Enhanced Normal Probability Plots

A.I. McLeod January 1997

The Splus function `nplot`, available in the Splus library `aimlm`, produces an *Enhanced Normal Probablity Plot.* Improvements include:

- 0.5% significance limits.
- Statistical tests useful in assessing non-normality are displayed.
- The sample size, n, is shown.
- The plot allows for interactive user identification of outliers using the mouse.

The enhanced normal probability plot provides much more information than the raw probability plot produced by Splus's `qqnorm` or the Trellis Graphics' `qqmath`.

The boxplot aids in assessing possible skewness in the distribution. The extra tests help one decide if apparent departures are statistically significant. Sometimes, this judgement is quite difficult to make when it is based only on the normal probablity plot especially when one bears in mind that the empirical quantiles are correlated so that spurious patterns in the plot may just be due to inherent randomness and not non-normality.

Let the observed data be denoted by $X_1, X_2, \ldots, X_n$. Then we plot the empirical data quantiles vs. the corresponding theoretical normal quantiles. The *quartile line* which is determined by the quartiles of the two distributions is shown.

This quartile line helps in the interpretation of tail behaviour (fat or thin) The normal probability plot sheds insights into the tail behaviour of the distribution. The tail, either right or left, of the observed distribution is thicker or thinner than the theoretical normal distribution according as the corresponding slope of the normal probability plot is $> \hat{\sigma}$ or $< \hat{\sigma}$.

The technical terms for thin and thick tails are platykurtosis and leptokurtosis. In other words, if the tail of the observed distribution lies below the fitted straight line, the empirical data distribution is said to be leptokurtic in that tail. Similarly, if the observed data distribution is above the straight line, the data distribution is said to be platykurtic in that tail. Actually leptokurtic means slender!! *Since there is less probability mass in the center of the distribution relative to the normal, the distribution is more slender in the center.* NOTE THIS IS THE REVERSE OF MODERN STANDARD USAGE WHICH CHARACTERIZES THE TAIL MASS RATHER THAN THE MASS IN THE CENTER. Leptokurtic distributions generate more outliers than the normal. Common examples of symmetric leptokurtic distributions include: the contaminated normal (with same means but different variances), the t-distribution especially on a low number of degrees of freedom, the Cauchy distribution, the family of stable distributions with parameter less than 2. Symmetric platykurtic distributions are somewhat less common in statistical practice but occasionally occur. Theoretical symmetric platykurtic distributions include members of the β distribution family such as the uniform distribution.

The quartile line also provides *robust* estimates of the mean and standard deviation of the data. In this setup with the data quantiles on the vertical axis, the quartile line vertical intercept estimates the mean of the data distribution and the slope estimates the standard deviation.

Some of the terms used on the the plot are defined below.

Empirical Quantiles

These are the ordered X-values: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$.

Plotting Positions

Each observed quantile $X_{(i)}$ corresponds to the percentile:

$$p_i = \frac{i - \frac{1}{2}}{n}.$$

Normal Quantiles

The normal quantiles $q_1, q_2, \ldots, q_n$ are given by

$$q_i = \Phi^{-1}(p_i), \quad i = 1, \ldots, n,$$

where, $\Phi^{-1}(p)$ denotes the inverse normal cumulative distribution function or for short, simply, the quantile function.

Skewness Coefficient

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^{\frac{3}{2}}}.$$

If $g_1 > 0$ this means the data is skewed to the right or positively skewed. Similarly, a value $g_1 < 0$ implies the data are skewed to the left. The skewness test has high power for alternative distributions possessing non-zero skewness. The two-sided significance level, under the null hypothesis that the data is NID(μ, σ^2), is calculated using the method of D'Agnostino (1970).

Michael's Statistic

Michael's D_{SP} test statistic (Michael, 1983) enables significance limits to be drawn on the normal probability plot and is much more powerful than the usual standard Kolmogoroff-Smirnov approach. Michael's statistic is derived by applying a variance-stabilization transformation to the Kolmogoroff-Smirnov method. Let $p_i = (i - \frac{1}{2})/n, \quad i = 1, \ldots, n$. Then $D_{SP} = \max|g(f_i) - g(p_i)|$, where $g(x) = (2/\pi)\sin^{-1}(\sqrt{x})$ and $f_i = \Phi(X_{(i)} - \bar{X})/\sqrt{v}$, where $v = \sum(X_i - \bar{X})^2/n$. Royston (1993) provides an algorithm for determining the significance level of an observed value of D_{SP} under the null hypothesis that the data are independent normal with constant variance. Royston also discusses plotting significance limits on the normal probability plot. The plot produced shows the 0.5% significance limits. The value of D_{SP} statistic and its two-sided significance level are also displayed.

Wilk-Shapiro Test

The Wilk-Shapiro statistic W measures the goodness-of-fit of the straight-line in the normal probability plot. Like R^2, the coefficient of determination in regression, W has the following properties:

$0 \leq W \leq 1$

W close to 1, implies a good-fit

W not close to 1, implies a poor-fit.

We test if the observed value of W is significantly smaller than that for normally distributed data. The Wilk-Shapiro test represents the most-powerful all-round test for normality. It is often very good even in small samples. The significance level of the Wilk-Shapiro test is calculated using the algorithm of Royston (1982).

References

D'Agnostino, R.B. (1970), "Transformations to normality of the null distribution of g_1", *Biometrika*, Vol.57, pp.679–680.

Michael, J.R. (1983), "The stabilized probability plot", *Biometrika*, Vol. 70, 11–17.

Royston (1982), "The W test for normality. Algorithm AS 181" *Applied Statistics*, V.31, pp.176–224.

Royston (1993), "Graphical detection of non-normality by using Michael's statistic", *Applied Statistics*, V.42, No.1, pp.153–158.