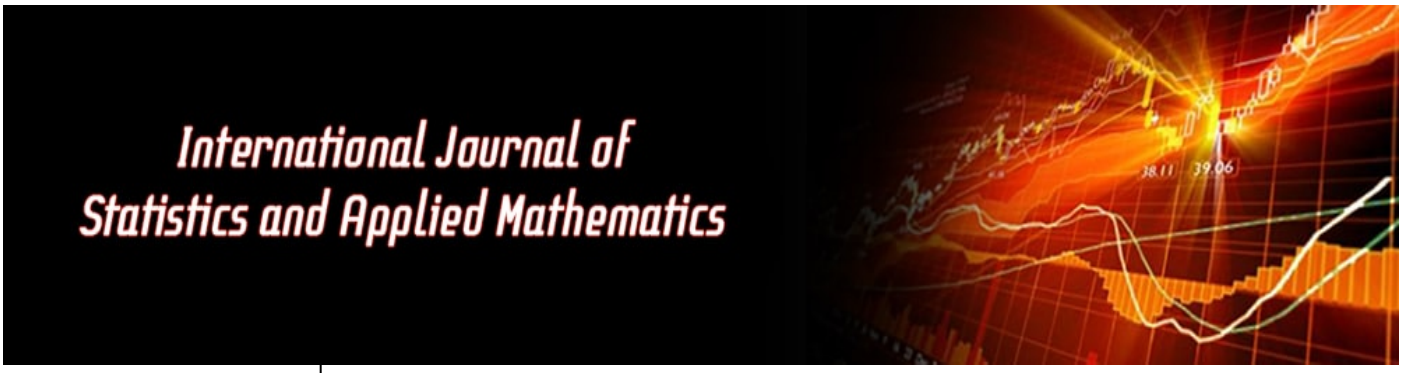
International Journal of Statistics and Applied Mathematics 2021; 6(6): 179-182



ISSN: 2456-1452
Maths 2021; 6(6): 179-182

www.mathsjournal.com
Received: 22-09-2021
Accepted: 24-10-2021

Mbanefo Solomon Madukaife
University of Nigeria,
Nsukka, Nigeria

Corresponding Author:
Mbanefo Solomon Madukaife
University of Nigeria,
Nsukka, Nigeria

Effect of window size (m) on entropy estimators

Mbanefo Solomon Madukaife

Abstract
Different estimators of the Shannon's measure of entropy of absolutely continuous distributions based on spacings have been proposed in the literature. It has also been noted that the performance of these estimators is dependent on the window size $(m : m \le n/2$, where n is the sample size$)$. In this paper, three different estimators are studied. A simulation study of their mean estimates and mean square errors (MSE) is performed at different window sizes for different sample sizes and probability distributions. In each case, the results obtained are discussed and conclusions made.

Keywords: Shannon's entropy, Monte Carlo simulation, mean square error, spacing, window size

1. Introduction
Shannon (1948) [9] introduced entropy as a measure of information and uncertainty. He defines the entropy of a random variable *X* as the amount of information contained in the variable. In this light, storage and transmission of information in such a random variable can intuitively be tied to the amount of information involved. Suppose that a random variable *X* has an absolutely continuous distribution function $F(x)$ and a probability density function $f(x)$, Shannon's entropy measure of the random variable *X*, denoted by $H(f)$, is defined by

$$H(f) = -\int_{-\infty}^{\infty} f(x,\theta)\log f(x,\theta)dx \qquad (1)$$

Since its introduction, entropy has played a vital role in statistical classification and goodness-of-fit tests. For instance, it has been used by Vasicek (1976) [11] and Arizono and Ohta (1989) [3] to propose tests for normality. Also, Grzegorzewski and Wieczorkowski (1999) [7] have used the concept of entropy to propose test for exponentiality. Dudewicz and van der Meulen (1981) [5] have used the concept to propose a test for uniformity of a random variable. Again, Zhu et al (1995) [13] have used the entropy measure of a multivariate normal distribution to propose a test for multivariate normality of a random vector.
All these uses and many more applications of entropy are based on the estimators of the population entropy measure. Vasicek (1976) [11] proposed an estimator of the population entropy measure given in (1). His estimator is given as

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

where $X_{(i+m)} = X_{(n)}$ if $(i+m) > n$; $X_{(i-m)} = X_{(1)}$ if $(i-m) < 1$. *n* is the sample size and *m* is an integar known as the window size which is chosen to be less than or equal to *n*/2, that is, $m : m \le n/2$. This estimator is based on spacings, provided by the window size, *m*. Its derivation was based on the fact that (1) can be presented as

$$H(f)=\int_0^1 \log\left(\frac{d}{dp}F^{-1}(p)\right)dp \tag{3}$$

Since after Vasicek's estimator, several other works have been carried out on this subject, either as an improvement to Vasicek's estimator or as an independent work. Some of them include van Es (1992) [10], Ebrahimi et al (1994) [6], Correa (1995) [4], Wieczorkowski and Grzegorzewski (1999) [12], Norghabi and Arghami (2010) [8], Al-Omari (2014) [1]. Particularly, Ebrahimi et al (1994) [6] modified Vasicek's estimator in order to assign smaller weights to observations in (2) that are replaced by $X_{(1)}$ and $X_{(n)}$. Their proposed entropy estimator is

$$HE_{mn}=n^{-1}\sum_{i=1}^{n}\log\left\{\frac{n}{\theta_i m}(X_{(i+m)}-X_{(i-m)})\right\} \tag{4}$$

where $\theta_i=\begin{cases}1+\frac{i-1}{m};1\le i\le m\\ 2;m+1\le n-m\\ 1+\frac{n-i}{m};n-m\le i\le n\end{cases}$; $X_{(i+m)}=X_{(n)}$ for $(i+m)>n$ and $X_{(i-m)}=X_{(1)}$ for $(i-m)<1$.

They showed by simulation that their estimator has smaller bias and mean squared errors than the Vasicek (1976) [11]. Also, they proved the consistency of the proposed modified estimator. Again, Al-Omari (2014) [1] suggested further modifications of the estimator of Shannon's entropy as given in (2). One of the modifications which is based on simple random sampling (SRS) is given as

$$HA_{mn}=n^{-1}\sum_{i=1}^{n}\log\left\{\frac{n}{w_i m}(X_{(i+m)}-X_{(i-m)})\right\} \tag{5}$$

where $w_i=\begin{cases}1+\frac{1}{2};1\le i\le m\\ 2;m+1\le n-m\\ 1+\frac{1}{2};n-m\le i\le n\end{cases}$; $X_{(i+m)}=X_{(n)}$ for $(i+m)>n$ and $X_{(i-m)}=X_{(1)}$ for $(i-m)<1$.

The author equally stated that the new estimator has smaller MSE than both the Vasicek (1976) [11] and Ebrahimi et al (1994) [6] estimators.

Now, since these estimators are based on spacings provided by the window size (m), how these estimators perform at different window sizes is the interest of this study. Wieczorkowski and Grzegorzewski (1999) [12] has suggested an optimal window size for each sample size as $m=\left[\sqrt{n}+0.5\right]$ which means the integer part of the right hand side. However, Al-Omari (2016) [2] maintains that the selection of the optimal values of the window size for a given value of *n* is as yet an open problem in entropy estimation, hence this work.

2. Simulation Study

To determine the effect of window size (*m*) on the performance of entropy estimators based on spacings, a simulation study was conducted. A total of 10,000 samples each were generated from standard normal, standard exponential and uniform (0,1) distributions at sample sizes 10, 20 and 30. From each sample, the entropy estimators according to Vasicek (1976), Ebrahimi et al (1994) and Al-Omari (2014) were computed using all the possible window sizes and MSEs of each estimator were obtained, adapted from Al-Omari (2016). The results are presented in Tables 1, 2, and 3 for each of the sample sizes considered respectively. In order to obtain the MSEs, the population (theoretical) entropy measures for each of the distributions is obtained. They are $\frac{1}{2}+\frac{1}{2}\log(2\pi\sigma^2)=1.4189$ for the standard normal distribution, $1-\log(\lambda)=1$ for the standard exponential distribution and $\log(b-a)=0$ for the uniform distribution in the interval (0, 1). With the theoretical entropy measures, the MSEs are obtained by:

$$\text{s}\,MSE=\frac{1}{10000}\sum_{i=1}^{10000}\left(H_{mn}-H(f)\right)^2 \tag{6}$$

where H_{mn} is the estimated and $H(f)$ is the corresponding theoretical entropy values.

Table 1: Empirical mean squared error of the HV_{mn}, *HEmn* and *HAmn* for the standard normal distribution, standard exponential distribution and uniform distribution, $U(0, 1)$, $n = 10$

m	Standard Normal Distribution			Standard Exponential Distribution			Uniform Distribution, U(0, 1)		
	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}
1	0.4577	0.3136	0.1849	0.4583	0.3181	0.3600	0.3249	0.2025	0.2401
2	0.3493	0.1849	0.1936	0.3272	0.1936	0.1962	0.2043	0.0784	0.0841
3	0.2737	0.1600	0.1444	0.3158	0.1600	0.1529	0.2061	0.0529	0.0454
4	0.4401	0.1521	0.1225	0.3306	0.1521	0.1376	0.2372	0.0484	0.0324
5	0.5171	0.1521	0.1089	0.3576	0.1444	0.1384	0.2788	0.0400	0.0282

Table 2: Empirical mean squared error of the HV_{mn}, *HEmn* and *HAmn* for the standard normal distribution, standard exponential distribution and uniform distribution, $U(0, 1)$, $n = 20$

m	Standard Normal Distribution			Standard Exponential Distribution			Uniform Distribution, U(0, 1)		
	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}
1	0.2337	0.1764	0.1849	0.2401	0.1849	0.1989	0.1747	0.1296	0.1421
2	0.1412	0.0841	0.0900	0.1414	0.0961	0.0961	0.0847	0.0400	0.0420
3	0.1332	0.0676	0.0625	0.1232	0.0784	0.0740	0.0729	0.0256	0.0213
4	0.1369	0.0625	0.0529	0.1246	0.0676	0.0635	0.0756	0.0169	0.0132
5	0.1600	0.0586	0.0484	0.1282	0.0640	0.0595	0.0835	0.0144	0.0090
6	0.1764	0.0576	0.0400	0.1296	0.0630	0.0615	0.0961	0.0130	0.0074
7	0.1936	0.0557	0.0361	0.1332	0.0630	0.0655	0.1102	0.0121	0.0067
8	0.2304	0.0562	0.0350	0.1429	0.0645	0.0756	0.1282	0.0117	0.0077
9	0.2500	0.0576	0.0346	0.1459	0.0676	0.0864	0.1475	0.0114	0.0096
10	0.2809	0.0576	0.0320	0.1482	0.0729	0.0891	0.1522	0.0110	0.0169

Table 3: Empirical mean squared error of the HV_{mn}, *HEmn* and *HAmn* for the standard normal distribution, standard exponential distribution and uniform distribution, $U(0, 1)$, $n = 30$

m	Standard Normal Distribution			Standard Exponential Distribution			Uniform Distribution, U(0, 1)		
	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}	HV_{mn}	HE_{mn}	HA_{mn}
1	0.1681	0.1369	0.1444	0.1789	0.1444	0.1560	0.1354	0.1063	0.1163
2	0.0900	0.0576	0.0625	0.0936	0.0676	0.0686	0.0557	0.0310	0.0317
3	0.0784	0.0441	0.0400	0.0795	0.0529	0.0497	0.0433	0.0169	0.0154
4	0.0784	0.0361	0.0320	0.0756	0.0441	0.0424	0.0424	0.0119	0.0088
5	0.0801	0.0324	0.0276	0.0745	0.0400	0.0392	0.0441	0.0094	0.0058
6	0.0841	0.0328	0.0250	0.0740	0.0396	0.0384	0.0488	0.0079	0.0040
7	0.0961	0.0313	0.0225	0.0756	0.0396	0.0392	0.0548	0.0067	0.0032
8	0.1024	0.0306	0.0228	0.0790	0.0388	0.0437	0.0620	0.0062	0.0032
9	0.1156	0.0299	0.0219	0.0795	0.0400	0.0488	0.0702	0.0059	0.0036
10	0.1225	0.0306	0.0219	0.0801	0.0441	0.0552	0.0801	0.0054	0.0045
11	0.1369	0.0309	0.0225	0.0818	0.0454	0.0640	0.0906	0.0054	0.0056
12	0.1521	0.0310	0.0231	0.0864	0.0484	0.0751	0.1017	0.0056	0.0072
13	0.1681	0.0306	0.0240	0.0888	0.0529	0.0520	0.1136	0.0054	0.0092
14	0.1849	0.0324	0.0240	0.0906	0.0576	0.0841	0.1267	0.0054	0.0112
15	0.1842	0.0318	0.0236	0.0902	0.0552	0.0840	0.1414	0.0053	0.0135

3. Discussion of Results

In order to discuss the results in comparison with Wieczorkowski and Grzegorzewski (1999) optimal window size for each sample size as $m = \left[\sqrt{n} + 0.5\right]$, optimal window sizes $m^* = 3$, 4 and 5 were obtained for sample sizes 10, 20 and 30 respectively. Under the sample size of $n = 10$, only the *HVmn* estimator conformed with the optimal window size of $m^* = 3$ only in the standard normal and standard exponential distributions. Under the sample size of $n = 20$, none of the estimators conformed with the optimal window size of $m^* = 4$ in all the different distributions considered. This was equally the case with the sample size of $n = 30$. Generally, while *HEmn* and *HAmn* tend to have better estimates as m approaches $n/2$, the *HVmn* maintains good estimates just before the middle of the range of values for m. The same pattern of result was observed under the uniform distribution except the *HAmn* whose good estimates approaches the middle of the range as the sample size increased. Under the standard exponential distribution, both *HEmn* and *HAmn* produced better estimates as m progressively approaches the middle of the range from $n/2$. However, *HVmn* maintained the same result as in the standard normal distribution.

4. Conclusion

Window size m has been seen to affect the entropy estimators. From this study, there is no basis to hold on to the optimal window size of $m = \left[\sqrt{n} + 0.5\right]$ as suggested by Wieczorkowski and Grzegorzewski (1999) without giving consideration to some other probable factor(s). This is more worrisome since the result fared worse with increasing sample size, as against the generally accepted theorem that the higher the sample size, the better an estimate. Finally, it is seen that the effect window size (m) has on the performance of the estimators depends on the distribution from where the sample is drawn.

5. References

1. 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.
2. Al-Omari AI. A new measure of entropy of continuous random variable. Journal of Statistical Theory and Practice. 2016;10:721-735. doi: 10.1080/15598608.2016.1217444.
3. Arizono I, Ohta H. A test for normality based on Kullback–Leibler information. American Statistician 1989; 43:20-23.
4. Correa JC. A new estimator of entropy. Communication in Statistics-Theory Methods. 1995;24(10):2439-2449. doi:10.1080/03610929508831626.
5. Dudewicz EJ, van der Meulen EC. Entropy-based tests of uniformity. Journal of the American Statistical Association. 1981;76:967-974.
6. 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.
7. Grzegorzewski P, Wieczorkowski R. Entropy-based goodness-of-fit test for exponentiality. Communications in Statistics-Theory and Methods. 1999;28(5):1183-1202.
8. Noughabi HA, Arghami NR. A new estimator of entropy. Journal of the Iranian Statistical Society. 2010;9(1):53-64.
9. Shannon CE. A mathematical theory of communications. Bell System Technical Journal 1948;27:379-423, 623–656. doi:10.1002/bltj.1948.27.issue-3.
10. Van Es B. Estimating functionals related to a density by class of statistics based on spacings. Scandinavian Journal of Statistics. 1992;19:61-72.
11. Vasicek O. A test for normality based on sample entropy. Journal of the Royal Statistical Society B. 1976;38:54-59.
12. Wieczorkowski R, Grzegorzewsky P. Entropy estimators improvements and comparisons. Communication in Statistics-Simulation and Computation. 1999;28(2):541-567. doi:10.1080/03610919908813564
13. Zhu L-X, Wong HL, Fang K-T. A test for multivariate normality based on sample entropy and projection pursuit. Journal of Statistical Planning and Inference. 1995;45:373-385.