Journal of the Korean Data & Information Science Society
2016, **27**(5), 1399–1412

http://dx.doi.org/10.7465/jkdi.2016.27.5.1399
한국데이터정보과학회지

Comprehensive comparison of normality tests: Empirical study using many different types of data[†]

Chanmi Lee[1] · Suhwi Park[2] · Jaesik Jeong[3]

[123]Department of Statistics, Chonnam National University

Received 4 July 2016, revised 22 July 2016, accepted 5 August 2016

Abstract

We compare many normality tests consisting of different sources of information extracted from the given data: Anderson-Darling test, Kolmogorov-Smirnov test, Cramer-von Mises test, Shapiro-Wilk test, Shaprio-Francia test, Lilliefors, Jarque-Bera test, D'Agostino' D, Doornik-Hansen test, Energy test and Martinzez-Iglewicz test. For the purpose of comparison, those tests are applied to the various types of data generated from skewed distribution, unsymmetric distribution, and distribution with different length of support. We then summarize comparison results in terms of two things: type I error control and power. The selection of the best test depends on the shape of the distribution of the data, implying that there is no test which is the most powerful for all distributions.

Keywords: Empirical power, empirical type I error, limiting distribution, normality tests.

1. Introduction

Many statistical methodologies, which have been developed so far, work very well under the assumption that the data come from normal distribution and it is very important to check for normality before any statistical analysis is done. Over several decades, various tests including goodness-of-fit test (Kang *et al.*, 2014; Lee, 2013) using different sources of information from data have been introduced by many researchers. More specifically, some methods exploit sample skewness or/and sample kurtosis (D'Agostino, 1970; D'Agostino and Pearson, 1973; Jarque and Bera, 1981) while other methods use the maximum distance between two distribution functions, i.e., target normal distribution function and empirical distribution function (Kolmogorov, 1933; Lilliefors, 1967, 1969; Smirnov, 1939). Other than that, some method uses the ratio of two estimators of variance (Martinez and Iglewicz, 1981).

[†] This study was financially supported by Chonnam National University (grant number: 2015-1848 to Jaesik Jeong)

[1] Undergraduate student, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

[2] Undergraduate student, Department of Statistics, Chonnam National University, Gwangju 61186, Korea.

[3] Corresponding author: Professor, Department of Statistics, Chonnam National University, Gwangju 61186, Korea. Email: jjs3098@jnu.ac.kr

Among many methods to test for normality, we select 11 of them for comparison: Anderson-Darling test, Kolmogorov-Smirnov test, Cramer-von Mises test, Shapiro-Wilk test, Shaprio-Francia test, Lilliefors, Jarque-Bera test, D'Agostino' D, Doornik-Hansen test, Energy test and Martinzez-Iglewicz test (Anderson, 1962; Cramer, 1928; von Mises, 1928; Shapiro and Wilk, 1965; Shapiro and Francia, 1972; Lilliefors, 1967; Jarque and Bera, 1981; D'Agostino, 1970; Doornik and Hansen, 2008; Szekely and Rizzo, 2005; Martinez and Iglewicz, 1981). Since each method makes use of different type of information extracted from the given data, the performance of each method depends heavily on the properties of the data: skewness, kurtosis, and the type of support. Accordingly, it is well known that there is no test which is the most powerful for all types of alternatives. In order to compare such methods, we apply those tests to the data generated from many different distributions and then find the best test for each alternative.

In this paper, we focus on the one sample test. That is, given the samples $X_1, \cdots, X_n$, we test for the normality of the given sample of data. Furthermore, for simplicity, our interest is restricted to the univariate normal data.

2. Normality tests

Let $X_1, \cdots, X_n$ be random samples. Ordered samples of $X_1, \cdots, X_n$ are denoted by $Y_1, \cdots, Y_n$. For simplicity, we denote the k-th sample moment with respect to sample mean, m_k by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$. That is, sample skewness and kurtosis are denoted by $S = \frac{m_3}{m_2^{3/2}}$ and $K = \frac{m_4}{m_2^2}$, respectively.

2.1. Kolmogorov-Smirnov test

The empirical distribution function F_n for n iid observations X_i's is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$$

where $I_{X_i \leq x}$ is the indicator function. Suppose that the F is the target normal distribution. Then the Kolmogorov-Smirnov (KS) statistic is defined as

$$T_{KS} = sup_x |F_n(x) - F(x)|$$

where sup_x is the supremum of the set of distances. The null hypothesis that the samples come from the normal distribution is rejected if the T_{KS} is larger than the tabulated values calculated by Smirnov in 1948 (Smirnov, 1948). Note that Glivenko-Cantelli theorem supports the theoretical background of the test.

2.2. Lilliefors test

The KS test is used when all parameters are completely specified while Lilliefors test can be used for the case parameters are not specified. The test statistic is defined as

$$T_L = sup_x |F_n(x) - F^*(x)|$$

where T_L is the supremum, over all x, of the absolute value of the difference $F^*(x) - F_n(x)$, and $F^*(x)$ is the cumulative distribution function of a normal distribution with sample

mean ($\bar{X}$) and variance ($\frac{n}{n-1}m_2$) , and $F_n(x)$ is the empirical distribution function of the values of $X_1, \cdots, X_n$. Of course, this test can be applied to the standardized values. The null hypothesis that the samples come from the normal distribution is rejected if the T_L is larger than the tabulated values calculated by Conover (1999).

2.3. Cramer-von Mises test

Let $Y_1, \cdots, Y_n$ be the ordered samples in increasing order. The test statistic is defined as

$$T_{CvM} = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(Y_i) \right]^2 .$$

Then the null hypothesis that the data come from the theoretical (normal) distribution (F) is rejected if the value of T_{CvM} is larger than the tabulated values calculated by Anderson (1962).

2.4. Anderson-Darling test

Anderson-Darling test is a modification of the CvM test, i.e., more weight is given to the tails compared to the CvM test. The test statistic is defined as

$$T_{AD} = -n - M$$

where

$$M = \sum_{i=1}^{n} \frac{2i-1}{n} [lnF(Y_i) + ln(1 - F(Y_{n+1-i}))]$$

and F is the cdf of normal distribution, and Y_i is the ordered data. The null hypothesis that the samples come from the normal distribution is rejected if the T_{AD} is larger than the tabulated values calculated by Stephens (Stephens, 1974, 1976; Stenphens, 1977).

2.5. Shapiro-Wilk test

The Shapiro-Wilk test was developed by Shapiro and Wilk (1965). The test statistic is defined as

$$T_{SW} = \frac{\left(\sum_{i=1}^{n} a_i y_i \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where the constants a_i are given by $(a_1, \cdots, a_n) = \frac{\mu^T \Sigma^{-1}}{(\mu^T \Sigma^{-1} \Sigma^{-1} \mu)^{1/2}}$. Here $y = (y_1, \cdots, y_n)$ are ordered random observations from normal distribution in increasing order, $\mu = (\mu_1, \cdots, \mu_n) = (E(Y_1), \cdots, E(Y_n))$ are their expectation, and Σ is the covariance matrix of Y. Then the null hypothesis that the data come from the theoretical normal distribution (F) is rejected if the value of T_{SW} is larger than the predetermined values (empirical percentage points). Such empirical percentage points are determined by Monte Carlo simulation by calculating the test statistic with the approximated elements of the vector a incorporated in the formula. Due to the difficulty of the approximation of the vector a, there was a limitation on the

sample size less than 50. However, such limitation was relaxed to the sample size of 2000 by Royston (1992) and further to the sample size of 5000 by Rahman and Govidarajulu (1997).

Compared to the Anderson-Darling test, the Shapiro-Wilk test is less affected by ties.

2.6. Shapiro-Francia test

The Shapiro-Francia test, a simplified version of Shapiro-Wilk test, was developed in 1972 (Shapiro and Francia, 1972). When calculating the test statistic of Shapiro-Wilk, the covariance matrix (Σ) of ordered normal variables should be calculated. However, calculating Σ for large n was burdensome and the elements of Σ were given only for sample sizes up to 20 in original paper. To avoid such difficulty, Shapiro and Francia suggested a simplified statistic using diagonal covariance matrix. The test statistic is defined as

$$T_{SF} = \frac{\left(\sum_{i=1}^{n} b_i y_i\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where the constants b_i are given by $(b_1, \cdots, b_n) = \frac{\mu^T}{(\mu^T \mu)^{1/2}}$ and μ is the vector of expected values of standard normal order statistics. Note that T_{SF} presents the squared product moment correlation coefficient between ordered observed data and expected values of standard normal order statistics with large values of T_{SF} indicating normality (Royston, 1983).

2.7. Jarque-Bera test

Jarque-Bera test is score test (also known as Lagrange multiplier test). The test statistic with asymptotic behavior of $\chi^2(2)$ distribution is defined as

$$JB \;=\; \frac{n}{6}\left(\; S^2 \;+\; \frac{1}{4}\,(K-3)^2\right)$$

where n is the number of observations, S and K are the sample skewness and kurtosis, respectively. That is,

$$S \;=\; \frac{m_3}{{m_2}^{3/2}} \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}}, \quad K \;=\; \frac{m_4}{{m_2}^2} \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2}$$

where $\overline{x}$ is the sample mean, m_2, m_3, and m_4 are the estimates of the second, third and fourth central moments, respectively.

It should be noted that the statistic was originally developed by Bowman and Shenton (1975). They, however, did not mention on the properties of the statistic such as large or finite sample properties. Later, Jarque and Bera recognized that the statistic is score test. Therefore, they uncovered its asymptotic efficiency. Also, they studied its finite sample properties computationally, not analytically. Notice that its finite sample property is still unknown. Approximation to the true distributions instead were investigated by D'Agostino and Tietjen (1973) for S, and by D'Agostino and Pearson (1973) for K.

2.8. D'Agostino's D

A transformation of the distribution of the standardized third sample moment into approximate normal distribution is developed by D'Agostino (1970). It is emphasized that the transformation can be used to perform tests of normality against skewed alternatives only. The standardized third sample moment is defined as

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}^{3/2}}$$

where $X_1, \cdots, X_n$ are random sample of n observations with mean $\bar{X}$.

For $n \geq 8$, let

$$\beta = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$W^2 = [2(\beta - 1)]^{1/2} - 1,$$

$$\delta = 1/\sqrt{lnW},$$

$$\alpha = \sqrt{\frac{\beta - 1}{2}} - 1,$$

$$Y = S\left\{\alpha\ \frac{(n+1)(n+3)}{6(n-2)}\right\}^{1/2}.$$

Then

$$Z = \delta\ ln\left\{Y + (Y^2 + 1)^{1/2}\right\}$$

is approximately a standard normal variable. Tables including critical values of S for some sample sizes are given in Pearson and Hartely (1966). However, we can calculate critical values for any sample size and any significance level by Monte carlo simulation.

2.9. Doornik-Hansen test

The Doornick-Hansen test for multivariate normality is a powerful alternative to the Shapiro-Wilk test. The test statistic is defined as

$$DH = Z_1'Z_1 + Z_2'Z_2$$

where Z_1 and Z_2 are approximate normal variate transformed from sample skewness (S) and kurtosis (K). The univariate skewness, S, is transformed into an approximately normal variate, z_1, as in D'Agostino (1970):

$$z_1 = \delta \cdot ln(y + \sqrt{1 + y^2})$$

where

$$\beta = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},$$

$$\omega^2 = -1 + \sqrt{2(\beta - 1)},$$
$$\delta = (ln\sqrt{\omega^2})^{-1/2},$$
$$y = S\left[\frac{(\omega^2 - 1)(n+1)(n+3)}{12(n-2)}\right]^{1/2}.$$

The univariate kurtosis, K, is transformed from a gamma variate into a χ^2-variate and then into a standard normal variable, z_2, using the Wilson-Hilferty transform (Wilson and Hilferty, 1931):

$$z_2 = \sqrt{9\alpha}\left[\left(\frac{\chi}{2\alpha}\right)^{1/3} - 1 + \frac{1}{9\alpha}\right]$$

where

$$\chi = 2f(K - 1 - S^2),$$
$$\alpha = a + S^2 c,$$
$$f = \frac{(n+5)(n+7)(n^3 + 37n^2 + 11n - 313)}{12\delta},$$
$$c = \frac{(n-7)(n+5)(n+7)(n^2 + 2n - 5)}{6\delta},$$
$$a = \frac{(n-2)(n+5)(n+7)(n^2 + 27n - 70)}{6\delta},$$
$$\delta = (n-3)(n+1)(n^2 + 15n - 4).$$

In k-variate case, the statistic $Z_1'Z_1 + Z_2'Z_2$ is approximately χ^2 distributed with $2k$ degree of freedom.

2.10. Energy test

Szekely and Rizzo provided a new test for multivariate normality, which is called energy test (Szekely and Rizzo, 2005). Energy distance is a statistical distance between two probability distributions. If X and Y are independent random vectors with cumulative distribution functions F and G respectively, then the energy between F and G is defined

$$D_E(F, G) = 2E||X - Y|| - E||X - X'|| - E||Y - Y'||$$

where $X,\ X'$ are independent and identically distributed and $Y,\ Y'$ are iid, $||\cdot||$ is Euclidean norm of a vector. Note that if $F = G$ then $D_E = 0$.

This concept was translated to statistical terminology by Szekely in the 1980s. Following the introduction by Szekely, energy statistic was used for Goodness-of-fit test. Given statistical samples $x_1, \cdots, x_n$ with distribution F, the energy statistic for testing $H_0 : F = F_0$ *v.s.* $H_1 : F \neq F_0$ is defined as

$$T_E = n\left[\frac{2}{n}\sum_{i=1}^{n} E\|x_i - X\| - E\|X - X'\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|\right]$$

where $X,\ X'$ are independent and identically distributed with distribution F_0, $\|\cdot\|$ is Euclidean norm of a vector $\|x\| = \sqrt{\sum_j x_j^2}$. Here $F_0 \sim N_d(\mu, \Sigma)$, d-variate normal distribution with mean μ and covariance Σ. Note that sample mean and sample covariance are used for the unknown parameters. That is, $\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^t$.

Theoretical property such as consistency of power, which is defined as

$$\lim_{n\to\infty} P(T_E \geq c) = 0$$

where c is critical region depending on three numbers such as level of significance, sample size, and dimension, was proved in the paper. The critical region is obtained by parametric bootstrap.

2.11. Martinez-Iglewicz test

Martinez-Iglewicz test for normality is based on the median and a robust estimator of dispersion. The test statistic, which is the ratio of two estimators of variance, is defined as

$$I \ = \ \frac{\sum_{i=1}^{n}(X_i - \tilde{X})^2}{(n-1)S_b^2}$$

where $\tilde{X}$ is the sample median, S_b^2 is a biweight estimator of scale

$$S_b^2 = \frac{n\sum_{|z_i|<1}(X_i - \tilde{X})^2(1-z_i^2)^4}{\left[\sum_{|z_i|<1}(1-z_i^2)(1-5z_i^2)\right]^2},$$

and $z_i = (X_i - \tilde{X})/(kA)$ for $|z_i| < 1$ and $z_i = 0$ otherwise and $A = med|X_i - \tilde{X}|$. Note that $k = 9$ was used in the original paper. A value of the test statistic that is close to one indicates that the distribution is normal. This test is available when n is greater than or equal to 3.

2.12. Relationships among tests

Anderson-Darling v.s. Cramer-von Mises

Both tests are based on the quadratic distance of two cumulative distribtuion functions (cdfs). Suppose that the hypothesized distribution is F, and empirical cdf is F_n. Then the basic form of test statistic is

$$TS = n\int_{-\infty}^{\infty}\left(F_n(x) - F(x)\right)^2 w(x)dF(x)$$

where $w(x)$ is a weight function. The statistic TS with $w(x) = 1$ is called the Cramer-von Mises statistic while Anderson-Darling statistic has the weight

$$w(x) = \frac{1}{F(x)[1-F(x)]}.$$

Shapiro-Wilks v.s. Shapiro-Francia

Both tests have similar types of test statistic defined as

$$T_{SW} = \frac{\left(\sum_{i=1}^{n} c_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

If the constants c_i are given by $(c_1, \cdots, c_n) = \frac{\mu^T \Sigma^{-1}}{(\mu^T \Sigma^{-1} \Sigma^{-1} \mu)^{1/2}}$, then the test is called Shapiro-Wilks test. However, if $(c_1, \cdots, c_n) = \frac{\mu^T}{(\mu^T \mu)^{1/2}}$, then the test is called Shapiro-Francia test. Here $y = (y_1, \cdots, y_n)$ are ordered random observations from normal distribution in increasing order, $\mu = (\mu_1, \cdots, \mu_n) = (E(Y_1), \cdots, E(Y_n))$ are their expectation, and Σ is the covariance matrix of Y.

D'Agostino D v.s. Jarque-Bera

D'Agostino D makes use of sample skewness information only while Jarque-Bera considers two sample moments such as skewness and kurtosis. Note that D'Agostino K-squared test (K^2 test), which is the modified version of D'Agostino D, also uses both. Main difference between K^2 and Jarque-Bera is that K^2 test first transforms sample skewness and kurtosis into normal variates, respectively, and then combine them into single test statistic. That is,

$$K^2 = [Z(S)]^2 + [Z(K)]^2.$$

The first normal transformation $Z(S)$ for skewness is developed by D'Agostino in 1970 while the second normal transformation $Z(K)$ for kurtosis is developed by Anscombe and Glynn (1983). In contrast, Jarque-Bera considers different transformation:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right).$$

Both test statistics have the common approximate distribution. That is, they have $\chi^2(2)$ distribution in asymptotic sense.

3. Simulation

For the purpose of comparison, we consider eleven normality tests: Anderson-Darling test (AD), Kolmogorov-Smirnov test (KS), Cramer-von Mises test (CV), Shapiro-Wilk test (SW), Shaprio-Francia test (SF), Lilliefors (LF), Jarque-Bera test (JB), D'Agostino' D (DA), Doornik-Hansen test (DH), Energy test (EG) and Martinzez-Iglewicz test (MI).

We validate those normality tests on many different data sets from various distributions at different sample sizes (n=20, 30, 50, 100). Regarding type I error control, we consider four different levels of α=0.01, 0.05, 0.1, and 0.2. For each level of α, all test are applied to the data generated from standard normal distribution. For power comparison, we consider eleven different alternative distributions: $t(2)$, $t(5)$, $\chi^2(2)$, $\chi^2(5)$, exp(1), uniform(0,1), Gamma(3,5), Beta(2,1), lognormal, standard Cauchy distribution, and normal mixture. As a normal mixture, we consider

$$f(x) = \frac{1}{2}\phi_1(x) + \frac{1}{2}\phi_2(x)$$

where ϕ_1 and ϕ_2 are pdfs of $N(2,1)$ and $N(-2,1)$, respectively.

4. Results

We summarize the results in two respects: type I error control and power comparison.

4.1. Type I error control

For different sample size of n=20, 30, 50,100, we generated 1000 random samples from normal distribution and calculated empirical rejection rates, which is used as an estimate of α of the test. In this case, rejection is wrong decision because the data follow normal distribution. Here we provide the values of empirical alpha for α=0.01 and 0.05. For each nominal alpha, we provide four empirical values, each corresponding to each sample size n=20, 30, 50, 100 in Table 4.1. We noticed that D'Agostino D test is the worst, i.e., the biggest difference between empirical alpha and nominal alpha. Other methods, however, show similar performance. Tables including results for other alphas (0.1 and 0.2) are provided in Supplementary Materials I.

Table 4.1 Empirical alphas for two nominal α=0.01, 0.05 at n=20, 30, 50, 100.

	$\alpha = 0.01$				$\alpha = 0.05$			
Test	$n = 20$	$n = 30$	$n = 50$	$n = 100$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
Lilliefors	0.015	0.013	0.005	0.014	0.055	0.045	0.053	0.068
Jarque-Bera	0.016	0.007	0.009	0.009	0.059	0.048	0.053	0.052
D'Agostino D	0.030	0.023	0.013	0.017	0.068	0.058	0.059	0.056
Doornik-Hansen	0.011	0.008	0.011	0.013	0.050	0.043	0.051	0.048
Energy	0.016	0.014	0.012	0.012	0.049	0.050	0.057	0.054
Martinez-Iglewicz	0.013	0.008	0.007	0.008	0.061	0.050	0.055	0.052
Shapiro-Wilk	0.009	0.012	0.009	0.011	0.055	0.053	0.059	0.049
Shapiro-Francia	0.016	0.009	0.010	0.009	0.053	0.058	0.054	0.055
Anderson-Darling	0.018	0.014	0.011	0.014	0.055	0.049	0.054	0.054
Cramer-von-Mises	0.008	0.012	0.011	0.010	0.049	0.067	0.048	0.062
Kolmogorov-Smirnov	0.013	0.015	0.013	0.008	0.048	0.048	0.050	0.040

Figure 4.1 includes plot of rejection rate (empirical alpha) all tests made for different sample sizes.

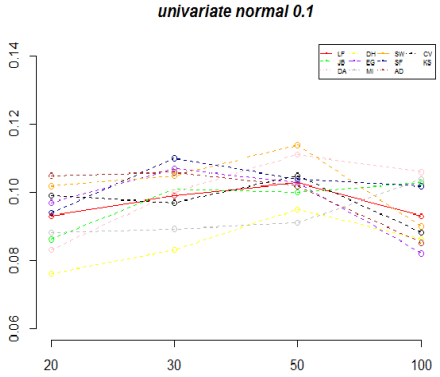

Figure 4.1 Empirical alpha for α=0.1 at different sample sizes (n=20, 30, 50, and 100).

Similar to the cases of α=0.01 and 0.05, we see that all methods control type I error very well for α=0.1.

4.2. Power comparison

We considered eleven different alternatives, which are divided into four groups of distribution: (i) Group 1: symmetric and heavy tail (t distribution and Cauchy distribution), (ii) Group 2: positive support and skewed (χ^2 distribution, exponential distribution, and log-normal distribution), (iii) Group 3: short support (uniform and beta distribution), and (iv) Group 4: bimodal (normal mixture).

For each combination of α and sample size, we generated 1000 random samples and calculated the average of 1000 values of empirical power through this section.

Group 1 distribution Here we consider t distribution ($t(2)$ and $t(5)$) and cauchy distribution. As mentioned before, both distributions have two things in common: symmetric and heavy tail compared to normal distribution.

Figure 4.2 includes three plots, each corresponding to the average of 1000 empirical power of $t(2)$, $t(5)$, and cauchy distribution, respectively. Each method is represented as different line in different color.

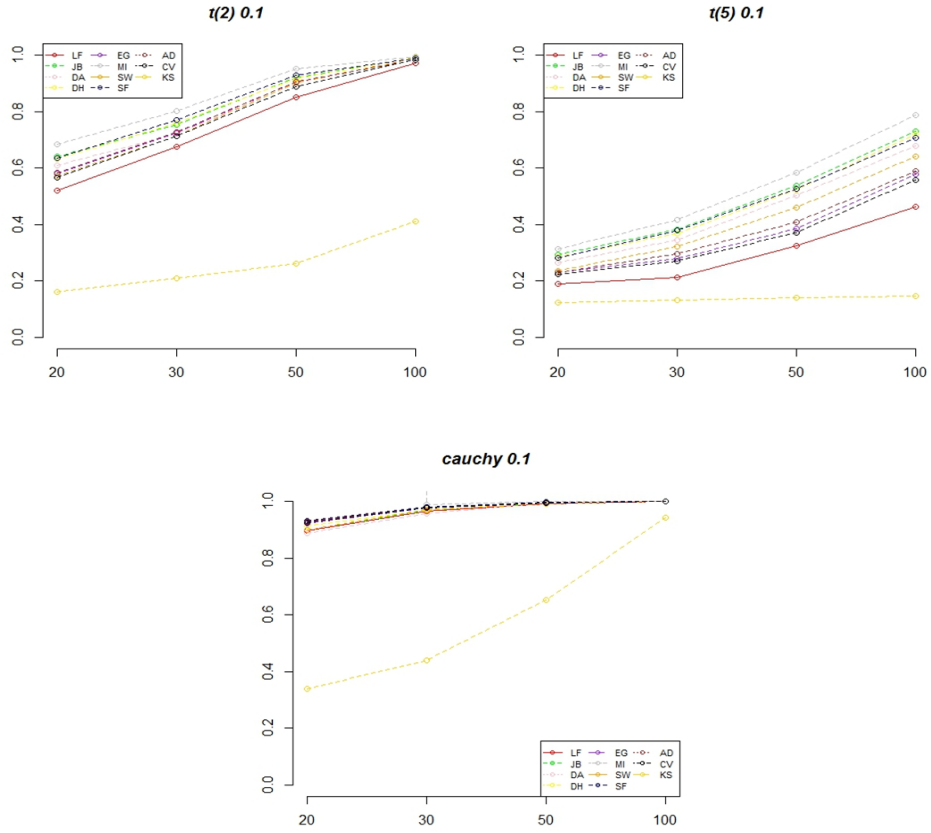

Figure 4.2 Empirical power for $t(2)$, $t(5)$ and Cauchy distribution. Cutoff value of 0.1 is used

All tests except DA (D'Agostino D test) have similar empirical power for Cauchy distribution because the distribution has big difference from the normal distribution in terms of decreasing rate in the tail part of the distribution. More specifically, Cauchy distribution has the decreasing rate of polynomial with degree of 2 while tail part of normal distribution decreases exponentially. That is, it is easy to tell Cauchy distribution from normal distribution. But, since cauchy distribution is symmetric and DA uses skewness information only, it has the worst power. For all tests, it is clear that the closer to the normal distribution, the smaller empirical power each test has. In other words, since $t(5)$ is closer than $t(2)$ to the normal distribution, all test have small empirical power for $t(5)$ distribution.

Group 2 distribution We consider skewed distributions with positive support, i.e., $\mathbf{R}^+$: $\chi^2(2)$, $\chi^2(5)$, exponential, and lognormal distribution. Figure 3 includes four plots, each corresponding to empirical power of $\chi^2(2)$, $\chi^2(5)$, exponential, and lognormal distribution, respectively.

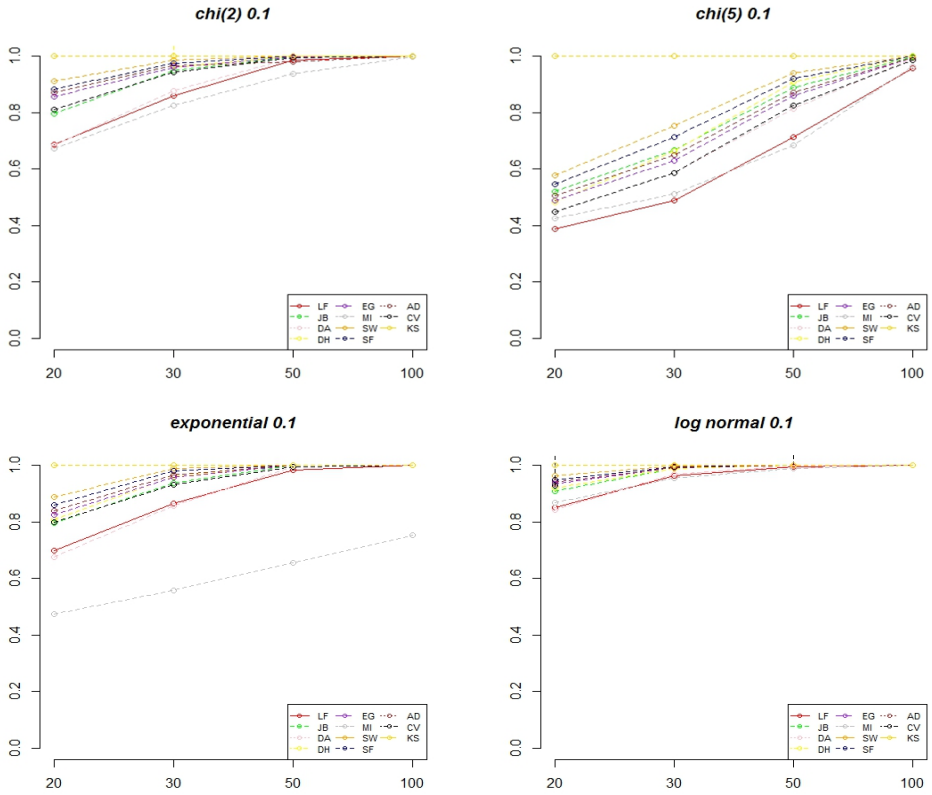

Figure 4.3 Empirical power for $\chi^2(2)$, $\chi^5(5)$, exponential, and lognormal distribution. Cutoff value of 0.1 is used.

For all distribution in Group 2, KS (Kolmogorov-Smirnov) test attains maximum empirical power of 1 and other 10 methods show similar performance except for exponential distribution. For exponential distribution, MI (Martinez-Iglewicz) test has the smallest empirical power compared to other tests.

Group 3 distribution We consider two distributions with support of $[0, 1]$: uniform and beta distribution. Figure 4.4 includes two plots, each corresponding to empirical power of uniform and beta distribution, respectively.

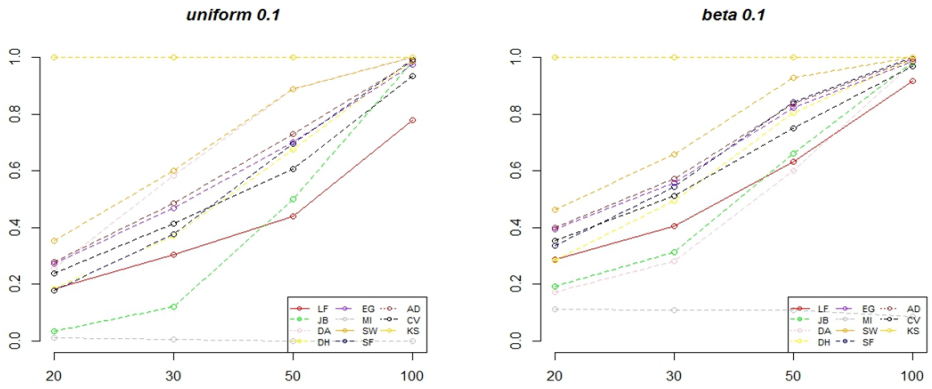

Figure 4.4 Empirical power for uniform (0,1), and beta distribution. Cutoff value of 0.1 is used.

Similar to the results in Group 2 distribution, KS test is the best regardless of sample size. However, the difference from previous results is that the performance of each method is relatively well-separated, ranging from 0 to 1 depending on the sample size. In other words, empirical power is very small for small sample size. But, it is incredibly getting bigger and bigger as the sample size increases.

Group 4 distribution As studied by Bajgier and Aggarwal (1991), we consider a balanced normal mixture as alternative distribution. Figure 4.5 includes the plot of empirical power of the distribution.

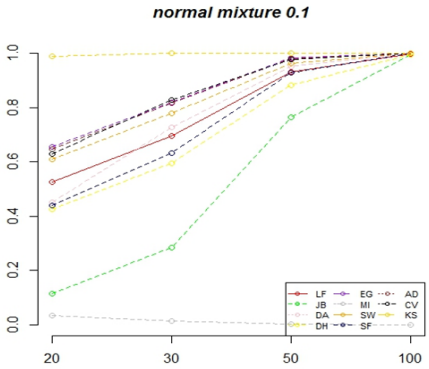

Figure 4.5 Empirical power for normal mixture, $0.5N(-2, 1)+0.5N(2, 1)$. Cutoff value of 0.1 is used.

For bimodal case, MI (Martinez-Iglewicz) test shows the worst performance in terms of empirical power. Furthermore, the empirical power doesn't even increase as sample size

increases. JB (Jarque-Bera) test is the second worst. It, however, shows increasing trend according to sample size. Just like Groups 2 and 3, KS is the best. Other nine tests except MI and JB show similar performance.

For all distributions (Group1 to Group 4), more results including average of 1000 empirical power for all combination of two parameters (α and sample size) are relegated in Supplementary Materials I. Also, corresponding plots are included in Supplementary Materials II.

5. Summary

We considered 11 normality tests and compared them in terms of two things: ability to control type I error and power. Regarding power, through the comprehensive simulation including many different types of distribution (different shape and different support), we found that, for the distribution with big difference in shape from normal distribution, Kolmogorov-Smirnov test showed the best performance because its test statistic consists of the maximum distance in two cumulative distribution functions when testing for normality. In contrast, for the distribution similar to the normal (i.e., similar shape and the same support $(-\infty, \infty)$, Kolmogorov-Smirnov was the worst and Martinez-Iglewicz was the best. Regarding alpha, all tests show similar ability to control type I error with a slight margin less than 0.02. Compared to power, type I error control is less affected by the sample size.

Omnibus test based on sample skewness and kurtosis has a problem. The sample skewness and kurtosis are not independent in finite samples, making the use of asymptotic distribution lead to under-rejection, i.e., making the test conservative. Even though Bowman and Shenton suggested a transformation to approximate χ^2 appropriate for small samples, the procedure is not used very much because of the computational cost.

References

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, **33**, 1148-1159.

Anscombe, F. J. and Glynn, W. J. (1983). Distribution of the kurtosis statistic b_2 for normal samples. *Biometrika*, **70**, 227-234.

Bajgier, S. M. and Aggarwal, L. K. (1991). Powers of goodness-of-fit tests in detecting balanced mixed normal distributions. *Educational and Psychological Measurement Summer*, **51**, 253-269.

Bowman, K. O. and Shenton, L. R. (1975). Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, **43**, 243-250.

Conover, W. J. (1999). *Practical nonparametric statistics*, 3rd Ed., Wiley, New York.

Cramer, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, **1**, 13-74.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika*, **57**, 679-681.

D'Agostino, R. B. and Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distribution of b_2 and $\sqrt{b_1}$, *Biometrika*, **60**, 613-622.

D'Agostino, R. B. and Tietjen, G. L. (1973). Approaches to the null distribution $\sqrt{b_1}$. *Biometrika*, **60**, 169-173.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, **70**, 927-939.

Jarque, C. M. and Bera, A. K. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, **7**, 313-318.

Kang, S. B., Han, J. T. and Cho, Y. S. (2014). Goodness of fit test for the logistic distribution based on multiply type II censored samples. *Journal of the Korean Data & Information Science Society*, **25**, 195-209.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzionc. *1st Itali Attuari*, **4**, 1-11.

Lee, H. Y. (2013). Goodness-of-fit tests for a proportional odds model. *Journal of the Korean Data & Information Science Society*, **24**, 1465-1475.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov tests with mean and variance unknown. *Journal of the American Statistical Association*, **62**, 399-402.

Lilliefors, H. (1969). On the Kolmogorov-Smirnov tests for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, **64**, 387-389.

Martinez, J. and Iglewicz, B. (1981). A test for departure from normality based on a biweight estimator. *Biometrika*, **68**, 331-333.

Pearson, E. S. and Hartely, H. O. (1966). *Biometrika tables for statisticians*, 3rd Ed., Cambridge, New York.

Rahman, M. and Govidarajulu, Z. (1997). A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics*, **24**, 219-236.

Royston, J. P. (1983). A simple method for evaluating the Shapiro-Francia W' test of non-normality. *The Statistician*, **32**, 297-300.

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

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

Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, **67**, 215-216.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for the two independent samples. *Bulletin Mathematique de L'Universite de Moscow*, **2**, 3-14.

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, **19**, 279-281.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**, 730-737.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, **4**, 357-369.

Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. *Biometrika*, **64**, 583-588.

Szekely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Aanalysis*, **93**, 58-80.

von Mises, R. E. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*, Julius Springer, Vienna, Austria.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, **17**, 684-688.