JIRSS (2024)
Vol. 23, No. 02, pp 69-96
DOI: 10.22034/jirss.2025.2013325.1040

Analysis of Restricted Mean Survival Time for Length-biased Data via Empirical Likelihood

Zahra Mohammadian[1], Vahid Fakoor[1], Arezou Habibirad[1] and Hadi Jabbari[1]

[1] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Islamic Republic of Iran.

Received: 10/12/2023, Accepted: 25/03/2025, Published online: 08/06/2025

Abstract. The restricted mean survival time (RMST) in the context of length-biased data is an important addition to clinical studies. The RMST is a widely used measurement for evaluating survival over a specific period, and the area under the survival function is a key component of this metric. However, when the data under study are length-biased, traditional parametric and classical methods for examining the RMST are not applicable. Nonparametric and semi-parametric methods are used to address this issue. We utilize the empirical likelihood (EL) method to investigate RMST. Our proposed EL procedure provides a reliable approach for inferential analysis of RMST in the presence of length-biased data. We have shown that the limiting distribution of the empirical log-likelihood ratio is a chi-square distribution with one degree of freedom. We also demonstrated that the likelihood ratio exhibits weak convergence to a mean-zero Gaussian process, which we used to construct a confidence band. In our simulation section, we compared the confidence intervals obtained from the normal approximation (NA) and EL methods. We showed that the EL method has a better coverage probability than the NA method. Additionally, we provided a real data application using bank customers' monthly taxes to illustrate further the effectiveness of our proposed method.

Keywords. Confidence band, Empirical likelihood, Length-biased data, Normal approximation, Restricted mean survival time.

Zahra Mohammadian (zahra.mohammadian@mail.um.ac.ir)
Corresponding Author: Vahid Fakoor (fakoor@um.ac.ir)
Arezou Habibirad (ahabibi@um.ac.ir)
Hadi Jabbari (jabbarinh@um.ac.ir).

MSC: ???, ???, ???.

1 Introduction

The RMST is a vital metric in clinical trials for cancer treatments. In these trials, researchers aim to determine the survival time of patients who receive a particular treatment. However, some patients may not survive long enough to complete the entire follow-up period, which can introduce bias into the results. The RMST allows researchers to calculate the average survival time up to a specific time (e.g., 2 years) for all patients, including those who did not survive until the end of the follow-up period. Therefore, the RMST provides a more reliable estimate of the treatment's effectiveness and can help inform clinical decision-making. By using the RMST, researchers can more effectively evaluate the impact of a given treatment on patient outcomes, even when the follow-up period is incomplete due to patient mortality. This is particularly important in cancer treatment trials, where time is of the essence and precise assessment of survival rates can make all the difference. In summary, the RMST is a valuable tool for researchers in the field of cancer treatment, as it helps provide a more comprehensive evaluation of the treatment's effectiveness and can ultimately improve patient outcomes.

The RMST is a critical statistic used to measure treatment effects in clinical studies. In cases where the proportional hazards assumption is violated, the RMST can serve as an alternative to the hazard ratio. The RMST is defined as the average survival time of all individuals in the study population up to a specific time point (t). It is equivalent to the area under the survival curve up to time t. Suppose the random variable X follows a distribution function F. The RMST function is defined as

$$\mu(t) = \int_0^t S(u)du, \qquad t > 0, \tag{1.1}$$

where $S(u) = 1 - F(u)$. The empirical estimator for $\mu(t)$ is defined by

$$\mu_n(t) = \int_0^t S_n(u)du, \tag{1.2}$$

where $S_n(u) = 1-F_n(u)$, and $F_n(\cdot)$ is an empirical distribution function based on random sample $X_1, X_2, \ldots, X_n$ of $F(\cdot)$.

The RMST has been the subject of numerous studies since it was first introduced by Irwin (1949) in 1949. Researchers have explored various methods for estimating the RMST function, including parametric, semi-parametric, and nonparametric approaches. For example, Royston and Parmar (2002) proposed a parametric method for estimating the RMST function under right-censored data, which involves integrating the survival function. Chen and Tsiatis (2001) presented estimators for the difference in the restricted mean lifetime between two groups, taking into account imbalances in

prognostic factors and assuming a proportional hazards relationship. Andersen et al. (2004) used pseudo-observations to estimate the RMST function, while Zhang and Schaubel (2011) proposed a method for estimating the difference in the RMST when the survival time is dependently censored. The flexibility of the risk model allows for the creation of baseline hazards and regression coefficients. Zhao et al. (2016) investigated the RMST curve and proposed methods for inference based on simultaneous confidence bands for a single RMST curve and the difference between two RMST curves.

The issue of biased sampling was first described in the well-known Wicksell corpuscle problem by Wicksell in 1925 (Wicksell (1925)). In this problem, spheres with larger radii were more likely to be sampled, leading to biased sampling. Although Fisher (1934) did not use the term "biased sampling," he did address these issues. The concept of weighted distributions was introduced by Rao (1965) in 1965, and more examples of weighted distributions can be found in his work as well as in Patil and Rao (1978). Length-biased sampling occurs when the probability of selecting a sample is proportional to its length, meaning that longer samples are more likely to be selected. Many scholars have addressed issues related to length-biased sampling, including Cox (2005), who presented some solutions to these problems in 2005, Vardi (1989), who discussed multiplicative bias in length-biased sampling in 1989, and Zelen (2004), who proposed a forward approach for estimating the distribution of length-biased samples in 2004. Overall, the issue of biased sampling is an important consideration in many fields, including statistics and epidemiology.

The estimation of length-biased sampling has been extensively studied using both nonparametric and semiparametric methods. Vardi (1982) proposed a nonparametric maximum likelihood estimate for the cumulative distribution function based on length-biased data, while in Vardi (1989), he presented the asymptotic properties of the nonparametric maximum likelihood estimator under length-biased sampling. Asgharian et al. (2002) derived the maximum likelihood estimation of the length-biased distribution and established its asymptotic properties. In recent years, Shi et al. (2018) proposed a nonparametric estimator for the quantile of the distribution function for length-biased data. This estimator was based on an improved product-limit estimator of the distribution function that incorporates auxiliary information about the length-biased sampling scheme. Semiparametric estimation procedures have also been developed for the proportional hazards model with length-biased data. Wang (1996) proposed such procedures, while Shen (2009) extended this work in two ways for length-biased data. They presented a semiparametric estimator for the proportional hazards model and proposed an analogous estimator for the additive hazards model. More recently, Lee et al. (2018) suggested nonparametric and semiparametric regression strategies for estimating the RMST under the setting of length-biased data. In 2020, He and Zhou (2020) proposed nonparametric and semiparametric estimators of the RMST with length-biased data.

The EL method is a nonparametric inference method based on the likelihood ratio function. In many cases, the EL method is more accurate than the NA method, par-

ticularly when the variance estimate is unstable. The EL method was introduced by Thomas and Grunkemeier (1975) when he used it to obtain a nonparametric confidence interval for the probability of survival in a model. Since then, the EL method has been substantially developed by many researchers, including Owen, who used it as a practical method in statistical fields. In particular, Owen's work in Owen (1988) and Owen et al. (1990) made significant contributions to the development of the EL method. These contributions included the extension of the EL method to the case of dependent data, the development of a more efficient algorithm for computing EL confidence intervals, and the establishment of the asymptotic properties of the EL method. Zhao and Qin (2006) proposed the EL inference procedure for the mean residual life (MRL) function. They showed that the limiting distribution of the EL ratio for the MRL function could be derived. The authors also demonstrated that the likelihood ratio exhibits weak convergence to a mean-zero Gaussian process. Based on this result, they constructed confidence intervals and bands for the MRL function, which were then compared with the NA method.

Shen (2011) utilized the EL method and a modified EM algorithm to obtain interval estimation for double-truncated data, and demonstrated that the EL approach is more efficient than the bootstrap method based on simulation results. The EL method has also been employed by Zhou and Jeong (2011) to provide confidence intervals for the median and mean residual lifetime. When it comes to estimating the cumulative hazard ratio with covariate adjustment, Dong and Matthews (2012) used the EL method to obtain nonparametric estimates. Additionally, Zhang et al. (2018) introduced a general kth correlation coefficient between the density function and distribution function of a continuous variable, and obtained the residual-based kth correlation coefficient estimator confidence interval using the EL method.

Moreover, several studies have extended the EL method to estimate confidence intervals under length-biased data. For example, Ning et al. (2013) used the EL method to obtain a confidence interval for length-biased data, covering both large and small sample cases. They also obtained a confidence interval for the mean, median, and survival function using the EL ratio. For the inference of MRL with length-biased data, Liang et al. (2016) proposed the EL procedure. However, under right censoring, the EL-based log-likelihood ratio has a scaled chi-squared distribution, and estimating the scale parameter can lead to lower coverage of confidence intervals. To address this issue, they introduced an algorithm that directly calculates the likelihood ratio (LR). They also discussed the convergence of the corresponding log-likelihood ratio to the standard chi-square distribution and showed that the corresponding confidence interval has better coverage probability. Finally, Mohammadian and Habibirad (2024) introduced both EL and adjusted EL methods for RMST under length-biased and right-censoring data. These studies demonstrate the versatility and effectiveness of the EL method in statistical analysis for length-biased data.

Previous investigations on length-biased sampling plans have suffered from overly simplistic strategies for estimating the margins of error in commonly utilized summary

statistics. However, ignoring this sampling plan can lead to fundamental overestimation in survival analyses. Therefore, using the EL method to obtain confidence intervals that are tailored to the RMST for length-biased data can address this gap.

In a recent study, Zhou (2021) proposed a confidence interval for the RMST utilizing the EL method. However, the inference of RMST with length-biased data using the EL method has not been studied yet. Therefore, in our article, we derive the confidence interval for the RMST under length-biased data via the EL method. In many instances, as widely identified in the literature, authors have conducted inferences in length-biased sampling, but constructing confidence bands for the $\mu(t)$ function is not yet available in the current literature. Therefore, we presented a novel approach to construct such confidence bands in our article.

The remainder of this paper is organized as follows: Section 2 provides a description of the data and notation used in this study. In Section 3, we introduce our methodology, which includes the EL method and the NA method, and present the confidence intervals and confidence bands for the RMST. We also conduct a simulation study to compare the efficiency of the proposed EL method to the NA method for length-biased data. Furthermore, two real datasets are presented in Section 4. In the last section, we provide proof of the main results.

2 Data and Notations

Let the positive random variable X have a distribution function F. When considering biased sampling, instead of drawing a random sample from F, the random sample $Y_1, \ldots, Y_n$, with a distribution function G, is observed. Under length-biased sampling, the distribution function G is

$$G(t) = \mu^{-1} \int_0^t y dF(y), \quad t \geq 0, \tag{2.1}$$

where

$$\mu = \int_0^\infty y dF(y), \quad y \geq 0,$$

where $\mu < \infty$. According to (2.1) and $F(0) = 0$, the distribution function F is obtained as follows:

$$F(t) = \mu \int_0^t y^{-1} dG(y), \quad t \geq 0. \tag{2.2}$$

Cox (1969) used this equation to estimate the function $F(\cdot)$. The empirical distribution function as an estimator of $G(\cdot)$, is

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq t), \tag{2.3}$$

where $I(A)$ denotes the indicator of event A. The empirical estimator of $F(\cdot)$ is defined by replacing (2.3) in (2.2), i.e.

$$\begin{aligned} F_n(t) &= \mu_n \int_0^t y^{-1} dG_n(y) \\ &= \frac{\mu_n}{n} \sum_{i=1}^{n} \frac{1}{Y_i} I(Y_i \le t), \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} \mu_n^{-1} &= \int_0^\infty y^{-1} dG_n(y) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{Y_i}. \end{aligned}$$

3 Methodology

In this section, the pointwise confidence interval and confidence band for RMST are obtained through the EL method and the NA method.

3.1 Empirical Likelihood

We introduce the EL method for the RMST with length-biased data and obtain a confidence interval for the RMST. Moreover, we present a confidence band for the RMST.

Under the setting of length-biased sampling, and considering Equations (1.1), we have

$$\mu(t) = \mu \int_0^\infty \Big(I(u \le t) + \frac{t}{u} I(u > t) \Big) dG(u),$$

where $\mu^{-1} = \int_0^\infty \frac{1}{u} dG(u)$. Therefore, we get

$$\mu(t) \int_0^\infty \frac{1}{u} dG(u) = \int_0^\infty \Big(I(u \le t) + \frac{t}{u} I(u > t) \Big) dG(u). \tag{3.1}$$

To compute the EL method for the RMST, under the setting of length-biased sampling, we utilize the following equation:

$$E\left[\frac{\mu(t) - Y I(Y \le t) - t I(Y > t)}{Y} \right] = 0. \tag{3.2}$$

Therefore, the estimation equation at a fixed time $0 < t < \tau$ is suggested by

$$U(\mu(t)) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mu(t) - Y_i I(Y_i \le t) - t I(Y_i > t)}{Y_i} \right] = 0. \tag{3.3}$$

Considering Equation (3.3), the EL ratio for the length-biased data is obtained in Equation (3.5). Suppose that $\mathbf{p} = (p_1, \ldots, p_n)$ is a probability vector for which $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for each $1 \leq i \leq n$. Define

$$D_i(\mu(t)) = \left[\frac{\mu(t) - Y_i I(Y_i \leq t) - tI(Y_i \geq t)}{Y_i} \right], \tag{3.4}$$

for $1 \leq i \leq n$ at a stable time t. Then, the evaluated EL is given by

$$L(\mu(t)) = \sup \Big\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i D_i(\mu(t)) = 0 \Big\}. \tag{3.5}$$

By using the Lagrange multiplier method, we obtain that

$$p_i = \{n(1 + \varrho(t) D_i(\mu(t)))\}^{-1}, \quad i = 1, \ldots, n,$$

where $\varrho(t)$ is obtained from solving of

$$\frac{1}{n} \sum_{i=1}^{n} \frac{D_i(\mu(t))}{1 + \varrho(t) D_i(\mu(t))} = 0. \tag{3.6}$$

Noting $\prod_{i=1}^{n} p_i$ and subjecting to the condition $\sum_{i=1}^{n} p_i = 1$, its maximum at $p_i = n^{-1}$ is n^{-n}. Afterwards, the EL ratio for $\mu(t)$ is defined by

$$Q(\mu(t)) = \prod_{i=1}^{n} (np_i) = \prod_{i=1}^{n} (1 + \varrho(t) D_i(\mu(t)))^{-1}. \tag{3.7}$$

Hence, the empirical log-likelihood ratio is given by

$$\kappa(\mu(t)) = -2 \log Q(\mu(t)) = 2 \sum_{i=1}^{n} \log(1 + \varrho(t) D_i(\mu(t))), \tag{3.8}$$

where $\varrho(t)$ is the solution of (3.6).

Theorem 3.1. *Assume* $E\left[\frac{1}{Y^2}\right] < \infty$. *Then, for all* $t \in [0, \tau)$,

$$\kappa(\mu(t)) \xrightarrow{\mathcal{D}} \chi_1^2.$$

□

We use Theorem 3.1, to obtain the confidence interval for $\mu(t)$ at a fixed time t, for $t \in [0, \tau)$. In this manner, an asymptotic $100(1 - \alpha)\%$ confidence interval for $\mu(t)$ is obtained from the following equation:

$$C_1(t) = \left\{ \mu(t) : \kappa(\mu(t)) \leq \chi_{1,\alpha}^2 \right\}, \tag{3.9}$$

where $\chi_{1,\alpha}^2$ is the upper α-quantile of χ_1^2.

Theorem 3.2. *Assume that* $\kappa(\mu(t)) = -2\log Q(\mu(t))$ *and* $E\left[\frac{1}{Y^2}\right] < \infty$. *Then, there exists a mean zero Gaussian process* $\{\vartheta(t), 0 \le t \le a\}$ *such that*

$$\kappa(\cdot) \xrightarrow{\mathcal{L}} \frac{\vartheta(\cdot)}{\sqrt{\nu(\cdot)}}, \tag{3.10}$$

where $\nu(t) = E(D_i(\mu(t)))^2$ *and* $\kappa(\cdot)$ *is defined in* $D[0,a]$, *the space of cadlag functions on* $[0,a]$, *and* $\xrightarrow{\mathcal{L}}$ *denotes weak convergence. The Gaussian process* $\vartheta(\cdot)$ *is given by*

$$\vartheta(t) = \int_0^\infty h(y,t)dB(G(y)), \tag{3.11}$$

where $B(\cdot)$ *is a Brownian Bridge on the unit interval, with the covariance function*

$$\mathrm{Cov}(\vartheta(t), \vartheta(s)) = \int_{t\vee s}^\infty h(y,t)h(y,s)dG(y), \tag{3.12}$$

where $h(y,t) = \left(\frac{\mu(t) - yI(y\le t) - tI(y>t)}{y}\right)$. □

Utilizing the continuous mapping theorem and Theorem 3.2, we have

$$\sup_{0\le t\le a}\{\kappa(\mu(t))\} \xrightarrow{\mathcal{L}} \sup_{0\le t\le a}\frac{\vartheta(t)}{\sqrt{\nu(t)}}. \tag{3.13}$$

The asymptotic $100(1-\alpha)\%$ confidence band for the RMST function is

$$C_2 = \{\mu(t) : \kappa(\mu(t)) \le q_\alpha, t \in [0,a]\}, \tag{3.14}$$

where q_α is the upper α-quantile of the distribution of

$$\sup_{0\le t\le a}\left\{\vartheta(t)/\sqrt{\nu(t)}\right\}. \tag{3.15}$$

Since it is difficult to evaluate the limiting distribution of (3.15) analytically, the Gaussian process may be approximated via the following method.

According to the assumptions of Theorem 3.2, for the nonrandom integrand $h(y,t)$, we have

$$\int_t^\tau h^2(y,t)dG(y) < \infty, \qquad 0 \le t \le a.$$

Therefore, the stochastic integral of $\vartheta(t)$ remains a Gaussian process. Moreover, for a fixed t, $\vartheta(t)$, as an Itô integral, is approximated by the following equation:

$$\widetilde{\vartheta}(t) = \lim_{n\to\infty}\sum_{i=0}^{n-1} h(k_i^n, t)\left[B(G(k_{i+1}^n)) - B(G(k_i^n))\right], \tag{3.16}$$

for every n, where $\{k_i^n, i = 0, \ldots, n\}$, is a partition of $[0, t]$. Also, the lim is in quadratic mean, taken over all partitions with

$$\sigma = \max_{1 \le i \le n}(t_{i+1}^n - t_i^n) \longrightarrow 0,$$

as $n \longrightarrow \infty$.

Equations (3.15) and (3.16) have unknown parameters $\nu(t)$ and $\mu(t)$, which should be replaced with their corresponding estimates. $\mu(t)$ and $\nu(t)$ may be estimated by their consistent estimators, i.e., $\mu_n(t)$ and

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mu_n(t) - Y_i I(Y_i \le t) - tI(Y_i > t)}{Y_i}\right)^2,$$

respectively.

Now, we could obtain, for example, N sample paths of $\widetilde{\vartheta}(t)$, denoted by $\{\widetilde{\vartheta}_k(t) : t \in [0, \tau]\}_{k=1}^N$ and calculate

$$q_k = \max_{1 \le j \le n} \frac{\widetilde{\vartheta}_k(t_j)}{\sqrt{\nu_n(t_j)}},$$

for a partition of the interval $[0, \tau]$. N should be large enough to stabilize the estimation. If you need a higher confidence level, you should make it even larger. This is because it becomes increasingly difficult to estimate the more extreme percentiles. Ultimately, q_α can be approximated by the empirical percentile of $\{q_1, \ldots, q_N\}$.

3.2 Normal Approximation

This subsection presents the limiting distribution of $\mu_n(t)$. We can construct NA-based confidence interval and confidence band for the RMST function. To construct an NA-based confidence interval, we present Theorem 3.3.

Theorem 3.3. *Assume* $E\left[\frac{1}{Y^2}\right] < \infty$. *Then, for all* $t \in [0, \tau)$,

$$\sqrt{n}\Big(\mu_n(t) - \mu(t)\Big) \xrightarrow{\mathcal{D}} N(0, \delta^2(t)), \tag{3.17}$$

where

$$\delta^2(t) = S^2(t)\mu\Big[E(Y^{-1}I(Y \le t)) + F^2(t)E(Y^{-1}) - 2F(t)E(Y^{-1}I(Y \le t))\Big].$$

□

Regarding $\delta^2(t)$ having a complex shape and being difficult to estimate, we introduce consistent and bootstrap estimators. Assume that $\delta_n^2(t)$ is a consistent estimator of $\delta^2(t)$, so the asymptotic confidence interval for the RMST function is as follows:

$$C_3(t) = \{\mu(t) : |\sqrt{n}(\mu_n(t) - \mu(t))| \le \delta_n^2(t) z_{\alpha/2}\}, \tag{3.18}$$

where $z_{\alpha/2}$ is the upper $\frac{\alpha}{2}$-quantile of the normal distribution.

We utilize the bootstrap method to construct an estimator for $\delta^2(t)$. Let $Y_1^*, \ldots, Y_n^*$ be iid random variables with distribution function $G_n(\cdot)$, where $Y_1, \ldots, Y_n$ are fixed. Additionally, suppose $G_n^*(\cdot)$ is the empirical distribution function of the random sample $Y_1^*, \ldots, Y_n^*$. Therefore, we define

$$\mu_n^* = \int_0^t (1 - F_n^*(u))du,$$

where

$$F_n^*(u) = \mu_n^* \int_0^u y^{-1} dG_n^*(y),$$

and

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

Defining $\theta^*(\cdot) = \sqrt{n}(\mu_n^*(\cdot) - \mu_n(\cdot))$, the above resampling procedure should be replicated $\mathcal{B}$ times, then the bootstrap estimate for $\delta^2(t)$ would be the sample variance of $\theta_1^*(t), \ldots, \theta_{\mathcal{B}}^*(t)$, that is

$$\widehat{Var}_{\mathcal{B}}(\theta^*(t)) = \frac{1}{\mathcal{B}-1} \sum_{i=1}^{\mathcal{B}} \left(\theta_i^*(t) - \frac{1}{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \theta_j^*(t) \right)^2.$$

To construct the confidence band for RMST, we present Theorem 3.4.

Theorem 3.4. *Assume $E\left[\frac{1}{Y^2}\right] < \infty$ and $B(\cdot)$ is a Brownian Bridge on the unit interval. Then, there is a mean zero Gaussian process $\mathcal{A}(\cdot)$, such that*

$$\sqrt{n}\Big(\mu_n(\cdot) - \mu(\cdot)\Big) \xrightarrow{\mathcal{L}} \mathcal{A}(\cdot), \tag{3.19}$$

where

$$\mathcal{A}(t) = \int_0^\infty \mu\Big(I(u \le t) + \frac{t}{u} I(u > t)\Big) dB(G(u)),$$

and

$$\mathrm{Cov}(\mathcal{A}(t), \mathcal{A}(s)) = \int_{t \vee s}^\infty \Big(I(u \le t) + \frac{t}{u} I(u > t)\Big)\Big(I(u \le s) + \frac{s}{u} I(u > s)\Big) dG(u).$$

□

Utilizing Theorem 3.4 and the continuous mapping theorem, we have

$$\sup_{0 \le t \le a} \sqrt{n}\Big(\mu_n(t) - \mu(t)\Big) \xrightarrow{\mathcal{L}} \sup_{0 \le t \le a} \mathcal{A}(t). \tag{3.20}$$

The asymptotic $100(1-\alpha)\%$ confidence band for the RMST function is

$$C_4 = \left\{ \mu(t) : \sqrt{n}|\mu_n(t) - \mu(t)| \le Q_\alpha, t \in [0, a] \right\}. \tag{3.21}$$

where Q_α is the upper α-quantile of the distribution of $\sup_{0 \le t \le a} \mathcal{A}(t)$. According to the proposed method to obtain q_α, we can obtain Q_α.

4 Simulation Study and Real Example

The present section examines and illustrates the performance of both the confidence interval and band via the EL and NA methods under various sample sizes of length-biased data. Two criteria, namely the average length of the confidence interval (Δ) and coverage probability (CP), are considered.

The performance of the proposed EL-based confidence intervals, which are presented in $C_1(t)$, has been compared with the proposed NA-based confidence intervals $C_3(t)$ in terms of CP and Δ. Having presented these results, the EL-based and NA-based confidence bands, which are defined in $C_2(t)$ and $C_4(t)$, are illustrated.

In this simulation, the data are simulated from the gamma and uniform distributions, and we illustrate our method using two sets of real data: the waiting times (in minutes) of 100 bank customers before service and the monthly actual tax.

In this regard, the numerical results are obtained based on two significance levels $\alpha = 0.10$ and $\alpha = 0.05$, three sample sizes $n = 50$, $n = 100$, and $n = 150$, and several statistical distributions including gamma and uniform distributions. Furthermore, the number of iterations used to calculate the average value has been set to 5000. This number guarantees the precision of all results to at least three decimal places. Iterations need to be large enough to stabilize the estimate, and they need to be even larger if a higher confidence level is desired because estimating extreme percentiles is more challenging. We considered the performance of the EL and NA pointwise confidence intervals and confidence bands under the condition $E[Y^{-r}]$ for some $r > 2$, and we generated data from length-biased distributions where this condition is satisfied.

The assumption $E\left[\frac{1}{Y^2}\right] < \infty$ is crucial for the validity of our theoretical results. To ensure this condition is met in our simulation studies, we choose distributions for Y such that the second inverse moment is finite. For example, consider the PDF of a gamma distribution X with a shape $\alpha = 4$ and rate $\beta = 1$. For a random variable X, the corresponding length-biased distribution Y has the PDF:

$$f_Y(y) = \frac{y \cdot f_T(y)}{E[T]} = \frac{y \cdot \frac{y^3 e^{-y}}{6}}{E[T]}.$$

Given that $X \sim \text{Gamma}(4, 1)$, the mean $E[T] = \frac{\alpha}{\beta} = 4$. Substituting this into the expression for $f_Y(y)$:

$$f_Y(y) = \frac{y \cdot \frac{y^3 e^{-y}}{6}}{4} = \frac{y^4 e^{-y}}{24}, \quad y > 0.$$

Thus, the length-biased distribution Y follows a Gamma distribution with shape parameter 5 and rate parameter 1. Now, let's verify that the second inverse moment of Y is finite:

$$E\left[\frac{1}{Y^2}\right] = \int_0^\infty \frac{1}{y^2} \cdot f_Y(y)\, dy = \int_0^\infty \frac{1}{y^2} \cdot \frac{y^4 e^{-y}}{24}\, dy = \frac{1}{24} \int_0^\infty y^2 e^{-y}\, dy.$$

The integral $\int_0^\infty y^2 e^{-y}\, dy$ is the expectation of Y^2 for a gamma distribution with shape parameter 3 and rate 1. This integral is finite, and its value is given by:

$$\int_0^\infty y^2 e^{-y}\, dy = \Gamma(3) = 2!.$$

Therefore:

$$E\left[\frac{1}{Y^2}\right] = \frac{\Gamma(3)}{24} = \frac{2}{24} = \frac{1}{12} < \infty.$$

To calculate the variance of the NA-based confidence intervals, we employed a bootstrap estimator. Specifically, we required $B = 500$ bootstrap samples to obtain valid estimates. However, for small sample sizes, it was necessary to generate data significantly more than 500 times to achieve stable results. This issue arises because, as time t increases, the variance of the NA-based estimator grows substantially, causing the length of the confidence interval to expand dramatically. This expansion leads to lower coverage probabilities, particularly at larger values of t, which compromises the reliability of the NA method. Consequently, the NA method becomes not only less accurate in terms of coverage probability but also more computationally burdensome, especially as time increases. This is because the variance inflates, demanding a greater number of iterations to stabilize the estimates. In contrast, the EL-based confidence intervals consistently provide accurate results across all time points, except for a few close-to-the-boundary observations. While the EL method involves more complex computations per iteration due to the need for Lagrange multipliers and the maximization of the empirical likelihood, its overall computational performance remains more efficient. This is because the EL method does not suffer from the same large variance issues at higher time points. Additionally, constructing the NA-based confidence intervals requires more time and computational resources than the EL method when accounting for the bootstrapping process. Thus, the EL method is less affected by time-induced variance inflation, making it the more computationally stable and reliable approach in practice.

4.1 Simulation 1

In this subsection, we present the analysis of length-biased data.

In Table 1, for the RMST function, we obtain CP and Δ of confidence intervals from the EL ($C_1(t)$) and NA ($C_3(t)$) methods, based on observations sampled from the Gamma$(5, 1)$ distribution and the unbiased case of the Gamma$(4, 1)$ distribution.

In Table 2, we consider the uniform distribution to obtain the confidence interval. For the RMST function, we obtain CP and Δ of confidence intervals from the EL ($C_1(t)$) and NA ($C_3(t)$) methods, based on observations sampled from the uniform$(2, 5)$ distribution. Since the samples are in the interval $(2, 5)$ and perform poorly at the beginning and end of the interval, the study interval is $(2.2, 4.8)$.

Table 1: 90% and 95% coverage probabilities and average lengths of confidence intervals for RMST of Gamma

		$1-\alpha = 0.90$				$1-\alpha = 0.95$			
Time	n	c.p.EL	Δ.EL	c.p.NA	Δ.NA	c.p.EL	Δ.EL	c.p.NA	Δ.NA
	50	0.737	0.143	0.643	0.160	0.790	0.172	0.680	0.184
2	100	0.778	0.122	0.732	0.129	0.827	0.145	0.788	0.151
	150	0.809	0.107	0.769	0.113	0.874	0.122	0.830	0.131
	50	0.764	0.251	0.733	0.257	0.851	0.301	0.805	0.304
2.5	100	0.829	0.203	0.799	0.208	0.883	0.241	0.844	0.244
	150	0.844	0.174	0.824	0.179	0.904	0.188	0.870	0.208
	50	0.813	0.370	0.795	0.371	0.873	0.437	0.845	0.445
3	100	0.841	0.285	0.835	0.292	0.905	0.339	0.874	0.351
	150	0.860	0.241	0.848	0.247	0.923	0.260	0.899	0.287
	50	0.832	0.479	0.811	0.481	0.902	0.566	0.865	0.572
3.5	100	0.857	0.361	0.838	0.370	0.921	0.435	0.902	0.438
	150	0.871	0.305	0.864	0.312	0.935	0.330	0.912	0.364
	50	0.840	0.581	0.832	0.587	0.917	0.690	0.893	0.692
4	100	0.864	0.435	0.857	0.443	0.933	0.519	0.914	0.526
	150	0.875	0.364	0.862	0.371	0.935	0.390	0.927	0.434
	50	0.849	0.671	0.841	0.670	0.906	0.790	0.898	0.801
4.5	100	0.873	0.499	0.866	0.504	0.937	0.588	0.918	0.599
	150	0.880	0.415	0.876	0.421	0.939	0.441	0.925	0.493
	50	0.863	0.745	0.851	0.745	0.917	0.876	0.904	0.883
5	100	0.871	0.552	0.868	0.557	0.932	0.657	0.932	0.661
	150	0.880	0.456	0.871	0.463	0.935	0.531	0.939	0.546

In the following, we describe Tables 1 and 2:

- For both distributions and a fixed significance level, the CP increases and Δ decreases for every strategy.

- For both distributions and a fixed significance level, the CP under the empirical method is significantly better than the normal one. At time 2.2 in Table 2, which is near the beginning of the interval, the coverage probability is slightly higher.

- For both distributions and a fixed significance level, Δ in both methods is almost the same. It is also worth mentioning that Δ decreases and CP increases with increasing sample size.

In Figure 1, we present the confidence intervals for the RMST function, obtained using both the EL method ($C_1(t)$) and the NA method ($C_3(t)$). These intervals are based on observations sampled from a Gamma$(5, 1)$ distribution, with the reference distribution being the unbiased Gamma$(4, 1)$. Notably, the confidence intervals generated by the EL method are slightly narrower than those produced by the NA method. This indicates that the EL method may provide more precise interval estimates for the RMST in this context, offering tighter bounds around the true survival function. Such a difference in interval width suggests that the EL approach is more efficient in capturing the uncertainty of the RMST function than the NA method, especially in scenarios where non-parametric techniques are preferred for improved accuracy.

Table 2: 90% and 95% coverage probabilities and average lengths of confidence intervals for RMST of Uniform

		$1-\alpha = 0.95$				$1-\alpha = 0.9$			
Time	n	c.p.EL	Δ.EL	c.p.NA	Δ.NA	c.p.EL	Δ.EL	c.p.NA	Δ.NA
	50	0.971	0.017	0.746	0.023	0.938	0.014	0.727	0.019
2.2	100	0.976	0.013	0.859	0.016	0.938	0.011	0.823	0.014
	150	0.971	0.011	0.887	0.013	0.944	0.011	0.889	0.012
	50	0.922	0.050	0.851	0.053	0.871	0.042	0.812	0.045
2.4	100	0.930	0.037	0.895	0.038	0.894	0.032	0.891	0.037
	150	0.943	0.031	0.917	0.031	0.899	0.026	0.891	0.031
	50	0.920	0.091	0.896	0.093	0.880	0.077	0.844	0.078
2.6	100	0.941	0.067	0.920	0.067	0.892	0.056	0.873	0.056
	150	0.945	0.055	0.930	0.055	0.900	0.047	0.892	0.055
	50	0.932	0.136	0.902	0.136	0.880	0.114	0.853	0.115
2.8	100	0.941	0.099	0.923	0.099	0.889	0.083	0.872	0.083
	150	0.946	0.082	0.933	0.091	0.902	0.068	0.893	0.082
	50	0.935	0.183	0.915	0.187	0.891	0.154	0.853	0.154
3	100	0.946	0.132	0.931	0.141	0.902	0.111	0.885	0.111
	150	0.947	0.109	0.935	0.119	0.911	0.091	0.894	0.109
	50	0.943	0.229	0.918	0.230	0.886	0.184	0.865	0.192
3.2	100	0.951	0.166	0.934	0.166	0.901	0.115	0.876	0.139
	150	0.947	0.136	0.939	0.136	0.897	0.114	0.894	0.136
	50	0.933	0.274	0.921	0.275	0.887	0.231	0.872	0.232
3.4	100	0.951	0.198	0.932	0.211	0.898	0.166	0.880	0.166
	150	0.948	0.163	0.941	0.163	0.899	0.137	0.893	0.163
	50	0.946	0.318	0.922	0.321	0.891	0.268	0.878	0.269
3.6	100	0.946	0.229	0.933	0.230	0.892	0.192	0.883	0.193
	150	0.954	0.188	0.942	0.188	0.903	0.158	0.894	0.188
	50	0.943	0.359	0.931	0.361	0.885	0.303	0.880	0.304
3.8	100	0.946	0.258	0.940	0.259	0.898	0.217	0.895	0.218
	150	0.948	0.212	0.945	0.212	0.898	0.178	0.893	0.212
	50	0.944	0.397	0.930	0.400	0.896	0.334	0.882	0.336
4	100	0.950	0.285	0.941	0.286	0.897	0.240	0.891	0.240
	150	0.954	0.234	0.943	0.234	0.899	0.196	0.894	0.234
	50	0.948	0.430	0.936	0.433	0.890	0.362	0.881	0.363
4.2	100	0.951	0.308	0.940	0.309	0.896	0.258	0.893	0.259
	150	0.953	0.253	0.950	0.253	0.899	0.212	0.894	0.253
	50	0.948	0.458	0.937	0.461	0.896	0.386	0.888	0.387
4.4	100	0.948	0.327	0.947	0.329	0.902	0.275	0.892	0.276
	150	0.952	0.269	0.950	0.269	0.892	0.226	0.951	0.269
	50	0.952	0.480	0.932	0.482	0.892	0.403	0.884	0.404
4.6	100	0.955	0.343	0.943	0.352	0.895	0.288	0.891	0.289
	150	0.947	0.281	0.940	0.281	0.896	0.236	0.913	0.278
	50	0.943	0.493	0.934	0.495	0.890	0.415	0.895	0.416
4.8	100	0.948	0.352	0.941	0.354	0.897	0.296	0.895	0.297
	150	0.952	0.289	0.945	0.311	0.903	0.243	0.898	0.321

In Table 3, the CP and Δ of the confidence bands for the RMST function are presented. These results are derived using the EL ($C_2(t)$) and NA ($C_4(t)$) methods, based on observations sampled from the Gamma(5, 1) distribution and the unbiased scenario of the Gamma(4, 1) distribution.

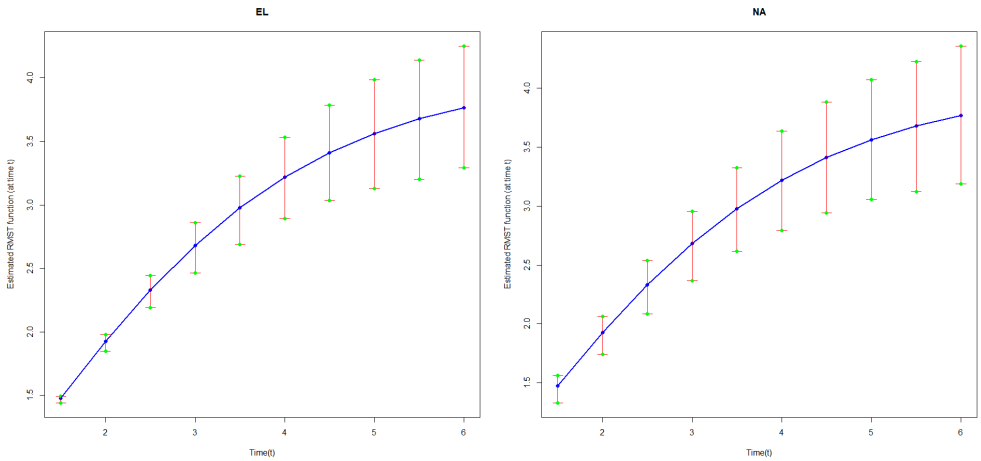

Figure 1: Confidence interval for RMST of gamma.

Table 3: 90% and 95% coverage probabilities and average lengths of confidence bands for RMST of Gamma

		$1-\alpha = 0.95$				$1-\alpha = 0.90$			
Time	n	c.p.EL	Δ.EL	c.p.NA	Δ.NA	c.p.EL	Δ.EL	c.p.NA	Δ.NA
	50	0.829	0.183	0.710	0.201	0.809	0.178	0.689	0.189
2	100	0.855	0.161	0.799	0.172	0.829	0.150	0.785	0.154
	150	0.903	0.138	0.838	0.151	0.878	0.129	0.822	0.139
	50	0.882	0.322	0.798	0.343	0.837	0.299	0.783	0.309
2.5	100	0.913	0.262	0.863	0.282	0.882	0.243	0.849	0.254
	150	0.942	0.226	0.886	0.240	0.916	0.210	0.870	0.218
	50	0.904	0.477	0.851	0.490	0.882	0.441	0.852	0.443
3	100	0.940	0.366	0.891	0.391	0.915	0.340	0.874	0.349
	150	0.945	0.311	0.920	0.331	0.916	0.289	0.902	0.300
	50	0.932	0.620	0.889	0.637	0.905	0.576	0.871	0.578
3.5	100	0.940	0.465	0.909	0.492	0.922	0.432	0.894	0.451
	150	0.947	0.392	0.930	0.419	0.929	0.364	0.918	0.380
	50	0.938	0.752	0.914	0.774	0.916	0.699	0.897	0.701
4.0	100	0.947	0.561	0.930	0.592	0.934	0.521	0.909	0.534
	150	0.963	0.466	0.927	0.499	0.945	0.433	0.915	0.449
	50	0.943	0.862	0.914	0.886	0.916	0.801	0.896	0.805
4.5	100	0.956	0.642	0.931	0.680	0.928	0.597	0.908	0.616
	150	0.965	0.531	0.939	0.571	0.934	0.493	0.919	0.509
	50	0.941	0.956	0.927	0.987	0.928	0.888	0.910	0.904
5	100	0.956	0.708	0.932	0.749	0.937	0.658	0.916	0.673
	150	0.964	0.585	0.925	0.945	0.934	0.544	0.923	0.563

In the following, we describe Table (3):

- For a given significance level, the Coverage Probability (CP) increases, while Δ decreases across all strategies.
- At a fixed significance level, the CP obtained using the empirical method is significantly higher than that of the normal method.
- For a fixed significance level, Δ remains nearly the same in both methods. Addi-

tionally, it is noteworthy that as the sample size increases, Δ decreases, and CP increases.

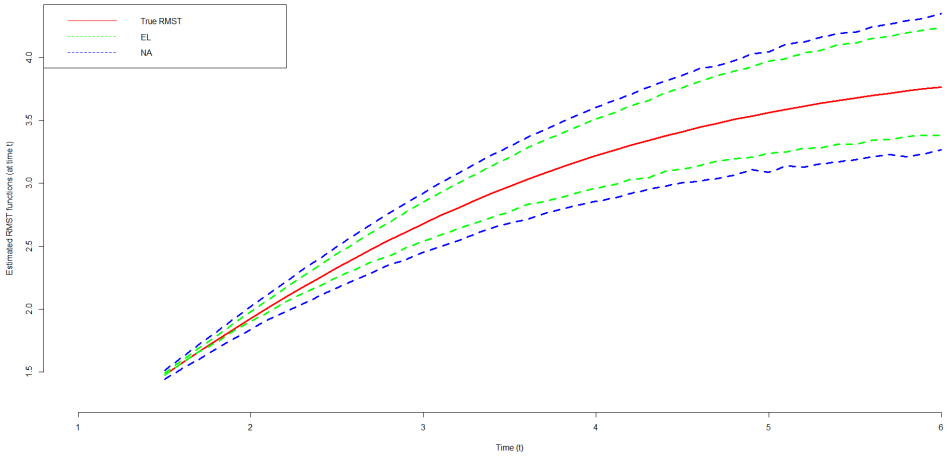

Figure 2: confidence band for RMST of gamma.

The observations depicted in Figure 2 provide valuable insights into the performance of the EL-based confidence band (C_2) and the NA-based confidence band (C_4) for the RMST of a gamma distribution in the context of length-biased sampling. Firstly, the figure highlights that C_2 exhibits a more favorable behaviour compared to its NA-confidence band counterpart. It demonstrates a better coverage of the true RMST values, as indicated by the band's proximity to the true values of interest. This suggests that the EL-based method provides more accurate and reliable estimates for the RMST under length-biased sampling. Furthermore, the narrower width of C_2 compared to C_4 implies a higher precision in the estimation of the RMST. This narrower band indicates that the EL-based approach achieves a smaller margin of error, resulting in more precise and efficient inference. Overall, these observations provide strong support for the EL-confidence band as a highly attractive alternative to the NA-confidence band in the analysis of RMST for length-biased data. The superior performance of the EL-based method in terms of coverage accuracy and precision inspires confidence in its application, suggesting that it can be a valuable tool for researchers and practitioners working with length-biased sampling in survival analysis.

We choose a distribution that violates the assumptions. Use a heavy-tailed distribution like the Pareto distribution, which is known to violate the assumption when the shape parameter $\alpha \leq 2$. The Pareto distribution is defined as $f_Y(y) = \frac{\alpha y_m^\alpha}{y^{\alpha+1}}$ for $y \geq y_m > 0$, where y_m is the scale parameter and α is the shape parameter. When $\alpha \leq 2$, $E\left[\frac{1}{Y^2}\right] = \infty$, which violates the assumption. Generate length-biased data by sampling from the Pareto distribution and then applying the length-biased transformation. For example, if T follows a Pareto distribution, the length-biased distribution for Y can be generated

by setting $Y = T \times W$, where W is an appropriate weight factor. Apply the empirical likelihood method to the generated data and compare the performance metrics (e.g., bias, variance, and coverage probability) with those obtained under the assumption $E\left[\frac{1}{Y^2}\right] < \infty$.

As expected, we observed an increase in the bias and variance of the parameter estimates when the assumption was not satisfied. The empirical likelihood confidence intervals may not maintain the nominal coverage probability. The empirical likelihood ratio may become unstable and unreliable, leading to questionable statistical inferences. These findings underscore the importance of the assumption $E\left[\frac{1}{Y^2}\right] < \infty$ for the reliability of the empirical likelihood method.

4.1.1 Real Data Analysis

In this section, we first study the waiting times (in minutes) of 100 bank customers before service. For this data, we obtain the confidence interval and confidence band via the EL method and the NA approach. The real dataset was presented by Ghitany et al. (2008), and they showed that the Lindley distribution can be a better model than one based on the exponential distribution. Rather and Subramanian (2018) analyzed the LB Sushila distribution and the Sushila distribution for this real data, and the models were compared with each other.

The bank dataset consists of customer records, where longer-tenured customers tend to be included in the dataset due to the nature of customer retention and data recording practices. This implies a length-biased sample, as customers with longer relationships with the bank have a higher probability of being sampled. To verify that the assumption $E\left[\frac{1}{Y^2}\right] < \infty$ holds, we analyzed the distribution of the tenure of customers. The distribution was assessed to ensure it aligns with a theoretical distribution that meets the assumption. For example, if the tenure distribution aligns with an exponential or Weibull distribution, it is straightforward to confirm that $E\left[\frac{1}{Y^2}\right]$ is finite.

Table 4: 95% confidence interval for RMST function for Bank customers

		T=2.52	T=5.56	T=10.12	T=13.16	T=16.20	T=23.80	T=26.84	T=32.92
	Lower	1.88	3.08	3.86	4.10	4.22	4.35	4.38	4.38
EL	Upper	2.32	4.15	5.40	5.80	6.00	6.22	6.27	6.29
	Lower	1.92	3.12	3.87	4.08	4.21	4.29	4.31	4.33
NA	Upper	2.35	4.19	5.43	5.85	6.00	6.25	6.27	6.34

Table 4 presents the 95% confidence interval via the EL method and the NA method for the bank customers' data, and T represents the waiting times. The NA-based confidence intervals for the RMST at selected times are slightly wider than the EL-based confidence intervals, and with increasing time, the differences between these intervals increase in both methods.

The second real dataset was used by Nassar and Nada (2011), which contains the

monthly actual tax revenue in Egypt from January 2006 to November 2010 (in 1000 million Egyptian pounds). Hassan and Assar (2017) presented a model referred to as the exponentiated Weibull power function distribution, based on the exponentiated Weibull-G family of distributions. They used the monthly actual tax revenue, and the practical importance of the proposed distribution was examined through these data.

The Tax dataset includes monthly records of businesses, where businesses with longer operational periods are more likely to have accumulated more tax records, increasing their likelihood of being sampled. This naturally results in a length-biased sample, as older businesses are overrepresented. We analyzed the operational duration distribution of businesses in the Tax dataset to ensure it meets the assumption $E\left[\frac{1}{Y^2}\right] < \infty$. Similar to the Bank data, if the distribution resembles a log-normal or Weibull distribution, the finiteness of $E\left[\frac{1}{Y^2}\right]$ can be confirmed.

Table 5: 95% confidence interval for RMST function for Monthly taxes

		T=6.36	T=10.10	T=11.97	T=15.71	T=19.44	T=23.18	T=28.78	T=36.26
	Lower	5.98	7.83	8.25	8.74	9.00	9.18	9.23	9.32
EL	Upper	6.33	8.81	9.40	10.29	10.79	10.98	11.20	11.43
	Lower	5.87	7.64	8.05	8.57	8.85	8.97	9.06	9.15
NA	Upper	6.29	8.85	9.53	10.47	11.02	11.27	11.46	11.68

In Table Taxes, the 95% confidence interval via the EL method and the NA method for the monthly taxes data are presented and T is the amount of tax. The length of the EL method is shorter than the NA approach and with increasing the amount of tax, these distances increase in both methods.

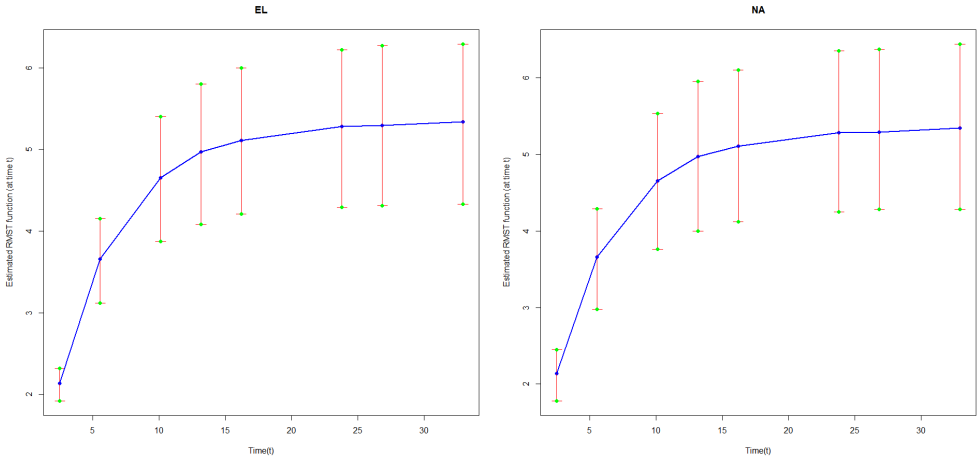

Figure 3: 95% confidence interval for RMST function for the bank data.

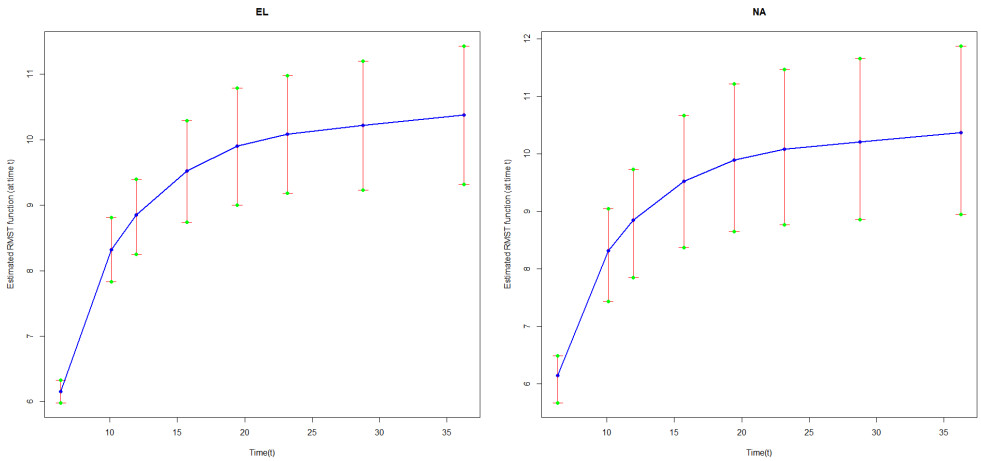

Figure 4: 95% confidence interval for RMST function for the monthly taxes.

In Figures 3 and 4, we demonstrate the 95% confidence intervals for the bank data and the monthly tax data via the EL and NA methods. According to these diagrams, it is estimated that the RMST of both datasets is increasing. Both datasets have an admissible interval, and the NA-based confidence intervals are wider than the EL-based confidence intervals.

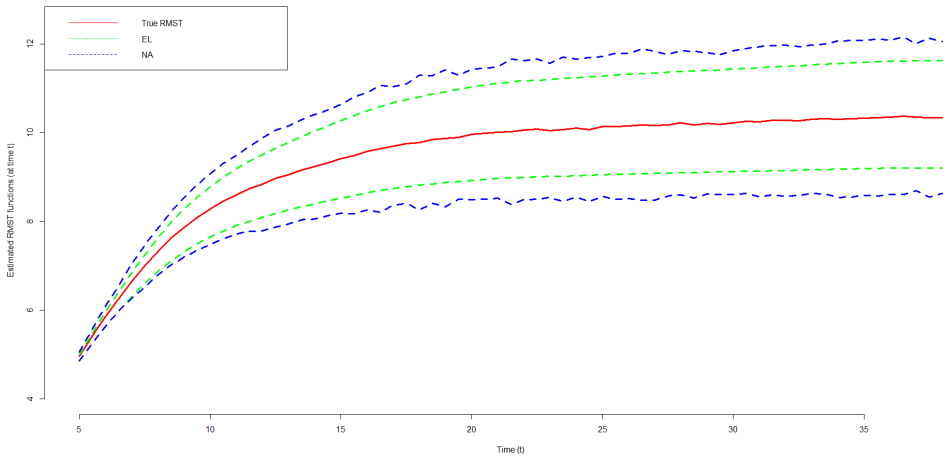

Figure 5: 95% confidence band for RMST function for the monthly taxes.

In Figures 5 and 6, we demonstrate the 95% confidence bands for the monthly tax data and the bank data via the EL and NA methods. According to these diagrams, it is estimated that the RMST of both datasets is increasing. Both datasets have an admissible band, and the NA-based confidence bands are wider than the EL-based confidence bands.

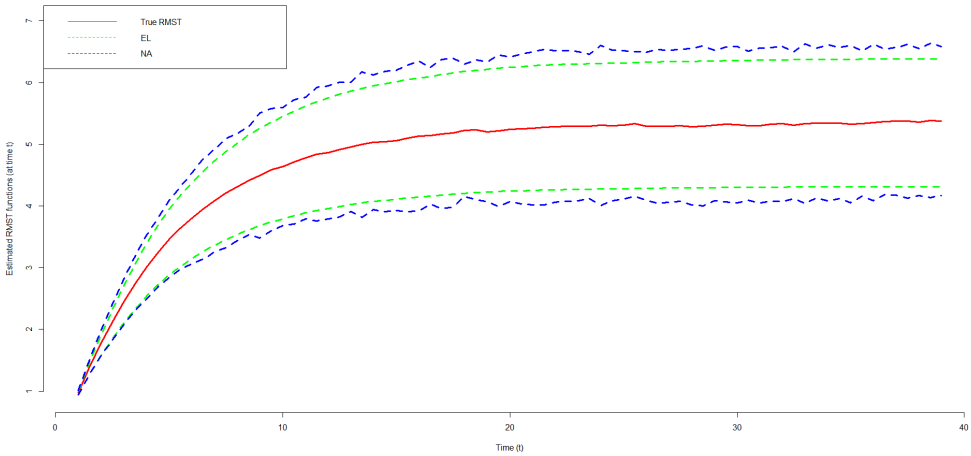

Figure 6: 95% confidence band for RMST function for the bank data.

Acknowledgments

The authors would like to thank the Editor-in-Chief, Dr. Alireza Nematollahi, and the anonymous reviewers for their valuable comments and constructive suggestions, which have significantly improved this article.

5 Proofs

This section contains the proof of Theorems.

Lemma 5.1. *Under the same condition as in Theorem 3.1, for every* $t \in [0, \tau)$*, we get*

$$E[D_i^2(\mu(t))] < \infty, \tag{5.1}$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i(\mu(t)) \xrightarrow{\mathcal{D}} N(0, \nu(t)), \tag{5.2}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} D_i^2(\mu(t)) \xrightarrow{\mathcal{P}} \nu(t), \tag{5.3}$$

where $\xrightarrow{\mathcal{P}}$ *denotes convergence in probability,* $D_i(\mu(t))$ *is given in* (3.4)*, and*

$$\nu(t) = E\left(\frac{\mu(t) - Y_i I(Y_i \le t) - tI(Y_i > t)}{Y_i}\right)^2 < \infty.$$

Proof. Let $t \in [0, \tau)$ be a fixed point. Under the assumption of the lemma, we have

$$\begin{aligned} E[D_i(\mu(t))]^2 &= E\left[\frac{\mu(t) - Y_i I(Y_i \le t) - tI(Y_i > t)}{Y_i}\right]^2 \\ &\le \left((\mu(t))^2 E(\frac{1}{Y_i^2}) + E(I(Y_i \le t))^2 + t^2 E(\frac{I(Y_i > t)}{Y_i})^2\right) \\ &\le \left((\mu(t))^2 E(\frac{1}{Y_i^2}) + 1 + t^2 E(\frac{1}{Y_i^2})\right) \\ &< \infty. \end{aligned}$$

According to the central limit theorem for i.i.d. random variables $D_1(\mu(t)), D_2(\mu(t)), \ldots, D_n(\mu(t))$ and $E\left[\frac{\mu(t) - Y_i I(Y_i \le t) - tI(Y_i > t)}{Y_i}\right] = 0$, we get

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i(\mu(t)) \xrightarrow{\mathcal{D}} N(0, \nu(t)).$$

Therefore (5.2) is obtained. Using the law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^{n} D_i^2(\mu(t)) \xrightarrow{\mathcal{P}} E\left[D_1^2(\mu(t))\right].$$

Thus the second part of the lemma is also proved. □

Proof of Theorem 3.1. Let $t \in [0, \tau)$. According to Lemma 5.1, since $E(D_1^2(\mu(t))) < \infty$, by using Lemma 3 of Owen et al. (1990), we have

$$\max_{1 \le i \le n} |D_i(\mu(t))| = o_p(n^{1/2}), \tag{5.4}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \left|D_i(\mu(t))\right|^3 = o_p(n^{1/2}). \tag{5.5}$$

Therefore, considering (5.4) and (5.5) and applying the same argumentations used in Owen (1991), we have

$$\left|\varrho(t)\right| = O_p(n^{-1/2}). \tag{5.6}$$

Using the Taylor expansion for (3.8), it is obvious that

$$\begin{aligned} \kappa(\mu(t)) &= 2 \sum_{i=1}^{n} \log(1 + \varrho(t) D_i(\mu(t))) \\ &= 2 \sum_{i=1}^{n} \left(\varrho(t) D_i(\mu(t)) - \frac{(\varrho(t) D_i(\mu(t)))^2}{2}\right) + R_n(t), \end{aligned} \tag{5.7}$$

by using Equations (5.5) and (5.6), it can be seen that

$$\begin{aligned} |R_n(t)| &\leq C\sum_{i=1}^{n}|\varrho(t)D_i(\mu(t))|^3 \\ &\leq C|\varrho(t)|^3\sum_{i=1}^{n}|D_i(\mu(t))|^3 \\ &= o_p(1). \end{aligned} \tag{5.8}$$

Considering (5.7) and (3.6), the following result is obtained:

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\frac{D_i(\mu(t))}{1+\varrho(t)D_i(\mu(t))} \\ &= \sum_{i=1}^{n}D_i(\mu(t))\Big[1-\varrho(t)D_i(\mu(t))+\frac{(\varrho(t)D_i(\mu(t)))^2}{1+\varrho(t)D_i(\mu(t))}\Big] \\ &= \sum_{i=1}^{n}D_i(\mu(t))-\Big(\sum_{i=1}^{n}D_i^2(t)\Big)\varrho(t)+\sum_{i=1}^{n}\frac{D_i(\mu(t))(\varrho(t)D_i(\mu(t)))^2}{1+\varrho(t)D_i(\mu(t))}. \end{aligned} \tag{5.9}$$

Considering (5.4) and (5.6) as well as the use of Lemma 5.1, we conclude from the Equation (5.9) that

$$\varrho(t)=\Big(\sum_{i=1}^{n}D_i(\mu(t))^2\Big)^{-1}\sum_{i=1}^{n}D_i(\mu(t))+o_p(1). \tag{5.10}$$

Now by remembering (3.6), we obtain

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\frac{\varrho(t)D_i(\mu(t))}{1+\varrho(t)D_i(\mu(t))} \\ &= \sum_{i=1}^{n}(\varrho(t)D_i(\mu(t)))-\sum_{i=1}^{n}(\varrho(t)D_i(\mu(t)))^2+\sum_{i=1}^{n}\frac{(\varrho(t)D_i(\mu(t)))^3}{1+\varrho(t)D_i(\mu(t))}. \end{aligned} \tag{5.11}$$

Furthermore, having (5.4) and (5.6), we get

$$\sum_{i=1}^{n}\frac{(\varrho(t)V_i(t))^3}{1+\varrho(t)V_i(t)}=o_p(n^{-1/2}). \tag{5.12}$$

Therefore, it can be concluded from (5.11) and (5.12) that

$$\sum_{i=1}^{n}(\varrho(t)D_i(\mu(t)))^2=\sum_{i=1}^{n}\varrho(t)D_i(\mu(t))+o_p(1).$$

Eventually, it follows from the equations (5.7) and (5.10) and Lemma 5.1 that

$$\begin{aligned}\kappa(\mu(t)) &= \sum_{i=1}^{n}(\varrho(t)D_i(\mu(t))) + o_p(1)\\ &= \frac{\left(\sum_{i=1}^{n} D_i(\mu(t))\right)^2}{\sum_{i=1}^{n} D_i^2(\mu(t))} + o_p(1)\\ &\xrightarrow{\mathcal{D}} \chi^2_{(1)}.\end{aligned}$$

□

For proving Theorem 3.2, two lemmas are presented.

Lemma 5.2. *Let that* $E[\frac{1}{Y^2}] < \infty$, *Then*

$$\frac{1}{n}\sum_{i=1}^{n} D_i^2(\cdot) \xrightarrow{\mathcal{P}} \nu(\cdot), \tag{5.13}$$

uniformly over $t \in [0, a]$, *where* $a < \tau$.

Remark 1 is required to prove Lemma 5.2.

Remark 1. (Theorem 3 of Van der Vaart and Wellner (2000)) Let $G(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}$ be a continuous function. Moreover, suppose that $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are P-Glivenko–Cantelli classes of functions. Then the class of functions $\mathcal{S} = G(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n)$ is also P-Glivenko–Cantelli, and it has an integrable envelope function. The class $\mathcal{S}$ is the collection of all functions $S(x)$, which are the form of $S(x) = G(F_1(x), F_2(x), \ldots, F_n(x))$, where F_i is in $\mathcal{F}_i$.

Proof of Lemma 5.2. Let $\mathcal{F}_1$ be the class $\{I(u \leq (t \wedge y)) : t > 0\}$. Using the usual Glivenko–Cantelli theorem, it is apparent that $\mathcal{F}_1$ is a P-Glivenko–Cantelli class. Additionally, suppose that $\mathcal{F}_2$ and $\mathcal{F}_3$ are the same class, that is,

$$\{(h(y, t) : t \in [0, a]\}.$$

Hence, through the using of the strong law of large numbers it seems that $\mathcal{F}_2$ and $\mathcal{F}_3$ are also P-Glivenko–Cantelli classes.

Now, let $G(x, y, z) = xyz$, which is obviously continuous. Recollecting the assumption of the theorem and considering the fact that for any arbitrary t such that $t \in [0, a]$. As $Y \epsilon [0, t]$, then $Y \leq t$, so we have

$$|(\frac{\mu(t) - Y_i I(Y_i \leq t) - tI(Y_i \geq t)}{Y_i})| \leq \frac{\mu(t) + 2t}{Y_i},$$

according to $\mu(t) = \int_0^t S(u)du$, then we get

$$\frac{\mu(t) + 2t}{Y_i} \leq \frac{3t}{Y_i},$$

there exists $C\epsilon[0,a]$ such that

$$(\frac{\mu(t)-Y_iI(Y_i\leq t)-tI(Y_i\geq t)}{Y_i})\leq(\frac{C}{Y_i}),$$

and the class of functions

$$\mathcal{S}:=\left\{(\frac{\mu(t)-Y_iI(Y_i\leq t)-tI(Y_i\geq t)}{Y_i})^2;t\in[0,a]\right\},$$

definitely has an integrable envelope function $(C/Y)^2$. Thus, considering the result in the recent argument, we have

$$\sup_{t\in[0,a]}\left|\frac{1}{n}\sum_{i=1}^{n}D_i^2(\mu(t))-E(D_i^2(\mu(t)))\right| \quad = \quad o_p(1),$$

uniformly over $[0,a]$. □

Lemma 5.3. *Let $E[\frac{1}{Y^2}]<\infty$. Then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_i(\cdot)$ converges weekly to a mean zero Gaussian process $\vartheta(\cdot)$ in the Skorokhod-space $D[0,a]$, where*

$$\vartheta(t)=\int_0^\infty h(y,t)dB\left(G(y)\right),$$

and

$$\mathrm{Cov}(\vartheta(t),\vartheta(s))=\int_{t\vee s}^{\infty}h(y,t)h(y,s)dG(y),$$

in which $B(\cdot)$ is a Brownain Bridge on the unit interval.

Proof. Considering the Equation (3.2), it follows that

$$\mu_n(t)-\mu(t) \quad = \quad -\frac{\mu}{\sqrt{n}}\sum_{i=1}^{n}D_i(\mu(t))+\mu\mu(t)\int_0^\infty\frac{1}{y}\Upsilon_nd(y), \tag{5.14}$$

where $\Upsilon_n(y)=\sqrt{n}(G_n(y)-G(y))$. Using $\mu_n(t)-\mu(t)=\mu\int_0^\infty I(y\leq t)+\frac{t}{y}I(y>t)\Upsilon_nd(y)$ and Equation (5.14), we can observe

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_i(\mu(t)) \quad = \quad \int_0^\infty(\frac{\mu(t)}{y}-z(y,t))\Upsilon_nd(y). \tag{5.15}$$

There exists a Gaussian process similar $\Upsilon(\cdot)$ such that

$$\Upsilon_n(\cdot)=\sqrt{n}\left(G_n(\cdot)-G(\cdot)\right)\xrightarrow{\mathcal{L}}B(\cdot).$$

Therefore

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D_i(\mu(t))\xrightarrow{\mathcal{L}}\int_0^\infty(\frac{\mu(t)}{y}-z(y,t))\Upsilon_nd(y). \tag{5.16}$$

□

Proof of Theorem 3.2. As $E[\frac{1}{Y^2}] < \infty$, therefore, considering the proof of Lemma 3 of Owen et al. (1990), we have

$$\max_{1\le i\le n}\Big|\frac{1}{Y_i}\Big| = o_p(n^{1/2}), \tag{5.17}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\Big|\frac{1}{Y_i}\Big|^3 = o_p(n^{1/2}). \tag{5.18}$$

By using (5.17), it can be seen that uniformly over $t \in [0, a]$,

$$\begin{aligned}\max_{1\le i\le n}|D_i(\mu(t))| &= \max_{1\le i\le n}\Big|\Big(\frac{\mu(t) - Y_i I(Y_i \le t) - tI(Y_i \ge t)}{Y_i}\Big)\Big| \\ &\le C\max_{1\le i\le n}\Big|\frac{1}{Y_i}\Big| \\ &= o_p(n^{1/2}).\end{aligned} \tag{5.19}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\Big|D_i(\mu(t))\Big|^3 = o_p(n^{1/2}). \tag{5.20}$$

According to (5.19) and (5.20) and similar to the proof of Theorem 3.1, we get

$$\begin{aligned}\kappa(\mu(t)) &= \sum_{i=1}^{n}(\varrho(t)D_i(\mu(t))) + o_p(1) \\ &= \frac{\Big(\sum_{i=1}^{n} D_i(\mu(t))\Big)^2}{\sum_{i=1}^{n} D_i^2(t)} + o_p(1),\end{aligned}$$

uniformly over $t \in [0, a]$. So, by using Lemmas 5.2 and 5.3, Equation (3.10) of Theorem 3.2 is proved. □

Proof of Theorem 3.3. According to Theorem 3.2 of de Uña-Álvarez (2002), we have

$$\sqrt{n}\Big(S_n(u) - S(u)\Big) \xrightarrow{\mathcal{D}} N(0, \varrho^2(u)),$$

where

$$\varrho^2(u) = \mu\{E(Y^{-1}I(Y \le u)) + F^2(u)E(Y^{-1}) - 2F(u)E(Y^{-1}I(Y \le u))\}. \tag{5.21}$$

Then, according to relation between $\mu(\cdot)$ and $S(\cdot)$ and delta method, we get

$$\sqrt{n}\Big(\mu_n(t) - \mu(t)\Big) \xrightarrow{\mathcal{D}} N(0, \delta^2(t)), \tag{5.22}$$

where

$$\delta^2(t) = S^2(t)\mu\{E(Y^{-1}I(Y \le t)) + F^2(t)E(Y^{-1}) - 2F(t)E(Y^{-1}I(Y \le t))\}.$$

□

Proof of Theorem 3.4. We have

$$\mu_n(t) - \mu(t) \quad = \quad \int_0^\infty \mu\Big(I(y \le t) + \frac{t}{y} I(y > t)\Big) \Upsilon_n d(y), \tag{5.23}$$

where

$$\Upsilon_n(\cdot) = \sqrt{n}\,(G_n(\cdot) - G(\cdot)) \xrightarrow{\mathcal{L}} B(\cdot).$$

Therefore $\sqrt{n}(\mu_n(t) - \mu(t))$ converges weakly to a mean zero Gaussian process $\mathcal{A}(\cdot)$, where

$$\mathcal{A}(t) = \int_0^\infty \mu\Big(I(u \le t) + \frac{t}{u} I(u > t)\Big) dB(G(u)).$$

□

References

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis*, **10**(4), 335–350.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, **97**(457), 201–209.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, **57**(4), 1030–1038.

Cox, D. (1969). Some sampling problems in technology», new developments in survey sampling (nl johnson et h. smith, éds.).

Cox, D. (2005). Some sampling problems in technology. *Selected Statistical Papers of Sir David Cox*, **1**, 81–92.

de Uña-Álvarez, J. (2002). Product-limit estimation for length-biased censored data. *Test*, **11**(1), 109–125.

Dong, B. and Matthews, D. E. (2012). Empirical likelihood for cumulative hazard ratio estimation with covariate adjustment. *Biometrics*, **68**(2), 408–418.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics*, **6**(1), 13–25.

Ghitany, M. E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, **78**(4), 493–506.

Hassan, A. S. and Assar, S. M. (2017). The exponentiated weibull power function distribution. *Journal of Data Science*, **16**(2), 589–614.

He, Y. and Zhou, Y. (2020). Nonparametric and semiparametric estimators of restricted mean survival time under length-biased sampling. *Lifetime Data Analysis*, **26**(4), 761–788.

Irwin, J. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiology & Infection*, **47**(2), 188–189.

Lee, C. H., Ning, J., and Shen, Y. (2018). Analysis of restricted mean survival time for length-biased data. *Biometrics*, **74**(2), 575–583.

Liang, W., Shen, J.-s., and He, S.-y. (2016). Likelihood ratio inference for mean residual life of length-biased random variable. *Acta Mathematicae Applicatae Sinica, English Series*, **32**(2), 269–282.

Mohammadian, Z. and Habibirad, A. (2024). Adjusted empirical likelihood analysis of restricted mean survival time for length-biased data. *Journal of Mahani Mathematical Research Center*, **13**(2).

Nassar, M. and Nada, N. (2011). The beta generalized pareto distribution. *Journal of Statistics: Advances in Theory and Applications*, **6**(1/2), 1–17.

Ning, J., Qin, J., Asgharian, M., and Shen, Y. (2013). Empirical likelihood-based confidence intervals for length-biased data. *Statistics in medicine*, **32**(13), 2278–2291.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 1725–1747.

Owen, A. et al. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**(1), 90–120.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2), 237–249.

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.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A*, 311–324.

Rather, A. A. and Subramanian, C. (2018). Length-biased sushila distribution. *Universal Review*, **7**, 1010–1023.

Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, **21**(15), 2175–2197.

Shen, P.-s. (2009). Hazards regression for length-biased and right-censored data. *Statistics & probability letters*, **79**(4), 457–465.

Shen, P.-S. (2011). Empirical likelihood ratio with doubly truncated data. *Journal of Applied Statistics*, **38**(10), 2345–2353.

Shi, J., Ma, H., and Zhou, Y. (2018). The nonparametric quantile estimation for length-biased and right-censored data. *Statistics & Probability Letters*, **134**, 150–158.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**(352), 865–871.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, **10**(2), 616–620.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, **76**(4), 751–761.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika*, **83**(2), 343–354.

Wicksell, S. D. (1925). The corpuscle problem: a mathematical study of a biometric problem. *Biometrika*, 84–99.

Zelen, M. (2004). Forward and backward recurrence times and length biased sampling: age specific models. *Lifetime Data Analysis*, **10**(4), 325–334.

Zhang, J., Zhang, J., Zhu, X., and Lu, T. (2018). Testing symmetry based on empirical likelihood. *Journal of Applied Statistics*, **45**(13), 2429–2454.

Zhang, M. and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*, **67**(3), 740–749.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M. A., Solomon, S. D., Trippa, L., and Wei, L. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics*, **72**(1), 215–221.

Zhao, Y. and Qin, G. (2006). Inference for the mean residual life function via empirical likelihood. *Communications in Statistics-Theory and Methods*, **35**(6), 1025–1036.

Zhou, M. (2021). Restricted mean survival time and confidence intervals by empirical likelihood ratio. *Journal of Biopharmaceutical Statistics*, **31**(3), 362–374.

Zhou, M. and Jeong, J.-H. (2011). Empirical likelihood ratio test for median and mean residual lifetime. *Statistics in Medicine*, **30**(2), 152–159.