

Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests

Nornadiah Mohd Razali¹

Yap Bee Wah¹

¹Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
E-mail: nornadiah@tmsk.uitm.edu.my, yapbeewah@salam.uitm.edu.my

ABSTRACT

The importance of normal distribution is undeniable since it is an underlying assumption of many statistical procedures such as *t*-tests, linear regression analysis, discriminant analysis and Analysis of Variance (ANOVA). When the normality assumption is violated, interpretation and inferences may not be reliable or valid. The three common procedures in assessing whether a random sample of independent observations of size n come from a population with a normal distribution are: graphical methods (histograms, boxplots, *Q-Q*-plots), numerical methods (skewness and kurtosis indices) and formal normality tests. This paper* compares the power of four formal tests of normality: Shapiro-Wilk (*SW*) test, Kolmogorov-Smirnov (*KS*) test, Lilliefors (*LF*) test and Anderson-Darling (*AD*) test. Power comparisons of these four tests were obtained via Monte Carlo simulation of sample data generated from alternative distributions that follow symmetric and asymmetric distributions. Ten thousand samples of various sample size were generated from each of the given alternative symmetric and asymmetric distributions. The power of each test was then obtained by comparing the test of normality statistics with the respective critical values. Results show that Shapiro-Wilk test is the most powerful normality test, followed by Anderson-Darling test, Lilliefors test and Kolmogorov-Smirnov test. However, the power of all four tests is still low for small sample size.

Keywords: normality test, Monte Carlo simulation, skewness, kurtosis

Introduction

Assessing the assumption of normality is required by most statistical procedures. Parametric statistical analysis is one of the best examples to show the importance of assessing the normality assumption. Parametric statistical analysis assumes a certain distribution of the data, usually the normal distribution. If the assumption of normality is violated, interpretation and inference may not be reliable or valid. Therefore it is important to check for this assumption before proceeding with any relevant statistical procedures. Basically, there are three common ways to check the normality assumption. The easiest way is by using graphical *methods*. The normal quantile-quantile plot (*Q-Q* plot) is the most commonly used and effective diagnostic tool for checking normality of the data. Other common graphical methods that can be used to assess the normality assumption include histogram, box-plot and stem-and-leaf plot. Even though the graphical methods can serve as a useful tool in checking normality for sample of n independent observations, they are still not sufficient to provide conclusive evidence that the normal assumption holds. Therefore, to support the graphical methods, more formal methods which are the *numerical methods* and formal *normality tests* should be performed before making any conclusion about the normality of the data.

ISBN 978-967-363-157-5

© 2010 Malaysia Institute of Statistics, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia

* An earlier version of this paper was presented at the Regional Conference on Statistical Sciences 2010 on 13 – 14 June 2010 at Kota Bharu, Kelantan, Malaysia, jointly organized by Institut Statistik Malaysia and Universiti Teknologi MARA (UiTM). It has been published in the Conference Proceeding

The numerical methods include the skewness and kurtosis coefficients whereas normality test is a more formal procedure whereby it involves testing whether a particular data follows a normal distribution. There are significant amount of normality tests available in the literature. However, the most common normality test procedures available in statistical software are the Shapiro-Wilk (*SW*) test, Kolmogorov-Smirnov (*KS*) test, Anderson-Darling (*AD*) test and Lilliefors (*LF*) test. Some of these tests can only be applied under a certain condition or assumption. Moreover, different test of normality often produce different results i.e. some test reject while others fail to reject the null hypothesis of normality. The contradicting results are misleading and often confuse practitioners. Therefore, the choice of test of normality to be used should indisputably be given tremendous attention. This study focuses on comparing the power of four normality tests; *SW*, *KS*, *AD* and *LF* tests via Monte Carlo simulation. The simulation process was carried out using FORTRAN programming language. Section 2 discusses the classification of normality tests. The Monte Carlo simulation methodology is explained in Section 3. Results and comparisons of the power of the normality tests are discussed in Section 4. Finally a conclusion is given in Section 5.

Methodology

There are nearly 40 tests of normality available in the statistical literature (Dufour et al., 1998). The effort of developing techniques to detect departures from normality was initiated by Pearson (1895) who worked on the skewness and kurtosis coefficients (Althouse et al., 1998). Tests of normality differ in the characteristics of the normal distribution they focus on, such as its skewness and kurtosis values, its distribution or characteristic function, and the linear relationship existing between the distribution of the variable and the standard normal variable, Z . The tests also differ in the level at which they compare the empirical distribution with the normal distribution, in the complexity of the test statistic and the nature of its distribution (Seier, 2002).

The tests of normality can sub-divided into two categories which are *descriptive statistics* and *theory-driven methods* (Park, 2008). Skewness and kurtosis coefficients are categorized as descriptive statistics whereas theory-driven methods include the normality tests such as *SW*, *KS* and *AD* tests. However, Seier (2002) classified the tests of normality into four major sub-categories which are skewness and kurtosis test, empirical distribution test, regression and correlation test and other special test. Arshad et al. (2003) also categorized the tests of normality into four major categories which are tests of chi-square types, moment ratio techniques, tests based on correlation and tests based on the empirical distribution function. The following sub-sections review some of the most well-known tests of normality based on EDF, regression and correlation and moments. The simulation procedure is then explained.

Empirical Distribution Function (EDF) Tests

The idea of the EDF tests in testing normality of data is to compare the empirical distribution function which is estimated based on the data with the cumulative distribution function (CDF) of normal distribution to see if there is a good agreement between them. Dufour et al. (1998) described EDF tests as those based on a measure of discrepancy between the empirical and hypothesized distributions. The EDF tests can be further subdivided into those belong to *supremum* and *square* class of the discrepancies. Arshad et al. (2003) and Seier (2002) claimed that the most crucial and widely known EDF tests are Kolmogorov-Smirnov, Anderson-Darling and Cramer Von Mises tests.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (referred to as *KS* henceforth) statistic belongs to the *supremum* class of EDF statistics and this class of statistics is based on the largest vertical difference between the hypothesized and empirical distribution (Conover, 1999). Given n ordered data points, $x_1 < x_2 < \dots < x_n$, Conover (1999) defined the test statistic proposed by Kolmogorov (1933) as,

$$T = \sup_x |F^*(x) - F_n(x)| \quad (1)$$

where ‘sup’ stands for supremum which means the greatest. $F^*(x)$ is the hypothesized distribution function whereas $F_n(x)$ is the EDF estimated based on the random sample. In *KS* test of normality, $F^*(x)$ is taken to be a normal distribution with known mean, μ , and standard deviation, σ .

The *KS* test statistic is meant for testing,

H_0 : $F(x) = F^*(x)$ for all x from $-\infty$ to ∞ (The data follow a specified distribution)

H_a : $F(x) \neq F^*(x)$ for at least one value of x (The data do not follow the specified distribution)

If T exceeds the $1-\alpha$ quantile as given by the table of quantiles for the Kolmogorov test statistic, then we reject H_0 at the level of significance, α . This simulation study used the *KSONE* subroutine given in the FORTRAN IMSL libraries.

Lilliefors Test

Lilliefors (*LF*) test is a modification of the Kolmogorov-Smirnov test. The *KS* test is appropriate in a situation where the parameters of the hypothesized distribution are completely known. However, sometimes it is difficult to initially or completely specify the parameters as the distribution is unknown. In this case, the parameters need to be estimated based on the sample data. When the original *KS* statistic is used in such situation, the results can be misleading whereby the probability of type I error tend to be smaller than the ones given in the standard table of the *KS* test (Lilliefors, 1967). In contrast with the *KS* test, the parameters for *LF* test are estimated based on the sample. Therefore, in this situation, the *LF* test will be preferred over the *KS* test (Oztuna, 2006). Given a sample of n observations, *LF* statistic is defined as (Lilliefors, 1967),

$$D = \max_x |F^*(X) - S_n(X)| \quad (2)$$

where $S_n(X)$ is the sample cumulative distribution function and $F^*(X)$ is the cumulative normal distribution function with $\mu = \bar{X}$, the sample mean and s^2 , the sample variance, defined with denominator $n - 1$.

Even though the *LF* statistic is the same as the *KS* statistic, the table for the critical values is different which leads to a different conclusion about the normality of a data (Mendes & Pala, 2003). The table of critical values for this test can be found in Table A15 of the textbook written by Conover (1999). If D exceeds the corresponding critical value in the table, then the null hypothesis is rejected. This simulation study used the *LILLF* subroutine given in the FORTRAN IMSL libraries.

Anderson-Darling Test

Anderson-Darling (*AD*) test is a modification of the Cramer-von Mises (*CVM*) test. It differs from the *CVM* test in such a way that it gives more weight to the tails of the distribution (Farrel & Stewart, 2006). According to Arshad *et al.* (2003), this test is the most powerful EDF tests. The *AD* test statistic belongs to the *quadratic* class of the EDF statistic in which it is based on the squared difference $(F_n(x) - F^*(x))^2$. Anderson and Darling (1954) defined the statistic for this test as,

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 \psi(F^*(X)) dF^*(x) \quad (3)$$

where ψ is a nonnegative weight function which can be computed by, $\psi = [F^*(x)(1 - F^*(x))]^{-1}$. In order to make the computation of this statistic easier, the following formula can be applied (Arshad *et al.*, 2003),

$$W_n^2 = -n - \frac{1}{n} \sum (2i - 1) \{ \log F^*(X_i) + \log (1 - F^*(X_{n+1-i})) \} \quad (4)$$

where $F^*(x_i)$ is the cumulative distribution function of the specified distribution
 x_i 's are the ordered data
 n is the sample size

This study used the following modified *AD* statistic given by D'Agostino and Stephens (1986) which takes into accounts the sample size n ,

$$W_n^{2*} = W_n^2 (1.0 + 0.75/n + 2.25/n^2) \quad (5)$$

Cramer-von Mises Test

Conover (1999) stated that the Cramer-von Mises test was developed by Cramer (1928), von Mises (1931) and Smirnov (1936). The *CVM* statistic uses the weight function, $\psi = 1$, so that the *AD* statistic in equation (2) becomes (Thadewald & Buning, 2007),

$$CVM = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 [F(x)] dF(x) \quad (6)$$

The *CVM* statistic can be computed as,

$$CVM = \frac{1}{12n} + \sum_{i=1}^n \left[F_0(x_{(i)}) - \frac{2i-1}{2n} \right]^2 \quad (7)$$

The test rejects H_0 if $CM \geq c_{1-\alpha}$. The approximate critical values $c_{1-\alpha}$ can be found in Anderson and Darling (1954). This test is not considered in this simulation study.

Regression and Correlation Tests

Dufour *et al.* (1998) defined correlation tests as those based on the ratio of two weighted least-squares estimates of scale obtained from order statistics. The two estimates are the normally distributed weighted least squares estimates and the sample variance from other population. Some of the regression and

correlation tests are Shapiro-Wilk test, Shapiro-Francia test and Ryan-Joiner test. Only the Shapiro-Wilk test is discussed in this paper.

Shapiro-Wilk Test

Shapiro and Wilk (1965) test was originally restricted for sample size of less than 50. This test was the first test that was able to detect departures from normality due to either skewness or kurtosis, or both (Althouse *et al.*, 1998). It has become the preferred test because of its good power properties (Mendes & Pala, 2003). Given an ordered random sample, $y_1 < y_2 < \dots < y_n$, the original Shapiro-Wilk test statistic (Shapiro, 1965) is defined as,

$$W = \frac{(\sum_{i=1}^n a_i y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

where y_i is the i^{th} order statistic,

\bar{y} is the sample mean,

$$\mathbf{a}_i = (a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

and $\mathbf{m} = (m_1, \dots, m_n)^T$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and \mathbf{V} is the covariance matrix of those order statistics.

The value of W lies between zero and one. Small values of W lead to the rejection of normality whereas a value of one indicates normality of the data. SW test was modified by Royston (1982a) to broaden the restriction of the sample size to 2000 and algorithm AS181 was then provided (1982b, 1982c). Later, Royston (1992) observed that Shapiro-Wilk's (1965) approximation for the weights \mathbf{a} used in the algorithms was inadequate for $n > 50$. He then gave an improved approximation to the weights and provided algorithm AS R94 (Royston, 1995) which can be used for any n in the range $3 \leq n \leq 5000$. This study used the algorithm AS R94.

Moment Tests

In addition to the types of normality test categorized by Seier (2002) above, there are also other types of normality test. One of these types is called the moment tests. Moment tests are those derived from the recognition that the departure of normality may be detected based on the sample moments which are the skewness and kurtosis. The procedures for individual skewness and kurtosis tests can be found in D'Agostino and Stephens (1986). The two most widely known are the tests proposed by D'Agostino-Pearson (1973) and Jarque-Bera (1987). The D'Agostino and Pearson test statistic is

$$DP = Z^2(\sqrt{b_1}) + Z^2(b_2) \quad (9)$$

where $Z(\sqrt{b_1})$ and $Z(b_2)$ are the normal approximations to sample skewness ($\sqrt{b_1}$) and kurtosis (b_2) respectively. The JB statistic is based on sample skewness ($\sqrt{b_1}$) and kurtosis (b_2) and is given as

$$JB = n \left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2 - 3)^2}{24} \right) \quad (10)$$

Simulation Procedures

In this study, Monte Carlo procedures was used to evaluate the power of *SW*, *KS*, *AD* and *LF* test statistics in testing if a random sample of n independent observations come from a population with a normal $N(\mu, \sigma^2)$ distribution. The null and alternative hypotheses are:

- H_0 : The distribution is normal
 H_1 : The distribution is not normal

Two levels of significance, $\alpha = 5\%$ and 10% were considered to investigate the effect of the significance level on the power of the tests. The critical values for each test vary with the sample size (Yazici & Yolacan, 2007). Therefore, first, appropriate critical values were obtained for each normality test statistic for sample sizes $n = 10, 15, 20, 25, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 1500$ and 2000 . The critical values were obtained based on 50,000 simulated samples from a standard normal distribution. The generated test statistics were then ordered to create an empirical distribution.

As the *SW* is a left-tailed test, their critical values are the $100(\alpha)^{th}$ percentiles of the empirical distributions of the test statistics. The *AD*, *KS*, and *LF* tests are right-tailed test, so their critical values are the $100(1 - \alpha)^{th}$ percentiles of the empirical distribution of the test statistics. In order to obtain the simulated power of the four normality tests at $\alpha = 5\%$ and 10% , for each sample size, a total of 10,000 samples were drawn from each of the 14 different non-normal distributions. The alternative distributions considered were seven symmetric distributions; U (0,1), Beta (2,2), t (300), t (10), t (7), Laplace and t (5) and seven asymmetric distributions; Beta (6,2), Beta (2,1), Beta (3,2), $\chi^2(20)$, Gamma (4,5), $\chi^2(4)$ and Gamma (1,5). These distributions were selected to cover various standardized skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) values. Simulation and computations were performed using FORTRAN compiler and the subroutines available in IMSL (*International Mathematical and Statistical Libraries*) libraries.

Results

The power of the tests varies with the significance level, sample size and alternative distributions. However, only the results of power for several sample sizes and selected distributions were presented in this paper due to space constraints. The sample sizes presented were selected at the point which the power dramatically changed.

Comparison of Power against the Symmetric Non-normal Distributions

Table 1 summarizes the simulated power for selected symmetric non-normal distributions for $\alpha = 5\%$ and 10% . Some plots are given in Figure 1. For symmetric distributions with kurtosis less than 3 that is platykurtic distributions, *SW* outperforms the other three tests. However, for sample size 30 or less the powers at 5% significance level for all four tests are less than 40%. Similarly, *SW* performs better than *AD*, *KS* and *LF* for symmetric distributions with kurtosis greater than 3 that is leptokurtic distributions. Again the performance of all tests is low for small sample sizes. Overall, generally for symmetric non-normal distributions, *SW* is the best test followed by *AD*, *LF* and *KS* tests. Results also show that *LF* test performs better than the *KS* test.

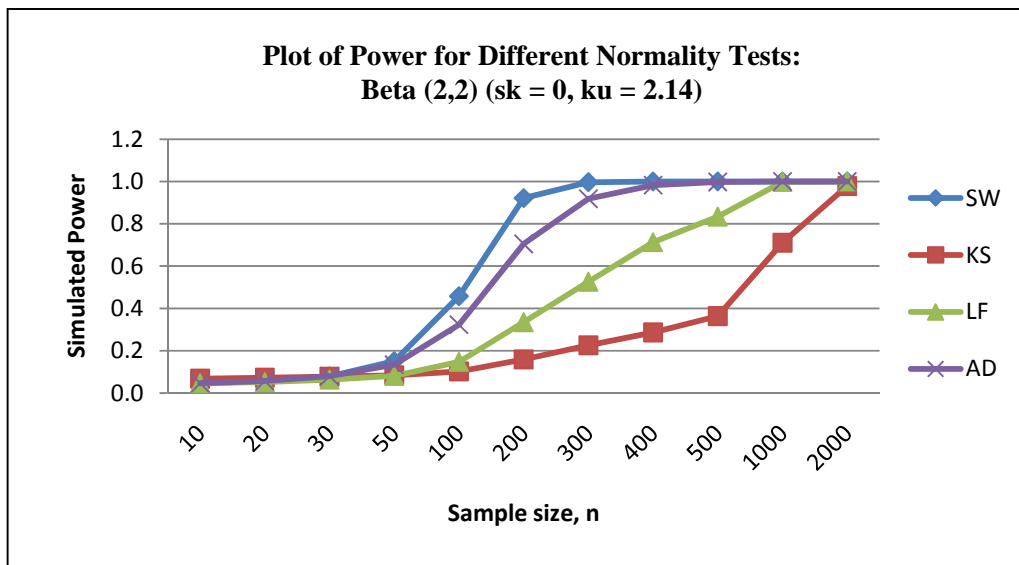


Figure 1(a): Comparison of Power for Different Normality Tests against Beta (2,2) Distribution ($\alpha = 0.05$)

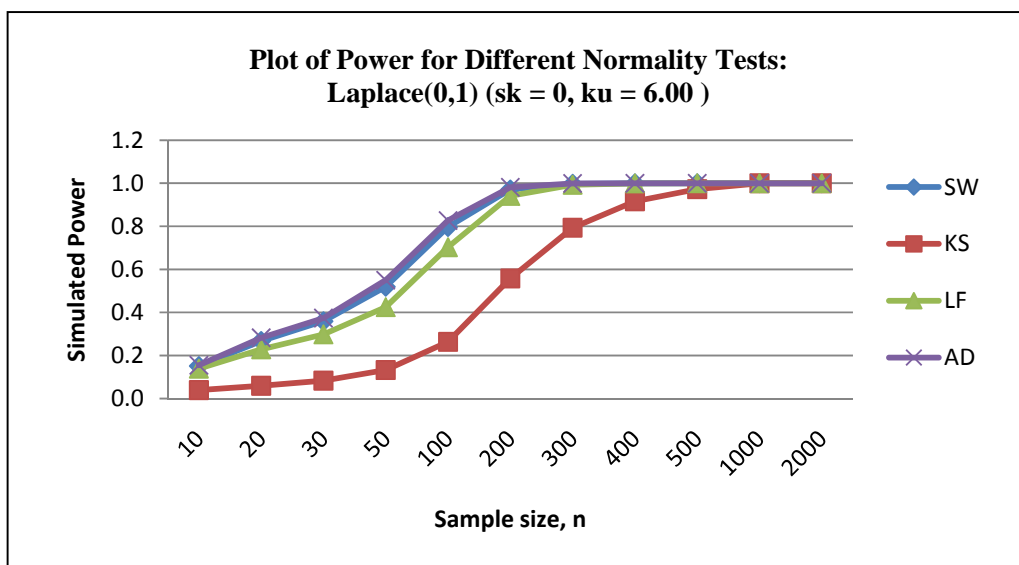


Figure 1(b): Comparison of Power for Different Normality Tests against Laplace (0,1) Distribution ($\alpha = 0.05$)

Table 1: Comparison of Power for Different Normality Tests against the Symmetric Non-normal Distributions

| Alternative Distribution | Skewness $\sqrt{\beta_1}$ | Kurtosis β_2 | Sample Size (n) | Power of Test | | | | | | | |
|--------------------------|------------------------------|-----------------------|-----------------|-----------------|--------|--------|--------|-----------------|--------|--------|--------|
| | | | | $\alpha = 0.05$ | | | | $\alpha = 0.10$ | | | |
| | | | | SW | KS | LF | AD | SW | KS | LF | AD |
| U(0,1) | 0 | 1.80 | 10 | 0.0920 | 0.0858 | 0.0671 | 0.0847 | 0.1821 | 0.1607 | 0.1283 | 0.1648 |
| | | | 20 | 0.2014 | 0.1074 | 0.1009 | 0.1708 | 0.3622 | 0.1785 | 0.1860 | 0.2926 |
| | | | 30 | 0.3858 | 0.1239 | 0.1445 | 0.3022 | 0.5764 | 0.2078 | 0.2578 | 0.4466 |
| | | | 50 | 0.7447 | 0.1618 | 0.2579 | 0.5817 | 0.8816 | 0.2653 | 0.4069 | 0.7314 |
| | | | 100 | 0.9970 | 0.2562 | 0.5797 | 0.9523 | 0.9996 | 0.3980 | 0.7530 | 0.9824 |
| | | | 200 | 1.0000 | 0.4851 | 0.9484 | 1.0000 | 1.0000 | 0.6604 | 0.9846 | 1.0000 |
| | | | 300 | 1.0000 | 0.7045 | 0.9974 | 1.0000 | 1.0000 | 0.8419 | 0.9996 | 1.0000 |
| | | | 400 | 1.0000 | 0.8446 | 0.9999 | 1.0000 | 1.0000 | 0.9332 | 1.0000 | 1.0000 |
| | | | 500 | 1.0000 | 0.9331 | 1.0000 | 1.0000 | 1.0000 | 0.9744 | 1.0000 | 1.0000 |
| | | | 1000 | 1.0000 | 0.9996 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| t(7) | 0 | 5.00 | 10 | 0.0892 | 0.0421 | 0.0797 | 0.0862 | 0.1458 | 0.0913 | 0.1396 | 0.1471 |
| | | | 20 | 0.1295 | 0.0437 | 0.0946 | 0.1177 | 0.1956 | 0.0948 | 0.1576 | 0.1834 |
| | | | 30 | 0.1697 | 0.0467 | 0.1060 | 0.1431 | 0.2372 | 0.0981 | 0.1771 | 0.2163 |
| | | | 50 | 0.2244 | 0.0529 | 0.1198 | 0.1785 | 0.3036 | 0.1107 | 0.1974 | 0.2632 |
| | | | 100 | 0.3698 | 0.0593 | 0.1761 | 0.2781 | 0.4569 | 0.1234 | 0.2800 | 0.3774 |
| | | | 200 | 0.5793 | 0.0935 | 0.2826 | 0.4496 | 0.6626 | 0.1808 | 0.4012 | 0.5581 |
| | | | 300 | 0.7278 | 0.1280 | 0.3872 | 0.5984 | 0.7941 | 0.2358 | 0.5214 | 0.7062 |
| | | | 400 | 0.8268 | 0.1625 | 0.4888 | 0.7115 | 0.8736 | 0.2888 | 0.6236 | 0.8007 |
| | | | 500 | 0.8982 | 0.2009 | 0.5755 | 0.8065 | 0.9296 | 0.3398 | 0.7033 | 0.8727 |
| | | | 1000 | 0.9937 | 0.4248 | 0.8740 | 0.9794 | 0.9967 | 0.6021 | 0.9364 | 0.9915 |
| | | | 2000 | 1.0000 | 0.8106 | 0.9947 | 0.9999 | 1.0000 | 0.9173 | 0.9982 | 0.9999 |

Comparison of Power against the Asymmetric Distributions

Table 2 summarizes the simulated power for selected asymmetric distributions for $\alpha = 5\%$ and 10% while Figure 2 show the plot of power for all tests against selected asymmetric distributions for 5% significance level. Again for asymmetric distributions, *SW* outperforms *AD*, *KS* and *LF* tests. *SW* achieved good power for sample size of at least 50 while *AD* and *LF* requires sample size of at least 100 to achieve good power. *KS* is the weakest test and requires much larger sample size to achieve comparable power with the other tests.

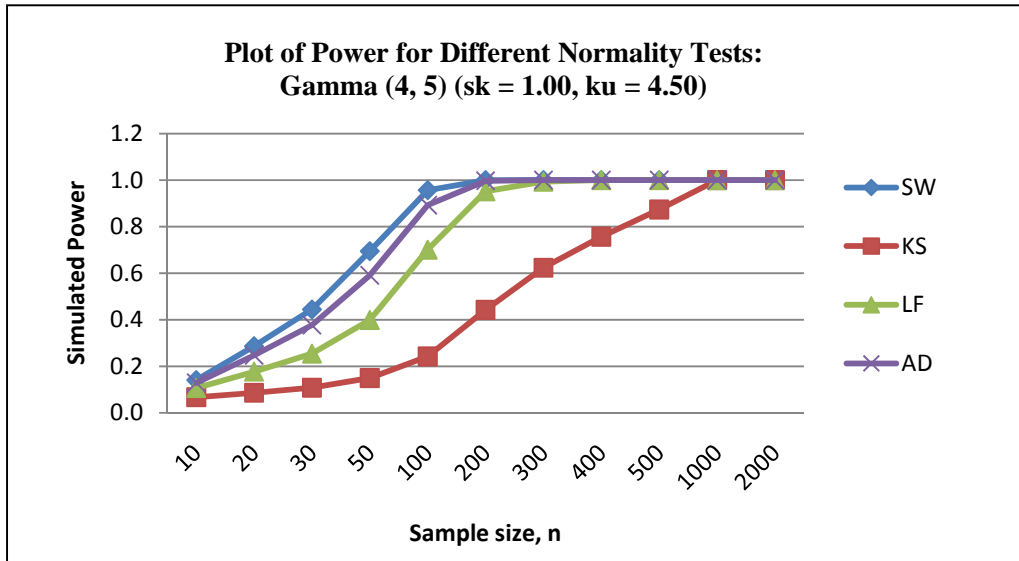


Figure 2(a): Comparison of Power for Different Normality Tests against Gamma (4,5) ($\alpha = 0.05$)

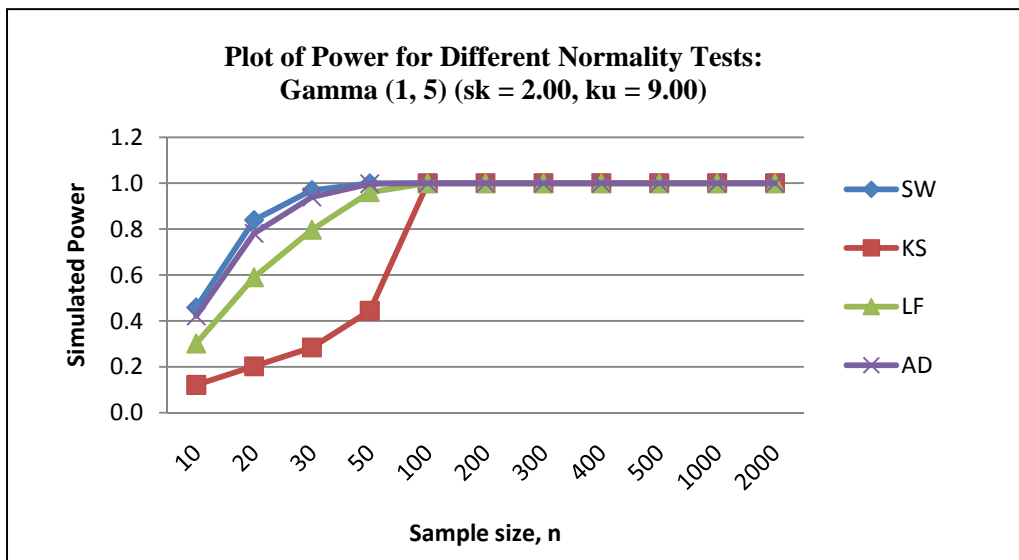


Figure 2(b): Comparison of Power for Different Normality Tests against Gamma (1,5) ($\alpha = 0.05$)

Table 2 Comparison of Power for Different Normality Tests against Asymmetric Distributions

| Alternative Distribution | Skewness $\sqrt{\beta_1}$ | Kurtosis β_2 | Sample Size (n) | Power of Test | | | | | | | |
|--------------------------|------------------------------|-----------------------|-----------------|-----------------|--------|--------|--------|-----------------|--------|--------|--------|
| | | | | $\alpha = 0.05$ | | | | $\alpha = 0.10$ | | | |
| | | | | SW | KS | LF | AD | SW | KS | LF | AD |
| Gamma (4,5) | 1.00 | 4.50 | 10 | 0.1407 | 0.0669 | 0.1065 | 0.1285 | 0.2153 | 0.1247 | 0.1809 | 0.2075 |
| | | | 20 | 0.2864 | 0.0861 | 0.1771 | 0.2469 | 0.3938 | 0.1502 | 0.2755 | 0.3462 |
| | | | 30 | 0.4442 | 0.1078 | 0.2545 | 0.3765 | 0.5628 | 0.1783 | 0.3697 | 0.4850 |
| | | | 50 | 0.6946 | 0.1495 | 0.3991 | 0.5908 | 0.7956 | 0.2337 | 0.5319 | 0.6979 |
| | | | 100 | 0.9566 | 0.2423 | 0.7008 | 0.8925 | 0.9802 | 0.3499 | 0.8107 | 0.9400 |
| | | | 200 | 0.9997 | 0.4424 | 0.9518 | 0.9970 | 1.0000 | 0.5759 | 0.9798 | 0.9992 |
| | | | 300 | 1.0000 | 0.6233 | 0.9929 | 1.0000 | 1.0000 | 0.7520 | 0.9980 | 1.0000 |
| | | | 400 | 1.0000 | 0.7568 | 0.9998 | 1.0000 | 1.0000 | 0.8725 | 0.9999 | 1.0000 |
| | | | 500 | 1.0000 | 0.8738 | 1.0000 | 1.0000 | 1.0000 | 0.9576 | 1.0000 | 1.0000 |
| | | | 1000 | 1.0000 | 0.9999 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| $\chi^2(4)$ | 1.41 | 6.00 | 10 | 0.2445 | 0.0801 | 0.1680 | 0.2196 | 0.3453 | 0.1484 | 0.2591 | 0.3190 |
| | | | 20 | 0.5262 | 0.1205 | 0.3184 | 0.4620 | 0.6525 | 0.1936 | 0.4433 | 0.5840 |
| | | | 30 | 0.7487 | 0.1584 | 0.4650 | 0.6617 | 0.8399 | 0.2465 | 0.5936 | 0.7624 |
| | | | 50 | 0.9484 | 0.2402 | 0.6841 | 0.8891 | 0.9761 | 0.3495 | 0.7991 | 0.9390 |
| | | | 100 | 0.9997 | 0.4391 | 0.9470 | 0.9971 | 0.9998 | 0.5732 | 0.9762 | 0.9992 |
| | | | 200 | 1.0000 | 0.8417 | 0.9997 | 1.0000 | 1.0000 | 0.9859 | 1.0000 | 1.0000 |
| | | | 300 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | | | 400 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | | | 500 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | | | 1000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 2000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | | |

In order to get a clearer picture of the performance of the different normality tests, the ranking procedure was used. The rank of 1 was given to the test with the highest power while rank of 4 (since there were four tests of normality considered in this study) was given to the test which has the lowest power. The ranks were then summed to get the grand total of ranks. As the lowest number was given to the test with the highest power, therefore the test which had the lowest total rank was nominated as the best test to detect the departure from normality. Table 3 and Table 4 show the rank of power based on the type of alternative distribution and sample size, respectively.

Table 3: Rank of Power Based on Types of Alternative Distribution

| Alternative Distributions | Total Rank | | | | | | | |
|---------------------------|-----------------|-------|-------|-------|-----------------|-------|-------|-------|
| | $\alpha = 0.05$ | | | | $\alpha = 0.10$ | | | |
| | SW | KS | LF | AD | SW | KS | LF | AD |
| Symmetric | 92.0 | 248.5 | 196.5 | 123.0 | 92.0 | 251.5 | 193.5 | 123.0 |
| Asymmetric | 120.5 | 270.5 | 218.5 | 160.5 | 119.0 | 271.0 | 217.0 | 163.0 |

Table 4: Rank of Power Based on Sample Size for All Alternative Distributions

| Sample size (n) | Total Rank | | | | | | | |
|-----------------|-----------------|------|------|-------|-----------------|-------|-------|------|
| | $\alpha = 0.05$ | | | | $\alpha = 0.10$ | | | |
| | SW | KS | LF | AD | SW | KS | LF | AD |
| 10 | 18.0 | 44.0 | 41.0 | 27.0 | 19.0 | 48.0 | 40.0 | 23.0 |
| 20 | 18.0 | 45.0 | 42.0 | 25.0 | 15.0 | 48.0 | 41.0 | 26.0 |
| 30 | 16.0 | 51.0 | 40.0 | 23.0 | 14.0 | 51.0 | 40.0 | 25.0 |
| 50 | 14.0 | 51.0 | 40.0 | 25.0 | 14.0 | 52.0 | 39.0 | 25.0 |
| 100 | 15.5 | 50.5 | 38.5 | 25.5 | 15.5 | 50.5 | 38.5 | 25.5 |
| 200 | 17.0 | 50.5 | 38.5 | 24.0 | 17.5 | 50.5 | 37.5 | 24.5 |
| 300 | 20.0 | 47.5 | 37.5 | 25.0 | 20.0 | 47.5 | 37.5 | 25.0 |
| 400 | 20.0 | 47.5 | 37.5 | 25.0 | 20.0 | 47.5 | 36.5 | 26.0 |
| 500 | 21.5 | 47.5 | 33.5 | 27.5 | 21.5 | 47.5 | 33.5 | 27.5 |
| 1000 | 24.5 | 44.5 | 34.5 | 26.5 | 26.0 | 41.5 | 34.5 | 28.0 |
| 2000 | 28.0 | 40.0 | 32.0 | 30.0 | 28.5 | 38.5 | 32.5 | 30.5 |
| Total | 212.5 | 519 | 415 | 283.5 | 211 | 522.5 | 410.5 | 286 |

From Table 3, it can be clearly seen that *SW* is the best test to be adopted for both symmetric non-normal and asymmetric distributions since it has the lowest total rank (for both 5% and 10% significance levels) among all the four tests considered. This is followed rather closely by the *AD* test. The results of the total rank based on sample size in Table 4 above also show that *SW* as the best test for all sample size since it consistently has the lowest total rank from $n = 10$ until $n = 2000$.

Conclusion

In general, it can be concluded that among the four tests considered, Shapiro-Wilk test is the most powerful test for all types of distribution and sample sizes whereas Kolmogorov-Smirnov test is the least powerful test. However, the power of Shapiro-Wilk test is still low for small sample size. The performance of Anderson-Darling test is quite comparable with Shapiro-Wilk test, and Lilliefors test always outperforms Kolmogorov-Smirnov test. The results of this study support the findings of Mendes and Pala (2003) and Keskin (2006) that Shapiro-Wilk test is the most powerful normality test. The results are also found to be similar to the one obtained by Farrel & Stewart (2006) which reported that simulated power for all tests increased as the sample size and significance level increased. As a concluding remark, practitioners should not depend solely on graphical techniques such as histogram to conclude about the distribution of the data. It is recommended that the graphical techniques be combined with formal normality test and inspection of shape parameters such as skewness and kurtosis coefficients. It is important to remember that skewness and kurtosis measures are also affected by sample size. Practitioners also need to be aware that these four normality tests do not perform well for small sample size (30 and below). Work is in progress to determine more recent normality tests which might work well for small sample size.

References

- Althouse, L.A., Ware, W.B. and Ferron, J.M. (1998). Detecting Departures from Normality: A Monte Carlo Simulation of A New Omnibus Test based on Moments. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Anderson, T.W. and Darling, D.A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, Vol 49, No. 268, 765-769.
- Arshad, M., Rasool, M.T. and Ahmad, M.I. (2003). Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Pakistan Journal of Applied Sciences* 3(2), pp. 85-88.
- Conover, W.J. (1999). *Practical Nonparametric Statistics*. Third Edition, John Wiley & Sons, Inc. New York, pp.428-433 (6.1).
- Cramer, H. (1928). On the composition of elementary errors, *Skandinavisk Aktuarietidskrift* 11, pp. 13–74, 141–180 (6.1).
- D'Agostino, R. and Pearson, E.S. (1973). Test for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, Vol. 60, No.3, pp. 613-622.
- D' Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-fit Techniques*, New York: Marcel Dekker.
- Dufour J.M., Farhat, A., Gardiol, L. and Khalaf, L. (1998). Simulation-based Finite Sample Normality Tests in Linear Regressions. *Econometrics Journal*, Vol. 1, pp. 154-173.
- Farrel, P.J. and Stewart, K.R. (2006). Comprehensive Study Of Tests For Normality And Symmetry: Extending The Spiegelhalter Test. *Journal of Statistical Computation and Simulation*, Vol. 76, No. 9, pp. 803–816.
- Jarque, C.M. and Bera, A.K. (1987). *A test for normality of observations and regression residuals*, Internat. Statist. Rev. 55(2), pp. 163–172.
- Keskin, S. (2006). Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation Based Small Samples. *Journal of Applied Science Research* 2(5), pp. 296-300.

- Kolmogorov, A.N. (1933). *Sulla determinazione empirica di una legge di distribuzione*, Giornale dell' Istituto Italiano degli Attuari 4, pp. 83–91 (6.1).
- Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of American Statistical Association*, Vol. 62, No.318, pp. 399-402.
- Mendes, M. and Pala, A. (2003). Type I Error Rate and Power of Three Normality Tests. *Pakistan Journal of Information and Technology* 2(2), pp. 135-139.
- Oztuna, D., Elhan, A.H. and Tuccar, E. (2006). Investigation of Four Different Normality Tests in Terms of Type I Error Rate and Power Under Different Distributions. *Turkish Journal of Medical Science*, 2006, 36(3), pp. 171-176.
- Park, H.M. (2008). *Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*. Technical Working Paper. The University Information Technology Services (UITSS) Center for Statistical and Mathematical Computing, Indiana University.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution,II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London*, 91, 343-414.
- Royston, J.P. (1982a). An Extension of Shapiro and Wilk's W Tests for Normality to Large Samples. *Applied Statistics*, 31, pp.115-124.
- Royston, J.P. (1982b). Algorithm AS 177: Expected Normal Order Statistics (Exact and Approximate), *Applied Statistics*, 31, pp.161-165.
- Royston, J.P. (1982c). Algorithm AS 181: The W Test for Normality. *Applied Statistics*, 31, pp.176-180.
- Royston, P. (1992). Approximating the Shapiro-Wilk W test for Non-normality [Abstract]. *Statistics and Computing*, 2, pp.117-119.
- Royston, P. (1995). Remark AS R94:A Remark on Algorithm AS181:The W -test for Normality. *Journal of the Royal Statistical Society*, Vol. 44, No. 4, pp. 547-551.
- Seier, E. (2002). Comparison of Tests for Univariate Normality. *InterStat Statistical Journal*, 1, pp.1-17.
- Shapiro, S.S. and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, Vol. 52, No. 3/4, pp. 591-611.
- Smirnov, N.V. (1936). *Sui la distribution de w^2 (Criterium de M.R.v. Mises)*, Comptes Rendus (Paris), 202, pp. 449–452 (6.1).
- Thadewald, T. and Buning, H. (2007). Jarque-Bera and its Competitors for Testing Normality. *Journal of Applied Statistics*, Vol. 34, No. 1, pp. 87-105.
- Von Mises, R. (1931). *Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theoretischen Physik*, F. Deuticke, Leipzig (6.1).
- Yazici, B. and Yolacan, S. (2007). A Comparison of Various Tests of Normality. *Journal of Statistical Computation and Simulation*, Vol. 77, No.2, pp. 175-183.