

Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation based Small Samples

Siddik Keskin

Department of Biostatistics, Faculty of Medicine, Yuzuncu Yil University, Van, Turkey.

Abstract: Several tests, such as Kolmogorov-Smirnov, Chi-square, Shaphiro-Wilk and D'Agostino-Pearson are available for testing the normality of data. It may be difficult to assess Type I error rate and power of the tests under the alternative hypothesis due to the many possibilities of choosing a particular alternative hypothesis, especially with increased sample size. Thus, an alternative way of simulating Type I error rate and power of the normality tests was used in this study. The third (g_1) and fourth sample moment (g_2), Shaphiro-Wilk (SW) and D'Agostino-Pearson (DP) tests were examined for Type I error rate and power of the test. In several sample sizes, 100.000 replications were done and performance of all tests were assessed 5% level for normal distribution and different types of departures from normality. As a result, Shaphiro-Wilk test was found as a decent performance test in considered conditions.

Key words: Shaphiro-Wilk test, D'Agostino-Pearson test, kurtosis, skewness

INTRODUCTION

Normality test is important for some theoretical and empirical research. The validity of several parametric statistical inference procedures depends on the underlying distributional assumptions. Several parametric statistical tests assumed to be distribution of data is normal. Iman^[6] reported that graphical methods could be used for testing normality. However these methods may failure to determine whether distribution of data is normal in some conditions. Therefore, several normality tests, such as Kolmogorov-Smirnov, Chi-square, Shaphiro-Wilk, D'Agostino-Pearson may be preferred. D'Agostino *et al.*^[4] suggested that Shaphiro-Wilk (SW), third sample moment (g_1), fourth sample moment (g_2) and D'Agostino-Pearson (DP) are excellent tests. SW and DP have good properties for non-normal distributions; g_1 and g_2 have excellent properties for detecting non-normality associated with skewness and kurtosis respectively. Also, same authors emphasized that Chi-square and Kolmogorov-Smirnov tests have poor properties and should not be used when testing for normality.

In this study, performance of the third (g_1) and fourth sample moment (g_2), Shaphiro-Wilk (SW) and D'Agostino-Pearson (DP) tests were examined in point of view Type I error rate and power of the test in simulation based small samples.

MATERIALS AND METHODS

Skewness (g_1) test: In skewness test, a null hypothesis is H_0 : normality versus alternative; H_1 : non-normality

due to skewness. Skewness statistic is computes as follows:

$$g_1 = k_3 / \sqrt{(S^2)^3}, \quad k_3 = \frac{n \sum (X_i - \bar{X})^3}{(n-1)(n-2)}, \quad \sqrt{b_1} = \frac{(n-2)g_1}{\sqrt{n(n-1)}}$$

$$A = \sqrt{b_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}}, \quad B = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$C = \sqrt{2(B-1)-1}, \quad D = \sqrt{C}, \quad E = \frac{1}{\sqrt{\ln D}}, \quad F = \frac{A}{\sqrt{\frac{2}{C-1}}}$$

$$Zg_1 = E \ln (F + \sqrt{F^2 + 1})^{[9]}.$$

Kurtosis (g_2) test: In kurtosis test, a null hypothesis is H_0 : normality versus alternative; H_1 : nonnormal due to kurtosis. Kurtosis statistic is computes as follows:

$$g_2 = k_4 \sqrt{S^4},$$

$$k_4 = \frac{\sum (X_i - \bar{X})^4 n(n+1)/(n-1)-3[\sum (X_i - \bar{X})^2]^2}{(n-2)(n-3)}$$

$$G = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}, \quad H = \frac{(n-2)(n-1)|g_2|}{(n+1)(n-1)\sqrt{G}}$$

$$J = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}}$$

$$K = 6 + 8/J \left[2/J \sqrt{1+4/J^2} \right]$$

$$L = \frac{1 - \frac{2}{K}}{1 + H \sqrt{\frac{2}{K-4}}}, \quad Z_{g2} = \frac{1 - \frac{2}{9K} - \sqrt[3]{L}}{\sqrt{\frac{2}{9K}}} \quad [9]$$

D’Agostino-Pearson test: D’Agostino-Pearson (DP) test statistic combines g_1 and g_2 to produce omnibus test of normality. By omnibus test, it is possible to detect deviation from normality due to either skewness or kurtosis. Test statistic is

$$DP = Z_{g1}^2 + Z_{g2}^2$$

Where, Z_{g1}^2 and Z_{g2}^2 are skewness and kurtosis test statistics, respectively. DP statistic has approximately a Chi-squared distribution with 2 degrees of freedom^[2,4].

Shapiro-Wilk test: Shapiro-Wilk test developed by Shapiro and Wilk is a powerful and omnibus test^[4]. This test calculates a W statistic that tests whether a random sample, x_1, x_2, \dots, x_n comes from (specifically) a normal distribution. Small values of W are evidence of departure from normality. The W statistic is defined as the ratio of the square linear combination of the ordered sample to the usual sum of squares of deviations from mean^[8]. The W statistic is calculated as follows:

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

Where the $x_{(i)}$ are the ordered sample values ($x_{(1)}$ is the smallest) and the a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution.

Random numbers were generated from standard normal, Student t [t(3)], Chi-square [c²(10)] and Beta [b(3,1.5)] distributions by using of IMSL subroutines and written by FORTRAN program^[1]. Because of having different means and variances, generated numbers were standardized for all distributions except standard normal.

Sample sizes (n) were determined as; 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125 and 150 in simulated samples. For each sample size, 100.000 replications were done and performance of all tests were

assessed 5% level. Type I error rates for standard normal distribution and power of the test other distributions were examined in each sample sizes. The proportion of rejected null hypotheses was considered as type I error rate for standard normal and power of the test other distributions.

RESULTS AND DISCUSSION

The empirical results of simulations are presented in Table 1.

Standard normal distribution: Performance of DP and SW for Type I error rate was found better than other two tests (Table 1). Type I error rate of g_2 test was a bit higher than 5 % and that of g_1 test was lower than 5% level. Thus, g_2 test tend to over-reject, while g_1 test under-reject in all sample sizes. However, type I error rate of DP and SW tests were around 5% level in all sample sizes.

t(3) distribution: When distribution was t(3), power of the all tests (proportion of rejected null hypotheses) ranged from 4.15 to 39.76%. Only g_2 test could reach small power (39.75%) when sample size = 125. g_1 test showed less power than other tests when sample size < 60. Power of SW test were lower in 60 and larger sample sizes. Therefore, it can be seen that all tests have low power for all sample sizes.

Chi square c² (10) distribution: In the case of c² (10), g_1 and SW tests had 80 % and larger power when sample size > 40. So, performance of these two tests appear adequate for sample size > 40. On the other hand, g_2 test could only reach 76.28% power even sample size = 150. DP test was also comparable to SW test for larger sample sizes. It was concluded that sample size larger than 100 has no significant effect on the power of the g_1 and SW tests in c²(10) distribution.

Beta b (3, 1.5) distribution: For b(3, 1.5) distribution, g_1 test was clearly biased and less powerful than other tests. The most powerful test was SW. SW test reached sufficient power with 60 and larger sample sizes. So, it can be seen that SW statistic have very good power especially when the sample size > 60.

Although, D’Agostino *et al.*^[4] point out that g_1 and g_2 tests are excellent and powerful test, it was found that performance of these tests was not adequate some conditions, especially in the case of b(3, 1.5) distribution. DP test was found as biased in the case of Beta distribution and small sample size in c² (10) distribution. This result is consistent with the results reported by Filliben^[5], Cho and Im^[3], Mendes and Pala^[7]. SW test showed good performance in most conditions.

Performance of all tests was found similar in normal distribution. For t(3) distribution, none of the test could

Table 1: Type I error rate and power of the tests for different distribution and sample sizes

n	test	Normal (0,1) Type 1 Error	t (3) Power of test	c ² (10) Power of test	b (3, 1.5) Power of test
10	g ₁	.02533	.04147	.19143	.00171
	b ₂	.06690	.10781	.17073	.05612
	DP	.04567	.07819	.15665	.04632
	SW	.05028	.07499	.20079	.08436
15	g ₁	.03471	.06525	.35339	.00066
	b ₂	.06148	.12584	.21221	.04320
	DP	.04823	.10357	.24055	.04941
	SW	.04796	.08481	.31283	.11599
20	g ₁	.04068	.07915	.49784	.00027
	b ₂	.06031	.14422	.25837	.03776
	DP	.05028	.12438	.32486	.05505
	SW	.04848	.09648	.43881	.16929
25	g ₁	.04577	.09245	.61265	.00011
	g ₂	.06140	.16328	.29704	.03351
	DP	.05224	.14311	.40365	.05905
	SW	.05041	.10795	.55156	.23118
30	g ₁	.04996	.09988	.70980	.00010
	g ₂	.06209	.18003	.33646	.03231
	DP	.05449	.15866	.48109	.06646
	SW	.05012	.10852	.64718	.29518
35	g ₁	.03858	.09351	.74274	.00001
	g ₂	.05973	.19838	.36858	.03354
	DP	.04628	.16205	.51315	.05765
	SW	.04987	.11128	.72861	.37805
40	g ₁	.04334	.10120	.80944	.00002
	g ₂	.06005	.21243	.39876	.03404
	DP	.04648	.17516	.57819	.06382
	SW	.04829	.11416	.80100	.46534
45	g ₁	.04586	.10741	.86018	.00001
	g ₂	.06150	.22618	.42833	.03843
	DP	.04871	.18821	.64144	.07499
	SW	.04997	.11419	.85508	.55041
50	g ₁	.04659	.11299	.90030	.00000
	g ₂	.06122	.23891	.45525	.04138
	DP	.04829	.20094	.70145	.08798
	SW	.05105	.11554	.89857	.62980

Table 1: (Continued)

n	test	Normal (0,1) Type 1 Error	t (3) Power of test	c ² (10) Power of test	b (3, 1.5) Power of test
60	g ₁	.04950	.11987	.94827	.00000
	g ₂	.06153	.26450	.50374	.05462
	DP	.04962	.22193	.80144	.12904
	SW	.05070	.11256	.94949	.76000
70	g ₁	.04630	.12020	.96958	.00000
	g ₂	.06146	.29075	.54614	.06752
	DP	.04618	.23546	.85796	.15545
	SW	.04978	.10924	.97688	.85523
80	g ₁	.04727	.12847	.98530	.00000
	g ₂	.06188	.31375	.58562	.08305
	DP	.04625	.25583	.91808	.23467
	SW	.05015	.10469	.98910	.91593
90	g ₁	.04984	.13192	.99275	.00000
	g ₂	.06279	.33577	.62190	.09946
	DP	.04750	.27478	.95444	.33619
	SW	.05160	.10259	.99509	.95304
100	g ₁	.04564	.12965	.99645	.00000
	g ₂	.06307	.35767	.65868	.11535
	DP	.04416	.28619	.97380	.39394
	SW	.04990	.09786	.99836	.97565
125	g ₁	.04981	.13726	.99914	.00000
	g ₂	.06592	.39759	.71399	.15017
	DP	.04650	.31994	.99374	.62367
	SW	.05027	.09258	.99966	.99347
150	g ₁	.04911	.14294	.99980	.00000
	g ₂	.06561	.43624	.76276	.18474
	DP	.04471	.34731	.99843	.77277
	SW	.05215	.08620	.99993	.99857

reached adequate power. g₁ and SW tests showed similar performance in c² (10) distribution. For b(3, 1.5) distribution, g₁ test was found as biased.

Results of this study showed that the SW test was the most powerful among mentioned tests. In addition, the performance of this test was found adequate. This test might be suggested for testing normality of data. It was

also concluded that performance of the normality tests was greatly affected by the distribution type and sample size.

REFERENCES

1. Anonymous, 1994. Fortran PowerStation version 4.0, Inc., Houston, USA.

2. Atwood, J.S. Shaik and M. Watts, 2003. Are crop yields normally distributed? A reexamination. *Am. J. Agric. Econ.*, 85: 888-901.
3. Cho, D.W. and K.S. Im, 2002. A Test of Normality Using Geary's Skewness and Kurtosis Statistics. Department of Economics, College of Business Administration, University of Central Florida, Orlando, pp: 1-11.
4. D'Agostino, R.B., A. Belanger and R.B. D'Agostino JR, 1990. A suggestion for using powerful and informative tests of normality. *American Statistician*, 44: 316-321.
5. Filliben, J. J., 1975. The probability plot correlation coefficient test for normality, *Technometrics*, 17: 111-117.
6. Iman, R.L., 1982. Graphs for use with the Lilliefors test for normal and exponential distribution. *Am. Statistician*, 36: 109-112.
7. Mendes, M. and A. Pala, 2003. Type I error rate and power of three normality tests. *Pak. J. Inform. Technol.*, 2: 135-139.
8. Shapiro, S.S. and M.B. Wilk, 1968. Approximations for the Null Distribution of W Statistic. *Technometrics*, 10: 861-866.
9. Zar, J.J., 1999. *Biostatistical Analysis*, Prentice Hall Inc. Upper Saddle River, New Jersey, USA, pp: 663.