Spring 5-3-2013

Nonparametric Inferences for the Hazard Function with Right Truncation

Haci Mustafa Akcin
Mathematics and Statistics

NONPARAMETRIC INFERENCES FOR THE HAZARD FUNCTION WITH RIGHT TRUNCATED DATA

by

HACI MUSTAFA AKCIN

Under the Direction of Dr. Xu Zhang

ABSTRACT

Incompleteness is a major feature of time-to-event data. As one type of incompleteness, truncation refers to the unobservability of the time-to-event variable because it is smaller (or greater) than the truncation variable. A truncated sample always involves left and right truncation.

Left truncation has been studied extensively while right truncation has not received the same level of attention. In one of the earliest studies on right truncation, Lagakos *et al.* [40] proposed to transform a right truncated variable to a left truncated variable and then apply existing methods to the transformed variable. The reverse-time hazard function is introduced

through transformation. However, this quantity does not have a natural interpretation. There exist gaps in the inferences for the regular forward-time hazard function with right truncated data. This dissertation discusses variance estimation of the cumulative hazard estimator, one-sample log-rank test, and comparison of hazard rate functions among finite independent samples under the context of right truncation.

First, the relation between the reverse- and forward-time cumulative hazard functions is clarified. This relation leads to the nonparametric inference for the cumulative hazard function. Jiang [32] recently conducted a research on this direction and proposed two variance estimators of the cumulative hazard estimator. Some revision to the variance estimators is suggested in this dissertation and evaluated in a Monte-Carlo study.

Second, this dissertation studies the hypothesis testing for right truncated data. A series of tests is developed with the hazard rate function as the target quantity. A one-sample log-rank test is first discussed, followed by a family of weighted tests for comparison between finite K-samples. Particular weight functions lead to log-rank, Gehan, Tarone-Ware tests and these three tests are evaluated in a Monte-Carlo study.

Finally, this dissertation studies the nonparametric inference for the hazard rate function for the right truncated data. The kernel smoothing technique is utilized in estimating the hazard rate function. A Monte-Carlo study investigates the uniform kernel smoothed estimator and its variance estimator. The uniform, Epanechnikov and biweight kernel estimators are implemented in the example of blood transfusion infected AIDS data.

INDEX WORDS: Right truncation, Reverse-time hazard, Kernel function, Hypothesis testing, Counting process

NONPARAMETRIC INFERENCES FOR THE HAZARD FUNCTION WITH RIGHT TRUNCATED DATA

by

HACI MUSTAFA AKCIN

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2013

NONPARAMETRIC INFERENCES FOR THE HAZARD FUNCTION WITH RIGHT TRUNCATED DATA

by

HACI MUSTAFA AKCIN

Committee Chair: Dr. Xu Zhang

Committee: Dr. Gengsheng Qin

Dr. Yichuan Zhao

Dr. Vijay Ganji

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2013

DEDICATION

To Simona, Amina, Omer and Davud

ACKNOWLEDGEMENTS

During my doctoral study and writing this dissertation, I have blessed by receiving help and support from many people. Although it is hard to thank everyone by name, I will still try to show my gratitude towards people that I received greater support.

First, I would like to thank my advisor, Dr. Xu Zhang. I should note that I have also completed my Masters degree under Dr. Zhang's supervision. I decided to have Dr. Zhang as my PhD advisor after experiencing her amazing guidance, in-debt knowledge of statistics, quality of research she does and her attention to details. Pursuing a PhD degree is a great challenge and Dr. Zhang's help and support motivated me tremendously. Thank you, Dr.Zhang.

I would also like to acknowledge other members of my dissertation committee, Dr. Gengsheng Qin, Dr. Yichuan Zhao and Dr. Vijay Ganji, for accepting to be part of my journey and their constructive comments. Thank you. I also would like to mention my Director and my mentor at CDC Dr. Carol Gotway-Crawford for her continuous support. Thank you for believing me.

My special thanks go to all of my friends that I had pleasure to spend time with such as Xin Huang, Haochuan (Harris) Zhou, Meng (Peter) Zhao and others too many to list.

Finally, I would like to thank my parents for their unconditional support and most importantly special thanks to love of my life Simona-Daniela who was with me during the worst days of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

BACKGROUND AND LITERATURE REVIEW

Survival analysis mainly deal with time-to-event data. The statistical interest is the survival quantities of the time-to-event variable such as the survival function, hazard rate function and cumulative hazard function. Statisticians often have to deal with incompleteness problem when analyzing time-to-event data. Censoring and truncation are the common reasons for incompleteness. Censoring refers to the scenario that the lifetime of an individual is known to stay in certain time interval but the exact lifetime is unknown. Truncation relates to the problem that the lifetime of an individual is unobservable because it is smaller (or greater) than the truncation time.

The counting process methodology has helped statisticians to develop inferences for censored and/or truncated time-to-event data. Basic properties of counting process and the martingale central limit theory are briefly presented in this chapter. This dissertation aims to develop inferences for right truncated data. Important works related to truncated data, especially analysis of right truncated data are reviewed in this chapter.

This chapter consists of three sections. The first section introduces the concepts of counting process, filtration, martingales and predictable variation process followed by the Nelson-Aalen estimator of the cumulative hazard function and the Kaplan-Meier estimator of the survival function. The second section presents some details of censoring and truncation, including three types of censoring, left and right truncation. The literature review on the inferences of truncated data is given in the third section.

1.1 Basic Properties

1.1.1 Filtrations

Define a time interval $[0, \tau]$ where $0 < \tau \leq \infty$. Let $(\Omega, \mathcal{F}, P)$ be a probability space. The σ-algebra $\mathcal{F}$ is a set of events (subsets of Ω), $\mathcal{F} \subseteq 2^{\Omega}$ and satisfies following conditions:

(1) $\emptyset \in \mathcal{F}$

(2) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, where $A \cup A^c = \Omega$

(3) If $A_i \in \mathcal{F}$ then $\cup_i A_i \in \mathcal{F}$, $i = 1, 2, ..., n$.

The probability measure P is a function on $\mathcal{F}$, $P : \mathcal{F} \to [0, 1]$. Let $t \in [0, \tau]$, then a filtration $\mathcal{F}_t$ is defined as an increasing right-continuous family of sub-σ-algebras of $\mathcal{F}$ on Ω [5](pp. 60-61). A filtration is also known as a history. Completeness is an important concept of filtration with the definition that σ-algebra $\mathcal{F}_t$ contains all σ-algebras on Ω for each $t \in [0, \tau]$. It was noted that assumption of completeness can be omitted. Jacod [31] discussed that the completeness assumption is not necessary while Von Weiszacker and Winkler [60] developed the whole theory without this assumption. Under the assumption of completeness, $\mathcal{F}_t$ is increasing and right-continuous.

Let $\mathcal{F}_t$ be a filtration on Ω then a random variable $S : \Omega \to [0, \tau]$ is called a stopping time if $(S \leq t) \in \mathcal{F}_t, \forall t \in [0, \tau]$. A stopping time is also called as a hitting time where a process hits a predetermined set for the first time [17]. One example of traffic lights can make stopping time easier to understand. Assume that there are only two lights, green and red. Once light runs to red, it stays red forever and it cannot be green (i.e. stopped once, stopped forever). Consider three options for the traffic light:

(1) it is red forever (stopped immediately)

(2) it is green at the beginning but turns red and stays red (stopped at some point)

(3) it stays green forever (never stopped)

The green-red process described here is a simple example for stopping time. The traffic light can only change at t based on the information up to t.

1.1.2 Martingale theory

In the second half of 18^{th} century, martingale referred to an unusual betting practice for a coin toss in Monte Carlo [45]. In this simple game, coin with head up represented winning the game. The strategy for the gambler was to double his bet after every loss to recover previous losses and win a sum equal to the first bet. The expected winnings will sum up to zero. As time approaches to infinity, probability of having a coin with head up would be 1, so winning was guaranteed. However, the strategy only works in case of unlimited resources of money and time. Because of bets were doubled for each game, gambler would eventually go bankruptcy.

The martingale theory in probability was introduced by Levy [42] and improved by Doob [13]-[14], Ito [28] and Meyer [48] among others. A sequence of random variables $X_1, X_2, \cdots$ is called a martingale if $E(|X_n|) < \infty$ and $E(X_{n+1}|X_1, \cdots, X_n) = X_n$. It is another way of saying that the expected value of next observation given all the past observations is equal to the last observation. Using linearity of expectation, $E(X_{n+1}|X_1, \cdots, X_n) - X_n = 0$ which means average winnings to be zero. In the case where the value of last observation X_n is no more than or no less than the expected value of next observation given all the past observations, we will have submartingales and supermartingales respectively. A sequence of random variables $X_1, X_2, \cdots$ is called a submartingale if $E(|X_n|) < \infty$ and $E(X_{n+1}|X_1, \cdots, X_n) \geq X_n$ and a supermartingale if $E(|X_n|) < \infty$ and $E(X_{n+1}|X_1, \cdots, X_n) \leq X_n$.

In more general form, a martingale M is integrable and the expected value of M at $t \in [0, \tau]$ given the filtration(history) is equal to the value right before t. In other words, $E(|M(t)|) < \infty$, and

$$E(M(t)|\mathcal{F}_s) = M(s) \tag{1.1}$$

where $\forall s \leq t$ and $\forall t \in [0, \tau]$. M is called a submartingale if $E(M(t)|\mathcal{F}_s) \geq M(s)$ and a supermartingale if $E(M(t)|\mathcal{F}_s) \leq M(s)$.

A martingale M is called a square integrable martingale if $\sup E(M^2(t)) < \infty$. Let

S_n be a sequence of increasing stopping times, $S_n : \Omega \longrightarrow [0, \infty]$. M is called a local martingale if the stopped process $M(t)^{S_n}$ is a martingale for each n. M is called a local square integrable martingale if $M(t)^{S_n}$ is a square integrable martingale. Similarly, M is called a local submartingale (or supermartingale) if $M(t)^{S_n}$ is a submartingale (or supermartingale) for each n. M is called a local square integrable submartingale (or supermartingale) if $M(t)^{S_n}$ is a square integrable submartingale (or supermartingale).

1.1.3 Predictable and optional variation processes

Let $M(t)$ and $M^{'}(t)$ be a local square integrable martingale and $V(t) = \langle M \rangle(t)$ be a process such that

(1) $V(t)$ is predictable (i.e. $V(t)$ is $\mathcal{F}_t$ measurable), and

(2) $M^2(t) - V(t)$ is a local martingale with respect to $\mathcal{F}_t$,

then $V(t)$ is called the predictable variation process of $M(t)$ and denoted as $\langle M \rangle(t)$. Similarly, let $V^{'}(t) = \langle M, M^{'} \rangle(t)$ be a process such that $V^{'}(t)$ is predictable and $M(t)M^{'}(t) - V^{'}(t)$ is a local martingale with respect to $\mathcal{F}_t$ then $\langle M, M^{'} \rangle(t)$ is called the predictable covariation process of $M(t)$ and $M^{'}(t)$.

The predictable variation process of $M(t)$ can also be written as $\sum \text{var}(M(t_i) - M(t_{i-1})|F_{t_{i-1}})$ where $i = 1, \cdots, n$. If we ignore the conditional expectation and only take the sums of squares, $\sum (M(t_i) - M(t_{i-1}))^2$, then the process is called the optional variation process. Formally, we can denote the optional variation process as $[M](t)$. Assume $M(t)$ is a local martingale ($M(t)$ does not have to be local square integrable),

$$[M](t) = M(t)^2 - 2 \int_0^t M(s^-) dM(s)$$

and the optional covariation process of $M(t)$ and $M^{'}(t)$

$$[M, M^{'}](t) = M(t)M^{'}(t) - \int_0^t M(s^-) dM^{'}(s) - \int_0^t M^{'}(s^-) dM(s),$$

where $s \leq t$ and $M(t)^2 - [M](t)$ is a local martingale.

1.1.4 Counting processes

A counting process is a stochastic process that counts the number of discrete events. Bremaud [7] was one of the pioneers who defined counting process by showing that integrated intensity process of counting process is actually its compensator. Aalen [1] studied the statistical inferences of counting processes. Developments of counting process theory was made by Jacod [29]-[30]. Andersen *et al.* [4] explained the notion of starting a counting process.

A counting process $N(t)$ satisfies the following conditions: nondecreasing with jumps of size 1, $N(0) = 0$, sample paths of $N(t)$ is right continuous and $P(N(t) < \infty) = 1$ [38] (p. 79).

The counting process $N(t)$ has a compensator $\Lambda(t)$ which also is a predictable process such that $M(t) = N(t) - \Lambda(t)$ is a local square integrable martingale. The predictable variation process of $M(t)$ is given by

$$\langle M \rangle(t) = \Lambda(t) - \int_0^t \Delta\Lambda(s) d\Lambda(s)$$

and $\langle M \rangle(t) = \Lambda(t)$ when $\Lambda(t)$ is continuous.

1.1.5 The martingale central limit theory

The martingale central limit theorem for discrete time was first considered by Billingsley [6] and followed by Brown [9] and Dvoretsky [15] among others. Aalen [1] extended this work to continuous-time context. Rebolledo [55] and Fleming and Harrington [19] were among the first mathematicians who studied a general continuous-time martingale central limit theorem.

Although there are many versions of martingale central limit theorem, the theorem proposed by Rebolledo [55] is commonly employed for the inferences related to survival data [5] (p. 83)

Let $M^{(n)} = (M_1^{(n)}, \cdots M_p^{(n)})$ be a vector of p local square integrable martingales

for each n and assume $M_\epsilon^{(n)}$ be a vector of p local square integrable martingales where $\epsilon > 0$ and $|\Delta M_k^{(n)} - \Delta M_{\epsilon,k}^{(n)}| \leq \epsilon$ where $k = 1, 2, \cdots, p$. Also, let $M^{(\infty)}$ be a Gaussian martingale where $\langle M^{(\infty)} \rangle = [M^{(\infty)}] = \sigma^2(t)$ and $M^{(\infty)}(t) - M^{(\infty)}(s) \sim N(0, \sigma^2(t) - \sigma^2(s))$. Further assume $\forall t \in T_0$ for $T_0 \subseteq T$. As $n \to \infty$, if

$$\langle M^{(n)} \rangle(t) \to^P \sigma^2(t) \tag{1.2}$$

and

$$\langle M_{\epsilon k}^{(n)} \rangle(t) \to^P 0, \tag{1.3}$$

then

$$(M^{(n)}(t_1), \cdots, M^{(n)}(t_q)) \to^D (M^{(\infty)}(t_1), \cdots, M^{(\infty)}(t_q)), \tag{1.4}$$

where $\forall t_1, \cdots, t_q \in T_0$.

1.1.6 The Nelson-Aalen estimator

Estimation of the cumulative hazard function with censored failure time data was first studied by Nelson [50]-[51] and Altshuler [3]. It was extended to counting process models by Aalen [1]-[2] and the proposed estimator is commonly known as the Nelson-Aalen estimator.

Consider a sample with random variables $T_1, T_2, \cdots, T_n$. Note that all event times are observed. We can define the counting processes $N_i(t) = I(T_i \leq t)$, $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ and let $Y_i(t) = I(T_i \geq t)$, then $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ is the risk set. Suppose that the counting process $\bar{N}(t)$ has the intensity process $\lambda(t) = \alpha(t)\bar{Y}(t)$, where $\alpha(t)$ is the hazard rate function. The compensator of $\bar{N}(t)$ is $\Lambda(t) = \int_0^t \lambda(s)ds$. Denote the cumulative hazard function as

$$A(t) = \int_0^t \alpha(s)ds.$$

The local square integrable martingale, $\bar{M}(t) = \sum_{i=1}^{n} M_i(t)$, is given by

$$\bar{M}(t) = \bar{N}(t) - \Lambda(t) = \bar{N}(t) - \int_0^t \alpha(s)\bar{Y}(s)ds$$

and $d\bar{N}(t) = \alpha(t)\bar{Y}(t) + d\bar{M}(t)$ where $\bar{M}(t)$ is a process of random noise. The Nelson-Aalen estimator of $A(t)$ is given by

$$\widehat{A}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)} \tag{1.5}$$

1.1.7 The Kaplan-Meier estimator

The Kaplan-Meier estimator [35] is used for estimating the survival function of a random variable. A heuristic explanation of the estimator is as follows: to estimate the survival function at a time point t, one needs to divide $[0, t]$ into smaller intervals based on distinct event times and find the proportions of survival for each interval. Multiplying these proportions together gives an estimate of the survival probability at t. Therefore, it is also known as the product-limit estimator.

The survival function of a continuous random variable T, $S(t)$, is defined by

$$S(t) = \prod_{s \le t}(1 - dA(s)) = \exp\left(-\int_0^t \alpha(u)du\right).$$

One can plug in the Nelson-Aalen estimator to obtain the Kaplan-Meier estimator

$$\widehat{S}(t) = \prod_{s \le t}(1 - d\widehat{A}(s)) = \prod_{s \le t}\left(1 - \frac{d\bar{N}(s)}{\bar{Y}(s)}\right).$$

1.2 Censoring and Truncation

Incompleteness is a common feature of time-to-event data. The reasons for incompleteness include censoring and truncation. In censoring, the lifetime of an individual is known to stay in certain interval but the exact lifetime is unknown. In truncation, the lifetime of an

individual is unobservable because it is smaller (or greater) than the truncation variable. In short, censoring means partial information about the lifetime while truncation means unobservability of the lifetime. Left censoring, right censoring and interval censoring are different types of censoring whereas left truncation and right truncation are categories of truncation. The details about censoring and truncation are presented in this section.

1.2.1 Censoring

Consider counting processes of n individuals $N_i(t) = I(T_i \leq t)$ where $i = 1, \cdots, n$. Right censoring happens if the event occurs after a random time C_i. We can observe $N_i(t)$ only if $T_i \leq C_i$ and it is censored if $T_i > C_i$.

Let R_i be a right-censoring process then we can define $R_i(t) = I(t \leq C_i)$. Let $N_i^{RC}(t)$ denote the right censored counting process, in other words, the observable part of counting process for individual i,

$$N_i^{RC}(t) = \int_0^t R_i(s) dN_i(s).$$

The simple idea behind the above equation is that the event is observed if it happens before C_i. Let $X_i = \min(T_i, C_i)$ be a random variable that gives the information on the smaller one between the event time and the censoring time. Let $\Delta_i = I(X_i = T_i)$. $\Delta_i = 1$ when the event happens before the censoring time, $T_i \leq C_i$; $\Delta_i = 0$ when the censoring is observed, $T_i > C_i$. The right-censored counting process can also be written as $N_i^{RC}(t) = I(X_i \leq t, \Delta_i = 1)$.

Right censoring is the most common type of incompleteness. It also branches out to two subcategories: Type I and Type II censoring. Assume that all subjects enter the study same time. Type I censoring occurs if the event of interest is observable only if it occurs before a predetermined time c_0. Here, the censoring time is same for each individual. So right-censoring process $R_i(t) = I(t \leq c_0)$ is nonrandom and predictable. A more common Type I censoring is related to the context that individuals enter the study at different times. In a clinical trial, researchers usually end the study at a predetermined date. Event times are observed on patients who have events before this study closing date while event times of remaining patients are censored. The observed censoring times differ by subjects.

Type II censoring arises when the study continues until occurrence of the rth event, where $r \leq n$. In other words, $C_i = T_{(r)}, i = 1, \cdots, n$ and $R_i = I(t \leq T_{(r)})$ is predictable where $T_{(r)}$ is the time to the rth event. In an electronics factory, engineers may want to analyze the life-time of a certain component. They continue to monitor a sample of n components until occurrence of rth events (in this case failure) [38](p. 67). We should note that in Type II censoring, the observed times $X_1, \cdots, , X_n$ are dependent.

One special type of right censoring occurs in the context of competing risks. Competing risks are exclusive causes of failure. Failing due to one cause precludes failing from any other causes. For example, during the follow up of a cohort of breast cancer patients, patients may die from non-cancer causes such as stroke or pneumonia. When the study emphasis is the hazard of cancer failure, deaths from non-cancer causes are treated as censoring.

Left censoring is not as common as right censoring in real applications. In left censoring, there is information that the event time T_i happens before the censoring time C_i but the exact T_i is unknown. We can define $X_i = \max(T_i, C_i)$ and $\Delta_i = I(X_i = T_i)$. Following the similar format as the right-censored counting process, we can write the left-censored counting process as $N_i^{LC}(t) = I(X_i \geq t, \Delta_i = 1)$. One example of left censoring is survey study conducted on high school students to find out the ages when they started smoking cigarettes. Some students may not remember the exact ages therefore the age of smoking cigarettes is left censored by the age at the study.

Interval censoring often occurs in longitudinal studies. Interval censoring can be understood as a general type of censoring with left and right censoring as special cases. Interval censoring refers to the scenario that the event time T_i falls in an interval (C_i^L, C_i^R) without observing exact T_i. Define $X_i = \max[\min(T_i, C_i^R), C_i^L]$ and $\Delta_i = I(X_i = T_i)$. $\Delta_i = 1$ when the exact event time is observed and $\Delta_i = 0$ when the event time is interval censored. Interval censoring becomes left censoring if the interval is $(0, C_i^L)$ and it becomes right censoring if it is (C_i^R, ∞).

1.2.2 Truncation

Truncation is another type of incompleteness occurring in time-to-event data. It is very different from censoring. An event time falling outside of an interval can not be observed, therefore it is not included in the sample and hence truncated. Concept of truncation differs from censoring in that partial information about event time is available for censoring but an event time is not observable for truncation. There are two types of truncation: left truncation and right truncation.

In left truncation, only the event time variable greater than the truncation variable is observable. The event time variable less than the truncation variable is unobservable and hence truncated. Right truncation occurs if one can observe an event time if it is less than the truncation variable. The event time variable greater than the truncation variable is unobservable and truncated.

The life tables constructed by Halley [27] was one of the earliest applications of left truncation. He recorded birth, death and cause of death of the individuals in the city of Breslow, United Kingdom. Let L be the time interval from date of birth till date of recording and L is the age at recording for a study participant. Let T be the age of death. Individuals could be recorded only if they were alive at the time of study and whomever died before this time were not captured in the study. In mathematical language, if $T < L$ then T is left truncated by L. Similarly, Kaplan and Meier [35] constructed a life table where they recorded entrance (L) and exit (T) ages of individuals. Additional information on exit was recorded to indicate whether the exit was due to death or right censoring. Kaplan and Meier named left truncation as *delayed entry* since individuals could only be observable since they enter the study. In this case, T is only observable if $L < T$ and T is left truncated by L.

Hald [25]-[26] was among the earliest statisticians discussing the concepts of censoring and truncation. Kaplan and Meier [35] were the first to contribute the survival function estimator with left truncated and right censored time-to-event data. There has been increasing attention on the truncation issue of survival data in recent years. There are still gaps for the inferences with truncated data.

1.3 Literature Review for the Inferences with Truncated Data

Incompleteness introduces challenges for analyzing time-to-event data. Let (L, T) be random variables with constraint of $L \leq T$. Under random truncation, T is left-truncated by L while L is right-truncated by T. Most of the literature about random truncation focuses on left truncation. In left truncation, T is the event time and of study interest while L is defined as study entrance time. Kaplan-Meier [35] described left truncation as late entrance.

Truncated samples involve biased sampling, since the probability of selection depends on the length of the variable. Kalbfleisch and Lawless [33] noted that the right truncation occurs in AIDS blood transfusion infected data that were originally collected by Centers for Disease Control and Prevention (CDC) in 1980's. CDC required all AIDS cases to be reported, after the (Morbidity and Mortality Weekly Report) MMWR of description of five cases defined as pneumocystis carinii pneumonia (PCP) which would later become to known as acquired immunodeficiency syndrome (AIDS) [49].

The blood transfusion infected AIDS data set contains information with diagnosis of AIDS cases up to July 1, 1986. The variable of interest to estimate is the incubation time of AIDS. The incubation time is defined as the duration between infection with HIV and onset of clinical AIDS. Infection date can only be determined if infection caused by blood transfusion. Since the closing date of study is July 1, 1986, the CDC data can only capture the cases if the diagnosis date of clinical AIDS is earlier than July 1, 1986. In other words, the incubation time of AIDS should be less than the duration between the infection date and the closing date. In mathematical framework, let L be the incubation time and T be the duration between infection date and the closing date then the individuals with AIDS only be observable if $L < T$ which is defined as right truncation.

Due to its truncated characteristic, the blood transfusion infected AIDS data set has been analyzed by many researchers worked in random truncation field such as Lui *et al.* [43], Medley *et al.* [46]-[47], Kalbfleisch and Lawless [33]-[34] among others. The truncated version of Kaplan-Meier [35] estimator routinely used to estimate the distribution of L and

T[66]. The consistency and asymptotic properties of truncated version of Kaplan-Meier [35] estimator studied by Woodroofe [66], Wang, Jewell and Tsai [63], Keiding and Gill [36], and Chen, Chao and Lo [11]. The weak convergence established by Chao and Lo [10] after presenting the independent and identically distributed representation of the left truncated version of the Kaplan-Meier estimator. Lai and Ying [41] modified the Kaplan-Meier estimator when distribution function is not continuous for data subject to truncation by an independent but not necessarily identically distributed random variable. Gurler, Stute and Wang [24] presented a strong representation of the empirical quantile function for left truncated data. Uzunogullari and Wang [59] studied the kernel estimators of the hazard rate for left truncated/ right censored data. They particularly choose adaptive bandwidth to get smoother curves and more precise estimation result. Regression analysis under left truncation and right censoring was studied by Klein and Zhang [39].

Right truncation has been routinely tackled by transforming it to left truncation. Let τ be a large constant. The transformed variable $\tau - L$ is left truncated by $\tau - T$. Using this relationship, the distribution function of L coincides with the survival function of the transformed variable $\tau - L$, and Kaplan-Meier estimator became the natural estimation method [40], [66], [36]. In recent years, Chi *et al.* [12] developed a test to compare integrated weighted differences between two survival functions. Another important survival quantity related to the transformed variable $\tau - L$ is its hazard rate function. This function is commonly interpreted as a hazard rate function with τ as the origin and counted backwards along the time axis. Therefore, it is known as reverse-time hazard or retro-hazard. Lagakos *et al.* [40] proposed a weighted log-rank test to compare the reverse-time hazard rates. Gross and Huber-Carol [23] and Kalbfleisch and Lawless [33] studied Cox regression of the reverse-time hazard rate. The standard Nelson-Aalen estimator is applicable for estimating the cumulative reverse-time hazard.

Interpretation of the reverse-time hazard function was noted to be difficult and unnatural [16]. In recent years, inferences about regular forward-time hazard draw more attention. Finkelstein *et al.*[16] studied the proportional hazards model. Chi *et al.*[12] developed a

two-sample test to compare survival functions. Shen [56] proposed a semiparametric test to compare weighted forward-time cumulative hazards functions where he suggested a resampling approach to estimate the variance. Jiang [32] studied two Nelson-Aalen type variance estimators of the forward-time cumulative hazard function.

Jiang [32] mentioned variance estimators increase dramatically when t approaches to the largest time of L. In Chapter 2, those variance estimators are slightly modified to improve the estimation and replicated the simulation study to validate it. One-sample weighted log-rank test is introduced in the following sections. In Chapter 3, the research is extended to hypothesis testing of K-sample and two-sample cases. The weighted log-rank test is developed for right truncated data in forward-time.

In general, inference about the hazard rate function under right truncation is scarce. In Chapter 4, the hazard rate function is directly estimated and subsequently the nonparametric inferences are developed. Our motivation for estimating the hazard rate function relies on the dynamic characteristic of this function and it gives more precise information about distribution than any other quantity such as cumulative distribution function, survival function or cumulative hazard function.

CHAPTER 2

NONPARAMETRIC INFERENCE FOR THE CUMULATIVE HAZARD FUNCTION WITH RIGHT TRUNCATED DATA

2.1 Motivation

Chapter 1 contains the literature review of the inferences related to truncated survival data. The existing methods for analysis of right-truncated data focus on the reverse-time hazard function. However, this quantity lacks of natural interpretation. In recent years, there has been an increasing interest on the inferences of the forward-time quantities. This chapter focuses on the nonparametric inferences of the forward-time cumulative hazard function.

This chapter organized as follows. Chapter 2.2 introduces the existing nonparametric inference for the reverse-time cumulative hazard function. Nonparametric inference of the cumulative hazard function was recently studied by Jiang [32]. Chapter 2.3 shows the similar procedure as Jiang but emphasizes on two modified variance estimators of estimated cumulative hazard function. Chapter 2.4 presents one-sample weighted log-rank test to compare the mortality rate of the study population to the known rate. Chapter 2.5 consists of two simulation studies designed to assess the performances of proposed methods. The final discussion is given in Chapter 2.6.

2.2 Nonparametric Inference for the Reverse-time Cumulative Hazard Function

One primary objective in survival analysis is to assess instantaneous, as well as cumulative, risk of failure. Censoring and truncation makes analysis of time-to-event data cumbersome. Analysis of truncated data is of primary interest in this dissertation. Throughout the dissertation, the univariate truncated sample is denoted as $\{L_i, T_i\}$, $i = 1, 2, ..., n$, and $L_i \leq T_i$. The variable of study interest, L, is right truncated by the truncation variable, T. Let $\alpha(t)$ and $A(t)$ and be the hazard rate and cumulative hazard function of L, respectively.

Their mathematical definitions are given by

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P[t \leq L < t + \Delta t | L \geq t]}{\Delta t} \tag{2.1}$$

and

$$A(t) = \int_0^t \alpha(u)du = \int_0^t \frac{dG(u)}{P(L \geq u)}, \tag{2.2}$$

respectively, where $G(t)$ is the distribution function of L that $G(t) = P(L \leq t)$.

In a truncated sample, right truncation can be easily transformed to become left truncation. Let τ be a large constant greater than $\max\{T_1, \cdots, T_n\}$ and consider the transformed random variables with $L^* = \tau - L$, $T^* = \tau - T$. For the newly constructed sample $\{L_i^*, T_i^*\}, i = 1, \cdots, n$, there is the constraint $L_i^* > T_i^*$. Therefore, the variable L^* is left truncated by the variable T^*. The hazard rate function of L^* is a quantity with τ as its origin and counted backwards along the time axis towards zero. As a result, such a quantity is called as "reverse-time hazard" by Lagakos *et al.* [40] or "retro hazard" by Keiding and Gill [36]. Let $\alpha^*(t)$ and $A^*(t)$ denote the reverse-time hazard rate and cumulative hazard function, respectively with the definitions [36]

$$\alpha^*(t) = \lim_{\Delta t \to 0} \frac{P[t \geq L^* > t - \Delta t | L^* \leq t]}{\Delta t} \tag{2.3}$$

and

$$A^*(t) = \int_t^\tau \alpha^*(u)du = \int_t^\tau \frac{dG(u)}{P(L \leq u)}. \tag{2.4}$$

$A^*(t)$ can be estimated by the Nelson-Aalen estimator. A clear definition about the reverse-time martingale is needed in order to establish the inference of the Nelson-Aalen estimator. For a truncated sample, define the following counting processes $N_i^L(t) = I(L_i \geq t)$, $Y_i(t) = I(L_i \leq t \leq T_i)$ and let $\bar{N}^L(t) = \sum_{i=1}^n N_i^L(t)$, $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$. The counting

process $N_i^L(t)$ is defined to count an event via the reversed time-axis. The corresponding martingale is given by

$$M_i^*(t) = N_i^L(t) - \int_\tau^t Y_i(u) dA^*(u). \tag{2.5}$$

It is the standard result that $M^*(t)$ is a local square integrable martingale [36]. Consider $A^{*+}(t) = \int_\tau^t \alpha^*(u) J(u) du$, where indicator process $J(t) = I(\bar{Y}(t) > 0)$. It is obvious that $A^{*+}(t)$ is almost equal to $A^*(t)$ if there is a small probability that $\bar{Y}(s) = 0$ for some $s \leq t$. Let $N_i(t) = I(L_i \leq t)$ and $\bar{N}(t) = \sum_{i=1}^n N_i(t)$. The Nelson-Aalen estimator of $A^*(t)$ is given by

$$\widehat{A}^*(t) = \int_\tau^t \frac{J(u)}{\bar{Y}(u)} d\bar{N}^L(u) = \int_t^\tau \frac{J(u)}{\bar{Y}(u)} d\bar{N}(u), \tag{2.6}$$

when $\bar{Y}(t) = 0$, $J(t)/\bar{Y}(t)$ is defined as 0. It follows that

$$\widehat{A}^*(t) - A^*(t) = \int_\tau^t \frac{J(u)}{\bar{Y}(u)} d\bar{M}^*(u), \tag{2.7}$$

where $\bar{M}^*(t) = \sum_{i=1}^n M_i^*(t)$. Keiding and Gill [36] studied the weak convergence of $\sqrt{n}[\widehat{A}^*(t) - A^*(t)]$ based on the martingale central limit theorem. $\sqrt{n}[\widehat{A}^*(t) - A^*(t)] \to_{\mathcal{D}} U_t$ where U_t is a Gaussian process with mean zero and variance $\int_t^\tau \alpha^*(u) du / y(u)$ where $y(u) = E[\bar{Y}(u)/n]$. Based on the properties of martingales, the predictable variation process for $\widehat{A}^*(t) - A^*(t)$ is

$$\langle \widehat{A}^* - A^* \rangle(t) = \int_\tau^t \frac{J(u)\alpha^*(u)}{\bar{Y}(u)} du. \tag{2.8}$$

Optional variation process can be used to estimate the variance of $\widehat{A}^*(t)$,

$$\hat{\text{var}}^{(1)}[\widehat{A}^*(t)] = \int_t^\tau \frac{J(u) d\bar{N}(u)}{\bar{Y}(u)^2}. \tag{2.9}$$

An alternative variance estimator developed by Klein [37] by assuming a binomial dis-

tribution for a jump in the event counting process. Using (2.5) and definition of predictable variation process in (1.1.3), the alternative variance estimator of $\widehat{A}^*(t)$ is

$$\hat{\text{var}}^{(2)}[\widehat{A}^*(t)] = \int_t^\tau J(u)\frac{\bar{Y}(u) - \Delta\bar{N}(u)}{\bar{Y}(u)^3}d\bar{N}(u). \tag{2.10}$$

2.3 Nonparametric Inference for the Cumulative Hazard Function

Estimation of the distribution function of L has been well studied. The cumulative distribution function of L, $G(t)$, can also be viewed as the survival function of L^* in reverse-time axis.

$$G(t) = P(L \leq t) = P(L^* \geq \tau - t). \tag{2.11}$$

Therefore, $G(t)$ can be estimated by Kaplan-Meier estimator [66], [36],

$$\widehat{G}(t) = \prod_{u>t}\left(1 - \frac{d[\sum_{i=1}^n I(L_i \leq u)]}{\bar{Y}(u)}\right). \tag{2.12}$$

It is known as the right truncated version of Kaplan-Meier estimator. Under the context of right truncation, Nelson-Aalen estimator of the cumulative hazard function is not applicable. Instead, one has to consider a plug-in estimator to estimate (2.2)

$$\widehat{A}(t) = \int_0^t \frac{d\widehat{G}(u)}{1 - \widehat{G}(u-)}, \tag{2.13}$$

where $\widehat{G}(u^-)$ is the Kaplan-Meier estimate of $P(L \leq t)$ prior to u.

It is necessary to establish a relationship between $A(t)$ and $A^*(t)$. Transforming the right truncated data to the left truncated data enables usage of existing inferences for left truncated data. The relation between reverse- and forward-time hazard functions was discussed by

Lagakos *et al.* [40] for two independent samples. They noted that

$$\frac{\alpha_1^*(t)}{\alpha_2^*(t)} = \frac{\alpha_1(t)\{1-G_1(t)\}G_2(t)}{\alpha_2(t)\{1-G_2(t)\}G_1(t)}, \tag{2.14}$$

where the subscript 1 or 2 is added to indicate that the quantity is associated with sample 1 or 2, respectively. Equation (2.14) suggests that whenever forward-time hazards have a constant ratio, proportionality of reverse-time hazards is violated. Based on the equations that $G(t) = \exp[-A^*(t)]$ and $1 - G(t) = \exp[-A(t)]$, we can clarify the relation between $A(t)$ and $A^*(t)$ as

$$A(t) = -\log(1 - \exp[-A^*(t)]). \tag{2.15}$$

It has been mentioned in the previous section that $\sqrt{n}\{\widehat{A}^*(t) - A^*(t)\} \to_{\mathcal{D}} U_t$, where U_t is a mean-zero Gaussian process. Applying the generalized delta method, we can get

$$\sqrt{n}\{\widehat{A}(t) - A(t)\} \to_{\mathcal{D}} \kappa(A^*(t))U_t \tag{2.16}$$

where

$$\kappa(\widehat{A}^*(t)) = -\frac{\exp(-A^*(t))}{1-\exp(-A^*(t))} = -\frac{G(t)}{1-G(t)}.$$

Based on this asymptotic result, the variance of $\widehat{A}(t)$ can be approximated by the following formula,

$$\text{var}[\widehat{A}(t)] \approx \left[\frac{G(t)}{1-G(t)}\right]^2 \text{var}[\widehat{A}^*(t)]. \tag{2.17}$$

The naive and alternative variance estimators of $\widehat{A}^*(t)$ given in Equations (2.9) and (2.10) can be plugged into (2.17), leading to the variance estimators of $\widehat{A}(t)$. Jiang [32] proposed the following two estimators,

$$\tilde{\text{var}}^{(1)}[\widehat{A}(t)] = \left[\frac{\widehat{G}(t)}{1-\widehat{G}(t)}\right]^2 \int_t^\tau \frac{J(s)d\bar{N}(s)}{\bar{Y}(s)^2}$$

and

$$\tilde{\text{var}}^{(2)}[\widehat{A}(t)] = \left[\frac{\widehat{G}(t)}{1-\widehat{G}(t)}\right]^2 \int_t^\tau J(s)\frac{\bar{Y}(s)-\Delta\bar{N}(s)}{\bar{Y}(s)^3}d\bar{N}(s).$$

A simulation study conducted by Jiang [32] suggested that the variance estimates using the above formulas overestimate the actual variance when t is large. This problem can be fixed by using $1-\widehat{G}(t^-)$ instead of $1-\widehat{G}(t)$. Please note that the same form, $1-\widehat{G}(t^-)$, was also used in the plug-in estimator given in Equation (2.13). Here, two modified variance estimators are presented and will be investigated in the simulation study. The naive variance estimator is given by

$$\hat{\text{var}}^{(1)}[\widehat{A}(t)] = \left[\frac{\widehat{G}(t)}{1-\widehat{G}(t^-)}\right]^2 \int_t^\tau \frac{J(s)d\bar{N}(s)}{\bar{Y}(s)^2} \tag{2.18}$$

and the alternative variance estimator is given by

$$\hat{\text{var}}^{(2)}[\widehat{A}(t)] = \left[\frac{\widehat{G}(t)}{1-\widehat{G}(t^-)}\right]^2 \int_t^\tau J(s)\frac{\bar{Y}(s)-\Delta\bar{N}(s)}{\bar{Y}(s)^3}d\bar{N}(s). \tag{2.19}$$

2.4 One-sample Weighted Log-rank Test

Let $\alpha_0(t)$ denote a known hazard rate function. The aim of study is to assess whether the hazard rate of the univariate sample equals to the known rate. The null hypothesis can be written as $H_0 : \alpha(t) = \alpha_0(t)$, where $\alpha_0(t)$ is the known hazard rate function. Define $A_0(t)$ as the cumulative hazard function associated with $\alpha_0(t)$. Let's revisit basic definitions, then

$$A(t) = \int_0^t \alpha(u)du \qquad \text{and} \qquad A_0(t) = \int_0^t \alpha_0(u)du.$$

We can define the test statistic as

$$Z(t) = \int_0^t W(u)d[\widehat{A}(u) - A_0(u)], \tag{2.20}$$

where $W(t)$ is a stochastic weight function. A common choice of $W(t)$ is the risk set process,

$\bar{Y}(t)$. When this weight function is chosen, the test becomes the one-sample log-rank test. For the complete data or the survival data subject to right censoring and/or left truncation, the log-rank test statistic is exactly the difference between observed number of events and expected number events. However, this type of interpretation is not obtainable under right truncation. The test discussed in this section is a closed form of log-rank test.

$Z(t)$ is a local square integrable martingale and we can derive the predictable variation process of $Z(t)$ with details given in Appendix A1. Based on the martingale central limit theorem, it can be proved that $Z(t) \to W_t$, where W_t is a Gaussian process with mean zero and variance $\sigma^2(t)$,

$$\sigma^2(t) = \int_t^0 \left[W(s) \frac{G_0(s)}{1 - G_0(s)} - \int_0^s W(u) d \left(\frac{G_0(u)}{1 - G_0(u)} \right) \right]^2 \frac{\alpha_0^*(s) ds}{y(s)}$$

$$+ \int_\tau^t \left[\int_0^t W(u) d \left(\frac{G_0(u)}{1 - G_0(u)} \right) \right]^2 \frac{\alpha_0^*(s) ds}{y(s)}. \tag{2.21}$$

The variance of $Z(t)$ can be estimated as

$$\widehat{\sigma}^2(t) = \int_0^t \left[W(s) \frac{G_0(s)}{1 - G_0(s^-)} - \int_0^s W(u) d \left(\frac{G_0(u)}{1 - G_0(u^-)} \right) \right]^2 \frac{d\bar{N}(s)}{\bar{Y}^2(s)}$$

$$+ \int_t^\tau \left[\int_0^s W(u) d \left(\frac{G_0(u)}{1 - G_0(u^-)} \right) \right]^2 \frac{d\bar{N}(s)}{\bar{Y}^2(s)}. \tag{2.22}$$

The asymptotic distribution of $Z(t)$ is discussed in Appendix A1. Distribution of the statistics $Z(t)/\widehat{\sigma}(t)$ is asymptotically normal when H_0 is true. Therefore, H_0 can be rejected when $|Z(t)/\widehat{\sigma}(t)| > 1.96$ for the significance level of 0.05.

2.5 Simulation Studies

This section consists of simulation studies that investigate performance of proposed variance estimators and the one-sample test. Distributions of the variables had to carefully

selected due to identifiability issue of truncated samples. Let (a_k, b_k) be the inner support of a distribution function $K(t)$, where $a_k = \inf\{z > 0 : K(z) > 0\}, b_k = \sup\{z > 0 : K(z) < 1\}$. Let F be the distribution function of T. G and F have the interior support (a_G, b_G) and (a_F, b_F), respectively. Based on a truncated sample, only the distribution function given $T \geq a_G, L \leq b_F$ are estimable [66]. Practically, one can choose $a^* = \min(L_1^0, \cdots, L_n^0), b^* = \max(T_1^0, \cdots, T_n^0)$ and estimate the conditional distribution functions $F^*(t) = P(T \leq t | T \geq a^*), \forall\ t \geq a^*$ and $G^*(t) = P(L \leq t | L \leq b^*), \forall\ t \leq b^*$ [38], [62]. For estimating distribution function of L, if it happens that $b^* > b_G$, then $G^*(t) = G(t)$.

When the parametric distribution of L is defined on $[0, \infty)$, the estimated conditional distributions vary by each sample, causing difficulties in assessing the outcome. This issue can be solved by choosing the distributions that are defined on a bounded interval, $[0, \pi]$ such as the uniform and truncated exponential distributions for L. The exponential distribution for T can be considered consequently, in the simulated settings, $b^* > \pi$ for all replicates, so that the underlying conditional distribution agrees with the $G(t)$, regardless of replicates.

Another disturbing issue may arise whenever the risk set equals to 1 at t but some events still occur after t. This issue causes the estimated distribution probability to reach 1 at t. Since there are still some events observed after t, the estimated distribution probability should not be evaluated as 1 at t. One solution is to set the risk set to a positive integer, $c > 1$, whenever it is less than c [66]. Keiding and Gill [36] referred to this problem as empty inner risk sets. Their Monte Carlo simulation of 10,000 samples yielded relative frequency of 0.003 for $n = 50$ and 0.0003 for $n = 100$. Large sample sizes, 200 and 400, were considered in this simulation study to avoid the problem of empty inner risk sets.

Two sets of simulation studies were conducted. Aim of Study I was to evaluate performances of proposed variance estimators by reporting bias and 95% coverage probabilities. Study II considered the one-sample context and evaluated the power of the test.

2.5.1 Study I

Uniform$[0, 1]$ and exponential distribution truncated at 1.2 were chosen to be distributions of L. The cumulative hazard functions are given by

$$A(t) = -\log(1-t), \quad 0 \le t \le 1$$

and

$$A(t) = -\log\left(1 - \frac{1-e^{-t}}{1-e^{-1.2}}\right), \quad 0 < t < 1.2.$$

Table 2.1 The simulation results of $A(t)$ when the underlying distribution of L is Uniform$[0, 1]$.

					Naive variance estimator		Alternative variance estimator	
n	$L\%$	t	Bias	Sample variance	Estimated variance	Coverage	Estimated variance	Coverage
200	25	0.20	0.000	0.0012	0.0012	0.943	0.0012	0.944
		0.50	-0.003	0.0059	0.0056	0.941	0.0056	0.941
		0.80	-0.005	0.0269	0.0253	0.950	0.0253	0.950
	50	0.20	-0.002	0.0014	0.0012	0.929	0.0012	0.933
		0.50	0.001	0.0081	0.0077	0.949	0.0076	0.950
		0.80	0.002	0.0445	0.0408	0.937	0.0404	0.937
400	25	0.20	0.000	0.0006	0.0006	0.950	0.0006	0.951
		0.50	0.002	0.0027	0.0029	0.952	0.0029	0.952
		0.80	0.000	0.0135	0.0128	0.939	0.0128	0.939
	50	0.20	0.000	0.0006	0.0006	0.947	0.0006	0.947
		0.50	0.003	0.0038	0.0038	0.959	0.0038	0.959
		0.80	0.004	0.0207	0.0207	0.947	0.0207	0.947

The exponential distribution with mean $1/\lambda$ was used as the distribution of the truncation variable T. The value of λ was selected to yield two truncation rates, 25% and 50%. The truncation rate is calculated by $(N-n)/N$, where N is the size of the pool from which the

Table 2.2 The simulation results of $A(t)$ when the underlying distribution of L is truncated exponential.

					Naive variance estimator		Alternative variance estimator	
n	$L\%$	t	Bias	Sample variance	Estimated variance	Coverage	Estimated variance	Coverage
200	25	0.15	0.000	0.0012	0.0012	0.947	0.0012	0.945
		0.43	-0.003	0.0059	0.0058	0.941	0.0058	0.941
		0.82	-0.005	0.0286	0.0279	0.942	0.0276	0.942
	50	0.15	-0.002	0.0012	0.0012	0.935	0.0012	0.933
		0.43	0.001	0.0090	0.0088	0.941	0.0086	0.937
		0.82	0.008	0.0620	0.0562	0.945	0.0552	0.944
400	25	0.15	0.000	0.0006	0.0006	0.953	0.0006	0.953
		0.43	0.002	0.0031	0.0029	0.954	0.0029	0.954
		0.82	0.000	0.0144	0.0142	0.933	0.0142	0.933
	50	0.15	0.000	0.0006	0.0006	0.948	0.0006	0.947
		0.43	0.002	0.0048	0.0045	0.951	0.0045	0.948
		0.82	0.004	0.0306	0.0286	0.948	0.0282	0.948

truncated sample is selected. Sample sizes were chosen to be 200 and 400. 1000 replicates were generated for each setting. Let $\widehat{A}^{(i)}(t)$ be the cumulative hazard estimate for the ith replicate at t. Let $\overline{\widehat{A}}(t)$ denote the average cumulative hazard estimate across 1000 replicates, where $\overline{\widehat{A}}(t) = \sum_{i=1}^{1000} \widehat{A}^{(i)}(t)$.

The bias was defined as the deviation between average cumulative hazard estimate and the true value, that is, Bias $= \overline{\widehat{A}}(t) - A(t)$. Sample variances were calculated by the following formula

$$\text{Sample variance} = \frac{1}{1000-1} \sum_{i=1}^{1000} \left(\widehat{A}^{(i)}(t) - \overline{\widehat{A}}(t) \right)^2.$$

Variance estimators that are given in (2.18) and (2.19) are evaluated and averages of 1000 replicates obtained from

$$\text{Estimated variance} = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\text{var}}^{(k)}[\widehat{A}^{(i)}(t)], \quad k = 1, 2.$$

95% confidence interval were calculated for each replicate and actual coverage rate across 1000 replicates were obtained.

Estimation results are reported at time points that correspond to 0.2, 0.5, 0.8 in $G(t)$. For the settings that the uniform distribution was used as the underlying distribution of L, we report the estimation result at $t = 0.2, 0.5, 0.8$ (see Table 2.1). For the settings using the truncated exponential distribution, we evaluated at $t = 0.15, 0.43, 0.82$, still relating to 0.2, 0.5, 0.8 in $G(t)$ (see Table 2.2).

In both tables, biases are very close to zero across the settings. The numerical values yielded from these two variance estimators are evaluated very close, and the averages match the variance among 1000 cumulative hazard estimates. The coverage percentages are close to the nominal level, with the exception for small t and heavy truncation, in which slight undercoverage is observed.

2.5.2 Study II

The performance of the one-sample test was evaluated in this study. We continue to use uniform and truncated exponential distribution for L. The truncation variable was generated from exponential distribution with different means to produce predetermined truncation rates. First, the known hazard rate function $\alpha_0(t)$ was assumed to be Uniform[0,1]. Three settings were generated from Uniform$[0, 1]$, Uniform$[0, 1.2]$ and Uniform$[0, 1.3]$, respectively (see Table 2.3). Second, $\alpha_0(t)$ was assumed to be exponential distribution with mean 1 and truncated at 1.2. The simulated settings were exponential distributions truncated at 1.2 with different means (see Table 2.4).

The test statistic given in Equation (2.20) was evaluated for each sample in one setting to construct the one-sample log-rank test. The weight function was chosen to be $\bar{Y}(t)$. The null hypothesis, H_0, was rejected at level 0.05. The proportion of rejection among 1000 samples is shown in Tables 2.3 and 2.4.

When sample distribution agrees with population distribution, the observed rejection rates are close to the significance level 0.05. Table 2.3 shows a trend of increasing power

Table 2.3 The proportions of rejection for one-sample test when $\alpha_0(t) \sim \text{Uniform}[0, 1]$.

			Proportion of rejecting H_0 at level 0.05		
n	$L\%$	t	Uniform[0,1]	Uniform[0,1.2]	Uniform[0,1.3]
200	25	0.20	0.049	0.233	0.388
		0.50	0.055	0.561	0.838
		0.80	0.049	0.966	0.999
	50	0.20	0.060	0.256	0.437
		0.50	0.055	0.505	0.774
		0.80	0.054	0.869	0.988
400	25	0.20	0.043	0.390	0.628
		0.50	0.037	0.836	0.986
		0.80	0.059	0.999	1.000
	50	0.20	0.047	0.409	0.677
		0.50	0.048	0.761	0.975
		0.80	0.040	0.990	1.000

by time when distributions differ from each other. The power is higher if the mean of the sample distribution differs more from the mean of distribution of $\alpha_0(t)$.

Table 2.4 shows a different trend. The power of test peaks for middle t. The reason to explain different trends with these two distributions has been explored in the first simulation study. The power also increases as expected when the difference between the means becomes greater. Generally, a higher truncation proportion leads to a lower power in both settings.

2.6 Discussion

This chapter emphasized on the nonparametric inference of the cumulative hazard function with right truncated data. The weak convergence properties of the plug-in estimator was derived and two variance estimators were presented. A weighted one-sample log-rank test was developed to compare the hazard rate function of the truncated sample to a given rate function. Two sets of simulation studies were conducted to investigate the practical performances of proposed variance estimators and the one-sample log-rank test. The variance estimators developed by Jiang [32] overestimated the variance when t is large. This

Table 2.4 The proportions of rejection for one-sample test when $\alpha_0(t) \sim \text{Exp}(1)$.

			Proportion of rejecting H_0 at level 0.05		
n	$L\%$	t	Exp(1.0)	Exp(1.5)	Exp(2.0)
200	25	0.15	0.047	0.199	0.661
		0.43	0.047	0.264	0.764
		0.82	0.048	0.169	0.524
	50	0.15	0.056	0.163	0.535
		0.43	0.046	0.167	0.468
		0.82	0.051	0.120	0.304
400	25	0.15	0.049	0.397	0.926
		0.43	0.047	0.550	0.972
		0.82	0.062	0.313	0.855
	50	0.15	0.058	0.355	0.875
		0.43	0.055	0.348	0.816
		0.82	0.054	0.215	0.573

issue was fixed by the variance estimators proposed in this chapter. It can be concluded from simulation studies that proposed variance estimators have satisfactory results.

An important extension of this research is the weighted two-sample and K-sample tests. There are various methods to assess survival outcomes between two independent samples. One may consider to compare the survival probabilities up to t, $H_0 : S_1(s) = S_2(s), \forall s \leq t$. Another option is to compare the survival probabilities at a selected time point, $H_0 : S_1(t) = S_2(t)$. Chi *et al.* [12] developed a nonparamteric test to compare two survival functions for the entire study period with right truncated data by finding the integrated weighted difference. A more common hypothesis for survival outcome comparison is to compare the hazard rate function up to t, $H_0 : \alpha_1(t) = \alpha_2(t), \forall s \leq t$. This type of test captures the direct risks of failure over the interval $[0, t]$. The tests developed for such hypothesis are the family of weighted log-rank tests. A few common choices of weight function lead to the well-known tests such as the log-rank, Gehan [21] as well as Tarone and Ware [57] tests. The family of weighted log-rank tests with right truncated samples will be considered in the next chapter.

CHAPTER 3

K-SAMPLE HYPTHESIS TESTING WITH RIGHT-TRUNCATED DATA

3.1 Motivation

This dissertation centers on the inferences of right truncated data. Chapter 2 contains the nonparametric inference for the cumulative hazard function, together with a one-sample weighted log-rank test. Real applications often involve risk assessment among finite groups. Although there are various methods to compare survival outcomes between K groups, assessment on the hazard rate function up to selected time point has the advantage of capturing the instantaneous failure rates in the chosen time interval. Therefore, a K-sample test is practically needed and this chapter focuses on this issue with right truncated data.

In this chapter, a K-sample test statistic is first developed for right truncated data. The test at the two-sample setting is subsequently considered. The family of weighted log-rank test contains several commonly used tests. Choices of different weight function leads to log-rank, Gehan and Tarone-Ware tests. Simulation studies are designed for the two-sample and three-sample settings to evaluate performance of proposed tests. AIDS blood transfusion data was analyzed to give a real example to illustrate the methods.

3.2 K-Sample Tests

The one-sample test developed in the previous chapter provides the foundation for the K-sample test. Suppose that there are K independent samples, denoted as $\{L_{ki}, T_{ki}\}$ where $k = 1, ..., K$, $i = 1, ..., n_k$ and with constraint $L_{ki} \leq T_{ki}$. L_k and T_k are the random variables associated with the kth sample and they have the distribution functions G_k and F_k, respectively. Let $\alpha_k(t)$ and $A_k(t)$ be hazard rate and cumulative hazard functions of L_k,

where

$$A_k(t) = \int_0^t \alpha_k(s)ds = \int_0^t \frac{dG_k(s)}{P(L_k \geq s)}. \tag{3.1}$$

Let τ be the largest observed time in the pooled samples. We can define $L_k^* = \tau - L_k$ and $T_k^* = \tau - T_k$ where L_k^* is left truncated by T_k^*. The concept of reverse-time hazard rate and cumulative hazard functions have been introduced in Chapter 2. Let $\alpha_k^*(t)$ and $A_k^*(t)$ be these two quantities associated with the kth sample. Their mathematical definitions are

$$\alpha_k^*(t)dt = P(t - dt < L_k \leq t | L_k \leq t)$$

and

$$A_k^*(t) = \int_\tau^t dA_k^*(s)ds = \int_t^\tau \alpha_k^*(s)ds = \int_t^\tau \frac{dG_k(s)}{P(L_k \leq s)}. \tag{3.2}$$

The relation between reverse-time and forward-time cumulative hazard functions have been clarified in Chapter 2. Here, for the kth sample,

$$A_k(t) = -\log[1 - \exp(-A_k^*(t))].$$

It is known that the reverse-time cumulative hazard function can be estimated by the Nelson-Aalen estimator. For the kth sample, the Nelson-Aalen estimator of $A_k^*(t)$ is given by

$$\widehat{A}_k^*(t) = \int_t^\tau J_k(s)\frac{d\bar{N}_k(s)}{\bar{Y}_k(s)}, \qquad k = 1, ..., K,$$

where $J_k(t) = I(\bar{Y}_k(t) > 0)$, $Y_{ki}(t) = I(L_{ki} \leq t \leq T_{ki})$, $N_{ki}(t) = I(L_{ki} \leq t)$, $\bar{Y}_k(t) = \sum_{i=1}^{n_k} Y_{ki}(t)$ and $\bar{N}_k(t) = \sum_{i=1}^{n_k} N_{ki}(t)$.

The reverse-time martingale was defined in Chapter 2. It is a fundamental quantity for establishing properties of the estimator of forward-time cumulative hazard function. The martingale can be similarly defined in the K-sample setting. Define the counting process,

$N_{ki}^L(t) = I(L_{ki} \geq t)$, which count event occurrence backwards from τ. The reverse-time martingale is defined as

$$M_{ki}^*(t) = N_{ki}^L(t) - \int_\tau^t Y_{ki}(s) dA_k^*(s).$$

Let $\bar{M}_k^*(t) = \sum_{i=1}^{n_k} M_{ki}^*(t)$. It can be shown that

$$\widehat{A}_k^*(t) - A_k^*(t) = \int_\tau^t \frac{J_k(s)}{\bar{Y}_k(s)} d\bar{M}_k^*(s). \tag{3.3}$$

The hypothesis that needs to be tested is $H_0 : \alpha_1(t) = \alpha_2(t) = ... = \alpha_K(t)$. Let $\alpha_\bullet(t)$ and $A_\bullet(t)$ be the assumed common hazard rate and cumulative hazard function. $A_\bullet(t)$ can be estimated by Formula (2.13) based on the pooled samples. For the hypothesized common value, the reverse-time Nelson-Aalen estimator can be constructed as

$$\widehat{A}_\bullet^*(t) = \int_\tau^t J(s) \frac{d\bar{N}_\bullet^L(s)}{\bar{Y}_\bullet(s)}, \tag{3.4}$$

where $J(t) = I(\bar{Y}_\bullet(t) > 0)$, $\bar{Y}_\bullet(t) = \sum_{k=1}^K \sum_{i=1}^{n_k} I(L_{ki} \leq t \leq T_{ki})$, $\bar{N}_\bullet^L(t) = \sum_{k=1}^K \sum_{i=1}^{n_k} I(L_{ki} \geq t)$. Also let $n = \sum_{k=1}^K n_k$. The test for the above hypothesis requires comparison between $\widehat{A}_k^*(t)$ and $\widehat{A}_\bullet^*(t)$. It is acceptable to compare $\widehat{A}_k^*(t)$ with $\widehat{A}_\bullet^*(t)$ only for the time when $\bar{Y}_k(t) > 0$. If the null hypothesis holds true,

$$\widehat{A}_k^*(t) - \widehat{A}_\bullet^*(t) \;=\; \int_\tau^t J_k(s) \frac{d\bar{M}_k^*(s)}{\bar{Y}_k(s)} - \int_\tau^t J_k(s) \frac{d\bar{M}_\bullet^*(s)}{\bar{Y}_\bullet(s)}, \tag{3.5}$$

where $\bar{M}_\bullet^*(t) = \sum_{k=1}^K \bar{M}_k^*(t)$. The difference between $\widehat{A}_k^*(t)$ and $\widehat{A}_\bullet^*(t)$ is a zero-mean random noise process related to martingales.

Let $W_k(t)$ be a stochastic weight process for the kth sample. Let $\widehat{A}_k(t)$ and $\widehat{A}_\bullet(t)$ be the plug-in estimators given in Equation (2.13) based on the kth sample and the pooled samples, respectively. Consider the following statistic,

$$Z_k(t) = \int_0^t W_k(s) d[\widehat{A}_k(s) - \widehat{A}_\bullet(s)]. \tag{3.6}$$

Suppose that $W(t)$ is a common weight function for K samples. $W(t)$ usually only depends on pooled counting process $\bar{N}_\bullet(t)$ and pooled risk set $\bar{Y}_\bullet(t)$. We further assume that $W(t) = 0$ and $W(t)/\bar{Y}_\bullet(t) = 0$ when $\bar{Y}_\bullet(t) = 0$. Let $W_k(t) = W(t)\bar{Y}_k(t)$ and then the above statistic can be written as

$$Z_k(t) = \int_0^t W(s)\bar{Y}_k(s)d[\widehat{A}_k(s) - \widehat{A}_\bullet(s)].$$

Asymptotic distribution of $Z_k(t)$ and its covariance matrix are established in Appendix A2. The asymptotic variance is

$$\int_t^0 \left[W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)} - \int_0^u W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{y_k(u)}{y_\bullet(u)}\left(\delta_{km} - \frac{y_m(u)}{y_\bullet(u)}\right)\alpha^*(u)y_\bullet(u)du$$

$$+\int_\tau^t \left[\int_0^t W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{y_k(u)}{y_\bullet(u)}\left(\delta_{km} - \frac{y_m(u)}{y_\bullet(u)}\right)\alpha^*(u)y_\bullet(u)du$$

Under the null hypothesis, mean of $Z_k(t)$ is zero with covariance $E\langle Z_k, Z_m\rangle(t)$. The covariance can be estimated by (details shown in Appendix A2)

$$\widehat{\sigma}^2_{km}(t) = \int_0^t \left[W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)} - \int_0^u W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(\delta_{km} - \frac{\bar{Y}_m(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u)$$

$$+\int_t^\tau \left[\int_0^t W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(\delta_{km} - \frac{\bar{Y}_m(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u) \tag{3.7}$$

and the variance can be estimated by

$$\widehat{\sigma}^2_{kk}(t) = \int_0^t \left[W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)} - \int_0^u W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(1 - \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u)$$

$$+\int_t^\tau \left[\int_0^t W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(1 - \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u). \tag{3.8}$$

Some options of $W(t)$ lead to a few standard tests. For example, the log-rank test can be obtained by choosing $W(t) = I(\bar{Y}_\bullet(t) > 0)$. With the choice of $W(t) = \bar{Y}_\bullet(t)$, the

test statistic is Gehan [21] generalization of Wilcoxon and Kruskal-Wallis test. The choice that $W(t) = g(\bar{Y}_\bullet(t))$ and $g(x) = \sqrt{x}$ leads to the Tarone and Ware [57] test. The log-rank test for complete or censored survival data can be interpreted as the differences between observed and expected number of events. However, such interpretation is not feasible for right truncated data.

Let $\widehat{\mathbf{\Sigma}}(\mathbf{t})$ denote the $K \times K$ matrix where the kth diagonal position is given by $\widehat{\sigma}^2_{kk}(t)$ and (k, m)th element is given by $\widehat{\sigma}^2_{km}$. Consider a column vector $\mathbf{Z}(t) = (Z_1(t)\ Z_2(t)\ ...\ Z_K(t))^T$. The test statistic for testing $H_0 : \alpha_1(t) = \alpha_2(t) = ... = \alpha_K(t)$ will have the following form

$$X^2 = \mathbf{Z}(t)^T \widehat{\mathbf{\Sigma}}^-(t)\mathbf{Z}(t) \sim \chi^2_{K-1}, \tag{3.9}$$

where $\widehat{\mathbf{\Sigma}}^-(t)$ is a generalized inverse and X^2 is asymptotically chi-squared distributed with $K-1$ degrees of freedom (details given in Appendix A2).

For all k and m, if there exists a time point where $\bar{N}^L_\bullet(t)$ jumps, also $W(t)$, $\bar{Y}_k(t)$ and $\bar{Y}_m(t)$ are positive then $\widehat{\mathbf{\Sigma}}(t)$ has rank $K-1$ [22]. One can reduce the $\widehat{\mathbf{\Sigma}}(t)$ to a $K-1 \times K-1$ full-rank matrix. One can delete the last row and last column of $\widehat{\mathbf{\Sigma}}(t)$ and denote it by $\widehat{\mathbf{\Sigma}}_d(t)$. Let $\mathbf{Z}_d(t) = (Z_1(t)\ Z_2(t)\ ...\ Z_{K-1}(t))^T$, which contains the first $K-1$ elements of $\mathbf{Z}(t)$. The test statistics (3.9) can be alternatively given as

$$X^2 = \mathbf{Z}_d(t)^T \widehat{\mathbf{\Sigma}}_d^{-1}(t)\mathbf{Z}_d(t), \tag{3.10}$$

where $\widehat{\mathbf{\Sigma}}_d^{-1}(t)$ is the ordinary inverse of a full-rank matrix.

3.3 Two-Sample Tests

Two-sample comparison appears frequently in real applications. It is useful to clarify the test procedure for the two-sample setting. In this section, the test procedure introduced in the previous section is studied for $K = 2$. For two-sample setting, $\widehat{\mathbf{\Sigma}}_d(t)$ is equal to $\widehat{\sigma}^2_{11}(t)$

and $\mathbf{Z}_d(t) = Z_1(t)$. The null hypothesis of $H_0 : \alpha_1(t) = \alpha_2(t)$ can be tested by the statistic

$$X^2 = (Z_1(t))^2 / \widehat{\sigma}^2_{11}(t).$$

The alternative test statistic is

$$U(t) = Z_1(t) / \widehat{\sigma}_{11}(t),$$

where $U(t)$ follows a standard normal distribution. Let $L(t) = \dfrac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)}$, the variance estimator of $Z_1(t)$ is

$$\widehat{\sigma}^2_{11}(t) = \int_0^t \left[W(u) \frac{G_\bullet(u)}{1 - G_\bullet(u)} - \int_0^u W(s) d \left(\frac{G_\bullet(s)}{1 - G_\bullet(s)} \right) \right]^2 \frac{L^2(u)}{\bar{Y}_1(u)\bar{Y}_2(u)} d\bar{N}_\bullet(u)$$

$$+ \int_t^\tau \left[\int_0^t W(s) d \left(\frac{G_\bullet(s)}{1 - G_\bullet(s)} \right) \right]^2 \frac{L^2(u)}{\bar{Y}_1(u)\bar{Y}_2(u)} d\bar{N}_\bullet(u). \tag{3.11}$$

At the significant level 0.05, the null hypothesis can be rejected if the absolute value of $U(t)$ is beyond 1.96.

3.4 Simulation Studies

This section contains two sets of simulation studies designed for the K-sample and two-sample tests. Rationales for choosing underlying distributions and sample size were discussed in Chapter 2.5. The uniform and exponential distributions were chosen for these simulation studies. Large sample sizes 200 and 400 were considered to avoid empty inner risk sets.

Two simulation studies were constructed. The first set of simulation evaluates the performances of two-sample tests and the second set of simulation considers the three-sample settings to evaluate the K-sample tests.

3.4.1 Study I

This study centers on the performance of the two-sample tests. The uniform and truncated exponential distributions were used for event time variable while exponential distribution was chosen for truncation variable. For the first set of the simulated settings, Uniform$[0, 1]$ was consistently used as the underlying distribution of the event time variable for Group 1, while Uniform$[0, 1]$, Uniform$[0, 1.2]$ and Uniform$[0, 1.3]$ were chosen for the distribution of the event time variable for Group 2 (see Table 3.1).

The truncation variables in Groups 1 and 2 were generated from exponential distributions with different means, to produce the same level of truncation rate in these two samples. For the second set of settings, the underlying distributions of the event time variables were the exponential distributions truncated at 1.2. The exponential distribution with mean 1 truncated at 1.2 was selected for Group 1. Different truncated exponential distributions were selected for Group 2. The explicit distributions of the event time variables for Groups 1 and 2 are provided in Table 3.2.

Selection of different weight functions lead to different types of test. The weight function $I(\bar{Y}_\bullet(t) > 0)$ leads to the log-rank test. Other choices of weight functions were $\bar{Y}_\bullet(t)$ and $\sqrt{\bar{Y}_\bullet(t)}$ and yielded Gehan and Tarone-Ware tests, respectively. The null hypothesis of equivalence in cumulative hazard was rejected at level 0.05 for each pair of samples. The proportion of rejection among 1000 pairs of samples is shown in Tables 3.1 and 3.2. In both tables, when the underlying distributions for Groups 1 and 2 are identical, the observed rejection rates are close to the significance level 0.05. When the underlying distributions in the two samples are different the observed power increases by time in Table 3.1, while Table 3.2 shows a different trend that the observed power increases for small t but declines when t gets large.

We depicted the underlying distributions of event time variables to find the plausible explanation for trends of observed power. When the event time variables follow different uniform distributions in two groups, the differences between two cumulative hazard functions monotonously increase by time. When the event time variables follow two truncated

exponential distributions, the differences on the cumulative hazard increase by time first, but start to decline when t is towards the end. In Table 3.1, the log-rank test has the highest power among all three tests. The explanation is that the log-rank test is most powerful when the hazard functions are proportional. When the underlying distributions are uniform, the hazard functions are close to proportional.

3.4.2 Study II

This study was designed to evaluate the performance of the K-sample tests. The uniform and truncated exponential distributions were selected for event time variables and three-sample settings were simulated for each. Truncation variables were generated from exponential distribution as described in Study I. The event time variables were generated from uniform distributions for first three settings, while exponential distributions truncated at 1.2 were used for the next three settings. The underlying distributions for each group can be found in Tables 3.3 through 3.6.

Tables 3.3 and 3.5 show the proportions of rejecting null hypothesis for the settings that the vent time variables followed uniform distributions. When the underlying distributions are identical for all three groups, observed rejection rates are consistently close to 0.05. The power of tests increases by time when the distributions are different among three groups. For the settings with same size, a higher truncation rate causes reduction in power.

Truncated exponential distributions were also used for the event time variables and the results are depicted in Tables 3.4 and 3.6. Observed rejection proportions are all around 0.05 when all three groups have the same underlying distributions. Unlike the first setting, the power of tests decreases by time when distributions vary among groups. Explanation for this discrepancy has been offered in Chapter 2.

3.5 The AIDS Latent Time Example

We used blood transfusion infected AIDS data set described in Section 1.3. Our analysis focused on nonparametric inference of the cumulative hazard functions. The data set contains

three variables: AIDS incubation time, infection time counted since January 1, 1978, and age at blood transfusion. Let L denote the incubation time. The truncation time T is the time from infection to the end of study, July 1, 1986. This data set was routinely divided into three subgroups: children (age range 1-4), adults (age range 5-59) and elderly patients (age $\geq$ 60), with the sizes 34, 120 and 141, respectively. The largest incubation times are respectively 43, 89, and 83 months for children, adults and elderly patients.

In Figure 3.1, we depicted the estimated cumulative hazard curves for each subgroup. We can conclude from Figure 3.1 that children have significantly higher cumulative hazard than adults and elderly patients which suggests that children has higher intensity of AIDS onset than adults and elderly patients.

The weighted log-rank tests were applied to compare the hazard functions between subgroups. In Table 3.7 shows the results of log-rank, Gehan and Taronea and Ware tests for comparing hazard functions up to 12, 24 and 36 months. Results of all tests indicate that the differences among subgroups are statistically significant at level 0.05. It can be easily recognized from Figure 3.1 that dramatically higher hazard function in children is the primary source of difference.

3.6 Discussion

This chapter extended the one-sample test developed in Chapter 2 to the K-sample context. The family of weighted log-rank tests was proposed and selection of different weight functions was discussed. The family of tests include several commonly used tests such as the log-rank test, Gehan and Tarone-Ware tests. Subsequently, two-sample test was particularly studied. The simulation studies were conducted to evaluate the performances of proposed tests for the two-sample and three-sample context. The log-rank, Gehan and Tarone-Ware tests were implemented for each setting. The simulation study yielded satisfactory result. Performances of three tests are slightly different, depending on the selected underlying distributions.

When there are multiple factors or continuous predictors are associated with survival

outcome, a regression model is needed to assess the association between covariates and survival. Regression analysis of survival data often models the hazard rate function. The Cox proportional hazards model is the most commonly used regression model because result is easy to interpret. Estimation of regression parameters in a Cox model and the inferences routinely rely on the partial likelihood. Finkelstein *et al.* [16] studied the Cox model for right truncated data using the full-likelihood approach. It is interesting to investigate the partial-likelihood-based solution of Cox analysis for right truncated data. Compared to the full-likelihood approach, estimation using the partial likelihood should have the advantage of computational efficiency.

Table 3.1 The proportion of rejecting H_0 when the underlying distributions are uniform.

			Uniform[0,1], Uniform[0,1]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.048	0.055	0.044
		0.50	0.044	0.050	0.038
		0.80	0.038	0.034	0.044
	50	0.20	0.049	0.045	0.047
		0.50	0.054	0.051	0.049
		0.80	0.049	0.037	0.046
400	25	0.20	0.044	0.045	0.048
		0.50	0.042	0.038	0.040
		0.80	0.045	0.038	0.048
	50	0.20	0.042	0.054	0.053
		0.50	0.053	0.041	0.051
		0.80	0.044	0.037	0.036
			Uniform[0,1], Uniform[0,1.2]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.212	0.218	0.213
		0.50	0.367	0.344	0.366
		0.80	0.709	0.686	0.703
	50	0.20	0.126	0.120	0.117
		0.50	0.172	0.166	0.169
		0.80	0.384	0.373	0.371
400	25	0.20	0.430	0.399	0.418
		0.50	0.675	0.640	0.659
		0.80	0.951	0.942	0.949
	50	0.20	0.172	0.143	0.146
		0.50	0.272	0.251	0.258
		0.80	0.660	0.657	0.665
			Uniform[0,1], Uniform[0,1.3]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.423	0.400	0.411
		0.50	0.632	0.602	0.620
		0.80	0.951	0.930	0.940
	50	0.20	0.137	0.145	0.144
		0.50	0.254	0.230	0.245
		0.80	0.585	0.574	0.579
400	25	0.20	0.749	0.724	0.739
		0.50	0.930	0.914	0.926
		0.80	1.000	1.000	1.000
	50	0.20	0.301	0.277	0.291
		0.50	0.536	0.505	0.525
		0.80	0.908	0.904	0.908

Table 3.2 The proportions of rejecting H_0 when the underlying distributions are exponential.

			Exp(1.0), Exp(1.0)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.056	0.048	0.048
		0.43	0.045	0.050	0.039
		0.82	0.045	0.044	0.040
	50	0.15	0.047	0.046	0.051
		0.43	0.046	0.049	0.048
		0.82	0.039	0.042	0.041
400	25	0.15	0.051	0.053	0.049
		0.43	0.056	0.051	0.042
		0.82	0.054	0.054	0.047
	50	0.15	0.053	0.055	0.043
		0.43	0.057	0.053	0.065
		0.82	0.039	0.044	0.044
			Exp(1.0), Exp(1.5)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.165	0.146	0.159
		0.43	0.173	0.133	0.150
		0.82	0.115	0.082	0.092
	50	0.15	0.093	0.104	0.091
		0.43	0.113	0.082	0.094
		0.82	0.064	0.069	0.061
400	25	0.15	0.292	0.269	0.281
		0.43	0.318	0.249	0.281
		0.82	0.209	0.140	0.160
	50	0.15	0.175	0.159	0.163
		0.43	0.172	0.150	0.164
		0.82	0.090	0.072	0.072
			Exp(1.0), Exp(2.0)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.543	0.513	0.521
		0.43	0.570	0.440	0.497
		0.82	0.374	0.244	0.301
	50	0.15	0.296	0.262	0.279
		0.43	0.287	0.213	0.239
		0.82	0.111	0.108	0.108
400	25	0.15	0.838	0.794	0.816
		0.43	0.872	0.772	0.835
		0.82	0.670	0.473	0.539
	50	0.15	0.562	0.496	0.530
		0.43	0.571	0.411	0.480
		0.82	0.216	0.165	0.180

Table 3.3 The proportions of rejection for three-sample settings when the underlying distributions are all Uniform[0,1].

n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.052	0.044	0.045
		0.50	0.049	0.046	0.049
		0.80	0.047	0.055	0.054
	50	0.20	0.050	0.059	0.060
		0.50	0.064	0.056	0.055
		0.80	0.054	0.055	0.053
400	25	0.20	0.043	0.043	0.047
		0.50	0.048	0.051	0.054
		0.80	0.055	0.051	0.052
	50	0.20	0.057	0.057	0.050
		0.50	0.050	0.058	0.052
		0.80	0.047	0.045	0.047

Table 3.4 The proportions of rejection for three-sample settings when the underlying distributions are all Exp(1.0).

n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.055	0.051	0.052
		0.43	0.038	0.048	0.041
		0.82	0.047	0.042	0.050
	50	0.15	0.047	0.056	0.057
		0.43	0.057	0.054	0.053
		0.82	0.050	0.057	0.055
400	25	0.15	0.043	0.048	0.047
		0.43	0.050	0.047	0.045
		0.82	0.045	0.040	0.042
	50	0.15	0.036	0.039	0.033
		0.43	0.052	0.056	0.053
		0.82	0.050	0.048	0.048

Table 3.5 The proportions of rejecting H_0 when the underlying distributions are uniform.

			Uniform[0,1], Uniform[0,1], Uniform[0,1.2]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.232	0.227	0.225
		0.50	0.372	0.352	0.365
		0.80	0.764	0.750	0.764
	50	0.20	0.084	0.084	0.089
		0.50	0.150	0.143	0.145
		0.80	0.379	0.383	0.387
400	25	0.20	0.454	0.444	0.457
		0.50	0.728	0.690	0.709
		0.80	0.978	0.981	0.984
	50	0.20	0.156	0.143	0.152
		0.50	0.313	0.302	0.312
		0.80	0.761	0.753	0.754
			Uniform[0,1], Uniform[0,1], Uniform[0,1.3]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.441	0.422	0.429
		0.50	0.669	0.651	0.672
		0.80	0.963	0.959	0.963
	50	0.20	0.153	0.146	0.156
		0.50	0.298	0.295	0.300
		0.80	0.709	0.700	0.704
400	25	0.20	0.769	0.772	0.777
		0.50	0.951	0.941	0.947
		0.80	1.000	1.000	1.000
	50	0.20	0.318	0.306	0.312
		0.50	0.605	0.569	0.589
		0.80	0.967	0.961	0.965
			Uniform[0,1], Uniform[0,1.2], Uniform[0,1.3]		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.20	0.339	0.330	0.335
		0.50	0.525	0.502	0.515
		0.80	0.887	0.857	0.874
	50	0.20	0.135	0.122	0.131
		0.50	0.205	0.195	0.204
		0.80	0.493	0.475	0.485
400	25	0.20	0.645	0.628	0.640
		0.50	0.873	0.751	0.869
		0.80	0.995	0.994	0.995
	50	0.20	0.242	0.222	0.223
		0.50	0.440	0.411	0.435
		0.80	0.861	0.854	0.857

Table 3.6 The proportions of rejecting H_0 when the underlying distributions are exponential.

			Exp(1.0), Exp(1.0), Exp(1.5)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.127	0.128	0.130
		0.43	0.153	0.126	0.135
		0.82	0.096	0.076	0.085
	50	0.15	0.091	0.089	0.090
		0.43	0.115	0.100	0.104
		0.82	0.058	0.055	0.057
400	25	0.15	0.294	0.276	0.298
		0.43	0.329	0.263	0.287
		0.82	0.176	0.112	0.135
	50	0.15	0.158	0.141	0.152
		0.43	0.174	0.122	0.137
		0.82	0.083	0.072	0.086
			Exp(1.0), Exp(1.0), Exp(2.0)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.527	0.489	0.505
		0.43	0.567	0.416	0.488
		0.82	0.312	0.180	0.226
	50	0.15	0.252	0.220	0.233
		0.43	0.243	0.164	0.199
		0.82	0.072	0.066	0.070
400	25	0.15	0.872	0.844	0.859
		0.43	0.901	0.763	0.832
		0.82	0.650	0.395	0.497
	50	0.15	0.542	0.451	0.498
		0.43	0.508	0.343	0.399
		0.82	0.171	0.127	0.139
			Exp(1.0), Exp(1.5), Exp(2.0)		
n	$L\%$	t	Log-rank	Gehan	Tarone-Ware
200	25	0.15	0.402	0.366	0.378
		0.43	0.434	0.326	0.369
		0.82	0.266	0.162	0.198
	50	0.15	0.208	0.283	0.200
		0.43	0.219	0.151	0.182
		0.82	0.109	0.098	0.096
400	25	0.15	0.770	0.721	0.746
		0.43	0.809	0.668	0.738
		0.82	0.577	0.381	0.466
	50	0.15	0.468	0.414	0.442
		0.43	0.456	0.318	0.379
		0.82	0.190	0.149	0.163

Table 3.7 The weighted log-rank tests for comparing the hazard rate functions between subgroups of the AIDS blood transfusion data set.

	Log-rank		Gehan		Tarone-Ware	
Time	X^2	p-value	X^2	p-value	X^2	p-value
12 months	87.67	< 0.001	80.74	< 0.001	84.34	< 0.001
24 months	110.74	< 0.001	84.39	< 0.001	96.93	< 0.001
36 months	87.48	< 0.001	51.82	< 0.001	65.91	< 0.001

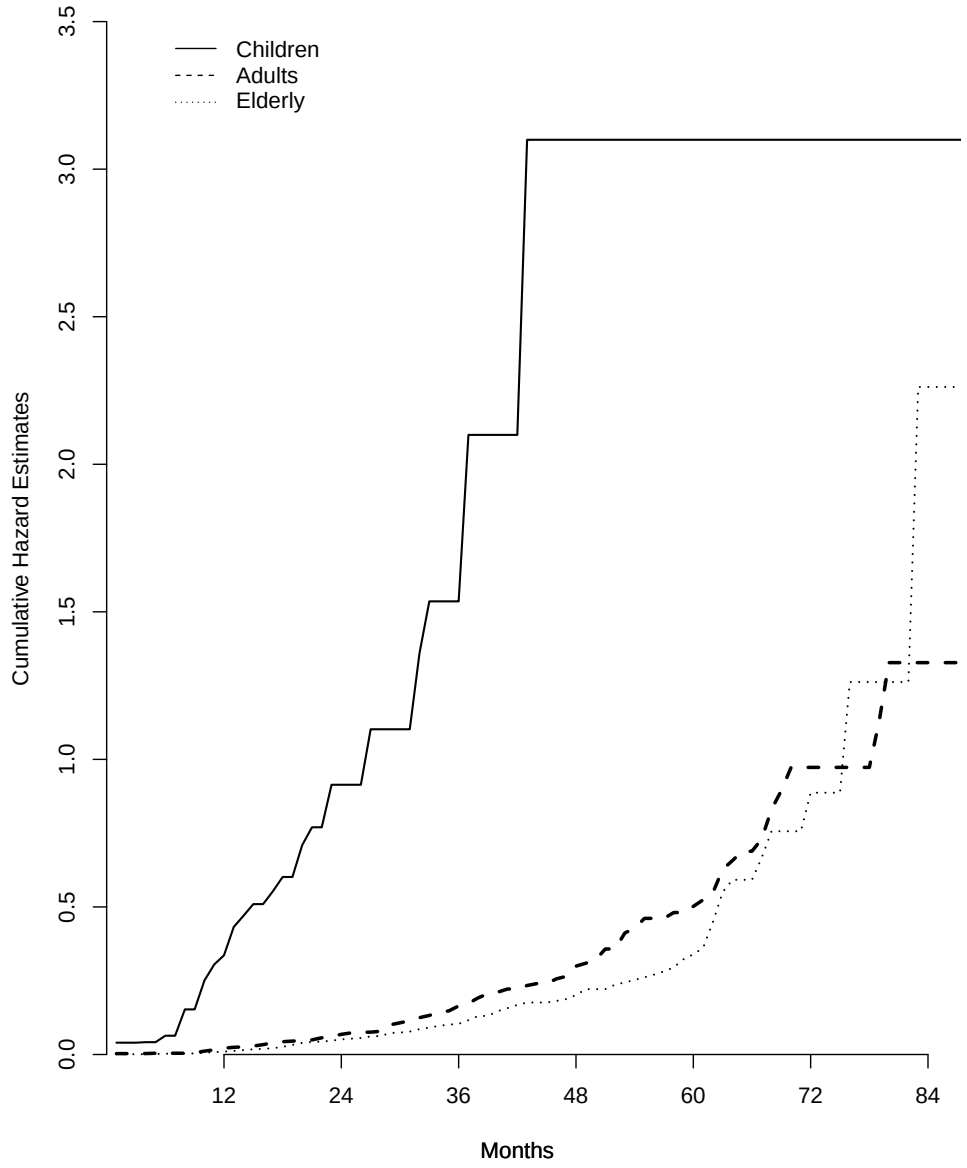

Figure 3.1 The comparisons of the cumulative hazard estimates between subgroups of AIDS data set

CHAPTER 4

NONPARAMETRIC INFERENCE FOR THE HAZARD RATE FUNCTION WITH RIGHT TRUNCATED DATA

4.1 Motivation

Chapters 2 and 3 center on inferences of cumulative hazard function with right truncated data. In some studies, the instantaneous risk of failure is of interest. If one can estimate the cumulative hazard function and plot the curve, the hazard rate function is the slope of the curve. Some smoothing technique needs to be employed to estimate the slope. The aim of this chapter is to develop the nonparametric inference of the hazard rate function with right truncated data. The commonly used kernel smoothing technique is chosen for estimating the hazard rate function.

Chapter 1 has explained that the reverse-time hazard function, not the regular hazard function, has been the study focus in the past. However, natural interpretation of the reverse-time hazard function does not exist. The hazard rate function is a dynamic measurement of the risk over time. It is an useful quantity for assessing short term treatment efficacy.

In this chapter, the kernel function estimator is first introduced for estimating the reverse-time hazard function. Common symmetric kernel functions such as uniform, Epanechnikov and biweight kernels are discussed. Subsequently, nonparametric inference of the forward-time hazard rate function is developed for right truncated data. The result from a simulation study is presented, showing satisfactory performance of the proposed inference. The AIDS blood transfusion data is revisited as an illustrative example of the inference of the hazard rate function.

4.2 Kernel Function Estimator of Reverse-Time Hazard Rate Function

We continue to use $\alpha(t)$ and $A(t)$ to denote hazard rate and cumulative hazard functions of random variable. When we have complete survival data or right censored (or left truncated) survival data, the method for estimating $\alpha(t)$ is to obtain the Nelson-Aalen estimator of $A(t)$, and then apply some smoothing technique to estimate the slope of the curve. A large number of smoothing methods have been developed for estimating the hazard rate function. Kernel smoothing, spline, and local polynomial regression are the most commonly used techniques. Kernel smoothing and local polynomial methods are theoretically more tractable than the spline approach [61]. The kernel smoothing method is considered in this chapter for estimating the hazard rate functions. Watson and Leadbetter [64]-[65] defined the kernel function estimator of the hazard rate function. Anderson *et al.* [5] (p. 231) summarized the general results of such estimator using the counting process notations, which was originally proposed by Ramlau-Hansen [53]-[54]. Let $\widehat{A}(t)$ be an estimator of $A(t)$. The kernel function estimator of $\alpha(t)$ is derived by smoothing the increments of $\widehat{A}(t)$,

$$\widehat{\alpha}(t) = \frac{1}{b}\int_0^t K\left(\frac{t-u}{b}\right) d\widehat{A}(u). \tag{4.1}$$

A kernel function is bounded in the interval $[-1, 1]$ and should be integrated to 1. The bandwidth b is a parameter taking positive values.

Some inferences related to the right truncated data have been developed on reverse-time quantities in the past. However, direct estimation of the reverse-time hazard rate function has not been studied before. The reason is the difficulty in interpreting this quantity. The fundamental aim of this chapter is to develop the inference of the regular hazard rate function. For the purpose of comparison, estimation of the reverse-time hazard rate function is discussed first. The univariate truncated sample has been defined as $\{L_i, T_i\}$ for $i = 1, \cdots, n$ and $L_i \leq T_i$. $\alpha^*(t)$ is the reverse-time hazard rate function with explanation and explicit definition given in Chapter 2. It is also defined in Chapter 2 that $A^{*+}(t) = \int_t^\tau \alpha^*(u)J(u)du$ where $J(u) = I(\bar{Y}(u) > 0)$. If $P(\bar{Y}(s) = 0)$ is really small for some $s \leq t$, then $A^{*+}(t)$ is

almost equivalent to $A^*(t)$. Introduce the quantity

$$\alpha^{*+}(t) = \frac{1}{b}\int_\tau^t K\left(\frac{t-u}{b}\right) dA^{*+}(u),$$

then $\alpha^{*+}(t)$ is very close to the smoothed version of $\alpha^*(t)$ which is

$$\alpha^{*s}(t) = \frac{1}{b}\int_\tau^t K\left(\frac{t-u}{b}\right) \alpha^*(u)du. \tag{4.2}$$

$A^*(t)$ can be estimated by the Nelson-Aalen estimator. The explicit expression is given by Equation (2.6). Similar to (4.1), the kernel function estimator of $\alpha^*(t)$ is given by

$$\widehat{\alpha}^*(t) = \frac{1}{b}\int_\tau^t K\left(\frac{t-u}{b}\right) d\widehat{A}^*(u). \tag{4.3}$$

The statistical properties of $\widehat{\alpha}^*(t)$ can be developed by using the fact that

$$\widehat{A}^*(t) - A^{*+}(t) = \int_\tau^t \frac{J(u)}{\bar{Y}(u)} d\bar{M}^*(u).$$

Then

$$\widehat{\alpha}^*(t) - \alpha^{*+}(t) = \frac{1}{b}\int_\tau^t K\left(\frac{t-u}{b}\right) d(\widehat{A}^* - A^{*+})(u) \tag{4.4}$$

$$= \frac{1}{b}\int_\tau^t K\left(\frac{t-u}{b}\right) \frac{J(u)}{\bar{Y}(u)} d\bar{M}^*(u).$$

$\widehat{\alpha}^*(t) - \alpha^{*+}(t)$ is a stochastic integral with respect to the local martingale $\bar{M}^*(t)$. The first- and second-order moments of $\widehat{\alpha}^*(t)$ exists if $E\{\widehat{\alpha}^*(t) - \alpha^{*+}(t)\}^2 < \infty$. The optional variation process of a martingale helps us to find a naive variance estimator of $\widehat{\alpha}^*(t)$,

$$\hat{\text{var}}[\widehat{\alpha}^*(t)] = \frac{1}{b^2}\int_\tau^t K^2\left(\frac{t-u}{b}\right) \frac{d\bar{N}^L(u)}{\bar{Y}^2(u)}. \tag{4.5}$$

Asymptotic normality can be established using the martingale central limit theorem.

4.3 Nonparametric Inference of Hazard Rate Function

Chapter 2.3 clarified the relation between forward- and reverse-time hazards, $A(t) = -\log(1 - \exp[-A^*(t)])$. Under the context of right truncation, the Nelson-Aalen estimator of the cumulative hazard function is not applicable. Instead, one may consider a plug-in estimator given in Equation (2.13).

In this section, the kernel-smoothed estimator of the hazard rate function is presented. The above relationship will be utilized to derive the variance of the estimator. Define $A^+(t) = \int_0^t \alpha(u)J(u)du$, we get

$$\alpha^+(t) = \frac{1}{b}\int_0^\tau K\left(\frac{t-u}{b}\right) dA^+(u) = \frac{1}{b}\int_0^\tau K\left(\frac{t-u}{b}\right) \frac{-G(u)}{1-G(u-)} dA^{*+}(u). \tag{4.6}$$

Plug in the Nelson-Aalen estimator of $A^*(t)$ and right truncated version of Kaplan-Meier estimator of $G(t)$. One will have the following estimate of $\alpha(t)$,

$$\widehat{\alpha}(t) = \frac{1}{b}\int_0^\tau K\left(\frac{t-u}{b}\right) \frac{-\widehat{G}(u)}{1-\widehat{G}(u-)} d\widehat{A}^*(u). \tag{4.7}$$

It is straightforward that

$$\widehat{\alpha}(t) - \alpha^+(t) = \frac{1}{b}\int_0^\tau K\left(\frac{t-u}{b}\right) d[A^+ - A](u) \tag{4.8}$$

$$= \frac{1}{b}\int_0^\tau K\left(\frac{t-u}{b}\right) \left[\frac{-\widehat{G}(u)}{1-\widehat{G}(u-)} d\widehat{A}^*(u) - \frac{-G(u)}{1-G(u-)} dA^{*+}(u)\right].$$

Appendix A3 shows that $(nb)^{1/2}[\widehat{\alpha}(t) - \alpha^+(t)]$ is asymptotically equivalent to the sum of functions of martingales,

$$(nb)^{1/2}[\widehat{\alpha}(t) - \alpha^+(t)] = \sqrt{\frac{1}{nb}}\int_\tau^0 H\left(\frac{t-u}{b}\right) J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)} \tag{4.9}$$

where

$$H\left(\frac{t-u}{b}\right)=\left[K\left(\frac{t-u}{b}\right)\frac{G(u)}{1-G(u-)}-\int_0^u K\left(\frac{t-x}{b}\right)d\left(\frac{G(x)}{1-G(x-)}\right)\right].$$

Through the martingale central limit theorem, $\widehat{\alpha}(t)-\alpha^{+}(t)$ converges in distribution to a normal random variable with mean zero and variance

$$\frac{1}{b^2}\int_\tau^0 H^2\left(\frac{t-u}{b}\right)\frac{\alpha^*(u)du}{y(u)}.$$

Based on above result, the variance of $\widehat{\alpha}(t)$ is estimated as

$$\widehat{\text{var}}[\widehat{\alpha}(t)]=\frac{1}{b^2}\int_\tau^0 \widehat{H}^2\left(\frac{t-u}{b}\right)\frac{d\bar{N}^L(u)}{\bar{Y}^2(u)}, \tag{4.10}$$

where

$$\widehat{H}\left(\frac{t-u}{b}\right)=\left[K\left(\frac{t-u}{b}\right)\frac{\widehat{G}(u)}{1-\widehat{G}(u-)}-\int_0^u K\left(\frac{t-x}{b}\right)d\left(\frac{\widehat{G}(x)}{1-\widehat{G}(x-)}\right)\right].$$

The kernel smoothed estimator of $\alpha(t)$ is a weighted average of crude hazard estimates over event times close to t. Most kernel functions allow the closer event times to t to have more weight than those farther from t. Bandwidth, b, is defined to control this closeness. b is chosen to include those events that are in $[t-b, t+b]$ interval. Symmetric kernel functions are commonly used such as uniform, Epanechnikov and biweight, with the following expressions:

$$K(x)=1/2, \qquad -1\le x\le 1 \quad \text{(Uniform kernel)},$$

$$K(x)=3(1-x^2)/4, \qquad -1\le x\le 1 \quad \text{(Epanecnikov kernel)},$$

$$K(x)=15(1-x^2)^2/16, \qquad -1\le x\le 1 \quad \text{(Biweight kernel)}.$$

The above kernels are applicable if $b\le t\le t_n-b$, where t_n is the biggest event time. If $t<b$, then adjustment is necessary because $t-b$ will be less than zero and inappropriate. In

this case symmetric kernels need to be modified and these modified or "asymmetric" kernels should be used. Gasser and Muller [20] suggested the boundary kernel method to modify kernels. The boundary kernel method uses linear multiples of the kernel function around the boundary, which chosen to minimize bias error.

The main problem is to find the best bandwidth to get a kernel smoothed estimate of hazard rate. There is a trade off between bias and variance in terms of choosing the bandwidth b. Generally speaking, small bandwidth will result less smooth curve; therefore, it will have smaller bias but larger variance. One way to choose an optimum bandwidth is to use mean integrated squared error (MISE) to see what value of b minimizes such error [38]. MISE of $\hat{\alpha}$ can be defined by

$$\text{MISE}(b) = E\left[\int_0^\tau [\hat{\alpha}(u) - \alpha(u)]^2 du\right]$$

$$\text{MISE}(b) = E\left[\int_0^\tau \hat{\alpha}^2(u)du\right] - 2E\left[\int_0^\tau \hat{\alpha}(u)\alpha(u)du\right] + E\left[\int_0^\tau \alpha^2(u)du\right]. \tag{4.11}$$

MISE(b) depends both on the kernel that used to estimate α and on the bandwidth b. Since the last term is independent from both kernel and bandwidth, it can be ignored. Let $t_1 < t_2 < ... < t_n$ be distinct event times, first term can be estimated by using trapezoidal rule, and the second term can be estimated by using cross-validation estimate given by Ramlau-Hansen [53]. Optimum bandwidth, b, minimizes following function [38],

$$g(b) = \sum_{i=1}^{n-1}\left(\frac{t_{i+1} - t_i}{2}\right)[\widehat{\alpha}^2(t_i) + \widehat{\alpha}^2(t_{i+1})] - \frac{2}{b}\sum_{i\neq j} K\left(\frac{t_i - t_j}{b}\right)\Delta\widehat{A}(t_i)\Delta\widehat{A}(t_j) \tag{4.12}$$

4.4 Simulation Study

A simulation study was constructed to assess the performance of the kernel smoothed estimator of the hazard rate function. Random variables (L, T) were generated with con-

straints of $L < T$. Two settings were considered for distribution of L: uniform [0,1] and exponential with mean 1 truncated at 1.2. The truncation variable T was generated from an exponential distribution with mean $1/\lambda$. Following steps were taken to generate a truncated sample with size n: First, random variables (L, T) were generated. This pair of variables would be discarded if $L > T$. Otherwise, we kept this pair in sample. Repeated this procedure until the desired sample size was obtained. Let N be the size of all generated pairs of random variables. The truncation rate is defined as $(N - n)/N$. We considered two levels of truncation rates 25% and 50%. In order to obtain a particular truncation rate, we searched for appropriate value for λ for the distribution of the truncation variable.

Each simulated setting contained 1000 replicates. For simplicity, a uniform kernel was used in a smoothing process. In order to obtain the optimum bandwidth, we searched for b, that minimized $g(b)$ given in (4.15) for each replicate. Searching for the optimum bandwidth can be computationally challenging when the sample size is large. Due to this limitation, the sample size used in simulation was chosen to be 200. Let $\bar{\widehat{\alpha}}(t)$ be the average of the kernel smoothed hazard estimates of 1000 replicates and $\widehat{\alpha}^{(i)}(t)$ be the kernel smoothed hazard estimate for the ith replicate, Then

$$\bar{\widehat{\alpha}}(t) = \sum_{i=1}^{1000} \widehat{\alpha}^{(i)}(t).$$

The relative bias provides a measure of the magnitude of the bias:

$$\text{Relative bias} = \frac{B[\bar{\widehat{\alpha}}(t)]}{\alpha(t)} = \frac{\bar{\widehat{\alpha}}(t) - \alpha(t)}{\alpha(t)}$$

where the bias, $B[\bar{\widehat{\alpha}}(t)]$, was defined as the deviation between the average kernel smoothed hazard estimate and the true value.

The variance estimator $\hat{\text{var}}[\widehat{\alpha}(t)]$ was evaluated for each replicate and the average of these values was calculated by

$$\text{Estimated variance} = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\text{var}}[\widehat{\alpha}^{(i)}(t)].$$

The sample variance was calculated using the formula

$$\text{Sample variance} = \frac{1}{1000-1}\sum_{i=1}^{1000}\left(\widehat{\alpha}^{(i)}(t) - \bar{\widehat{\alpha}}(t)\right)^2.$$

The 95% confidence interval for the hazard rate function for each replicate was calculated and the actual coverage fraction across 1000 replicates was obtained. The estimation results were reported at time points that corresponds to 0.2, 0.5, 0.8 in $G(t)$. For this reason, results were evaluated at $t = 0.2, 0.5, 0.8$ for the uniform distribution and settings at $t = 0.15, 0.43, 0.82$ for the truncated exponential distribution (see Table 4.1).

The relative biases are very small for all the settings. Although there is no clear trend but larger time points are associated with smaller relative biases in general. Estimated variances are very close to sample variances indicating good performance of the variance estimator. The observed coverage proportions are slightly below the nominal level but the result is acceptable in general.

Table 4.1 The simulation results for estimating the hazard rate function based on 1000 replicates with size 200.

Distribution of L	Truncation rate	t	Relative bias (%)	Sample variance	Estimated variance	Coverage
Uniform[0,1]	25	0.20	1.2	0.492	0.484	0.930
		0.50	1.8	0.860	0.844	0.926
		0.80	1.2	2.194	2.249	0.926
	50	0.20	-0.9	0.389	0.377	0.920
		0.50	-0.8	0.735	0.742	0.937
		0.80	0.4	2.304	2.262	0.919
Exponential(1.0)	25	0.15	0.5	0.391	0.380	0.937
truncated at 1.2		0.43	-1.7	0.570	0.559	0.918
		0.82	0.3	1.352	1.278	0.908
	50	0.15	0.1	0.314	0.301	0.924
		0.43	0.2	0.488	0.498	0.942
		0.82	0	1.385	1.322	0.911

4.5 The AIDS Latent Time Example

In this simulation study, the blood transfusion infected AIDS data is analyzed. Details about data given in Chapters 1.3 and 3.5. Three subgroups of data set considered for analysis: children (age range 1-4 years), adults (age range 5-59 years) and elderly patients (age $\geq$ 60). Sample sizes are 34 for children, 120 for adults and 141 for elderly people. The largest incubation times recorded are 43, 89, and 83 months for children, adults and elderly patients, respectively. Our goal is to get kernel smoothed hazard rate estimates and compare them between different groups. Comparisons between adults vs children and elderly vs children are graphed until 40 months as the largest incubation time for children was 43 months. For similar reasons, comparison between adults vs elderly goes up to 80 months.

We have used kernel smoothing to get a smoothed hazard rate function for right truncated data. We looked for the optimum bandwidth for each group using three different kernels. For the uniform kernels, optimum bandwidth selections were $b = 5$ for adults, $b = 8$ for the elderly and $b = 8$ for children. In Figure 4.1, smoothed hazard functions for three types of kernels are plotted for each group. Epanechnikov and biweight kernels assign more weight in the middle and less weight towards the tails where the uniform kernel assigns a homogeneous weight. For weight homogeneity, illustration purposes and simplicity, we chose a uniform kernel to smooth the hazard function for right-truncated data.

Figure 4.1 shows that the kernel-smoothed hazard rate estimates for adults increase by time for all three kernels. There is a sudden decrease towards the end when Epanechnikov or biweight kernels are used. In elderly patients, the Epanechikov kernel increases up to 50 months and levels off afterwards. The hazard rate smoothed with the biweight and uniform kernels shows similar trends, increasing after 50 months. Children had higher smoothed hazard rates compared with the other two groups. All hazard rate estimates show a sudden jump around 5 months and increase slowly up to 30 months. The uniform and biweight kernel smoothed hazard rate estimates increase after 30 months where the Epanechnikov kernel smoothed hazard estimate stays flat for children. Using a uniform kernel for smoothing

distributes weight evenly for all time points.

Figure 4.2 shows the uniform kernel smoothed hazard rate functions and pointwise 95% confidence intervals for each group. The pointwise confidence intervals are very wide, even include negative values for adults and elderly after 60 months. The estimated hazard rates of these two groups are associated with low degree of precision. The children group had much higher hazard rate estimates and the 95% pointwise confidence intervals are slightly narrower, compared to the result in other two groups.

Figure 4.3 shows the estimated differences between two kernel smoothed hazard rate functions and 95% pointwise confidence intervals. The differences of smoothed hazard rates between adults and elderly is not significant since 95% confidence intervals includes zero; The differences of smoothed hazard rates between children and the other two groups are significant, indicating higher instantaneous risks of AIDS onset in infected groups.

4.6 Discussion

The aim of this chapter was to study one important survival quantity, the hazard rate function for right-truncated data. The reverse-time hazard rate has been studied by many researchers but the forward-time hazard rate has not received the same degree of attention. One of the earliest researches on the forward-time hazard was done by Finkelstein, Moore and Schoenfeld [16]. They studied the Cox model for right truncated data and proposed to use the full likelihood to estimate regression parameters. Estimations of the hazard rate function helps one to examine the shape of the function and gives a direct assessment of proportional assumption in case of multiple samples. Nonparametric inference makes it feasible to compare hazard rate functions of different groups without any time transformation. Pointwise comparison of hazard rates between two samples can be implemented by finding a confidence interval for the differences of the hazard rate.

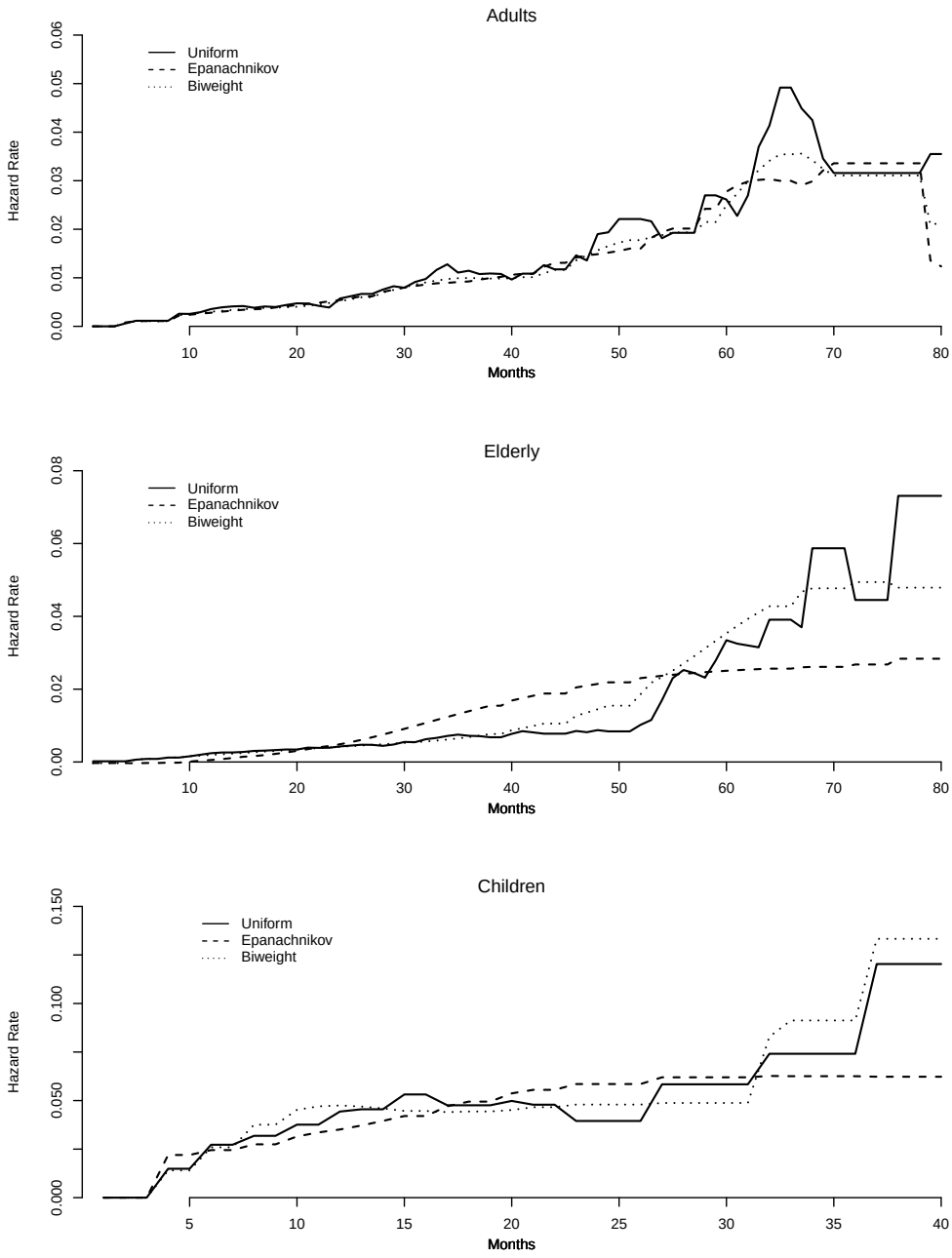

Figure 4.1 Smoothed hazard rate curves using three kernels

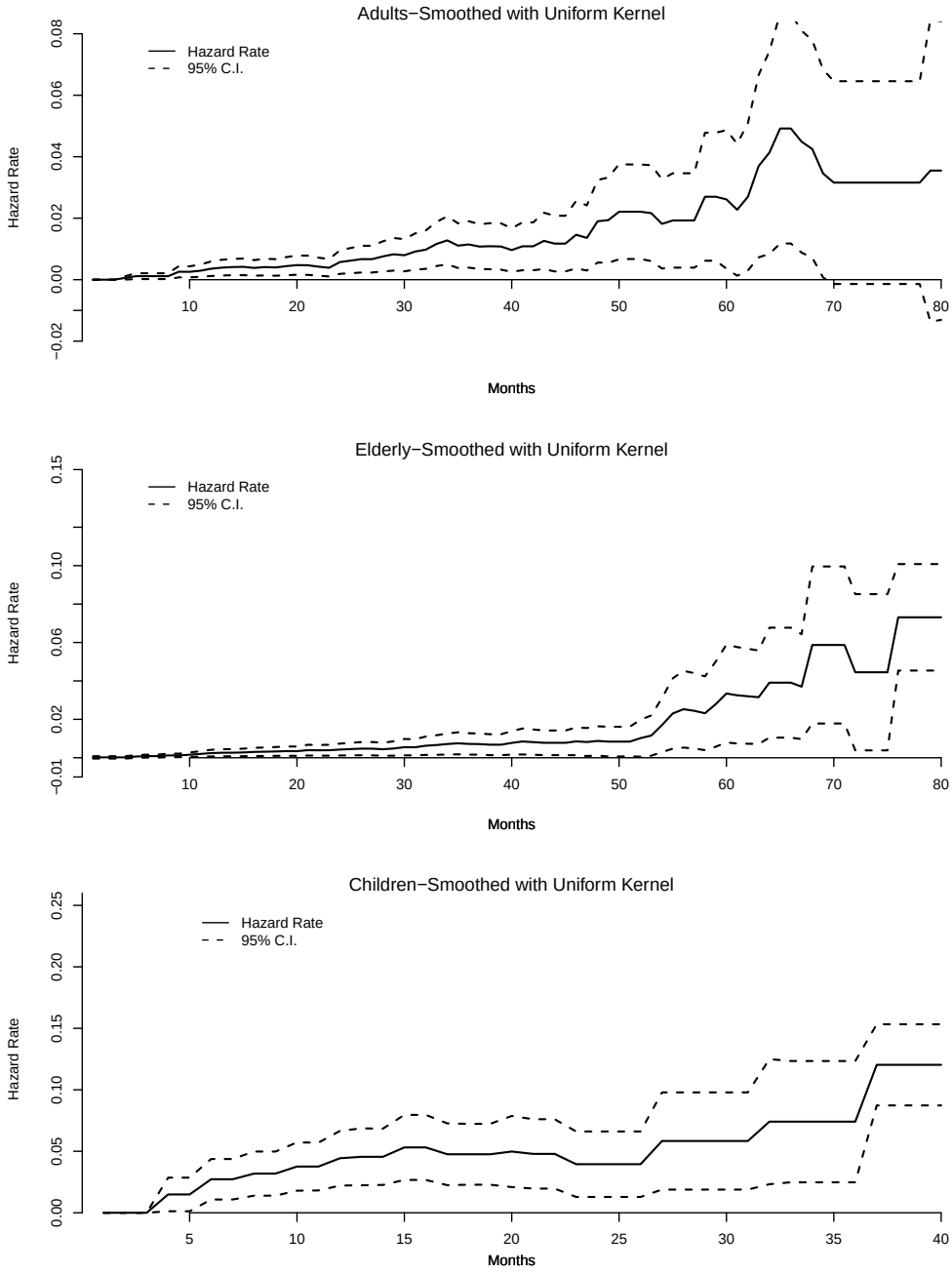

Figure 4.2 Uniform-kernel smoothed hazard rate curves and 95% confidence intervals

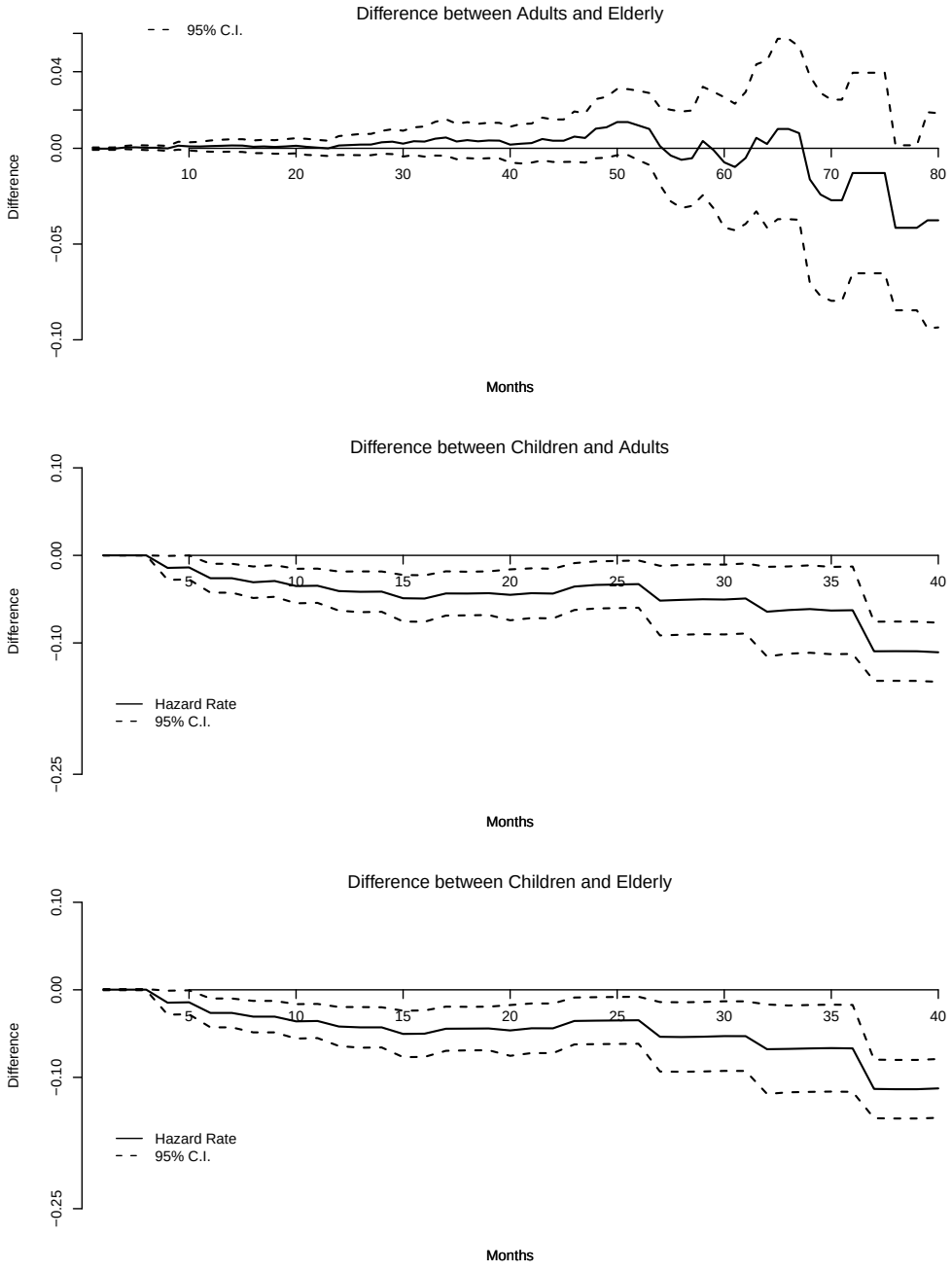

Figure 4.3 Differences of uniform-kernel smoothed hazard rate estimates and 95% confidence intervals

CHAPTER 5

CONCLUSIONS

The fact that few researches have been done on the inferences for the hazard function with right truncated data was my motivation to study this subject. Chapter 1 of this dissertation describes different types of incompleteness in time-to-event data, as well as the subcategories in censoring and truncation. Between two types of truncation, left and right truncation, left truncation has received more attention. One of the earliest studies about random truncation model was published by Lynden-Bell [44], who noted the truncation issue in astronomic data. In time-to-event data, left truncation takes the form of left entrance. Analysis of cancer or bone-marrow transplant registry data often involves the complexity of left truncation.

Lagakos *et al.* [40] were one of the pioneers to suggest transforming the right-truncated variable to the left-truncated one and then applying the existing statistical inferences. The reverse-time hazard function was introduced by this type of transformation. Inferences such as log-rank test, Cox regression model were developed on this quantity. However, lack of natural interpretation remains an unsolved issue for the reverse-time hazard function. A few researches have been done in the recent years on the forward-time hazard function, including the full-likelihood-based Cox model by Finkelstein, Moore and Schoenfeld [16] and the semi-parametric log-rank test by Shen [56].

Chapter 1 introduces basic concepts and provides a literature review for analysis of truncated data. First, the concepts and properties of filtration, martingales and counting process were reviewed, following by a discussion of predictable and optional variation process of a martingale process. Rebolledo's [55] version of the martingale central limit theorem was presented. The Nelson-Aalen estimator of the cumulative hazard function and the Kaplan-Meier estimator of the survival function were discussed briefly. Another major component

of this chapter is a description of censoring and truncation. Right censoring is almost an inevitable issue in analysis of survival data. Occurrence of truncation is not as frequent as censoring. The mechanism of truncation is not well understood by researchers in other disciplines. The issue of truncation may be ignored even though it is truly present. In this chapter, the difference between truncation and censoring as well as the relation between left and right truncation are clarified. The last component of this chapter is a review of the literature related to statistical analysis of truncated data.

Chapter 2 develops nonparametric inference for forward-time cumulative hazard function. The existing inference for the cumulative reverse-time hazard function is first presented, including an explicit definition of the reverse-time martingale. Using the relation between forward-time and reverse-time hazards, weak convergence of estimated cumulative hazard is derived. Two existing variance estimators are revised to correct the problem of overestimation when t is large. The weighted one-sample log-rank test is the new development. The revised variance estimators of the cumulative hazard and the one-sample test show satisfactory performances in the simulation studies.

Chapter 3 studies a family of weighted log-rank tests for comparing survival outcomes among independent samples. A test statistic is proposed and its asymptotic normality is studied. Selection of weight function leads to different types of tests, including the well-known log-rank, Gehan and Tarone-Ware tests. Simulation studies designed for two-sample and three-sample tests show satisfactory results. Application of the proposed tests has been demonstrated on the AIDS blood transfusion data set, for which the hazard rate functions of three age subgroups are compared.

Chapter 4 studies the nonparametric inference of the hazard rate function of right truncated data. The kernel smoothed estimator of the forward-time hazard rate is proposed. Different choices of kernel function such as uniform, Epanechnikov and biweight kernels are discussed. Weak convergence of the kernel smoothed estimator of the hazard rate function is provided in the appendix. The estimator of the hazard rate function using the uniform kernel s investigated in the simulation study, yielding a low level of relative bias. The criterion

to select optimum bandwidth is presented and implemented in both simulation study and example. The AIDS blood transfusion data set is revisited to illustrate the developed methods. Three kernels, uniform, Epanechnikov and biweight are all implemented to estimate the hazard rate functions of three age subgroups.

Future work on analysis of right truncated data can be Cox regression analysis. The Cox proportional hazards model is the most commonly used regression model for survival data because the result is easy to interpret. Finkelstein, Moore and Schoenfeld [16] studied the Cox model for right truncated data using the full likelihood. It should be interesting to investigate a solution based on the partial likelihood of Cox model with right truncated data. The tentative solution is a weighted score estimating equation which stems from the partial likelihood. Proper weigh function should be employed to compensate probabilities of selection, which vary among subjects in the truncated sample. This approach is expected to have the advantage of computational efficiency.

Another path for future research is to consider other weight functions to extend the tests given in Chapter 3. Peto and Peto [52] suggested using a weight function close to the Kaplan-Meier estimator of the survival function of the pooled samples. Fleming and Harrington [18] proposed a weight function which is the product of the power functions of the pooled Kaplan-Meier estimator and its complement. One can consider tests using these weight functions to analyze right truncated data. Following Peto and Peto approach, one can utilize the weight function $\tilde{G}_\bullet(t)$ for the K-sample test where

$$\tilde{G}_\bullet(t) = \prod_{u>t} \left(1 - \frac{d[\sum_{k=1}^{K} \sum_{i=1}^{n_k} I(L_{ki} \leq u)]}{\bar{Y}_\bullet(u) + 1} \right). \tag{5.1}$$

Note that $\tilde{G}_\bullet$ is close to the right truncated version Kaplan-Meier estimator of $P(L \leq t)$. Similarly, Fleming and Harrington approach leads to a weight function $W_{p,q}(t) = [\hat{G}_\bullet(t)]^p[1 - \hat{G}_\bullet(t)]^q$ where $p \geq 0$ and $q \geq 0$, where $\hat{G}_\bullet(t)$ denotes the right truncated version Kaplan-Meier estimator of pooled survival function. The tests developed in Chapter 3, together with Peto and Peto's test and Fleming and Harrington's test have little power in detecting

the differences if hazard rate functions cross. Renyi type tests can be constructed for right truncated data to achieve better power for the context of crossing hazard functions.

One meaningful research extended from this dissertation is to compare the performances of the weighted log-rank tests to other tests. One candidate is the pointwise comparison of the survival probabilities, $H_0 : S_1(t) = S_2(t)$. The hypothesis can be tested by a Wald test using the Kaplan-Meier estimates of the distribution probabilities at t. Another candidate is the test proposed by Chi *et al.* [12] for the hypothesis $H_0 : S_1(u) = S_2(u)$, $0 \leq u \leq \tau$. Their test statistic is the integrated weighted differences between distribution probability estimates. In addition, one can consider a two-sample median test for right truncated data. Such a test was initially studied by Brookmeyer and Crowley [8] for censored survival data and the test has acceptable power to detect the differences between survival functions when the hazard rate functions cross. More investigation is needed to construct the test statistic suitable for right truncated samples. It is very interesting and practically useful to design a Monte-Carlo study to evaluate the aforementioned tests, together with the tests developed in Chapter 3 of this dissertation.

REFERENCES

[1] Aalen, O.O. Statistical inference for a family of counting processes. *PhD thesis, University of Berkeley, California*, 1975.

[2] Aalen, O.O. Nonparametric inference for a family of counting processes. *Annals of Statistics*, Vol. 6, pp. 701-726, 1978.

[3] Altshuler, B. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, Vol. 6, pp. 1-11, 1970.

[4] Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N. Censoring, truncation and filtering in statistical models based on counting processes. *Contemporary Mathematics*, Vol. 80, pp. 19-60, 1988.

[5] Andersen, P.K., Borgan, O., Gill R.D., Keiding, N. Statistical Models Based on Counting Processes. *New York: Springer Series in Statistics*, 1993.

[6] Billingsley, P. Statistical Inference for Markov Processes. *University of Chicago Press, Chicago*, 1961.

[7] Bremaud, P. A martingale approach to point processes. *PhD thesis, Electrical Research Laboratory, Berkeley*, 1972.

[8] Brookmeyer, R., Crowley, J.J. A k-sample median test for censored data. *Journal of the American Statistical Association*, Vol. 77, pp. 433-440, 1982.

[9] Brown, B.M. Martingale central limit theorems. *Annals of Mathemetical Statistics*, Vol. 42, pp. 59-66, 1971.

[10] Chao, M.T., Lo, S.H. Some representations of the non-parametric maximum likelihood estimators with truncated data. *The Annals of Statistics*, Vol. 16, pp. 661-668, 1988.

[11] Chen, K., Chao, M.T., Lo, S.H. On strong uniform consistency of the Lynden-Bell estimator for truncated data. *The Annals of Statistics*, Vol. 23, pp.440-449, 1995.

[12] Chi, Y., Tsai, W.Y., Chiang, C.L. Testing the equality of two survival functions with right truncated data. *Statistics in Medicine*, Vol. 26, pp. 812-827, 2007.

[13] Doob, J.L. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, Vol. 47, pp. 455-486, 1940.

[14] Doob, J.L. Stochastic Processes. *Wiley, New York*, 1953.

[15] Dvoretsky, A. Asymptotic normality for sums of dependent random variables. *In Proceedings of Sixth Berkeley Symposium of Mathematics, Statistics and Probability, University of California Press, Berkeley*, Vol. 2, pp. 513-535, 1972.

[16] Finkelstein, D.M., Moore, D.F., Schoenfeld, D.A. A proportional hazards model for truncated AIDS data. *Biometrics*, Vol. 49, pp. 731-740, 1993.

[17] Fischer, T. On simple representations of stopping times and stopping time sigma-algebras. *Statistics and Probability Letters*, Vol. 83, pp. 345-349, 2013.

[18] Fleming, T.R. and Harrington, D.P. A class of hypothesis tests for one and two samples of censored survival data. *Communications in Statistics*, Vol. 10, pp. 763-794, 1981.

[19] Fleming, T.R. and Harrington, D.P. Counting Processes and Survival Analysis. *Wiley, New York*, 1991.

[20] Gasser, T.H., Muller H.G. Kernel estimation of regression functions. *Lecture Notes in Mathematics*, Vol. 757, pp. 23-68, 1979.

[21] Gehan, E.A. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, Vol. 52, pp. 203-223, 1965.

[22] Gill, R.D. On estimating transition intensities of a Markov process with aggregated data of a certain type: " Occurrences but no exposures". *Scandinavian Journal of Statistics*, Vol. 13, pp. 113-134, 1986.

[23] Gross, S.T., Huber-Carol, C. Regression models for truncated survival data. *Scandinavian Journal of Statistics*, Vol. 19, pp. 193-213, 1992.

[24] Gurler, U., Stute, W., Wang, J.L. Weak and strong quantile representations for randomly truncated data with applications. *Statistics and Probability Letters*, Vol. 17, pp. 139-148, 1993.

[25] Hald, A. Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Scandinavian Actuarial Journal*, Vol. 1949, pp. 119-134, 1949.

[26] Hald, A. Statistical Theory with Engineering Applications. *Wiley, New York*, 1952.

[27] Halley, E. An estimate of the degrees of the mortality of mankind drawn from curious tables of the births and funerals at the city Breslaw. *Philosophical Transactions Royal Society of London*, Vol. 17, pp. 596-610, 1693. [Reprinted in *Journal of Institute of Actuaries*, Vol. 112, pp. 278-301, 1985.]

[28] Ito, K. Stochastic Integral. *Proceedings of Imperial Academy, Tokyo*, Vol. 20, pp. 519-524, 1944.

[29] Jacod, J. On the stochastic intensity of a random point process over the half-line. *Technical Report 15, Department of Statistics, Princeton University*, 1973.

[30] Jacod, J. Multivariate point processes: Predictable projection, Radon-Nikodym derivatives, representation of martingales. *Z.Wahrsch. verw. Gebiete* , Vol. 31, pp. 235-253, 1975.

[31] Jacod, J. and Shiryaev,A.N. Limit Theorems for Stochastic Processes. *Springer-Verlag, Berlin*, 1987.

[32] Jiang, Y. Estimation of hazard function for right truncated data. *Georgia State University, Master Thesis*, 2010.

[33] Kalbfleisch, J.D., Lawless, J.F. Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, Vol. 84, pp. 360-372, 1989.

[34] Kalbfleisch, J.D., Lawless, J.F. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, Vol. 1, pp. 19-32, 1991.

[35] Kaplan, E.L., Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Vol. 53, pp. 457-481, 1958.

[36] Keiding, N., Gill, R.D. Random truncation models and Markov process. *The Annals of Statistics*, Vol. 18, pp. 582-602, 1990.

[37] Klein, J.P. Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. *Scandinavian Journal of Statistics*, Vol. 18, pp. 333-340, 1991.

[38] Klein, J.P., Moeschberger, M.L. Survival Analysis: Techniques for Censored and Truncated Data. *New York: Springer Series Statistics for Biology and Health*, 2003.

[39] Klein, J.P., Zhang, M.J. Statistical challenges in comparing chemotherapy and bone-marrow transplantation as a treatment for leukemia. *Life Data: Models in Reliability and Survival Analysis*, Vol. ed. NP et al., pp. 175-185, 1996.

[40] Lagakos, S.W., Barraj, L.M., Gruttola, V. Nonparametric analysis of truncated survival data with applications to AIDS. *Biometrika*, Vol. 75, pp. 515-523, 1998.

[41] Lai, T.L., Ying, Z. Estimating a distribution function with truncated and censored data. *The Annals of Statistics*, Vol. 19, pp. 417-442, 1991.

[42] Levy, P. Proprietees asymptotiques des sommes de variables aleatoires en chainees. *Bulletin de la Societe Mathematique de France*, Vol. 59, pp. 84-96, 1935.

[43] Lui, K.J., Lawrence, D.L., Morgan, W.M., Peterman, T.A., Haverkos, H.W., Bregman, D.J. A model-based approach for estimating the mean incubation period of transfusion-associated Acquired Immunodeficiency Syndrome. *Proceedings of National Academy of Sciences USA*, Vol. 83, pp. 3051-3055, 1986.

[44] Lynden-Bell, D. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Royal Astronomical Society*, Vol. 155, pp. 95-118, 1974.

[45] Mansuy, R. The origins of the word "martingale". *Electronic Journal for History of Probability and Statistics*, Vol. 5,Retrieved 02-27-2013, 2009.

[46] Medley, G.F., Anderson, R.M., Cox, D.R. and Billiard, L. Incubation period of AIDS in patients infected via blood transfusion, *Nature,London*, Vol. 328, pp. 719-721, 1987.

[47] Medley, G.F., Billiard, L., Cox, D.R. and Anderson, R.M. The distribution of the incubation period for the Acquired Immunodeficiency Sydrome (AIDS). *Proceedings of Royal Sciety of London, Series B*, Vol. 233, pp. 367-377, 1988.

[48] Meyer, P.-A. A decomposition theorem for supermartingales. *Illinois Journal of Mathematics*, Vol. 6, pp. 193-205, 1962.

[49] CDC. Pneumocystis pneumonia-LosAngeles. *MMWR*, Vol. 30, pp. 1-3, 1981.

[50] Nelson, W. Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, Vol. 1, pp. 27-52, 1969.

[51] Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, Vol. 14, pp. 945-965, 1972.

[52] Peto, R., Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, Vol. A 135, pp. 185-206, 1972.

[53] Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Annals of Statistics*, Vol. 11, pp. 453-466, 1983.

[54] Ramlau-Hansen, H. The choice of a kernel function in the graduation of counting process intensities. *Scandinavian Actuarial Journal*, pp. 165-182, 1983.

[55] Rebolledo, R. Central limit theorems for local martingales. *Z.Wahrsch. verw. Gebiete*, Vol. 51, pp. 269-286, 1980.

[56] Shen, P. A class of semiparametric rank-based tests for right-truncated data. *Statistics and Probability Letters*, Vol. 80, pp. 1459-1466, 2010.

[57] Tarone, R.E. and Ware, J.H. On distribution-free tests for equality for survival distributions. *Biometrika*, Vol. 64, pp. 156-160, 1977.

[58] Tsai, W.Y. Testing the assumption of independence between truncated time and failure time. *Biometrika*, Vol. 77, pp. 169-178, 1990.

[59] Uzunogullari, U. and Wang, J.L. A comparison of hazard rate estimators for left truncated and right censored data. *Biometrika*, Vol. 79, pp. 297-310, 1992.

[60] Von Wiezsacker, H. and Winkler, H. Stochastic Integrals. An Introduction. *Vieweg, Braunschweig*, 1990.

[61] Wang, J.L. Smoothing hazard rate. *Encyclopedia of Biostatistics, 2nd Edition*, Vol. 2, pp. 4986-4997, 2005.

[62] Wang, M.C. A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* , Vol. 84, pp. 742-748, 1989.

[63] Wang, M.C., Jewell, N.P., Tsai, W.Y. Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, Vol. 14, pp. 1597-1605, 1986.

[64] Watson, G.S., Leadbetter, M.R. Hazard Analysis I. *Biometrika*, Vol. 51, pp. 175-184, 1964.

[65] Watson, G.S., Leadbetter, M.R. Hazard Analysis II. *Sankhya Series A*, Vol. 26, pp. 101-116, 1964.

[66] Woodroofe, M. Estimating a distribution function with truncated data. *Annals of Statistics*, Vol. 13, pp. 163-177, 1985.

Appendix A

ASYMPTOTIC PROPERTIES

A1

The one-sample log-rank test was discussed in Chapter 2. This appendix sketches the asymptotic distribution of the proposed test statistic. Let $\alpha_0(t)$ and $A_0(t)$ be the true hazard rate and cumulated hazard functions. $\alpha_0^*(t)$ and $A_0^*(t)$ are the corresponding reverse-time hazard rate and cumulative hazard functions. The test statistics for $H_0 : \alpha(t) = \alpha_0(t)$ is given by $Z(t) = \int_0^t W(u)d[\widehat{A}(u) - A_0(u)]$, where Equation (2.13) presents the explicit expression of $\widehat{A}(t)$. Then the test statistic formula can be evaluated as

$$\sqrt{n}\int_0^t W(u)d[\widehat{A}(u) - A_0(u)] = \sqrt{n}\int_0^t W(u)[\frac{-\widehat{G}(u)}{1-\widehat{G}(u)}d\widehat{A}^*(u) - \frac{-G_0(u)}{1-G_0(u)}dA_0^*(u)].$$

Add and subtract an interim term, $\dfrac{\widehat{G}(u)}{1-\widehat{G}(u)}dA_0^*(u)$, there will be

$$\sqrt{n}\int_t^0 W(u)\frac{\widehat{G}(u)}{1-\widehat{G}(u)}d[\widehat{A}^*(u) - A_0^*(u)] + \sqrt{n}\int_t^0 W(u)[\frac{\widehat{G}(u)}{1-\widehat{G}(u)} - \frac{G_0(u)}{1-G_0(u)}]dA_0^*(u).$$

In the following context, $\approx$ indicates asymptotic equivalence. Let $\int_\tau^t Y_i(u)\alpha_0^*(u)du$ be the compensator of counting process $N_i^L(t)$ and $M_i^*(t) = N_i^L(t) - \int_\tau^t Y_i(u)\alpha_0^*(u)du$ is a martingale. Under the null hypothesis, the first term is asymptotically equal to the sum of martingales

$$\sqrt{n}\int_t^0 W(u)\frac{\widehat{G}(u)}{1-\widehat{G}(u)}d[\widehat{A}^*(u) - A_0^*(u)] \approx \sqrt{n}\int_t^0 W(u)\frac{G_0(u)}{1-G_0(u)}[J(u)\frac{d\bar{N}^L(u)}{\bar{Y}(u)} - \alpha_0^*(u)du]$$

$$\approx \sqrt{n}\int_t^0 W(u)\frac{G_0(u)}{1-G_0(u)}[J(u)\frac{d\bar{M}^*(u) + \alpha_0^*(u)\bar{Y}(u)du}{\bar{Y}(u)} - \alpha_0^*(u)du]$$

$$= \sqrt{n}\int_t^0 W(u)\frac{G_0(u)}{1-G_0(u)}J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}.$$

For the second term, Taylor series expansion can be applied

$$\begin{aligned}\sqrt{n}\int_t^0 W(u)[\frac{\widehat{G}(u)}{1-\widehat{G}(u)} - \frac{G_0(u)}{1-G_0(u)}]dA_0^*(u) &\approx \sqrt{n}\int_t^0 W(u)[\widehat{A}^* - A_0^*](u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\\ &= \sqrt{n}\int_t^0 W(u)\left[\int_\tau^s J(u)\frac{d\bar{M}^*(u)+\alpha_0^*(u)\bar{Y}(u)du}{\bar{Y}(u)} - \int_\tau^s \alpha_0^*(u)du\right]d\left(\frac{G_0(u)}{1-G_0(u)}\right)\\ &= \sqrt{n}\int_t^0 W(u)\left[\int_\tau^s J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}\right]d\left(\frac{G_0(u)}{1-G_0(u)}\right).\end{aligned}$$

Change the order of integration in above double integrals,

$$\sqrt{n}\int_0^t\left[W(u)\int_0^s d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)} + \sqrt{n}\int_t^\tau\left[W(u)\int_0^s d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}.$$

Combining the first and second terms leads to the following result,

$$\begin{aligned}\sqrt{n}\int_0^t W(u)d[\widehat{A}(u)-A_0(u)] \approx \sqrt{n}\int_t^0 &\left[W(u)\frac{G_0(u)}{1-G_0(u)} - \int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}\\ &-\sqrt{n}\int_\tau^t\left[\int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}\end{aligned}$$

$\sqrt{n}\int_0^t W(u)d[\widehat{A}(u)-A_0(u)]$ converges in distribution to a zero-mean normal random variable with variance

$$\begin{aligned}&\int_t^0\left[W(u)\frac{G_0(u)}{1-G_0(u)} - \int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]^2\frac{\alpha_0^*(u)du}{y(u)}\\ &+\int_\tau^t\left[\int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]^2\frac{\alpha_0^*(u)du}{y(u)}.\end{aligned}$$

The variance of $Z(t)$ can be estimated by

$$\widehat{\sigma}^2(t) = \int_0^t\left[W(u)\frac{G_0(u)}{1-G_0(u)} - \int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]^2\frac{d\bar{N}(u)}{\bar{Y}^2(u)}$$

$$+\int_t^\tau \left[\int_0^s W(u)d\left(\frac{G_0(u)}{1-G_0(u)}\right)\right]^2 \frac{d\bar{N}(u)}{\bar{Y}^2(u)}.$$

Based on the above weak convergence result, $U(t) = Z(t)/\widehat{\sigma}(t)$ follows standard normal distribution.

A2

The asymptotic properties of K-sample test is given in this appendix. In Chapter 3, the test statistic for the K-sample context is given as $Z_k(t) = \int_0^t W_k(s)d[\widehat{A}_k(s) - \widehat{A}_\bullet(s)]$. Using the relation in (2.15),

$$\sqrt{n}\int_0^t W_k(s)d[\widehat{A}_k(s) - \widehat{A}_\bullet(s)] = \sqrt{n}\int_0^t [\frac{-\widehat{G}_k(s)}{1-\widehat{G}_k(s)}W_k(s)d\widehat{A}^*_k(s) - \frac{-\widehat{G}_\bullet(s)}{1-\widehat{G}_\bullet(s)}W_k(s)d\widehat{A}^*_\bullet(s)].$$

Add and subtract the interim term $\dfrac{\widehat{G}_k(s)}{1-\widehat{G}_k(s)}d\widehat{A}^*_\bullet(s)$ to above equation,

$$\sqrt{n}\int_t^0 \frac{\widehat{G}_k(s)}{1-\widehat{G}_k(s)}W_k(s)d[\widehat{A}^*_k(s) - \widehat{A}^*_\bullet(s)] + \sqrt{n}\int_t^0 \left[\frac{\widehat{G}_k(s)}{1-\widehat{G}_k(s)} - \frac{\widehat{G}_\bullet(s)}{1-\widehat{G}_\bullet(s)}\right] W_k(s)d\widehat{A}^*_\bullet(s).$$

Under the null hypothesis, the first term is asymptotically equal out to the following expression

$$\sqrt{n}\int_t^0 W_k(u)\frac{G_\bullet(u)}{1-G_\bullet(u)}d[\widehat{A}^*_k(s)-\widehat{A}^*_\bullet(s)] = \sqrt{n}\int_t^0 W_k(u)\frac{G_\bullet(u)}{1-G_\bullet(u)}J_k(u)\left[\frac{d\bar{M}^*_k(u)}{\bar{Y}_k(u)} - \frac{d\bar{M}^*_\bullet(u)}{\bar{Y}_\bullet(u)}\right].$$

The similar technique can be applied to the second term. Add and subtract the interim term $\dfrac{G_\bullet(s)}{1-G_\bullet(s)}$, then we will have

$$\sqrt{n}\int_t^0 W_k(s)\left[\left(\frac{\widehat{G}_k(s)}{1-\widehat{G}_k(s)} - \frac{G_\bullet(s)}{1-G_\bullet(s)}\right) - \left(\frac{\widehat{G}_\bullet(s)}{1-\widehat{G}_\bullet(s)} - \frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right] d\widehat{A}^*_\bullet(s).$$

Using the plug-in estimator $d\widehat{A}^*_\bullet(s) = \dfrac{-d\widehat{G}_\bullet(s)}{\widehat{G}_\bullet(s)}$ and applying Taylor series expansions

$$\left(\frac{\widehat{G}_k(s)}{1-\widehat{G}_k(s)} - \frac{G_\bullet(s)}{1-G_\bullet(s)}\right) \approx \frac{-G_\bullet(s)}{(1-G_\bullet(s))^2}(\widehat{A}^*_k - A^*_\bullet)(s)$$

and

$$\left(\frac{\widehat{G}_\bullet(s)}{1-\widehat{G}_\bullet(s)} - \frac{G_\bullet(s)}{1-G_\bullet(s)}\right) \approx \frac{-G_\bullet(s)}{(1-G_\bullet(s))^2}(\widehat{A}^*_\bullet - A^*_\bullet)(s).$$

For the second term, we now have

$$\sqrt{n}\int_t^0 W_k(s)\left[(\widehat{A}^*_k - A^*_\bullet)(s) - (\widehat{A}^*_\bullet - A^*_\bullet)(s)\right] d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)$$

Note that under the null hypothesis

$$(\widehat{A}^*_k - A^*_\bullet)(s) = \int_\tau^s J_k(u)\frac{d\bar{M}^*_k(u)}{\bar{Y}_k(u)} \qquad \text{and} \qquad (\widehat{A}^*_\bullet - A^*_\bullet)(s) = \int_\tau^s J_k(u)\frac{d\bar{M}^*_\bullet(u)}{\bar{Y}_\bullet(u)}$$

then

$$\sqrt{n}\int_t^0 W_k(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\left[\int_\tau^s J_k(u)\frac{d\bar{M}^*_k(u)}{\bar{Y}_k(u)} - \int_\tau^s J_k(u)\frac{d\bar{M}^*_\bullet(u)}{\bar{Y}_\bullet(u)}\right].$$

Changing the order of the double integral leads to

$$\sqrt{n}\int_0^t \left[\int_0^u W_k(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right] J_k(u)\left[\frac{d\bar{M}^*_k(u)}{\bar{Y}_k(u)} - \frac{d\bar{M}^*_\bullet(u)}{\bar{Y}_\bullet(u)}\right]$$

$$+\sqrt{n}\int_t^\tau \left[\int_0^t W_k(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right] J_k(u)\left[\frac{d\bar{M}^*_k(u)}{\bar{Y}_k(u)} - \frac{d\bar{M}^*_\bullet(u)}{\bar{Y}_\bullet(u)}\right].$$

Combining the above results and let $W_k(t) = W(t).\bar{Y}_k(t)$ where $W(t)$ is a locally bounded, non negative weight process. $W(t)$ depends on the process $(\bar{N}^L_\bullet(t), \bar{Y}_\bullet(t))$ and it is assumed $W(t)$ is zero when $\bar{Y}_\bullet(t)$ is zero, then

$$Z_k(t) = \sqrt{n}\int_t^0 \left[W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)} - \int_0^u W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right] J_k(u)\left[d\bar{M}_k^*(u) - \bar{Y}_k(u)\frac{d\bar{M}_\bullet^*(u)}{\bar{Y}_\bullet(u)}\right]$$

$$-\sqrt{n}\int_\tau^t \left[\int_0^t W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right] J_k(u)\left[d\bar{M}_k^*(u) - \bar{Y}_k(u)\frac{d\bar{M}_\bullet^*(u)}{\bar{Y}_\bullet(u)}\right].$$

Let

$$C = W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)}, \qquad D = W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right),$$

using Kronecker delta we can rewrite the equation as

$$Z_k(t) = \sum_{p=1}^K \sqrt{n}\int_t^0 \left[C - \int_0^u D\right]\left[\delta_{kp} - \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\right] J_k(u)d\bar{M}_p^*(u)$$

$$-\sum_{p=1}^K \sqrt{n}\int_\tau^t \left[\int_0^t D\right]\left[\delta_{kp} - \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\right] J_k(u)d\bar{M}_p^*(u). \tag{A.1}$$

Based on the martingale central limit theorem, $\sqrt{n}\int_0^t W_k(s)d[\widehat{A}_k(s) - \widehat{A}_\bullet(s)]$ converges in distribution to a mean zero Gaussian martingale with covariance

$$\int_t^0 \left[C - \int_0^u D\right]^2 \frac{y_k(u)}{y_\bullet(u)}\left(\delta_{km} - \frac{y_m(u)}{y_\bullet(u)}\right)\alpha^*(u)y_\bullet(u)du$$

$$+\int_\tau^t \left[\int_0^t D\right]^2 \frac{y_k(u)}{y_\bullet(u)}\left(\delta_{km} - \frac{y_m(u)}{y_\bullet(u)}\right)\alpha^*(u)y_\bullet(u)du. \tag{A.2}$$

Under the null hypothesis, $Z_k(t)$ has mean zero and the covariance between $Z_k(t)$ and $Z_m(t)$ can be estimated by

$$\widehat{\sigma}_{km}^2(t) = \int_0^t \left[W(u)\frac{G_\bullet(u)}{1-G_\bullet(u)} - \int_0^u W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(\delta_{km} - \frac{\bar{Y}_m(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u)$$

$$+\int_t^\tau \left[\int_0^t W(s)d\left(\frac{G_\bullet(s)}{1-G_\bullet(s)}\right)\right]^2 \frac{\bar{Y}_k(u)}{\bar{Y}_\bullet(u)}\left(\delta_{km}-\frac{\bar{Y}_m(u)}{\bar{Y}_\bullet(u)}\right)d\bar{N}_\bullet(u). \tag{A.3}$$

Let $\widehat{\mathbf{\Sigma}}(t)$ denote the $k\times k$ matrix for which the (k,m)th element is given by $\widehat{\sigma}^2_{km}(t)$ and let $\mathbf{Z}(t)=(Z_1(t)\ Z_2(t)\ ...\ Z_K(t))^T$. The test statistic for $H_0:\alpha_1(t)=\alpha_2(t)=...=\alpha_K(t)$ has the following form

$$\mathbf{Z}(t)\widehat{\mathbf{\Sigma}}^{-1}(t)\mathbf{Z}^T(t)\sim\chi^2_{K-1}, \tag{A.4}$$

where $\widehat{\mathbf{\Sigma}}^{-1}(t)$ is a generalized inverse. This test statistic followsa Chi-square distribution with $K-1$ degrees of freedom, χ^2_{K-1}.

A3

Asymptotic consistency of a kernel estimator has been routinely established under the condition that $n\to\infty$, the bandwidth $b\to 0$ and $nb\to\infty$ [53], [54]. In this study, we try to exploratively investigate the limiting distribution of $\widehat{\alpha}(t)$, and we do not give a proof of asymptotic consistency. In the following context, "$\approx$" indicates asymptotic equivalence. Please note that $(nb)^{1/2}[\widehat{\alpha}(t)-\alpha^+(t)]$ can be expressed as

$$(nb)^{1/2}[\widehat{\alpha}(t)-\alpha^+(t)]=\frac{(nb)^{1/2}}{b}\int_0^\tau K\left(\frac{t-u}{b}\right)\left[\frac{-\widehat{G}(u)}{1-\widehat{G}(u)}d(\widehat{A}^*-A^{*+})(u)\right]$$

$$-\frac{(nb)^{1/2}}{b}\int_0^\tau K\left(\frac{t-u}{b}\right)\left(\frac{\widehat{G}(u)}{1-\widehat{G}(u-)}-\frac{G(u)}{1-G(u-)}\right)dA^{*+}(u)$$

For the first term on the right hand side of the above equation, it can be shown that

$$\begin{aligned}
&\frac{(nb)^{1/2}}{b}\int_0^\tau K\left(\frac{t-u}{b}\right)\frac{-\widehat{G}(u)}{1-\widehat{G}(u-)}d(\widehat{A}^*-A^{*+})(u)\\
\approx\ &\sqrt{\frac{n}{b}}\int_\tau^0 K\left(\frac{t-u}{b}\right)\frac{G(u)}{1-G(u-)}d(\widehat{A}^*-A^{*+})(u)\\
=\ &\sqrt{\frac{n}{b}}\int_\tau^0 K\left(\frac{t-u}{b}\right)\frac{G(u)}{1-G(u-)}J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)}
\end{aligned}$$

To investigate the second term on the right hand side, we first consider the Taylor series expansion,

$$\frac{\widehat{G}(u)}{1-\widehat{G}(u^-)} - \frac{G(u)}{1-G(u-)} \simeq \frac{d}{dA^*(u)}\left(\frac{G(u)}{1-G(u-)}\right)(\widehat{A}^* - A^{*+})(u).$$

Then we will have

$$\begin{aligned}
& \frac{(nb)^{1/2}}{b}\int_0^\tau K\left(\frac{t-u}{b}\right)\left[-\left(\frac{\widehat{G}(u)}{1-\widehat{G}(u-)} - \frac{G(u)}{1-G(u-)}\right)dA^{*+}(u)\right] \\
\approx\ & \sqrt{\frac{n}{b}}\int_0^\tau K\left(\frac{t-u}{b}\right)\left[-d\left(\frac{G(u)}{1-G(u-)}\right)(\widehat{A}^* - A^{*+})(u)\right] \\
=\ & \sqrt{\frac{n}{b}}\int_0^\tau K\left(\frac{t-u}{b}\right)\left[-d\left(\frac{G(u)}{1-G(u-)}\right)\int_\infty^u J(x)\frac{d\bar{M}^*(x)}{\bar{Y}(x)}\right] \\
=\ & \sqrt{\frac{n}{b}}\int_0^\tau K\left(\frac{t-u}{b}\right)\left[d\left(\frac{G(u)}{1-G(u-)}\right)\int_0^\infty I(x\geq u)J(x)\frac{d\bar{M}^*(x)}{\bar{Y}(x)}\right] \\
=\ & \sqrt{\frac{n}{b}}\int_0^\tau \left[\int_0^u K\left(\frac{t-y}{b}\right)d\left(\frac{G(y)}{1-G(y-)}\right)\right]J(x)\frac{d\bar{M}^*(x)}{\bar{Y}(x)} \\
=\ & \sqrt{\frac{n}{b}}\int_\tau^0 \left[-\int_0^u K\left(\frac{t-y}{b}\right)d\left(\frac{G(y)}{1-G(y-)}\right)\right]J(x)\frac{d\bar{M}^*(x)}{\bar{Y}(x)}
\end{aligned}$$

Combining the above results, we get $(nb)^{1/2}[\widehat{\alpha}_n(t) - \alpha_n^+(t)]$ to be equal to

$$\sqrt{\frac{1}{nb}}\int_\tau^0 \left[K\left(\frac{t-u}{b}\right)\frac{G(u)}{1-G(u-)} - \int_0^u K\left(\frac{t-y}{b}\right)d\left(\frac{G(y)}{1-G(y-)}\right)\right]J(u)\frac{d\bar{M}^*(u)}{\bar{Y}(u)/n}$$

Through the martingale central limit theorem, when $n\to\infty, b\to 0, nb\to\infty$, $(nb)^{1/2}[\widehat{\alpha}(t) - \alpha^+(t)]$ converges in distribution to a normal random variable with mean zero and the following variance function,

$$\frac{1}{b}\int_\tau^0 \left[K\left(\frac{t-u}{b}\right)\frac{G(u)}{1-G(u-)} - \int_0^u K\left(\frac{t-x}{b}\right)d\left(\frac{G(x)}{1-G(x-)}\right)\right]^2 \frac{\alpha^*(u)du}{y(u)}.$$

In addition, it needs to be confirmed that $(nb)^{1/2}[\alpha_n^+(t) - \alpha(t)]$ is asymptotically negligible. Some regularity conditions for establish such a result can be found in Ramlau-Hansen [53](§4). We do not further investigate this problem in this study.

Appendix B

THE AIDS DATA SET

Table B.1 AIDS Blood Transfusion Data Set

Age	L	T	Age	L	T	Age	L	T	Age	L	T	Age	L	T
1	11	38	3	23	29	29	68	72	39	17	57	50	39	62
1	10	57	3	21	48	29	61	99	39	13	44	50	25	48
1	10	54	4	43	52	29	12	26	39	5	28	50	10	26
1	10	17	4	37	71	29	4	35	41	31	45	51	49	55
1	10	13	4	37	53	30	69	87	41	23	27	51	48	67
1	8	31	4	27	79	30	46	81	41	22	24	51	44	47
1	8	26	4	27	40	32	41	68	42	48	74	51	34	69
1	8	22	4	11	19	32	32	43	42	26	41	51	33	45
1	8	16	5	51	53	32	10	36	42	10	28	51	31	37
1	4	35	6	68	87	33	79	80	44	29	31	52	53	83
1	4	11	11	41	45	33	53	85	44	24	49	52	29	43
2	23	63	17	70	83	33	33	44	44	24	30	52	24	52
2	20	37	20	34	79	34	37	65	45	14	35	52	17	27
2	20	35	21	60	90	34	29	68	46	50	61	53	65	73
2	18	33	21	36	44	34	16	33	46	43	49	53	54	55
2	17	26	22	47	66	35	39	42	46	38	74	53	36	43
2	15	22	23	35	55	35	18	38	46	36	52	53	29	69
2	14	64	23	18	45	36	15	21	46	34	77	53	21	32
2	13	52	24	29	50	36	12	16	46	17	60	54	80	85
2	13	40	25	30	65	36	4	5	46	12	49	54	55	73
2	13	34	26	48	67	37	53	63	46	4	20	54	51	54
2	12	49	26	32	68	37	46	49	47	43	48	54	29	69
2	6	38	26	30	46	38	89	90	48	63	76	54	23	40
3	33	54	27	51	54	38	22	40	49	64	65	54	13	29
3	32	38	28	58	62	38	16	17	49	40	42	55	39	42
3	32	33	28	36	41	38	10	19	49	17	18	55	12	19
55	11	18	60	49	54	63	20	28	66	17	46	68	6	27
56	48	50	60	32	57	63	15	54	66	13	26	69	67	73
56	38	76	60	20	34	63	13	34	67	64	73	69	62	63
56	38	66	60	18	20	63	12	32	67	63	66	69	58	83

NOTE: Study starting date is January 1, 1978 and closing date is July 1, 1986 (102 months). L is the incubation period in months, T is the duration between infection date and closing date and Age is the age in years at the time of transfusion.

Table B.2 AIDS Blood Transfusion Data Set Cont'd.

Age	L	T	Age	L	T	Age	L	T	Age	L	T	Age	L	T
56	32	44	61	57	61	63	0	36	67	63	65	69	38	52
56	20	49	61	48	83	64	56	57	67	42	56	69	38	39
56	15	64	61	26	59	64	52	60	67	41	69	69	31	45
57	63	66	61	25	62	64	48	63	67	29	43	69	28	56
57	37	67	61	19	27	64	40	56	67	21	59	69	13	42
57	28	75	61	18	45	64	23	44	67	20	37	70	62	80
57	22	53	61	14	33	64	18	45	67	20	36	70	41	44
57	9	15	61	11	24	65	62	75	67	18	36	70	27	30
58	62	90	61	10	18	65	59	61	67	17	23	70	24	25
58	53	61	62	63	76	65	47	51	67	10	34	70	21	54
58	29	37	62	43	59	65	36	43	68	54	64	70	19	46
58	25	48	62	42	61	65	35	56	68	46	61	70	19	42
58	19	40	62	37	70	65	34	44	68	38	42	70	14	26
59	67	80	62	35	53	65	32	36	68	32	37	71	53	57
59	63	65	62	33	60	65	29	35	68	27	47	71	49	69
59	55	59	62	29	67	65	25	33	68	27	31	71	33	34
59	38	53	62	29	57	65	23	27	68	24	60	71	32	34
59	16	35	62	24	38	65	18	66	68	22	26	71	31	32
59	11	27	62	21	22	66	83	94	68	20	41	71	26	46
59	11	17	62	16	32	66	33	41	68	19	47	71	14	23
60	68	73	63	61	66	66	32	46	68	15	27	71	12	31
60	59	75	63	37	39	66	23	37	68	11	35	72	52	61
72	40	41	72	29	37	72	29	35	72	16	35	73	72	73
73	42	60	73	40	42	73	34	50	73	30	46	73	15	17
73	8	37	74	41	42	74	19	34	76	37	59	76	24	30
77	49	57	77	20	30	77	19	25	78	76	85	78	38	57
78	34	45	78	29	39	78	20	49	80	55	65	80	27	40
81	19	29	81	10	27	82	37	60	84	25	20	85	38	39

NOTE: Study starting date is January 1, 1978 and closing date is July 1, 1986 (102 months). L is the incubation period in months, T is the duration between infection date and closing date and Age is the age in years at the time of transfusion.