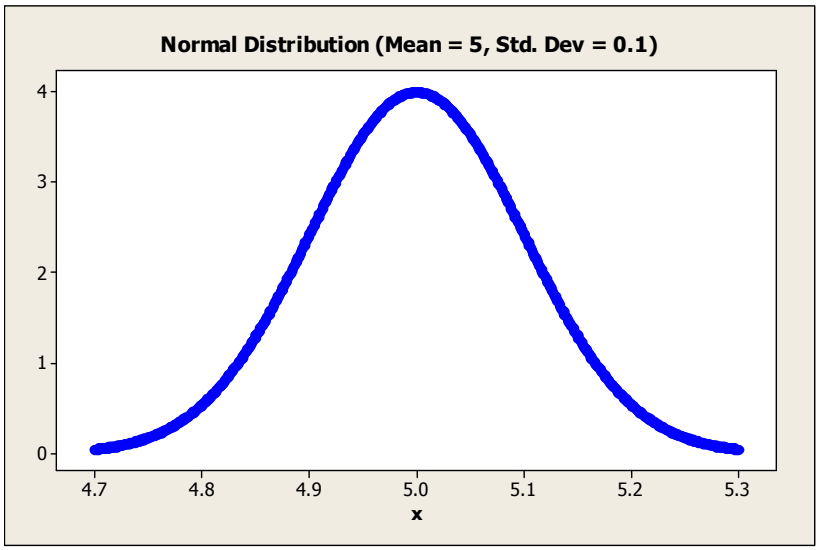
How do I Test my Data for Normality?

Steven Wachs
Principal Statistician
Integral Concepts, Inc.

Many statistical tests and procedures assume that data follows a normal (bell-shaped) distribution.



For example, all of the following statistical tests, statistics, or methods assume that data is normally distributed:

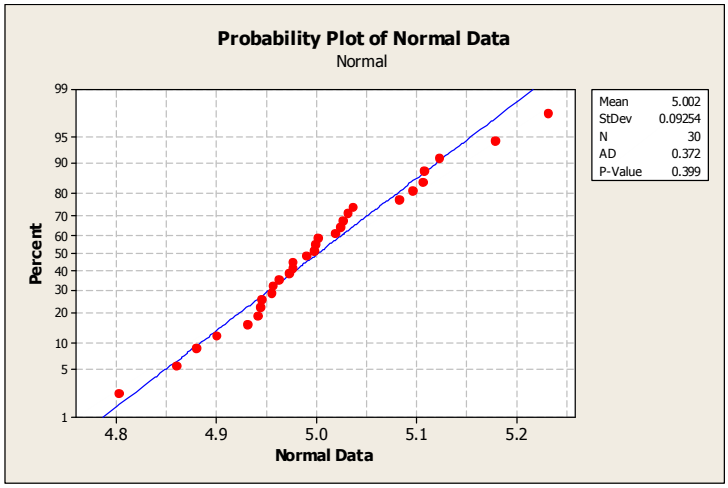
- Hypothesis tests such as t tests, Chi-Square tests, F tests
- Analysis of Variance (ANOVA)
- Least Squares Regression
- Control Charts of Individuals with 3-sigma limits
- Common formulas for process capability indices such as C_p and C_{pk}

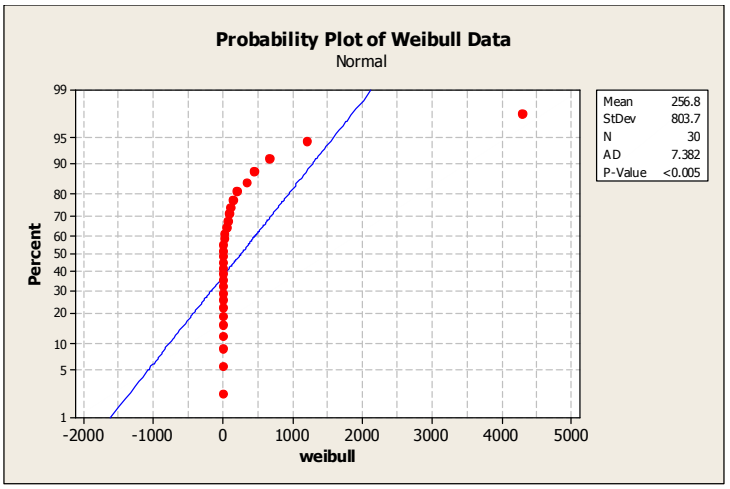
Before applying statistical methods that assume normality, it is necessary to perform a normality test on the data (with some of the above methods we check *residuals* for normality). We hypothesize that our data follows a normal distribution, and only reject this hypothesis if we have strong evidence to the contrary.

While it may be tempting to judge the normality of the data by simply creating a histogram of the data, this is not an objective method to test for normality – especially with sample sizes that are not very large. With small sample sizes, discerning the shape of the histogram is difficult. Furthermore, the shape of the histogram can change significantly by simply changing the interval width of the histogram bars.

Normal probability plotting may be used to objectively assess whether data comes from a normal distribution, even with small sample sizes. On a normal probability plot, data that follows a normal distribution will appear linear (a straight line). For example, a random sample of 30 data points from a normal distribution results in the first normal probability plot (below left). Here, the data points fall close to the straight line. The second normal probability plot (below right) illustrates data that does not come from a normal distribution.

PO Box 251652 • West Bloomfield, MI 48325 • Tel. 248-421-7590 • Fax. 248-539-3858 • www.integral-concepts.com





The "P-Value" on the plot legend is based on the computed Anderson-Darling statistic and provides an objective assessment of the hypothesis that the data is normally distributed (comes from a normal distribution). How the p-value is formally defined is an advanced subject and is not discussed in this article. Here, we focus on the interpretation.

If the p-value is "small" (usually less than 0.05), then we have strong evidence that the data is **not normal** (does not come from a normal distribution). If the p-value is "large" (usually more than 0.10), then we assume that the data is normal (comes from a normal distribution). P-values larger than 0.10 tell us that there is not sufficient evidence to reject the normality assumption (hypothesis). In this case, we are justified in using statistical methods that assume normality.

Many methods are available to handle non-normal data and these should be utilized when necessary. Applying methods which assume the normal distribution when this assumption is not valid often results in incorrect conclusions.

PO Box 251652 • West Bloomfield, MI 48325 • Tel. 248-421-7590 • Fax. 248-539-3858 • www.integral-concepts.com