IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

Comparison of Normality Tests in Terms of Type-I Error and Power with Different Sample Sizes and Distributions

Mustafa Çağatay Büyükuysal[1] Vildan Sümbüloğlu[2]

[1]Zonguldak Bulent Ecevit University, School of Medicine, Department of Biostatistics
e-mail: cbuyukuysal@gmail.com
[2]Sanko University, School of Medicine, Department of Biostatistics
e-mail: vildansumbul@gmail.com

Abstract

The aim of the study was to compare the power and Type-I error rates of 5 different normality tests with different distributions and sample sizes.
For each sample sizes (n=7, 10, 30, 50, 100), 500,000 data sets generated from a normal distribution to test the Type-I error rates of Anderson-Darling, Cramer-von-Mises, Jarque-Bera, Kolmogorov-Smirnov and Shapiro-Wilk tests. Type-I errors of the tests were calculated by computing the rejected null hypothesis over 500,000 samples for each distribution. To evaluate the power of the tests, Chi-square $(\chi_1^2,\ \chi_5^2)$, beta (22, 3), beta (2, 5) gamma (3, 1), weibull (2, 4) and exponential distributed 500,000 data set were generated for each sample sizes. The power values of the tests were calculated by computing the accepted null hypothesis over 500,000 samples for each distribution and sample sizes.
According to the simulation results, the Shapiro-Wilk test provided the best results and Anderson-Darling also showed good results as the sample size increased.
The power values of all the normality tests decreased as sample size decreased. Nevertheless, at that situation after all normality tests we suggest to evaluate the test results with graphical methods.
Key Words: Normal Distribution, Normality tests, Type-I Error, Power

Corresponding Author:
Mustafa Çağatay Büyükuysal, Zonguldak Bulent Ecevit University
School of Medicine Department of Biostatistics, Zonguldak / TURKEY
Phone: +90 555 625 00 55, Fax: +90 372 261 02 64
e-mail: cbuyukuysal@gmail.com

Introduction

All statistical tests are based on the occurrence of certain assumptions such as normality, homogeneity of variance, linearity, independence, and multicollinearity. If any of these assumptions for a statistical test is absent, the results will be incorrect and biased. Most of the powerful significance tests works well under the assumption of normality. Therefore, testing

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

normality is becoming the most important step for statistical analysis and many normality tests have been developed and described in the literature.

Studies on normality test began at the beginning of the 20th century (1). The first study was by K.Pearson in 1900 (2), followed in 1928 by Cramer and von Mises (3), in 1933 by Kolmogorov and Smirnov (4), in 1954 by Anderson and Darling (5), in 1965 by Shapiro and Wilk (6), and in 1987 by Jarque and Bera (7).

a. Kolmogorov-Smirnov test

Kolmogorov-Smirnov test compares the theoretical cumulative distribution function F(x) with the empirical distribution function, which is an estimate of the cumulative distribution function based on the data. If the distribution fit with the specific distribution tested then theoretical and empirical cumulative distribution functions should be close to each other. Arrange the observations in increasing order and label them as $X_{(1)}, X_{(2)},\ldots,X_{(n)}$, so that $X_{(1)}$ and X(n) are the smallest and largest observations, respectively (8). KS statistic is defines as the formula below:

$$D = \underbrace{max}_{1<i<n} \left[F\left(X_{(i)}\right) - \frac{i-1}{n}, \frac{i}{n} - F(X_{(i)})\right] \quad (1)$$

Finally if the of the D statistic is smaller than the critical value at a given level of significance, then we accept the hypothesis that the sample of the data fit to a specific distribution.

b. Anderson-Darling test

Anderson-Darling (A-D) test is a modification of the Kolmogorov-Smirnov (K-S) test which gives more weight to tail and is used to test if a sample of data came from a population with a specific distribution. The Anderson-Darling test is defined with the null hypothesis "the data fit to a specific distribution" and with the alternative hypothesis "the data doesn't fit to a specific distribution".

Test statistic for Anderson-Darling is defines as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln(F_0(Z_i)) + \ln(F_0(Z_{n-i+1}))] \quad (2)$$

Where F_0 (Z_i) shows the cumulative probability value of the standart normal distribution at the point Z_i, Z_i is the standardized convert values of X_i observed values, n is the sample size and i defines the i^{th} observation which calculated the cumulative probability value (9).

c. Cramér-von-Mises test

Cramér-von-Mises test is a nonparametric test based on a cumulative empirical density function and used to test the null hypothesis states that random variables $X_1, X_2,\ldots,X_n$ have a given specific continuous distribution function. The Cramér-von-Mises test statistic W^2 is defined as:

$$W^2 = \sum_{i=1}^{n}\left(F_i(X) - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n} \quad (3)$$

where i=1 to n; $F_n(X)$ is observed empirical distribution function.

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

d. Jarque-Bera test

Most of the normality tests are based on whether testing empirical cumulative distributions with theoretical cumulative distributions or comparing empirical and theoretical values. Unlike the others, Jarque-Bera uses uses kurtosis and skewness for testing if the data set from a sample fits to a specific distribution. Jarque-Bera test statistics is defined as:

$$JB = \left[\frac{(skewness)^2}{6/n} + \frac{(kurtosis)^2}{24/n}\right] \quad (4)$$

e. Shapiro-Wilk test

Shapiro-Wilk is the one of the most common used test among normality tests which is developed in 1965 by Shapiro and Wilk. The W test statistic is calculated as follows by a_i weights generated from means, variances and covarianes of the order statistics of a samples size n from a normal distribution (9).

The equation defined as:

$$W = \frac{(\sum_{i=1}^{n} a_i X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (5)$$

With the improvements in computer and/or software technology in recent decades, the development of various normality tests has been accompanied by simulation studies to compare these tests. In a 2006 study with a 1000 reputation study by Öztuna et. al., the Kolmogorov-Smirnov test was found to have more powerful results if the data fit to any theoretical distribution. Instead of Kolmogorov-Smirnov, the Lilliefors corrected Kolmogorov-Smirnov test was suggested, although the Shapiro-Wilk test performed well in the detection of deviations from normal distribution (10).

In same year (2006), with a 100,000 reputation study by Yazıcı et. al. found that the power of all the tests increased with an increase in sample size, and that the Kolmogorov-Smirnov and Anderson-Darling tests perform better with small sample sizes and symmetric distributions, where as the Shapiro-Wilk test is least powerful with an increase in sample size (1).

In a 2011 study by Razali et. al. with a 50,000 reputation, the Shapiro-Wilk test was reported to be the most powerful test with all sample sizes and distributions and Kolmogorov-Smirnov was the weakest. The Anderson-Darling and Shapiro-Wilk tests showed similar results and the Lilliefors test was much better than the Kolmogorov-Smirnov test (11).

In this study, an evaluation was made of power and Type-I error rates of the five normality tests (Anderson-Darling, Cramér–von Mises, Jarque-Bera, Kolmogorov-Smirnov and Shapiro-Wilk) which which are commonly used and available at statistical softwares are compared for the populations from normal, chi-square, beta, gamma, weibull and exponential distributions and with different samples sizes (n=7, 10, 30, 50, 100). The following sections includes methodologies simulation, comparison results and conclusions.

Type-I error rate is the probability rejecting a true H_0 hypothesis and where power of the test is defined as the probability of rejecting a false H_0 hypothesis. Type-I error and power of the test has a relationship which, one increases the other decreases. Statistical tests are commonly designed to minimize the Type-II error for a fixed Type-I error where α is 0.05 usually (12).

Methods

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

All scenarios about the simulation are summarized in Table-1.

Table 1. Simulation scenarios of the study

Distribution Generated From	Sample Sizes	Reputation Number	Normality Tests	Tested For
Normal	n=7 n=10 n=30 n=50 n=100	500.000	Kolmogorov-Smirnov Shapiro-Wilk Anderson-Darling Cramér-von-Mises Jarque-Bera	Power
Chi-Square χ_1^2 Chi-Square χ_5^2 Beta (22, 3) Beta (2, 5) Gamma (3, 1) Weibull (2, 4) Exponential	n=7 n=10 n=30 n=50 n=100	500.000	Kolmogorov-Smirnov Shapiro-Wilk Anderson-Darling Cramér-von-Mises Jarque-Bera	Type-I Error

First, a normal distributed 500.000 data sets were generated for each sample sizes (n=7, 10, 30, 50, 100) were taken to test Type-I error of the five normality tests stated at the end of introduction section. Type-I errors of the tests were calculated by computing the rejected null hypothesis over 500,000 samples for each distribution. The related results are summarized in Table 2.

In the second part of the simulation study, the 5 normality tests were applied to non-normal distributed populations to test the power of the tests. For this purpose, Chi-square (χ_1^2, χ_5^2), Beta (22, 3), Beta (2, 5) Gamma (3, 1), Weibull (2, 4) and Exponential distributed 500,000 data set were generated for each sample sizes (n=7, 10, 30, 50, 100). The power values of the tests were calculated by computing the accepted null hypothesis over 500,000 samples for each distribution and sample sizes. The related results are summarized in Tables 3 - 9.

The simulation study was applied using the SAS Software 9.4. with "RAND" function which uses a Mersenne-Twister random number generator. The Mersenne-Twister random number generator was developed by Matsumoto and Nishimura in 1998 with the period 2^{19937}-1 and 623-dimensional equidistribution property, and seems to be the best among all the generators ever implemented (13). The name Mersenne Twister is derived from the fact that it uses a period which is a Mersenne prime. It is a modification of a Twisted Generalized Feedback Shift Register(TGFSR) which takes in an incomplete array to realize a Mersenne prime as its period and uses an inversive-decimation method for primitivity testing of a characteristic polynomial of a linear recurrence with a computational complexity of $O(p^2)$ where p is the degree of the polynomial (14).

Results

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

The Type-I error rates (Table-2) and power of the related tests (Table-3-9) for different theoretical distributions are presented in the tables.

Table-2. Type-I errors of normality tests for normal distribution

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.0502	0.0504	0.0500	0.0501	0.0505
Cramér-von-Mises	0.0470	0.0488	0.0503	0.0502	0.0505
Jarque-Bera	0.0000	0.0047	0.0266	0.0334	0.0393
Kolmogorov- Smirnov	0.0527	0.0507	0.0493	0.0496	0.0512
Shapiro-Wilk	0.0503	0.0495	0.0505	0.0503	0.0496

Table-3. Power of normality tests for χ_1^2

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.4845	0.6976	0.9983	1.0000	1.0000
Cramér-von-Mises	0.4541	0.6619	0.9954	1.0000	1.0000
Jarque-Bera	0.0000	0.2254	0.9137	0.9983	1.0000
Kolmogorov- Smirnov	0.3752	0.5357	0.9823	0.9999	1.0000
Shapiro-Wilk	0.5100	0.7334	0.9996	1.0000	1.0000

Table-4. Power of normality tests for χ_5^2

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.1255	0.1851	0.5549	0.8043	0.9871
Cramér-von-Mises	0.1150	0.1699	0.4952	0.7367	0.9679
Jarque-Bera	0.0000	0.0424	0.3842	0.6661	0.9727
Kolmogorov- Smirnov	0.1107	0.1453	0.3810	0.5918	0.8931
Shapiro-Wilk	0.1300	0.1982	0.6450	0.8885	0.9979

IJBCS
ISSN: 2147-1428

Original Article

International Journal of Basic and Clinical Studies (IJBCS)
2021; 10(2): 57-65 Büyükuysal MÇ and Sümbüloğlu Vildan

Table-5. Power of normality tests for Beta (22, 3)

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.0878	0.1188	0.3353	0.5439	0.8730
Cramér-von-Mises	0.0802	0.1097	0.2931	0.4740	0.8023
Jarque-Bera	0.0000	0.0205	0.2133	0.4083	0.8178
Kolmogorov- Smirnov	0.0814	0.0992	0.2281	0.3652	0.6595
Shapiro-Wilk	0.0902	0.1261	0.4096	0.6622	0.9536

Table-6. Power of normality tests for Beta (2, 5)

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.0690	0.0849	0.2241	0.3936	0.7572
Cramér-von-Mises	0.0628	0.0796	0.1960	0.3323	0.6519
Jarque-Bera	0.0000	0.0084	0.0717	0.1485	0.5101
Kolmogorov- Smirnov	0.0664	0.0755	0.1575	0.2551	0.5055
Shapiro-Wilk	0.0707	0.0884	0.2727	0.5005	0.8966

Table-7. Power of normality tests for Gamma (3, 1)

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.1112	0.1602	0.4766	0.7199	0.9644
Cramér-von-Mises	0.1020	0.1471	0.4226	0.6492	0.9307
Jarque-Bera	0.	0.0355	0.3327	0.5924	0.9428
Kolmogorov- Smirnov	0.0994	0.1286	0.3254	0.5121	0.8258
Shapiro-Wilk	0.1151	0.1710	0.5624	0.8203	0.9912

Table-8. Power of normality tests for Weibull (2, 4)

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.0664	0.0805	0.1871	0.3105	0.6115
Cramér-von-Mises	0.0612	0.0751	0.1640	0.2625	0.5114
Jarque-Bera	0.0000	0.0100	0.1043	0.2006	0.5005
Kolmogorov- Smirnov	0.0650	0.0720	0.1349	0.2067	0.3945
Shapiro-Wilk	0.0676	0.0835	0.2356	0.4137	0.7898

Table-9. Power of normality tests for Exponential

	n=7	n=10	n=30	n=50	n=100
Anderson-Darling	0.2642	0.4137	0.9338	0.9968	1.0000
Cramér-von-Mises	0.2433	0.3829	0.8977	0.9904	1.0000
Jarque-Bera	0.0000	0.1135	0.7098	0.9526	1.0000
Kolmogorov- Smirnov	0.2095	0.3023	0.7808	0.9605	0.9999
Shapiro-Wilk	0.2789	0.4449	0.9673	0.9995	1.0000

First, Type-I error rates of the 5 normality tests (Anderson-Darling, Cramér-von-Mises, Jarque-Bera, Kolmogorov-Smirnov and Shapiro-Wilk) were given in Table 2. According to the

simulation results, the lowest Type-I error rate (0%) of the normality tests was achieved for Jarque-Bera test when sample size is 7 (n=7). This result may suggest that Jarque-Bera is much more reliable for a sample size of 7. Such a low Type-I error will increase the Type-II error, therefore the test will be not be usefull. The expected result for this situation is for all normality tests to have a close Type-I error rate of 5%. In addition, the Type-I error rates of other 4 normality tests were close to 5% with the Anderson-Darling and Kolmogorov-Smirnov tests much closer to the expected value of 5% (5.04% and 4.95% respectively).

When n=30, the Jarque-Bera test had the lowest Type-I error rate (2.66%) and the other 4 normality tests had an approximately 5% Type-I error rate. When sample size was increased to 50 and 100 respectively, the Anderson-Darling, Cramér-von-Mises, Kolmogorov-Smirnov and Shapiro-Wilk tests had similar Type-I error rates with the lowest Type-I error rate for Jarque-Bera, despite the increased sample size.

The power comparisons of the normality tests are presented in Tables 3-9. According to the simulation results, for n=10 with all non-normal distributed samples, the power of all 5 normality tests was found to be very low. For n= 30, which is recommended in literature as a critical value for normality, for all the types of distributions studied in this paper, the Shapiro-Wilk test had better results than the other 4 normality tests but was still not at the expected and desired power level for a normality test for some distributions. For example, the Shapiro-Wilk test showed very powerful results for the samples χ_1^2 and exponential distributed samples (99.96% and 96.73% respectively). However, the power of the Shapiro-Wilk test, which was the strongest test for n=30, for χ_5^2 , Beta (22, 3), Beta (2, 5), Gamma (3, 1) and Weibull (2, 4) was 64.5% , 40.96%, 27.27%, 56.4% and 23.56% respectively. In addition, the Anderson-Darling was the second strongest test after Shapiro-Wilk, whereas the Kolmogorov-Smirnov and Jarque-Bera tests were the weakest.

The power of all normality tests increased with sample size increased from 30 to 50, but the Shapiro-Wilk test was still the strongest test with power values (100%, 88.85%, 66.22%, 50.05%, 82.03%, 41.37%, 99.95%) for all distributions respectively. After the Shapiro-Wilk test, Anderson-Darling was the strongest test with power values of 100%, 80.43%, 54.39%, 39.36%, 71.99%, 31.05%, 99.68% for all distributions respectively. In the results for n=30, Kolmogorov-Smirnov was the weakest of the normality tests. For n=100 the Shapiro-Wilk test was still the strongest of all normality tests and Kolmogorov-Smirnov was the weakest.

Discussion

In this study, it was to compare to compare normality tests under different conditions. Type-I error and power values of the tests were calculated for different sample sizes and distributions. According to the results for Type-I errors of the normality tests, the Type-I error rate of the Jarque-Bera test was too low. When the sample size was increased from 10 to 100, the Type-I error rate increased from 0.47% to %3.93 which was not sufficient and therefore the Jarque-Bera test should not be used to test normality in these situations. When the sample sizes were increased to 500 and 1000, the Type-I error rates increased 4.62% and 4.82% respectively.

This result shows that even with a sample size over 500, the Jarque-Bera test hardly reached the theoretical expected value of 5%. Therefore, the Jarque-Bera test is not recommended for small samples. When n=30, the Type-I error of the Anderson-Darling test was 5% and the other tests had approximately the same results which are close to 5% with the Kruskal Wallis test most distant at (%4.93). In this situation, the Anderson-Darling test can be recommended for normality testing. When the samples sizes were increased to 50 and 100, the Kolmogorov-Smirnov test showed the worst result, followed by the Anderson-Darling, Cramér-von-Mises and the best results determined in the Shapiro-Wilk test with close Type-I error rates.

According to the simulation results for the power values of normality tests, for a sample size of 10, the power results of all normality tests were too low. With an increase of sample size from 10 to 100, the power of the tests also increased and for all types of distribution and sample sizes the Shapiro-Wilk test had the best power results. This was followed by the Anderson-Darling test then the Cramér-von-Mises, Kolmogorov-Smirnov and Jarque-Bera tests respectively.

Conclusion

In contrast to the results of this study, Berna Yazıcı et. al. (2007) found that Shapiro-Wilk had low power results with an increase in sample size and under different distributions such as Beta, Gamma and Weilbull (1). Ahmad & Khan (2015) found with the exception of alternate logistic distribution, the Shapiro-Wilk and Shapiro-Francia tests had better results in almost all cases (12). Chanmi et. al. (2016) simulated 11 different normality test according to type-I error and power with different distributions. Kolmogorov-Smirnov test found the best performed test for the distributions with big difference in shape from normal distribution. In contrast to this result Kolmogorov-Smirnov found the worst for the distributions similar to normal (15). The simulation results of the current study showed that in all the normality tests except the Jarque-Bera test, the Type-I error rates increased with sample size and were close to the expected value. The Type-I error rate of the Jarque-Bera test was too far from the expected value even at sample sizes of 500 and 1000 and should therefore not be recommended to test normality under any conditions. However, in all types of distributions and sample sizes, the Shapiro-Wilk test had the best results followed by the Anderson- Darling test.

References

1. Yazıcı B, Yolacan S. A comparison of various tests of normality. J Stat Comput Simul. 2007; 77(2): 175-183.
2. Barnard GA. “Introduction to Pearson (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling”. Breakthroughs in Statistics, pages 1-10, Springer New York, 1992.
3. Sandor C, Faraway JJ. “The exact and asymptotic distributions of Cramer-von Mises Statistics”. Journal of the Royal Statistical Society, 58 (1), 221-234.
4. Facchinetti S. “A procedure to find exact crıtıcal values of kolmogorov-smırnov test”. Statistica Applicata – Italian Journal of Applied Statistics.2009; 21(3-4).

5. Anderson TW, Darling DA."A test of goodness of fit". Journal of the American Statistical Association. 1954; 49 (268), 765-769.
6. Shapiro SS, Wilk MB. "An analysis of variance test for normality". Biometrika. 1965; 52(3-4), 591-611.
7. Jarque CM, Bera AK, "A test of normality of observations and regression residuals". International Statistical Review. 1987; 55(2), 163-172.
8. Neil J. Salking. Encyclopedia of Measurements ans Statistics. SAGE publications, 2007.
9. National Institute of Standards and Technology (NIST) and SEMATECH. Engineering Statistics Handbook, Springer, 2003.
10. Öztuna D, Elhan AL, Tüccar E. "Investigation of four different normality tests in terms of type-1 error rate and power under different distributions". Turkish Journal of Medical Sciences. 2006; 36(3), 171-176.
11. Razali NM, Wah YB. "Power comparion of Shapiro Wilk, Kolmogorov Smirnov, Lilliefors, and Anderson Darling tests". Journal of Statistical Modelling and Analytics. 2011; 2(1), 21-33.
12. Arnastauskaitė J, Ruzgas T, Bražėnas M. An Exhaustive Power Comparison of Normality Tests. Mathematics. 2021; 9(7).
13. Ahmad F, Khan RA. Power Comparison of various normality tests". Pakistan Journal of Statistics and Operation Research. 2015; 11(3).
14. Lilliefors HW. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown". Journal of the American Statistical Association. 1967; 62(318), 399-402.
15. Chanmi L, Suhwi P, Jaesik J. Comparison of various types of normality tests. Journal of The Korean Data & Information Science Society. 2016; 27(5).