Thailand Statistician
July 2007; 5: 57-68
http://statassoc.or.th
Contributed paper

A Power Comparison of Goodness-of-fit Tests for Normality Based on the Likelihood Ratio and the Non-likelihood Ratio

Rinnakorn Chaichatschwal, and Kamon Budsaba*

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Phathum Thani, 12121, Thailand.

***Author for correspondence**, e-mail: kamon@mathstat.sci.tu.ac.th

Received: 8 November 2006

Accepted: 7 March 2007

Abstract

The goal of the study is to select the best goodness-of-fit test among six tests; the Z_A statistic, the Z_C statistic, the Z_K statistic, the Anderson-Darling (A^2) statistic, the Shapiro-Wilk (W) statistic and the Shapiro-Francia statistic (W'). The tests were compared when the normal parameters are unknown and sample sizes are 10, 30, 50, 70 and 100 each with 0.05 level of significance. With 1,000 Monte Carlo replications, the probability of type I error of all six statistics can be controlled for all sample sizes under study. Both sample sizes and types of the distribution affect the power of the test.

Keyword: Anderson-Darling, Goodness-of-fit, Shapiro-Francia, Shapiro-Wilk.

1. Introduction

Inferential statistics can be categorized into parametric and nonparametric statistics. Most parametric tests are more reliable than nonparametric tests because of the known distribution and having some assumptions to be stated. Normal distribution is also an important assumption in parametric statistics, especially in parameter estimation and hypothesis testing. If the data set is normally distributed and conform with its assumptions, parametric tests generally will have more efficiency and power than nonparametric tests. In order to test whether population follows normal distribution, some graphical methods are plotted and displayed. However the graphical methods might have some risk for miss conclusion since they are subjective tools. Nowadays,

statisticians have proposed various tests for checking the form of a distribution. Those tests can be grouped into likelihood ratio test statistic and non-likelihood ratio test statistic which are more efficient than the visual displays.

There are many test statistics for testing normality. In various scenarios, such as different sample sizes and/or different underline distributions, the power of each statistic might be different. The researcher would like to use the statistic which is the most appropriate powerful test. Therefore, the purpose of this research is to compare the power of the test in different scenarios. Only the following six statistics for normality test are considered in the study, that is the Anderson-Darling statistic (A^2), the Shapiro-Wilk statistic (W), the Shapiro-Francia statistic (W'), the Z_A statistic, the Z_C statistic and the Z_K statistic.

The Z_A, Z_C and Z_K statistics are recently developed for testing normality based on the likelihood ratio. The W and W' statistics are based on the non-likelihood ratio. Originally the W' statistic was proposed for fixing the problem of being unable to extent the test to the sample size of 50 or more of the W statistic. W' and A^2 statistics are the most favored statistics which are familiar in some popular software packages such as SAS, P-Stat and Minitab. The expected utility of this study is to guide a researcher in selecting the appropriate statistics for testing normality in different practical situations.

2. Scope of the study

This study will convert all simulated data to have the population mean equal zero and population variance equal one in order to test the hypotheses as following:

H_0 : The data distribution is a standard normal

H_1 : The data distribution is non-standard normal

The power of the statistics will be calculated if and only if the tests can control the type I error. The statistics will be compared for various choices of sample sizes:

10, 30, 50, 70 and 100 for Z_A, Z_C, Z_K and Anderson-Darling (A^2) tests

10, 30, and 50 for Shapiro-Wilk (W) test

50, 70, and 100 for Shapiro-Francia (W') test

The normal parameters are unknown and the tests will be compared at 0.05 level of significance.

3. Goodness-of-fit tests for normality

Let X be a continuous random variable with distribution function $F(x)$, and $X_1, X_2, .., X_n$ be a random sample from X with order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$. To the test hypothesis

$$H_0 \ : \ F(x) \ = \ F_0(x), \quad \text{for all} \ \ x \in (-\infty, \infty)$$

against the general alternative

$$H_1 \ : F(x) \ \neq \ F_0(x), \quad \text{for some} \ x \in (-\infty, \infty)$$

where $F_0(x)$ is a hypothetical distribution function which is completely specified. If $F_0(x)$ is a family of distribution with unknown parameters, it need to estimate the parameters first and then apply the tests. When testing the goodness-of-fit for the family of normal distributions, μ and σ are estimated by the sample mean $\overline{X} = \sum_{i=1}^{n} X_i / n$ and the sample standard deviation $S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \Big/ (n-1)}$, respectively. The power of the tests also depend on the estimators of μ and σ. Good estimators should induce powerful tests of normality. For normal distribution, as well known $\overline{X}$ and S^2 are the uniformly minimum variance unbiased estimators of μ and σ.

Anderson and Darling [1] proposed the Anderson-Darling statistic (A^2). It is defined as:

$$A^2 \ = \ -n - \frac{1}{n}\left[\sum_{i=1}^{n}(2i-1)\left[\ln F(X_{(i)}) + \ln\left(1 - F(X_{(n-i+1)})\right)\right]\right]$$

which is modified by computing $A^{2*} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$, where the $F(X_{(i)})$ provide cumulative distribution function of the normal distribution and the n provide the sample size.

The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published in

Upton and Cook [2] for a few specific distributions (e.g. normal, exponential). The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A^{2*}, is greater than the critical value.

Shapiro and Wilk [4] proposed the Shapiro-Wilk statistic (W), in 1965, that tests whether a random sample, $X_1, X_2, .., X_n$ comes from (specifically) a normal distribution. Small values of W are evidence of departure from normality and percentage points for the W test statistic, obtained via Monte Carlo simulations. The Shapiro-Wilk statistic is defined as:

$$W = \frac{\left[\sum_{i=1}^{k} a_{n-i+1}\left(X_{(n-i+1)} - X_{(i)}\right)\right]^2}{\sum_{i=1}^{n}\left(X_{(i)} - \overline{X}\right)^2}$$

where the $X_{(i)}$ are the ordered sample values ($X_{(1)}$ is the smallest) and the k is the greatest integer in $n/2$. Shapiro and Wilk [4] gives tabled values that can be used to compute the coefficients (a_i) and the percentage point of the W statistic. The null hypothesis will be rejected if the test statistic, W , is less than the percentage point. It is note that this test statistic can be served for sample as small as 3, or as large as 50.

Shapiro and Francia [3] proposed the Shapiro-Francia statistic (W'), The W' test statistic was modified of the Shapiro-Wilk statistic for testing normality which can be used with large samples.

The Shapiro-Francia statistic is defined as:

$$W' = \frac{\left[\sum_{i=1}^{n} m_i X_{(i)}\right]^2}{\sum_{i=1}^{n} m_i^2 \cdot \sum_{i=1}^{n} (X_{(i)} - \overline{X})^2}$$

where the m_i are expected value of normal order statistics. The null hypothesis will be rejected if the statistic, W', is less than the empirical percentage point of the approximate W' test.

Zhang and Wu [6] proposed the statistics, Z_A, Z_C and Z_K for testing normality. The Zhang-Wu statistics are defined as:

$$Z_A = -\sum_{i=1}^{n}\left[\frac{\log F_0(X_{(i)})}{n-i+0.5}+\frac{\log\left[1-F_0(X_{(i)})\right]}{i-0.5}\right],$$

$$Z_C = \sum_{i=1}^{n}\left[\log\left(\frac{\frac{F_0(X_{(i)})^{-1}-1}{(n-0.5)}}{\frac{}{(i-0.75)}-1}\right)\right]^2$$

and $Z_K = \max_{1\le i\le n}\left[(i-0.5)\log\left[\frac{i-0.5}{nF_0(X_{(i)})}\right]+[n-i+0.5]\log\frac{n-i+0.5}{n\left(1-F_0(X_{(i)})\right)}\right]$

where $F_0(X_{(i)})$ are cumulative distribution function of the normal distribution. The null hypothesis will be rejected if the statistics, Z_A, Z_C and Z_K, are greater than the table of percentage point for Z_A, Z_C and Z_K in Zhang and Wu [6].

4. Simulation study

The simulation study to compare the six statistics has 4 steps.

Step 1 : Find the probability of type I error

- Simulate the standard normal distribution data set by Minitab 14 for windows for testing the null hypothesis

H0: The data distribution is a standard normal

against the alternative

H1 : The data distribution is non-standard normal

- Calculate each statistic, then find the empirical type I error rate collected from 1,000 Monte-Carlo replications.

Step 2 : Test the controllable of the probability of type I error

- Binomial test is used to test the null hypothesis

H0: $\alpha \leq 0.05$

against the alternative

H1: $\alpha > 0.05$. The statistics can control the type I error at 0.05 level of significance if the empirical type I error rate in step 1 is between 0 and 0.061.

Step 3 : Find the power of each test

- Simulate the non-normal distribution data set with Minitab 14 for windows, then standardized the data set to have mean zero and variance one for testing the null hypothesis

H0: The data distribution is a standard normal

against the alternative

H1 : The data distribution is non-standard normal

- Calculate each statistic, then find the empirical power of each test collected from 1,000 Monte Carlo replications.

Step 4 : Compare the powers of the tests

Find the statistic which has the highest power for each scenario.

The Non-normal distributions under study are t-distribution, Lognormal distribution, Beta distribution and Weibull distribution with parameters set by the coefficient of skewness (γ_1) and the coefficient of kurtosis (γ_2). Shapiro et al. [5] proposed the classification of distributions from their skewness and kurtosis into five categories, as the following:

1. Near normal distribution.
2. Symmetric long-tailed distribution.
3. Symmetric short-tailed distribution.
4. Asymmetric long-tailed distribution.
5. Asymmetric short-tailed distribution.

The parameters γ_1 and γ_2 of each non-normal distributions are shown in Table 1.

Table 1. Classification of the distributions under study.

Case	Skewness : γ_1, Kurtosis : γ_2		Distributions used in the study
1	$\gamma_1 = 0$, $2.5 \le \gamma_2 \le 4.5$	Near normal	t(34), t(14), t(10), t(8), Beta(13.5,13.5)
2	$\gamma_1 = 0$, $\gamma_2 > 4.5$	Symmetric long-tailed	t(7), t(6), t(5)
3	$\gamma_1 = 0$, $\gamma_2 < 2.5$	Symmetric short-tailed	Beta(1,1), Beta(1.5,1.5), Beta(2.25,2.25), Beta(3.5,3.5)
4	$\lvert\gamma_1\rvert > 0.3$, $\gamma_2 > 3.0$	Asymmetric long-tailed	Weibull(2.211,1), Weibull(1.563,1), Weibull(1.211,1), Weibull(1,1), Weibull(0.896,1), Lognormal(0,0.0269), Lognormal(0,0.0988), Lognormal(0,0.1967), Lognormal(0,0.3040), Lognormal(0,0.4108), Beta(7,2), Beta(5,1)
5	$\lvert\gamma_1\rvert > 0.3$, $\gamma_2 < 3.0$	Asymmetric short-tailed	Beta(2,1.08), Beta(2.28,5), Beta(2,1), Beta(2,5), Beta(0.5,1), Beta(2,4)

5. Results

The results of this study are as following:

1. The empirical type I error rate at 0.05 level of significance are presented in Table 2. All six statistics, Z_A, Z_C, Z_K, A^2, W and W' can control the type I error at 0.05 significant level. The type I error decreases if the sample size increases as expected except for W' statistic (Figure1).

Table 2. The empirical type I error rate at 0.05 level of significance.

Sample size (n)	The empirical type I error rate					
	Z_A	Z_C	Z_K	A^2	W	W^*
10	0.046	0.045	0.047	0.048	0.045	N/A
30	0.059	0.053	0.048	0.054	0.049	N/A
50	0.056	0.050	0.051	0.057	0.044	0.043
70	0.053	0.055	0.049	0.050	N/A	0.053
100	0.039	0.042	0.048	0.044	N/A	0.060

N/A : Not Applicable

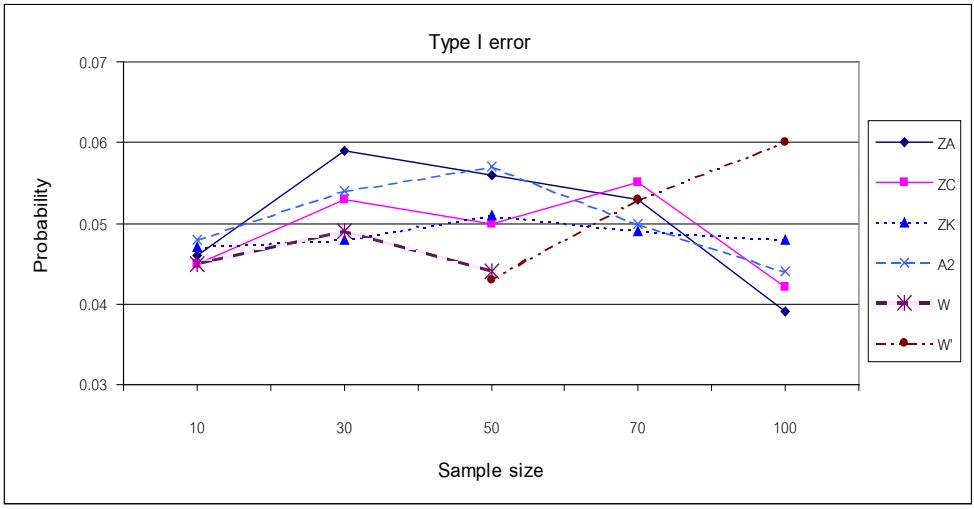

Figure 1. The empirical type I error rate of six test statistics for the sample size n equals 10, 30 50, 70, 100 at 0.05 level of significance.

2. The empirical power of all six statistics are presented in Table 3. The power of Z_A and Z_C statistics are higher than the power of Z_K and A^2 statistics. The powers of the test presented here are the average for all sample sizes and for all distributions. Based on the non-likelihood ratio tests, if the sample size equals 50, the power of W^* statistic is higher than the W statistic for the case of near normal distributions and symmetric long-tailed distributions. On the other hand for other distributions, such as ; (i)

symmetric short-tailed distribution, (ii) asymmetric long-tailed distribution and (iii) asymmetric short-tailed distribution, the power of W statistic is higher than $W^{'}$ statistic.

To compare the power of all scenarios at each specific sample size, the results are as following:

(i): For n = 10, Z_A statistic has the highest power for near normal distribution and for both symmetric and asymmetric long-tailed distributions. Z_C statistic has the highest power for both symmetric and asymmetric short-tailed distributions.

(ii): For n = 30, Z_C statistic has the highest power for near normal distribution and for symmetric both short-tailed and long-tail distributions. Z_A statistic has the highest power for asymmetric both long-tailed and short-tailed distributions.

(iii): For n = 50, $W^{'}$ statistic has the highest power for near normal distribution and symmetric long-tail distribution. W statistic has the highest power for symmetric short-tailed distribution. Z_A statistic has the highest power for asymmetric both long-tailed and short-tailed distributions.

(iv): For n = 70, $W^{'}$ statistic has highest power for near normal distribution and symmetric long-tailed distribution. Z_C statistic has the highest power for asymmetric short-tailed distribution. Z_A statistic has the highest power for asymmetric both long-tailed and short-tailed distributions.

(v): For n = 100, $W^{'}$ statistic has the highest power for near normal distribution and symmetric long-tailed distribution. Z_A statistic has the highest power for symmetric short-tail distribution and asymmetric for both long-tailed and short-tailed distributions.

Notice that the power of the test for all distribution increase if the sample size increases (Figure 2).

Table 3. The empirical power of all six statistics.

	Distribution	Power of test					
		Z_A	Z_C	Z_K	A^2	W	W'
10	Near Normal	0.072	0.066	0.065	0.066	0.062	-
	Symmetric Long-tailed	0.102	0.097	0.086	0.097	0.086	-
	Symmetric Short-tailed	0.039	0.059	0.051	0.057	0.054	-
	Asymmetric Long-tailed	0.413	0.408	0.321	0.384	0.398	-
	Asymmetric Short-tailed	0.118	0.131	0.096	0.121	0.123	-
30	Near Normal	0.094	0.098	0.082	0.084	0.081	-
	Symmetric Long-tailed	0.213	0.219	0.177	0.181	0.192	-
	Symmetric Short-tailed	0.139	0.184	0.106	0.140	0.180	-
	Asymmetric Long-tailed	0.831	0.813	0.793	0.778	0.812	-
	Asymmetric Short-tailed	0.446	0.411	0.377	0.361	0.420	-
50	Near Normal	0.112	0.120	0.104	0.095	0.086	0.133
	Symmetric Long-tailed	0.271	0.297	0.257	0.238	0.214	0.336
	Symmetric Short-tailed	0.312	0.342	0.178	0.240	0.393	0.150
	Asymmetric Long-tailed	0.915	0.899	0.886	0.866	0.897	0.880
	Asymmetric Short-tailed	0.678	0.611	0.604	0.535	0.647	0.507
70	Near Normal	0.113	0.125	0.108	0.096	-	0.156
	Symmetric Long-tailed	0.310	0.347	0.301	0.290	-	0.435
	Symmetric Short-tailed	0.481	0.490	0.306	0.359	-	0.321
	Asymmetric Long-tailed	0.952	0.945	0.938	0.929	-	0.925
	Asymmetric Short-tailed	0.841	0.779	0.770	0.684	-	0.712
100	Near Normal	0.141	0.166	0.136	0.127	-	0.218
	Symmetric Long-tailed	0.395	0.447	0.386	0.373	-	0.537
	Symmetric Short-tailed	0.607	0.605	0.429	0.475	-	0.459
	Asymmetric Long-tailed	0.971	0.960	0.954	0.939	-	0.954
	Asymmetric Short-tailed	0.951	0.909	0.905	0.821	-	0.872

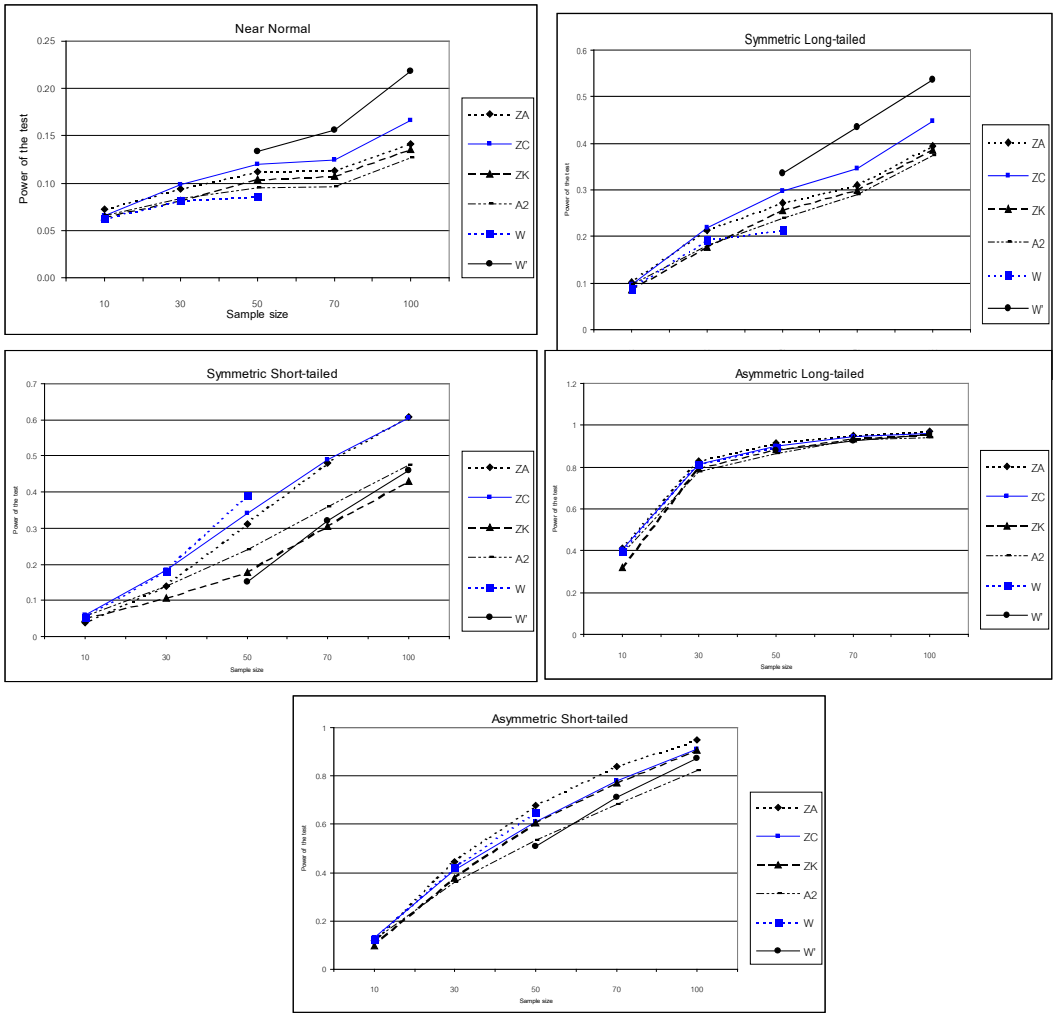

Figure 2. The power of the test of six test statistics at 0.05 level of significance for each underline distribution.

6. Conclusion and discussion

All six tests can control the probability of type I error for all sample sizes under study. The power of Z_A and Z_C statistics are higher than the power of Z_K and A^2 statistics. Based on the non-likelihood ratio tests, if the sample size equals 50, the power of $W^{'}$ statistic is higher than the W statistic for near normal distributions and symmetric long-tailed distributions. For other distributions, such as symmetric short-tailed

distribution and asymmetric for both long-tailed short-tailed distributions, the power of W statistic is higher than W' statistic.

In this study the data sets were simulated with known coefficient of skewness and coefficient of kurtosis. In practical situation, the researcher has never known the shape of the data distribution in hand so some visual displays and some measures of skewness and kurtosis should be provided before choosing the most appropriate goodness of fit test. Furthermore other robust estimators of population mean and variance could be used to see the performance of these six tests and the relationship between skewness, kurtosis and the power of each statistic should be under studied in the future.

References

[1] Anderson, T.W., and Darling, D.A., A Test of Goodness of fit, *Journal of the American Statistical Association,* **49**;1954: 765-769.

[2] Upton, G.H., and Cook, I., A Dictionary of Statistics. 2004.

[3] Shapiro, S.S., and Francia, R.S., An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, **67**;1972: 215-216.

[4] Shapiro, S.S., and Wilk, M.B., An Analysis of Variance Test for Normality, *Biometrika*, **52**;1965: 591-611.

[5] Shapiro, S.S., Wilk, M.B. and Chen, H.J., A Comparative Study of Various Tests for Normality, *Journal of the American Statistical Association,* **63**;1968: 1343-1370.

[6] Jin, Z., and Yuehua, W., Likelihood-ratio Tests for Normality, *Computational Statistics and Data Analysis*, **49**;2005: 709-721.