ZHAO, GUOLIN, M.A. Nonparametric and Parametric Survival Analysis of Censored Data with Possible Violation of Method Assumptions. (2008)
Directed by Dr. Kirsten Doehler. 55pp.

Estimating survival functions has interested statisticians for numerous years. A survival function gives information on the probability of a time-to-event of interest. Research in the area of survival analysis has increased greatly over the last several decades because of its large usage in areas related to biostatistics and the pharmaceutical industry. Among the methods which estimate the survival function, several are widely used and available in popular statistical software programs. One purpose of this research is to compare the efficiency between competing estimators of the survival function.

Results are given for simulations which use nonparametric and parametric estimation methods on censored data. The simulated data sets have right-, left-, or interval-censored time points. Comparisons are done on various types of data to see which survival function estimation methods are more suitable. We consider scenarios where distributional assumptions or censoring type assumptions are violated. Another goal of this research is to examine the effects of these incorrect assumptions.

NONPARAMETRIC AND PARAMETRIC SURVIVAL ANALYSIS OF CENSORED DATA WITH POSSIBLE VIOLATION OF METHOD ASSUMPTIONS

by

Guolin Zhao

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
2008

Approved by

_____

Committee Chair

This thesis is dedicated to

My parents and grandfather,

without whose support and inspiration

I would never have had the courage to follow my dreams.

I love you and I miss you.

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair_____

Committee Members_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

First of all, I would especially like to thank Dr. Kirsten Doehler for her guidance and encouragement throughout the preparation of this thesis. I greatly appreciate the time and effort that Kirsten dedicated to helping me complete my Master of Arts degree, as well as her support and advice over the past two years.

I would also like to thank Dr. Sat Gupta and Dr. Scott Richter. It is Dr. Gupta who helped me to quickly adjust to the new study environment in this strange country, and it is also him who provided me a lot of suggestions and advice. Thanks also to Dr. Gupta and Dr. Richter for their recommendations.

Last but not the least, I would like to thank the thesis committee members for their time and efforts in reviewing this work.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER I
INTRODUCTION

Survival analysis is widely applied in many fields such as biology, medicine, public health, and epidemiology. A typical analysis of survival data involves the modeling of time-to-event data, such as the time until death. The time to the event of interest is called either survival time (section 1.1) or failure time. In this thesis, we use words "failure" and "event of interest" interchangeably. The survival function (section 1.2) is a basic quantity employed to describe the probability that an individual "survives" beyond a specified time. In other words, this is the amount of time until the event of interest occurs.

In survival analysis, a data set can be exact or censored, and it may also be truncated. Exact data, also known as uncensored data, occurs when the precise time until the event of interest is known. Censored data arises when a subject's time until the event of interest is known only to occur in a certain period of time. For example, if an individual drops out of a clinical trial before the event of interest has occurred, then that individual's time-to-event is right censored at the time point at which the individual left the trial. The time until an event of interest is truncated if the event time occurs within a period of time that is outside of the observed time period. For example, if we consider time until death, and we look only at individuals who are in a nursing home for people who are 65 or older, then all individuals who die before age 65 will not be observed. This is an example of left-truncated data, because no

deaths will be observed before the age of 65. In this thesis, only exact and censored data (section 1.3) are considered.

The objective of this thesis is to examine the efficiency of several methods that are commonly used to estimate survival functions in the presence of different types of censored data. We compare different techniques used to estimate survival functions under various scenarios. In some scenarios we considered, we purposely used techniques when assumptions were not met in order to find out how this affected the results.

1.1 Survival Time

Definition: Survival time is a variable which measures the time from a particular starting point (e.g. the time at which a treatment is initiated) to a certain endpoint of interest (e.g. the time until development of a tumor).

In most situations, survival data are collected over a finite period of time due to practical reasons. The observed time-to-event data are always non-negative and may contain either censored (section 1.3) or truncated observations [6].

1.2 Survival Function

Definition: The survival function models the probability of an individual surviving beyond a specified time x. We denote X as the random variable representing survival time, which is the time until the event of interest. In other words, the probability of experiencing the event of interest beyond time x is modeled by the survival function [11]. The statistical expression of the survival function is shown in Equation (1.1):

$$S(x) = Pr(X > x). \tag{1.1}$$

Since X is a continuous random variable, the survival function can be presented as it is in Equation (1.2), where $S(x)$ is the integral of the probability density function (PDF), $f(x)$:

$$S(x) = Pr(X > x) = \int_x^\infty f(t)\, dt. \tag{1.2}$$

Therefore, by taking the negative of the derivative of Equation (1.2) with respect to x, we have

$$f(x) = -\frac{dS(x)}{dx}. \tag{1.3}$$

The quantity $f(x)dx$ might be considered an "approximate" probability that the event will occur at time x. Since the derivative of the survival function with respect to x is negative, then the function $f(x)$ represented in Equation (1.3) will be non-negative [11]. The survival curve, $S(x)$, can be plotted to graphically represent the probability of an individual's survival at varying time points. All survival curves have the following properties:

1. $S(x)$ is monotone;
2. $S(x)$ is non-increasing;
3. When time $x = 0$, $S(x) = 1$;
4. $S(x) \to 0$ as $x \to \infty$.

Although survival curves can take a variety of shapes, depending on the underlying distribution of the data, they all follow the four basic properties previously mentioned. Figure 1.1 shows three survival functions resulting from distributions that

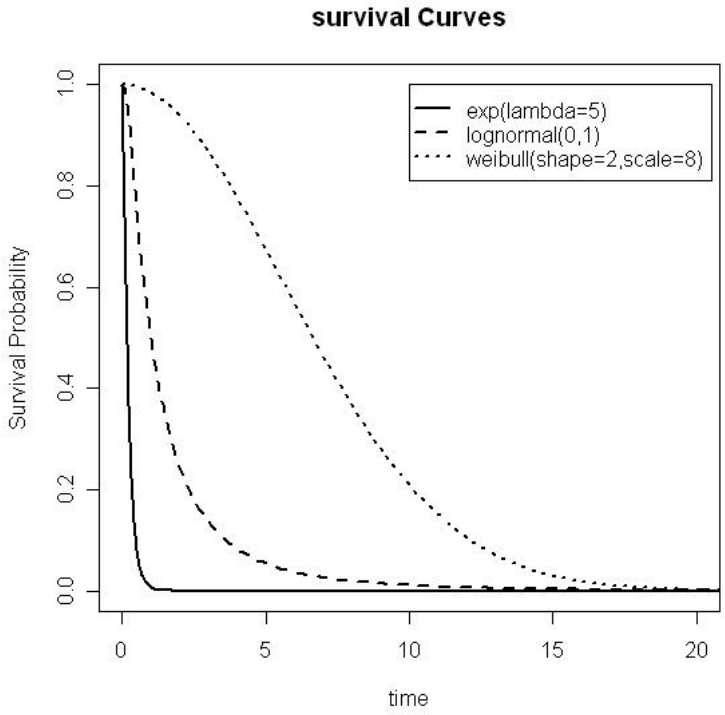

Figure 1.1: exponential, lognormal and Weibull survival functions.

are commonly used in survival analysis [1; 3]. These distributions are the *exponential* with shape parameter $\lambda = 5$, the standard *lognormal*, and the *Weibull* distribution with shape and scale parameters of 2 and 8, respectively. In Figure 1.1, we can see that as time increases, the *exponential* survival probability decreases fastest among the three distributions, and the probability of survival associated with the *Weibull* distribution decreases much more slowly compared to the other two distributions.

1.3 Censored Data

In the beginning of this chapter, it was mentioned that survival data may consist of censored or truncated observations. In this section, we will focus on discussing

censored data. Censored data arises when the exact time points at which failures occur are unknown. However, we do have knowledge that each failure time occurs within a certain period of time. We will now introduce right-censored and interval-censored data, which are both common types of data encountered in real life scenarios.

1.3.1 Right-Censored Data

Right-censored data occurs when an individual has a failure time after their final observed time. For example, we may want to know how long individuals will survive after a kidney transplant. If we set up a 10-year follow-up study, it is possible that an individual moves away before the end of the study. In this case we are not able to obtain information regarding the time to death. However, we do know the time that he or she moved away, and this is defined as the time point at which the true survival time is right-censored. It is also possible that at the end of a study period, individuals are still alive. In this case, the time when the study ends can be considered a right-censored data point for all of the individuals in the study who are still alive. The reason that the data points in both of the previous examples are considered to be right-censored data is because the exact survival times for those subjects are unknown, but we do know that each individual's time of death will occur after some specified time point.

A typical right-censored data set includes a variable for an individual's time on study and an indicator of whether the associated time is an exactly known or a right-censored survival time. Usually we use an indicator variable of 1 if the exact survival time is known, and an indicator of 0 for right-censored times.

Consider a simple data set that has 5 individuals enrolled in a 10-year follow-up study. The raw data are listed as follows:

$$8^+,\ 10^+,\ 6,\ 9^+,\ 5.$$

In this data set, there are three numbers having a "+" in the superscript, which is commonly used as an indication that these are right-censored data points. When doing a typical analysis of this data using statistical software, we will set the indicator to be 1 if there is no "+" in the superscript and 0 if there is a "+" present.

This data set can also be written as (t_i, δ_i) values:

$$(8, 0),\ (10, 0),\ (6, 1),\ (9, 0),\ (5, 1).$$

In this format, t_i is the variable representing the time associated with the i^{th} individual and δ_i is an indicator of whether the survival time for individual i is exact or right-censored.

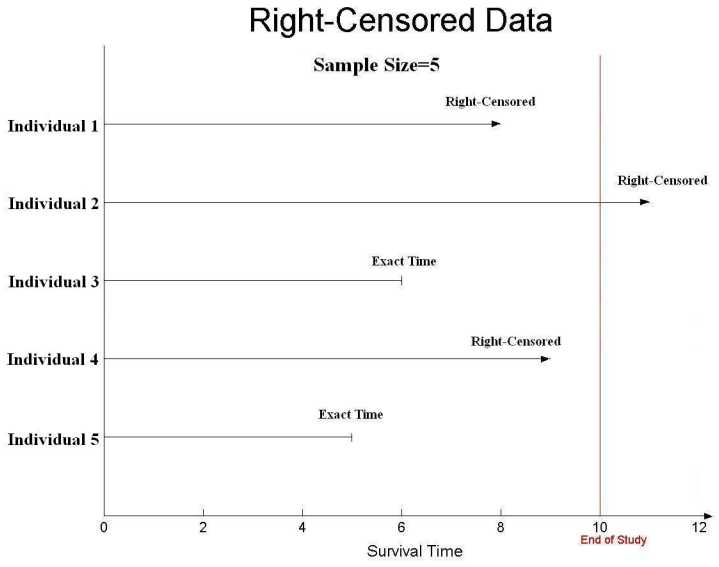

Figure 1.2: Data set with exact and right-censored survival times.

Figure 1.2 shows a graphical representation of the survival information which is known about each of the five individuals in our example. The survival information

is represented by a horizontal line for each individual. An arrow at the end of an individual's line indicates a right-censored survival time. Therefore, the failure time for individuals 3 and 5 is exactly known, but the failure times associated with individuals 1, 2, and 4 are right-censored. Individuals 1 and 4 were lost to follow-up at a certain time before the end of this study. Individual 2 is still alive at 10 years, which is the end of the study period. Therefore this individual has a survival time which is right-censored at 10 years.

1.3.2 Interval-Censored Data

Interval-censoring occurs when the event of interest is known only to occur within a given period of time. Both left-censored and right-censored data are special cases of interval-censored data, where the lower endpoint is 0 and the upper endpoint is ∞, respectively.

Consider the example data set mentioned earlier, in which patients received a kidney transplant (section 1.3.1). The object of this type of study is to observe how long people will survive after a kidney transplant. Suppose that as part of the study, individuals are requested to make a clinic visit once a year. An individual may die sometime after his or her last visit and before the next time that they are supposed to come in for another visit. Other things could also occur which cause the individual to be lost to follow-up between two visit times. For example, an individual may move away and drop out of a study. Another possibility is that an individual may die in an event such as a car accident, which is unrelated to the event of interest. In this case the time of death is considered a right-censored time point, as the individual did not die from kidney failure.

Figure 1.3 shows us another example with five individuals in the sample. The

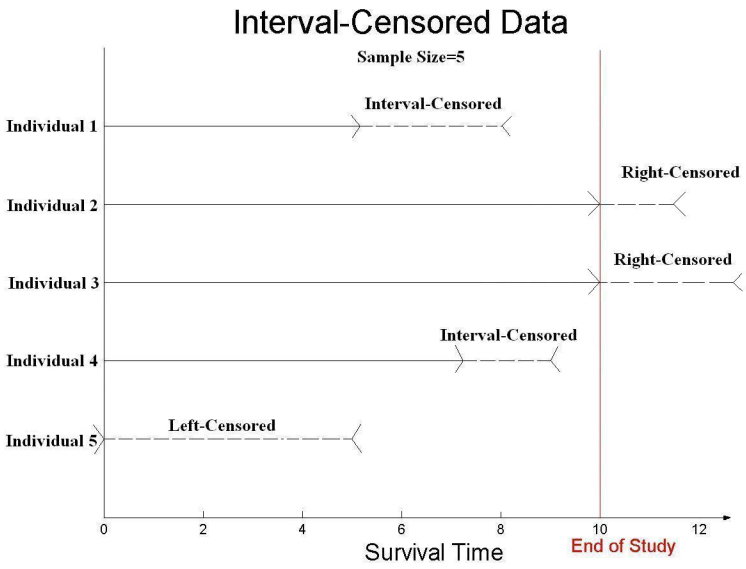

Figure 1.3: Data set with interval-censored and right-censored survival times.

unknown survival times for each of the five individuals are interval-censored between a lower and upper endpoint. In this data set, individuals 2 and 3 have a survival time which is right-censored, since they are both still alive at the end of our study. As mentioned earlier, left-censored data are also a special kind of interval-censored data. Individual 5 in Figure 1.3 has a survival time which is left-censored.

CHAPTER II
METHODS

Both parametric and nonparametric estimation of the survival function will be introduced in this chapter. For right-censored data and interval-censored data, a parametric estimator (section 2.1) is sometimes used for estimation of the survival function. However, in real life scenarios the exact distribution of the data is usually unknown. In such cases, using a nonparametric method is a common alternative, since the nonparametric estimator does not assume that the data come from a specified distribution. The Kaplan-Meier estimator (section 2.2) is widely employed as the nonparametric estimator in the presence of right-censored data. Similarly, nonparametric estimation of the survival function with interval-censored data is available with the Turnbull estimator (section 2.3). Several other survival function estimators (section 2.4) for right-censored and interval-censored data will also be mentioned.

2.1 Parametric Estimator

Although nonparametric estimation is more widely used, it is still necessary to discuss parametric estimation in which the distribution of the survival data is assumed known. Distributions that are commonly used in survival analysis are the *exponential*, *Weibull*, and *lognormal* [1; 3].

Because of its historical significance, mathematical simplicity and important properties, the *exponential* distribution is one of the most popular parametric models.

It has a survival function, $S(x) = \exp[-\lambda x]$, for $x \geqslant 0$. Although the *exponential* distribution is popular in some situations, it is restrictive in many real applications due to its functional features [11].

The *Weibull* distribution has a survival function, $S(x) = \exp[-\lambda x^{\alpha}]$, for $x \geq 0$. Here $\lambda > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter. The *exponential* distribution is a special case of the *Weibull* distribution when $\alpha = 1$.

Another distribution that is frequently used to model survival times is the *lognormal*. If X has a *lognormal* distribution, then $\ln X$ has a *normal* distribution. For time-to-event data, this distribution has been popularized not just because of its relationship to the *normal* distribution, but also because some authors have observed that the *lognormal* distribution can approximate survival times or age at the onset of certain diseases [5; 9]. Like the *normal* distribution, the *lognormal* distribution is completely specified by two parameters, μ and σ [11].

2.2 Kaplan-Meier Estimator

The standard nonparametric estimator of the survival function is the Kaplan-Meier (K-M) estimator, also known as the product-limit estimator. This estimator is defined as:

$$\hat{S}(x) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leqslant t}[1 - \frac{d_i}{Y_i}] & \text{if } t_1 < t, \end{cases} \tag{2.1}$$

where t_1 denotes the first observed failure time, d_i represents the number of failures at time t, and Y_i indicates the number of individuals who have not experienced the event of interest, and have also not been censored, by time t.

From the function given in Equation (2.1), we notice that before the first failure happens, the survival probability is always 1. As failures occur, the K-M

estimator of the survival function decreases. A step function with jumps at the observed event times will be obtained by using Kaplan-Meier method to estimate the survival function. The jumps on the survival curve depend not only on the number of events observed at each event time, but also on the pattern of the censored observations before the event time.

Consider again the 10-year follow-up study, where we are interested in knowing how long people will survive after a kidney transplant. Suppose there are a total of 50 patients in the 10-year study. Also suppose that six of them died at 0.5 years, and two are lost to follow up during the half year after transplant. Therefore, at 0.5 years after the transplant, there are 42 patients still in this study. Similarly, we have some deaths at 1 year after transplant and so on, until the end of the study period, at which time there are 22 patients still alive and enrolled in the study.

Table 2.1: Construction of the Kaplan-Meier estimator.

Time t_i	Number of events d_i	Number at risk Y_i	K-M Estimator $\hat{S}(t) = \prod_{t_i \leqslant t}[1 - \frac{d_i}{Y_i}]$
0.5	6	42	$[1 - \frac{6}{42}] = 0.857$
1	5	35	$[0.857](1 - \frac{5}{35}) = 0.735$
2	3	32	$[0.735](1 - \frac{3}{32}) = 0.666$
3.5	2	30	$[0.666](1 - \frac{2}{30}) = 0.622$
5	1	28	$[0.622](1 - \frac{1}{28}) = 0.600$
6.5	1	27	$[0.600](1 - \frac{1}{27}) = 0.578$
8.5	2	25	$[0.578](1 - \frac{2}{25}) = 0.532$
9.5	2	22	$[0.532](1 - \frac{2}{22}) = 0.484$

Data from this hypothetical study are given in Table 2.1, along with K-M estimates of the survival function at the various death times. Table 2.2 shows the K-M estimates for all times, and the corresponding graph of the K-M function is given in Figure 2.1.

Table 2.2: Kaplan-Meier survival estimates.

Time on study (t)	K-M Estimator $\hat{S}(t)$
$0 \leqslant t < 0.5$	1.000
$0.5 \leqslant t < 1$	0.857
$1 \leqslant t < 2$	0.735
$2 \leqslant t < 3.5$	0.666
$3.5 \leqslant t < 5$	0.622
$5 \leqslant t < 6.5$	0.600
$6.5 \leqslant t < 8.5$	0.578
$8.5 \leqslant t < 9.5$	0.532
$9.5 \leqslant t < 10$	0.484

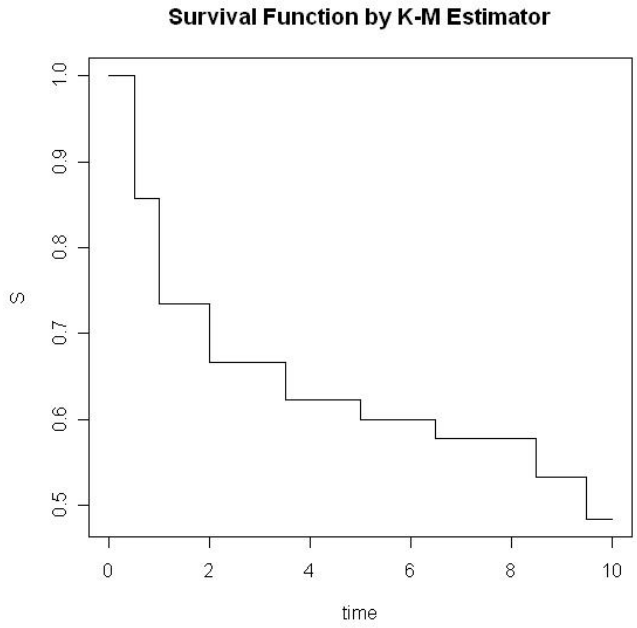

Figure 2.1: Kaplan-Meier survival function for right-censored data.

The K-M estimator is a common nonparametric estimator. It is efficient and easy to use, and it is available in many statistical software programs such as SAS, S-Plus and R. The Greenwood formula [8] to estimate the variance for the K-M

estimator is represented as:

$$\widehat{var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

The standard error of a K-M estimate, $\sqrt{\widehat{var}[\hat{S}(t)]}$, can be obtained directly from the LIFETEST procedure in SAS.

2.3 Turnbull Estimator

An estimator of the survival function is available for interval-censored data. Richard Peto developed a Newton-Raphson method to estimate the nonparametric maximum likelihood estimator (NPMLE) for interval-censored data in 1973 [17]. Then in 1976 Richard Turnbull formulated an EM algorithm which also estimated the NPMLE for interval-censored data, but this estimator could accommodate truncated data too [18]. We have implemented Turnbull's algorithm in R software to calculate the NPMLE for interval-censored data, and we use the terms "Turnbull Estimator" and "NPMLE" interchangeably. The NPMLE for interval-censored data is based on n independent, arbitrarily interval-censored and/or truncated observations $(x_1, x_2, ..., x_n)$. Here, as in earlier sections, we only discuss the situation where there is no truncation. We assume that x_i is interval-censored, so that x_i is only known to lie in an interval $[L_i, R_i]$.

The derivation of the Turnbull estimator will be introduced here with a simple scenario. Suppose we have 5 failures which occur in a study. The survival times for the 5 patients in this hypothetical study are interval-censored. The following data set shows the $n = 5$ interval censored data points.

$$[L_1, R_1] = [1, 2],\ [L_2, R_2] = [2, 5],\ [L_3, R_3] = [4, 7],\ [L_4, R_4] = [3, 8],\ [L_5, R_5] = [7, 9]$$

Before the NPMLE can be estimated using Turnbull's algorithm, equivalence classes must be defined to determine at what time points the survival function takes jumps. Equivalence classes are defined by each "L" that is immediately followed by "R" once the endpoints are ordered. To find the equivalence classes, we need to consider all of the $[L_i, R_i]$ intervals, for $i = 1, 2, \ldots, n$, and then order the $2n$ endpoints from smallest to largest.

Table 2.3: Construction of equivalence classes for an interval-censored data set.

Initial Endpoints	1	2	2	5	4	7	3	8	7	9
Corresponding Labels	L_1	R_1	L_2	R_2	L_3	R_3	L_4	R_4	L_5	R_5
Ordered Endpoints	1	2	2	3	4	5	7	7	8	9
Corresponding Labels	L_1	L_2	R_1	L_4	L_3	R_2	L_5	R_3	R_4	R_5
Labels	L	L	R	L	L	R	L	R	R	R

Table 2.3 shows how we obtain the equivalence classes for the hypothetical data set. The first two lines in the table are the initial data. The next two lines represent the ordered endpoints and their corresponding labels. The fifth line denotes only whether the corresponding point is a left (L) or right (R) endpoint. Therefore, we have $m = 3$ equivalence classes which are $[2, 2], [4, 5], [7, 7]$. Only within these equivalence classes can the survival function have jumps. The Turnbull estimator of the CDF is given in Equation (2.2):

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < q_1 \\ \hat{s}_1 + \hat{s}_2 + \cdots + \hat{s}_j & \text{if } p_i < x < q_{j+1} \quad (1 \leqslant j \leqslant m-1) \\ 1 & \text{if } x > p_m, \end{cases} \tag{2.2}$$

where q_1 is the lower bound of the first equivalence class and p_m is the upper bound of the last equivalence class. The interval $[q_j, p_j]$ represents the j^{th} equivalence class. Therefore, $\hat{F}$ is undefined for $x \in [q_j, p_j]$, for $1 \leqslant j \leqslant m$, which means that $\widehat{F}$ is

defined only in between the equivalence classes.

Before maximizing the likelihood of F, the CDF, it is necessary to be familiar with the alpha matrix, α. This is an $n \times m$ matrix of indicator variables. As stated earlier, each $[L_i, R_i]$ interval represents the censoring interval which contains the failure time for individual i and $[q_j, p_j]$ represents the j^{th} equivalence class. The $(ij)^{th}$ element of the alpha matrix is defined in Equation (2.3):

$$\alpha_{ij} = \begin{cases} 1 & \text{if } [q_j, p_j] \subseteq [L_i, R_i] \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The maximum likelihood estimator for the CDF is represented as:

$$L(F) = \prod_{i=1}^{n} [F(R_{ij}+) - F(L_{ij}-)]. \tag{2.4}$$

Once the CDF of the survival distribution is obtained, integration techniques can be used to calculate the PDF of the survival times, and the survival function is simply $S(x) = 1 - F(x)$. As an alternative to maximizing the likelihood in Equation (2.4), we can also maximize the equivalent likelihood given in Equation (2.5):

$$L(s_1, s_2, ..., s_m) = \prod_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} s_j}{\sum_{j=1}^{m} s_j} \right), \tag{2.5}$$

where α_{ij} is the $(ij)^{th}$ element of the α matrix defined earlier, and s_j represents the jump amount within the j^{th} equivalence class.

The process of maximizing Equation (2.5) is the procedure used in Turnbull's method of finding the NPMLE. This technique involves finding the expected value of I_{ij}, a matrix that has the same dimensions as α_{ij}, and is defined as:

$$I_{ij} = \begin{cases} 1 & \text{if } x_i \subseteq [L_i, R_i] \\ 0 & \text{otherwise.} \end{cases}$$

The expected value of I_{ij} is first calculated under an initial s matrix, which is an $m \times 1$ matrix with elements equal to $\frac{1}{m}$. $E_s[I_{ij}]$ is the expected value of I_{ij} as a function of s, and is also denoted as $\mu_{ij}(s)$. When we treat μ_{ij} as an observed quantity, we can estimate the proportion of observations in the j^{th} equivalence class, π_{ij}, as the following:

$$\pi_j(s) = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(s).$$

The Turnbull method is an example of an Expectation-Maximization (EM) algorithm since the final s matrix of jump amounts is found by alternating between the calculation of $E_s[I_{ij}]$ and obtaining an s matrix which maximizes π_{ij}. The algorithm stops when the difference between successive values for s is less than a given tolerance [15].

Using the Turnbull algorithm on the data set in our example, we obtain the NPMLE of the survival function given in Figure 2.2. Notice that jumps in the survival function occur only in the 3 equivalence classes, represented by the intervals of $[2, 2], [4, 5], [7, 7]$.

2.4 Other Survival Function Estimators

Besides the parametric, K-M, and Turnbull estimators mentioned already, there are several other survival function estimators which are also available.

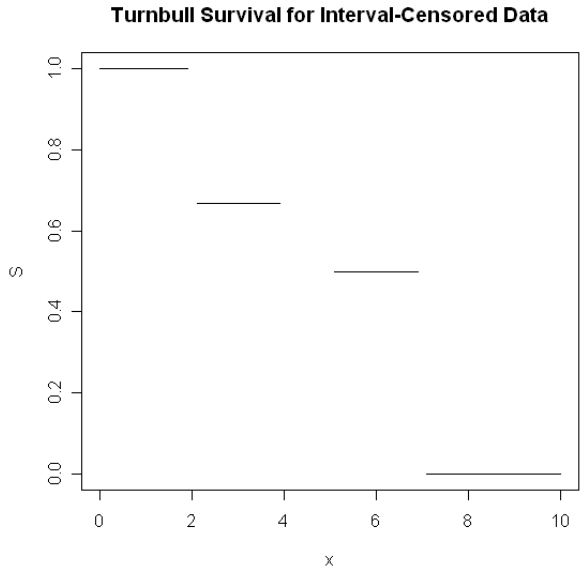

Figure 2.2: Turnbull survival function for interval-censored data.

A simple technique that is used in the case of interval-censored data is to wrongfully assume that the data is actually right-censored, and then apply the K-M method by assuming that the left, right, or middle points of the intervals are the known failure times. These methods may give biased results and misleading inferences [15]. The major benefit of using the K-M estimator on a data set that is interval-censored is that the K-M method is easily available if interest lies only in obtaining a "rough" estimate of the survival function.

To obtain a smooth estimate of the survival function with exact or censored data, the logspline method of Kooperberg and Stone can be applied [13]. This method involves fitting smooth functions to the log density within subsets of the time axis. This logspline smoothing technique uses the Newton-Raphson method to implement maximum likelihood estimation [14].

Keiding [10] and Anderson et al. [2] have studied kernel smoothing, which uses a kernel function to smooth increments of the Kaplan-Meier estimator (or NPMLE for interval-censored data) of the survival function. Pan has done comparisons of the kernel smoothing method and the logspline method [16]. He notes that in many cases the kernel smoothing method cannot recover all of the information that is lost when the discrete NPMLE of the survival function is smoothed with kernel smoothing techniques. Pan also says that it may be more efficient to directly model a smooth survival function or its density, rather than to use kernel smoothing.

Komárek, Lesaffre, and Hilton have developed a semiparametric procedure involving splines to smooth the baseline density in an accelerated failure time (AFT) model [12]. The result of this technique is an expression for the density in terms of a fixed and relatively large number of normal densities. Although Komárek et al. have focused on the estimation of the density itself, or on estimation of the regression parameters in an AFT model, their methods can also be applied to the survival function.

Doehler and Davidian use a SemiNonParametric (SNP) density to directly estimate the density of survival times under mild smoothness assumptions [4]. The SNP density is represented by a squared polynomial multiplied by a base density, which can be any density which has a moment generating function. The SNP method results in efficiency gains over standard nonparametric and semiparametric methods. Another advantage of the SNP technique of Doehler and Davidian is the ability to accommodate data with arbitrary censoring and/or truncation. Also, common survival densities such as the lognormal and exponential are special cases of the SNP density.

CHAPTER III
SIMULATIONS: RIGHT-CENSORED DATA

In the following two chapters, simulation procedures and results are discussed. In this chapter, simulations involving right-censored data are introduced. The following chapter considers interval-censored data.

As stated earlier, one of the goals of this thesis is to examine various estimates of the survival function, sometimes in the case of violated method assumptions. Therefore, in this chapter involving right-censored data, we have focused on estimating the survival function using both nonparametric and parametric estimates, where sometimes the parametric distributional assumptions are incorrect. In each simulation discussed, we have compared biases, variances, Mean Squared Errors (MSE), and Relative Efficiency (RE) values. RE is used to measure the efficiency of one estimator to another. We follow reference [4] and consider the times associated with equally spaced survival probabilities. Specifically, for each simulation, we have considered times when the true survival probabilities are approximately 0.9, 0.75, 0.5, 0.35 and 0.2.

3.1 Design of Simulations

Before comparing different estimation methods, we first discuss how we generated the data, and how we obtained bias, variance, MSE, and RE values.

3.1.1 Generating the Data

The simulations we considered are based on three commonly used distributions: exponential, lognormal, and Weibull. Specifically we generated data from an exponential distribution with shape parameter $\lambda = 5$, a standard lognormal, and a Weibull distribution with shape and scale parameters of 2 and 8, respectively. For each simulation, we generated $N = 100$ data sets, each with $n = 100$ individuals. Each individual has approximately a 30% chance of having a right-censored survival time. We implemented SAS code and procedures, including PROC LIFETEST and PROC LIFEREG.

The steps used to generate a right-censored data set are shown below. For this example we assume that the true survival time has an exponential distribution.

♠ STEP 1: Generate a variable t for true survival time from an exponential distribution with $\lambda = 5$.

♠ STEP 2: Generate another variable for censoring time from a uniform distribution on the interval $(0, 0.6303)$ to obtain approximately 30% censoring for each data set. This censoring variable is denoted as c.

♠ STEP 3: Compare the true survival time t with the censoring time c. We denote the time and indicator variables as follows.

$$\text{time}[i] = \begin{cases} t[i] & \text{if } t[i] \leq c[i], \\ c[i] & \text{if } t[i] > c[i] \end{cases}$$

$$\text{indicator}[i] = \begin{cases} 0 & \text{if time}[i] = c[i], \\ 1 & \text{if time}[i] = t[i] \end{cases}$$

We used the variables *time* and *indicator* to run the simulations. Similarly, we generated data sets when the true survival time had a standard lognormal or Weibull($\alpha = 2$, $\lambda = 8$) distribution. Table 3.1 shows the uniform distributions generated for each data set, which resulted in approximately 30% of the survival times being right-censored.

Table 3.1: Uniform censoring distributions used to achieve approximately 30% censoring in right-censored data sets.

Distribution of Data	Censoring Distribution
Exponential($\lambda = 5$)	$U(0, 0.6303)$
Standard Lognormal	$U(0, 4.784)$
Weibull($\alpha = 2$, $\lambda = 8$)	$U(0, 23.41)$

3.1.2 Nonparametric Estimation

With the generated right-censored data, we first used the K-M estimator to nonparametrically estimate the survival function. There were 5 survival time points of interest for each distribution considered. We estimated the K-M survival probability at each of these 5 times using the LIFETEST procedure in SAS. This was done for each of the $N = 100$ data sets in each simulation. Therefore, each simulation had 100 estimated survival probabilities at each time point. Once these estimated survival probabilities were obtained, we calculated the mean and variance of the estimates at each time point of interest.

3.1.3 Parametric Estimation

We also applied parametric estimation to estimate the survival function for the same exponential, lognormal and Weibull data sets we had generated. For each distribution

we looked at the estimated survival probabilities at the same time points of interest mentioned in section 3.1.2. The SAS LIFEREG procedure was used to obtain the parametric survival estimates.

3.1.4 Generating Table Output

Using nonparametric and parametric estimation methods, we obtained tables of values similar in form to Table 3.2 for each of 3 distributions that we considered.

Table 3.2: Output table obtained from each estimation method used in a simulation.

Time t	True $S(t)$	Estimated Survival Probabilities	Avg. of Survival Estimates	Variance of Survival Estimates
t_1	0.90	$\hat{S}_1(t_1), \hat{S}_2(t_1), ..., \hat{S}_{100}(t_1)$	$\bar{S}(t_1)$	$Var\{S_1(t_1), \ldots, S_{100}(t_1)\}$
t_2	0.75	$\hat{S}_1(t_2), \hat{S}_2(t_2), ..., \hat{S}_{100}(t_2)$	$\bar{S}(t_2)$	$Var\{S_1(t_2), \ldots, S_{100}(t_2)\}$
t_3	0.50	$\hat{S}_1(t_3), \hat{S}_2(t_3), ..., \hat{S}_{100}(t_3)$	$\bar{S}(t_3)$	$Var\{S_1(t_3), \ldots, S_{100}(t_3)\}$
t_4	0.35	$\hat{S}_1(t_4), \hat{S}_2(t_4), ..., \hat{S}_{100}(t_4)$	$\bar{S}(t_4)$	$Var\{S_1(t_4), \ldots, S_{100}(t_4)\}$
t_5	0.20	$\hat{S}_1(t_5), \hat{S}_2(t_5), ..., \hat{S}_{100}(t_5)$	$\bar{S}(t_5)$	$Var\{S_1(t_5), \ldots, S_{100}(t_5)\}$

After each simulation was completed, we calculated average bias as the difference between the true survival probability at time t and the average of the $N = 100$ estimated survival probabilities at time t. This is denoted as $\text{Bias}\{\hat{S}(t)\} = \bar{S}(t) - S(t)$.

We used Equation (3.1) to calculate the Mean Square Error Estimates.

$$MSE = [Bias\{\hat{S}(t)\}]^2 + Var\{\hat{S}(t)\} = \{\bar{S}(t) - S(t)\}^2 + Var\{\hat{S}(t)\} \tag{3.1}$$

In this equation, $S(t)$ is the true survival probability at time t, $\bar{S}(t)$ is the average of the $N = 100$ estimates of survival at time t, and $Var\{\hat{S}(t)\}$ is the variance of the $N = 100$ estimates at time t.

For right-censored data simulations, we compared the K-M and parametric estimators by examining Relatively Efficiency (RE) values defined in Equation (3.2).

$$RE = \frac{MSE(\text{K–M Estimation})}{MSE(\text{Parametric Estimation})}. \tag{3.2}$$

Tables 3.3, 3.5, and 3.7 show the results for each distribution. Moreover, we give three tables showing confidence interval coverages. For each distributional assumption, we estimated 100 nonparametric and parametric survival probabilities for each of the 5 time points of interest, along with the corresponding standard errors. For nonparametric K-M estimation we obtained standard errors using the Greenwood method. However, to obtain standard errors of the parametric survival estimates, we needed to use the Delta method which uses a gradient vector and the variance-covariance matrix of the model parameters. The Delta Method formula is given in Equation (3.3).

$$SE(\hat{S}) = \left(\frac{\partial S}{\partial \theta}\right) \times \hat{\Sigma} \times \left(\frac{\partial S}{\partial \theta}\right)' \tag{3.3}$$

$SE(\hat{S})$ represents the vector of standard errors, θ is the vector of parameters, and $\hat{\Sigma}$ is the variance-covariance matrix obtained from the parametric estimation method. The first term in the equation for $SE(\hat{S})$ is a row vector, and the third term is its transpose. Once we had obtained standard errors, we were able to construct 95% Wald Confidence Intervals as

$$(\hat{S}_i(t_j) - 1.96 \times SE\{\hat{S}_i(t_j)\},\ \hat{S}_i(t_j) + 1.96 \times SE\{\hat{S}_i(t_j)\})$$
$$\text{for}\ \ \mathrm{i} = 1, \ldots, 100;\ \ \mathrm{j} = 1, \ldots, 5.$$

We show the percent of the confidence intervals that include the true survival probability. This proportion is called the Monte Carlo coverage. Standard errors for the Monte Carlo coverage probabilities are given in Tables 3.4, 3.6, and 3.8.

3.2 Exponential Data

Table 3.3: Simulation results based on 100 Monte Carlo data sets, 30% right-censored exponential($\lambda = 5$) data, $n = 100$. RE is defined in Equation (3.2) as MSE(K-M)/MSE(Parametric).

Time t	True $S(t)$	K-M Bias×100	Par Bias×100	K-M Var×100	Par Var×100	RE
0.021	0.90	0.078	-0.167	0.1071	0.0134	7.820
0.058	0.75	-0.515	-0.350	0.1950	0.0688	2.824
0.139	0.50	-0.751	-0.458	0.3075	0.1743	1.776
0.210	0.35	-0.569	-0.390	0.3290	0.1939	1.700
0.322	0.20	0.061	-0.212	0.2480	0.1481	1.670

Table 3.3 shows results under an exponential distribution. All of the RE values are greater than 1, which indicates that the parametric estimator is more efficient than the K-M estimation. Both the K-M and parametric estimators seem to have biases that are closest to zero when the true survival probabilities are 0.9 and 0.2. At the first and last time point of interest the K-M bias values are lower than the parametric bias values. Although neither estimator has the lower absolute bias value at all time points, the parametric estimator always gives smaller variance values than the K-M. From this point on in this thesis, unless it is stated otherwise, we will assume that the statements made on bias refer to the absolute value of the bias.

Table 3.4 gives the 95% confidence interval coverages with the exponential distribution. Both the K-M and parametric estimation methods provided acceptable coverage. At all 5 times of interest, at least 90% of the K-M confidence intervals and at least 95% of the parametric intervals contained the true survival probabilities.

Table 3.4: Simulation results based on 100 Monte Carlo data sets, 30% right-censored exponential($\lambda = 5$) data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries ≈ 0.022.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M	0.90	0.97	0.94	0.93	0.95
Parametric	0.96	0.97	0.95	0.95	0.95

Table 3.5: Simulation results based on 100 Monte Carlo data sets, 30% right–censored standard lognormal data, $n = 100$. RE is defined in Equation (3.2) as MSE(K-M)/MSE(Parametric).

Time t	True $S(t)$	K-M Bias×100	Par Bias×100	K-M Var×100	Par Var×100	RE
0.278	0.90	-0.362	-0.098	0.0867	0.0561	1.57
0.509	0.75	-0.612	-0.331	0.1888	0.1310	1.46
1.000	0.50	-0.405	-0.785	0.2634	0.1980	1.30
1.470	0.35	-0.918	-0.906	0.2761	0.2010	1.36
2.320	0.20	-0.579	-0.775	0.2415	0.1544	1.53

3.3 Lognormal Data

Table 3.5 shows the simulation results under the lognormal distribution. We again observe that all RE values are greater than 1, which shows that parametric estimation is more efficient than nonparametric estimation. However, all RE values are less than 1.6, which shows that the two estimators have more similar MSE values than they did in the exponential data case, where all RE values were greater than 1.6. Again the average estimates from the nonparametric estimation sometimes have a smaller bias, but they always have a larger variance.

Table 3.6 gives the 95% confidence interval coverages for the lognormal distribution. K-M estimation provided good coverage, which was sometimes higher than the parametric estimation. At all 5 times of interest, around 95% of the confidence

intervals contained the true survival probabilities. Although the parametric method shown in Table 3.5 is more efficient compared to K-M estimation, it has 95% coverages which are similar to the K-M method.

Table 3.6: Simulation results based on 100 Monte Carlo data sets, 30% right-censored standard lognormal data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries ≈ 0.022.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M	0.94	0.94	0.97	0.95	0.92
Parametric	0.93	0.93	0.96	0.94	0.92

3.4 Weibull Data

Table 3.7: Simulation results based on 100 Monte Carlo data sets, 30% right-censored Weibull($\alpha = 2$, $\lambda = 8$) data, $n = 100$. RE is defined in Equation (3.2) as MSE(K-M)/MSE(Parametric).

Time t	True $S(t)$	K-M Bias×100	Par Bias×100	K-M Var×100	Par Var×100	RE
2.597	0.90	0.157	0.074	0.1166	0.0647	1.804
4.291	0.75	-0.004	0.161	0.2941	0.1748	1.680
6.660	0.50	-0.570	0.030	0.3154	0.2406	1.325
8.197	0.35	-0.253	-0.135	0.3176	0.2170	1.466
10.149	0.20	0.278	-0.247	0.2650	0.1544	1.715

Table 3.7 shows simulation results with Weibull distribution data. We again conclude that parametric estimation is more efficient. The average estimates from the nonparametric estimation method have smaller biases than the parametric estimator at some time points. Also, at time 4.291, the K-M estimator has a bias almost equal to zero, while at time 6.66, the parametric estimator has a bias that is very close to zero. As usual, the parametric estimator gives less variable estimates than the K-M

at all 5 time points.

Table 3.8 gives the 95% confidence interval coverages for the Weibull distribution. Both the K-M and the parametric methods show acceptable coverage levels, although since neither estimator had any coverages above 0.95 at any of the time points, the overall results err on the side of lower coverage than 0.95.

Table 3.8: Simulation results based on 100 Monte Carlo data sets, 30% right-censored Weibull($\alpha = 2$, $\lambda = 8$) data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries ≈ 0.022.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M	0.92	0.89	0.94	0.95	0.91
Parametric	0.91	0.92	0.92	0.92	0.95

As mentioned earlier, the benefits of using a parametric estimator can not always be taken advantage of, as there is often uncertainty about the true underlying distribution. We have carried out numerous simulations where we have used parametric estimation to estimate the survival function while assuming an incorrect distribution. We report here on a subset of these simulations which use Weibull data.

Table 3.9 shows simulation results when we incorrectly assumed that the underlying distribution was exponential when it was actually Weibull. In the table, the biases from the parametric estimator are much larger than those resulting from the K-M estimation. The parametric estimator greatly overestimates the true survival probability at early time points when the true survival is near 0.9 and 0.75. As time approaches 6.66, when the true survival is 0.5, the parametric estimates seem to get better, but as the true survival probability approaches 0.35 and 0.2, the parametric estimates begin to underestimate the true survival probability.

It is interesting to note that the parametric estimation still shows less variabil-

ity compared to the K-M at all time points. Although the RE value at time 8.197 is larger than one, the rest are less than one and sometimes very close to zero, indicating that the K-M estimator is often much more efficient than the parametric estimator that assumes an incorrect distribution. The large RE value at time 8.197 indicates that the MSE for the parametric estimator is less than the MSE for the K-M estimator. This is due to the very small variance of the parametric estimator compared to the K-M estimator at this time point. Table 3.9 demonstrates that incorrect knowledge of the underlying distribution can have drastic negative effects when parametric estimation of the survival function is applied.

Table 3.9: Simulation results based on 100 Monte Carlo data sets, 30% right-censored Weibull($\alpha = 2$, $\lambda = 8$) data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the parametric estimation method uses the incorrect assumption that the data is exponentially distributed.

Time t	True $S(t)$	K-M Bias×100	Par Bias×100	K-M Var×100	Par Var×100	RE
2.597	0.90	0.157	-17.226	0.1166	0.0341	0.039
4.291	0.75	-0.004	-15.834	0.2941	0.0613	0.115
6.660	0.50	-0.570	-5.687	0.3154	0.0824	0.785
8.197	0.35	-0.253	1.749	0.3176	0.0856	2.739
10.149	0.20	0.278	8.980	0.2650	0.0813	0.299

We also considered estimation of the survival function with this same Weibull data set when we incorrectly assumed that the true distribution was lognormal. The corresponding simulation results are given in Table 3.10. Except for the first time point of interest, when the true survival probability is 0.9, the biases for the non-parametric estimator are always much smaller than those of the parametric estimator which assumes an incorrect distribution. At the three earliest time points considered, the RE is smaller than one, indicating that the K-M is the more efficient estimator. However, at the last two time points, the MSE of the K-M estimator is larger than the

corresponding MSE of the parametric estimator, because the parametric estimator has much smaller variance.

Table 3.10: Simulation results based on 100 Monte Carlo data sets, 30% right-censored Weibull($\alpha = 2, \lambda = 8$) data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the parametric estimation method uses the incorrect assumption that the data is lognormally distributed.

Time t	True $S(t)$	K-M Bias×100	Par Bias×100	K-M Var×100	Par Var×100	RE
2.597	0.90	0.157	-0.111	.1166	.1345	0.868
4.291	0.75	-0.004	-4.552	.2941	.2587	0.631
6.660	0.50	-0.570	-4.881	.3154	.1788	0.764
8.197	0.35	-0.253	-1.857	.3176	.1310	1.924
10.149	0.20	0.278	2.424	.2650	.0992	1.682

The results shown in Table 3.10 demonstrate that incorrect knowledge of the underlying distribution may still result in higher efficiency of the incorrect parametric estimator at some time points. However, if one is concerned more with bias than with MSE, then an incorrect parametric estimator in this case would be a poor choice because it has much larger biases than the K-M estimator.

Figure 3.1 shows a graphical representation of how the correct and incorrect parametric distributional assumptions fit the true survival function. The parametric estimator can provide accurate estimates under correct distributional assumptions, as it can be seen from the graph that the average of these estimates is almost identical to the true survival function. However, when the distribution is incorrectly assumed, the estimated survival probabilities will have large biases. In our example, when we incorrectly assumed weibull survival times were lognormally distributed, the biases which resulted from the parametric estimation were smaller than when we incorrectly assumed that the Weibull survival times were exponentially distributed.

As you can see from Figure 3.1, the true survival function and the survival

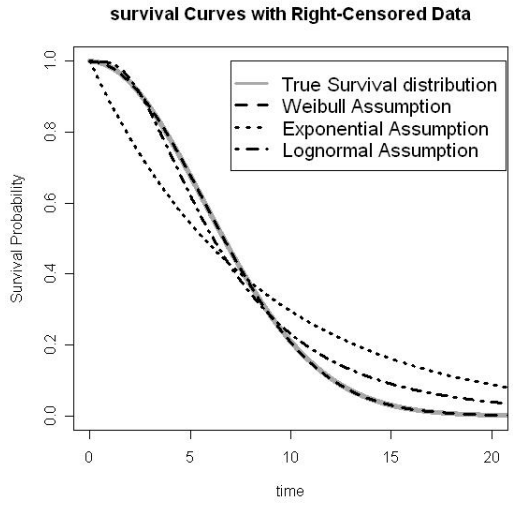

Figure 3.1: Multiple estimates of a Weibull survival function.

function that is estimated from the incorrect exponential distributional assumption have different shaped survival curves. However, both of these functions cross at approximately time 8. Therefore, we would expect that the bias values for the incorrect parametric estimation are smaller in absolute value at times near time 8. Looking back at Table 3.9, we see that the bias at time 8.197 has the smallest absolute value of the 5 parametric bias values.

The true survival function and the survival function that is estimated from the incorrect lognormal distributional assumption are also slightly different, and these two functions cross at approximately time 9. Again we would expect that the bias values for the incorrect parametric estimation are smaller in absolute value at times near time 9. Looking at Table 3.10, we see that the bias values for the incorrect parametric estimation are smallest in absolute value at times 2.597, 8.197, and 10.149. Times 8.197 and 10.149 are close to the time of 9, where the true survival and the lognormal survival functions cross. Also, looking at Figure 3.1, we see that the true and average

lognormal survival functions are very similar at time 3 and earlier.

In this thesis, we have only reported on simulations involving incorrect distributional assumptions when the true survival distribution is Weibull. However, we have also looked at other scenarios where the true distribution was exponential or lognormal. Of all the simulations we considered where we purposely used an incorrect distribution to estimate the survival function using parametric methods, the only case where negative effects were not seen was when we estimated exponential data while assuming the correct distribution was Weibull. However, this was not surprising, since the exponential distribution is a special case of the Weibull distribution when the shape parameter is 1.

CHAPTER IV

SIMULATIONS: INTERVAL-CENSORED DATA

Similar to the right-censored data simulations, we also considered nonparametric and parametric estimators with data that was interval-censored. To correctly estimate the survival function for interval-censored data using a nonparametric technique, we applied the Turnbull estimator, which was calculated through a program we wrote in R. We used the LIFEREG Procedure in SAS to obtain the parametric survival function estimates. Besides these two estimators, we also obtained an adhoc nonparametric estimate of the survival function using the K-M estimator through the LIFETEST procedure in SAS. This was obviously a violation of method assumptions, since we incorrectly assumed that interval-censored data was right-censored. For each estimation method considered, we obtained estimated survival probabilities at each of the 5 times of interest for each of the $N = 100$ data sets.

In our interval-censored simulations we considered several sample sizes ranging from 30 to 200, and found that results were often similar. We chose to follow Reference [7] and report on simulations using a sample size of $n = 100$ for each data set. Besides different sample sizes, we experimented with different percentages of interval and right-censoring as well. Also, in the case of parametric estimation, we again considered both incorrect and correct distributional assumptions. In this chapter we report on a subset of the simulations involving interval-censored data that we have done.

4.1 Design of Simulations

Generating the interval-censored data involved more steps compared to right-censored data. The details of how to generate the data are discussed here.

4.1.1 Generating the Data

We considered the same three distributions for the interval-censored simulations that we used for the right-censored simulations. Each data set contained 100 interval-censored observations, of which approximately 25% were right-censored. Some observations were left-censored, but here we do not specify left-censored from interval-censored data.

The steps used to generate the interval-censored data with 25% right-censoring are shown below. As we did in Chapter 3, we assume that the true survival time follows an exponential distribution to demonstrate the data generation process.

- ♠ STEP 1: Generate a variable t for true survival time from an exponential distribution with $\lambda = 5$.

- ♠ STEP 2: Assume there are 5 clinic visits for each individual. The time interval between two visits has a uniform distribution on the interval $(0, 0.115371)$. We used this particular uniform distribution to obtain approximately 25% right-censored data.

- ♠ STEP 3: Generate a 100×2 empty matrix named "bounds" for each data set. Compare the true survival time with the 5 visit times to find the visit times which bound the survival time. We set a large number for the upper bound (e.g. 1000) when the true survival time is larger than the last visit time (when the

true survival is right-censored). The formulas for the lower and upper bounds are given here.

$$\text{bounds}[i,1] = \begin{cases} 0 & \text{if } t < V[i,1], \\ V[i,j] & \text{if } V[i,j] < t < V[i,j+1] \text{ where j=1,2,3,4,} \\ V[i,5] & \text{if } t > V[i,5], \end{cases}$$

$$\text{bounds}[i,2] = \begin{cases} V[i,1] & \text{if } t < V[i,1], \\ V[i,j+1] & \text{if } V[i,j] < t < V[i,j+1] \text{ where j=1,2,3,4,} \\ 1000 & \text{if } t > V[i,5]. \end{cases}$$

♠ STEP 4: Generate a 100×1 empty matrix named indicator. Based on the bounds, let

$$\text{indicator}[i] = \begin{cases} 0 & \text{if bounds}[i,2] = 1000, \\ 1 & \text{otherwise.} \end{cases}$$

In the case of the exponential distribution, a $U(0,\ 0.115)$ distribution was used to generate the first visit time, c_1. Then a $U(c_1,\ c_1 + 0.115)$ distribution was used to generate the second visit time. The last three visit times were generated in a similar manner. The specific uniform distributions used to generate the visit times in each simulation are listed in Table 4.1. These uniform distributions resulted in interval-censored data sets with approximately 25% right-censoring. We used similar steps to those given here to generate interval censored data sets where the true survival time was standard lognormal and Weibull($\alpha = 2,\ \lambda = 8$).

Table 4.1: Uniform censoring distributions used to achieve approximately 25% right–censoring in interval-censored data sets.

Distribution of Data	Censoring Distribution
Exponential($\lambda = 5$)	$U(0, 0.115371)$
Standard Lognormal	$U(0, 0.836)$
Weibull($\alpha = 2$, $\lambda = 8$)	$U(0, 3.9625)$

4.1.2 Estimation Methods Applied for Interval-Censored Data

For each interval-censored data set that we generated, we applied the nonparametric Turnbull estimation method and the parametric estimation technique. We used the R program that we wrote to carry out the Turnbull estimation, which involved finding the corresponding equivalence classes and α matrix for each data set generated. The parametric estimation was again done using the SAS LIFEREG procedure, which can accommodate both right- and interval-censored data.

We also applied the adhoc method of incorrectly assuming that the data was right-censored so that we could use the K-M estimator by assuming that either the lower bound, upper bound, or the midpoint of the interval was the exactly known survival time when the indicator variable was equal to 1. If the indicator variable was 0, we correctly assumed that the lower endpoint of the interval was the time point at which the survival time was right-censored. We clearly used the K-M estimator under false assumptions, as this nonparametric estimator is meant to be used for right-censored data. The reason that somebody might use this improvised technique with interval-censored data is because the K-M method is widely employed and easily accessible in common statistical software programs. As discussed in Chapter 3, inferential methods can be erroneous when assumptions are violated. However, we wanted to see what effect these violations had on the overall estimates when we wrongfully

assumed that the survival times were known.

For all three survival estimation techniques considered in the interval-censored data simulations, the identical 5 survival time points from the right-censored data simulations were of interest for each distribution considered. As in the right-censored data case, we calculated the average of estimated survival probabilities and the variance of the survival probabilities at each time point of interest. We also constructed tables similar to Table 3.2 for each simulation involving interval-censored data. The construction of relative efficiency values is discussed in the following section.

4.1.3 Generating Table Output

To compare K-M and parametric estimation techniques, we calculated a Relative Efficiency (RE) value which is defined in Equation (3.2). To compare Turnbull Estimation and Parametric Estimation, we defined RE as

$$RE = \frac{MSE(\text{Turnbull Estimation})}{MSE(\text{Parametric Estimation})},$$

which is similar to Equation (3.2), as the Mean Squared Error of the parametric estimator is still in the denominator.

Moreover, when comparing the Turnbull estimator and the K-M estimator with the midpoint of the interval as the "true" failure, we used the following definition for Relative Efficiency (RE).

$$RE = \frac{MSE(\text{Turnbull Estimation})}{MSE(\text{K–M : MID Estimation})}.$$

Similar RE equations were used when we compared the Turnbull estimator to the

K-M estimator when the left or right endpoint of the interval was assumed to be the "true" failure time.

Tables similar to tables 3.3, 3.5, 3.7, 3.9, and 3.10 were constructed to show the results of the interval-censored data simulations. As with the right-censored data case, we also generated tables which show 95% coverage probabilities in the case of interval-censoring. However, the coverages based on the appropriate nonparametric estimation, or Turnbull Estimation, are not provided here. This is because the nonparametric maximum likelihood estimation for interval-censored data does not follow standard asymptotic theory [15], and besides using the computer-intensive bootstrap method, there is no alternative option to obtain standard errors. We do show coverages for when the incorrect K-M estimation technique is used with interval-censored data. However, these coverages result from using the Greenwood formula for standard errors, which is appropriate only for right-censored data, and not for interval-censored data.

4.2 Exponential Data

The following three tables are the results when we compared the K-M estimator with parametric estimation using exponential data.

Table 4.2 shows results when we compared parametric estimation assuming the correct exponential distribution to K-M estimation under the assumption that the lower bound of the interval is the exactly known survival time. We refer to this latter method as the K-M:LEFT estimation method. All RE values are greater than 1, indicating that the parametric estimator is more efficient than the K-M estimator. Based on the average estimates from both procedures, the parametric estimator always has smaller variances and biases. As we did in Chapter 3, we will again assume

Table 4.2: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the left interval endpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	K-M:LEFT Bias×100	Par Bias×100	K-M:LEFT Var×100	Par Var×100	RE
0.021	0.90	-19.640	-0.098	0.2563	0.0150	268.453
0.058	0.75	-13.191	-0.187	0.2528	0.0769	25.469
0.139	0.50	-8.813	-0.177	0.2821	0.1946	5.376
0.210	0.35	-5.148	-0.074	0.2579	0.2164	2.398
0.322	0.20	0.457	0.089	0.2388	0.1651	1.455

that any statements made about bias refer to the absolute value of the bias, unless stated otherwise. The biases obtained from K-M estimation are much larger when we assume that the left endpoint of the interval is the exact failure time. These biases are most extreme at early time points, and they tend to decrease as time increases. Also, as time increases, the RE values decrease. It is interesting to note that at all 5 time points considered, either both estimates have negative bias, or both estimates have positive bias.

Table 4.3: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	K-M:MID Bias×100	Par Bias×100	K-M:MID Var×100	Par Var×100	RE
0.021	0.90	5.940	-0.098	0.0373	0.0150	25.463
0.058	0.75	-4.152	-0.187	0.2401	0.0769	5.272
0.139	0.50	0.148	-0.177	0.2367	0.1946	1.203
0.210	0.35	-0.507	-0.074	0.2777	0.2164	1.285
0.322	0.20	0.303	0.089	0.2857	0.1651	1.732

Table 4.3 shows results when we compared parametric estimation assuming the correct exponential distribution to K-M estimation under the assumption that

the midpoint of the interval is the exactly known survival time. We refer to this latter method as the K-M:MID estimation method. All RE values are greater than 1, which indicates that the parametric estimator still is more efficient than the K-M estimator. However, the RE values at the later time points are all less than 2, and even early RE values are not as large as they were in Table 4.2. This indicates that the K-M:MID estimation method is more efficient than the K-M:LEFT method. Based on the average estimates by the parametric and K-M procedures, sometimes the parametric estimator has a smaller bias and at other time points the K-M estimator has a smaller bias. The K-M bias values get closer to zero as time increases. This pattern is not seen with the parametric biases. As usual, parametric estimation always gives a smaller variance than the nonparametric estimation.

Table 4.4: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the right interval endpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	K-M:RIGHT Bias×100	Par Bias×100	K-M:RIGHT Var×100	Par Var×100	RE
0.021	0.90	9.010	-0.098	0.0094	0.0150	53.592
0.058	0.75	17.290	-0.187	0.0910	0.0769	39.368
0.139	0.50	10.179	-0.177	0.2372	0.1946	6.466
0.210	0.35	6.620	-0.074	0.2600	0.2164	3.202
0.322	0.20	0.747	0.089	0.2388	0.1651	1.477

Table 4.4 shows results when we compared parametric estimation assuming the correct exponential distribution to K-M estimation under the assumption that the right endpoint of the interval is the exactly known survival time. We refer to this latter method as the K-M:RIGHT estimation method. From these results, we can again conclude that the parametric estimation is more efficient in terms of bias and variance. Also, comparing these results to Table 4.2 shows that using the left

endpoint as the exactly known survival time is even less efficient than using the right endpoint at all time points except the first time of interest. This is because the RE values in Table 4.4 are all larger than those given in Table 4.2 for the four later time points.

By comparing the parametric estimator to the incorrect application of K-M:LEFT, K-M:MID, or K-M:RIGHT methods, we see that the parametric estimator almost always has a smaller absolute bias, especially at times when the true survival probability is 0.9, when the biases obtained from all three K-M applications are much larger. It is interesting to note that at time 0.139, the K-M:MID method actually gives a lower absolute value for bias than the parametric method.

Table 4.5 shows results when we compared the Turnbull estimation with parametric estimation, assuming the correct exponential distribution. Through the information given in this table, we can see that the parametric estimator has better RE values, especially when the time is 0.021. Based on the average estimates by both procedures, sometimes the Turnbull estimator has a smaller bias and sometimes the parametric estimator has a smaller bias. Parametric estimation always gives a smaller variance than the Turnbull estimation, especially at the first two time points. The lowest RE value in the table occurs at time 0.210. At this time, the Turnbull estimator has a bias value that is closer to zero than the Turnbull bias values at the other four time points.

It is not surprising that the parametric estimator is more efficient compared to the nonparametric estimators considered. However, it is important to remember that it is often unrealistic to assume that the true underlying distribution of the data is known. To compare the Turnbull estimator with the estimator resulting from the K-M:MID method, we generated Table 4.6, where RE is defined as

Table 4.5: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined in Equation (4.1) as MSE(Turnbull)/MSE(Parametric).

Time t	True $S(t)$	Turnbull Bias×100	Par Bias×100	Turnbull Var×100	Par Var×100	RE
0.021	0.90	0.940	-0.098	0.7637	0.0150	50.416
0.058	0.75	-0.957	-0.187	0.7482	0.0769	9.679
0.139	0.50	-0.726	-0.177	0.4127	0.1946	2.122
0.210	0.35	-0.025	-0.074	0.3914	0.2164	1.795
0.322	0.20	0.867	0.089	0.3944	0.1651	2.428

MSE(Turnbull)/MSE(K-M:MID).

Table 4.6: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(K-M), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
0.021	0.90	0.940	5.940	0.7637	0.0373	1.980
0.058	0.75	-0.957	-4.152	0.7482	0.2401	1.836
0.139	0.50	-0.726	0.148	0.4127	0.2367	1.764
0.210	0.35	-0.025	-0.507	0.3914	0.2777	1.396
0.322	0.20	0.867	0.303	0.3944	0.2857	1.402

Table 4.6 shows us that all RE values are greater than 1, which indicates that the K-M estimator is more efficient than the Turnbull estimator. Based on the average estimates by both procedures, sometimes the Turnbull estimator has a smaller bias and sometimes the K-M estimator has a smaller bias. The lowest RE value in the table again occurs at time 0.210. It is at this time that the Turnbull estimator shows a bias value that is very close to zero. However, the Turnbull estimator always has a larger variance than the K-M estimator, which results in larger MSE values compared to the K-M estimator. This shows that in some cases, using the quick and easy K-M method with interval censored data may be a viable option to the Turnbull method.

However, this is only true if the midpoint of the interval is assumed to be the known survival time.

Table 4.7 gives 95% coverages when exponential data are used. Due to extreme biases, the K-M:LEFT and K-M:RIGHT methods exhibit large undercoverage at early time points, but acceptable coverage at the largest time point of interest, when the true survival probability is 0.2. The K-M:MID estimation method shows acceptable coverages at time points corresponding to true survival probabilities of 0.75 or less. However, low coverage when the true survival probability is 0.9 is not surprising, as bias values are large to begin with. Also, the Greenwood formula, intended for right-censored data, was used to calculate standard errors in this interval-censored data situation. This could have added to the low coverage seen at the first time, but it does not appear to affect the K-M:MID coverages at the later time points. We again applied the delta method to obtain the standard errors needed to calculate coverages for the parametric case. The parametric method still provided adequate coverages, as at least 90% of the confidence intervals contain the true survival probabilities.

Table 4.7: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries $\approx .022$.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M:LEFT	0.00	0.24	0.48	0.73	0.94
K-M:MID	0.21	0.88	0.94	0.96	0.90
K-M:RIGHT	0.00	0.00	0.42	0.79	0.95
Parametric	0.93	0.93	0.93	0.92	0.91

As stated earlier in the Table 4.6 discussion, we found that the RE values comparing the K-M:MID method to the Turnbull method indicated that the latter

is less efficient. However, this was not the case in all of the exponential simulations that we examined. Below we will show one example which demonstrates that Turnbull estimation can be more efficient. Using a different number of visits for each individual (2 visits instead of 5 visits) and different censoring percentages (5% are right-censored instead of 25%), we obtained the results shown in Table 4.8. Most RE values are less than 1, indicating that the Turnbull estimator has a higher level of efficiency compared to the K-M:MID method. Although the Turnbull estimator has larger variances, it is still more efficient due to bias values, which are often much smaller than those of the K-M:MID method. The only time point where the RE value is greater than one is at time 0.021, and this is due to the much smaller variance of the KM:MID method compared to the Turnbull method.

Table 4.8: Simulation results based on 100 Monte Carlo data sets, 5% right-censoring in exponential($\lambda = 5$) interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(K-M), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
0.021	0.90	6.741	9.530	0.8291	0.0051	1.405
0.058	0.75	5.459	21.757	3.4343	0.0362	0.782
0.139	0.50	0.701	34.781	4.1784	0.1504	0.342
0.210	0.35	0.708	34.743	1.4617	0.1953	0.120
0.322	0.20	-1.352	21.178	0.5945	0.2598	0.129

4.3 Lognormal Data

This section shows results from when we compared survival estimation methods using lognormal data.

Table 4.9 shows the comparison of the K-M:MID and parametric methods. It is unusual for a nonparametric method to have a smaller variance than a parametric

method, but this phenomenon is actually observed at the first time point of interest. Similar to previous parametric estimation results, RE values are greater than one. However, it is interesting to note that for the lognormal distribution, the RE values are much closer to one compared to those seen in the exponential case. This shows that for lognormal data, the parametric estimation method is only slightly more efficient than the K-M:MID estimation method. Although not shown here, the very large MSE values that resulted from comparing the K-M:LEFT and K-M:RIGHT techniques to the parametric method indicate that these methods should not be used. Therefore the K-M method should only be considered if the midpoints of the intervals are assumed to be the true failure times.

Table 4.9: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in standard lognormal interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	K-M:MID Bias×100	Par Bias×100	K-M:MID Var×100	Par Var×100	RE
0.278	0.90	1.460	-0.404	0.0849	0.0945	1.11
0.509	0.75	-2.653	-0.199	0.1676	0.1491	1.59
1.000	0.50	0.713	0.069	0.2139	0.1553	1.41
1.470	0.35	0.546	0.218	0.1837	0.1691	1.10
2.320	0.20	0.849	0.400	0.2346	0.1580	1.51

Table 4.10 shows results when we compared Turnbull estimation with parametric estimation under the lognormal distribution. RE values are still greater than 1, as they were in the exponential case. There is usually smaller biases with the parametric method, but at one time point the Turnbull estimator shows a smaller bias. Parametric estimation always gives a smaller variance than Turnbull estimation.

Table 4.11 shows results comparing the Turnbull estimator and the K-M:MID estimator. Sometimes the Turnbull estimator has a smaller bias and sometimes the

Table 4.10: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in standard lognormal interval-censored data, $n = 100$. RE is defined in Equation (4.1) as MSE(Turnbull)/MSE(Parametric).

Time t	True $S(t)$	Turnbull Bias×100	Par Bias×100	Turnbull Var×100	Par Var×100	RE
0.278	0.90	2.120	-0.404	0.3293	0.0945	3.89
0.509	0.75	0.257	-0.199	0.5448	0.1491	3.65
1.000	0.50	-1.631	0.069	0.4660	0.1553	3.17
1.470	0.35	-0.143	0.218	0.3708	0.1691	2.19
2.320	0.20	1.452	0.400	0.3033	0.1580	2.03

K-M estimator has a smaller bias. However, Turnbull estimation always gives a larger variance, which results in larger MSE values than those seen with the K-M estimation method.

Table 4.11: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in standard lognormal interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(K-M), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
0.278	0.90	2.120	1.460	0.3293	0.0849	3.522
0.509	0.75	0.257	-2.653	0.5448	0.1676	2.292
1.000	0.50	-1.631	0.664	0.4660	0.2043	2.252
1.470	0.35	-0.143	0.546	0.3708	0.1837	1.987
2.320	0.20	1.452	0.849	0.3033	0.2346	1.342

Table 4.12 gives the 95% coverages for lognormal distribution. The K-M:MID coverages seen in the lognormal case are higher than those seen in the exponential case. However, the K-M:MID method still shows low coverage at the first time point, when the true survival is 0.9. It is at this time point that the Turnbull estimator has a large bias. The parametric method shows coverages that are similar to what we expected.

As we did with the exponential data, we also wanted to compare the Turnbull

Table 4.12: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in standard lognormal interval-censored data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries $\approx .022$.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M:LEFT	0.00	0.15	0.56	0.86	0.88
K-M:MID	0.85	0.97	0.96	0.99	0.93
K-M:RIGHT	0.00	0.00	0.30	0.66	0.92
Parametric	0.94	0.97	0.98	0.93	0.94

method to the K-M:MID method under the extreme scenario with only 2 visit times and 95% interval-censoring. The results in Table 4.13 indicate that the K-M:MID method is less efficient than the Turnbull method at the later time points. At these later time points, the K-M:MID method is even more biased than it is at earlier time points. The MSE values greater than one at the first two time points are again the result of the K-M:MID method having lower variance than the Turnbull estimator.

Table 4.13: Simulation results based on 100 Monte Carlo data sets, 5% right-censoring in standard lognormal interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
0.278	0.90	-9.046	8.560	4.4475	0.0120	7.071
0.509	0.75	-10.27	21.459	4.8753	0.0292	1.291
1.000	0.50	-0.317	41.126	2.4182	0.0827	0.142
1.470	0.35	9.746	50.044	1.9067	0.1254	0.113
2.320	0.20	14.064	49.973	1.4450	0.1969	0.136

4.4 Weibull Data

The Weibull data simulations showed similar results to those seen in the exponential and lognormal data cases. Table 4.14 shows results from comparing the K-M:MID

estimation method to the parametric method, Table 4.15 gives the results from comparing the Turnbull and parametric methods, and Table 4.16 shows output from the simulation using the Turnbull and K-M:MID methods.

Again parametric estimation proves to be more efficient than both nonparametric estimators. It is interesting to note that in Table 4.14 the bias for the K-M:MID estimator at time 4.291 is almost zero. This could be because some of the interval midpoints were very close to the actual failure time. However, this same thing was observed in the right-censored Weibull case in Chapter 3. Table 4.15 shows RE values greater than 1, although there are two time points where the Turnbull estimates have lower absolute bias than the parametric estimates. From Table 4.16 we can see that the Turnbull estimator is less efficient that the K-M:MID method, although the Turnbull estimator sometimes has a smaller bias.

Table 4.14: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Parametric), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	K-M:MID Bias×100	Par Bias×100	K-M:MID Var×100	Par Var×100	RE
2.597	0.90	-1.992	-0.349	0.1029	0.0694	2.020
4.291	0.75	0.006	-0.431	0.2200	0.1643	1.324
6.660	0.50	-0.770	-0.499	0.2656	0.2176	1.234
8.197	0.35	-0.596	-0.481	0.2614	0.2194	1.195
10.149	0.20	-0.086	-0.308	0.2647	0.1896	1.390

Table 4.17 gives 95% coverage results with the Weibull distribution. Again, the K-M:MID coverages are slightly lower than the parametric coverages.

As with the other two distributions, we also show results from the extreme scenario of 2 visit times and 5% interval-censoring with the Weibull data. Table 4.18 shows that the biases obtained by using Turnbull estimation are always lower than

Table 4.15: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(Parametric).

Time t	True $S(t)$	Turnbull Bias×100	Par Bias×100	Turnbull Var×100	Par Var×100	RE
2.597	0.90	0.264	-0.349	0.2074	0.0694	2.948
4.291	0.75	0.636	-0.431	0.4342	0.1643	2.639
6.660	0.50	-1.451	-0.499	0.3869	0.2176	1.854
8.197	0.35	-0.128	-0.481	0.5381	0.2194	2.428
10.149	0.20	-0.560	-0.308	0.3542	0.1896	1.875

Table 4.16: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined as MSE(K-M)/MSE(Turnbull), where the interval midpoint is assumed to be the exact survival time in the K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
2.597	0.90	0.264	-1.992	0.2074	0.1029	1.460
4.291	0.75	0.636	0.006	0.4342	0.2200	1.992
6.660	0.50	-1.451	-0.770	0.3869	0.2656	1.503
8.197	0.35	-0.128	-0.596	0.5381	0.2614	2.031
10.149	0.20	-0.560	-0.086	0.3542	0.2647	1.349

Table 4.17: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. 95% Wald confidence interval coverage probabilities for Kaplan-Meier (Greenwood standard errors) and parametric (Delta method standard errors) estimation of the survival function. Standard errors of Monte Carlo coverage entries $\approx .022$.

True Survival	0.9	0.75	0.5	0.35	0.2
K-M:LEFT	0.23	0.15	0.34	0.70	0.90
K-M:MID	0.92	0.91	0.95	0.93	0.89
K-M:RIGHT	0.11	0.21	0.28	0.49	0.84
Parametric	0.94	0.91	0.95	0.94	0.90

those of the K-M:MID method, and sometimes less than 10% of the bias shown by the K-M:MID method. RE values at later time points are less than 1, indicating that the Turnbull estimator has a higher level of efficiency compared to the K-M:MID method

at these later times.

Table 4.18: Simulation results based on 100 Monte Carlo data sets, 5% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined in Equation (4.1) as MSE(K-M)/MSE(Turnbull) assuming midpoint of interval as exactly survival time by K-M estimation.

Time t	True $S(t)$	Turnbull Bias×100	K-M:MID Bias×100	Turnbull Var×100	K-M:MID Var×100	RE
2.597	0.90	5.977	7.341	0.8273	0.0335	2.069
4.291	0.75	5.728	14.917	2.8169	0.0953	1.355
6.660	0.50	2.839	21.495	2.3501	0.1752	0.507
8.197	0.35	0.548	21.555	2.1152	0.2089	0.436
10.149	0.20	-1.961	16.424	1.0533	0.2749	0.367

As we did in Chapter 3 with right-censored data, we show the effect of an incorrect distributional assumption on parametric estimation of the survival function in the presence of interval-censored Weibull data. Table 4.19 shows simulation results when we incorrectly assumed that the exponential was the correct distribution. The RE values are close to zero at most time points, indicating that the parametric estimator may not result in efficiency gains if the distribution is misspecified.

Table 4.19: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(Parametric), where the parametric estimation method uses the incorrect assumption that the data is exponentially distributed.

Time t	True $S(t)$	Turnbull Bias×100	Par Bias×100	Turnbull Var×100	Par Var×100	RE
2.597	0.90	0.264	-17.060	0.2074	0.0498	0.070
4.291	0.75	0.636	-15.601	0.4342	0.0897	0.174
6.660	0.50	-1.451	-5.402	0.3869	0.1211	0.988
8.197	0.35	-0.128	2.051	0.5381	0.1261	3.202
10.149	0.20	-0.560	9.287	0.3542	0.1202	0.364

Table 4.20 shows results from a simulation in which we incorrectly assumed that the underlying Weibull distribution was lognormal. Surprisingly the RE values

are all greater than one. This is the result of the Turnbull estimator having larger variability than the parametric estimator. However, there are much larger biases with the parametric estimator at each time of interest.

Table 4.20: Simulation results based on 100 Monte Carlo data sets, 25% right-censoring in Weibull($\alpha = 2$, $\lambda = 8$) interval-censored data, $n = 100$. RE is defined as MSE(Turnbull)/MSE(Parametric), where the parametric estimation method uses the incorrect assumption that the data is lognormally distributed.

Time t	True $S(t)$	Turnbull Bias×100	Par Bias×100	Turnbull Var×100	Par Var×100	RE
2.597	0.90	0.264	1.2675	0.2074	0.1079	1.679
4.291	0.75	0.636	-2.8225	0.4342	0.2346	1.395
6.660	0.50	-1.451	-4.0159	0.3869	0.2011	1.126
8.197	0.35	-0.128	-1.5346	0.5381	0.1798	2.647
10.149	0.20	-0.560	2.3217	0.3542	0.1565	1.698

Figure 4.1 shows the average estimates under each of the three distributional assumptions compared to the true survival function. As we saw in the right-censored case, we again observe that parametric estimation may not provide acceptable estimates when the distribution assumptions are incorrect.

As we did with right-censored data, we also considered some other interval-censored scenarios involving parametric estimation of the survival function where the true distribution was exponential or lognormal. We again found that we were successfully able to estimate an exponential survival function under the assumption that the true distribution was Weibull, since the exponential is a special case of the Weibull.

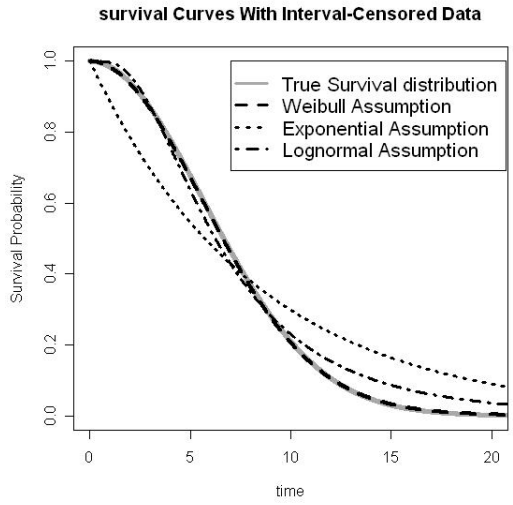

Figure 4.1: Multiple estimates of a Weibull survival function.

CHAPTER V
CONCLUSION

5.1 Conclusion

We have investigated different techniques to estimate the survival function with both right- and interval-censored data. We considered both parametric and nonparametric estimators, and we examined results when incorrect information was used either on the type of censoring of the data, or in the case of parametric estimation, the underlying distribution. Due to time limitations, we were only able to examine and compare some of the survival function estimators that are available. Note also that we only considered three specific distributions in our simulations.

In Chapter 3 we considered right-censored data simulations. We were able to conclude from observing RE values and 95% confidence interval coverage probabilities that the nonparametric survival function estimator for right-censored data, known as the Kaplan-Meier estimator, provides good estimates. We also saw that the parametric estimation of the survival function seems to consistently provide better estimates than the K-M estimator if the correct distribution is assumed. However, when an incorrect distribution is assumed, the parametric estimator is often very biased, and not always as efficient as the K-M estimator.

In Chapter 4 we considered interval-censored data. We were again able to conclude that the parametric estimator is more efficient than nonparametric estimates when distributional assumptions are correct. Although parametric estimates with in-

correct distributional assumptions may result in MSE values that are similar to those resulting from nonparametric estimators, it is also possible that incorrect distributional assumptions can result in very inefficient estimators that are both biased and have large variances.

From results given in Chapters 3 and 4, we saw that with both right- and interval-censored Weibull data, it is better to wrongfully assume that the data is from a lognormal distribution than an exponential distribution. Also, as mentioned in the previous two chapters, it is not problematic to estimate exponential survival functions by using a parametric estimator with the assumption of Weibull data. In summary, for both right- and interval-censored data, unless you are sure that the distribution of you data is a subset of another distribution, or you have definite information on the exact underlying distribution of your data, we recommend that one of the more conservative nonparametric estimation methods be applied.

From the numerous interval-censored simulations we ran, we found that as the number of visits decreased and the percentage of right-censoring in the data decreased, the efficiency of the Turnbull method compared to the K-M:MID method increased. When considering MSE values, the Turnbull estimator does not show improvement over the K-M:MID estimation method until both the number of visits is low and the percentage of right-censoring is low. Therefore, in this case we recommend the Turnbull method over the K-M:MID method. When the percentage of right-censoring is high and/or there are a large number of visits, the K-M:MID method will likely provide more efficient estimates. Although the K-M:MID method seems like a reasonable option to obtain an adhoc and often more efficient estimate of the survival function, we do not recommend use of the K-M:LEFT or K-M:RIGHT methods. This is because survival probabilities are often severely overestimated or

underestimated, and estimates resulting from these methods were never shown to be more efficient than a competing estimator.

Besides the results shown in the previous 2 chapters, we ran other simulations to complement the results we have shown, and to better understand what assumptions can be violated in survival function estimation, and what assumptions should not be violated. In many cases we found similar results to those presented in the previous two chapters. However, when using $n = 30$ as the sample size for each data set, RE values comparing nonparametric and parametric estimators often became smaller compared to when the sample size was $n = 100$. This was mostly due to changes in variance. Although both the parametric estimation methods and the competing nonparametric method showed increases in variances of estimates when the sample size decreases to 30, the parametric method shows a larger increase. This results in smaller RE values when $n = 30$.

We have given some confidence interval results in this thesis which confirm that Wald confidence intervals have good coverages when the correct parametric estimates are used. Also, we have shown that using K-M estimates with Greenwood standard errors results in good coverages at all time points, but only with right-censored data. Although we have mentioned confidence interval coverages, it is important to keep in mind that the main suggestions we have given in this thesis apply to survival function estimation, and that the suggestions given are not about any inferencial methods involving testing, which may rely on survival function estimates.

Due to the difficulty in calculating standard errors for the NPMLE in the case of interval-censored data, we have not given any inference results concerning the Turnbull method. Also, as the K-M:MID method is an adhoc method, without any recommended standard errors, a computer intensive bootstrap method would be one

technique to consider to obtain standard errors. Lindsey and Ryan warn that the K-M:MID method may give biased or misleading inferences [15]. Again, we have shown that these biases can sometimes occur, but have not given any results to support the use of the K-M:MID method for inferencial procedures, even if bootstrap standard errors were available.

5.2 Future Work

In this thesis, three of the most common survival distributions are considered. In the future, it would also be of interest to look at other distributions or the same distributions that we used with different parameters. An alternative distribution that might be informative to consider is a non-standard mixture distribution, such as the mixture of lognormal and an exponential. Also, instead of using the uniform distribution to generate censoring times, we could examine an alternative distribution, such as the exponential. Using an alternative censoring distribution other than the uniform may result in the Turnbull estimator showing more efficient estimates compared to the K-M:MID method in the case of interval-censored data. Moreover, consideration of a larger number of observations in each data set, and a larger number of data sets in each simulation would be beneficial. Using larger values for N and n is especially attractive, as this may eliminate some of the anomalies that we have seen in the results.

Regarding standard errors, we may also want to consider the bootstrap method, both for right- and interval-censored data. This would be especially useful in the case of interval-censored data, as the Turnbull estimator follows nonstandard asymptotic theory. Lastly, future research involving other estimation methods besides those considered in this thesis would nicely complement the results we have obtained.

BIBLIOGRAPHY

[1] Allison, Paul D. Survival Analysis Using SAS. Cary, NC: SAS Publishing, 1995.

[2] Anderson et al. Statistical Models Based on Counting Processes. New York: Springer-Verlag, 1993.

[3] Cantor, Alan B. SAS Survival Analysis Techniques for Medical Research. Cary, NC: SAS Publishing, 2003.

[4] Doehler, K. and Davidian, M. "'Smooth' Inference for Survival Functions with Arbitrarily Censored Data." Statistics in Medicine, in press, (2008).

[5] Feinleib, M. "A method of Analyzing Log Normally Distributed Survival Data with Incomplete Follow-Up." Journal of the American Statistical Association, Vol.55 (1960):534-545.

[6] Fox, J. "Introduction to Survival Analysis."
Lecture Notes for 'Statistical Applications in Social Research' 2006,
<http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf>.

[7] Goodall, Dunn, and Babiker. "Interval-censored survival time data: confidence intervals for the non-parametric survivor function." Statistics in Medicine, Vol.23 (2004):1131-1145.

[8] Greenwood, M. "The errors of sampling of the survivorship tables." Reports on Public Health and Statistical Subjects, No. 33, Appendix 1. London: H.M. Stationery Office, 1926.

[9] Horner, R.D. "Age at Onset of Alzheimer's Disease: Clue to the Relative Importance of Etilogic Factors?" American Journal of Epidemiology, Vol.126 (1987):409-414.

[10] Keiding, N. "Age-Specific Incidence and Prevalence: A Statistical Perspective." Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol.154, No. 3 (1991):371-412.

[11] Klein, John P. and Moeschberger, L.
Survival Analysis: Techniques for censored and truncated data). 2rd ed. New York : Springer,2003.

[12] Komáek, Lesaffre, and Hilton. "Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution."
Journal of Computational and Graphical Statistics, Vol.14, No. 3 (2005):726-745.

[13] Kooperberg, C. and Stone, Charles J. "A study of logspline density estimation."
Computational Statistics and Data Analysis, Vol.2, No. 3 (1991):327-347.

[14] Kooperberg, C. and Stone, Charles J. "Logspline Density Estimation for Censored Data." Journal of Computational and Graphical Statistics, Vol.1, No. 4 (1992):301-328.

[15] Lindsey, Jane C. and Ryan, Louise M. "Tutorial in Biostatistics Methods for Interval-Censored Data." Statistics in Medicine, Vol.17 (1998):219-238.

[16] Pan, W. "Smooth estimation of the survival function for interval censored data."
Statistics in Medicine, Vol.19 (2000):2611-2624.

[17] Peto, R. "Experimental Survival Curves for Interval-censored Data."
Journal of the Royal Statistical Society: series C (Applied Statistics), Vol.22, No. 1 (1973):86-91.

[18] Turnbull, Bruce W. "The Empirical Distrubution Function with Arbitrarily Grouped, Censored and Truncated Data."
Journal fo the Royal Statistical Society, Series B, Vol.38 (1976):290-295.