TÜBİTAK

ORIGINAL ARTICLE

Turk J Med Sci
2006; 36 (3): 171-176
© TÜBİTAK
E-mail: medsci@tubitak.gov.tr

Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions

Derya ÖZTUNA
Atilla Halil ELHAN
Ersöz TÜCCAR

Department of Biostatistics,
Faculty of Medicine, Ankara University,
06100 Ankara - TURKEY

Background: An important aspect of the "description" of a variable is the shape of its distribution, which tells you the frequency of values from different ranges of the variable. Typically, as most of the statistical tests are based on the normality assumption, a researcher is interested in how well the distribution can be approximated by the normal distribution. Unless there are extreme violations of the normality assumptions, approved statistical tests usually provide accurate results. Although simple descriptive statistics can provide some information relevant to this issue, more precise information can be obtained by performing one of the tests of normality to determine whether the sample comes from a normally distributed population or not.

Aim: Lilliefors corrected Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson and Jarqua-Bera tests were aimed to be compared in terms of Type I error and power of the tests.

Materials and Methods: The simulation was run 1000 times for 23 different sample sizes and for 8 different distributions. Lilliefors corrected Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson and Jarqua-Bera tests were compared in terms of Type I error and power of the tests.

Results: The most powerful results for normal distributions were given by the Jarqua-Bera and for non-normal distributions by the Shapiro-Wilk test.

Conclusions: As it had the lowest Type I error rate, the Jarqua-Bera test was superior for normal and standard normal distributions. For nonnormal distributions, achieving sufficient power at smaller sample sizes, the Shapiro-Wilk was the most powerful.

Key Words: Lilliefors corrected Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson and Jarqua-Bera tests

Dört Farklı Normallik Testinin Farklı Dağılımlar İçin Tip 1 Hata ve Güç Açısından İncelenmesi

Giriş: Bir değişken, sahip olduğu değerlerin frekansına ilişkin bilgi veren dağılım şekli ile tanımlanabilir. İstatistiksel testlerin çoğunluğu genellikle normallik varsayımına bağlı olduğu için, araştırmacı dağılımın normal dağılıma yakın olup olmadığı ile ilgilenir. Normallik varsayımından çok ciddi sapmalar olmadığı sürece, uygun istatistiksel testler genellikle doğru sonuçlar verir. Basit tanımlayıcı istatistikler, normallik ile ilgili önemli bilgiler sağlasa da, daha kesin bilgi, örneklemin normal dağılıma sahip olan bir kitleden çekilip çekilmediği hakkında bilgi veren mormallik testlerinden birisinin uygulanması ile elde edilebilir.

Amaç: Lilliefors düzeltmeli Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson ve Jarqua-Bera testlerinin, Tip I hata ve güç açısından karşılaştırılması.

Yöntem ve Gereç: Simülasyon çalışması 23 farklı örneklem büyüklüğü ve 8 farklı dağılım için 1000 defa uygulanmış ve Lilliefors düzeltmeli Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson ve Jarqua-Bera testleri, Tip I hata ve güç açısından karşılaştırılmıştır.

Bulgular: Normal dağılımlar için Jarqua-Bera testi, normal olmayan dağılımlar için Shapiro-Wilk testi en güçlü testler olarak belirlenmiştir

Sonuç: En küçük Tip I hata oranı nedeniyle, normal dağılım ve standart normal dağılımlar için Jarqua-Bera testi en iyi sonucu vermiştir. Normal olmayan dağılımlar için küçük örneklem büyüklüklerinde yeterli gücü sağlaması bakımından Shapiro-Wilk en güçlü test olarak belirlenmiştir.

Anahtar Sözcükler: Lilliefors düzeltmeli Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino Pearson ve Jarqua-Bera testleri

Received: September 13, 2005
Accepted: January 27, 2006

Correspondence
Derya ÖZTUNA
Department of Biostatistics,
Faculty of Medicine, Ankara
University, 06100
Ankara - TURKEY

dgokmen2001@yahoo.com

Introduction

It is necessary to apply statistical methods in all phases of the study from collecting data to evaluating its results in medical sciences. Although researchers commonly use statistical analyses as computer technology develops, it is known that some of them do not test parametric test assumptions, especially the "normality assumption". This assumption is crucial for the reliability of test results. In statistical package programs, there are several tests for normality. However, the important point is to assess which test should be used under which condition. We consider that this study will be a guide for researchers in medical sciences to decide the most appropriate normality test for their data set.

Many data analysis methods depend on the assumption that data were sampled from a normal distribution. There are several methods in order to see whether or not continuous data are distributed normally. In general, the normality assumption can be evaluated by graphical and test methods. However, graphical methods provide us with some information about the shape of the distribution, but do not guarantee that the distribution is normal and do not test whether the difference between the normal distribution and the sample distribution is significant. Moreover, there are potential problems with normality tests. Because of a small sample size, normality tests have little power to reject the null hypothesis that the data come from a normal distribution. Therefore, small samples always pass normality tests. With large samples, minor deviations from normality may be flagged as statistically significant, even though small deviations from a normal distribution will not affect the results of a parametric test (1). Thus, the best way to decide whether data are normal or not is to evaluate graphs together with an appropriate normality test.

In the literature, the main tests that assess the assumption of normality are the chi-square goodness of fit test, Kolmogorov-Smirnov (K-S) test, Lilliefors corrected Kolmogorov-Smirnov test, Anderson-Darling test, Cramer-von Mises test, Shapiro-Wilk test, D'Agostino skewness test, Anscombe-Glynn kurtosis test, D'Agostino Pearson omnibus test and Jarqua-Bera test.

The aim of this study was to evaluate the performance of the Lilliefors corrected Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino-Pearson and Jarqua-Bera tests, which are commonly used in the SPSS program.

Graphical Methods

Histogram

The simplest and perhaps the oldest graphical display for one-dimensional data is the histogram, which divides the range of the data into bins and plots bars corresponding to each bin, the height of each bar reflecting the number of data points in the corresponding bin. Unfortunately, the way in which histograms depict the distribution of the data is somewhat arbitrary, depending heavily on the choice of bins and bin widths. The histogram graphically summarizes the distribution of a data set such as the center of the data, spread of the data, skewness of the data, presence of outliers, and presence of multiple modes in the data (2).

Stem and Leaf Plot

A stem-and-leaf plot is a variant on histograms that combines the features of a graphic and a table in that the original data values are explicitly shown in the display as a "stem" and a "leaf" for each value. The stems determine a set of bins into which leaves are sorted, and the resulting list of leaves for each stem resembles a bar in a histogram (2).

Boxplot

A boxplot provides an excellent visual summary of many important aspects of a distribution. Tukey developed the boxplot display, based on the 5-number summary (minimum, first quartile, median, third quartile, maximum) of the data (3). Suspected outliers appear in a boxplot as individual points o or x outside the box. If these appear on both sides of the box, they suggest the possibility of a heavy-tailed distribution. If they appear on only one side, they suggest the possibility of a skewed distribution (4).

Each of the aforementioned displays tries to answer the question of how the data are distributed by showing what the data distribution "looks like", but they do not deal with the issue of how the data distribution compares with some theoretical distributions.

Normal Quantile Quantile Plot (Q-Q Plot)

The normal Q-Q plot may be the single most valuable graphical aid in diagnosing how a population distribution appears to differ from a normal distribution. Normal Q-Q plots plot the quantiles of a variable's distribution against the quantiles of the normal distribution. For values sampled from a normal distribution, the normal Q-Q plot

has the points all lying on or near the straight line drawn through the middle half of the points. Scattered points lying away from the line are suspected outliers that may cause the sample to fail a normality test.

Normal Probability Plot (P-P Plot)

A normal probability plot plots observed cumulative probabilities of occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree line will appear when the observed conforms to the normally expected and the assumption of normally distributed error is met.

Test Methods

Kolmogorov-Smirnov (KS) Test

The Kolmogorov Smirnov test is an "empirical distribution function (EDF)" test in which the theoretical cumulative distribution function of the test distribution is contrasted with the EDF of the data (5). The KS test was first proposed by Kolmogorov and then developed by Smirnov. This test compares the cumulative distribution of the data with the expected cumulative normal distribution, and bases its P value on the largest discrepancy.

The test statistic is defined by $D = \sup_x |F_n(x) - F(x, \mu, \sigma)|$, where $F(x, \mu, \sigma)$ is theoretical cumulative distribution function of the normal distribution function, and $F_n(x)$ is the empirical distribution function of the data. Large values of D indicate nonnormality. If the population parameters (μ and σ) are known, then the original KS test can be used. When they are not known they can be replaced by sample estimates (5).

Lilliefors Corrected Kolmogorov-Smirnov Test

The Lilliefors corrected KS test compares the cumulative distribution of data to the expected cumulative normal distribution. This test is different from the KS test because the population parameters that are unknown are estimated, while the statistic is the same. The Table values of the two tests are different, which results in different decisions (6).

Shapiro-Wilk (SW) test

The Shapiro-Wilk test, developed by Shapiro and Wilk, is the most powerful and omnibus test in most situations. In recent years, the SW test has become the preferred test of normality because of its good power properties as compared to a wide range of alternative tests (6). The SW test depends on the correlation between given data and their corresponding normal scores. A significant W statistic causes the researcher to reject the assumption that the distribution is normal (7). The test statistic for this test is

$$W = \frac{\left\{\sum_{i=1}^{n} a_i x_{(i)}\right\}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (1)$$

where $x_{(i)}$ is the *i*-th largest order statistic, $\bar{x}$ is the sample mean, and *n* is the number of observations (6).

D'Agostino-Pearson (DAP) Omnibus Test

The D'Agostino-Pearson test first analyzes data to determine skewness (to quantify the asymmetry of the distribution) and kurtosis (to quantify the shape of the distribution). It then calculates how far each of these values differs from the value expected with a normal distribution, and computes a single P value from the sum of the squares of these discrepancies (1). This test is a combination of the D'Agostino skewness test and Anscombe-Glynn kurtosis test. The DAP test statistic is $K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2)$, where $Z^2(\sqrt{b_1})$ and $Z^2(\sqrt{b_2})$ are the standard normal deviates equivalent to observing $\sqrt{b_1}$(skewness) and b_2 (kurtosis) (5). The K^2 statistic has approximately a chi-squared distribution, with 2 degrees of freedom when the population is normally distributed. For sample sizes $n \geq 8$, a normal approximation that is easily computerized is available (8).

Jarqua-Bera (JB) Test

The Jarqua-Bera test depends on skewness and kurtosis statistics. The JB test statistic is:

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

The T statistic has approximately a chi-squared distribution, with 2 degrees of freedom (9). If the JB test statistic equals zero, it means that the distribution has zero skewness and 3 kurtosis, and so it can be concluded that the normality assumption holds. Skewness values far from zero and kurtosis values far from 3 lead to an increase in JB values.

Table 1. Type I error rates and power of tests for different distributions and sample sizes.

	Normal				Standard Normal				Gamma				Exponential			
Tests/ Sample Size	Type I Error Rate				Type I Error Rate				Test Power				Test Power			
	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB
5	3.8	3.8	-	-	4.8	4.5	-	-	8.9	11	-	-	11.9	15	-	-
6	5.1	4.7	-	-	3.6	3.8	-	-	11.1	12.6	-	-	16.8	21.7	-	-
7	5.9	5.8	-	-	4.6	4.4	-	-	11.2	14.5	-	-	21.6	29.3	-	-
8	5	5	5.3	0.6	5.5	4.5	5.4	0.2	14.1	18.4	18.5	2.6	24	34.6	26.7	5.9
9	4.9	4.9	5.1	0.5	5.6	5.6	6.9	0.4	14.4	19.1	18	3.9	28.7	38.1	31	11
10	5.9	5.6	6.5	0.8	4.7	4.1	5.2	0.3	16.6	23.5	23.1	7.4	30.8	46.7	34	17
11	5.3	4.7	5.1	1.1	6.9	7.3	7.6	3.9	18.4	27.7	23.3	11.1	32.7	51.4	37.3	19
12	4.2	4.7	5.3	1	5	4.5	4.5	1.2	20.6	30.2	25.6	11.8	37.5	54.6	41.6	23.7
13	4.4	3.8	6	1.3	4.7	4.3	4.5	1	22.4	32.9	29.1	15.4	41.1	60.2	44.3	27
14	4.5	5.4	5.9	1.6	3.9	4	4.5	1.4	20.5	34.8	30.1	17	37.6	60.6	44.2	27.2
15	6	4.7	6.1	2	4.6	5	5.3	1.8	23.6	38.8	31.4	18.6	42.4	64.1	47.7	31.1
20	4.2	5	5.9	2.2	3.8	5	6.2	2	31	54.8	42.7	32	58.2	83.3	60.4	49.5
25	5	5.6	6.4	2.9	4.9	4.6	4.6	2.2	37.5	64.4	48.3	39	70.8	91.7	69.5	59.8
30	4.1	5.2	5.2	2.5	4.8	4.7	5.2	2.8	46.4	74.6	54.2	47.8	77.4	96.8	79.4	72.7
35	4.6	5.1	6.5	3.2	5.1	5.6	7	4.3	50.8	84.3	64.6	57.8	86.2	99	85.3	81.9
40	5.6	5.8	6.1	4.1	4.9	4.2	5.4	3.1	59.5	87.6	68.7	63.4	91	99.7	90.2	88.2
45	5.3	5	6.4	4.4	4.9	4.7	5.1	3.1	63.9	91.5	74.8	70.3	93.5	100	93.9	92.5
50	6.1	5.3	2.5	1.2	4.6	5.2	5.5	3.6	69.9	94.6	79	75.5	96.2	99.9	96.8	95.9
75	3.9	3.9	5	3.5	5.2	5.8	5.3	4.1	87	99.7	96.2	95.6	99.9	100	99.9	99.8
100	4.6	3	4	3.2	5.6	6.4	5.1	3.9	95.5	100	99.3	99.3	100	100	100	99.9
150	4.5	4.2	6.4	4.4	5.6	5.5	6	4.4	99.7	100	100	100	100	100	100	100
175	5.3	5.8	6.2	4.4	6	5.6	6.7	5.3	99.9	100	100	100	100	100	100	100
200	6.1	5.2	6.1	4.4	6.3	4.9	5.2	4.7	99.9	100	100	100	100	100	100	100

	T				Beta				Chi-square				Uniform			
Tests/ Sample Size	Test Power				Test Power				Test Power				Test Power			
	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB
5	3.6	3.4	-	-	4.7	5	-	-	4.5	5.2	-	-	3.5	5.2	-	-
6	4.8	4.4	-	-	6.9	6.8	-	-	6.1	6.2	-	-	4.1	6.2	-	-
7	5.1	4.6	-	-	5.6	5.6	-	-	5.6	6.9	-	-	6.1	6.7	-	-
8	4.7	4.7	5.7	0.3	6.7	9.2	8.2	0.2	5.9	6.7	8.7	0.6	5.2	7.3	2.4	0
9	6	6.7	8.1	0.8	7.8	9.2	8.8	1.9	6.4	6.5	7.4	0.8	5.8	7.7	3.3	0.4
10	5	5.7	6.9	1.2	7.2	9.5	9.1	1.4	5.6	7	8.9	1.5	6.6	7.7	2.8	0.2
11	5.8	7	8	1.9	8	9.3	6.7	1.8	6.7	7.5	8.2	1.5	5.9	9.9	1.8	0
12	6.7	5.3	7.2	1.7	7.6	8.7	8	1.8	6.7	7.5	8.4	3.3	6.1	9.9	3	0.2
13	5.5	6.4	7.5	2.3	8.1	9.9	8.5	2.2	6.5	10.6	10.1	4	5.9	11.6	5.4	0.1
14	4.4	5.1	6.1	1.7	9.2	12.1	8	2.8	6.8	9.2	8.3	3.1	8.5	12.8	7.3	0
15	5.1	6.7	7.4	3.6	8.6	10.3	7.9	2.3	6.5	9	9	4.5	8.1	13.8	7.3	0
20	5.4	4.9	5.8	2.5	11.5	16.6	11.7	5.1	7.8	11.6	12.1	6.7	8.9	19.6	14.3	0
25	6.1	6.3	8.8	5.5	12.4	21.9	12.1	5.9	9.8	14	14.3	8.7	10.4	26.9	27.9	0.1
30	5.3	6.3	8.2	5.4	16	29.2	17.4	9.9	9.8	14.8	13.5	9.6	13.1	36	39.2	0
35	5.5	6.1	8.2	5.7	16.9	31.8	17.2	10	11.7	17.1	16.7	12.1	13.8	46.2	53	0.2
40	6	7.7	8.6	7	19.9	41	20.7	12.5	12.1	19.8	16.7	12.9	20	58.3	64.2	0
45	5	5.6	7.7	5.1	21.5	43.9	21.9	12.7	13.9	21.6	17.5	14.3	22.6	65.5	70.4	0
50	4.8	7.1	9.5	7.6	26.8	48.6	22.9	15.2	4	11	12.3	11.5	27.9	75.7	82.4	0
75	5.1	8.1	8.2	8.2	39.2	76.1	40.2	30.1	18.4	35.9	30.8	26.5	41.7	95.1	97.3	7.5
100	5.8	7.8	9.2	8.5	50.4	89	59.2	48.5	22.7	43.7	37.9	34.5	57.3	99.7	99.5	55
150	6	7.9	10.6	12	72.6	99	90.4	86.3	34.4	60.5	54.4	51.6	84.4	100	100	98
175	6.1	8.8	10.7	12	77.2	99.6	96	94.2	39.1	68.9	62.2	61	89.9	100	100	100
200	4.9	8.5	10.4	12	83.1	100	98.7	97.7	44.8	74.6	68.7	66.1	93.7	100	100	100

Materials and Methods

In order to compare the performances of the 4 tests considered above, for each of the samples of sizes n = 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 150, 175 and 200, 1000 samples were generated from Normal (100, 4), Standard Normal (0, 1), Gamma (2, 1), Exponential (1), t (30), Beta (2, 5), Chi-square (30) and Uniform (40, 60) distributions. Random samples were generated by using the functions RV.NORMAL, RV.GAMMA, RV.EXP, RV.T, RV.BETA, RV.CHISQ and RV.UNIFORM in "Transform-Compute Menu" in SPSS for Windows 11.5. Then each pair was compared by the four tests. When samples were taken from normal (0, 1) populations, the number of rejected null hypotheses was declared as the probability for a Type I error. When samples were taken from populations with non-normal distributions, the number of rejected null hypotheses was declared as the test's power. Therefore, to compute the empirical Type I error rate and test power, the program ran each condition 1000 times and recorded the proportion of significant statistics. DAP and JB test statistics were calculated for sample sizes $n \geq 8$.

Results

Empirical results of 1000 simulation runs are given in Table 1. Since the DAP and JB test statistics could not be calculated for sample sizes smaller than 8, 4 tests were compared for 8 or larger sample sizes. For samples of size $n < 8$, no differences were observed between the Lilliefors corrected KS and SW tests for normal, standard normal, t or beta distributions; however, the SW test seemed more powerful for other distributions.

When the distribution was normal, especially for $n \leq 30$, the JB test yielded the smallest Type I error rate, followed by the Lilliefors corrected KS and SW tests.

When the distribution was gamma, SW and DAP were the most powerful tests. Although there was no difference between these tests for sample sizes smaller than 20, for sample sizes larger than 20 the power of the SW test was greater than that of the DAP test. The powers of Lilliefors corrected KS and JB tests were quite low for sample sizes smaller than 35.

For exponentially distributed samples, the SW test was the most powerful for all sample sizes. When the sample size was greater than 30, the DAP and Lilliefors corrected KS tests were similar in power, but for 30 or smaller sample sizes the DAP test was more powerful.

When the distribution was t, none of the tests had enough power to reject the null hypothesis As the t distribution approaches the normal distribution for sample sizes larger than 30, tests were compared in terms of Type I error and the Lilliefors corrected KS test had the best performance.

When the distribution was beta, the SW test was most powerful; for the sample size of 75, it reached a power of 80%. The DAP and Lilliefors corrected KS tests were the most powerful tests after the SW test, but there was no difference between them.

For the chi-square distributed samples, the SW and DAP tests were similar in power, but they had more power compared to the other tests. Their powers were less than 80% even when the sample size was 200. Although the JB test and Lilliefors corrected KS test had similar performance, for sample sizes larger than 50, the Lilliefors corrected KS test had less power.

When the distribution was uniform, the SW and DAP tests had similar power, but were more powerful than the other tests. They reached the power of 80% for 50 or larger sample sizes. The JB test was the weakest one among the 4 tests; there were even samples with zero power.

When the results were evaluated generally, we concluded that for normal distributions the JB and for non-normal distributions the SW test gave the most powerful results.

Discussion

Many statistical tests require the data to be approximately normally distributed. Usually, the first step of data analysis is to test the normality. There are several tests that provide an easy way to test this.

The KS test can be applied to test whether the data follow any specified distribution, not just the normal distribution. As a general test, it may not be as powerful as a test specifically designed to test for normality. Moreover, because of the difficulty in specifying the mean and/or variance beforehand, in practice the Lilliefors corrected KS test is used instead of the KS test.

The SW test is difficult for non-mathematicians to understand, and it does not work well when several values are the same in the data set. In contrast, the DAP omnibus test is easy to understand. Unlike the SW test, this test is not affected if the data contain identical values. The SW and DAP tests are specifically designed to detect departures from normality, without requiring that the mean or variance of the hypothesized normal distribution be specified in advance. These tests tend to be more powerful than the KS test. Furthermore, Monte Carlo simulations studies have indicated that the SW test has good power properties for a wide range of alternative distributions (6). As the measures of skewness and kurtosis are based on the moments of the data, the JB test has zero breakdown. In other words, a single outlier can make the test worthless (9). The JB test is an asymptotic test in which reliability increases with the number of observations.

In conclusion, because it had the lowest Type I error rate, the JB test was superior for normal and standard normal distributions. The SW and Lilliefors corrected KS tests can also be used for practical purposes. For nonnormal distributions, except for the t distribution, achieving sufficient power at smaller sample sizes, the SW was the most powerful. For the t distribution, all of the tests had low and insufficient power.

References

1. http://www.graphpad.com/library/BiostatsSpecial/article_197.htm
2. Armitage P, Colton T. Encyclopedia of Biostatistics, Wiley. New York 1998, Vol: 2, pp: 1759-1760.
3. Tukey JW. Exploratory Data Analysis. Addison-Wesley, Reading.
4. http://www.basic.northwestern.edu/statguidefiles/ttest_unpaired_exam_res.html#Normality%20tests
5. Armitage P, Colton T. Encyclopedia of Biostatistics, Wiley. New York 1998, Vol:4, pp: 3075-3079.
6. Mendes M, Pala A. Type I Error Rate and Power of Three Normality Tests. Pakistan Journal of Information and Technology, 2003; 2: 135-139.
7. http://www2.chass.ncsu.edu/garson/pa765/assumpt.htm
8. D'Agostino RB, Belanger A, D'Agostino Jr. A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician, 1990; 44: 316-321.
9. Brys G, Hubert M, Struyf A. A Robustification of the Jarqua-Bera Test of Normality. COMPSTAT 2004 Symposium, Section: Robustness, 2004.