An Extension of the Normal Probability Plot

Dr. Wolfgang A. Rolke

April 8, 2003

We will discuss an extension of the standard normal probability plot that combines the graphical nature of the plot with a formal hypothesis test for normality and thereby helps in assessing the severity of the departure from the normal distribution. We perform a simulation study which shows that the performance of this method is comparable to other tests for normality.

Tests for Normality, Quantile Plots, Simulation, Normal Probability Plot

1 Introduction

Assessing distributional assumptions of a univariate data set is one of the most common tasks in Statistics. Many standard statistical methods such as the one or two sample t-tests, tests for equality of variances or ANOVA tests depend to some degree on the assumption that the data has an approximate normal distribution. Assessing normality is also an important problem in other areas of science. For example, the classical paper by Shapiro and Wilk (1956), which first described the test later named after them, appears with over 100 citations in the 1994 science citation index, all of them from the applied sciences.

Over the years a number of methods have been proposed for checking normality, all of which fall in one of two categories, namely formal hypothesis tests on the one hand and graphical displays on the other. Both of these approaches have their strengths and their weaknesses. Graphical displays such as histograms, boxplots and normal probability plots are fairly easy to understand and interpret. Besides just showing a departure from the normal distribution they also allow an assessment of the approximate shape of the

distribution which is really present. This of course is quite important because different statistical methods are sensitive to different kinds of departures from normality. On the other hand it takes quite a bit of experience to decide whether a certain graph shows a sufficient closeness to the normal distribution, or whether the departure is too severe. This is especially difficult for people who are not trained as statisticians such as scientists in other fields. Formal hypotheses tests on the other hand do not require such a judgment, the decision is a simple yes or no. This simplicity is also the major drawback of hypotheses tests because it means that they are not very flexible: the answer is a yes/no with a rather artificial cutoff. Moreover, different tests pick up on different kinds of departures from normality and can therefore give different answers.

In this paper we discuss a method that combines these two approaches: it is a graphical display and it also performs a formal hypothesis test. While we were mostly interested in making the assessment of the graphical display less subjective, we will also give the results of a brief simulation study that shows that the performance of this method as a formal hypothesis test is competitive with those tests currently in use.

2 Envelopes

The idea of this method is simple: we begin with the most popular graphical display for assessing normality, the normal probability plot. In this graph one plots the sample quantiles versus the population quantiles of a standard normal distribution. Say we have n observations $X_1, .., X_n$. Then we define the x coordinates $p_i, i = 1.., n$ by

$$p_i = \begin{cases} \Phi^{-1}\left(\frac{i-0.375}{n-0.375}\right) \\ \Phi^{-1}\left(\frac{i-0.5}{n}\right) \end{cases} \quad \text{if} \quad \begin{cases} n \leq 10 \\ n > 10 \end{cases}$$

and the y coordinates $y_i, i = 1, .., n$ by $y_i = X_{[i]}$. Here $\Phi^{-1}$is the inverse of the standard normal distribution function and $X_{[i]}$ is the i^{th} order statistic of the sample. In other words we plot the population quantiles of a standard normal versus the sample quantiles. If the data was in fact generated by a normal distribution this graph should be roughly linear. Different departures from linearity point out different distributions. For example a leptokurtic distribution will appear S-shaped rather than linear, and a platykurtic distribution

will appear as a reverse S shape. For more on normal probability plots see Chambers (1983) and Tukey (1983).

In the next step we add an envelope to the graph. The idea of an envelope was discussed in Ripley (1981) and in Atkinson (1985). Venables and Ripley (1994) discuss an Splus function for computing an envelope based on random sampling from a normal distribution. An actual confidance band for the normal probability plot was discussed in Bickel and Doksum (1977), p.384, which is based on inverting the Kolmogorov-Smirnoff statistic D_n^*. Unfortunately their method as a test has such poor performance that they themselves do not advocate its use. We will take a slightly different route. For each sample quantile we plot upper and lower limits such that the overall coverage probability for all quantiles combined is equal to a prechosen α. In this paper we will always use $\alpha = 0.05$. We then combine the upper and the lower limits into two curves, or an envelope, for easier assessment. Then, if the confidence band contains all the sample quantiles the hypothesis test fails to reject the assumption of normality. Often a straight line through the 25^{th} and the 75^{th} percentiles is added to the plot to aid the assessment of linearity but because of the presence of the envelope this makes the graph appear somewhat crowded.

Clearly, this method combines the graphical and the hypothesis testing approach to assessing normality: The graph still contains all the information usually included in a normal probability plot, and we perform a hypothesis test for normality at the α level of significance.

Some care needs to be taken in finding the envelope. Say we have a sample $X_1, .., X_n$ from an normal distribution with mean μ and standard deviation σ. We denote the standardized order statistic by $X'_{[1]}, .., X'_{[n]}$, that is

$$X'_{[i]} = \frac{X_{[i]} - \overline{X}}{s}$$

where $\overline{X}$ and s are the sample mean and the sample standard deviation of the data, respectively. We want to find lower and upper limits $\{(l_i, u_i)\}_{i=1}^n$ such that

$$P\left(l_i \leq X'_{[i]} \leq u_i;\ i = 1, .., n\right) = 1 - \alpha \tag{1}$$

that is, we want the overall rejection probability to be α. If the data were generated by a standard normal we could use the fact that $\Phi(X_{[k]}) \sim \beta(k, n-k+1)$ for any $k = 1, .., n$, where Φ is the cumulative distribution function of a standard normal random variable and $\beta(k, n-k+1)$ is the

beta distribution with k and $n-k+1$ degrees of freedom. This follows from the probability integral transform, see for example Gibbons (1985, pp. 23). We could therefore choose

$$l_k(\alpha; n) = \Phi^{-1}(\beta^{-1}(\alpha/2; k, n-k+1)) \tag{2}$$

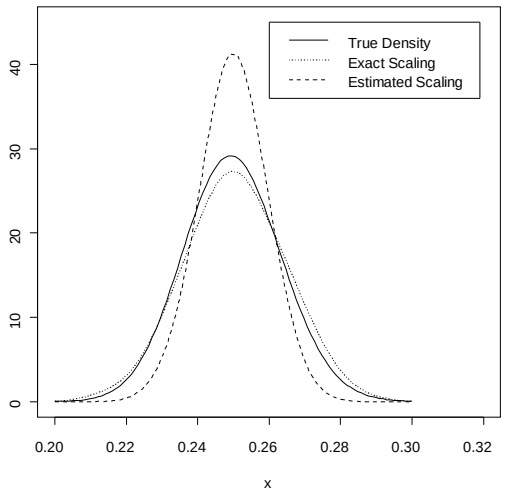
$$u_k(\alpha; n) = \Phi^{-1}(\beta^{-1}(1-\alpha/2; k, n-k+1))$$

where Φ^{-1} is the inverse of the standard normal distribution function and $\beta^{-1}(\cdot; k, n-k+1)$ is the inverse of the beta distribution function with parameters k and $n-k+1$. This would give the correct limits for the individual unstandardized order statistics. In our case, though, we need to standardize the data because we want to test for composite normality. Unfortunately this standardization results in a distribution other than the β even when the sample size is huge. As an example consider the following simulation: We generate 1000 samples of size 1000 from a standard normal distribution. For each sample we find the 250^{th} order statistic. Then we standardize this order statistic, on the one hand with the exact mean and standard deviations (here 0 and 1, respectively) and on the other hand with the sample mean and the sample standard deviation. Finally we transform the observations using the standard normal distribution function. In figure 1 we show density estimates (using a kernel density estimator) for the correctly standardized and transformed data (called Exact Scaling) and for the data where the sample mean and the sample standard deviation was used for the standardization (called Estimated Scaling). We also include the correct density, a $\beta(250, 751)$. Clearly even for this sample size, where we would expect to estimate the true mean and standard deviation quite well, the β is not the correct distribution for the standardized order statistic.

In addition to the problem with the standardization we are also performing n simultaneous hypothesis tests which are clearly not independent. As we shall see the situation is helped somewhat by the fact that the true distribution of the standardized order statistic is independent of the true mean and standard deviation.

Because we are not aware of an explicit formula for the joint distribution of the standardized order statistic we resorted to a Monte Carlo simulation to approximate the overall probability in (1). We use limits of the kind in (2), only the α there is chosen in such a way as to make the overall level of significance α. The algorithm used is as follows: Let the function

Figure 1: True density and estimated densities for the 250th order statistic of a sample of size 1000 from a standard normal distribution. If the scaling is done using the sample mean and the sample standard deviation the probability transform does not result in a beta distribution.



$\psi : [0,1] \to \Re$ be defined by

$$\psi(p) = P\left(l_i(p;n) \leq X'_{[i]} \leq u_i(p;n)\,;\, i = 1,..,n\right)$$

for a fixed sample size n, where $l_i(p;n)$ and $u_i(p;n)$ are defined as in (2). We estimate $\psi(p)$ by generating 50000 standard normal variates and finding the percentage of $X'_{[i]}$ in the interval $(l_i(p;n), u_i(p;n))$ for all $i = 1,..,n$, that is the true percentage of hypothesis tests that would have accepted the null hypothesis of normality. ψ is a strictly increasing function in p, and there is therefore a unique solution of the equation $\psi(p) = \alpha$. For a fixed α we find this solution α_n using a simple bisection algorithm, i.e we start with $p_L = 0, p_H = 1$, find their midpoint m, compute $\psi(m)$ and set $p_L = m$ if $\psi(m) < \alpha$ or $p_H = m$ if $\psi(m) > \alpha$. This is repeated until $p_H - p_L < 0.001$.

In figure 2 we show the estimated α_n such that $\alpha = 0.05$ for $n = 5(1)100(5)150$ together with the least squares fit of a simple power function of the form $\alpha_n = a \cdot n^b$. We found that the function $\alpha_n = 0.7 \cdot n^{-0.72}$ is reasonably good for sample sizes up to 150. This approximation is included here because the computation of α_n is somewhat time consuming. As an example, computing α_{10} on a 266 Mh PC using Splus 4.0 to within 0.001 takes about 1.5 minutes. The graph of α_n is also included in figure 2.Table 1 shows the values of α_n obtained from the Monte Carlo simulation as well as the values given by the approximation $\alpha_n = 0.7 \cdot n^{-0.72}$ for $n = 10(10)100$. As can be seen, the approximation is quite acceptable for this range of sample sizes.

Sample Size n	MC α_n	$0.7 \cdot n^{-0.72}$
10	0.137	0.133
20	0.076	0.081
30	0.058	0.060
40	0.048	0.049
50	0.040	0.042
60	0.037	0.037
70	0.032	0.033
80	0.030	0.030
90	0.028	0.027
100	0.028	0.025

As an example for this method consider figure 3. Here we have the normal probability plot of the infant mortality in Swiss provinces at about 1888. This data set was also considered in Venables and Ripley (1994) where it is used

Figure 2: Adjusted alpha vs. sample size with a least square fitted power function added.

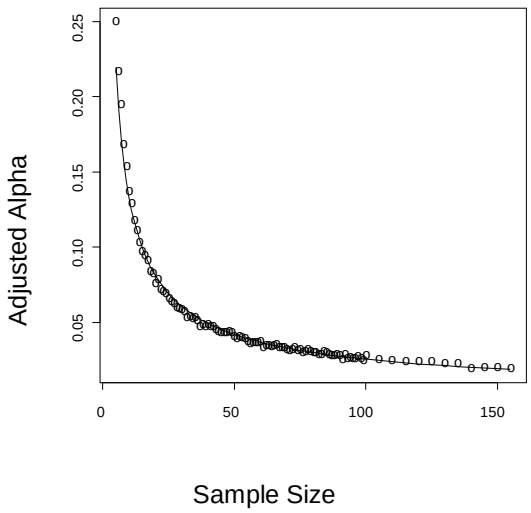
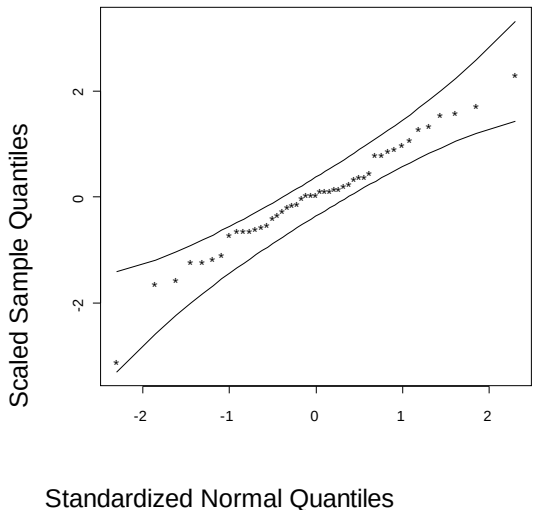


Figure 3: Normal probability plot with envelope of infant mortality in Switzerland around 1888. All the quantiles are within the envelope, and so the assumption of normality seems justified.



to illustrate a normal probability plot with an envelope found by random sampling from a normal distribution. All the sample quantiles are within the envelope and we can therefore treat this data set as approximately normal.

3 Power Considerations

Although we were more interested in extending the graphical display we also performed a number of simulations to study the performance of this method as a formal hypothesis test. This simulation study is not meant to be comprehensive but rather to give a feel for the performance of this method. As comparisons we used the Chi-Square goodness-of-fit test, the Kolmogorov - Smirnov test for normality and the test by D'Agostino and Pearson.

The Chi-Square goodness-of-fit test is probably the most widely known test for normality in the applied sciences. In our simulation we used the test implemented in Splus 4.0. The sample mean and the sample variance were

used in the test, with $m-3$ degrees of freedom where m is the number of bins. The binning is done so as to make the expected cell counts equal. For more details on this test see Moore (1986). The Kolmogorov - Smirnov test for composite normality is the best known test if the data is assumed to be normal. Because the mean and the standard deviation are computed from the data this test is no longer distribution-free but methods for computing correct p-values are known. See Lilliefors (1967) and Stephens (1986) for details. The D'Agostino and Pearson test is based on measures of symmetry and kurtosis. It was chosen because of its performance, which is known to be quite excellent. For a detailed discussion of this test see Zar (1996). There are of course quite a number of other tests available, from general goodness-of-fit tests such as the Cramer-von Mises test or the Anderson-Darling test (Anderson and Darling 1954) to tests especially designed to test for normality such as the Shapiro-Wilk test (Shapiro and Wilk 1965), the skewness test $\sqrt{b_1}$ and the kurtosis test b_2 (Anscombe and Glynn 1983) as well as Filliben's probability plot correlation test r_F (Filliben 1975). The Shapiro-Wilk test and Filliben's r_F test in fact take as their starting point the normal probability plot and exploit the fact that for normally distributed data the graph should be approximately linear. They are, however, standard hypothesis tests without an associated graph. The D'Agostino-Pearson test is a combination of the skewness test $\sqrt{b_1}$ and the kurtosis test b_2. D'Agostino and Pearson (1973) argues that the D'Agostino - Pearson test is, on balance, preferable to the Shapiro - Wilk test.

The first simulation compares the performance of these four tests when the true distribution is a t-distribution, with 5 and 10 degrees of freedom, respectively. These distributions serve as examples for a heavy tailed symmetric distribution and are often more realistic than the normal distribution. Figures 4 and 5 show the estimated power of the four tests by sample size, based on 2500 simulations, for the t distribution with 5 and 10 degrees of freedom, respectively. We use the abbreviations E for the envelope test, AP for the D'Agostino-Pearson, KS for the Kolmogorov-Smirnov test and CHI for the Chi-Square goodness-of-fit test. As we can see, in both cases the envelope test is quite competitive with the D'Agostino-Pearson test, whereas the Kolmogorov-Smirnov test and the Chi-Square test are inferior to both. In the next simulation we generate uniform random variables, which can be used as a model for a thin tailed symmetric distribution. Figure 6 shows the results, again based on 2500 runs.Here for sample sizes over 50 the envelope test is clearly superior to the other two, especially to the D'Agostino-Pearson

Figure 4: The true distribution is a t distribution with 5 d.f. The AP and the E test have nearly identical performance whereas the KS and CHi tests perform poorly.

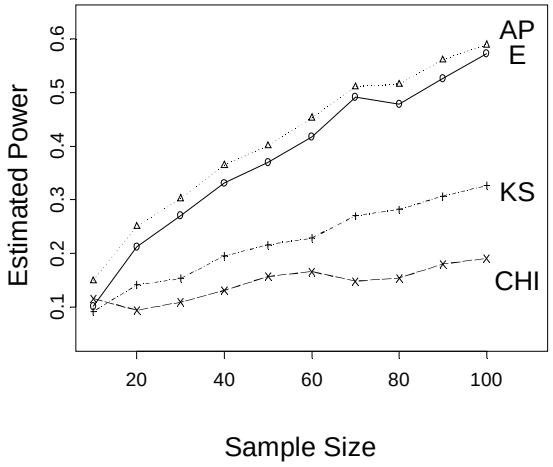


Figure 5: The true distribution is a t distribution with 10 d.f. As with 5 d.f. the AP and the E test have nearly identical performance whereas the KS and CHi tests perform poorly

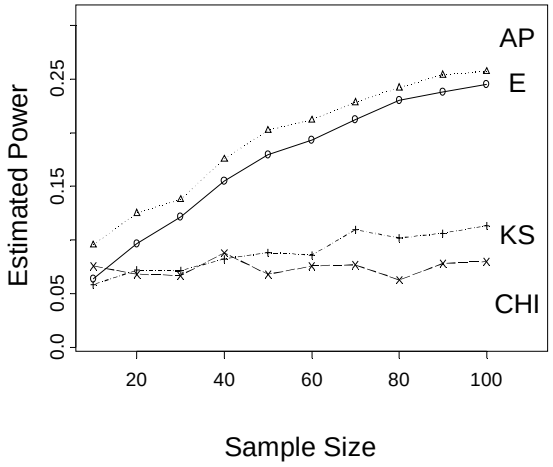


Figure 6: The data has a uniform distribution. The AP test is unable to detect this departure from normality, and the E test is superior to the others.

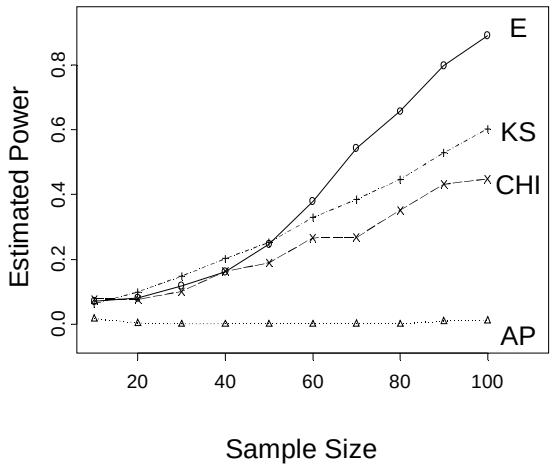
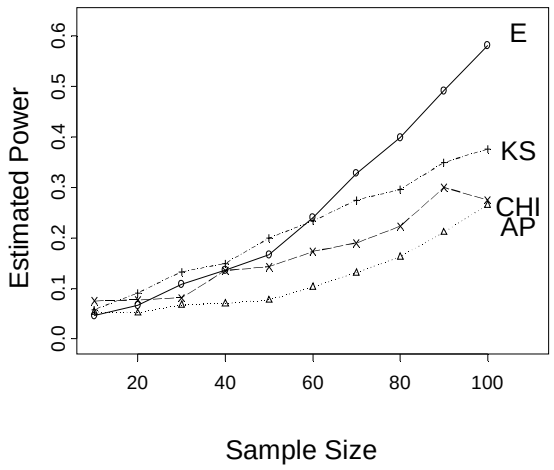
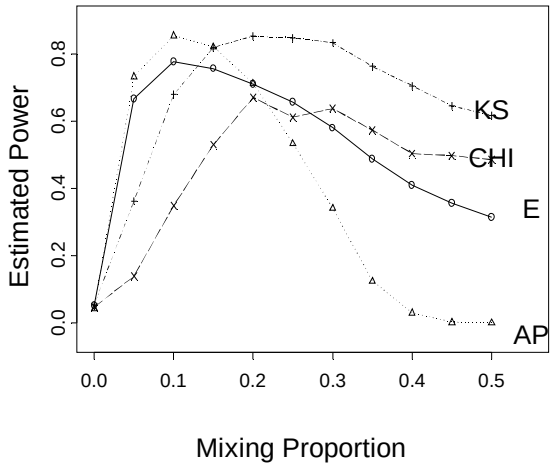


Figure 7: The true distribution is a beta with shape parameters 2 and 4. The E test is superior, with the KS test second. Poor performance of the AP test.



test which fails miserably. The power of the D'Agostino-Pearson test is actually less than 5%, or the α used. This is due to the fact that the uniform distribution is perfectly symmetric with a kurtosis of -1.2. Our next example in figure 7 uses a β distribution with shape parameters 2 and 4, which is an example for a skewed distribution without heavy tails. Here the envelope test is the best performer, at least for sample sizes larger than 50, with the Kolmogorov-Smirnov test second and the D'Agostino-Pearson test last. The bad performance of the D'Agostino-Pearson test is not surprising because the skewness of this distribution is rather slight. Next up is an example of a data set with a potential outlier, the so called 1-wider distribution, see i.e. Hoaglin, Mosteller and Tukey (1983). This is an example where $n-1$ observations come from a standard normal distribution and 1 observation has a normal distribution with mean 0 and standard deviation 9. Here the envelope test and the D'Agostino test show a power of about 70% whereas the Kolmogorov-Smirnov test has a power of about 55% and the Chi-Square goodness-of-fit test has power of about 48%, regardless of the sample size

Figure 8: Power of the tests for a two component normal mixture by mixing proportion. KS is most powerful, with AP rather weak for 50-50 mixtures.



n. Given that the probability that the "outlier" is within the usual range of the standard normal random variable is about 0.25 this means that both the D'Agostino and the envelope test pick up on the presence of the outlier quite well. Finally we consider the example of a mixture of two normal distributions. This type of departure was studied in detail in Mendell, Finch and Thode Jr. (1993). We have 100 observations, with $100p\%$ from a standard normal and $100(1-p)\%$ form a normal with mean three and variance one, where the mixing proportion p ranges from zero to 0.5. As can be seen in figure 8 the performance of all three methods depends strongly on the mixing proportion, with the Kolmogorov-Smirnov test in general being best and the D'Agostino-Pearson test worst.This study of the power of the envelope test is of course far from comprehensive but it is sufficient to show that even as a formal hypothesis test it is quite competitive. In fact, it did not place last in any of the simulations and came close to first several times.

4 Conclusions and Remarks

We described an extension to the usual normal probability plot that is designed to make the judgement of the severity of the departure from the normal distribution easier. All the features of the normal probability plot are preserved, with a confidence band added to the plot. This method can also serve as a formal hypothesis test for normality, and its power for some of the types of departures from normality often seen in practice is quite good, certainly comparable to the D'Agostino-Pearson test and in general superior to the Kolmogorov-Smirnov test and the Chi-Square goodness-of-fit test. The procedure described here uses a fixed level of significance α rather than computing a p value. It would certainly be possible to estimate the p value but it would have to be done via a Monte Carlo simulation and would be computationally quite expensive.

All the simulations and computations were performed using Splus 4.0 on PC's and Splus 3.3 on a DEC alpha workstation. Splus programs for the envelope test, the graph as well as the estimation of α_n are available from us at w_rolke@rumac.upr.clu.edu

References

[1] Anderson, T.W. and Darling, D.A. (1954), "A test of goodness-of-fit", *Jounal of the American Statistical Association*", Vol. 49, p. 765-769.

[2] Anscombe, F.J. and Glynn, W.J. (1983), "Distribution of the kurtosis statistic b_2 for normal statistics", *Biometrika*, Vol. 70, p. 227-234.

[3] Atkinson, A.C. (1985), *Plots, Transformations and Regression*, Oxford, Oxford University Press.

[4] Bickel, P.J. and Doksum, K.A. (1977), *Mathematical Statistics*, Holden-Day.

[5] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, California.

[6] David, H.A. (1981), *Order Statistics*, New York, Wiley and Sons.

[7] D'Agostino, R.B. and Pearson, E.S. (1973), "Tests for Departures from Normality. Empirical Results for the Distribution of b_2 and $\sqrt{b_1}$", *Biometrika*, Vol. 60, p.613-622.

[8] Filliben, J.J. (1975), "The probability plot coefficient test for normality", *Technometrics*, Vol. 17, p.111-117.

[9] Gibbons, J.D. (1985), *Nonparametric Statistical Inference*, New York, Marcel Dekker.

[10] Lilliefors, H.W. (1967), "On the Kolmogorov-Smirnoff test for normality with mean and variance unknown", *Journal of the American Statistical Association*", Vol. 62, p. 399-402.

[11] Mendell, N.R., Finch, S.J. and Thode Jr., H.C. (1993), "Where is the likelihood ratio test powerful for detecting two component normal mixtures?", *Biometrics,* Vol. 49, p. 907-915.

[12] Moore, D. S. (1986). *Goodness-of-Fit Techniques*. R. B. D'Agostino and M. A. Stevens, eds. New York: Marcel Dekker.

[13] Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (ed.) (1983), *Understanding Robust and Exploratory Data Analysis*, New York, Wiley and Sons.

[14] Ripley, B.D. (1981), *Spatial Statistics*, New York, Wiley and Sons.

[15] Shapiro, S.S. and Wilk, M.B. (1965), "An Analysis of Variance Test for Normality", *Biometrika*, Vol. 52, p. 591-611.

[16] Stephens, M. A. (1986). *Goodness-of-Fit Techniques*. D'Agostino, R. B. and Stevens, M. A., eds. New York: Marcel Dekker.

[17] Venables, W.N., Ripley, B.D. (1994), *Modern Applied Statistics with S-Plus*, New York, Springer Verlag.

[18] Zar, J.H. (1996), *Biostatistical Analysis,* New Jersey, Prentice-Hall.