Original Research Article

A New Test for Normality Based on Balakrishnan – Sanghvi Divergence Measure

Abstract

In this paper, a more robust estimator of the Shannon entropy is applied in place of an earlier estimator, to obtain an improved goodness-of-fit test to normality which is based on the Balakrishnan-Sanghvi measure of divergence. The statistic is affine invariant and consistent against fixed alternatives. The critical values of the new statistic and those of a competing statistic as well as their power comparisons are obtained through extensive simulation study. The result of the power comparison showed that the statistic can be recommended as a good test for normality especially at small samples and against symmetric alternative distributions.

Keywords: Balakrishnan-Sanghvi divergence measure, empirical critical value, power of a test, Shannon entropy, test for normality.

1. Introduction

Goodness-of-fit test to normality has attracted the attention of scores of researchers, both in statistical methods and in applications. This is so because of the importance of the normal distribution in classical statistical analysis, where most statistical techniques depend on the assumption of normality of datasets. This may also be so due to the fact that most datasets approximate to normality especially at large sample sizes. As a result, there are in existence more than 80 known statistical techniques for assessing the normality of datasets. These range from suggestive procedures like the graphical approaches to pure statistical tests with known statistics.

Suppose a random variable X follows a normal distribution F_X with probability function f_X, given by:

$$f_X\left(x,\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\};-\infty<x<\infty,-\infty<\mu<\infty,\sigma^2>0 \qquad (1.1)$$

where μ and σ^2 are the parameters of the normal distribution , respectively known as the mean and variance of the distribution. A good number of unique characterizations have been obtained in the literature for the normal distribution in (1.1), such as, the characteristic function, the moment generating function, the entropy measure, measures of symmetry and kurtosis, and the behavior of its various transforms, to mention but a few. These unique characterizations have been employed extensively in developing goodness-of-fit statistics for assessing normality of datasets. One such characterization that is of interest in the present work is the entropy measure of the normal distribution.

Shannon [1] has obtained the entropy measure of a random variable X, with probability density function $f_X(x)$ as:

$$H(f_X) = -\int_{-\infty}^{\infty} f_X(x)\log\{f_X(x)\}dx \tag{1.2}$$

For the function in (1.1), the Shannon [1] measure in (1.2) has been obtained as:

$$H(f_X) = \frac{1}{2} + \log\left\{\sqrt{2\pi\sigma^2}\right\} = \frac{1}{2}\log\left\{2\pi\sigma^2 e\right\} \tag{1.3}$$

Now, a good number of authors have considered the problem of estimation of (1.2) in general and (1.3) in particular. Some of them include Vasicek [2], Dudewicz and van der Meulen [3], van Es [4], Ebrahimi et al. [5], Correa [6], Wieczorkowski and Grzegorzewski]7], Pasha et al. [8], Noughabi and Arghami [9], Park and Shin [10], Zamanzade and Arghami [11], Kohansal and Rezakhal [12], and Chaji and Zografos [13]. Most of these estimators, which have been obtained mainly based on spacings, nearest neighbor, kernels and quantile density, have been applied to goodness-of-fit tests to statistical distributions such as normal, beta, and exponential distributions.

In what appears to be an extension of the entropy measure in (1.2) for determining divergence between two statistical distributions, researchers have obtained entropy-based phi-divergence measure between two statistical distributions, f and g, as:

$$D_\Phi(f_X, g_X) = \int_{-\infty}^{\infty} \Phi\left(\frac{g_X(x)}{f_X(x)}\right)\{f_X(x)\}dx \tag{1.4}$$

Notable among the phi divergence measures in the literature include the Kullback and Leibler [14] and the Balakrishnan and Sanghvi [15], which are given as:

$$\begin{aligned} D_{KL}(f_X, g_X) &= \int_{-\infty}^{\infty} \log\left(\frac{f_X(x)}{g_X(x)}\right) f_X(x)dx; \text{ and} \\ D_{BS}(f_X, g_X) &= \int_{-\infty}^{\infty} \left(\frac{f_X(x) - g_X(x)}{f_X(x) + g_X(x)}\right)^2 f_X(x)dx \end{aligned} \tag{1.5}$$

respectively.

Recently, Tavakoli et al. [16] and Tavakoli et al. [17] have employed these divergence measures in testing for normality of datasets. They achieved this by taking $g_X(x)$ in (1.4) and (1.5) to be normally distributed, having the pdf in (1.1) with μ and σ^2 estimated from the dataset and taking $f_X(x)$ to have an unknown continuous distribution whose entropy measure is estimated by the Vasicek [2] estimator such that the resultant statistic is obtained as an integral of the appropriate phi-divergence measure. They showed that the statistics are both affine invariant and consistent against fixed alternatives, which are desirable properties of a good goodness-of-fit statistic. They also showed through empirical study that the statistics have relatively good power performances.

Now, it has been shown in the literature that Vasicek [2] estimator of the entropy measure of a distribution, $H(f_X)$, is not as good as some other newer estimators, see for instance, van Es [4]. As a result, it is expected that a test for normality that is based on the Vasicek [2] estimator will not be as good as that is based on an improved estimator. This therefore is the essence of this work. The rest of the paper is organized as follows: the statistic is developed in section two while its empirical critical values are obtained in section three. Section four gives the empirical power comparison of the statistic with other competing statistics while the paper is concluded in section five.

2. The statistic

Vasicek [2] showed that the entropy measure of a certain distribution as given in (1.2) can be rewritten as:

$$H(f_X) = \int_0^1 \log\left\{\frac{d}{dp}F^{-1}(p)\right\}dp \qquad (2.1)$$

where $p \in (0,1)$ is such that the inverse distribution function, $F^{-1}(p) = x$ for a strictly continuous x. By replacing *F* in (2.1) with F_n and using the difference operator in place of the differential operator, Vasicek [2] obtained an estimator of (1.2) as:

$$H_{mn} = \frac{1}{n}\sum_{j=1}^{n}\log\left\{\frac{n}{2m}\left(X_{(j+m)} - X_{(j-m)}\right)\right\} \qquad (2.2)$$

where $X_{(j)} = X_{(1)}$ for $j < 1$ and $X_{(j)} = X_{(n)}$ for $j > n$, and *m* is an integer such that $m < \frac{n}{2}$. Now, Tavakoli et al. [16] employed the Vasicek [2] estimator in (2.2) to obtain a statistic for testing normality in what follows.

From the Balakrishnan and Sanghvi [15] divergence measure, given in (1.5), it is obvious that $D_{BS}(f_X, g_X) \geq 0$ where equality holds if and only if $f_X(x) = g_X(x)$. Suppose $g_X(x)$ is a normal distribution with parameters μ and σ^2, which are estimated by $\hat{\mu} = n^{-1}\sum_{j=1}^{n} x_j$ and $\hat{\sigma}^2 = n^{-1}\sum_{j=1}^{n}(x - \hat{\mu})^2$ respectively, such that $g_X(x)$ is approximated by:

$$\hat{g}_X(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left\{-\frac{1}{2}\frac{(x-\hat{\mu})^2}{\hat{\sigma}^2}\right\}; -\infty < x < \infty \qquad (2.3)$$

By substituting (2.3) for $g_X(x)$ in (1.5), Tavakoli et al. [16] obtained:

$$D_{BS}(f_X, g_X) = \int_{-\infty}^{\infty} \left(\frac{f_X(x) - \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}}{f_X(x) + \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}} \right)^2 f_X(x)dx \tag{2.4}$$

and by taking $f_X(x)$ to have an unknown continuous distribution whose entropy measure is estimated by Vasicek [2] in (2.2) through the transformation in (2.1), they obtained

$$D_{BS}(f_X, g_X) = \int_{0}^{1} \left(\frac{\left(\frac{dF^{-1}(p)}{dp}\right)^{-1} - \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}}{\left(\frac{dF^{-1}(p)}{dp}\right)^{-1} + \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}} \right)^2 dp \tag{2.5}$$

which by replacing *F* by F_n and employing (2.2) gave rise to a statistic for testing the null distribution of the normality of $f_X(x)$, given by:

$$T(BS)_{mn} = \frac{1}{n}\sum_{j=1}^{n} \left(\frac{\frac{2m}{n\left(X_{(j+m)} - X_{(j-m)}\right)} - \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}}{\frac{2m}{n\left(X_{(j+m)} - X_{(j-m)}\right)} + \left(1\big/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2 \big/ 2\hat{\sigma}^2\right\}} \right)^2 \tag{2.6}$$

The statistic rejects the null distribution of normality for large values of the $T(BS)_{mn}$ since it is expected to tend to zero if $f_X(x)$ is normal.

Now, van Es [4], Ebrahimi et al. [5], and Noughabi and Arghami [9] have all obtained estimators of the Shannon [1] entropy which are improvements on the Vasicek [2] estimator. More recently, Al-Omari [18] obtained yet another improved estimator of (2.2), which is shown to have least absolute bias and smallest root mean square error (RMSE). It is given by:

$$HA_{mn} = \frac{1}{n}\sum_{j=1}^{n} \log\left\{ \frac{n}{c_j m}\left(X_{(j+m)} - X_{(j-m)}\right) \right\} \tag{2.7}$$

where $c_j = \begin{cases} 1+\frac{1}{2}; 1 \le j \le m, \\ 2; \quad m+1 \le j \le n-m, \\ 1+\frac{1}{2}; n-m+1 \le j \le n, \end{cases}$

$X_{(j-m)} = X_{(1)}$ for $j \le m$, $X_{(j+m)} = X_{(n)}$ for $j \le n-m$ and *m* has its usual meaning. As a result, an improved statistic for testing the normality of a dataset is obtained in this paper, based on the Al-Omari [18] entropy estimator, as:

$$TBSI_{mn} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\frac{c_j m}{n\left(X_{(j+m)} - X_{(j-m)}\right)} - \left(1/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2/2\hat{\sigma}^2\right\}}{\frac{c_j m}{n\left(X_{(j+m)} - X_{(j-m)}\right)} + \left(1/\sqrt{2\pi\hat{\sigma}^2}\right)\exp\left\{-(x-\hat{\mu})^2/2\hat{\sigma}^2\right\}}\right)^2 \qquad (2.8)$$

where c_j is as given in (2.7). It is obvious, from Tavakoli et al. [16], that the test is both affine invariant and consistent against fixed alternatives. The test rejects the null distribution of normality of datasets for large values of the statistic.

3. Empirical critical values of the test

Since the exact null distribution of the $TBSI_{mn}$ is not known as it is not the interest in this work, the empirical critical values of the statistic shall be obtained in this section. They are computed for different sample sizes, n, through extensive simulation study. Precisely, the critical values at the level of significance, α = 0.05 for n = 5, 6, 7, 8, 9, 10 (5) 100 are evaluated. N = 100,000 samples were generated from a standard normal distribution and the N values of the $TBSI_{mn}$ statistic are obtained from each generated set of N samples under each specified n and m. The α-level critical value of the test for each n and m is then obtained as the 100(1−α) percentile of the N values. In order to carry out this, an appropriate window size, m, for each sample size is needed to avoid a too lengthy computation for all the possible m. Wieczorkowski and Grzegorzewski [7] have obtained it as $m = \left[\sqrt{n} + 0.5\right]$, where $[y]$ is the integer part of y. This measure has however been criticized. This led Tavakoli et al. [16] to obtain appropriate m for each range of values. Since this work is similar but a mere improvement to their work, their method of determining appropriate m is adopted in this paper. The percentile values are presented in Table 1. It is important to note that no effort is made to obtain the critical values of the statistic for all the sample sizes n which may be encountered in real-life applications as such will definitely be an effort in futility. This is because it will be practicably impossible to have all of them computed and listed. As a result, the percentile values presented in Table 1 may be regarded appropriate for demonstration purposes, especially for power comparisons in section four.

Table 1. Empirical critical values of the *T(BS)* and the *TBSI* statistics

n	*T(BS)*	*TBSI*
5	0.2462	0.1934
6	0.2422	0.2018
7	0.2381	0.2002
8	0.2301	0.1950
9	0.2205	0.1878
10	0.1888	0.1548
15	0.1575	0.1332
20	0.1311	0.1097
25	0.1145	0.0995
30	0.1151	0.0997
35	0.1037	0.0925
40	0.0947	0.0863
45	0.0876	0.0815
50	0.0999	0.0925
55	0.0931	0.0879
60	0.0873	0.0837
65	0.0825	0.0801
70	0.0783	0.0769
75	0.0748	0.0739
80	0.0905	0.0894
85	0.0865	0.0861
90	0.0828	0.0834
95	0.0795	0.0807
100	0.0766	0.0781

4. Empirical power comparison

The power of a test is the ability of the test to take the right decision of rejecting a wrong null hypothesis. It is therefore expected that a statistic for assessing goodness-of-fit for normality will reject a null hypothesis of the normality of a dataset which is not drawn from a normal distribution. In this section, we shall investigate the relative ability of two statistics to reject normality when the true distribution is non-normal. The statistics are the *T(BS)* of Tavakoli et al. [16] and the *TBSI,* proposed in this paper. In order to carry out the power comparison objective of this section, seven different distributions are considered. We simulated 10,000 samples of each of the seven distributions at sample sizes n = 5, 10, 25, 50 and 100. For each simulated sample under each sample size, we calculated the values of the two competing statistics, giving rise to 10,000 values for each statistic, and estimated the power of each of the statistics as the percentage of the 10,000 samples that is rejected by the statistic at α = 0.05.

Three of the seven distributions considered are symmetric while the remaining four distributions are skewed. They include the standard normal, Laplace, and student's t as the symmetric distributions. Others are the exponential, Weibull, lognormal and the beta as skewed distributions. The results are presented in Table 2.

Table 2. Empirical powers of T(BS) and TBSI statistics with higher power values in bold, α = 0.05

Distribution	*n*	*T(BS)*	*TBSI*
Normal (0, 1)	5	4.8	4.9
	10	5.4	4.9
	25	5.0	4.9
	50	4.9	5.2
	100	5.0	4.8
Laplace (0, 1)	5	9.0	9.0
	10	9.7	**12.9**
	25	23.4	**38.7**
	50	27.1	**64.7**
	100	80.5	**90.6**
t (2)	5	13.9	**14.1**
	10	22.6	**26.6**
	25	54.3	**67.0**
	50	71.7	**91.0**
	100	98.3	**99.6**
Exponential (1)	5	17.7	**17.9**
	10	**49.9**	49.4
	25	**95.1**	92.9
	50	**100.0**	99.6
	100	100.0	100.0
Weibull (1, 2)	5	17.6	**18.1**
	10	49.8	**50.0**
	25	**95.2**	92.9
	50	**100.0**	99.6
	100	100.0	100.0
Lognormal (0, 1)	5	24.9	**25.4**
	10	63.4	**64.4**
	25	**98.1**	97.9
	50	100.0	100.0
	100	100.0	100.0
Beta (1, 5)	5	12.1	**12.5**
	10	33.2	**33.6**
	25	**85.6**	77.8
	50	**99.5**	95.4
	100	**100.0**	99.8

From Table 2, the two statistics have averagely the same power performance of 5% under the standard normal distribution. The null distribution in this case is the normal distribution, as such, the power performance should equal the level of significance, α = 5%. This is attained by the two competing statistics and as a result, they are said to have good control over type-one-error.

Under the remaining two symmetric alternative distributions of Laplace (Laplace (0, 1)) and student's t (t (2)), it is clear that the new proposed statistic is more powerful than its counterpart in almost all the sample sizes considered. Under the four skewed alternative distributions considered however, the new proposed statistic appeared to be more powerful than the T(BS) statistic at small sample sizes while it is either at par or less powerful than the T(BS) statistic at large sample sizes. This by extension could suggest that the Vasicek [2] estimator of the Shannon entropy approaches the Al-Omari [18] estimator at large sample sizes under skewed distributions.

5. Conclusion

In this paper, we have developed an alternative version of the Tavakoli et al. [16] statistic for assessing the normality of datasets. The affine invariance and consistency of the statistic are drawn from the Tavakoli et al. [16] statistic. The power performance of the test shows that is has a very good control over type-one-error and that it has relative good power performance, especially under symmetric alternative distributions. As a result, it can be regarded as a good statistic for testing normality of datasets.

References

[1] Shannon, CE. A mathematical theory of communications. Bell System Technical Journal. 1948;27:379- 423, 623656. doi:10.1002/bltj.1948.27.issue-3.

[2] Vasicek O. A test for normality based on sample entropy. Journal of the Royal Statistical Society B. 1976;38:54-59.

[3] Dudewicz EJ, van der Meulen EC. Entropy-based tests of uniformity. Journal of the American Statistical Association. 1981;76:967-974.

[4] Van Es B. Estimating functionals related to a density by class of statistics based on spacings. Scandinavian Journal of Statistics. 1992;19:61-72.

[5] Ebrahimi N, Pflughoeft K, Soofi ES. Two measures of sample entropy. Statistics & Probability Letters. 1994;20:225-234. doi:10.1016/0167-7152(94)90046-9.

[6] Correa JC. A new estimator of entropy. Communication in Statistics-Theory and Methods. 1995;24(10):2439-2449. doi:10.1080/03610929508831626.

[7] Wieczorkowski R, Grzegorzewsky P. Entropy estimators improvements and comparisons. Communication in Statistics – Simulation and Computation. 1999;28(2):541-567. doi:10.1080/03610919908813564.

[8] Pasha E, Kokabi Nezhad M, Mohtashami GR. A version of the entropy estimator via spacing. Iranian International Journal of Sciences. 2005;6(1):119–129.

[9] Noughabi HA, Arghami NR. A new estimator of entropy. Journal of the Iranian Statistical Society. 2010;9(1):53-64.

[10] Park S, Shin DW. On the choice of nonparametric entropy estimator in entropy-based goodness-of-fit test statistics. Communications in Statistics - Theory and Methods. 2012;41(5):809-819, DOI: 10.1080/03610926.2010.531365.

[11] Zamanzade E, Arghami NR. Testing normality based on new entropy estimators. Journal of Statistical Computation and Simulation. 2012;82(11):1701-1713.

[12] Kohansal A, Rezakhah S. Modified entropy estimators for testing normality. Journal of Statistical Computation and Simulation. 2015; DOI: 10.1080/00949655.2015.1025270.

[13] Chaji A, Zografos K. An estimator of Shannon entropy of beta-generated distributions and a goodness of fit test. Communications in Statistics - Simulation and Computation. 2017; DOI: 10.1080/03610918.2017.1381739.

[14] Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics. 1951;22:79–86.

[15] Balakrishnan V, Sanghvi LD. Distance between populations on the basis of attribute. Biometrics. 1968;24:859–865.

[16] Tavakoli M, Arghami N, Abbasnejad M. A goodness of fit test for normality based on Balakrishnan-Sanghvi information. Journal of the Iranian Statistical Society, 2019;18(1):177-190.

[17] Tavakoli M, Alizadeh Noughabi H, Borzadaran GRM. An estimation of phi divergence and its application in testing normality. Hacettepe Journal of Mathematics and Statistics. 2020;49 (6):2104-2118.

[18] Al-Omari AI. Estimation of entropy using random sampling. Journal of Computation and Applied Mathematics. 2014;261:95-102. doi:10.1016/j.cam.2013.10.047.361.