

A presmooth estimator of unbiased distributions with length-biased data

Reza Heidari, Vahid Fakoor & Ali Shariati

Mathematical Sciences

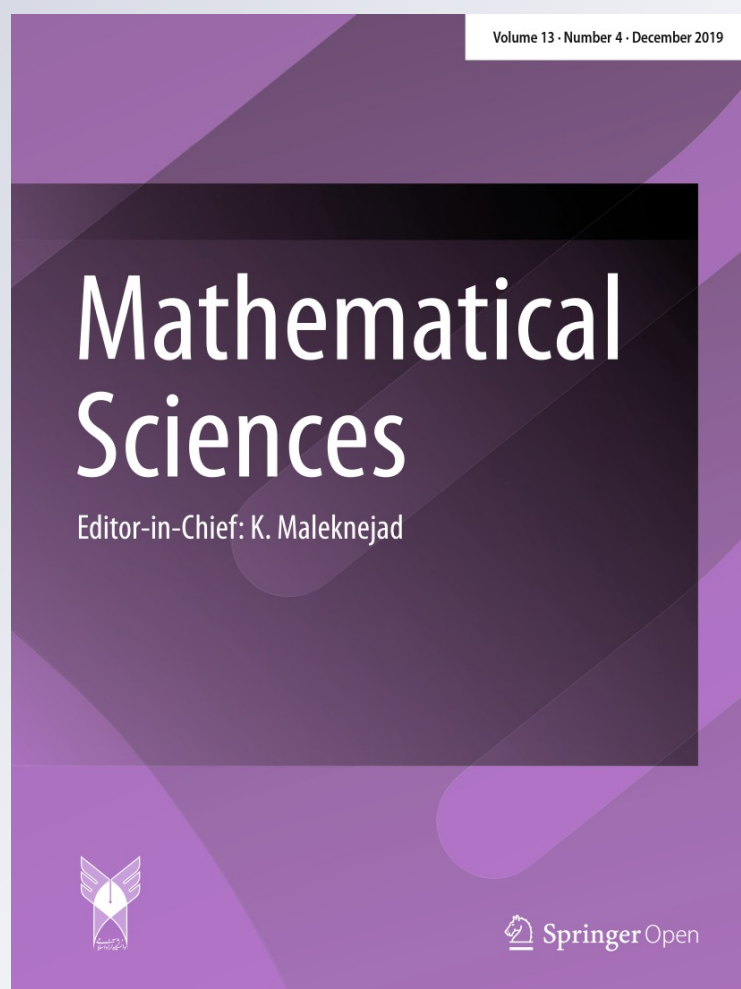
ISSN 2008-1359

Volume 13

Number 4

Math Sci (2019) 13:317-323

DOI 10.1007/s40096-019-00301-z



Your article is published under the Creative Commons Attribution license which allows users to read, copy, distribute and make derivative works, as long as the author of the original work is cited. You may self-archive this article on your own website, an institutional repository or funder's repository and make it publicly available immediately.



A presmooth estimator of unbiased distributions with length-biased data

Reza Heidari¹ · Vahid Fakoor¹ · Ali Shariati¹

Received: 19 July 2019 / Accepted: 10 September 2019 / Published online: 18 September 2019
© The Author(s) 2019

Abstract

In this paper, we propose a presmooth product-limit estimator to draw statistical inference on the unbiased distribution function representing the population of interest. The strong consistency of the estimator proposed is investigated. The finite sample performance of the proposed estimator is evaluated using simulation studies. It is observed that the proposed estimator exhibits greater efficiency in comparison with the alternative method in de Uña-Álvarez (Test 11(1):109–125, 2002).

Keywords Kaplan–Meier estimator · Length-biased data · Presmooth estimator · Random right censoring

Introduction

In practical studies, it is not feasible sometimes to collect a random sample from the population of interest, called the target population. In such occasions, a biased or weighted sample from the underlying population is obtained. In other words, the data are collected from another distribution, which may induce a type of bias to our inferences. Thus, the biased sampling problem is to make inference about the target population, while our samples are generated from another distribution function. The most common case, called length-biased, is when the biased data are collected according to the chance proportional to their value (measures, lengths, etc.), arising in many types of sampling. There can be found many examples of such situations in various articles and books from different disciplines, e.g., biology, biotechnology, genetics, forestry, economics, industry and medical science. For more information, we refer readers to the following studies: Rao [19, 20], Patil and Ord [18], Zelen and Feinleib [28], Song et al. [22] and Kvam [12].

Wicksell [26] found in the study blood cells microscope that only those cells which are bigger than a threshold are detectable and the existing smaller cells are not visible. At that point in time, he called this phenomenon corpuscle problem, which was later known as the length-biased sampling. McFadden [15], Blumenthal [2] and Cox [5] are of the first scientists who address this phenomenon in statistics. In the past two decades, many more studies have been performed to extend the statistical inferences in regard to the problem of length-biased sampling [21, 14, 17].

For example, Efromovich [8] attempted to draw statistical inference for the police investigation, in which they looked at the ratio of alcohol in liquor-intoxicated drivers. Since there is a higher chance for drunker drivers to be suspected by the police, their data collection suffers from the length-biased problem.

Definition 1 Suppose that $F(\cdot)$ is an absolutely continuous cumulative distribution function. The random variable Y has the length-bias distribution corresponding to $F(\cdot)$, if it satisfies the following distribution function.

$$G(t) = \mu^{-1} \int_0^t u dF(u), \quad t \geq 0, \quad (1.1)$$

$$\text{where } \mu = \int_0^\infty u dF(u).$$

It is easy to obtain Eq. (1.1) for the distribution function $F(\cdot)$ as follows

✉ Vahid Fakoor
fakoor@um.ac.ir

Reza Heidari
reza.heidary@mail.um.ac.ir

Ali Shariati
ali-shariati@mail.um.ac.ir

¹ Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

$$F(t) = \frac{\int_0^t u^{-1} dG(u)}{\int_0^\infty u^{-1} dG(u)}, \quad t \geq 0. \quad (1.2)$$

Here, the problem is to draw statistical inference about the underlying population ($F(\cdot)$), while the available data are collected from length-bias distribution $G(\cdot)$. Accordingly, when encountering the length-biased sampling problem, we obtain a data collection which may be used to estimate the distribution function $G(\cdot)$ empirically. Having obtained the empirical estimation of $G(\cdot)$, the distribution function $F(\cdot)$ can be estimated using Eq. (1.2). For a sample consisting of complete observations (excluding censored data), the non-parametric estimation of the distribution function $F(\cdot)$ under length bias has been investigated by Vardi et al. [25], Vardi [24], Horváth et al. [9] and Jones [10].

The other obstacle commonly faced, specially in survival analysis, is that some of the subjects are not completely observed owing to right censoring. There have been many studies in the literature concerning length-biased and right-censored data. Let Y_1, Y_2, \dots, Y_n denote iid random variables with distribution function $G(\cdot)$. Also, suppose that C_1, C_2, \dots, C_n are iid random variables from the distribution function $\bar{G}(\cdot)$ and independent from Y_1, Y_2, \dots, Y_n . Define $I(A)$ as the indicator function of the event A . Under the right random censorship model, observations are $\{(Z_i, \delta_i); i = 1, 2, \dots, n\}$ in which the variables $Z_i = \min(Y_i, C_i)$ are iid copies of distribution function $H(\cdot)$ and $\delta_i = I(Y_i \leq C_i)$. C_i denotes the censored random variable, and $\delta_i = 0$ indicates that the i th subject is censored, while $\delta_i = 1$ denotes for the uncensored observations. We are interested in estimating the distribution function $F(\cdot)$ based on the pair of observations (Z_i, δ_i) .

In the presence of length bias and right censoring, Winter and Földes [27] introduced a conditional method to estimate the distribution function $F(\cdot)$. Their proposed estimator is applicable under the left-truncation of observations in general, when the values of truncation variable for each subject are specified. de Uña-Álvarez [6] showed that ignoring the extra information of the length-bias model (unconditional approach) in the structure of conditional estimators, like Winter and Földes [27], results in lower efficiency.

When a sample only includes censored data, but not length-biased or truncated observations, the nonparametric maximum likelihood product-limit estimator [11] may be used. The Kaplan–Meier estimator of distribution function $F(\cdot)$ is defined as follows

$$F_n^{\text{KM}}(t) = 1 - \prod_{i=1: Z_{(i)} \leq t} \left[1 - \frac{\delta_{[i]}}{n - i + 1} \right], \quad (1.3)$$

where the $Z_{(i)}$ variables are the ordered observations of Z_i and $\delta_{[i]}$ is the responsible value of δ_i for the variable $Z_{(i)}$. Equation (1.3) could be simply rewritten as follows

$$F_n^{\text{KM}}(t) = \sum_{i=1}^n w_i I_{(Z_{(i)} \leq t)}, \quad (1.4)$$

where

$$w_i = \frac{\delta_{[i]}}{n - i + 1} \prod_{j=1}^{i-1} \left[1 - \frac{\delta_{[j]}}{n - j + 1} \right].$$

To study the statistical properties of the estimator $F_n^{\text{KM}}(\cdot)$, we refer readers to Andersen et al. [1]. It is deduced from Eq. (1.4) that the Kaplan–Meier estimator is a step function with jumps at the non-censored observations. The sizes of jumps in each step depend not only on the complete observations, but also on the number of censored observations prior to the step. It can be easily checked that when the sample does not consist of any censored data, the Kaplan–Meier estimator is equivalent to the empirical distribution function estimator.

de Uña-Álvarez [6] first estimated the distribution function $G(\cdot)$ on the basis of biased observations using the Kaplan–Meier estimator. Afterward, by substituting the estimator for $G(\cdot)$ in Eq. (1.2), they have obtained the following estimator, called the length-bias-corrected product-limit estimator, for the target distribution.

$$\hat{F}_n(t) = \frac{\int_0^t u^{-1} dG_n^{\text{KM}}(u)}{\int_0^\infty u^{-1} dG_n^{\text{KM}}(u)}, \quad t \geq 0. \quad (1.5)$$

It is simply followed from (1.5) that

$$\begin{aligned} \hat{F}_n(t) &= \frac{\sum_{i=1}^n w_i Z_{(i)}^{-1} I_{(Z_{(i)} \leq t)}}{\sum_{i=1}^n w_i Z_{(i)}^{-1}} \\ &= \sum_{i=1}^n \tilde{w}_i I_{(Z_{(i)} \leq t)}, \end{aligned} \quad (1.6)$$

where

$$\tilde{w}_i = \frac{w_i Z_{(i)}^{-1}}{\sum_{i=1}^n w_i Z_{(i)}^{-1}}.$$

\tilde{w}_i is calculated applying the jumps of Kaplan–Meier estimator in (1.4).

In the rest of this study, we first introduce the presmooth estimator in “The presmooth estimator” section. Next, by substituting the presmooth estimator for the estimator $G_n^{\text{KM}}(\cdot)$ in Eq. (1.5), we obtain a smoother method

for the distribution function $F(\cdot)$, which is expected to possess higher efficiency in predicting the true distribution function. In “**Strong consistency**” section, the strong consistency of the proposed estimator is investigated. Finally, the simulations studies for two estimators are presented and their behaviors are discussed in “**Simulation**” section.

The presmooth estimator

In this section, inspired by the expression of the product-limit estimator of de Uña-Álvarez [6], we will propose the presmooth estimator for the distribution function using length-biased right-censored data.

Presmooth estimator in LB distribution

Suppose that

$$p(z) := P(\delta = 1 \mid Z = z) = E(\delta \mid Z = z)$$

denote the conditional probability of non-censored event given the observation $Z = z$. Cao et al. [4] introduced a presmooth estimator by substituting the nonparametric regression estimator proposed in Nadaraya [16] for the censoring indicator variables. Following that, Cao and Jácome [3] assessed the limit distribution and the asymptotic mean squared error of the estimator presented in Cao et al. [4]. Stute and Wang [23] discussed the significance of $p(\cdot)$ to prove the consistency of an integral $\int \varphi dF_n^{\text{KM}}$ where φ is a measurable function over \mathbb{R} and F_n^{KM} is the Kaplan–Meier estimator. Given the pairs observations $\{(Z_i, \delta_i); i = 1, 2, \dots, n\}$, we propose the following estimator for $p(z)$,

$$p_n(z) = \frac{\sum_{i=1}^n K_b(z - Z_i) \delta_i}{\sum_{i=1}^n K_b(z - Z_i)}, \quad (2.1)$$

where $K(\cdot)$ is a kernel function and $K_b(\cdot) = \left(\frac{1}{b}\right)K\left(\frac{\cdot}{b}\right)$, in which $\{b \equiv b_n, n = 1, 2, \dots\}$ is the sequence of bandwidth.

By replacing $\delta_{[.]}$ in Eq. (1.4) with $p_n(\cdot)$, the presmooth estimator of the distribution function $F(\cdot)$ is defined as follows:

$$F_n^P(t) = \sum_{i=1}^n v_i I_{(Z_{(i)} \leq t)}, \quad (2.2)$$

where

$$v_i = \frac{p_n(Z_{(i)})}{n - i + 1} \prod_{j=1}^{i-1} \left[1 - \frac{p_n(Z_{(j)})}{n - j + 1} \right].$$

Remark 1 Even though the proposed presmooth estimator seems similar to the Kaplan–Meier estimator in (1.4), they mainly differ in $\delta_{[.]}$ which has been replaced with the smoothing function $p_n(Z_{(.)})$ in (2.2). Adapting the Kaplan–Meier estimator to the smoother method by plugging in $p_n(Z_{(.)})$ might be invaluable, as the presmooth estimators exhibit superior accuracy in estimating the distribution function. By contrast, the Kaplan–Meier estimator only assigns equal jumps to all of the complete observations, excluding the censored data.

Remark 2 It is of note that when $n \rightarrow \infty$, the bandwidth defined in (2.1) decreases and converges to 0, and very small values of $b \simeq 0$ implies $p_n(Z_i) \simeq \delta_i$. Accordingly, when $n \rightarrow \infty$, the presmooth and the Kaplan–Meier estimators become equal.

The presmooth estimator of the unbiased distribution

In this section, we present a presmooth product-limit estimator for a distribution function in the length-biased and random right-censored sampling. For this purpose, we substitute the presmooth estimator of the right-censored data, say $G_n^P(\cdot)$, for the distribution function $G(\cdot)$ in Eq. (1.5). Hence, we obtained the presmooth product-limit estimator through

$$\hat{F}_n^P(t) = \frac{\int_0^t u^{-1} dG_n^P(u)}{\int_0^\infty u^{-1} dG_n^P(u)}, \quad t \geq 0. \quad (2.3)$$

Given Eq. (2.3), it can be deduced that

$$\begin{aligned} \hat{F}_n^P(t) &= \frac{\sum_{i=1}^n v_i Z_{(i)}^{-1} I_{(Z_{(i)} \leq t)}}{\sum_{i=1}^n v_i Z_{(i)}^{-1}} \\ &= \sum_{i=1}^n \tilde{v}_i I_{(Z_{(i)} \leq t)}, \end{aligned} \quad (2.4)$$

with

$$\tilde{v}_i = \frac{v_i Z_{(i)}^{-1}}{\sum_{i=1}^n v_i Z_{(i)}^{-1}}.$$

\tilde{v}_i is calculated from the jumps of a presmooth estimator of distribution function in Eq. (2.2).

Corollary 1 Cumulative hazard rate function of $F(\cdot)$ is defined as

$$\Lambda(t) = \int_0^t \frac{1}{1 - F(u^-)} dF(u). \quad (2.5)$$

By plugging in $\hat{F}_n^P(\cdot)$ in Eq. (2.5), the presmooth estimator of the cumulative hazard rate can be obtained by means of

$$\begin{aligned} \hat{\Lambda}_n^P(t) &= \int_0^t \frac{1}{1 - \hat{F}_n^P(u^-)} d\hat{F}_n^P(u) \\ &= \sum_{i=1}^n \frac{\tilde{v}_i}{1 - \hat{F}_n^P(Z_{(i)}^-)} I_{(Z_{(i)} \leq t)} \\ &= \sum_{i=1}^n \frac{v_i Z_{(i)}^{-1}}{\sum_{j=1}^n v_j Z_{(j)}^{-1} I_{(Z_{(i)} \leq Z_{(j)})}} I_{(Z_{(i)} \leq t)}. \end{aligned} \quad (2.6)$$

Strong consistency

In this section, we study the strong consistency of $\hat{F}_n^P(\cdot)$. For this purpose, we define

$$\tau_F = \inf\{t : F(t) = 1\}.$$

Similarly, τ_G , $\tau_{\tilde{G}}$ and τ_H are defined for the distribution functions $G(\cdot)$, $\tilde{G}(\cdot)$ and $H(\cdot)$. Apparently, it could be claimed that $\tau_F = \tau_G$ and $\tau_H = \min(\tau_F, \tau_{\tilde{G}})$.

Theorem 1 Let $\varphi(\cdot)$ be a measurable function, $\varphi_1(x) = x^{-1}$ and $\varphi_2(x) = x^{-1}\varphi(x)$ are $G(\cdot)$ -integrable. Assuming that $\tau_H = \tau_F$ is reasonable, then we have

$$\lim_{n \rightarrow \infty} \int \varphi(x) d\hat{F}_n^P(x) = \int \varphi(x) dF(x) \quad a.s. \quad (3.1)$$

Proof Suppose

$$\hat{S}_\varphi = \int \varphi(x) d\hat{F}_n^P(x),$$

and

$$S_\varphi = \int \varphi(x) dF(x).$$

Using Eq. (2.4), it is obtained that

$$\begin{aligned} \hat{S}_\varphi &= \sum_{i=1}^n \tilde{v}_i \varphi(Z_{(i)}) \\ &= \frac{\sum_{i=1}^n v_i Z_{(i)}^{-1} \varphi(Z_{(i)})}{\sum_{i=1}^n v_i Z_{(i)}^{-1}}. \end{aligned} \quad (3.2)$$

According to Theorem 2.1 of de Uña-Álvarez and Rodríguez-Campos [7], regardless of the presence of covariates in this case, when $n \rightarrow \infty$, we have

$$\tilde{\mu} = \frac{1}{\sum_{i=1}^n v_i Z_{(i)}^{-1}} \xrightarrow{a.s.} \frac{1}{E(\varphi_1(Y))} = \frac{1}{\int \varphi_1(y) dG(y)} = \mu, \quad (3.3)$$

and

$$\sum_{i=1}^n v_i Z_{(i)}^{-1} \varphi(Z_{(i)}) \xrightarrow{a.s.} E(\varphi_2(Y)) = \int \varphi_2(y) dG(y). \quad (3.4)$$

Now, considering Eqs. (3.2)–(3.4), it can be deduced that

$$\hat{S}_\varphi \xrightarrow{a.s.} S_\varphi.$$

Therefore, the proof of Theorem 1 is completed. \square

Corollary 2 Let $0 < \mu < \infty$. Using Theorem 1, we have

$$\hat{F}_n^P(t) \xrightarrow{a.s.} F(t)$$

for any $t > 0$, when $n \rightarrow \infty$.

Moreover, given the relation $\Lambda(t) = -\ln(1 - F(t))$, if $n \rightarrow \infty$, we have

$$\Lambda_n^P(t) \xrightarrow{a.s.} \Lambda(t).$$

Simulation

In this section, simulation studies are carried out to inspect the finite sample performance of the presmooth limit-product estimator. For better illustration, we have examined the proposed estimator behavior (2.3) in comparison with the product-limit estimator introduced in (1.5). For this purpose, we have studied the relative efficiency of the two estimators (RE) through the ratio of their values of $MSE(\cdot)$, which is defined as follows.

$$RE(t) = \frac{MSE(\hat{F}_n(t))}{MSE(\hat{F}_n^P(t))}, \quad (4.1)$$

where

$$MSE(\hat{F}_n(t)) = E(\hat{F}_n(t) - E(\hat{F}_n(t)))^2,$$

and

$$MSE(\hat{F}_n^P(t)) = E(\hat{F}_n^P(t) - E(\hat{F}_n^P(t)))^2.$$

We have approximated the value of RE in (4.1) using Monte Carlo simulations. The amounts of RE are calculated based on $B = 5000$ replications of estimations using samples sizes $n = 20, 50$ and 100 . Suppose distribution $F(\cdot)$ is a member of gamma family with density

$$f(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} e^{-\frac{t}{\beta}}, \quad \alpha > 0, \beta > 0, t \geq 0.$$

Given Eq. (1.1), it can be easily deduced that if the distribution of target population is $\text{Gamma}(\alpha, \beta)$, the resulting length-biased distribution is $\text{Gamma}(\alpha + 1, \beta)$. Similarly, the corresponding length-biased distribution to the $\text{Weibull}(p, \lambda)$ is a generalized gamma distribution, $G(\mu, \sigma, \nu)$, with the probability density function:

$$f(y) = \frac{\nu y^{\nu-1}}{(\sigma/\mu)^\nu \Gamma(\mu)} y^{\nu(\mu-1)} \exp(-(y\mu/\sigma)^\nu),$$

where $\mu = 1 + 1/p$, $\sigma = 1/\lambda$, and $\nu = p$ are the shape, scale and family parameters, respectively.

Figure 1 compares the performance of the presmooth estimator with that of the product-limit estimator in predicting the survival function of the $U(0, 4)$ distribution. The figure consists in 5000 iterations for the moderate sample scenario ($n = 50$). It can be observed that applying the proposed method has considerably reduced the amount of deviations from the true survival function by comparison with the product-limit estimator.

Figures 2 and 3 illustrate the approximated values of $\text{RE}(\cdot)$ for the $\text{Weibull}(0.5, 1)$ and $\text{Gamma}(1, 1)$ (the exponential distribution, $\text{EXP}(1)$) unbiased distributions, respectively. The diagrams were estimated based on 5000 iterations

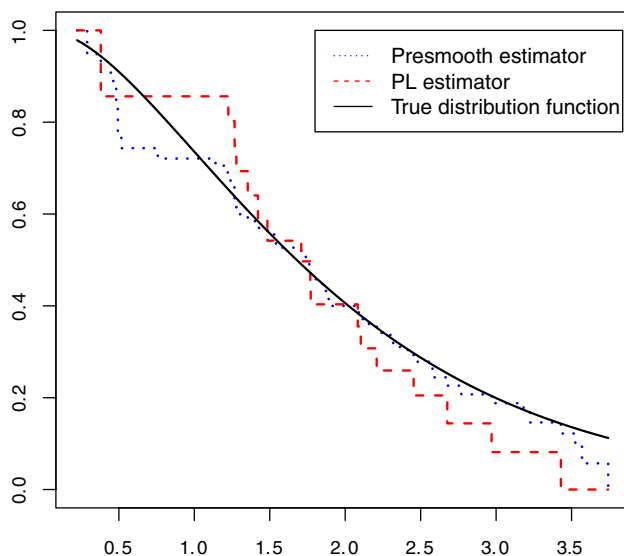


Fig. 1 Presmooth estimator versus PL-estimator for $\text{Exp}(1)$ with $n = 50$ and $C \sim U(0, 4)$

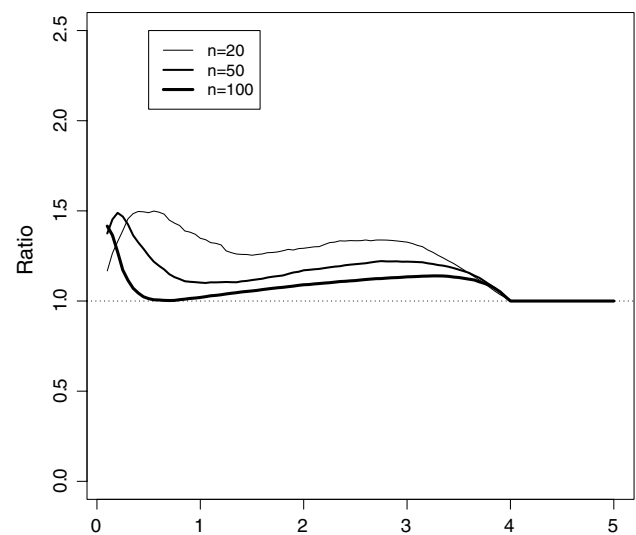


Fig. 2 Curves obtained for simulated AR with different sample size taken from $\text{Weibull}(0.5, 1)$ and 24% censoring

of the different sample sizes generated from the corresponding length-biased observations. The $U(0, 4)$ distribution was considered as the censoring distribution, resulting in about 24% incomplete data in all surveys. Considering the sample sizes $n = 20$ and 50 in the both diagrams, it can be observed that the proposed method for all values of t exhibited superiority over the product-limit estimator in terms of MSE . Similarly, the proposed presmooth estimator indicated much better results for the large sample scenario ($n = 100$) of the $\text{Weibull}(0.5, 1)$ target population (Fig. 2). However, turning to the large sample scenario ($n = 100$) of the $\text{Gamma}(1, 1)$

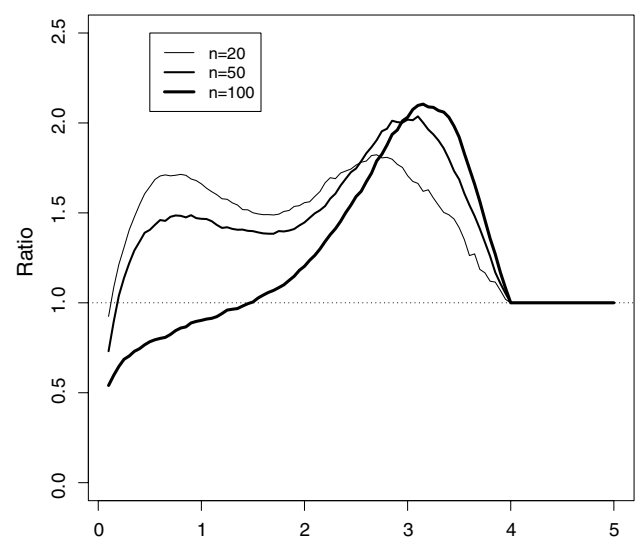


Fig. 3 Curves obtained for simulated AR with different sample size and 32% censoring

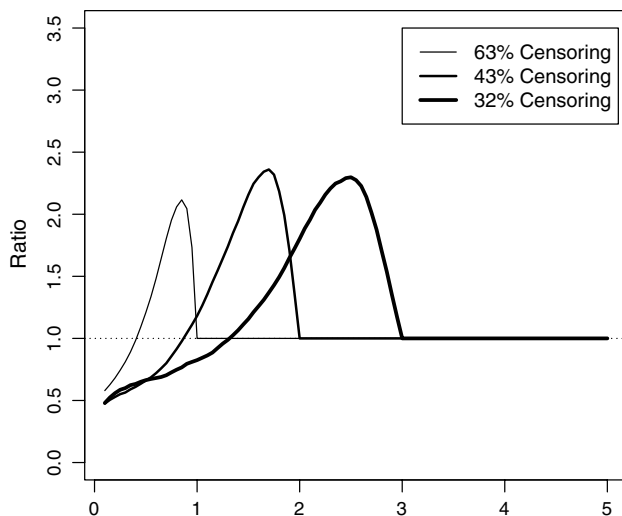


Fig. 4 Efficiency comparison for Exp(1) with $n = 100$ and different values of censoring rate

unbiased population, while for the values of t smaller than roughly 0.5 the product-limit estimator performed better, the proposed method revealed better efficiency for t larger than 0.5.

For better illustration, we have compared the efficiency of the estimators under different levels of censoring via RE in Fig. 4. For this purpose, we calculated the values of RE based on 5000 replications of the large sample scenario ($n = 100$). The data were generated from the length-biased distribution corresponding to the Exp(1) target population. The censoring times were generated from the $U(0, \beta)$ distribution, in which the values of 3, 2 and 1 were chosen for β resulting in 63%, 43% and 32% incomplete observations, respectively. Accordingly, it can be obtained that although for the small values of t the product-limit estimator exhibited better results, the presmooth estimator has significantly reduced the amount of MSE in all levels of censoring as the value of t increased. Broadly, the presmooth estimator exhibited superior efficiency in comparison with the product-limit method.

It has been revealed in the all above RE diagrams that the figures have become closer to the line $RE = 1$ by rising in the sample sizes. This increasing tendency for the two methods to perform similar as the sample size increased is clearly justified by Remark 2.

To calculate the estimated $\hat{F}_n^P(\cdot)$, the Nadaraya–Watson estimator method, presented in Eq. (2.1), has been used to estimate $p(\cdot)$. Thus, it is crucial to select the bandwidth properly as it plays a key role the Nadaraya–Watson estimator. For this purpose, Cao and Jácome [3] and Cao et al. [4] obtained an optimal bandwidth for their presmooth pug-in method by minimizing the asymptotic mean integrated squared error, avoiding any bias in results. Later, the

simulation results for this estimator was studied through the Simpson’s rule using the **survPresmooth** package López-de Ullibarri and Jácome Pumar [13].

Conclusion

In this paper, we have proposed a presmooth estimator by adapting the product-limit estimator of de Uña-Álvarez [6] for length-biased and (random) right-censored data. The limit properties of the proposed estimator have been investigated. To inspect the performance of the method, simulations studies were conducted to make comparisons between the proposed estimator and the product-limit estimator of de Uña-Álvarez [6]. As mentioned, it is very important to select the bandwidth for the Nadaraya–Watson regression estimation appropriately. We have overcome this issue by applying the $MISE(\cdot)$ estimator.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. Springer, Berlin (2012)
- Blumenthal, S.: Proportional sampling in life length studies. *Technometrics* **9**(2), 205–218 (1967)
- Cao, R., Jácome, M.: Presmoothed kernel density estimator for censored data. *Nonparametr. Stat.* **16**(1–2), 289–309 (2004)
- Cao, R., López-de Ullibarri, I., Janssen, P., Veraverbeke, N.: Presmoothed Kaplan–Meier and Nelson–Aalen estimators. *J. Nonparametr. Stat.* **17**(1), 31–56 (2005)
- Cox, D.R.: Some sampling problems in technology. In: Johnson, N.L., Smith Jr., H. (eds.) *New Developments in Survey Sampling*, pp. 506–527. Wiley, New York (1969)
- de Uña-Álvarez, J.: Product-limit estimation for length-biased censored data. *Test* **11**(1), 109–125 (2002)
- de Uña-Álvarez, J., Rodríguez-Campos, M.C.: Strong consistency of presmoothed Kaplan–Meier integrals when covariables are present. *Statistics* **38**(6), 483–496 (2004)
- Efromovich, S.: *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer, Berlin (2008)
- Horváth, L., et al.: Estimation from a length-biased distribution. *Stat. Decis.* **3**, 91–113 (1985)
- Jones, M.: Kernel density estimation for length biased data. *Biometrika* **78**(3), 511–519 (1991)
- Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958)
- Kvam, P.: Length bias in the measurements of carbon nanotubes. *Technometrics* **50**(4), 462–467 (2008)
- López-de Ullibarri, I., Jácome Pumar, M.A.: *survpresmooth: An R Package for Presmoothed Estimation in Survival Analysis*. American Statistical Association, Alexandria (2013)

14. McCullagh, P.: Sampling bias and logistic models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **70**(4), 643–677 (2008)
15. McFadden, J.A.: On the lengths of intervals in a stationary point process. *J. R. Stat. Soc. Ser. B (Methodol.)* **24**(2), 364–382 (1962)
16. Nadaraya, E.A.: On estimating regression. *Theory Probab. Appl.* **9**(1), 141–142 (1964)
17. Ning, J., Qin, J., Shen, Y.: Non-parametric tests for right-censored data with biased sampling. *J. R. Stat. Soc. Ser. B (Methodol.)* **72**(5), 609–630 (2010)
18. Patil, G.P., Ord, J.K.: On size-biased sampling and related form-invariant weighted distributions. *Sankhy Indian J. Stat. Ser. B (1960–2002)* **38**(1), 48–61 (1976)
19. Rao, C.R.: On discrete distributions arising out of methods of ascertainment. *Sankhy Indian J. Stat. Ser. A (1961–2002)* **27**(2/4), 311–324 (1965)
20. Rao, C.R.: A natural example of a weighted binomial distribution. *Am. Stat.* **31**(1), 24–26 (1977)
21. Scheike, T.H., Keiding, N.: Design and analysis of time-to-pregnancy. *Stat. Methods Med. Res.* **15**(2), 127–140 (2006)
22. Song, R., Karon, J.M., White, E., Goldbaum, G.: Estimating the distribution of a renewal process from times at which events from an independent process are detected. *Biometrics* **62**(3), 838–846 (2006)
23. Stute, W., Wang, J.-L.: The strong law under random censorship. *Ann. Stat.* **21**(3), 1591–1607 (1993)
24. Vardi, Y.: Empirical distributions in selection bias models. *Ann. Stat.* **13**(1), 178–203 (1985)
25. Vardi, Y., et al.: Nonparametric estimation in the presence of length bias. *Ann. Stat.* **10**(2), 616–620 (1982)
26. Wicksell, S.D.: The corpuscle problem. A mathematical study of a biometric problem. *Biometrika* **17**(1–2), 84–99 (1925)
27. Winter, B., Földes, A.: A product-limit estimator for use with length-biased data. *Can. J. Stat.* **16**(4), 337–355 (1988)
28. Zelen, M., Feinleib, M.: On the theory of screening for chronic diseases. *Biometrika* **56**(3), 601–614 (1969)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.