Sains Malaysiana 40(6)(2011): 637–641

Sensitivity of Normality Tests to Non-normal Data

(Kepekaan Ujian Kenormalan Terhadap Data Tidak Normal)

NOR AISHAH AHAD, TEH SIN YIN*, ABDUL RAHMAN OTHMAN & CHE ROHANI YAACOB

ABSTRACT

In many statistical analyses, data need to be approximately normal or normally distributed. The Kolmogorov-Smirnov test, Anderson-Darling test, Cramer-von Mises test, and Shapiro-Wilk test are four statistical tests that are widely used for checking normality. One of the factors that influence these tests is the sample size. Given any test of normality mentioned, this study determined the sample sizes at which the tests would indicate that the data is not normal. The performance of the tests was evaluated under various spectrums of non-normal distributions and different sample sizes. The results showed that the Shapiro-Wilk test is the best normality test because this test rejects the null hypothesis of normality test at the smallest sample size compared to the other tests, for all levels of skewness and kurtosis of these distributions.

Keywords: Monte Carlo simulation; sample size; sensitivity; tests of normality

ABSTRAK

Dalam kebanyakan analisis statistik, data perlu tertabur secara normal atau menghampiri taburan normal. Empat ujian statistik yang digunakan secara meluas untuk memeriksa kenormalan data adalah Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises dan Shapiro-Wilk. Salah satu daripada faktor yang mempengaruhi ujian-ujian ini ialah saiz sampel. Untuk sebarang ujian kenormalan seperti yang dinyatakan, kajian ini akan menentukan saiz sampel dan ujian-ujian menunjukkan bahawa data tersebut adalah tidak normal. Prestasi ujian-ujian ini dinilai pada pelbagai spektrum data yang tidak normal dan saiz sampel yang berbeza. Keputusan kajian menunjukkan bahawa ujian Shapiro-Wilk adalah ujian kenormalan terbaik kerana ujian ini menolak hipotesis nol bagi ujian kenormalan pada saiz sampel terkecil berbanding dengan ujian-ujian yang lain, untuk semua peringkat kepencongan dan kurtosis setiap taburan.

Kata kunci: Kepekaan; saiz sampel; simulasi Monte Carlo; ujian kenormalan

INTRODUCTION

Prior to using any statistical analyses (e.g. *t*-test, ANOVA, and correlation) it is important to check that any of the 'assumptions' incurred on individual tests are not violated. A common assumption is that the random sample is normally distributed. Normally distributed data have a symmetric bell-shaped curve, which has highest frequency in the middle, with lower frequencies towards the extremes (Gravetter & Wallnau 2000). In many statistical analyses, normality is often conveniently assumed without any empirical evidence or test. Indeed, normality is crucial in many parametric statistical methods. Furthermore, understanding the distribution of data could provide more information on the underlying mechanisms for generating the data (Chambers et al. 1983). When this assumption is violated, the interpretation and inference made be invalid.

There are four statistical tests that are widely used for checking normality, namely, the Kolmogorov-Smirnov test (Kolmogorov 1956; Smirnov 1936), Anderson-Darling test (Anderson & Darling 1952), Cramer-von Mises test (Anderson 1962), and Shapiro-Wilk test (Shapiro & Wilk 1965). These tests are well known for their simplicity and availability in most statistical softwares (e.g. SAS, PASW (formerly SPSS), STATA, Minitab, etc.).

The null hypothesis of normality test state that the data are sampled from a normal distribution. When the *p*-value is greater than the predetermined critical value (α=0.05), the null hypothesis is not rejected and thus we conclude that the data is normally distributed. Sample size is a factor that can influence the outcome of the statistical tests mentioned earlier. For example, the Shapiro-Wilk test requires the sample size to be between 3 to 50 (Shapiro & Wilk 1965). Moreover, Shapiro and Wilk did not extend their test beyond samples size of 50 (D'Agostino 1971).

When a normality test is conducted on any non-normal data especially with small sample sizes, there is a possibility of concluding that the data are normal when in fact they are not. This is because, when the data are few, the test which is usually based upon plotting the empirical cumulative density function (cdf) and the normal cdf tends to form a straight line and thus leading to a conclusion of normality. Therefore, given any test of normality on non-normal distributions, this study intends to determine the sample size at which the tests would indicate that the data are non-normal. Hence, the objective of this study is to determine the sensitivity of rejecting the tests of normality on non-normal data. Monte Carlo simulations were conducted on different non-normal distributions,

ranging from symmetric to skew, and kurtosis ranging from platykurtic (light-tailed) to normal-tailed to leptokurtic (heavy-tailed) distributions.

The rest of the paper is organized as follows. In the second section after the first section of the Introduction, the statistics of normality tests are discussed. The design specifications of the data are described in the third section while the fourth section discusses the results. The conclusion of our study is in the final section.

Statistics for Normality Tests

Normality can be assessed to some extend by obtaining skewness and kurtosis levels which are usually part of the descriptive statistics output. The skewness value provides an indication of departure from symmetry in a distribution. A distribution, or data set, is symmetric if the median divides the left side and the right side into two identical areas. Skewness is measured with the following equation (Kenney & Keeping 1962):

$$\text{Skewness} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^3}{\left(N-1\right)s^3} \tag{1}$$

where $\bar{X}$ is the mean, N is the number of data points, and, s is the standard deviation.

A symmetric distribution has a skewness value of zero. Negative values indicate data that are left skewed and positive values indicate data that are right skewed.

Kurtosis, on the other hand, is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. Kurtosis is measured with the following equation (Miles & Shevlin 2001):

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^4}{\left(N-1\right)s^4} \tag{2}$$

where $\bar{X}$ is the mean, N is the number of data points, and s is the standard deviation.The kurtosis for a standard normal distribution is three. Thus, the kurtosis is redefined as

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^4}{\left(N-1\right)s^4} - 3 \tag{3}$$

so that the standard normal distribution has a kurtosis of zero. If the distribution is perfectly normal, skewness and kurtosis values of zero will be obtained. Positive kurtosis indicates a leptokurtic distribution. The word 'leptokurtic' is derived from the Greek word 'leptos', meaning small or slender. Negative kurtosis indicates a platykurtic distribution. The term 'platykurtic' is derived from the French word 'plat', meaning flat (Miles & Shevlin 2001).

The numerical methods for testing normality compare empirical data with a theoretical distribution. Suppose there are n independent observations $X_1, X_2, \ldots, X_n$ with a common distribution function $F(x) = P(X_i \le x)$. The ordered statistics can be represented by $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. The empirical distribution function (edf), $F_n(x)$which is used to estimate $F(x)$, is defined as

$$F_n\left(x\right) = \begin{cases} 0, & x \le X_{(1)} \\ \dfrac{i}{n}, & X_{(i)} \le x < X_{(i+1)},\ i = 1,\ldots,n-1 \\ 1, & x < X_{(n)} \end{cases} \tag{4}$$

Notice $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. The function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations less than or equal to x, while $F(x)$ is the probability of an observation less than or equal to x.

The Kolmogorov-Smirnov statistic, D, which is defined as

$$D = \sup_x\left[\left|F_n\left(x\right) - F\left(x\right)\right|\right] \tag{5}$$

is an edf statistic because it is a measure of the discrepancy between the edf, $F_n(x)$ and $F(x)$. Kolmogorov (1956) and Smirnov (1936) states that D belongs to the supremum class of the edf statistics.

The random variable $U_{(i)} = F(X_{(i)})$ is computed by applying the probability integral transformation and $U_{(i)}$ follows a uniform distribution between 0 and 1. The Kolmogorov-Smirnov statistic, D, is thus the maximum of D^+ and D^-, that is

$$D = \max(D^+, D^-) \tag{6}$$

where $D^+ = \max_i\left[\frac{i}{n} - U_{(i)}\right]$ is the largest vertical distance between the edf and the distribution function when the edf is larger than the distribution function, and $D^- = \max_i\left[U_{(i)} - \frac{i-1}{n}\right]$ is the largest vertical distance between the edf and the distribution function when the edf is smaller than the distribution function.

The Anderson-Darling statistic, A^2, is defined as

$$A^2 = n\int_{-\infty}^{+\infty}\left[F_n\left(x\right) - F\left(x\right)\right]^2\psi\left(x\right)\mathrm{dF}\left(x\right) \tag{7}$$

where function $\Psi(x)$ weights the squared difference $(F_n(x) - F(x))^2$. The Anderson-Darling statistic (Anderson & Darling 1952) is computed as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[\left(2i-1\right)\log U_{(i)} + \left(2n+1-2i\right)\log\left(1-U_{(i)}\right)\right] \tag{8}$$

where the weight function is taken as $\Psi(x) = [F(x)(1 - F(x))]^{-1}$.

The Cramer-von Mises statistic, W^2, is similar to Anderson-Darling statistic. The weight function here is $\Psi(x) = 1$, instead of $\Psi(x) = [F(x)(1 - F(x))]^{-1}$. The Cramer-von Mises statistic (Anderson 1962) is computed as

$$W^2 = \sum_{i=1}^{n}\left(U_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}. \quad (9)$$

The Shapiro-Wilk statistic, W_n, is the ratio of the best estimator of the variance to the corrected sum of square estimator of the variance, where $0 < W_n \leq 1$ and $7 \leq n \leq 2{,}000$ (Shapiro & Wilk 1965). The statistic is given by

$$W_n = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad (10)$$

where $a' = (a_1, a_2, \ldots, a_n) = m'V^{-1}[m'V^{-1}V^{-1}m]^{-0.5}$, $m' = (m_1, m_2, \ldots, m_n)$ is the vector of expected values of standard normal order statistics, $V = n \times n$ covariance matrix and $X' = (X_1, X_2, \ldots, X_n)$ is a random sample.

Design Specificaton

In this section, we will show that the rejection of the four statistical tests mentioned in the previous section depends on sample size. That is, as the sample size increases, the tests will reject the null hypothesis. Most statistical tests have small statistical power, which is the probability of detecting non-normal data, unless the sample size is large. In this study, we test the sample sizes for normality tests from $n = 5, 6, \ldots, 500$ on different non-normal distributions, ranging from symmetric to skew, and kurtosis ranging from platykurtic (light-tailed) to normal-tailed to leptokurtic (heavy-tailed) distributions (See Table 1).

Data from the first distribution, chi-square (3) was generated directly using the SAS RANGAM function while data for Uniform distribution (the fifth distribution) was generated using the RANUNI function. Data for the other distributions cannot be generated directly. We need to generate a standard normal data first using the RANNOR function and then converted it to the Fleishman distribution (the third distribution), *g*-distribution (the second distribution) and *h*-distribution (the fourth distribution) respectively. The Fleishman (1978) power transformation is of the form

$$Y = a + bZ + cZ^2 + dZ^3, \quad (11)$$

where Z are standard normal variates. Fleishman (1978) provided a table of values for the coefficients, b, c, and d that enable the transformation of the standard normal distribution to a non-normal distribution, with mean zero and variance one, but with different degrees of skewness and kurtosis. The extra coefficient a is easily obtained through the relation $a = -c$, that was a direct result of constraining $E(Y) = 0$. One set of coefficients (b, c, d) was selected from Fleishman (1978) with skewness 0.5 and kurtosis -0.5, which we used in the preceding equation to generate the normal variates to produce skewed platykurtic distribution.

The other two distributions were obtained from the *g*-distribution and *h*-distribution from Hoaglin (1985). The general form of the distribution with constant g is represented by the following equation.

$$Y = \frac{e^{gZ} - 1}{g} e^{\frac{hZ^2}{2}} \quad (12)$$

When g is zero,

$$Y = Ze^{\frac{hZ^2}{2}}. \quad (13)$$

While, when h is zero,

$$Y = \frac{e^{gZ} - 1}{g}. \quad (14)$$

Since sample size is believed to influence the normality test, we would like to determine the sample size at which the normality tests would indicate that the data are not normal for the non-normal data. The non-normal data are generated by Monte Carlo simulation as follows:

1. Generate a data set of sample size n=5 from a χ^2_3 distribution using RANGAM function (SAS Institute, 2004).
2. Repeat Step 1 one thousand times.
3. For each replication in Step 2, perform these normality tests: Kolmogorov-Smirnov, Shapiro-Wilk, Cramer-von Mises and Anderson-Darling tests.
4. If p-value of the normality test is greater than 0.05, then increase the number of count by one (count = count + 1). Initial count has been set equal to 0.
5. Obtain the average Type I error rates by dividing count by 1000 for each normality tests.
6. Repeat Step 1 to Step 5 for $n = 6, 7, \ldots, 500$.

TABLE 1. Skewness and kurtosis coefficients for various types of distributions

Type of distribution	Distribution	Skewness coefficient	Kurtosis coefficient
Skewed leptokurtic	Chi-Square (3)	3	6
Skewed mesokurtic	*g*-distribution (*g*=.5, *h*=0)	1.75	8.9
Skewed platykurtic	Fleishman	0.5	-0.5
Symmetric leptokurtic	*h*-distribution (*g*=0, *h*=.225)	0	154.84
Symmetric platykurtic	Uniform	0	-6/5

7. Repeat this simulation for the other types of non-normal distributions: Skewed mesokurtic, skewed platykurtic, symmetric leptokurtic and symmetric platykurtic.

Results and Discussion

The purpose of this study is to determine at what sample sizes the tests would indicate that the data are non-normal when they are really non-normal. The results for the normality tests with various number of sample sizes on different type of distributions are presented in Table 2. When the *p*-value is less than the significance level (α=0.05), the null hypothesis was rejected in favor of the alternative hypothesis that the data is not normal. The *p*-values in bracket showed the right conclusion of non-normal data corresponding to the sample size.

In Table 2, the pattern of rejection of null hypothesis is the same for all type of distributions. For example, for skewed leptokurtic, which was represented by χ^2_3 distribution shows that Shapiro-Wilk test rejects the hypothesis that data is normal when the sample size is equal to 40. Anderson-Darling test gives the same conclusion at sample size of 48 followed by Cramer-von Mises at sample size of 55. Kolmogorov-Smirnov needs a larger sample size (*n*=77) to indicate that the data are not normal. This pattern of rejection is the same in other non-normal distributions. In conclusion, we can say that Shapiro-Wilk is the best normality test compared to the other tests because this test rejects the null hypothesis of normality at the smallest

TABLE 2. Sensitivity of sample sizes for normality tests of continuous data by $P(E_N|O_{NN})$

	Normality Test	Chi-Square (3) -- Skewed Leptokurtic							
		n = 39	*n* = 40	*n* = 47	*n* = 48	*n* = 54	*n* = 55	*n* = 76	*n* = 77
$P(E_N\|O_{NN})$	Anderson-Darling	0.1120	0.0940	0.0520	(0.0470)	(0.0250)	(0.0240)	(0.0030)	(0.0020)
	Cramer-von Mises	0.1530	0.1330	0.0890	0.0850	0.0510	(0.0490)	(0.0080)	(0.0080)
	Kolmogorov- Smirnov	0.2840	0.2820	0.2110	0.1770	0.1520	0.1480	0.0530	(0.0490)
	Shapiro-Wilk	0.0570	(0.0430)	(0.0180)	(0.0130)	(0.0090)	(0.0090)	(0.0020)	(0.0010)
	Normality Test	*g*-distribution (*g*=.5, *h*=0) -- Skewed Mesokurtic							
		n = 54	*n* = 55	*n* = 65	*n* = 66	*n* = 74	*n* = 75	*n* = 103	*n* = 104
$P(E_N\|O_{NN})$	Anderson-Darling	0.1010	0.0810	0.0590	(0.0490)	(0.0370)	(0.0310)	(0.0100)	(0.0030)
	Cramer-von Mises	0.1390	0.1220	0.0900	0.0750	0.0510	(0.0460)	(0.0110)	(0.0080)
	Kolmogorov- Smirnov	0.2420	0.2570	0.1910	0.1690	0.1280	0.1170	0.0510	(0.0390)
	Shapiro-Wilk	0.0540	(0.0500)	(0.0320)	(0.0210)	(0.0170)	(0.0140)	(0.0010)	(0.0000)
	Normality Test	Fleishman -- Skewed Platykurtic							
		n = 91	*n* = 92	*n* = 129	*n* = 130	*n* = 178	*n* = 179	*n* = 214	*n* = 215
$P(E_N\|O_{NN})$	Anderson-Darling	0.1970	0.2060	0.0570	(0.0480)	(0.0050)	(0.0050)	(0.0040)	(0.0010)
	Cramer-von Mises	0.3610	0.3750	0.1750	0.1740	0.0520	(0.0470)	(0.0200)	(0.0180)
	Kolmogorov- Smirnov	0.5310	0.5360	0.3330	0.3170	0.1550	0.1470	0.0570	(0.0490)
	Shapiro-Wilk	0.0560	(0.0490)	(0.0030)	(0.0020)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	Normality Test	*h*-distribution (*g*=0, *h*=.225) -- Symmetric Leptokurtic							
		n = 129	*n* = 130	*n* = 141	*n* =142	*n* = 153	*n* =154	*n* = 189	*n* = 190
$P(E_N\|O_{NN})$	Anderson-Darling	0.0660	0.0630	0.0520	(0.0480)	(0.0460)	(0.0350)	(0.0230)	(0.0150)
	Cramer-von Mises	0.0870	0.0720	0.0630	0.0600	0.0600	(0.0440)	(0.0300)	(0.0240)
	Kolmogorov- Smirnov	0.1480	0.1350	0.1330	0.1250	0.1090	0.0960	0.0590	(0.0490)
	Shapiro-Wilk	0.0530	(0.0490)	(0.0430)	(0.0390)	(0.0390)	(0.0300)	(0.0180)	(0.0110)
	Normality Test	Uniform -- Symmetric Platykurtic							
		n = 73	*n* = 74	*n* = 99	*n* = 100	*n* = 133	*n* = 134	*n* = 200	*n* = 201
$P(E_N\|O_{NN})$	Anderson-Darling	0.1940	0.1790	0.0670	(0.0500)	(0.0140)	(0.0060)	(0.0000)	(0.0000)
	Cramer-von Mises	0.3350	0.3440	0.1610	0.1610	0.0560	(0.0480)	(0.0030)	(0.0040)
	Kolmogorov- Smirnov	0.5670	0.5800	0.4160	0.4080	0.2250	0.2370	0.0560	(0.0480)
	Shapiro-Wilk	0.0510	(0.0450)	(0.0060)	(0.0070)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

*$P(E_N|O_{NN})$=*P*(predict distribution is normal lobserved distribution is not normal).

sample size. Next is Anderson-Darling, followed by Cramer-von Mises and then Kolmogorov-Smirnov test. Apparently, the level of skewness and kurtosis do not affect the sensitivity of the normality tests.

CONCLUSIONS

The performances of the normality tests, namely, the Kolmogorov-Smirnov test, Anderson-Darling test, Cramer-von Mises test, and Shapiro-Wilk test, were evaluated under various spectrums of non-normal distributions and different sample sizes. The results showed that the Shapiro-Wilk test is the most sensitive normality test because this test rejects the null hypothesis of normality at the smallest sample sizes compared to the other tests, at all levels of skewness and kurtosis. Thus, when the four normality tests are available in a statistical package, we would recommend practitioners to use the Shapiro-Wilk normality to test the normality of data.

ACKNOWLEDGEMENTS

We greatly appreciate the helpful comments of the anonymous referees and editor. Their comments have contributed in the improvement of this article. The work that led to the publication of this paper was funded by the Short Term Grant Scheme of Universiti Sains Malaysia (USM), and supported by the USM Fellowship.

REFERENCES

Anderson, T.W. & Darling, D.A. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Ann. Math. Stat.* 23: 193-212.

Anderson, T.W. 1962. On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Stat.* 33(3): 1148-1169.

Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. 1983. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth International Group.

D'Agostino, R.B. & Stephens, M.A. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.

D'Agostino, R.B. 1971. An omnibus test of normality for moderate and large size samples. *Biometrika* 58(2): 341-348.

Fleishman, A.I. 1978. A method for simulating non-normal distributions. *Psychometrika*. 43: 521-532.

Gravetter, F.J. & Wallnau, L.B. 2000. *Statistics for the Behavioral Sciences*, 5th ed. Belmont, CA: Wadsworth.

Hoaglin, D.C. 1985. *Summarizing Shape Numerically: The g-and-h Distributions*, 461-513. In *Exploring Data Tables, Trends, and Shapes*. D. Hoaglin, F. Mosteller and J. Tukey (Eds), New York: Wiley.

Kenney, J.F. & Keeping, E.S. 1962. *Skewness. §7.10 in Mathematics of Statistics, Pt. 1,* 3rd ed. Princeton, NJ: Van Nostrand.

Kolmogorov, A. 1956. *Foundations of the Theory of Probability*, 2nd ed. Chelsea: New York.

Miles, J. & Shevlin, M. 2001. *Applying Regression and Correlation*. London: Sage.

SAS Institute Inc. 2004. *SAS OnlineDoc® 9.1.2*., SAS Institute Inc., Cary, NC.

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

Smirnov, S. 1936. Beschreibung einer neuen Acartia-Art aus dem Japanischen Meer nebst einiger Bemerkungen über die Untergattung Euacartia Steuer. *Zool Anz.* 114: 87-92.

Nor Aishah Ahad, Teh Sin Yin*, Abdul Rahman Othman & Che Rohani Yaacob
Robust Statisties Computational Laboratory
School of Distance Education
Universiti Sains Malaysia
11800 Minden
Penang, Malaysia

Nor Aishah Ahad
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 Sintok
Kedah, Malaysia

Teh Sin Yin* & Che Rohani Yaacob
School of Mathematical Sciences
Universiti Sains Malaysia
11800 Minden
Penang, Malaysia

Abdul Rahman Othman
Institute of Postgraduate Studies
Universiti Sains Malaysia
11800 Minden
Penang, Malaysia

*Corresponding author; email: syin.teh@gmail.com

Received: 9 June 2010
Accepted: 29 September 2010