JIRSS (2022)
Vol. 21, No. 01, pp 1-18
DOI: 10.22034/JIRSS.2022.703882

Nonparametric Estimation of the Residual Entropy Function with Length-Biased Data

Farzaneh Oliazadeh [1], **Anis Iranmanesh** [1], **Vahid Fakoor** [2]

[1] Department of Mathematics and Statistics, Mashhad Branch, Islamic Azad University, Mashhad, Iran.
[2] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashahd, Iran.

Received: 23/12/2020, Accepted: 05/10/2022, Published online: 03/05/2023

Abstract. We propose a nonparametric estimator for the residual entropy function based on length-biased data. Some asymptotic results have been proved. The strong consistency and asymptotic normality of the proposed estimator are established under suitable regularity conditions. Monte Carlo simulation studies are carried out to evaluate the performance of the estimator using the bias and mean-squared error. A real data set is considered, and we show that the data follow a length-biased distribution. Moreover, the proposed estimator yields a better value for the estimated residual entropy in comparison to the competitor estimator.

Keywords. Asymptotic Normality, Length-Biased Data, Kernel Density Estimation, Residual Entropy, Strong Consistency.

MSC: 94A17, 62B10.

Farzaneh Oliazadeh (farz.olia@gmail.com)
Corresponding Author: Anis Iranmanesh (iranmanesh0030@mshdiau.ac.ir)
Vahid Fakoor (fakoor@um.ac.ir)

1 Introduction

Let X be a positive absolutely continuous random variable with density function $f(\cdot)$. Shannon's information measure or the differential entropy of X is given by

$$H(f) = -\int_0^\infty f(x)\log f(x)dx. \tag{1.1}$$

Therefore, $H(f)$ measures the expected uncertainty contained in $f(\cdot)$ about the predictability of an outcome of X. In survival analysis and life testing, since the current age of the system under consideration is also taken into account, the Shannon entropy is not suitable for calculating uncertainty of a system that has survived for some unit of time. Therefore, the notion of residual entropy has been introduced in the literature. The residual lifetime of the system when it is still operating at time t is $X_t = (X - t|X > t)$. It can be readily shown that the comulative distribution function and probability density function of X_t is given by

$$\begin{aligned} F_t(x) &= 1 - \frac{S(x+t)}{S(t)}, \\ f_t(x) &= \frac{f(x+t)}{S(t)}, \end{aligned}$$

respectively, where $f(\cdot)$ denotes the density function of X and $S(t) = P(X > t)$ is the reliability (or survival) function. Ebrahimi (1996) proposed the entropy of the residual lifetime X_t as

$$\begin{aligned} H(t) &:= -\int_t^\infty \log(f_t(x))dF_t(x) \qquad (1.2) \\ &= -\int_t^\infty \frac{f(x)}{S(t)}\log\left(\frac{f(x)}{S(t)}\right)dx \\ &= \log(S(t)) - \frac{1}{S(t)}\int_t^\infty f(x)\log f(x)dx. \qquad (1.3) \end{aligned}$$

The residual entropy function, introduced in (1.3), is viewed as a dynamic measure of uncertainty, since this measure finds applications in modeling and analysis of lifetime data. In the literature, several estimators of residual entropy have been proposed for a random sample. More recently, Ebrahimi (1997) considered the problem of testing the monotonicity of this measure. Belzunce et al. (2001) proposed a kernel type estimation of the residual entropy function in the case of independent complete data sets. Also

Belzunce et al. (2004) established that if $H(t)$ is increasing in t, then $H(t)$ determines the distribution uniquely. Given that an item has survived up to time t, $H(t)$ measures the uncertainty about its remaining life. Maya et al. (2014) proposed nonparametric estimators for the Rényi information measure for the residual lifetime distribution based on complete and censored data. Rajesh et al. (2015) discussed a nonparametric estimation of the residual entropy function with censored dependent data. They also investigated asymptotic properties of the estimator under suitable regularity conditions.

Kumar and Taneja (2015) handled the study on a length-biased dynamic measure of past inaccuracy. Taneja et al. (2009) studied a dynamic measure of inaccuracy between two residual life distributions. Tahmasebi and Daneshi (2018) considered a measure of inaccuracy between distributions of the nth record value and parent random variable. They also proposed the measure of residual inaccuracy of record values and investigated characterization results of dynamic cumulative residual inaccuracy measure. Kayal et al. (2017) introduced a generalized measure of inaccuracy between two residual and past lifetime distributions of a system. Rajesh et al. (2017) proposed nonparametric estimators for the inaccuracy measure for the lifetime distribution based on censored data. Kayal (2015) dealt with a generalized residual entropy of record values and weighted distributions. Also, some results on monotone behavior of generalized residual entropy in record values were obtained. For more discussion on the properties and characterization results using the notion of residual entropy, we refe to Nair and Rajesh (1998), Sankaran and Gupta (1999), and Asadi and Ebrahimi (2000).

For representing population, it is assumed that a sample has the same basic properties. However, the sample may not be entirely representative of the population in practice and it is known as weighted data when bias is introduced in the sampling scheme. The observed sample will not be representative of the population of interest, once an appropriate randomization cannot be attained. Since in the real world this biased sampling issue appears frequently, truly random sampling is not easily achievable or practically feasible. Weighted data appear in many sampling processes. This type of sample is produced when the probability of choosing an observation depends on its value and/or other covariates of interest; see Patil and Rao (1978). Such data are observed in a variety of fields such as biomedicine (Chakraborty and Rao, 2000), epidemiology (Simon, 1980), textile fibers (Cox, 1969), social sciences, economics, and quality control. It is worth knowing that biased sampling problems may occur even if the sampling is unbiased. For a comprehensive overview of biased sampling, we refer to Qin (2017).

Let X be a non-negative random variable with density function and distribution

function of $f(\cdot)$ and $F(\cdot)$, respectively. The random variable Y has the length-biased distribution of $F(\cdot)$ if the density function of Y reads as follows:

$$G(y) = \frac{1}{\mu}\int_0^y xf(x)dx, \quad y \geq 0, \tag{1.4}$$

where $\mu = \int_0^\tau xf(x)dx < \infty$ and $\tau = \sup\{x : F(x) < 1\}$. We assume that τ is finite. Let $Y_1, \ldots, Y_n$ be a random sample from G. Then the empirical distribution function (edf) of Y is given by

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq t),$$

where $I(A)$ denotes the indicator of the event A. The empirical estimator of F can be written in the form

$$F_n(t) = \mu_n \int_0^t y^{-1} dG_n(y), \tag{1.5}$$

where

$$\mu_n^{-1} = \int_0^\infty y^{-1} dG_n(y).$$

For various nonparametric estimations of f and F based on a length-biased sampling, we refer to Guillamón et al. (1998), Vardi (1982), De Uña-álvarez (2004), Chaubey et al. (2010), Jones (1991), Ajami et al. (2013), and references therein.

The importance of the biased sampling problem and residual entropy appears in many fields of research; we are interested in the estimation of the residual entropy of X in the length-biased setting. Although ample information about estimation of the residual entropy with complete and censored data is available, there is few research in the literature studying the estimation of the residual entropy with length-biased data.

Since Shannon's entropy plays an important role in the context of the information theory, Oliazadeh et al. (2021) proposed an estimator of Shannon entropy based on kernel estimators of density function in a length-biased setting. Also, they proved the strong consistency of the proposed estimator. Since this entropy is not applicable to a system that is known to have survived beyond some time, say t, the concept of residual entropy was later developed. Accordingly, in this paper, we study information measures for residual lifetime distributions based on this measure. Considering the fact that both residual entropy and length-biased data have wide applications in different fields, for example in reliability for measuring uncertainty about the remaining lifetime of the unit, in the present article, we extend the work of Oliazadeh et al. (2021) to

an estimation of the residual entropy of X based on length-biased data and study the strong consistency and asymptotic normality of the proposed estimator.

Given a random sample $Y_1, \ldots, Y_n$, drawn from the distribution G, the plug-in estimator for $H(t)$ can be defined as

$$\begin{aligned} H_n(t) &:= -\int_t^\tau \frac{f_n(x)}{S_n(t)} \log\Big(\frac{f_n(x)}{S_n(t)}\Big) dx \\ &= \log(S_n(t)) - \frac{1}{S_n(t)} \int_t^\infty f_n(x) \log(f_n(x)) dx, \end{aligned} \tag{1.6}$$

where

$$f_n(t) = \frac{\mu_n}{nh_n} \sum_{i=1}^{n} \frac{1}{Y_i} K\Big(\frac{t - Y_i}{h_n}\Big). \tag{1.7}$$

The estimator (1.7) was proposed by Jones (1991). Its asymptotic behaviour considered by Jones (1991) and Ajami et al. (2013). Our objective in this paper is to establish some asymptotic behaviors of $H_n(t)$.

We organize our article as follows. In Section 2, we propose some assumptions and prove the main theorems on the strong consistency and asymptotic normality of the proposed estimator. In Section 3, a simulation study to illustrate the behaviour of the new estimator is carried out, and the results are compare with the competitor estimator. Also, the new estimator and its competitor are applied to a real data set to evaluate the new estimator. The conclusions of this study and suggestions for future work, especially in the application, are given briefly in Section 4.

2 Main Result

In this section, we discuss strong consistency and asymptotic normality of the proposed estimator for the residual entropy under the length-biased scheme. First, we list some assumptions, which will be used in our results below.

(A1) Let K be a kernel function of bounded variation, vanishing outside of the interval $(-1, 1)$, with

$$\int_{-1}^{1} K(x)dx = 1, \quad \int_{-1}^{1} xK(x)dx = 0, \quad \int_{-1}^{1} x^2 K(x)dx < \infty.$$

(A2) Let f be twice differentiable with a continuous and bounded second derivative f''.

(A3) $\int_0^\tau |\log f(x)|dx < \infty$.

(A4) $v(r) = \int_0^\infty u^{-2}G^{1/r}(u)du < \infty$ for some $r > 2$.

(H1) $\log n/(n^{1/2+\rho}h_n) = o(1)$, for any $0 < \rho < 1/2 - 1/r$ for some $r > 2$.

(H2) $h_n^{-1/2}n^{-\rho} = o(1)$.

(H3) $h_n^{5/2}n^{1/2} = o(1)$.

Remark 1. Assumption **(A1)** is a commonly used smoothness condition. We need Assumption **(A2)**, because we use a second-order kernel. Assumptions **(A3)** and **(A4)** are required to prove consistency and asymptotic normality of the proposed estimator. Assumption **(H1)** gives a wide range, from $O((\log n)^{-\alpha})$, $\alpha > 0$, to $O(n^{-\beta})$, $0 < \beta < \frac{1}{2} + \rho$, of bandwidths that include the optimal bandwidth in kernel density function, that is, $O(n^{-1/5})$. The conditions on the bandwidth in Assumptions **(H2)** and **(H3)** are also not restrictive. Considering $h_n \sim n^{-\beta}$, if we choose $0 < \beta < \frac{1}{5} \wedge 2\rho$, then Assumptions **(H1)** − **(H3)** are satisfied.

In the below theorems, we let T be such that $S(T) > \delta$ with some $\delta > 0$. The first theorem gives the strong uniform consistency of $H_n(\cdot)$.

Theorem 2.1. *Let Assumptions **(A1)–(A4)** and **(H1)** be fulfilled. Then*

$$\sup_{0<t\leq T} |H_n(t) - H(t)| \to 0 \quad a.s.$$

Proof. First, we have

$$H_n(t) - H(t) = (m_n(t) - m(t)) - \left(\frac{\mathbb{H}_n(t)}{S_n(t)} - \frac{\mathbb{H}(t)}{S(t)}\right), \tag{2.1}$$

where

$$m(t) = \log(S(t)), \quad m_n(t) = \log(S_n(t)),$$

$$\mathbb{H}(t) = \int_t^\tau f(x)\log f(x)dx, \quad \mathbb{H}_n(t) = \int_t^\tau f_n(x)\log f_n(x)dx.$$

Using the Taylor expansion for the logarithm function $S(t)$, we have

$$m_n(t) - m(t) = \frac{S_n(t) - S(t)}{S_n^*(t)},$$

where $\min\{S_n(t), S(t)\} < S_n^*(t) < \max\{S_n(t), S(t)\}$ for $0 < t \le T$. Recall that the mean values $S_n^*(\cdot)$ on the line segment between $S(\cdot)$ and $S_n(\cdot)$ are bounded from below by δ^{-1} on $(0, T]$. Thus, Theorem 2.1. of Horváth (1985) implies that

$$\begin{aligned}\sup_{0<t\le T} |m_n(t) - m(t)| &\le \delta^{-1} \sup_{0<t\le T} |S_n(t) - S(t)| \\ &= o(1), \quad a.s. \quad n \to \infty. \end{aligned} \tag{2.2}$$

Note that for every $t \ge 0$,

$$\begin{aligned}\frac{\mathbb{H}_n(t)}{S_n(t)} - \frac{\mathbb{H}(t)}{S(t)} =& \frac{\mathbb{H}_n(t) - \mathbb{H}(t)}{S(t)} + \frac{\mathbb{H}_n(t)(F_n(t) - F(t))}{S^2(t)} \\ &+ \frac{\mathbb{H}_n(t)(F_n(t) - F(t))^2}{S_n(t)S^2(t)}. \end{aligned} \tag{2.3}$$

Clearly

$$\begin{aligned}|\mathbb{H}_n(t) - \mathbb{H}(t)| \le& \left| \int_t^\tau f_n(x)[\log f_n(x) - \log f(x)]dx \right| \\ &+ \left| \int_t^\tau \log f(x)[f_n(x) - f(x)]dx \right| \\ \le& \int_t^\tau f(x) \left| \frac{f_n(x)}{f(x)} \log \frac{f_n(x)}{f(x)} \right| dx \\ &+ \|f_n - f\|_\infty \int_t^\tau |\log f(x)|dx \\ =:& I_{1n} + I_{2n}. \end{aligned} \tag{2.4}$$

where $\|\cdot\|_\infty$ denotes the usual sup norm, that is, $\|H\|_\infty = \sup\{|H(x)| : x > 0\}$. To deal with I_{1n}, since $|z \log z| \le |z - 1|(1 + z)$ for all $z \ge 0$, we have

$$\left| \frac{f_n(x)}{f(x)} \log \frac{f_n(x)}{f(x)} \right| \le \left| \frac{f_n(x)}{f(x)} - 1 \right| \left(1 + \frac{f_n(x)}{f(x)} \right).$$

Hence, with finite τ and almost sure convergence of f_n, for every $0 < t \le T$, we have

$$\begin{aligned} I_{1n} &\le \int_t^\tau \left| f_n(x) - f(x) \right| \left(1 + \frac{f_n(x)}{f(x)} \right) dx \\ &\le \|f_n - f\|_\infty \int_0^\tau \left(1 + \frac{f_n(x)}{f(x)} \right) dx \\ &\le \|f_n - f\|_\infty \left(2\tau + \frac{\|f_n - f\|_\infty}{\inf_{0<t\le T} f(t)} \right) \to 0, \quad a.s. \end{aligned} \tag{2.5}$$

To deal with I_{2n}, using Assumption **(A3)** and Theorem 2 in Ajami et al. (2013), we have

$$I_{2n} \leq C\|f_n - f\|_\infty \to 0 \quad a.s. \tag{2.6}$$

Then, using (2.4)–(2.6), we have

$$\sup_{0<t\leq T} |\mathbb{H}_n(t) - \mathbb{H}(t)| = o(1) \quad a.s. \quad n \to \infty. \tag{2.7}$$

By the strong uniform consistency of F_n (Horváth, 1985) for the second and third part of the right side of equation (2.3) we have

$$\frac{\mathbb{H}_n(t)}{S_n(t)} - \frac{\mathbb{H}(t)}{S(t)} \to 0 \quad a.s. \tag{2.8}$$

uniformly on $(0, T]$. In view of (2.1), (2.2) and (2.8) together imply that

$$\sup_{0<t\leq T} |H_n(t) - H(t)| \to 0 \quad a.s.$$

□

The next theorem proves that $H_n(\cdot)$ is an asymptotically normally distributed estimator.

Theorem 2.2. *Suppose that Assumptions **(A1)–(A4)** and bandwidth conditions **(H1)–(H3)** hold. Then, $\sqrt{nh_n}(H_n(t) - H(t))$ converges weakly to a normal distribution with mean zero and variance $\sigma^2(t) := R(t,t)$, that $R(u,v)$ has been given in* (2.10).

Proof. Utilizing (2.3) and (2.1), we have

$$\begin{aligned}\sqrt{nh_n}(H_n(t) - H(t)) &= \sqrt{nh_n}(m_n(t) - m(t)) - \sqrt{nh_n}\frac{\mathbb{H}_n(t) - \mathbb{H}(t)}{S(t)} \\ &\quad - \sqrt{nh_n}\frac{\mathbb{H}_n(t)(F_n(t) - F(t))}{S^2(t)} - \sqrt{nh_n}\frac{\mathbb{H}_n(t)(F_n(t) - F(t))^2}{S_n(t)S^2(t)} \\ &=: J_1 + J_2 + J_3 + J_4.\end{aligned}$$

The Taylor expansion for logarithm functions gives us

$$J_1 = -\sqrt{nh_n}\frac{(F_n(t) - F(t))}{S_n^*(t)},$$

where $\min\{S_n(t), S(t)\} < S_n^*(t) < \max\{S_n(t), S(t)\}$ for $0 < t \leq T$. Using Theorem 4.1. in Horváth (1985), we have $\sqrt{n}(F_n(t) - F(t)) = O_p(1)$. Now, since $S_n^*(t) \to S(t)$ as $n \to \infty$, we get $J_1 = o_p(1)$. Similarly, utilizing Theorem 2.1, $J_3 = o_p(1)$, and $J_4 = o_p(1)$. To deal with J_2, we first obtain the limiting distribution of $\sqrt{nh_n}(\mathbb{H}_n(t) - \mathbb{H}(t))$. Using the Taylor expansion for the logarithm function and Theorem 2.2 in Ajami et al. (2013), we have

$$\sqrt{nh_n}(\mathbb{H}_n(t) - \mathbb{H}(t)) = \sqrt{nh_n} \int_t^\tau (f_n(x) - f(x))(1 + \log f(x))dx + o_p(1). \tag{2.9}$$

It follows, from Lemma A.3 in Akbari et al. (2019) that there exists a sequence of a mean zero Gaussian processes $\{\varrho(x, n),\ x > 0\}$ such that, $0 < \rho < 1/2 - 1/r$ for some $r > 2$.

$$\sup_{x>0} \left| \sqrt{nh_n}(f_n(x) - f(x)) - \varrho(x, n) \right| = O(h_n^{-1/2} n^{-\rho} \vee n^{1/2} h_n^{5/2}), \quad a.s.,$$

where

$$\varrho(x, n) = -h_n^{-1} \int_0^\infty \Gamma(t, n) d_t K\left(\frac{t - x}{h_n}\right) dx,$$

in which, $\Gamma(x, n)$ is a two-parameter Gaussian process with mean zero and covariance function $E(\Gamma(x, n)\Gamma(y, m))$ with complicated structure which mentioned in Lemma A.3 in Akbari et al. (2019). This result and (2.9) yields

$$\sup_{t>0} \left| \sqrt{nh_n}(\mathbb{H}_n(t) - \mathbb{H}(t)) - \int_t^\tau \varrho(x, n)(1 + \log f(x))dx \right| = o_p(1).$$

Thus, it is immediate that the process $\sqrt{nh_n}\frac{(\mathbb{H}_n(t) - \mathbb{H}(t))}{S(t)}$ is approximated in probability by a Gaussian process with mean zero and covariance function

$$R(u, v) = E\left(\frac{1}{S(u)}\frac{1}{S(v)} \int_u^\tau \varrho(x, n)(1 + \log f(x))dx \int_v^\tau \varrho(y, n)(1 + \log f(y))dy\right). \tag{2.10}$$

From this result $\sqrt{nh_n}(H_n(t) - H(t))$ converges weekly to a normal random variable with mean zero and variance $R(u, u)$. □

Remark 2. Due to the complex form of the asymptotic variance of $H_n(\cdot)$, it is difficult to directly obtain a consistent variance estimator for $H_n(\cdot)$. Alternatively, the bootstrap procedure can be used to estimate $\sigma^2(t)$.

Remark 3. It should be mention that $H(t) \to H(f)$ as $t \to 0$, where $H(f)$ is a Shannon's information measure of X mentioned in (1.1). This measure of uncertainty has been investigated by Oliazadeh et al. (2021) in the length-biased setting. To estimate $H(f)$, they considered $H(f_n)$, defined as,

$$H(f_n) := -\int_{A_n} f_n(x) \log f_n(x) dx, \tag{2.11}$$

where $f_n(\cdot)$ is given in (1.7), $A_n = \{x : f_n(x) \geq \gamma_n\}$, and $\gamma_n \downarrow 0$ is a sequence of positive constant. To circumvent problem related to a singularity at zero, A_n has been considered to exclude the small values of $f_n(\cdot)$. The choice of γ_n close to zero guarantees the closeness of $H(f_n)$ to $H(f)$. Also, they proved strong consistency of the sequence of estimators $H(f_n)$, i.e.,

$$|H(f_n) - H(f)| \to 0, \quad a.s. \quad as \quad n \to \infty. \tag{2.12}$$

Although $H(t)$ is an extension of $H(f)$, its estimator, $H_n(t)$, as $t \to 0$ is not the same as $H(f_n)$. Since Oliazadeh et al. (2021) considered the estimator $H(f)$ over the set A_n, the problem of singularity at the origin is avoided. The entropy of the residual lifetime X_t, i.e., $H(t)$, is defined for $t > 0$, therefore plug-in estimator for $H(t)$ resolve the mentioned problem. Strong uniform consistency of random function $H_n(t)$ was proved in Theorem 2.1. In Theorem 2.2, asymptotic normality of the estimator $H_n(t)$ was investigated. This asymptotic property is missing in the work of Oliazadeh et al. (2021) for $H(f_n)$.

3 Illustration

3.1 Simulations

To evaluate finite sample performances of the proposed estimator, we conduct simulation studies in three scenarios.

Scenario 1: Let X be a random variable having a Weibull distribution with probability density function

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left\{-\left(\frac{x}{b}\right)^{a}\right\}, \quad x > 0, \quad a, b > 0.$$

Scenario 2: Let X be a random variable having a log-normal distribution with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0, \ \ \mu \in R, \ \ \sigma > 0,$$

where μ and σ are the mean and standard deviation of the logarithm of X, respectively.

Scenario 3: Let X be a random variable having a beta distribution with probability density function

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 \le x \le 1, \;\; a, b > 0,$$

where $\Gamma(\cdot)$ is the gamma function.

We use the notation $W(a,b)$, $LN(\mu,\sigma)$, and $Beta(a,b)$ to show that X has a Weibull, log-normal and beta distribution, respectively. We should mention that these three distributions are common in the reliability theory and survival analysis. We choose these distributions because Weibull and log-Normal distributions are two distributions with positive values that are important, suitable, common and useful for lifetime data. Beta distribution is also used as a distribution with limited support.

Length-biased data are generated from Equation (1.4) through the inverse transformation sampling method. We note that, if X has $W(a,b)$, then the corresponding length-biased form is $GG(1+1/a, 1/b, a)$, where $GG(a,b,c)$ follows the generalized gamma distribution with probability density function

$$f(x) = \frac{1}{\Gamma(a)} bc(bx)^{ac-1} \exp\{-(bx)^c\}, \quad x > 0, \quad a, b, c > 0.$$

Given a log-normal distributed random variable X with parameters μ and σ, then the transformation $Y = \exp\{2\mu + \sigma^2\}/X$ connects the length-biased version to the original variable X. Also, in the $Beta(a,b)$ case, the length-biased form is $Beta(a+1,b)$. In the evaluation process of the proposed estimator, we can consider $H_n^*(t)$ as another estimator for the baseline function $H(t)$ as follows

$$H_n^*(t) = -\frac{1}{\sum_{i=1}^n Y_i^{-1}(t)} \sum_{i=1}^n Y_i^{-1}(t) \log \hat{f}_{n,i}(Y_i), \tag{3.1}$$

in which $Y_i(t) = Y_{(i)} I(Y_{(i)} > t)$ and

$$\hat{f}_{n,i}(x) = \frac{1}{h_n \sum_{j \neq i} Y_j^{-1}(t)} \sum_{j \neq i} K\Big(\frac{Y_j(t) - x}{h_n}\Big).$$

$H_n^*(t)$ provided with ideas from Belzunce et al. (2001) and Equation (1.2). We replaced Cox empirical estimator $F_n(t)$ from (1.5) in (1.2) only by considering Y_i greater than t, then we got the following estimator

$$\begin{aligned}\widehat{I(t)} &= \int_t^{\infty} f_t(x)dF_n(x) \\ &= \frac{\mu_n}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\log f_t(Y_i)I(Y_i > t).\end{aligned}$$

Because f is unknown, $\widehat{I(t)}$ cannot be used in practice. Hence we substitute the Jones's estimator $f_n(\cdot)$ for $f_t(\cdot)$ in $\hat{I}(t)$. This estimator is also based on Y_i greater than t. The result is the estimator $H_n^*(t)$ in (3.1).

For these simulations, we use various sample sizes $n = 50, 100, 200$. The Epanechnikov density function $K(x) = 3/4(1 - x^2)I(|x| < 1)$ is used as the kernel function. The Monte Carlo mean square error (MSE) and bias are calculated for each estimator at three points based on 1000 replications with bandwidths selected by the minimum MSE method. We considered $MSE(H_n)$ and $MSE(H_n^*)$ as a function of h_n and choose the value of h_n that minimize the estimated $MSE(H_n)$ and $MSE(H_n^*)$. We considered this value as a optimal value for h_n. Figure 1. shows estimated MSE as a function of h_n for different distributions. Table 1 summarizes MSE and bias (in the parentheses) of two estimators for the Weibull distribution with the shape parameters $a = 0.5, 1$ and 2 and the scale parameters $b = 1$, at $t = 0.5, 1.5$ and 2.5. The estimator H_n^* has been constructed based on observation greater that t value. Thus, in Table 1, some values of H_n^* at point $t = 2.5$ are not specified because length-biased data were not larger than 2.5, so the estimator (3.1) cannot be calculated. Infact for $W(2, 1)$ the length-biased version is $GG(3/2, 1, 2)$ with pdf $f(x) = 4/\sqrt{\pi}x^2\exp(-x^2)$ and $P(X > 2.5) = 0.00585$ which is very small so it is rare to have data after this point. Table 2 summarizes MSE and bias (in the parentheses) of two estimators for the log-normal distribution with $\mu = 0$ and $\sigma = 0.5, 1$ and 2 at $t = 0.5, 1.5$ and 2.5. In this table, for $LN(0, 0.5)$ at point 2.5, H_n^* cannot be calculated for the same reason described earlier for $W(2, 1)$. Table 3 summarizes MSE and bias (in the parentheses) of two estimators for the Beta(1,1), Beta(0.5,0.5) and Beta(2,2) at $t = 0.5, 1.5$ and 2.5.

As shown in Tables 1–3, the plug-in estimator H_n outperforms the estimator H_n^* in most cases. In view of Tables 1 and 2, we notice that it may be impossible to compute the estimator H_n^* for the moderate t.

Table 1: MSE and bias (in the parentheses) of Weibull distribution with different parameters

t	Estimator	n	W(0.5,1)	W(1,1)	W(2,1)
0.5	H_n	50	0.173(-0.071)	0.015(0.011)	0.006(0.003)
		100	0.094(-0.019)	0.008(0.011)	0.003(0.002)
		200	0.056(-0.007)	0.004(0.009)	0.002(0.002)
	H_n^*	50	0.397(-0.612)	0.092(-0.131)	0.116(0.308)
		100	0.379(-0.608)	0.098(0.224)	0.099(0.298)
		200	0.371(-0.605)	0.099(0.279)	0.089(0.290)
1.5	H_n	50	0.096(-0.016)	0.032(-0.026)	0.064(-0.008)
		100	0.050(0.014)	0.015(-0.018)	0.028(-0.018)
		200	0.028(0.014)	0.008(-0.006)	0.014(0.012)
	H_n^*	50	1.683(-1.293)	0.116(-0.180)	0.069(0.203)
		100	1.664(-1.288)	0.157(0.318)	0.048(0.183)
		200	1.664(-1.289)	0.343(0.557)	0.042(0.187)
2.5	H_n	50	0.097(-0.034)	0.096(-0.091)	0.582(0.402)
		100	0.048(-0.001)	0.041(-0.046)	0.807(0.180)
		200	0.027(-0.007)	0.02(-0.025)	0.124(-0.101)
	H_n^*	50	2.767(-1.659)	0.254(0.235)	-
		100	2.734(-1.652)	0.167(0.272)	-
		200	2.730(-1.651)	0.740(0.833)	-

Table 2: MSE and bias (in the parentheses) of Lognormal distribution with different parameters

t	Estimator	n	LN(0,0.5)	LN(0,1)	LN(0,2)
0.5	H_n	50	0.009(0.009)	0.021(0.017)	0.650(-0.652)
		100	0.005(0.006)	0.010(0.013)	0.615(-0.745)
		200	0.003(0.023)	0.006(0.008)	0.429(-0.598)
	H_n^*	50	0.013(0.033)	0.031(0.012)	0.314(-0.169)
		100	0.006(0.019)	0.016(0.006)	0.172(-0.013)
		200	0.003(0.014)	0.008(0.001)	0.118(0.027)
1.5	H_n	50	0.0489(-0.009)	0.045(-0.040)	0.747(-0.770)
		100	0.023(0.0001)	0.022(-0.020)	0.750(-0.835)
		200	0.01(-0.001)	0.010(-0.006)	0.664(-0.754)
	H_n^*	50	0.652(0.784)	0.056(0.026)	0.361(-0.231)
		100	0.628(0.781)	0.030(0.042)	0.213(-0.037)
		200	0.614(0.778)	0.016(0.040)	0.145(0.025)
2.5	H_n	50	0.506(-0.251)	0.076(-0.071)	1.497(-1.179)
		100	0.170(-0.114)	0.035(-0.04)	1.323(-1.127)
		200	0.072(-0.106)	0.017(-0.023)	1.246(-1.118
	H_n^*	50	-	0.107(0.012)	0.313(-0.064)
		100	2.786(1.637)	0.052(-0.004)	0.204(-0.050)
		200	2.726(1.638)	0.099(0.282)	0.124(-0.092)

Table 3: MSE and bias (in the parentheses) of Beta distribution with different parameters

t	Estimator	n	Beta(1,1)	Beta(0.5,0.5)	Beta(2,2)
0.25	H_n	50	0.002(-0.004)	0.006(0.01)	0.003(-0.0004)
		100	0.001(-0.003)	0.003(0.008)	0.001(-0.001)
		200	0.0003(-0.001)	0.002(0.004)	0.001(-0.001)
	H_n^*	50	0.419(0.638)	0.066(0.158)	0.493(0.696)
		100	0.382(0.614)	0.147(0.340)	0.467(0.680)
		200	0.354(0.593)	0.246(0.481)	0.450(0.669)
0.5	H_n	50	0.002(-0.003)	0.007(0.012)	0.006(-0.002)
		100	0.001(-0.002)	0.004(0.008)	0.002(-0.009)
		200	0.0003(-0.001)	0.002(0.004)	0.001(0.001)
	H_n^*	50	0.186(0.424)	0.083(0.193)	0.246(0.492)
		100	0.156(0.392)	0.125(0.336)	0.218(0.464)
		200	0.136(0.368)	0.092(0.295)	0.197(0.443)
0.75	H_n	50	0.004(-0.002)	0.01(0.013)	0.014(0.027)
		100	0.001(-0.001)	0.005(0.008)	0.006(0.019)
		200	0.001(-0.001)	0.003(0.022)	0.003(0.013)
	H_n^*	50	0.100(0.300)	0.107(0.291)	0.158(0.360)
		100	0.067(0.252)	0.035(0.173)	0.079(0.273)
		200	0.048(0.218)	0.035(0.173)	0.079(0.273)

3.2 Real Data Analysis

In this section, we apply the proposed estimator (H_n) and its competitor (H_n^*) to analyse the automobile brake pads dataset from Lawless (2011). The data set includes the 98 automobile brake pads (in 1000-km units) for which each car brake pad lifetime is left truncated at the current odometer reading. By analyzing several parametric models, Lawless (2011) graphically showed that the log-normal distribution fits the data well. Considering this distribution, we have obtained $H(1) = 0.643$ as a residual entropy estimation for real data. Applying the test statistics W_n proposed by Addona and Wolfson (2006) for testing the stationarity assumption, we found that the underlying truncation time of brake pad data is uniformly distributed. Thus, the test does not reject the null hypothesis that the data follow a length-biased distribution at significance level $\alpha = 0.05$. We calculate $H_n(1)$ and $H_n^*(1)$, for the data. The observed values of these estimators were approximately 0.395 and 1.291. It can be seen that for these data the estimated residual entropy due to H_n is closer to H than H_n^*.

4 Conclusions

In this paper, we introduced an estimator for the residual entropy under length-biased samples, which will be useful for analysis and modeling of lifetime data. We considered the estimator in (1.6) as the plug-in estimator for $H(t)$, where f_n was proposed by Jones

(1991). In continuation, the asymptotic properties of $H_n(t)$ were studied. To this goal the strong convergence and asymptotic normality of the proposed estimator was derived. To evaluate that, we simulated length-biased samples from three different distributions. The Monte Carlo mean square error and bias were calculated for each estimator and revealed that the proposed estimator has a lower bias and mean square error than the competitor estimator. Finally for a real data set, we have shown that the data follow a length-biased distribution and the proposed estimator provides a better value for the estimated residual entropy in comparison to the competitor estimator.

More properties and applications of the new proposed estimator can be considered in future researches. Also due to the charm of censorship, this work can be extended to length-biased and right censored data. Further, the characterization problem of distributions can be considered by the new proposed estimator with length-biased data. Some new properties of the residual entropy in connection to order statistics and record values can be derived in the length-biased setting.

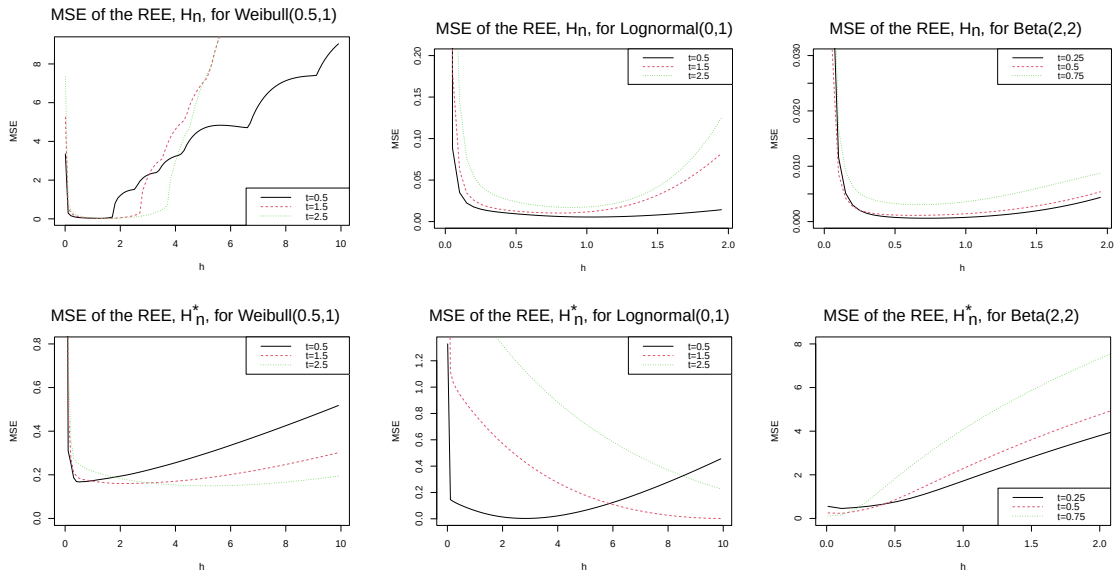

Figure 1: MSE as a function of bandwidth for proposed residual entropy estimator (REE) and its competitor.

Acknowledgements

The authors thank the Editor, Professor Nematollahi, the anonymous Associate Editor and the referees for their exceptionally insightful and constructive comments.

References

Addona, V., and Wolfson, D. B. (2006), A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime data analysis*, **12**(3), 267-284.

Ajami, M., Fakoor, V., and Jomhoori, S. (2013), Some asymptotic results of kernel density estimator in length-biased sampling. *Journal of Sciences, Islamic Republic of Iran*, **24**.

Akbari, M., Rezaei, M., Jomhoori, S., and Fakoor, V. (2019), Nonparametric estimators for quantile density function under length-biased sampling. *Communications in Statistics-Theory and Methods*, **48**(19), 4918-4935.

Asadi, M., and Ebrahimi, N. (2000), Residual entropy and its characterizations in terms of hazard function and mean residual life function. *Statistics & probability letters*, **49**(3), 263-269.

Belzunce, F., Guillamón, A., Navarro, J., and Ruiz, J. (2001), Kernel estimation of residual entropy. *Communications in Statistics-Theory and Methods*, **30**(7):1243-1255.

Belzunce, F., Navarro, J., Ruiz, J. M., and del Aguila, Y. (2004), Some results on residual entropy function. *Metrika*, **59**(2),147-161.

Chakraborty, R., and Rao, C. R. (2000), 23 selection biases of samples and their resolutions. *Handbook of statistics*, **18**, 675-712.

Chaubey, Y., Sen, P., and Li, J. (2010), Smooth density estimation for length-biased data. *Journal of the Indian Society of Agricultural Statistics*, **64(2)**,145-155.

Cox, D. R. (1969), Some sampling problems in technology. *New developments in survey sampling*, 506-527.

De Uña-álvarez, J. (2004), Nonparametric estimation under length-biased sampling and type i censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, **56**(4), 667-681.

Ebrahimi, N. (1996), How to measure uncertainty in the residual life time distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, 48-56.

Ebrahimi, N. (1997), Testing whether lifetime distribution is decreasing uncertainty. *Journal of statistical planning and inference*,**64**(1), 9-19.

Guillamón, A., Navarro, J., and Ruiz, J. (1998), Kernel density estimation using weighted data. *Communications in Statistics-Theory and Methods*, **27**(9), 2123-2135.

Horváth, L. (1985), Estimation from a length-biased distribution. *Statistics and Decisions*, **3**, 91-113.

Jones, M. C. (1991), Kernel density estimation for length biased data. *Biometrika*, **78**(3), 511-519.

Kayal, S. (2015), Generalized residual entropy and upper record values. *Journal of Probability and Statistics*, 2015.

Kayal, S., Madhavan, S. S., and Ganapathy, R. (2017), On dynamic generalized measures of inaccuracy. *Statistica*, **7**7(2):133-148.

Kumar, V., and Taneja, H. (2015), Dynamic cumulative residual and past inaccuracy measures. *Journal of Statistical Theory and Applications*, **14**(4), 399-412.

Lawless, J. F. (2011), *Statistical models and methods for lifetime data*, **362**. John Wiley & Sons.

Maya, R., Abdul-Sathar, E., Rajesh, G., and Nair, K. M. (2014), Estimation of the renyi's residual entropy of order α with dependent data. *Statistical Papers*, **55**(3), 585-602.

Nair, K., and Rajesh, G. (1998), Characterization of probability distributions using the residual entropy function. *J. Indian Statist. Assoc*, **36**, 157-166.

Oliazadeh, F., Iranmanesh, A., and Fakoor, V. (2021), A note on the strong consistency of nonparametric estimation of shannon entropy in length-biased sampling. *Communications in Statistics-Theory and Methods*, **50**(24), 5779-5791.

Patil, G. P. and Rao, C. R. (1978), Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 179-189.

Qin, J. (2017), *Biased sampling, over-identified parameter problems and beyond*. Springer.

Rajesh, G., Abdul-Sathar, E., Maya, R., Nair, K. M., et al. (2015), Nonparametric estimation of the residual entropy function with censored dependent data. *Brazilian Journal of Probability and Statistics*, **29**(4), 866-877.

Rajesh, G., Abdul Sathar, E., and Viswakala, K. (2017), Estimation of inaccuracy measure for censored dependent data. *Communications in Statistics-Theory and Methods*, **46**(20), 10058-10070.

Sankaran, P., and Gupta, R. (1999), Characterization of lifetime distributions using measure of uncertainty. *Calcutta Statistical Association Bulletin*, **49**(3-4), 159-166.

Simon, R. (1980), Length biased sampling in etiologic studies. *American Journal of Epidemiology*, **111**(4), 444-452.

Tahmasebi, S., and Daneshi, S. (2018), Measures of inaccuracy in record values. *Communications in Statistics-Theory and Methods*, **47**(24), 6002-6018.

Vardi, Y. (1982), Nonparametric estimation in renewal processes. *The Annals of Statistics*, 772-785.