Empirical Investigation of Type 1 Error Rate of Some Normality Test Statistics

[1]John O. Kuranga, [2]Kayode Ayinde, [3]Gbenga S. Solomon

ABSTRACT--*Normality assumption is important in many parametic statistical tests. Either the varibles or the error terms in the model have to be assumed to be normally distributed before statistical conclusions can be made. Various statistical tests which include that ofPearson (1900, 1905),Kolmogorov–Smirnov (1933), Anderson-Darling (1954), Shapiro–Wilk (1965),Lilliefor (1967),D'Agostino and Pearson (1973), Jarque-Bera (1987),Shapiro-Franca (1992),Energy (Szekeley and Rizzo, 2005)and Cramer-von Mises (Thadewalid and Buning, 2007) have been developed to test for normality of a set of data. However, when applied in practice, they hardly lead to the same conclusion. This is a serious challenge to practioners. Consequently, this research work aims at investigating the Type1 error rate of some of the nomality statistics so as to identify the best one and recommed the same for statistics users. Monte Carlo experiments were conductedfive thousand (5000) times with six sample sizes (n =20, 50, 100, 250 and 500) at three pre-selected levels of significance ($\alpha =$ 0.01, 0.05 and0.1). A statistic was considered good if its estimated Type 1 error rate approximated the pre-selected level of significance, and was considered best if its number of counts at which it was good over the three (3) levels of significance and six (6) sample sizes was the highest. Results show that Type 1 error rate of all the statistics are goodexcept that of Kolmogorov–Smirnov, Pearson Unadjusted and Jarque-Bera. The Ominibus test statistics is only good at 0.1 level of significance. In general, the Type 1 error rate of Anderson-Darling,Shapiro-Wilk,Energy, Cramer-vonMises test statistics are best. These are followed by that of Shapiro-Franca and Lilliefortest statistics . Consequently, Anderson-Darling, Shapiro–Wilk, Energy and Cramer-VonMises test statistics are recommended for use in test of normality of a data set.*

Keywords--*Parametric test statistics, Monte Carlo experiments, Type 1 error rate, Inferencial statistics tests, Levels of signficance.*

I INTRODUCTION

Normality assumption is an underlying assumption in many statistical parametric tests. It is used in many statistical procedures include Time series, Discriminant analysis and Analysis of Variance (ANOVA) and in virtually all the parametric statistical tests. Assessing the assumption of normality is required before proceeding with any relevant statistical inferences. There are three common techniques for checking the normality status of independent observations. These are Graphical, Numerical and the formal normality test statistic methods.The graphical is the easiest and it requires the normalquantile-quantile (Q-Qplot) and Histogram plots.Genarally, graphical methods areinformal approach. The Numericalmethods includeSkewness and Kurtosis indices

[1]*Department of Statistics,Kwara State Polytechnic,* P.M.B. 1375*, Ilorin, Kwara State, Nigeria,olatundej22@yahoo.com*
[2]*Department of Statistics,*Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria, *kayinde@lautech.edu.ng,*
[3]*Department of Statistics,*Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria, *gssolomon@outlook.com*

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

(coefficience)which are generally refered to as standardized moments. The Formal Normality testis a scientific test in that test statisticsaredeveloped. The procedure involves testing whether a particular data set follows a normal distribution and computing the probablities of how likely underlying data set is normally distributed. In this study, attention is on thirteen(13) test statistics of univariate normality which areUnadjusted Pearson (1900, 1905),Kolmogorov–Smirnov (1933), Anderson-Darling (1954), Shapiro–Wilk (1965),Lilliefor (1967),D'Agostino (Skewness and Kurtosis, 1970),Adjusted Pearson (1973), Omnibus (1973) Jarque-Bera (1987), Shapiro-Franca (1992),Energy(Szekeley and Rizzo, 2005) and Cramer-von Mises (Thadewalid and Buning, 2007)test statistics. It is intended to evalate their Type 1 error rate and identify the best ones among them for inferencial usefulness.

II LITERATURE REVIEW

Statistical investigation and inference require making correct decision. These have led to the development of the various emperical and simulation studies in order to identify the test statistic which can provide a good Type 1 error rate and the one that is most poweful, sensitive to departure from normality. The Ominbus test statistic has been reported to be questionable because of the dependence stature of the transformed Skweness and Kurtosis of the statistic (Shestion and Bowman, 1977).

Judge, et al. (1988), and Gujarati (2002)recommended the use of Jarque–Bera test statistics. However, Jarque-Bera test statistics was observed to have low power for distributions with short tails, more importantly if it is bimodal distributions. ThadewaldandBüning (2007), and Sürücü(2008)did not use Jarque-Bera teststatistics in their own studies because it's poor in overall performance.

Mendes and Pala (2003) in a study on Type 1 error rate and power of three normality test statistics. The Shapiro-Wilk test is reported to have good power.

A research by Nornadiah and Yap Bee(2011) and Ogunleye(2013) examined "the power of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling test".(Razali and Wah, 2011) They concluded that Shapiro-Wilk has the best power for a given significance, it's followed by Anderson-Darling whereas Kolmogorov-Simirnov test is least powerful.

According to Jarque and Bera (1987), if data come from a normal distribution, the Jarque-Bera statistics asymptotically has chi-square distribution with two degrees of freedom. For small samples the chi-square approximation is too sensitive, this will lead to rejecting the null hypothesis when actually it is true. In addition, the distribution of p-value departs from a uniform distribution and become a right-skewed uni-modal distribution, more importantly for small p-values. which leads to a large Type 1 error rate.

However, in this review, it should be noted that no study has ever compared a large number normality test staistics like the one done in this study; and more importantly the performance of test statistics like Energy(Szekeley and Rizzo, 2005) and Cramer-von Mises criterion (Thadewalid and Buning, 2007) test statistics have not been examined with others.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

III METHODOLOGY

In order to emperically ivestigate the Type 1 error rate of the normality test statistics, Monte Carlo experiments were conducted by generating data from normally distributed population five thousand times, $X_i \sim N(0,1)$, i =1,2,...,5000 for six sample sizes namely; n = 20, 30, 50, 100, 250,500**.** The R- Statistical software was used for the simulation study, and R-codes were written for all the thirteen (13) normality test statistics. These statistis are Unadjusted Pearson (1900, 1905),Kolmogorov–Smirnov (1933), Anderson-Darling (1954), Shapiro–Wilk (1965),Lilliefor (1967),D'Agostino'sK-squard (Skewness and Kurtosis, 1973),Adjusted Pearson (1973), Omnibus (1973) Jarque-Bera (1987), Shapiro-Franca (1992),Energy(Szekeley and Rizzo, 2005) and Cramer-von Mises (Thadewalid and Buning, 2007).Three pre-selected levels of significance used are 0.1, 0.05 and 0.01.

At a particular sample size, the number of times the true hypothesis is rejected is counted and the total is divided by the number of replication to estimate the Type 1 error rate of each statistic at the three levels of significance. Statistic whose error rateapproximated the true error rate was considered good. That is, statistic whose error rate felt into the preffered intervals as specified in Table 1. Each peferred intervals are such that the values therein approximate the true level of significance.

Table 1: The true/pre-selectedlevels of significance and the preferred interval of levels of significance

True levels of significance	Preferred interval
0.01	0.005 − 0.014.
0.05	0.045 − 0.054
0.1	0.095 − 0.14

Source: Self motivationed.

Furthermore, the number of times each statistic was considered good was counted over the three (3) levels of significance and six (6) sample sizes. Thus, a total number of eighteen (18) counts were expected. A statistic was considered best if it has the highest number of total counts.

IV RESULTS AND DISCUSSION

4.1 Results Of **Type1 ErrorRates Of The Statistics***at 0.1 level of significance*

Table 2 shows the result of the Type 1 error rate at which each of the thirteen (13) normality test statistics reject a true null hypothesis at 0.1 level of significance. In order to get a better view of the performance of the thirteen13 normality tests, the value of the test whose Type 1 error rate is closest to 0.1, using the preffered interval, are bold and presented in Table 2.

From Table 2, it can be seen that all the statistics generally have good Type 1 error rate except Kolmogorov–Smirnov, Pearson Unadjusted and Jarque-Bera. It should be noted that the Jarque-Bera test statistic under estimate the Type 1 error rates even though the error rate is not as bad as Pearson Unadjusted test statistic. The Type 1 error rate of Kolmogorov–Smirnov test statistic is worst. Moreover, the Ominibus, Skewnwss and Kurtosis statistics are also good except when the sample size is small, n=20. This is further shown in Figure 1.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

Table2: Type1 errorrateofthestatistics at 0.1 level of significance

Statistics	Sample Size						Total Count
	20	30	50	100	250	500	
Anderson	**0.0986**	**0.102**	**0.1058**	**0.1036**	**0.1056**	**0.0982**	**6**
Kolmogorov	2.00E-04	0.0012	0.001	6.00E-04	8.00E-04	6.00E-04	0
Shapiro-Wilk	**0.0986**	**0.1016**	**0.1034**	**0.1036**	**0.1002**	**0.1054**	**6**
Cramer-vonMises	**0.0994**	**0.1048**	**0.1018**	**0.1012**	**0.1072**	**0.0972**	**6**
Shapiro-Franca	**0.099**	**0.1022**	**0.1018**	**0.1044**	**0.1016**	**0.1054**	**6**
Jarque-Bera	0.08	0.0908	0.081	0.0826	0.088	0.0886	**0**
Lilliefors	**0.0954**	**0.0972**	**0.1064**	**0.104**	**0.1004**	**0.1038**	**6**
Omnibus	0.0916	**0.0976**	**0.1002**	**0.0984**	**0.1018**	**0.1016**	**5**
Energy	**0.0984**	**0.1036**	**0.1032**	**0.1036**	**0.1072**	**0.1008**	**6**
Skewness	0.0934	**0.103**	**0.1014**	**0.1034**	**0.1076**	**0.1064**	**5**
Kurtosis	0.0928	**0.0952**	**0.0962**	**0.1008**	**0.1034**	**0.0976**	**5**
Pearson	**0.1218**	**0.122**	**0.1066**	**0.1052**	**0.1052**	**0.1002**	**6**
UnPearson	0.0336	0.0402	0.05	0.0558	0.057	0.065	0

Source: Computer Output

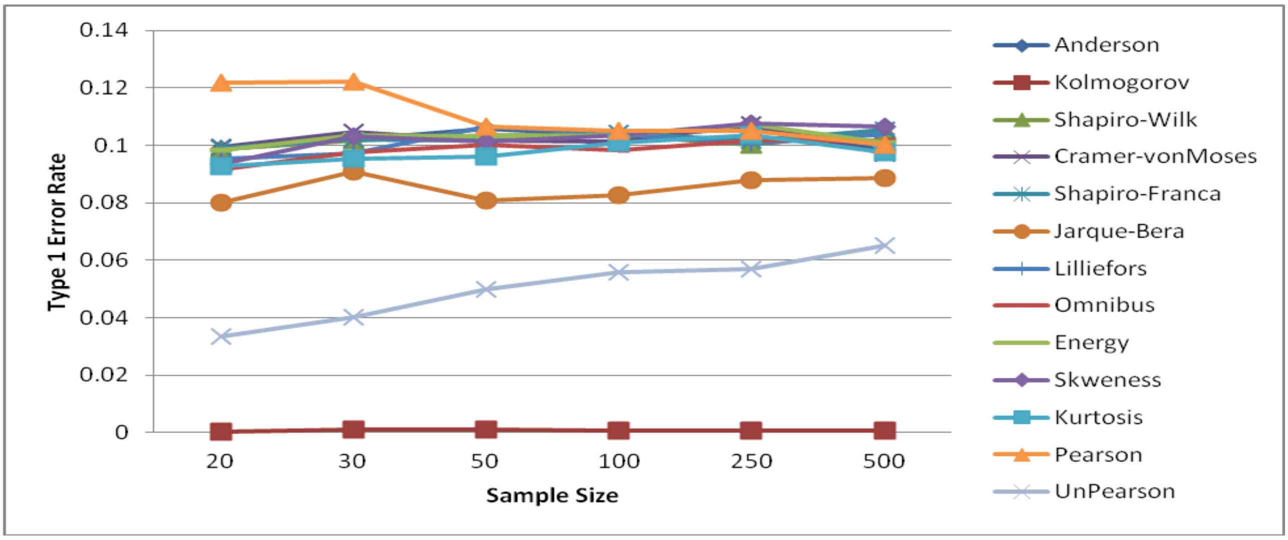

Figure 1: Type1 error rate of the statistics at 0.1 level of significance

Source: Table 2

From Figure 1, it can be seen that the Type 1 error rate of Anderson-Darling, Shapiro–Wilk, Energy and Cramer-VonMises test statistics are genearally good at all the sample sizes.

4.2 Results Of **Type1 ErrorRates Of The Statistics** *at 0.05 level of significance*

Table 3 shows the result of the Type 1 error rate at which each of the thirteen (13) normality test statistics reject a true null hypothesis at 0.05 level of significance. In order tohave a better understanding of the performance of the thirteen (13) normality tests, the value of the test whoseType 1 error rate is closest to 0.05, using the preffered interval, are bold and presented in Table 3.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

Table 3: Type1 errorrateofthestatistics at 0.05 level of significance

Statistics	Sample Size						Total Count
	20	30	50	100	250	500	
Anderson	**0.0476**	**0.0508**	**0.0516**	**0.0522**	**0.0498**	**0.0506**	**6**
Kolmogorov	0	2.00E-04	0	0	0	0	0
Shapiro-Wilk	**0.049**	**0.0502**	**0.0518**	**0.0504**	**0.0504**	**0.052**	**6**
Cramer-vonMises	**0.0488**	**0.0518**	**0.0516**	**0.051**	**0.05**	**0.0508**	**6**
Shapiro-Franca	**0.052**	**0.0546**	**0.0502**	**0.0522**	**0.0522**	0.0554	5
Jarque-Bera	0.0614	0.0668	0.058	0.0586	**0.0536**	**0.0478**	**2**
Lilliefors	**0.0488**	**0.0492**	**0.0546**	**0.0476**	0.0448	**0.0474**	**5**
Omnibus	0.0558	0.0592	**0.0544**	0.0556	**0.0534**	0.056	2
Energy	**0.0468**	**0.0498**	**0.052**	**0.052**	**0.051**	**0.0508**	**6**
Skewness	**0.0486**	**0.0508**	**0.0482**	**0.0496**	**0.054**	0.0574	5
Kurtosis	**0.0464**	**0.0478**	**0.0518**	0.057	**0.0488**	**0.0516**	**5**
Pearson	0.044	**0.0524**	0.056	0.0558	**0.0518**	0.0554	2
UnPearson	0.0136	0.0178	0.022	0.0252	0.0234	0.0286	0

Source: Computer Output

From Table 3, it can be seen that all the statistics generally have good Type 1 error rate except Kolmogorov–Smirnov, Pearson adjusted and Unadjusted, Ominibus and Jarque-Bera. It should be noted tha the Jarque-Bera test statistic over estimate the Type 1 error rates except when the sample size is large, $n \geq 250$. Its error rate is also not as bad as Pearson Unadjusted test statistic. The Type 1 error rate of Kolmogorov–Smirnov test statistic is worst. Moreover, that of Skewnwss and Kurtosis statistics are better than that of Ominibus. This is further shown in Figure 2.

From Figure 2, it can be seen that the Type 1 error rate of Anderson-Darling, Shapiro–Wilk, Energy and Cramer-VonMises test statistics are genearally good at all the sample sizes.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

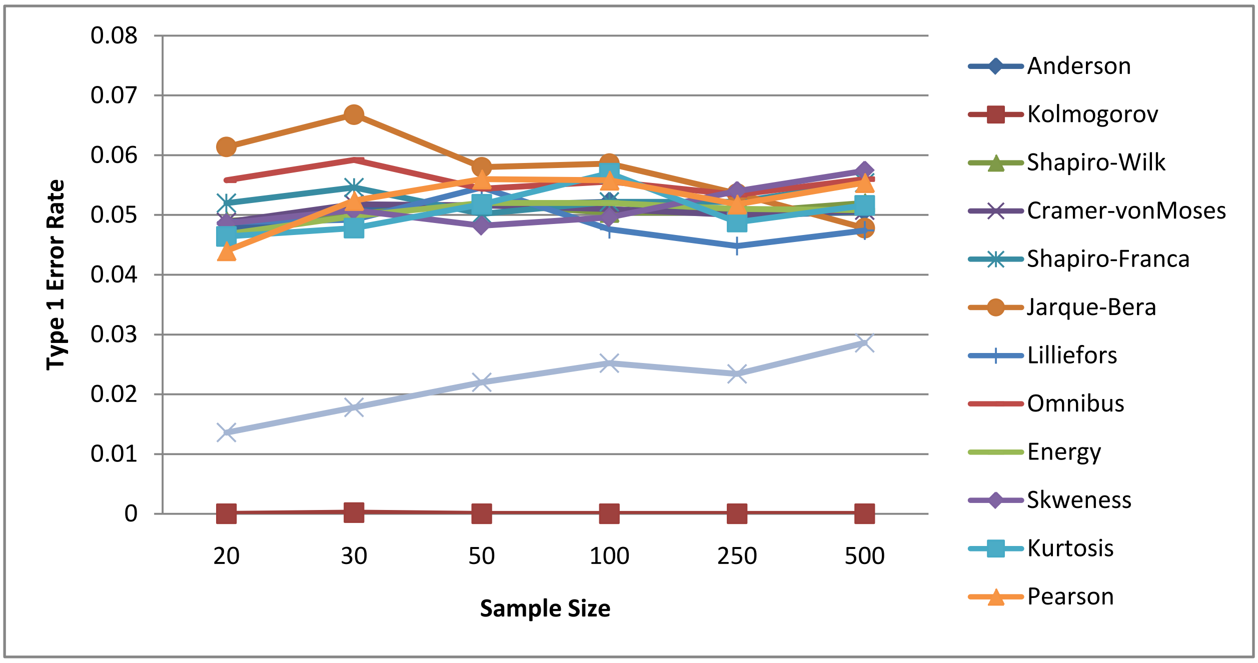

Figure 2: Type1 error rate of the statistics at 0.05 level of significance

Source: Table 4

4.3 Results of **Type1 ErrorRates of the Statistics** *at 0.01 level of significance*

Table 4 shows the result of the Type 1 error rate at which each of the thirteen (13) normality test statistics reject a true null hypothesis at 0.01 level of significance. In order to get a better view of the performance of the thirteen (13)normality tests, the value of the test whose Type 1 error rate is closest to 0.01, using the preffered interval, are bold and presented in Table 4.

Table4: Type 1 error rate of thestatistics at 0.01 level of significance

Statistics	Sample Size						Total Count
	20	30	50	100	250	500	
Anderson	**0.0084**	**0.0094**	**0.0114**	**0.0104**	**0.009**	**0.0108**	**6**
Kolmogorov	0	0	0	0	0	0	0
Shapiro-Wilk	**0.0096**	**0.01**	**0.0084**	**0.01**	**0.0114**	**0.012**	**6**
Cramer-vonMises	**0.0088**	**0.0098**	**0.0108**	**0.0088**	**0.0088**	**0.0108**	**6**
Shapiro-Franca	**0.0106**	**0.0118**	**0.009**	**0.0122**	**0.0126**	**0.0124**	**6**
Jarque-Bera	0.0364	0.04	0.0304	0.0288	0.0232	0.0162	0
Lilliefors	**0.01**	**0.01**	**0.0128**	**0.01**	**0.0106**	**0.0088**	**6**
Omnibus	0.0202	0.019	0.0154	0.018	0.0178	0.0138	0
Energy	**0.009**	**0.0098**	**0.012**	**0.0108**	**0.0096**	**0.011**	**6**
Skewness	**0.0104**	**0.0106**	**0.008**	**0.0108**	**0.0122**	**0.01**	**6**
Kurtosis	**0.0084**	**0.0112**	**0.0122**	0.0152	**0.0132**	**0.0118**	**5**
Pearson	**0.0098**	**0.0134**	**0.011**	**0.011**	**0.0086**	**0.009**	**6**
UnPearson	0.0024	0.0034	0.004	0.0034	0.0034	0.004	0

Source: Computer Output

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

From Table 4, it can be seen that all the statistics generally have good Type 1 error rate except Kolmogorov–Smirnov, Pearson Unadjusted Ominibus and Jarque-Bera. It should be noted that the Jarque-Bera and Ominibustest statistics over estimate the Type 1 error rates . Pearson Unadjusted test statistic under estimate the Type 1 error rates. The Type 1 error rate of Kolmogorov–Smirnov test statistic is worst. This is further shown in Figure 3

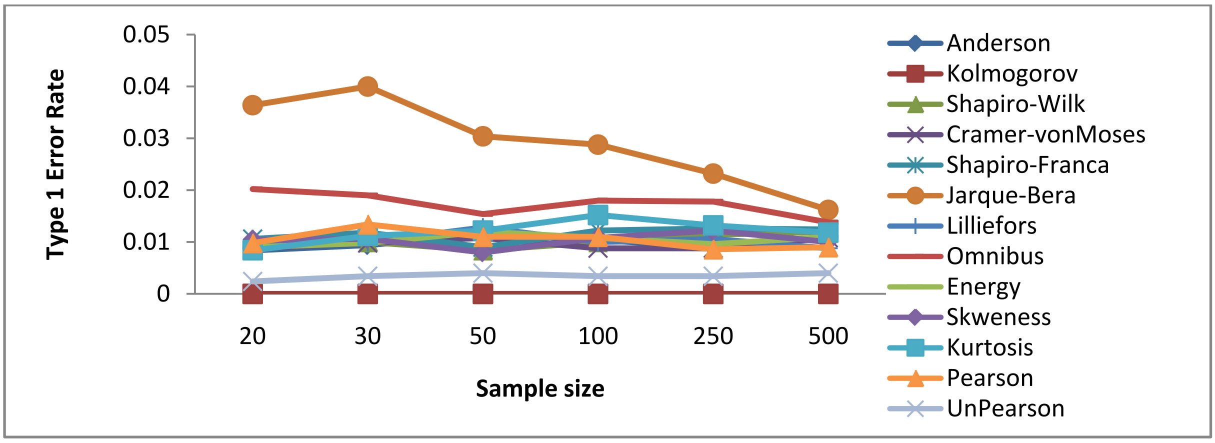

Figure 3: Type1 error rate of the statistics at 0.01 level of significance

Source: Table 4.

4.4 Overall performance of the Normality Test Statistics

In order to see the performance of the statistics clearly, the number of times the Type 1 error rate falls into the preferred interval is counted over the three levels of significance. This is referred to as number of good performance of the nomality test statistics. This is given in Table 5

Table 5: Number of good performance of the Normality Test Statistics

Statistics	Sample Size						Total counts
	20	30	50	100	250	500	
Anderson	3	3	3	3	3	3	18
Kolmogorov	0	0	0	0	0	0	0
Shapiro-Wilk	3	3	3	3	3	3	18
Cramer-vonMises	3	3	3	3	3	3	18
Shapiro-Francia	3	3	3	3	3	2	17
Jarque-Bera	0	0	0	0	1	1	2
Lilliefors	3	3	3	3	2	3	17
Omnibus	0	1	2	1	2	1	7
Energy	3	3	3	3	3	3	18
Skewness	2	3	3	3	3	2	16
Kurtosis	2	3	3	1	3	3	15
Pearson	2	3	2	2	2	3	14
Unadjusted Pearson	0	0	0	0	0	0	0

Source: Table 2, 3 and 4.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

From Table 5,it can be generally observed that the Type 1 error rate of Anderson-Darling,Shapiro-Wilk,Energy, Cramer-vonMises are best since they have the highest number of times the estimated Type 1 error rate fall into the preferred intervals . These are followed by those of Shapiro-Franca and Lilliefor . This is futher illustrated in Figure 4. Consequently, Anderson-Darling, Shapiro–Wilk, Energy and Cramer-VonMises test statistics are recommended for use in test of normality of a data set.

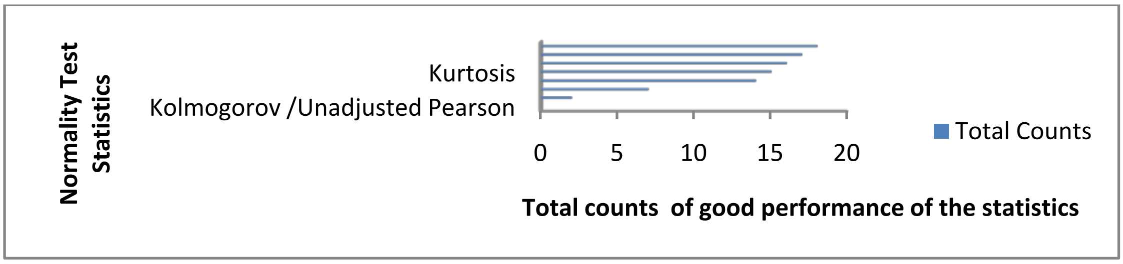

Figure 4: Bar chart showing Total Number of good performance of the Normality Test Statistics

Source: Table 4.4a.

V CONCLUSION

In this study, the Type 1 error rate of the thirteen (13) normality test statistics have been examined. Results have revealed that Type 1 error rate of Anderson-Darling,Shapiro-Wilk,Energy, Cramer-vonMises test statistics arebest. These are followed by those of Shapiro-Franca and Lilliefor test statistics. Skewness, Kurtosis, Pearson and Ominibus, in this order, follow the Shapiro-Franca and Lilliefor test statistics. The performance of Jarque-Bera test Statistic is not good while that of Unadjusted Pearson and Kolmogorov–Smirnov are worst Consequently, Anderson-Darling, Shapiro–Wilk, Energy and Cramer-VonMises test statistics are recommended for use in test of normality of a data set.

REFERENCES

1. Anderson, T. W. and Darling, D. A. (1954): A Test of Goodness of Fit. Journal of the Anerican statistical Association, 49, 268, 765−769
2. D'Agostino, R. B. (1970): Transformation to normality of the null distribution of g_1. Biometrika 57, 679–681.
3. D Agostino, R. and Pearson, E. S. (1973): Test for Departure from Normality. Empirical Results for the Distributions ofb_2 and $\sqrt{b_1}$. Biometrika, 60, 3, 613-622.
4. Gujarati, D. N. (2002): Basic Econometrics, Fourth Edition, 147–148, McGraw Hill. ISBN 0-07-123017-3.
5. Jarque, C. M. and Bera, A. K. (1987): A test for Normality of observations and regression residual,Internat. Statst. Rev, 55, 2, 163 – 172..
6. Judge, G. G; Griffiths W. E; Hill, R. C; Lütkepohl, H. and Lee, T. (1988): Introduction to the Theory and Practice of Econometrics, Second Edition, 890–892, Wiley. ISBN 0-471- 08277-5.
7. Kolmogorov, A. N. (1933): Sulla determinazione empirica di una lagge di distribuzione, Giornale dell Instituto Italiano degli Attuari 4, 83-91.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020

8. Lilliefors, H. W. (1967): Onthe Kolmogorov – Smirnov Test for Normality with mean andvariance unknown. Journal of American statistical Assocition, 62, 318, 399- 402.
9. Mendes, M. and Pala, A. (2003): Type 1 Error rate and power of Three Normality Test.Pakistan Journal of information and Technology 2, 2, 135 – 139.
10. Normadiah, M, R. and Yap, B,R.(2011): Power comperisons of Shapiro-Wilk,Kolmogorov- Smirnov, Lilliefors and Anderson-Darling test.Journal of Statistical Modeling and Analysis. 2,1,21-33.
11. Ogunleye, L.A. (2013): Comparision of some common tests for normality. Unpublished M.Sc. Thesis,University of Ilorin.
12. Pearson, K. (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. Philosophical Magazine Serie*s 5*50,302, 157–175.
13. Pearson, K.(1905):On general theory of skew correlation and Non-linear regression. London: Dulau and Co.
14. Shapiro, S. S. and Francia, R. S. (1972): An Approximate analysis of varinace test for normality. Journal of American Statistical Association. 67,215-216.
15. Shapiro, S. S. and Wilk, M. B. (1965): An Analysis of variance Test for Normalty Biometrika, 52, 3, 591 – 611.
16. Sürücü, B. (2008): A power comparison and simulation study of goodness-of-fit tests. Computers & Mathematics with Applications56, 6, 1617-1625.
17. Székely, G. J. and Rizzo, M. L. (2005): A new test for multivariate normality, Journal of MultivariateAnalysis 93, 58–80.
18. Thadewald,T. and Buning, H.(2007): Jarque – Bera and its Competitors for Testing Normality. Journal of Applied Statistics, 34, 1, 87 – 91.
19. Firas Hassan, Salam Abd AlQadeem Mohammed, Anil Philip, Ayah Abdul Hameed, Emad Yousif. "Gold (III) Complexes as Breast Cancer Drug." Systematic Reviews in Pharmacy 8.1 (2017), 76-79. Print. doi:10.5530/srp.2017.1.13
20. Başar, E.Multiple oscillations and phase locking in human gamma responses: An essay in search of Eigenvalues(2012) NeuroQuantology, 10 (4), pp. 606-618.
21. Clark, K.B.Bioreaction quantum computing without quantum diffusion(2012) NeuroQuantology, 10 (4), pp. 646-654.

DOI: 10.37200/IJPR/V24I4/PR201037
Received: 22 Oct 2019 | Revised: 13 Sep 2019 | Accepted: 15 Jan 2020