Comparison of Tests for Univariate Normality

Edith Seier
Department of Mathematics
East Tennessee State University
Johnson City, TN 37614
E-mail: seier@access.etsu.edu

Abstract

Tests for univariate normality, some of them not included in previous comparisons, are compared according to their power and simplicity, the validity of their reported p-values, their behavior under rounding, the information they provide and their availability in software. The power of each test was estimated by computer simulation for small, moderate and large sample sizes and a wide range of symmetric, skewed, contaminated and mixed distributions. A new omnibus test based on skewness and kurtosis is discussed.

Key Words: Skewness, kurtosis, L-skewness, scale contaminated, mixed distributions.

1 Introduction

The large number of available tests for normality can be overwhelming for students and practitioners. The tests are based on different characteristics of the normal distribution and the power of these tests differs depending on the nature of the nonnormality. D'Agostino and Stephen (1986) presented a comprehensive description of tests for normality. Gan and Koehler (1990) compared several normality tests in terms of estimated power, based on 1,000 Monte Carlo samples, for a wide variety of distributions. The Shapiro-Wilk test for normality emerged as the best overall test, but was not uniformly most powerful against all alternatives considered.

In the past decade, computers have become faster permitting more acurate Monte Carlo simulations, new extentions of the Shapiro-Wilk test have been developed, new skewness and kurtosis measures have been defined (Hosking, 1990) and new normality tests have been proposed (Chen and Shapiro, 1995; Park, 1999; Zhang, 1999). A new comparison of normality tests will be informative.

The power of the tests is studied using 50,000 Monte Carlo samples of small (20), moderate (50) and large (100) sizes from symmetric, skewed, scale contaminated distributions and balanced mixtures of normal distributions. Other issues discussed are the simplicity of the calculations, the need for special critical values and the behavior of tests when the observations have been rounded. When the test statistic is assumed to have a certain standard distribution we examine how well that distribution approximates the true distribution. The comparison of regression type tests is emphasized as well as the effect of the way in which kurtosis is measured in the simplicity and the power of the skewness-kurtosis tests of normality.

2 Tests for Normality

Normality tests differ in the characteristic of the normal distribution they focus in, such as its skewness and kurtosis values (which can be measured in more than one way), its distribution or characteristic function, and the linear relationship existing between a normally distributed variable and the standard normal z. The tests also differ in the level at which they compare the empirical distribution with the normal distribution (compare and summarize vs. summarize and compare), in the complexity of the test statistic and the nature of its distribution (a standard distribution or an ad-hoc one).

2.1 Tests based on Skewness and Kurtosis

The empirical distribution is summarized through its skewness and kurtosis statistics and compared to the skewness and kurtosis of the normal distribution. Tests based on kurtosis statistics would have very small power against the symmetric alternatives which kurtosis coincides or is close to that of the normal distribution. Pearson's classical measure of kurtosis $\beta_2 = E(x-\mu)^4/(E(x-\mu)^2)^2$ has value 3 not only for the normal distribution but for other distributions as well. Some examples are Tukey (0.135) and Double Gamma $((\sqrt{13}+1)/2)$ as mentioned by Balanda and MacGillivray (1988), and a wide family of symmetric distributions defined by Kale and Sebastian (1996).

In general, $\sqrt{b}_1 = (1/n)\sum((x_i-\bar{x})/s)^3$ and $b_2 = (1/n)\sum((x_i-\bar{x})/s)^4$ are not independent, there are known bounds for the second one in terms of the first for unimodal distributions as specified in Teuscher and Guiard (1995), their correlation is 0 under normality (D'Agostino and Stephen, 1986) but $|\sqrt{b}_1|$ and b_2 are correlated. For $10,000$ samples from the normal distribution the correlation between $|\sqrt{b}_1|$ and b_2 is around 0.65, 0.5 and 0.4 for n=20, 50, 100. Tests defined with $\sqrt{b}_1$ and b_2 can be found in D'Agostino (1971) and Bowman and Shenton (1975) and require special critical values. D'Agostino *et al* (1990) proposed a test statistic that has $\chi^2_{(2)}$ distribution and is defined as $K^2 = [Z(\sqrt{b}_1)]^2 + [Z(b_2)]^2$ where $Z(\sqrt{b}_1)$ and $Z(b_2)$ are normalizing transformations of skewness and kurtosis. The correlation between either $Z(\sqrt{b}_1)$ or $|Z(\sqrt{b}_1)|$ and $Z(b_2)$ is zero.

An alternative measure of skewness and a normality test against skewed alternatives are defined by Hosking(1990). L-moments are defined in terms of expected values of order statistics, $\lambda_2 = (1/2)E(X_{2:2} - X_{1:2})$ and $\lambda_3 = (1/3)E(X_{3:3} - 2X_{2:3} + X_{3:3})$ where $X_{i:n}$ denotes the i^{th} order statistic in a sample of size n. L-skewness is defined as $\tau_3 = \lambda_3/\lambda_2$. The test statistic $[(0.1866n^{-1} + 0.8n^{-2})^{-1/2}]\hat{\tau}_3$ has an approximately normal distribution . The power of this test is unkonwn and will be examined here.

An alternative measure of kurtosis (G-kurtosis) based on Geary's (1935) test for normality is defined by Bonett and Seier (2002) as $\omega = 13.29(ln(\sigma) - ln(\tau))$. G-kurtosis respects van Zwet's kurtosis ordering, gives more importance to the central peak of the distribution than to the tails as opposed to β_2, is equal to 3 for the normal distribuiton, and is less correlated with skewness than β_2. The correlation of $|\sqrt{b_1}|$ and $\hat{\omega}$ for 10,000 samples from the normal distribution was 0.4, 0.29, and 0.22 for n=20, 50, 100, the correlation under normality of $\hat{\omega}$ with $\sqrt{b_1}$ is practically zero. To test for G-kurtosis=3 the statistic $z_\omega = (n+2)^{1/2}(\hat{\omega} - 3)/3.54$ is used, where $\hat{\omega} = 13.29(ln(\hat{\sigma}) - ln(\hat{\tau})), \hat{\tau} = \sum |x_i - \bar{x}|/n$, and $\hat{\sigma} = \sqrt{\sum (x_i - \bar{x})^2/n}$. Two omnibus tests, G^2_ω and G^{2*}_ω, can be defined using z_ω. The first statistic, $G^2_\omega = [Z(\sqrt{b_1})]^2 + [z_\omega]^2$, uses the rather complex normalizing transformation of skewness. The second statistic, noteworthy for its computational simplicity, is defined as $G^{2*}_\omega = [a\sqrt{b_1}]^2 + [z_\omega]^2$ where $a = n/[(n-2)\sqrt{6/(n+1)}]$. Some new measures of kurtosis E[g(z)], where g is a special type of function of the standardized variable z, with properties similar to those of G-kurtosis are defined by Seier and Bonett (2002). Tests similar to G^2_ω and G^{2*}_ω using the new kurtosis statistics instead of ω were considered in the simulations. The behavior of their power was similar to that of the tests based on G-kurtosis.

2.2 EDF tests

The two more popular EDF tests are the one defined by Kolmogorov (1933), which summarizes the comparison of the EDF with the CDF of the normal distribution through the maximum difference, and the test by Anderson and Darling (1954), which involves a combination of all the differences.

2.3 Regression and Correlation Tests

The regression and correlation tests are based on the fact that a variable $y \sim N(\mu, \sigma^2)$ can be expressed as $y = \mu + \sigma x$ where $x \sim N(0,1)$ and are naturally associated to probability plots. Regression tests focus on the slope of the line when the order statistics of the sample are confronted with their expected value under normality and correlation tests focus on the strength of the relationship. The most well known of the regresion tests is the one defined by Shapiro and Wilk (1965), originally restricted for $n \leq 50$. Royston (1992) gives an extention for $n \leq 2000$ and provides a normalizing transformation for the test statistic. Rahman and Govidarajulu (1997) extend the Shapiro-Wilk test to sample sizes $n \leq 5000$ but special critical values are required. Labrecque (1977) defines a family of regression tests (F_1,F_2,F_3) in which models with additional terms are considered as alternatives to $y = \mu + \sigma x$, the test statistic depending on the alternative model considered. The calculation of the test statistic requires special constants that are provided for $n = 4, ..., 63$ and needs special critical values, determined by simulation, which depend on n and the type of alternative model considered. For small samples (n=12 and n=30) F_1 has higher power than the Shapiro-Wilk test against symmetric distributions of high kurtosis, but not for skewed distributions. F_3 has higher power than the S-W test against the Uniform distribution but its power is quite low for highly skewed distributions. Chen and Shapiro (1995) defined a test of simple calculation, QH^*, that requires special critical values which depend on n. The test compares the spacings between the order statistics with the spacings of their expected values under normality. A test of normality Q, also based in the spacings, is defined by Zhang (1999) as the ratio of two linear combinations of the order statisics; the critical values depend on n and are calculated using the Cornish-Fisher expansion.

D'Agostino (1971, 1972) defined a correlation test, Y, which according to Krouse (1994) can be more powerful than others when the alternative is a scale contaminated normal. The correlation test defined by Filliben(1975) is known to have power above average for skewed and long tailed symmetric distributions and power below average for symmetric short tailed distributions.

2.4 Other tests of Normality

Park (1999) defined a test against skewed distributions only based on the sample entropy of order statistics that has higher power than the Shapiro-Wilk test when the alternative is $\chi^2_{(q)}$ with low values of q. Its application requires special critical values that have been reported only for n=10, 20, 30, 40, 50, 100. Tiku (1974) defined a test based on the comparison of the standard deviation of the trimmed data with the standard deviation of the whole sample. The trimming is done differently if the alternative distribution is skewed or symmetric. Tests based on the empirical characteristic function have been defined by Murota and Takeuchi (1981), Hall and Welsh (1983) and Epps and Pulley (1983); all require special critical values. The first test is restricted to symmetric alternatives, the second to long-tailed alternatives and the last one is less powerful than the Shapiro-Wilk test for the ten distributions reported. A test based on U-statistics was defined by Oja (1983)for which critical values determined by simulation are not included. These tests are not considered in the simulations.

2.5 Empirical Type I Error Rates

The empirical alphas, based on $100,000$ Monte Carlo samples, for nominal α=0.01, 0.05, 0.10 appear in Table 1. The tests considered are D (Kolmogorov, 1933), A^{2*} (Anderson and Darling, 1952), Y (D'Agostino 1972), W (Shapiro and Wilk, 1965), z (Royston, 1992), QH^* (Chen and Shapiro, 1995), Q (Zhang, 1999), K^2 (D'Agostino, 1990) , G^2_ω and G^{2*}_ω. The last six were not included in Gan and Koehler (1990). The K^2 and Q tests are liberal at $\alpha = 0.05$ and $\alpha = 0.01$. The G^2_ω and G^{2*}_ω tests are liberal at $\alpha = 0.01$.

3 Power against symmetric alternatives

Table 2 summarizes the empirical powers at $\alpha = 0.05$ based on $50,000$ Monte Carlo samples of size n=20, 50, and 100 for most of the symmetric distributions considered in Gan and Koehler (1990) and some symmetric Lambda distributions.

Overall A^{2*} (Anderson-Darling) is more powerful than D (Kolmogorov-Smirnov) probably due to the more detailed comparison it does. Both tend to have in general lower power than the bes of the other two types of tests considered. Regression tests that use expected values of order statistics have higher power than skewness-kurtosis tests for symmetric bimodal distributions. For the symmetric short tailed distributions the relative superiority of either type depends on the sample size. For the distributions that have slightly or definitely higher kurtosis than the normal, the skewness-kurtosis based tests are more powerful than the other types of test.

Table 1. Empirical Alpha of Tests for Normality

Test	n	0.1	0.05	0.01
Empirical Distribution Function Tests				
D	20	0.1020	0.0493	0.0108
	50	0.1020	0.0489	0.0113
	100	0.1020	0.0505	0.0115
A^{2*}	20	0.1020	0.0503	0.0100
	50	0.0996	0.0493	0.0099
	100	0.1000	0.0505	0.0102
Regression and Correlation Tests				
Y	20	0.0992	0.0520	0.0113
	50	0.1000	0.0493	0.0093
	100	0.1020	0.0512	0.0102
W	20	0.1010	0.0511	0.0105
	50	0.0962	0.0432	0.0076
	100	N/A	N/A	N/A
z	20	0.1020	0.0499	0.0097
	50	0.0991	0.0500	0.0104
	100	0.0980	0.0497	0.0105
QH^*	20	0.1020	0.0500	0.0101
	50	0.0988	0.0498	0.0097
	100	0.1000	0.0497	0.0102
Q	20	0.101	0.0533	0.0134
	50	0.102	0.0554	0.0139
	100	0.104	0.0560	0.0143
Omnibus Skewness and Kurtosis Tests				
K^2	20	0.0948	0.0576	0.0206
	50	0.0982	0.0575	0.0196
	100	0.0995	0.0568	0.0177
G^2_ω	20	0.0854	0.0454	0.0160
	50	0.0920	0.0481	0.0139
	100	0.0956	0.0493	0.0126
G^{2*}_ω	20	0.0838	0.0463	0.0177
	50	0.0914	0.0496	0.0159
	100	0.0956	0.0504	0.0141

The pattern of the power of the correlation test Y is in general more similar to the skewness-kurtosis tests than to the regression tests. The extension z (Royston, 1992) of the Shapiro-Wilk test (W) has lower power for n=50 than the original test when the distribution is bimodal or short tailed but higher power in the case of distributions with kurtosis slightly higher than 3. For the short-tailed unimodal distribution G^2_ω and G^{2*}_ω have low power. QH^*, W and K^2 have the highest power depending if the sample size is small, moderate or large. For distributions with slightly higher kurtosis than the normal the skewness-kurtosis tests are more powerful. The detection of distributions with high kurtosis is important because high kurtosis affects the inference about spread. The power of the Q test decreases when n increases for the $Tukey(10)$ distribution.

Table 2. Power Against Symmetric Non-Normal Distributions

Distribution	n	D	A^{2*}	Y	W	z	QH^*	Q	K^2	G^2_ω	G^{2*}_ω
Bimodal distributions											
Arc sine	20	0.32	0.62	0.03	0.73	0.73	0.75	0.73	0.52	0.29	0.29
	50	0.80	0.99	0.12	1.00	1.00	1.00	1.00	0.99	0.90	0.88
	100	0.99	1.00	0.37	N/A	1.00	1.00	1.00	1.00	1.00	1.00
SB(0,.5)	20	0.19	0.37	0.07	0.45	0.44	0.48	0.46	0.33	0.16	0.16
	50	0.54	0.90	0.39	0.99	0.97	0.98	0.96	0.97	0.73	0.72
	100	0.92	1.00	0.85	N/A	1.00	1.00	1.00	1.00	0.99	0.99
Tukey(1.5)	20	0.11	0.21	0.09	0.26	0.25	0.28	0.23	0.19	0.08	0.08
	50	0.31	0.68	0.54	0.92	0.84	0.88	0.82	0.87	0.48	0.48
	100	0.69	0.98	0.95	N/A	1.00	1.00	1.00	1.00	0.90	0.90
Short tailed distributions											
Uniform	20	0.10	0.17	0.10	0.20	0.20	0.22	0.18	0.15	0.07	0.07
	50	0.26	0.58	0.56	0.86	0.75	0.80	0.71	0.80	0.40	0.40
	100	0.59	0.95	0.96	N/A	1.00	1.00	1.00	1.00	0.84	0.84
Lambda(0.5,0.6)	20	0.06	0.09	0.09	0.09	0.09	0.10	0.08	0.07	0.03	0.03
	50	0.13	0.26	0.46	0.48	0.34	0.41	0.27	0.47	0.18	0.18
	100	0.29	0.62	0.90	N/A	0.83	0.89	0.76	0.91	0.51	0.51
SB(.707)	20	0.08	0.13	0.10	0.15	0.15	0.16	0.12	0.11	0.05	0.05
	50	0.21	0.45	0.55	0.72	0.58	0.65	0.48	0.69	0.32	0.31
	100	0.48	0.87	0.95	N/A	0.97	0.99	0.95	0.99	0.75	0.75
Tukey(0.7)	20	0.07	0.11	0.10	0.12	0.11	0.13	0.10	0.09	0.04	0.04
	50	0.16	0.35	0.52	0.63	0.48	0.55	0.40	0.60	0.24	0.24
	100	0.37	0.77	0.94	N/A	0.94	0.96	0.92	0.97	0.64	0.64
Tukey(3)	20	0.05	0.06	0.08	0.07	0.06	0.07	0.06	0.04	0.02	0.02
	50	0.08	0.17	0.39	0.40	0.27	0.33	0.26	0.34	0.09	0.09
	100	0.15	0.44	0.81	N/A	0.77	0.84	0.81	0.81	0.29	0.29
Beta(2,2)	20	0.05	0.06	0.08	0.06	0.05	0.06	0.05	0.04	0.02	0.02
	50	0.08	0.13	0.30	0.24	0.15	0.19	0.12	0.24	0.09	0.09
	100	0.15	0.32	0.70	N/A	0.45	0.55	0.36	0.65	0.27	0.27
Distributions with kurtosis slightly higher than the normal											
SU(0,3)	20	0.06	0.07	0.08	0.08	0.08	0.08	0.08	0.10	0.09	0.09
	50	0.07	0.08	0.11	0.07	0.11	0.10	0.11	0.13	0.12	0.13
	100	0.08	0.10	0.15	N/A	0.14	0.12	0.14	0.17	0.16	0.17
t(10)	20	0.07	0.09	0.10	0.10	0.10	0.10	0.11	0.13	0.11	0.12
	50	0.09	0.12	0.16	0.11	0.15	0.14	0.16	0.19	0.18	0.19
	100	0.11	0.16	0.26	N/A	0.23	0.20	0.22	0.27	0.26	0.26
Logistic	20	0.08	0.11	0.12	0.12	0.12	0.11	0.13	0.15	0.14	0.14
	50	0.11	0.16	0.21	0.13	0.20	0.18	0.19	0.24	0.24	0.24
	100	0.16	0.24	0.37	N/A	0.31	0.27	0.26	0.35	0.36	0.36
SU(0,2)	20	0.09	0.11	0.12	0.13	0.13	0.12	0.13	0.16	0.15	0.15
	50	0.12	0.17	0.23	0.15	0.22	0.20	0.21	0.26	0.26	0.26
	100	0.17	0.26	0.40	N/A	0.34	0.30	0.30	0.38	0.39	0.39

Table 2. *(Continued)* Power Against Symmetric Non-Normal Distributions

Distribution	n	D	A^{2*}	Y	W	z	QH^*	Q	K^2	G_ω	G_ω
Distributions with high kurtosis											
Tukey(10)	20	0.90	0.91	0.82	0.81	0.81	0.79	0.50	0.51	0.79	0.79
	50	1.00	1.00	1.00	0.99	1.00	1.00	0.38	0.70	0.99	0.99
	100	1.00	1.00	1.00	N/A	1.00	1.00	0.24	0.91	1.00	1.00
Laplace	20	0.22	0.27	0.28	0.26	0.26	0.25	0.26	0.30	0.33	0.34
	50	0.43	0.55	0.60	0.40	0.52	0.48	0.40	0.51	0.65	0.65
	100	0.71	0.83	0.87	N/A	0.80	0.76	0.53	0.74	0.90	0.90
Lambda(-.33,-.14)	20	0.18	0.24	0.26	0.25	0.25	0.24	0.26	0.30	0.30	0.30
	50	0.33	0.45	0.54	0.38	0.49	0.46	0.44	0.52	0.56	0.56
	100	0.54	0.70	0.81	N/A	0.74	0.71	0.61	0.75	0.80	0.80
Lambda(-.55,-.2)	20	0.24	0.31	0.33	0.32	0.32	0.31	0.32	0.37	0.37	0.38
	50	0.45	0.58	0.66	0.50	0.61	0.58	0.55	0.63	0.68	0.69
	100	0.70	0.83	0.90	N/A	0.85	0.83	0.73	0.85	0.89	0.89
SU(0,1)	20	0.35	0.43	0.45	0.43	0.43	0.42	0.42	0.47	0.49	0.49
	50	0.65	0.76	0.81	0.68	0.77	0.74	0.67	0.76	0.83	0.83
	100	0.89	0.95	0.98	N/A	0.96	0.94	0.85	0.94	0.97	0.97
SU(0,0.9)	20	0.43	0.51	0.52	0.50	0.50	0.49	0.49	0.53	0.57	0.55
	50	0.76	0.85	0.89	0.78	0.85	0.83	0.75	0.84	0.90	0.88
	100	0.95	0.98	0.99	N/A	0.98	0.98	0.91	0.97	0.99	0.99
t(4)	20	0.17	0.23	0.25	0.25	0.24	0.24	0.26	0.29	0.29	0.29
	50	0.31	0.42	0.51	0.37	0.47	0.44	0.44	0.51	0.53	0.53
	100	0.49	0.65	0.78	N/A	0.71	0.68	0.62	0.73	0.76	0.76
t(2)	20	0.45	0.53	0.55	0.53	0.53	0.52	0.52	0.57	0.59	0.59
	50	0.78	0.86	0.90	0.81	0.87	0.85	0.80	0.86	0.90	0.90
	100	0.96	0.99	0.99	N/A	0.99	0.98	0.94	0.98	0.99	0.99
t(1)	20	0.85	0.89	0.88	0.87	0.87	0.86	0.83	0.86	0.90	0.90
	50	0.99	1.00	1.00	0.99	1.00	1.00	0.98	0.99	1.00	1.00

D'Agostino and Stephen (1986) mention that the K^2 and Y tests are good against symmetric alternatives with kurtosis higher than 3. In Table 2 it can be observed that if the distribution has kurtosis slightly or moderately higher than the normal distribution then K^2 is more powerful, but if the kurtosis is very large then G^2_ω and G^{2*}_ω have higher power.

Uthoff (1973) proved that the normality test defined by Geary (1936) is similar to the most powerful test against the Laplace distribution. Tiku (1974) reports the power (for α=0.05, 0.10) of the test T based on symmetric trimming for some distributions (t-student, SB, SU, Tukey, Uniform, Logistic). In all the reported cases the power of T is smaller than the power of G^{2*}_ω. There is very little difference in power between G^2_ω and G^{2*}_ω. This means that when using G-kurtosis it is possible to define more simple omnibus tests, in the sense of not needing complicated normalizing transformations neither for kurtosis nor for skewness, without sacrificing power. Spiegelhalter (1983) showed that the standard measure of kurtosis is an asymptocally locally optimal test statistic for normality against the Student t family. Comparing the power for K^2,G^2_ω and G^{2*}_ω in Table 2 and from results for the kurtosis statistics defined in Seier and Bonett (2002), we notice that tests based on other kurtosis statistics also have a very good performance against the Student t family.

4 Power against Scale Contaminated and Mixtures of Normal Distributions

Table 3. Empirical Power against Scale Contaminated Normal Distributions

p	σ	n	D	A^{2*}	Y	W	z	QH^*	Q	K^2	G^2_ω	G^{2*}_ω
0.05	0.14	20	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.07
		50	0.07	0.07	0.06	0.04	0.06	0.06	0.06	0.07	0.08	0.08
		100	0.09	0.08	0.07	N/A	0.07	0.06	0.06	0.07	0.10	0.10
0.1		20	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.09	0.09	0.10
		50	0.13	0.12	0.09	0.06	0.09	0.08	0.06	0.09	0.15	0.15
		100	0.21	0.18	0.14	N/A	0.12	0.10	0.07	0.10	0.23	0.23
0.2		20	0.18	0.17	0.13	0.13	0.13	0.12	0.10	0.13	0.18	0.18
		50	0.37	0.35	0.26	0.15	0.23	0.20	0.10	0.16	0.38	0.37
		100	0.65	0.63	0.45	N/A	0.41	0.37	0.10	0.20	0.65	0.65
0.05	3	20	0.13	0.17	0.19	0.20	0.19	0.19	0.21	0.23	0.21	0.21
		50	0.20	0.29	0.38	0.32	0.38	0.37	0.41	0.43	0.40	0.40
		100	0.28	0.42	0.59	N/A	0.59	0.57	0.61	0.63	0.58	0.58
	5	20	0.28	0.34	0.36	0.36	0.36	0.36	0.38	0.40	0.38	0.38
		50	0.49	0.59	0.66	0.62	0.66	0.65	0.68	0.69	0.67	0.67
		100	0.69	0.81	0.88	N/A	0.88	0.87	0.89	0.89	0.87	0.87
	7	20	0.38	0.43	0.45	0.45	0.45	0.45	0.46	0.48	0.46	0.46
		50	0.65	0.72	0.76	0.74	0.77	0.76	0.78	0.78	0.77	0.77
		100	0.86	0.91	0.94	N/A	0.94	0.94	0.95	0.95	0.94	0.94
0.1	3	20	0.19	0.26	0.29	0.29	0.29	0.28	0.30	0.34	0.32	0.32
		50	0.32	0.46	0.58	0.47	0.56	0.54	0.56	0.61	0.59	0.59
		100	0.49	0.68	0.82	N/A	0.80	0.78	0.77	0.83	0.80	0.80
	5	20	0.43	0.52	0.55	0.55	0.55	0.54	0.55	0.58	0.58	0.58
		50	0.73	0.83	0.88	0.83	0.87	0.86	0.86	0.80	0.88	0.88
		100	0.92	0.97	0.99	N/A	0.98	0.98	0.98	0.99	0.98	0.98
	7	20	0.59	0.66	0.68	0.68	0.68	0.67	0.62	0.70	0.69	0.70
		50	0.88	0.92	0.95	0.93	0.94	0.94	0.94	0.95	0.95	0.95
		100	0.99	0.99	1.00	N/A	1.00	1.00	1.00	1.00	1.00	1.00
0.2	3	20	0.26	0.35	0.38	0.38	0.37	0.36	0.38	0.43	0.43	0.43
		50	0.48	0.65	0.75	0.59	0.71	0.68	0.62	0.73	0.75	0.75
		100	0.74	0.89	0.95	N/A	0.92	0.91	0.79	0.93	0.93	0.93
	5	20	0.61	0.71	0.73	0.71	0.71	0.70	0.66	0.71	0.75	0.76
		50	0.92	0.97	0.98	0.95	0.97	0.97	0.89	0.97	0.98	0.98
		100	1.00	1.00	1.00	N/A	1.00	1.00	0.97	1.00	1.00	1.00
	7	20	0.79	0.86	0.87	0.85	0.85	0.84	0.79	0.82	0.87	0.88
		50	0.98	0.99	1.00	0.99	1.00	1.00	0.96	0.99	1.00	1.00

Table 3 displays the power of the tests to detect contamination with probability p of a normal distribution N(0,1) with another normal distribution $N(0,\sigma^2)$ when σ=0.14, 3, 5, 7 and p=0.05, 0.1, 0.2.

When the variance of the contaminating distribution is smaller than the variance of the main distribution, the new tests based on G-kurtosis, G^2_ω and G^{2*}_ω have higher power followed by the EDF tests. Murota and Takeuchi (1981) point that their test is also better than the one based in b_2 when the contaminating variance is smaller. When the variance of the contaminating distribution is larger, the skewness and kurtosis tests in general tend to have higher power with superiority of one or the other depending on the probability of contamination.

Table 4. Power against Balanced Mixtures of Normal distributions

m	n	D	A^{2*}	Y	W	z	QH^*	Q	K^2	G^2_ω	G^{2*}_ω
10	20	1.00	1.00	0.18	1.00	1.00	1.00	0.84	0.90	0.98	0.98
	50	1.00	1.00	0.47	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	100	1.00	1.00	0.98	N/A	1.00	1.00	1.00	1.00	1.00	1.00
5	20	0.65	0.81	0.02	0.76	0.75	0.78	0.31	0.59	0.62	0.61
	50	0.99	1.00	0.01	1.00	1.00	1.00	0.63	0.99	1.00	1.00
	100	1.00	1.00	0.01	N/A	1.00	1.00	0.88	1.00	1.00	1.00
4	20	0.35	0.46	0.04	0.41	0.41	0.44	0.18	0.31	0.29	0.28
	50	0.83	0.94	0.11	0.93	0.90	0.92	0.40	0.91	0.92	0.92
	100	1.00	1.00	0.42	N/A	1.00	1.00	0.67	1.00	1.00	1.00
3	20	0.12	0.15	0.07	0.14	0.13	0.15	0.09	0.10	0.07	0.07
	50	0.31	0.44	0.27	0.45	0.38	0.43	0.18	0.46	0.39	0.39
	100	0.65	0.82	0.70	N/A	0.75	0.80	0.33	0.84	0.83	0.83
2	20	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.02	0.02
	50	0.07	0.08	0.10	0.09	0.07	0.08	0.06	0.08	0.05	0.05
	100	0.10	0.13	0.22	N/A	0.11	0.14	0.17	0.09	0.11	0.11
1	20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04
	50	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.05	0.04	0.04
	100	0.05	0.05	0.05	N/A	0.05	0.05	0.05	0.05	0.04	0.04
Average	20	0.37	0.42	0.07	0.40	0.40	0.41	0.25	0.33	0.34	0.34
	50	0.54	0.58	0.17	0.59	0.56	0.58	0.38	0.58	0.57	0.56
	100	0.63	0.67	0.40	N/A	0.65	0.67	0.50	0.67	0.66	0.66

Bajgier and Aggarwal (1991) concluded that a test based on b_2 is more powerful than several other tests including Anderson-Darling and Shapiro-Wilk to detect balanced mixtures of normal distributions. Their study included mixtures of 2, 3, and 4 distributions, distances between the means of 0.5 to 5 standard deviations and sample sizes of 30, 60 and 90. Gan and Koehler (1990) considered the distribution $LoConN(0.5, m)$ for m=1, 2, 3, 4, 5, 10, where a N(0,1) distribution gets contaminated with probability 0.5 with a N(m,1) distribution. Table 4 shows that when m=1 the mixture is almost undetectable with any test. The behavior of Y is not consistent along the values of m. The Q test has relatively low power against mixtures. The average power of the other tests is similar for medium and large samples. For small samples, A^{2*} and QH^* tend to have the highest power while Y has very low power.

5 Power against skewed alternatives

The skewed distributions considered by Gan and Koehler (1990) appear in Table 5, except the Triangle and the Truncated Normal distributions. The power of two normality tests against skewed alternatives, $Z(\sqrt{b_1})$ (D'Agostino et. al. 1990) and the L-skewness test (Hosking, 1990) appear in the first two columns, the second is more powerful than the first one against skewed distributions with low kurtosis. These tests should not be understood as symmetry tests in a general sense, the power for symmetric nonnormal distributions can be quite different from α.

The tests based on skewness and kurtosis tend to have lower power than the Anderson-Darling test for skewed distributions when the kurtosis is low. The three skewness-kurtosis tests have similar average power. The Shapiro-Wilk and the Chen-Shapiro tests have higher average power. The Q test has low power against location contaminated normals.

6 P-values

When performing a test of hypothesis, the p-value is found using the distribution assumed for the test statistic under the null hypothesis of normality. The accuracy of the p-value depends on how close the assumed distribution is to the true distribution of the test statistic under the null hypothesis. An alternative, computer intensive, way of finding p-values for tests of normality is to generate a large number of samples from the normal distribution and find in what proportion of them the value of the statistic is larger (or smaller for lower tail cases) than the value of the statistic for the data. The theoretical p-values obtained with the distribution will be compared to the empirical p-values obtained through simulations for the data set in Chen and Shapiro (1995) consisting of the SiO_2 content of plutonic rocks of 21 samples of plutonic rocks of the Alaska-Aleutian Range batholith. Using the Kolmogorov-Smirnov (D) test the normality hypothesis had not been rejected. Using the Anderson-Darling (A^{2*}) test the normality hypothesis is kind of a borderline case and had not been rejected at the $\alpha = 0.05$ level. All the other tests considered here reject the hypothesis of normality. The empirical p-values reported, rounded to 4 decimal places, were obtained with $100,000$ samples from the normal distribution.

	p-value			**p-value**	
test	theoretical	empirical	test	theoretical	empirical
QH^*	0.0319	0.0326	z	0.0294	0.0295
A^{2*}	0.0800	0.0794	K^2	0.0023	0.0087
W	0.0254	0.0279	G^2_ω	0.0117	0.0175
D	> 0.15	0.2212	G^{2*}_ω	0.0075	0.0153

Table 5. Power against Skewed distributions

Dist	n	$Z(\sqrt{b_1})$	τ_3	A^{2*}	Y	W	z	QH^*	Q	K^2	G_ω^{2*}
Unimodal, low kurtosis											
Beta(2,1)	20	0.12	0.26	0.26	0.06	0.31	0.31	0.32	0.36	0.11	0.08
	50	0.32	0.62	0.72	0.07	0.88	0.84	0.86	0.95	0.34	0.42
	100	0.71	0.92	0.98	0.07	N/A	1.00	1.00	1.00	0.92	0.92
Beta(3,2)	20	0.03	0.06	0.07	0.06	0.07	0.07	0.08	0.07	0.04	0.03
	50	0.05	0.13	0.17	0.14	0.26	0.20	0.23	0.22	0.14	0.08
	100	0.11	0.27	0.40	0.28	N/A	0.53	0.60	0.63	0.40	0.27
Weibull(4)	20	0.03	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.03
	50	0.03	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.03
	100	0.03	0.06	0.06	0.06	N/A	0.05	0.06	0.05	0.05	0.03
Weibull(3.6)	20	0.03	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.03
	50	0.02	0.04	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.03
	100	0.02	0.04	0.05	0.07	N/A	0.04	0.05	0.05	0.05	0.03
SB(.533,.5)	20	0.20	0.50	0.65	0.11	0.73	0.72	0.74	0.82	0.23	0.22
	50	0.48	0.87	0.99	0.14	1.00	1.00	1.00	1.00	0.83	0.88
	100	0.85	0.99	1.00	0.18	N/A	1.00	1.00	1.00	1.00	1.00
SB(1,1)	20	0.20	0.32	0.27	0.09	0.31	0.30	0.31	0.28	0.16	0.12
	50	0.55	0.75	0.69	0.11	0.81	0.80	0.81	0.82	0.38	0.48
	100	0.91	0.97	0.97	0.14	N/A	1.00	1.00	1.00	0.89	0.94
SB(1,2)	20	0.05	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.05	0.04
	50	0.09	0.11	0.09	0.06	0.10	0.10	0.11	0.09	0.08	0.06
	100	0.16	0.21	0.16	0.07	N/A	0.19	0.21	0.15	0.13	0.12
Weibull(2.2)	20	0.10	0.12	0.10	0.07	0.11	0.11	0.11	0.09	0.10	0.07
	50	0.24	0.28	0.20	0.09	0.27	0.27	0.28	0.24	0.19	0.17
	100	0.50	0.54	0.41	0.10	N/A	0.57	0.60	0.56	0.38	0.41
Weibull(2)	20	0.15	0.17	0.13	0.09	0.16	0.15	0.15	0.13	0.13	0.10
	50	0.37	0.42	0.31	0.12	0.41	0.41	0.42	0.37	0.28	0.26
	100	0.70	0.74	0.61	0.14	N/A	0.79	0.81	0.80	0.57	0.62
Half N(0,1)	20	0.33	0.45	0.37	0.16	0.45	0.44	0.45	0.47	0.27	0.22
	50	0.76	0.87	0.84	0.30	0.94	0.93	0.94	0.98	0.61	0.71
	100	0.98	0.99	1.00	0.49	N/A	1.00	1.00	1.00	0.97	0.99
Bimodal, Location contaminated normal											
LoConN(.2,3)	20	0.20	0.28	0.27	0.11	0.27	0.26	0.26	0.15	0.16	0.15
	50	0.50	0.68	0.65	0.17	0.60	0.62	0.62	0.23	0.33	0.37
	100	0.85	0.95	0.94	0.28	N/A	0.92	0.92	0.34	0.71	0.74
LoConN(.2,5)	20	0.58	0.78	0.88	0.47	0.87	0.87	0.87	0.45	0.46	0.48
	50	0.95	1.00	1.00	0.80	1.00	1.00	1.00	0.67	0.94	0.94
LoConN(.2,7)	20	0.74	0.92	0.99	0.77	0.99	0.99	0.99	0.72	0.63	0.72
	50	0.98	1.00	1.00	0.98	1.00	1.00	1.00	0.91	1.00	1.00
LoConN(.1,3)	20	0.27	0.25	0.23	0.17	0.25	0.25	0.24	0.18	0.26	0.23
	50	0.59	0.55	0.51	0.36	0.50	0.56	0.54	0.25	0.52	0.54
	100	0.87	0.83	0.81	0.62	N/A	0.85	0.84	0.32	0.83	0.85

Table 5. *(Continued)* Power Against Skewed Distributions

Dist	n	$Z(\sqrt{b_1})$	τ_3	A^{2*}	Y	W	z	QH^*	Q	K^2	G_ω^{2*}
LoConN(.1,5)	20	0.75	0.69	0.73	0.65	0.77	0.76	0.76	0.52	0.69	0.68
	50	0.99	0.95	0.97	0.97	0.98	0.99	0.98	0.69	0.99	0.99
LoConN(.1,7)	20	0.86	0.85	0.87	0.86	0.88	0.88	0.88	0.74	0.82	0.83
LoConN(.05,3)	20	0.21	0.17	0.15	0.15	0.18	0.18	0.18	0.17	0.22	0.20
	50	0.43	0.31	0.30	0.31	0.32	0.39	0.37	0.27	0.44	0.42
	100	0.68	0.51	0.50	0.53	N/A	0.64	0.61	0.36	0.69	0.67
LoConN(.05,5)	20	0.57	0.46	0.50	0.52	0.55	0.55	0.54	0.48	0.58	0.56
	50	0.88	0.74	0.80	0.87	0.85	0.88	0.87	0.78	0.90	0.89
	100	0.98	0.92	0.96	0.98	N/A	0.99	0.98	0.91	0.99	0.99
LoConN(.05,7)	20	0.66	0.61	0.64	0.65	0.66	0.65	0.65	0.62	0.65	0.65
	50	0.93	0.87	0.92	0.93	0.92	0.93	0.93	0.93	0.93	0.93
	100	1.00	0.98	0.99	1.00	N/A	1.00	1.00	0.99	1.00	1.00
Unimodal, high skweness and/or kurtosis											
SU(1,1)	20	0.71	0.74	0.71	0.61	0.73	0.72	0.72	0.62	0.67	0.65
	50	0.97	0.98	0.98	0.93	0.97	0.98	0.98	0.87	0.96	0.96
SU(1,2)	20	0.23	0.21	0.19	0.17	0.21	0.21	0.20	0.19	0.23	0.21
	50	0.47	0.42	0.38	0.33	0.37	0.44	0.42	0.33	0.45	0.44
	100	0.71	0.66	0.62	0.55	N/A	0.69	0.67	0.47	0.70	0.69
Gumbel	20	0.32	0.34	0.27	0.20	0.31	0.31	0.31	0.25	0.29	0.25
	50	0.70	0.72	0.60	0.40	0.66	0.69	0.69	0.52	0.60	0.60
	100	0.95	0.95	0.89	0.62	N/A	0.94	0.94	0.77	0.90	0.91
Chi-square(4)	20	0.47	0.56	0.47	0.29	0.54	0.53	0.54	0.47	0.40	0.36
	50	0.89	0.94	0.89	0.56	0.95	0.95	0.95	0.95	0.80	0.84
	100	1.00	1.00	1.00	0.83	N/A	1.00	1.00	1.00	0.99	1.00
Exponential	20	0.70	0.83	0.78	0.51	0.84	0.84	0.84	0.86	0.60	0.58
	50	0.99	1.00	1.00	0.88	1.00	1.00	1.00	1.00	0.97	0.99
Chi-square(1)	20	0.89	0.97	0.97	0.81	0.98	0.98	0.99	0.99	0.81	0.83
Lognormal	20	0.87	0.94	0.91	0.76	0.93	0.93	0.93	0.92	0.80	0.79
	50	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00
Weibull(.5)	20	0.98	1.00	1.00	0.96	1.00	1.00	1.00	1.00	0.95	0.96

The tests in the left use special distributions, either derived or obtained by simulation, so it is natural that the empirical p-values are very close to the p-values obtained from the ad-hoc distribution. The tests on the right use either the normal or the Chi-square(2) distribution to find the p-values. G_ω^{2*} does not involve elaborate normalizing transformations which Royston's z test and K^2 require. The empirical alphas in Table 6 have been calculated based on $100,000$ samples from the normal distribution for small and large sample sizes (20 and 100).

7 Normality tests and rounding

Procedures based on order statistics tend to be sensitive to coarse rounding and normality tests are not an exception. Pearson *et al.* (1977) showed that the Shapiro-Wilk test could be seriously

Table 6. Empirical alphas for tests with approximate distributions

α	n	z	K^2	G_ω^{2*}
0.20	20	0.2032	0.1702	0.1734
	100	0.1971	0.1909	0.1904
0.15	20	0.1504	0.1318	0.1262
	100	0.1489	0.1446	0.1423
0.08	20	0.0806	0.0807	0.0684
	100	0.0797	0.0825	0.0770
0.06	20	0.0613	0.0654	0.0540
	100	0.059	0.0648	0.0586
0.04	20	0.0396	0.0485	0.0386
	100	0.0397	0.0472	0.0404
0.02	20	0.0192	0.0311	0.0254
	100	0.0209	0.0281	0.0233
0.005	20	0.0048	0.0137	0.0129
	100	0.0053	0.0114	0.0085

sensitive to rounding when the ratio (grouping interval)/(standard deviation) exceeds 0.1. Royston (1989) proposed a correction for the test in case of ties. Chen and Shapiro (1995) suggest that the same correction could be applied to the QH^* test but no results were included. We performed simulations and found that, in effect, the correction works well and should be used for QH^* if the data have been rounded. The test statistic is $QH^* = (1 - QH)\sqrt{n}$ where $QH = [\sum_i^{n-1}(x_{(i+1)} - x_{(i)})/(m_{(i+1)} - m_{(i)})]/((n-1) * s)$ and the $m_{(i)}$ are the expected values of the order statistics assuming normality. The correction involves a modification to the standard deviation. When the data have been rounded to the nearest integer, the correction simply consists in using $\sqrt{(s^2 - 1/12)}$ instead of s. The tests based on skewness and kurtosis are affected very little by rounding.

8 Normality tests in statistical software

SPSS performs the Shapiro Wilk test only for $n \leq 50$ and the Kolmogorov-Smirnov test for larger sample sizes. The UNIVARIATE procedure in SAS has the option of the Kolmogorov-Smirnov test for $n > 2000$ and the Shapiro-Wilk test for $n \leq 2000$; an improvement in the algorithm has been reported starting with Version 7.1. MINITAB includes the Anderson-Darling, Ryan-Joiner (similar to the Shapiro-Wilk test) and Kolmogorov-Smirnov tests. Skeweness and kurtosis statistics, $\sqrt{b_1}$ and b_2, are calculated by SAS, Minitab, SPSS and STATISTICA, but generally the statistics do not automatically accompany the normality tests. Normality tests based on skewness and kurtosis statistics are not usually provided by commercial software.

DeCarlo (1997) includes an SPSS Macro for testing for Univariate and Multivariate Skewness and Kurtosis. A SAS macro is available also in the SAS library in the Web. Care should be taken with these and other macros in the normalizing transformation $Z(b_2)$ when the kurtosis is very low as in the Beta(0.5,0.5) distribution because of the way computers calculate cubic roots.

9 Summary

Normality tests vary not only in power but also in simplicity in terms of calculation and need for special critical values. The average power for different types of distributions is displayed in Tables 7 and 4.

Table 7. Average Power

n	D	A^{2*}	Y	W	z	QH^*	Q	K^2	G^2_ω	G^{2*}_ω
Symmetric, bimodal (3)										
20	0.21	0.40	0.07	0.48	0.47	0.50	0.47	0.34	0.18	0.20
50	0.55	0.86	0.35	0.97	0.94	0.95	0.93	0.94	0.70	0.69
100	0.87	0.99	0.72	*N/A*	1.00	1.00	1.00	1.00	0.96	0.96
Symmetric, short tailed (6)										
20	0.07	0.10	0.09	0.11	0.11	0.13	0.10	0.08	0.04	0.04
50	0.15	0.32	0.46	0.55	0.43	0.49	0.37	0.52	0.22	0.22
100	0.34	0.66	0.88	*N/A*	0.82	0.87	0.80	0.89	0.55	0.55
Symmetric, kurtosis slightly higher than normal (4)										
20	0.08	0.10	0.10	0.11	0.11	0.10	0.11	0.14	0.12	0.12
50	0.10	0.13	0.18	0.11	0.17	0.15	0.17	0.21	0.20	0.20
100	0.13	0.19	0.29	*N/A*	0.25	0.22	0.23	0.29	0.29	0.29
Symmetric, high kurtosis (9)										
20	0.42	0.47	0.48	0.47	0.47	0.46	0.43	0.47	0.51	0.51
50	0.63	0.71	0.77	0.66	0.73	0.71	0.60	0.70	0.78	0.78
100	0.80	0.88	0.93	*N/A*	0.89	0.88	0.71	0.87	0.92	0.92
Scale Contaminated Normal, smaller σ (3)										
20	0.11	0.10	0.09	0.09	0.09	0.08	0.07	0.09	0.11	0.11
50	0.19	0.18	0.14	0.09	0.13	0.11	0.07	0.11	0.20	0.20
100	0.32	0.30	0.22	*N/A*	0.20	0.18	0.08	0.12	0.33	0.33
Scale contaminated Normal, larger σ (9)										
20	0.41	0.48	0.50	0.50	0.49	0.49	0.48	0.52	0.52	0.52
50	0.63	0.71	0.77	0.72	0.76	0.75	0.77	0.74	0.78	0.78
100	0.77	0.85	0.91	*N/A*	0.90	0.89	0.88	0.91	0.90	0.90
Skewed, unimodal, low kurtosis (10)										
20	0.14	0.20	0.08	0.23	0.22	0.23	0.24	0.12	0.09	0.09
50	0.30	0.41	0.11	0.48	0.46	0.47	0.47	0.29	0.30	0.31
100	0.47	0.56	0.16	N/A	0.62	0.63	0.62	0.54	0.53	0.53
Skewed, Location Contaminated Normal, moderate kurtosis (8)										
20	0.51	0.58	0.48	0.60	0.60	0.60	0.43	0.50	0.49	0.50
50	0.72	0.79	0.71	0.80	0.82	0.81	0.61	0.78	0.78	0.79
100	0.85	0.91	0.82	N/A	0.93	0.93	0.71	0.91	0.91	0.92
Skewed, high skewness and/or kurtosis (9)										
20	0.56	0.66	0.54	0.69	0.69	0.69	0.66	0.59	0.56	0.58
50	0.79	0.86	0.76	0.87	0.88	0.88	0.83	0.85	0.85	0.85
100	0.90	0.94	0.87	N/A	0.95	0.95	0.90	0.95	0.95	0.95

If the purpose of testing for normality is to simply determine if the distribution is normal or not, some regression tests seem to be the best option from the point of view of power for a wide diversity of alternatives; the appropriate correction should be applied when the observations have been rounded. The QH^* test (Chen-Shapiro, 1995) has a more consistent performance than the other regression test, Q (Zhang, 1999), based on normal spacings. If the purpose of the test is to identify symmetric distributions with high kurtosis, the skewness-kurtosis tests have higher power. G_{ω}^{2*} seems appropriate to identify scale contaminated normal distributions in which the standard deviation of the contaminating distribution can be smaller or larger than the standard deviation of the main distribution.

The way in which kurtosis is measured makes a difference not only in the power of the test for different types of distributions, but also in the way in which omnibus tests can be defined. Using kurtosis statistics other than b_2, it is possible to define more simple omnibus tests that do not include complicated transformations of the skewness and kurtosis statistics and still use the Chi-square distribution to find p-values. These tests do not have a good performace detecting distributions with kurtosis lower than the normal but have higher power when the distribution is more peaked than the normal.

Statistical software should incorporate the Chen-Shapiro test. The results of a test for normality should not only report a p-value but they should be accompanied by a careful interpretation of the probability plot and skewness and kurtosis statistics for a complete diagnosis.

References

Anderson, T. W. and Darling, D. A. (1952) A test of goodness-of-fit. *Journal of the American Statitical Association* **49**, 765-769.

Bajgier, S. M. and Agarwal, L. K. (1991) Powers of Goodness-of-Fit Tests in Detecting Balanced Mixed Normal Distributions. *Educational and Psychological Measurement* **51**, 253-269.

Balanda, K. P. and MacGillivray, H. L. (1988) Kurtosis: A Critical Review. *The American Statistician* **42**, 111-119.

Bonett, D. G. and Seier, E. (2002) A test of Kurtosis with High Uniform Power. Computational Statistics and Data Analysis, *conditionally accepted and final version under review.*

Bowman, K. and Shenton, L. R. (1975) Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2 . *Biometrika* **62**, 243-250.

Chen, L. and Shapiro, S. (1995) An Alternative test for normality based on normalized spacings. *Journal of Statistical Computation and Simulation* **53**, 269-287.

D'Agostino, R. B. (1971) An omnibus test of normality for moderate and large size samples. *Biometrika* **58**, 341-348.

D'Agostino, R. B. and Stephen, M. A. (1986) Goodness of Fit Techniques, Marcel Dekker, New York.

D'Agostino, R. B., Belanger, A., and D'Agostino Jr., R. B. (1990) A suggestion for using powerful and informative tests of normality. *The American Statistician* **44**, 316-322.

DeCarlo, L. T. (1997) On the Meaning and Use of Kurtosis. *Psychological Methods* **2**, 292-307.

Epps, T.W. and Pulley, L.B. (1983) A Test of Normality Based on the Empirical Characteristic Function. *Biometrika* 70, 723-726.

Filliben, J.J. (1975) The Probability Plot Correlation Coefficient Test for Normality *Technometrics* **17**, 111-117

Geary, R.C. (1935) The Ratio of the Mean Deviation to the Standard Deviation as a test of normality. *Biometrika* **27**, 310-332.

Gan, F. F. and Koehler, K. J. (1990) Goodness of Fit Tests Based on P-P Probability Plots. *Technometrics* **32**, 289-303.

Hall, P. and Welsh, A.H. (1983) A Test for Normality Based on the Empirical Characteristic Function. *Biometrika* **70**, 485-489.

Hosking, J. R. M. (1990) L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society, Series B* **52**, 105-124.

Kale B. K. and Sebastian, G. (1996) On a Class of Symmetric Non-normal Distributions with a Kurtosis of Three, in *Statistical Theory and Applications: Papers in Honor of Herbert A. David*, Nagarija, H. N., Sen, P. K., and Morrison, D. F. eds. New York: Springer.

Kolmogorov, A. N. (1933) Sulla determinazione empirica di una legge di distribuzione. *Giorna. Ist. Attuari.* **4**, 83-91.

Krouse, Donal P. (1994) The Power of D'Agostino's D Test of Normality against a Normal Mixture Alternative. *Communications in Statistics Theory and Methods* **23**, 45-57.

LaBrecque, J. (1977) Goodness-of-fit Tests based on Non-linearity in Probability Plots. *Technometrics* **19**, 293-306.

Murota, K. and Takeuchi, K. (1981) The Studentized Empirical Characteristic Function and Its Application to Test for the Shape of Distribution. *Biometrika* **68**, 55-65.

Oja, H. (1983) New Tests for Normality. *Biometrika* **70**, 297-299.

Park, S. (1999) A goodness-of fit test for normality based on the sample entropy of order statistics. *Statistics and Probability Letters* **44**, 359-363.

Pearson, E. S., D'Agostino, R. B. and Bowman, L. R. (1977) Tests for departure from normality: Comparison of powers. *Biometrika* **64**, 231-246.

Rahman, M. M. and Govindarajulu, Z. (1997) A Modification of the test of Shapiro and Wilk for Normality. *Journal of Applied Statistics* **24**, 219-235.

Royston, P. (1989) Correcting the Shapiro Wilk W for Ties. *Journal of Computation and Simulation* **31**, 237-249.

Royston, P. (1992) Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing* **2**, 117-119.

Seier, E. and Bonett, D. G. (2002) A Family of Kurtosis Measures. Metrika, *conditionally accepted and final version under review.*

Shapiro, S. and Wilk, M. B. (1965) An analysis of variance test for normality. *Biometrika* **52**, 591-611.

Spiegelhalter, D.J. (1983) Diagnostic tests of distributional shape. *Biometrika* **70**, 401-408.

Teuscher, F. and Guiard, V. (1995) Sharp inequalities between skewness and kurtosis for unimodal distribuitons. *Statistics and Probability Letters* **22**, 257-260.

Tiku, M. L. (1974) A New Statistic for Testing Normality. *Communications in Statistics* **3**, 223-232.

Uthoff, V. A. (1973) The Most Powerful Scale and Location Invariant Test of the Normal versus the Double Exponential. *The Annals of Statistics* **1**, 170-174.

van Zwet, W. R. (1974) Convex Transformations on Random Variables. *Mathematical Centre Tracts*. Mathematisch Centrum. Amsterdan, Holland.

Zhang, P. (1999) Omnibus test of normality using the Q statistic. *Journal of Applied Statistics* **26**, 519-528.