Nonparametric Inference in the Presence of Biased Sampling

By

Ali Shariati **(45037663)**

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Statistics
July 2019

Typeset in LaTeX 2_ε.

Statement of Originality

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Ali Shariati (45037663)

Acknowledgements

I would like to specially thank my supervisors Dr. Hassan Doosti and Dr. Justin Wishart. Since the beginning, Hassan provided me with guidance, support and friendly discussions so that I could adapt myself to the new situations. Great thanks without doubt would have to be given to Justin for his assistance and encouragement that gave me a constant support and motivation. Also, I would like to thank my first M.Sc. supervisor Dr. Vahid Fakoor who took me under his wings and developed my interest and research skills in the field of Nonparametric Statistics. I would also like to thank Macquarie University and the Department of Mathematics and Statistics for their financial support and providing such a pleasant work environment.

With great heartfelt thanks, I would like to dedicate this thesis to my parents Mr. Reza Shariati and Mrs. Zohreh Fatemi for their constant love, encouragement and sacrifice. I would also like to warmly thank my family, Sanaz, Omid, my late beloved brother Hessam and Hannaneh, for supporting my decisions and giving their wisdom and guidance in all facets of my life over the years.

List of Conference participation and Publications

- Ali Shariati, Hassan Doosti and Justin Wishart. *Empirical Likelihood Inference for Length-biased and Right-censored Data via the Unconditional NPMLE*. Under preparation.
- Ali Shariati. *Empirical Likelihood-based Confidence Band in Length-biased Sampling*. Joint International Society for Clinical Biostatistics and Australian Statistical Conference, Melbourne, Australia (2018).
- Ali Shariati and Hassan Doosti. *Empirical Likelihood-based Inference for Right-censored Length-biased Survival Data*. BioNetwork 2018 'Killing it in Science', Macquarie University, Sydney, Australia (2018).
- Vahid Fakoor, Ali Shariati and Majid Sarmad. *The MRL function inference through empirical likelihood in length-biased sampling*. Journal of Statistical Planning and Inference. **196**, 115-131 (2018).

Abstract

Life expectancy is a key concept in survival analysis. When communicating with non-statisticians, average remaining lifespan is a more meaningful and comprehensible measure than the survival probability or the hazard rate. Therefore our research is centered on the mean residual lifetime function.

Survival data collected in a cohort of prevalent cases may be used to draw statistical inference. Since non-random sampling of subjects is involved, the data collected in this sampling scheme are biased. The most common case of this bias, occurring when the so-called stationarity assumption is satisfied, is called length-bias. While prospective prevalent cohort studies are commonly conducted to evaluate the progression of some disease over time, observations of many other sampling schemes have been reported to be length-biased. It is often necessary to take into account loss to follow-up of subjects, that is, the presence of censored data.

In this thesis, we study the problem of statistical inference (i.e. confidence interval) for length-biased data via the empirical likelihood method. The results are extended to construct a confidence interval for length-biased random censored data. The performance of these methods are illustrated through a simulation study and a data set obtained from a study of shrubs.

Contents

1

Preliminaries and Background

In this chapter, we present definitions and concepts required for this thesis. The discussions given in this chapter include concepts from probability theory, stochastic processes, time series, survival analysis and statistical inference. Although these topics are widely investigated in the literature, we have collated an overview which provides the background necessary for this thesis. We also provide some examples from the current literature to illustrate the topic presented in each section. A comprehensive overview of the literature relevant to this thesis may be found in Section 1.8.

1.1 Stochastic Process

DEFINITION 1.1.1 *A stochastic process is a collection of random variables X_t which is indexed by t such that $\{X_t, t \in \mathcal{T}\}$ and is defined on a probability space, when the indexing set $\mathcal{T}$ may be an arbitrary continuous or discrete set.*

For every $\omega \in \Omega$, the set $\{X_t(\omega), t \in \mathcal{T}\}$ is called the sample path of the stochastic process X_t. If the indexing set $\mathcal{T}$ is discrete, the stochastic process is defined as a set of discrete random variable indicated by X_j, in which j is an integer. Alternatively, $X(t)$ is used for stochastic process with a continuous index set. However, X_t denotes the stochastic process in general regardless of it is discrete or continuous.

We define two important stochastic process below that will be used in our analysis in the following chapters.

DEFINITION 1.1.2 *A stochastic process X_t is stationary if and only if for any possible values of n, s, and time sequence $t_1, \ldots, t_n$, the random variables $X_{(t_1)}, \ldots, X_{(t_n)}$ and $X_{(t_1+s)}, \ldots, X_{(t_n+s)}$ have the same cumulative distribution function.*

DEFINITION 1.1.3 *The stochastic process $X(t_1), \ldots, X(t_n)$ is a Gaussian process if for arbitrary time sequence $t_1, \ldots, t_n$, the random variables $X(t_1), \ldots, X(t_n)$ follow a multivariate normal distribution.*

1.1.1 Renewal Process

Renewal theory–the study of probability methods for analyzing renewal processes–originally arose from the study of some particular probability problems in the are of reliability that focus on failure and replacement of components. However, it has developed over the decades into the investigation of a wide range of practical probability problems. One of these problems (which is of concern in this thesis) is that of length-bias; this will be discussed in the next chapter. Before defining a renewal process, we require the definition of a counting process.

DEFINITION 1.1.4 *Let $\{N(t), t \geq 0\}$ be a stochastic process with $N(t)$ representing the number of events which have occurred up to time t and $N(0) = 0$. Then $\{N(t), t \geq 0\}$ is a counting process if it satisfies the following conditions:*

(I) $N(t) \in \mathbb{N} \cup \{0\}$ *for all values of t, where $\mathbb{N}$ is the field of natural numbers.*

(II) If $s < t$, *then* $N(s) \leq N(t)$.

(III) For $s < t$, *the number of events occurring in the interval* $(s, t]$ *is given by* $N(t) - N(s)$.

We now define the ordinary renewal process as a special kind of counting process.

DEFINITION 1.1.5 *Let $\{X(n), n \in \mathbb{N}\}$ denote a sequence of non-negative identical random variables with a common distribution function $F(\cdot)$. To avoid trivial cases, suppose that $F(0) = P\{X(n) = 0\} < 1$. Then we define $X(n)$ as the interval time in between event number n and event $n + 1$ in a counting process (Definition 1.1.4). Suppose that*

$$\begin{aligned} \mu &:= E\left(X(n)\right) \\ &= \int_0^\infty x dF(x), \end{aligned}$$

is the mean of time duration between successive events.

Considering the assumptions that $X(n) \geq 0$ and $F(0) < 1$, it can be concluded that $0 < \mu \leq \infty$. Now, if we define

$$S_0 := 0, \quad S_n := \sum_{i=1}^{n} X(i), \quad (n \geq 1),$$

then S_n is the time of observing the nth event. Suppose that $N(t)$ denotes the total number of events up to time t. Since $N(t)$ is the maximum amount of n satisfying $S_n \leq t$, we have

$$N(t) = \sup\{n : S_n \leq t\}.$$

This counting process is called a renewal process.

1.1.2 Brownian Motion

In 1826, Robert Brown, a Scottish botanist and palaeobotanist, while examining grains of pollen of the plant Clarkia pulchella suspended in water under a microscope, observed minute particles, now known to be amyloplasts (starch organelles) and spherosomes (lipid organelles), ejected from the pollen grains, executing a continuous jittery motion. Later, he observed the same pattern of motion in particles of inorganic matter, enabling him to rule out the hypothesis that the effect was life-related. This event is now known as Brownian motion,

although Brown did not provide a theory to explain the motion and even he was not the first person to report this phenomenon.

A French mathematician, Louis Jean-Baptiste Alphonse Bachelier, was the first person to introduce a mathematical model of Brownian motion in 1900. In his Doctoral thesis in Finance, he applied his method of modeling the Brownian motion for valuing stock options. Einstein (1905) proved the probabilistic nature of Brownian motion. He was the first person explained precisely how the motion that Brown had obtained was a consequence of the pollen being moved by individual water molecules (the immersed particle was continuously bombarded by the surrounding molecules), which was one of his first significant contributions to science. However, the definition presented here for a Brownian motion was given by Norbert Wiener in several articles he published in 1918.

There exist numerous examples of phenomena that may be modeled by Brownian motion. One simple example which could be used to describe Brownian motion is random walk. A random walk is a stochastic process which represents a path that consists of a succession of random steps on some mathematical space such as the integers. An elementary example of a random walk is the random walk on the integer number line $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Suppose that an object is at the origin of the integer number line at the time 0. It then moves right or left by taking a unique step at each time unit when all the possible steps (toward right or left) have exactly the same probabilities. Let X_i denote the distance that the object travels at time i. Accordingly, $\{X_i : i \in \mathbb{N}\}$ is a sequence of independent and identically distributed random variables with probability mass function

$$P(X_i = +1) = P(X_i = -1) = \frac{1}{2}.$$

Define $S_n := \sum_{i=1}^{n} X_i$ which is the position of the object at the nth step. Following the above paragraph, it is apparent that $S_0 = 0$. The series $\{S_n : n \in \mathbb{N}\}$ is called the simple random walk on $\mathbb{Z}$. It is of note that $\{S_n : n \in \mathbb{N}\}$ represents the distant walked with $E(S_n) = 0$ and $Var(S_n) = n$. Generally, any stochastic process $\{S_n : n \in \mathbb{N}\}$ with

$$S_n = S_{n-1} + X_n \quad (n \in \mathbb{N}); \quad S_0 = 0$$

is a random walk process, when the random variables X_i $(i \in \mathbb{N})$ are independent and identically distributed such that $E(X_i) = 0$ and $Var(X_i) = \sigma^2 < \infty$.

However, in reality an object exhibiting Brownian motion fluctuates constantly and continuously, and therefore a continuous timescale is required. Consequently, the continuity of time should be considered when investigating the asymptotic behavior of the object location S_n. For this purpose, we consider the partial summation $S_{[nt]}$, where $[nt]$ denotes the greatest integer less than or equal to nt and $1/n < t < \infty$. Bear in mind that when $0 \leq t \leq 1/n$, the value of $S_{[nt]}$ is assumed to be zero. Now, we define the random variable $W_n(t)$ as follows

$$W_n(t) := n^{-1/2} \frac{S_{[nt]}}{\sigma}.$$

Given the central limit theorem, it is easy to conclude that

$$W_n(t) \xrightarrow{\mathcal{L}} N(0, t),$$

where $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution (convergence in law). It is apparent that the variable $W_n(t)$ is random since it depends on the random variables $\{X_i : i \in \mathbb{N}\}$. On the

other hand, it is a function of the variable t. Therefore, in this situation the random function $W_n(\cdot)$ converges weakly to a random function called W, according to Donsker's theorem. The continuous-time stochastic process W is named the Wiener process (the standard Brownian motion) in honor of Norbert Wiener. He presented the mathematical fundamental of the theory of the random walk processes (Brownian paths) using Fourier series (Wiener, 1923, 1924).

DEFINITION 1.1.6 *Any stochastic process* $\left\{W_t, t \in \mathbb{R}^+\right\}$ *is a Wiener process (the standard Brownian motion) if it satisfies the following properties:*

(I) $P\left\{W_0 = 0\right\} = 1$,

(II) For any $0 < s < t$: $P\left\{W_t - W_s \leq x | W_u, 0 \leq u < s\right\} = P\left\{W_t - W_s \leq x\right\}$

(III) For any $0 < s < t$: $W_t - W_s \xrightarrow{\mathcal{L}} N(0, t-s)$,

(IV) The sampling path of the stochastic process is almost surely a continuous function t.

However, we may define a Brownian motion process in general by using directly a random walk process. Let we speed up the simple random walk on $\mathbb{Z}$ by taking smaller and smaller steps in shorter and shorter time intervals. Precisely, suppose we take a step of size Δx at each time interval Δt. In addition, let $\{X_i : i \in \mathbb{N}\}$ be the sequence of independent and identically distributed random variables defined above. Therefore, for the position of the object in the new random walk process we have

$$\begin{aligned} X_t &= S_{[t/\Delta t]} \\ &= \Delta x \left(X_1 + \cdots + X_{[t/\Delta t]}\right) \\ &= S_{[t/\Delta t]-1} + \Delta x\, X_{[t/\Delta t]} \end{aligned}$$

where $[t/\Delta t]$ denotes the greatest integer less than or equal to $t/\Delta t$. Then, it can be checked that

$$E\left(X_t\right) = 0,$$

$$Var\left(X_t\right) = (\Delta x)^2 \left[t/\Delta t\right].$$

To avoid trivial situations, let $\Delta x = c\left(\Delta t\right)^{1/2}$ for some constant $c \in \mathbb{R}^+ := [0, \infty)$. Given this relation for Δx, let $\Delta t \to 0$. Then,

$$Var\left(X_t\right) \longrightarrow c^2 t.$$

The stochastic process $\left\{X_t, t \in \mathbb{R}^+\right\}$ has Brownian motion when $\Delta t \to 0$.

DEFINITION 1.1.7 *Any stochastic process* $\left\{X(t), t \in \mathbb{R}^+\right\}$ *is a Brownian motion process if*

(I) $P\left\{X(0) = 0\right\} = 1$,

(II) The increments of $\left\{X(t), t \in \mathbb{R}^+\right\}$ *are independent and stationary.*

(III) For any $t > 0$: $X(t)$ *is normally distributed with* $E\left(X(t)\right) = 0$ *and* $Var\left(X(t)\right) = c^2 t$.

Given Definition 1.1.7, when $c = 1$ the process is called the standard Brownian motion (Wiener process) often.However, any Brownian motion like $X(t)$ can be converted to the standard Brownian motion by considering $X(t)/c$.

As mentioned, there are many real examples, including the path traced by a molecule as it moves in a liquid or a gas, the search path of a foraging animal, the price of a fluctuating stock and the financial status of a gambler, that may be approximated by random walk models. Therefore, random walk processes have a variety of applications in many disciplines such as physics, chemistry, economics, computer science, biology, ecology and psychology. Thus, random walks could be used to explain the behaviors of many processes, particularly in nonparametric studies. Here, we define two important processes that will be used in asymptotic study of the statistics obtained.

DEFINITION 1.1.8 *Let $\{X(t) : t \in \mathbb{N}\}$ be a stochastic process. If for any sequence $t_1, \ldots, t_n$, the random variables $X(t_i), \ldots, X(t_n)$ satisfy the multivariate normal distribution, then $\{X(t) : t \in \mathbb{N}\}$ is a Gaussian process.*

DEFINITION 1.1.9 *Define a stochastic process $\{B(t), 0 \leq t \leq 1\}$ on the probability space (Ω, ν, P). We call this stochastic process a Brownian bridge whenever it satisfies the following conditions:*

(I) $E\big(B(t)\big) = 0$

(II) For any $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq 1$ *($n \in \mathbb{N}$), the sequence of random variables $B(t_1), \ldots, B(t_n)$ is a Gaussian process.*

(III) The covariance of $B(t)$ is given by

$$Cov(s,t) = E\big(B(s)B(t)\big) = \min(s,t) - st.$$

(IV) The sampling path of the stochastic process is almost surely a continuous function of t.

Given Definition 1.1.9, let $\{X_t, t \in \mathbb{R}^+\}$ be a Brownian motion. If $B(t) = X(t) - tX(1)$, then $\{B(t), 0 \leq t \leq 1\}$ is a Brownian bridge process.

1.1.3 Empirical Process

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.)[1] random variables with a common distribution function $F(\cdot)$. Thus, the empirical distribution function of these random variables is

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \quad x \in \mathbb{R},$$

in which $I(\cdot)$ is the indicator function.

Accordingly, the strong law of large numbers indicates that

$$F_n(x) \xrightarrow{a.s.} F(x).$$

From the Glivenko-Cantelli theorem, it can be obtained that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

[1]Independent and identically distributed (i.i.d.)

DEFINITION 1.1.10 *A function $g(x,t)$ is an empirical process if*

$$g(x,t) := [t]\left(F_{[t]}(x) - F(x)\right) \quad x \in \mathbb{R},\ t \geq 0,$$

where $[t]$ is the largest integer that is less than or equal to t.

The theory of empirical distribution functions and empirical processes has a long history in probability and statistics. There have been many studies in this area (e.g. Csörgo and Révész (1981) and Gaenssler and Stute (1979)), among which, the strong approximation presented by Kiefer (1972), that is now being known as the Kiefer process, has attracted significant attention.

1.2 Survival analysis

A problem frequently faced by statisticians is the analysis of survival data. Survival data, sometimes called time-to-event data, spans the time between two events, namely initiating and terminating (or failure) events. A verity of statistical and probabilistic methods may be applied for analyzing time-to-event data. Survival analysis, that is the study of time-to-event data, is composed of the all statistical and probabilistic methods used for analyzing non-negative random variables.

Examples of survival data arise in numerous scientific fields, such as medicine, biology, public health, epidemiology, engineering, economics and demography. Therefore, the statistical tools we shall present are applicable to all these disciplines, albeit our focus in this thesis is mainly on the applications of the techniques proposed to biology and medicine.

Examples of initiating event from the literature include a medical diagnosis, appearance of a tumor, becoming a resident of a retirement community, and onset of addiction to a substance. Instances of terminating events comprise of death from a particular disease, a threshold of tumor or disease progression, mortality from natural causes, or even positive events such as cessation of substance abuse or disease remission. Some real examples of studies conducted on survival data are given below.

Analyzing data on the time to death for patients with psychiatric disorder, elderly residents of a retirement community, male laryngeal cancer, cancer of the tongue, and two categories of dementia including vascular dementia and Alzheimer's disease presented by Tsuang and Woolson (1977), Nahman et al. (1992), Kardaun (1983), Sickle-Santanello et al. (1988) and Wolfson et al. (2001), respectively. The survival time from marrow transplants to partial or complete remission in patients with acute myeloctic/lymphoblastic leukemia was studied by Copelan et al. (1991). Lagakos et al. (1988) were interested in drawing inference on the lifespan to Acquired Immune Deficiency Syndrome (AIDS) among patients with Human Immunodeficiency Virus (HIV).

Hamburg (1975) conducted a study on the time taken from the beginning of high school to first using of marijuana for Californian students. This issue had been first addressed by Turnbull and Weiss (1978).

Freireich et al. (1963) investigated the time to partial/complete remission in children with acute leukemia. Another study measured the time from percutaneous/surgical placement of a catheter until the onset of renal insufficiency infection in kidney dialysis patients (Nahman et al. (1992)).

1.2.1 Survival functions

There are four functions that are used to model the random survival time X. The first and the most commonly used function is the probability density function (or probability mass function), say $f(x)$. This indicates the probability of observing the terminating event at a time x. The second function is the survival function, denoted $S(x)$. This represents the probability a subject survives beyond time x. Thus,

$$\begin{aligned} S(x) &:= P(X > x) \\ &= 1 - F(x). \end{aligned}$$

The hazard rate function, or just the hazard function, indicates the probability that a subject with age x will experience the terminating event in the next instant. The mean residual lifetime (MRL)[2] function of a subject that has survived beyond time x, say $M(x)$, is another important function. We will discuss in detail this function in the next section. Theoretically, given one of these four functions, the other three functions may be derived uniquely. Along with the cumulative hazard function, these functions are the most often used to explain different characteristics of the random variable X.

1.2.2 The Mean Residual Lifetime Function

As mentioned previously, the mean residual lifetime function is very useful as it used in statistical studies to characterize the survival time of interest. Over the years, the MRL function has attracted considerable researchers' attention. The MRL function indicates the expected lifetime remaining for a subject at age x.

For any distribution function (d.f)[3] such as $G(\cdot)$, let τ_G be the right endpoint of its support. Thus $\tau_G := \inf\left\{x : G(x) = 1\right\}$. Given this definition, let $F(\cdot)$ be an arbitrary distribution such that $\tau_F < \infty$, then the mean residual function $M_F(\cdot) = M(\cdot)$ at any point $x > 0$ is defined by

$$\begin{aligned} M(x) &:= E(X - x \mid X > x) \\ &= \frac{I_{[0,\tau_F)(x)}}{S(x)} \int_x^\infty (t - x) dF(t) \\ &= \frac{I_{[0,\tau_F)(x)}}{1 - F(x)} \int_x^\infty (1 - F(t)) dt. \end{aligned} \tag{1.1}$$

Note that the last line in the equation above was obtained using integration by parts.

There are many studies in literature concerning the mean residual lifetime function. The MRL function estimation on a fixed interval $0 \le t \le \tau < \infty$ was studied by Yang (1978). He proved that the proposed estimator is strongly uniform consistent over the mentioned interval. By using compact topology, Yang (1978) obtained that the stochastic process $\sqrt{n}(\tilde{M}_n(t) - M(t))$ is weakly convergent to a certain Gaussian process on $[0, \tau]$ where $\tilde{M}_n(\cdot)$ is the estimator he proposed. After that, Hall and Wellner (1979) extended the findings of Yang (1978) over the positive half-line $\mathbb{R}^+$ by means of an appropriate metric.

Csörgo and Zitikis (1996) studied the mean residual life process over the whole positive half-line $\mathbb{R}^+$. They presented an approximation of the empirical mean residual life process by employing special weight functions. Csörgo and Zitikis (1996) also revealed the strong

[2]Mean residual lifetime (MRL)

[3]Distribution function (d.f.)

uniform consistency and the weak convergence of the proposed estimation over the positive half-line $\mathbb{R}^+$. By investigating the empirical mean residual life process as an integral form, Bae and Kim (2006) proved uniform asymptotic behaviours of the process over the positive half line. Under length-biased sampling with *Type I* censoring (defined in the next section), pointwise consistency of MRL was established by De Uña-álvarez (2004), when uncensoring is a special case. A comprehensive review of the recent statistical studies with regard to the MRL function inference in the presence of different biased sampling is presented in Section 1.8.

1.2.3 Censoring

Survival data collection is frequently associated with many restrictions which cause different obstacles in analysis. One of the most important problem in obtaining survival data is censoring. Censoring occurs when all we know about some of the underlying subjects is that they have experienced the event of interest (either the terminating event or the initiating event) in a certain period of time, but we do not know the exact time. In other words, survival data is censored when we are not able to observe the complete lifetime for some subjects owing to a variety of reasons, which is why the censored data is sometimes called "incomplete data" as well. There are many reasons that results in observing censored data, some of which are end of the study due to time limitation, loss to follow-up and leaving the study by some subjects.

Different types of censoring may be obtained depending on the sampling mechanism and different aspects of data and the underlying study restrictions. The possible types of censoring are *type I* censoring, right-censoring, left-censoring, interval-censoring, random censoring and multiplicative censoring. We discuss below a few types of possible censoring that may obtain in the data collection procedures discussed in this thesis.

Type I Censoring

In *type I* censoring, we only observe the terminating event if the subject experiences it before a fixed time. The fixed time is named the censoring time. However, since underlying subjects do not experience the initiating event at the same time necessarily, time of censoring for each subject is unique and different than the others.

Suppose we are conducting a statistical study involving a model organism or a clinical trial on the patients of a hospital with a specific disease. Assume that the study has commenced with a fixed number of cases. After a period of time, the researcher is obligated to finish the study due to time or money restrictions, reporting the results, while, some of the underlying subjects are yet to experience the terminating event. In this case that the calender time of censoring is the same for all of the censored cases, rather than being random, and all these subjects have been studying for the whole duration of study we face *type I* censoring.

DEFINITION 1.2.1 *Suppose that X denotes the lifetime of an underlying subject and C_0 is a fixed censoring time. The exact value of X is obtained if and only if X is less than or equal to C_0. Whereas, when X is greater than C_0, we are not able to observe the value of X and the subject is censored.*

For the simplicity of notations, the pair of random variables (T, δ) is used to indicate the lifespans of subjects under *type I* censoring. In this occasion, δ is the censoring indicator which is equal to 0 ($\delta = 0$), when the subject is censored, and is equal to 1, once the subject is not censored ($\delta = 1$). Also, the random variable T is defined to be $T := \min(X, C_0)$.

Random Censoring

Assume that we conduct a survey of patients with a special type of disease, like a specific cancer. The study commenced with a fix number of cases and continues until the last subject experience the underlying terminating event, e.g. death due to that kind of cancer. Now, if we lose some of the subjects during the study due to any other reasons different than the terminating event of interest (e.g. heart attack, accident, leaving the study by some subjects etc.), then these subjects experience random censoring.

Random censoring, called random right-censoring sometimes, is a type of "competing risks" censoring. This type of censoring occurs when some recruited subjects may experience some competing risks events which results in leaving the study by those cases.

DEFINITION 1.2.2 *Let $X_1, \ldots, X_n$ denote independent and identically distributed random survival times with a distribution function indicated by $F(\cdot)$. Moreover, suppose that $C_1, \ldots, C_n$ are non-negative i.i.d. random censoring variables from distribution function $G(\cdot)$. Under random censoring, a pair of random variables (Z_i, δ_i) $(i = 1, \ldots n)$ arises, when, for each i, $Z_i := \min(X_i, C_i)$, $\delta_i := I(X_i \leq C_i)$, and $I(\cdot)$ is the indicator function.*

It can be easily concluded from Definition 1.2.2 that the random variables Z_i are independent and identically distributed.

Multiplicative Censoring

The multiplicative censoring was initially introduced by Vardi (1989) as an artificial model which could be used to model several practical situations. Consequently, Vardi (1989) showed that the multiplicative censoring model could be applied, for example, to model survival data with right-censoring under some specific circumstances as well.

DEFINITION 1.2.3 *Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two sequences of i.i.d. random variables with the common distribution function $F(\cdot)$. While we observe the complete lifetime for the random variables $X_1, \ldots, X_n$, we are not able to observe the complete lifespans for random variables $Y_1, \ldots, Y_m$ owing to censoring. Instead, we obtain the random variables $Z_1, \ldots, Z_m$. Under multiplicative censoring, for each observable random variable Z_i $(i = 1, \ldots, m)$ there exist two random variable Y_i defined and U_i such that U_i possesses $U(0,1)$[4] and $Z_i = Y_i U_i$.*

It is worth mentioning that the numbers of uncensored and censored subjects (n and m) are random in most of the practical situations which could be modeled via multiplicative censoring. Nonetheless, Vardi (1989) has considered them to be fixed numbers that are not random.

1.2.4 Truncation

The other very common feature associated with the survival data is truncation. Different types of truncation may obtain in survival data collection depending on the sampling mechanism and study restrictions such as time and budget. Survival analysis becomes more complex under truncation, which sometimes is confused with censoring. Truncation is observed when the subjects are observed if and only if they have sufficient lifetime, distance, or any other

[4] $U(a, b)$ indicates the continuous uniform distribution defined on interval (a, b)

measures, depending on the study details, in order to be observable. Furthermore, truncation also occurs when the subjects recruited in the sampling procedure have experienced an event (initiating event or terminating event) by a fixed point in time. Depending on the sampling mechanism, we may obtain different types of truncation in the data set collected. The possible types of truncation are left-truncation, right-truncation, or both.

Left-truncation is obtained if there is a time delay between experiencing the initiating event of interest and recruitment of subjects. Thus, each individual enters at a special age, but subjects are not observed when they experience the initiating event. The subjects are then followed from the recruitment until experiencing the terminating event or the censoring time. Basically, truncation is obtained frequently in the study of time-to-event data. However, the application of truncated data is not restricted to survival analysis. Another possible situation in which truncation is obtained is when only the objects whose measures are greater than a specific amount are observable. Consequently, those that are smaller than that specific measure are not observable.

Right-truncation occurs when only those individuals that have experienced the terminating event before a definite time are detectable. For instance, in estimation of the distribution of the distance of a galaxy from the Earth, only those stars whose distances from the Earth are less than an specific amount depending on the accuracy of the telescope are discernible. And those stars that are beyond that specific distance are not detectable owing to right-truncation, albeit they exist. This is the principal difference between censorship and truncation that we do not have any information about the truncated data, while we know at least that the censored subjects have existed and we even obtained some information about them by observing them for a period of time.

In general, the truncation arises in survival analysis when we only obtain the subjects whose either initiating events or terminating events occur within a specific time interval, say (T_L, T_R). Thus, any subject that does not experience the event within this period of time is not observable and we do not receive any information about this individual. The scale of the interval (T_L, T_R) could be adapted to any required measure in order to explain truncation in different situations.

DEFINITION 1.2.4 *When the value of T_R in the defined interval is large enough (theoretically says T_R goes to infinity), we obtain left-truncation in the underlying sampling. In this situation, we only observe the subjects whose terminating events are experienced after the time T_L. In other words, in order to be discernible, ones lifetime, say X, should be greater than the left-truncation time T_L, $X > T_L$.*

Thus, all the individuals that have experienced the terminating event before the time T_L are not observable under left-truncation. Woodroofe (1985) illustrated left-truncation through several real examples in economic and astronomy. Moreover, numerous practical applications of different types of truncation in biology, epidemiology and medical sciences are presented in Klein and Moeschberger (2003).

Truncation Variable

In survival analysis, studying on time-to-event data is based on the initiating and terminating events. Under left truncation, since the time of experiencing the initiating event is random, it implies that the age of the subjects at the sampling time is random as well.

DEFINITION 1.2.5 *Let S (Start) and E (End) indicate the initiating and terminating events times, respectively. It is apparent that $S \leq E$ and the survival time $X = E - S$ is independent of S. Now, suppose that t indicates the recruitment time. Under left-truncation, any individual for that $X = E - S$ is greater than $A := t - S$ is observed. The time A which is the delay time for an individual to enter to the sample is called the truncation variable.*

Considering Definition 1.2.5, it is of note that, since S is random, $A = t - S$ is random as well. For better illustration, Figure 1.1 reveals the truncation times for two subjects that one is observed, while the other one is left-truncated. As can be seen, any subject, like X_1, whose lifetime is greater than truncation time is discernible ($X_1 > A_1$). Whereas, those individuals such as X_2 for that the truncation time is greater than the lifetime ($A_2 > X_2$) do not have any chance of being observed.

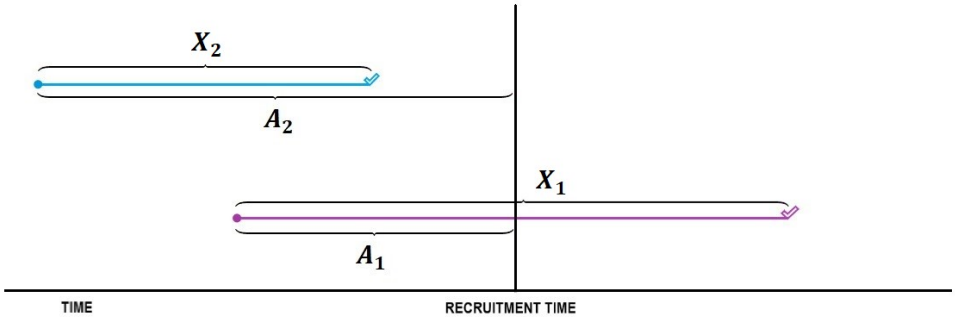

Figure 1.1: Truncation Variable

1.3 Length-Bias

In practical studies and real situations, obtaining an independent and identically distributed sample from the population of interest, called *target population* in this thesis, is impossible. Instead, a weighted sample, known as biased sample, is observed which implies a distribution that is different from the target population. There are many practical applications of similar bias in different disciplines such as forestry, genetic, economic, industry, biology, epidemiology and medical sciences. One can find several real examples of length-biased data in Patil and Ord (1976), Rao (1965) and Rao (1977).

There exists a special case of such bias that have a more prominent role in practice. Length-bias, which is the most frequent case of the discussed bias, occurs when the probability of being collected for a subject is proportional to its lifetime, length, or any other related measure. In other words, if the subjects whose lifetime, lengths, or measures are greater than the others have more chance of being collected in the sampling procedure, then the sample obtained suffers from length-bias.

The problem of length-bias was initially discovered by Wicksell (1925) while doing research into anatomy. When he was observing the corpuscles of organs under microscope, he noticed that only the corpuscles whose measures is greater than the magnification of the microscope are discernible. However, those cells that are smaller than this level are truncated and could not be observed. After that, the probabilistic model under the length-bias was investigated by McFadden (1962) and Cox (1962). In estimating the distribution of the lengths of fabric fibers, Cox (1969) figured out that linger fibers have more chance of being

collected and therefore the sample is subject to length-bias. Patil and Rao (1978) studied various types of bias commonly associated with the studies in demography and wildlife biology. Cristóbal and Alcalá (2001) have provided an invaluable overview of nonparametric studies on the statistical inferences when the data arises from a weighted distribution function.

DEFINITION 1.3.1 *Suppose that X is an arbitrary random variable with the continuous cumulative distribution function $F(\cdot)$. The random variable Y is length-biased with respect to the random variable X if the distribution function of Y has the representation,*

$$G(y) := \int_0^y \frac{x}{\mu_X} dF(x), \quad y \geq 0,$$

where μ_X is the mean of the random variable X.

The distribution $G(\cdot)$ presented in Definition 1.3.1 is called length-biased distribution, and the corresponding distribution $F(\cdot)$ is known as the unbiased distribution or the target distribution.

1.4 Cohort: Prevalent and Incident Cases

One of the problems which is frequently faced by researchers is the analysis of survival data arising form a population of interest. To analyze the data collected, what we need to do is estimate one of the survival functions defined. Survival data and the related variables arise in many disciplines such as medical tests, clinical trials, cohort studies, or prevalent cohort studies. Having defined the population of interest by specifying the related initiating and terminating events, the researcher needs to collect data from a previous data bank or by sampling from the existing subjects. Here, a cohort emerges–a group of subjects that have experienced the same initiating event of interest shapes a cohort. After recruitment of individuals, the members of a cohort are followed over time until they experience either the terminating(s) event or censoring. It is worth mentioning that in some studies, despite having the same initiating event (e.g. bone marrow transplant for acute leukemia patients), the patients may experience different terminating events (e.g. platelet recovery, relapse, acute Graft-versus-host disease, and death) in addition to censoring.

There are three different types of cohort that are used to collect survival data. Cohort of incidence cases is one of these cohorts in which we only recruit incident cases. Incident cases are the subjects that newly experience the initiating event (being diagnosed with the underlying disease over the course of sampling). Such cases experience the initiating event after the commencement of the study. However, prevalent cases are the subjects that have already experienced the initiating event (being diagnosed by the underlying disease before recruitment), but are yet to experience the terminating event or censoring. Cohort of prevalent cases is an alternative way of studying survival data. The third type of cohort consists of both prevalent and incident cases. Studying on prevalent cases is more interesting and practical, since it raising the number of potential cases. Indeed, it is extremely more time- and cost-efficient to study prevalent cases. However, various bias is associated with prevalent cohort study due mainly to left-truncation.

1.5 Stationarity Assumption

As mentioned, the survival data collected through a prevalent cohort study are left-truncated. The stationarity assumption of incidence event holds if and only if the number initiating

event, which is known as incident rate in literature, satisfies an stationary Poisson process (see Asgharian et al. (2006)). In other words, the rate of incidence process is constant under stationarity assumption. When this assumption satisfies, it can be simply proven that the truncation variable defined in Definition 1.2.5 follows a uniform distribution defined on an interval greater than the length of time from recruitment to initiating event.

Left-truncated survival data is basically analyzed by conditioning on the truncation times, which are the values realized for truncation variables. This approach is known as the conditional approach[5] in literature. A huge body of research has been done during the recent decades using the conditional approach. When the stationarity assumption is violated the model is not identifiable, and therefore conditional methods are mostly used. However, if there are adequate reasons that the stationarity assumption holds, it implies the observations follows the length-bias distribution. Consequently, the normal empirical estimator is the unconditional nonparametric maximum likelihood estimator[6], and it is not inevitable to condition on the observed truncation times. The following lemma indicates that how stationarity assumption implies the length-biased distribution.

LEMMA 1.5.1 *Let X denote the random variable in regard to survival time of the subjects that follows probability density function $f(\cdot)$. Assuming the stationarity assumption of incidence, which equivalently means that the truncation random variable T have a uniform distribution defined on an interval $(0, \theta)$, then for the distribution function of the observed survival data X under left-truncation T we have*

$$\begin{aligned} g(x) &:= f(x|X > T) \\ &= \frac{xf(x)}{\mu}, \end{aligned}$$

in which μ is the mean of the random variable X. This is the length-bias distribution presented in Definition 1.3.1.

Proof. *See the Appendix I for the proof of this lemma.*

1.6 Cross-Sectional Sampling

Cross-sectional sampling is a method for collecting data in which we only obtain the subjects in an instant or a very short period of time. However, it does not necessarily means that all the data are collected in that point in time. Instead, it means that the recruitment is done in an instant or a very short period of time. In survival analysis, the recruited subjects are then monitored over time in order to record the terminating event or censoring time for all of them.

The type of sampling which leads to a cohort of prevalent cases is cross-sectional sampling. Because, a prevalent sampling design only collect survival data from individuals who have already been diagnosed with a condition or disease but have yet to experience the failure event. In other words, we only select the subjects who are in the middle of initiating and terminating events at sampling time which is an instant or a very short period.

Figure 1.2 illustrates a cross-sectional sampling procedure for survival data. The sampling time is the start of the study. It is revealed that only violet, red and green subjects are observed. Whereas, the blue subjects do not have any chance of being collected due to left-truncation. Moreover, the subjects who will experience the initiating event after recruitment

[5]Conditional approach
[6]Unconditional approach

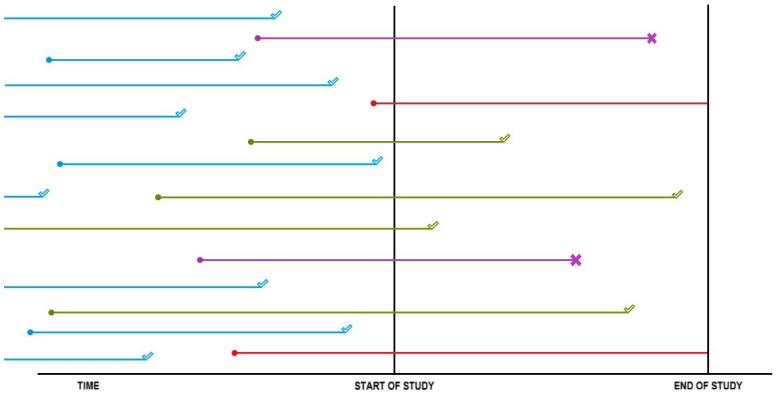

Figure 1.2: Cross-sectional sampling in the presence of censoring

are not detectable. Accordingly, the random sampling of subjects is violated (non-random sampling[7]) in cross-sectional surveys and prevalent cohort studies. It is worth mentioning that, among all the subjects recruited, only the green cases have experienced the terminating event. However, while the cases shown violet are randomly censored, those shown red have experienced the *type I* censoring.

1.7 EM algorithm

The problem investigators frequently face is maximizing the likelihood function while the function and even its logarithm do not result in closed forms. To deal with this problem, the Expectation Maximization (EM)[8] algorithm is a practical approach that provide researchers with an accurate approximation of the maximum likelihood estimator (MLE)[9]. This method is an iterative procedure that was generalized to analysis of incomplete data by Dempster et al. (1977). Having figured out that a similar technique had been used many times in prior studies, Dempster et al. (1977) presented a general framework and named it EM algorithm for the first time.

Suppose that $Y_1, \ldots, Y_n$ is a random sample with d.f. $g(\cdot|\theta)$ and θ is the underlying parameter that we want to estimate. Let Y_i $(i = 1, \ldots, n)$ be a function of the random variable X_i with probability density function $f(\cdot|\theta)$. We assume the set of random variables $(Y_1, \ldots, Y_n)$ denotes incomplete data and $(X_1, \ldots, X_n)$ is the corresponding set of complete data. In practice, the set of $(X_1, \ldots, X_n)$ is not observable, instead, the set $(y_1, \ldots, y_n)$ is realized for $(Y_1, \ldots, Y_n)$ as a result of a biased sampling procedure. Each iteration of the algorithm consists of two separate steps, namely Expectation step (E-step) and Maximization step (M-step). By selecting an arbitrary amount of θ, say $\theta^{(0)}$, the algorithm commences. In E-step, we calculate the following conditional expectation.

$$\begin{aligned} Q\left(\theta|\theta^{(n)}\right) &:= E\left[\log\left\{f\left(X_1, \ldots, X_n|\theta\right)\right\} \mid y_1, \ldots, y_n, \theta^{(n)}\right] \\ &= E\left[\log\left\{L(\theta)\right\} \mid y_1, \ldots, y_n, \theta^{(n)}\right] \\ &= E\left[l(\theta)|\, y_1, \ldots, y_n, \theta^{(n)}\right] \quad (n = 0, 1, 2, \ldots), \end{aligned} \tag{1.2}$$

[7]Non-random sampling
[8]Expectation Maximization (EM) algorithm
[9]Maximum likelihood estimator (MLE)

where $\theta^{(n)}$ is the estimation of θ in the nth repetition. In the next step, say M-step, given a fixed amount of $\theta^{(n)}$, we should find the maximum of (1.2) as a function of the variable θ, obtaining $\theta^{(n+1)}$.

By means of the following equation, Dempster et al. (1977) have proven that the EM algorithm leads to maximizing the likelihood function.

$$\begin{aligned} H\left(\theta|\theta^{(n)}\right) &:= E_X\left[\log\left\{f\left(X_1,\ldots,X_n|\, y_1,\ldots,y_n,\theta\right)\right\} \mid y_1,\ldots,y_n,\theta^{(n)}\right] \\ &= Q\left(\theta|\theta^{(n)}\right) - \log\left\{g\left(y_1,\ldots,y_n|\,\theta\right)\right\}, \end{aligned}$$

where the variables $y_1,\ldots,y_n$ and also $\theta^{(n)}$ are assumed to be known. In addition, $X_1,\ldots,X_n$ are unknown random variables, and θ is an unknown variable that is not random. Dempster et al. (1977) showed that the function $H\left(\theta|\theta^{(n)}\right)$ hits its maximum at the point $\theta = \theta^{(n)}$. Furthermore, for any $n \in \mathbb{N} \cup \{0\}$, the likelihood function of observations satisfies the following equation.

$$L\left[g(y_1,\ldots,y_n|\,\theta^{(n+1)}\right] \geq L\left[g(y_1,\ldots,y_n|\,\theta^{(n)}\right].$$

Thus, the likelihood increases in each repetition of the EM algorithm, which continues until the EM algorithm reaches an optimized point. Under some circumstances, this optimized point is the maximum likelihood estimation. It is of note that the convergence of EM algorithm to the optimized point does not depends on the selection of the initial point $\theta^{(0)}$. Therefore, the invaluable and distinct advantage of the EM algorithm is detect of complete data $X_1,\ldots,X_n$ by using incomplete observations.

1.8 Literature Review

In this section, we conduct a comprehensive review on the recent literature regarding nonparametric inference for the survival data. It was mentioned that in many practical situation we are not able to observe an i.i.d. sample of complete data from the target population. Consequently, we restrict our attention to the researches including censoring and biased sampling. On the other hand, utilizing the empirical likelihood (EL)[10] method, one can draw statistical inference for a population parameter, say θ, in general. This is one of the most important advantages of the empirical likelihood method that provides researchers with a flexible framework for making nonparametric inference (i.e. hypothesis testing, or constructing confidence interval) for a parameter generally. Specifying a relation for the parameter of interest, an EL type theorem which is a nonparametric alternative of Wilks's theorem is required in order to obtain the asymptotic behavior of the empirical likelihood ratio.

Accordingly, empirical likelihood is another issue that is highly relevant to our researches for this thesis. There has been a lot of research centred on survival functions based on an i.i.d. sample of the population of interest (e.g. Owen (2001)). However, there have been fewer studies done on survival data when considering censoring and truncation. Zheng et al. (2014) proposed an adjusted EL method for constructing confidence interval for a population parameter. Although adjusted EL method had already been used for unbiased data, Zheng et al. (2014) presented an extension for analysis of the right-censored data by using an influence function. It was observed that the introduced confidence interval improved the performance of the empirical likelihood method particularly for small sample sizes.

[10]Empirical likelihood (EL)

By applying the unconditional nonparametric maximum likelihood estimator (NPMLE)[11] of the length-biased d.f., Ning et al. (2013) studied the problem of constructing confidence interval for summary statistics using empirical likelihood method. They considered the possibility of observing censored subjects in their sample. Ning et al. (2013) derived that the empirical log-likelihood under some regularity conditions has a limiting chi-square distribution with one degree of freedom. The method proposed may be used to do hypothesis testing as well. They investigated the application of the method presented for the survival function at a fixed point in time, say x_0.

There has been lots of work on survival functions by adopting conditional approach. But, different studies indicate that the empirical log-likelihood of a population parameter like θ using censored data asymptotically goes to a scaled chi-square distribution (See e.g. Zhao and Qin (2007), Wang and Jing (2001) and Hjort et al. (2009)). Since the scale parameter is a function of the unknown asymptotic variance, the scale parameter is required to be estimated separately which results in the decreasing the coverage probability for the parameter of interest. However, under some normal circumstances, He et al. (2016) proved that the empirical log-likelihood of the parameter goes asymptotically to chi-square distribution with one degree of freedom by means of a special influence function as an estimating function. Moreover, the estimating function derived in He et al. (2016) exhibited a smaller asymptotic variance than those in Wang and Jing (2001) and Qin and Zhao (2007). Thus, the coverage probability of their confidence intervals indicated better results by comparison with other alternative methods.

As mentioned, life expectancy is a very important concept in survival analysis. There has been increasing tendency to do research into the residual lifetime. Reporting the average remaining lifespan to researchers from other disciplines, physicians, and patients is a more meaningful and easier to understand measure than the survival chance or the hazard rate. Therefore, our main focus is on the researches concerning statistical inference for the mean residual lifetime. Zhao and Qin (2007) pointed out that statistical inference on the linear functional of cumulative hazard function may result in a flexible framework that could be applied for analysis of survival data. Using this framework, Zhao and Qin (2007) studied the problem of constructing confidence interval and band for two linear functionals of cumulative hazard function, namely partial mean lifetime and distribution function. They proved confidence bands for these functions based on independent right-censored data. Moreover, they discussed the application of the method proposed for the cumulative hazard function as well.

Qin and Zhao (2007) propose an EL-based method for drawing statistical inference on mean residual lifetime function using right-censored data. They derived that the logarithm of the EL ratio for the mean residual lifetime function converges in distribution to a scaled chi-square distribution. They applied this limiting distribution to construct confidence interval. After that, Chaubey and Sen (2008) discussed the problem of point estimation for the MRL function using right-censored data. Inspired by techniques introduced by Chaubey and Sen (1998), Chaubey and Sen (2008) proposed a smoothing method to improve the performance of the product-limit estimator. It was revealed that the proposed method does not result in boundary bias, despite the standard kernel smoothing estimator.

Zhou (2011) investigated the problem of testing hypothesis for the MRL function using random censored time-to-event data. They introduced an empirical likelihood-based method

[11] Nonparametric maximum likelihood estimator (NPMLE)

which may be used for constructing confidence interval and confidence band as well. Following this, Chan et al. (2012) proposed a proportional mean residual life regression model for analyzing the right-censored and length-biased survival data collected from a prevalent survey. The method they presented was an extension of the technique had been introduced earlier by Oakes and Dasu (1990). Afterwards, Zhao et al. (2013) discussed the problem of nonparametric estimation of the MRL function with left-truncated and right-censored data. They proved that their estimator is weakly convergent to a zero mean Gaussian process. Then, Wu and Luan (2014) proposed a new estimator of the MRL function based on length-biased and right-censored data. It was observed that the proposed estimator converges in distribution to a normal random variable under some circumstances. They also illustrated that the proposed estimator is more efficient that other alternatives for length-biased and right-censored.

In Fakoor (2015), the problem of statistical inference for a non-parametric estimator of the mean residual lifetime function using length-biased data was discussed. The estimator proposed is the unconditional NPMLE of the MRL function. He proved the strong uniform consistency and weak converge of the estimator. Following this, given the wide applications of Cox model in cohort studies, Bai et al. (2016) presented a semiparametric approach for the proportional residual life model in length-biased setting. The possibility of observing censored data was considered in the model. The introduced method may be used for the construction of confidence band for the MRL function when the Cox model for data is assumed.

Liang et al. (2016) made statistical inference on mean residual lifetime using empirical likelihood. They considered the possibility of censoring in their model. They also discussed the problem that the observation of many survival data are length-biased. However, the mean residual lifetime function discussed in this article is not the common MRL function defined in (1.1). Additionally, although they discussed the problem of statistical inference using length-biased observations, the proposed method was not based on length-biased data and the statistical inference was not made on the target population. Instead, they used unbiased data for they analysis, while unbiased observations are not available in the problem of length-bias. These issues extensively limits the application of this article. Afterwards, Chen et al. (2017) proposed a nonparametric method by means of Jackknife empirical likelihood to test the equality of two MRL functions. Their Jackknife empirical likelihood was derived using a a U-statistic. They proved that the empirical log-likelihood ratio is convergent in distribution to a chi-square distribution with one degree of freedom.

Ultimately, Fakoor et al. (2018) is the most recent article related to the underlying problem. The paper involved studying the problem of making statistical inference for MRL function under length-bias. An EL-based procedure was presented for constructing confidence interval for the MRL function through length-biased data. Utilizing the findings of Fakoor (2015), Fakoor et al. (2018) introduced another normal approximation method for obtaining confidence interval for MRL function. Furthermore, the empirical log-likelihood of the MRL function was derived to be convergent weakly to a Gaussian process. As a consequence, a confidence band for the MRL function using length-biased data was obtained by means of a Gaussian process approximation procedure.

In the rest of this thesis: Chapter 2 commences with reviewing two real examples in which length-biased data have arisen. Then, the preliminaries and the unconditional NPMLEs of survival functions are discussed. Following that, the fundamental theoretical framework of the empirical likelihood is discussed. An empirical likelihood-based method is then proposed for constructing confidence interval for the MRL function under length-bias. The method proposed is inspired by Fakoor et al. (2018). It is noticed that the method can be

alternatively applied for hypothesis testing. Afterwards, the finite sample performance of the introduced method is inspected using a simulation study. For better comparison, another normal approximation method has been applied to the simulated data, exhibiting superiority of the EL method. A real data application for the method is presented at the end.

Chapter 3 involves investigating into the length-biased right-censored data. This chapter starts by introducing another real study in which length-biased and right-censored data has arisen. Afterwards, this chapter continues with discussing the preliminaries and the NPMLEs for the survival functions using incomplete length-biased data. Then, a new method for constructing confidence interval for the length-biased right-censored data has been proposed by means of the empirical likelihood method. The method involves applying the mean residual life function to draw statistical inference on the survival data. Following that, a simulation study is conducted to reveal the performance of the method proposed.

Eventually, Chapter 4 presents a detailed overview of the discussions presented in this dissertation. The chapter continues with discussing the future of the thesis and the possible further investigations using the results obtained toward this monograph.

2

Confidence Interval for the MRL Function Based on Length-biased Data

2.1 Introduction

A frequent problem statisticians face is the analysis of survival data. Survival data arises in various disciplines, such as reliability, engineering, economics, demography, biology, epidemiology and public health. In various studies, we are frequently obliged to deal with the obstacle of analyzing the observation obtained from a bias sampling procedure. In the past decades, bias statistical inferences caused by different bias sampling procedures have been reported extensively in literature. Among various bias sampling procedures, there is a very important type of bias called length-bias which has turned many researchers' attention for many years. Length-biased data obtained when the data arises in a sampling procedure are not randomly selected, instead, the sample observations are collected with probabilities proportional to their measures.

In this thesis, we are interested to make statistical inference on studying survival data concerning progression of a disease. Usually, prospective prevalent cohort surveys are conducted to assess the history of a disease (e.g. time to onset of acquired immune deficiency syndrome or death) among recruited individuals who have already experienced the initiating event (e.g. diagnosis of human immunodeficiency virus, cancer or Alzheimer's disease). A frequent used sampling scheme to estimate the lifetime between the initiating event and the terminating event is a cross-sectional follow-up study. On average, the prevalent subjects observed in such studies are confirmed to possess a longer lifetime, as the sample does not consist of a random selection from the target population. The most common case of this bias, occurring when the so-called stationarity assumption is satisfied, is called length-bias.

However, the application of the length-biased data is not restricted to biology, epidemiology, medical sciences, and public health. Efromovich (2008) studied the distribution of the ratio of alcohol in the blood of liquor-intoxicated drivers in England. The data was collected from routine police reports on arrested drivers charged with driving under the influence of alcohol. In this study, it was observed that more intoxicated drivers had more chance of being identified and arrested by the police. So the data set was reported to be length-biased.

Another biased sampling procedure was reported in a study into the lifetime of automobile brake pads by Kalbfleisch and Lawless (1992). A real data was initially given by Lawless (2011). The number of units (distance in kilometers) driven until the brake pads were reduced to a specific minimum thickness was considered as the lifetime. Therefore, it apparent that the special minimum thickness was the terminating event in this study. To study the survival distribution, a manufacturer selected a random sample of vehicles sold over the course of proceeding 12 months in a known group of dealers. Accordingly, only those cars that still had been using the initial brake pads could be selected. Because of the non-random sampling procedure (cross-sectional sampling) of subjects, the brake-pads data was reported to be length-biased due to uniform left truncation of data. therefore the presence of left-truncation. There may be found many real situations in which such survival data arises.

Although biased inferences due to length-biased sampling have been widely identified in the statistical and epidemiological literature, there has not been any adequate solution until recently for the problem of constructing the confidence band for the MRL function (Fakoor et al. (2018)), in which an empirical likelihood (EL) procedure for the MRL function based on length-biased data is proposed.

In this chapter, we study the problem of drawing statistical inference for the length-biased survival data. An empirical likelihood-based method is proposed for statistical analysis. We applied this method for constructing confidence interval for the MRL function through length-biased data. The method proposed may be used for hypothesis testing on the MRL function as well. We introduce a normal approximation method for interval estimation of the MRL function as well. In addition to the theoretical viewpoint, we consider the application of our findings for epidemiological sciences. The performance of the proposed method is illustrated using a real data set on the widths of shrubs collected through the line intercept sampling method.

2.2 Preliminaries and NPMLE

Suppose that associated with each subject in a target population we have a double (X', T'), in which X' denotes the lifetime (or more generally the failure time), T' is the truncation time. It is often reasonable to assume that X' is independent of T'. Suppose that the distribution function of X', say $F_{X'}(\cdot) = P(X' \leq \cdot)$, is defined on $\mathbb{R}^+ = [0, \infty)$ with a finite mean μ. In cross-sectional sampling, only the subjects satisfying the condition $X' > T'$ are observable. Under the stationarity assumption, the probability density function of the observed survival time, say $g(\cdot)$, is related to the unbiased density ($f_{X'}(\cdot)$) by means of the following equation

$$\begin{aligned} g(x) &= f_{X'|X'>T'}\left(x|X' \geq T'\right) \\ &= \frac{x f_{X'}(x)}{\mu_{X'}}. \end{aligned} \tag{2.1}$$

As mentioned, (2.1) is the length-biased density function.

Let X denote an observation in cross sectional sampling under the stationarity assumption. Similar to the above equation, we can deduce for the relation between the distribution of the length-biased observation and that of the target population that

$$\begin{aligned} G(x) &= F_{X'|X'>T'}\left(x|X' \geq T'\right) \\ &= \mu_{X'}^{-1} \int_0^x s dF_{X'}(s), \quad x \geq 0. \end{aligned} \tag{2.2}$$

For the random variable X' with distribution $F_{X'}(\cdot)$, let $F_{X'}(x^-)$ denote $P(X' < x)$. Therefore $P(X' = x) = F_{X'}(x) - F_{X'}(x^-)$. Suppose that $I(A)$ denotes the indicator of the event A. Given the real valued random variables $X'_1, X'_2 \ldots, X'_n$ assumed to be independent with the common cumulative distribution function $F_{X'}(\cdot)$, it is well known that the nonparametric likelihood for $F_{X'}(\cdot)$ is

$$L(F_{X'}) = \prod_{i=1}^{n} \left(F_{X'}(X'_i) - F_{X'}(X'^-_i)\right). \tag{2.3}$$

The value $L(F_{X'})$ is the probability of obtaining exactly the observed sample values $X'_1, \ldots, X'_n$, from the distribution function $F_{X'}(\cdot)$. As a direct consequence, $L(F_{X'}) = 0$ when $F_{X'}(\cdot)$ is continuous. Accordingly, in order to have a positive nonparametric likelihood, a distribution $F_{X'}$ must possess positive probability for every one of the observed data values. The following theorem proves that the nonparametric likelihood function $L(F_{X'})$ is maximized by the empirical distribution function, and thus, the empirical distribution function is the nonparametric maximum likelihood estimator[1] (NPMLE) of $F_{X'}$.

THEOREM 2.2.1 *Suppose that $X'_1, X'_2 \ldots, X'_n$ are real valued and independent random variables with a common distribution function $F_{X'}(\cdot)$. Let $F_n(\cdot)$ denote their empirical distribution function and $F(\cdot)$ be any distribution function. If $F_n(\cdot) \neq F(\cdot)$, then $L(F) < L(F_n)$*

Proof. See the Appendix II for the proof.

Theorem 2.2.1 indicates that the classical NPMLE of $F_{X'}(\cdot)$ is simply the empirical distribution function using independent and identically distributed random variables sampled from the population of interest. However, such a sample is inaccessible in the presence of biased sampling. The empirical distribution function of $G(\cdot)$ is

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \tag{2.4}$$

which is the NPMLE of the length-biased distribution function. Following the procedure proposed by Cox (1969), for the empirical estimator of $F_{X'}$ we have

$$\begin{aligned} F_n(x) &= \mu_n \int_0^x s^{-1} dG_n(s) \\ &= \frac{\mu_n}{n}\sum_{i=1}^{n} \frac{1}{X_i} I(X_i \leq x). \end{aligned} \tag{2.5}$$

where

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

Let $M_{F_{X'}}(x)$ be the MRL function defined in (1.1) corresponding to the distribution function $F_{X'}(\cdot)$. So, when our observations are length-biased, for the empirical counterpart of the MRL function we have

$$\begin{aligned} M_n(x) &:= E_{X'}(X - x \mid X > x) \\ &= \frac{I_{[0,X_{(n)})(x)}}{1 - F_n(x)} \int_x^\infty (1 - F_n(s))ds. \end{aligned} \tag{2.6}$$

[1]Nonparametric maximum likelihood estimator (NPMLE)

2.3 Nonparametric Empirical Likelihood Ratios

The likelihood ratio method plays a crucial role in hypothesis testing and constructing confidence intervals for parametric models. The confidence interval region for γ_0 or the hypothesis test that $\gamma_0 = \gamma$ is rejected, once $L(\gamma)$ is much smaller than $L(\hat{\gamma})$. According to Wilks's theorem, under some regularity conditions, $-2\log(L(\gamma_0)/L(\hat{\gamma}))$ converges to a chi-squared distribution as $n \to \infty$. This issue can be used to specify the exact confidence region or critical region. In other words, we are interested in obtaining how small $L(\gamma)$ should be in order to reject the hypothesis test that $\gamma_0 = \gamma$. For the parameter of the asymptotic distribution, Wilks's theorem indicates that the dimension of the set of γ values is the degrees of freedom in the chi-squared distribution. When we are interested in constructing a confidence region for the parameter γ, the image of a confidence region for γ may be created by

$$\left\{\theta(\gamma) \mid L(\gamma) \geq cL(\hat{\gamma})\right\},$$

where the threshold c is chosen by using Wilks's theorem with a degree of freedom equal to the dimension of the set of γ values.

Similarly, we may also use a ratio of the nonparametric likelihood as a basis for a hypothesis test or a confidence interval. For a distribution F, we define the likelihood ratio function as

$$R(F) = \frac{L(F)}{L(F_n)}, \tag{2.7}$$

when $L(F)$ is the nonparametric likelihood defined in (2.3). Now we can deal with nonparametric likelihood for making statistical inferences similar to the method for the parametric likelihood.

Let θ be some function of the distribution function $F_{X'}(\cdot)$, say $\theta = T(F_{X'})$ for some function $T(\cdot)$. We consider $F_{X'}(\cdot)$ as a member of a set of distribution function indicated by $\mathcal{F}$. We might consider $\mathcal{F}$ to be the set of all possible distribution functions on $\mathbb{R}$. However, we often consider a more restricted set of distribution function. The likelihood ratio profile is defined as follows.

$$\mathcal{R}(\theta) = \sup\left\{R(F) \mid T(F) = \theta, F \in \mathcal{F}\right\}. \tag{2.8}$$

In the above profile the supremum is taken on the set of all the distribution functions on $\mathcal{F}$ restricted to $T(F) = \theta$. In this thesis, $\mathcal{F}$ indicates all possible distribution functions on $\mathbb{R}$.

Accordingly, the hypothesis $H_0 : T(F_0) = \theta_0$ is rejected through empirical likelihood test, once $\mathcal{R}(\theta_0) < c_0$ for some threshold value c_0. Thus, an empirical likelihood-based confidence region may be obtained by analogy with the following relation.

$$\left\{\theta \mid \mathcal{R}(\theta) \geq c_0\right\}.$$

An empirical likelihood type theorem[2] (ELT) which is a nonparametric alternative of Wilks's theorem is needed in order to reach the precise threshold c_0 depends on the corresponding settings. From now onward, we will discuss the problem of drawing inference for the MRL function under a length-bias setting, and will present the appropriate ELTs for each case separately.

[2]Empirical Likelihood Theorem (ELT)

2.4 Empirical Likelihood for Length-biased Data

The fundamental principle for inferring a confidence interval for the survival functions through empirical likelihood method is to obtain the EL ratio statistics under the specified constraints using the Lagrange multiplier method. In this section, we are interested in testing $T(F_{X'})$, where $T(\cdot)$ is a function of the distribution function, while the available data arises from a length-biased sampling procedure. More precisely, we are interested in testing $T(F_{X'}) = M(x)$ here, when $M(x)$ is the MRL function defined in (1.1). As mentioned, if an i.i.d. sample from the underlying distribution had been observable, by substituting $F_{X'}(\cdot)$ with the classical empirical distribution function in $T(F_{X'})$, we would have been able to obtain the maximum likelihood estimator. However, the available data in length-biased sampling does not consist of i.i.d. copies of the target population. But, the $F_{X'}(\cdot)$ could be estimated using the empirical distribution of the observations by (2.5), and therefore $M(x)$ is related to the observations through (2.6). Accordingly, by plugging $F_n(\cdot)$ defined in (2.5) in $T(F_{X'})$, the maximum likelihood estimation of θ may be obtained. The principal aim of this thesis is to prove that an empirical likelihood-based confidence interval for survival data. For this purpose, the essentials of an EL-based confidence interval is to prove an appropriate ELT under different setting in each chapter. Accordingly, it will be observed that the confidence intervals under different setting possess sets of the following form.

$$\left\{\theta \mid \mathcal{R}(\theta) \geq c\right\},$$

where the threshold values c for different cases will be obtained.

We need to find a unique function $T(\cdot)$ of distribution $F_{X'}(\cdot)$, so that $T(F) = \theta$, in order to define the special profile likelihood ratio function for the mean residual function under length-biased data.

$$H_0 : \quad T(F) = \int_0^\tau \xi(s) dF_{X'}(s) = \theta.$$

The above null hypothesis may be alternatively considered as

$$\int_0^\tau \xi(s) - \theta dF_{X'}(s) = 0.$$

Since we do not have an independent and identically distributed random sample under length-biased sampling, the empirical likelihood approach consists in the estimation of the biased distribution function $G(\cdot)$ instead of the unbiased distribution $F(\cdot)$. Consequently, we should restrict our attention, equivalently, to the following hypothesis test for length-biased data.

$$H_0 : \quad T(G) = \mu_{X'} \int_0^\tau \xi(s) s^{-1} dG(s) = \theta, \tag{2.9}$$

which implies the following constraint:

$$\int_0^\tau \left(\xi(s) - \theta\right) s^{-1} dG(s) = 0.$$

Defining $\eta(s, \theta) := s^{-1}\left(\xi(s) - \theta\right)$, we can rewrite the constraint H_0 as follows.

$$H_0 : \quad E_X \eta(X, \theta) \quad = \quad \int_0^\tau \eta(s, \theta) dG(s)$$

$$= \quad 0. \tag{2.10}$$

When we are interested in constructing a confidence interval for the mean residual lifetime function at an arbitrary but fixed point x_0, we have

$$\begin{aligned} M(x_0) &= E_{X'}(X - x_0 \mid X > x_0) \\ &= \frac{E_{X'}\left[(X - x_0)\, I_{[x_0,\tau)}(X)\right]}{1 - F_{X'}(x_0)}. \end{aligned} \tag{2.11}$$

Now, assuming $\theta_0 = \theta(x_0) = M(x_0)$ that is the true value of the MRL function at time x_0, it can be deduced that

$$\begin{aligned} \frac{E_{X'}\left[(X - x_0)\, I_{[x_0,\tau)}(X)\right]}{1 - F_{X'}(x_0)} - \theta_0 &= \mu_{X'} \frac{E_X\left[\left(1 - \frac{x_0+\theta_0}{X}\right) I_{[x_0,\tau)}(X)\right]}{1 - \mu_{X'} \int_0^{x_0} s^{-1} dG(s)} \\ &= E_X\left[\left(1 - \frac{x_0 + \theta_0}{X}\right) I_{[x_0,\tau)}(X)\right] \\ &= 0. \end{aligned} \tag{2.12}$$

It is worth mentioning that the parameter of interest here is the MRL function at a fix point x_0, and therefore it is defined and closely related to the point x_0 ($M(x_0) = \theta(x_0)$). However, from now onward, for the simplicity of notation we only use θ_0 instead of $\theta(x_0)$ to emphasize that it is a function of x_0.

One can draw statistical inference for the MRL function at a fixed time like x_0. Considering (2.10), the equation (2.12) implies the following representation for the function $\eta(\cdot, \theta)$ with respect to the MRL function at x_0,

$$\eta_{x_0}(s, \theta_0) = \left(1 - \frac{x_0 + \theta_0}{s}\right) I_{[x_0,\tau)}(s). \tag{2.13}$$

It is of note that although in the above expression x_0 is fixed and $\eta_{x_0}(s, \theta_0)$ is a function of s and θ_0, it is defined and affected by x_0, and thus we have different function $\eta_{x_0}(s, \theta_0)$ for various points x_0, which is why we have used x_0 as a subscript for emphasis.

Accordingly, to test the null hypothesis defined or construct a confidence interval for the MRL function, it is necessary to maximize the likelihood function (2.8) under the specific constraint H_0 defined in (2.10), in which $\eta_{x_0}(s, \theta_0)$ is equal to (2.13). Hence, for the length-biased sample $X_1, X_2, \ldots, X_n$ the following estimation equation might be applied for calculating the maximum likelihood estimating practically.

$$H_0 : \quad \frac{1}{n} \sum_{i=1}^{n} \eta_{x_0}(X_i, \theta_0) = 0. \tag{2.14}$$

where

$$\eta_{x_0}(X_i, \theta_0) = \left(1 - \frac{x_0 + \theta_0}{X_i}\right) I_{[x_0, X_{(n)})}(X_i),$$

for $i = 1, \ldots n$, in which $X_{(n)}$ is the maximum of the underlying sample.

Considering the discussion presented in Section 2.3, we shall construct our inference based on the nonparametric likelihood function for the distribution function $F_{X'}(\cdot)$. According

to Theorem 2.2.1, based on the random length-biased sample $X_1, X_2, \ldots, X_n$, the classical empirical distribution of the observation is the nonparametric maximum likelihood estimator for $G(\cdot)$, but not for $F_{X'}(\cdot)$. For the likelihood function $L(G')$, suppose that the distribution $G'(\cdot)$ places a probability mass $p_i = G'(X_i) - G'(X_i^-) \geq 0$ on the value of $X_i \in \mathbb{R}$. Then, $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ is a probability vector which assign the probability p_i to the value of $X_i \in \mathbb{R}$, and therefore $\eta_{x_0}(X_i, \theta_0)$, where $\sum_{i=1}^n p_i = 1$ and $L(G') = \prod_{i=1}^n p_i$. Thus, the likelihood ratio function of $G'(\cdot)$ is a function of $\boldsymbol{p}$, and we have:

$$\begin{aligned} R(G') &= \frac{L(G')}{L(G_n)} \\ &= \left\{ \prod_{i=1}^n np_i : \sum_{i=1}^n p_i = 1, p_i \geq 0,\ i = 1, 2, \ldots, n \right\}. \end{aligned} \tag{2.15}$$

As mentioned in Section 2.3, the distribution $G'(\cdot)$ in (2.15) is a member of the set of all distribution in $\mathcal{R}$, indicated by $\mathcal{F}$ in the previous sections. However, we use a smaller set of distributions by applying the constraint (2.14), since we are interested in drawing statistical inference for the parameter θ_0 here. Accordingly, define the profile likelihood ratio as follows.

$$\begin{aligned} \mathcal{R}(\theta_0) &= \sup \left\{ R(G') \mid T(G') = \theta_0, G' \in \mathcal{F} \right\} \\ &= \sup \left\{ \prod_{i=1}^n np_i : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \eta_{x_0}(X_i, \theta_0) = 0,\ p_i \geq 0,\ i = 1, 2, \ldots, n \right\} \end{aligned} \tag{2.16}$$

Given (2.14), we can deduce that the second equality in the above likelihood ratio to satisfy the condition $T(G') = \theta_0$. To maximize the likelihood function under the extra constraint H_0, the Lagrange multiplier method can be applied. Thus, it is obtained that the above likelihood function attains its maximum at $p_i = \left[n \left(1 + \lambda_{x_0} \eta_{x_0}(X_i, \theta_0) \right) \right]^{-1}$, in which λ_{x_0} is the solution of the following equation.

$$\frac{1}{n} \sum_{i=1}^n \frac{\eta_{x_0}(X_i, \theta_0)}{1 + \lambda_{x_0} \eta_{x_0}(X_i, \theta_0)} = 0.$$

Consequently, the likelihood ratio function (2.16) for the parameter θ_0 is concluded to be

$$\mathcal{R}(\theta_0) = \prod_{i=1}^n \left(1 + \lambda_{x_0} \eta_{x_0}(X_i, \theta_0) \right)^{-1}.$$

Bear in mind that we have used subscript x_0 for λ_{x_0} because it depends solely on $\eta_{x_0}(\cdot, \cdot)$, which is defined for, and is a function of x_0.

Therefore, the corresponding empirical log-likelihood ratio is defined as

$$\begin{aligned} l(\theta_0) &:= -2 \log \mathcal{R}(\theta_0) \\ &= 2 \sum_{i=1}^n \log \left\{ 1 + \lambda_{x_0} \eta_{x_0}(X_i, \theta_0) \right\}. \end{aligned} \tag{2.17}$$

The following theorem presented by Fakoor et al. (2018) indicates the asymptotic distribution of the empirical log-likelihood ratio.

THEOREM 2.4.1 *Assume that* $E\,(X)^{-2} < \infty$. *Then, for all* $x_0 \in [0, \tau)$, *the limiting distribution of* $l(\theta_0)$ *is a chi-square distribution with* 1 *degree of freedom. That is,*

$$l(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_{(1)}.$$

Proof. See Fakoor et al. (2018).

This limiting distribution may be used to obtain the following EL ratio confidence interval for θ_0. Therefore, an asymptotic $100(1-\alpha)\%$ confidence interval for the MRL function $\theta_0 = M(x_0)$ at a fixed time x_0, when $x_0 \in [0, \tau)$, may be obtained using the following relation:

$$C_1(x_0) = \left\{\theta(x_0) = \theta_0 : l\,(\theta_0) \leq \chi^2_{1,\alpha}\right\},$$

where $\chi^2_{1,\alpha}$ is the upper α-quantile of the distribution $\chi^2_{(1)}$.

2.5 Simulation

A simulation study was carried out to evaluate and check the performance of the EL method proposed. The code for the simulation presented was written in *R* by the author and may be found in Appendix II. The performance of the method was inspected through computing the coverage probability and the lengths of the EL-ratio confidence intervals. We used the small sample size 50, the moderate sample size 100, and the large sample size 200 to illustrate the results of the method. In each scenario, the data was generated from the related length-biased distribution function. The resulting lengths of confidence intervals and coverage probabilities was calculated for various scenarios based on the 5000 iterations. In each repetition, while the data are simulated from a length-biased distribution, the confidence interval is estimated for the corresponding target population of interest. The coverage probabilities were computed as the ratio of the number of confidence intervals covering the real value of the mean residual lifetime function for the unbiased distribution out of 5000, the total number of repetitions. Similarly, the lengths of the confidence intervals (Δ) were calculated as the average of the lengths of intervals estimated for 5000 data sets simulated. Also, we have considered two separate nominal levels $(1-\alpha)$, which are 0.95 and 0.90, for estimating the confidence intervals.

Having generated the simulated data according to the above scenarios, we obtained the empirical likelihood ratio confidence intervals via $C_1(x_0)$. For better illustration and comparison, one can apply the weak convergence resulted in Fakoor (2015) to obtain another confidence interval for the MRL function in length-bias setting through the normal approximation method. According to Fakoor (2015), under the condition that $E\,[X^p] < \infty$ $(p > 2)$ for the unconditional NPMLE of the MRL function defined in (2.6) we have

$$\sqrt{n}\left(M_n(x_0) - M(x_0)\right) \xrightarrow{\mathcal{L}} N(0, \sigma_0^2), \quad \left(x_0 \in [0, \tau)\right)$$

where $\sigma_0^2 = \sigma^2(x_0)$ is the asymptotic variance of the above empirical process.

However, the relation presented for the variance σ_0^2 takes a sophisticated form and could not be easily estimated. Alternatively, a resampling method may be applied for estimating the variance of the stochastic process $\sqrt{n}\left(M_n(\cdot) - M(\cdot)\right)$ at each point $x_0 \in [0, \tau]$. Thus, we applied the following bootstrap procedure to estimate the variance σ_0^2. For a fixed $X_1, \ldots, X_n$ random variables from the distribution function (2.1), let $X_1^*, \ldots, X_n^*$ denote random variables

from the distribution function $G_n(\cdot)$, defined in (2.4). It is worth mentioning that $G_n(\cdot)$ is the empirical distribution function of the length-biased sample $X_1, \ldots, X_n$. Also, similar to (2.4), we obtain the empirical distribution $G_n^*(\cdot)$ for the sample $X_1^*, \ldots, X_n^*$. Now, define

$$\begin{aligned} F_n^*(x) &= \mu_n^* \int_0^x s^{-1} dG_n^*(s) \\ &= \frac{\mu_n^*}{n} \sum_{i=1}^{n} \frac{1}{X_i^*} I(X_i^* \leq x). \end{aligned}$$

where

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

Consequently, the NPMLE of the MRL function based on the length-biased resamples, that is the bootstrap alternative to (2.6), is as follows

$$M_n^*(x) := \frac{I_{[0, X_{(n)}^*)}(x)}{1 - F_n^*(x)} \int_x^\infty (1 - F_n^*(s)) ds.$$

Define the stochastic process $\delta^*(\cdot) := \sqrt{n} \left(M_n^*(\cdot) - M(\cdot) \right)$. Suppose that we have iterated the above bootstrap procedure B times.The bootstrap estimator of the variance $\sigma_0^2 = \sigma^2(x_0)$ is the sample variance of the stochastic process $\delta_1^*(x_0), \ldots, \delta_B^*(x_0)$, which may be computed using the following equation.

$$\sigma_{boot}^{*2} := \frac{1}{B-1} \sum_{i=1}^{B} \left(\delta_i^*(x_0) - \frac{1}{B} \sum_{j=1}^{B} \delta_j^*(x_0) \right)^2$$

Here, to obtain the variance σ_0^2, we have considered $B = 500$ in each iteration. The code for the normal approximation simulation presented was also written in *R* and could be found in Appendix II.

Table 2.1 compares the performance of the EL confidence interval with that of the normal approximation method for the target population $Weibull(2, 2)$. It can be checked that the underlying population fulfills the requirement of Theorem 2.4.1. It was observed that the values of CP for the EL method was moderately larger that those for the normal approximation method except for $x_0 = 3.7$ and $n = 50$ the reason for which will be discussed. Moreover, both methods have almost preserved the nominal levels, albeit the EL method have exhibited superiority. Furthermore, the values of Δ for the confidence intervals declined tremendously as the sample sized increased in both methods. Additionally, the EL-based confidence intervals has slightly decreased the lengths of the confidence intervals in comparison with the normal approximation method. There was not observed any noticeable difference between two separate values of $(1 - \alpha)$.

The other significant advantage of the empirical likelihood method is that it is considerably more efficient in terms of time and calculation. The normal approximation method is more computationally demanding and time consuming. In addition, over the course of estimating the confidence intervals, we pointed out that the natural procedure of the bootstrap estimator

Table 2.1: 90 % and 95 % Confidence Intervals for MRL of Weibull(2,2)

Time	Sample	$1-\alpha = 90\%$				$1-\alpha = 95\%$			
		EL		NA		EL		NA	
x_0	n	Δ	C.P.	Δ	C.P.	Δ	C.P.	Δ	C.P.
	50	0.443	0.879	0.443	0.873	0.529	0.941	0.529	0.934
0.5	100	0.317	0.897	0.318	0.892	0.380	0.945	0.379	0.942
	200	0.227	0.899	0.226	0.895	0.270	0.950	0.270	0.947
	50	0.377	0.898	0.379	0.891	0.450	0.948	0.452	0.942
1.0	100	0.268	0.897	0.268	0.894	0.319	0.949	0.320	0.945
	200	0.190	0.905	0.191	0.903	0.227	0.950	0.227	0.945
	50	0.352	0.901	0.354	0.897	0.420	0.947	0.422	0.941
1.5	100	0.250	0.901	0.251	0.901	0.297	0.948	0.298	0.945
	200	0.176	0.900	0.177	0.897	0.210	0.946	0.210	0.944
	50	0.357	0.889	0.360	0.886	0.430	0.945	0.432	0.438
2.0	100	0.254	0.895	0.255	0.891	0.303	0.950	0.304	0.948
	200	0.180	0.899	0.180	0.897	0.215	0.948	0.215	0.945
	50	0.395	0.888	0.413	0.886	0.474	0.935	0.481	0.928
2.5	100	0.280	0.897	0.282	0.893	0.336	0.950	0.337	0.941
	200	0.199	0.895	0.199	0.893	0.238	0.952	0.238	0.947
	50	0.458	0.850	0.689	0.881	0.547	0.921	0.853	0.951
3.0	100	0.337	0.882	0.343	0.875	0.401	0.935	0.410	0.925
	200	0.238	0.900	0.239	0.895	0.285	0.947	0.285	0.938

does not present a valid result for large values of x_0 (here $x_0 = 3$ and sometime $x_0 = 2.5$) when the sample size is small ($n = 50$). Because, the MRL function $M_n^*(\cdot)$ is defined on $[0, X_{(n)}^*)$, and therefore, $\delta^*(\cdot)$ is defined on $[0, X_{(n)}^*)$ as well. However, it is very unlikely for the large amounts of x_0 that $X_{(n)}^* > x_0$ is satisfied for all the generated bootstrap sample of the distribution function $G_n(\cdot)$, when $G_n(\cdot)$ is calculated based on the sample size $n = 50$. Indeed, that we have considered x_0 up to roughly 90%-quantile. Accordingly, in order to obtain $B = 500$ valid observations $\delta_1^*(x_0), \ldots, \delta_B^*(x_0)$ for large values of x_0, we had to generate higher number of bootstrap samples. Therefore, only those samples for which $X_{(n)}^* > x_0$ could be used for the calculation of σ_{boot}^{*2}. This issue, for small sample scenario ($n = 50$) and large values of x_0, leads to an illusive dramatic increase in the variance σ_{boot}^{*2}, which increases the lengths of the normal approximation confidence intervals hugely, while growing the values of CP moderately. This issue also makes the normal approximation method computationally more demanding. It is worth mentioning that Fakoor et al. (2018) applied the EL confidence intervals for some different Uniform and Gamma family distributions and it was observed that the EL method even increased the coverage probabilities more than what we observed for Weibull distribution here.

We estimated the El-based confidence intervals for various amounts of x_0 using length-biased observations corresponding to the target distribution $Weibull(2, 2)$. Figure 2.1 shows the performance of the average 95% confidence intervals of the large sample scenario ($n = 200$) based on 5000 number of repetitions. The true MRL curve for the unbiased distribution function $Weibull(2, 2)$ is plotted simultaneously for better illustration. The EL method was observed to preserve the nominal level very accurately for all amounts of x_0. Additionally, it was revealed that the lengths of EL confidence intervals for the MRL function of target

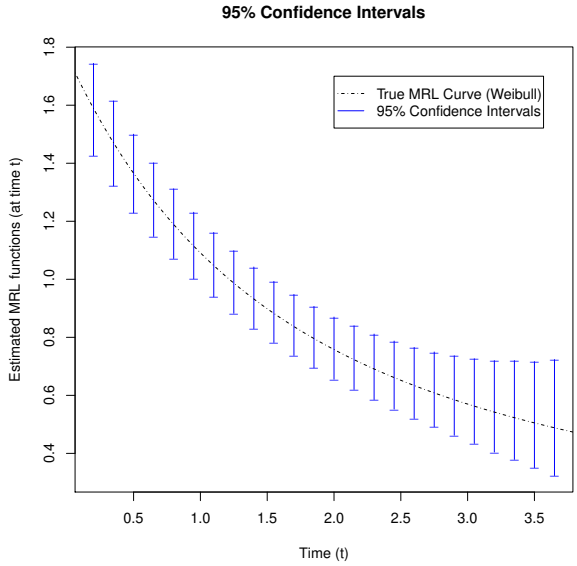

Figure 2.1: Weibull(2, 2) MRL Function with 95% Confidence Intervals

distribution $Weibull(2, 2)$ grows noticeably as the value of x_0 increases.

It was observed that, under the condition of Theorem 2.4.1, not only the EL procedure results in an accurate confidence interval that preserve the nominal levels even for small sample sizes, but also it exhibits superior coverage probability and shorter length of interval in comparison to the normal approximation method. However, one may be interested in investigating whether the proposed method is robust when the condition of Theorem 2.4.1 is ignored–although this condition is not unrealistic or overly restrictive.

Table 2.2 illustrates the simulation results for the target distribution $Weibull(0.5, 2)$. It can be easily checked that length-biased observations corresponding to the unbiased distribution $Weibull(0.5, 2)$ does not meet the requirement of Theorem 2.4.1. It can be obtained that the EL method has improved lengths of Δ moderately although the values of coverage probabilities were quite similar in two methods. This issue indicates more accuracy of the EL method. As expected, the confidence intervals widths decreased by increasing the sample sizes. We had the same problem as in Table 2.1 here for estimating the variance of the stochastic process $\sqrt{n}\left(M_n(\cdot) - M(\cdot)\right)$ for $x_0 = 8.5$ and $x_0 = 10.5$. However, the justification given for the slight reduction in CP of the point $x_0 = 3$ and sample size $n = 50$ in Table 2.1 applies here as well. There was not any clear differences for simulation results in the two separate nominal levels. Comparing the results in Table 2.2 and those in Table 2.1, the most interesting result is that both empirical likelihood ratio and and normal approximation method are robust when the Theorem 2.4.1 condition is ignored.

Figure 2.2 illustrates the confidence intervals of the MRL function with respect to the unbiased d.f. $Weibull(0.5, 2)$. The confidence intervals were calculated based on the 5000 iterations of the large sample scenario (200) using the empirical likelihood ration method. In each repetition, the length-biased observations were generated and the point-wise confidence interval was estimated for the target population. The nominal level of 95% was considered to estimate the intervals. For better illustration, the true MRL curve of the distribution function $Weibull(0.5, 2)$ has been plotted as well. Again the coverage probabilities of the confidence intervals out of 5000 repetitions were absolutely close to the nominal level, which is not reported here due to space limitation. The confidence intervals widened as the value of x_0

Table 2.2: 90 % and 95 % Confidence Intervals for MRL of Weibull(0.5,2)

Time	Sample	$1-\alpha = 90\%$				$1-\alpha = 95\%$			
		EL		NA		EL		NA	
x_0	n	Δ	C.P.	Δ	C.P.	Δ	C.P.	Δ	C.P.
	50	4.476	0.843	4.892	0.843	5.184	0.883	5.648	0.900
0.5	100	3.338	0.882	3.488	0.893	3.901	0.931	4.078	0.932
	200	2.390	0.886	2.443	0.887	2.844	0.941	2.903	0.938
	50	5.101	0.890	5.317	0.899	6.077	0.946	6.320	0.952
2.5	100	3.596	0.898	3.668	0.900	4.264	0.954	4.346	0.953
	200	2.529	0.898	2.553	0.895	3.017	0.953	3.044	0.953
	50	5.798	0.888	6.002	0.897	6.882	0.945	7.100	0.942
4.5	100	4.075	0.900	4.141	0.902	4.882	0.943	4.952	0.947
	200	2.877	0.900	2.898	0.904	3.433	0.950	3.455	0.951
	50	6.651	0.889	6.863	0.889	7.494	0.949	7.870	0.942
6.5	100	4.589	0.897	4.658	0.903	5.497	0.951	5.559	0.950
	200	3.254	0.896	3.276	0.895	3.874	0.950	3.894	0.949
	50	7.498	0.898	7.891	0.896	8.252	0.948	8.513	0.946
8.5	100	5.149	0.899	5.224	0.901	6.188	0.946	6.260	0.948
	200	3.636	0.895	3.660	0.896	4.329	0.947	4.351	0.948
	50	8.197	0.898	10.117	0.899	9.271	0.947	11.054	9.952
10.5	100	5.761	0.900	5.848	0.910	7.134	0.948	7.223	0.962
	200	4.020	0.900	4.047	0.904	4.794	0.951	4.818	0.951

rose. It is of note that the MRL function for $Weibull(0.5, 2)$ is an increasing function. The results of Figure 2.2 indicated once again the robustness of the proposed EL method with respect to underlying condition (see Theorem 2.4.1).

2.6 Real Data Application

In this section, we apply the proposed method for the construction of confidence interval using a set of real data. We used a real collection of data presented in Muttlak and McDonald (1990). The shrubs data was collected using the line-intercept sampling procedure as a part of a biological study. The shrub data set comprises the widths of shrubs for three different intersect. In each sampling time, they were only able to obtain three separate widths of shrubs. The sampling procedure was subject to the circumstances similar to those discussed in Section 1.6 due to truncation. Muttlak and McDonald (1990) figured out that the chance of observing a shrub in their sample was proportional to the width of that subject. Thus, the wider shrubs had more chances of being included in the sample. The data set includes three different transects: 18 complete widths in transect I, 22 complete observations in transect II, and those in transect III were equal to 6. The data was initially analyzed by Muttlak and McDonald (1990); after which several studies investigated the difference in widths of shrubs observed in three transects. According to Wang (1996) and Ning et al. (2010), there exist statistically significant differences between widths of shrubs in transect I and transect II, and also between transect I and transect III. However, both researches using separate methods indicated that there were not any statistically meaningful difference between widths of shrubs in transect II and transect III.

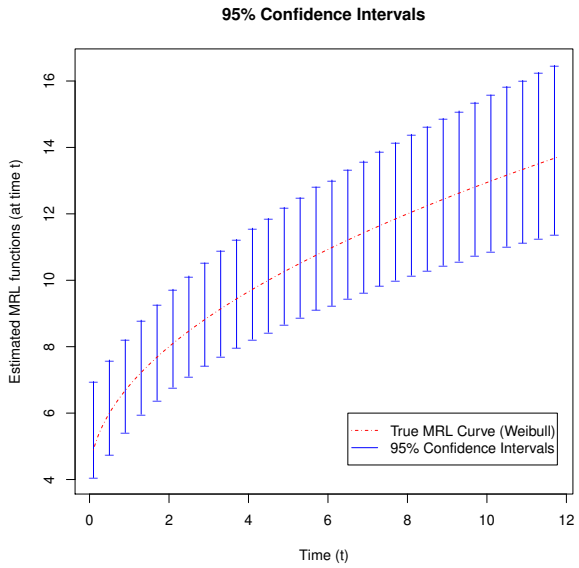

Figure 2.2: Weibull$(0.5, 2)$ MRL Function with 95% Confidence Intervals

Consequently, we pooled the widths of shrubs in transect I and II together to construct empirical likelihood-based confidence interval. The pooled sample includes 28 complete widths of shrubs. The range of the data varies from 0.35 to 2.54. We have considered separate values for $x_0 \in [0.2, 2.2]$ to estimate the point-wise confidence intervals of the MRL function. Figure 2.3 illustrates the results of the 95% confidence intervals for the MRL function using the empirical likelihood ratio method. For better illustration, the point estimation of the mean residual lifetime function was plotted simultaneously using (2.6).

It can be observed that both point and interval estimations of the MRL function have indicated that the average remaining widths of shrubs should decreases gradually as their widths increase over time. For instance, we are 95% confident that a shrub with a width equal to 0.5 should grow until it becomes thicker by approximately $\theta_0 \in [0.4, 0.8]$ unites. Whereas, the 95% EL-based confidence interval for a shrub with a width equal to 2 indicated that on average it ought to grow until it become something between 0.15 and 0.52 units thicker.

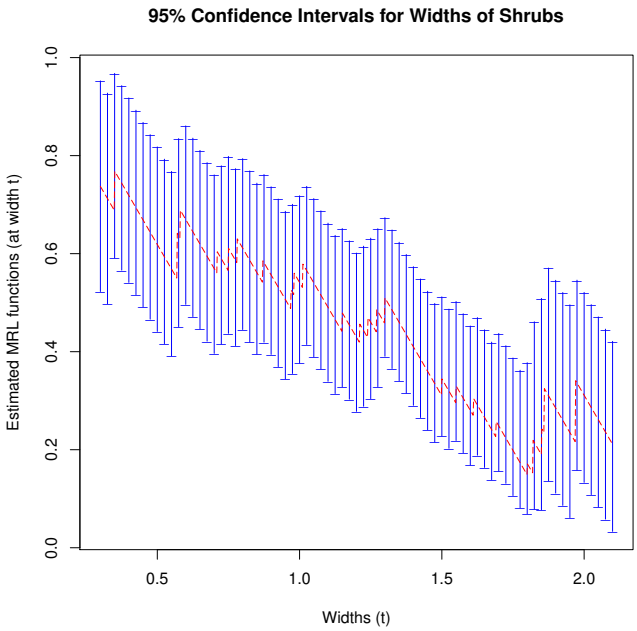

Figure 2.3: Estimated MRL for Widths of Shrubs with 95% Confidence Intervals

3

Confidence Interval Based on Length-biased and Right-censored Data

3.1 Introduction

As discussed in Section 1.6, the data collected in cross-sectional sampling with follow-up is commonly known as a cohort of prevalent cases in biosciences (e.g. epidemiology, biology or biomedical sciences). In such sampling procedures only the subjects that have already experienced an initiating event but are yet to obtain the terminating event at the enrollment time, are recruited. It has been mentioned that, on average, the prevalent subjects observed in such studies are confirmed to possess a longer lifetime owing to non-random selection of subjects from the target population. This means our sample does not represent the behavior of the population of interest. Instead, it consists of a non-representative (biased) sample of the target population. The most common case of this bias, discussed in Chapter 1 and 2, is called length-bias, occurring when the so-called stationarity assumption is satisfied.

In Chapter 2, we discussed some motivating problems which result in observing length-biased data. To emphasis the importance of the study of length-biased data, we also presented several real examples in which such data have arisen. Additionally, we briefly discussed in sections 1.2, 1.3, 1.4 and 1.6 how important it is to study residual lifetime in survival analysis. Accordingly, we defined the mean residual lifetime function to model the survival data and we reviewed different studies concerning the MRL function in sections 1.2.2 and 1.8.

Prospective prevalent cohort studies are conducted to assess the history of a disease among cases that have already experienced the initiating event (e.g. onset of disease). The recruited subjects are then followed-up over time until they experience the ending event. However, it is often impossible to follow-up on some of the recruited subjects (for variety of reasons some of which has been reviewed in Chapter 1) which leads to censored data. The obstacle posed by censored data is frequently associated with the study of survival data. The analysis of survival data is very complicated in the presence of censoring, where all we know is that the terminating event for a subject has occurred in a certain period of time, but not the exact time. The feasible censoring schemes in a prevalent cohort study with follow-up are random right-censoring and *type I* censoring. However, this issue was missed in Chapter 2, which is

why, over the course of this chapter, we are adapting the statistical tools presented in Chapter 2 to analyze the length-biased data in the presence of right-censoring.

To emphasize the application of length-biased and right-censored data, we present a real example here, but many more applications may be find in the literature. Wolfson et al. (2001) conducted a study in epidemiology to explore the survival time among patients with probable or possible dementia, in which the patients whose age were at least 65 years being recruited. The enrolled subjects were then screened over time for dementia as the terminating event. The dates of death or censoring, whichever happened first, were prospectively recorded for all recruited cases. In this survey, the data collected was proven to be biased as the time intervals recorded (from onset of dementia up to death) tended to be longer for the subjects observed in the cohort in comparison to subjects in the target population.

In Section 1.8 we undertook a comprehensive review on recent articles in regard to inference on a population parameter using the empirical likelihood. In particular, we reviewed all the existing literature concerning nonparametric inference about the mean residual lifetime function in the presence of biased sampling. It was observed that utilizing the empirical likelihood method, one can draw statistical inference on a family of population parameters, say θ, in general. This is one of the most important advantages of the empirical likelihood method that we are able to make nonparametric inference (e.g. hypothesis testing, constructing confidence interval) for a parameter generally, sometimes even without specifying a relation for the parameter. Fakoor et al. (2018) revealed that the empirical likelihood method may be adapted for the MRL function as well. However, the possibility of observing censored data is not considered in this paper.

There are some articles in which the problem of making inference for an unknown parameter θ via the empirical likelihood method in the presence of censorship has been discussed. However, there has not been any solution for the problem of estimating the margin of errors of commonly used summary statistics when our sample consist of length-biased right-censored data until Ning et al. (2013). They have presented an empirical likelihood-based method for constructing confidence intervals for summary statistics as well as the survival function at a fixed point x_0. By considering the mean residual function at any arbitrary but fixed time like x_0, one can deal with the MRL function ($M(x_0)$) as a parameter (θ) in the empirical likelihood method. Considering the importance of right-censoring and length-bias in survival analysis, the motivating question here is how to draw statistical inference about residual lifetime function using length-biased right-censored data. This issue along with the advantages of empirical likelihood in statistical inference inspired us to carry out this chapter.

3.2 Preliminaries and NPMLE

Given the motivation problem, it was discussed that a prevalent sampling survey is generally more practical and efficient than using incident sampling. Accordingly, we consider a sampling mechanism similar to Chapter 2, however, sometimes we are not able to follow some of the subjects during the study due to a variety of reasons. This issue leads to observing incomplete data (censored data) in our sample, arising when all we know is that the terminating event for an individual has occurred in a certain period of time. Right-censoring is the only feasible censoring scheme in cross-sectional sampling with follow-up.

Assuming the possibility of censoring in our sample, which was ignored in the data setting discussed in Section 2.2, the following notation is needed. All of the variables below relate

to backward recurrence times, forward recurrence times, total lifespans and censoring times. Let X' denote the total lifetime of interest. Associated with each X' we observe two separate random variables, namely T' and C', representing the truncation variable and the censoring time, respectively. Therefore, we observe a triple random variable (X', T', C') with regard to each of the subjects in our sample. Similar to the notation in Chapter 2, let $F_{X'}(\cdot) = P(X' \leq \cdot)$ and $f_{X'}(\cdot)$, defined on $\mathbb{R}^+ = [0, \infty)$, denote the respective cumulative distribution and density functions of the lifetime variable X'. As mentioned in Section 2.2, in prevalent cohort studies, only those subjects whose $X' \geq T'$ are observable. Suppose that we denote the observation subject to the condition $X' \geq T'$ by X, regardless of whether it is censored or not. Similar to Chapter 2, assuming the stationarity assumption for the incidence rate holds, the density of the observation is $g(\cdot) = f_{X'|X'>T'}\left(\cdot|X' \geq T'\right)$ defined in the equation (2.1).

In other words, X is the observed lifespan under left truncation. That is the obtained length of life X' subject to the condition $X' \geq T'$. The random variables T and C associated with X are defined similarly as well. Thus, in the presence of right-censoring, we have the following triple related to the lifetime X in cross-sectional sampling with follow-up,

$$(A, R \wedge C, \delta),$$

in which A, R and C, are the backward recurrence time, the forward recurrence time and the residual censoring time for the individual obtained. Also, δ is the censoring indicator variable, which is $\delta = I(R \leq C)$. It is worth mentioning that the residual censoring time is the censoring time remaining from the recruitment for a subject observed in cross-sectional sampling. It is often reasonable to assume X is independent of (T, C), when $P(C' \geq T') = 1$ (See Wang (1991)). Moreover, it is a reasonable assumption in most real situations to consider C is independent of (A, R).

Since our sample is consists of both censored and uncensored observations, denote the observed subject X by Y if it is uncensored, and by Z once it is censored. That means once the underlying lifetime variable X' satisfies the condition $X' \geq T'$, we indicate the length of time either with Y, if it is a complete data, or with Z, when it is censored. Accordingly, using the recurrence times, it is easy deduced that $Y = A + R$ and $Z = A + C$. Bear in mind that although C is independent of (A, R), the complete lifetime observed Y, and the censoring time observed Z are not independent. To show this issue,

$$\begin{aligned} Cov(Y, Z) &= E\left[(A + R)(A + C)\right] - E\left[A + R\right] E\left[A + C\right] \\ &= Var\left(A\right) + Cov\left(A, R\right) \\ &= Var\left(A\right)\left[1 + Corr(A, R)\sqrt{(Var(R)/Var(A))}\right]. \end{aligned}$$

It is clear that $Corr(A, R) \neq 0$. Also, when the initiating event satisfies a stationary Poisson process, it implies A is conditionally $U(0, X)$, given X. Consequently, $Var(A) = Var(X-A) = Var(R)$. Accordingly, it is obvious that, excluding for trivial situations, the above equation indicates that $Cov(Y, Z) > 0$ which apparently implies that the censoring mechanism in a prevalent of cohort study is informative. This fact leads the censoring time to not be independent of the lifetime of interest in a prevalent cohort study (cross-sectional sampling) with follow-up, which is why the Kaplan-Meier estimator is not the unconditional NPMLE for the length-biased distribution here. In other words, owing to the bias induced by the sampling mechanism, the subject's failure time must be longer than its truncation time ($X' \geq T'$), that is, be observable (X), in order to be censored.

Let $A_1, \ldots, A_n$ and $R_1, \ldots, R_n$ denote, respectively, backward recurrence times (the current ages) and the corresponding forward recurrence times for the random sample observed in

a prevalent cohort study, which is $X_1, \ldots, X_n$. Associated with each pair of (A_i, R_i), for $i = 1, \ldots, n$, there is C_i which is the residual censoring time, and therefore δ_i, the censoring indicator. Accordingly, we obtain a triple like

$$(A_i, R_i \wedge C_i, \delta_i), \quad i = 1, 2, \ldots n.$$

in related with each subject observed in cross-sectional sampling with follow-up. Considering the above explanation of the sampling procedure, it can be deduced that the vectors $(A_i, R_i \wedge C_i, \delta_i)$ for $i = 1, 2, \ldots, n$ are independent.

We discussed in Section 1.3, 1.6 and 2.2 that the density function and the cumulative distribution function of the truncated observations, say the respective $g(\cdot)$ and $G(\cdot)$, under stationarity assumption imply the equation 2.1 and 2.2, respectively. Under the data setting discussed in Chapter 2, we obtained that the classical empirical distribution of the observations is the NPMLE for the distribution function $G(\cdot)$. However, it is not possible to use the classical empirical counterpart distribution here, since our sample consists of censored observations as well. Furthermore, as discussed, the Kaplan-Meier estimator may not be used due to the informative mechanism of censoring here.

Vardi (1989) derived a nonparametric maximum likelihood estimator for the common lifespan distribution function of the n independent and identically distributed stationary renewal processes. It has been proven that prevalent cohort cases under the stationarity assumption and observations of n independent and identically distributed stationary renewal processes started a long time ago have exactly the same probabilistic characteristics (see e.g. Cox (1969, 1962)). To study the NPMLE for renewal processes that commenced along time ago, Vardi (1989) initially proposed a multiplicative censorship model for which he deduced the NPMLE. Then, Vardi (1989) proved that the multiplicative censorship model proposed may generalize several interesting and important statistical problems such as estimating in renewal processes, nonparametric deconvolution of an exponential random variable, and estimating a distribution function under a decreasing density constraint. Having presented the common NPMLE for these models by Vardi (1989), Vardi and Zhang (1992) studied the asymptotic behavior of the nonparametric maximum likelihood estimator proposed under multiplicative censoring. However, they realized that, despite the same maximum likelihood estimates of the common likelihood function for all these models, the asymptotic behavior of the NPMLE depends specifically on the characteristics of the model defined. Hence, the asymptotic behavior of the NPMLE using the data obtained from a prevalent cohort study with follow-up was studied by Asgharian and Wolfson (2005).

In this section, we will present and discuss the NPMLE of the length-biased right-censored survival data for the underlying model in this thesis. Vardi (1989) commenced their study by proving the NPMLE for the multiplicative censoring. However, we use a more straightforward strategy for a cohort of prevalent cases (or, alternatively, independent and identically distributed stationary renewal processes). We will also discuss briefly why the asymptotic behavior of the subjects' lifetime in a prevalent cohort differ from the multiplicative censoring introduced by Vardi (1989).

Imagine that the number of subjects observed at recruitment time is equal to n. Unfortunately, only a percentage of these n subjects experience the failure during the survey; while a proportion of subjects survive until the end of the study, we lose the chance to follow-up on the remaining subjects. It is a reasonable assumption that the remaining lifetimes of the recruited subjects are randomly censored which is why we lose the chance to follow-up on some of the subjects. In other words, the forward recurrence times of the recruited subjects are subject to random right-censoring. Also, those subjects that survive until the end of the

study experience *type I* censoring. Nonetheless, from now onwards we will treat both the random censored subjects and *type I* censoring as the censored subjects since there is not any difference between this two groups in the underlying model. It is reasonably assumed that the number of censored and uncensored subjects are random. This issue is the principal difference between a cohort of prevalent cases with follow-up and the Vardi's multiplicative censorship model. Because he assumed a fixed number of subjects to be censored while it is random in our setting. Thus, let N_1 denote the random number of uncensored observations, then $N_2 = n - N_1$ is the number of censored subjects.

To complete the rest of analysis, we need the joint distribution of the backward and the forward recurrence times, which has been proven in the study of renewal processes initially. However, considering the same probabilistic characteristics of a sample collected using cross-sectional sampling under the stationarity assumption, this joint distribution can be used for modeling the current age and the residual lifetime of the subjects collected (See e.g. Vardi (1989) or Gilbert et al. (1999)). The joint distribution is given by,

$$F_{A,R}(a,r) = \frac{F_{X'}(a+r)}{\mu_{X'}}.$$

Denote $\mathcal{U}$ and C (also C), respectively, for uncensored and censored in this thesis. Then $G_{\mathcal{U}}(t) = P(A + R \leq t|\delta = 1)$. Then, by assuming $p = P(\delta = 1) = P(R \leq C)$ and $S_C(r) = 1 - F_C(r) = 1 - P(C \leq r)$, we have

$$\begin{aligned} G_{\mathcal{U}}(y) &= \int_0^y \int_0^s \frac{1}{p} f_{A,R}(s-r,r) S_C(r) dr ds \\ &= \frac{1}{p\mu_{X'}} \int_0^y f_{X'}(s) \int_0^s S_C(r) dr ds \\ &= \frac{1}{p} \int_0^y \frac{g(s)}{s} \int_0^s S_C(r) dr ds. \end{aligned} \tag{3.1}$$

If $g_{\mathcal{U}}(t)$ is the density function corresponding to $G_{\mathcal{U}}(\cdot)$, then it is obtained from the above equation that

$$g_{\mathcal{U}}(y) = \frac{g(y)}{py} \int_0^y S_C(r) dr. \tag{3.2}$$

Let $G_C(t) = P(A + C \leq t|\delta = 0)$, then we have

$$\begin{aligned} G_C(z) &= \frac{1}{1-p} \int_0^z \int_0^s \int_c^\infty f_{A,R}(s-c,r) dr dF_C(c) ds \\ &= \frac{1}{1-p} \int_0^z \int_0^s \int_c^\infty \frac{f_{X'}(s+r-c)}{\mu_{X'}} dr dF_C(c) ds \\ &= \frac{1}{1-p} \int_0^z \int_0^s \int_s^\infty \frac{f_{X'}(u)}{\mu_{X'}} du dF_C(c) ds \\ &= \frac{1}{1-p} \int_0^z \frac{1 - F_{X'}(s)}{\mu_{X'}} \int_0^s dF_C(c) ds \\ &= \frac{1}{1-p} \int_0^z \frac{1 - F_{X'}(s)}{\mu_{X'}} F_C(s) ds. \end{aligned} \tag{3.3}$$

It is worth mentioning that $\left(1 - F_{X'}(\cdot)\right)/\mu_{X'}$, which appeared in the last equation, is the common and well known probability density function of the backward and forward recurrence

times (see Cox (1962)). Let $f_r(\cdot)$ denote the recurrence density function, in which the subscript r stands for the recurrence times, it can be deduced that

$$\begin{aligned} f_r(x) &= \frac{1}{\mu_{X'}}\left(1 - F_{X'}(x)\right) \\ &= \int_x^{\infty} s^{-1} dG(s). \end{aligned}$$

Accordingly, the cumulative distribution function (3.3) could be rewritten as

$$G_C(z) = \frac{1}{1-p}\int_0^z f_r(s)F_C(s)ds. \tag{3.4}$$

As a consequence, for the density function corresponding to the distribution function $G_C(t)$ we have

$$\begin{aligned} g_C(z) &= \frac{F_C(z)\left(1 - F_{X'}(s)\right)}{\mu_{X'}(1-p)} \\ &= \frac{f_r(z)F_C(z)}{1-p} \\ &= \frac{F_C(z)}{1-p}\int_z^{\infty} s^{-1} dG(s). \end{aligned} \tag{3.5}$$

Given Theorem 2.2.1, it can be easily deduced that the classical NPMLE of $F_{X'}(\cdot)$ is the empirical distribution function using independent and identically distributed random variables sampled from the population of interest. However, we do not have access to i.i.d. data in a cohort of prevalent cases with follow-up due to non-random sampling. Since the observable sample under uniform left-truncation in a prevalent cohort study is $X_1, \ldots, X_n$, it satisfies the distribution function (2.2), but not $F_{X'}(\cdot)$. Thus, we can employ the sample observed to draw inference on the likelihood function of the distribution $G(\cdot)$. Then, having reached the nonparametric maximum likelihood estimation of $G(\cdot)$, we can estimate the target distribution $F_{X'}(\cdot)$ by means of (2.2).

For this purpose, let n_1 and n_2 be the values realized for the random variables N_1 (the random number of uncensored observations) and N_2 (the random number of censored observations), respectively. Consequently, given n_1 and n_2, it is obvious that $n_1 + n_2 = n$ and our sample $X_1, \ldots, X_n$ consists of $Y_1, \ldots, Y_{n_1}$ and $Z_1, \ldots, Z_{n_2}$. Suppose that $y_1, \ldots, y_{n_1}$ and $z_1, \ldots, z_{n_2}$ denote the uncensored and censored realized values for the random variables $Y_1, \ldots, Y_{n_1}$. $Z_1, \ldots, Z_{n_2}$, respectively. Then, the full likelihood based on the information of the total lifespans observed is as follows.

$$\begin{aligned} \mathcal{L} &= \left(\prod_{i=1}^{n_1} dG_{\mathcal{U}}(y_i)\right)\left(\prod_{j=1}^{n_2} dG_C(z_j)\right) \\ &= \left(\prod_{i=1}^{n_1} \frac{dG(y_i)}{pt}\int_0^{y_i} S_C(r)dr\right)\left(\prod_{j=1}^{n_2} \frac{F_C(z_j)}{1-p}\int_{z_j}^{\infty} s^{-1}dG(s)\right) \\ &\propto \left(\prod_{i=1}^{n_1} dG(y_i)\right)\left(\prod_{j=1}^{n_2}\int_{z_j}^{\infty} s^{-1}dG(s)\right). \end{aligned} \tag{3.6}$$

To obtain the NPMLE, all we need to do is to find a distribution function $G(\cdot)$ in the set $\mathcal{F}$, defined in Section 2.3, that maximize (3.6). For this purpose, it is apparent that $G(\cdot)$

should be a kind of distribution which assign probability masses to the set of all observations $\{(y_1, \ldots, y_{n_1}) \cup (z_1, \ldots, z_{n_2})\}$. But, if $G(\cdot)$ is a kind of distribution function assigning probability masses to any set beyond the sampled observations $\{(y_1, \ldots, y_{n_1}) \cup (z_1, \ldots, z_{n_2})\}$, then, by moving the mass(es) to the nearest observation(s) on the left side, we obtain another distribution function possessing higher likelihood. Furthermore, if $G(\cdot)$ is a distribution that assigns any probability mass to the left side of the smallest observation, adding this mass to the smallest observation results in growing the likelihood. We obtain this increase in the likelihood, since the final integral in (3.6) for different values of z_i is right-tailed with a limit from left. Thus, it is concluded that all we need to do, in order to find the NPMLE, is restrict our attention to the problem of finding a discrete distribution which assigns probability masses to the mentioned set and maximizes the likelihood function (3.6).

For this aim, suppose that $t_1, \ldots, t_m$ are the ordinal values of the set of all of observations $\{(y_1, \ldots, y_{n_1}) \cup (z_1, \ldots, z_{n_2})\}$ so that $0 < t_1 < t_2 < \cdots < t_m$. It is of note that if the distribution function $F_{X'}(\cdot)$ and as a direct consequence $G(\cdot)$ are completely continuous, we should not observe any tie in our sample and therefore theoretically $m = n_1 + n_2 = n$ with probability 1. However, it is common in practice to observe ties in our real data sets due to possibility of discrete distribution of the underlying survival time, which is why we have used m for the number of observations instead of $n = n_1 + n_2$ indicating the possibility of multiplicity. It is apparent that $m \leq n_1 + n_2$ is always true. Let ξ_j and ζ_j for $j = 1, \ldots, m$ denote the multiplicity of uncensored (Y values) and censored (Z values) data, respectively. Therefore,

$$\xi_j = \sum_{i=1}^{n_1} I(y_i = t_j) \quad \text{and} \quad \zeta_j = \sum_{i=1}^{n_2} I(z_i = t_j), \quad (j = 1, \ldots, m).$$

Given the above definition, the problem of maximizing likelihood function (3.6) is equivalent to the problem of maximizing the following likelihood.

$$L(G) := \prod_{i=1}^{m} p_i^{\xi_i} \left(\sum_{j=i}^{m} \frac{1}{t_j} p_j \right)^{\zeta_i}, \tag{3.7}$$

in which p_i indicates the weight that the NPMLE of the distribution $G(\cdot)$ assigns to the point t_i, for $j = 1, \ldots, m$. In other words, $p_i := p(t_i) = G(dt_i)$. Accordingly, $\boldsymbol{p} := (p_1, \ldots, p_m)$ must satisfy the following conditions:

$$p_i \geq 0 \quad (j = 1, \ldots, m), \quad \sum_{i=1}^{m} p_i = 1. \tag{3.8}$$

Considering the form of (3.7), it is hard to maximize the likelihood analytically as it does not have a closed form. However, following studies such as Vardi (1989) we can employ some optimization method to obtain the point $\hat{\boldsymbol{p}}$ at which the likelihood function (3.7) hits its maximum.

For this purpose, we consider $\{(Y_1, \ldots, Y_{n_1}) \cup (Z_1, \ldots, Z_{n_2})\}$ as the set of data consisting of incomplete and complete data for which we observe the pooled sample $T_1 < T_2 < \cdots < T_m$. Moreover, we think of $\{(Y_1, \ldots, Y_{n_1}) \cup (Y_{21}, \ldots, Y_{2n_2})\} = \left\{\cup_{i=1}^{n} X_i\right\}$ as the set of complete data which is not observable due to censoring of the subjects $Y_{21}, \ldots, Y_{2n_2}$. Similar to Section 2.2, if observing a complete sample was possible, the nonparametric maximum likelihood estimator of the distribution function $G(\cdot)$ would be simply the empirical counterpart of the data set

$(X_1, \ldots, X_n)$. However, the NPMLE of the distribution function $G(\cdot)$ here is a nonparametric distribution function that assigns the weights p_i $(i = 1, \ldots, m)$ to the combined sample $(T_1, \ldots, T_m)$ using only the information provided by the sample $\{(Y_1, \ldots, Y_{n_1}) \cup (Z_1, \ldots, Z_{n_2})\}$. Consequently, the EM algorithm is an invaluable tool by considering different aspects of the problem. Indeed, the EM algorithm takes more advantages in comparison to the other existing methods when the underlying problem is to deal with incomplete data.

Since we have considered the possibility of observing ties in our data, suppose that for the complete sample $X_1, \ldots, X_n$ we have observed $x_1, \ldots, x_n$ and the corresponding ordinal values $t_1, \ldots, t_m$. Let ϑ_i $(i = 1, \ldots, m)$ denote the number of multiplicity corresponding to the sample value t_i,

$$\vartheta_i = \sum_{j=1}^{n} I(x_j = t_i) \quad (i = 1, \ldots, m).$$

Then, recalling the vector of probability masses assigned by distribution $G(\cdot)$ to the observation points t_i $(i = 1, \ldots, m)$, we can rewrite the likelihood function (3.7) based on the sample consists of all complete data as follows,

$$L'(G) := \prod_{i=1}^{m} p_i^{\vartheta_i}. \tag{3.9}$$

This likelihood has a closed form and the maximum point could be observed through Lagrange multiplier method. Given the likelihood function (3.9), we have the following Lagrangian for the log-likelihood function (3.9) under the constraint (3.8).

$$\mathcal{L}(G, \lambda_0) \;:=\; \sum_{i=1}^{m} \vartheta_i \log\left(p_i\right) + \lambda_0 \left(\sum_{i=1}^{m} p_i - 1 \right) \tag{3.10}$$

By considering the Lagrangian (3.10), the following EM algorithm may be applied to obtain the NPMLE of $G(\cdot)$ using the sample consisting of incomplete and complete data $\{(y_1, \ldots, y_{n_1}) \cup (z_1, \ldots, z_{n_2})\}$. Given an initial arbitrary probability mass vector, denote $\boldsymbol{p}^0 = (p_1^0, \ldots, p_n^0)$, satisfying the conditions (3.8), we can calculate the conditional expectation vector of the multiplicities of observations at observation points t_i $(i = 1, \ldots, m)$ indicated by $\boldsymbol{\vartheta}^1 = (\vartheta_1^1, \ldots, \vartheta_n^1)$. Having obtained $\boldsymbol{\vartheta}^1$, we can maximize the Lagrangian (3.10) reaching $\boldsymbol{p}^1 = (p_1^1, \ldots, p_n^1)$. For the remaining repetitions of the EM algorithm , all we need to do in each iteration is substitute the new estimated weight vector $\boldsymbol{p}^l$ for the previous probability mass vector $\boldsymbol{p}^{l-1}$ and calculate the new conditional expectation, say $\boldsymbol{\vartheta}^{l+1}$, based on the new weights $(\boldsymbol{p}^l)$, and then calculating $\boldsymbol{p}^{l+1}$ by means of $\boldsymbol{\vartheta}^{l+1}$ $(l = 1, 2, \ldots)$.

Considering the underlying sampling setting, we present below the computational phases for the EM algorithm in order to estimate the NPMLE $G(\cdot)$.

Expectation Step: For any arbitrary $i \in (1, \ldots, n_1)$, $j \in (1, \ldots, n_1)$ and $k \in (1, \ldots, m)$, we have

$$\begin{aligned}
E_{p^{l-1}}\left\{I(Y_i = t_k) \;+\; I(Y_{2j} = t_k) \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2})\right\} & \\
= P\left\{Y_i = t_k \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2}),\, p^{l-1}\right\} & \\
+ P\left\{Y_{2j} = t_k \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2}),\, p^{l-1}\right\} & \\
= \frac{P\left\{Y_i = t_k\,,\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2}),\, p^{l-1}\right\}}{P\left\{(y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2}),\, p^{l-1}\right\}} &
\end{aligned}$$

$$
\begin{aligned}
&+\frac{P\{Y_{2j}=t_k\,,\,(y_1,\,\ldots\,,\,y_{n_1},\,z_1,\,\ldots\,,\,z_{n_2}),\,p^{l-1}\}}{P\{\,(y_1,\,\ldots\,,\,y_{n_1},\,z_1,\,\ldots\,,\,z_{n_2}),\,p^{l-1}\}}\\
&= \frac{P\{Y_i=t_k\,,\,Y_i=y_i,\,p^{l-1}\}}{P\left\{\,Y_i=y_i,\,p^{l-1}\right\}}+\frac{P\{Y_{2j}=t_k\,,\,Z_j=z_j,\,p^{l-1}\}}{P\left\{\,Z_j=z_j,\,p^{l-1}\right\}}\\
&= I(y_i=t_k)+\frac{\int P\left\{A_{2j}=a\,,\,Y_{2j}=t_k\,,\,Z_j=z_j,\,p^{l-1}\,|\,\delta=0\right\}da}{P\left\{\,Z_j=z_j,\,p^{l-1}\right\}}\\
&= I(y_i=t_k)+\frac{\int P\left\{A_{2j}=a\,,\,R_{2j}=t_k-a\,,\,C_j=z_j-a,\,p^{l-1},\,\delta=0\right\}da}{(1-p)P\left\{\,Z_j=z_j,\,p^{l-1}\right\}}\\
&= I(y_i=t_k)+\frac{\int_0^{z_j}P\left\{A_{2j}=a\,,\,R_{2j}=t_k-a\,,\,p^{l-1}\right\}P\left\{C_j=z_j-a,\,p^{l-1}\right\}da}{(1-p)\tilde{g}_C(t_r)}\\
&= I(y_i=t_k)+\frac{\int_0^{z_j}\tilde{f}_{A,R}(a,t_k-a)f_C(z_j-a)da}{F_C(z_j)\int_{x\geq z_j}x^{-1}\,d\tilde{G}(x)}\\
&= I(y_i=t_k)+\frac{\tilde{g}(t_k)\int_0^{t_r}f_C(a')da'}{t_kF_C(t_r)\int_{x\geq t_r}x^{-1}\,d\tilde{G}(x)}\\
&= I(y_i=t_k)+\frac{p_k^{l-1}}{t_k\sum_{s=r}^m\frac{p_s^{l-1}}{t_s}}\\
&= I(y_i=t_k)+\frac{p_k^{l-1}}{t_k}\left\{\frac{I(z_j=t_1)}{\sum_{s=1}^m\frac{p_s^{l-1}}{t_s}}+\frac{I(z_j=t_2)}{\sum_{s=2}^m\frac{p_s^{l-1}}{t_s}}+\cdots+\frac{I(z_j=t_k)}{\sum_{s=k}^m\frac{p_s^{l-1}}{t_s}}\right\}. \qquad (3.11)
\end{aligned}
$$

Now, we can obtain the following expectation by applying the most recent estimation of probability masses vector, say $\boldsymbol{p}^{l-1}$. Given (3.11), for $l=1,2,\ldots$ we have

$$
\begin{aligned}
\vartheta_k^l &:= E_{p^{l-1}}\left(\vartheta_k\,|\,(y_1,\,\ldots\,,\,y_{n_1},\,z_1,\,\ldots\,,\,z_{n_2})\right)\\
&= E_{p^{l-1}}\left(\sum_{i=1}^n I(X_i=t_k)\,|\,(y_1,\,\ldots\,,\,y_{n_1},\,z_1,\,\ldots\,,\,z_{n_2})\right)\\
&= E_{p^{l-1}}\left\{\sum_{i=1}^{n_1}I(Y_i=t_k)+\sum_{j=1}^{n_2}(Y_{2j}=t_k)\,|\,(y_1,\,\ldots\,,\,y_{n_1},\,z_1,\,\ldots\,,\,z_{n_2})\right\}\\
&= \sum_{i=1}^{n_1}I(y_i=t_k)+\frac{p_k^{l-1}}{t_k}\left\{\frac{\sum_{j=1}^{n_2}I(z_j=t_1)}{\sum_{s=1}^m\frac{p_s^{l-1}}{t_s}}+\frac{\sum_{j=1}^{n_2}I(z_j=t_2)}{\sum_{s=2}^m\frac{p_s^{l-1}}{t_s}}\right.\\
&\qquad\left.+\cdots+\frac{\sum_{j=1}^{n_2}I(z_j=t_k)}{\sum_{s=k}^m\frac{p_s^{l-1}}{t_s}}\right\}
\end{aligned}
$$

$$= \xi_k + \frac{p_k^{l-1}}{t_k}\left\{\frac{\zeta_1}{\sum_{s=1}^{m}\frac{p_s^{l-1}}{t_s}} + \frac{\zeta_2}{\sum_{s=2}^{m}\frac{p_s^{l-1}}{t_s}} + \cdots + \frac{\zeta_k}{\sum_{s=k}^{m}\frac{p_s^{l-1}}{t_s}}\right\}$$

$$= \xi_k + \frac{p_k^{l-1}}{t_k}\sum_{j=1}^{k}\left\{\frac{\zeta_j}{\sum_{s=j}^{m}\frac{p_s^{l-1}}{t_s}}\right\}.$$

Maximization Step: In this step, by using the vector $\boldsymbol{\vartheta}^l$, we can obtain a new probability vector, say $\boldsymbol{p}^l$, which maximize the Lagrangian (3.10).

$$\begin{aligned}\boldsymbol{p}^l &= \left(p_1^l, p_2^l, \ldots, p_m^l\right)\\ &= \left(\frac{\vartheta_1^l}{\sum_{i=1}^{m}\vartheta_i^l}, \frac{\vartheta_2^l}{\sum_{i=1}^{m}\vartheta_i^l}, \ldots, \frac{\vartheta_m^l}{\sum_{i=1}^{m}\vartheta_i^l}\right)\\ &= \frac{1}{n}(\vartheta_1^l, \ldots, \vartheta_m^l).\end{aligned}$$

The last relation in the above equation is satisfied because the summation of all multiplicities ϑ_i $(i = 1, \ldots, m)$ must be equal to the sample size, $\sum_{i=1}^{m}\vartheta_i = n$.

It was expected that the maximization phase would lead to probability masses ϑ_i/n $(i = 1, \ldots, n)$ because (3.10) is the likelihood function corresponding to the empirical counterpart distribution of the sample of complete data $x_1, \ldots, x_n$ by considering the possibility of observing ties in our observation. It is of note that the probability reached in this step is not only to maximize the Lagrangian (3.10) and therefore the likelihood function (3.9), but also the likelihood function (3.7).

Properties of the EM algorithm

1 There exists a unique $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ which maximize the likelihood function (3.7).

2 The likelihood function (3.7) grows in each repetition of the EM algorithm.

3 The algorithm converges to $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$, the unique maximizer of the likelihood function (3.7).

4 Suppose that

$$\hat{G}_n(x) := \sum_{i=1}^{m}\hat{p}_i I_{[0,x]}(t_i) \tag{3.12}$$

be the nonparametric maximum likelihood estimator of $G(\cdot)$. Then, $\hat{G}_n(\cdot)$ is a consistent estimate of $G(\cdot)$ for $x > 0$.

Proof. For proof of these properties see Vardi (1989) for an analogous argument.

Efron (1967) initially introduced a similar algorithm for another problem, after which Vardi (1989) proposed another similar EM algorithm which was deduced for the multiplicative censoring initially and it is validity for the underlying model, however, has been noted. Here, we presented a more straight forward method for a prevalent cohort study with follow-up.

COROLLARY 3.2.1 *If the value realized for the random variable N_2 is equal to zero, indicating the number of the censored data is equal to zero ($n_2 = 0$), then the likelihood function (3.6) leads to the same problem discussed in Chapter 2. That means the NPMLE obtained here using the above EM algorithm will be equal to the classical empirical counterpart of the observations, (2.4).*

NPMLE: More informative structure

It is of note that we have not used the information about the current age and the residual lifetime for the likelihood function (3.6). Instead, we have used the total lifetime obtained for censored and uncensored subjects by applying distribution functions (3.1) and (3.4). However, the information of the backward and forward recurrence times are basically accessible in the sampling mechanism of interest in this chapter. Accordingly, the aim of this section is to use the more informative structure of the data to obtain the likelihood function. It will be observed that both procedures lead to the same maximization problem which is a function of $G(\cdot)$. This means that there is a unique maximum estimator for both likelihood functions.

For this purpose, let A_i and R_i, for $i = 1, \ldots, n_1$, denote the backward recurrence time and the forward recurrence time of the random variables $Y_1, \ldots, Y_{n_1}$ such that $Y_i = A_i + R_i$, for $i = 1, \ldots, n_1$. Also, suppose that A_j and C_j ($j = 1, \ldots, n_2$) denote the backward recurrence time and the residual censoring time in regard to the random variables $Z_1, \ldots, Z_{n_2}$, and therefore $Z_j = A_j + C_j$ ($j = 1, \ldots, n_2$). Furthermore, suppose that a_i (a_j) and r_i (r_j) are the values realized for A_i (A_j) and R_i (R_j), respectively. As a consequence, it is apparent that $y_i = a_i + r_i$ and $z_j = a_j + c_j$, when $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Then,

$$
\begin{aligned}
\mathcal{L} &= \left(\prod_{i=1}^{n_1} f_{A,R}(A_i = a_i, R_i = r_i|\delta_i = 1)\right)\left(\prod_{j=1}^{n_2} f_{A,C}(A_j = a_j, C_j = c_j|\delta_j = 0)\right) \\
&= \left(\prod_{i=1}^{n_1} \frac{1}{p} S_C(r_i) f_{A,R}(a_i, r_i)\right)\left(\prod_{j=1}^{n_2} \frac{1}{1-p} f_C(c_j) \int_{c_j}^{\infty} f_{A,R}(a_j, r) dr\right) \\
&= \left(\prod_{i=1}^{n_1} \frac{1}{p} S_C(r_i) \frac{f_{X'}(a_i + r_i)}{\mu_{X'}}\right)\left(\prod_{j=1}^{n_2} \frac{f_C(c_j)}{1-p} \int_{c_j}^{\infty} \frac{f_{X'}(a_j + r)}{\mu_{X'}} dr\right) \\
&= \left(\prod_{i=1}^{n_1} \frac{S_C(r_i)}{p y_i} dG(y_i)\right)\left(\prod_{j=1}^{n_2} \frac{f_C(c_j)}{1-p} \int_{z_j}^{\infty} s^{-1} dG(s)\right) \\
&\propto \left(\prod_{i=1}^{n_1} dG(y_i)\right)\left(\prod_{j=1}^{n_2} \int_{z_j}^{\infty} s^{-1} dG(s)\right). \qquad (3.13)
\end{aligned}
$$

It can be observed that the likelihood functions (3.6) and (3.13) have led to the same maximizing problem which results in the same unique NPMLE.

NPMLE: Random number of random censoring subjects

Bear in mind that in our analysis, without loss of generality, we have obtained the likelihood functions (3.13), (3.6) by conditioning on observing $N_1 = n_1$ and $N_2 = n_2$. However, as we have insisted previously, N_1 and N_2 are random and not fixed variables in the underlying model. Our conditional approach does not have any impact on the NPMLE proposed here,

albeit N_1 and N_2 must be considered as random variables in the asymptotic study of the proposed NPMLE of $G(\cdot)$ since it leads to different limiting properties.

Here we present the likelihood for the random number of censoring to indicate that our conditional procedure is not affecting the result for the NPMLE.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left(f_{A,R}(A_i = a_i, R_i = r_i)\right)^{\delta_i} \left(P(A_j = a_j, c_j < R_j)\right)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\frac{f_{X'}(y_i)}{\mu_{X'}}\right)^{\delta_i} \left(\int_{c_j}^{\infty} f_{A,R}(a_j, r)dr\right)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\frac{G(y_i)}{y_i}\right)^{\delta_i} \left(\int_{z_j}^{\infty} s^{-1} dG(s)\right)^{1-\delta_i} \\
&\propto \prod_{i=1}^{n} \left(G(y_i)\right)^{\delta_i} \left(\int_{z_j}^{\infty} s^{-1} dG(s)\right)^{1-\delta_i} . \qquad (3.14)
\end{aligned}$$

It is apparent that the likelihood functions (3.6), (3.13) and (3.13) result in the same maximizing problem, and consequently a unique NPMLE.

3.3 Empirical Likelihood for Length-biased Right-censored Data

In Section 2.3, we reviewed the basic principle of constructing a confidence interval for a parameter. We discussed a method for constructing a confidence interval for length-biased data using empirical likelihood in Section 2.4. Similar to the previous chapter, let $T(\cdot)$ denote a function $T(\cdot) : \mathcal{F} \to \mathbb{R}$. As mentioned in Chapter 2, we are precisely interested in testing hypothesis or constructing a confidence interval when the EL constraint is $T(F_{X'}) = M(x)$, in which $M(x)$ is the MRL function defined in (1.1). However, while Chapter 2 presents statistical inference on the MRL function using the length-biased survival data, we consider here the possibility of observing censored data in a length-biased sampling procedure.

If an i.i.d. sample from the target distribution had been observable, we would have been able to obtain an NPMLE for the MRL function by substituting $F_{X'}(\cdot)$ with the classical empirical distribution function in $T(F_{X'})$. However, the available data in cross-sectional sampling does not consist of i.i.d. copies of the target population because of non-random sampling of subjects. Furthermore, the method presented in Section 2.4 is not applicable here as our sample is comprised of complete and incomplete data.

However, our statistical inference may be based on the NPMLE presented in Section 3.2. Having assumed the stationarity assumption for incidence rate, all the equations presented from (2.1) to (2.13) are valid here as well. Therefore, we need to maximize the likelihood function (3.7) under the constraint H_0 defined in (2.10) by substituting $\eta_{x_0}(s, \theta_0)$ from (2.13) . Hence, given (2.8), for the length-biased right-censored sample of survival data $\left\{\cup_{i=1}^{n} X_i\right\} = \left\{(Y_1, \ldots, Y_{n_1}) \cup (Z_1, \ldots, Z_{n_2})\right\}$ defined in Section 3.2, we are looking for a distribution $G \in \mathcal{F}$ that satisfies the following estimation equation,

$$H_0 : \quad \int \eta_{x_0}(s, \theta_0) dG(s) = 0, \qquad (3.15)$$

where $\eta_{x_0}(X_i, \theta_0)$ is defined by (2.13). As mentioned in Section 2.4, although $\eta_{x_0}(s, \theta_0)$ is a function of s and θ_0, it is defined and affected by x_0. Nonetheless, x_0 was assumed to be a fixed point in Section 2.4. Consequently, we have different function $\eta_{x_0}(s, \theta_0)$ for various points x_0, which is why we have used x_0 as a subscript of $\eta_{x_0}(s, \theta_0)$ for more emphasis.

Inspired by the discussion given in Section 2.3, we may make statistical inference on $T(F)$ by constructing an alternative likelihood ratio to (2.7). The method presented in Chapter 2 is not applicable here since the possibility of observing censored data is missed. However, the discussion presented in Section 3.2, and therefore the NPMLE $\hat{G}_n$ obtained in (3.12) can be applied alternatively for constructing the likelihood ratio.

Consider the sample variables $T_1 < T_2 < \cdots < T_m$ corresponding to the set of all observations $\{(Y_1, \ldots, Y_{n_1}) \cup (Z_1, \ldots, Z_{n_2})\}$. For the likelihood function $L(\tilde{G})$, we shall suppose that $\tilde{G}(\cdot)$ is a distribution which assign a probability masses w_i $(i = 1, \ldots, m)$ to the random variable $T_i \in \left\{\cup_{i=1}^m T_i\right\}$. Then, $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ is a probability vector assigned by $\tilde{G}(\cdot)$ to the random variables $T_i \in \mathbb{R}$, and therefore $\eta_{x_0}(T_i, \theta_0)$ $(i = 1, \ldots, m)$. Thus, $w_i = w(T_i) = d\tilde{G}(T_i)$ $(i = 1, \ldots, m)$, where $\sum_{i=1}^m w_i = 1$. Following relation (3.7), it is apparent that for $L(\boldsymbol{w})$ we have

$$L(\tilde{G}) = \prod_{i=1}^{m} w_i^{\xi_i} \left(\sum_{j=i}^{m} \frac{1}{t_j} w_j \right)^{\zeta_i}. \tag{3.16}$$

Note that we have considered the possibility of observing multiplicity in our sample ($m \leq n = n_1 + n_2$) for deriving distribution function $\tilde{G}(\cdot)$. The likelihood ratio function $\tilde{G}(\cdot)$, which is a function of $\boldsymbol{w}$, has the representation

$$\begin{aligned} R(\tilde{G}) &= \frac{L(\tilde{G})}{L(\hat{G}_n)} \\ &= \left\{ \prod_{i=1}^{m} \left(\frac{w_i}{\hat{p}_i} \right)^{\xi_i} \left(\frac{\sum_{j=i}^m \frac{1}{t_j} w_j}{\sum_{j=i}^m \frac{1}{t_j} \hat{p}_j} \right)^{\zeta_i} : \sum_{i=1}^{m} w_i = 1,\ w_i \geq 0,\ i = 1, \ldots, m \right\}. \end{aligned} \tag{3.17}$$

The last equation in (3.17) is obtained by means of (3.7) and (3.16).

As mentioned in Section 2.3, the distribution $\tilde{G}(\cdot)$ in (3.17) is a member of the set of all distributions on $\mathbb{R}$, indicated by $\mathcal{F}$. However, since we are interested in drawing statistical inference for the parameter θ_0, we use a smaller set of distributions by applying the constraint H_0. Therefore, given (3.15), the following estimation equation may be used in order to find the NPMLE under the constraint H_0,

$$H_0 : \quad \sum_{i=1}^{m} \eta_{x_0}(T_i, \theta_0) w_i = 0. \tag{3.18}$$

where

$$\eta_{x_0}(T_i, \theta_0) = \left(1 - \frac{x_0 + \theta_0}{T_i} \right) I_{[x_0, X_{(n)})}(T_i),$$

for $i = 1, \ldots m$, in which $X_{(n)} = T_m$ is the maximum of the underlying sample and $\theta_0 = \theta(x_0) = M(x_0)$, where $M(x_0)$ is the value of the MRL function (2.11) at point x_0.

Based on the ordinal random sample $T_1, \ldots, T_m$, (3.18) is the alternative constraint to (3.15). Therefore, we can form the profile likelihood ratio for the length-biased right-censored sample by applying the constraint (3.18) as follows.

$$\begin{aligned}\mathcal{R}(\theta_0) &= \sup\left\{R(\tilde{G}) \mid T(\tilde{G}) = \theta_0, \tilde{G} \in \mathcal{F}\right\} \\ &= \sup\left\{\prod_{i=1}^{m}\left(\frac{w_i}{\hat{p}_i}\right)^{\xi_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}w_j\right)^{\zeta_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)^{-\zeta_i} : \sum_{i=1}^{m} w_i = 1,\right. \\ &\quad \left.\sum_{i=1}^{m} w_i\eta_{x_0}(T_i, \theta_0) = 0, w_i \geq 0, i = 1, \ldots, m\right\}. \end{aligned} \tag{3.19}$$

It is not easy to maximize the above likelihood ratio analytically since it does not have a closed form. Therefore, we can employ a suitable optimization method to find the maximizer of the likelihood ratio $\mathcal{R}(\theta_0)$. Inspired by discussion resulting in (3.9), suppose that $X_1, \ldots, X_n$ is the complete sample, a proportion of which is not observable owing to random censoring. Instead, let $t_1, \ldots, t_m$ denote the ordinal values we obtained for the length-biased right-censored sample $\left\{(Y_1, \ldots, Y_{n_1}) \cup (Z_1, \ldots, Z_{n_2})\right\}$. Suppose that ρ_i, $(i = 1, \ldots, m)$ denotes the number of ties for the sample value t_i,

$$\rho_i = \sum_{j=1}^{n} I(x_j = t_i) \quad (i = 1, \ldots, m).$$

Consequently,

$$\begin{aligned}\mathcal{R}(\theta_0) &= \sup\left\{\prod_{i=1}^{m} w_i^{\rho_i}\hat{p}_i^{-\xi_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)^{-\zeta_i} : \sum_{i=1}^{m} w_i = 1,\right. \\ &\quad \left.\sum_{i=1}^{m} w_i\eta_{x_0}(T_i, \theta_0) = 0, w_i \geq 0, i = 1, \ldots, m\right\} \end{aligned} \tag{3.20}$$

Thus, the log-likelihood ratio profile under the constraint (3.18) is equal to

$$\begin{aligned} l(\theta_0) &= -2\log\mathcal{R}(\theta_0) \\ &= 2\log\left\{\prod_{i=1}^{m}\hat{p}_i^{\xi_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)^{\zeta_i}\right\} \\ &\quad -2\sup\left\{\log\prod_{i=1}^{m} w_i^{\rho_i} : \sum_{i=1}^{m} w_i = 1, \sum_{i=1}^{m} w_i\eta_{x_0}(T_i, \theta_0) = 0, w_i \geq 0, i = 1, \ldots, m\right\} \\ &= 2\sum_{i=1}^{m}\left[\xi_i\log\left(\hat{p}_i\right) + \zeta_i\log\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)\right] - 2\sup\left\{\sum_{i=1}^{m}\rho_i\log\left(w_i\right) : \sum_{i=1}^{m} w_i = 1,\right. \\ &\quad \left.\sum_{i=1}^{m} w_i\eta_{x_0}(T_i, \theta_0) = 0, w_i \geq 0, i = 1, \ldots, m\right\} \end{aligned} \tag{3.21}$$

Considering the form of equation (3.20) and (3.21), the Lagrange multiplier method can be used to maximize the likelihood ratio profile under the extra constraint H_0. So, the Lagrangian is defined by

$$\begin{aligned}\mathcal{L}(\tilde{G}, \lambda_{x_0}, \lambda) &:= \sum_{i=1}^{m} \rho_i \log(w_i) + \lambda_{x_0}\left(\sum_{i=1}^{m} w_i \eta_{x_0}(T_i, \theta_0)\right) \\ &\quad + \lambda\left(\sum_{i=1}^{m} w_i - 1\right) + c, \end{aligned} \tag{3.22}$$

where c is a constant.

Adopting the method proposed to obtain the maximizer of the Lagrangian (3.10), we can apply a constrained EM algorithm to reach the NPMLE of the likelihood ratio function (3.19) under the restriction (3.18). To start the algorithm, select an arbitrary weight vector $\boldsymbol{w}^0 = (w_1^0, \ldots, w_n^0)$, which should satisfy the conditions $\sum_{i=1}^{m} w_i = 1$ and $w_i \geq 0\,(i = 1, \ldots, m)$. Then, calculate the conditional expectation of the number of multiplicity of observations at points t_i $(i = 1, \ldots, m)$ that is $\boldsymbol{\rho}^1 = (\rho_1^1, \ldots, \rho_m^1)$. Having calculated $\boldsymbol{\rho}^1$, maximize the Lagrangian (3.22), obtaining $\boldsymbol{w}^1 = (w_1^1, \ldots, w_m^1)$. Thereafter, the lth iteration estimates the conditional expectation $\boldsymbol{\rho}^l$ by means of the last weight vector $\boldsymbol{w}^{l-1}$ and will terminates by maximizing the Lagrangian, reaching the new weight vector $\boldsymbol{w}^l$.

We introduce below the computational procedure of the EM algorithm for obtaining the NPMLE of $G(\cdot)$ under the constraint 3.18.

Expectation Step: In this phase, we can obtain the following expectation by applying the estimation of probability weight vector presented in the previous step, say $\boldsymbol{p}^{l-1}$. Given (3.11), we possess

$$\begin{aligned}\rho_k^l &:= E_{w^{l-1}}\left(\rho_k \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2})\right) \\ &= E_{w^{l-1}}\left(\sum_{i=1}^{n} I(X_i = t_k) \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2})\right) \\ &= E_{w^{l-1}}\left\{\sum_{i=1}^{n_1} I(Y_i = t_k) + \sum_{j=1}^{n_2} I(Y_{2j} = t_k) \,|\, (y_1, \ldots, y_{n_1}, z_1, \ldots, z_{n_2})\right\} \\ &= \xi_k + \frac{w_k^{l-1}}{t_k}\left\{\frac{\zeta_1}{\sum_{s=1}^{m} \frac{w_s^{l-1}}{t_s}} + \frac{\zeta_2}{\sum_{s=2}^{m} \frac{w_s^{l-1}}{t_s}} + \cdots + \frac{\zeta_k}{\sum_{s=k}^{m} \frac{w_s^{l-1}}{t_s}}\right\} \\ &= \xi_k + \frac{w_k^{l-1}}{t_k}\sum_{j=1}^{k}\left\{\frac{\zeta_j}{\sum_{s=j}^{m} \frac{w_s^{l-1}}{t_s}}\right\}, \end{aligned}$$

where $k = 1, \ldots, m$.

Maximization Step: In this step, we can maximize the Lagrangian (3.22) by substituting the $\boldsymbol{\rho}^l$ estimated in above for $\boldsymbol{\rho}$. Thus,

$$w_k^l = \frac{\rho_k}{\sum_{i=1}^{m} \rho_i - \lambda_{x_0}\eta_{x_0}(T_i, \theta_0)}, \quad (k = 1, \ldots, m), \tag{3.23}$$

where λ_{x_0} is the solution of the following equation.

$$\sum_{i=1}^{n} \frac{\rho_i \eta_{x_0}(T_i, \theta_0)}{\sum_{i=1}^{m} \rho_i - \lambda_{x_0} \eta_{x_0}(T_i, \theta_0)} = 0.$$

Therefore, we can conclude the following new weight vector, say $\boldsymbol{w}^l$, which maximize the likelihood function (3.20) based on the $\boldsymbol{\rho}^l$.

$$\begin{aligned} \boldsymbol{w}^l &= \left(w_1^l, w_2^l, \ldots, w_m^l\right) \\ &= \left(\frac{\rho_1}{n - \lambda_{x_0}\eta_{x_0}(T_i, \theta_0)}, \frac{\rho_2}{n - \lambda_{x_0}\eta_{x_0}(T_i, \theta_0)}, \ldots, \frac{\rho_m}{n - \lambda_{x_0}\eta_{x_0}(T_i, \theta_0)}\right). \end{aligned}$$

The last equation holds because the sum of all multiplicities ρ_i $(i = 1, \ldots, m)$ must be equal to n, $\sum_{i=1}^{m} \rho_i = n$.

Properties of the EM algorithm

1 There is a unique $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_m)$ which maximize the likelihood function (3.16).

2 The likelihood function (3.16) climbs by increasing the number of iterations of the EM algorithm.

3 The algorithm converges to the unique maximizer of the likelihood function (3.16), say $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_m)$.

4 Suppose that

$$\tilde{G}_n(x) := \sum_{i=1}^{m} \hat{w}_i I_{[0,x]}(t_i)$$

denote the nonparametric maximum likelihood estimator of $G(\cdot)$ under the constraint (3.18) using length-biased right-censored data. Then, $\tilde{G}_n(x)$ is a consistent estimate of the distribution function $G(x)$ for $x > 0$ under the constraint H_0 defined in (3.15). In other words, $\tilde{G}_n(\cdot)$ is a consistent estimate of $\tilde{G}(\cdot) \in \mathcal{F}$ that achieves the supremum of (3.19).

Proof. For the proof of these properties see Ning et al. (2013) for an analogous argument.

Thus, by plugging in the weigh vector $\hat{\boldsymbol{w}}$ in (3.19) we can obtain the following equation for the likelihood ratio profile of the parameter θ_0.

$$\mathcal{R}(\theta_0) = \prod_{i=1}^{m} \left(\frac{\hat{w}_i}{\hat{p}_i}\right)^{\xi_i} \left(\sum_{j=i}^{m} \frac{1}{t_j} \hat{w}_j\right)^{\zeta_i} \left(\sum_{j=i}^{m} \frac{1}{t_j} \hat{p}_j\right)^{-\zeta_i}. \tag{3.24}$$

Accordingly, the empirical log-likelihood ratio corresponding to (3.24) is equal to

$$\begin{aligned} l(\theta_0) &= -2 \log \mathcal{R}(\theta_0) \\ &= 2 \sum_{i=1}^{m} \left[\xi_i \log\left(\hat{p}_i\right) + \zeta_i \log\left(\sum_{j=i}^{m} \frac{1}{t_j} \hat{p}_j\right)\right] \end{aligned}$$

$$-2\sum_{i=1}^{m}\left[\xi_i \log\left(\hat{w}_i\right) + \zeta_i \log\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{w}_j\right)\right]. \tag{3.25}$$

In the following theorem, we have studied the limiting distribution of the empirical log-likelihood ratio.

THEOREM 3.3.1 *Assume that* $E\left(X^{-2}\right) < \infty$. *For all* $x_0 \in [0, \tau)$, *the limiting distribution of* $l(\theta_0)$ *is a chi-square distribution with* 1 *degree of freedom. That is,*

$$l(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_{(1)}.$$

Proof. See Appendix I for this proof.

The limiting distribution presented in Theorem 3.3.1 can be used to construct the following EL ratio-based confidence interval for θ_0 using the length-biased right-censored data. Thus, an asymptotic $100(1-\alpha)\%$ confidence interval for the MRL function $\theta_0 = M(x_0)$ at a fixed time x_0, when $x_0 \in [0, \tau)$, could be obtained through the following relation.

$$C_3(x_0) = \left\{\theta(x_0) = \theta_0 : l\left(\theta_0\right) \leq \chi^2_{1,\alpha}\right\},$$

where $\chi^2_{1,\alpha}$ is the upper α-quantile of the distribution of χ^2_1.

3.4 Simulation

A Monte Carlo simulation was undertaken to inspect and illustrate the performance of the empirical likelihood ratio confidence interval proposed. Three separate sample sizes 60, 100 and 200 were considered for constructing the confidence intervals, representing small, moderate and large sample sizes, respectively. Two different values for the nominal level, that are $1-\alpha = 0.95$ and 0.90, were used to estimate the confidence intervals. The performance of the EL method was evaluated based on 5000 repetitions of the confidence intervals for each sample size. Consequently, the coverage probability of the confidence interval in each scenario was computed as the proportion of the number of intervals covering the real value of the unbiased MRL function out of the 5000 repetitions.

For each scenario, the pairs of independent random variables (X_i', T_i') $(i = 1, 2, \ldots, k)$ were generated, where T_i' is the truncation variable from a uniform distribution $U(0, b)$ and X_i' is the i.i.d. copies from the target population of interest. The truncation random variables were generated from a uniform distribution to ensure that the stationarity assumption was satisfied. For each scenario, only those pairs of random variables (X_i', T_i) satisfying the condition $X_i' > T_i'$ were collected, and the remaining data was considered as the unobserved left-truncated subjects. The amount of k was considered large enough to ensure that the required sample size was provided. Also, the value of b was chosen in regard to the target population of interest. We call the observed truncated sample $X_1, \ldots, X_n$ in this thesis. It is apparent that the obtained data was length-biased. For the observed sample, the backward and forward recurrence times were computed using (X_i, T_i) $(i = 1, 2, \ldots, k)$, which means that, for an observed individual, $A_i = T_i$ and $R_i = X_i - T_i$. Afterwards, the censoring variables C_i was generated from some uniform distributions $U(0, b)$. Two separate levels of censoring 12% and 30% were taken for each simulation scenario. The censoring variables were applied for

Table 3.1: 90 % and 95 % Confidence Intervals for MRL of Uniform(1,4)

Scenario:		$1-\alpha = 90\%$				$1-\alpha = 95\%$			
		Censoring Level				Censoring Level			
Time	Sample	15%		30%		15%		30%	
x_0	n	Δ	C.P.	Δ	C.P.	Δ	C.P.	Δ	C.P.
	60	0.402	0.899	0.435	0.886	0.479	0.953	0.519	0.955
1.2	100	0.315	0.897	0.343	0.895	0.375	0.951	0.407	0.948
	200	0.225	0.897	0.245	0.895	0.267	0.952	0.290	0.952
	60	0.333	0.893	0.363	0.887	0.397	0.949	0.432	0.936
1.7	100	0.261	0.893	0.284	0.897	0.311	0.941	0.338	0.952
	200	0.186	0.895	0.203	0.897	0.221	0.945	0.241	0.943
	60	0.277	0.891	0.301	0.886	0.328	0.945	0.357	0.949
2.2	100	0.216	0.897	0.236	0.889	0.258	0.950	0.281	0.940
	200	0.154	0.903	0.168	0.905	0.183	0.950	0.199	0.949
	60	0.223	0.891	0.243	0.890	0.265	0.950	0.288	0.941
2.7	100	0.174	0.900	0.190	0.892	0.208	0.948	0.227	0.950
	200	0.124	0.900	0.136	0.899	0.148	0.950	0.161	0.950
	60	0.167	0.896	0.181	0.883	0.198	0.949	0.216	0.937
3.2	100	0.130	0.902	0.143	0.889	0.156	0.944	0.170	0.944
	200	0.093	0.896	0.102	0.897	0.110	0.943	0.121	0.945
	60	0.099	0.860	0.110	0.854	0.119	0.929	0.133	0.919
3.7	100	0.077	0.885	0.083	0.866	0.093	0.935	0.099	0.922
	200	0.055	0.920	0.150	0.899	0.065	0.954	0.071	0.943

the remaining lifetime (the forward recurrence times) to meet the requirement of informative censoring structure. The code for the simulation presented was written in *R* by the author and can be found in Appendix II.

Table 3.1 illustrates the performances of the 90% and 95% confidence intervals for the MRL function using the empirical likelihood method. The length-biased and right-censored data corresponding to the target population $U(1, 4)$ was considered for analysis, which meets the condition given in Theorem 3.3.1. It can be observed that the EL method preserved the nominal level very well. As expected, the confidence intervals narrowed enormously by increasing in the sample sizes. Moreover, the lengths of confidence intervals widened noticeably as the share of censored subjects increased. However, the observed coverage probabilities for both levels of censoring were almost similar and fortunately there was not considerable negative impact between the results for the censoring levels. This issue can be justified by considering the average lengths of confidence intervals. In addition, the widths of intervals dived swiftly by increasing the value of x_0. A quite slight decline in the coverage probabilities were only observed for small ($n = 60\%$)sample scenario and $x_0 = 3.7$. This issue is completely acceptable since $x_0 = 3.7$ is exactly the 90%-quantile of the target distribution (the censored sample provide less information for tails). Furthermore, there was not any statistically meaningful disparity between two groups of the nominal levels, 90% and 95%.

Table 3.2 summarizes the simulation results for the MRL confidence intervals of the $Gamma(2, 4)$ distribution as the unbiased population. However, the length-biased and right-censored data, according to the setting presented above, was simulated to construct the confidence intervals. The proposed empirical likelihood method was applied to obtain the

Table 3.2: 90 % and 95 % Confidence Intervals for MRL of Gamma(2,4)

Scenario:		$1-\alpha = 90\%$				$1-\alpha = 95\%$			
		Censoring Level				Censoring Level			
Time	Sample	15%		30%		15%		30%	
x_0	n	Δ	C.P.	Δ	C.P.	Δ	C.P.	Δ	C.P.
	60	0.412	0.902	0.443	0.899	0.491	0.956	0.528	0.951
1.0	100	0.320	0.907	0.344	0.894	0.382	0.952	0.411	0.943
	200	0.227	0.905	0.245	0.895	0.273	0.959	0.291	0.948
	60	0.393	0.895	0.423	0.897	0.469	0.951	0.505	0.942
1.5	100	0.305	0.899	0.327	0.900	0.364	0.945	0.391	0.947
	200	0.217	0.906	0.232	0.897	0.264	0.955	0.278	0.951
	60	0.412	0.895	0.445	0.890	0.494	0.941	0.532	0.943
2.0	100	0.321	0.903	0.346	0.889	0.382	0.944	0.411	0.945
	200	0.228	0.899	0.244	0.893	0.275	0.956	0.291	0.951
	60	0.467	0.886	0.464	0.884	0.558	0.939	0.560	0.941
2.5	100	0.363	0.894	0.389	0.888	0.435	0.945	0.465	0.949
	200	0.261	0.898	0.276	0.894	0.323	0.958	0.329	0.948
	60	0.552	0.874	0.558	0.869	0.668	0.925	0.667	0.928
3.0	100	0.430	0.887	0.465	0.893	0.519	0.942	0.559	0.942
	200	0.337	0.904	0.332	0.898	0.409	0.959	0.397	0.954
	60	0.670	0.832	0.683	0.825	0.782	0.894	0.784	0.878
3.5	100	0.533	0.872	0.576	0.860	0.635	0.929	0.687	0.913
	200	0.406	0.913	0.415	0.890	0.498	0.960	0.499	0.945

values reported in this table. It was revealed that the lengths of confidence intervals decreased tremendously as the sample size increased. But, the coverage probabilities were fortunately similar for all sample sizes excluding $x_0 = 3.7$. The observed moderate decrease in values of CP for $x_0 = 3.7$ is not surprising since this point is roughly 90%-quantile of the underlying distribution and censored data results in some instability in tails. In comparison to the lengths of intervals for 30% censoring level, those for 15% censoring level were smaller marginally. In addition, there did not exist any noticeable disparity in the coverage probabilities of the two groups of censoring, 90% and 95%. Once again, this issue can be explained by considering the lengths of intervals and coverage probabilities simultaneously. Moreover, the EL-based confidence intervals narrowed significantly by increasing the value of x_0 in the MRL function. The proposed method almost achieved the nominal levels 90% and 95% for variouse sample sizes and all censoring levels. Broadly, the widths of MRL function confidence intervals were observed to reduce considerably as the values of x_0 grew.

4 Discussion and Future Directions

4.1 Contributions

In this thesis, we have proposed applying the empirical likelihood method for analysis of the survival data collected from a biased sampling procedure. The significant advantage of this method is to provide researchers with a flexible framework for drawing statistical inference. As mentioned, during the recent years there has been an increasing tenancy to study into residual life expectancy. The benefits of investigating the MRL function in comparison to the other survival functions were mentioned. Following this, recent research studies concerning statistical inference about mean residual lifetime were reviewed. Prospective prevalent cohort study was introduced as a practical and efficient method to evaluate the progression of a disease of interest. It was observed that the prevalent cases have longer lifetime on average, that is, length-bias. We considered this issue as the main motivating problem of this thesis. In addition, it was highlighted that several common other sampling procedures in practice result in obtaining length-biased data. Recent literature with regard to statistical inference on length-biased or right-censored survival data was reviewed comprehensively.

In chapter 2, observations from a length-biased sampling procedure were considered for analysis. We investigated the problem of estimating different survival functions using a sample of length-biased data. In addition, the concept of the empirical likelihood ratio method was studied in detail. We drew statistical inference for the MRL function using length-biased data through the EL method. The limiting distribution of the empirical log-likelihood was presented. A simulation study was conducted to inspect the finite sample performance of the method proposed. Another method for constructing confidence interval for the length-biased observations was introduced. Comparing the simulation results of the methods proposed, we mentioned some superiority for the EL method over the normal approximation procedure. A real data example was used for better illustration of the proposed EL method.

The prospective subjects recruited in a prevalent cohort study or other cross-sectional sampling surveys ought to be followed over time until recording the terminating event for all of them. However, it was mentioned that it is always necessary to expect that we may loss to follow-up of some recruited individuals (for a variety of reasons) leads to censoring. In section 3, we considered the possibility of obtaining censored data in a length-biased

sampling procedure. It was indicated that the underlying non-random sampling procedure induces the informative censoring structure to the observed data. The advantages of the unconditional NPMLE approach over the conditional methods in which statistical inference are drawn by conditioning on the observed truncation times was discussed briefly. However, there is no study in the current literature on MRL which uses an unconditional approach to analyze length-biased right-censored data. For this reason, we proposed applying the full likelihood model of the observations to make inference on the MRL function. We established the empirical likelihood of the MRL function using the NPMLE of the length-biased right-censored data. We studied the limiting distribution of the empirical log-likelihood ratio. This limiting distribution was applied for constructing a confidence interval for the MRL function. A simulation study was carried out to investigate the performance of the introduced empirical likelihood method for different finite sample sizes. The simulation results indicated the efficiency and accuracy of the EL method proposed for length-biased right-censored data.

4.2 Future Directions

In this thesis, we proposed a new and advanced method for constructing a confidence interval for the MRL function using empirical likelihood method. But, another very interesting question, specially from a theoretical point of view, is how to obtain an EL-based confidence band for the length-biased right-censored survival data. There is not any study based on the full likelihood function and the unconditional approach in the current literature regarding length-biased and right-censored data. Fakoor et al. (2018) investigated this question for length-bias setting, but they did not considered the possibility of observing censored data in their sample. They established the weak convergence of the empirical log-likelihood ratio stochastic process (2.17). The covariance of the resulting limiting process depends on the following,

$$\psi(x) = E\left[\eta_x^2(X_i, \theta)\right].$$

THEOREM 4.2.1 *Assume that* $E\left[X^{-2}\right] < \infty$. *Then, there exists a mean zero Gaussian process* $\{\varepsilon(x), 0 \leq x \leq a\}$ *such that*

$$l(\theta) \xrightarrow{\mathcal{W}} \frac{\varepsilon^2(\cdot)}{\psi(\cdot)},$$

in $D[0, a]$, *the space of cadlag functions on* $[0, a]$, *where* $\xrightarrow{\mathcal{W}}$ *denotes weakly convergence. The Gaussian process* $\varepsilon(\cdot)$ *is given by*

$$\varepsilon(x) := \int_x^\infty \eta_x(u, \theta) dB(G(u)),$$

with covariance function

$$Cov(\varepsilon(x_1), \varepsilon(x_2)) = \int_{x_1 \vee x_2}^\infty \left(1 - \frac{x_2 + \theta(x_2)}{u}\right)\left(1 - \frac{x_1 + \theta(x_1)}{u}\right) dG(u),$$

where $B(\cdot)$ *is a Brownian bridge in a unit interval.*

Proof. See Fakoor et al. (2018) for this proof.

Using Theorem 4.2.1 and the continuous mapping theorem, it is obtained that

$$\sup_{0\leq x\leq a}\left\{l(\theta)\right\} \xrightarrow{\mathcal{L}} \sup_{0\leq x\leq a}\frac{\varepsilon^2(x)}{\psi(x)}.$$

Consequently, for $x \in [0, a]; a \leq \tau$, an asymptotic $100(1-\alpha)\%$ confidence band for the MRL function is

$$C_4 = \left\{\theta : l\left(\theta\right) \leq q_\alpha, x \in [0, a]\right\},$$

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

$$\sup_{0\leq x\leq a}\left\{\varepsilon^2(x)/\psi(x)\right\}. \tag{4.1}$$

Since it is difficult to evaluate the limiting distribution of (4.1) analytically, they suggested a method to approximate this Gaussian process, and therefore estimate the α-quantile of $\sup_{0\leq x\leq a}\left\{\varepsilon^2(x)/\psi(x)\right\}$ to obtain the confidence band.

We have applied the results of Theorem 4.2.1 on the real data set of the widths of shrubs studied in Section 2.6. The obtained confidence band is presented in Appendix I (see Section 5.2, Figure 5.1). We considered extending the methods presented in Fakoor et al. (2018) (Theorem 4.2.1) and Hollander et al. (1997) for length-biased right-censored data. However, proving the weak convergence for the $l(\theta)$ defined in (3.25) is much more complex and beyond the scope of this thesis. We are studying this problem as our main plan for the future.

5 Appendix I

5.1 Proofs

The aim of this appendix is to present the proof of some theorem and lemma presented throughout of the previous sections.

Proof of Lemma 1.5.1.

$$f\left(x|X \geq T\right) dx \approx \frac{P\left(x \leq X \leq x+dx, X \geq T\right)}{P\left(X \geq T\right)}, \tag{5.1}$$

Now for the denominator of (5.1) we have

$$\begin{aligned} P\left(X \geq T\right) &= \int_0^{\infty}\int_0^{x} f_{X,T}\left(x,t\right)\, dt\, dx \\ &= \int_0^{\infty}\int_0^{x} f\left(x\right) f_T\left(t\right)\, dt\, dx \\ &= \int_0^{\infty}\int_0^{x} f\left(x\right)\frac{1}{\theta}\, dt\, dx \\ &= \frac{1}{\theta}\int_0^{\infty}\int_0^{u} f\left(u\right)\, dt\, dx \\ &= \frac{\mu}{\theta}, \end{aligned} \tag{5.2}$$

for the numerator of (5.1) we have

$$\begin{aligned} P\left(x \leq X \leq x+dx, X \geq T\right) &= \int_x^{x+dx}\int_0^{x} f\left(x\right)\frac{1}{\theta} dt dx \\ &= \frac{1}{\theta}\int_x^{x+dx} x f\left(x\right) dx \\ &\approx \frac{1}{\theta} x f(x)\, dx, \end{aligned} \tag{5.3}$$

Accordingly, given (5.1), (5.2) and (5.3), the underlying equation is proven.

□

The proof of Theorem 2.2.1 has been presented in Owen (2001) and shown here for completeness.

Proof of Theorem 2.2.1. Let $z_1 < z_2 < \ldots < z_m$ denote the distinct values observed for $X_1', X_2', \ldots, X_n'$, and let $n_j \geq 1$ be the number of X_i that are equal to z_j. Let $p_j = F(z_j) - F(z_j^-)$ and define $\hat{p}_j = n_j/n$. If $p_j = 0$ for any $j = 1, \ldots, m$, then $L(F) = 0 < L(F_n)$. But, if we suppose that all $p_j > 0$ and for at least one j, $p_j \neq \hat{p}_j$, then

$$\begin{aligned} \log\left(\frac{L(F)}{L(F_n)}\right) &= \sum_{j=1}^{m} n_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\ &= n\sum_{j=1}^{m} \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \\ &< n\sum_{j=1}^{m} \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right) \qquad (5.4) \\ &\leq 0, \qquad (5.5) \end{aligned}$$

and therefore $L(F) < L(F_n)$. It is worth mentioning that the inequality (5.4) is held because $\log(x) \leq x - 1$ for all $x > 0$ with equality only when only when $x = 1$.

□

In order to complete the proof of Theorem 3.3.1, we initially need to prove the following lemmas. Before we state the first lemma, we need to define the following function :

$$\psi(x) := E\left[\eta_x^2(X, \theta)\right].$$

LEMMA 5.1.1 *Assume that* $E\left(X^{-2}\right) < \infty$. *Then, for all* $x_0 \in [0, \tau)$ *we have*

$$\sqrt{n}\sum_{i=1}^{m} \hat{p}_i \eta_{x_0}(t_i, \theta_0) \xrightarrow{\mathcal{L}} N\left(0, \psi(x_0)\right),$$

and

$$\sum_{i=1}^{m} \hat{p}_i \eta_{x_0}^2(t_i, \theta_0) \xrightarrow{\mathcal{P}} \psi(x_0),$$

where $\xrightarrow{\mathcal{P}}$ *denotes convergence in probability and*

$$\psi(x_0) < \infty.$$

Proof. Suppose x_0 is an arbitrary but fixed point that $x_0 \in [0, \tau)$. Then, under the assumption of the lemma, we have

$$\begin{aligned} \psi(x_0) &= E\left[\eta_{x_0}^2(X, \theta)\right] \\ &= E\left[\left(1 - \frac{x_0 + \theta_0}{X}\right) I_{[x_0,\tau)}(X)\right]^2 \\ &\leq E\left[I_{[x_0,\tau)}(X)\right]^2 + E\left[\frac{x_0 + \theta_0}{X} I_{[x_0,\tau)}(X)\right]^2 \end{aligned}$$

$$\begin{aligned} &\leq 1 + (x_0 + \theta_0)^2 \, E\left[X^{-2}\right] \\ &< \infty. \end{aligned} \tag{5.6}$$

On the other hand, since $E\left[\eta_{x_0}(X, \theta_0)\right] = \int_0^\tau \eta_{x_0}(t, \theta_0) dG(t) \equiv 0$,

$$\sqrt{n} \sum_{i=1}^{m} \hat{p}_i \eta_{x_0}(t_i, \theta_0) = \int_0^\tau \eta_{x_0}(t, \theta_0) \, \alpha_n \, (dt),$$

where $\alpha_n := \sqrt{n}\left(\hat{G}_n(t) - G(t)\right)$.

Now, according to Asgharian and Wolfson (2005), the α_n is weakly convergence to $B\left(G(\cdot)\right)$ where $B(\cdot)$ is a Brownian bridge on the unit interval. Then, by continuous mapping theorem it is concluded that

$$\int_0^\tau \eta_{x_0}(t, \theta_0) \alpha_n \, (dt) \xrightarrow{\mathcal{W}} \mathcal{G}(t) := \int_0^\tau \eta_{x_0}(t, \theta_0) dB\left(dtG(t)\right),$$

and

$$Cov\left(\mathcal{G}(t), \mathcal{G}(s)\right) := \int_{t \vee s}^{\infty} \left(1 - \frac{x_0 + \theta_0}{t}\right)\left(1 - \frac{x_0 + \theta_0}{s}\right) dG(u),$$

where $\mathcal{W}$ indicates weakly converges.

Consequently, given (5.6), it is obtained that

$$\int_0^\tau \eta_{x_0}(t, \theta_0) \alpha_n \, (dt) = \sqrt{n} \sum_{i=1}^{m} \hat{p}_i \eta_{x_0}(t_i, \theta_0) \xrightarrow{\mathcal{L}} N\left(0, \psi(x_0)\right).$$

Regarding the second part of Lemma 5.1.1, using the Asgharian and Wolfson (2005)'s findings and the continuous mapping theorem we have

$$\sum_{i=1}^{m} \hat{p}_i \eta_{x_0}^2(t_i, \theta_0) = \int_0^\infty \eta_{x_0}^2(t, \theta_0) d\hat{G}_n(t) \xrightarrow{\mathcal{L}} \psi(x_0),$$

Thus,

$$\sum_{i=1}^{m} \hat{p}_i \eta_{x_0}^2(t_i, \theta_0) \xrightarrow{\mathcal{P}} \psi(x_0),$$

and thus the proof of Lemma 5.1.1 is completed. □

LEMMA 5.1.2 *Let $h(\cdot)$ be a continuous function. Under the same condition as in Theorem 3.3.1, for all $x_0 \in [0, \tau)$ we have*

$$\sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) h(t_i) \hat{p}_i \xrightarrow{\mathcal{P}} \iota(x_0),$$

and

$$\frac{1}{n}\left[-\sum_{i=1}^{m} \xi_i h^2(t_i) + 2n \sum_{j=1}^{m} \hat{p}_j h^2(t_j) - n\left(\sum_{j=1}^{m} \hat{p}_j h(t_j)\right)^2\right.$$

$$-2\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)+\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)^2\Bigg] \xrightarrow{\mathcal{P}} \varphi_{1h},$$

where

$$\iota(x_0) := \int_0^\tau \eta_{x_0}(t,\theta_0)h(t)dG(t),$$

and

$$\begin{aligned}\varphi_{1h} \;:=\; & \int_0^\tau t^{-1}h^2(t)\int_0^t S_C(s)dsdG(t) - \left(\int_0^\tau h(t)dG(t)\right)^2 \\ & + \int_0^\tau F_C(t)\frac{\left(\int_t^\tau s^{-1}h(s)dG(s)\right)^2}{\int_t^\tau s^{-1}dG(s)}dt.\end{aligned}$$

Proof. In order to present the proof of this lemma, we need to make some statements. Let n_1 and $n_2 = n - n_1$ be the realized values for N_1 and $N_2 = n - N_1$. Then, for the length-biased right-censored sample $\left\{\cup_{i=1}^{n} X_i\right\} = \left\{(Y_1,\ldots,Y_{n_1}) \cup (Z_1,\ldots,Z_{n_2})\right\}$ define $\hat{p} := n_1/n$ $\left(1-\hat{p} = n_2/n\right)$

$$\hat{G}_{\mathcal{U}}(t) := \frac{1}{n_1}\sum_{i=1}^{n_1} I_{[0,t]}(Y_i)$$

and

$$\hat{G}_C(t) := \frac{1}{n_2}\sum_{i=1}^{n_2} I_{[0,t]}(Z_i).$$

Then,

$$\hat{p} \xrightarrow{a.s.} p, \tag{5.7}$$

$$\hat{G}_{\mathcal{U}}(\cdot) \xrightarrow{\mathcal{W}} G_{\mathcal{U}}(\cdot), \tag{5.8}$$

$$\hat{G}_C(\cdot) \xrightarrow{\mathcal{W}} G_C(\cdot) \tag{5.9}$$

and, following Asgharian and Wolfson (2005),

$$\hat{G}_n(\cdot) \xrightarrow{\mathcal{W}} G(\cdot), \tag{5.10}$$

where $\xrightarrow{a.s.}$ indicates convergence almost surely (with probability one), and (5.7) is concluded using the law of large numbers.

Considering the first part of the lemma, $h(\cdot)$, $\eta_{x_0}(\cdot,\theta_0)$, and therefore their product are continuous. Thus,

$$\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)h(t_i)\hat{p}_i = \int_0^\tau \eta_{x_0}(t,\theta_0)h(t)d\hat{G}_n(t) < \infty.$$

Given (5.10), by using the continuous mapping theorem we have

$$\int_0^\tau \eta_{x_0}(t,\theta_0)h(t)d\hat{G}_n(t) \xrightarrow{\mathcal{L}} \int_0^\tau \eta_{x_0}(t,\theta_0)h(t)dG(t).$$

Therefore,

$$\int_0^\tau \eta_{x_0}(t,\theta_0)h(t)d\hat{G}_n(t) \xrightarrow{\mathcal{P}} \iota(x_0).$$

Turning to the second part of the lemma, following (3.3), (5.7) and (5.8) we have

$$\begin{aligned}
\frac{1}{n}\left[\sum_{i=1}^m \xi_i h^2(t_i)\right] &= \sum_{i=1}^m \frac{n_1}{n} d\hat{G}_{\mathcal{U}}(t_i)h^2(t_i) \\
&= \int_0^\tau h^2(t)\hat{p}d\hat{G}_{\mathcal{U}}(t) \\
&\xrightarrow{\mathcal{L}} \int_0^\tau h^2(t)pdG_{\mathcal{U}}(t) \\
&= \int_0^\tau t^{-1}h^2(t)\int_0^t S_C(s)dsdG(t) \\
&= \int_t^\tau h^2(t)dG(t) - \int_0^\tau F_C(s)\int_s^\tau t^{-1}h^2(t)dG(t)ds. \qquad (5.11)
\end{aligned}$$

Furthermore, by considering (3.4) and (3.5), and then continuous mapping and Slutsky's theorems, it is obtained that

$$\begin{aligned}
&\frac{1}{n}\left[2n\sum_{j=1}^m \hat{p}_j h^2(t_j) - 2\sum_{i=1}^m \zeta_i\left(\frac{\sum_{j=i}^m \hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^m \hat{p}_j t_j^{-1}}\right) + \sum_{i=1}^m \zeta_i\left(\frac{\sum_{j=i}^m \hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^m \hat{p}_j t_j^{-1}}\right)^2\right] \\
= \; & 2\sum_{j=1}^m \hat{p}_j h^2(t_j) - 2\sum_{i=1}^m \frac{(n-n_1)}{n} d\hat{G}_C(t_i)\left(\frac{\int_{t_i}^\tau h^2(s)s^{-1}d\hat{G}_n(s)}{\int_{t_i}^\tau s^{-1}d\hat{G}_n(s)}\right) \\
& + \sum_{i=1}^m \frac{(n-n_1)}{n} d\hat{G}_C(t_i)\left(\frac{\int_{t_i}^\tau h(s)s^{-1}d\hat{G}_n(s)}{\int_{t_i}^\tau s^{-1}d\hat{G}_n(s)}\right)^2 \\
= \; & 2\int_0^\tau h^2(t)d\hat{G}_n(t) - 2\int_0^\tau (1-\hat{p})\left(\frac{\int_t^\tau h^2(s)s^{-1}d\hat{G}_n(s)}{\int_t^\tau s^{-1}d\hat{G}_n(s)}\right)d\hat{G}_C(t) \\
& + \int_0^\tau (1-\hat{p})\left(\frac{\int_t^\tau h(s)s^{-1}d\hat{G}_n(s)}{\int_t^\tau s^{-1}d\hat{G}_n(s)}\right)^2 d\hat{G}_C(t) \\
& \xrightarrow{\mathcal{L}} 2\int_0^\tau h^2(t)dG(t) - 2\int_0^\tau (1-p)\left(\frac{\int_t^\tau h^2(s)s^{-1}dG(s)}{\int_t^\tau s^{-1}dG(s)}\right)dG_C(t) \\
& + \int_0^\tau (1-p)\left(\frac{\int_t^\tau h(s)s^{-1}dG(s)}{\int_t^\tau s^{-1}dG(s)}\right)^2 dG_C(t) \\
= \; & 2\int_0^\tau h^2(t)dG(t) - 2\int_0^\tau F_C(t)\int_t^\tau h^2(s)s^{-1}dG(s)dt \\
& + \int_0^\tau F_C(t)\frac{\left(\int_t^\tau h(s)s^{-1}dG(s)\right)^2}{\int_t^\tau s^{-1}dG(s)}dt, \qquad (5.12)
\end{aligned}$$

Given (5.11) and (5.12), since it can be easily checked that $\varphi_{1h}(x_0) < \infty$, we can conclude

that

$$\frac{1}{n}\left[-\sum_{i=1}^{m}\xi_i h^2(t_i)+2n\sum_{j=1}^{m}\hat{p}_j h^2(t_j)-n\left(\sum_{j=1}^{m}\hat{p}_j h(t_j)\right)^2\right.$$
$$\left.-2\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)+\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)^2\right] \xrightarrow{\mathcal{P}} \varphi_{1h}(x_0).$$

□

The proof of Theorem 3.3.1 is mainly inspired by the method discussed in Pan and Zhou (1999) and that in Ning et al. (2013).

Proof of Theorem 3.3.1. In order to prove the asymptotic behavior of the empirical log-likelihood (3.25), we need to define a one-parameter sub-family of all distribution functions $\mathcal{F}$. We first look for a distribution in this sub-family maximizing the empirical log-likelihood, which helps us to obtain the limiting distribution of $l(\theta_0)$.

For any fixed but arbitrary x_0 (and so θ_0), let $h(\cdot)$ be a continuous function such that $h(t)\eta_{x_0}(t,\theta_0) \geq 0$ for all $t > 0$. Then we define $\mathcal{H}_{\theta_0}^G$ to be the class of all functions $h(\cdot)$. Therefore,

$$\mathcal{H}_{\theta_0}^G := \left\{h(\cdot)|\ h(\cdot) \text{ is continuous and } h(\cdot)\eta_{x_0}(\cdot,\theta_0) \geq 0 \text{ a.s. } G(\cdot)\right\}.$$

We also assume the following sub-family of one-parameter distribution function.

$$\mathcal{A}_h^G := \left\{\tilde{G}'(x)|\ \tilde{G}'(x) = \sum_{i=1}^{m} w_i' I_{[0,x]}(t_i);\ w_i' := \frac{\hat{p}_i}{1+\lambda h(t_i)}\left(\sum_{j=1}^{m}\frac{\hat{p}_j}{1+\lambda h(t_j)}\right)^{-1}\right\},$$

where $\hat{p}_i$ $(i = 1,\ldots,n)$ is defined in (3.12).

Let $\boldsymbol{w}' := (w_1',\ldots,w_n')$ defined in $\mathcal{A}_h^G$. Following this, we can define the likelihood ratio function and the profile likelihood ratio for family $\mathcal{A}_h^G$ similar to those in (3.16) and (3.19) as follows, respectively.

$$\begin{aligned} R_h(\tilde{G}') &:= \frac{L(\tilde{G}')}{L(\hat{G}_n)} \\ &= \left\{\prod_{i=1}^{m}\left(\frac{w_i'}{\hat{p}_i}\right)^{\xi_i}\left(\frac{\sum_{j=i}^{m}\frac{1}{t_j}w_j'}{\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j}\right)^{\zeta_i} : \sum_{i=1}^{m} w_i' = 1,\ w_i' \geq 0,\ i = 1,\ldots,m\right\}, \end{aligned}$$

and

$$\begin{aligned} \mathcal{R}_h(\theta_0) &= \sup\left\{R_h(\tilde{G}')\ |\ \int \eta_{x_0}(s,\theta_0)d\tilde{G}'(s) = 0, \tilde{G}' \in \mathcal{A}_h\right\} \\ &= \sup\left\{\prod_{i=1}^{m}\left(\frac{w_i'}{\hat{p}_i}\right)^{\xi_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}w_j'\right)^{\zeta_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)^{-\zeta_i} : \sum_{i=1}^{m} w_i' = 1,\right. \\ &\qquad \left.\sum_{i=1}^{m} w_i'\eta_{x_0}(T_i,\theta_0) = 0, w_i' \geq 0, i = 1,\ldots,m\right\}. \end{aligned}$$

The estimation equation $\sum_{i=1}^{m} w'_i \eta_{x_0}(T_i, \theta_0) = 0$ for the ordinal random sample $T_1, \ldots, T_m$ has been used in practice to impose the constraint $\int \eta_{x_0}(s, \theta_0) dG(s) = 0$. But, this estimation equation has one unique solution which is denoted $\hat{\boldsymbol{w}}' = (\hat{w}'_1, \ldots, \hat{w}'_m)$, obtained by substituting λ'_{x_0} for λ in $\boldsymbol{w}'$. Therefore, similar to (3.25), the corresponding empirical log-likelihood is equal to:

$$\begin{aligned} l_h(\theta_0) &:= -2 \log \mathcal{R}_h(\theta_0) \\ &= 2 \sum_{i=1}^{m} \left[\xi_i \log\left(\hat{p}_i\right) + \zeta_i \log\left(\sum_{j=i}^{m} \frac{1}{t_j} \hat{p}_j \right) \right] \\ &\quad -2 \sum_{i=1}^{m} \left[\xi_i \log\left(\hat{w}'_i\right) + \zeta_i \log\left(\sum_{j=i}^{m} \frac{1}{t_j} \hat{w}'_j \right) \right]. \end{aligned} \tag{5.13}$$

It is of note that the condition $h(\cdot)\eta_{x_0}(\cdot, \theta_0) \geq 0$ almost surely on $G(\cdot)$ used in $\mathcal{H}^G_{\theta_0}$ was required to make sure that $\sum_{i=1}^{m} w'_i \eta_{x_0}(T_i, \theta_0)$ is a monotonic function of λ in order to avoid multiple solutions for $\sum_{i=1}^{m} w'_i \eta_{x_0}(T_i, \theta_0) = 0$.

Before we continue with the rest of the proof, we need to make the following statements. Since $E\left(\eta^2_{x_0}(T_i, \theta_0)\right) < \infty$ by Lemma 5.1.1, according to Lemma 3 of Owen (1990) we have

$$\max_{1 \leq i \leq m} \left| \eta_{x_0}(T_i, \theta_0) \right| = o_p\left(n^{1/2}\right). \tag{5.14}$$

Moreover, given the definition of $h(\cdot)$, it is obtained that

$$\max_{1 \leq i \leq m} \left| h(T_i) \right| = o_p\left(n^{1/2}\right). \tag{5.15}$$

Since $\sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) w'_j = 0$ has one unique solution, say λ'_{x_0},

$$\begin{aligned} 0 &\equiv \left| \sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) \frac{\hat{p}_i}{1 + \lambda'_{x_0} h(t_i)} \right| \\ &\geq \frac{\left|\lambda'_{x_0}\right|}{\max_{1 \leq j \leq m} \left|1 + \lambda'_{x_0} h(t_i)\right|} \left| \sum_{i=1}^{m} h(t_i) \eta_{x_0}(t_i, \theta_0) \hat{p}_i \right| - \left| \sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) \hat{p}_i \right| \\ &\geq \frac{\left|\lambda'_{x_0}\right|}{1 + \left|\lambda'_{x_0}\right| \max_{1 \leq j \leq m} \left|h(t_i)\right|} \left| \sum_{i=1}^{m} h(t_i) \eta_{x_0}(t_i, \theta_0) \hat{p}_i \right| - \left| \sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) \hat{p}_i \right|. \end{aligned} \tag{5.16}$$

According to Lemma 5.1.1, for the second term of the right side of the recent inequality, we have

$$\left| \sum_{i=1}^{m} \eta_{x_0}(t_i, \theta_0) \hat{p}_i \right| = O_p\left(n^{-1/2}\right).$$

Turning to the other term of the mentioned inequality, by considering Lemma 5.1.2, it is deduced that

$$\frac{\left|\lambda'_{x_0}\right|}{1+\left|\lambda'_{x_0}\right|\max_{1\le j\le m}\left|h(t_i)\right|} = O_p\left(n^{-1/2}\right).$$

By applying (5.15), this equation results in

$$\left|\lambda'_{x_0}\right| = O_p\left(n^{-1/2}\right). \tag{5.17}$$

Considering the first equality in (5.16), use a Taylor expansion of each function $f_i(\lambda) := \hat{p}_i\left(1+\lambda h(t_i)\right)^{-1}$ around the origin, we have

$$\begin{aligned}\sum_{i=1}^{m}\frac{\eta_{x_0}(t_i,\theta_0)\hat{p}_i}{1+\lambda'_{x_0}h(t_i)} &= \sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i - \lambda'_{x_0}\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i h(t_i)\\ &\quad +\lambda'^2_{x_0}\sum_{i=1}^{m}\frac{\eta_{x_0}(t_i,\theta_0)\hat{p}_i h^2(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3},\end{aligned} \tag{5.18}$$

where $\tau' \in [0,1]$.

But, for the last term of the above equation we have

$$\begin{aligned}\left|\lambda'_{x_0}\sum_{i=1}^{m}\frac{\eta_{x_0}(t_i,\theta_0)\hat{p}_i h^2(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3}\right| &\le \max_{1\le i\le m}\left|\frac{\lambda'_{x_0}h(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3}\right|\left|\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i h(t_i)\right|\\ &\le \frac{1}{\left|1-\tau'\left|\lambda'_{x_0}\right|\max_{1\le i\le m}\left|h(t_i)\right|\right|^3}\left|\lambda'_{x_0}\right|\max_{1\le i\le m}\left|h(t_i)\right|\left|\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i h(t_i)\right|\\ &= o_p(1).\end{aligned} \tag{5.19}$$

The last inequality is true because $\tau'\left|\lambda'_{x_0}\right|\max_{1\le i\le m}\left|h(t_i)\right| = o_p(1)$.

Consequently, (5.18) and (5.19) together result in the following relation.

$$\sum_{i=1}^{m}\frac{\eta_{x_0}(t_i,\theta_0)\hat{p}_i}{1+\lambda'_{x_0}h(t_i)} = \sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i - \lambda'_{x_0}\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i h(t_i) + o_p(\lambda'_{x_0}).$$

Following this, since $\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)w'_j = 0$, it can be easily obtained that

$$\lambda'_{x_0} = \frac{\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i}{\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)\hat{p}_i h(t_i)} + o_p\left(n^{-1/2}\right)$$

Given Lemma 5.1.1 and 5.1.2, by applying Slutsky's theorem, we have

$$n\lambda'^2_{x_0} \xrightarrow{\mathcal{L}} \varphi_{2h}(x_0)\chi^2_1, \tag{5.20}$$

where $\varphi_{2h}(\cdot)$ is defined as follows,

$$\varphi_{2h}(x_0) := \frac{\psi(x_0)}{\iota^2(x_0)}.$$

Now, an alternative to the $L(\boldsymbol{w})$ in (3.16) is obtained by substituting $\hat{w}'$ for w as follows:

$$\begin{aligned} U(\lambda) &:= \log\left[L\left(\boldsymbol{w}'\right)\right] \\ &= \log\left\{\prod_{i=1}^{m} w_i'^{\xi_i}\left(\sum_{j=i}^{m}\frac{1}{t_j}w_j'\right)^{\zeta_i}\right\} \\ &= \sum_{i=1}^{m}\xi_i\log\hat{p}_i - \sum_{i=1}^{m}\xi_i\log\left(1+\lambda h(t_i)\right) - \sum_{i=1}^{m}\left(\xi_i+\zeta_i\right)\log\left(\sum_{j=1}^{m}\frac{\hat{p}_j}{1+\lambda h(t_j)}\right) \\ &\quad + \sum_{i=1}^{m}\zeta_i\log\left(\sum_{j=i}^{m}\frac{\hat{p}_j}{t_j\left(1+\lambda h(t_j)\right)}\right). \end{aligned}$$

Bear in mind that in the above equation $\sum_{i=1}^{m}\left(\xi_i+\zeta_i\right) = n$. It is apparent that w_i' $(i = 1,\ldots,n)$, and therefore $L\left(\boldsymbol{w}'\right)$ are functions of λ. Moreover, as mentioned, the equation $\sum_{i=1}^{m}\eta_{x_0}(t_i,\theta_0)w_j' = 0$ has one unique solution that is $\hat{\boldsymbol{w}}'$ which obtained by substituting λ_{x_0}' for λ in $\boldsymbol{w}'$. Accordingly, by applying a Taylor expansion for $U(\lambda) = L\left(w'\right)$ around the origin, it can be obtained that

$$\begin{aligned} U(\lambda_{x_0}') &= \sum_{i=1}^{m}\xi_i\log\left(\hat{p}_i\right) + \sum_{i=1}^{m}\zeta_i\log\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right) \\ &\quad +\lambda_{x_0}'\left[-\sum_{i=1}^{m}\xi_i h(t_i) + n\left(\sum_{j=1}^{m}\hat{p}_j h(t_j)\right) - \sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)\right] \\ &\quad +\frac{\lambda_{x_0}'^2}{2}\left[\sum_{i=1}^{m}\xi_i h^2(t_i) - 2n\sum_{j=1}^{m}\hat{p}_j h^2(t_j) + n\left(\sum_{j=1}^{m}\hat{p}_j h(t_j)\right)^2\right. \\ &\quad \left. +2\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right) - \sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)^2\right] \\ &\quad +\frac{\lambda_{x_0}'^3}{6}\left[-\sum_{i=1}^{m}\xi_i\frac{h^3(t_i)}{\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^3} + 6n\frac{\sum_{i=1}^{m}\hat{p}_i h^3(t_i)\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-4}}{\sum_{i=1}^{m}\hat{p}_i\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-1}}\right. \\ &\quad -2n\frac{\sum_{i=1}^{m}\hat{p}_i h(t_i)\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-2}\sum_{i=1}^{m}\hat{p}_i h^2(t_i)\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-3}}{\left\{\sum_{i=1}^{m}\hat{p}_i\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-1}\right\}^2} \\ &\quad -4n\frac{\sum_{i=1}^{m}\hat{p}_i h^2(t_i)\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-3}}{\left\{\sum_{i=1}^{m}\hat{p}_i\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-1}\right\}^2} + 2n\frac{\left\{\sum_{i=1}^{m}\hat{p}_i h(t_i)\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-2}\right\}^3}{\left\{\sum_{i=1}^{m}\hat{p}_i\left(1+\tau'\lambda_{x_0}'h(t_i)\right)^{-1}\right\}^4} \\ &\quad \left. -6\sum_{i=1}^{m}\zeta_i\frac{\sum_{j=1}t_j^{-1}\hat{p}_j h^3(t_j)\left(1+\tau'\lambda_{x_0}'h(t_j)\right)^{-4}}{\sum_{j=1}t_j^{-1}\hat{p}_j\left(1+\tau'\lambda_{x_0}'h(t_j)\right)^{-1}}\right. \end{aligned} \tag{5.21}$$

$$+2\sum_{i=1}^{m}\zeta_i\frac{\sum_{j=1}t_j^{-1}\hat{p}_j h(t_j)\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-2}\sum_{j=1}t_j^{-1}\hat{p}_j h^2(t_j)\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-3}}{\left\{\sum_{j=1}t_j^{-1}\hat{p}_j\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-1}\right\}^2}$$

$$+4\sum_{i=1}^{m}\zeta_i\frac{\sum_{j=1}t_j^{-1}\hat{p}_j h^2(t_j)\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-3}\sum_{j=1}t_j^{-1}\hat{p}_j h(t_j)\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-2}}{\left\{\sum_{j=1}t_j^{-1}\hat{p}_j\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-1}\right\}^2}$$

$$-2\sum_{i=1}^{m}\zeta_i\frac{\left\{\sum_{j=1}t_j^{-1}\hat{p}_j h(t_j)\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-2}\right\}^3}{\left\{\sum_{j=1}t_j^{-1}\hat{p}_j\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-1}\right\}^3}\Bigg] \tag{5.22}$$

Given (3.12), it is apparent that $L(\boldsymbol{w}')$ attains its maximum in general at $\boldsymbol{w}' = \hat{\boldsymbol{p}}$ (when $\lambda' = 0$), which results in $U'(0) = 0$, and consequently (5.21) is equal to zero. On the other hand, in view of the fact that

$$\left(1+\tau'\lambda'_{x_0}h(t_j)\right)^{-1} = \left(1+o_p(1)\right)^{-1} = 1+o_p(1),$$

by using an argument analogous with Lemma 5.1.2, we have

$$\begin{aligned}
U(\lambda'_{x_0}) &= \sum_{i=1}^{m}\xi_i\log\left(\hat{p}_i\right)+\sum_{i=1}^{m}\zeta_i\log\left(\sum_{j=i}^{m}\frac{1}{t_j}\hat{p}_j\right)\\
&\quad+\frac{\lambda'^2_{x_0}}{2}\left[\sum_{i=1}^{m}\xi_i h^2(t_i)-2n\sum_{j=1}^{m}\hat{p}_j h^2(t_j)+n\left(\sum_{j=1}^{m}\hat{p}_j h(t_j)\right)^2\right.\\
&\quad\left.+2\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)-\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)^2\right]\\
&\quad+\frac{\lambda'^3_{x_0}}{6}\left[-\sum_{i=1}^{m}\xi_i\frac{h^3(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3}\right]+o_p(n^{-1/2}).
\end{aligned} \tag{5.23}$$

However,

$$\begin{aligned}
\left|\lambda'^3_{x_0}\sum_{i=1}^{m}\xi_i\frac{h^3(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3}\right| &\leq \left|\lambda'_{x_0}\right|^3\max_{1\leq i\leq m}\left|\frac{\xi_i}{\hat{p}_i}\right|\left|\sum_{i=1}^{m}\hat{p}_i\frac{h^3(t_i)}{\left(1+\tau'\lambda'_{x_0}h(t_i)\right)^3}\right|\\
&= \left|\lambda'_{x_0}\right|^3\max_{1\leq i\leq m}\left|\frac{(n-n_1)\,d\hat{G}_{\mathcal{U}}(t_i)}{\hat{p}_i}\right|o_p(1)\\
&\leq o_p(n^{-1/2}).
\end{aligned} \tag{5.24}$$

Following this, by applying (5.24) and (5.23) we obtain for $l_h(\theta_0)$ obtained in (5.13) that

$$l_h(\theta_0) = -2\left\{\log\left[L\left(\hat{\boldsymbol{w}}'\right)\right]-\log\left[L\left(\hat{\boldsymbol{p}}\right)\right]\right\}$$

$$
\begin{aligned}
&= -2\left(U(\lambda'_{x_0}) - U(0)\right) \\
&= \lambda'^2_{x_0}\left[-\sum_{i=1}^{m}\xi_i h^2(t_i) + 2n\sum_{j=1}^{m}\hat{p}_j h^2(t_j) - n\left(\sum_{j=1}^{m}\hat{p}_j h(t_j)\right)^2\right. \\
&\quad \left. -2\sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h^2(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right) + \sum_{i=1}^{m}\zeta_i\left(\frac{\sum_{j=i}^{m}\hat{p}_j h(t_j)t_j^{-1}}{\sum_{j=i}^{m}\hat{p}_j t_j^{-1}}\right)^2\right] + o_p(1).
\end{aligned}
$$

Now, given Lemma 5.1.2 and (5.20), it can be concluded by applying Slutsky's theorem that

$$l_h(\theta_0) \xrightarrow{\mathcal{L}} \varphi_{1h}(x_0)\varphi_{2h}(x_0)\chi^2_1.$$

However, by an analogous argument of Pan and Zhou (1999) and Zhou and Li (2008), it can be deduced that, for each point $x_0 \in [0, \tau)$,

$$\inf_h \varphi_{1h}\varphi_{2h}(x_0) = 1$$

Accordingly, applying the continuous mapping theorem

$$\inf_h l_h(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_1. \tag{5.25}$$

Thus, there exists an $h(\cdot)$ for which the empirical log-likelihood ratio $l_h(\theta)$ converges to a chi-square random variable with one degree of freedom. But, there is only one $\tilde{G}'(\cdot)$ maximizing the log-likelihood ratio $\mathcal{R}_h(\theta_0)$ which is equivalent the unique $\tilde{G}(\cdot)$ that maximizes $\mathcal{R}(\theta_0)$. This issue alongside with (5.25) result in

$$l(\theta_0) = -2\log\mathcal{R}(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_1,$$

and therefore the proof is completed.

□

5.2 Real Data Application

As mentioned in Chapter 4, the weak convergence presented in Theorem 4.2.1 can be used to construct a confidence band for a set of length-biased data. We have applied these results for the real data set on the widths of shrubs discussed in Section 2.6 for better illustration.

Figure 5.1 reveals the empirical likelihood ratio-based 95% confidence band for the MRL function of the widths of shrubs. In addition, the estimated MRL curve has been plotted simultaneously using the consistent estimator (2.6) for better comparison. According to this diagram, the MRL of shrubs are anticipated to decline gradually as their widths increases. For instance, while it is estimated for the shrubs with more that 0.1 widths to grow gradually until they become between 0.75 to 1.35 thicker, the MRL of the shrubs with widths equal to roughly 1.9 is from almost 0 to 0.8.

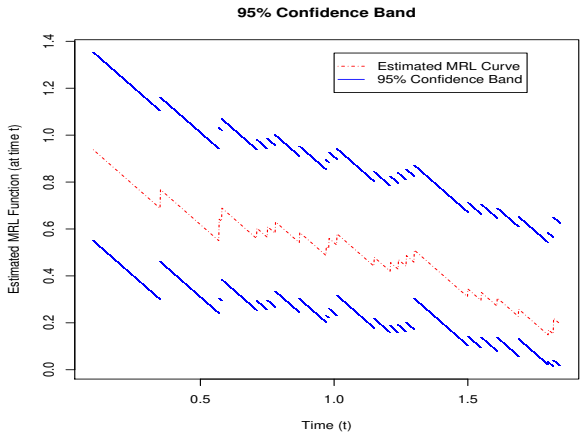

Figure 5.1: Confidence Band for the MRL function of the widths of the shrubs

6
Appendix II

This appendix comprises the R code for simulation study.

Confidence Intervals for Length-biased Survival data using Empirical Likelihood and Normal Approximation Methods.

```
library(boot)

library(rootSolve)
\vspace{0.5 cm}

final=function(n,nn,alpha) {

c=seq(0.01, 5, 0.01)

cc=seq(0.01, 10, 0.01)

tt=seq(0.5, 10.5, 2)

lucb=c()


M= function(t0) {
as.numeric(integrate(function(x) {
pweibull(x, shape=0.5, scale = 2,lower.tail =FALSE)},t0,20000)[1])
/pweibull(t0, shape=0.5, scale = 2,lower.tail =FALSE)
}

FF1=function(m,q) mean((((t0+m)/yy)-1)/(1+q*(((t0+m)/yy)-1)))
```

```
FF2=function(m,q) {
2*sum(log(1+ q* (((t0+m)/yy)-1)))
- qchisq(p=alpha, df=1, ncp = 0,lower.tail = FALSE, log.p = FALSE)
}

model <- function(s)c( FF1(m=s[1],q=s[2]) , FF2(m=s[1],q=s[2]) )

FF3=function(o){ (n/sn)*((mn-o)^2)
-qchisq(p=alpha, df=1, ncp = 0, lower.tail = FALSE, log.p = FALSE)
}

FF4=function(y,i) {xx=y[i]

mun2=(sum(1/xx)/n)^-1

yo2=c(xx[xx<t0])

fnt2=(mun2/n)*sum(1/yo2)

xx=c(y[xx>t0])

mn2=(mun2/n)*length(xx)

mn2=(mn2/(1-fnt2))-t0

sqrt(n)*((mn2)-mn)

}

FF5=function(y,a) as.numeric( integrate( function(x) {
(sqrt(x)*exp(-sqrt(x/2))/(8*sqrt(2)))}, 0, upper=y)[1])-a

FF6=function(y) uniroot(F5,c(0,2699),tol = 0.00001, a =y)

F5=Vectorize(FF5)

F6=Vectorize(FF6)

for (t0 in tt){

cb=cb0=c()

j=0

mr1=M(t0)

while ( j< nn ){
```

```
y=runif(n, 0,1 )

y=c(as.numeric(F6(y)[1,]))

y=sort(y)

yy=c(y[y>t0])

r=r0=c()

for(i in c) r=c(r,multiroot(f=model, start=c(i,0.25))\$root[1])

r=r[!is.na(r)]

r=sort(r)

while ( r[1]<0 ) r=r[-1]

while ( (r[2]-r[1]) > 0.01) r=r[-1]

while ( (r[length(r)]-r[length(r)-1]) > 0.01) r=r[-length(r)]

j=j+1

mun=(sum(1/y)/n)^-1

yo=c(y[y<t0])

fnt=(mun/n)*sum(1/yo)

mn=(mun/n)*length(yy)

mn=(mn/(1-fnt))-t0

sn=sd(boot(y,FF4,R=500)\$t)

r0=c(qnorm(alpha/2,0,1)*(sn/sqrt(n))+mn ,
-qnorm(alpha/2,0,1)*(sn/sqrt(n))+mn )


cb =rbind(cb, c( r[1], r[length(r)], r[length(r)]-r[1],
sign((mrl-r[1])*(r[length(r)]-mrl))+1))

cb0=rbind(cb0, c(r0[1], r0[2], r0[2]-r0[1],
sign((mrl-r0[1])*(r0[2]-mrl))+1))
}

cat("time=",t0,"\n")
```

```
cat("MRL(t=",t0,")",mrl,"\n")

cat("EL: C.I.=(", mean(cb[,1]),",", mean(cb[,2]),"); Delta=",
mean(cb[,3]),"and CP=",mean(cb[,4]) /2 )

cat("\n")

cat("NA: C.I.=(", mean(cb0[,1]),",", mean(cb0[,2]),"); Delta =",
mean(cb0[,3]),"  and CP=",mean(cb0[,4])/2 )

cat("\n")

}

}

final(n=100, nn=5000, alpha=0.05)
```

Confidence Intervals for Length-biased and Right-censored Survival Data using Empirical Likelihood Method.

The following R code represents the EM algorithm for the unconditional NPMLE of the distribution function using right-censored length-biased survival data.

```
Mn=function(x, p) (sum((p[1,]>x)*p[2,])/sum((p[1,]>x)*(p[2,]/p[1,]))) - x

p.em=function(y, z, tol ){

m=length(y)

n=length(z)

xi=zeta=p=c()

tt=c(y,z)

names(tt)=rep(c('y','z'),c(m,n))

t=sort(tt)

xi=names(t)=='y'

zeta=!xi

p=rep(1/(m+n),m+n)

pt=matrix(p/t,m+n,m+n)
```

```
pt[upper.tri(pt)]=0

pnew=(xi+p*cumsum(zeta/apply(pt,2,sum))/t)/(m+n)

while(sum((p-pnew)^2)>tol){

p=pnew

pt=matrix(p/t,m+n,m+n)

pt[upper.tri(pt)]=0

pnew=(xi+p*cumsum(zeta/apply(pt,2,sum))/t)/(m+n)

}

pnew=rbind(t,pnew)

return(pnew)

}
```

The EM algorithm below can be used to obtain the unconditional NPMLE of the distribution function under the constraint defined for the MRL function using right-censored length-biased survival data.

```
w.em=function(y, z, t0, theta, tol){

m=length(y)

n=length(z)

xi=zeta=p=c()

tt=c(y,z)

names(tt)=rep(c('y','z'),c(m,n))

t=sort(tt)

xi=names(t)=='y'

zeta=!xi

w=rep(1/(m+n),m+n)

wt=matrix(w/t,m+n,m+n)
```

```
wt[upper.tri(wt)]=0

rho=xi+w*cumsum(zeta/apply(wt,2,sum))/t

eta=(1-((t0+theta)/t))*(t>=t0)

lambda=multiroot(f = Vectorize(function(x) {sum(rho*eta/(m+n+x*eta))}),
start= c(0))\$root[1]

wnew=(rho)/(m+n+lambda*eta)

while(sum((w-wnew)^2)> tol ){

w=wnew

wt=matrix(w/t,m+n,m+n)

wt[upper.tri(wt)]=0

rho=xi+w*cumsum(zeta/apply(wt,2,sum))/t

lambda=multiroot(f = Vectorize(function(x) {sum(rho*eta/(m+n+x*eta))}),
start= c(0))\$root[1]

wnew=(rho)/(m+n+lambda*eta)

}

wnew=rbind(t,wnew)

return(wnew)

}
```

For the Length-biased sample size m=n_1 and $n = n_2$, the nominal level alpha=α and

```
t0=$x_0$ we have

library(rootSolve)

Model=function(y, z, t0, p, theta, tol, alpha){

w=w.em(y=y, z=z, t0=t0, theta=theta, tol=tol)

if (all(w[2,]>0)==TRUE) {
return(2*sum(xi*log(p[2,]/w[2,])+
```

```
zeta*log( rev(cumsum(rev(p[2,]/t)))/rev(cumsum(rev(w[2,]/t)))))
-qchisq(p=alpha, df=1, ncp = 0, lower.tail=FALSE,log.p = FALSE))

}

}

MRL=as.numeric(integrate(function(x) pgamma(x, shape=4,
scale = 1/2,lower.tail =FALSE),t0,1000)[1])
/pgamma(t0, shape=4, scale = 1/2,lower.tail =FALSE)

#MRL=(16-(8*t0)+t0^2)/ (8-(2*t0))

j=count=0
```

We need to consider y and z as the vectors of length-biased, and length-biased right-censored observations corresponding the target population of interest.

```
count=count+1

xi=zeta=w=c()

tt=c(y,z)

names(tt)=rep(c('y','z'),c(m,n))

t=sort(tt)

xi=names(t)=='y'

zeta=!xi

LB=UB=c()

p=p.em(y=y, z=z, tol=tol, t=t)

M=Mn(x=t0, p=p)

d2=M-0.05

Ld=Model(y=y, z=z, t0=t0, p=p, theta=d2, tol=tol, alpha=alpha)

while(Ld<0 || is.null(Ld)){

d2=d2-0.25

Ld=Model(y=y, z=z, t0=t0, p=p, theta=d2, tol=tol, alpha=alpha)

}
```

```
LB=uniroot(function (x) {Model(y=y, z=z, t0=t0, p=p, theta=x, tol=tol,
alpha=alpha)}, interval=c(d2, M))$root

d=M+0.05

Ud=Model(y=y, z=z, t0=t0, p=p, theta=d, tol=tol, alpha=alpha)

while( Ud<0 || is.null(Ud)){

d=d+0.25

Ud=Model(y=y, z=z, t0=t0, p=p, theta=d, tol=tol, alpha=alpha,
m=m, n=n, t=t, xi=xi, zeta=zeta)

}

UB=uniroot(function (x) {Model(y=y, z=z, t0=t0, p=p, theta=x, tol=tol,
alpha=alpha, m=m, n=n, t=t, xi=xi, zeta=zeta)}, interval=c(M, d))$root

CI=cbind(CI, c(LB, UB, LB<=MRL & MRL<=UB, UB-LB, M) )
```

References

Asgharian, M. and Wolfson, D. B. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics*, 33(5):2109–2131.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine*, 25(10):1751–1767.

Bae, J.-S. and Kim, S.-Y. (2006). Uniform asymptotics in the empirical mean residual life process. *Journal of the Korean Mathematical Society*, 43(2):225–239.

Bai, F., Huang, J., and Zhou, Y. (2016). Semiparametric inference for the proportional mean residual life model with right-censored length-biased data. *Statistica Sinica*, 26(3):1129–1158.

Chan, K. C. G., Chen, Y. Q., and Di, C.-Z. (2012). Proportional mean residual life model for right-censored length-biased data. *Biometrika*, 99(4):995–1000.

Chaubey, Y. P. and Sen, A. (1998). *On Smoothed Functional Estimation Under Random Censoring*, pages 83–97. WORLD SCIENTIFIC.

Chaubey, Y. P. and Sen, A. (2008). Smooth estimation of mean residual life under random censoring. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, Volume 1:35–49.

Chen, Y. J., Ning, W., and Gupta, A. K. (2017). Jackknife empirical likelihood test for mean residual life functions. *Communications in Statistics - Theory and Methods*, 46(7):3111–3122.

Copelan, E., Biggs, J., Thompson, J., Crilley, P., Szer, J., Klein, J., Kapoor, N., Avalos, B., Cunningham, I., and Atkinson, K. (1991). Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with bucy2. *Blood*, 78(3):838–843.

Cox, D. (1962). *Renewal Theory*. Methuen's monographs on applied probability and statistics. Methuen.

Cox, D. R. (1969). Some sampling problems in technology. *In New Developments in Survey Sampling*, pages 506–527.

Cristóbal, J. A. and Alcalá, J. T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. *Test*, 10(2):309–332.

Csörgo, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press.

Csörgo, M. and Zitikis, R. (1996). Mean residual life processes. *Ann. Statist.*, 24(4):1717–1739.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Efromovich, S. (2008). *Nonparametric curve estimation: methods, theory, and applications*. Springer Science & Business Media.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853.

Einstein, A. (1905). Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 322(8):549–560.

Fakoor, V. (2015). On the nonparametric mean residual life estimator in length-biased sampling. *Communications in Statistics - Theory and Methods*, 44(3):512–519.

Fakoor, V., Shariati, A., and Sarmad, M. (2018). The MRL function inference through empirical likelihood in length-biased sampling. *Journal of Statistical Planning and Inference*, 196:115–131.

Freireich, E. J., Gehan, E., Fret, E., Schroder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B., and Lee, S. (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, 21(6):699–716.

Gaenssler, P. and Stute, W. (1979). Empirical processes: A survey of results for independent and identically distributed random variables. *Ann. Probab.*, 7(2):193–243.

Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika*, 86(1):27–43.

Hall, W. J. and Wellner, J. A. (1979). Estimation of mean residual life. *University of Rochester, Department of Statistics, Technical Report*.

Hamburg, B. A., K.-H. C. . J. W. (1975). A hierarchy of drug use in adolescene: behavioral and attitudinal correlates of substantial drug use. *American Journal of Psychiatry*, 132(11):1155–1163. PMID: 1166892.

He, S., Liang, W., Shen, J., and Yang, G. (2016). Empirical likelihood for right censored lifetime data. *Journal of the American Statistical Association*, 111(514):646–655.

Hjort, N. L., McKeague, I. W., and Keilegom, I. V. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, 37(3):1079–1111.

Hollander, M., McKeague, I. W., and Yang, J. (1997). Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association*, 92(437):215–226.

Kalbfleisch, J. D. and Lawless, J. F. (1992). Some useful statistical methods for truncated data. *Journal of Quality Technology*, 24(3):145–152.

Kardaun, O. (1983). Statistical survival analysis of male larynx-cancer patients - a case study. *Statistica Neerlandica*, 37(3):103–125.

Kiefer, J. (1972). Skorohod embedding of multivariate rv's, and the sample df. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 24(1):1–35.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, chapter : Examples of Survival Data, pages 1–20. Springer New York, New York, NY.

Lagakos, S. W., Barraj, L. M., and Gruttola, V. D. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75(3):515–523.

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

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.

McFadden, J. A. (1962). On the lengths of intervals in a stationary point process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):364–382.

Muttlak, H. A. and McDonald, L. L. (1990). Ranked set sampling with size-biased probability of selection. *Biometrics*, 46(2):435–445.

Nahman, N. S., Middendorf, D. F., Bay, W. H., McElligott, R., Powell, S., and Anderson, J. (1992). Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *Journal of the American Society of Nephrology*, 3(1):103–107.

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.

Ning, J., Qin, J., and Shen, Y. (2010). Non-parametric tests for right-censored data with biased sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):609–630.

Oakes, D. and Dasu, T. (1990). A note on residual life. *Biometrika*, 77(2):409–410.

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

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Pan, X.-R. and Zhou, M. (1999). Using one-parameter sub-family of distributions in empirical likelihood ratio with censored data. *Journal of Statistical Planning and Inference*, 75(2):379 – 392.

Patil, G. P. and Ord, J. K. (1976). On size-biased sampling and related form-invariant weighted distributions. *Sankhy: The Indian Journal of Statistics, Series B (1960-2002)*, 38(1):48–61.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34(2):179–189.

Qin, G. and Zhao, Y. (2007). Empirical likelihood inference for the mean residual life under random censorship. *Statistics and Probability Letters*, 77(5):549 – 557.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 27(2/4):311–324.

Rao, C. R. (1977). A natural example of a weighted binomial distribution. *The American Statistician*, 31(1):24–26.

Sickle-Santanello, B. J., Farrar, W. B., Decenzo, J. F., Keyhani-Rofagha, S., Klein, J., Pearl, D., Laufman, H., and O'Toole, R. V. (1988). Technical and statistical improvements for flow cytometric dna analysis of paraffin-embedded tissue. *Cytometry*, 9(6):594–599.

Tsuang, M. T. and Woolson, R. F. (1977). Mortality in patients with schizophrenia, mania, depression and surgical conditions: A comparison with general population mortality. *British Journal of Psychiatry*, 130(2):162166.

Turnbull, B. W. and Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 34(3):367–375.

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

Vardi, Y. and Zhang, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics*, 20(2):1022–1039.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143.

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

Wang, Q. H. and Jing, B. Y. (2001). Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53(3):517–527.

Wicksell, S. D. (1925). The corpuscle problem. a mathematical study of a biometric problem. *Biometrika*, 17(1-2):84–99.

Wiener, N. (1923). Differential-space. *Journal of Mathematics and Physics*, 2(1-4):131–174.

Wiener, N. (1924). Un problème de probabilités dénombrables. *Bull. Soc. Math., France*, 52:569–578.

Wolfson, C., Wolfson, D. B., Asgharian, M., M'Lan, C. E., Østbye, T., Rockwood, K., and Hogan, D. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116. PMID: 11297701.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.*, 13(1):163–177.

Wu, H. and Luan, Y. (2014). An efficient estimation of the mean residual life function with length-biased right-censored data. *Mathematical Problems in Engineering*, 2014(937397):5.

Yang, G. L. (1978). Estimation of a biometric function. *Ann. Statist.*, 6(1):112–116.

Zhao, M., Jiang, H., and Liu, X. (2013). A note on estimation of the mean residual life function with left-truncated and right-censored data. *Statistics and Probability Letters*, 83(10):2332–2336.

Zhao, Y. and Qin, G. (2007). Inference for a linear functional of cumulative hazard function via empirical likelihood. *Communications in Statistics - Theory and Methods*, 36(2):313–327.

Zheng, J., Shen, J., and He, S. (2014). Adjusted empirical likelihood for right censored lifetime data. *Statistical Papers*, 55(3):827–839.

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

Zhou, M. and Li, G. (2008). Empirical likelihood analysis of the buckleyjames estimator. *Journal of Multivariate Analysis*, 99(4):649 – 664.