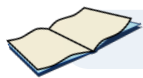


Tips and Tricks for Analyzing Non-Normal Data

Many statistical analyses are based on an assumed distribution—in other words, they assume that your data resemble a certain shape. And the most commonly assumed distribution, or shape, is the normal distribution. However, normally distributed data isn't always the norm.



A **normal distribution** has a symmetric bell shape and is centered at the mean.

Some measurements naturally follow a non-normal distribution. For example, non-normal data often results when measurements cannot go beyond a specific point or boundary. Consider wait times at a doctor's office or customer hold times at a call center where it's not possible to wait a negative amount of time. These scenarios have a hard boundary at 0, which can skew the data to the right.

This article will cover various methods for detecting non-normal data, and will review valuable tips and tricks for analyzing non-normal data when you have it.

Normal or Not

Several graphical and statistical tools can be used to assess whether your data follow a normal distribution, including:

Histogram

Do your data resemble a bell-shaped curve?

Normality Test

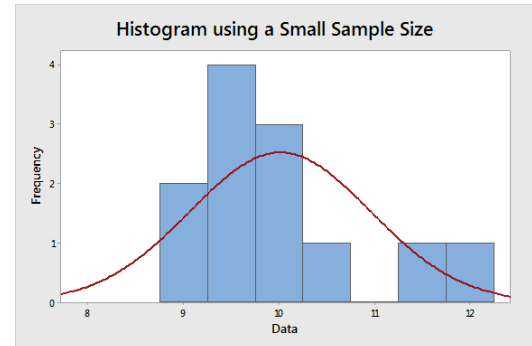
Is the p-value greater than your α -level (e.g. $\alpha=0.05$)?

Probability Plot

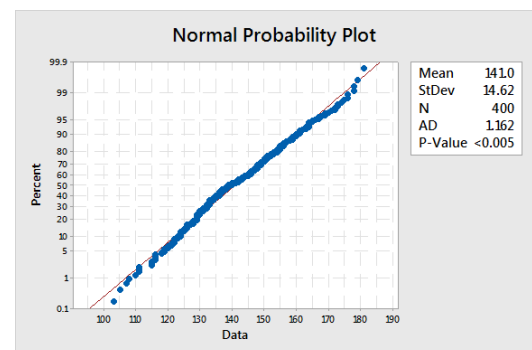
Do the plotted points follow a straight line?

Answering "yes" to the questions above typically indicates that your data follow a normal distribution. However, these tools can be misleading.

If you have a small sample size ($n < 30$), a histogram may falsely suggest the data are skewed or even bimodal. Similarly, if you have a large sample size ($n > 200$), the Anderson-Darling normality test can detect small but meaningless departures from normality, yielding a significant p-value even when the normal distribution is a good fit.



The 10 data points graphed here were sampled from a normal distribution, yet the histogram appears to be skewed. **Helpful hint:** Avoid histograms for small sample sizes.



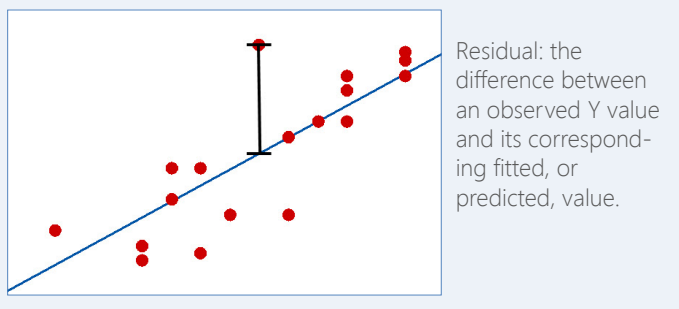
The Anderson-Darling normality test p-value for these 400 data points indicates non-normality, yet the probability plot reveals a normal distribution. **Helpful hint:** Beware of significant p-values for large sample sizes.

Sometimes Normality Simply Doesn't Matter

For some statistical analyses, it doesn't actually matter if your data follow a normal distribution or not. They are described by statisticians as "robust to the normality assumption," and include:

- t-tests
- Xbar control charts
- Gage R&R
- Equivalence tests

And for certain analyses, it's not the actual data that should follow a normal distribution, but rather the residuals.



Take regression, design of experiments (DOE), and ANOVA, for example. You don't need to check Y for normality because any significant X's will affect its shape—inherently lending itself to a non-normal distribution.

Analyzing Non-Normal Data

When you do have non-normal data and the distribution does matter, there are several techniques available to properly conduct your analysis.

1. Nonparametrics

Suppose you want to run a 1-sample t-test to determine if a population's average equals a specific target value.

Although t-tests are robust to the normality assumption, suppose you have a small sample size and are concerned about non-normality. Or, suppose you have a sufficient sample size, but you don't believe the average is the best measure of central tendency for your data.

Instead of a parametric test such as the t-test, which is based on the mean of the data, you can use a nonparametric, distribution-free test such as the 1-sample sign test to test if the *median* is on target.

Nonparametric alternatives are available for the most commonly used parametric hypothesis tests:

Parametric Tests	Nonparametric Tests
1-Sample t	1-Sample Sign or Wilcoxon
2-Sample t	Mann-Whitney
One-Factor ANOVA	Kruskal-Wallis or Mood
Two-Factor ANOVA	Friedman
Pearson Correlation	Spearman Correlation

Note there are also nonparametric approaches to analyses beyond hypothesis tests, such as tolerance intervals.

2. Alternative Distributions

For certain analyses that are highly sensitive to the normality assumption, such as reliability and survival probabilities, finding a distribution that adequately fits the data is critical. And for reliability applications, it's quite normal to have non-normally distributed data.

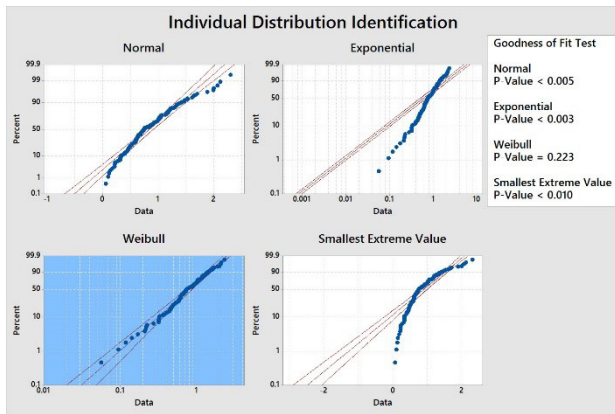
For instance, the Weibull distribution is quite common when modeling time-to-failure data. This versatile distribution can be skewed left, skewed right, or even approximately symmetric.

When analyzing data where the risk of failure does not depend on the age of the unit, the exponential distribution may be most suitable. For example, failure times for many electrical components typically follow an exponential distribution.

Other common distributions for reliability data include the lognormal distribution for modeling cycles-to-failure in fatigue and load testing, and the extreme value distribution for modeling breaking strength.

And if you're not sure what distribution best fits your data, then you can use tools like Minitab's Individual Distribution Identification to find out.

When your data follow a Weibull, exponential, or some other non-normal distribution, you don't have to be restricted to using the normal distribution to run your analysis. Instead, use the distribution that best fits your data.



The probability plots and p-values reveal that a Weibull distribution provides the best fit for this data set.

3. Transformations

Capability analysis, which is used to determine if a process falls within specifications, is also highly sensitive to the distribution assumption.

Like reliability analysis, you can use a non-normal distribution to calculate process capability, or alternatively, you can try to transform your data to follow a normal distribution using either the Box-Cox or Johnson transformation.

When you transform your data, you modify the original data using a function of a variable. Functions used in the Box-Cox transformation are simple, and include taking the square root, inverse, or natural log of the original data.

Functions used in the Johnson transformation are more complex than Box-Cox, but they are also more powerful. Here's an example of what a Johnson transformation function looks like:

$$-3.80937 + 1.67123 * \ln(X + 0.160537)$$

Although the Box-Cox and Johnson transformations often successfully transform non-normally distributed data to normally distributed data, they are not foolproof.

Michelle Paret, product marketing manager, Minitab Inc. michelle.paret@minitab.com

Minitab is the leading provider of statistical software for quality improvement and statistics education. More than 90% of Fortune 100 companies and more than 4,000 colleges and universities use Minitab 17, our flagship product. Visit www.minitab.com to learn more or contact us at sales@minitab.com.

Sometimes the transformed data will not follow a normal distribution, just like the original data. In that case, consider using an alternative distribution, as described for reliability analysis.

The End

Non-normal data can occur for many reasons. Perhaps your data:

- Were sampled from different populations (locations, genders, seasons, etc.).
- Shift and drift over time.
- Contain extreme outliers.
- Have insufficient resolution (too few significant digits).

And some data sets exhibit none of the above, but instead are inherently non-normal, as we previously discussed.

Whatever the case may be, the first step in analyzing non-normal data is to understand *why* it's non-normal, especially if you have reason to believe it should be normal.

For example, if you discover your data are in fact sampled from different populations, consider analyzing each population independently. Perhaps each group when treated separately is normally distributed.

Once you understand the distribution and associated properties of your data, there are many tools—all available in Minitab Statistical Software—to help you properly analyze it, gain valuable insight, and find meaningful solutions to your toughest problems.