Journal of Computational and Applied Mathematics 261 (2014) 95–102

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Estimation of entropy using random sampling

Amer Ibrahim Al-Omari *

Department of Mathematics, Faculty of Science, Al al-Bayt University, Mafraq 25113, Jordan

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form 8 September 2013

Keywords:
Entropy
Root mean square error
Simple random sampling
Ranked set sampling
Double ranked set sampling

ABSTRACT

In this paper, three new entropy estimators of continuous random variables are proposed using simple random sampling (SRS), ranked set sampling (RSS) and double ranked set sampling (DRSS) techniques. The new estimators are obtained by modifying the estimators suggested by Noughabi and Arghami (2010) and Ebrahim et al. (1994). In terms of the root mean square error (RMSEs) and bias values, a numerical comparison is considered to compare the suggested estimators with Vasicek's (1976) estimator. Our results reveal that the suggested estimators have smaller mean squared error than Vasicek's estimator. Also, the suggested estimators under double ranked set sampling are more efficient than other suggested estimators based on SRS and RSS.

1. Introduction

Assume that the random variable X has a continuous probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$. Shannon [1] defined the differentiable entropy $H(f)$ of the random variable X as

$$H(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx. \tag{1}$$

The entropy is a measure of uncertainty and dispersion. Many authors have been considered the problem of estimating the entropy of the continuous random variables. See for example [2–5].

Vasicek [4] showed that the estimator in (1) can be written as

$$H(f) = \int_0^1 \log\left(\frac{d}{dp}F^{-1}(p)\right) dp. \tag{2}$$

The estimator given in (2) is estimated by Vasicek by replacing the cdf $F(x)$ with the empirical cdf $F_n(x)$, and using a difference operator instead of the differential operator. Therefore, the derivative of $F^{-1}(p)$ is estimated by a function of the order statistics.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample of size n from $F(x)$, and let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of this sample. Vasicek [4] suggested an estimator of H as

$$\mathrm{HV}_{(m,n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Log}\left\{\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\}, \tag{3}$$

where m is a positive integer known as the window size, $m < \frac{n}{2}$ and $X_{(i)} = \begin{cases} X_{(1)}, & \text{if } i < 1 \\ X_{(n)}, & \text{if } i > n. \end{cases}$

* Tel.: +962 777906433.
E-mail addresses: alomari_amer@yahoo.com, amerialomari@gmail.com.

0377-0427/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cam.2013.10.047

Vasicek showed that $\mathrm{HV}_{(m,n)}$ converges in probability to $H(f)$ as $n \to \infty$, $m \to \infty$, and $m/n \to 0$.

Van Es [6] suggested an entropy estimator as

$$\mathrm{HVE}_{(m,n)} = \frac{1}{n-m}\sum_{i=1}^{n-m}\left\{\frac{n+1}{m}\left(X_{(i+m)} - X_{(i)}\right)\right\} + \sum_{k=m}^{n}\frac{1}{k} + \log\left(\frac{m}{n+1}\right), \tag{4}$$

and under some conditions, proved the consistency and asymptotic normality of the estimator.

Ebrahimi et al. [2] adjusted the weight $\frac{n}{2m}$ in [4] estimator to assign smaller weights and proposed the following estimator

$$\mathrm{HE}_{(m,n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Log}\left\{\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\}, \tag{5}$$

where

$$c_i = \begin{cases} 1 + \dfrac{i-1}{m}, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1 + \dfrac{n-i}{m}, & n-m+1 \le i \le n. \end{cases}$$

Based on simulations, Ebrahimi et al. [2] showed that their estimator has a smaller bias and mean square errors. Also, they proved that $\mathrm{HE}_{(m,n)}$ converges in probability to $H(f)$ as $n \to \infty$, $m \to \infty$ and $m/n \to 0$.

Noughabi and Arghami [7] suggested a modified version of Ebrahimi et al. [2] entropy estimator and proved that it performs better than the Vasicek [4] and Ebrahimi et al. [2] estimators. Their proposed estimator is given by

$$\mathrm{HNA}_{(m,n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Log}\left\{\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\}, \tag{6}$$

where

$$c_i = \begin{cases} 1, & 1 \le i \le m, \\ 2, & m+1 \le i \le n-m, \\ 1, & n-m+1 \le i \le n. \end{cases}$$

They proved the consistency of the estimator, $\mathrm{HNA}_{(m,n)} \xrightarrow{P} H(f)$ as $n \to \infty$, $m \to \infty$, $m/n \to 0$.

Noughabi and Noughabi [5] suggested a new estimator of entropy of an unknown continuous probability density function as

$$\mathrm{HNN}_{(m,n)} = -\frac{1}{n}\sum_{i=1}^{n}\mathrm{Log}\left\{s_i(n,m)\right\}, \tag{7}$$

where

$$s_i(n,m) = \begin{cases} \hat{f}\left(X_{(i)}\right), & 1 \le i \le m, \\ \dfrac{2m/n}{X_{(i+m)} - X_{(i-m)}}, & m+1 \le i \le n-m, \\ \hat{f}\left(X_{(i)}\right), & n-m+1 \le i \le n, \end{cases} \quad \text{and} \quad \hat{f}(X_i) = \frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{X_i - X_j}{h}\right),$$

where h is bandwidth and k is a kernel function satisfying $\int_{-\infty}^{\infty} k(x)dx = 1$. They proved that $\mathrm{HNN}_{(m,n)} \xrightarrow{P} H(f)$ as $n \to \infty$, $m \to \infty$, $m/n \to 0$.

Note that the kernel function in [5] is chosen to be the standard normal distribution and the bandwidth h is chosen to be the normal smoothing formula, $h = 1.06\, sn^{-1/5}$, where s is the sample standard deviation.

Correa [3] proposed a modification of Vasicek's estimator with a smaller mean square error by

$$\mathrm{HC}_{mn} = -\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{\sum\limits_{j=i-m}^{i+m}(j-i)\left(X_{(j)} - \bar{X}_{(i)}\right)}{n\sum\limits_{j=i-m}^{i+m}\left(X_{(j)} - \bar{X}_{(i)}\right)^2}\right], \tag{8}$$

where $\bar{X}_{(i)} = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} X_{(j)}$.

For more about entropy estimators, see [8–13].

The rest of this paper is organized as follows. In Section 2, the suggested estimators using SRS, RSS and DRSS are introduced. Numerical comparisons between the suggested estimators and that of Vasicek [4] are given in Section 3. Finally, Section 4 summarizes our conclusions.