Nonparametric Analysis of Interval-Censored Failure Time Data

---

A Doctoral Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

Jeremy Gorelick

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

July, 2009

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled.

Nonparametric Analysis of

Interval-Censored Failure Time Data

presented by Jeremy Gorelick

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

Dr. (Tony) Jianguo Sun _____

Dr. Nancy Flournoy _____

Dr. Lori Thombs _____

Dr. Athanasios Micheas _____

Dr. Wade Davis _____

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my mentor and adviser, Dr. Jianguo Sun, for his guidance, encouragement and patience throughout this project. This work would never have been done without his inspiration and enthusiasm. I am also very grateful to the other members of my advisory committee Dr. Nancy Flournoy, Dr. Lori Thombs, Dr. Athanasios Micheas, and Dr. Wade Davis.

I would like to thank the Statistics Department faculty members, staff, and my fellow graduate students for making the department a strong one and a friendly atmosphere.

I am forever indebted to my family for their understanding and moral support throughout my time at school.

Nonparametric Analysis of Interval-Censored Failure Time Data

Jeremy Gorelick

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

This thesis considers the problem of treatment comparisons when only interval-censored failure time data are available. This type of data occurs frequently in clinical trials and other follow-up studies. We study several nonparametric procedures developed previously and compare them under different situations. In particular, we study the situation where the difference between the groups occurs at an early or late time period. For this problem, we generalize the log-rank tests developed for interval-censored data in Zhao and Sun (2004) and the weighted log-rank test presented in Kalbfleisch (2002). Numerical studies are conducted to evaluate the proposed test and compare it with the unweighted log-rank test, which indicate that the proposed method works well.

This thesis also considerers the problem of finding an appropriate sample size to achieve a desired power. We present a simple-to-use formula to find the sample size for a prespecified power and level of significance for the case of interval-censored data.

Since many researchers use missing data techniques such as imputation along with right censored methods to analyze interval-censored data, we also compare an imputed Kaplan-Meier Estimate of the survival function to Turnbull's Self Consistent Estimate. We present a large numerical study to show that these estimates often disagree at late time points when the mean survival time is large.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Interval-Censored Failure Time Data

In recent years, a great deal of attention has been given to the study of interval-censored failure time data. Many new methods have been developed to analyze this type of data, but there has not been much work done to compare the different techniques. In this thesis, we are interested in nonparametric estimation of a survival function and a nonparametric comparison of several survival function estimates. One goal is to investigate several of these new methods under different situations and to identify the method that performs better under each circumstance. Eventually, we would like to be able to decide the method that will yield the most power for a particular set of data.

What is interval-censored data? It is, simply put, data that has been censored into an interval. So, instead of seeing an event of interest, such as a patient's death time, or the time at which a person is infected with the human immunodeficiency virus (HIV),

we only see an interval within which the time falls. For example, we have a person who was screened six months ago for HIV and the results were negative, but when he was screened again last week his test was positive. We don't know exactly when he was infected with the virus only that it occurred sometime between 6 months ago and last week. This interval of time is where interval censored data get their name.

We usually denote the interval into which an event of interest is censored by $(L_i, R_i]$. So, L_i, the left endpoint, was the latest observation time where the event of interest had not occurred, and R_i, the right endpoint, is the first observation time after the event of interest has occurred. We have an open left endpoint because we know the event had not occurred at time L, but only after that time. Similarly, we don't know that the event of interest did not occur in the instant we tested the subject, so we have a closed right end point.

There are two special cases of interval-censored data that are often of interest. The first is called right-censored data. A right censored data point is one whose left endpoint is known, but we do not know the right endpoint, so the subject is censored into the interval (L_i, ∞). A right-censored data set is one where the failure times are either known exactly or right-censored. This type of data occurs frequently in medical studies where we are not able to follow the subjects for an infinite amount of time. Suppose, for example, that we are interested in the time at which a laboratory rat develops a tumor and we have 6 months to perform the experiment. If at the end of the 6 months a rat did not have a tumor, we can only say that it would develop a tumor some time after 6 months, so the censoring interval would be $(6, \infty)$.

In the same study, if we were to test the rats for the first time when they were 2 months old, and if we were to observe a 2 month old rat who had a tumor, we do not know when the rat got the tumor, only that it occurred before 2 months. The censoring interval for this rat would be $(0, 2]$. When a subject has experienced the event of interest before our first observation, we call these types of observations left-censored times. It is very clear that both left censored data and right censored data are special cases of interval-censored data.

Finklestein and Wolf (1985) presented a set of interval-censored data from a retrospective study to compare two different treatments for breast cancer. The goal was to compare early breast cancer patients who were treated with primary radiation therapy and adjuvant chemotherapy to women who were treated with radiation therapy alone. In the study, breast cancer patients were seen at the clinics at intervals of 4 to 6 months, with increased time after the primary radiation treatment or for people living in rural areas. The variable of interest was the time until the cosmetic deterioration defined as the appearance of breast retraction. All of the patients were treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980.

Since the doctors were not able to constantly monitor the women, they could only say that the cosmetic deterioration occurred after the previous visit and before the current one. The data for several women were right censored. This could be the result of women failing to return because they no longer lived in the Boston area (i.e. they changed clinics), or they no longer experienced any symptoms of the cancer, or a variety of other reasons.

Fischel et al. (1990) presented another set of interval-censored data. Zidovudine (AZT) is a thymidine analogue that inhibits the replication of HIV. The authors showed that AZT effectively delayed the onset of acquired immunodeficiency syndrome (AIDS) and prolonged the survival of patients with the AIDS virus. The study was composed of people who were diagnosed with mildly symptomatic HIV infections. All the people were evaluated twice before entering the study, and all subjects were monitored every two weeks for the first sixteen weeks of the study and then every month thereafter. In this study, the event that the researchers were interested in was the onset of AIDS or advanced AIDS-related complexes. An advanced AIDS-related complex is the presence of two or more symptoms as well as a CD4 count less than 200 $cells/mm^3$.

In order to test for the onset of AIDS, the doctors need to draw a sample of blood. It is not possible to sample a person's blood every day, so the doctors could only test the subjects on a monthly basis and sometimes longer. Also, since the study was terminated early, many of the subjects did not experience the onset of AIDS. These subject's failure time were subject to right censoring.

1.2 Analysis of Failure Time Data

When analyzing failure time data, there are generally three things that are of interest. The first is estimation of the survival function. Often, we need to estimate the survivor function before we can make any inferences about the population. In addition, many methods used for treatment comparison also require the survivor function to be estimated. The second is regression analysis. We use this in order to make inferences

about the survivor function as well as the effects of various covariates on the survivor function. The last task is treatment comparison. This is generally performed when we need to test the effectiveness of several different treatments. An example of treatment comparison is needed to investigate the effectiveness of a new cold medication. The doctors can administer a placebo and a newly developed cold remedy to two groups of patients with colds. Since their goal is to determine if the cold infection's survival time is lower for the group receiving the remedy, they would need methods to test this. One of the goals of this thesis is to analyze methods developed for interval-censored data.

1.2.1 Estimation of a Survival Function

For interval-censored failure time data, many early methods for estimation of a distribution function used constrained Newton-Raphson methods. Although this method worked, it was extremely cumbersome. In 1972, Bruce Turnbull developed a simpler and more straight-forward method for estimating the distribution function in his paper "The empirical Distribution Function with Arbitrarily Grouped, Censored, and Truncated Data." In this paper, Turnbull developed the so-called self-consistency algorithm. This algorithm is a special case of the EM algorithm and is used to estimate the survival function when the data is censored and/or truncated.

Turnbull's self consistency algorithm is based upon the maximum likelihood estimator (MLE) of the distribution function. It is important to note that the MLE, $\hat{F}$, of F will put probability mass only at a finite number of points. In other words, it will be a step-function. This means that we only need to estimate F at the points where

the jumps occur. To do this, we obtain initial estimates of $\underline{s} = (s_1, s_2, ..., s_m)$, where $s_j = F(p_j+) - F(q_j-)$. We can use these initial estimates to find the probability that the i^{th} observation lies in $[q_j, p_j]$ and the expected number in the group corresponding to the i^{th} observation which have values in $[q_j, p_j]$. We can then obtain improved estimates of $\underline{s}$. We repeat this process until the convergence is achieved. The main advantage of this algorithm is the fact that it is "automatic, simple to implement, and is intuitively appealing" (Trunbull, 1976).

Although it is a very powerful technique, the self-consistency algorithm is often very slow to converge. Gentleman and Geyer (1990) developed an improved method for estimating a survivor function in which the standard convex optimization technique is applied. They also provide easily verifiable conditions for the self-consistent estimator proposed by Turnbull to be the maximum likelihood estimator and for checking whether the MLE is unique. A sufficient condition is given for almost sure convergence of the MLE to the true underlying distribution function.

The method proposed by Gentleman and Geyer requires two assumptions: the censoring must be noninformative and failures cannot coincide with observation times. The assumptions are required in order for the probability of the observation times to not involve any of the parameters of interest. We need this so that we can consider the likelihood conditional upon the observed intervals. Gentleman and Geyer then considered the problem of estimating $\underline{s}$, where $s_j = F_0(p_j-) - F_0(p_{j-1})$. In order to find the MLE of p, five conditions called the Kuhn-Tucker conditions must hold. The MLE of $\underline{s}$, $\underline{\hat{s}}$, will be the solution to the Kuhn-Tucker equations.

These methods of estimating the survival function can be difficult to compute. However, in the case of right-censored data, the MLE is given by the Kaplan-Meier or Product Limit Estimator (Kalbfleisch 2002). The Kaplan-Meier estimate is a generalization of the empirical distribution function for complete data. The Kaplan-Meier estimate is such that the estimated probability of failure agrees exactly with the observed proportion of deaths at a particular time and the number of individuals still in the study at that time.

Since this method is relatively easy to calculate, many researchers also use this method when the data are interval-censored. However, this requires using a missing data technique such as imputation. Sun (2006) suggests a simple imputation scheme. He pointed out that a common way to perform this imputation is to simply use the right endpoint or the left endpoint of the interval. Sun also pointed out that a major advantage of the single imputation methods is they are simple to compute and when the intervals of observation are narrow the right-imputed estimate and the left-imputed estimate will be similar.

If a single imputation is not adequate, many multiple imputation methods are also available. Several of these can be found in Sun (2006).

1.2.2 Regression Analysis of Interval-Censored Data

One of the first papers discussing regression analysis of interval-censored failure time data was given by Finklestein in "A Proportional Hazard Model for Interval-Censored Failure Time Data" in 1986. In the paper, she suggested estimating unknown

parameters by using the maximum likelihood approach. In order to use this approach, two assumptions are necessary: i) the censoring time is independent of the covariates and the time of the event of interest, and ii) given an infinite amount of time, each subject will experience the event of interest. These assumptions are not very difficult to achieve.

For example, in a follow-up study where scientists only have funding for six months of follow-up, the censoring time for all subjects would be at the six month mark. This does not have any effect on the failure time, nor would the covariates such as the treatment group. The second assumption is also very common in real world applications. In a study where researchers are interested in the amount of time a patient has AIDS before he/she dies, we know that all people will eventually die no matter what treatment they receive.

A drawback of the likelihood approach developed by Finklestein is that it could involve the inversion of potentially large matrices. To deal with these sorts of situations, several different types of methods were developed. In 2000, Pan, for example, developed a method based on the multiple imputation approach. Other methods can be found in Sun (2006).

1.2.3 Nonparametric Treatment Comparison

Methods like Finklestein's involve making assumptions about the distribution from which the failure times occur. For example, she made the assumption that the survival times follow the Cox model. In many situations, we do not know anything about the

underlying distribution of survival times. Even when we do not know the shape of the survival function, we may wish to compare several different treatments to see if the survival time for one treatment is longer than another. In these cases, we use tests which do not make any assumptions about the shape of the distribution, or nonparametric tests. A major advantage of using these tests is that no matter what the underlying distribution actually is, they will give the correct level of significance.

Peto and Peto (1972) developed a rank-based testing procedure for comparing treatment groups when right-censored data are present. They focus on three tests in particular: the log-rank test, the probit-rank test, and Wilcoxson's rank sum test. The Wilcoxson test proposed is a generalization of the usual version (Wilcoxson 1945). The scoring system is modified to take into account the censoring of the data. The authors noted that their generalization is preferred to previous generalizations because under right-censoring the expectation of their test statistic remains the same as Wilcoxson's.

The log-rank test is a test which uses the number of patients still in the study and the number of patients who fail. They simply compute the observed numbers and the number we would expect to see if the treatment groups were from an identical distribution. The test is based on the difference between these observed numbers and the expected ones.

The authors pointed out that the decision to use a particular test is based not only on the power of the test in question, but also of the power of other tests for the same situation. They showed that under a normal alternative, the probit-rank test is the most efficient of the three tests and the log-rank test is the most efficient under

Lehmann alternatives.

Kalbfleisch and Prentice (2002) pointed out that there are important extensions of the log-rank test. A stratified log-rank test allows for heterogeneity in the populations being compared. They also noted that the log-rank test is sensitive when the hazard ratios are constant over time. There are many cases where this may not hold. They gave a weighted log-rank test to help handle this type of situation.

Sun (1996) further generalized the log-rank test for the case where interval-censored data are present. In his paper, Sun suggested a procedure to estimate the number of patients still at risk and the number of deaths at each time in order to approximate the log-rank test for right censored data. He also suggested two methods for estimating the variance. One method is to use the method proposed by Louis (1982) and the other method is to use a logistic approximation. Sun (2004) developed a further generalization of the log-rank test. In his 2004 paper, he proposed a model that would reduce to the usual log-rank test when right-censored data are present.

1.3 Sample Size Calculations

In many medical studies, it can be costly to follow patients for an extended period of time. Many researchers would like to ensure they can achieve a desired level of power for a test of $H_0 : S_1(t) = S_2(t) \, \forall t$, while keeping costs low. This involves making sure enough patients are recruited for the study but not too many. Because of this, a power analysis is often run before a study begins or after a brief pilot study.

There have been many methods developed to find an appropriate sample size.

Schoenfeld (1981) presented methods for computing the asymptotic mean and variance of a generalized version of the log-rank statistic, the modified Wicoxson statistic, and many other commonly used methods for comparing the survival curve. Using these results, he also presented a sample size calculation for the log-rank statistic.

Schoenfeld's methods are good under a specific set of assumptions. Lakatos (1988) developed a more general method for computing the sample size. As in Schoenfeld's paper, Lakatos computed the sample size required to compare two survival functions using the log-rank test. However, Lakatos's calculation removed some of the more restrictive assumptions from the earlier calculations. His method does not require the proportional hazards assumption. It also has the benefit of allowing for some more common clinical trial designs such as a trial with staggered entry or stratification. Lakatos's method is general and easy enough to use that major software developers such as PASS use it for their power analysis programs.

1.4 Outline of the Thesis

In recent years many methods have been developed to analyze interval-censored failure time data. In particular, there has been a great deal of attention given to treatment comparison. However, not much effort has been put into comparing these methods. One of the goals of this thesis is to analyze several of the new methods that have been developed in recent years and decide which approach is best under a practical, specific situation. Chapter 2 describes several of these methods and presents the results of a large simulation study for the evaluation of the methods. We also apply

the methods to a real-world example.

While these methods perform well in a general setting, in some studies we may expect that the difference in the survival functions occurs at early or late time points. Since these methods may not be sensitive to these early or late differences, Chapter 3 introduces a new weighted log-rank test applicable when interval-censored data are present. This test is more sensitive to early and late differences in the survival function than previously developed nonparametric tests.

Often, clinical trials can be expensive to run. Researchers would like to keep the costs of these trials as small as possible while ensuring that the power of the tests is adequate. This means that they need to recruit enough patients into the study to achieve their desired power, but not too many. Chapter 4 presents a method for computing the sample size required for interval-censored data.

In some cases, researchers are unfamiliar with techniques to analyze interval-censored data, but they are comfortable using techniques developed for right-censored data. When this situation arises, the researcher may use a missing data technique such as imputation to approximate a right-censored data set, and then use a method such as Kaplan-Meier to estimate the survival function. Chapter 5 considers such a situation and specifically investigates the difference between the imputed Kaplan-Meier estimate and the self consistency estimate of a survival function.

Chapter 2

Statistical Methods for Comparing Survival Functions

2.1 Notation

Consider a survival study involving two treatment groups. Let T denote the time of the event of interest, $S(t) = P(T \geq t) = 1 - F(t)$ the survival function, $\lambda(t) = -\frac{d}{dt} log S(t)$ the hazard function of the survival time, and $\Lambda(t) = \int_0^t \lambda(u)du$ the cumulative hazard function. Then $S_j(t)$, $\lambda_j(t)$, and $\Lambda_j(t)$ are the survival function, hazard function, and cumulative hazard function for treatments $j = 1, 2$ and T_i is the event time of interest for subject i. Let n_1 and n_2 be the number of subjects in treatment groups 1 and 2, respectively, and $n = n_1 + n_2$ the total number of subjects. Let $\hat{S}(t)$ denote the maximum likelihood estimator (MLE) of $S(t)$ and $\hat{S}_j(t)$ the MLE of $S_j(t)$. Since our overall goal is to test whether the two treatment groups have the same survival function, we will be testing $H_0 : S_1(t) = S_2(t)\ \forall t$.

2.2 Method 1: A Generalized Log-Rank Test for Interval-Censored Failure Time Data

This method was developed by Zhao and Sun (2004). It was developed as an improvement over the method developed by Sun (1996), which could overestimate the numbers of risks and failures. Thus, when right-censored data are available, the test may not reduce to the log-rank statistic. The new method overcomes these faults and reduces to the log-rank statistic when right-censored data are available.

In Zhao and Sun (2004), the authors started by defining $\delta_i = 0$ if the observation on the failure time T_i is right censored and $\delta_i = 1$ otherwise, and $\rho_{ij} = I(\delta_i = 0, L_i \geq s_j)$. So, ρ_{ij} is the indicator of the the event that T_i is right-censored and subject i is still at risk at time s_j-. Define $\alpha_{ij} = I(s_j \in [L_i, R_i])$ the indicator of event $s_j \in [L_i, R_i]$. The authors estimated the total number of failures and the total number of subjects at risk at time s_j by

$$d'_j = \sum_{i=1}^{n} \delta_i \frac{\alpha_{ij}(\hat{S}(s_j) - \hat{S}(s_j+))}{\sum_{u=1}^{m+1} \alpha_{iu}(\hat{S}(s_u) - \hat{S}(s_u+)}$$

and

$$n'_j = \sum_{r=j}^{m+1} \sum_{i=1}^{n} \delta_i \frac{\alpha_{ir}(\hat{S}(s_r) - \hat{S}(s_r+))}{\sum_{u=1}^{m+1} \alpha_{iu}(\hat{S}(s_u) - \hat{S}(s_u+))} + \sum_{i=1}^{n} \rho_{ij} \ ,$$

and the numbers of failures and subjects at risk in treatment group l at time s_j by

$$d'_{jl} = \sum_{i}^{l} \delta_i \frac{\alpha_{ij}(\hat{S}(s_j) - \hat{S}(s_j+))}{\sum_{u=1}^{m+1} \alpha_{iu}(\hat{S}(s_u) - \hat{S}(s_u+))}$$

and

$$n'_{jl} = \sum_{r=j}^{m+1} \sum_{i}^{l} \delta_i \frac{\alpha_{ir}(\hat{S}(s_r) - \hat{(}S)(s_r+))}{\sum_{u=1}^{m+1} \alpha_{iu}(\hat{S}(s_u) - \hat{S}(s_u+))} + \sum_{i=1}^{n} \rho_{ij} \ ,$$

where $\sum_i^l$ denotes the summation over all subjects in population l, $l = 1, 2$, $j = 1, \ldots, m$. If the data were right-censored these would be exactly the numbers of failures and subjects at risk.

To test H_0 the authors use the test statistic $\underline{U} = (U_1, U_2)^t$, where

$$U_l = \sum_{j=1}^{m} d'_{jl} - n'_{jl} \frac{d'_j}{n'_j} \ .$$

When the data are right censored this U is exactly the same as the usual log-rank statistic. Since the usual log-rank statistic is a special case of this test, the authors called this the generalized log-rank test.

To carry out the test, the authors suggested using $U^* = U^t V^- U$ which follows a χ^2_1 distribution asymptotically under the null hypothesis. Here V^- is the generalized inverse of the estimate of the covariance matrix of U. Unfortunately, using the fisher information matrix to estimate V is extremely complicated, so the authors proposed using a multiple imputation to estimate it. The idea of the imputation is to impute the failure times for subjects whose failure times are not right censored. Then the covariance matrix can be estimated by summing the within-imputation covariance and the between-imputation covariance. They use a 3-step procedure to accomplish this.

Let M be a pre-specified number of resamplings for the bootstrap. Then for each r in $1, \ldots, M$,

Step 1: If $\delta_i = 0$, then let $T_i^r = L_i$ and $\delta_i^r = 0$. Otherwise, $\delta_i^r = 1$ and T_i^r is a realization from the conditional survival function

$$f_i(s) = P(T_i^r = s) = (\hat{S}(s) - \hat{S}(s+))/(\hat{S}(L_i) - \hat{S}(R_i+)), \;\; s \in [L_i, R_i].$$

So, the δ_i^r are always going to be the same. Step 2: Use the new data to find the number of failures and risks, and estimate U as before, then compute the covariance estimates $\hat{V}^r = \hat{V}_1^r + \hat{V}_m^r$ where

$$(\hat{V}_j^r)_{ll} = \frac{n_{jl}^r(n_j^r - n_{jl}^r)d_j^r(n_j^r - d_j^r)}{(n_j^r)^2(n_j^r - 1)} \;\; ,$$

$$(\hat{V}_j^r)_{l_1 l_2} = \frac{n_{jl_1}^r n_{jl_2}^r d_j^r(n_j^r - d_j^r)}{(n_j^r)^2(n_j^r - 1)}, \;\; l_1 \neq l_2.$$

Step 3: Repeat steps 1 and 2 for $r = 1, \ldots, M$ the number of imputations, and estimate V by $\hat{V} = \hat{V}_1 + \hat{V}_2$, where

$$\hat{V}_1 = \frac{1}{M}\sum_{r=1}^{M} \hat{V}^r \;\; ,$$

$$\hat{V}_2 = (1 + \frac{1}{M})\frac{\sum_{r=1}^{M}[U^r - \bar{U}][U^r - \bar{U}]^t}{(M-1)} \;\; ,$$

and

$$\bar{U} = \sum_{r=1}^{M} \frac{U^r}{M} \;\; .$$

If we had right-censored data, all of the $\hat{V}^r$'s would be the same, and $\hat{V}_2$ would be $\mathbf{0}$,

and $\hat{V}$ would be the usual estimate for the Log-Rank Statistic.

2.3 Method 2: A Generalized Log-Rank Tests for Interval Censored Failure Time Data II

This method was developed by Sun, Zhao, and Zhao (2005). To test H_0 the authors proposed using the following test statistic

$$U_\xi = \sum_{i=1}^{n} x_i \frac{\xi[\hat{S}(L_i)] - \xi[\hat{S}(R_i)]}{\hat{S}(L_i) - \hat{S}(R_i)} \ ,$$

where x_i is the 2x1 vector of treatment indicators associated with subject i whose l^{th} element is equal to 1 if it is from population l and 0 otherwise, and ξ is a known function over (0, 1). When $\xi(t) = tlogt$, and the data are right censored, this will give us the usual log-rank statistic.

In order to establish the asymptotic distribution of U_ξ, we first must define several things. $\eta(x) = 1 - \xi(1 - x)$ and we assume that $lim_{x->0}\eta(x) = lim_{x->1}\eta(x) = c_0$. Then we let H and h denote the distribution and density functions of (U_i, V_i) , and $F(t) = 1 - S(t)$. So, we can rewrite U_ξ as

$$U_\eta = \delta \frac{\eta[F(U)] - c_0}{F(U)} + \Gamma \frac{\eta[F(V)] - \eta[F(U)]}{F(V) - F(U)} + (1 - \Delta - \Gamma) \frac{c_0 - \eta[F(V)]}{1 - F(V)} \ .$$

If we let λ_2 and ν_2 denote the Lebesgue measure on R^2 and counting measure on the

set $[(0,1),(1,0),(0,0)]$, respectively. Define

$$q_{F,H}(u,v,\delta,\gamma) = h(u,v)(F(u))^{\delta}(F(v)-F(u))^{\gamma}(1-F(v))^{1-\delta-\gamma} \ ,$$

the density function of $(U_i, V_i, \Delta_i, \Gamma_i)$, and define $dQ_0 = q_{FH}d(\lambda_2 \otimes \nu_2)$

$$Q_n(u,v,\delta,\gamma) = \frac{1}{n}\sum_{i=1}^{n} I[(U_i,V_i) \leq (u,v), (\Delta_i,\Gamma_i) = (\delta,\gamma)] \ ,$$

$$K_0(u,v,\delta,\gamma) = \delta\frac{\eta[F(u)]-c_0}{F(u)} + \gamma\frac{\eta[F(v)]-\eta[F(u)]}{F(v)-F(u)} + (1-\delta-\gamma)\frac{c_0-\eta[F(v)]}{1-F(v)} \ .$$

Now, if the regularity conditions of Groenboom and Wellner (1992) hold for the strong consistency of $\hat{F}_n = 1-\hat{S}(t)$ and $F(t)$ has a support on [0,M] with continuous density function, then U_ξ has an asymptotic distribution given by theorem 1 in Sun et al. (2005) with $k=2$. So, U_η has an asymptotic normal distribution with mean 0 and covariance matrix $\Sigma = [\sigma_{lr}]_{2x2}$, where

$$\sigma_{ll} = p_l(1-p_l)Q_0(K_n^2)$$

and

$$\sigma_{lr} = -p_l p_r Q_0(K_n^2), \ \ l \neq r.$$

So, we can consistently estimate Σ by $\hat{\Sigma} = [\hat{\sigma}_{lr}]_{2x2}$, where

$$\hat{\sigma}_{ll} = \frac{n_l(n-n_l)}{n^2} Q_n(\hat{K}_n^2)$$

and

$$\hat{\sigma}_{lr} = \frac{-n_l(n_r)}{n^2} Q_n(\hat{K}_n^2) \ .$$

It is clear that the sum of the components of U_η will be zero, and $\hat{\Sigma}$ will be singular. So, if we let U_0 denote the first component of U_η and $\hat{\Sigma}_0 = \hat{\sigma}_{11}$, we can test H_0 with $\chi_0 = U_0^t \hat{\Sigma}_0^{-1} U_0$ which has a χ_1^2 asymptotic distribution under the null hypothesis.

For this thesis, we will restrict ourselves to the same class of functions for ξ as the authors did, namely $\xi(x) = (xlogx)x^\rho(1-x)^\gamma$ where ρ and γ are constants. If $\rho = \gamma = 0$ and the data were right censored, then we would have the usual log-rank test.

As noted by the authors, many methods for comparing survival functions when data is interval censored do not have known asymptotic properties. A key advantage of this method is that they derived the asymptotic properties for their test. Here we know that the results will hold regardless of the distribution of the survival function. Other methods may work in the situations the authors propose, but when applied to another situation (i.e. a survival curve that does not follow the proportional hazard model) the method may fail.

2.4 Imputed Log-Rank Test: A Log-Rank Test Based on a Single Imputation

This test utilizes the usual log-rank test. However, since the log-rank test is only applicable to right-censored data, we must use the interval censored data to simulate right censored data. We accomplish this task with a single imputation of the failure time.

We first estimate the common survival function $\hat{S}(t)$ under $H_0 : S_1(t) = S_2(t) = S(t)$, $\forall t$. Next, we use a single imputation procedure to impute right-censored data. Let $\delta_i = 0$ if the i^{th} subject's failure time is right censored and $\delta_i = 1$ if it is not right censored. So, δ is the indicator that the data point is not right censored. Now, the imputation is two steps.

Step 1: If $\delta_i = 0$, then let $T_i^r = L_i$. Otherwise, when $\delta_i^r = 1$ and T_i^r is a realization from the conditional survival function

$$f_i(t) = P(T_i^r = s) = \frac{\hat{S}(s) - \hat{S}(s+)}{\hat{S}(L_i) - \hat{S}(R_i+)} \ .$$

So, the δ_i^r are always going to be the same.

Step 2: Use the right-censored data to find the numbers of failures and risks at each time point, t_i. We call d_{lj} the number of risks at time t_j from the l^{th} population, r_{lj} the number of subjects at risk at time t_j^- from the l^{th} population, d_j the number of failures from all populations at time t_j, and r_j the total number of subjects at risk at time t_j^-. Also, define the expected number of failures from population l at time t_{lj}

$\omega_{lj} = E[d_{lj}|d_j] = r_{lj}\frac{d_j}{r_j}$.

So, we let D_l be the total number of failures from population l and E_l the expected number of failures from population l. That is,

$$D_l = \sum_{j=1}^{m} d_{lj}$$

and

$$E_l = \sum_{j=1}^{m} \omega_{lj} \ .$$

We can also estimate the variance of $d_{lj}|d_j$ by $V_j = [V_{ij}]$, where

$$V_j^{(ll)} = \frac{r_{lj}(r_j - r_{lj})d_j(r_j - d_j)}{r_j^2(r_j - 1)}$$

and

$$V^{(l_1 l_2)} = -\frac{r_{l_1} r_{l_2} d_j(r_j - d_j)}{r_j^2(r_j - 1)}, \ \ l_1 \neq l_2.$$

Then the log-rank statistic is defined as $\underline{\nu} = (D_1 - E_1, D_2 - E_2)^t$.

We estimate the covariance of $\underline{\nu}$ by $V = V_1 + V_2 + \ldots + V_m$. We can then test H_0 with $\chi^2 = \underline{\nu}^t V^{-1} \underline{\nu}$ based on a χ^2 distribution with 1 degree of freedom. However, since $\sum_i \nu_i = 0$ we will use $\chi^{2*} = (\nu^*)^t (V^*)^{-1} (\nu^*)$ to test H_0. Where $\nu^* = \nu_1$ and $V^* = V_{11}$.

A major drawback of this log-rank test is that it may not perform well when the hazard functions cross. This is because we can rewrite $D_l - E_l$ as

$$D_l - E_l = \sum_{j=1}^{m}[r_{lj}(\frac{d_{lj}}{r_{lj}} - \frac{d_j}{r_j})] = \sum_{j=1}^{m} r_{lj}(\hat{\lambda}_{lj} - \hat{\lambda}_j) = \int_0^\infty \omega_i(t)[d\tilde{\Lambda}_i(t) - d\tilde{\Lambda}(t)] \ .$$

So, if the hazards cross, this could integrate out to 0 even if there is a large difference between the survival functions. However, this is counterbalanced by the fact that if the survival functions follow the proportional hazards model, the log-rank test is the most efficient test for right-censored data. Another advantage that this test has is the ease of use. Many researchers are familiar with the log-rank test, so if a statistician were to inform the researcher that they were using the log-rank test, the researcher would feel at ease knowing a powerful and useful tool was implemented.

2.5 The Kolmogorov Approach

Many of the methods developed earlier in this chapter as well as methods not considered by this thesis perform poorly in the situation when the hazard functions cross each other. For example, the method proposed in Zhang et al. (2001) involves an integral of the difference between the survival functions. If the hazards cross, then the survival functions may also cross and will result in a small test statistic, even though we may be far from $H_0 : S_1(t) = S_2(t) \, \forall t$. Consequently, methods like this typically have a high probability to cause a type II error.

The Kolmogorov approach is very simple in its design. We use the test statistic $K = sup_{t \geq 0}|S_1(t) - S_2(t)|$. So, we first estimated the survival function using the

self-consistency algorithm, and used this to find the test statistic $K^* = max_{t\geq 0}|\hat{S}_1(t) - \hat{S}_2(t)|$. We will reject H_0 if K* is large. In order to determine if K^* is large, a bootstrap procedure is employed. This is broken down into 3 steps.

Step 1: Re-sample the data. To accomplish this, all the data are placed in a population and two random samples of sizes n_1 and n_2 of these combined data are taken with replacement.

Step 2: Use the new data to estimate the survival functions $\hat{S}_1^r$ and $\hat{S}_2^r$, and find the test statistic K_r^* with these new data.

Step 3: Repeat steps 1 and 2 for $r = 1, 2, ..., M$ the bootstrap size.

Now, if the regularity conditions from Fang, Sun, and Lee (2002) hold, we can apply the theorem from that paper, and when M is large the bootstrap samples $K_1^*, ..., K_M^*$ follow a normal distribution. So, the p-value of the test can be calculated to be the proportion of the K_r^*'s whose values are greater than K^*.

The main advantage of the Kolmogorov method is that it performs well when the hazard functions cross. This is due to the fact that this test looks only at the maximum of the difference between the two estimated survival functions. It is also easy to use. It does not require any complicated formulas. Unfortunately, it is very computationally expensive. This is because it requires the estimation of two survival curves for each bootstrap resample and many algorithms used to estimate the survival function, like the self-consistency algorithm, are very slow to converge. That is, this method will take a fairly long time to run when the bootstrap size is large.

2.6 Method 5: A Simple Nonparametric Two-sample Test

In many studies, patients are examinened several times. The previous methods only take into consideration the time of the examination before the failure occurred and the observation after it occurred. Zhang et. al (2003) developed a simple method for analyzing interval-censored data that takes into account all observation time points for each patient. The authors note that this type of method is more realistic in its application.

To construct the test statistic, define $\alpha_{0,i} = \int(1 - S_0(t))dG_i$ where G_i is the distribution function of the i^{th} observation time. The authors note that $\alpha_{0,i} = P(T \leq T_i) = E[I(T \leq T_{i,j})]$. So, we can estimate $\alpha_{0,i}$ with the simple empirical estimate

$$\hat{\alpha}_{n,i} = \sum_{[j:K_j \geq i]} \frac{I(T_j \leq T_{i,j})}{n^{(i)}} \ ,$$

where $n^{(i)} = \sum_{j=1}^{n} I(K_j \geq i)$ is the number of subjects who have at least i observations, and K_j is the number of observations of the j^{th} subject, and T_{ij} is the i^{th} observation time for the j^{th} subject ($j = 1, \ldots, n$). Zhang et al. (2003) proved that

$$\sqrt{n}(\underline{\hat{\alpha}}_n - \underline{\alpha_0}) \rightarrow_p N(0, A) \ ,$$

where $\underline{\hat{\alpha}}_n = (\hat{\alpha}_{n,1}, \hat{\alpha}_{n,2}, ...\hat{\alpha}_{n,K_0})^t$ and $\underline{\alpha}_0 = (\alpha_{0,1}, \alpha_{0,2}, ...\alpha_{0,K_0})^t$. Let $\nu_i = P(K < i)$ the probability that a subject drops out by the i^{th} observation. Then we have the asymptotic covariance matrix $A = [a_{i,j}]_{K_0 X K_0}$ with

$$a_{i,j} = \alpha_{0,j} \frac{1 - \alpha_{0,i}}{1 - \nu_j} \ .$$

The authors used the asymptotic results above to construct the nonparametric test statistic. So, under H_0 we have

$$\sqrt{n_1}(\underline{\hat{\alpha}}_{n_1} - \underline{\alpha_0}) \rightarrow_p N(0, A)$$

and

$$\sqrt{n_2}(\hat{\alpha}_{n_2} - \alpha_0) \rightarrow_p N(0, A) \ ,$$

where $\underline{\hat{\alpha}}_{n_1}$ and $\underline{\hat{\alpha}}_{n_2}$ are empirical estimates of $\underline{\hat{\alpha}}_0$ for the two samples. Then if $n_1/n \rightarrow \mu$, the above yields

$$\sqrt{n}(\underline{\hat{\alpha}}_{n_1} - \underline{\hat{\alpha}}_{n_2}) \rightarrow_d N(0, \frac{A}{\mu(1-\mu)}) \ .$$

Now, let $\underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_{K_0})^t$ be the weights, adjusting for attrition in follow-up studies. So, let ω_i be the percentage of subjects in the combined sample who have at least i observations. Then, we have

$$\sqrt{n} \sum_{i=1}^{K_0} \omega_i (\hat{\alpha}_{n_1,i} - \hat{\alpha}_{n_2,i}) \rightarrow_d N(0, \frac{\underline{\omega}^t A \underline{\omega}}{\mu(1-\mu)})$$

and

$$\frac{\underline{\omega}^t A \underline{\omega}}{\mu(1-\mu)} = \sum_{i=1}^{K_0} c_i \omega_i^2 \alpha_{0,i}(1 - \alpha_{0,i}) + 2 \sum_{i=1}^{K_0 - 1} \sum_{j=i+1}^{K_0} c_i \omega_i \omega_j \alpha_{0,i}(1 - \alpha_{0,j}) \ ,$$

with $c_i = 1/[\mu(1-\mu)(1-\nu_i)]$ for $i = 1, \ldots, K_0$, and we have the test statistic

$$Z = \frac{n^{1/2} \sum_{i=1}^{K_0} w_i(\hat{\alpha}_{n_1,i} - \hat{\alpha}_{n_2,i})}{(\sum_{i=1}^{K_0} c_{n,i} w_i^2 \hat{\alpha}_{n,i}(1-\hat{\alpha}_{n,i}) + 2\sum_{i=1}^{K_0-1} \sum_{j=i+1}^{K_0} c_{n,i} w_i w_j \hat{\alpha}_{n,i}(1-\hat{\alpha}_{n,i}))^{1/2}} \ .$$

H_0 is rejected when $|Z| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ point of a standard normal distribution.

A main advantage of this test is that it is that although "it does not use information about the actual observation times, it is not a rank-based test." Also, "although it is designed for continuous observation data, it is applicable to discrete observational data as well." Many methods were created with type II interval censoring strictly in mind. This method, however, was designed to be used with type k interval censoring. So, it is applicable no matter how many observations you have for each subject.

2.7 A Simulation Study

2.7.1 Simulation Set-ups

The simulation study analyzed the five different techniques given in the previous sections. In the study, we compared the methods under both proportional hazard setups as well as non-proportional hazard setups. For the proportional hazard setup, we used the Weibull model with shape and scale parameters 1 and a, respectively. So, $S_1(t) = exp(-t^a)$ and $S_2(t) = S_1^{\beta}$, with $\beta = 1.73$, 2, and 2.25.

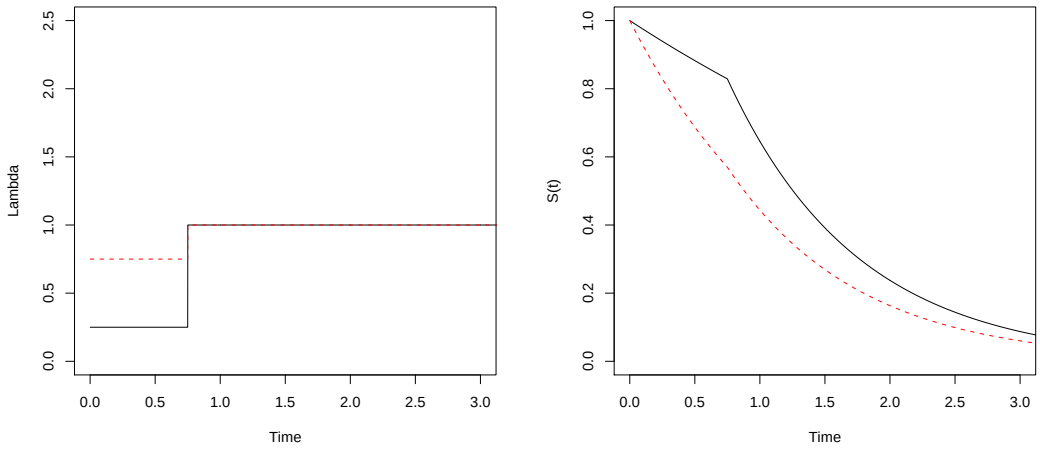
For the non-proportional hazard model, we looked at piecewise exponential func-

tions, with an early difference between the survival functions, a late difference between the survival functions, and most importantly a situation where the hazard functions cross. Specifically, we considered three setups with the following hazard functions and plots for the corresponding hazard and survival functions.

Early difference: Hazard A

$$\lambda_1(t) = 0.25I_{t\leq 0.75} + I_{t>0.75}$$

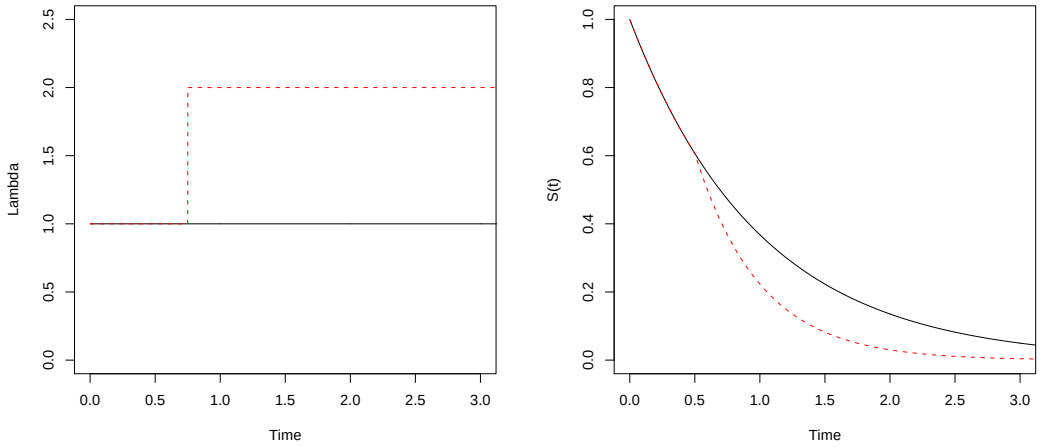
$$\lambda_2(t) = 0.75I_{t\leq 0.75} + I_{t>0.75}$$



Late Difference: Hazard B

$$\lambda_1(t) = I_{t\leq 0.5} + I_{t>0.5}$$

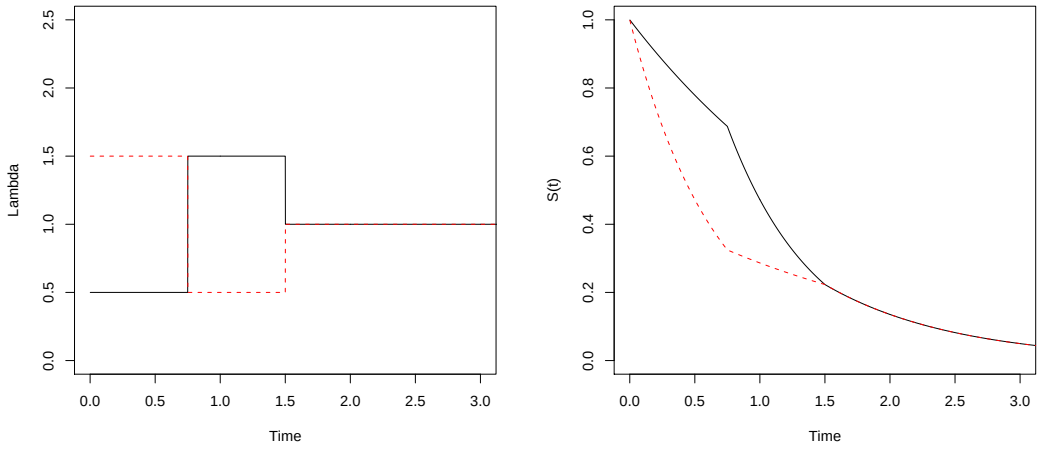
$$\lambda_2(t) = I_{t\leq 0.75} + 2I_{t>0.75}$$



Crossing Hazards: Hazard C

$$\lambda_1(t) = 0.5I_{t\leq 0.75} + 1.5I_{0.75<t\leq 1.5} + I_{t>1.5}$$

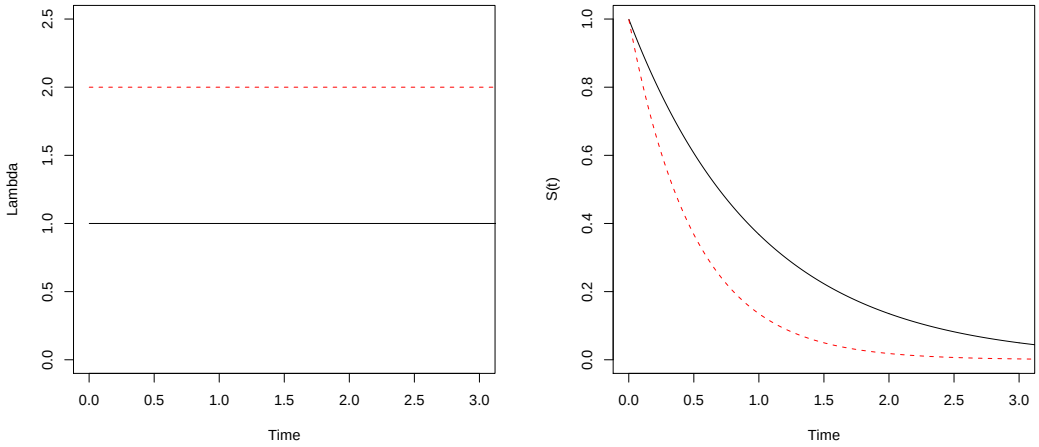
$$\lambda_2(t) = 1.5I_{t\leq 0.75} + 0.5I_{0.75<t\leq 1.5} + I_{t>1.5}$$



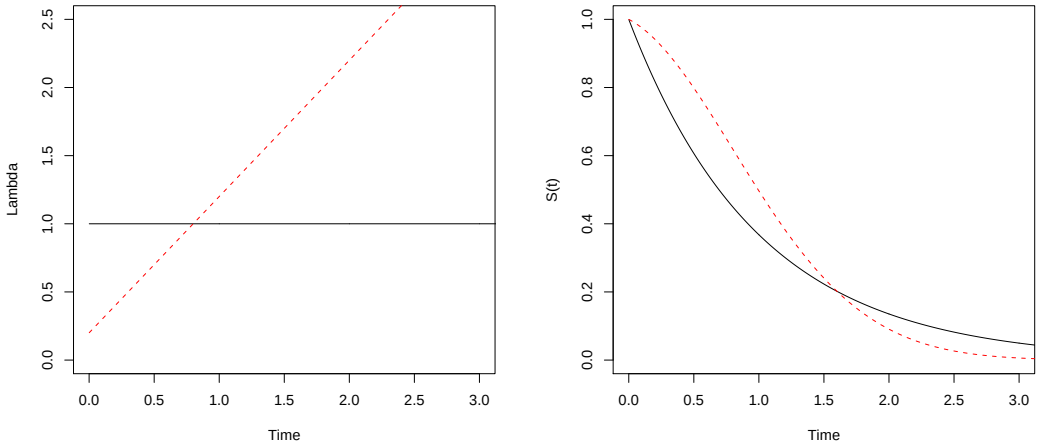
In addition to the piecewise exponential, We studied the linear hazard functions

given by $\lambda(t) = \alpha_0 + \alpha_1 x$. These hazard functions can give parallel hazard functions, crossing hazard functions, and neither parallel nor crossing (NPNC) hazard functions.

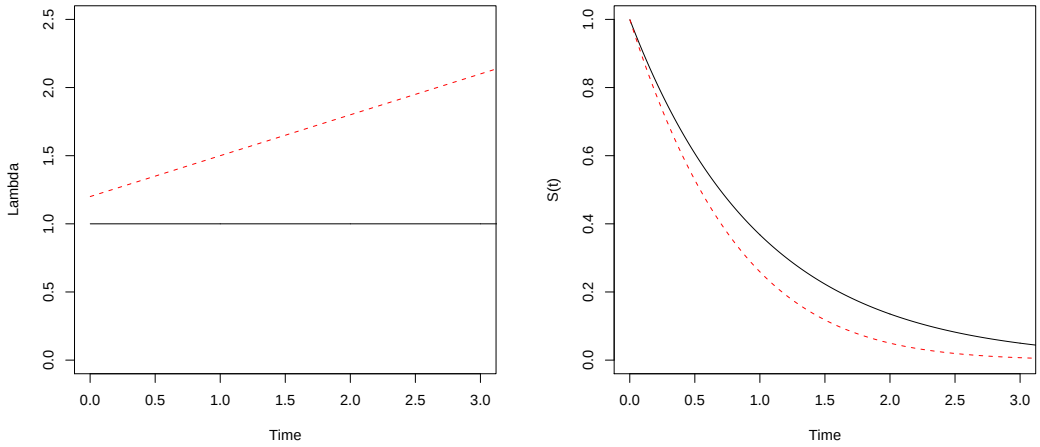
Parallel Hazards, Hazard D: $\lambda_1(t) = 1$ and $\lambda_2(t) = 2$



Crossing Hazards, Hazard E: $\lambda_1(t) = 1$ and $\lambda_2(t) = 0.2 + t$



NPNC Hazards, Hazard F: $\lambda_1(t) = 1$ and $\lambda_2(t) = 1.2 + .3t$



We also considered two hazards to check the size of the tests.

Exponential(1) Hazards: Hazard S

$$\lambda_1(t) = \lambda_2(t) = 1$$

Weibull$(2, 2)$ Hazards: Hazard W

$$\lambda_1(t) = \lambda_2(t) = 8t$$

2.7.2 Generation of Interval-Censored Data

We used R to draw a sample of fifty subjects from populations whose survival functions were described above. We then drew a random sample from a uniform distribution on the interval [0, 5] and found the order statistics of the sample. If the failure time was between two observations, U and V, then the censoring interval was taken to

be [U, V). If it was below the lowest observation U, then the interval was [0, U), and if it was greater than the highest observation V, the censoring interval was taken to be $[V, \infty)$.

While the theory behind many of the methods described above require the censoring time to be drawn from a continuous distribution, we knew that this assumption was not practical in many real world applications. So, in addition to the uniform censoring, we also analyzed the data using a discrete censoring mechanism. To accomplish this, we assumed that the subject was supposed to be observed at 8 times (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4). For each subject at each time point we performed a Bernoulli trial with probability of success 0.5. If the trial was a success, we observed the individual at that time, and if it was a failure we did not observe the subject at that time. The censoring interval was then determined to be the observation before the failure time to the observation after the failure time. If the failure time was before the first observation, the left endpoint was taken to be 0, and if the failure time was after the last observation time, the right endpoint was taken to be ∞.

In all situations, we used Turnbull's self-consistency algorithm to estimate the survival function $S(t)$.

After collecting our data as described above, we applied the methods to analyze the data and decide whether or not the test will reject the null hypothesis $H_0 : S_1(t) = S_2(t)$ using the 0.05 level of significance. This was done for one thousand replications of each hazard/method combination. The empirical power or size of the test was defined as the proportion of times we rejected H_0.

2.7.3 Results

When deciding which method is best, we looked at two properties. The first is the power. The more power a method has the more desirable it is. The second property we are interested in is robustness. This is important because we will seldomly know exactly which type of hazard we have. A more robust method will give us more confidence in the results of a test especially when we know very little about the underlying hazard. In addition to comparing the methods to each other, we also needed a benchmark to make sure they performed well in general. In order to accomplish this, we compared them to the parametric score test.

Tables 1 and 2 give the power and sizes for the discrete censoring case. Here, when we do not know anything about the shapes of the hazard functions of the two survival functions, the Kolmogorov approach is the method we would choose. While the power for the Kolmogorov approach is only the best for hazards C and E, it is the most robust of all the methods we are considering. All of the other methods do not handle the crossing hazard cases very well. In the case of hazard F, for example, each method has a power below .10, with the Kolmogorov approach being the sole exception.

If we knew, however, that the proportional hazards assumption holds, then method 5 would be the clear choice. If we knew that there were an early difference or a late difference in the hazard functions, we would also choose method 5 since it would yield the highest power for these cases as well. However, if the hazards cross, we would want to use the Kolmogorov method since it has the most power of the methods we are considering when the hazards cross.

Tables 3 and 4 give the power and sizes for the case where the censoring times were drawn from the uniform distribution. In this case, with little or no information about the hazard functions, we would want to use method 2 with $\gamma = 1$ and $\rho = 1$. This method performs well under all of the hazard setups we proposed, and it performs the best under the situation where the hazards crossed or there was an early difference in the hazard functions.

Again, for the continuous hazards case, we would use a different method if we had more information about the underlying hazard functions. If the proportional hazards assumption holds, we would want to use method 2 with $\gamma = 0$ and $\rho = 0$. For many of the proportional-hazards cases we considered, this method has the highest power. In the few cases where it does not have the highest power, it is relatively close to the highest. If, however, we had the case where there was a late difference between the hazard functions, it appears that method 2 with $\gamma = 1$ and $\rho = 0$ would be the best choice. When the hazard functions cross, we would want to use the Kolmogorov approach. Although it does not yield the most power for both of our crossing hazard cases, it is very close to the highest power for Hazard E and it is the best for hazard C.

We can see from Tables 1 and 3 that the size of some of the tests may be different from the expected 0.05. For example, method 1 has a size of 0.064 for hazard S. In order to see if this was a problem, we conducted a test $H_0 : p = 0.05$ against $H_a : p \neq 0.05$. The p-value of this test was 0.59. So, at the 0.05 level, there is not sufficient evidence to conclude that the size is different from 0.05. Therefore, we can conclude that the

empirical size of these tests is ok.

If we look at our results, we can see that many of the methods we are considering compare favorably to the score test. For example, if we look at Table 2, the proportional hazards with discrete censoring time, we can see that the power of method 5, the method we chose as the best for this case, has at least 86% of the power that the score test had. For the non-proportional hazard case, the Kolmogorov approach had at least 55% of the power of the score test. When the censoring is continuous, the new methods did not perform quite as well. When the models follow a Cox model, method 2 with $\gamma = 0$ and $\rho = 0$ has at least 59% of the power of the score test. Finally, when the proportional hazards assumption does not hold, method 2 with $\gamma = 0$ and $\rho = 0$ has at least 37% of the power of the score test.

We also performed the simulation for the sample size n = 200 using the discrete censoring mechanism. The results of this are given in Tables 5 and 6. This confirms the results we had seen for the smaller sample size study. When we do not have any information about the underlying hazard function, we would still wish to use the Kolmogorov approach. It gives a large power for all of the hazard functions we considered, especially in hazard E where many of the other tests failed to give adequate power. If we look at Hazard E, for example, we can see that most Methods still had a power below .10. The Kolmogorov approach and Method 2 with $\gamma = 1$ and $\rho = 1$ both had power greater than .10, but the power of the Kolmogorov approach was more than 1.6 times the power of method 2. When we know the proportional hazard assumptions hold, method 5 would be the best choice. It had the most power of the methods we

considered. So, the results for our study still hold when the sample size was changed, but as we expected, the power for all of the tests increased when we doubled the sample size for each method for each hazard.

2.8 An Illustrative Example

We applied all of the methods we were testing to the data from the breast cosmesis study given in Section 1.1. Recall that Finklestein and Wolf (1985) presented a set of interval-censored data from a retrospective study to compare two different treatments for breast cancer. The goal was to compare early breast cancer patients who were treated with primary radiation therapy and adjuvant chemotherapy to women who were treated with radiation therapy alone. The variable of interest was the time until the cosmetic deterioration occurred, which was determined by the appearance of breast retraction.

We used Turnbull's self-consistency algorithm to estimate the survival curves of the two groups. A plot of the estimated survival functions is given in Figure 1. The p-value for each method was computed, and they are given in Table 7.

Method 1 yielded a p-value of .0052, which tells us that there is sufficient evidence to conclude that women who received primary radiation therapy and adjuvant chemotherapy experienced cosmetic deterioration later than women treated with radiation therapy alone.

Method 2 with $\gamma = 0$ and $\rho = 0$ yielded a p-value of .007. When $\gamma = 1$ and $\rho = 0$ the p-value was .002, and when $\gamma = 1$ and $\rho = 1$ the p-value was .0004. So, for all

three cases at $\alpha = 0.05$, we would reject H_0 and conclude that the time to cosmetic deterioration was different for the two groups.

The log-rank test yielded a p-value of .0057, so the results from this test agree with the results from method 1 and method 2. The p-value from method 5 was .0007, so this also agrees with method 1 and method 5's results.

However, the Kolmogorov approach gave a p-value of 0.14. So, at $\alpha = 0.05$, we do not have sufficient evidence to conclude that the rates of cosmetic deterioration differed for the two groups. So, the Kolmogorov approach differed from the other four methods we are considering.

So, why did the Kolmogorov approach differ from the others? If we look at Figure 1, we can see that the difference between the two survival curves is very small when $t < 20$, but when $t \geq 20$ the difference is very large. So, it appears we have data that displayers a late difference in their survival functions. If we look at Table 3, we can see that the Kolmogorov approach has power = 0.080 for hazard B, which corresponds to a late difference in the hazards. So, it is probably not a very reliable method in the circumstances of this example.

2.9 Discussion

This chapter discussed nonparametric comparison of two survival functions when interval-censored failure time data are available. We looked at five methods that have been developed in recent years. In order to decide which method was the best, we considered them under numerous hazard function setups including proportional-hazard

models as well as nonproportional-hazard models. For each setup, we also considered the methods when the failure time data was drawn from a continuous distribution or from a set of discrete time points.

To carry out the comparison, we ran a large simulation study and found the empirical power and size for each hazard setup. We showed that when the censoring times are discrete and nothing is known about the underlying hazard functions, the Kolmogorov approach was the best choice, and method 2 with $\gamma = 1$, and $\rho = 1$ is the best choice when the censoring was continuous. If the proportional hazards assumptions hold, method 5 was the best method for discrete censoring and method 2 with $\gamma = 0$, and $\rho = 0$ was best for the case where the censoring was continuous.

Following the comparison, we illustrated our results using the breast cosmesis data set. The results for four of the models agreed with Finklestein and Wolf's (1985) conclusions: the time to cosmetic deterioration was different for the two groups. The Kolmogorov approach was unable to detect a difference between since it has a very low power for the late difference situation.

Chapter 3

A Weighted Generalized Log-Rank Test for Interval Censored Failure Time

3.1 Introduction

In Zhao and Sun (2004) the authors developed a GLRT that reduces to the usual log-rank test. This test performs well, but it may not be sensitive to early and/or late differences in the survival functions.

To detect such early and late differences in the survival functions, a test statistic should place greater emphasis on the early or late differences in the survival function. When right-censored data are present, many researchers utilize a weighted log-rank test. In this chapter, we will develop a weighted generalized log-rank test (WGLRT) that is sensitive to either early or late differences in the survival function and is applicable to interval-censored data.

In section 3.2 we will present a test procedure we can use to test to see if two populations have the same survival time; in section 3.3, we present a large simulation

study to examine the empirical properties of the model; in section 3.4 we analyze a real world example; and in section 3.5 we offer some concluding remarks.

3.2 Methods

Consider a survival study involving k treatment groups. Let T denote the time of the event of interest, $S(t) = P(T \geq t) = 1 - F(t)$, the survival function, $\lambda(t) = -\frac{d}{dt} log S(t)$, the hazard function of the survival time, and $\Lambda(t) = \int_0^t \lambda(u)du$ the cumulative hazard function. Then $S_j(t)$, $\lambda_j(t)$, and $\Lambda_j(t)$ are the survival, hazard, and cumulative hazard functions, respectively, for subjects in treatment group $j = 1, 2, \ldots, k$. Let T_i denote the event time of interest for subject i, n_i the number of subjects in treatment group i, and $n = n_1 + n_2 + \ldots + n_k$, the total number of subjects. Also let $\hat{S}(t)$ denote the maximum likelihood estimator (MLE) of $S(t)$ and $\hat{S}_j(t)$ the MLE of $S_j(t)$. Let $t_1 < t_2 < \ldots < t_n$ denote the ordered points of the smallest subset of $\{L_i, U_i : 1 = 1 \ldots n\}$. Our problem focuses on the problem of testing the hypothesis that the survival functions for each of the groups is the same, $H_0 : S_1(t) = S_2(t) = \ldots = S_k(t)$ for all t.

When a data set is right censored, a popular nonparametric test is the weighted log-rank test given in Kalbfleisch and Prentice (2002)

$$\sum_{j=1}^{k} w(t_j) \left(d_{lj} - \frac{n_{lj} d_j}{n_j} \right),$$

where n_{lj} is the number of subjects at risk in population l at time t_j, d_{lj} is the number of failures in population l at time t_j, n_j is the total number of subjects still at risk at

time s_j, d_j is the total number of failures at time t_j, and w is a weight function.

However, when the data is interval-censored we do not know the exact number of subjects still at risk nor the number of subjects who experience the event of interest at a given time point. Zhao and Sun's (2004) proposed method for estimating these quantities was given in Chapter 2. In order to detect early and/or late differences in the survival function, Wu and Gilbert (2002) proposed the weight function

$$w_1(t) = \left[\hat{S}(t^-) - (a\hat{S}(\tau^-) + 1 - a)\right]^2, \ a \in [0, 1]$$

for the weighted log-rank statistic in the presence of right-censored data. When $a = 0.5$ this weight function will weight early and late differences equally. As a approaches 0, the function will put more weight on late differences and less on early differences, and as it approaches 1 it will put more weight on early differences in the survival functions and less emphasis on late differences. Another common weight function that is sensitive to early and late differences is

$$w_2(t) = \left[\hat{S}(t)\right]^b \left[1 - \hat{S}(t)\right]^{1-b}.$$

When $b = 0.5$ this weight function will weigh early and late differences equally. As b approaches 0, the function will put more weight on late differences and less on early differences, and as it approaches 1 it will put more weight on early differences in the survival functions and less emphasis on late differences.

We propose the following test statistic using the weight functions above:

$$U = \sum_{j=1}^{k} w(t_j) \left[\frac{n'_{1j} n'_{2j}}{n'_{1j} + n'_{2j}} \right] \left[\frac{d'_{1j}}{n'_{1j}} - \frac{d'_{2j}}{n'_{2j}} \right].$$

Now, as in Zhao and Sun (2004), we need to estimate the covariance matrix. To do this, we employed the same 3-step bootstrap procedure. Let M be a pre-specified number of resamplings for the bootstrap. Then for each r in $1, \ldots, M$, Step 1: If $\delta_i = 0$, then let T_i^r and $\delta_i^r = 0$. Otherwise, $\delta_i^r = 1$ and T_i^r is a realization from the conditional survival function

$$f_i(t) = P(T_i^r = t) = (\hat{S}(t) - \hat{S}(t+))/(\hat{S}(L_i) - \hat{S}(R_i+)), \;\; s \in [L_i, R_i].$$

Since δ_i is the right censoring indicator, the δ_i^r are going to be the same for every r.

Step 2: Use the new data to find the number of failures and risks, and estimate U as before, then compute the covariance estimates $\hat{V}^r = \hat{V}_1^r + \hat{V}_m^r$ where

$$(\hat{V}_j^r)_{ll} = (w^j)^2 \frac{n_{jl}^r (n_j^r - n_{jl}^r) d_j^r (n_j^r - d_j^r)}{(n_j^r)^2 (n_j^r - 1)} \;\;,$$

$$(\hat{V}_j^r)_{l_1 l_2} = (w^j)^2 \frac{n_{jl_1}^r n_{jl_2}^r d_j^r (n_j^r - d_j^r)}{(n_j^r)^2 (n_j^r - 1)} \;\;,$$

and w^j is the appropriate weight function at time t_j.

Step 3: Repeat steps 1 and 2 for $r = 1, \ldots, M$ the number of resamplings, and estimate V by $\hat{V} = \hat{V}_1 + \hat{V}_2$, where

$$\hat{V}_1 = \frac{1}{M}\sum_{r=1}^{M}\hat{V}^r \ ,$$

$$\hat{V}_2 = (1+\frac{1}{M})\frac{\sum_{r=1}^{M}[U^r-\bar{U}][U^r-\bar{U}]^t}{(M-1)} \ ,$$

and

$$\bar{U} = \sum_{r=1}^{M} U^r/M \ .$$

The test of hypothesis can be carried out using the statistic $U^* = U^{'}V^{-}U$ which follows a χ^2_{k-1} distribution under the null hypothesis.

3.3 A Numerical Study

A large scale simulation study was conducted to assess the properties of the proposed WGLRT. We looked at the proposed test under several different hazard function setups: piecewise exponential functions, with an early difference between the survival functions, a late difference between the survival functions, and a situation where the hazard functions cross. Specifically, we considered setups with the following hazard functions.

Early difference: Set-up A

$$\lambda_1(t) = 0.25I_{t\leq 0.75} + I_{t>0.75}$$

$$\lambda_2(t) = 0.75I_{t\leq 0.75} + I_{t>0.75}$$

Late Difference: Set-up B

$$\lambda_1(t) = I_{t\leq 0.75} + I_{t>0.75}$$

$$\lambda_2(t) = I_{t\leq 0.75} + 2I_{t>0.75}$$

Crossing Hazards: Set-up C

$$\lambda_1(t) = 0.5I_{t\leq 0.75} + 1.5I_{0.75<t\leq 1.5} + I_{t>1.5}$$

$$\lambda_2(t) = 1.5I_{t\leq 0.75} + 0.5I_{0.75<t\leq 1.5} + I_{t>1.5}$$

In addition to the piecewise exponential, We studied the linear hazard functions given by $\lambda(t) = \alpha_0 + \alpha_1 t$. These hazard functions can give parallel hazard functions, crossing hazard functions, and neither parallel nor crossing hazard functions, respectively.

Proportional Hazards: Set-up D

$$\lambda_1(t) = 1$$

$$\lambda_2(t) = 2$$

Crossing Hazards: Set-up E

$$\lambda_1(t) = 1$$

$$\lambda_2(t) = 0.2 + t$$

Set-up F:

$$\lambda_1(t) = 1$$

$$\lambda_2(t) = 1.2 + .3t$$

We also analyzed several piecewise linear hazard functions.

Early Difference: Set-up G

$$\lambda_1(t) = (t+1)I_{t\leq 1} + 2I_{t>1}$$

$$\lambda_2(t) = 0.5I_{t\leq 1} + 2I_{t>1}$$

Late Difference: Set-up H

$$\lambda_1(t) = 1I_{t\leq 2} + 1I_{t>2}$$

$$\lambda_2(t) = 1I_{t\leq 2} + (t-1)I_{t>2}$$

Early Difference: Set-up I

$$\lambda_1(t) = (2-t)I_{t\leq 1} + 1I_{t>1}$$

$$\lambda_2(t) = tI_{t\leq 1} + 1I_{t>1}$$

Late Difference: Set-up J

$$\lambda_1(t) = 1I_{t\leq 2} + \frac{t}{2}I_{t>2}$$

$$\lambda_2(t) = 1I_{t\leq 2} + \frac{1}{t-1}I_{t>2}$$

Crossing Hazards: Set-up K

$$\lambda_1(t) = (1.2 - t)I_{t\leq 1} + 0.2I_{t>1}$$

$$\lambda_2(t) = (0.2 + t)I_{t\leq 1} + 1.2I_{t>1}$$

Both Early and Late Differences: Setup L

$$\lambda_1(t) = 1$$

$$\lambda_2(t) = 0.5I_{t\leq 1} + 1I_{1<t\leq 2} + 0.5I_{t>2}$$

We also considered two set-ups to check the size of the tests.

Set-up S

$$\lambda_1(t) = \lambda_2(t) = 1$$

Set-up W

$$\lambda_1(t) = \lambda_2(t) = 2t$$

3.3.1 Generation of Interval-Censored Data

To examine the small sample properties of the proposed statistics, we performed extensive simulation studies. We drew a sample of fifty subjects from populations whose survival functions were described above. Each subject was supposed to be observed at 8 times (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4). For each subject at each time point we performed a Bernoulli trial with probability of success 0.5. If the trial was a success, we observed the individual at that time, and if it was a failure we did not observe the subject at that time. The censoring interval was then determined to be the observation before the failure time to the observation after the failure time. If the failure time was before the first observation, the left endpoint was taken to be 0, and if the failure time was after the last observation time, the right endpoint was taken to be ∞.

In all situations, we used Turnbull's self-consistency algorithm (Turnbull 1976) to estimate the survival function $S(t)$. We applied the WGLRT to the data and decided whether or not the test will reject the null hypothesis $H_0 : S_1(t) = S_2(t)$ using the 0.05 level of significance. This was done for one thousand replications of each set-up. The empirical power or size of the test were defined as the proportion of times we rejected H_0.

3.3.2 Numerical Results

Table 8 and Table 9 gives the power and size of the proposed test statistic using $w_1(t)$, Table 10 and Table 11 gives the power and size of the proposed statistic using $w_2(t)$, and Table 12 gives the size and power using the unweighted test statistic. All

of the tests give the correct approximate size. The WGLRT performs better in most situations than the unweighted GLRT. The only exception to this is when there are proportional hazards. This is to be expected however since there is an equal difference between the hazard functions for all times, a weight of 1 will weight all times the same, so the unweighted test is superior under the proportional hazards assumption.

The weight function $w_1(t)$ performed best when the hazard functions cross each other and when there was an early difference in the hazard functions. $w_2(t)$ performed best when there was a late difference in the hazard functions, and it performed better under the proportional hazards setup than $w_1(t)$.

In practice we rarely know what type of situation we have. So, in addition to the powers found earlier, we found the average power given by each test. We found that the weight function $w_1(t)$ had an average power of 0.360 while the weight function $w_2(t)$ had an average power of 0.386. So, while both tests perform well, if we did not know anything about the underlying hazard functions, w_2 is the best among the proposed statistics.

3.4 An Example

Goggins and Finklestein (2000) discussed a data set arising from an AIDS clinical trial. During the trial, blood and urine samples were collected from patients and tested for the presence of cytomegalovirus (CMV). We applied the methods presented in the previous section to compare the urine shedding times of CMV for patients infected with HIV. 91 patients with high CD4 counts ($CD4 > 75 cells/\mu l$) and 121 patients

with low CD4 count were measured. The time to shedding the virus is of interest. Since HIV is a progressive disease and CMV is an opportunistic infection caused by the weakened immune system, the difference between the hazard functions for the patients are expected to be large for early time periods since those with a low cell count will be more susceptible to infection.

The hazard functions for both the low CD4 count and the high CD4 count were estimated using the Gaussian kernel based method presented in Sun (2006) and are given in Figure 2. At low times, $t < 5$, we observe a large difference in the two groups. The hazard rate for the low CD4 group is much larger than that for the high CD4 group.

Table 13 gives the test statistics and p-value for testing whether there is a difference between the time to urine shedding of CMV in patients with a low CD4 cell count and those with a non low CD4 count. Since we observed an early difference in the hazard functions, we will focus on the results for $a = 1.0$ since this is particularly sensitive to an early difference. We can see that using $w_1(t)$ we have a test statistic of -6.151 and a p-value below 0.001. Using $w_2(t)$ we got a test statistic of -7.14 and a p-value below 0.001. In both cases, we reject the null hypothesis and conclude that the survival curves are different. From Figure 3, we can tell that the survival time for the low CD4 group is significantly lower than that for the non low CD4 count group.

3.5 Discussion

Early and late differences occur often in many clinical trial studies. We have proposed a modified weighted log-rank test that offers significant increases in power over previously proposed statistics. When there is a good reason to expect early or late differences in the survival curve, the researcher can utilize an appropriate value for a. We may also plot the estimated hazard functions and survival curves to determine if there may be an early or late difference, or crossing hazard functions and the researcher can use an appropriate value of a to see if the difference is statistically significant.

In addition to testing whether there is simply a difference in the survival curves, it may be of interest to determine where the difference lies. Which group has a lower survival rate? Is it an early difference, a late difference, both, neither? This is important for interpreting the results in terms of a real-world problem. Wu (2002) suggests that this can be carried out by a secondary analysis. He utilizes a linear combination of test statistics, and proposes that the secondary analysis can be used to interpret differences that may be observed among the results of the tests.

Chapter 4

Sample Size Calculation for Interval-Censored Failure Time Data

4.1 Introduction

In many medical studies, it can be costly to follow patients for an extended period of time. Many researchers would like to ensure they can achieve a desired level of power for a test of $H_0 : S_1(t) = S_2(t)$, while keeping costs low. This involves making sure enough patients are recruited for the study but not too many. Because of this, a power analysis is often run before a study begins or after a brief pilot study.

There have been many methods developed to find an appropriate sample size. Schoenfeld (1981) presented methods for computing the asymptotic mean and variance of a generalized version of the log-rank statistic, the modified Wiocoxson statistic, and many other commonly used methods for comparing the survival curve. Using these results, he presented a sample size calculation formula for the log-rank statistic.

Schoenfeld's methods are good under a specific set of assumptions. Lakatos (1988)

developed a more general method for computing the sample size. As in Schoenfeld's paper, Lakatos computed the sample size required to compare two survival functions using the log-rank test. However, Lakatos's calculation removed some of the more restrictive assumptions from the earlier calculations. His method does not require the proportional hazards assumption. It also has the benefit that allows for some more common clinical trial designs such as a trial with staggered entry or stratification.

Lakatos's method is general and easy enough to use such that major software developers such as PASS use it for their power analysis programs. However, the major drawback to this method and others that have been developed is they are only appropriate for computing the sample size when right-censored data are present. There are not many methods for the case when the data are interval-censored. In this chapter, we develop an easy-to-use method for computing the sample size for the interval-censored data case.

4.2 Sample Size Calculation

Assume that we are studying two groups of patients ($j = 1, 2$) and observe each patient at l fixed time points, $0 = t_0, t_1, \ldots, t_l$. Also assume that each group follows an exponential distribution with hazard functions λ_1 and λ_2 respectively. Let $\delta_{i,k}$ be the indicator that subject i fails after t_{k-1} but before t_k. Then, $\Delta_{i,j} = (\delta_{i,1}, \delta_{i,2}, \ldots \delta_{i,l})'$ follows a multinomial distribution with probabilities $\underline{\theta_j} = (\theta_{j,1}, \theta_{j,2}, \ldots, \theta_{j,l})$, where

$$\theta_{j,k} = P(t_{k-1} \leq T \leq t_k) = e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j}$$

Since the j^{th} group follows an exponential distribution with hazard λ_j we have $S_j(t) = e^{-t\lambda_j}$ as the survival function for group j and the likelihood function

$$\mathscr{L}_j = \prod_{i=1}^{n_j} \sum_{k=1}^{l} \delta_{i,k} \left[S_j(t_{k-1}) - S_j(t_k) \right] = \prod_{i=1}^{n_j} \sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right].$$

Then, the log-likelihood function is

$$\eta_j = log(\mathscr{L}_j) = \sum_{i=1}^{n_j} log \left(\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right] \right)$$

with first derivative

$$\frac{\partial \eta_j}{\partial \lambda_j} = \sum_{i=1}^{n_j} \frac{\sum_{k=1}^{l} \delta_{i,k} \left[-t_{k-1} e^{-t_{k-1}\lambda_j} + t_k e^{-t_k\lambda_j} \right]}{\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right]}$$

and second derivative

$$\frac{\partial^2 \eta_j}{\partial \lambda_j^2} = \sum_{i=1}^{n_j} \frac{\left(\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right] \right) \left(\sum_{k=1}^{l} \delta_{i,k} \left[t_{k-1}^2 e^{-t_{k-1}\lambda_j} - t_k^2 e^{-t_k\lambda_j} \right] \right)}{\left(\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right] \right)^2} -$$

$$- \frac{\left(\sum_{k=1}^{l} \delta_{i,k} \left[-t_{k-1} e^{-t_{k-1}\lambda_j} + t_k e^{-t_k\lambda_j} \right]^2 \right)}{\left(\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\lambda_j} - e^{-t_k\lambda_j} \right] \right)^2}$$

We want to find the maximum likelihood estimator (MLE) for λ, $\hat{\lambda}$, so we set

$$\left(\frac{\partial \eta_j}{\partial \lambda_j} \right) |_{\lambda = \hat{\lambda}} = 0$$

which implies

$$\sum_{i=1}^{n_j} \frac{\sum_{k=1}^{l} \delta_{i,k} \left[-t_{k-1} e^{-t_{k-1}\hat{\lambda}_j} + t_k e^{-t_k \hat{\lambda}_j} \right]}{\sum_{k=1}^{l} \delta_{i,k} \left[e^{-t_{k-1}\hat{\lambda}_j} - e^{-t_k \hat{\lambda}_j} \right]} = 0$$

However, there is no analytical solution to this, so we must use a numerical algorithm such as the Newton-Raphson algorithm to find a solution to this problem.

Once we find the MLE, we need to know its distribution. We know that asymptotically $\hat{\lambda}_j \sim N(\lambda_j, I_x^F(\lambda_j)^{-1})$, where $I_x^F(\lambda_j) = -E\left(\frac{\partial^2 \eta_j}{\partial \lambda_j^2}\right)$. Then let

$$g_{j,k} = \left(e^{-t_{k-1}\lambda_j} - e^{-t_k \lambda_j}\right)\left(t_{k-1}^2 e^{-t_{k-1}\lambda_j} - t_k^2 e^{-t_k \lambda_j}\right) - \left(-t_{k-1} e^{-t_{k-1}\lambda_j} + t_k e^{-t_k \lambda_j}\right)^2$$

and

$$\frac{\partial^2 \eta_j}{\partial \lambda_j^2} = \sum_{i=1}^{n_j} \frac{\sum_{k=1}^{l} \delta_{i,k} g_{j,k}}{\sum_{k=1}^{l} \delta_{i,k} \theta_{j,k}^2} = \sum_{i=1}^{n_j} \sum_{k=1}^{l} \delta_{j,k} \frac{g_{i,k}}{\theta_{j,k}^2}.$$

So, we have

$$E\left(\frac{\partial^2 \eta_j}{\partial \lambda_j^2}\right) = E\left(\sum_{i=1}^{n_j} \sum_{k=1}^{l} \delta_{j,k} \frac{g_{i,k}}{\theta_{j,k}}\right)$$

but both $g_{j,k}$ and $\theta_{j,k}$ are constants, so

$$E\left(\frac{\partial^2 \eta_j}{\partial \lambda_j^2}\right) = \sum_{i=1}^{n_j} \sum_{k=1}^{l} \frac{g_{i,k}}{\theta_{j,k}^2} E(\delta_{i,k})$$

and we know that $E(\delta_{i,k})$ is $\theta_{j,k}$. Therefore,

$$E\left(\frac{\partial^2 \eta_j}{\partial \lambda_j^2}\right) = \sum_{i=1}^{n_j} \sum_{k=1}^{l} \frac{g_{i,k}}{\theta_{j,k}} = n_j \sum_{k=1}^{l} \frac{g_{i,k}}{\theta_{j,k}}.$$

Thus,

$$I_x^F(\hat{\lambda}_j) = -n_j \sum_{k=1}^{l} \frac{g_{i,k}}{\theta_{j,k}}$$

and the asymptotic variance of $\hat{\lambda}_j$ is

$$Var(\hat{\lambda}_j) = \frac{-1}{n_j} \left(\sum_{k=1}^{l} \frac{g_{i,k}}{\theta_{j,k}} \right)^{-1}.$$

Now, we would like to test H_0: $\lambda_1 = \lambda_2$ versus H_a: $\lambda_1 > \lambda_2$. Under the null hypothesis the variable $\hat{D} = \hat{\lambda}_1 - \hat{\lambda}_2$ has an asymptotic normal distribution with mean 0 and variance $V(\hat{\lambda}_1) + V(\hat{\lambda}_2)$. Call this variance s^2. We will reject H_0 in favor of H_a if $\hat{\lambda}_1 - \hat{\lambda}_2$ is large. So, under H_0

$$P(\hat{\lambda}_1 - \hat{\lambda}_2 > c) = P(\hat{D} > c) = \alpha$$

which implies

$$P(Z > c/s) = \alpha$$

and

$$\Phi\left(\frac{c}{s}\right) = 1 - \alpha$$

where Z is a standard normal random variable and Φ is the the CDF of a standard normal random variable. Then we know

$$\frac{c}{s} = Z_\alpha$$

$$c = Z_\alpha s$$

Now, if we would like to have at least power $\pi = 1 - \beta$ while being able to detect a difference $\lambda_1 - \lambda_2 = D_a$ then under the alternative we have

$$P(\hat{D} > c) = 1 - \beta$$

which implies

$$P\left(\frac{\hat{D} - D_a}{s} > \frac{c - D_a}{s}\right) = P\left(Z > \frac{c - D_a}{s}\right) = 1 - \beta$$

and

$$\Phi\left(\frac{c - D_a}{s}\right) = \beta$$

so

$$\frac{c - D_a}{s} = Z_{1-\beta} = -Z_\beta.$$

Plugging in the previous result we get

$$\frac{Z_\alpha s - D_a}{s} = -Z_\beta$$

which implies

$$s = \frac{D_a}{Z_\alpha + Z_\beta}$$

and

$$s^2 = \left(\frac{D_a}{Z_\alpha + Z_\beta}\right)^2.$$

Now, if we plug the variance into this equation

$$-\frac{1}{n_1}\left(\sum_{k=1}^{l}\frac{g_{1,k}}{\theta_{1,k}}\right)^{-1} - \frac{1}{n_2}\left(\sum_{k=1}^{l}\frac{g_{2,k}}{\theta_{2,k}}\right)^{-1} = \left(\frac{D_a}{Z_\alpha + Z_\beta}\right)^2$$

and if we assume that $n_1 = n_2 = n/2$ then we get

$$-\frac{2}{n}\left(\sum_{k=1}^{l}\frac{g_{1,k}}{\theta_{1,k}}\right)^{-1} - \frac{2}{n}\left(\sum_{k=1}^{l}\frac{g_{2,k}}{\theta_{2,k}}\right)^{-1} = \left(\frac{D_a}{Z_\alpha + Z_\beta}\right)^2$$

or

$$n = -2\left(\frac{Z_\alpha + Z_\beta}{D_a}\right)^2\left[\left(\sum_{k=1}^{l}\frac{g_{1,k}}{\theta_{1,k}}\right)^{-1} + \left(\sum_{k=1}^{l}\frac{g_{2,k}}{\theta_{2,k}}\right)^{-1}\right].$$

4.3 A Simulation Study

A large scale simulation study was conducted to ensure that the sample size given in the previous section provides adequate power. In addition, the size of the test must be assessed to ensure that it is accurate as well. For the simulation, we let λ_1 take values from 0.5 to 2.0 with an increasing increment of 0.1, and λ_2 take values from 0.5 to λ_1 with an increasing increment of 0.1. For all of the simulations, we let α be 0.05 and we let β be from $\{0.05, 0.1, 0.2\}$. The follow-up for the study was a period of 5 years observing each individual every six months.

In each case we took a small pilot sample of 20 individuals, 10 in each group, and

computed the required sample size based on the MLE found from the pilot study. We then drew an appropriate sized sample and performed the test to see if there was a significant difference between the hazards. For each hazard combination, we repeated this test 1000 times. The empirical size or power of the test is the proportion of the tests we rejected.

Table 15 gives the power and size of the proposed test statistic using $\beta = 0.2$, Table 16 gives the power and size of the proposed statistic using $\beta = 0.1$, and Table 17 gives the size and power using $\beta = 0.05$. We can see that in each case where $\lambda_1 \neq \lambda_2$ the power is not significantly different from the true power. When $\lambda_1 = \lambda_2$ the size is not significantly different from the true size, 0.05.

Table 18 gives the sample size required to achieve a power of $\beta = 0.05$, Table 19 gives the sample size required to achieve a power of $\beta = 0.1$, and Table 20 gives the sample size required to achieve a power of $\beta = 0.2$. When λ_1 and λ_2 are close, the required sample size can be very large. If we look at Figure 4 and Figure 5, we can see when the hazards are close to each other, the survival functions are also very close to each other. The sample size required to be able to detect this difference will be quite large, and we can see that reflected in these tables. However, when the hazards are far apart, the difference in the survival function is also large. So, a much smaller sample can be taken to achieve the same power.

Chapter 5

A Comparison of the Imputed Kaplan-Meier Estimate and The Self-Consistency Estimate for Interval-Censored Failure Time Data

5.1 Introduction

In some cases researchers are unfamiliar with proper techniques needed to analyze interval-censored data, but they are comfortable using techniques developed for right-censored data. When this situation arises, the researcher may use a missing data technique such as imputation in order to use a method appropriate for right-censored data.

Sun (2006) described several possible ways to perform the imputation for interval-censored failure time data. Possibly the simplest and most commonly used method is to simply use the mid-point, the right endpoint, or the left endpoint of the interval as the failure time. Sun pointed out that if the intervals are narrow then the estimate of

the survival function using the mid-point, left endpoint, or right endpoint imputation will be similar. In studies with long follow-up periods or studies where it is not possible to have narrow intervals, the left-imputed estimate and the right-imputed estimate may be quite different. Additionally, in studies where the sample size is small, even if the intervals are narrow, the estimates may be significantly different.

In order to investigate whether this method of imputation is reasonable for analyzing interval-censored data, we compare the right endpoint and left endpoint imputed estimates of the survival function to the self consistency estimate of the survival function. As suggested in Sun (2006), with the right endpoint and left endpoint imputation, we used the Kaplan-Meier (KM) estimate of the survival curve

$$\hat{S}(t) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right).$$

If we use the left endpoint in the imputation, we are assuming the patient dies at the earliest possible time in the interval. This means that the estimated survival function using this imputation should be a lower bound. Similarly, if we use the right endpoint we assume that they die at the latest possible time in the interval, so this should provide an upper bound for the estimated survival function. It would be reasonable then that any estimated survival function should fall between the left-imputed and the right-imputed estimates.

Suppose, for example, that we are interested in the time at which a laboratory rat develops a tumor. The survival function using the left-imputed KM estimate as well as

the self consistency estimate are given in Figure 6. We can see that at 14 weeks, the self consistent method drops significantly below the left-imputed KM estimate. For this case, if we had used the imputed KM estimate we would be significantly overestimating the survival probability for the rats using the imputed KM estimate.

In this chapter, we are primarily interested in how the right-imputed KM and left-imputed KM estimates compared to the self consistent estimate of the survival curve, and if the self consistent estimate crosses over either of these estimates. If the self consistent estimate crosses one of the imputed KM estimates, then a research study that uses an imputed method may be underestimating or overestimating the survival probability.

If the imputed KM estimate crosses the self consistent estimate of the survival function, we would like to investigate what is causing them to disagree. To study this, we looked at the probability of the self consistent estimate crossing over either the left-imputed KM estimate or the right-imputed KM estimate. Do different hazard functions lead to a different probability of the imputed KM estimates crossing over the self consistent estimate? If so, what causes this difference? We will look at the following possibilities:

1. Does the probability of these estimates crossing depend on the average length of time between the observations?

2. Does the probability of these estimates crossing depend on the percentage of observations that are right-censored?

3. Does the probability of these estimates crossing depend on the mean survival time?

4. Does the probability of these estimates crossing depend on the number of patients in the study?

5. Does the probability of these estimates crossing depend on the proportion of visits a patient misses over the course of the study?

5.2 A Numerical Study

We ran a large scale simulation to see how often the self-consistency estimate for the survival function crosses either the left-imputed KM estimate or the right-imputed KM estimate under situations with a wide variety of hazard functions.

We used all of the hazard setups A through L from Chapter 3. Since each hazard setup has two hazard functions, this gave a total of 24 hazards. We sampled 1000 data sets for each hazard using several different interval lengths. The study period was over the interval $[0, 4]$. For each setup the patients were scheduled to be observed $4, 8, 10$, or 20 times with observations spaced evenly throughout the study period. At each scheduled observation time, we assumed the patient missed the observation with probability 0.5. The censoring interval was then determined to be the observation before the failure time to the observation after the failure time. If the failure time was before the first observation, the left endpoint was taken to be 0, and if the failure time was after the last observation time, the right endpoint was taken to be ∞.

Since we were interested in determining the probability that the self-consistency estimate for the survival curve crossed the right-imputed or left-imputed KM estimate of the survival curve, we looked at the proportion of the time these estimates crossed. The results of this study can be found in Table 21 and Table 23. We can see that there was very little difference between the $4, 8, 10$, and 20 observations intervals. However, there was a large difference amongst the different hazard functions.

In order to investigate what was causing this difference, we asked, "does the percent of right censored observations have an effect on the proportion of times self-consistency estimate crosses one of the imputed KM estimates?" To answer this question, we reran the simulation. We ran them with an exponential distribution with mean 1, and we altered the follow-up period to adjust the probability of observations that would be right-censored. These probabilities were from $0.10, 0.25, 0.50, 0.75$, and 0.90.

These results can be found in Table 25. From this table, we can see that there are no large differences in the proportion of times the estimates cross.

It was also of interest to determine if the average survival time had an effect on the proportion of times the self-consistency estimate crosses one of the imputed KM estimates. To explore this, we ran the original simulation using several exponential distributions with means of $0.5, 1, 2, 3, 5$, and 6. The results of this simulation can be found in Table 27.

We can see that as the mean increases so does the crossing proportion. To determine what is causing this, we can look at Figure 8 and Figure 11. We can see that for the mean 5 group, the survival functions cross each other at time point 3.4. This late

crossing is common when the mean survival time is large relative to the study period. The average time where the hazards cross for the mean 5 group was 3.62. This late crossing occurs more often when there are many patients still in the study at large time points.

Tables 22 and 24 give the average time when the self-consistency estimate first crosses the KM estimate. We can see from the table that this often occurs at later time points. This may be due to the fact that if the maximum time where the imputed KM estimate jumps is greater than the maximum time for the self-consistency estimate, the KM estimate will continue it's estimate to infinity and vice versa. This can lead to situations where the self-consistency estimate crosses the Kaplan-Meier after the last time where a jump occurs in the KM estimate.

In order to determine if this is causing a problem, we ran another simulation using the exponential hazards with the same means as before, but we ignored later time points. If the self-consistency estimate did not cross the imputed KM estimates before $t = 3$, we said they didn't cross. The results of this are given in Table 28. If we compare the results to those of of Table 27, we can see that there is a significant reduction in times the estimates crossed between the two simulations.

It seems reasonable that if we were to sample more patients that the left-imputed and right-imputed KM estimates would be closer to each other. Does the increase in sample size also mean that the self consistent estimate would cross the imputed KM estimates less often? To investigate this, we reran the simulation using the exponential hazards and sample sizes of 10, 100, and 1000 patients.

Table 29, Table 30, and Table 31 give the proportion of times the self-consistent estimate crosses the imputed KM estimates when we have a sample of 10, 100, and 1000 respectively. When the sample size dropped, studies with only four observation times did not provide enough information to predict the survival curve so it was dropped from these simulations.

We can see that the proportion of crossings seems to increase as the sample size increases. For the sample size 10 case, Table 29 shows that the proportion of times the estimated survival functions cross is below that for the sample size 50 case in table 29 and Tables 30 and 31 are both higher. In fact, Table 31 shows that they cross in almost every case. If we look at Figure 13 we can see that due to the large sample sizes, at large time points, the estimated survival functions are very similar. These crossings all occur at very late time points, with an average crossing time of 3.72.

Again, since the crossings are occurring due to the self consistent estimate jumping after the largest time point for the imputed KM estimate, we reran this simulation ignoring any crossings that occur after time 3. The results of these simulations can be found in Tables 32 and 33. We can see that the proportion of crossings has been reduced significantly. In fact for the sample size 1000 case, if the mean was greater than 1 the estimated curves never crossed. There was also a very significant reduction in crossing from our sample size 50 method of Table 28.

In addition to determining if the size of the sample had such a large impact on the proportion of times the estimates crossed, we also wanted to determine if the average length of the interval have an effect on the proportion of times the self consistent

estimate crossed the imputed KM estimate. In order to investigate this, we modified the probability that an individual was observed at each time point. We ran the simulation again using the exponential hazards, but instead of using a 0.5 probability of observing the patients, we allowed the probability to be from $\{0.2, 0.9, 1.0\}$. The higher the probability of observation, the narrower the average interval was.

Table 34 gives the proportion of times the self consistent estimate crossed the imputed KM estimate when the probability of observation was 0.2. Table 35 gives the proportion of times the self consistent estimate crossed the imputed KM estimate when the probability of observation was 0.9. Table 36 gives the proportion of times the self consistent estimate crossed the imputed KM estimate when the probability of observation was 1.0. When we compare these three tables, we find that as the probability of observation increases so does the proportion of times the estimates cross. Although this may seem counter intuitive at first, we are running into the same problem as before. The self consistent estimate changes after the last jump from the left-imputed KM estimate.

We reran these simulations again ignoring all of the crossings that occurred at late time points. The results for these can be found in Tables 37, 38, and 39. Once again, when we removed the late crossings, we had a significant reduction in the proportion of times the estimates crossed. We can see that with the wider intervals the self consistent estimate still crossed the imputed KM estimate fewer times. However, these proportions were not significantly different. So, the probability of observing the patients at each scheduled observation time did not have a significant effect on the proportion of

times the KM estimate and the self consistent estimate crossed.

5.3 Two Examples

In order to see the effects of using an imputed Kaplan-Meier estimate versus the self consistent estimate on a real-world data set, we applied these techniques to the AIDS study given in Goggins and Finklestein (2000) as well as the breast cosmesis study given by Finklestein and Wolf (1985).

A plot of the estimated survival curves for the radiation group from the Finklestein and Wolf (1985) data set is given in Figure 19. We can clearly see that the self-consistency estimate goes below the left-imputed KM estimate at 40 months. However, the last jump for the left-imputed KM estimate occurred at 35 months. This is a similar situation that arose during the simulation study. The estimated survival curves for the chemotherapy group from the Finklestein and Wolf (1985) data set are given in Figure 20. At 35 months, the self consistent estimate of the survival function dropped below the left-imputed KM estimate.

In this example, it would probably not be appropriate to use the imputed KM estimates because at large time points the imputed KM estimates are probably over estimating the probability of survival. For a situation such as this, the researchers should use an interval-censored technique.

Next, we considered the time to shedding in blood from the AIDS study given in Goggins and Finkelstein (2000). Figure 21 and Figure 22 are plots of the estimated survival curves for patients with a non-low and a low CD4 count respectively. For the

non-low CD4 group, we can see that the self consistent estimate does not cross either of the imputed KM estimates at any time. For the low CD4 group, the self consistent estimate only crossed the left-imputed KM estimate after the the left-imputed KM estimates's final jump.

In studies such as this one, it would be acceptable to use the imputed KM estimates. The researcher could take a conservative approach by using the left-imputed value knowing that they would be under estimating the patient's survival probability.

5.4 Conclusions

In this chapter, we analyzed cases where the self consistent estimate of the survival curve crossed the left-imputed or the right-imputed Kaplan-Meier estimates of the survival curve. We found that different hazards lead to different probabilities of these estimates crossing. We investigated 5 possible causes listed in Section 5.1 above and found:

1. The probability of these estimates crossing did not depend on the average length of time between the observations.

2. The probability of these estimates crossing did not depend on the percentage of observations that are right-censored.

3. The probability of these estimates crossing depended on the mean survival time. As the survival time increased, so did the proportion of times the self-consistent estimate crossed the imputed KM estimates.

4. The probability depended on the number of patients in the study. As the sample size increased, so did the proportion of times the self consistent estimate crossed the imputed KM estimated. This effect, however, was due to the fact that at late time points the self consistent estimate can jump after the last left-imputed time point. When we controlled for this effect, we discovered that a larger sample size does lead to a smaller proportion of times the estimated survival curves cross.

5. The probability of the estimates crossing depended on the proportion of visits a patient misses over the course of the study. As the probability of observation increased, so did the proportion of times the estimates crossed. As before, this effect was caused by the self consistent estimate jumping after the last jump from the left-imputed KM estimate. When we controlled for this, we found that the probability of the estimates crossing did not depend on the proportion of visits a patient missed over the course of the study.

It is clear that the Kaplan-Meier right and left imputed estimates of the survival function do not always contain the self consistent estimate. However, when the self consistent estimate crosses one of the KM estimates, it usually occurs at a late time point. Because the KM estimate is constant after the largest failure time, these late crossings are often a result of the self consistent estimator changing after the largest imputed time point. So, at late time points the imputed KM estimates can overestimate the probability of survival. Researchers who wish to use an imputed Kaplan-Meier estimate to perform an analysis of an interval-censored data set should be cautious of

these drawbacks. The effect is small in studies where few patients survive until the end, so if only a small number of patients, relative to the total number in the study, are remaining researchers can use an imputed method without fear of under estimating the survival rate by too much.

Chapter 6

Future Work

This chapter discusses several directions for future work. We will discuss an analysis of interval-censored data when the times we observe patients depends on the failure time. We will discuss a more general method for computing the sample size when interval-censored data are present. We will also propose some general guidelines for using an imputed Kaplan-Meier estimate for interval-censored data.

6.1 Analysis With Dependent Censoring

Consider a study of patients who suffer from HIV and we are interested in the time until these patients have AIDS. In many cases, if the patients become more ill they may go for treatment more often. However, if they get better their trips may become more infrequent or stop all together. It is clear that in such a situation, the times we observe the patients and the time when the patients get AIDS, the failure time, are related to each other. All of the methods we have considered here involve the assumption that

the censoring times and the time of the event of interest are independent of each other. In the future we would like to further generalize the GLRTs to allow for situations where these distributions are not independent of each other.

6.2 Sample Size Calculations

Lakatos (1988) pointed out that in many clinical trials the hazard functions for the patients are not a constant. This can occur when the treatment may increase or decrease in effectiveness over time. Even if they are constant, there can be many other difficulties such as staggered entry, patients dropping out of a study early, or patients who may miss several follow-up observations. The method for computing the required sample size presented in Chapter 4 makes many assumptions about the populations of interest. We would like to further generalize these calculations. Removing the assumption of proportional hazards is the first step, and we would like to remove several of the other assumptions as well.

6.3 Comparison of the Kaplan-Meier and Self Consistent Estimates

In Chapter 5, we presented some findings to compare the imputed Kaplan-Meier estimates of the survival function to the self consistency estimate of the survival function. We showed that under some circumstances using the imputed KM tests was appropriate while in others the researcher may under estimate or over estimate the

survival function. We would like to develop some general guidelines that can tell a researcher when using the imputed KM methods are appropriate and when they are not appropriate.

Chapter 7

References

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Fang, H., Sun J. and Lee, M. L. T. (2002). Nonparametric survival comparisons for interval-censored continuous data. *Statist. Sinica* **12**, 1073-83.

Fay, M. (1999). Comparing Several Score Tests for Interval-Censored Data. *Statistics in Medicine* **19** 273-285.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-54.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval censored failure time data. *Biometrics* **41**, 933-45.

Fischl, M. MD et al. (1990) The safety and efficacy of Zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (HIV) infection. *Annals of Internal Medicine* **112**, 727-737.

Gentleman, R. and Geyer, C. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika* **81**, 618-623.

Goggins, W. and Finkelstein, D. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56**, 940-943.

Groeneboom, P. and Wellner, J.A. (1992). Information bounds and nonparametric maximum likelihood estimation. DMV Seminar, Band 19, Birkhauser, New York.

Huang, J. and Lee, C. (). Covariance Estimation of a Log-rank Test for Interval-Censored Failure Time Data. *Statistics in Medicine* 1-14.

Kalbfleisch, J. and Prentice, R. (2002). The statistical analysis of failure time data. *Hoboken, NJ* John Wiley & Sons, Inc.

Kim, J. et al. (2005). Logrank-type tests for comparing survival curves with interval-censored data. *Computational Statistics and Data Analysis* **50**, 3165-3178.

Lakatos, E. et al. (1988). Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials. *Biometrics* **44**, 229-241.

Lim, H. and Sun, J. (2003) Nonparametric tests for interval-censored failure time data. *Biometrical Journal* **45**, 263-276.

Louis T.A. (1982) Finding the Observed Information Matrix Using the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226-233.

Pan, W. (1999). A comparison of some two-sample tests with interval-censored data. *Journal of Nonparametric Statistics* **12**, 133-146.

Pan, W. (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics* **56**, 199-203.

Pan, W. (2000). A two-smple test with interval-censored data via multiple imputation. *Statistics in Medicine* **19**, 1-11.

Peto, R. and Peto, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistics Society. Sieries A* **2**, 185-207.

Ren, J. (2003) Goodness of fit tests with interval-censored data. *Scandinavian Journal of Statistics* **30** 211-226.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biomterika* **68**, 316-319.

Sun, J. (2006). The statistical analysis of interval-censored failure time data. *New York, NY* Springer Science + Business Media Inc.

Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS. *Statistics in Medicine* **15**, 1378-1395.

Sun, J., Zhao, Q. and Zhao, X. (2005). Generalized Log-Rank Tests for Interval-Censored Failure Time Data. *Scandinavian Journal of Statistics* **32**, 49-57.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38**, 290-295.

Wilcoxson, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1** 80-83.

Wolfe, R. and Petroni, G. (1994). A Two-Sample Test for Stochastic Ordering with Interval-Censored Data. *Biometrics* **1** 77-87.

Wu, L. and Gilbert, P. (2002). Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* **58**, 997-1004.

Yuen, K., Shi, J., and Zhu, L. (2006) A k-sample test with interval censored data. *Biometrika* **93** 315-328.

Zhang, Y., Liu, W. and Wu, H. (2003). A simple nonparametric two-sample test for the distribution function of event time with interval censored data. *Nonparametric Statistics* **15**, 643-652.

Zhang, Y., Liu, W. and Zhan, Y. (2001). A nonparametric two-sample test of the failure function with interval-censoring case. *Biometrika* **88**, 677-686.

Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine* **23**, 1621-1629.

Hazard	A Early	B Late	C Crossing	D Parallel	E Crossing	F NPNC	S Same	W Same
Method 1	0.191	0.227	0.087	0.845	0.016	0.184	0.064	0.062
Method 2 $\gamma = 0$ $\rho = 0$	0.153	0.290	0.090	0.730	0.024	0.325	0.059	0.052
Method 2 $\gamma = 1$ $\rho = 0$	0.026	0.324	0.017	0.571	0.035	0.160	0.039	0.051
Method 2 $\gamma = 1$ $\rho = 1$	0.291	0.154	0.371	0.673	0.178	0.131	0.044	0.049
Logrank	0.275	0.222	0.131	0.805	0.047	0.190	0.039	0.035
Kolmogorov	0.393	0.208	0.624	0.691	0.290	0.145	0.052	0.051
Method 5	0.407	0.426	0.268	0.900	0.094	0.369	0.052	0.058
Score	0.576	0.752	0.419	0.951	0.212	0.673	0.053	0.022

Table 1: Discrete Censoring and Non-proportional Hazards

	a = 0.5 b = 1.73	a = 1 b = 1.73	a = 2 b = 1.73	a = 0.5 b = 2	a = 1 b = 2	a = 2 b = 2	a = 0.5 b = 2.25	a = 1 b = 2.25	a = 2 b = 2.25
Method 1	0.681	0.459	0.080	0.864	0.686	0.339	0.932	0.795	0.246
Method 2 $\gamma = 0$ $\rho = 0$	0.447	0.492	0.460	0.650	0.690	0.897	0.838	0.847	0.800
Method 2 $\gamma = 1$ $\rho = 0$	0.348	0.398	0.378	0.566	0.597	0.825	0.769	0.773	0.710
Method 2 $\gamma = 1$ $\rho = 1$	0.445	0.413	0.385	0.631	0.606	0.833	0.772	0.780	0.763
Logrank	0.643	0.489	0.242	0.816	0.685	0.494	0.940	0.787	0.787
Kolmogorov	0.505	0.426	0.346	0.688	0.582	0.700	0.841	0.706	0.608
Method 5	0.713	0.762	0.696	0.887	0.924	0.989	0.948	0.978	0.978
Score	0.828	0.859	0.460	0.945	0.937	0.764	0.981	0.975	0.737

Table 2: Discrete Censoring and Proportional Hazards

Hazard	A Early	B Late	C Crossing	D Parallel	E Crossing	F NPNC	S Same	W Same
Method 1	0.292	0.345	0.183	0.896	0.038	0.264	0.050	0.049
Method 2 $\gamma = 0$ $\rho = 0$	0.219	0.358	0.154	0.781	0.046	0.290	0.035	0.046
Method 2 $\gamma = 1$ $\rho = 0$	0.060	0.377	0.033	0.589	0.062	0.256	0.066	0.050
Method 2 $\gamma = 1$ $\rho = 1$	0.388	0.251	0.432	0.781	0.239	0.218	0.069	0.060
Logrank	0.195	0.218	0.146	0.769	0.037	0.168	0.040	0.043
Kolmogorov	0.319	0.081	0.497	0.417	0.192	0.079	0.041	0.047
Method 5	0.279	0.276	0.234	0.797	0.058	0.244	0.043	0.048
Score	0.443	0.671	0.301	0.963	0.197	0.560	0.049	0.063

Table 3: Continuous Censoring and Non-proportional Hazards

	a = 0.5 b = 1.73	a = 1 b = 1.73	a = 2 b = 1.73	a = 0.5 b = 2	a = 1 b = 2	a = 2 b = 2	a = 0.5 b = 2.25	a = 1 b = 2.25	a = 2 b = 2.25
Method 1	0.597	0.389	0.070	0.811	0.588	0.310	0.911	0.736	0.212
Method 2 $\gamma = 0$ $\rho = 0$	0.477	0.591	0.527	0.703	0.770	0.837	0.841	0.873	0.835
Method 2 $\gamma = 1$ $\rho = 0$	0.366	0.465	0.447	0.597	0.666	0.824	0.749	0.784	0.758
Method 2 $\gamma = 1$ $\rho = 1$	0.490	0.500	0.473	0.662	0.688	0.823	0.811	0.794	0.762
Logrank	0.588	0.426	0.198	0.764	0.569	0.477	0.874	0.715	0.379
Kolmogorov	0.178	0.175	0.181	0.206	0.257	0.432	0.290	0.352	0.376
Method 5	0.643	0.489	0.233	0.817	0.685	0.494	0.940	0.787	0.787
Score	0.816	0.799	0.766	0.927	0.932	0.973	0.975	0.956	0.948

Table 4: Continuous Censoring and Proportional Hazards

Sample	A Early	B Late	C Crossing	D Parallel	E Crossing	F NPNC	S Same	W Same
Method 1	0.398	0.561	0.178	0.990	0.014	0.457	0.055	0.043
Method 2 $\gamma = 0$ $\rho = 0$	0.333	0.607	0.160	0.972	0.020	0.455	0.049	0.051
Method2 $\gamma = 1$ $\rho = 0$	0.700	0.739	0.120	0.931	0.078	0.440	0.035	0.042
Method 2 $\gamma = 1$ $\rho = 1$	0.600	0.255	0.548	0.944	0.334	0.265	0.045	0.040
Logrank	0.466	0.526	0.279	0.991	0.046	0.434	0.046	0.051
Kolmogorov	0.739	0.388	0.619	0.923	0.542	0.226	0.049	0.054
Method 5	0.715	0.735	0.488	0.997	0.109	0.675	0.047	0.053

Table 5: Discrete Censoring and $n = 200$

	a = 0.5 b = 1.73	a = 1 b = 1.73	a = 2 b = 1.73	a = 0.5 b = 2	a = 1 b = 2	a = 2 b = 2	a = 0.5 b = 2.25	a = 1 b = 2.25	a = 2 b = 2.25
Method 1	0.923	0.853	0.314	0.986	0.958	0.841	0.999	0.991	0.728
Method 2 $\gamma = 0$ $\rho = 0$ 0 0	0.807	0.839	0.811	0.966	0.964	0.998	0.992	0.998	0.987
Method 2 $\gamma = 1$ $\rho = 0$	0.749	0.782	0.724	0.942	0.941	0.987	0.985	0.990	0.971
Method 2 $\gamma = 1$ $\rho = 1$	0.736	0.683	0.682	0.920	0.885	0.981	0.979	0.962	0.953
Logrank	0.899	0.808	0.594	0.991	0.930	0.931	0.999	0.986	0.880
Kolmogorov	0.764	0.704	0.559	0.919	0.858	0.949	0.980	0.940	0.886
Method 5	0.941	0.968	0.936	0.996	0.995	1.000	1.000	1.000	1.000

Table 6: Continuous Censoring and $n = 200$

Method	p-value
Method 1	0.0052
Method 2 $\gamma = 0$ $\rho = 0$	0.007
Method 2 $\gamma = 1$ $\rho = 0$	0.002
Method 2 $\gamma = 1$ $\rho = 1$	0.0004
Logrank	0.0057
Kolmogorov	0.14
Method 5	0.0007

Table 7: Results

Setup	A	B	C	D	E	F	S
	Early	Late	Crossing	Parallel	Crossing	NPNC	Same
$a = 0.0$	0.107	0.482	0.105	0.619	0.212	0.380	0.056
$a = 0.5$	0.438	0.137	0.343	0.756	0.158	0.237	0.061
$a = 1.0$	0.532	0.186	0.532	0.753	0.247	0.248	0.051

Table 8: Power and Size using $w_1(t)$

Setup	G	H	I	J	K	L	W
	Early	Late	Early	Late	Crossing	Both	Same
$a = 0.0$	0.375	0.107	0.158	0.158	0.148	0.330	0.065
$a = 0.5$	0.787	0.062	0.975	0.061	0.090	0.566	0.049
$a = 1.0$	0.861	0.076	0.983	0.062	0.067	0.547	0.055

Table 9: Power and Size using $w_1(t)$

Setup	A	B	C	D	E	F	S
	Early	Late	Crossing	Parallel	Crossing	NPNC	Same
$b = 0.0$	0.130	0.632	0.100	0.801	0.198	0.414	0.062
$b = 0.5$	0.215	0.569	0.100	0.843	0.094	0.438	0.058
$b = 1.0$	0.476	0.259	0.432	0.828	0.161	0.251	0.052

Table 10: Power and Size using $w_2(t)$

Setup	G	H	I	J	K	L	W
	Early	Late	Early	Late	Crossing	Both	Same
$b = 0.0$	0.527	0.104	0.393	0.138	0.115	0.390	0.052
$b = 0.5$	0.699	0.081	0.692	0.094	0.086	0.574	0.055
$b = 1.0$	0.857	0.058	0.980	0.056	0.074	0.574	0.044

Table 11: Power and Size using $w_2(t)$

Setup	A	B	C	D	E	F	S
	Early	Late	Crossing	Parallel	Crossing	NPNC	Same
GLRT	0.191	0.227	0.087	0.845	0.016	0.184	0.059

Table 12: Power and Size using the GLRT proposed in Sun 2006

Weight	w_1 Test Stat	p-value		w_2 Test Stat	p-value
$a = 0.0$	-.077	0.034		-.2066	0.000
$a = 0.5$	-1.305	0.000		-1.310	0.000
$a = 1.0$	-6.151	0.000		-7.142	0.00

Table 13: Test Statistics and p-values for testing whether there is a difference in time to shedding for patients with a low CD4 count and those with a non low CD4 count.

Weight	w_1 Test Stat	p-value		w_2 Test Stat	p-value
$a = 0.0$	-.077	0.034		-.2066	0.000
$a = 0.5$	-1.305	0.000		-1.310	0.000
$a = 1.0$	-6.151	0.000		-7.142	0.00

Table 14: Test Statistics and p-values for testing whether there is a difference in time to shedding for patients with a low CD4 count and those with a non low CD4 count.

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 0.5$	0.061	0.823	0.834	0.862	0.850	0.807	0.912	0.837	0.863	0.902	0.935	0.950	0.914	0.807	1.003	0.952
0.6	-	0.049	0.866	0.776	0.867	0.870	0.853	0.848	0.892	0.876	0.896	0.946	0.873	0.926	1.005	0.954
0.7	-	-	0.042	0.805	0.786	0.799	0.807	0.876	0.807	0.806	0.887	0.884	0.872	0.930	0.898	0.925
0.8	-	-	-	0.044	0.831	0.737	0.829	0.791	0.858	0.845	0.853	0.859	0.897	0.871	0.899	0.927
0.9	-	-	-	-	0.056	0.825	0.799	0.764	0.828	0.861	0.819	0.874	0.842	0.826	0.836	0.904
1.0	-	-	-	-	-	0.048	0.883	0.779	0.817	0.757	0.825	0.771	0.852	0.831	0.875	0.799
1.1	-	-	-	-	-	-	0.056	0.830	0.796	0.729	0.789	0.797	0.831	0.874	0.800	0.855
1.2	-	-	-	-	-	-	-	0.060	0.807	0.432	0.808	0.770	0.808	0.842	0.818	0.806
1.3	-	-	-	-	-	-	-	-	0.069	0.742	0.823	0.808	0.821	0.805	0.838	0.896
1.4	-	-	-	-	-	-	-	-	-	0.046	0.731	0.814	0.804	0.771	0.820	0.804
1.5	-	-	-	-	-	-	-	-	-	-	0.064	0.839	0.788	0.858	0.798	0.832
1.6	-	-	-	-	-	-	-	-	-	-	-	0.051	0.828	0.742	0.809	0.864
1.7	-	-	-	-	-	-	-	-	-	-	-	-	0.051	0.854	0.769	0.711
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	0.050	0.754	0.835
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053	0.832
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053

Table 15: Power and Size using $\beta = 0.2$

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 0.5$	0.046	0.905	0.923	0.920	0.940	0.938	0.923	0.945	0.960	0.958	0.985	0.958	0.973	0.983	0.978	0.993
0.6	-	0.059	0.908	0.915	0.903	0.918	0.930	0.930	0.935	0.965	0.943	0.965	0.968	0.988	0.983	0.978
0.7	-	-	0.041	0.890	0.905	0.915	0.925	0.905	0.930	0.933	0.930	0.955	0.970	0.963	0.940	0.950
0.8	-	-	-	0.050	0.900	0.903	0.925	0.923	0.900	0.945	0.903	0.943	0.925	0.948	0.973	0.965
0.9	-	-	-	-	0.048	0.920	0.895	0.928	0.883	0.920	0.953	0.935	0.945	0.923	0.920	0.958
1.0	-	-	-	-	-	0.052	0.913	0.908	0.900	0.938	0.918	0.925	0.933	0.940	0.923	0.928
1.1	-	-	-	-	-	-	0.048	0.893	0.895	0.910	0.895	0.918	0.930	0.938	0.933	0.935
1.2	-	-	-	-	-	-	-	0.045	0.913	0.640	0.893	0.910	0.930	0.898	0.928	0.935
1.3	-	-	-	-	-	-	-	-	0.048	0.913	0.903	0.903	0.920	0.923	0.928	0.908
1.4	-	-	-	-	-	-	-	-	-	0.042	0.878	0.920	0.910	0.903	0.915	0.918
1.5	-	-	-	-	-	-	-	-	-	-	0.053	0.898	0.918	0.930	0.878	0.900
1.6	-	-	-	-	-	-	-	-	-	-	-	0.050	0.925	0.890	0.898	0.918
1.7	-	-	-	-	-	-	-	-	-	-	-	-	0.056	0.903	0.895	0.903
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	0.043	0.875	0.923
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.056	0.913
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.048

Table 16: Power and Size using $\beta = 0.1$

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 0.5$	0.043	0.948	0.953	0.950	0.968	0.973	0.975	0.968	0.985	0.980	0.988	0.988	0.943	0.985	0.993	0.941
0.6	-	0.054	0.948	0.958	0.958	0.965	0.973	0.978	0.978	0.983	0.978	0.980	0.988	0.990	0.988	0.985
0.7	-	-	0.045	0.963	0.980	0.963	0.955	0.960	0.960	0.978	0.980	0.988	0.980	0.980	0.985	0.988
0.8	-	-	-	0.056	0.933	0.970	0.945	0.955	0.970	0.965	0.965	0.985	0.970	0.978	0.993	0.995
0.9	-	-	-	-	0.058	0.935	0.955	0.955	0.970	0.968	0.963	0.963	0.970	0.980	00.951	0.968
1.0	-	-	-	-	-	0.055	0.963	0.935	0.935	0.958	0.950	0.973	0.955	0.978	0.968	0.973
1.1	-	-	-	-	-	-	0.048	0.950	0.930	0.960	0.955	0.958	0.953	0.963	0.963	0.970
1.2	-	-	-	-	-	-	-	0.053	0.958	0.708	0.938	0.960	0.963	0.945	0.968	0.955
1.3	-	-	-	-	-	-	-	-	0.058	0.948	0.938	0.963	0.940	0.963	0.948	0.960
1.4	-	-	-	-	-	-	-	-	-	0.051	0.960	0.948	0.965	0.948	0.973	0.960
1.5	-	-	-	-	-	-	-	-	-	-	0.045	0.955	0.948	0.953	0.955	0.950
1.6	-	-	-	-	-	-	-	-	-	-	-	0.047	0.955	0.928	0.940	0.955
1.7	-	-	-	-	-	-	-	-	-	-	-	-	0.046	0.963	0.965	0.950
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	0.051	0.953	0.955
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.048	0.945
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.052

Table 17: Power and Size using $\beta = 0.05$

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 0.5$	296	1419	424	225	150	113	91	78	69	62	57	53	50	48	46	44
0.6	-	413	1931	564	292	191	141	112	94	82	73	67	62	58	55	52
0.7	-	-	552	2535	727	370	238	173	136	113	97	86	77	71	66	62
0.8	-	-	-	715	3233	914	459	291	209	163	133	114	100	89	81	75
0.9	-	-	-	-	901	4028	1125	558	350	249	192	156	132	115	102	93
1.0	-	-	-	-	-	1112	4922	1361	669	415	293	224	181	152	131	116
1.1	-	-	-	-	-	-	1348	5916	1623	791	488	342	259	208	174	149
1.2	-	-	-	-	-	-	-	1609	7014	849	925	567	394	298	238	197
1.3	-	-	-	-	-	-	-	-	1897	8217	2225	1071	652	452	339	269
1.4	-	-	-	-	-	-	-	-	-	2211	9528	2567	1229	745	514	384
1.5	-	-	-	-	-	-	-	-	-	-	2552	10950	2937	1401	845	580
1.6	-	-	-	-	-	-	-	-	-	-	-	2922	12487	3336	1585	953
1.7	-	-	-	-	-	-	-	-	-	-	-	-	3321	14142	3766	1783
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	3750	15920	4226
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4209	17824
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4702

Table 18: Sample Size Required Using $\alpha = 0.05$ and $\beta = 0.05$

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 0.5$	234	1123	336	178	118	89	72	62	54	49	45	42	40	38	36	35
0.6	-	326	1528	446	231	151	112	89	75	65	58	53	49	46	43	41
0.7	-	-	437	2006	575	293	188	137	108	89	77	68	61	56	52	49
0.8	-	-	-	565	2558	723	363	230	165	129	106	90	79	71	64	60
0.9	-	-	-	-	713	3187	890	442	277	197	152	124	104	91	81	73
1.0	-	-	-	-	-	880	3895	1077	529	329	232	177	143	120	104	92
1.1	-	-	-	-	-	-	1066	4682	1284	626	386	270	205	165	137	118
1.2	-	-	-	-	-	-	-	01273	5550	672	732	448	312	236	188	156
1.3	-	-	-	-	-	-	-	-	1501	6502	1761	848	516	357	268	213
1.4	-	-	-	-	-	-	-	-	-	1749	7540	2031	973	590	406	304
1.5	-	-	-	-	-	-	-	-	-	-	2020	8665	2324	1108	669	459
1.6	-	-	-	-	-	-	-	-	-	-	-	2312	9881	2640	1254	754
1.7	-	-	-	-	-	-	-	-	-	-	-	-	2628	11191	2980	1411
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	2967	12598	3344
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3331	14105
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3720

Table 19: Sample Size Required Using $\alpha = 0.05$ and $\beta = 0.10$

λ_1	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\lambda_2 = 169$	0.059	811	242	128	86	64	52	44	39	35	33	30	29	27	26	25
0.6	-	235	1103	322	167	109	81	64	54	47	42	38	35	33	31	30
0.7	-	-	315	1448	415	211	136	99	78	64	55	49	44	41	38	36
0.8	-	-	-	408	1847	522	262	166	119	93	76	65	57	51	47	43
0.9	-	-	-	-	515	2301	643	319	200	142	110	89	75	66	58	53
1.0	-	-	-	-	-	635	2812	778	382	237	168	128	103	87	75	66
1.1	-	-	-	-	-	-	770	3380	927	452	279	195	148	119	99	85
1.2	-	-	-	-	-	-	-	919	4007	485	528	324	225	170	136	113
1.3	-	-	-	-	-	-	-	-	1083	4694	1271	612	373	258	194	154
1.4	-	-	-	-	-	-	-	-	-	1263	75443	1466	702	426	293	219
1.5	-	-	-	-	-	-	-	-	-	-	1458	6256	1678	800	483	332
1.6	-	-	-	-	-	-	-	-	-	-	-	1669	7134	1906	905	544
1.7	-	-	-	-	-	-	-	-	-	-	-	-	1897	8079	2151	1019
1.8	-	-	-	-	-	-	-	-	-	-	-	-	-	2142	9095	2414
1.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2405	10183
2.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2686

Table 20: Sample Size Required Using $\alpha = 0.05$ and $\beta = 0.20$

Hazard	4	8	10	20
A	0.682	0.675	0.676	0.684
B	0.563	0.508	0.492	0.507
C	0.565	0.474	0.486	0.539
D	0.566	0.496	0.484	0.506
E	0.582	0.474	0.478	0.499
F	0.570	0.509	0.486	0.503
G	0.127	0.049	0.063	0.126
H	0.533	0.503	0.501	0.465
I	0.458	0.370	0.387	0.426
J	0.365	0.300	0.307	0.323
K	0.370	0.284	0.308	0.320
L	0.931	0.927	0.920	0.896

Table 21: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate (Part 1 of hazards)

Hazard	4	8	10	20
A	3.356	3.616	3.593	3.490
B	3.246	3.476	3.479	3.221
C	3.171	3.492	3.485	3.343
D	3.192	3.527	3.476	3.284
E	3.216	3.481	3.485	3.327
F	3.201	3.503	3.506	3.270
G	2.387	3.125	2.878	2.318
H	3.189	3.505	3.460	3.258
I	2.848	3.374	3.321	3.137
J	3.051	3.441	3.410	3.066
K	3.049	3.433	3.423	3.071
L	3.185	3.495	3.520	3.419

Table 22: Average time where the Self-Consistency Estimate first crosses either the right or left-imputed Kaplan-Meier Estimate.Using 1st hazard)

Hazard	4	8	10	20
A	0.601	0.571	0.571	0.576
B	0.000	0.034	0.041	0.095
C	0.579	0.464	0.504	0.527
D	0.986	0.979	0.984	0.959
E	0.099	0.028	0.037	0.091
F	0.155	0.053	0.068	0.117
G	0.197	0.087	0.100	0.134
H	0.271	0.159	0.195	0.225
I	0.700	0.646	0.669	0.660
J	0.829	0.751	0.712	0.688
K	0.802	0.752	0.717	0.641
L	0.563	0.486	0.504	0.505

Table 23: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate (part 2 of hazards)

Hazard	4	8	10	20
A	3.236	3.556	3.524	3.376
B	2.307	3.163	2.944	2.337
C	3.140	3.489	3.447	3.290
D	3.460	3.549	3.549	3.465
E	2.467	3.110	3.150	2.595
F	2.200	3.025	2.880	2.412
G	2.648	3.374	3.205	3.000
H	2.941	3.313	3.350	2.851
I	3.397	3.587	3.577	3.469
J	3.144	3.447	3.449	3.373
K	3.158	3.452	3.412	3.397
L	3.180	3.500	3.499	3.319

Table 24: Average time where the Self-Consistency Estimate first crosses either the right or left-imputed Kaplan-Meier Estimate. (Using second hazard)

Percent Right-Censored	4	8	10	20
90.00%	0.980	0.972	0.950	0.831
75.00%	1.000	0.999	0.998	0.979
66.00%	1.000	1.000	0.998	0.997
50.00%	1.000	1.000	1.000	0.992
33.00%	1.000	1.000	1.000	0.993
25.00%	0.999	0.999	1.000	0.984
10.00%	0.945	0.960	0.962	0.932

Table 25: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate to compare different proportions of right-censored observations.

Percent Right-Censored	4	8	10	20
90.00%	0.069	0.067	0.066	0.070
75.00%	0.177	0.206	0.207	0.277
66.00%	0.265	0.305	0.314	0.332
50.00%	0.519	0.583	0.592	0.607
33.00%	0.893	0.943	0.950	0.944
25.00%	1.185	1.230	1.233	1.205
10.00%	1.976	2.057	2.051	1.982

Table 26: Average time the Self-Consistency Estimate first crosses either the right or left imputed Kaplan-Meier Estimate.

Mean	4	8	10	20	continuous
0.5	0.000	0.035	0.034	0.075	0.008
1	0.559	0.499	0.497	0.447	0.251
2	0.987	0.982	0.987	0.951	0.882
3	1.000	1.000	0.997	0.991	0.991
5	1.000	1.000	1.000	0.991	1.000
6	1.000	1.000	0.999	0.991	1.000

Table 27: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate (exponential with various means)

Mean	4	8	10	20	continuous
0.5	0.000	0.035	0.031	0.074	0.001
1.0	0.353	0.131	0.085	0.189	0.008
2.0	0.485	0.230	0.091	0.149	0.028
3.0	0.430	0.213	0.095	0.205	0.042
5.0	0.442	0.321	0.175	0.325	0.060
6.0	0.444	0.332	0.202	0.377	0.080

Table 28: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate ignoring all time points greater than 3

Mean	8	10	20	continuous
0.5	0.000	0.000	0.000	0.000
1.0	0.000	0.164	0.228	0.102
2.0	0.579	0.549	0.713	0.501
3.0	0.778	0.721	0.848	0.712
5.0	0.870	0.826	0.804	0.819
6.0	0.867	0.846	0.781	0.846

Table 29: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Sample Size 10

Mean	8	10	20	continuous
0.5	0.051	0.052	0.094	0.010
1.0	0.718	0.731	0.685	0.375
2.0	1.000	0.998	0.999	0.989
3.0	1.000	1.000	1.000	0.999
5.0	1.000	1.000	0.999	1.000
6.0	1.000	1.000	1.000	1.000

Table 30: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Sample Size 100

Mean	8	10	20	continuous
0.5	0.340	0.250	0.170	0.080
1.0	1.000	1.000	1.000	1.000
2.0	1.000	1.000	1.000	1.000
3.0	1.000	1.000	1.000	1.000
5.0	1.000	1.000	1.000	1.000
6.0	1.000	1.000	1.000	1.000

Table 31: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Sample Size 1000

Mean	8	10	20
0.5	0.016	0.010	0.040
1.0	0.071	0.034	0.057
2.0	0.046	0.015	0.021
3.0	0.036	0.006	0.024
5.0	0.045	0.016	0.039
6.0	0.040	0.018	0.048

Table 32: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Sample Size 100, ignoring all time points greater than 3

Mean	8	10	20
0.5	0.031	0.009	0.018
1.0	0.001	0.000	0.000
2.0	0.000	0.000	0.000
3.0	0.000	0.000	0.000
5.0	0.000	0.000	0.000
6.0	0.000	0.000	0.000

Table 33: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Sample Size 1000, ignoring all time points greater than 3

Mean	8	10	20
0.5	0.000	0.132	0.017
1.0	0.589	0.373	0.143
2.0	0.841	0.634	0.328
3.0	0.865	0.678	0.395
5.0	0.856	0.697	0.477
6.0	0.838	0.713	0.524

Table 34: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 0.2

Mean	8	10	20
0.5	0.274	0.269	0.401
1.0	0.338	0.412	0.497
2.0	0.259	0.317	0.521
3.0	0.245	0.325	0.518
5.0	0.284	0.351	0.602
6.0	0.342	0.378	0.634

Table 35: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 0.9

Mean	8	10	20
0.5	0.469	0.448	0.632
1.0	0.368	0.431	0.426
2.0	0.266	0.312	0.519
3.0	0.262	0.299	0.478
5.0	0.300	0.355	0.597
6.0	0.320	0.387	0.637

Table 36: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 1.0

Mean	8	10	20
0.5	0.469	0.134	0.016
1.0	0.594	0.397	0.146
2.0	0.836	0.621	0.311
3.0	0.867	0.706	0.425
5.0	0.864	0.690	0.497
6.0	0.856	0.726	0.494

Table 37: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 0.2, ignoring crossings at time greater than 3.

Mean	8	10	20
0.5	0.257	0.267	0.432
1.0	0.342	0.418	0.444
2.0	0.258	0.322	0.524
3.0	0.271	0.348	0.516
5.0	0.306	0.345	0.601
6.0	0.312	0.383	0.660

Table 38: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 0.9, ignoring crossings at time greater than 3.

Mean	8	10	20
0.5	0.469	0.448	0.632
1.0	0.368	0.431	0.426
2.0	0.266	0.312	0.519
3.0	0.262	0.299	0.478
5.0	0.300	0.355	0.597
6.0	0.320	0.387	0.637

Table 39: Proportion of Times Self Consistent Estimate Crosses Kaplan-Meier Estimate: Probability of observation 1.0, ignoring crossings at time greater than 3.

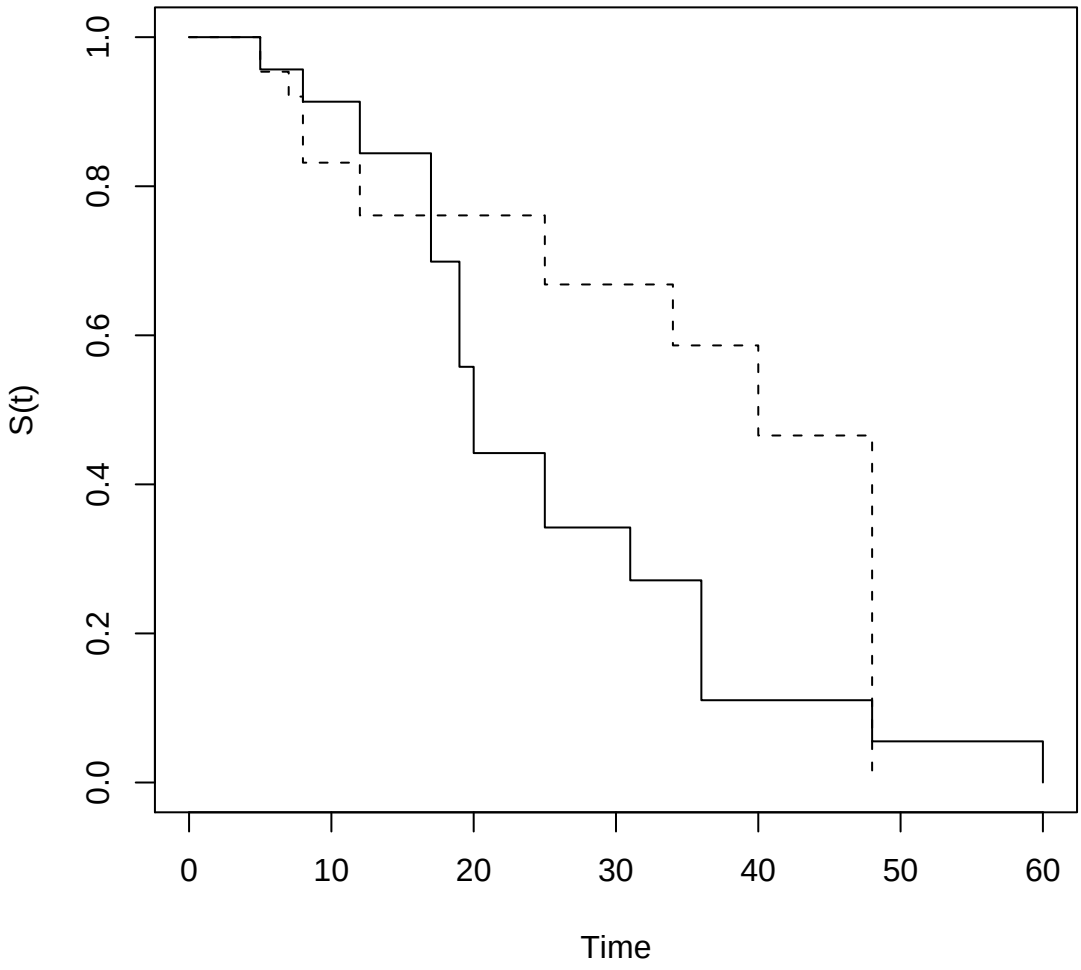


Figure 1: Survival curves by treatment groups in breast cosmesis data.

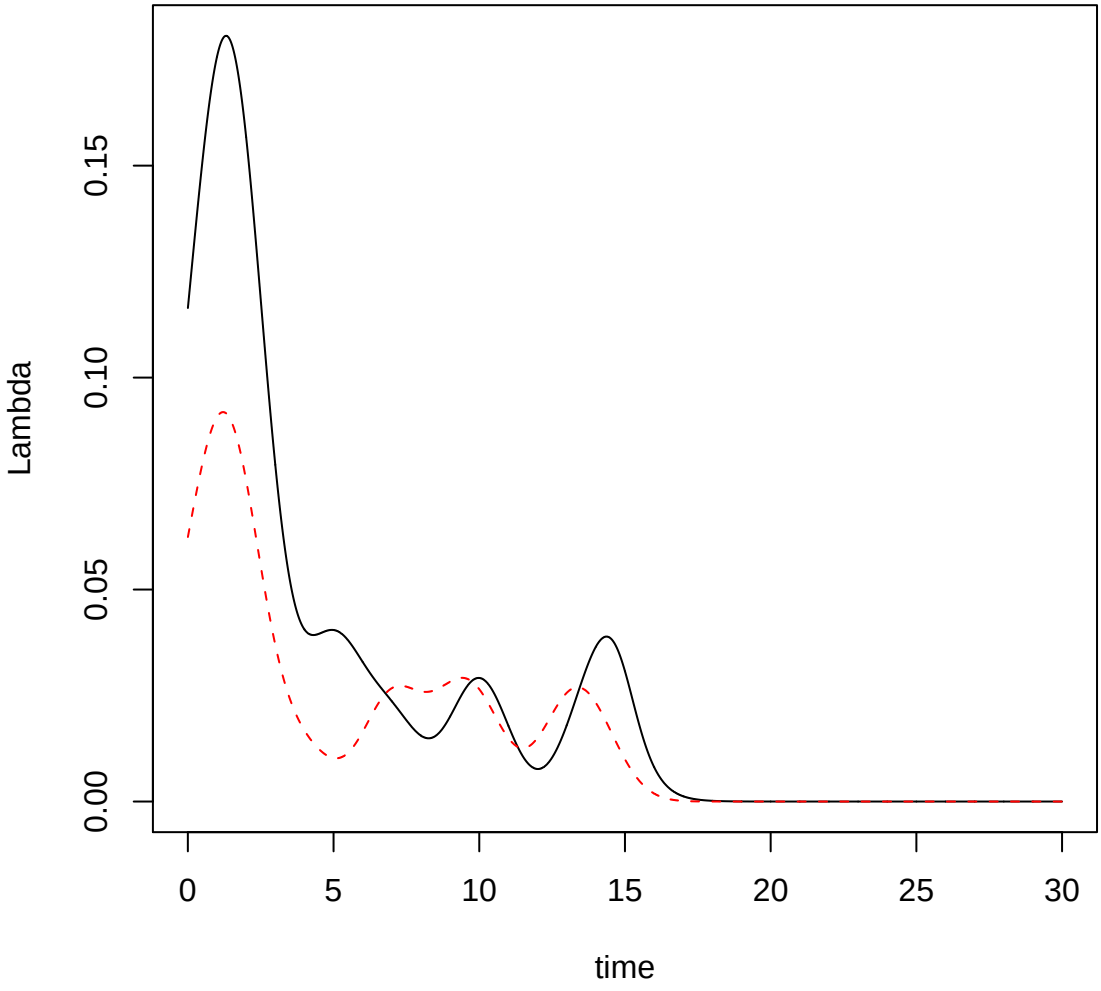


Figure 2: Smoothed Hazard functions for the shedding time of CMV in patients with HIV.

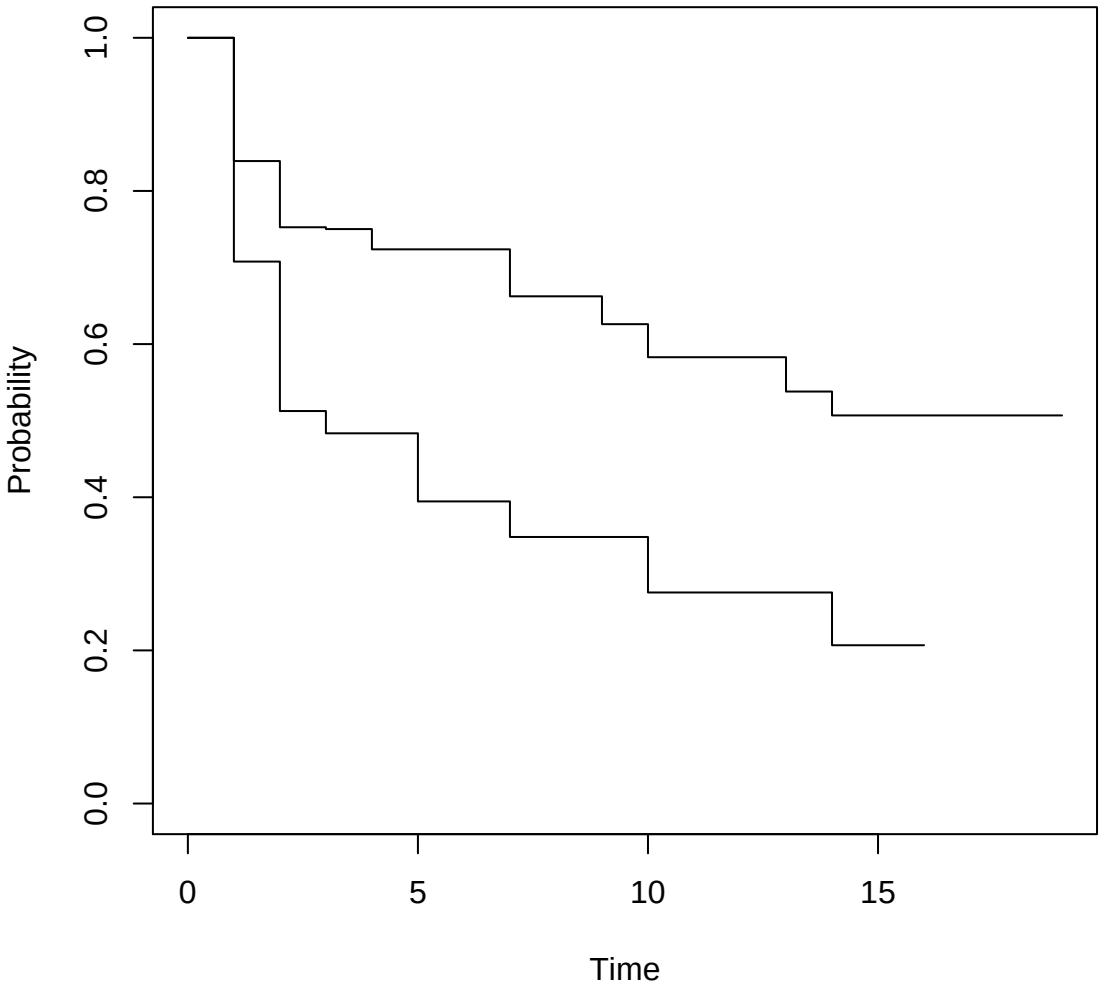


Figure 3: Estimated survival functions for the shedding time of CMV in patients with HIV.

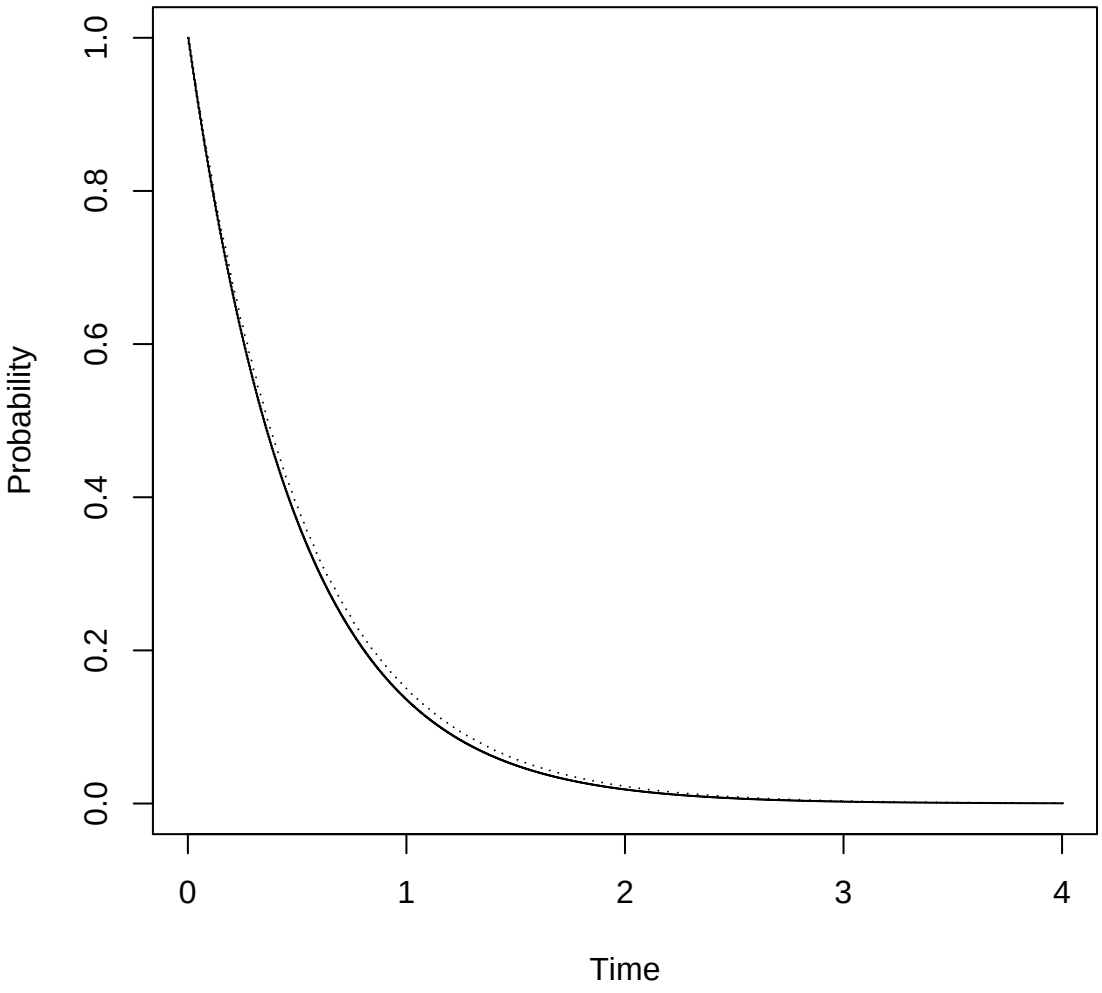


Figure 4: Plot of the survival functions for an exponential with $\lambda = 2.0$ and with $\lambda = 1.9$

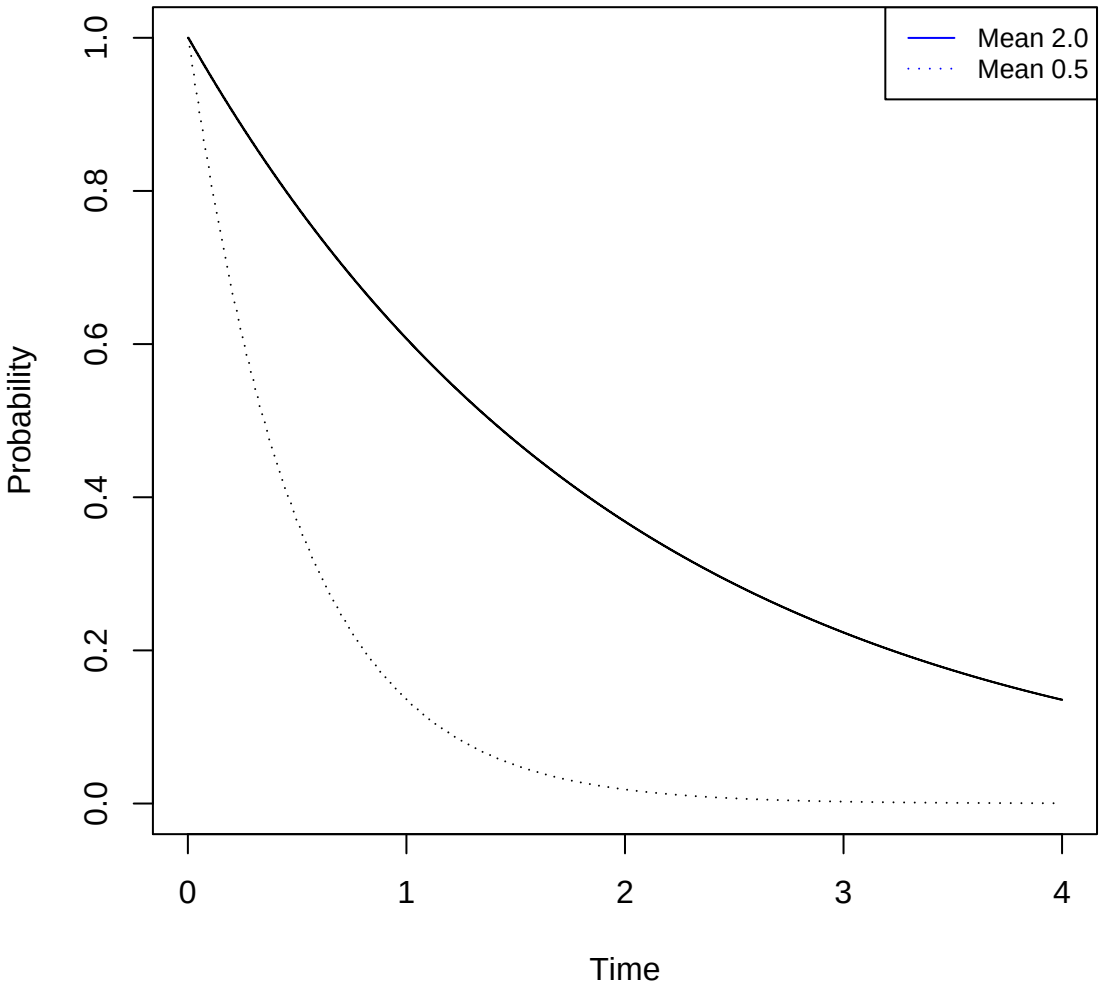


Figure 5: Plot of the survival functions for an exponential with $\lambda = 2.0$ and with $\lambda = 0.5$

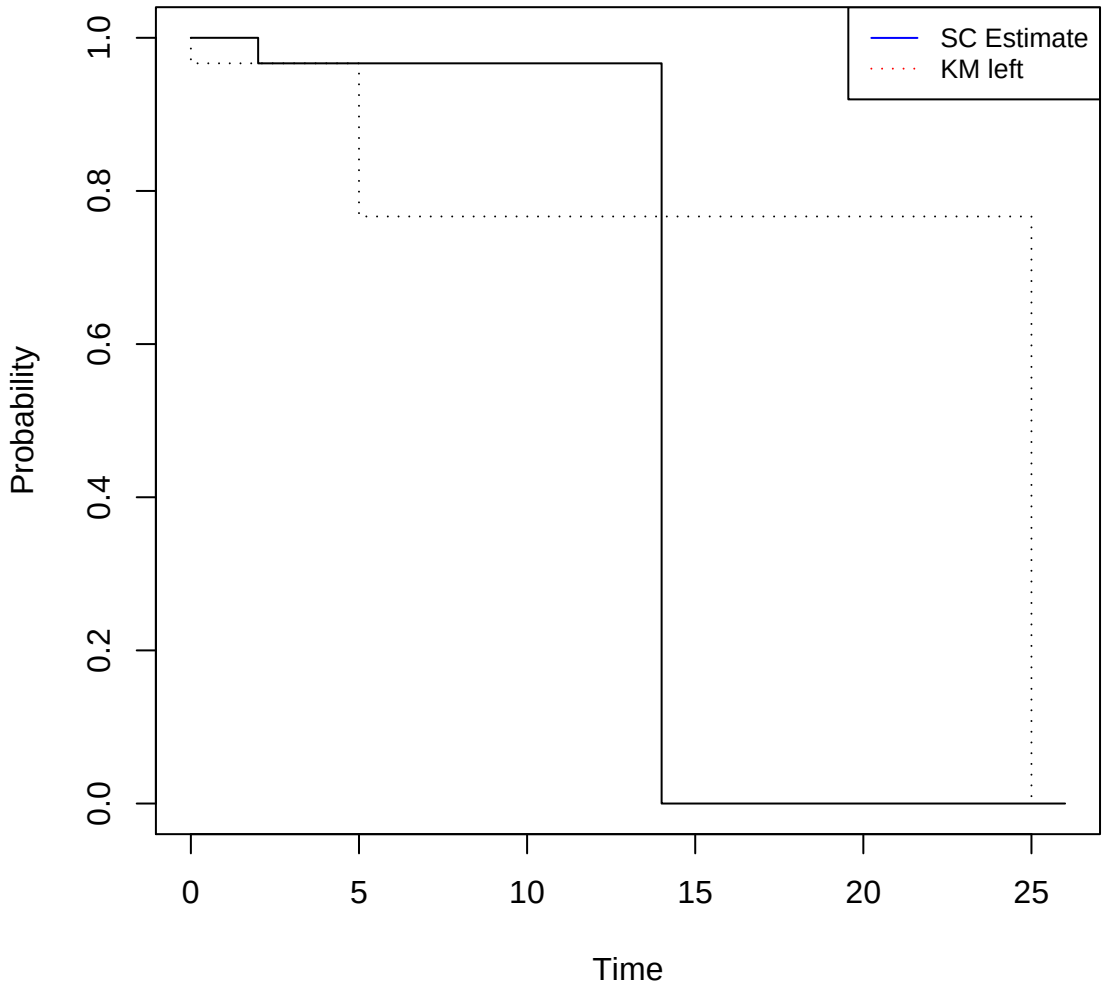


Figure 6: Plot of the Self Consistency Estimate and the Left-Imputed Kaplan-Meier Estimate for Rats

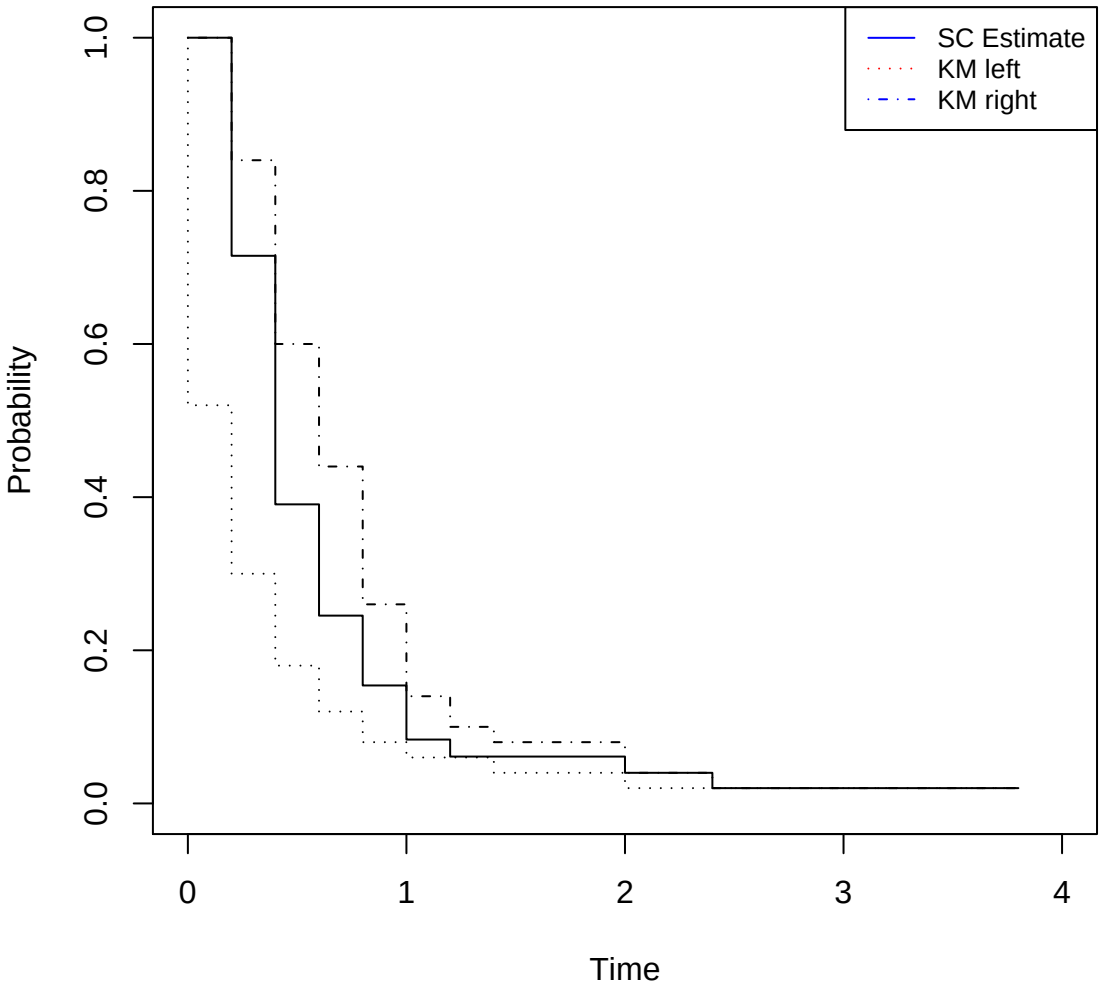


Figure 7: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 0.5

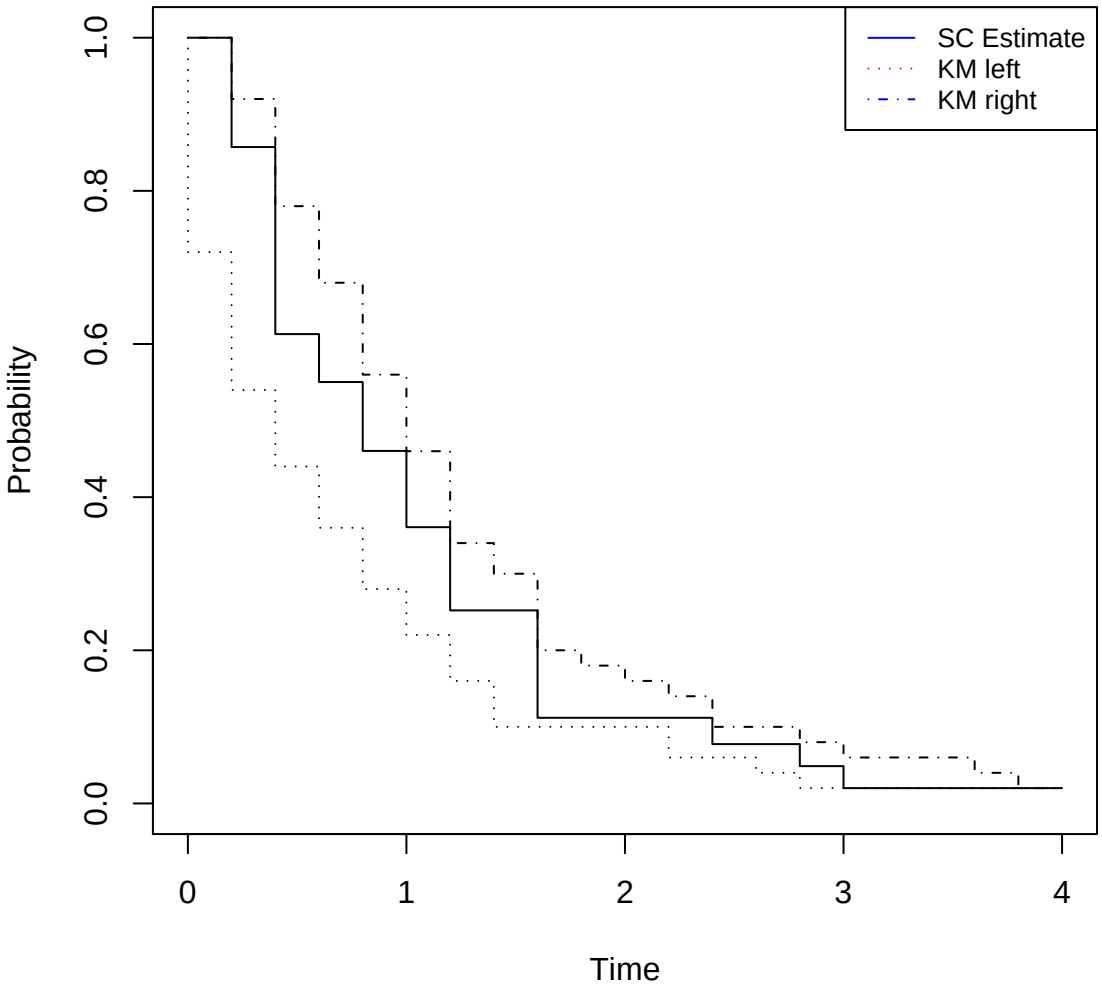


Figure 8: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 1

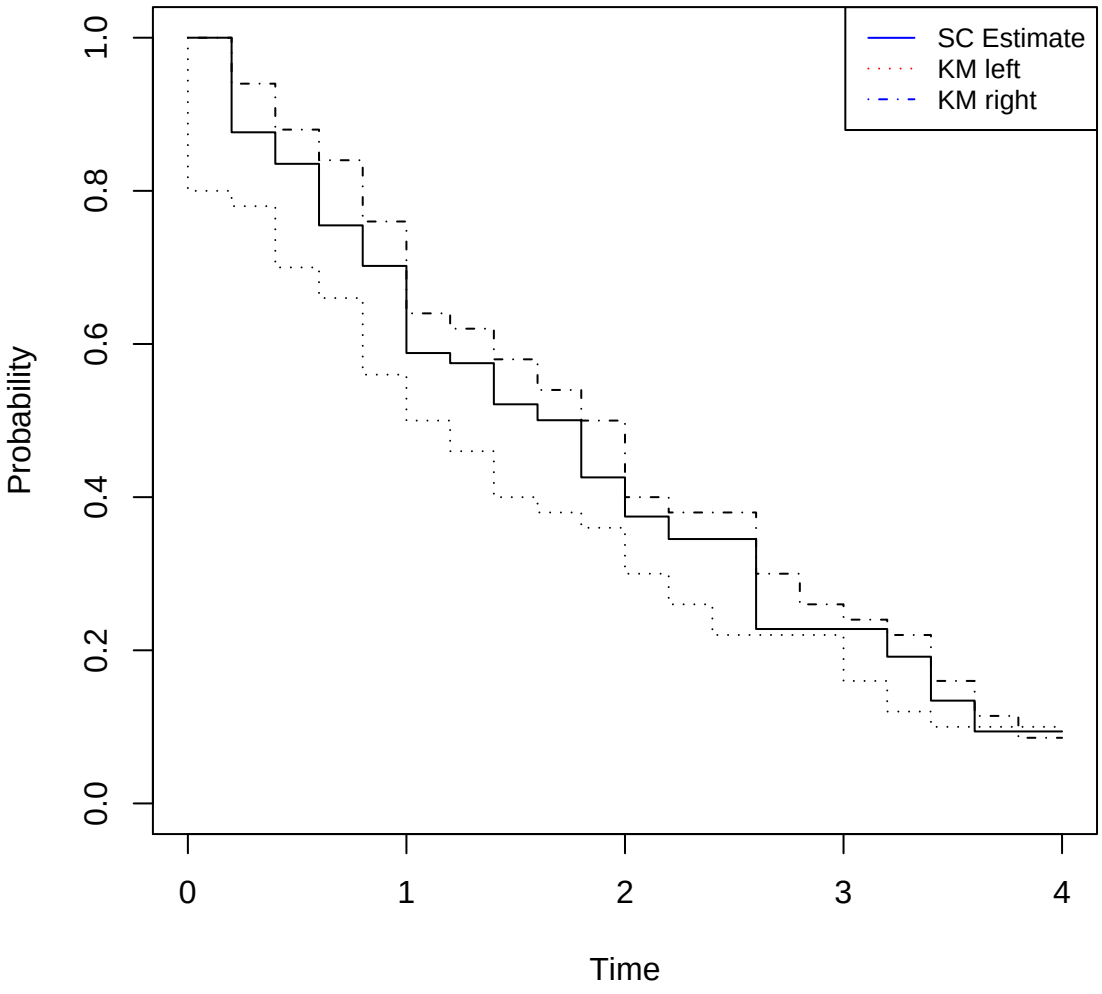


Figure 9: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 2

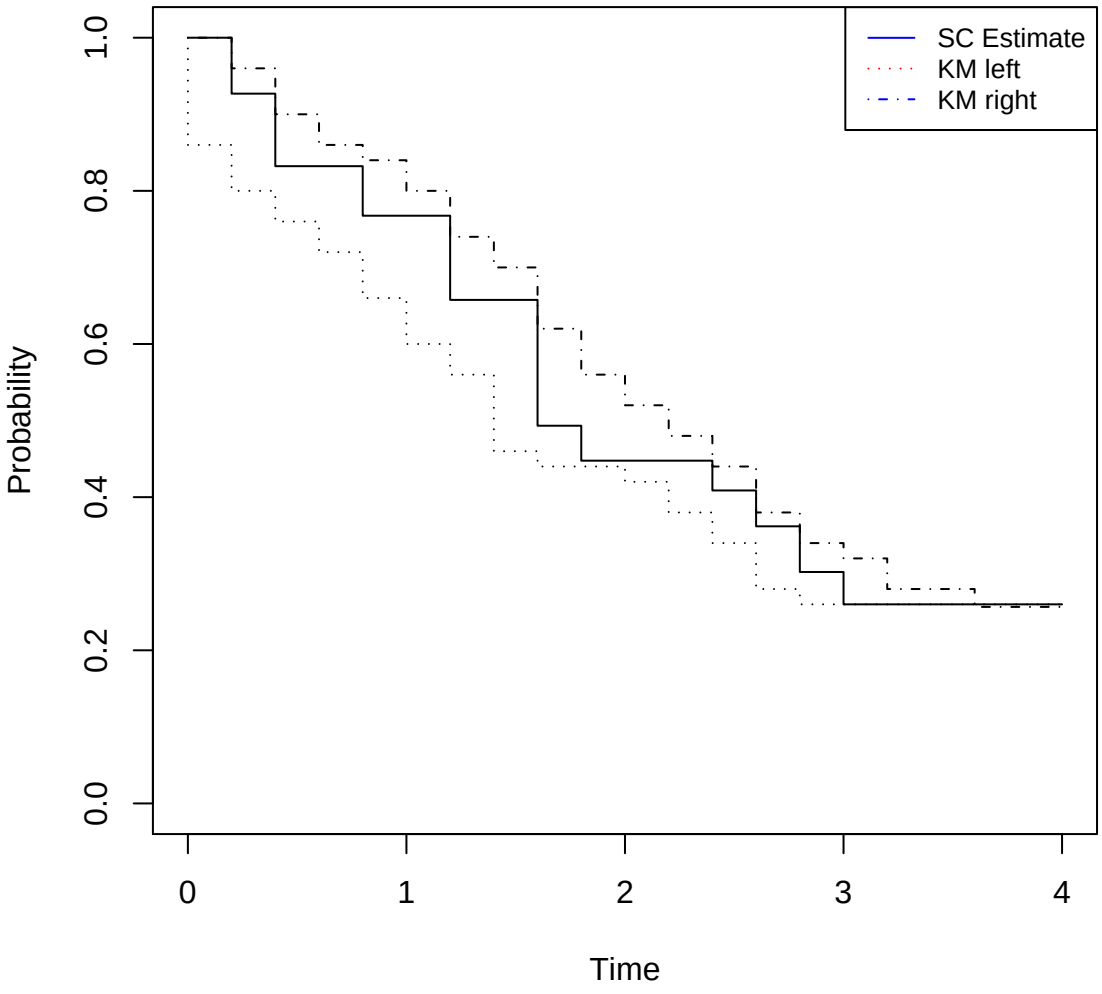


Figure 10: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 3

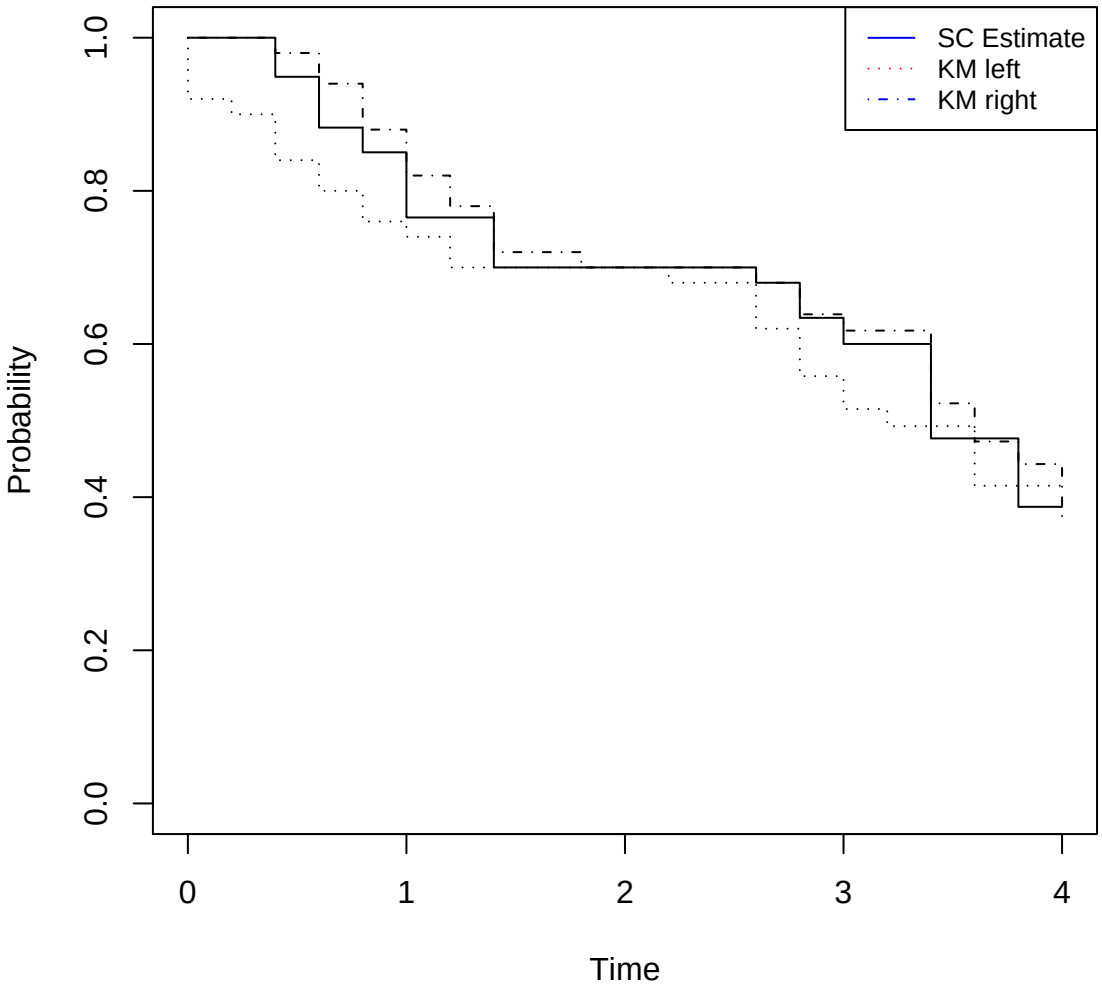


Figure 11: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 5

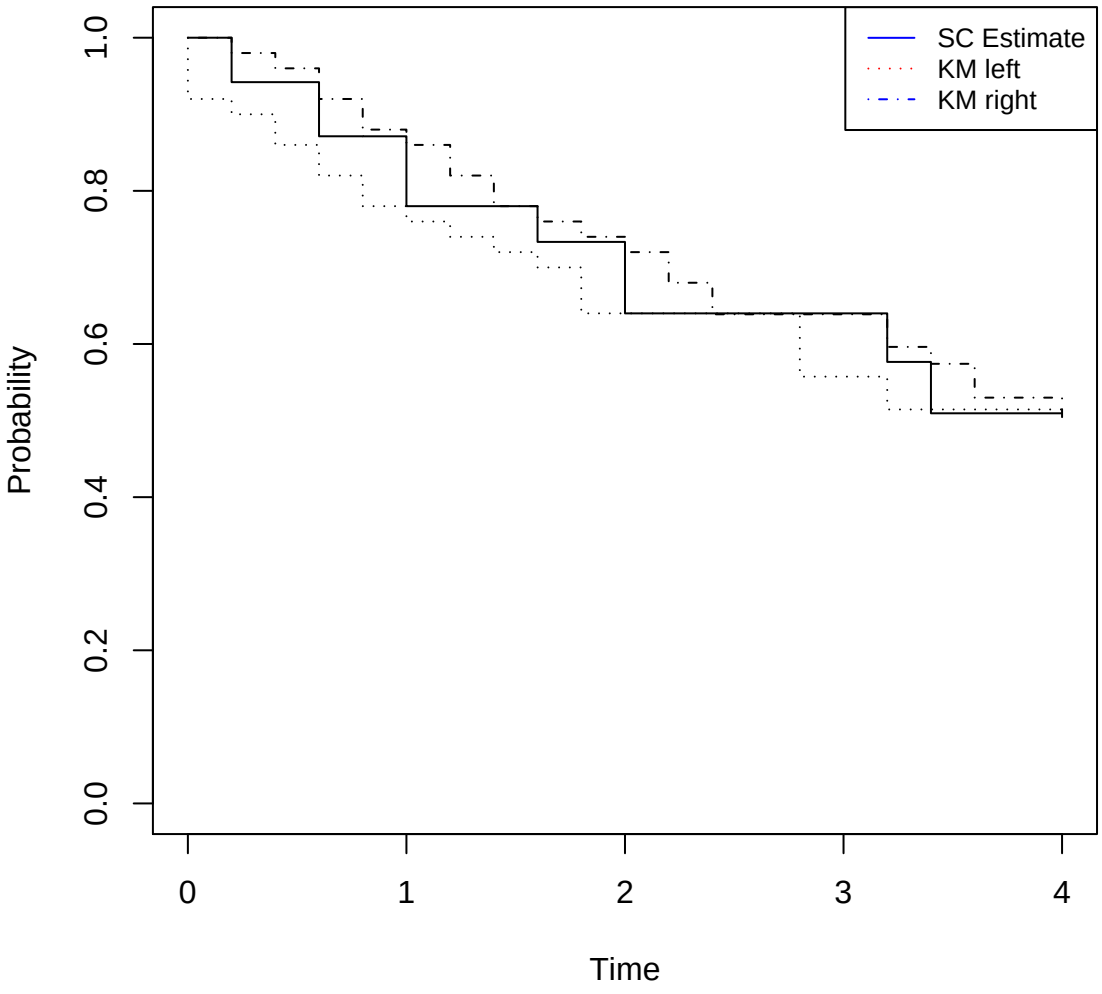


Figure 12: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 6

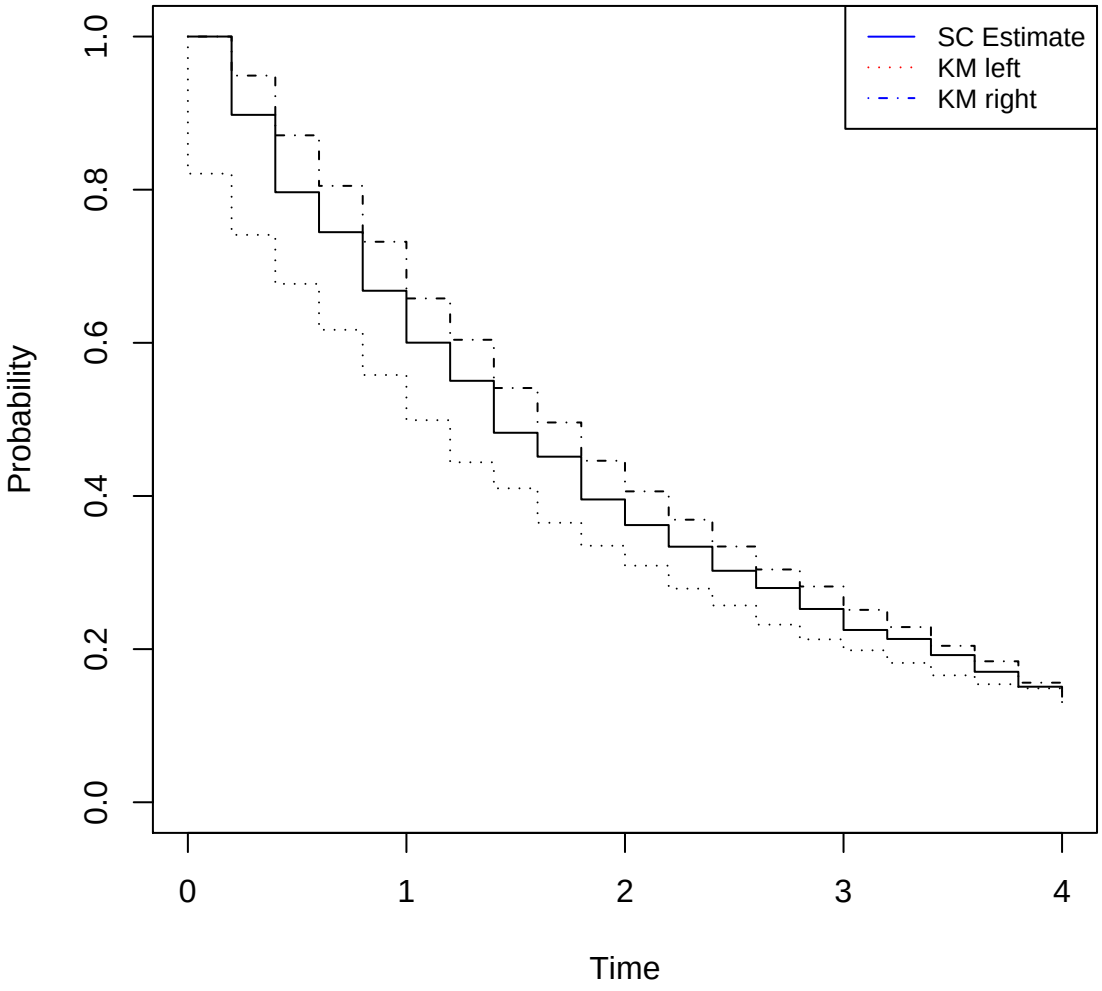


Figure 13: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 2 and sample size 1000

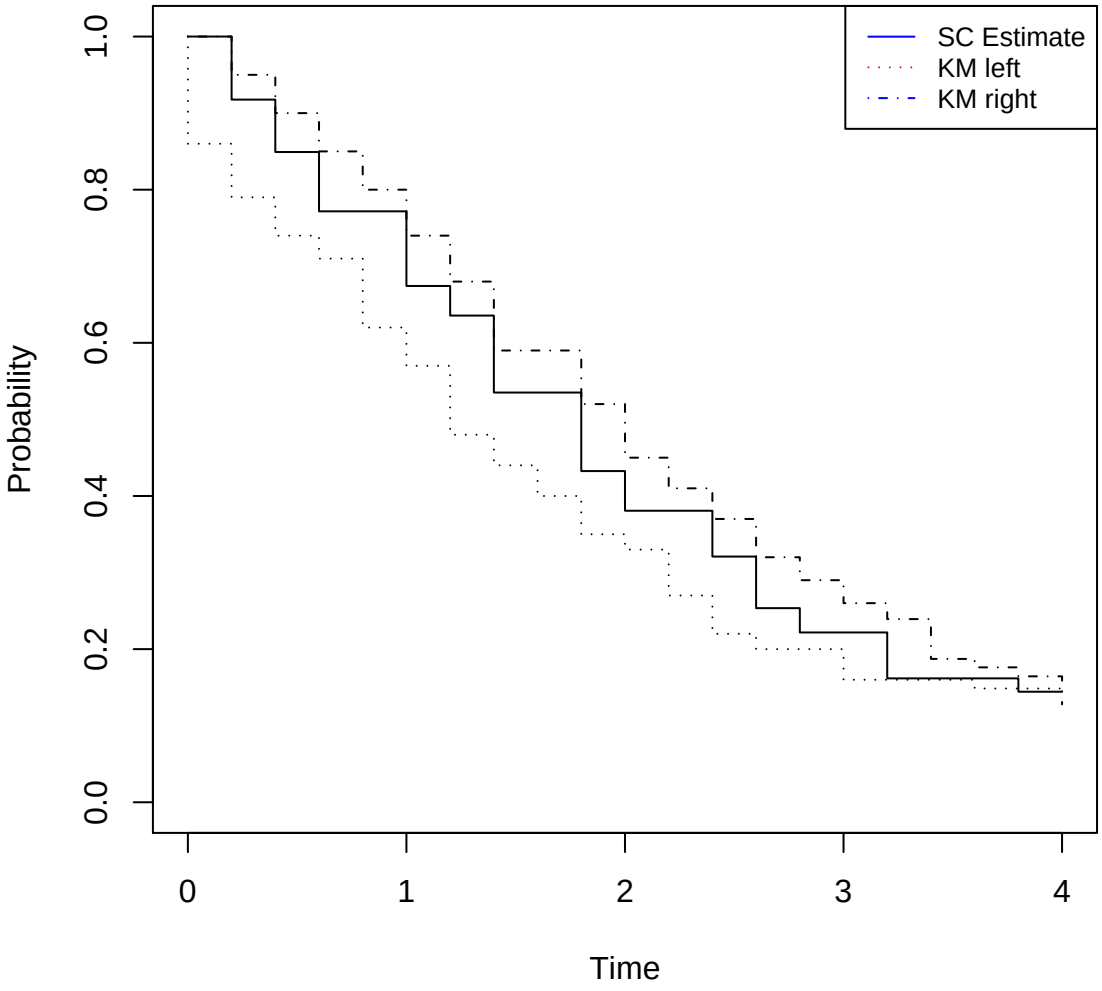


Figure 14: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 2 and sample size 100

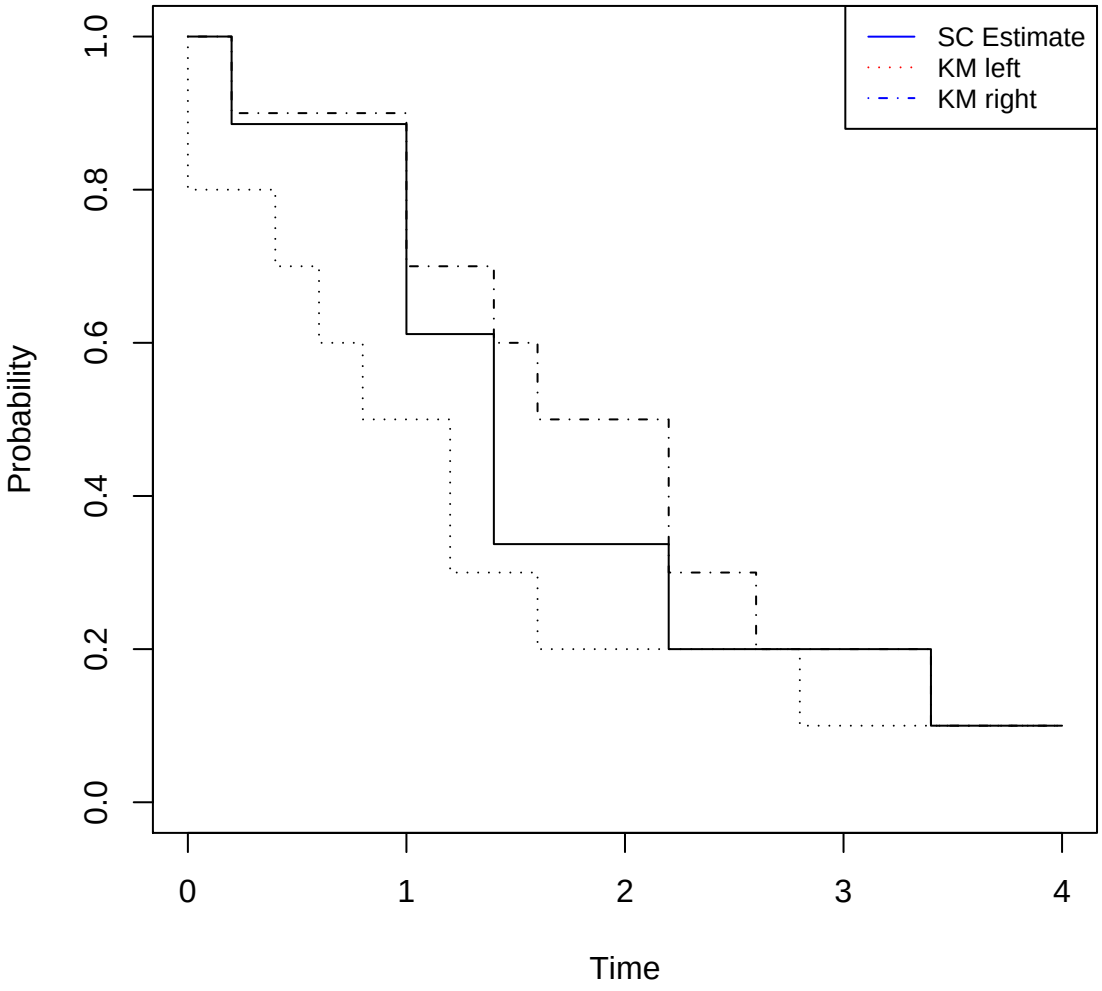


Figure 15: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 2 and sample size 10

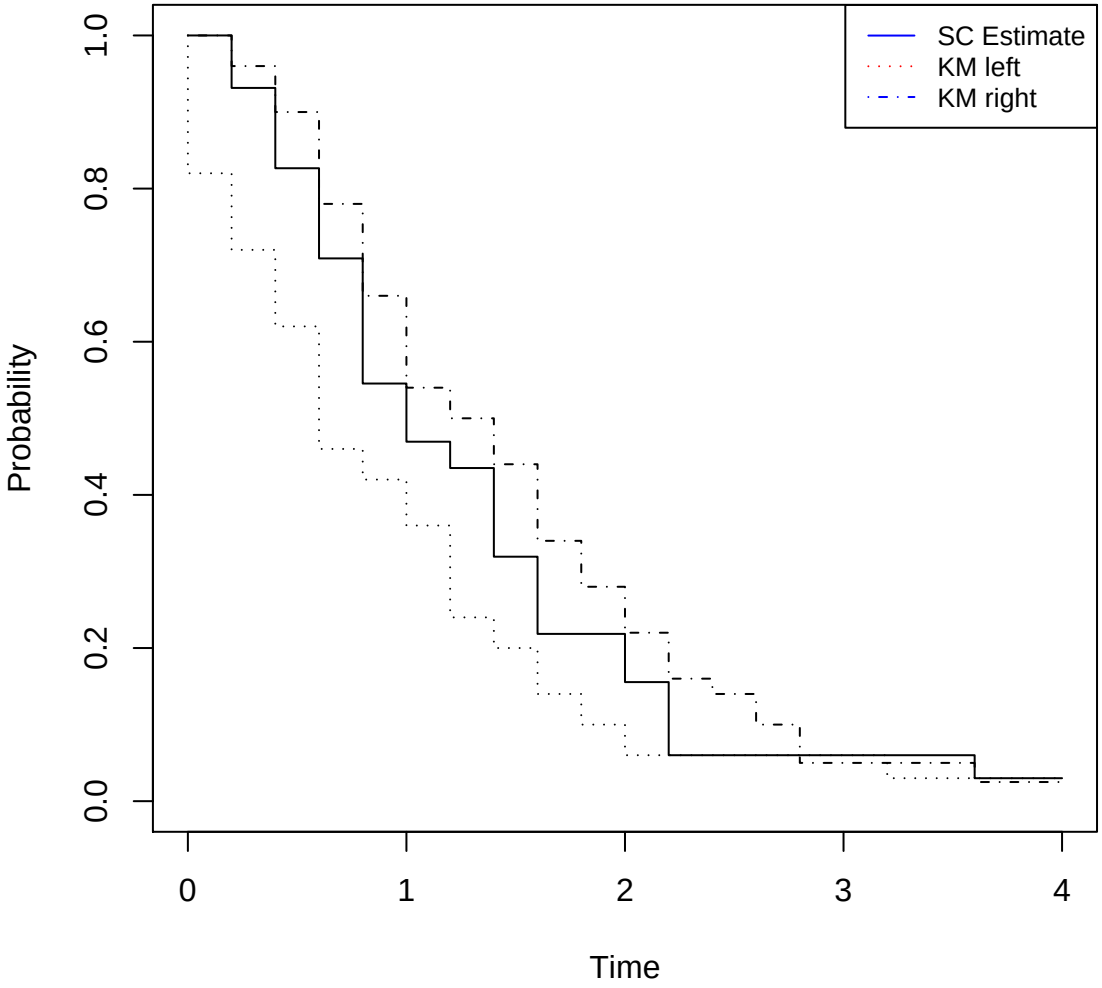


Figure 16: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 1. This is an example where they cross

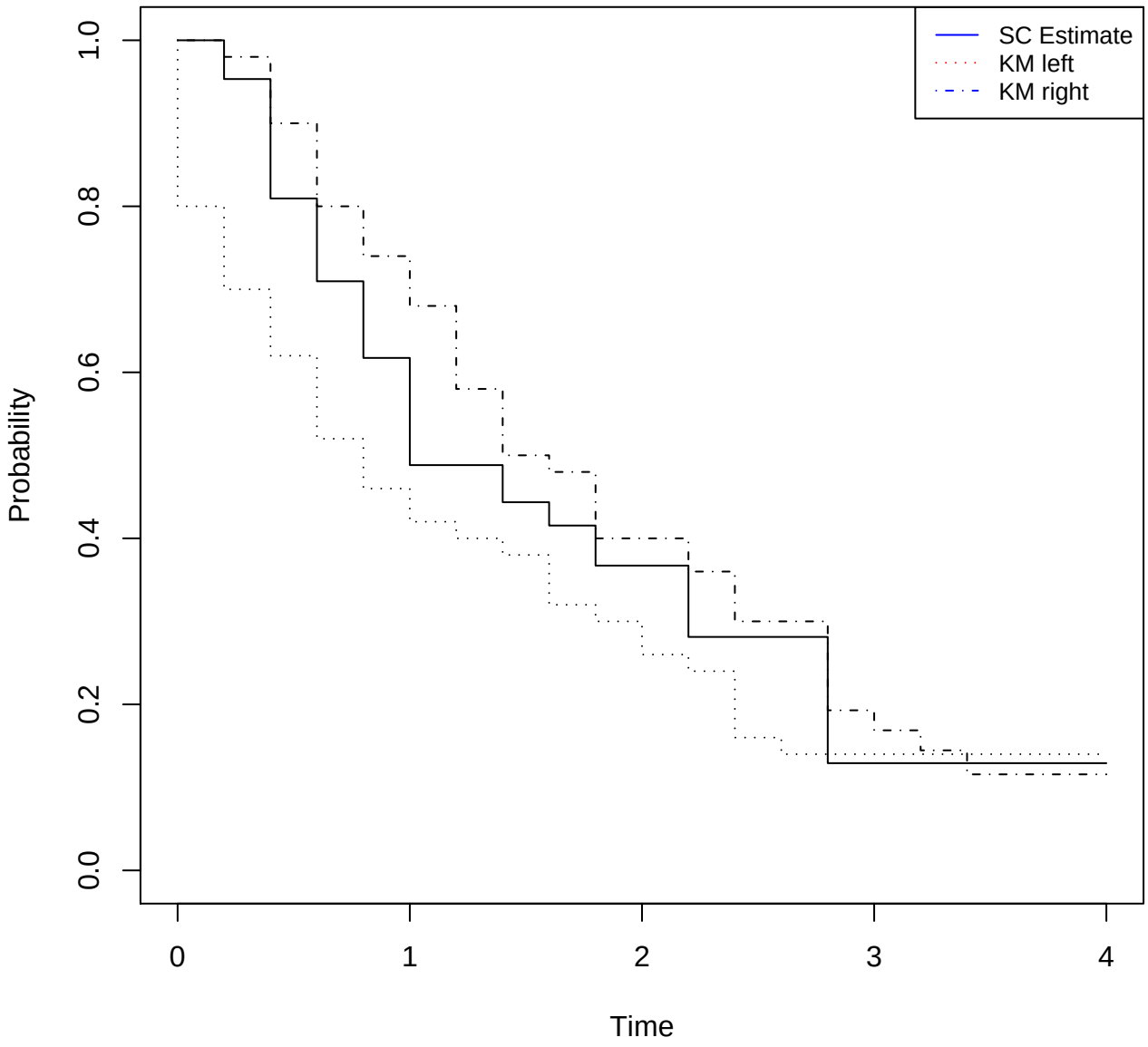


Figure 17: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 2. This is an example where they cross

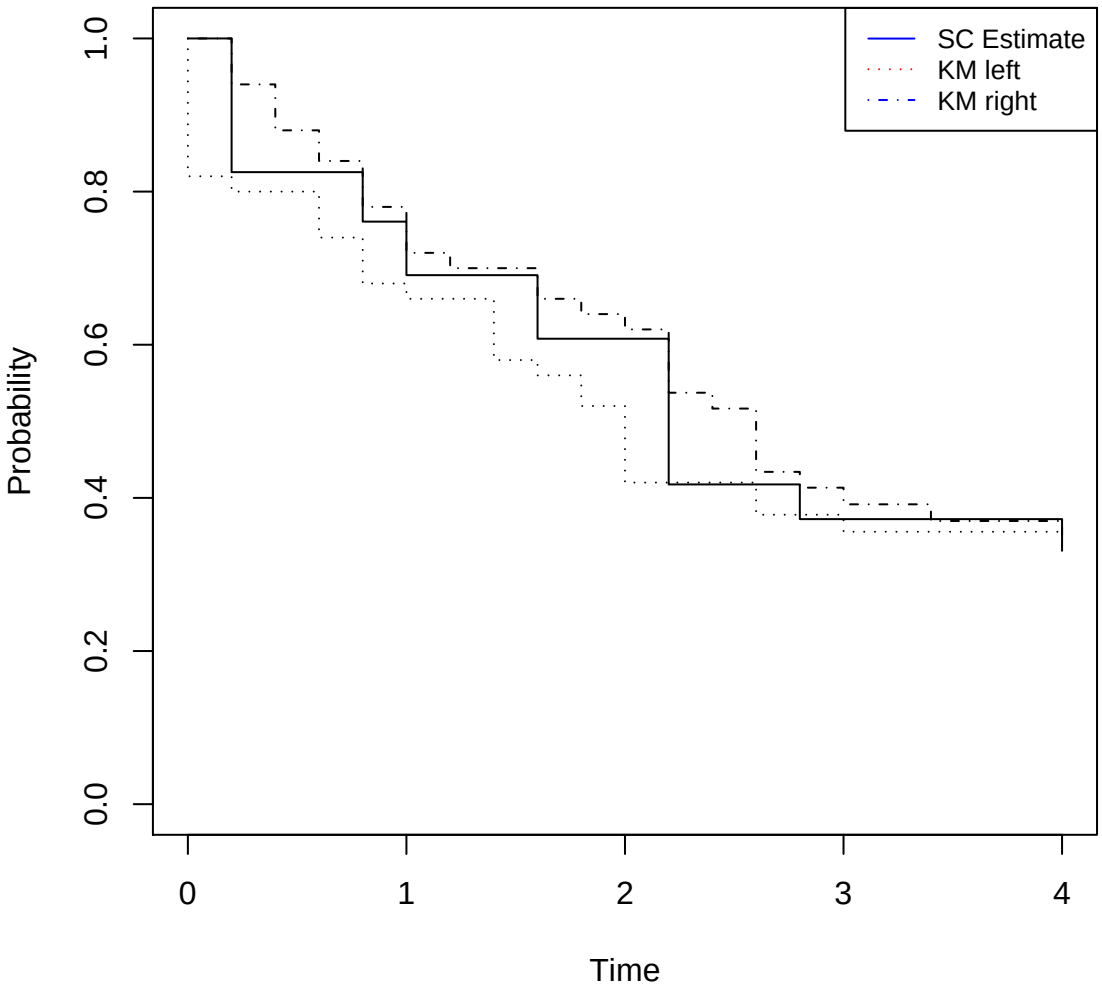


Figure 18: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates using an Exponential Hazard Function with mean 3. This is an example where they cross

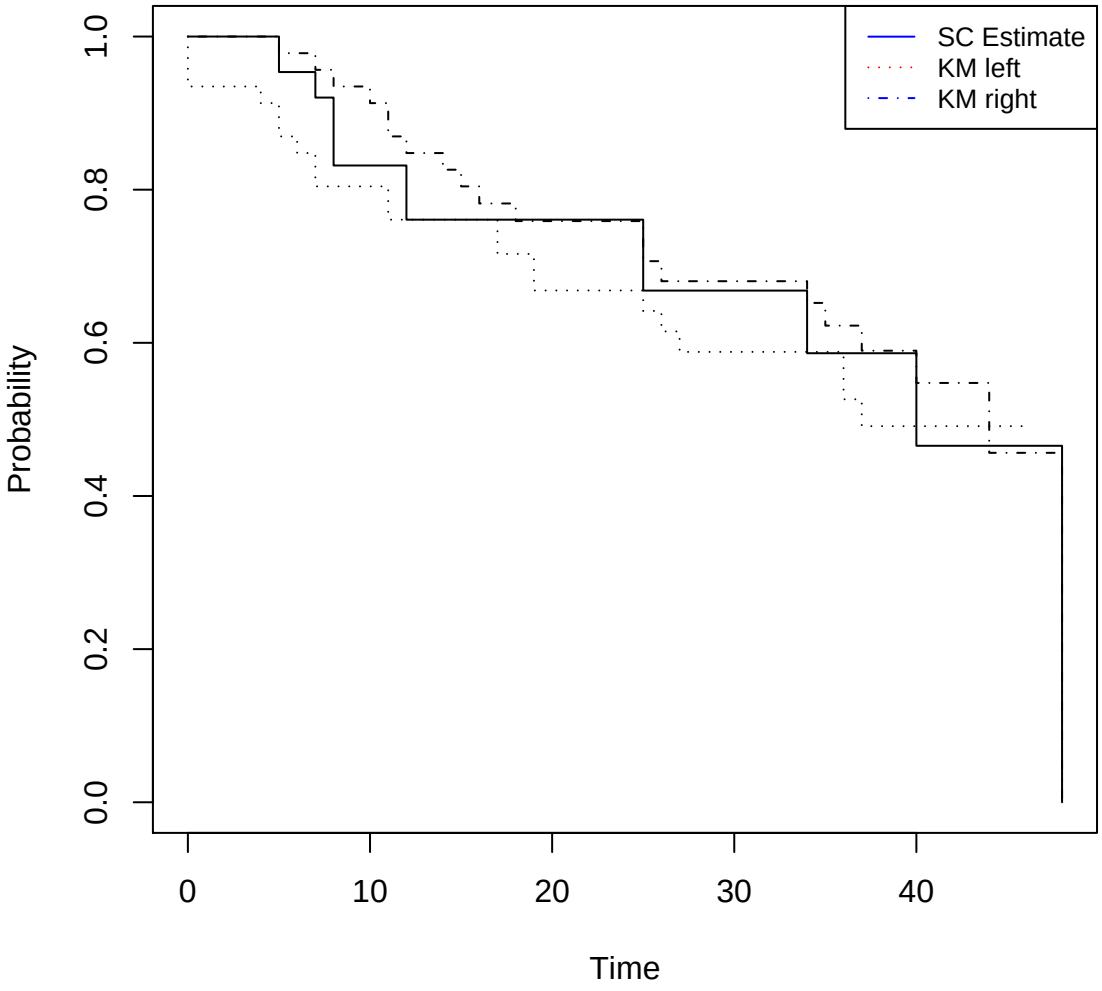


Figure 19: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates for the Radiation Group from the Caner Study in Finkelstein and Wolf

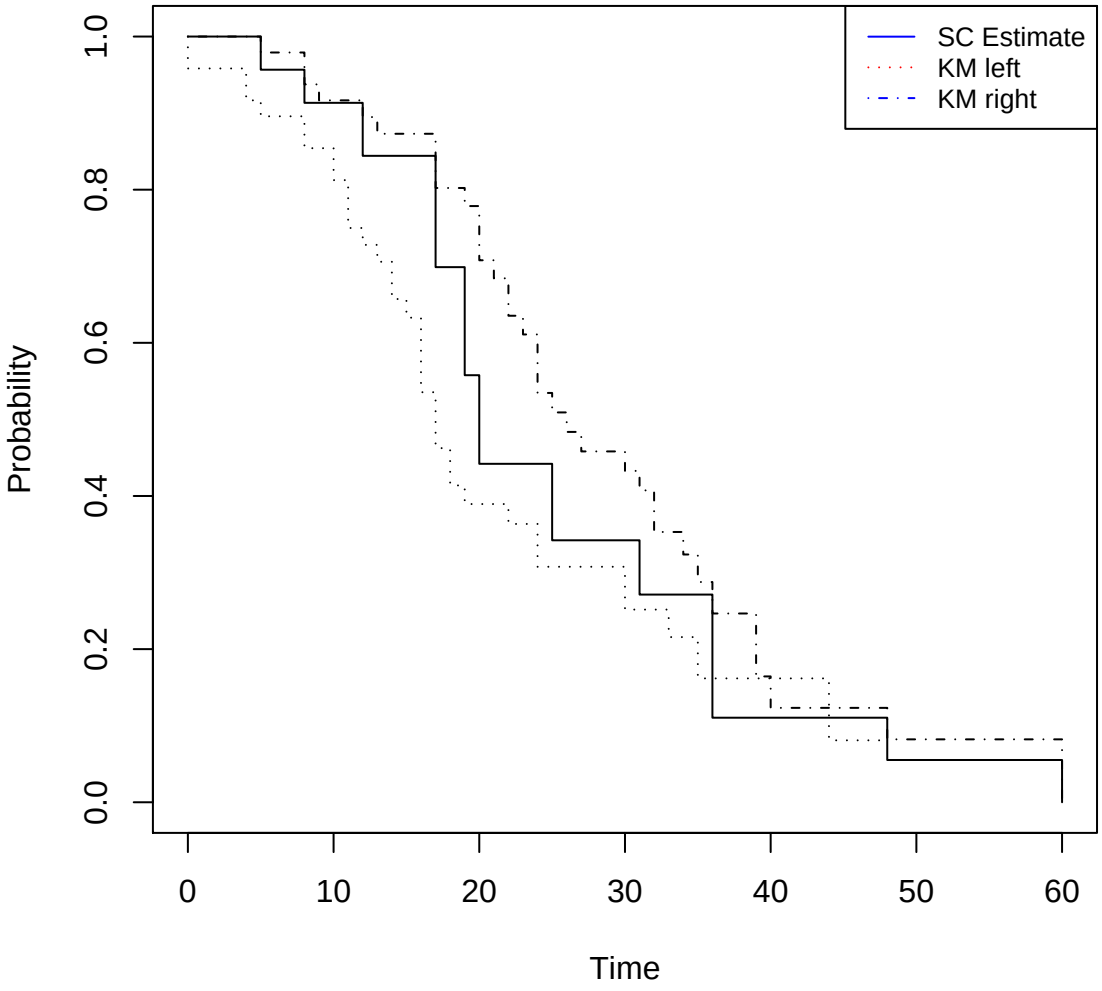


Figure 20: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates for the Chemotherapy Group from the Caner Study in Finkelstein and Wolf

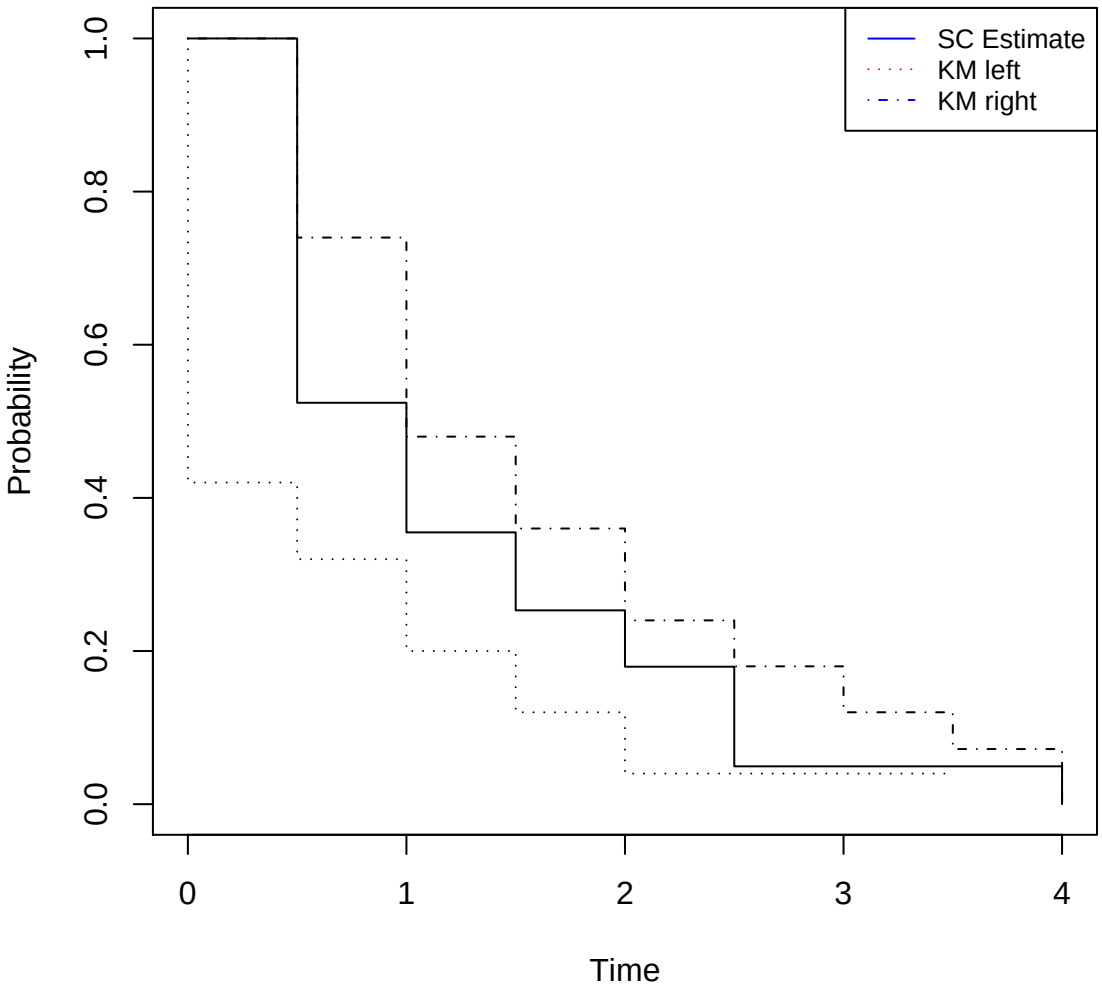


Figure 21: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates for patients with a non-low CD4 count from the AIDS study in Goggins and Finklestein

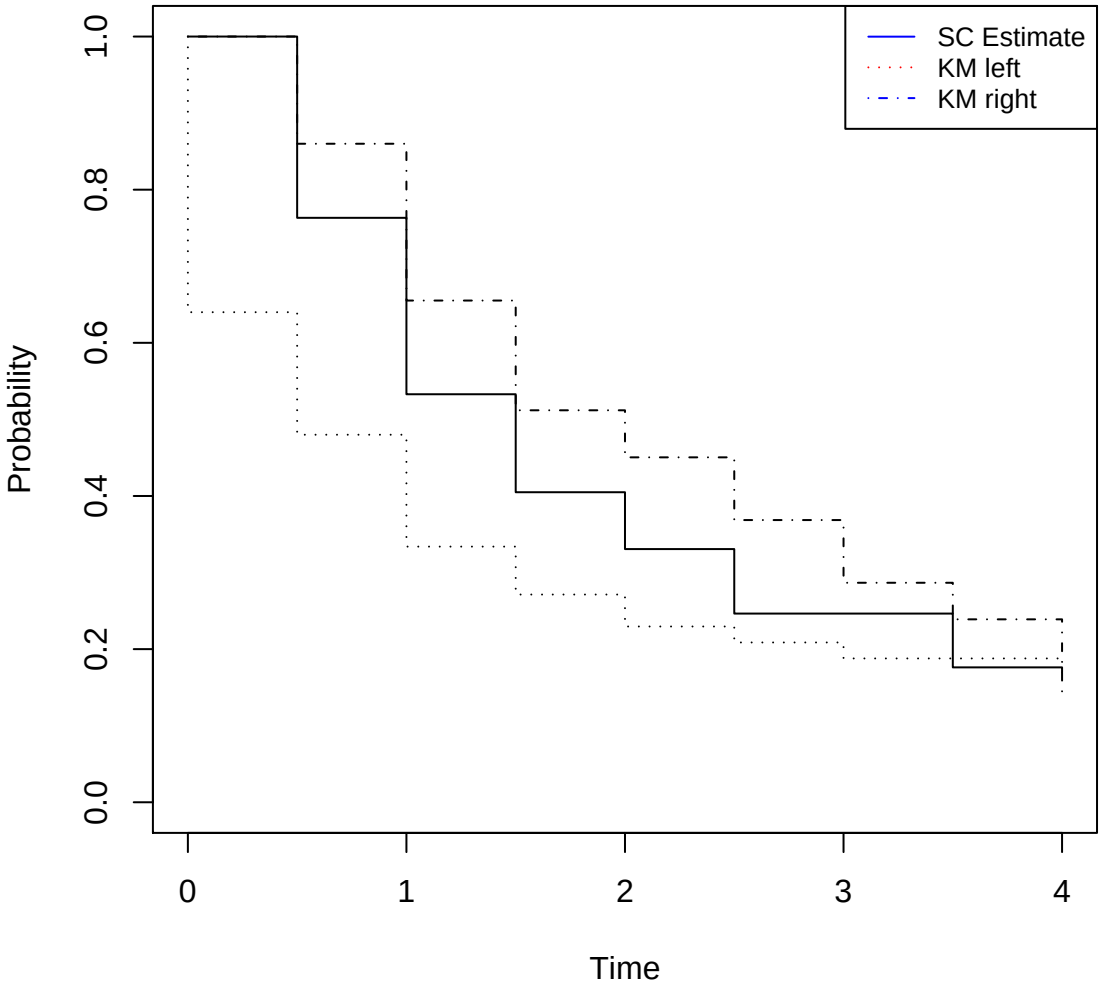


Figure 22: Plot of the Self Consistency Estimate and the Kaplan-Meier Right and Left Imputed Estimates for patients with a low CD4 count from the AIDS study in Goggins and Finklestein

VITA

Jeremy Gorelick was born on August 15, 1981 in Minneapolis Minnesota. He received his B.S. in applied mathematics from the University of Missouri-Rolla in 2004. Then, he became a graduate student of statistics at the University of Missouri-Columbia. He received his M.A. in statistics in 2007, and continued his study for a doctoral degree. He is going to graduate in 2009.