MIKAILALSYS
Journal of Multidisciplinary Sciences

e-ISSN : 2987-2286
p-ISSN : 2987-3924

Index : Dimensions, Scilit, Crossref, Garuda, Semantic, Lens,Google Scholar,Base, etc

https://doi.org/10.58578/mikailalsys.v3i1.5234

Type 1 Error Rate of Some Normality Tests

Ademola Abiodun Adetunji
Federal Polytechnic, Ile-Oluji, Nigeria
adeadetunji@fedpolel.edu.ng

Article Info:

Submitted:	Revised:	Accepted:	Published:
Feb 24, 2025	Mar 9, 2025	Mar 21, 2025	Mar 26, 2025

Abstract

Most Statistical procedures require normality of data for a reasonable interpretation and inference. Failure of this assumption invariably undermines the applicability of such data. Using a statistical method that requires normality of the data when the data is not normal will lead to misleading inference. Quite a number of procedures exist in literature in verifying the normality assumption. This paper examines simulated 10,000 normally distributed data and assessed them for normality using Anderson-Darling test (AD-test), Kolmogorov-Smirnov test (KS-test) and Shapiro-Wilk test (SW-test). The results of the analysis show that AD-test outperforms the other two tests. However, for sample size 20 or less, KS-test outperforms the rest. The sum of ranks of type 1 error rate showed that the AD-test is the best because it has the lowest rank sum. It is followed by the KS-test and the SW- test has the largest sum of ranks and hence the poorest. The type 1 error rate of each test does not show any consistent pattern as the sample size increases. Hence, it can be inferred that sample size does not have significant impact on the type I error of a test.

Keywords: Normality Test, Type 1 Error, Anderson-Darling Test, Kolmogrov-Smirnov Test, Shapiro-Wilk Test

Mikailalsys is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

INTRODUCTION

Recent developments in statistical data analysis necessitate developments of new distributions (Ademuyiwa *et al.,* 2023 & Adetunji and Sabri, 2023). Normal distribution represents the distribution of many random variables as a symmetrical bell-shaped graph. The normal distribution is informally called the bell curve, however, many other distributions such as Cauchy, students and logistic are bell shaped . The probability density of normal distribution is:

$$f(x/\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

In particular, the distribution with μ=0 and σ=1 is called standard normal distribution or unit normal distribution denoted by:

$$f(x/0,1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\sim N(0,1) \quad (2)$$

The normal distribution is remarkably useful because of the central limit theorem. In the most general form, under some condition (which include infinite variance), it states the averages of random variables independently drawn from independent distributions converges in distribution to the normal, that is, become normally distributed when the number random variables is sufficiently large. There is a very strong connection between the size of a sample ***(n)*** and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large ***N*** can be approximated by the normal distribution even though the population distribution itself is not normal.

The distribution is the most important and most widely used distribution in statistics and sometimes called the "Gaussian distribution". Strictly speaking, it is not correct to talk about *"the normal"* distribution since there are many normal distributions. The distributions can differ in their means and in their standard deviations. It is a commonly used continuous probability distribution function in probability theory. Normal distributions are also important in statistics and social sciences for real-valued random variables whose distributions are not known because of the central limit theorem, which states that, *under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normal, irrespective of the form of the original distribution*. Many results and methods can be derived analytically in explicit form when the relevant variables are normally distributed.

Assessing the normality assumption is required by most statistical procedures. If the assumption of normality is violated, interpretation and inference may not be reliable or valid. Therefore it is importance to validate this assumption before proceeding with any relevant statistical procedure. In assessing random samples of independent observations for normality, Graphical Methods (histogram, boxplots, Q-Q plots), Numerical Methods (Skewness and Kurtosis), and Formal Normality tests (D'Agostino's K-squared test, Jarque–Bera test, Anderson–Darling test, Cramér–von Mises criterion, Lilliefors test for normality Kolmogorov–Smirnov test, Shapiro–Wilk test, Pearson's chi-squared test, and Shapiro–Francia test) are used. In this research, normality test is limited to the Shipiro Wilk test (SW), Kolmogorov Smirnov (KS) test, and Anderson Darling (AD) test.

METHODS

In statistical literature, there are over forty tests of normality available and researchers are coming up with new procedures and tests (Dufour *et al*., 1998). Pearson (1895) commenced the process of developing techniques to detect departures of observations from normality when he worked on the skewness and kurtosis coefficients (Althouse *et al*., 1998). Various normality tests differ in the characteristics *(skewness* & *kurtosis; characteristic function; and the linear relationship existing between the distribution of the variable and the standard normal variable, Z)* in terms of the normal distribution they focus on, Nornadiah and Bee (2011). The tests also differ in the level at which they compare the empirical distribution with the normal distribution, in the complexity of the test statistic and the nature of its distribution (Seier, 2002).

While Park (2008) categorised formal normality tests into two classes namely: (i) *descriptive statistics* (skewness and kurtosis coefficients) and (ii) *theory-driven methods* (*SW*, *KS* and *AD*). Seier (2002) classified the tests into four sub-divisions viz skewness & kurtosis test; empirical distribution test; regression & correlation test; and other special test. Arshad, Rasool and Ahamad (2003) categorization was into four categories which are (i) tests of chi-square types; (ii) moment ratio techniques; (iii) tests based on correlation; and (iv) tests based on the Empirical Distribution Function (EDF).

Empirical Distribution Function (EDF) Tests

The EDF tests of data normality is to compare the empirical distribution function which is estimated based on the data with the Cumulative Distribution Function (CDF) of normal distribution to see if there is a good agreement between them. EDF tests are techniques that are based on a measure of discrepancy between the empirical and hypothesized distributions. It is subdivided into those belong to ***supremum*** and ***square*** class of the discrepancies. Among most widely known EDF tests are KS-test and AD-test Arshad *et al.*, (2003) and Seier (2002).

Anderson Darling (AD) Test

It is a statistical test of whether or not a dataset comes from a certain pre-specified population. The test assumes that there are no parameters to be estimated in the distribution being tested, i.e. the test and its set of critical values is distribution free. For example, when testing if a set of random observations come from a normal distribution, we need not specifying the parameters (mean and standard deviation) of the distribution. When testing if a normal distribution adequately describes a set of data AD-test is one of the most powerful statistical tools for detecting most departure from normality and it gives more weight to the tail of the distribution, Farrel and Stewart (2006). In addition to its use as a test of fit for distribution, it is used in parameter estimation as the basis for a form of minimum distance estimation procedure. It is the most powerful of the EDF tests which is based on the squared difference $\{F^*(x) - F_n(x)\}^2$, Arshard *et al.* (2003). It tests the hypothesis:

H_0: The data follows the normal distribution

H_1: The data do not follow the normal distribution

Anderson and Darling (1954) gave the AD-statistic as:

$$AD = n\int_{-\infty}^{\infty}[F_x(x) - F^*(x)]^2 \emptyset[F^*(x)]dF^*(x) \quad (3)$$

where $\emptyset$ is a non-negative weight function which can be computed by $\emptyset = [F^*(x)(1 - F^*(x)]^{-1}$. Arshard *et al.* (2003) however simplify the formula to:

$$AD = -n - \frac{1}{n}\sum (2i-1)\ln F^*(x_i) + \ln\left(1 - F^*(x_{n+1-i}\right) \quad (4)$$

where $\boldsymbol{n}$ = sample size

$F^*(x_i)$ is the cumulative distribution function for the specified distribution

$x_i's$ are the ordered data (i[th] sample when the data is sorted in ascending order)

A modified AD-statistic was given by D'Agostino and Stephens (1986) which take the sample size $\boldsymbol{n}$ into consideration is given as:

$$AD^* = AD\left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \quad (5)$$

Kolmogorov Smirnov (KS) Test

KS-test belongs to the *supremum* class of empirical distribution function (EDF) statistics. Given $\boldsymbol{n}$ ordered data points, $x_1 < x_2 < x_3 \ldots < x_n$, KS-test is based on the maximum difference between the observed distribution and expected cumulative normal distribution, Conover (1999). The statistic quantifies a distance between the empirical distribution of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. Smaller maximum difference indicates the distribution is more likely to be normal. Given $\boldsymbol{n}$ ordered data points $x_1 < x_2 < x_3 \ldots < x_n$, the test statistic proposed by Kolmogorov (1933) is defined by Conover (1999) as:

$$D_n = Sup_x\left[F^*(x) - F_n(x)\right] \quad (6)$$

where *'sup'* indicates the supremum, i.e. the greatest. $F^*(x)$ is the hypothesized distribution function where as $F_n(x)$ is the EDF estimated based on random sample. In K-S test of normality $F^*(x)$ is taken to be a normal distribution with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$.

The K-S test statistic is meant for testing

H_0: Data is normally distributed.

H_1: Data is not normally distributed

Shapiro Wilk (SW) Test

Developed by Shapiro and Wilk (1965), SW-test was initially limited to sample size of less than 50 but recent applications has however shown that it competes favourably well with other normality tests. The SW-test however performs better if values in the data set are as different as possible. According to Althouse *et al.* (1998), the test was the first to detect departures from normality due to skewness, kurtosis or both. Also, because of its good power properties, it has become the preferred test, Mendels and Pala (2003). Given an ordered random sample $x_1 < x_2 < x_3 \ldots < x_n$, the SW-statistic tests the hypothesis:

H_0: Data is normally distributed.

H_1: Data is not normally distributed

The SW-test statistic given by Shapiro (1965) is defined as:

$$SW = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

where $\bar{x}$ is the sample mean

$x_{(i)}$ is the i[th] ordered statistic (order of the sample from the smallest to the largest)

$$a_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$$

where $m = (m_1, \ldots, m_n)^T$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and $\boldsymbol{V}$ is the covariance matrix of those order statistics.

The value of $\boldsymbol{SW}$ lies between zero and one with small values leading to the rejection of normality whereas a value of one indicates normality of the data. *SW* test was modified by Royston (1995) to broaden the restriction of the sample size by improving the approximation to the weights $\boldsymbol{a_i}$ which can be used for any n in the range $3 \leq n \leq 5000$.

$\boldsymbol{a_i}$ for different $\boldsymbol{n}$ can be obtained from the table of Shapiro-Wilk

RESULTS

Algorithm for Simulation/Analysis

- Generate $\boldsymbol{n}$ samples that are standard normally distributed.
- Test the $\boldsymbol{n}$ samples for normality using AD-test, KS-test and SW-test.
- Repeat the process for the $\boldsymbol{n}$ 10,000 times.
- Note the number of times $\boldsymbol{H_0}$ is rejected (x)
- Find $\frac{x}{10{,}000}$, the probability of rejecting a true $\boldsymbol{H_0}$
- Repeat the process by varying size of $\boldsymbol{n}$

Levels of significance (α = 5%) is considered in the study for sample sizes n =10, 20, 30, 40, 50, 100, 200, 300, 400, 500, and 1000. The sample sizes were selected at the point where the probability of error drastically changes.

Table 1: Probability of empirical type I error rate

		Normality Tests		
Sample size	n	AD	KS	SW
	10	0.0484	0.0454	0.0490
	20	0.0462	0.0456	0.0492
	30	0.0490	0.0520	0.0500
	40	0.0524	0.0528	0.0502
	50	0.0538	0.0522	0.0574
	100	0.0404	0.0478	0.0428
	200	0.0506	0.0528	0.0494
	300	0.0472	0.0494	0.0478
	400	0.0445	0.0453	0.0449
	500	0.0466	0.0490	0.0506
	1000	0.0496	0.0496	0.0502

Table 1 summarizes the empirical probability of type 1 for α = 5%. From the plot given in Figure 1, *AD* outperforms the other two tests. However, for sample size 20 or less, *KS* outperforms the rest.

The type *I* error rate of each test does not show any consistent pattern as the sample size increases. Hence, it can be inferred that sample size does not have significant impact on the

type *I* error of a test. The inference is justified based on the fact that α - the significance level is usually predetermined prior to a parametric test

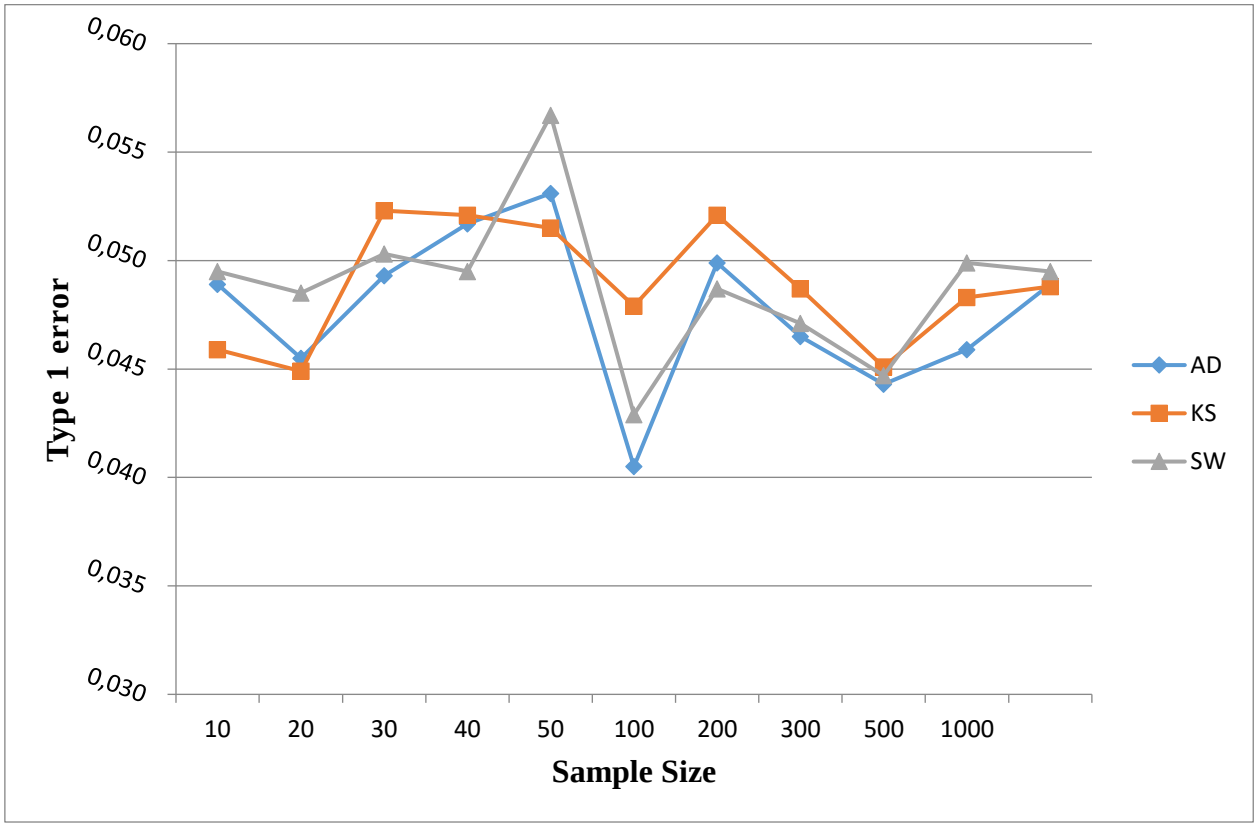


Figure 1: Probability of Type 1 Error for the Three Tests of Normality

Table 2: Rank of Probability of Empirical type I error

		Normality Tests		
Sample size	n	AD	KS	SW
	10	2	1	3
	20	2	1	3
	30	1	3	2
	40	2	3	1
	50	2	1	3
	100	1	3	2
	200	2	3	1
	300	1	3	2

	400	1	3	2
	500	1	2	3
	1000	1	1	3
	Sum of Ranks	16	24	25

The sum of ranks above showed that the *AD-test* is the best among the three normality tests because it has the lowest rank. It is followed by the *KS-test* and the *SW- test* has the largest sum of ranks and hence the poorest.

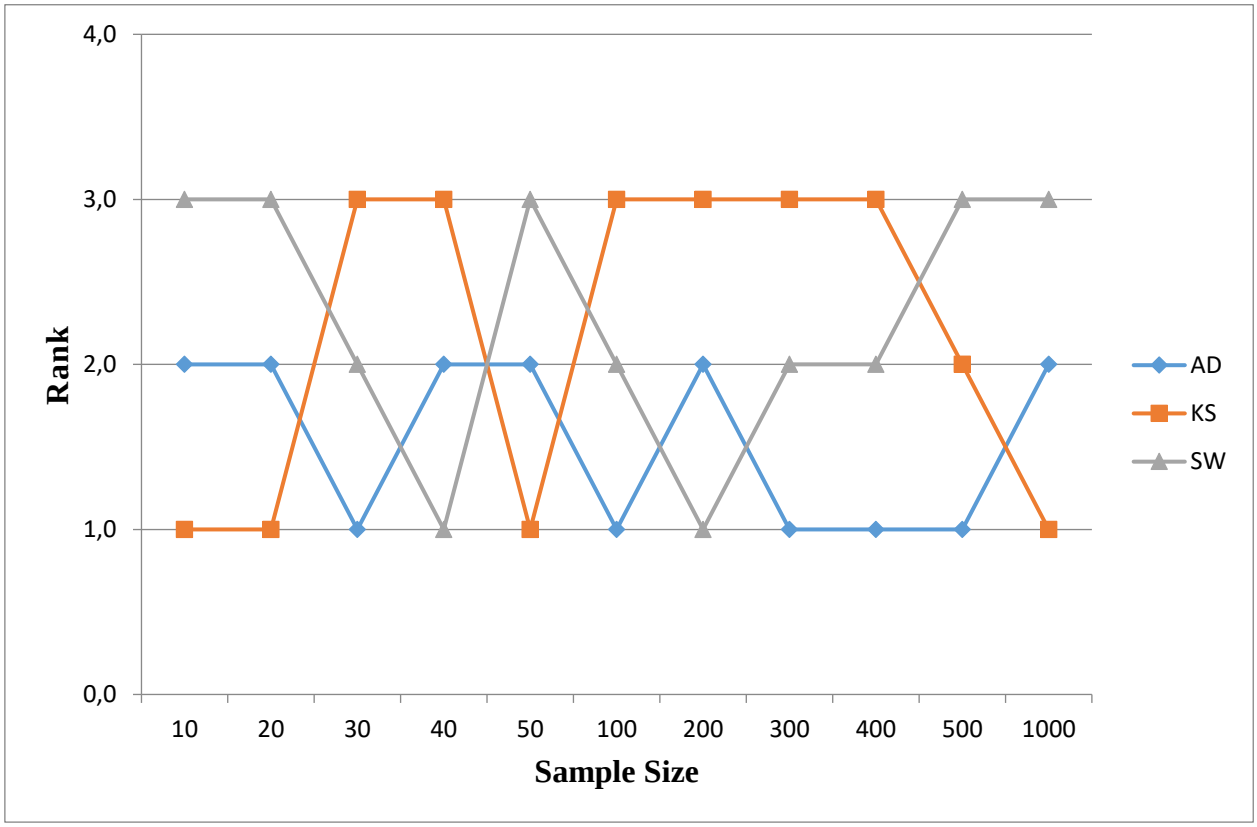


Figure 2: Rank of Probability of Type 1 Error

CONCLUSION

Random variables generated are normally distributed but there are discrepancies in the examination of the data for normality. The implication of this for researchers and users of data is to ensure verification of various assumptions underlying different statistical techniques in order for a valid conclusion of analysis. In general, it is noted that there is none among the tests that is uniformly most powerful. However, the Kolmogorov-

Smirnov test outperforms the other techniques for relatively small sample size while the Anderson-Darling test outperforms the other two tests on the average.

In conclusion, users of statistical data should not depend only on graphical techniques (like histogram and Q-Q plot) to determine normality of the data set. It is therefore recommended effort should be made to use any of the formal normality tests technique.

REFERENCES

Ademuyiwa, J. A., Sabri, S. R. M., & Adetunji, A. A. (2023). Modelling Count Variables: A Comparative Analysis of two Discretization Techniques. Asian Journal of Probability and Statistics, 25(2): 37-51. 10.9734/AJPAS/2023/v25i2551

Adetunji, A. A. & Sabri, S. R. M. (2023). An Alternative Count Distribution for Modelling Dispersed Observations, Pertanika Journal of Science and Technology, 31(3), 1587–1603. 10.47836/pjst.31.3.25

Althouse, L.A., Ware, W.B. & Ferron, J.M. (1998). Detecting Departures from Normality: A Monte Carlo Simulation of A New Omnibus Test based on Moments. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Anderson, T.W. & Darling, D.A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, *49*(268), 765-769.

Arshad, M., Rasool, M.T. & Ahmad, M.I. (2003). Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Pakistan Journal of Applied Sciences, 3*(2), 85-88.

Conover, W.J. (1999). *Practical Nonparametric Statistics*. Third Edition, John Wiley & Sons, Inc. New York, 428-433.

D' Agostino, R.B. & Stephens, M.A. (1986). *Goodness-of-fit Techniques*, NewYork: Marcel Dekker.

Dufour J.M., Farhat, A., Gardiol, L. & Khalaf, L. (1998). Simulation-based Finite Sample Normality Tests in Linear Regressions. *Econometrics Journal*, 1, 154-173.

Farrel, P.J. & Stewart, K.R. (2006). Comprehensive Study Of Tests For Normality And Symmetry: Extending The Spiegelhalter Test. *Journal of Statistical Computation and Simulation*, *76*(9), 9803–816.

Kolmogorov, A.N. (1933). *Sulla determinazione empirica di una legge di distribuzione*, Giornale dell' Instituto Italiano degli Attuari 4, 83–91.

Mendes, M. & Pala, A. (2003). Type I Error Rate and Power of Three Normality Tests. *Pakistan Journal of Information and Technology, 2*(2), 135-139.

Nornadiah, M.R. & Bee, W.Y. (2011): Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests, *Journal of Statistical Modeling and Analytics*, *2*(1), 21-33

Park, H.M. (2008). *Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*. Technical Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.

Royston, P. (1995). Remark AS R94:A Remark on Algorithm AS181:The W-test for Normality. *Journal of the Royal Statistical Society*, *44*(4), 547-551.

Seier, E. (2002). Comparison of Tests for Univariate Normality. *InterStat Statistical Journal*, 1, 1-17.

Shapiro, S.S. & Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4), 591-611. 10.1093/biomet/52.3-4.591