distribution fits the transformed data well.

It is always a good practice to know the distribution of your data before analysing them. Minitab's Individual Distribution Identification is an easy-to-use tool that can help you identify the distribution of your data and eliminate the consequences of an analysis conducted using an inappropriate distribution. You can use this feature to check the fit of a single distribution, or use it to compare the fits of several distributions, selecting the one that fits best. If you prefer to work with normal data, you can even use Individual Distribution Identification to transform your nonnormal data to follow a normal distribution. ❍

Submitted by Minitab's Documentation Department. This is part of a series of articles entitled ***Accessing the Power of Minitab.*** *Visit www.minitab.com/accessingthepower to learn more*

This article first appeared in *ASIA PACIFIC ENGINEER,* December 2007, visit www.engineerlive.com