Pakistan Journal of Information and Technology 2 (2): 135-139, 2003
ISSN 1682-6027

Type I Error Rate and Power of Three Normality Tests

Mehmet Mendes and [1]Akin Pala
Department of Genetics, Faculty of Agriculture Biometry, [1]Department of Animal Science,
Canakkale Onsekiz Mark University, Canakkale, Turkey

Abstract: In this study, Shapiro-Wilks, Lilliefors and Kolmogorov-Smirnov tests were compared for Type I error and for power of the tests. The simulation was run 100, 000 times for different situations and for different types of departures from normality. For all different sample sizes and distributions, Shapiro-Wilks gave the most powerful results, followed by the Lilliefors test. Kolmogorov-Smirnov test results were the weakest among all three tests. All three test were most powerful when ran on data with exponential distribution.

Key words: Type I error, shapiro-wilks, lilliefors and kolmogorov-smirnov tests

Introduction

In most statistical analyses, such as F, Z, T-tests, data is assumed to be normally distributed. The main three tests that assess assumption of normality are Shapiro-Wilks, Lilliefors and Kolmogorov-Smirnov. Data can be viewed with graphical methods to roughly assess normality. However, graphical methods do not test if the differences between normal distribution and the sample distribution are significant. Tests used for assessing normality are Chi-square, Anderson Darling, Ryan Joiner, Kolmogorov-Smirnov, Shapiro-Wilks and Lilliefors. The last three are the most frequently used tests. In most situations, data deviates from normality. Previous studies did not attempt to determine which testing method gives higher power for different cases of sample sizes and distributions and they had low simulation runs (Oja, 1983; Ohta and Arizono, 1989; Lin and Mudholkar, 1980). The major objective of this study was to evaluate Shapiro-Wilks (Shapiro and Wilk, 1965); Lilliefors (Lilliefors, 1967) and Kolmogorov-Smirnov (Kolmogorov, 1933 and Smirnov, 1939) methods for Type I error rates and for power of the tests.

Shapiro-wilk W test

This test for normality, developed by Shapiro and Wilk (1965) is the most powerful and omnibus test in most situations (D'Agostino and Stevens, 1986). In recent years, the Shapiro-Wilks SW test has become the preferred test of normality because of its good power properties as compared to a wide range of alternative tests (Shapiro *et al.,* 1968).

The test statistic for this test is;

$$SW = \frac{\left\{ \sum_{i=1}^{n} a_i X(i) \right\}^2}{\sum_{i=1}^{n} (s_I - \bar{X})^2}$$

where *x*(i) is the *i*-th largest order statistic, 0 is the sample mean, and *n* is the number of observations. Royston (1982) gives approximations and tabled values which may be used to compute the coefficients *a*i, *I* = 1, K, *n*, and obtain the significance level of the *SW* statistic.

Kolmogorov-smirnov test

Kolmogorov-Smirnov test was first proposed by Kolmogorov (1933) and then developed by Smirnov (1939). The test statistic is defined as $D = |F_0(X) - S_n(X)|$; where F_0 (X) is function of the random variable X (expected) and S_n (X) is the observed frequency of the variable X from sample. If resulting D statistic is significant, then the hypotheses that sample comes from a normally distributed population is rejected.

Lilliefors test

Lilliefors test is different than Kolmogorov-smirnov test because the parameters are estimated; while the statistic is the same. The table values of the two tests are different, which results in different decisions.

Materials and Methods

A computer simulation program was used to study Monte Carlo techniques. Fortran was used to write the program for Intel Pentium III processor. Type I error rates and statistical power of Shapiro-Wilk, Lilliefors, and Kolmogorov-Smirnov tests were measured for different situations. Samples with various sample sizes were taken from the Normal (0, 1), t (30), P^2 (30), Gamma (2,3), Wiebul (1.5), Exp (0.50), Beta (2,5) and P^2 (3) distributions. Random numbers were generated using generators from IMSL (functions RNNOA, RNSTT, RNCHI, RNWIB, RNEXP and RNBET) (Anonymous, 1994). Sample sizes were chosen as n_i = 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, 150, 200 for each distribution. This allowed assessment of the Type I error rates and power of statistical tests under small, moderate and large sample size conditions. In each case 100,000 pairs of data sets were generated. Each pair was then compared by each of the two tests. The populations were standardized because they have different means and variances. When samples were taken from normal (0,1) populations, the number of rejected H_0 hypotheses was declared as the probability for Type I error. When samples were taken from populations with non-normal distributions, the number of rejected H_0 hypotheses was declared as the test's power. So, to compute empirical Type I error rate and test power, the program ran each condition 100,000 times and kept tract of proportion of significant statistics.

Results and Discussion

Empirical results of 100 000 simulation runs are given in Table 1. When the distribution was normal, all three tests had similar Type I error rates. When the distribution was t (30), number of rejected null hypotheses (power of the tests) were similar to the number of rejected null hypotheses (Type I error rate) for normally distributed data (0, 1) for all three tests. The only exception of this was the data where sample size was 100. It may be suggested that t-student distributed data with 30 degrees of freedom or more can be treated as normally distributed data.

When the tests were run on data with 30 degrees of freedom and chi square distribution, they all presented low power. Even the tests ran on data with sample sizes as large as 200 had low power when the data was distributed with chi-square. Powers of tests were 63.3% for Shapiro-Wilk, 43.6% for Lilliefors and 14.4% for Kolmogorov-Smirnov. When the simulated data was distributed with Gamma (2, 3), Shapiro-Wilk test was most powerful for 50 or larger sample sizes while Lilliefors test was most powerful when sample size was 100 or larger (Table 1). Kolmogorov-Smirnov test could only reach small power levels (55.1%) even when large sample sizes (n = 200) were used. When the data was distributed with Weibul (1.5, 1), Shapiro-Wilk test was most powerful for 45 and larger sample sizes, and Lilliefors test was most powerful for 100 and larger sample sizes. Kolmogorov-Smirnov test was the weakest one; it could not reach sufficient power levels (80%) even for the largest sample size.

For Gamma (2, 3) or Weibull (1.5, 1) distributions, Lilliefors and Kolmogorov-Smirnov tests were similar in power. Exponential (0.50) distribution had a positive effect on power levels of all three tests, especially Kolmogorov-Smirnov test. Shapiro-Wilk test reached sufficient power levels with 20 and larger sample sizes while Lilliefors test required 35 or larger, and Kolmogorov-Smirnov test required 75 or larger sample sizes. Lin and Mudholkar (1980) reported that Shapiro-Wilk test was more powerful than Kolmogrov-Smirnov for exponential distribution. All tests had low power levels

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

	Normal			t (30)			P^2 (30)			Gamma (2,3)		
	Type I error rate			Test power			Test power			Test power		
Distributions												
tests	SW	LF	KS	SW	LF	KS	SW	LF	KS	SW	LF	KS
7	5.13	5.21	4.98	5.39	5.24	5.06	6.33	6.12	5.38	11.15	9.69	6.30
8	5.09	5.26	5.06	5.27	5.38	4.96	6.33	6.00	5.47	13.05	10.90	6.73
9	4.75	5.00	5.24	5.38	5.43	5.21	6.89	6.50	5.24	14.99	11.55	7.13
10	5.10	5.21	5.06	5.39	5.00	5.10	7.02	6.27	5.17	17.32	12.38	7.23
11	4.96	4.93	4.88	5.42	5.29	5.00	7.99	6.81	5.85	19.63	13.47	7.63
12	4.96	4.88	5.22	5.88	5.61	5.09	8.40	7.15	5.61	21.19	14.57	7.41
13	4.83	4.96	4.92	5.83	5.47	4.94	8.23	6.86	5.51	23.19	15.55	8.02
14	4.90	4.96	4.92	5.54	5.36	5.24	8.79	7.09	5.64	24.52	16.97	8.11
15	4.67	5.05	5.06	5.30	5.42	5.00	8.80	7.13	5.70	27.02	17.61	8.16
20	4.71	4.72	4.96	5.79	5.18	5.07	11.22	8.24	5.85	36.88	22.23	9.77
25	5.00	4.71	4.98	5.83	5.33	5.12	13.79	9.38	6.24	47.16	27.64	10.66
30	4.82	5.00	5.03	5.88	5.42	5.36	15.15	9.96	6.50	56.11	32.52	12.47
35	4.93	4.84	4.94	6.01	5.54	5.12	17.18	11.03	6.53	63.90	36.96	13.39
40	4.92	4.78	4.92	6.05	5.75	5.22	18.92	12.12	6.97	71.94	42.56	15.62
45	5.00	4.77	4.77	6.03	5.43	5.22	20.26	12.79	7.33	77.44	45.98	16.16
50	5.05	4.91	5.01	5.79	5.47	5.25	22.60	13.55	7.36	83.03	50.84	17.18
75	5.13	4.98	4.91	5.33	5.72	5.17	31.66	19.03	8.60	95.89	69.91	24.02
100	4.92	5.09	5.22	4.65	5.88	4.72	39.00	24.41	9.31	99.11	82.55	29.71
150	5.05	5.05	4.57	3.91	7.60	4.96	52.42	10.20	11.66	99.98	86.97	42.81
200	5.41	5.16	4.92	3.52	6.04	4.79	63.32	43.60	14.40	100.0	98.87	55.12
Mean	4.97	4.97	4.98	5.41	5.58	5.08	18.72	11.71	7.02	50.17	35.99	15.68
Std. Error	0.038	0.036	0.035	0.151	0.119	0.035	3.60	1.98	0.531	7.45	6.30	2.93

Table 1: Continue

	Wiebul (1.5,1)			Exp (0.50)			Beta (2,5)			P^2 (3)		
	Test power			Test power			Test power			Test power		
Distributions												
tests	SW	LF	KS	SW	LF	KS	SW	LF	KS	SW	LF	KS
7	12.02	9.92	6.79	27,96	20,36	9,07	7,49	6,90	6,46	19,37	14,56	7,87
8	13.82	10.70	7.18	33,17	23,99	10,47	7,40	6,97	6,51	22,92	17,27	8,47
9	14.92	11.16	7.38	38,91	27,25	11,12	8,15	7,03	6,37	27,51	19,27	8,79
10	17.39	12.45	7.67	43,68	30,00	11,86	8,80	7,34	6,77	31,27	21,49	9,22
11	19.56	13.53	7.71	49,41	32,50	12,11	9,36	7,99	6,98	16,14	11,42	6,30
12	21.21	14.42	7.95	53,87	35,63	13,54	10,15	8,02	6,76	17,96	12,73	6,54
13	23.27	15.06	8.16	57,99	38,38	14,04	10,10	8,16	7,12	20,23	13,41	6,64
14	25.25	16.34	8.54	62,87	41,60	14,85	11,43	8,57	7,22	35,09	23,14	9,79
15	27.40	17.46	8.84	66,76	44,15	15,35	11,68	8,82	7,08	38,20	25,29	10,12
20	40.00	22.77	10.53	83,75	57,56	20,13	17,09	10,95	7,87	41,87	27,28	10,83
25	51.20	27.68	11.30	92,64	68,40	24,20	22,09	13,00	8,38	45,71	29,45	11,40
30	60.58	31.96	12.28	97,13	78,55	28,36	28,17	15,93	9,20	49,37	31,39	12,42
35	69.84	37.67	14.25	98,83	84,77	32,35	34,37	18,07	9,69	66,25	40,83	14,91
40	78.22	43.32	15.54	99,64	90,19	35,81	41,09	20,14	10,30	78,38	49,85	17,38
45	84.56	47.63	16.55	99,87	93,70	39,81	48,32	22,61	10,82	87,04	58,60	20,33
50	89.39	52.02	17.73	99,97	96,14	44,27	55,89	25,82	11,96	92,44	65,79	22,75
75	98.50	71.94	25.29	100,0	99,73	99,90	80,67	38,31	14,64	96,11	72,91	25,26
100	99.92	85.32	30.40	100,0	99,99	99,97	93,00	50,65	16,72	98,15	77,85	28,25
150	100.0	89.50	43.92	100,0	100,0	99,98	99,39	34,22	24,19	98,97	82,63	30,84
200	100.0	99.42	55.73	100,0	100,0	99,99	99,98	84,61	30,74	99,97	95,34	44,28
Mean	52.35	36.5	16.19	75.32	63.15	36.86	35.23	20.20	10.79	54.15	39.52	15.62
Std. Error	7.70	16.48	2.96	5.99	6.85	7.59	7.42	4.35	1.44	7.12	5.97	2.26

when the distribution was Beta (2, 5). Shapiro-Wilk test reached sufficient power levels with 75 or larger sample sizes, while Lilliefors test required at least a sample size of 200 to reach such a level of power. Kolmogorov-Smirnov test was weak in all situations for the Beta (2, 5) distribution. Tests ran on data with chi-square distribution and 3 degrees of freedom accomplished slightly higher power levels than tests ran on data with Beta distribution. Shapiro-Wilk test required 45 and larger sample sizes and Lilliefors test required a sample size of 200 to reach a sufficiently large power level. Kolmogorov-Smirnov test had small power levels in all sample sizes.

Implications

Shapiro-Wilk was the most powerful test regardless of distribution and sample size and it should be used when testing for normality. Kolmogorov-Smirnov had the smallest rejection rates, so this method should be used with strong caution when assessing normality. All tests were more powerful when used on data with exponential distribution.

The results of 100,000 simulation runs showed that;

1 When the distribution is normal (0, 1), any of these tests can be used to compare Type I error rates,

2 Regardless of the distribution and sample size, Shapiro-Wilk test gave higher power levels than the other two tests,

3 In all situations, Kolmogorov-Smirnov test achieved smallest power levels,

4 Power levels for t (30) distribution and the type I error rates for normal (0, 1) distribution were similar, indicating that tests done on t (30) distributed populations can be used to have an idea on Type I error rates of normal populations,

5 Smallest power levels were achieved in samples with Beta (2, 5) distributions,

6 Shapiro-Wilk and Liliefors tests performed similar in samples with Gamma (2, 3) and Weibull (1.5, 1) distributions,

7 All tests were more powerful when used on data with exponential distribution (0.50).

References

Anonymous, 1994. Fortran subroutines for Mathematical Application. IMSL MATH/LIBRARY. Vol. 1=2. Visual Numerics, Inc., Houston, USA.

Lilliefors, H.W., 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. ASA J., pp: 399-402.

Lin, C.C. and G.S. Mudholkar, 1980. A Simple Test For Normality Against Asymetric Alternatives Biometrika, 67: 455.

Shapiro, S.S. and M.B. Wilk, 1965. An Analysis of Variance Test for Normality (Complete Samples), Biometrika, 52: 591.

Shapiro, S.S., M.B. Wilk and H.J. Chan, 1968. A Comparative Study of Various Tests for Normality JASA, 63: 324.

Ohta, H. and I. Arizono, 1989. A Test For Normality Based On Kullback-Leibler Information. The Am Stat, 43: 20-22.

Oja and Hannu, 1983. New tests for normality, Biometrika, 70: 297.