Comparing the Seminonparametric Survival Function Estimator to the Kaplan-Meier Estimator and Equivalent Parametric Methods for Right-Censored Data

Ali Miller

11/30/11

Abstract

When it comes to the medical field, and particularly pharmaceutical research, survival analysis is frequently used with clinical trial data. Often the objective of a clinical trial is to determine whether or not a given drug or treatment is effective, or whether it is more effective than the treatment already on the market. To accomplish this, an experiment must be conducted to determine how long individuals given a specific treatment will survive, or how quickly they get better. This "time to event" data is known as survival data. In order for a clinical trial to be of any use, it must be used to make generalizations about the greater population of individuals who may use the treatment being tested. To project the results of a clinical trial onto a larger population, we must use survival analysis to create an estimate. Survival analysis is used to determine what proportion of a population experiences an event of interest after a given time point, or at what rate individuals within a population experience an event. Often the event of interest is death. Unfortunately, there is no one foolproof method of estimating the survival function. Researchers must therefore attempt to judge which of the methods is most appropriate for the data they are working with. Choosing the wrong method of estimation may lead to an incorrect conclusion. In the pharmaceutical field, this may mean putting a less effective drug on the market. My research will assess the strengths and weaknesses of several different estimation methods. I will particularly look at the Kaplan-Meier estimator, the most frequently used method of survival estimation, comparing it to the seminonparametric method and equivalent parametric methods of survival estimation. My work will in turn help researchers in these other fields make the most accurate decisions possible and hopefully increase the lifespan or quality of life of their patients.

Introduction of Field

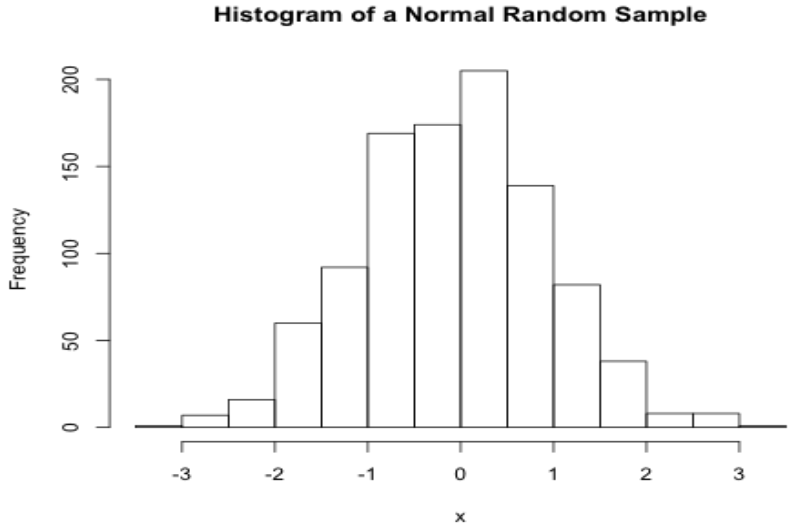
"Survival analysis is a class of statistical methods for studying the occurrence and timing of events" (Allison 1). For this reason, survival data can also be known as "time to event data." While the original purpose of survival analysis may have been to study time to death, it has since branched out to be applied to many other fields of interest. For example, the analysis of survival data arises in a number of fields, including medicine, biology, public health, epidemiology, engineering, economics, and demography (Klein et al, 1). Any time that one is interested in analyzing the time until a specified event of interest, survival analysis is appropriate for use. This may even include comparisons of average times to event between two different groups, as is often the case when comparing multiple treatments. Survival analysis forms its own branch of statistical analysis and must use methods different from other forms of analysis due to the nature of the data it speculates upon. One of the most prominent characteristics of survival data is the presence of censoring.

As is unfortunately often the case when collecting survival data, sometimes we do not know the exact time that a participant (human or otherwise) experiences the event of interest. Sometimes we can only identify an interval of time in which the time of interest occurred. This is known as censored data.

There are multiple categories of censoring, including right censoring, left censoring, and interval censoring. Right censoring of data occurs when a subject leaves the sample before the time of interest has occurred. For example, if a participant moves away and looses contact with researchers, he or she is removed from the study and we do not know at what time their event of interest occurred. All we know is that the event of interest did not occur by the time they left the study. Left censoring occurs when a subject experiences the event of interest between time zero and the first time that the patient was officially observed after the study began. For example, suppose that you enroll cancer patients into a study at time zero to examine the time until these individuals develop a tumor. At the start of the study no patient has a tumor. If three months later at the first official clinic visit you notice that an individual has already developed a tumor, all you know is that the tumor developed sometime between 0 and 3 months. Therefore the time until the event of interest is left censored at 3 months. Finally, interval censoring occurs when we only know that the event of interest has occurred within an interval of time. In many clinical trials, patients are examined at follow-up visits to determine whether or not the event of interest has yet occurred. These follow-up visits may be years apart. Therefore, when a patient comes in and the event of interest has occurred, we may only know that it has occurred within an interval between the previous visit and the current visit. This is interval censored data. Both left and right censoring are special forms of interval censoring. Left censoring occurs when we know that the event of interest occurred before an observed endpoint, and right censoring occurs when we know that the event of interest occurred after an observed endpoint.

The most widely used methods of approaching survival analysis are probability-based. This means that "the times at which events occur are assumed to be realizations of some random process" (Allison 14). Therefore, t, the time until the event of interest, is a random variable with a probability distribution. There are several different probability distributions that t may follow, and the distribution plays an important role in determining which method is appropriate for analyzing a given data set. Each distribution of a random variable has an associated set of parameters that serve as an appropriate means of describing the distribution and estimating the larger population. For example, if we know that our population of interest follows a normal distribution our data could look like something similar to Figure 1.

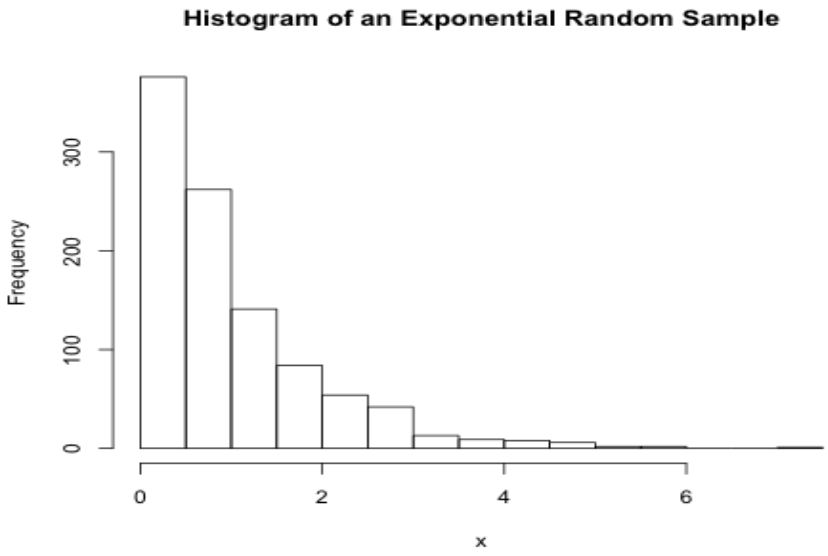
Figure 1: Histogram of a Normal Random Sample



When dealing with a normal distribution, it would be appropriate to describe the distribution in terms of the mean (μ) and the standard deviation (σ). If I took a sample of data from a normal distribution, I could calculate the average and standard deviation of the sample values, and use those statistics to estimate the true values of those parameters for the population.

However, we may instead have a population that follows an exponential distribution. In this case, our data could look something like Figure 2. In this case, the parameters μ and σ are no longer appropriate to give an accurate description of this data. Exponential data instead uses the parameter β to describe rate. Since each distribution is associated with a parameter or set of parameters that are appropriate to describe and estimate it, the distribution of a population plays a large role in determining how we go about analyzing the data from that population. However, when dealing with survival data, our population of interest may be all individuals on Earth with skin cancer. When dealing with such large and complex populations, it can be impossible to get an accurate idea of what the distribution of their survival data may be. This is a problem that statisticians have been trying to come up with a solution to for a long time.

Figure 2: Histogram of an Exponential Random Sample



One way of analyzing survival data is using the survival function. When time is a random continuous variable, the survival function is the complement of the cumulative distribution function (CDF) (Klein 22). This means that the survival function tells us the probability of an individual surviving beyond time t, whereas in the context of time to event data, a CDF represents the probability that an individual experiences an event before time t.

While there are many different types of survival curves, they all share the same properties. As you can see from Figure 3, the survival function is nonincreasing. This is true for all survival functions. The value of the survival function is assumed to be 1 at time zero, and the value of the function approaches zero as time approaches positive infinity. This is because the probability of survival past time t=0 is always 1, and the probability of surviving past a time of infinity is of course zero. In this illustration of survival curve, the continuous, curved line represents the survival function for a random

variable. The step function, as seen in Figure 4, represents the estimated survival function when no distributional assumptions are made on the data.

Figure 3: An Exponential Survival Curve

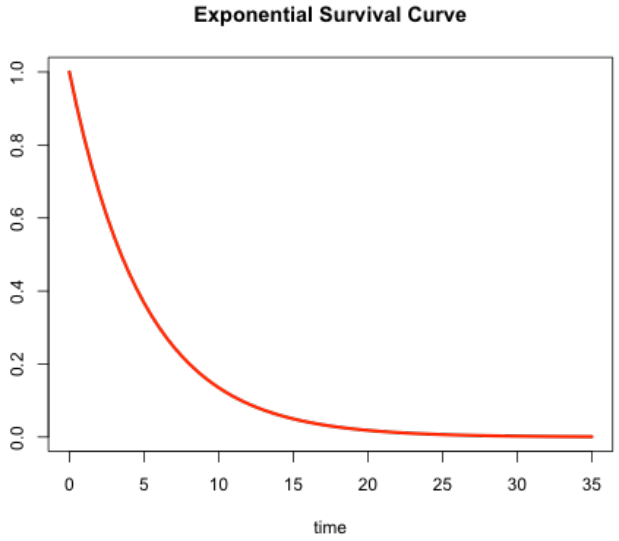
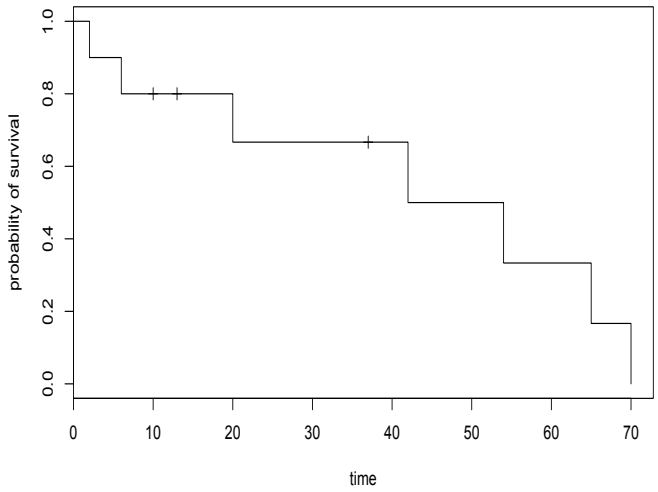


Figure 4: A Generic Step Function



Methods for analyzing and estimating survival data and survival functions fall into three major categories: parametric methods, nonparametric methods, and seminonparametric methods. Generally speaking, a good estimator has two characteristics: it is unbiased and it is accurate. Bias occurs when an estimator consistently either underestimates or overestimates the true population value. Accuracy simply means that the estimator has a small associated variance, which leads to a smaller margin of error in a confidence interval. All currently known methods of analysis and estimation have different benefits and flaws. There is no definite, foolproof method. Statisticians are constantly coming up with new ways to alter the accepted methods to reduce bias and error. At the current moment, choosing between methods of analysis and estimation can be somewhat of a tossup (Cantor). However, this is not necessarily because one method is not better than the others, but rather because not enough research has yet been done comparing these methods.

Parametric methods are often the easiest to implement. If we assume that our survival data follows a known distribution, then making estimates or inferences about the population is not difficult. We can use our sample as a means of obtaining estimates of parameters for the assumed distribution. For example, if we assumed that our data followed a normal distribution, we can use our sample mean to generate and make inferences about the population mean of the survival. However, making any assumptions about the underlying distribution of data can be risky (Cantor 15). Also, making assumptions about the distribution of data can lead to large amounts of bias in our estimations if we are not correct. Since we will often not know the exact distribution a population follows, it is often safer to use methods of estimation and inference that do not require any assumptions to be made about the distribution. This is why, when dealing with survival data, it is extremely popular for nonparametric methods to be used for analysis.

Nonparametric survival analysis methods are different in that they do not make any assumptions whatsoever about the underlying distribution of the data. Like parametric methods, nonparametric

methods can be used to estimate survival probability, and they can also be used to compare survival rates between groups. Since nonparametric methods do not rely on the use of specified distributional parameters, they can be used to analyze any type of population regardless of distribution. Unfortuantely there are certain drawbacks associated with nonparametric estimators. Nonparametric methods tend to be associated with higher rates of bias and higher variances, which means less accuracy. However, the associated benefit of not having to make assumptions about the distribution of the data may or may not outweigh these drawbacks.

Seminonparametric methods may be considered a good compromise between parametric and nonparametric techniques. Seminonparametric estimation contains both parametric and nonparametric pieces, which can be manipulated to reduce variance. For the seminonparametric (SNP) method we are considering, we only have to assume the distribution of a base density. Also, the SNP function is very flexible due to the presence of a tuning parameter. This flexibility will allow the survival function to be estimated for numerous different distributions. However, this is a relatively new method in the world of statistics, and not enough work has yet been completed to determine the strength of this method under different conditions of distribution and censoring. Hopefully, future research will be able to shed some light on the strengths and weaknesses of all of these methods comparatively under various conditions, giving all those who use survival analysis a greater sense of confidence in their results.

Build to your Question

Currently, the Kaplan-Meier estimator is the most popular method of estimating the survival curve in the field of biostatistics. This is not necessarily because it is the most accurate method, but instead because it is the most convenient since data in the biostatistics field tends to generally lend itself to this type of estimation. Comparable parametric tests and newly available seminonparametric methods may be just as effective if not more effective at estimating the survival curve. Since survival analysis is such a staple of the medical and pharmaceutical fields, it is important to ensure that researchers are using the most accurate methods possible to analyze their data.

The Kaplan-Meier estimator is a nonparametric method of estimating the survival curve (Kaplan and Meier 1958). Being a nonparametric method simply means that this estimator does not require any assumptions to be made about the distribution of the data. This is convenient and helpful for researchers, because more often than not we do not know the distribution that a given data set comes from. Also, sample data will rarely follow a specific distribution perfectly, so the Kaplan-Meier eliminates any guesswork that would otherwise have to be done. Another benefit of the Kaplan-Meier estimator is that it easily accounts for right-censored data, which occurs when the censoring time for an individual in the study occurs before the time of interest (Kaplan and Meier 1958). This is the most frequently occurring type of censoring, so this is a large benefit for the Kaplan-Meier. The Kaplan-Meier also has the benefit of simplicity. It is relatively simple to calculate, and most statistical software has programming written for it that allows researchers to simply plug in their data and get an estimate of the survival function. Also, while the Kaplan-Meier does not have an unreasonably large amount of variability, it may not be the most accurate estimation possible.

If a distribution of the data can be reasonably assumed, we may get a more accurate estimation using parametric methods. Such methods use an estimate of an appropriate parameter of the distribution, such as the mean, the standard deviation, or the rate, and simply base the estimate of the survival

function off of the distribution and the associated parameter or parameters (Cantor). This is an even simpler method than the Kaplan-Meier, and in some instances is also more accurate. However, making assumptions about the distribution of data can be very risky (Cantor 15). If we incorrectly identify the underlying distribution, any results we find will be completely invalid. For this reason, nonparametric methods are generally preferred in practice.

When it comes to comparing the abilities of the Kaplan-Meier estimator to those of a parametric equivalent, both sides have had equal support and opposition. Many make the logical assumption that since the nonparametric method of the Kaplan-Meier does not require as much information to be known about the data being analyzed, it will ultimately result in a less accurate estimate than the parametric counterpart. In 1983, Rupert Miller compared the Kaplan-Meier to parametric testing by looking at asymptotic efficiency. Simply put, asymptotic efficiency is calculated as a ratio of the variances of the estimators being compared. Since variance is an indication of accuracy, this method is appropriate to compare the accuracy of estimation methods. He concluded that in many cases, the Kaplan-Meier estimator sacrifices efficiency for ease of use. In response to these findings, Paul Meier revisited the bias and efficiency of the Kaplan-Meier method. In 2004, Meier asserted that the nonparametric approach is unbiased and results in little to no loss in efficiency. If this is true, we should then look at the accuracy of the Kaplan-Meier compared to other estimation methods.

More recently, a method of estimating the survival curve has been introduced which may combine the strengths of both of the previously mentioned estimators. This method is known as the seminonparametric method of estimation. The seminonparametric method, or SNP, lies between a completely parametric and fully nonparametric procedure (Gallant and Nychka 1987). Like the nonparametric method, the SNP does not require any assumptions about the underlying distribution. Another benefit of the SNP is that it is extremely flexible, and therefore can be used with virtually any data. Additionally, the SNP may or may not have less variability than the Kaplan-Meier. Since variance is our measurement of accuracy in Statistics, this may mean that the SNP is a more accurate method of estimation than the KM. However, the SNP is extremely complex, and not easy to work with. Since it was not created until well after the Kaplan-Meier was already well established, the SNP does not have any type of universally available computational code written for it to be used in statistical software. For this reason, the SNP is not a popular method of estimation.

The research that has been done so far concerning the SNP has shown that it may be preferable over the Kaplan-Meier estimator. In 2008, Doehler and Davidian looked at the SNP density as a means of estimating the survival curve when dealing with arbitrarily censored data. They used a combination of simulations and real medical data sets and ran a number of trials using the KM estimator and the SNP density function. They then calculated comparisons both graphically and by computing relative efficiency. Ultimately, they found that the SNP offers "efficiency gains over completely nonparametric methods" (Doehler et al. 2008, p.5437). Since the Kaplan-Meier would be considered a completely nonparametric method, this means there is reason to believe that the SNP may be more accurate than the KM in various cases.

In the field of survival curve estimation, convenience may be creating a decrease in accuracy of estimations. If accessible and easy to use code could be written for the SNP and the SNP was shown to be more accurate than the Kaplan-Meier estimator, the SNP could possibly replace the Kaplan-Meier as the most popular method of survival curve estimation. However, there is still much that has gone

untested to determine under which conditions the SNP is more accurate than the Kaplan-Meier, as well as what conditions may or may not make it more accurate than a parametric equivalent. As stated previously, the SNP is extremely flexible and can be used with a wide variety of data types, so it stands to reason that it may the most reliable method of estimation. Our question is:

How does the SNP compare in accuracy to the Kaplan-Meier estimator and relevant parametric estimators under varying levels of censoring?

In the process of answering this question, we will create code for the utilization of the SNP density function in R software. If it is shown that the SNP is truly superior to the KM and parametric testing, there may be more incentive to develop a macro in R which would allow others to easily apply the SNP estimation method. Once code for the SNP is readily available, individuals may be more likely to apply SNP methods.

References

Allison, Paul D. *Survival Analysis Using the SAS System a Practical Guide*. Cary, NC: SAS Publ., 2003. Print.

Cantor, Alan. *SAS Survival Analysis Techniques for Medical Research*. Cary, NC, USA: SAS Pub., 2003. Print.

Doehler, Kirsten, and Marie Davidian. "'Smooth' Inference for Survival Functions with Arbitrarily Censored Data." *Statistics in Medicine* 27.26 (2008): 5421-439. Print.

This is the research done by my mentor, and the principle study that we are building off of in our research. We will be relying on the work done in this research, in which the seminonparametric distribution function is manipulated to function as a smooth survival curve estimator. We will be building off of the findings of this research to further test the abilities and relative efficiency of the seminonparametric estimation method.

Gallant, A. Ronald, and Douglas W. Nychka (1987), ``Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

This research was the basis for the implementation of seminonparametric methods that my mentor tested previously, and which we are continuing to further test with this research. It was the findings of this research that was adapted into a seminonparametric means of estimating the survival curve. The method that was created based on this research is the method that I will be testing the accuracy and efficiency of in my research.

Kaplan, E. L., and Paul Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53.282 (1958): 457-81. Print.

The research that I am doing is comparing seminonparametric methods of analysis to nonparametric and parametric methods. More specifically, I am looking at parametric, nonparametric, and

seminonparametric methods of survival function estimators. My mentor and I have chosen to use the Kaplan-Meier estimator as our nonparametric method for comparison. We chose this estimator because it is considered to be the most popular nonparametric method, as it creates a simple solution to a problem known as right-censoring which researchers commonly run into. This article also delves somewhat into the very reasons my mentor and I have for choosing this topic of research. While nonparametric methods can at times be less accurate than parametric methods, there are many instances in which parametric methods, which create assumptions about populations, are simply not appropriate for a given data set. In these situations, it is convenient to have alternative methods which do not require one to make any assumptions about the underlying distribution of the data.

Klein, John P., and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer, 1997. Print.

Meier, Paul, Theodore Karrison, Rick Chappell, and Hui Xie. "The Price of Kaplan–Meier." *Journal of the American Statistical Association* 99.467 (2004): 890-96. Print.

Miller, Rupert G. "What Price Kaplan-Meier?" *Biometrics* 39.4 (1983): 1077-081. Print.

Methods/Approach

Using R software, we will create simulations which generate data from various known distributions. We will then analyze each simulated data set using all three of the survival estimation methods. We will graph all of the resulting survival curves, and compare them to the true survival curve. A large portion of our time will be spent understanding and formulating code to create seminonparametric estimates. This is the one area in which code is not currently available, so we must write it ourselves.

Before we delve into programming, we must first take the time to understand the concepts and theories being applied in our research, and the level of appropriateness of their applications. This has taken up most of our time so far.

Timeline

Spring 2012:
- Take 1 credit hour of 499
- Have a good amount of programming written and generated
- March: Attend and tentatively present at NC Symposium for Women in Mathematics and Statistics
- Complete Junior year progress report

Fall 2012:
- Take 1 credit hour of 499
- Complete any other remaining programming
- Begin formal write-up of results
- November: Attend and present at UNCG Regional Mathematics and Statistics Conference

Spring 2013:
- Take 1 credit hour of 499
- Complete final paper and Fellows Final Reflection paper
- Present at SURF

Budget/Budget Justification

Items	Fellows
• Poster	$38.52
Proposal poster for the Elon College Fellows Junior poster session.	
• Proposed Travel Expenses	$50
There are two conferences which we would like to attend (and perhaps present at), the North Carolina Symposium for Women in Mathematics and Statistics and the UNCG Regional Mathematics and Statistics Conference. These are both free and do not require posters, so the only expenses would be for gas and perhaps lunch.	
• Introductory Statistics with R, 2nd edition	$65
This book will help with the coding of the software which I will be using for my research, and contains information about the specific methods I will be using.	
• Data Analysis and Graphics Using R- An Example-Based Approach	$66
This book will also help me in using the R software to perform my research.	
• Introducing Monte Carlo Methods with R	$65
This book will aid in my understanding of the specifics of performing simulations using R.	
• Total	$284.52