Fall 2017

Survival Analysis: A Modified Kaplan-Meir Estimator

Justin A. Bancroft
Missouri State University, Justin145@live.missouristate.edu

SURVIVAL ANALYSIS: A MODIFIED KAPLAN-MEIR ESTIMATOR

A Masters Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Mathematics

By

Justin A. Bancroft

December 2017

SURVIVAL ANALYSIS: A MODIFIED KAPLAN-MEIR ESTIMATOR

Mathematics

Missouri State University, December 2017

Master of Science

Justin A. Bancroft

ABSTRACT

The popular Kaplan-Meir estimator has traditionally been used to great effect as a survival function estimator. However, the Kaplan-Meir estimator is dependent upon a maximum likelihood parameter estimator which may not be the best estimator in all cases. We modify the Kaplan-Meir estimator, based on a Bayes parameter estimation, in hopes of providing a more accurate survival estimator for small sample sizes. Core elements of survival analysis are presented, acting as a foundation from which to construct and compare our modified Kaplan-Meir estimator. It is hypothesized that our modified Kaplan-Meir estimator is generally more accurate than the standard Kaplan-Meir estimator for smaller sample sizes, while the standard Kaplan-Meir estimator remains appropriate for larger sample sizes. Both Kaplan-Meir estimators are compared to theoretical distributions, with the traditional expectation that theoretical distributions will model data best if data can be fitted to a theoretical distribution. In order to show validity for our hypothesis one smaller data set and one larger data set were analyzed. The results of the analysis appeared to agree with our hypothesis.

KEYWORDS: survival analysis, Kaplan-Meir estimator, modified Kaplan-Meir estimator, censored data, hazard plotting, life table

This abstract is approved as to form and content

_____

Yingcai Su, PhD
Chairperson, Advisory Committee
Missouri State University

SURVIVAL ANALYSIS: A MODIFIED KAPLAN-MEIR ESTIMATOR

By

Justin A. Bancroft

A Masters Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Mathematics

December 2017

Approved:

_____
Yingcai Su, PhD

_____
George Mathew, PhD

_____
Songfeng Zheng, PhD

_____
Julie Masterson, PhD: Dean, Graduate College

ACKNOWLEDGEMENTS

The completion of this work would have not been possible without the guidance and support of family, friends and mentors. I am fortune to have received their encouragement and contributions. Of the many people who supported me in this process, I would like to acknowledge the group who has been the most influential.

Dr. Yingcai Su's patience and extensive knowledge was invaluable in this process. He gave generously of his time and focus. His insight's served as a core foundation for my work. I am sincerely thankful for his role as my advisor.

I would like to thank all the professors who have made an impact in shaping my educational development. My knowledge of statistics and mathematics has grown proportional to their passion and dedication in sharing that knowledge.

Lastly, I would like to thank my family and friends for their unyielding encouragement. I would have not come this far without them. Specifically, my parents were eternally patient is this process, and continued in their steadfast support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Importance of the Kaplan-Meir Estimator

Assuming no theoretical distribution when modeling data, the Kaplan-Meir estimator serves an important role in survival analysis. As outlined by Lee (1992), the Kaplan-Meir estimator can be used for a variety of purposes. It serves as a starting point from which to choose an appropriate theoretical distribution, and it can be used as a predictive distribution when no known theoretical distribution can be modeled.

Because the Kaplan-Meir estimator is one of the most frequently used methods in survival analysis, we seek to improve upon the current model by modifying the estimator. Our modification of the Kaplan-Meir estimator makes use of a Bayes parameter estimation. Although the Bayes estimation has bias and thus extends that bias to our modified Kaplan-Meir estimator, we believe that the estimator is still more accurate in certain circumstances, in particular, when the sample size is small.

To fully understand our modified Kaplan-Meir estimator, a background in survival analysis is necessary. We will begin by presenting the core concepts and foundations of survival analysis, which will ultimately lead to the presentation of our modified Kaplan-Meir estimator. Understanding survival analysis starts with the most fundamental of questions, "what is survival analysis?”

Survival Analysis

Survival Analysis is a branch of statistics that attempts to predict the time until one or more events takes place. It is an essential part of medical research. It’s statistical

methods have been extended to fields such as sociology, industry, economics, and ecology. Researchers use survival analysis to answer questions regarding time related data. As an example, a medical researcher may be interested in which treatment is most effective over a period of time or how long a condition lasts. In order to answer researcher questions, data is gathered, statistical methods are employed, and the results are analyzed.

Definitions

In order to establish the fundamentals of survival analysis, we first present some essential definitions. In survival analysis, survival data refers to information regarding the data set, this includes, survival times, patient characteristics, response to treatments, censoring, etc. The most crucial components of survival data are the observed values or the survival times.

Survival time is the measure of how long a subject has "survived" from a starting point to an ending point. Survival does not have to be literally interpreted. Here survival means a subject is in one state or condition that represents the default condition. The subject will remain in that state until an event of interest occurs.

Failure is the event of interest that marks the end of the survival time. Failure is usually death or some negative experience. However in some cases, failure may be positive such as disease remission. Failure may also be referred to as "the event", or "death" if death represents the failure.

Censoring

A sometimes challenging aspect of survival analysis is censoring. Censoring occurs when the exact survival time of a subject is unknown. It may happen that a subject drops out of a study, survives until the end of the study, or for some other reason is not precisely observed at their failure time.

An observation that does not have an exact survival time is called a censored observation. If an observation does have an exact survival time, then it is called an exact or uncensored observation. A censored observation's survival time is measured as the time under observation. The censored survival time usually receives a plus next to its survival time to indicate that its true survival time may surpass the recorded time. If a data set contains censored observations, then it is referred to as censored data. Data that has no censored observations is referred to as uncensored data.

There are several different types of censoring including, right censoring, left censoring, and interval censoring. Among these, right censoring is the most common. Right censoring occurs when the time of failure is unknown. This may happen if a subject experiences failure after the end of the study, a subject is withdrawn from the study, or a subject is lost to follow-up. Left censoring takes place if the true beginning of a subject's survival time is unknown. It may be that the subject had a condition that started at an unknown time or that it is not known when the subject entered the study. Interval censoring happens when survival data is organized in intervals such that it is impossible to determine when in the interval a subject might have failed.

Since right censoring is very common, there are methods specifically designed to handle right censored data. These methods are Type I, Type II, and Type III censoring. Each method is a way of handling right censored data for a specific study format.

In Type I censoring, a study starts with a fixed number of subjects that each have the beginning of the study as the starting point of a subject's observation period. The study lasts for a fixed amount of time and then ends. Any subject that experiences failure within the time frame of the study is given an uncensored survival time. If a subject fails accidentally, is withdrawn, or is lost to follow up, then that subject's survival time is censored. Any subject that has survived until the end of the study is also given censored survival times.

For Type II censoring, the subjects all start being observed at the beginning of the study, but the study ends when a certain amount of subjects experience failure. If a subject is not observed for the entire study, aside from the cause of failure, then that subject's survival time is censored. Censoring also occurs for any subject who has survived until the end of the study.

In one of the most common type of studies, the study period starts and ends at a specific time, but subjects may enter the study at any time. This means that survival times may start at different points in real time. Type III censoring occurs in this study type of study. The censoring occurs in the exact same way as in the previous studies presented. However, in this study, it is possible that two subjects start the study at different times. If they both survive until the end of the study, then their respective censored survival times will be different even though the study ended at the same time for both of them. In the previous type of studies, this situation would have not occurred since all subjects entered

the study at the beginning, and thus would have the exact same survival times if censored due to survival until the end of the study.

Censored data generally complicates survival analysis and needs to be carefully addressed. If there exists censored data in a study, then it often will affect how a distribution models the data. For the most part censored data only affects a modeled function in a very limited fashion. Examples of this will be shown latter.

Survival Time Functions

The beginning of the survival analysis process starts with survival time functions. Although survival time can refer to how long a specific subject has survived, it also can refer to how long a random subject might survive. This means that survival time is sometimes used to describe a random variable, denoted T. For clarity, we will use the term "survival time variable" in reference to T.

Probability Function. Like any other statistical variable, the survival time variable follows a probability distribution. The distribution is the probability that T will occur at a specific time t or close to a specific time t, for the continuous case. In essence, this distribution measures the probability of failure at any given time. The variable T follows the probability distribution

$$f(t) = P(T = t), \text{ for the discrete case,}$$

and

$$f(t) = \lim_{dt \to 0} \frac{P(t \le T \le t + dt)}{dt}, \text{ for the continuous case.}$$

Survival Function. While the probability function serves as a basis for survival time functions, the most commonly used and analyzed distribution in survival analysis is the survival distribution. The name suggests the meaning. The survival distribution measures the probability that a subject survives at a particular time. This is the same as the probability a subject will fail at a later time. The survival distribution is mathematically defined as

$$S(t) = P(T > t).$$

It can also be defined as

$$S(t) = 1 - F(t).$$

Kleinbaum and Klein (1996) describe several important attributes of the theoretical continuous survival distribution. The theoretical $S(t)$ should be a strictly decreasing function, with domain $(0, \infty)$. At $t = 0$, $S(t) = 1$ and at $t = \infty$, $S(t) = 0$. This would indicate that, at the beginning of the study, no subject has experienced failure. But, if the study continued indefinitely every subject will have experienced failure. Note, however, that these are the theoretical attributes of a continuous $S(t)$.

In practice, $S(t)$ must be estimated. The estimate of $S(t)$ is referred to as the survival estimate, and most survival analysists denote this estimation as $\hat{S}(t)$. It is possible for $\hat{S}(t)$ to be discrete and not strictly abide by its theoretical attributes. Examples of $\hat{S}(t)$ being a step function rather than a smooth curve will later be see. Additionally, since the end of a study is finite, the domain of $\hat{S}(t)$ may have a finite right end point. Often, t values beyond the end of the study are not considered. This means that a study may end and some subjects will have never experienced failure.

Hazard Function. The hazard rate function is another important survival time function. For any particular time, the hazard function gives the rate that a failure might occur per unit time, given that the subject has survived up until that same time. In essence the hazard function measures the likelihood of death at some time, given survival of previous times. The hazard function is defined mathematically as

$$h(t) = P(T = t \mid T \geq t), \text{ for the discrete case,}$$

and

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} \text{ for the continuous case.}$$

Sometimes it can be more useful to analyze the cumulative hazard function, which is defined as

$$H(t) = \int_0^t h(x)dx .$$

It will shortly be shown that the cumulative hazard function can also be written as

$$H(t) = -\log_e S(t) .$$

Thus, at $t = 0$, $H(t) = 0$ and at $t = \infty$, $H(t) = \infty$.

The cumulative hazard function is somewhat difficult to describe and comprehend. It is the sum of the risks up to time t. Another way to describe it would be to define $H(t)$ as the probability that a subject has survived to time t, by surviving the cumulative risks described by $h(x)$ along the way.

Regardless of the cumbersome interpretation, the cumulative hazard function serves a crucial role as an opposite to the survival function. Both are cumulative in

nature, but one describes the chance of survival over time, while the other describes the chance of death over time.

Relationship between Survival Functions. As outlined by Lee (1992), there are close relationships between all the survival functions presented thus far. If one of the survival functions is given, then the others can be derived. These relationships can be useful in theoretical analysis and in practice.

If the definition of $h(t)$ is consider, it can be seen that all of the survival functions are mathematically related. Recall that, for the discrete case,

$$h(t) = P(T = t \mid T \geq t).$$

One can expand $h(t)$ so that

$$h(t) = P(T = t \mid T \geq t) = \frac{P(T = t, T \geq t)}{P(T \geq t)} = \frac{P(T = t)}{P(T \geq t)} = \frac{f(t)}{S(t)}.$$

(This relationship is also true of the continuous case.) Also, remember that $S(t) = 1 - F(t)$. Thus,

$$F(t) = 1 - S(t)$$

and

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t).$$

Then $h(t)$ can be written as

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\ln S(t).$$

Since $H(t) = \int_0^t h(x)dx$,

$$H(t) = \int_0^t -\frac{d}{dt} \ln S(x) dx = -\ln S\left(t\right).$$

It is now clearly seen that each survival time function has a mathematical equivalency.

Exponential Survival Time Functions. Suppose, for example, that one wanted to find the survival time functions of a survival time T which followed an exponential distribution. For the exponential distribution, it is known that

$$f(t) = \lambda e^{-\lambda t} \text{ and } F(t) = 1 - e^{-\lambda t}$$

where $t \geq 0$ and $\lambda > 0$. The survivor function is

$$S(t) = 1 - F(t) = 1 - \left(1 - e^{-\lambda t}\right) = e^{-\lambda t}.$$

Thus,

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

and

$$H(t) = -\ln S\left(t\right) = -\ln\left(e^{-\lambda t}\right) = \lambda t.$$

The survival time distributions for every other probability distribution can be found in this same manner.

ESTIMATION OF SURVIVAL FUNCTIONS

As was mentioned previously, the survival function is the most commonly used of the survival time functions. We know that in determining one survival time function, the others can be determined. Thus, most of our focus will be spent on the survival function estimates.

There are various methods of estimating survival functions. These methods are typically split into two categories, non-parametric and parametric. Parametric methods are based on estimating parameters of known distributions such as the exponential, Weibull, log-normal, gamma, and normal distributions. The parametric methods assume the data follow a known distribution. The non-parametric methods do not assume any known distribution. While there are parameters to be estimated, the shape of the distribution is not restricted to a predetermined outcome. It only follows the approximate trend of the data.

Both methods are good in particular circumstances. The nonparametric distribution always has some value. It will be better than a parametric method if no known distribution can be chosen to fit the data or if the fit is not very good. The nonparametric method can also be beneficial in choosing an appropriate known distribution based on the shape of the nonparametric distribution.

The parametric method is only useful if the trend of the data roughly matches the shape of a known distribution. While this can be limiting in certain circumstances, it is also very powerful when a known distribution can be matched. If the data matches well,

then a known distribution should be far better at modeling the data than a non-parametric distribution.

Parametric Methods

We first focus on the parametric method. There are two main components of the parametric process. The first is finding the appropriate known distribution to model the data. The second is choosing an appropriate method to estimate the parameter.

In order to find an appropriate known distribution, the hazard plot will be presented and later it will be shone how to analyze the shape of a non-parametric distribution during our data analysis. If censoring was not present, then a probability plot could be utilized. However, in most studies, censoring exists. Thus, a probability plot will not be presented.

There is a large collection of methods to choose from in estimating a parameter. These include linear regression, logistic regression, poison regression, the method of moment estimation, and the maximum likelihood estimation. Each method has its respective benefits and value, but our focus will remain on the frequently used maximum likelihood estimation (MLE).

Hazard Plotting

Hazard plotting was first presented by Nelson (1972). A hazard plot is a method that compares graphs and functions. An estimated cumulative hazard function is derived from the data values. A graph is constructed by plotting data values against an estimated

cumulative hazard function. Then the estimated cumulative hazard function and corresponding graph can be compared to theoretical functions and their respective graphs.

Hazard Table. A table can be created to organize the data and find the estimated cumulative hazard values. Table 1 is the format of such a table. A description of each column is shown below.

Table 1. Hazard Table Format

Survival Times	Number at Risk	Hazard Values	Cumulative Hazard Values
t_1	n_1	$\hat{h}(t_1)$	$\hat{H}(t_1)$
t_2	n_2	$\hat{h}(t_2)$	$\hat{H}(t_2)$
.	.	.	.
.	.	.	.
.	.	.	.
t_i	n_i	$\hat{h}(t_i)$	$\hat{H}(t_i)$
.	.	.	.
.	.	.	.
.	.	.	.
t_{n-1}	n_{n-1}	$\hat{h}(t_{n-1})$	$\hat{H}(t_{n-1})$
t_n	n_n	$\hat{h}(t_n)$	$\hat{H}(t_n)$

The first column is survival times, including censored times, listed from least to greatest. That is, if we let t_i be a survival time, then $t_1 \le t_2 \le ....t_n$, where n is the sample size. If two of the survival times are the same, then they are listed in random order regardless of censoring. In practice a plus is given to any censored survival time.

In the second column, n_i is the number of subjects alive and at risk of dying at t_i just before t_i, where $0 \le n_i < n$. Naturally, n_i would become smaller with each subsequent t_i so that $n_1 \ge n_2 \ge ....n_n$. Note that $n_n \ne n$. The value n_n is the number of subjects still at risk at the t_nth observation, and would be the smallest of the n_i. Some researchers and authors might use a different notation for n_i to avoid this confusion.

The third column is the percentage of individuals who have failed at time t_i. These percentages are estimations of the hazard values for the data set and are designated, $\hat{h}(t_i) = 100/n_i$. For $\hat{h}(t_i)$, $1/n_i$ is multiplied by 100 to convert from a decimal to a percentage. The multiplication by 100% is a matter of preference. Censored survival times will not have hazard values.

The fourth column is the cumulative hazard values. A cumulative hazard value is denoted $\hat{H}(t_i)$, and is the sum of the hazard value at t_i and all previous hazard values. Censored survival times will not have cumulative hazard values.

Hazard Graphs. Once all the cumulative hazard values are found, t_i can be plotted against the $\hat{H}(t_i)$ values to form one of our graphs. The x-values of the graph are the $\hat{H}(t_i)$ values, while the y-values of the graph are the t_i values. Notice that the x-values and y-values are switched from what might be expected.

The inverse cumulative hazard function of theoretical distributions, denoted $G_T\left(H_T\left(t\right)\right)$, can be compared with the graph of the estimated cumulative hazard values. Researchers should determine if the function $G_T\left(H_T\left(t\right)\right)$ is linear, exponential, logarithmic, or some other shape. Similarities between the shape of the estimated cumulative hazard graph and $G_T\left(H_T\left(t\right)\right)$ are desired, not specific accuracies.

The general shape of $G_T\left(H_T\left(t\right)\right)$ can be plotted against H. The plots of some Weibull cumulative hazard distributions are shown in Figure 1.

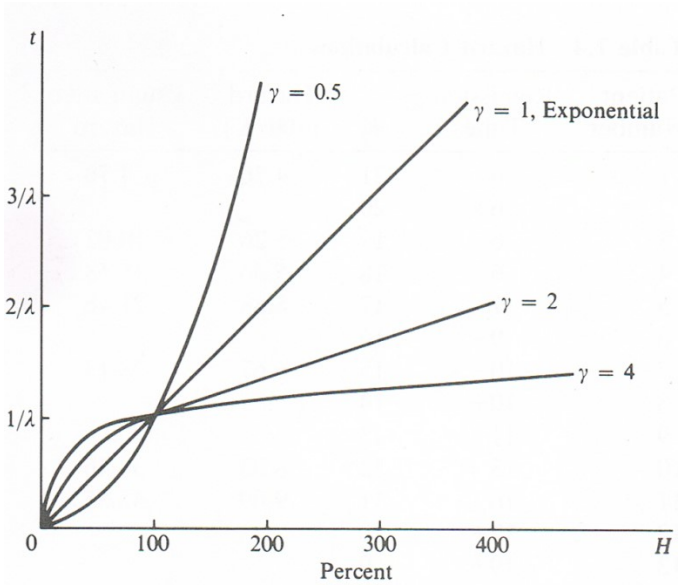


Figure 1. Weibull Cumulative Hazard Functions (Lee 1992, pg. 175). Permission was given to reproduce this image.

Weibull distributions are commonly chosen to model survival data. Notice that the exponential distribution is a Weibull distribution where $\gamma = 1$ and is linear in nature.

If the general shape of the estimated cumulative hazard distribution and theoretical cumulative hazard distributions match in their curvature, then the theoretical

distribution analyzed could be chosen. If they do not match, then a different theoretical distribution should be analyzed.

It should be noted that, when comparing the cumulative hazard graphs, some cases provide graphs with obvious conclusions about the appropriateness of a distribution. Other cases are less clear. It is up to the researcher to determine if the data fits "well enough" to model the data for a particular distribution.

As an example of determining an appropriate distribution, the shape of a cumulative hazard graph for an exponential function is known to be linear. This can be seen by analyzing the inverse cumulative hazard function for the exponential distribution

Suppose we have the exponential distribution $f(t) = \lambda e^{-\lambda t}$. It has been seen previously that $h(t) = \lambda$. Thus,

$$H(t) = \int_0^t H(x)\,dx = \int_0^t \lambda\,dx = \lambda t$$

Solving for t, we have

$$t = \frac{1}{\lambda} H(t)$$

Thus,

$$G(H(t)) = \frac{1}{\lambda} H(t)$$, which is linear.

Additionally, it can be seen from Figure 1 that the exponential distribution has a linear cumulative hazard graph. In this case, if the cumulative hazard graph estimated from the data is linear, then one could choose the exponential distribution to model the data.

Now that a method to finding an appropriate known distribution has been presented, an estimate for a specific parameter can be found.

Maximum Likelihood Estimation for a Parametric Distribution

As stated previously, the maximum likelihood estimation is a method of estimating a parameter. The MLE method here is similar to the MLE method presented in many statistical courses. The maximum likelihood function must be found which is defined as

$$L = \prod_{i=1}^{n} f(x_i \mid \theta)$$

where each x_i is independent and identically distributed (I.I.D) from a distribution $f(x)$ that belongs to a particular family of distributions, such as the exponentials. A parameter estimator is then found through differentiation that maximizes the likelihood function. Sometimes the log of the likelihood function is used instead with the same results.

It would be expedient if this process was all that was needed. However, in survival analysis, censoring complicates this process so that some preliminary work must be done.

Parametric Probability Distribution. Before the MLE is applied to determine an appropriate estimator, it needs to show what a parametric distribution would look like when censoring is considered, and in particular how a study with type III censoring would affect that distribution. The set up and construction of the distribution is shown by Le (1997). An adaption of this process is shown as follows.

Assume there is a data set with type III censored data and a sample size of n. If death/failure is observed for the ith subject, then the associated time variable is designated as T_i for $i = 1, 2, ..., n$. It is assumed that T_i follows a probability distribution of $f(t)$ with one or more parameters $\theta_1, \theta_2, ..., \theta_n$ and has a survival function of $S(t)$.

If death/failure has not occurred for the ith survival observation and the data is censored, then the associated survival variable is designated as C_i, for $i = 1, 2, ..., n$. It is assumed that C_i follows a uniform probability distribution of $g(t)$.

Most studies have an enrollment period of $(0, \pi_1)$ and a follow up period of (π_1, π_2). The uniform distribution intervals for $g(t)$ are usually $(\pi_2 - \pi_1, \pi_2)$. It is assumed that $g(t)$ has a survival function of $R(t)$.

The survival time functions are mathematically defined as follows:

$$S(t) = P(T > t)$$

$$f(t)dt = P(t < T < t + dt)$$

$$R(t) = P(C > t)$$

$$g(t)dt = P(t < C < t + dt)$$

Let t_i be the lifetimes of the data determined either by death/failure or censoring. Define δ_i as a censoring indicator such that

$$\delta_i = \begin{cases} 0 & \text{if censored due to study termination} \\ 1 & \text{if death/failure was observed.} \end{cases}$$

It would follow that if $\delta_i = 0$, then censoring took place for the ith sample and $t_i = C_i$. If $\delta_i = 1$, then death/failure was observed for the ith sample and $t_i = T_i$. It is assumed that the deaths or censoring for each sample have no relation to one another so that each T_i and C_i would be stochastically independent of each other.

The objective is to determine the probability distribution for both censored and uncensored data. To do that, two different probability distributions need to be considered.

After that a combination of the two distributions will be shown. The first distribution, $f(t,\delta_i=1)$, is designed for uncensored observations, while the second distribution, $f(t,\delta_i=0)$, is designed for censored observations.

Probability Distribution for Uncensored Observations. In order to find $f(t,\delta_i=1)$, $P(t\le T\le t+dt,\delta_i=1)$ needs to be determined. If $\delta_i=1$ is observed for the ith sample, this means that death\failure occurred at time t_i or $t_i=T_i$. It also means that the maximum observation time C must have occurred later then the death or failure. That is $C_i>T_i=t_i$. Thus, the probability that T occurs with a lack of censoring is

$$\begin{aligned}P(t\le T\le t+dt,\delta_i=1) &= P(t\le T\le t+dt, C>t)\\ &= P(t\le T\le t+dt)P(C>t)\\ &= f(t)dt\cdot R(t).\end{aligned}$$

Therefore,

$$f(t,\delta=1)=\lim_{dt\to 0}\frac{f(t)dt\cdot R(t)}{dt}=f(t)\cdot R(t).$$

Probability Distribution for Censored Observations. In order to find $f(t,\delta_i=0)$, consider $P(t\le T\le t+dt,\delta_i=0)$. If $\delta_i=0$ is observed for the ith sample, then censoring of the study occurred at time t_i or $t_i=C_i$. It also means that death\failure , T, must have occurred later than the maximum observable time (or end of the study), C. That is $T_i>C_i=t_i$. Thus, the probability that T occurs with censoring is

$$\begin{aligned}P(t\le T\le t+dt,\delta_i=0) &= P(t\le C\le t+dt, T>t)\\ &= P(t\le C\le t+dt)P(T>t)\\ &= g(t)dt\cdot S(t).\end{aligned}$$

Therefore,

$$f(t,\delta_i=0)=\lim_{dt\to 0}\frac{g(t)dt\cdot S(t)}{dt}=g(t)\cdot S(t).$$

Probability Distribution for All Observations. The distributions for uncensored and censored observations can be combined into one function based on whether a particular observation is censored or uncensored. The combined probability distribution is

$$f(t,\delta_i)=\left[f(t)g(t)\right]^{\delta_i}\left[R(t)S(t)\right]^{1-\delta_i}.$$

This probability distribution can be simplified further. Recall that, $h(t)=\frac{f(t)}{S(t)}$ and hence $f(t)=h(t)S(t)$. Thus, $h(t)S(t)$ is substituted for $f(t)$. (This will later be crucial when finding estimators using the MLE.)

$$\begin{aligned}f(t,\delta_i)&=\left[f(t)g(t)\right]^{\delta_i}\left[R(t)S(t)\right]^{1-\delta_i}\\&=\left[h(t)S(t)g(t)\right]^{\delta_i}\left[R(t)S(t)\right]^{1-\delta_i}\\&=h(t)^{\delta_i}S(t)^{\delta_i}g(t)^{\delta_i}R(t)^{1-\delta_i}S(t)^{1-\delta_i}\\&=g(t)^{\delta_i}R(t)^{1-\delta_i}h(t)^{\delta_i}S(t)^{\delta_i}S(t)\\&=g(t)^{\delta_i}R(t)^{1-\delta_i}h(t)^{\delta_i}S(t).\end{aligned}$$

Therefore,

$$f(t,\delta_i)=g(t)^{\delta_i}R(t)^{1-\delta_i}h(t)^{\delta_i}S(t).$$

Since the parameters of $g(t)$ and $R(t)$ are not the parameters to be estimated, it turns out that $g(t)^{\delta_i}$ and $R(t)^{1-\delta_i}$ will bear no relevance when finding the MLE and thus will be dropped. The final simplified version of $f(t,\delta_i)$ is

$$f(t,\delta_i) = h(t)^{\delta_i} S(t).$$

Estimating an Exponential Parameter. Our purpose in finding $f(t,\delta_i)$, was to use the MLE to find estimators for parameters when censoring was considered. In order to demonstrate how to find the MLE in practice, the MLE for the exponential distribution will be found.

Suppose that there is a survival time T, which belongs to an exponential distribution. Let $t_1, t_2, ..., t_n$ be the censored or uncensored observed lifetimes, with a sample size of n. It is known that $f(t) = \lambda e^{-\lambda t}$ and from previous work it was found that $S(t) = e^{-\lambda t}$ and $h(t) = \lambda$. Thus, we know that the probability distribution which considers censoring is

$$f(t,\delta_i) = h(t)^{\delta_i} S(t) = \lambda^{\delta_i} e^{-\lambda t}.$$

Now the MLE of the parameter λ can be found. The likelihood function of $f(t_i,\delta_i)$ is

$$L(\lambda) = \prod_{i=1}^{n} f(t_i,\delta_i) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i} = \lambda^{\sum_{i=1}^{n}\delta_i} e^{-\lambda\sum_{i=1}^{n} t_i}.$$

Taking the log of $L(\lambda)$ results in

$$\log L(\lambda) = \log\left[\lambda^{\sum_{i=1}^{n}\delta_i} e^{-\lambda\sum_{i=1}^{n} t_i}\right]$$

$$= \log \lambda^{\sum_{i=1}^{n}\delta_i} + \log e^{-\lambda\sum_{i=1}^{n} t_i}$$

$$= \log \lambda^{\sum_{i=1}^{n}\delta_i} + \log e^{-\lambda\sum_{i=1}^{n} t_i}$$

$$= \sum_{i=1}^{n} \delta_i \log \lambda - \lambda \sum_{i=1}^{n} t_i \, .$$

Taking the partial derivative of $\log L(\lambda)$ gives

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = \frac{\partial}{\partial \lambda} \left(\sum_{i=1}^{n} \delta_i \log \lambda - \lambda \sum_{i=1}^{n} t_i \right) = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i \, .$$

Setting $\frac{\partial}{\partial \lambda} \log L(\lambda)$ equal to zero and solving for λ we have

$$\frac{\partial}{\partial \lambda} \log L(\lambda) = 0$$

$$\frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

$$\frac{\sum_{i=1}^{n} \delta_i}{\lambda} = \sum_{i=1}^{n} t_i$$

$$\lambda = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} \, .$$

Thus, the MLE estimator for λ is

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

indicating that

$$\hat{\lambda} = \frac{\text{number of deaths observed}}{\text{Sum of all observed suvival times}} \, .$$

Note that if all the observations are uncensored, then each $\delta_i = 1$, resulting in

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}.$$

However, if at least one observation is censored then $\hat{\lambda}$ becomes larger since, $\sum_{i=1}^{n} \delta_i \le n$.

Anytime an estimator or estimate is used a $\wedge$ will be placed above the function or parameter estimator(estimate). Additionally if any function comes from a theoretical distribution a subscript T will be placed next to the representative letter of the function. For example, $\hat{f}_T(t)$, $\hat{S}_T(t)$, and $\hat{h}_T(t)$ would all be theoretical functions using an estimate for λ. The subscript T will be used to differentiate theoretical functions from Kaplan-Meir functions, which will be presented latter.

Variance of the Exponential Survival Estimator The survival function $S(t)$ is a common function that has many uses. Latter discrete estimates of $S(t)$ will be shown and compared to theoretical estimators of $S(t)$. One useful way to compare estimates of $S(t)$ is by comparing their respective variances. The variance of $\hat{S}_T(t)$ can be used to create confidence intervals.

Here Var $\hat{S}_T(t)$ will be determined for the popular exponential distribution. First, we need to take a look at the distribution of $\hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i$. Bartholomew (1957) found that $\hat{\lambda} \xrightarrow{D} N\left(\lambda, \lambda^2 / \sum_{i=1}^{n}\left(1 - e^{-\lambda T_i}\right)\right)$, where T_i is computed as the time between the ith subject entering the study until the end of the study. In many instances, this information is not available. Sometimes only the survival time since entering the study is

recorded. Thus, we will focus on an alternate distribution for $\hat{\lambda}$ that does not require knowledge about T_i. This alternative is $\hat{\lambda} \xrightarrow{D} N\left(\lambda, \lambda^2 / \sum_{i=1}^{n} \delta_i\right)$.

The delta method can be used to find the variance of different distributions based on the distribution of the parameters or variables. The delta method states

If $Y_n \xrightarrow{D} N(\mu, \sigma^2)$, $g'(\mu)$ exists, and $g'(\mu) \neq 0$, then

$$g(Y_n) \xrightarrow{D} N\left(g(\mu), \left[g'(\mu)\right]^2 \sigma^2\right).$$

We proceed to find $\operatorname{Var} \hat{S}_T(t) = \operatorname{Var}\left(e^{-\hat{\lambda} t}\right)$ using the delta method.

Let $Y_n = \hat{\lambda}$ and $g(Y_n) = e^{-\hat{\lambda} t}$. We know that $\hat{\lambda} \xrightarrow{D} N\left(\lambda, \lambda^2 / \sum_{i=1}^{n} \delta_i\right)$. Thus,

$$g'(Y_n) = -te^{-\hat{\lambda} t} = -te^{\left(\sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i\right) t} \text{ and } g'(\mu) = -te^{-\lambda t}.$$

It can clearly be see that g' exists if $\sum_{i=1}^{n} t_i \neq 0$, and $g'(\mu) \neq 0$ if $t \neq 0$. In practice, $\sum_{i=1}^{n} t_i \neq 0$ and $t \neq 0$. Thus,

$$e^{-\hat{\lambda} t} \xrightarrow{D} N\left(g(\mu), \left[g'(\mu)\right]^2 \sigma^2\right).$$

The values for $g(\mu)$ and $\left[g'(\mu)\right]^2 \sigma^2$ are

$$g(\mu) = e^{-\lambda t}$$

and

$$\left[g'(\mu)\right]^2 \sigma^2 = \left(-te^{-\lambda t}\right)^2 \frac{\lambda^2}{\left[\sum_{i=1}^{n} \delta_i\right]} = \frac{(\lambda)^2 t^2 e^{-2\lambda t}}{\left[\sum_{i=1}^{n} \delta_i\right]}.$$

Hence,

$$e^{-\hat{\lambda} t} \xrightarrow{D} N\left(e^{\lambda t}, \frac{(\lambda)^2 t^2 e^{-2\lambda t}}{\left[\sum_{i=1}^{n} \delta_i\right]}\right).$$

Therefore,

$$\operatorname{Var} \hat{S}_T(t) = \frac{(\lambda)^2 t^2 e^{-2\bullet}}{\left[\sum_{i=1}^{n} \delta_i\right]}.$$

Since λ_i is unknown, the estimator $\hat{\lambda}_i = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i$ can be used. In some theoretical circumstances, it may not be desirable to estimate λ_i. However, this substitution is useful in practice. The variance resulting from the substitution is

$$\operatorname{Var} \hat{S}_T(t) = \frac{\left(\sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i\right)^2 t^2 e^{-2\bullet}}{\left[\sum_{i=1}^{n} \delta_i\right]} = \frac{\left(\sum_{i=1}^{n} \delta_i\right) t^2 e^{\bullet}}{\left(\sum_{i=1}^{n} t_i\right)^2}.$$

The $\operatorname{Var} \hat{S}_T(t)$ for a different theoretical distribution could be found using the same method presented. The variance of the survival estimate can be found once the distribution(s) of the parameter(s) are determined and the delta method applied.

Confidence Interval for the Exponential Survival Estimator. A straightforward confidence interval for the exponential $\hat{S}_T(t)$ would be

$$\hat{S}_T(t) \pm z_{0.975}\text{Var}\,\hat{S}_T(t).$$

However, this confidence interval could result in impossible values outside of $[0,1]$. Recall that $S(t)$ can never be larger than 1 nor can it be less than 0. If $\hat{S}_T(t)$ is already close to 1 or 0, it is possible that adding or subtracting $z_{0.975}\text{Var}\,\hat{S}_T(t)$ would result in values greater than 1 or less than 0. To avoid these impossible results, a new confidence interval needs to be found that will always fit within $[0,1]$. Xu (2016) shows a method for finding a confidence interval that safely lies within $[0,1]$. A similar approach is adopted below.

First a modification of $\hat{S}_T(t)$ is made so that the range is between $(-\infty,\infty)$ instead of $[0,1]$.

$$0 \le \hat{S}_T(t) \le 1$$

$$-\infty \le \log \hat{S}_T(t) \le 0$$

$$0 \le -\log \hat{S}_T(t) \le \infty$$

$$-\infty \le \log\left(-\log \hat{S}_T(t)\right) \le \infty$$

If we add or subtract $z_{0.975}\text{Var}\left[\log\left(-\log_T \hat{S}(t)\right)\right]$ to $\log\left(-\log \hat{S}_T(t)\right)$ we arrive at a 95% confidence interval between $-\infty$ and ∞. That is,

$$-\infty \le \log\left(-\log \hat{S}_T(t)\right) \pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right] \le \infty .$$

We now have a confidence interval such that when using exponentials to convert back, the confidence limits safely lie in the range $[0,1]$. The resulting confidence interval will not be equivalent to the confidence interval for $\hat{S}(t)$, but will still be a fairly good representation.

$$\text{Let } \text{C.I.}\left[\log\left(-\log \hat{S}(t)\right)\right] = \log\left(-\log \hat{S}_T(t)\right) \pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right].$$

Working backwards using exponentials we have

$$-\infty \le \text{C.I.}\left[\log\left(-\log \hat{S}_T(t)\right)\right] \le \infty$$

$$0 \le e^{\text{C.I.}\left[\log\left(-\log \hat{S}_T(t)\right)\right]} \le \infty$$

$$-\infty \le -e^{\text{C.I.}\left[\log\left(-\log \hat{S}_T(t)\right)\right]} \le 0$$

$$0 \le \exp\left[-e^{\text{C.I.}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}\right] \le 1 .$$

Now we can find $\exp\left[-e^{\text{C.I.}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}\right]$.

$$\exp\left[-e^{\log\left(-\log \hat{S}_T(t)\right) \pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}\right] = \exp\left[\log \hat{S}_T(t)\, e^{\pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}\right]$$

$$= \hat{S}_T(t)^{e^{\pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}}$$

Note here that this process is valid for any estimate of $S(t)$. Thus, in the future the above confidence interval will be used as our standard confidence interval, where $S(t)$ can be replaced by any estimator.

Now we substitute $\hat{S}_T(t)$ for $e^{-\hat{\lambda}t}$, resulting in

$$\hat{S}_T(t)^{e^{\pm z_{0.975}\mathrm{Var}\left[\log\left(-\log \hat{S}_T(t)\right)\right]}} = \left(e^{-\hat{\lambda}t}\right)^{e^{\pm z_{0.975}\mathrm{Var}\left[\log\left(-\log\left(e^{-\hat{\lambda}t}\right)\right)\right]}}$$

$$= e^{-\hat{\lambda}t \cdot e^{\pm z_{0.975}\mathrm{Var}\left[\log\left(\hat{\lambda}t\right)\right]}}.$$

The delta method can be applied again to get the distribution of $\log\left(\hat{\lambda}t\right)$, resulting in

$$\mathrm{Var}\left[\log\left(\hat{\lambda}t\right)\right] \approx \frac{1}{\sum_{i=1}^{n}\delta_i}.$$

Thus, a suitable 95% confidence interval for the exponential $\hat{S}_T(t)$ would be

$$\left[e^{-\hat{\lambda}t \cdot e^{1.96\left(1/\sum_{i=1}^{n}\delta_i\right)}}, e^{-\hat{\lambda}t \cdot e^{-1.96\left(1/\sum_{i=1}^{n}\delta_i\right)}}\right].$$

Theoretical Survival Table. The theoretical survival table we introduce shows information regarding any theoretical survival estimate. Since $\hat{S}_T(t)$ is continuous, it is not ordinary to show the survival function at specific values. However, we show the survival function at specific values here because latter we will see discrete estimates of

$S(t)$ so that the table will be useful for comparison purposes. Table 2 shows the format of our Theoretical Survival Table.

Table 2. Theoretical Survival Table Format

Survival Times	$\hat{S}_T(t)$	Var $\hat{S}_T(t)$
t_1	$\hat{S}_T(t_1)$	Var $\hat{S}_T(t_1)$
t_2	$\hat{S}_T(t_2)$	Var $\hat{S}_T(t_2)$
.	.	.
.	.	.
.	.	.
t_i	$\hat{S}_T(t_i)$	Var $\hat{S}_T(t_i)$
.	.	.
.	.	.
.	.	.
t_{n-1}	$\hat{S}_T(t_{n-1})$	Var $\hat{S}_T(t_{n-1})$
t_n	$\hat{S}_T(t_n)$	Var $\hat{S}_T(t_n)$

Non-Parametric Methods

As was discussed previously, the crucial difference between a parametric and non-parametric distribution is that the parametric distribution assumes the data follows a

specific predetermined shape. A non-parametric distribution assumes no specific shape. Because of this, the non-parametric distributions will most often be discrete rather than continuous. This more readily allows for the data to match any sort of pattern without limitations. Additionally, since non-parametric methods are not limited to known distributions, parameter estimation receives exclusive attention.

There are many non-parametric methods we could present. However our focus will remain on two methods, the Kaplan-Meir (KM) estimation, and the modified Kaplan-Meir (MKM) estimation. The two distributions are essentially the same, except that the parameters are different. We will spend much of our focus on the modified Kapan-Meir estimation. In the future, this estimator will be compared to other survival estimators. However, we first present the standard Kaplan-Meir estimation

Kaplan-Meir Estimator

Developed by Kaplan and Meier (1958), the Kaplan-Meier estimator gives a discrete step function estimate that attempts to model survival data. It is formed based on finding discrete probabilities of survival at each observation, given survival occurred previously. These probabilities are basic and do not assume any pattern or shape. The product limit variation of the Kaplan Meir method will be presented since it is compatible with censored observations.

In the Kaplan-Meir method, the survival time of each subject is measured as the survival time since enrollment, so that each subject starts at $t = 0$. This avoids problems that could arise from patients enrolling at different times. Additionally, subjects who are censored for any reason are considered to have survived the study, and thus will not have

a measured survival time. The Kaplan-Meir process also has a specific way of handling data which is shown below.

If we let t_i be an arbitrary survival time, n be the total sample size including the censored observations, and k be the number of uncensored observations, such that $k \leq n$, then the survival times should be ordered from least to greatest. That is,

$$t_1 < t_2 < t_3 < ... < t_k .$$

Equivalent t_is are merged into one, so as to not be redundant, when finding values such as $f(t_i)$ and $S(t_i)$. However, equivalent t_is still each count towards totals such as k and n.

Note again, that the censored t_is will not be shown, since they do not have measured survival times. However, they are still relevant. In particular, n represents information about all observations.

Probability Mass Function. Before we present the survival function for the Kaplan-Meir process, the origins and definition of the probability mass function will be shown. We can determine $f(t)$ at each t_i by considering the definition of a discrete probability distribution function.

We know that

$$f(t_i) = P(T = t_i) = P(\text{Death occurs at time } t_i) .$$

If death occurs at t_i, then it is assumed that the subject was alive at all previous times since enrollment in the study. That is,

$P(\text{Death occurs at time } t_i)$

$$= P(\text{Death occurs at time } t_i, \text{ Death does not occur at } t_{i-1}, t_{i-2}, ...t_1) .$$

It is assumed that death at t_i, and survival at $t_{i-1}, t_{i-2}, ...t_1$ are statistically unrelated events, and thus are independent. Hence,

$$P(\text{Death occurs at time } t_i,\ \text{Death does not occur at } t_{i-1}, t_{i-2}, ...t_1)$$

$$= P(\text{Death occurs at time } t_i) \bullet \prod_{j=1}^{i-1} P(\text{Death does not occur at } t_j).$$

Since $P(\text{Death occurs at time } t_i)$ is not known, it will be treated it as a parameter, and designated λ_i. Since $P(\text{Death does not occur at } t_j) = 1 - P(\text{Death occurs at time } t_j)$,

$$P(\text{Death does not occur at } t_j) = 1 - \lambda_j.$$

Therefore,

$$f(t_i) = P(\text{Death occurs at time } t_i) \bullet \prod_{j=1}^{i-1} P(\text{Death does not occur at } t_j)$$

$$= \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j).$$

Survival Function. To find the survival function $S(t)$ at each value t_i consider that $S(t_i) = P(T > t_i)$. Thus,

$$S(t_i) = P(\text{Death occurs after } t_i)$$

$$= P(\text{Death has not occured at } t_i, t_{i-1}, ...t_1)$$

$$= \prod_{j=1}^{i} (1 - \lambda_j).$$

Estimated Survival Function. Our ultimate goal is to estimate $S(t)$. One can do that by estimating the unknown λ_j values, and those values can be estimated using the MLE method. However, trying to find the MLE of $S(t)$ would not produce any results.

Thus, we have to rely upon finding the MLE of a distribution associated with $S(t)$ that exists in the same parameter space. That distribution is defined as the probability that d_i of n_i subjects die at t_i, given that all of the n_i subjects were known to have survived previously. That is,

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i).$$

The random variable D_i represents the possible number of deaths at t_i. The expression $n_i - D_i$ represents the number of subjects that have survived at t_i. The observation d_i is the number of deaths at t_i, and n_i is the number of subjects alive and at risk of dying at t_i just before t_i. Note that the closest time before t_i would be t_{i-1}. Also note that if a subject is censored before the study ends, then the subject is subtracted from the appropriate n_i.

The distribution $P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i)$ can be simplified down to a form, which should be more recognizable.

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{P(D_i = d_i, n_{i-1} - D_{i-1} = n_i)}{P\left(n_{i-1} - D_{i-1} = n_i\right)}.$$

If d_i deaths are observed at t_i, then that also means $n_i - d_i$ subjects have survived at t_i. Thus,

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{P(D_i = d_i, n_i - D_i = n_i - d_i, n_{i-1} - D_{i-1} = n_i)}{P\left(n_{i-1} - D_{i-1} = n_i\right)}.$$

Since the event of dying at t_i assumes survival at previous times, and the event of surviving at t_i also assumes survival at previous times

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{P\left(D_i = d_i, n_i - D_i = n_i - d_i\right)}{P\left(n_{i-1} - D_{i-1} = n_i\right)}.$$

The numerator follows a binomial distribution. Thus,

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{\binom{n_i}{d_i} P\left(D_i = d_i\right) P\left(n_i - D_i = n_i - d_i\right)}{P\left(n_{i-1} - D_{i-1} = n_i\right)}.$$

Since $P(\text{subject dies at } t_i) = f(t_i)$, $P(\text{subject survives at } t_i) = S(t_i)$, and each subject's survival is considered independent from another's,

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{\binom{n_i}{d_i} \left[f(t_i)\right]^{d_i} \left[S(t_i)\right]^{n_i - d_i}}{\left[S(t_{i-1})\right]^{n_i}}.$$

Finally, using the definitions of $f(t)$ and $S(t)$ we have

$$P(D_i = d_i \mid n_{i-1} - D_{i-1} = n_i) = \frac{\binom{n_i}{d_i} \lambda_i^{d_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{d_i} \left[\prod_{j=1}^{i}(1-\lambda_j)\right]^{n_i - d_i}}{\left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{n_i}}$$

$$= \frac{\binom{n_i}{d_i} \lambda_i^{d_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{d_i} \left[1-\lambda_i\right]^{n_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{n_i} \left[1-\lambda_i\right]^{-d_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{-d_i}}{\left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{n_i}}$$

$$= \frac{\binom{n_i}{d_i} \lambda_i^{d_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{d_i} \left[1-\lambda_i\right]^{n_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{n_i} \left[1-\lambda_i\right]^{-d_i}}{\left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{n_i} \left[\prod_{j=1}^{i-1}(1-\lambda_j)\right]^{d_i}}$$

$$= \binom{n_i}{d_i} \lambda_i^{d_i} \left[1-\lambda_i\right]^{n_i} \left[1-\lambda_i\right]^{-d_i}$$

$$= \binom{n_i}{d_i} \lambda_i^{d_i} \left[1-\lambda_i\right]^{n_i-d_i}.$$

Therefore, $P(D_i = d_i \mid D_{i-1} = n-n_i) = \binom{n_i}{d_i} \lambda_i^{d_i} \bullet$ $^{-d_i}$. It can clearly be seen that $P(D_i = d_i \mid D_{i-1} = n-n_i)$ is a binomial distribution. Thus, $D_i \sim B(n_i, \lambda_i)$.

Now an estimator for λ_i can be found using the maximum likelihood method. We let $L_i = \prod_{i=1}^{k} P(D_i = d_i \mid D_{i-1} = n-n_i)$ be the likelihood function, and proceed to find a λ_i that maximizes L_i. This λ_i we find will also maximum $S(t_i)$, which is our desired result. Taking the log of L_i we have

$$\log L_i = \log\left[\prod_{i=1}^{k} \binom{n_i}{d_i} \lambda_i^{d_i} \bullet \quad ^{-d_i}\right]$$

$$= \sum_{i=1}^{k}\left[\log\binom{n_i}{d_i} + \log \lambda_i^{d_i} + \log\left[1-\lambda_i\right]^{n_i-d_i}\right]$$

$$= \sum_{i=1}^{k}\left[\log\binom{n_i}{d_i} + \log \lambda_i^{d_i} + \log\left[1-\lambda_i\right]^{n_i-d_i}\right]$$

$$= \sum_{i=1}^{k}\left[\log\binom{n_i}{d_i} + d_i \log \lambda_i + \left(n_i-d_i\right)\log\left[1-\lambda_i\right]\right]$$

$$= \sum_{i=1}^{k}\left[\log\binom{n_i}{d_i}\right] + \sum_{i=1}^{k}\left[d_i \log \lambda_i\right] + \sum_{i=1}^{k}\left[\left(n_i-d_i\right)\log\left[1-\lambda_i\right]\right].$$

We now differentiate L_i with respect to λ_i, noting that the differentiation does not iterate with the summation.

$$\frac{\partial}{\partial \lambda_i} \log L_i = \frac{d_i}{\lambda_i} - \frac{\left(n_i-d_i\right)}{\left(1-\lambda_i\right)}.$$

Setting $\frac{\partial}{\partial \lambda_i} \log L_i$ equal to zero and solving for λ_i we have

$$\frac{d_i}{\lambda_i} - \frac{(n_i - d_i)}{(1-\lambda_i)} = 0 .$$

$$\frac{(n_i - d_i)}{(1-\lambda_i)} = \frac{d_i}{\lambda_i}$$

$$\lambda_i n_i - \lambda_i d_i = d_i - \lambda_i d_i$$

$$\lambda_i = \frac{d_i}{n_i} .$$

Therefore,

$$\hat{\lambda}_i = \frac{d_i}{n_i} .$$

Note that the estimator $\hat{\lambda}_i$ can be used for any given λ_i. If we change the subscript from i to another subscript, say j, the estimator does not change. For instance $\hat{\lambda}_j$ is defined as $\hat{\lambda}_j = d_j / n_j$.

Recall that $S(t_i) = \prod_{j=1}^{i} (1-\lambda_j)$. If we use $\hat{\lambda}_j$ to estimate λ_j, we arrive at

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j} \right).$$

The generalized version of the estimator for any value of t would be

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i} \right).$$

This estimator is referred to as the Kaplan-Meier estimator.

Variance of the Kaplan-Meier Estimator. Finding the variance of $\hat{S}(t)$ directly would be challenging. Instead, we can find the mean and variance of $\log \hat{S}(t)$ first, and then proceed to find the variance of $\hat{S}(t)$. The mean of $\log \hat{S}(t)$ is

$$\mathrm{E}\left[\log \hat{S}(t)\right] = \mathrm{E}\left[\log\left[\prod_{t_i \le t}\left(1-\hat{\lambda}_i\right)\right]\right]$$

$$= \mathrm{E}\left[\sum_{t_i \le t}\log\left(1-\hat{\lambda}_i\right)\right]$$

$$= \sum_{t_i \le t}\mathrm{E}\left[\log\left(1-\hat{\lambda}_i\right)\right].$$

Similarly, the variance of $\log \hat{S}(t)$ is

$$\mathrm{Var}\left[\log \hat{S}(t)\right] = \mathrm{Var}\left[\log\left[\prod_{t_i \le t}\left(1-\hat{\lambda}_i\right)\right]\right]$$

$$= \mathrm{Var}\left[\sum_{t_i \le t}\log\left(1-\hat{\lambda}_i\right)\right]$$

$$= \sum_{t_i \le t}\mathrm{Var}\left[\log\left(1-\hat{\lambda}_i\right)\right].$$

The delta method is used below to determine a convergent distribution for $\log\left(1-\hat{\lambda}_i\right)$ which in turn will help us find a convergent distribution for $\log \hat{S}(t)$.

Let $Y_n = d_i$ and $g(Y_n) = \log\left(1-\frac{Y_n}{n_i}\right) = \log\left(1-\hat{\lambda}_i\right)$. We know that $D_i \sim B\left(n_i, \lambda_i\right)$, and that $D_i \xrightarrow{D} N\left(n_i\lambda_i, n_i\lambda_i\left(1-\lambda_i\right)\right)$. Thus,

$$g'(Y_n) = \frac{1}{-n_i\left(1-\frac{d_i}{n_i}\right)} \text{ and } g'(\mu) = \frac{1}{-n_i\left(1-\frac{n_i\lambda_i}{n_i}\right)} = \frac{1}{-n_i\left(1-\lambda_i\right)}.$$

We can clearly see that g' exists if $d_i \neq n_i$, and $g'(\mu) \neq 0$. If $d_i = n_i$, then

$\text{Var}\left(\log\left(1-\hat{\lambda}_i\right)\right)$ would be very easy to find. It would be zero for that ith iteration.

Thus,

$$\log\left(1-\lambda_i\right) \xrightarrow{D} N\left(g\left(\mu\right), \left[g'\left(\mu\right)\right]^2 \sigma^2\right).$$

The values for $g\left(\mu\right)$ and $\left[g'\left(\mu\right)\right]^2 \sigma^2$ are

$$g\left(\mu\right) = \log\left(1-\frac{n_i\lambda_i}{n_i}\right) = \log\left(1-\lambda_i\right)$$

and

$$\left[g'\left(\mu\right)\right]^2 \sigma^2 = \left(\frac{1}{-n_i\left(1-\lambda_i\right)}\right)^2 n_i\lambda_i\left(1-\lambda_i\right)$$

$$= \frac{n_i\lambda_i\left(1-\lambda_i\right)}{{n_i}^2\left(1-\lambda_i\right)^2}$$

$$= \frac{\lambda_i}{n_i\left(1-\lambda_i\right)}.$$

Hence,

$$\log\left(1-\hat{\lambda}_i\right) \xrightarrow{D} N\left(\log\left(1-\lambda_i\right), \frac{\lambda_i}{n_i\left(1-\lambda_i\right)}\right).$$

Since each $\log\left(1-\hat{\lambda}_i\right)$ would be I.I.D,

$$\log\hat{S}(t) = \log\left[\prod_{t_i \le t}\left(1-\hat{\lambda}_i\right)\right] = \sum_{t_i \le t}\log\left(1-\hat{\lambda}_i\right) \xrightarrow{D} N\left(\sum_{t_i \le t}\log\left(1-\lambda_i\right), \sum_{t_i \le t}\frac{\lambda_i}{n_i\left(1-\lambda_i\right)}\right).$$

We use the delta method again to find $\operatorname{Var}\hat{S}(t)$.

Let $Y_n = \log\hat{S}(t)$ and $g(Y_n) = \exp\left(\log\hat{S}(t)\right) = \hat{S}(t)$. We know that

$\log\hat{S}(t) \xrightarrow{D} N\left(\sum_{t_i \le t}\log\left(1-\lambda_i\right), \sum_{t_i \le t}\frac{\lambda_i}{n_i\left(1-\lambda_i\right)}\right)$ as shown previously. Thus,

$$g'(Y_n) = \exp\left(\log\hat{S}(t)\right) = \hat{S}(t) \text{ and } g'(\mu) = \exp\left[\sum_{t_i \le t}\log\left(1-\lambda_i\right)\right] = S(t).$$

We see that g' exists, and $g'(\mu) \neq 0$. Thus,

$$\hat{S}(t) \xrightarrow{D} N\left(g(\mu), \left[g'(\mu)\right]^2\sigma^2\right).$$

The values of $g(\mu)$ and $\left[g'(\mu)\right]^2\sigma^2$ are

$$g(\mu) = \exp\left[\sum_{t_i \le t}\log\left(1-\lambda_i\right)\right] = S(t) \text{ and } \left[g'(\mu)\right]^2\sigma^2 = \left[S(t)\right]^2\sum_{t_i \le t}\frac{\lambda_i}{n_i\left(1-\lambda_i\right)}.$$

It is clear now that,

$$\operatorname{Var}\hat{S}(t) \approx \left[S(t)\right]^2\sum_{t_i \le t}\frac{\lambda_i}{n_i\left(1-\lambda_i\right)}.$$

Since λ_i is unknown, an estimate for λ_i is needed to find the variance in practice. The estimator $\hat{\lambda}_i = d_i / n_i$, which we used before, should be suitable in practical applications. The variance resulting from the substitution is

$$\operatorname{Var}\hat{S}(t) \approx \left[\hat{S}(t)\right]^2\sum_{t_i \le t}\frac{\frac{d_i}{n_i}}{n_i\left(1-\frac{d_i}{n_i}\right)} = \left[\hat{S}(t)\right]^2\sum_{t_i \le t}\frac{d_i}{n_i\left(n_i-d_i\right)}.$$

The variance above is known as Greenwood's formula produced by Greenwood (1926).

Confidence Intervals for the Kaplan-Meier Estimator. If we initially look at the confidence interval for the Kaplan-Meier estimator it appears straightforward. A 95% confidence interval for $\hat{S}(t)$ would be

$$\hat{S}(t) \pm z_{0.975} \operatorname{Var} \hat{S}(t).$$

However, like with the initial confidence interval for $\hat{S}_T(t)$, this confidence interval could also result in impossible values outside of $[0,1]$. Thus we have to use the confidence interval formula derived previously shown as

$$\left[S(t)^{e^{z_{0.975}\operatorname{Var}\left[\log(-\log S(t))\right]}}, S(t)^{e^{-z_{0.975}\operatorname{Var}\left[\log(-\log S(t))\right]}} \right]$$

Substituting $\hat{S}(t)$ for $S(t)$, we have

$$\left[\hat{S}(t)^{e^{z_{0.975}\operatorname{Var}\left[\log\left(-\log \hat{S}(t)\right)\right]}}, \hat{S}(t)^{e^{-z_{0.975}\operatorname{Var}\left[\log\left(-\log \hat{S}(t)\right)\right]}} \right].$$

We can apply the delta method to get the distribution of $\log\left(-\log \hat{S}(t)\right)$, and as a result,

$$\operatorname{Var}\left[\log\left(-\log \hat{S}(t)\right)\right] \approx \frac{1}{\left[\log \hat{S}(t)\right]^2} \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}.$$

Thus, a suitable 95% confidence interval for $\hat{S}(t)$ would be

$$\left[\hat{S}(t)^{e^{\frac{1.96}{\left[\log \hat{S}(t)\right]^2}\sum_{t_i \le 1} \frac{d_i}{n_i(n_i - d_i)}}}, \hat{S}(t)^{e^{-\frac{1.96}{\left[\log \hat{S}(t)\right]^2}\sum_{t_i \le 1} \frac{d_i}{n_i(n_i - d_i)}}} \right].$$

Kaplan-Meir Table. The Kaplan-Meir table shows information regarding the Kaplan-Meir estimate for any given set. Table 3 is a format of the Kaplan-Meir table. Note that Table 3 is not the same as a general lifetable which will be discussed in the future.

Table 3. Kaplan-Meir Table Format

Survival Times	Number at Risk	Number of Failures	$\hat{S}(t)$	Var $\hat{S}(t)$	Lower 95% CI Bound	Upper 95% CI Bound
t_1	n_1	d_1	$\hat{S}(t_1)$	$\text{Var}\,\hat{S}(t_1)$	$L\left[\hat{S}(t_1)\right]$	$U\left[\hat{S}(t_1)\right]$
t_2	n_2	d_2	$\hat{S}(t_2)$	$\text{Var}\,\hat{S}(t_2)$	$L\left[\hat{S}(t_2)\right]$	$U\left[\hat{S}(t_2)\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_i	n_i	d_i	$\hat{S}(t_i)$	$\text{Var}\,\hat{S}(t_i)$	$L\left[\hat{S}(t_i)\right]$	$U\left[\hat{S}(t_i)\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_{n-1}	n_{n-1}	d_{n-1}	$\hat{S}(t_{n-1})$	$\text{Var}\,\hat{S}(t_{n-1})$	$L\left[\hat{S}(t_{n-1})\right]$	$U\left[\hat{S}(t_{n-1})\right]$
t_n	n_n	d_n	$\hat{S}(t_n)$	$\text{Var}\,\hat{S}(t_n)$	$L\left[\hat{S}(t_n)\right]$	$U\left[\hat{S}(t_n)\right]$

Determining a Theoretical Distribution. Instead of hazard plotting, the Kaplan-Meir method can be used to determine an appropriate theoretical distribution to model data. The Kaplan-Meir method provides an estimated survival function that can be analyzed for similarities to theoretical survival function. The graph of the Kaplan-Meir estimate can be compared to the graph of theoretical survival estimates.

The Kaplan-Meir step function may be somewhat difficult to compare to a smooth function. Thus, for the Kaplan-Meir graph we can draw lines between each $\left(t_{i,}S\left(t_i\right)\right)$ for a smoother representation. This representation is called a Kaplan-Meir curve. If the Kaplan-Meir curve is similar in shape to a theoretical survival function, then one might use that theoretical distribution to model the data.

Modified Kaplan-Meir Estimator

We now turn our attention to the main focus of our work. That is modifying the Kaplan-Meir estimator. It was seen previously with the standard Kaplan-Meir estimator that $\hat{\lambda}_i$ was an estimator of λ_i based on the MLE. Although the MLE is a relatively good method of estimation, there might be other methods of estimation which would produce better results under the right circumstances. We explore such a method here.

The tradition Kaplan-Meier estimator can modified by using a binomial Bayes estimator to estimate λ_i rather than a MLE estimator. We define the modified Kaplan-Meir estimator as

$$\hat{S}_B(t)=\prod_{t_i\le t}\left(1-\hat{\lambda}_{Bi}\right)$$

where $\hat{\lambda}_{Bi}$ is the Bayes estimation for λ_i. There are many important attributes of this estimation, as well as reasons why one might want to use it instead of the standard Kaplan Meir estimation in certain circumstances.

We can investigate this function further by focusing on its parameter $\hat{\lambda}_{Bi}$. The origins of $\hat{\lambda}_{Bi}$ will be shown and eventually defined it in order to have a more complete picture of $\hat{S}_B(t)$. In order to do that, we start with a review of Bayes estimators.

Bayes Estimators. A Bayes estimator is different from most estimation techniques in that it assumes there is some prior knowledge about the distribution of the unknown parameter. The experimenter may subjectively have some idea or intuition about what the distribution for a parameter is before the data is viewed. This prior distribution, as it is called, is an important part of constructing a Bayes estimations.

For a parameter θ, $\pi(\theta)$ denotes the prior distribution. Once the prior distribution is chosen, the data is then observed and fitted to an appropriate distribution called the sampling distribution, denoted $f(x \mid \theta)$. Then, the prior distribution is updated to more closely align with the observed data. This updated distribution is called the posterior distributed and is denoted $f(\theta \mid x)$. It is the conditional distribution of θ, given the sample, x. Once the posterior distributed is found, the mean of $f(\theta \mid x)$ can be calculated to determine the Bayes estimator for θ, denoted $\hat{\theta}_B$.

If the prior distribution and sampling distribution are known, the posterior distribution can be determined by making use of the relationship $f(\theta \mid x) = f(x \mid \theta)\pi(\theta) / f(x)$. We can multiply $f(x \mid \theta)$ and $\pi(\theta)$ to get $f(x, \theta)$. Then

we can find the margin distribution $f(x)$, and divide $f(x,\theta)$ by $f(x)$ to arrive at $f(\theta \mid x)$.

The Bayes estimator we are searching for is then found by taking the expected value of $f(\theta \mid x)$. That is, $\hat{\theta}_B = E\left[f(\theta \mid x)\right]$.

Binomial Bayes Estimators. We have interest in finding the Bayes estimator for a binomial distribution, since we desire to find an estimator for λ_i, which is a parameter of the binomial distribution, $P(D_i = d_i \mid D_{i-1} = n - n_i) = \binom{n_i}{d_i} \lambda_i^{d_i} \bullet \qquad ^{-d_i}$. We replicate the process of finding a binomial Bayes estimator shown by Casella and Berger (2002).

Using the Bayes estimation method described in the previous section, we first find the Bayes estimator of any given parameter p belonging to an arbitrary binomial distribution. Suppose that θ has a prior distribution $\text{Beta}(\alpha,\beta)$ and the sampling distribution is found to be $B(n,p)$. Then,

$$f(x,\theta) = f(x \mid \theta)\pi(\theta) = \left[\binom{n}{x} p^x (1-p)^{n-x}\right]\left[\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$

$$= \binom{n}{x}\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}.$$

$$f(x) = \int_0^1 f(x,p)dp = \int_0^1 \binom{n}{x}\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\,dp$$

$$= \binom{n}{x}\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\,dp$$

$$= \binom{n}{x}\frac{\Gamma(\alpha,\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}.$$

Hence,

$$f(p \mid x) = \frac{f(x,p)}{f(x)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}.$$

Since the expected value of $\text{Beta}(\alpha,\beta)$ is $1/\left(1+\frac{\beta}{\alpha}\right)$ and $f(p \mid x)$ is

$\text{Beta}(x+\alpha, n-x+\beta)$,

$$\mathrm{E}[f(p \mid x)] = \frac{1}{1+\frac{n-x+\beta}{x+\alpha}}$$

$$= \frac{1}{\frac{n+\alpha+\beta}{x+\alpha}}$$

$$= \frac{x+\alpha}{n+\alpha+\beta}.$$

Therefore, the binomial Bayes estimator for a parameter θ is

$$\hat{\theta} = \frac{x+\alpha}{n+\alpha+\beta}.$$

Bayes Estimator for λ_i. Now that we have seen the definition of a binomial Bayes estimator, we can find the Bayes estimator for λ_i. It is known that d_i follows the binomial distribution, $P(D_i = d_i \mid D_{i-1} = n - n_i) = \binom{n_i}{d_i} \lambda_i^{d_i} \bullet \qquad ^{-d_i}$. If we assume the prior distribution of λ_i is $\text{Beta}(\alpha, B)$. Then

$$\lambda_{Bi} = \frac{d_i+\alpha}{n+\alpha+\beta}.$$

Notice that α and β must still be chosen in order to fully identify λ_{Bi}. If one has sufficient knowledge about the prior distribution, one might have prior knowledge of what α and β might be. However, if one doesn't have enough knowledge about the prior distribution, α and β must be chosen. An appropriate α and β will be revealed as we discuss the comparison between the two estimators, $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$.

Comparing $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$. The Bayes estimator can be compared to the tradition MLE estimator in order to determine which one might be a better estimator and in what circumstances. One way to do this is to compare their mean square errors.

The mean square error (MSE) is defined as $E_\theta\left(\hat{\theta}-\theta\right)^2$, where $\hat{\theta}$ is an estimator for the parameter θ. In general, it is desirable for the MSE to be small. However, sometimes determining how small an MSE is can be challenge. In order to help to make this challenge easier and provide more detailed information, the MSE can be split to two parts as shown below.

$$E_\theta\left(\hat{\theta}-\theta\right)^2 = \mathrm{Var}_\theta\,\hat{\theta}+\left(E_\theta\,\hat{\theta}-\theta\right)^2 = \mathrm{Var}_\theta\,\hat{\theta}+\left(\mathrm{Bias}_\theta\,\hat{\theta}\right)^2$$

The bias of $\hat{\theta}$ measures how closely the estimator fits the data. The Variance of $\hat{\theta}$ is the variability of the estimator. It is an expression of how spread out each $\hat{\theta}$ is from $E\left(\hat{\theta}\right)$.

If the bias of the MSE is 0, then the estimator is referred to as unbiased. This would indicate that the estimator fits the parameter very well and usually means that the MSE is small. However, sometimes if the bias of the MSE is small, the variance becomes too large. In such a case, it can be acceptable to have a larger bias in exchange for a smaller variance and an overall smaller MSE.

Now let us look at the mean square errors for both estimators. The MSE of $\hat{\lambda}_i$ is defined as

$$\text{MSE}\,\hat{\lambda}_i = \text{Var}_{\lambda_i}\,\hat{\lambda}_i + \left(\text{Bias}_{\lambda_i}\,\hat{\lambda}_i\right)^2.$$

This implies,

$$E_{\lambda_i}(\hat{\lambda}_i - \lambda_i)^2 = \text{Var}_{\lambda_i}\,\hat{\lambda}_i + \left(E_{\lambda_i}\,\hat{\lambda}_i - \lambda_i\right)^2.$$

$$= \text{Var}_{\lambda_i}\,\frac{d_i}{n_i} + \left(E_{\lambda_i}\,\frac{d_i}{n_i} - \lambda_i\right)^2$$

$$= \frac{\text{Var}_{\lambda_i} d_i}{n_i^{\,2}} + \left(\lambda_i - \lambda_i\right)^2$$

$$= \frac{n_i\lambda_i\left(1-\lambda_i\right)}{n_i^{\,2}}$$

$$= \frac{\lambda_i\left(1-\lambda_i\right)}{n_i}.$$

Thus,

$$\text{MSE}\,\hat{\lambda}_i = \frac{\lambda_i\left(1-\lambda_i\right)}{n_i}.$$

The MSE of $\hat{\lambda}_{Bi}$ is defined as

$$\text{MSE}\,\hat{\lambda}_{Bi} = \text{Var}_{\lambda_i}\,\hat{\lambda}_{Bi} + \left(\text{Bias}_{\lambda_i}\,\hat{\lambda}_{Bi}\right)^2.$$

This implies,

$$E_{\lambda_i}\left(\hat{\lambda}_{Bi} - \lambda_i\right)^2 = \text{Var}_{\lambda_i}\,\hat{\lambda}_{Bi} + \left(E_{\lambda_i}\,\hat{\lambda}_{Bi} - \lambda_i\right)^2$$

$$= \mathrm{Var}_{\lambda_i}\left(\frac{d_i+\alpha}{\alpha+\beta+n_i}\right)+\left(E_{\lambda_i}\left(\frac{d_i+\alpha}{\alpha+\beta+n_i}\right)-\lambda_i\right)^2$$

$$= \frac{\mathrm{Var}_{\lambda_i} d_i}{\left(\alpha+\beta+n_i\right)^2}+\left(\frac{E\left(d_i\right)+\alpha}{\alpha+\beta+n_i}-\lambda_i\right)^2$$

$$= \frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\alpha+\beta+n_i\right)^2}+\left(\frac{n_i\lambda_i+\alpha}{\alpha+\beta+n_i}-\lambda_i\right)^2.$$

Thus,

$$\mathrm{MSE}\,\hat{\lambda}_{Bi} = \frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\alpha+\beta+n_i\right)^2}+\left(\frac{n_i\lambda_i+\alpha}{\alpha+\beta+n_i}-\lambda_i\right)^2.$$

These same results were derived by Casella and Berger (2002) for the mean square errors of general proportions. Suppose that we were to graph the MSE of $\hat{\lambda}_i$ as a function of λ_i. There would exist a λ_i-axis, a MSE axis, and the graph would be quadratic as shown in Figure 2.

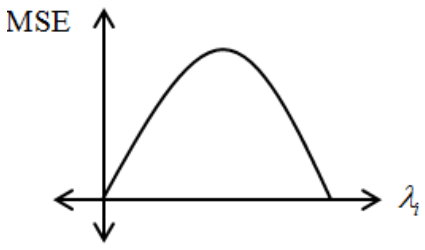


Figure 2. General Shape of MSE $\hat{\lambda}_i$

If on wanted a better estimator than λ_i, one might seek to find a MSE of an estimator which has a graph closer to the λ_i-axis. A constant MSE far below the maximum of the

quadratic function shown above would indicate a seemingly better estimator. It would also produce a consistent MSE for each value of λ_i.

Casella and Bergers (2002) show a method to arrive at a constant MSE of a general Binomial Bayes estimator. We adopt a similar method to arrive at a constant MSE $\hat{\lambda}_{Bi}$. If there is no good prior information regarding λ_i, choosing α and β to both be $\sqrt{n_i}/2$ would result in

$$\hat{\lambda}_{Bi} = \frac{d_i + \sqrt{n_i}/2}{n_i + \sqrt{n_i}}$$

and thus,

$$E_{\lambda_i}\left(\hat{\lambda}_{Bi} - \lambda_i\right)^2 = \frac{n_i\lambda_i(1-\lambda_i)}{(\alpha+\beta+n_i)^2} + \left(\frac{n_i\lambda_i + \alpha}{\alpha+\beta+n_i} - \lambda_i\right)^2$$

$$= \frac{n_i\lambda_i(1-\lambda_i)}{\left(\sqrt{n_i}/2 + \sqrt{n_i}/2 + n_i\right)^2} + \left(\frac{n_i\lambda_i + \sqrt{n_i}/2}{\sqrt{n_i}/2 + \sqrt{n_i}/2 + n_i} - \lambda_i\right)^2$$

$$= \frac{n_i\lambda_i - n_i\lambda_i^2}{\left(\sqrt{n_i} + n_i\right)^2} + \left(\frac{n_i\lambda_i + \sqrt{n_i}/2 - \lambda_i\sqrt{n_i} - \lambda_i n_i}{\left(\sqrt{n_i} + n_i\right)}\right)^2$$

$$= \frac{n_i\lambda_i(1-\lambda_i)}{\left(\sqrt{n_i} + n_i\right)^2} + \frac{n_i(1/2 - \lambda_i)^2}{\left(\sqrt{n_i} + n_i\right)^2}$$

$$= \frac{n_i\lambda_i(1-\lambda_i) + n_i(1/2 - \lambda_i)^2}{\left(\sqrt{n_i} + n_i\right)^2}$$

$$= \frac{n_i\lambda_i - n_i\lambda_i^2 + n_i/4 - \lambda_i n_i + \lambda_i^2 n_i}{\left(\sqrt{n_i} + n_i\right)^2}$$

$$= \frac{n_i}{4\left(\sqrt{n_i} + n_i\right)^2}$$

We can clearly see that our choice of α and β, resulted in a constant $\text{MSE}\,\hat{\lambda}_{Bi}$.

The two estimators $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ can be compared by analyzing the graphs of their mean square errors. But first, we have to make an observation about the MSEs of our two estimators,

$$\text{MSE}\left(\hat{\lambda}_i\right) = \frac{\lambda_i\left(1-\lambda_i\right)}{n_i} \text{ and } \text{MSE}\left(\hat{\lambda}_{Bi}\right) = \frac{n_i}{4\left(\sqrt{n_i} + n_i\right)^2}.$$

Notice that for both of the MSEs there are n_i terms. This means that there will be different graphs for the functions depending on the n_i values. This will have an effect as n_i changes. However, for the sake of simplicity, we will analyze n_i at just one value, its value at n. That is $n_i = n$. Figure 3 shows graphs of the MSEs with different respective values for n. A similar figure was shown by Casella and Berger (2002).

Without any knowledge of the true λ_i, it turns out that if n is small, then $\hat{\lambda}_{Bi}$ is more likely a better estimator. Most values of $\text{MSE}\left(\hat{\lambda}_{Bi}\right)$ will be smaller, except for values close to $\lambda_i = 0$ or $\lambda_i = 1$. If n is large, then $\hat{\lambda}_i$ is a relatively better estimator, since most of the values of $\text{MSE}\left(\hat{\lambda}_i\right)$ will be smaller except for values close to $\lambda_i = 1/2$. An attempt can be made to determine the true value of $\hat{\lambda}_i$. This will be discussed in a future section.

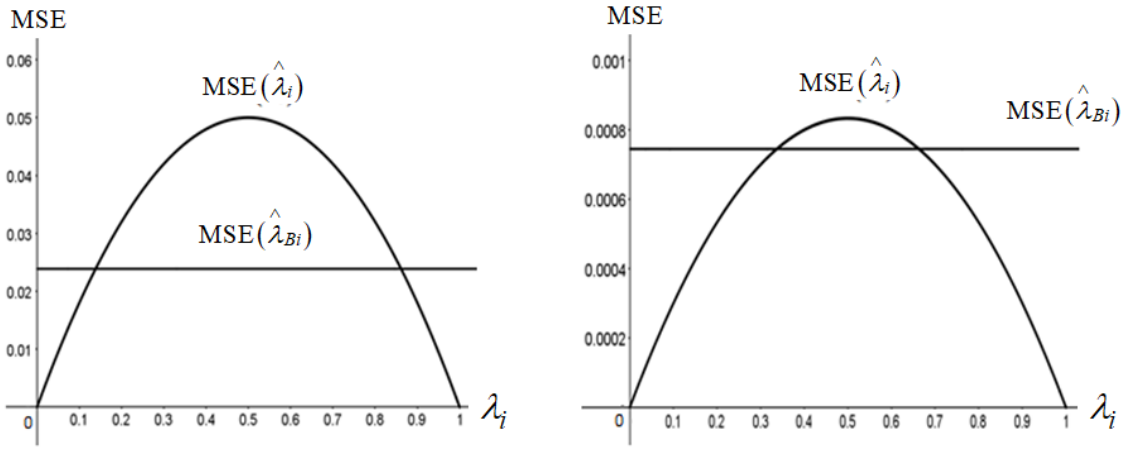


Figure 3. Comparison of $\text{MSE}\left(\hat{\lambda}_i\right)$ and $\text{MSE}\left(\hat{\lambda}_{Bi}\right)$ at $n = 5$ and $n = 300$.

A large or small value of n is rather subjective. A more objective way of determining size can be seen by observing the graph of the MSEs at $n = 42$ shown in Figure 4.

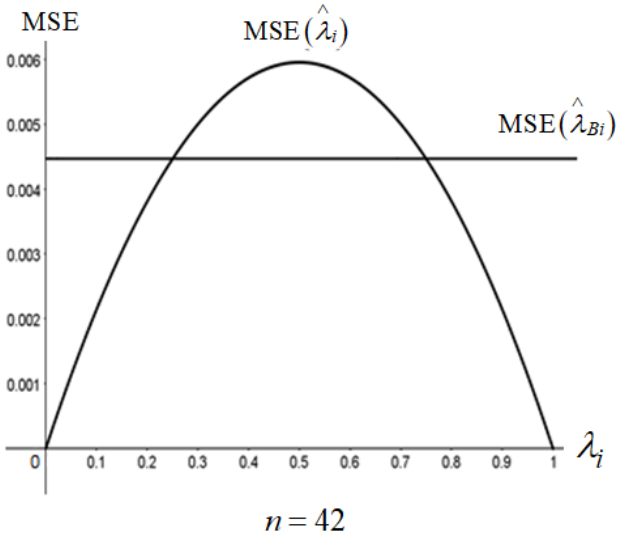


Figure 4. Comparison of $\text{MSE}\left(\hat{\lambda}_i\right)$ and $\text{MSE}\left(\hat{\lambda}_{Bi}\right)$ at $n = 42$

At $0.250258 \le \lambda_i \le 0.749742$, $\text{MSE}\left(\hat{\lambda}_i\right) \ge \text{MSE}\left(\hat{\lambda}_{Bi}\right)$ and at $\lambda_i \le 0.250258$ and $0.749742 \le \lambda_i$, $\text{MSE}\left(\hat{\lambda}_i\right) \le \text{MSE}\left(\hat{\lambda}_{Bi}\right)$. About 49.9% of the graph shows that the estimator $\hat{\lambda}_i$ should be favored. At $n = 41$, about 50.1% of the corresponding graph shows that $\hat{\lambda}_i$ should be favored. Thus, $n = 42$ marks the turning point where more values of the true λ_i favor $\hat{\lambda}_i$. Theoretically speaking, either estimator should be fine around $n = 42$ if there is no information regarding λ_i. This is an important case, but one should be cautious about what a sample size of 42 represents. The true value of λ_i is not known. In practice, a sample size of 41 or 42 may not be an appropriate boundary for a respectively small or large sample size. In practice, certain data sets may have other boundaries base on the general trend of λ_i. For instance some data, may favor a smaller λ_i, so that the boundary for a small sample size may be much lower than 41.

It is important to address the effect of n_i on choosing between $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$. Notice that $n_i \le n$. If n is small, there is no reason to consider a choice between $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ for each n_i, because $\hat{\lambda}_{Bi}$ is the natural choice for smaller sample sizes. If n starts small, each successive n_i will naturally be smaller. If n is large, then there is a need to see how much smaller each n_i will be. If enough subsequent n_i values are less than 30, then $\hat{\lambda}_{Bi}$ may be a better choice even if n is large.

Our comparison of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ shows why we have chosen $\hat{\lambda}_{Bi}$ as the estimator for λ_i in our modified Kaplan-Meir estimator. While $\hat{\lambda}_{Bi}$ does not universally

outperform $\hat{\lambda}_i$, it does generally do better for smaller sample sizes, theoretically under 42. This shows the potential value of the modified Kaplan-Meir estimator. If one was only considering the parameters, the modified Kaplan-Meir estimator would do best for smaller sample sizes compared to the standard Kaplan-Meir estimator. This is not a certainty, but the suspicion warrants a further analysis which will be discussed latter.

A More Detailed Definition of $\hat{S}_B(t)$. Now that we have discussed and defined $\hat{\lambda}_{Bi}$, a more complex and detailed definition of $\hat{S}_B(t)$ can be analyzed. Since it was shown that $\hat{\lambda}_{Bi} = \dfrac{d_i + \sqrt{n}/2}{\sqrt{n_i} + n_i}$, we can now define $\hat{S}_B(t)$ as,

$$\hat{S}_B(t) = \prod_{t_i \le t}\left(1 - \hat{\lambda}_{Bi}\right) = \prod_{t_i \le t}\left[1 - \left(\frac{d_i + \sqrt{n}/2}{\sqrt{n_i} + n_i}\right)\right].$$

We reorganize $\dfrac{d_i + \sqrt{n}/2}{\sqrt{n_i} + n_i}$ so that $\hat{\lambda}_{Bi}$ is a linear combination of $\hat{\lambda}_i$.

$$\frac{d_i + \sqrt{n_i}/2}{\sqrt{n_i} + n_i} = \frac{d_i}{\sqrt{n_i} + n_i} + \frac{\sqrt{n_i}/2}{\sqrt{n_i} + n_i}$$

$$= \frac{n_i}{\sqrt{n_i} + n_i} \bullet \left(\frac{\quad}{n_i}\right) + \frac{\sqrt{n_i}/2}{\sqrt{n_i} + n_i}$$

$$= \frac{n_i}{\sqrt{n_i} + n_i} \bullet \quad + \frac{\quad/2}{\sqrt{n_i} + n_i}$$

Thus,

$$\hat{S}_B(t) = \prod_{t_i \le t}\left[1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \bullet \quad + \frac{\quad/2}{\sqrt{n_i} + n_i}\right)\right].$$

This version of $\hat{S}_B(t)$ will be used so that our parameter might be $\hat{\lambda}_i = d_i / n_i$.

This gives us a closer association with $\hat{S}(t)$ in that they have the same parameters. We will also use $\hat{\lambda}_i$ as an estimator for λ_i when the variance and expectation is computed.

Variance of the Modified Kaplan-Meir Estimate. The variance for $\hat{S}_B(t)$ can be found just as the variance for $\hat{S}(t)$ was found. First, the variance and expectation of $\log \hat{S}_B(t)$ are found.

$$\mathrm{E}\left[\log \hat{S}_B(t)\right] = \sum_{t_i \le t} \mathrm{E}\left[\log\left(1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right)\right]$$

and

$$\mathrm{Var}\left[\log \hat{S}_B(t)\right] = \sum_{t_i \le t} \mathrm{Var}\left[\log\left(1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right)\right].$$

The delta method can now be used to find a convergent distribution for

$$\log\left[1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right].$$

Let $Y_n = \hat{\lambda}_i$ and $g(Y_n) = \log\left[1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right]$. Simplifying $g(Y_n)$ we have

$$g(Y_n) = \log\left[1 - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right]$$

$$= \log\left[\frac{\sqrt{n_i} + n_i}{\sqrt{n_i} + n_i} - \left(\frac{n_i}{\sqrt{n_i} + n_i} \cdot \hat{\lambda}_i \sqrt{n_i}/2\right)\right]$$

$$= \log\left(\frac{\sqrt{n_i}/2 + n_i - n_i \hat{\lambda}_i}{\left(\sqrt{n_i} + n_i\right)}\right).$$

Since $D_i \xrightarrow{D} N\left(n_i\lambda_i, n_i\lambda_i\left(1-\lambda_i\right)\right)$, and since we are assuming each D_i is identically independent,

$$Y_n = \frac{d_i}{n_i} \xrightarrow{D} N\left(\lambda_i, \frac{\lambda_i\left(1-\lambda_i\right)}{n_i}\right).$$

Thus,

$$g'(Y_n) = \frac{1}{\frac{\sqrt{n_i}/2 + n_i - n_i\hat{\lambda}_i}{\left(\sqrt{n_i}+n_i\right)}} \bullet \left(\sqrt{n_i}+n_i\right) = \frac{-n_i}{\sqrt{n_i}/2 + n_i - n_i\hat{\lambda}_i}$$

and

$$g'(\mu) = \frac{-n_i}{\sqrt{n_i}/2 + n_i - n_i\lambda_i}.$$

If $n_i + \sqrt{n_i}/2 - n_i\lambda_i = 0$, then $\lambda_i = \frac{1}{2\sqrt{n_i}} + 1$. We are assuming that $0 \le \lambda_i \le 1$ so that this case will not occur in practice. Thus, $g'(\mu)$ exists and $g'(\mu) \neq 0$ since $1 \le n$. Thus,

$$\log\left(1 - \left(\frac{n_i}{\sqrt{n_i}+n_i} \bullet \frac{\sqrt{n_i}/2}{\sqrt{n_i}+n_i}\right)\right) \xrightarrow{D} N\left(g(\mu), \left[g'(\mu)\right]^2\sigma^2\right).$$

The values of $g(\mu)$ and $\left[g'(\mu)\right]^2\sigma^2$ are

$$g(\mu) = \log\left(\frac{\sqrt{n_i}/2 + n_i - n_i\lambda_i}{\left(\sqrt{n_i}+n_i\right)}\right)$$

and

$$\left[g'(\mu)\right]^2\sigma^2 = \frac{\left(-n_i\right)^2}{\left(\sqrt{n_i}/2 + n_i - n_i\lambda_i\right)^2} \bullet \frac{n_i\lambda_i\left(1-\lambda_i\right)}{n_i^2} = \frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}/2 + n_i - n_i\lambda_i\right)^2}.$$

Hence,

$$\log\left(1-\frac{n_i}{\sqrt{n_i}+n_i}\bullet\quad\frac{/2}{\sqrt{n_i}+n_i}\right)\xrightarrow{D}N\left(\log\left(\frac{\sqrt{n_i}/2+n_i-n_i\lambda_i}{\left(\sqrt{n_i}+n_i\right)}\right),=\frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}/2+n_i-n_i\lambda_i\right)^2}\right).$$

Since each $\log\left(1-\frac{n_i}{\sqrt{n_i}+n_i}\bullet\quad\frac{/2}{\sqrt{n_i}+n_i}\right)$ is I.I.D,

$$\log\hat{S}_B(t)=\log\left[\prod_{t_i\le t}\left(1-\frac{n_i}{\sqrt{n_i}+n_i}\bullet\quad\frac{/2}{\sqrt{n_i}+n_i}\right)\right]$$

$$=\sum_{t_i\le t}\log\left(1-\frac{n_i}{\sqrt{n_i}+n_i}\bullet\quad\frac{/2}{\sqrt{n_i}+n_i}\right)$$

$$\xrightarrow{D}N\left(\sum_{t_i\le t}\log\left(\frac{\sqrt{n_i}/2+n_i-n_i\lambda_i}{\left(\sqrt{n_i}+n_i\right)}\right),\sum_{t_i\le t}\frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}/2+n_i-n_i\lambda_i\right)^2}\right).$$

We use the delta method again to find $\mathrm{Var}\,\hat{S}_B(t)$.

Let $Y_n=\log\hat{S}_B(t)$ and $g(Y_n)=\exp\left(\log\hat{S}_B(t)\right)=\hat{S}_B(t)$. We know that

$$\log\hat{S}_B(t)\xrightarrow{D}N\left(\sum_{t_i\le t}\log\left(\frac{\sqrt{n_i}/2+n_i-n_i\lambda_i}{\left(\sqrt{n_i}+n_i\right)}\right),\sum_{t_i\le t}\frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}/2+n_i-n_i\lambda_i\right)^2}\right).$$

Thus,

$$g'(Y_n)=\exp\left(\log\hat{S}_B(t)\right)=\hat{S}_B(t)$$

and

$$g'(\mu)=\exp\left[\sum_{t_i\le t}\log\left(\frac{\sqrt{n_i}/2+n_i-n_i\lambda_i}{\left(\sqrt{n_i}+n_i\right)}\right)\right]$$

$$=\exp\left[\sum_{t_i\le t}\log\left(1-\frac{n_i}{\sqrt{n_i}+n_i}\bullet\quad\frac{/2}{\sqrt{n_i}+n_i}\right)\right].$$

It is clear that $g'(\mu)$ exists, and $g'(\mu) \neq 0$. Thus, $\hat{S}_B(t) \xrightarrow{D} N\left(g(\mu), \left[g'(\mu)\right]^2 \sigma^2\right)$.

The values of $g(\mu)$ and $\left[g'(\mu)\right]^2 \sigma^2$ are

$$g(\mu) = \exp\left[\sum_{t_i \le t} \log\left(1 - \frac{n_i}{\sqrt{n_i} + n_i} \bullet \right.\right.$$

and

$$\left[g'(\mu)\right]^2 \sigma^2 = \left[\prod_{t_i \le t}\left(1 - \frac{n_i}{\sqrt{n_i} + n_i} \bullet \right)\right]^2 \sum_{t_i \le t} \frac{n_i \lambda_i (1-\lambda_i)}{\left(\sqrt{n_i}/2 + n_i - n_i\lambda_i\right)^2}.$$

Therefore,

$$\operatorname{Var} \hat{S}_B(t) \approx \left[\prod_{t_i \le t}\left(1 - \frac{n_i}{\sqrt{n_i} + n_i} \bullet \right)\right]^2 \sum_{t_i \le t} \frac{n_i \lambda_i (1-\lambda_i)}{\left(\sqrt{n_i}/2 + n_i - n_i\lambda_i\right)^2}.$$

Since λ_i is unknown, $\hat{\lambda}_i = d_i / n_i$ will be used as the estimator. Like with previous estimators, this is useful in practice and may not be desirable for theoretical analysis. The variance resulting from the substitution is

$$\operatorname{Var} \hat{S}_B(t) \approx \left[\hat{S}_B(t)\right]^2 \sum_{t_i \le t} \frac{n_i \hat{\lambda}_i \left(1 - \hat{\lambda}_i\right)}{\left(\sqrt{n_i}/2 + n_i - n_i \hat{\lambda}_i\right)^2}$$

$$= \left[\hat{S}_B(t)\right]^2 \sum_{t_i \le t} \frac{n_i \frac{d_i}{n_i}\left(1 - \frac{d_i}{n_i}\right)}{\left(\sqrt{n_i}/2 + n_i - n_i \frac{d_i}{n_i}\right)^2}$$

$$= \left[\hat{S}_B(t)\right]^2 \sum_{t_i \le t} = \frac{d_i (n_i - d_i)}{\left(\sqrt{n_i}/2 + n_i - d_i\right)^2 n_i}.$$

Confidence Intervals for the Modified Kaplan-Meier Estimator. A 95% confidence interval for $\hat{S}_B(t)$ would be

$$\hat{S}_B(t) \pm z_{0.975} \operatorname{Var} \hat{S}_B(t).$$

However, it was seen before that this could lead to results outside of the interval $[0,1]$. Thus, we make use of the formula we derived previously as shown below.

$$\left[S(t)^{e^{z_{0.975}\operatorname{Var}\left[\log(-\log S(t))\right]}}, S(t)^{e^{-z_{0.975}\operatorname{Var}\left[\log(-\log S(t))\right]}} \right]$$

Substituting $\hat{S}_B(t)$ for $S(t)$, we have

$$\left[\hat{S}_B(t)^{e^{z_{0.975}\operatorname{Var}\left[\log\left(-\log \hat{S}_B(t)\right)\right]}}, \hat{S}_B(t)^{e^{-z_{0.975}\operatorname{Var}\left[\log\left(-\log \hat{S}_B(t)\right)\right]}} \right].$$

The delta method is applied again to derive the distribution of $\log\left(-\log \hat{S}_B(t)\right)$, resulting in

$$\operatorname{Var}\left[\log\left(-\log \hat{S}_B(t)\right)\right] \approx \frac{1}{\left[\log \hat{S}_B(t)\right]^2} \sum_{t_i \le t} \frac{d_i(n_i - d_i)}{\left(\sqrt{n_i}/2 + n_i - d_i\right)^2 n_i}.$$

Thus, a suitable 95% confidence interval for $\hat{S}_B(t)$ would be

$$\left[\hat{S}_B(t)^{e^{\frac{1.96}{\left[\log \hat{S}_B(t)\right]^2} \sum_{t_i \le t} \frac{d_i(n_i-d_i)}{\left(\sqrt{n_i}/2+n_i-d_i\right)^2 n_i}}}, \hat{S}_B(t)^{e^{-\frac{1.96}{\left[\log \hat{S}_B(t)\right]^2} \sum_{t_i \le t} \frac{d_i(n_i-d_i)}{\left(\sqrt{n_i}/2+n_i-d_i\right)^2 n_i}}} \right].$$

Modified Kaplan-Meir Table. A modified Kaplan-Meir table contains information regarding the modified Kaplan-Meir estimate. The table is similar to the one created for the standard Kaplan-Meir estimate. Table 4 shows the format for a modified Kaplan-Meir table.

Table 4. Modified Kaplan-Meir Table Format

Survival Times	Number at Risk	Number of Failures	$\hat{S}_B(t)$	Var $\hat{S}_B(t)$	Lower 95% CI Bound	Upper 95% CI Bound
t_1	n_1	d_1	$\hat{S}_B(t_1)$	$\text{Var}\,\hat{S}_B(t_1)$	$L\left[\hat{S}_B(t_1)\right]$	$U\left[\hat{S}_B(t_1)\right]$
t_2	n_2	d_2	$\hat{S}_B(t_2)$	$\text{Var}\,\hat{S}_B(t_2)$	$L\left[\hat{S}_B(t_2)\right]$	$U\left[\hat{S}_B(t_2)\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_i	n_i	d_i	$\hat{S}_B(t_i)$	$\text{Var}\,\hat{S}_B(t_i)$	$L\left[\hat{S}_B(t_i)\right]$	$U\left[\hat{S}_B(t_i)\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_{n-1}	n_{n-1}	d_{n-1}	$\hat{S}_B(t_{n-1})$	$\text{Var}\,\hat{S}_B(t_{n-1})$	$L\left[\hat{S}_B(t_{n-1})\right]$	$U\left[\hat{S}_B(t_{n-1})\right]$
t_n	n_n	d_n	$\hat{S}_B(t_n)$	$\text{Var}\,\hat{S}_B(t_n)$	$L\left[\hat{S}_B(t_n)\right]$	$U\left[\hat{S}_B(t_n)\right]$

Comparing Survival Estimators

We now have the necessary elements to properly compare the survival estimators presented thus far. In order to determine which survival estimate models a given data set the best, we can compare the mean square errors of their respective survival estimators. The distribution with the smallest MSE survival estimator would be the optimal choice to model the data.

For any given survival estimator, $\hat{S}_E(t)$, its mean square error, is defined as

$$\begin{aligned}\mathrm{MSE}\,\hat{S}_E(t) &= \mathrm{E}\left(\hat{S}_E(t) - S(t)\right)^2 \\ &= \mathrm{Var}\,\hat{S}_E(t) + \left(\mathrm{Bias}\,\hat{S}_E(t)\right)^2 \\ &= \mathrm{Var}\,\hat{S}_E(t) + \left(\mathrm{E}\left(\hat{S}_E(t)\right) - S(t)\right)^2 .\end{aligned}$$

Here $S(t)$ is the true survival function that fits the data. Since we do not know the true survival function it is difficult to compute $\mathrm{MSE}\,\hat{S}_E(t)$. We could leave it as a constant. However, by doing so, it provides a difficulty in interpreting $\mathrm{MSE}\,\hat{S}_E(t)$. Instead, from this point forward, we will assume that the $S(t)$ chosen to model the data, corresponding to the given $\hat{S}_E(t)$, is the true survival function or acts as an estimator for the true survival function. While this sacrifices accuracy, it gives us more to work with in understanding $\mathrm{MSE}\,\hat{S}_E(t)$.

An important part of $\mathrm{MSE}\,\hat{S}_E(t)$ is the bias. For many estimators, $\mathrm{E}\left(\hat{S}_E(t)\right) = S\left(t\right)$, and thus $\mathrm{Bias}\,\hat{S}_E(t) = 0$. In these cases, one only needs to evaluate

$\operatorname{Var}\hat{S}_E(t)$. However, if $\mathrm{E}\left(\hat{S}_E(t)\right) \neq S(t)$, then the bias must be considered.

There are three survival estimators that are of particular importance and we would like to compare their mean square errors. The first survival estimator is the popular exponential survival function, denoted $\hat{S}_T(t)$. The mean square error of $\hat{S}_T(t)$ is

$$\operatorname{MSE}\hat{S}_T(t) = \mathrm{E}\left(\hat{S}_T(t) - S(t)\right)^2 = \operatorname{Var}\hat{S}_T(t) + \left(\mathrm{E}\left(\hat{S}_T(t)\right) - S(t)\right)^2.$$

When the $\operatorname{Var}\hat{S}_T(t)$ was determined previously, we also found that $\mathrm{E}\left(\hat{S}_T(t)\right) = S(t)$.

Thus,

$$\operatorname{MSE}\hat{S}_T(t) = \operatorname{Var}\hat{S}_T(t) + \left(S(t) - S(t)\right)^2 = \operatorname{Var}\hat{S}_T(t).$$

This is an important result. Since there is no bias for $\operatorname{MSE}\hat{S}_T(t)$, we only need to find the $\operatorname{Var}\hat{S}_T(t)$ to determine the effectiveness of the distribution in modeling the data.

The second survival estimator is the Kaplan-Meir estimator, denoted $\hat{S}(t)$. The mean square error of $\hat{S}(t)$ is

$$\operatorname{MSE}\hat{S}(t) = \mathrm{E}\left(\hat{S}(t) - S(t)\right)^2 = \operatorname{Var}\hat{S}(t) + \left(\mathrm{E}\left(\hat{S}(t)\right) - S(t)\right)^2.$$

When the $\operatorname{Var}\hat{S}(t)$ was determined previously, we also found that $\mathrm{E}\left(\hat{S}(t)\right) = S(t)$.

Thus,

$$\operatorname{MSE}\hat{S}(t) = \operatorname{Var}\hat{S}(t) + \left(S(t) - S(t)\right)^2 = \operatorname{Var}\hat{S}(t).$$

Again this is an important result because there is no bias.

The third survival estimator is the modified Kaplan-Meir estimator, denoted $\hat{S}_B(t)$. The mean square error $\hat{S}_B(t)$ is

$$\mathrm{MSE}\,\hat{S}_B(t) = \mathrm{E}\left(\hat{S}_B(t) - S(t)\right)^2 = \mathrm{Var}\,\hat{S}_B(t) + \left(\mathrm{E}\left(\hat{S}_B(t)\right) - S(t)\right)^2.$$

When the $\mathrm{Var}\,\hat{S}_B(t)$ was determined previously, we also found that

$$\mathrm{E}\left(\hat{S}_B(t)\right) = \prod_{t_i \le t}\left[1 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \bullet \frac{\lceil / 2}{\sqrt{n_i + n_i}}\right)\right].$$

Thus,

$$\mathrm{MSE}\,\hat{S}_B(t) = \mathrm{Var}\,\hat{S}_B(t) + \left(\prod_{t_i \le t}\left[1 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \bullet \frac{\lceil / 2}{\sqrt{n_i + n_i}}\right)\right] - S(t)\right)^2$$

$$= \mathrm{Var}\,\hat{S}_B(t) + \left(\prod_{t_i \le t}\left[1 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \bullet \frac{\lceil / 2}{\sqrt{n_i + n_i}}\right)\right] - S(t)\right)^2$$

$$= \mathrm{Var}\,\hat{S}_B(t) + \left(\prod_{t_i \le t}\left[1 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \bullet \frac{\lceil / 2}{\sqrt{n_i + n_i}}\right)\right] - \prod_{t_i \le t}\left(1 - \lambda_i\right)\right)^2.$$

Notice that $\mathrm{MSE}\,\hat{S}_B(t)$ has a bias which is lacking in the other MSE survival estimators we have derived. The bias here is of great importance. However, it is quite difficult to determine how the bias behaves, because each n_i and λ_i may change within the products. It is not an easy task to interpret the graph of the bias in a theoretical sense.

The method of trying to evaluate the bias empirically, or by example, is also a challenge. The true parameter for λ_i is not known. The purpose of the bias is to judge how close $\hat{S}_B(t)$ is to the true survival function $S(t)$. An attempt could be made to

estimate λ_i. However, estimating λ_i for the bias would be hazardous. The choice of an estimator for λ_i would dramatically influence the bias. For instance, if $\hat{\lambda}_i$ was used as an estimator for $\mathrm{E}\left(\hat{S}_B(t)\right)$ and $\hat{\lambda}_{Bi}$ was used as an estimator for $S(t)$, then there would be no bias. On the other hand, if $\hat{\lambda}_i$ was used as the estimator for both $\mathrm{E}\left(\hat{S}_B(t)\right)$ and $S(t)$, the bias might be quite large. Because of these dramatically varying outcomes, we make no attempt to determine an estimator for λ_i in trying to determine the bias.

The question then remains, how do we consider the bias of $\hat{S}_B(t)$? The simple answer is that we only have information about the bias of the parameter $\hat{\lambda}_{Bi}$. We can try to determine how the bias of $\hat{S}_B(t)$ will behave based on how the bias of $\hat{\lambda}_{Bi}$ behaves.

Our overall objective is to determine which survival estimator is the best. Just observing the behavior of $\mathrm{Bias}\,\hat{\lambda}_{Bi}$ is useful, but it can sometimes be difficult in accomplishing our overall objective.

The accuracy with which one can determine which survival estimate is best depends on the survival estimates being compared. It is often difficult to compare the survival estimators of $\hat{S}_B(t)$ and $\hat{S}_T(t)$ with the greatest accuracy. One might be tempted to compare their MSE parameters. However, since $\hat{\lambda}_T$ is a rate and $\hat{\lambda}_{Bi}$ is a proportion, it is difficult to accurately compare the two. Thus, we have to rely upon a comparison of $\mathrm{Var}\,\hat{S}_B(t)$ and $\mathrm{Var}\,\hat{S}_T(t)$. If a bias is relevant, then one would have to guess how relatively large or small the bias of $\hat{S}_B(t)$ might be according to the parameter $\hat{\lambda}_{Bi}$.

Comparing the survival estimators $\hat{S}_B(t)$ and $\hat{S}(t)$ is easier since their respective

parameters are both proportions and they come from the same distribution. We have knowledge about the $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ comparisons which were shown earlier. We can see how the bias affects the estimators, and we can make a fairly good determination about how the bias affects the comparison of $\hat{S}_B(t)$ and $\hat{S}(t)$.

In practice, for any survival distributions, we can compare their variances. If a survival distribution has a bias, then we have to consider the bias separately.

Analyzing the Bias of $\hat{\lambda}_{Bi}$. Since we cannot easily determine the bias of $\hat{S}_B(t)$ directly, we must instead analyze how λ_{Bi} is affected by its bias. Recall that MSE λ_{Bi} can be written as

$$\text{MSE}\,\hat{\lambda}_{Bi} = \text{Var}_{\lambda_i}\,\hat{\lambda}_{Bi} + \left(\text{Bias}_{\lambda_i}\,\hat{\lambda}_{Bi}\right)^2.$$

This implies,

$$E_{\lambda_i}\left(\hat{\lambda}_{Bi} - \lambda_i\right)^2 = \text{Var}_{\lambda_i}\,\hat{\lambda}_{Bi} + \left(E_{\lambda_i}\,\hat{\lambda}_{Bi} - \lambda_i\right)^2$$

which simplifies to

$$\frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}+n_i\right)^2} + \frac{n_i\left(1/2-\lambda_i\right)^2}{\left(\sqrt{n_i}+n_i\right)^2}.$$

Therefore,

$$\text{Var}_{\lambda_i}\,\hat{\lambda}_{Bi} = \frac{n_i\lambda_i\left(1-\lambda_i\right)}{\left(\sqrt{n_i}+n_i\right)^2} \quad \text{and} \quad \left(\text{Bias}_{\lambda_i}\,\hat{\lambda}_{Bi}\right)^2 = \frac{n_i\left(1/2-\lambda_i\right)^2}{\left(\sqrt{n_i}+n_i\right)^2}.$$

Here the bias and the variances are compared at different sample sizes as shown in Figure 5. We will use two of the sample sizes we used before when comparing the

MSEs of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$.

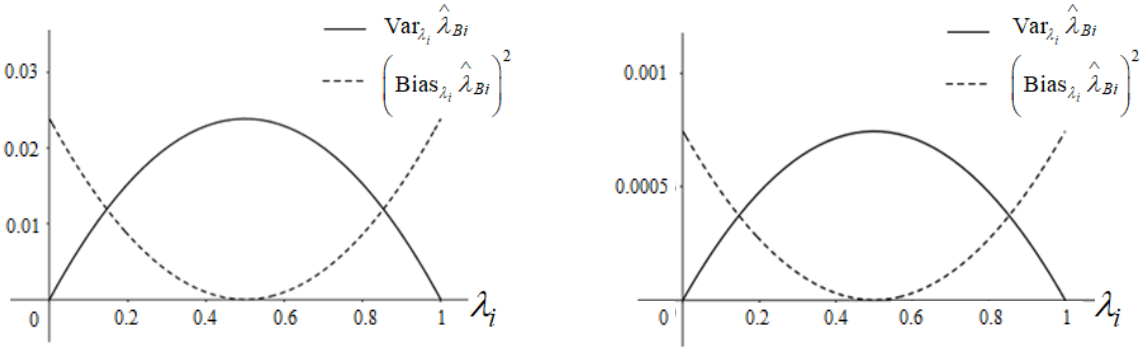


Figure 5. Comparison of $\text{Var}_{\lambda_i}\hat{\lambda}_{Bi}$ and $\left(\text{Bias}_{\lambda_i}\hat{\lambda}_{Bi}\right)^2$ at $n = 5$ and $n = 300$

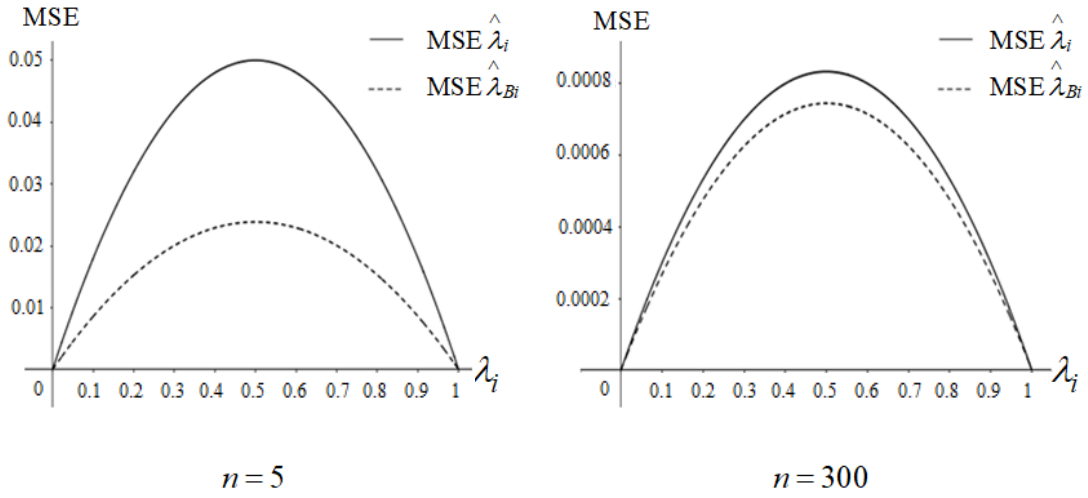
A couple of things can be observed from Figure 5. The first observation is that the relationship of the bias and the variance does not appear to change with sample size. Another observation is that the rate of increase and decrease for the variance and bias are inversely proportional. Note that since the variance is parabolic, this would have to be the case in order for the MSE of $\hat{\lambda}_{Bi}$ to be constant. The final observation is that the bias is largest near λ_i values of 1 and 0.

Unfortunately, these observations do not tell us with precision how the bias would influence the effectiveness of $\hat{S}_B(t)$ compared to some arbitrary survival estimation. It could be that the bias is enough to make $\hat{S}_B(t)$ not as effective. One cannot ignore the fact that when the variance is at its smallest, the bias is at its largest. However, it could also be that even when the bias is at its largest, it is still not enough to cause serious issue. While these are not definite answers, there is useful information here that should be noted

and considered in any practical application.

Another Comparison of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$. We will now turn our attention back to a comparison of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$. Two comparisons of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ will be explored. The first delves more deeply into an analysis of the bias while the second leads to a method of considering the true value of λ_i, in practice.

Revisiting the Bias. We desire to see how the bias affects the comparison of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ this time by removing the bias from the MSE of $\hat{\lambda}_{Bi}$. Figure 6 shows graphs of the $\text{MSE}\,\hat{\lambda}_{Bi}$ and $\text{MSE}\,\hat{\lambda}_i$ at different sample sizes, without the $\hat{\lambda}_{Bi}$ bias.



$n = 5$ $n = 300$

Figure 6. Comparison of $\text{MSE}\,\hat{\lambda}_{Bi}$ and $\text{MSE}\,\hat{\lambda}_i$ without Bias

It turns out that regardless of the sample size and the value of λ_i, $\text{MSE}\,\hat{\lambda}_{Bi}$ outperforms $\text{MSE}\,\hat{\lambda}_i$ if the bias for $\hat{\lambda}_{Bi}$ is not considered. However, we saw previously from Figure 6 that the $\text{MSE}\,\hat{\lambda}_i$ generally outperformed $\text{MSE}\,\hat{\lambda}_{Bi}$ when the sample size was large if the bias is considered. This might indicate something important about our

comparison of $\hat{S}(t)$ and $\hat{S}_B(t)$. If the sample size is small one might expect $\hat{S}_B(t)$ to be a better estimator than $\hat{S}(t)$. However, if the sample size is large, particular care must be taken about how the bias of $\hat{S}_B(t)$ would affect the results. If it turns out that $\text{Var}\,\hat{S}_B(t) \leq \text{Var}\,\hat{S}(t)$, by a relatively small margin, then one must be cautious in believing $\hat{S}_B(t)$ is the better estimator. It would likely be the case that the bias of $\hat{S}_B(t)$ would be large enough so that the bias would make $\text{MSE}\,\hat{S}_B(t)$ greater than $\text{MSE}\,\hat{S}(t)$. This might lead us to belief that $\hat{S}(t)$ would be the better estimator rather than $\hat{S}_B(t)$.

Estimating the True Value of λ_i. Up until now, we have operated under the assumption that the true value of λ_i is unknown when comparing $\text{MSE}\,\hat{\lambda}_i$ and $\text{MSE}\,\hat{\lambda}_{Bi}$. For instance, in Figure 3 and Figure 4 we assumed that the true value of λ_i was unknown and thus could be treated as a variable.

While our assumption is correct, it fails to offer enough insight into our comparison. In practice it is possible to have some knowledge and understanding about the true value of λ_i. Estimators of λ_i can be used in an attempt to understand the possible size of λ_i and thus help us decide which estimator is a better choice. While the estimators cannot be expected to fully represent λ_i, we would like to believe they have some measure of accuracy if we will ultimately choose one of the estimators to model the data.

Since we have interest in comparing $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$, $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ can be used to estimate λ_i. However, $\hat{\lambda}_i$ has traditionally suffered from inaccuracy when the sample size is small. In particular, the value of $\hat{\lambda}_i$ is often smaller than it should be. Thus, for

smaller sample sizes we use a modification of $\hat{\lambda}_i$, defined as

$$\tilde{\lambda}_i = \frac{d_i + 2}{n_i + 4}.$$

The above modification seeks to attain more accuracy by increasing the number of failures by 2 and the sample size by 4. This also results in a better confidence interval for the proportion. The modification of $\hat{\lambda}_i$ is based on a general modification of the binomial proportion $\hat{p}$ as presented by Agresti and Coull (1998).

Now that we have established which estimators to use for λ_i, $\text{MSE}\,\hat{\lambda}_i$ and $\text{MSE}\,\hat{\lambda}_{Bi}$ can be graphically compared with more insight. Figure 7 shows the graph of a general $\text{MSE}\,\hat{\lambda}_i$ and $\text{MSE}\,\hat{\lambda}_{Bi}$ comparison.

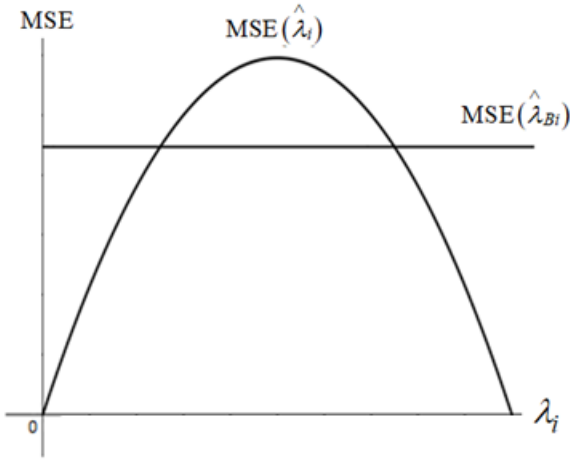


Figure 7. A General $\text{MSE}\,\hat{\lambda}_i$ and $\text{MSE}\,\hat{\lambda}_{Bi}$ Comparison

One can determine exactly when the true value of λ_i favors one estimator over the other by finding the λ_i-values where the two MSEs intersect. The points of intersection can be

found for any n_i by setting the two MSEs equal to each other and solving for λ_i. That is we solve

$$\frac{\lambda_i(1-\lambda_i)}{n_i} - \frac{n_i}{4\left(\sqrt{n_i}+n_i\right)^2} = 0$$

for λ_i.

The left intersection point turns out to be

$$\frac{\sqrt{2\bullet}}{2\bullet} \quad \frac{\sqrt{n_i}+1}{}$$

while the right intersection point turns out to be

$$\frac{-\sqrt{2\bullet}}{2\bullet} \quad \frac{\sqrt{n_i}-1}{}$$

For a given data set, once the intersection points are determined they can be compared to the λ_i estimates. Table 5 shows a format of presenting the λ_i estimates and intersection value comparisons.

In Table 5, the values of $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ can be compared to the $LI[\lambda_i]$ and $RI[\lambda_i]$ values. Each $Diff[\lambda_i]$ value is the difference between the estimate and the nearest point of intersection. If the λ_i estimate is between its corresponding $LI[\lambda_i]$ and $RI[\lambda_i]$, then the difference value is measured as positive, otherwise it is measured as negative. The positive and negative signs hold no significance except to judge whether the individual value favors $\text{MSE}\,\hat{\lambda}_i$ or $\text{MSE}\,\hat{\lambda}_{Bi}$. The sum of the estimate difference gives an overall idea of which parametric estimate might be appropriate for the data. If the sum is positive, then there is evidence that the $\hat{\lambda}_i$ parameter estimate is more accurate,

Table 5. Estimated λ_i Table Format. Note that $\hat{\lambda}_i$ is replaced with $\tilde{\lambda}_i$ if the sample size is small.

Survival Times	$\hat{\lambda}_i$	$\hat{\lambda}_{Bi}$	Right Point of MSE Intersection	Left Point of MSE Intersection	Difference Between Estimator and Nearest Intersection Point	
					$\hat{\lambda}_i$	$\hat{\lambda}_{Bi}$
t_1	$\hat{\lambda}_1$	$\hat{\lambda}_{B1}$	$LI[\lambda_1]$	$RI[\lambda_1]$	$Diff\left[\hat{\lambda}_1\right]$	$Diff\left[\hat{\lambda}_{B1}\right]$
t_2	$\hat{\lambda}_2$	$\hat{\lambda}_{B2}$	$LI[\lambda_2]$	$RI[\lambda_2]$	$Diff\left[\hat{\lambda}_2\right]$	$Diff\left[\hat{\lambda}_{B2}\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_i	$\hat{\lambda}_i$	$\hat{\lambda}_{Bi}$	$LI[\lambda_i]$	$RI[\lambda_i]$	$Diff\left[\hat{\lambda}_i\right]$	$Diff\left[\hat{\lambda}_{Bi}\right]$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_{n-1}	$\hat{\lambda}_{n-1}$	$\hat{\lambda}_{Bn-1}$	$LI[\lambda_{n-1}]$	$RI[\lambda_{n-1}]$	$Diff\left[\hat{\lambda}_{n-1}\right]$	$Diff\left[\hat{\lambda}_{Bn-1}\right]$
t_n	$\hat{\lambda}_n$	$\hat{\lambda}_{Bn}$	$LI[\lambda_n]$	$RI[\lambda_n]$	$Diff\left[\hat{\lambda}_n\right]$	$Diff\left[\hat{\lambda}_{Bn}\right]$
Totals					Sum of $\hat{\lambda}_i$ Differences	Sum of $\hat{\lambda}_{Bi}$ Differences

indicating that the Kaplan-Meir estimate might prove best in modeling the data. If the sum is negative, then there is evidence that the $\hat{\lambda}_{Bi}$ parameter estimate is more accurate, and thus indicates that the modified Kaplan-Meir estimate might be best in modeling the data.

Before we move on to the next section, we make a note of why a sum of differences is used in our comparison of the λ_i estimates and intersection values. The sum of the differences is used because the difference values have weighted effects on λ_i. If the distance between an estimate and an intersection point is large, then choosing the inappropriate estimate between $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ could result in a very inaccurate λ_i estimate. On the other hand, if the distance between an estimate and an intersection point is small, the choice between $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ would likely not be as consequential.

Life Tables

The life table method is a presentation of survival information that is modeled using the Kaplan-Meir estimates. It is designed to be used for relatively large sample sizes. The larger data sets are organized in intervals rather than considered at specific survival times. To accommodate the interval format, certain aspects of the Kaplan-Meir process are changed. These changes will be shown latter. There are three types of life tables, cohort lifetables, current life tables, and clinical life tables.

A cohort life table shows statistical information about a population that was born at the same time or started a process at the same time. The study follows the subjects throughout their survival times until death. Cohort studies are less common due to trouble with observation over long time periods.

A current life table is a life table that shows information about a population over a given time period without concern for when the subjects of the population were born or started the process. The rates and statistical information gathered are then used to guess the behavior of a hypothetical cohort starting at birth or year one and continuing until death.

Unlike the cohort and current life tables, a clinical life table shows statistical information about a specific study or experiment, rather than a population. These life tables are measured over a fixed amount of time and are usually focused on measuring the effects of a stimuli or condition. Clinical life tables are a common tool for researchers to analyze the effects of a specific illness or treatment. We will be focusing our attention on clinical studies and life tables, rather than cohort or current life tables.

The clinical life table has many variations. However, most lifetables have common elements. We focus on presenting those common elements as well as additional information which may be useful to a clinical researcher. Table 6 shows an example of a clinical lifetable format. Each column is described as follows.

The first column gives $\left[t_i - t_{i+1}\right)$, where $i = 1, ..., s$. These are the intervals in which the survival information is distributed. Recall that survival data is analyzed in fixed intervals for a life table rather than analyzed as single data values. The intervals start at and include t_i and continue until t_{i+1} which is not included. The time t_s marks the end of the study so that the last interval $\left[t_s - t_{s+1}\right)$ is infinite. The last interval is excluded in its analysis for some of the remaining columns because of its infinite nature. This exclusion is apparent when i is defined to be $i = 1, ..., s-1$.

Life Table Format. A similar life table was produced by Lee (1992), pg. 91.
ion was granted to use this adaptation.

Interval	Midpoint	Width	Number Lost to Follow-up	Number Withdrawn Alive	Number Dying	Number Entering Interval	Number Exposed to Risk	Interval Death Rate	Interval Survival Rate	$\hat{S}(t)$	$\hat{f}(t_m)$	$\hat{h}(t_m)$	$\mathrm{Var}\,\hat{S}(t)$	$\mathrm{Var}\,\hat{f}(t_m)$	$\mathrm{Var}\,\hat{h}(t_m)$	Lower 95% CI Bound	Upper 95% CI Bound
$[t_1 - t_2)$	t_{m1}	b_1	l_1	w_1	d_1	n'_1	n_1	$\hat{\lambda}_1$	$\hat{\omega}_1$	1	$\hat{f}(t_{m1})$	$\hat{h}(t_{m1})$		$\mathrm{Var}\,\hat{f}(t_{m1})$	$\mathrm{Var}\,\hat{h}(t_{m1})$		
$[t_2 - t_3)$	t_{m2}	b_2	l_2	w_2	d_2	n'_2	n_2	$\hat{\lambda}_2$	$\hat{\omega}_2$	$\hat{S}(t_2)$	$\hat{f}(t_{m2})$	$\hat{h}(t_{m2})$	$\mathrm{Var}\,\hat{S}(t_2)$	$\mathrm{Var}\,\hat{f}(t_{m2})$	$\mathrm{Var}\,\hat{h}(t_{m2})$	$L\left[\hat{S}(t_2)\right]$	$U\left[\hat{S}(t_2)\right]$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$[t_i - t_{i+1})$	t_{mi}	b_i	l_i	w_i	d_i	n'_i	n_i	$\hat{\lambda}_i$	$\hat{\omega}_i$	$\hat{S}(t_i)$	$\hat{f}(t_{mi})$	$\hat{h}(t_{mi})$	$\mathrm{Var}\,\hat{S}(t_i)$	$\mathrm{Var}\,\hat{f}(t_{mi})$	$\mathrm{Var}\,\hat{h}(t_{mi})$	$L\left[\hat{S}(t_i)\right]$	$U\left[\hat{S}(t_i)\right]$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$[t_{s-1} - t_s)$	$t_{m,s-1}$	b_{s-1}	l_{s-1}	w_{s-1}	d_{s-1}	n'_{s-1}	n_{s-1}	$\hat{\lambda}_{s-1}$	$\hat{\omega}_{s-1}$	$\hat{S}(t_{s-1})$	$\hat{f}(t_{m,s-1})$	$\hat{h}(t_{m,s-1})$	$\mathrm{Var}\,\hat{S}(t_{s-1})$	$\mathrm{Var}\,\hat{f}(t_{m,s-1})$	$\mathrm{Var}\,\hat{h}(t_{m,s-1})$	$L\left[\hat{S}(t_{s-1})\right]$	$U\left[\hat{S}(t_{s-1})\right]$
$[t_s - \infty)$			l_s	w_s	d_s	n'_s	n_s	1	0	$\hat{S}(t_s)$			$\mathrm{Var}\,\hat{S}(t_s)$			$L\left[\hat{S}(t_s)\right]$	$U\left[\hat{S}(t_s)\right]$

The second column contains the midpoint of the interval and is denoted t_{mi}, where $t_{mi} = (t_i + t_{i+1})/2$ and $i = 1, ..., s-1$. The midpoint is needed for plotting the hazard and probability density functions.

The third column is the width of each interval b_i, where $b_i = t_{i+1} - t_i$ and $i = 1, ..., s-1$. The width is used in finding the hazard and probability density functions in subsequent columns.

The fourth column is the number of individuals lost to follow-up in the ith interval and is designated l_i, where $i = 1, ..., s$.

The fifth column is the number of individuals withdrawn alive at the end of the study in the ith interval. It is denoted w_i, where $i = 1, ..., s$. In a life table, the time intervals are seen as time since entering the study. Thus, subjects that have entered the study late may withdraw alive at an earlier interval than others that have entered at a previous time.

The six column represents the number of deaths in the ith interval, designated d_i where $i = 1, ..., s$.

The seventh column represents the number of subjects at the beginning of the ith interval, denoted n_i', where $n_1' = n$ and $i = 1, ..., s$. One can determine n_i' by determining n_{i-1}' and subtracting those that die, are lost to follow up, or that are withdrawn alive. That is, $n_i' = n_{i-1}' - l_{i-1} - w_{i-1} - d_{i-1}$.

The eighth column is the number of subjects that are at risk of death, denoted n_i where $i = 1, ..., s$. Since the data is organized in intervals, there is no specific information shown regarding when in the interval a subject might withdraw alive or be lost to follow

up. The times for loss or withdraw are assumed to be uniformly distribution and thus a subject that withdraws or is lost to follow up is considered at risk for only half the interval. This results in $n_i = n_i' - (1/2)\bullet$.

The ninth column shows the estimated probability of death within the *i*th interval, denoted $\hat{\lambda}_i$ where $i = 1,....,s$, and $\hat{\lambda}_s = 1$. This is the same $\hat{\lambda}_i$ found in the Kaplan Meir estimator. That is $\hat{\lambda}_i = d_i / n_i$. However, in this case, d_i and n_i are defined over an interval rather than at a single time. Note that $\hat{\lambda}_i$ is the estimated probability of death only considering risk within the *i*th , and not before it. The terminology rate of death within the *i*th interval will be used to describe $\hat{\lambda}_i$, since it is the proportion of subjects whom have died in the *i*th interval.

The tenth column is the estimated probability of survival within the *i*th interval, denoted $\hat{\omega}_i$ where $i = 1,....,s$, and $\hat{\omega}_s = 0$. This is defined as $\hat{\omega}_i = 1 - \hat{\lambda}_i$. The terminology used for $\hat{\omega}_i$ is rate of survival within the *i*th interval.

The eleventh column is the estimated survival function, $\hat{S}(t_i)$, where $i = 1,....,s$ and $\hat{S}(t_1) = 1$. This is the Kaplan-Meir estimator for $S(t_i)$ as shown previously. However, $\hat{S}(t_i)$ defined for data values over an interval is different than $\hat{S}(t_i)$ defined for single data values. For single values, $\hat{S}(t_i) = \hat{S}(t_{i-1})\hat{\omega}_i$, since the probability of surviving at t_i is equal to the probability of surviving at t_i and surviving all previous times. However, over an interval, $\hat{S}(t_i) = \hat{S}(t_{i-1})\hat{\omega}_{i-1}$. Here $\hat{S}(t_i)$ measures the probability of survival to the start of the interval $[t_i, t_{i+1})$. This is equivalent to surviving to the start of $[t_{i-1}, t_i)$ and surviving that interval itself. If $\hat{S}(t_i)$ was defined as

$\hat{S}(t_i) = \hat{S}(t_{i-1})\hat{\omega}_i$ we might assume survival through previous intervals and the entire interval $[t_i, t_{i+1})$ up until t_{i+1}. This would indicate survival past the time t_i, which would not be appropriate for survival at t_i.

The twelfth column is an estimation of the probability function at the midpoint, denoted $\hat{f}(t_{mi})$, where $i = 1,....,s-1$. Recall that $\hat{\lambda}_i$ denotes the probability of death within the *i*th interval. However, it is not known where in the interval death might occur. We can choose to believe that death is likely to occur in the middle of the interval so that our error is never to large with respect to where death actually occurs. Hence, we choose to estimate $f(t_{mi})$ rather than some arbitrary value $f(t_h)$, where $t_h \in [t_i, t_{i+1})$. However, we are assuming that the probability of death at any point in the interval is equal. Therefore, there will be no real difference between the estimate of $f(t_{mi})$ and the estimate of any other $f(t_h)$, such as $\hat{f}(t_i)$. It is known that $\hat{f}(t_i) = \hat{\lambda}_i \prod_{j=1}^{i-1}\left(1-\hat{\lambda}_j\right)$ and does not change due to the data being organized in intervals. However, we know that $\hat{S}(t_i)$ does. We can define $\hat{S}(t_i)$ as

$$\hat{S}(t_i) = \prod_{j=1}^{i-1}\left(1-\hat{\lambda}_j\right).$$

Thus,

$$\hat{f}(t_{mi}) = \hat{f}(t_i) = \hat{\lambda}_i \prod_{j=1}^{i-1}\left(1-\hat{\lambda}_j\right) = \hat{\lambda}_i \hat{S}(t_i).$$

It is preferable to divide $\hat{\lambda}_i \hat{S}(t_i)$ by b_i so that the probability is given per unit width, leading to

$$\hat{f}(t_{mi}) = \hat{\lambda}_i \hat{S}(t_i) / b_i .$$

The thirteenth column presents an estimate of the hazard function at the midpoint, denoted $\hat{h}(t_{mi})$ where $i = 1,....,s-1$. Since $h(t_i) = f(t_i) / S(t_i)$ we define $\hat{h}(t_{mi})$ as

$$\hat{h}(t_{mi}) = \frac{\hat{f}(t_{mi})}{\hat{S}(t_{mi})} .$$

We have defined $\hat{f}(t_{mi})$ previously, but we still need to derive $\hat{S}(t_{mi})$. Since $S(t_i)$ is the probability of survival at the beginning of the interval, rather than the midpoint, we define $\hat{S}(t_{mi})$ as

$$\hat{S}(t_{mi}) = \frac{1}{2}\left[\hat{S}(t_i) + \hat{S}(t_{i+1})\right].$$

Thus,

$$\hat{h}(t_{mi}) = \frac{\frac{\hat{\lambda}_i \hat{S}(t_i)}{b_i}}{\frac{1}{2}\left[\hat{S}(t_i) + \hat{S}(t_{i+1})\right]} = \frac{2\hat{\lambda}_i \hat{S}(t_i)}{b_i\left[\hat{S}(t_i) + \hat{S}(t_{i+1})\right]} = \frac{2\hat{\lambda}_i \hat{S}(t_i)}{b_i\left[\hat{S}(t_i) + \hat{S}(t_i)\hat{\omega}_i\right]} = \frac{2\hat{\lambda}_i}{b_i\left[2 - \hat{\lambda}_i\right]} .$$

The fourteenth, fifteenth, and sixteenth columns are the variances of $\hat{S}(t_i)$, $\hat{f}(t_{mi})$, and $\hat{h}(t_{mi})$ respectively. The equation for $\mathrm{Var}\,\hat{S}(t_i)$ was presented by Greenwood (1926). Gehan (1969) first derived $\mathrm{Var}\,\hat{f}(t_{mi})$, and $\mathrm{Var}\,\hat{h}(t_{mi})$. The variance for $\hat{S}(t_i)$ with data organized in intervals is the same as the variance of $\hat{S}(t_i)$ with single data values, except that $\hat{S}(t_i)$ is defined differently. The variances for $\hat{f}(t_{mi})$ and

$\hat{h}(t_{mi})$ can be found using the delta method. We rederive the three survival time functions below.

As we have seen before,

$$\operatorname{Var}\hat{S}(t_i) \approx \left[\hat{S}(t_i)\right]^2 \sum_{j=1}^{i-1} \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

Note that $\operatorname{Var}\hat{S}(t_1)$ does not exist since $\sum_{j=1}^{i-1} \frac{d_j}{n_j\left(n_j - d_j\right)}$ can't be found. When substituting i for 1, there does not exist d_js and n_js for j values less than 1.

We find $\operatorname{Var}\hat{f}(t_{mi})$ by first finding $\operatorname{Var}\left[\log \hat{S}(t_i)\,\hat{\lambda}_i\right]$.

$$\operatorname{Var}\left[\log \hat{S}(t_i)\,\hat{\lambda}_i\right] = \operatorname{Var}\left[\log \hat{S}(t_i)\right] + \operatorname{Var}\left[\log \hat{\lambda}_i\right]$$

By the delta method,

$$\operatorname{Var}\left[\log \hat{\lambda}_i\right] \approx \frac{\left(1-\hat{\lambda}_i\right)}{\hat{\lambda}_i\, n_i}.$$

Thus,

$$\operatorname{Var}\left[\log \hat{S}(t_i)\right] + \operatorname{Var}\left[\log \hat{\lambda}_i\right] \approx \sum_{j=1}^{i-1} \frac{\hat{\lambda}_j}{n_j\left(1-\hat{\lambda}_j\right)} + \frac{\left(1-\hat{\lambda}_i\right)}{\hat{\lambda}_i\, n_i}.$$

By a second use of the delta method,

$$\operatorname{Var}\left[\hat{S}(t_i)\,\hat{\lambda}_i\right] \approx \left[\hat{f}(t_{mi})\right]^2 \left[\sum_{j=1}^{i-1} \frac{\hat{\lambda}_j}{n_j\,\hat{\omega}_j} + \frac{\hat{\omega}_i}{\hat{\lambda}_i\, n_i}\right].$$

Therefore,

$$\operatorname{Var}\hat{f}\left(t_{mi}\right) \approx \frac{\left[\hat{f}\left(t_{mi}\right)\right]^2}{b_i^{\,2}}\left[\sum_{j=1}^{i-1}\frac{\hat{\lambda}_j}{n_j\,\hat{\omega}_j}+\frac{\hat{\omega}_i}{\hat{\lambda}_i\,n_i}\right]=\frac{\left[\hat{f}\left(t_{mi}\right)\right]^2}{b_i^{\,2}}\left[\sum_{j=1}^{i-1}\frac{d_j}{n_j\left(n_j-d_j\right)}+\frac{\left(n_i-d_i\right)}{n_i d_i}\right].$$

We found previously that $\hat{h}\left(t_{mi}\right)=2\hat{\lambda}_i/b_i\left[2-\hat{\lambda}_i\right]$. Using the delta method once again, we have

$$\operatorname{Var}\hat{h}\left(t_{mi}\right) \approx \frac{16\,\hat{\lambda}_i\,\hat{\omega}_i}{n_i b_i^{\,2}\left(2-\hat{\lambda}_i\right)^4}=\frac{16 n_i d_i\left(n_i-d_i\right)}{b_i^{\,2}\left(2n_i-d_i\right)^4}.$$

The seventeenth and eighteenth columns are the upper and lower limits of the confidence interval for $\hat{S}(t_i)$. One could also add columns to display the confidence intervals for $\hat{f}\left(t_{mi}\right)$ and $\hat{h}\left(t_{mi}\right)$ one so desired.

The confidence interval for $\hat{S}(t_i)$ here is similar to the confidence interval derived previously for $\hat{S}(t_i)$ with single data values. It is

$$\left[\hat{S}\left(t_i\right)^{e^{\frac{1.96}{\left[\log\hat{S}(t_i)\right]^2}\sum_{j=1}^{i-1}\frac{d_j}{n_j\left(n_j-d_j\right)}}},\ \hat{S}\left(t_i\right)^{e^{-\frac{1.96}{\left[\log\hat{S}(t_i)\right]^2}\sum_{j=1}^{i-1}\frac{d_j}{n_j\left(n_j-d_j\right)}}}\right].$$

This confidence interval does not exist for $\hat{S}\left(t_1\right)$ since $\sum_{j=1}^{0}\frac{d_j}{n_j\left(n_j-d_j\right)}$ does not make sense.

Life Table for the Modified Kaplan-Meir Estimate. A life table requires a relatively large sample size so that the data can be grouped into intervals. The modified Kaplan-Meir estimate would likely not be the best estimate for a large sample size. However, it is still possible to create a life table using the modified Kaplan-Meir estimate for the purpose of comparing the results with the standard life table. Additionally, it is possible for the modified Kaplan-Meir estimate to be a more efficient estimator, even with a large sample size, if the true values of the λ_i parameters are close to 0.5.

The two life tables would be identical in nature except for columns 11 onward. Columns 11-13 have estimated survival, probability, and hazard functions similar to the standard life table, except that the survival time functions use the modified Kaplan-Meir estimators. Those functions are

$$\hat{S}_B(t_i) = \prod_{j=1}^{i-1}\left(1 - \frac{n_j}{\sqrt{n_j + n_j}} \cdot \frac{\sqrt{n_j}/2}{\sqrt{n_j} + n_j}\right),$$

$$\hat{f}_B(t_{mi}) = \hat{\lambda}_{i_B}\hat{S}_B(t_i) = \left[\frac{n_i}{\sqrt{n_i + n_i}} \cdot \frac{\sqrt{n_i}/2}{\sqrt{n_i} + n_i}\right]\hat{S}_B(t_i),$$

and

$$\hat{h}_B(t_{mi}) = \frac{2\hat{\lambda}_{i_B}}{b_i\left[2 - \hat{\lambda}_{i_B}\right]} = \frac{2\left(\frac{n_i}{\sqrt{n_i + n_i}} \cdot \frac{\sqrt{n_i}/2}{\sqrt{n_i} + n_i}\right)}{b_i\left[2 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \cdot \frac{\sqrt{n_i}/2}{\sqrt{n_i} + n_i}\right)\right]}.$$

We can make use of the delta method as we did previously to find the variances in the final columns. The variances are

$$\operatorname{Var}\hat{S}_B(t_i) \approx \left[\hat{S}_B(t_i)\right]^2 \sum_{j=1}^{i-1} \frac{d_j\left(n_j - d_j\right)}{\left(\sqrt{n_j}/2 + n_j - d_j\right)^2 n_j},$$

$$\operatorname{Var}\hat{f}_B(t_{mi}) \approx \frac{\left[\hat{f}_B(t_{mi})\right]^2}{b_i^2}\left[\sum_{j=1}^{i-1} \frac{n_j \hat{\lambda}_j \hat{\omega}_j}{\left(\sqrt{n_j}/2 + n_j - n_j \hat{\lambda}_j\right)^2} + \frac{n_i \hat{\lambda}_i \hat{\omega}_i}{\left(\sqrt{n_i}/2 + n_i \hat{\lambda}_i\right)^2}\right]$$

$$= \frac{\left[\hat{f}_B(t_{mi})\right]^2}{b_i^2}\left[\sum_{j=1}^{i-1} \frac{d_j\left(n_j - d_j\right)}{n_j\left(\sqrt{n_j}/2 + n_j - d_j\right)^2} + \frac{d_j\left(n_j - d_j\right)}{n_j\left(\sqrt{n_i}/2 + d_j\right)^2}\right],$$

and

$$\operatorname{Var}\hat{h}_B(t_{mi}) \approx \frac{16\, n_i \hat{\lambda}_i \hat{\omega}_i \left(\sqrt{n_i} + n_i\right)^2}{b_i^2\left(\frac{3}{2}\sqrt{n_i} + 2n_i - n_i \hat{\lambda}_i\right)^4} = \frac{16\, d_i\left(n_i - d_i\right)\left(\sqrt{n_i} + n_i\right)^2}{n_i b_i^2\left(\frac{3}{2}\sqrt{n_i} + 2n_i - d_i\right)^4}.$$

The final columns show the upper and lower confidence intervals for $\hat{S}_B(t_i)$.

$$\left[\hat{S}_B(t_i)^{e^{\frac{1.96}{\left[\log \hat{S}_B(t_i)\right]^2}\sum_{j=1}^{i-1}\frac{d_j\left(n_j-d_j\right)}{\left(\sqrt{n_j}/2+n_j-d_j\right)^2 n_j}}}, \hat{S}_B(t_i)^{e^{-\frac{1.96}{\left[\log \hat{S}_B(t_i)\right]^2}\sum_{j=1}^{i-1}\frac{d_j\left(n_j-d_j\right)}{\left(\sqrt{n_j}/2+n_j-d_j\right)^2 n_j}}}\right].$$

Note that the confidence interval for $\hat{S}_B(t_1)$ does not exist.

Table 7 shows the lifetable format for the modified Kaplan-Meir estimate.

d Kaplan-Meir Life Table Format. Adapted from Lee (1992), pg. 91

Interval	Midpoint	Width	Number Lost to Follow-up	Number Withdrawn Alive	Number Dying	Number Entering Interval	Number Exposed to Risk	Interval Death Rate	Interval Survival Rate	$\hat{S}_B(t)$	$\hat{f}_B(t_m)$	$\hat{h}_B(t_m)$	$\widehat{\mathrm{Var}}\,S_B(t)$	$\widehat{\mathrm{Var}}\,\hat{f}_B(t_m)$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_m)$	Lower 95% CI Bound	Upper 95% CI Bound
$[t_1 - t_2)$	t_{m1}	b_1	l_1	w_1	d_1	n'_1	n_1	$\hat{\lambda}_1$	$\hat{\omega}_1$	1	$\hat{f}_B(t_{m1})$	$\hat{h}_B(t_{m1})$		$\widehat{\mathrm{Var}}\,\hat{f}_B(t_{m1})$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_{m1})$		
$[t_2 - t_3)$	t_{m2}	b_2	l_2	w_2	d_2	n'_2	n_2	$\hat{\lambda}_2$	$\hat{\omega}_2$	$\hat{S}_B(t_2)$	$\hat{f}_B(t_{m2})$	$\hat{h}_B(t_{m2})$	$\widehat{\mathrm{Var}}\,S_B(t_2)$	$\widehat{\mathrm{Var}}\,\hat{f}_B(t_{m2})$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_{m2})$	$L\left[\hat{S}_B(t_2)\right]$	$U\left[\hat{S}_B(t_2)\right]$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$[t_i - t_{i+1})$	t_{mi}	b_i	l_i	w_i	d_i	n'_i	n_i	$\hat{\lambda}_i$	$\hat{\omega}_i$	$\hat{S}_B(t_i)$	$\hat{f}_B(t_{mi})$	$\hat{h}_B(t_{mi})$	$\widehat{\mathrm{Var}}\,S_B(t_i)$	$\widehat{\mathrm{Var}}\,\hat{f}_B(t_{mi})$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_{mi})$	$L\left[\hat{S}_B(t_i)\right]$	$U\left[\hat{S}_B(t_i)\right]$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$[t_{s-1} - t_s)$	$t_{m,s-1}$	b_{s-1}	l_{s-1}	w_{s-1}	d_{s-1}	n'_{s-1}	n_{s-1}	$\hat{\lambda}_{s-1}$	$\hat{\omega}_{s-1}$	$\hat{S}_B(t_{s-1})$	$\hat{f}_B(t_{m,s-1})$	$\hat{h}_B(t_{m,s-1})$	$\widehat{\mathrm{Var}}\,S_B(t_{s-1})$	$\widehat{\mathrm{Var}}\,\hat{f}_B(t_{m,s-1})$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_{m,s-1})$	$L\left[\hat{S}_B(t_{s-1})\right]$	$U\left[\hat{S}_B(t_{s-1})\right]$
$[t_s - \infty)$			l_s	w_s	d_s	n'_s	n_s	1	0	$\hat{S}_B(t_s)$			$\widehat{\mathrm{Var}}\,\hat{S}_B(t_s)$			$L\left[\hat{S}_B(t_s)\right]$	$U\left[\hat{S}_B(t_s)\right]$

DATA ANALYSIS

Overview

It is important to see how the different methods for survival estimation compare and function in practice. For this reason, we will analyze two sets of data. The first data set will have a sample size of 21. The first data set has a relatively small sample size and is less than 42. Thus, we might expect that the modified Kaplan-Meir estimate would be a far better estimate than the standard Kaplan-Meir estimate for that data set. The second data will have a sample size of 2418. Since this data set has a relatively large sample, being far greater than 42, it is likely that the standard Kaplan-Meir estimate would be better than the modified Kaplan-Meir estimate. The parametric method may or may not be better than both the Kaplan-Meir estimates, and is determined by how well the data matches the shape of the known distribution. If the data fits the the general shape well, then we might expect the parametric estimations to do best.

The statistical package R in conjunction with Microsoft Excel was used in the analysis of our data sets.

Acute Leukemia Data Analysis

The first data set is from a study to assess the effectiveness of the drug 6-mercaptopurine (6-MP) on patients with acute leukemia cancer. The remission times of the patients were organized and reported by Freireich et al. (1963). The remissions times are considered the "failure times." The study consists of 42 patients split into two groups of 21, and ended after one year. One group was given the drug, and another received a

placebo. For our purposes, we will only analyze the group given the drug. The following are the remission times (in weeks) of the patients given 6-MP.

6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+

A plus indicates that the observation was censored. The subjects were enrolled at different times. Study termination is the cause for each censored value.

Parametric Estimate. First we will use the parametric method to model the data set. Recall that the hazard plot will be needed with the estimated hazard values. Table 8 shows the cumulative hazard calculations.

Table 8. Hazard Table for Patients with Acute Leukemia

Survival Times	Number at Risk	Hazard Values	Cumulative Hazard Values
6	21	4.76	4.76
6	19	5.26	10.02
6	18	5.56	15.58
7	17	5.88	21.46
10	15	6.67	28.13
13	12	8.33	36.46
16	11	9.09	45.55
22	7	14.29	59.84
23	6	16.67	76.51

From Table 8, a hazard graph is created and shown in Figure 8.

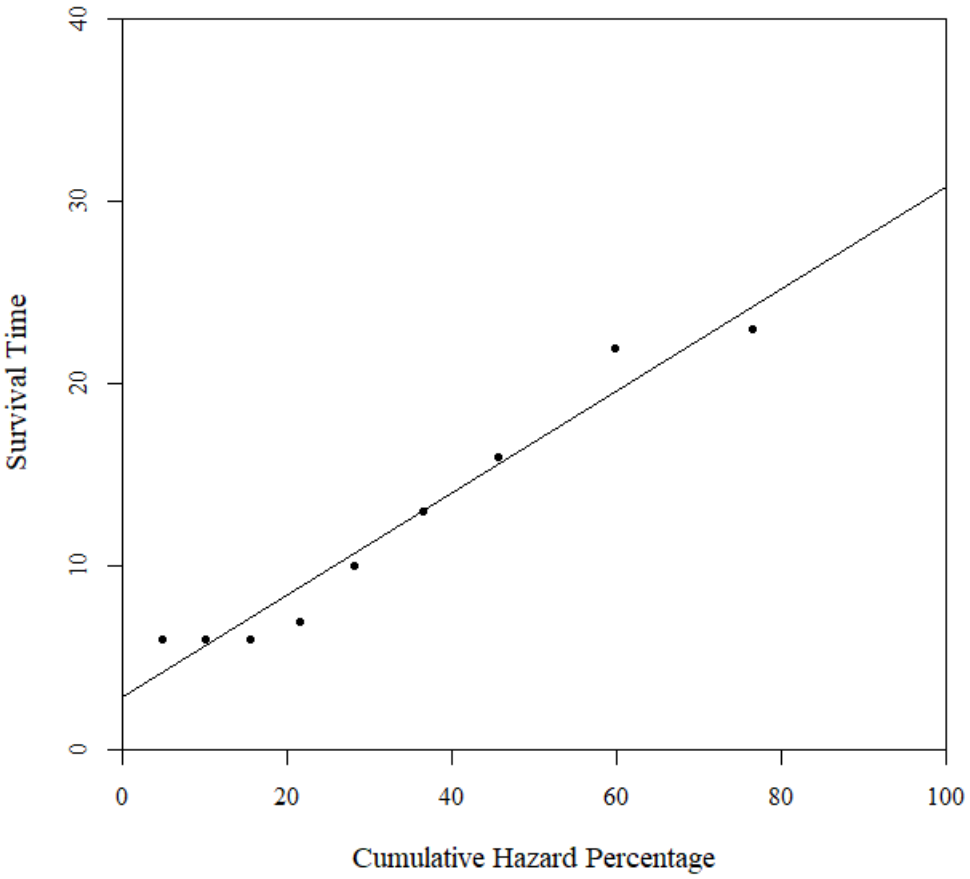


Figure 8. Hazard Graph for Patients with Acute Leukemia

The data appears to be linear in nature, thus we choose an exponential distribution to model the data. It was found previously that a parametric estimate of λ for an exponential distribution is $\hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i$, where $\delta_i = 0$ if censored or $\delta_i = 1$ if uncensored. Thus, we calculate that $\hat{\lambda} = 9/359 = 0.02507$ and $f(t) = 0.02507e^{-0.02507t}$ would be a good parametric model for the data set.

Choosing $f(t) = 0.02507e^{-0.02507t}$ as our parametric distribution, the

corresponding survival function would be $\hat{S}_T(t) = e^{-0.02507t}$. Recall that

$$\operatorname{Var}\hat{S}_T(t) = \frac{\left(\sum_{i=1}^{n}\delta_i\right)t^2 e^{-2\bullet\left(\sum_{i=1}^{n}t_i\right)t}}{\left(\sum_{i=1}^{n}t_i\right)^2}.$$

Thus,

$$\operatorname{Var}\hat{S}_T(t) = \frac{9t^2e^{-2\bullet}}{(359)^2} = 0.00007t^2e^{-0.050139t}.$$

Table 9 displays the theoretical survival table for $\hat{S}_T(t)$. This will be used for comparison purposes latter.

Table 9. Theoretical Survival Table for Patients with Acute Leukemia

Survival Times	$\hat{S}_T(t)$	Var $\hat{S}_T(t)$
6	0.8603466	0.001861
7	0.839045788	0.002409
10	0.778255813	0.004230
13	0.721870153	0.006150
16	0.669569708	0.008015
22	0.576061991	0.011216
23	0.561799643	0.011660

Recall that the confidence interval for any exponential $\hat{S}_T(t)$ is

$$\left[\left(e^{-\hat{\lambda}t}\right)^{e^{1.96\left(1/\sum_{i=1}^{n}\delta_i\right)}}, \left(e^{-\hat{\lambda}t}\right)^{e^{-1.96\left(1/\sum_{i=1}^{n}\delta_i\right)}}\right].$$

Thus, the confidence interval in this case would be

$$\left[\left(e^{-0.02507t}\right)^{e^{1.96*(1/9)}}, \left(e^{-0.02507t}\right)^{e^{-1.96*(1/9)}}\right] = \left[e^{-0.03117t}, e^{-0.020164t}\right].$$

A graph of $S_T(t)$ and its 95% confidence interval are shown in Figure 9.

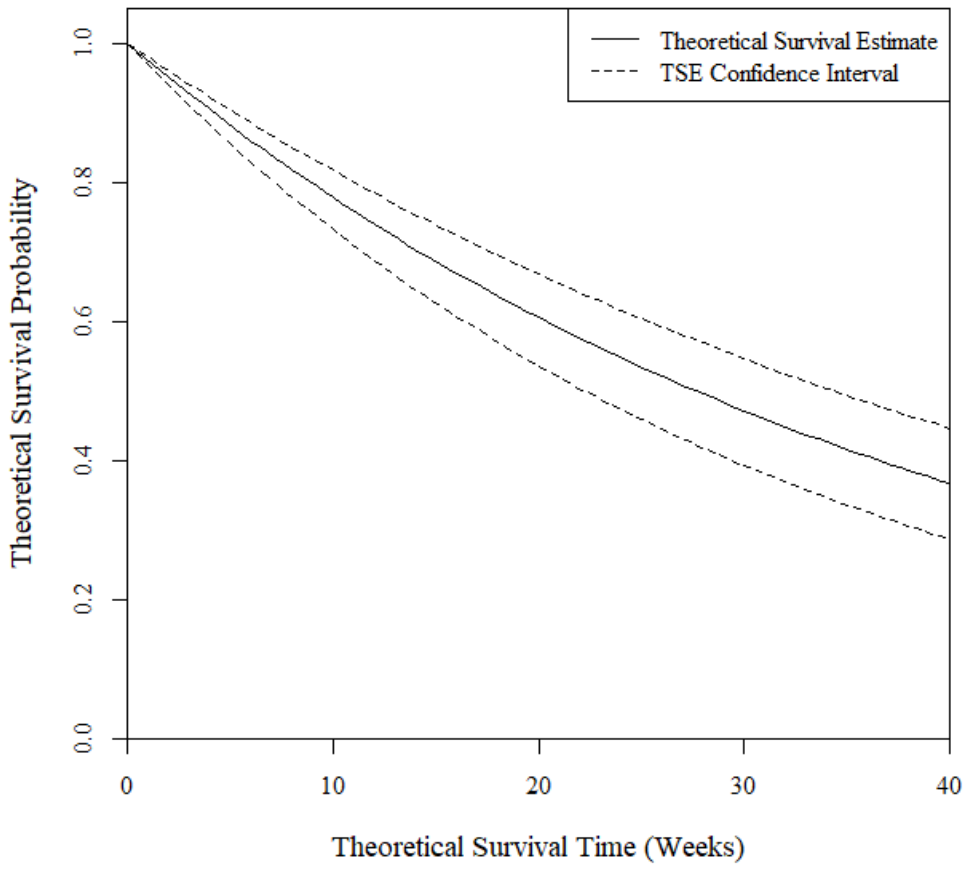


Figure 9. Theoretical Survival Estimate for Patients with Acute Leukemia

Kaplan-Meir Estimate. The data set will now be modeled using the Kaplan-Meir process. Recall that the survivor function, $S(t)$, can be estimated as $\hat{S}(t) = \prod_{t_i \le t} \left(1 - \hat{\lambda}_i\right)$, where $\hat{\lambda}_i = d_i / n_i$. The computation of each $\hat{S}(t_i)$, along with its variance, confidence intervals, and other survivor information is shown in Table 10.

Table 10. Kaplan-Meir Table for Patients with Acute Leukemia. A traditional life table is not used since the sample size is small.

Survival Times	Number at Risk	Number of Failures	$\hat{S}(t)$	Var $\hat{S}(t)$	Lower 95% CI Bound	Upper 95% CI Bound
6	21	3	0.8571	0.005830	0.7743	0.9230
7	17	1	0.8067	0.007558	0.7034	0.8771
10	15	1	0.7529	0.009281	0.6553	0.8265
13	12	1	0.6902	0.011408	0.5935	0.7683
16	11	1	0.6275	0.013009	0.5337	0.7076
22	7	1	0.5378	0.016442	0.4367	0.6286
23	6	1	0.4482	0.018116	0.3479	0.5434

From Table 10, the graph of $\hat{S}(t)$ and its 95% confidence interval is constructed in Figure 10.

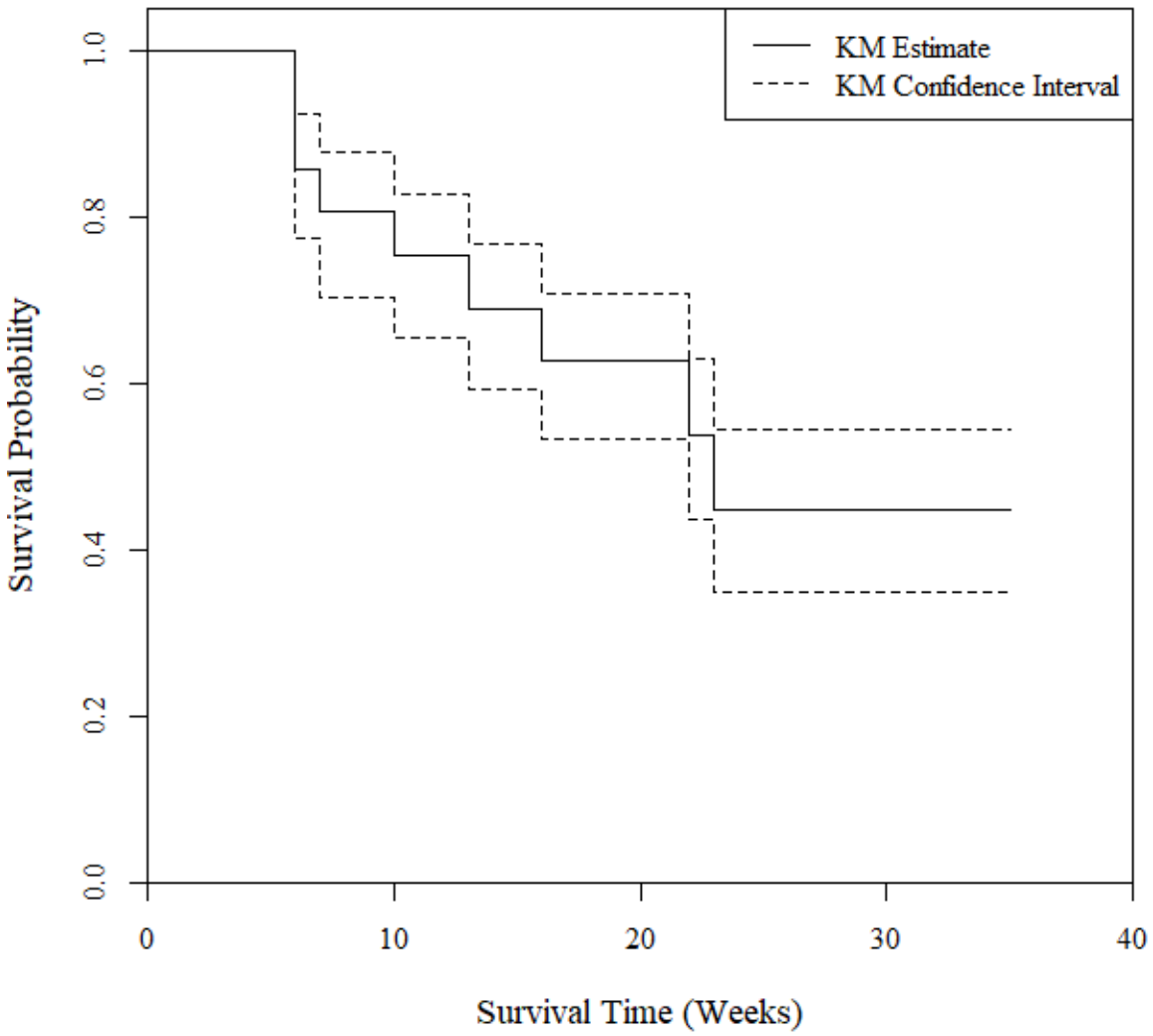


Figure 10. Kaplan-Meir Estimate for Patients with Acute Leukemia

Modified Kaplan-Meir Estimate. Finally, we model the data set using the modified Kaplan-Meir process. For the modified Kaplan-Meir process, recall that $S(t)$ is estimated as $\hat{S}_B(t) = \prod_{t_i \le t} \left[1 - \left(\frac{n_i}{\sqrt{n_i + n_i}} \cdot \frac{\lambda_i}{\sqrt{n_i + n_i}} \right) \right]$, where $\hat{\lambda}_i = d_i / n_i$. Table 11 shows the important survival information, including $\text{Var}\ \hat{S}_B(t)$ and confidence intervals for $\hat{S}_B(t)$.

Table 11. Modified Kaplan-Meir Table for Patients with Acute Leukemia

Survival Times	Number at Risk	Number of Failures	$\hat{S}_B(t)$	Var $\hat{S}_B(t)$	Lower 95% CI Bound	Upper 95% CI Bound
6	21	3	0.7932	0.003929	0.7475	0.8316
7	17	1	0.6782	0.004200	0.6458	0.7083
10	15	1	0.5727	0.004200	0.5465	0.5980
13	12	1	0.4715	0.004104	0.4486	0.4940
16	11	1	0.3840	0.003708	0.3642	0.4037
22	7	1	0.2915	0.003495	0.2726	0.3107
23	6	1	0.2148	0.002889	0.1979	0.2322

The graph of $\hat{S}_B(t)$ and its 95% confidence interval are shown in Figure 11.

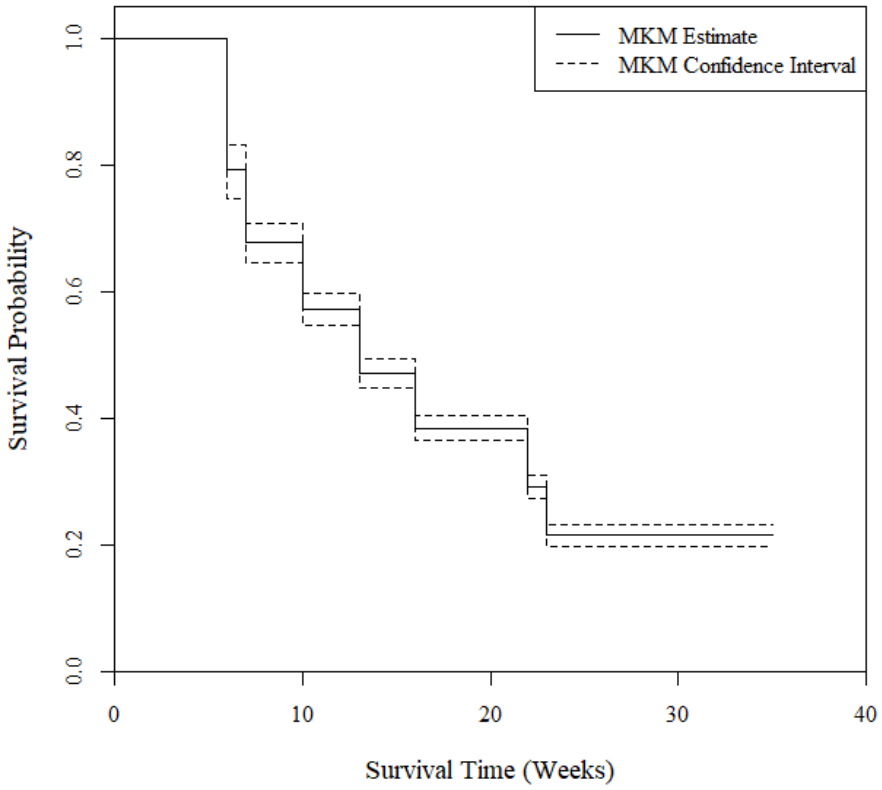


Figure 11. Modified Kaplan-Meir Estimate for Patients with Acute Leukemia

Comparing Survival Variances. Here we will compare the variances of the three estimates. We will latter consider the bias of the modified Kaplan-Meir estimate in a discussion of the results.

Since there is more than one variance for each estimate, the overall sums of the respective survival function variances are compared. That is, we can compare $\sum_{i=1}^{n} \text{Var } \hat{S}_T(t_i)$, $\sum_{i=1}^{n} \text{Var } \hat{S}(t_i)$, and $\sum_{i=1}^{n} \text{Var } \hat{S}_B(t_i)$. We wish to see which summation is smaller. A smaller summation is one indication of a better corresponding survival estimate. Table 12 shows the survival function variances at each t_i along with their summations.

Table 12. Comparison of Survival Variances for Patients with Acute Leukemia

Survival Times	Var $\hat{S}_T(t)$	Var $\hat{S}(t)$	Var $\hat{S}_B(t)$
6	0.001861	0.005830	0.003929
7	0.002409	0.007558	0.004200
10	0.004230	0.009281	0.004200
13	0.006150	0.011408	0.004104
16	0.008015	0.013009	0.003708
22	0.011216	0.016442	0.003495
23	0.011660	0.018116	0.002889
Totals	0.045541	0.081644	0.023825

Table 12 shows that $\sum_{i=1}^{n} \text{Var}\ \hat{S}_B(t_i) \le \sum_{i=1}^{n} \text{Var}\ \hat{S}_T(t_i) \le \sum_{i=1}^{n} \text{Var}\ \hat{S}(t_i)$.

Estimating the True Value of the Kaplan-Meir λ_i. In order to compare $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$, a λ_i estimate table is created. Table 13 shows a comparison between the λ_i estimates and the intersection points of MSE $\hat{\lambda}_i$ and MSE $\hat{\lambda}_{Bi}$.

Table 13. Estimated λ_i Table for Patients with Acute Leukemia. A plus sign is given to difference values between the right and left intersection points. A minus sign is given to difference values outside of this interval.

Survival Times	$\tilde{\lambda}_i$ [1]	$\hat{\lambda}_{Bi}$	Right Point of MSE Intersection	Left Point of MSE Intersection	Difference Between Estimator and Nearest Intersection Point	
					$\tilde{\lambda}_i$	$\hat{\lambda}_{Bi}$
6	0.2000	0.2068	0.2144	0.7856	-0.0144	-0.0076
7	0.1429	0.1449	0.2032	0.7968	-0.0603	-0.058
10	0.1579	0.1556	0.1966	0.8034	-0.0387	-0.0410
13	0.1875	0.1767	0.1800	0.8200	0.0075	-0.0080
16	0.2000	0.1857	0.1800	0.8200	0.0200	0.0057
22	0.2727	0.2408	0.1560	0.8440	0.1167	0.0848
23	0.3000	0.2633	0.1480	0.8520	0.1520	0.1153
Totals					0.1828	0.0911

[1] Due to small sample size, $\tilde{\lambda}_i$ is used.

Positive difference summations in table 13 show good evidence that $\hat{\lambda}_{Bi}$ would be the more accurate estimate for λ_i.

Angina Pectoris Data Analysis

The second data set is from a study measuring the time until death of 2418 males with angina pectoris. The data was organized and reported by Parker et al. (1946). An observation is measured as survival time since diagnosis until death. The data was organized in 16 intervals, with each interval being one year. The study ended after 15 years. Right censoring occurred for some of the observations, due to loss of follow-up. The observation of subjects started at different times. The data will be shown in lifetables, due to the quantity of observations.

For this data set, we will take a different approach in finding an appropriate parametric method. Recall, that one can make a choice concerning which parametric function is best based on the shape of the non-parametric graphs. We will use this method here. The non-parametric methods will be presented first. Then, the general shape of the graph will be observed and compared to the shape of theoretical survival functions. An appropriate parametric function will then be found and compared to the Kaplan-Meir estimates.

Kaplan-Meir Estimate. We first model the data using the Kaplan-Meir process for life tables. Recall that each $S(t_i)$, can be estimated as $\hat{S}(t_i) = \prod_{j=1}^{i-1}\left(1-\hat{\lambda}_j\right)$, where $\hat{\lambda}_j = d_j / n_j$. The lifetable using the Kaplan-Meir process is shown in Table 14. For our purposes, the values of $\hat{f}(t_m)$, $\hat{h}(t_m)$, Var $\hat{f}(t_m)$, and Var $\hat{h}(t_m)$ will not be analyzed.

r Patients with Angina Pectoris. A similar life table format was produced by Lee (1992), pg. 91.

Interval	Midpoint	Width	Number Lost to Follow-up	Number Withdrawn Alive	Number Dying	Number Entering Interval	Number Exposed to Risk	Interval Death Rate	Interval Survival Rate	$\hat{S}(t)$	$\hat{f}(t_m)$	$\hat{h}(t_m)$	$\widehat{\mathrm{Var}}\,\hat{S}(t)$	$\widehat{\mathrm{Var}}\,\hat{f}(t_m)$	$\widehat{\mathrm{Var}}\,\hat{h}(t_m)$	Lower 95% CI Bound	Upper 95% CI Bound
[0–1)	0.5	1	0	0	456	2418	2418.0	0.1886	0.8114	1.0000	0.1886	0.2082		0.000064	0.000094		
[1–2)	1.5	1	39	0	226	1962	1942.5	0.1163	0.8837	0.8114	0.0944	0.1235	0.000064	0.000036	0.000067	0.8107	0.8121
[2–3)	2.5	1	22	0	152	1697	1686.0	0.0902	0.9098	0.7170	0.0646	0.0944	0.000085	0.000026	0.000058	0.7163	0.7177
[3–4)	3.5	1	23	0	171	1523	1511.5	0.1131	0.8869	0.6524	0.0738	0.1199	0.000094	0.000029	0.000085	0.6517	0.6531
[4–5)	4.5	1	24	0	135	1329	1317.0	0.1025	0.8975	0.5786	0.0593	0.1080	0.000102	0.000024	0.000086	0.5780	0.5792
[5–6)	5.5	1	107	0	125	1170	1116.5	0.1120	0.8880	0.5193	0.0581	0.1186	0.000106	0.000025	0.000110	0.5187	0.5199
[6–7)	6.5	1	133	0	83	938	871.5	0.0952	0.9048	0.4611	0.0439	0.1000	0.000108	0.000022	0.000121	0.4605	0.4617
[7–8)	7.5	1	102	0	74	722	671.0	0.1103	0.8897	0.4172	0.0460	0.1167	0.000110	0.000027	0.000182	0.4166	0.4178
[8–9)	8.5	1	68	0	51	546	512.0	0.0996	0.9004	0.3712	0.0370	0.1048	0.000112	0.000025	0.000216	0.3706	0.3718
[9–10)	9.5	1	64	0	42	427	395.0	0.1063	0.8937	0.3342	0.0355	0.1123	0.000114	0.000028	0.000299	0.3336	0.3348
[10–11)	10.5	1	45	0	43	321	298.5	0.1441	0.8559	0.2987	0.0430	0.1552	0.000109	0.000040	0.000557	0.2981	0.2993
[11–12)	11.5	1	53	0	34	233	206.5	0.1646	0.8354	0.2557	0.0421	0.1794	0.000123	0.000046	0.000936	0.2550	0.2564
[12–13)	12.5	1	33	0	18	146	129.5	0.1390	0.8610	0.2136	0.0297	0.1494	0.000130	0.000045	0.001232	0.2128	0.2144
[13–14)	13.5	1	27	0	9	95	81.5	0.1104	0.8896	0.1839	0.0203	0.1169	0.000139	0.000042	0.001513	0.1830	0.1848
[14–15)	14.5	1	23	0	6	59	47.5	0.1263	0.8737	0.1636	0.0207	0.1348	0.000151	0.000064	0.003014	0.1626	0.1646
[15–∞)			0	0	0	30	30.0	1.0000	0	0.1429			0.000177			0.1417	0.1441

The graph of $\hat{S}(t)$ is shown in Figure 12.

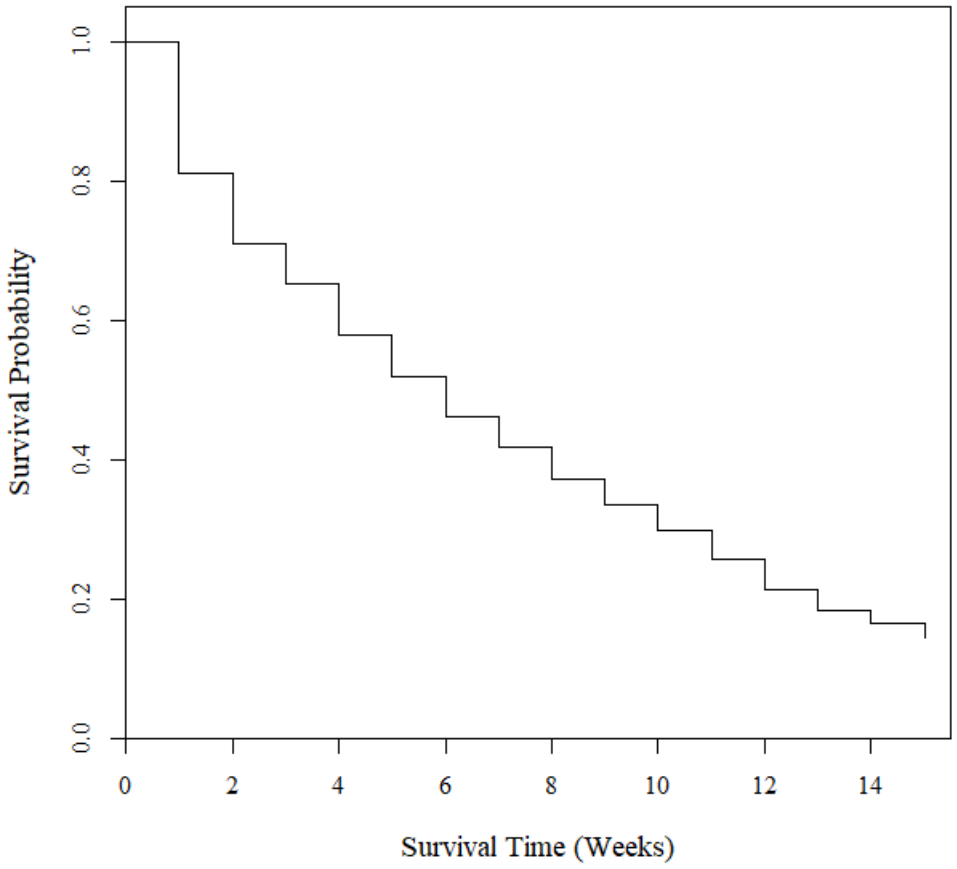


Figure 12. Kaplan-Meir Estimate for Patients with Angina Pectoris. The confidence intervals are not included since they are too small to observe.

Modified Kaplan-Meir Estimate. For the Modified Kaplan-Meir process, recall that each $\hat{S}_B(t_i)$ is estimated as $\hat{S}_B(t) = \prod_{j=1}^{i-1}\left[1 - \left(\frac{n_j}{\sqrt{n_j + n_j}} \cdot \frac{\sqrt{n_j}/2}{\sqrt{n_j} + n_j}\right)\right]$, where $\hat{\lambda}_j = d_j / n_j$. The modified Kaplan-Meir life table is shown in Table 15. The values of $\hat{f}_B(t_m)$, $\hat{h}_B(t_m)$, Var $\hat{f}_B(t_m)$, and Var $\hat{h}_B(t_m)$ will not be analyzed.

n-Meir Life Table for Patients with Angina Pectoris. A similar life table was produced by Lee (1992),

Interval	Midpoint	Width	Number Lost to Follow-up	Number Withdrawn Alive	Number Dying	Number Entering Interval	Number Exposed to Risk	Interval Death Rate	Interval Survival Rate	$\hat{S}_B(t)$	$\hat{f}_B(t_m)$	$\hat{h}_B(t_m)$	$\widehat{\mathrm{Var}}\,\hat{S}_B(t)$	$\widehat{\mathrm{Var}}\,\hat{f}_B(t_m)$	$\widehat{\mathrm{Var}}\,\hat{h}_B(t_m)$	Lower 95% CI Bound	Upper 95% CI Bound
[0–1)	0.5	1	0	0	456	2418	2418.0	0.1886	0.8114	1.0000	0.1948	0.2158		0.000061	0.000092		
[1–2)	1.5	1	39	0	226	1962	1942.5	0.1163	0.8837	0.8052	0.1005	0.1332	0.000061	0.000034	0.000065	0.8045	0.8059
[2–3)	2.5	1	22	0	152	1697	1686.0	0.0902	0.9098	0.7047	0.0704	0.1052	0.000079	0.000024	0.000057	0.7040	0.7053
[3–4)	3.5	1	23	0	171	1523	1511.5	0.1131	0.8869	0.6343	0.0779	0.1309	0.000087	0.000027	0.000081	0.6337	0.6349
[4–5)	4.5	1	24	0	135	1329	1317.0	0.1025	0.8975	0.5564	0.0630	0.1200	0.000093	0.000022	0.000084	0.5558	0.5569
[5–6)	5.5	1	107	0	125	1170	1116.5	0.1120	0.8880	0.4934	0.0608	0.1313	0.000093	0.000022	0.000108	0.4929	0.4939
[6–7)	6.5	1	133	0	83	938	871.5	0.0952	0.9048	0.4326	0.0469	0.1147	0.000092	0.000018	0.000126	0.4321	0.4331
[7–8)	7.5	1	102	0	74	722	671.0	0.1103	0.8897	0.3857	0.0481	0.1331	0.000091	0.000022	0.000175	0.3852	0.3861
[8–9)	8.5	1	68	0	51	546	512.0	0.0996	0.9004	0.3375	0.0393	0.1238	0.000090	0.000020	0.000204	0.3371	0.3380
[9–10)	9.5	1	64	0	42	427	395.0	0.1063	0.8937	0.2982	0.0373	0.1335	0.000088	0.000021	0.000282	0.2977	0.2987
[10–11)	10.5	1	45	0	43	321	298.5	0.1441	0.8559	0.2609	0.0427	0.1781	0.000087	0.000027	0.000519	0.2604	0.2614
[11–12)	11.5	1	53	0	34	233	206.5	0.1646	0.8354	0.2182	0.0407	0.2056	0.000086	0.000031	0.000861	0.2177	0.2187
[12–13)	12.5	1	33	0	18	146	129.5	0.1390	0.8610	0.1775	0.0299	0.1836	0.000085	0.000027	0.001110	0.1770	0.1781
[13–14)	13.5	1	27	0	9	95	81.5	0.1104	0.8896	0.1477	0.0220	0.1613	0.000083	0.000023	0.001332	0.1471	0.1482
[14–15)	14.5	1	23	0	6	59	47.5	0.1263	0.8737	0.1256	0.0218	0.1902	0.000081	0.000030	0.002548	0.1250	0.1262
[15–∞)			0	0	0	30	30.0	1.0000	0	0.1038			0.000084			0.1031	0.1045

The graph of $\hat{S}_B(t)$ is shown in Figure 13.

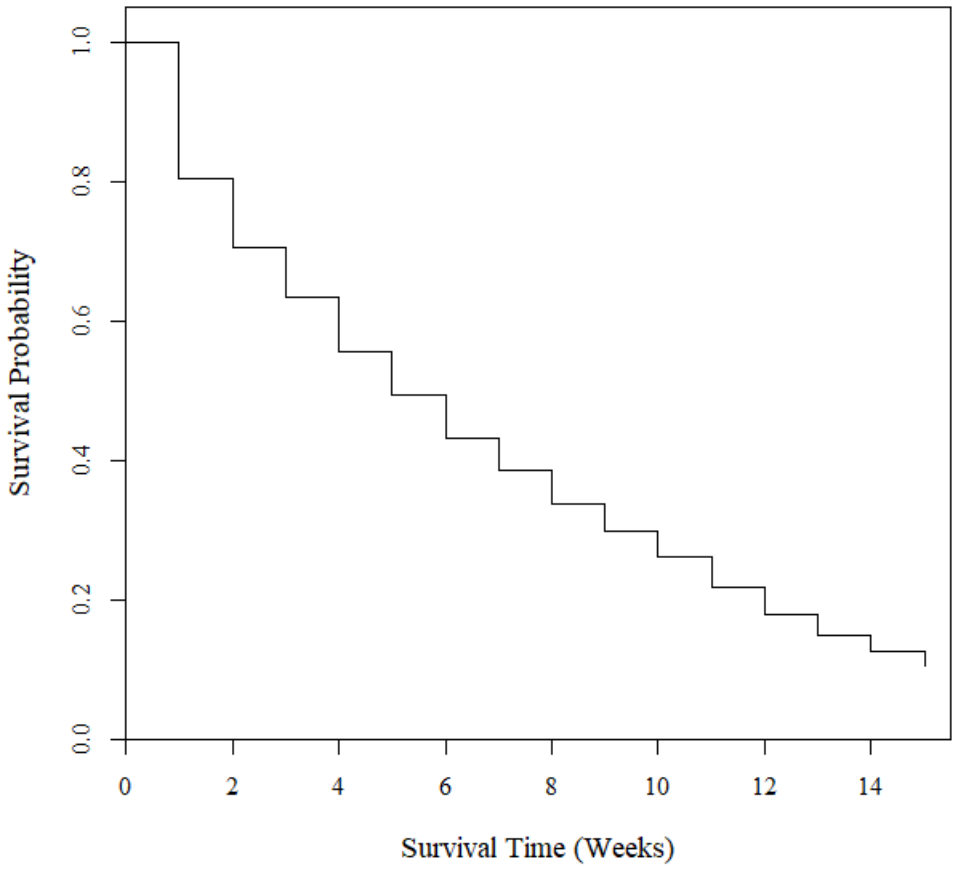


Figure 13. Modified Kaplan-Meir Estimate for Patients with Angina Pectoris. The confidence intervals are not included since they are too small to observe.

Parametric Method. In order to model an appropriate parametric model, we must choose which non-parametric estimate to analyze. In this case, the modified Kaplan-Meir estimate and standard Kaplan-Meir estimate are similar enough in shape that either would be sufficient to choose. In this case, it turns out that the Kaplan-Meir estimate is likely the more accurate estimate. Thus, we will analyze the Kaplan-Meir estimate. A smooth representation of $\hat{S}(t)$ is shown in Figure 14.

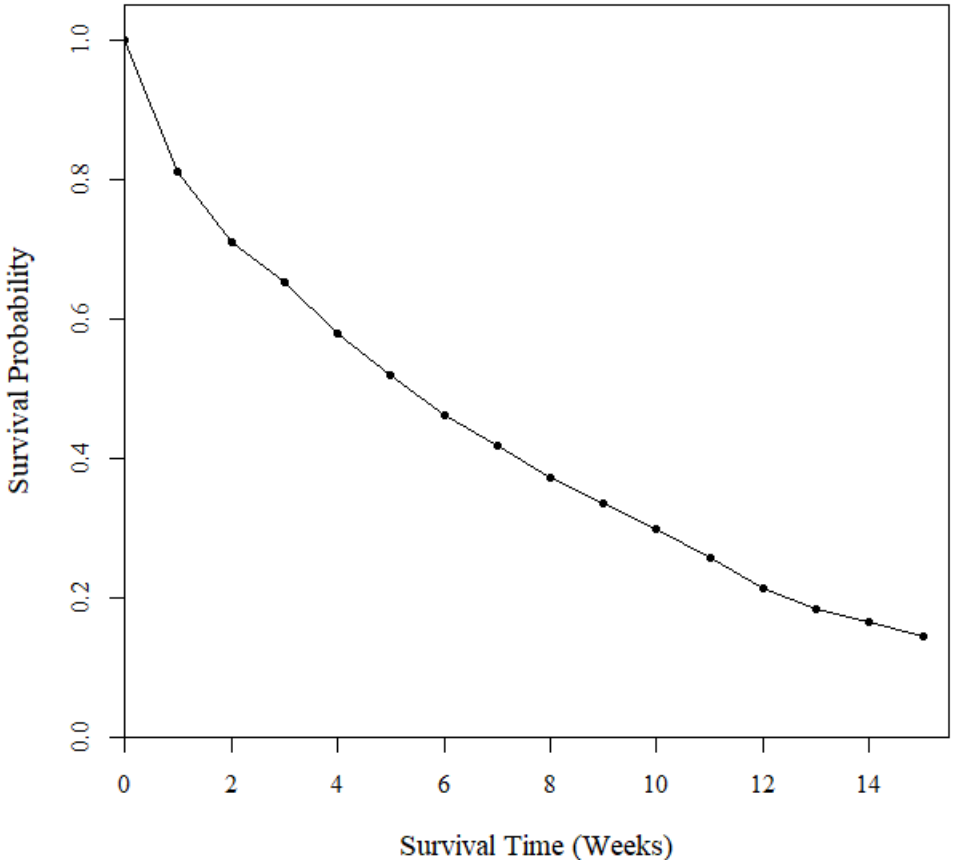


Figure 14. Kaplan-Meir Curve for Patients with Angina Pectoris

If we compare the above Kaplan-Meir curve with the graph of survival functions from known distributions, it appears that the exponential survival function would fit the data well.

Since an exponential survival function is used for our parametric model, an estimation of the parameter λ needs to be found. A suitable modification of $\hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} t_i$ for a lifetable is used as an estimator for λ. The indicator variable, δ_i is replaced with $n - \sum_{i=1}^{s} l_i + w_i$, n is replaced with s, and each t_i is replaced with t_{mi}• w_i). Since a lifetable does not show the exact failure times or times of

censoring within the interval, the failures or censorships are estimated as taking place at the midpoint in each interval, Thus,

$$\hat{\lambda} = \frac{n - \sum_{i=1}^{s} l_i + w_i}{\sum_{i=1}^{s} t_{mi} \bullet \qquad w_i)} = 0.138587 .$$

We can conclude that the function $f(t) = 0.138587e^{-0.138587t}$ would be a good parametric model for the data set. The exponential survival function is $\hat{S}_T(t) = e^{-0.138587t}$.

It is known that

$$\operatorname{Var} \hat{S}_T(t) = \frac{\left(\sum_{i=1}^{n} \delta_i \right) t^2 e^{-2 \bullet \left(\sum_{i=1}^{n} t_i \right) t}}{\left(\sum_{i=1}^{n} t_i \right)^2} .$$

Making the substitutions given above, we have

$$\operatorname{Var} \hat{S}_T(t) = \frac{\left(n - \sum_{i=1}^{s} l_i + w_i \right) t^2 e^{-2 \bullet \left(\sum_{i=1}^{s} t_{mi} \bullet \qquad w_i \right) t}}{\left[\sum_{i=1}^{s} t_{mi} \bullet \qquad w_i) \right]^2}$$

$$= \frac{(1655) t^2 e^{-2 \bullet \qquad t}}{[11942]^2}$$

$$= 0.000012 t^2 e^{-0.277174t} .$$

Table 16 shows $\hat{S}_T(t)$ at each uncensored t_i, along with the respective variances. This will be used for comparison purposes latter.

Table 16. Theoretical Survival Table for Patients with Angina Pectoris

Survival Times	$\hat{S}_T(t)$	Var $\hat{S}_T(t)$
1	0.870588	0.000009
2	0.757923	0.000027
3	0.659838	0.000045
4	0.574447	0.000061
5	0.500106	0.000073
6	0.435386	0.000079
7	0.379042	0.000082
8	0.329989	0.000081
9	0.287284	0.000078
10	0.250106	0.000073
11	0.217739	0.000067
12	0.189561	0.000060
13	0.165030	0.000053
14	0.143673	0.000047
15	0.125080	0.000041

The confidence intervals for $\hat{S}_T(t)$ are

$$\left[\left(e^{-\hat{\lambda}t}\right)^{e^{1.96\left(1/n-\sum_{i=1}^{s} l_i+w_i\right)}},\left(e^{-\hat{\lambda}t}\right)^{e^{-1.96\left(1/n-\sum_{i=1}^{s} l_i+w_i\right)}}\right]$$

$$=\left[\left(e^{-0.063892t}\right)^{e^{1.96(1/1655)}},\left(e^{-0.063892t}\right)^{e^{-1.96(1/1655)}}\right]$$

$$= \left[e^{-0.13875t}, e^{-0.138422t} \right].$$

The graph of $\hat{S}_T(t)$ is shown in Figure 15.

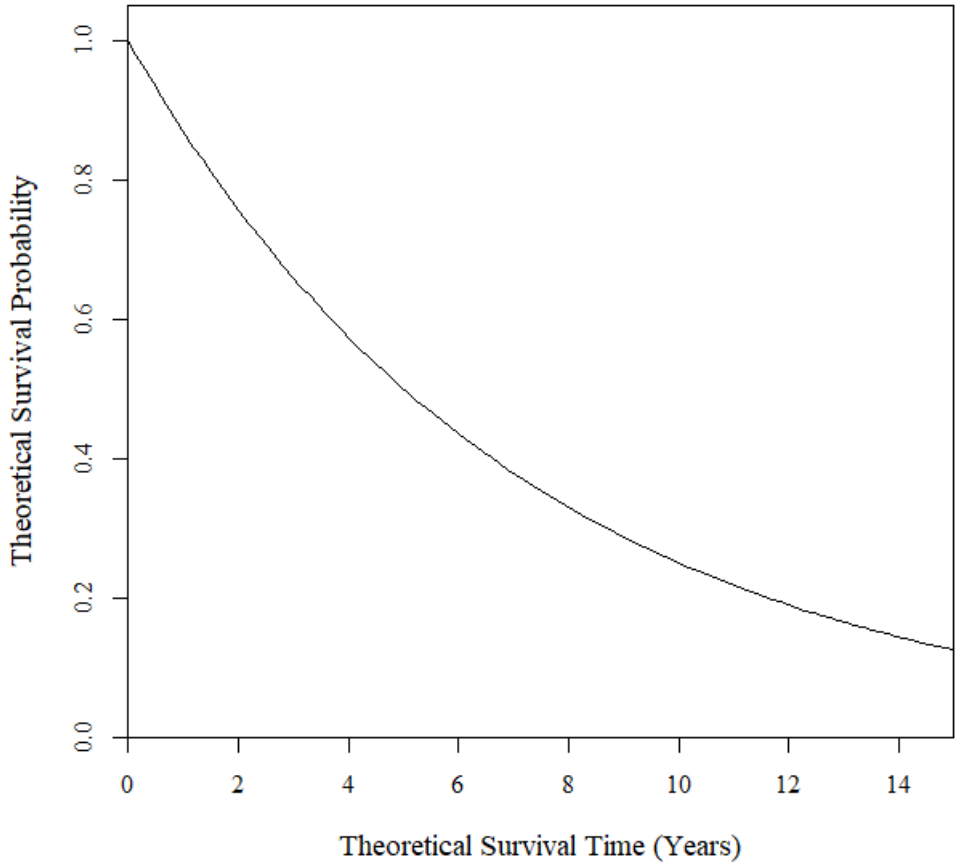


Figure 15. Theoretical Survival Estimate for Patients with Angina Pectoris. The confidence intervals are left out since they are too small to observe.

Comparing Survival Variances. Here the summation of each estimate variance is analyzed. A consideration of the bias of will be addressed in a discussion of the results. The variances of the respective survival functions are shown in Table 17.

Table 17. Comparison of Survival Variances for Patients with Angina Pectoris

Survival Intervals	Var $\hat{S}_T(t)$	Var $\hat{S}(t)$	Var $\hat{S}_B(t)$
$[1-2)$	0.000009	0.000064	0.000061
$[2-3)$	0.000027	0.000085	0.000079
$[3-4)$	0.000045	0.000094	0.000087
$[4-5)$	0.000061	0.000102	0.000093
$[5-6)$	0.000073	0.000106	0.000093
$[6-7)$	0.000079	0.000108	0.000092
$[7-8)$	0.000082	0.000110	0.000091
$[8-9)$	0.000081	0.000112	0.000090
$[9-10)$	0.000078	0.000114	0.000088
$[10-11)$	0.000073	0.000109	0.000087
$[11-12)$	0.000067	0.000123	0.000086
$[12-13)$	0.000060	0.000130	0.000085
$[13-14)$	0.000053	0.000139	0.000083
$[14-15)$	0.000047	0.000151	0.000081
$[15-\infty)$	0.000041	0.000177	0.000084
Totals	0.000875	0.001724	0.001280

Table 17 shows that $\sum_{i=1}^{n} \text{Var } \hat{S}_T(t_i) \le \sum_{i=1}^{n} \text{Var } \hat{S}_B(t_i) \le \sum_{i=1}^{n} \text{Var } \hat{S}(t_i)$.

Estimating the True Value of the Kaplan-Meir λ_i. The estimates $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ are compared by creating and analyzing an estimate λ_i table. Table 18 shows a comparison between the λ_i estimates and the intersection points of $\text{MSE}\,\hat{\lambda}_i$ and $\text{MSE}\,\hat{\lambda}_{Bi}$

Table 18. Estimated λ_i Table for Patients with Angina Pectoris. A plus sign is given to difference values between the right and left intersection points. A minus sign is given to difference values outside of this interval.

Survival Times	$\hat{\lambda}_i$	$\hat{\lambda}_{Bi}$	Right Point of MSE Intersection	Left Point of MSE Intersection	Difference Between Estimator and Nearest Intersection Point	
					$\hat{\lambda}_i$	$\hat{\lambda}_{Bi}$
[0-1)	0.1886	0.1948	0.4007	0.5993	-0.2121	-0.2059
[1-2)	0.1163	0.1249	0.3953	0.6047	-0.2790	-0.2704
[2-3)	0.0902	0.0999	0.3916	0.6083	-0.3014	-0.2917
[3-4)	0.1131	0.1228	0.3887	0.6113	-0.2756	-0.2659
[4-5)	0.1025	0.1132	0.3850	0.6150	-0.2825	-0.2718
[5-6)	0.1120	0.1232	0.3803	0.6197	-0.2683	-0.2571
[6-7)	0.0952	0.1085	0.3731	0.6269	-0.2779	-0.2646
[7-8)	0.1103	0.1248	0.3649	0.6351	-0.2546	-0.2402
[8-9)	0.0996	0.1166	0.3561	0.6439	-0.2565	-0.2395
[9-10)	0.1063	0.1252	0.3471	0.6529	-0.2408	-0.2219
[10-11)	0.1441	0.1635	0.3369	0.6631	-0.1928	-0.1734
[11-12)	0.1646	0.1865	0.3226	0.6774	-0.1580	-0.1361
[12-13)	0.1390	0.1682	0.3031	0.3031	-0.1641	-0.1350
[13-14)	0.1104	0.1493	0.2823	0.7177	-0.1719	-0.1331
[14-15)	0.1263	0.1737	0.2564	0.7436	-0.1301	-0.0827
Totals					-3.4656	-3.1893

The negative difference summations in Table 18 show good evidence that $\hat{\lambda}_i$ would be the more accurate estimate for λ_i.

DISCUSSION

The results for the Acute Leukemia data analysis are partially what we expected. If we only analyze the variances, the modified Kaplan-Meir estimate appears to be better than the standard Kaplan-Meir estimate. Additionally, the modified Kaplan-Meir estimate seems to outperform the exponential estimate. However, if we consider the bias, we might be able to get a more accurate understanding of which estimation is best.

In our analysis of the Acute Leukemia data, we don't know exactly how large the bias of $\hat{S}_B(t)$ is, however an analysis about the overall nature of the parameters, might suggest something about the size of the bias. When the true value of λ_i was estimated and compared to the intersection of the $\hat{\lambda}_i$ and $\hat{\lambda}_{Bi}$ MSE, we found that the results favored the $\hat{\lambda}_{Bi}$ estimate. While this does not directly show that the bias of $\hat{S}_B(t)$ is small, it does show evidence that the bias of $\hat{\lambda}_{Bi}$ is small so that $\text{MSE}\,\hat{\lambda}_{Bi}$ is less than $\text{MSE}\,\hat{\lambda}_i$. We know that the bias $\hat{S}_B(t)$ is affected by the bias of $\hat{\lambda}_{Bi}$ so that it is likely the bias $\hat{S}_B(t)$ is also small. Thus, we believe with some confidence that the bias $\hat{S}_B(t)$ is small enough to allow $\text{MSE}\,\hat{S}_B(t)$ to still be less than $\text{MSE}\,\hat{S}(t)$. We have less knowledge about how an exponential survival estimate compares to a Kaplan-Meir estimate, but we know that a theoretical estimate should outperform a non-parametric estimate if the data can be accurately modeled with a theoretical distribution. Thus, we might guess that the bias is small, but large enough so that $\text{MSE}\,\hat{S}_T(t)$ is less than $\text{MSE}\,\hat{S}_B(t)$.

After considering the bias in our mean square errors, we make the conclusion that the following is likely true.

$$\text{MSE}\,\hat{S}_T(t) \le \text{MSE}\,\hat{S}_B(t) \le \text{MSE}\,\hat{S}(t).$$

This would result in the exponential survival function being the best estimate followed by the modified Kaplan-Meir estimate, and the traditional Kaplan-Meir estimate being the worst of the three.

While our conclusion for the Acute Leukemia data analysis is likely true, there are a number of possible scenarios that would lead to a conclusion other than the one we suggested. The possibilities are discussed as follows.

The bias of the modified Kaplan-Meir estimate might be smaller or larger than expected. The first reason for this might be due to parameter estimation. While using an estimate of the Kaplan-Meir λ_i gave us some understanding of λ_i, the estimate was not the true value itself. It could be that our estimate may have been insufficiently. The second reason the $\text{bias}\,\hat{S}_B(t)$ might be different than expected is that we only know information about the variance of the survival function, not the bias. Since we do not know the $\text{bias}\,\hat{S}_B(t)$, it is possible that the parameter bias affected the corresponding survival function bias in ways we did not predict. Two things could result from a $\text{bias}\,\hat{S}_B(t)$ different than expected. The first is that it could be too small to affect the MSE in any way. This would result in the Kaplan-Meir estimate being the most accurate survival estimate. Another possibility is that the $\text{bias}\,\hat{S}_B(t)$ is large enough to make the modified Kaplan-Meir estimate worse than all the other estimates.

It is possible that the data may not sufficiently be modeled by an exponential distribution. There does appear to be a slight variation from the exponential curve. This might be enough to cause the modified Kaplan-Meir estimate to be more accurate. However, we note that this is a difficult option to assess, and it is unlikely that this variation from the exponential curve would be enough to affect the results.

A final possibility is that the approximations used in our methods skew the results so that an unexpected outcome arises. Of considerable concern is the use of the delta method in the construction of our estimators. The delta method is optimally used with large sample sizes. The sample size for the Acute Leukemia data is small. This inaccuracy due to sample size is difficult to avoid, because alternatives to the delta method are hard to find.

In our analysis of the angina pectoris data, it appears that the theoretical estimate is the most accurate survival function estimate followed by the modified Kaplan-Meir estimate, and lastly the standard Kaplan-Meir estimate. However, this assessment does not consider the bias of the modified Kaplan-Meir estimate. In our interpretation of the Acute Leukemia data, we discussed that the bias would likely cause the standard Kaplan-Meir estimate to be better than the modified Kaplan-Meir estimate. This might especially be the case here since the difference between the respective variances is very small. In our comparison of the estimated λ_i and MSE points of intersection, we found that the results favored the $\hat{\lambda}_i$ estimate. Thus, there is evidence that the bias for the modified Kaplan-Meir would be large enough so that $\text{MSE}\,\hat{S}_B(t) \leq \text{MSE}\,\hat{S}(t)$. Based on this, we make the conclusion that the standard Kaplan-Meir estimate would be more accurate than

the modified Kaplan-Meir estimate, while the theoretical estimate would remain the most accurate.

Like with the acute leukemia data set, there are suspicions that our conclusions might not be correct. The suspicions are shared for similar reasons except that for this data set, the sample size is large enough that we don't have to worry too much about the delta method approximations. One alternative conclusion is that the modified Kaplan-Meir bias is smaller than expected so that in fact, $\text{MSE}\,\hat{S}(t) \le \text{MSE}\,\hat{S}_B(t)$. This would lead to the conclusions that the modified Kaplan-Meir estimate is more accurate than the standard Kaplan-Meir estimate.

Our analysis encouraged us to consider a number of different scenarios and alternative methods where more certainty in the results might be attained. We first discuss some different scenarios followed by two alternative methods that deserve future consideration.

From our analysis of the Acute Leukemia data, we observed that the estimated $\lambda_i s$ were relatively close to the left intersection point. The reason for this lies in the nature of the data. Except for the first survival time, there was one observed failure at each time. This meant that our estimations of each λ_i, except for values with very small n_is, would be small enough to cause a relative amount of uncertainty. The values were closer to zero than was desirable so that choosing between the standard and modified Kaplan-Meir estimate was a relatively difficult choice.

Data sets with high death rates at each observed time would have less uncertainty and would tend to favor the modified Kaplan-Meir estimates. For instance, suppose we had a data set with the following survival times.

6, 6, 6, 6, 6, 6, 6+, 7, 7, 7, 7, 7, 7, 10, 10, 10, 10, 10, 13, 13

If we did analysis on this data set, then we would find that the estimated λ_i values would be much further away from the intersection points, because there are more observed failures at each survival time. In this case, the modified Kaplan-Meir estimate would outperform the standard Kaplan-Meir estimate with more certainty.

Data sets with smaller samples sizes will also tend to favor the modified Kaplan-Meir estimate. In an earlier analysis we discussed that a sample size less than 42 might be considered small. However, that is not a concrete marker. The smaller the sample size, the more accurate the modified Kaplan-Meir estimate will usually be. In our Acute Leukemia data analysis, it was seen that the latter survival times were more clearly in favor of the modified Kaplan-Meir estimates. This is because the n_is became increasingly smaller and acted as a local sample size. Based on our analysis, it may be that the sample size needs to be significantly smaller than 42, depending on other factors such as death rate per survival time. We guess that a sample size of 25 or less may be a better marker for favoring the modified Kaplan-Meir estimate.

A possible alternative method for estimating $S(t)$ for small sample sizes, may be to use the standard Kaplan-Meir estimate for some values of t_i and the modified Kaplan-Meir estimate for others. In our Acute Leukemia data analysis, we found that the first 3-4 comparisons of the estimated $\lambda_i s$ and the nearest intersection points, favored the

parametric estimate $\hat{\lambda}_i$, while the latter comparisons favored $\hat{\lambda}_{Bi}$. This suggests that the Kaplan-Meir estimate would likely be more accurate for the first 3-4 survival times, and the modified Kaplan-Meir estimate would probably be more accurate for the latter survival times. Future researcher might explore estimating each $S(t_i)$ with non-uniform estimates. For instance, in our Acute Leukemia analysis, it might be beneficial to use $\hat{S}(t)$ as an estimator for $S(t_1)$, $S(t_2)$, $S(t_3)$, and $S(t_4)$, while using $\hat{S}_B(t)$ to estimate the rest of the survival functions. The effects and complications of this sort of method deserve more scrutiny.

Another alternative method for estimating $S(t)$ for small sample sizes might be to use $\tilde{\lambda}_i$ as an estimate for λ_i. We discussed previously that $\tilde{\lambda}_i$ is a better estimator when the sample size is small, and used it to estimate the true value of λ_i. However, we did not explore how a modification of the Kaplan-Meir estimate based on $\tilde{\lambda}_i$ might perform. It is reason to believe it might be a good alternative for smaller sample sizes.

CONCLUSION

Evidence was shown, both theoretical and empirical, that suggests our modified Kaplan-Meir estimate is likely more accurate than the standard Kaplan-Meir estimate for smaller sample sizes. We conclude that our hypothesis is correct, but with some degree of uncertainty.

More analysis should be conducted to verify and confirm our hypothesis. The comparison of $\mathrm{MSE}\,\hat{S}(t)$ and $\mathrm{MSE}\,\hat{S}_B(t)$ should be explored in more detail. It could be useful to create a multidimensional graph of $\mathrm{MSE}\,\hat{S}_B(t)$ or some other mechanism for analyzing $\mathrm{MSE}\,\hat{S}_B(t)$ to determine the effect of the $\hat{S}_B(t)$ bias. Also, alternative estimates to the Delta method might be considered to address estimation error for smaller sample sizes.

Assuming that our conclusions are verified and confirmed, our results would help alleviate the particularly troublesome problem of inaccuracy due to small sample size. A lack of participants in a study may be mitigated by our modified Kaplan-Meir estimate. This would prove to be a boon to research when larger sample sizes are sometimes not available or even feasible.

REFERENCES

Agresti, A. and Coull, B. A. (1998) Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2), 119-126.

Bartholomew, D. J. (1957). A Problem in Life Testing. *Journal of the American Statistical Association*, 52, 350-355.

Casella, G. and Berger, R. (2002). *Statistical Inference* (2nd ed.). Brooks/Cole Cengage Learning.

Freireich E. J., Gehan, E. A., Frei, E., et al. (1963). The Effect of 6-Mercap-topurine on the Duration of Steroid-Induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potential Useful Therapy. *Blood*, 21(6), 699-716.

Gehan, E. A. (1969). Estimating Survival Function from the Life Table. *Journal of Chronic Diseases*, 21, 629-644.

Greenwood, M. (1926). The Natural Duration of Caner. *Reports on Public Health and Medical subjects, Her Majesty's Stationery Office*, London, 33, 1-26.

Kaplan, E. L., and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53, 457-481.

Kleinbaum, D. G. and Klein, M. (1996). *Survival Analysis: A Self-Learning Text* (3rd ed.). Springer.

Le, C. T. (1997). *Applied Survival Analysis.* Wiley-Interscience, Inc.

Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis* (2nd ed.). Wiley-Interscience, Inc.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14, 945-966.

Parker, R. L., Dry, T.J., Willius, F. A., and Gage, R. P. (1946). Life Expectancy in Angina Pectoris. *Journal of the American Medical Association*, 131, 95-100.

Xu, R. **(**2016)**.** *Lecture 2 Estimating the Survival Function*. UCSD, La Jolla, CA. Retrieved from http://www.math.ucsd.edu/~rxu/math284/slect2.pdf