Original Research Article

A new estimator of Shannon entropy with application to goodness-of-fit test to normal distribution

Abstract

In this paper, a new estimator of the Shannon entropy of a random variable X having a probability density function $f(x)$ is obtained. Under the standard normal, standard exponential and uniform distributions, the estimator is shown to have relative low bias and low RMSE through extensive simulation study at sample sizes 10, 20, and 30. Based on the results, it is recommended as a good estimator of the entropy. Also, the new estimator is applied in goodness-of-fit test to normality and the results show that it is a good statistic for assessing univariate normality of datasets.

Keywords: entropy estimator, window size spacing, bias of an estimator, root mean square error of an estimator, test for normality.

1 Introduction

The entropy of a random variable, X, which has a distribution function is function $F(x)$ and a probability density function $f(x)$ is a function which measures the amount of information in the random variable. It is denoted by $H(f)$ and Shannon [11] defines it as:

$$H(f) = -\int_{-\infty}^{\infty} f(x) log \{f(x)\} \, dx \tag{1.1}$$

It has been established that different random variables give different entropy functions obtained from (1.1). For instance, suppose X is a normal random variable with mean, μ and variance σ^2, the entropy of X has been obtained as $H(f) = \frac{1}{2} + ln\left(\sigma\sqrt{2\pi}\right) = \frac{1}{2}ln\left(2\pi e\sigma^2\right)$ and that of a random variable X which is exponential with probability function $f(x) = \lambda e^{-\lambda x}; x > 0, \lambda > 0$ is $H(X) = -ln\lambda + (1 - ln\lambda)$. As a result, entropy function can be used to characterize different random variables (different distributions).

In real life however, it is usually not known the distribution from where datasets are generated. Hence, it is of utmost importance to obtain an estimator of the population parameter. Several researchers such as Vasicek [12], Ebrahimi et al. [6], Wieczorkowski and Grzegorzewski [13], Alizadeh Noughabi [1], Alizadeh Noughabi and Arghami [2], Zamanzede and Arghami [14], Al-Omari [4], Lombardi and Pant [9], Kohansal and Rezakhah [8], and Bitaraf et al. [5], to mention but a few, have devoted research attention to this direction. These estimators are obtained using different nonparametric approaches such as window size (m) spacings, kernel density estimation, nearest neighbour technique, and so on. The window size approach, no doubt, has attracted more researchers the rest of other approaches put together. Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a random variable X whose probability law is $f(x)$ with distribution function $F(x)$. Also, let the sample order statistics obtained from the random sample be $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ such that $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. For a non-negative integar m such that $m = 0, 1, 2, \ldots, \frac{n}{2}$, the sample m spacing is defined on jth order statistic by $X_{(j+m)} - X_{(j-m)}$. Using such spacings, Vasicek [12] in his pioneer work in this direction, estimated the derivative of $F^{-1}(p)$ by the slope given as:

$$\left[X_{(j+m)} - X_{(j-m)}\right] \frac{n}{2m} \tag{1.2}$$

for $\frac{(j-1)}{n} < p < \frac{j}{n}; j = m+1. m+2, \ldots, n-m$ after transforming (1) in the form:

$$H(f) = \int_0^1 log\left\{\frac{d}{dp}F^{-1}(p)\right\} dp\,;\, p \in (0,1) \tag{1.3}$$

to birth an estimator given by:

$$HV_{mn} = \frac{1}{n}\sum_{j=1}^{n} log\left\{\frac{n}{2m}\left[X_{(j+m)} - X_{(j-m)}\right]\right\}; \tag{1.4}$$

where $X_{(j)} = X_{(1)}$ for $j < 1$ and $X_{(j)} = X_{(n)}$ for $j > n$. Although this estimator is biased, like all the other non-parametric estimators, it is conceived to have a theoretically sound background to making a good estimator of $H(f)$.

Soon after the development of the estimator in (1.4), a number of researchers criticised the slope in (1.2) as incorrect when $j \leq m$ or $j > n-m+1$. This triggered different modifications of the slope to arrive at different estimators of the Shannon entropy with comparatively better results. Examples of such estimators obtained in this light include the Ebrahimi et al. [6], Alizadeh Noughabi and Arghami [2] and Al-Omari [3, 4]. Till present however, there has not been a generally accepted estimator with respect to correctness of the slope when $j \leq m$ or $j > n - m + 1$, hence this research paper.

In this work, a modification is introduced to the Bitaraf et al. [5] approach to obtain a competitively new estimator of the Shannon entropy which is applied to goodness-of-fit test to normality of datasets. In what follows, the new estimator is obtained in Section 2. Section 3 gives the Monte Carlo comparison of the estimator with some other existing ones. The application to testing for normality is carried out in Section 4 while the paper is concluded in Section 5.

2 The new Estimator

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a continuous distribution $F(x)$ having a probability function $f(x)$. Also, let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of the sample such that $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Using the sample order statistics, Ebrahimi et al. [6] modified the Vasicek estimator to propose a new estimator given by:

$$HEb_{mn}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} log\left\{\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i)}\right)\right\}, \tag{2.1}$$

where $c_i = \begin{cases} 1+\frac{i-1}{m}, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1+\frac{n-i}{m}, & n-m+1 \le i \le n, \end{cases}$ $X_{(i-m)} = X_{(1)}$ for $i \le m$ and $X_{(i+m)} = X_{(n)}$ for $i \ge n-m$. Then, Bitaraf et al. [5] intoduced an internal jth spacings, in addition to the mth spacings, thereby changing the normal slope in (1.2) to have:

$$T_{i.} = \frac{1}{2}\sum_{j=0}^{1} T_{ij};\ T_{ij} = \frac{n}{w_j(m-j)}\left\{X_{(i+m-j)} - X_{(i-m+j)}\right\} \tag{2.2}$$

for $\frac{(i-1)}{n} < p < \frac{i}{n}$; $i = m+1. m+2, \ldots, n-m$, where $w_j = \begin{cases} 1, & 1 \le i \le m-j \\ 2, & m-j+1 \le i \le n-m+j \\ 1, & n-m+j+1 \le i \le n \end{cases}$ and $j = 0, 1$. Based on (2.2), they obtained an estimator of the Shannon entropy as:

$$HB_{mn} = \frac{1}{n}\sum_{i=1}^{n} log\left\{T_{i.}\right\} \tag{2.3}$$

and showed empirically that the estimator has smaller RMSE and bias than the Ebrahimi et al. [6]. Due to the interesting properties of the jth internal spacing estimator of Bitaraf et al. [5], a new estimator of the entropy is proposed, which modifies the weighting factor of the Bitaraf et al. [5] estimator. The proposed estimator is given by:

$$HM_{mn} = \frac{1}{n}\sum_{i=1}^{n} log\left\{\frac{1}{2}\left(T_{i0} + T_{i1}\right)\right\} \tag{2.4}$$

where $T_{ij} = \frac{n}{w_{ij}(m-2j)}\left\{X_{(i+m-j)} - X_{(i-m+j)}\right\}$ and $w_{ij} = \begin{cases} 1+\frac{1}{3}, & 1 \le i \le m-j \\ 2, & m-j+1 \le i \le n-m+j \\ 1+\frac{1}{3}, & n-m+j+1 \le i \le n \end{cases}$; $X_{(i-m+j)} = X_{(1)}$ for $i \le m-j$ and $X_{(i+m-j)} = X_{(n)}$ for $i \ge n-m+j$, $\frac{n}{m} \to 0$, $j = 0, 1$

Theorem 2.1: For a random sample $X_1, X_2, \ldots, X_n$ from a continuous distribution $F(x)$ with a probability density function $f(x)$, $HM_{mn} \overset{Pr}{\to} H(f)$ as $n \to \infty$, $m \to \infty$, and $\frac{n}{m} \to 0$.

Proof: Using a property of the convex function, Bitaraf et al. [5] obtained the inequality:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(loga_i + logb_i\right) \le \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{1}{2}(a_i + b_i)\right) \le \frac{1}{n}\sum_{i=1}^{n} max\left\{loga_i, logb_i\right\}; \tag{2.5}$$

where $a_i, b_i \in R^+$. Taking T_{i0} for a_i and T_{i1} for b_i and following the same principle with Bitaraf et al. [5], we have the extreme terms of the inequality in (2.5) presented as:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(log\left\{\frac{n\left(X_{(i+m)} - X_{(i-m)}\right)}{2m}\right\} + log\left\{\frac{n\left(X_{(i+m-1)} - X_{(i-m+1)}\right)}{2(m-1)}\right\} + log\left\{\frac{2}{w_{i0}}\right\} + log\left\{\frac{2}{w_{i1}}\right\}\right)$$
$$\le \frac{1}{n}\sum_{i=1}^{n} max\left(log\left\{\frac{n\left(X_{(i+m)} - X_{(i-m)}\right)}{2m}\right\}, log\left\{\frac{n\left(X_{(i+m-1)} - X_{(i-m+1)}\right)}{2(m-1)}\right\}\right)$$

Since $w_{ij} \in (1,2)$, $(2n)^{-1}\sum_{i=1}^{n} log\left\{\frac{2}{w_{i0}}\right\}$ and $(2n)^{-1}\sum_{i=1}^{n} log\left\{\frac{2}{w_{i1}}\right\}$ converges to 0. Also, there exists $X_{ij} \in \left(X_{(i-m+j)}, X_{(i+m-j)}\right)$ such that

$$\frac{F\left(X_{(i+m-j)}\right) - F\left(X_{(i-m+j)}\right)}{X_{(i+m-j)} - X_{(i-m+j)}} = f\left(X_{ij}\right); j = 0, 1.$$

The rest of the proof collapses into the proof of theorem 2.1 of Bitaraf et al. [5]. Hence, the proof is completed.

3 Empirical comparison of entropy estimators

In this section, the performance of some selected three estimators of the Shannon entropy is compared with the proposed estimator through simulation studies. The selected estimators include the Vasicek [12], Ebrahimi et al. [6], and the Bitaraf et al. [5] estimators, which are presented in this work as HV_{mn}, HE_{mn}, and HB_{mn} respectively. Precisely, the empirical root mean square error (RMSE) and the bias of these three competing estimators are compared with those of the proposed HM_{mn} estimator based on 10,000 samples of sizes 10, 20, and 30. These samples are obtained from three different continuous distributions, namely: the standard normal, uniform in the interva (0, 1), and the standard exponential. In each of the sample sizes, the study is conducted for $m = 1, 2, \ldots, \frac{n}{2}$. Although Grzegorzewski and Wieczorkowski [7] have obtained an optimal m as $m^* = [\sqrt{n} + 0.5]$, but it has been criticised as being incorrect, hence our choice of using all the possible m for each sample size.

From each sample size under each distribution with a known measure of entropy θ, 10,000 samples are generated and the estimated entropy measure, $\hat{\theta}$, is computed from each of the samples using the four competing estimators. Then, the RMSE and bias of each of the estimators is obtained by:

$$RMSE = \left(\frac{1}{10000} \sum_{j=1}^{10000} \left\{ \hat{\theta}_j - \theta \right\}^2 \right)^{\frac{1}{2}} ;\ bias = \frac{1}{10000} \sum_{j=1}^{10000} \hat{\theta}_j - \theta. \tag{3.1}$$

The results are presented in Tables 1, 2, and 3 for normal, uniform and exponential distributions respectively.

From the results in Tables 1, 2, and 3, a number of important facts about the general behaviour of the Shannon entropy estimators as well as specific estimators are established. Firstly, the efficiency behaviour of the estimators depends on the distribution from where the sample is obtained. This is because the minimum RMSE of each of the competing estimators under a specified sample size varies across the distributions. For instance, at sample size of 10, HV_{mn}, recorded minimum RMSEs of 0.5955, 0.5673, and 0.4520 under normal, exponential and uniform distributions respectively. At a sample size of 20, the observed minimum RMSEs are 0.3662, 0.3563, and 0.2702 under normal, exponential and uniform distributions respectively. In a similar manner, minimum RMSEs of 0.2734, 0.2719, and 0.2043 were obtained for the same HV_{mn} at sample size of 30 under the same respective distributions. For the HE_{mn}, ordered sets of minimum RMSEs, in the manner of the HV_{mn} were obtained as (0.3852, 0.3794, 0.2058); (0.2353, 0.2484, 0.1061); and (0.1733, 0.1988, 0.0724) respectively for sample sizes 10, 20, and 30 respectively. For the HB_{mn} estimator, the ordered sets of the minimum RMSEs are (0.2591, 0.3619, 0.1704); (0.1800, 0.2418, 0.0878) and (0.1432, 0.1957, 0.0619) respectively for sample sizes $n = 10$, 20, and 30 while those of the new estimator, HV_{mn}, are (0.2747, 0.3727, 0.1664); (0.1781, 0.2531, 0.0959) and (0.1429, 0.2102, 0.0735) respectively for sample sizes $n = 10$, 20, and 30.

Secondly, the RMSEs for each of the estimators across the distributions considered decrease with increasing sample size. This is however expected for every good estimator. Again, it is observed that the window size m at which the RMSE is a minimum (also known as the optimal window size) for a specified sample size varies from one distribution to another and from one estimator to another. This is against Grzegorzewski and Wieczorkowski [7], who proposed optimal window size $m^* = [\sqrt{n} + 0.5]$, where $[x]$ is the integral part of x. For each of the estimators, the optimal window size, m^*, is presented in Table 4 for each sample size under different distributions with that of Grzegorzewski and Wieczorkowski [7] denoted by $m^*(GW)$. The results in Table 4 show that all the estimators, except the HV_{mn}, maintained optimal window size generally greater than the $m^*(GW)$ in almost all the distributions considered.

In addition to the above, the results in Tables 1, 2, and 3 show that HE_{mn}, HB_{mn}, and HM_{mn} generally have smaller RMSEs and bias values than the HV_{mn} in all the sample sizes and distributions considered. As a result, they can be regarded as bias corrected and generally more efficient estimators

Table 1: Empirival root mean square errors and biases of some Shannon entropy estimators under the standard normal distribution with $\theta = 1.4189$

n	m	RMSEs				Biases			
		HV_{mn}	HE_{mn}	HB_{mn}	HM_{mn}	V_{bias}	E_{bias}	B_{bias}	M_{bias}
10	1	0.6719	0.5559	–	1.2042	-0.6718	-0.3508	–	-1.1068
	2	0.5955	0.4297	0.4115	–	-0.1791	-0.4041	-0.1882	–
	3	0.6190	0.4045	0.3239	0.2789	-0.6090	-0.0066	-0.1761	-0.2118
	4	0.6630	0.3936	0.2781	0.2747	-0.4730	-0.3832	0.2768	-0.4054
	5	0.7165	0.3852	0.2591	0.2864	-0.2858	-0.1560	0.3527	-0.2146
20	1	0.4859	0.4243	–	1.1013	-0.4881	-0.4188	–	-1.1863
	2	0.3770	0.2957	0.2964	–	-0.2671	-0.1690	-0.6345	–
	3	0.3662	0.2631	0.2220	0.2711	0.2124	0.3404	-0.2391	0.4778
	4	0.3762	0.2485	0.1912	0.1894	-0.4731	-0.3147	-0.0624	0.2219
	5	0.3924	0.2395	0.1800	0.1782	-0.8076	-0.6187	-0.0106	0.0491
	6	0.4198	0.2408	0.1818	0.1781	-0.2402	-0.0207	-0.1301	0.0298
	7	0.4453	0.2375	0.1898	0.1808	-0.7344	-0.4843	0.0316	0.0711
	8	0.4736	0.2389	0.2063	0.1806	-0.1426	0.1381	0.3092	-0.2776
	9	0.5018	0.2380	0.2290	0.1803	-0.5254	-0.2141	0.3924	-0.0604
	10	0.5268	0.2353	0.2549	0.1791	-0.6992	-0.3573	0.0472	0.1024
30	1	0.4130	0.3710	–	1.0550	-0.5199	-0.4737	–	-1.1980
	2	0.2987	0.2441	0.2515	–	-0.3975	-0.3321	-0.3971	–
	3	0.2749	0.2072	0.1879	0.2752	-0.4684	-0.3830	-0.0931	0.5519
	4	0.2734	0.1918	0.1587	0.1735	-0.1492	-0.0436	-0.0146	0.0348
	5	0.2831	0.1854	0.1474	0.1488	-0.0057	0.1202	0.1881	0.2138
	6	0.2904	0.1788	0.1432	0.1443	-0.2438	-0.0972	-0.2126	0.0689
	7	0.3051	0.1770	0.1528	0.1429	-0.3850	-0.1683	0.0965	-0.1639
	8	0.3209	0.1768	0.1620	0.1478	-0.3329	-0.1458	-0.1017	-0.0268
	9	0.3376	0.1753	0.1807	0.1466	-0.6350	-0.4275	0.1052	-0.1991
	10	0.3553	0.1762	0.1977	0.1510	-0.3276	-0.0996	0.0971	-0.3528
	11	0.3728	0.1758	0.2139	0.1506	-0.3077	-0.0593	0.0383	-0.0390
	12	0.3871	0.1733	0.2362	0.1527	-0.4180	-0.1492	0.2273	-0.2422
	13	0.4091	0.1741	0.2597	0.1555	-0.4504	-0.1611	0.2060	-0.0603
	14	0.4274	0.1748	0.2805	0.1552	-0.5827	-0.2730	0.1564	0.0003
	15	0.4479	0.1764	0.3037	0.1579	-0.4534	-0.1233	0.2449	-0.0288

Table 2: Empirical root mean square errors and biases of some Shannon entropy estimators under the standard exponential distribution with $\theta = 1$

n	m	RMSEs				Biases			
		HV_{mn}	HE_{mn}	HB_{mn}	HM_{mn}	V_{bias}	E_{bias}	B_{bias}	M_{bias}
10	1	0.6742	0.5704	–	1.2278	-0.6411	-0.2621	–	-1.0727
	2	0.5777	0.4376	0.4271	–	-0.2224	0.0063	0.2789	–
	3	0.5673	0.3984	0.3619	0.4002	-0.3692	-0.2055	0.2221	0.5587
	4	0.5738	0.3905	0.3633	0.3737	-0.0775	0.2163	0.3187	0.5347
	5	0.5970	0.3794	0.4183	0.3727	-0.5236	-0.1298	0.0118	-0.5849
20	1	0.4874	0.4297	–	1.0949	-0.4684	-0.4558	–	-1.4015
	2	0.3774	0.3107	0.3201	–	-0.3003	-0.2719	0.3241	–
	3	0.3583	0.2790	0.2528	0.3363	-0.3846	-0.0740	-0.3377	0.2598
	4	0.3555	0.2613	0.2418	0.2671	-0.3429	-0.3365	-0.3971	-0.1285
	5	0.3563	0.2529	0.2517	0.2531	-0.1613	-0.3875	0.0975	0.0736
	6	0.3572	0.2496	0.2799	0.2643	-0.4417	0.0610	0.3004	0.1359
	7	0.3675	0.2484	0.3176	0.2679	-0.1151	0.0599	0.2417	0.0084
	8	0.3769	0.2566	0.3605	0.2773	-0.1169	0.0393	0.2599	0.1986
	9	0.3818	0.2628	0.4066	0.2960	-0.0756	-0.0904	0.3443	0.3686
	10	0.3954	0.2719	0.4680	0.3139	-0.1185	0.1554	0.4483	0.5568
30	1	0.4207	0.3802	–	1.0530	-0.5638	-0.2902	–	-0.9760
	2	0.3065	0.2629	0.2720	–	-0.1007	0.0792	-0.1688	–
	3	0.2827	0.2285	0.2150	0.3254	-0.2235	-0.2701	-0.1336	0.4435
	4	0.2736	0.2125	0.1975	0.2283	-0.2235	-0.1100	-0.3329	0.0932
	5	0.2722	0.2048	0.1957	0.2132	-0.5552	-0.2931	0.3241	0.3032
	6	0.2719	0.1996	0.2073	0.2102	-0.2572	-0.1826	-0.1514	-0.0942
	7	0.2752	0.1988	0.2300	0.2128	-0.1886	-0.2609	0.0404	0.1135
	8	0.2802	0.2026	0.2543	0.2162	-0.2200	0.1126	0.0583	-0.1477
	9	0.2777	0.2021	0.2859	0.2305	-0.0324	0.2507	0.0171	-0.0296
	10	0.2861	0.2082	0.3218	0.2388	-0.4808	0.0820	0.1876	-0.1442
	11	0.2896	0.2141	0.3571	0.2492	-0.4216	-0.0712	-0.0542	0.1880
	12	0.2953	0.2211	0.3936	0.2652	0.0961	0.2786	0.5240	0.2406
	13	0.2979	0.2312	0.4384	0.2811	-0.0932	0.1771	0.1063	0.0724
	14	0.3044	0.2422	0.4731	0.2983	-0.2753	0.3789	0.3739	0.6182
	15	0.3031	0.2543	0.5213	0.3179	0.0411	-0.2888	0.3556	0.1823

Table 3: Empirical root mean square errors and biases of some Shannon entropy estimators under the uniform distribution in the interval 0 and 1, with $\theta = 0$

n	m	RMSEs				Biases			
		HV_{mn}	HE_{mn}	HB_{mn}	HM_{mn}	V_{bias}	E_{bias}	B_{bias}	M_{bias}
10	1	0.5698	0.4461	–	1.1567	-0.5854	-0.2439	–	-1.1061
	2	0.4520	0.2839	0.2712	–	-0.2184	0.0144	-0.1331	–
	3	0.4534	0.2308	0.1769	0.2590	-0.5832	-0.2226	0.0942	0.2940
	4	0.4833	0.2165	0.1704	0.1769	-0.3333	-0.0827	0.1686	-0.0642
	5	0.5305	0.2058	0.2255	0.1664	-0.5598	0.1308	0.2110	0.0939
20	1	0.4193	0.3555	–	1.0585	-0.2526	-0.2007	–	-0.9864
	2	0.2884	0.2017	0.2081	–	-0.1701	-0.1565	-0.2188	–
	3	0.2702	0.1556	0.1195	0.2666	-0.4173	-0.0834	-0.1081	0.1787
	4	0.2751	0.1324	0.0878	0.1421	-0.1140	-0.0069	0.0012	0.2632
	5	0.2904	0.1231	0.0990	0.1104	-0.1942	-0.0315	0.1244	0.0452
	6	0.3105	0.1131	0.1298	0.1001	-0.3145	-0.0148	0.0864	0.1174
	7	0.3326	0.1086	0.1677	0.0959	-0.3174	-0.2053	0.1870	0.0574
	8	0.3599	0.1083	0.2084	0.0969	-0.2569	-0.0794	0.1652	0.1306
	9	0.3858	0.1077	0.2479	0.1002	-0.3297	-0.0170	0.2803	-0.1262
	10	0.4115	0.1061	0.2900	0.1038	-0.4326	-0.0760	0.26462	0.1206
30	1	0.3697	0.3245	–	1.0264	-0.3457	-0.3587	–	-1.0021
	2	0.2352	0.1757	0.1883	–	-0.1430	-0.1442	-0.1214	–
	3	0.2080	0.1314	0.1062	0.2758	-0.1615	-0.0652	-0.1184	0.1452
	4	0.2043	0.1094	0.0678	0.1402	-0.1870	-0.0786	-0.1437	0.1991
	5	0.2116	0.0963	0.0619	0.0996	-0.1363	0.0008	0.0485	0.1163
	6	0.2217	0.0895	0.0781	0.0837	-0.3114	-0.0610	-0.0716	0.1146
	7	0.2344	0.0835	0.1019	0.0781	-0.2111	0.0190	0.0827	0.1617
	8	0.2490	0.0784	0.1288	0.0745	-0.2545	-0.0812	0.1460	0.0842
	9	0.2658	0.0764	0.1563	0.0735	-0.2719	-0.2071	0.1949	0.0441
	10	0.2827	0.0756	0.1828	0.0761	-0.3348	-0.0527	0.1564	0.0251
	11	0.3013	0.0727	0.2101	0.0775	-0.3395	-0.0082	0.2102	0.1134
	12	0.3194	0.0724	0.2376	0.0800	-0.3262	-0.0182	0.3247	0.0174
	13	0.3369	0.0732	0.2656	0.0831	-0.3025	-0.0229	0.1635	0.1199
	14	0.3557	0.0728	0.2924	0.0887	-0.3883	-0.1183	0.2449	0.1052
	15	0.3740	0.0726	0.3179	0.0915	-0.3555	-0.0375	0.2767	0.0189

Table 4: Optimal window size of different estimators in comparison with $m^*(GW) = [\sqrt{n} + 0.5]$

n	$m^*(GW)$	$m^*(HV_{mn})$	$m^*(HE_{mn})$	$m^*(HB_{mn})$	$m^*(HM_{mn})$
10	3	2	5	5	4
20	5	3	10	5	6
30	6	4	12	6	7
10	3	3	5	3	5
20	5	6	7	4	5
30	6	6	7	5	6
10	3	2	5	4	5
20	5	3	10	4	7
30	6	4	12	5	9

than the HV_{mn}. Specifically, the HB_{mn} and HM_{mn} recorded least RMSEs in all the sample sizes and distributions considered except in the standard exponential distribution, where the HE_{mn} recorded lower RMSEs than the HM_{mn} in sample sizes $n = 20$ and 30. Finally, the new HM_{mn} estimator recorded higher RMSEs than the HB_{mn} in the sample size of $n = 10$ of the normal distribution and $n = 20$ and 30 of the uniform distribution. Based on these results, the proposed estimator can be regarded as a good estimator of the Shannon entropy.

4 Test of normality based on the new estimator

It is well known that among all the statistical distributions that possess a density function $f(x)$ and have a given variance, σ^2, the entropy, $H(f)$, is maximized by the normal distribution. Based on this, Vasicek [12] obtained a statistic for testing the assumption of normality. Now, following the procedure of Vasicek [12], a new test for testing the normality of datasets is obtained by

$$M_{mn} = \frac{1}{s} \prod_{i=1}^{n} \left\{ \frac{n\left(X_{(i+m)} - X_{(i-m)}\right)}{2w_{i0}m} + \frac{n\left(X_{(i+m-1)} - X_{(i-m+1)}\right)}{2w_{i1}(m-2)} \right\}^{\frac{1}{2}} \tag{4.1}$$

where $s^2 = n^{-1} \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2$. The statistic is consistent and affine invariant.

The M_{mn} statistic is applied to test for the normality of four different datasets. The first three datasets are simulated from the standard normal, standard exponential, and standard lognormal distributions, havine sample size of 30 in each case while the last dataset is the petal length of the iris setosa dataset, see Seber [10]. Expectedly, the test rejected the normality of all the datasets at 5% level of significance, except the statndard normal distribution. The results therefore show that the statistic is a good test for testing normality of datasets.

5 Conclusion

Several estimators of the Shannon entropy already exists in the literature via window size spacings. All the estimators are however biased with varied degrees of biasedness. Also, the ability of these estimators to be appropriate for the true population entropy depends largely on their relative efficiencies, measured in terms of the MSEs (or RMSEs). In this study, the new bias corrected estimator was obtained by a modification of the Bitaraf et al. [5] estimator. In addition to having relatively low bias, the new estimator also has relatively low RMSE compared to some other estimators. As a result, it can be recommended as a good estimator of the Shannon entropy. Also, the M_{mn} statistic obtained for testing normality which is based on new estimator can be recommended as a good test for normality of datasets.

References

[1] Alizadeh Noughabi, H. (2010). A new estimator of entropy and its application in testing normality. Journal of Statistical Computation and Simulation 80(10), 1151–1162.

[2] Alizadeh Noughabi, H., & Arghami, N. R. (2010). A new estimator of entropy. Journal of the Iranian Statistical Society 9(1), 53–64.

[3] Al-Omari, A. I. (2014). Estimation of entropy using random sampling. Journal of Computation and Applied Mathematics 261, 95–102, doi:10.1016/j.cam.2013.10.047.

[4] Al-Omari, A. I. (2016). A new measure of entropy of continuous random variable. Journal of Statistical Theory and Practice 10, 721–735, 10.1080/15598608.2016.1217444.

[5] Bitaraf, M., Rezaei, M., & Yousefzadeh, F. (2017). Test for normality based on two new estimators of entropy. Journal of Statistical Computation and Simulation 87(2), 280–294, DOI: 10.1080/00949655.2016.1208201.

[6] Ebrahimi, N., Pflughoeft, K., & Soofi, E. S. (1994). Two measures of sample entropy. Statistics and Probability Letters 20, 225-234.

[7] Grzegorzewski P, Wieczorkowski R. (1999). Entropy-based goodness-of-fit test for exponentiality. Commun Stat – Theory Methods 28, 1183–1202.

[8] Kohansal, A., & Rezakhah, S. (2016). Two new estimators of entropy for testing normality. Communications in Statistics - Theory and Methods 45(18), 5392–5411, DOI: 10.1080/03610926.2014.942436 .

[9] Lombardi, M. & Pant, S. (2016). Nonparametric k-nearest-neighbor entropy estimator. Physical Review E 93, 013310.

[10] Seber, W. A. F. (2004). Multivariate observations, 2nd edition. John Wiley and Sons Inc., New York.

[11] Shannon, C. E. (1948). A mathematical theory of communications. Bell System Technical Journal 27, 379–423, 623–656, doi:10.1002/bltj.1948.27.issue-3.

[12] Vasicek, O. (1976). A test for normality based on sample entropy. Journal of the Royal Statistical Society B 38, 54–59.

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

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