

NONPARAMETRIC INFERENCE IN SURVIVAL ANALYSIS

By

HUMAIRA KHAIR

A Thesis

Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

Department of Statistics
The University of Manitoba
Winnipeg, Manitoba

© August, 2003

THE UNIVERSITY OF MANITOBA
FACULTY OF GRADUATE STUDIES

COPYRIGHT PERMISSION PAGE

NONPARAMETRIC INFERENCE IN SURVIVAL ANALYSIS

BY

Humaira Khair

**A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University
of Manitoba in partial fulfillment of the requirements of the degree**

of

Master of Science

HUMAIRA KHAIR © 2003

Permission has been granted to the Library of The University of Manitoba to lend or sell copies of this thesis/practicum, to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film, and to University Microfilm Inc. to publish an abstract of this thesis/practicum.

The author reserves other publication rights, and neither this thesis/practicum nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

ABSTRACT

In practice, the presence of censored and correlated data is common, especially in long-term medical follow-up studies. There are not enough tools to handle such data but the field of study is open for research and there are various possibilities to extend existing results. In this thesis, the available nonparametric estimators are described and illustrated using some censored data sets. To incorporate the correlation of censored data, modification of the existing Nelson-Aalen estimator is discussed. We have found that the modified Nelson-Aalen estimator performs better than any other estimators for censored correlated data in the sense that this estimator minimizes the sum of squares of forecast errors. Using the modified Nelson-Aalen estimator, an analogue of the product limit estimator is also introduced. A functional version of the delta method is discussed which is used to derive the variance of the product limit estimator. Some hypothesis testing examples are presented with censored and complete samples. Examples based on the Kruskal-Wallis test and Puri's normal score test are of special interest since these tests are nonparametric analogue of the one-way analysis of variance test (F -test). For normal samples, Puri's test is as efficient as the parametric F -test and more efficient than the Kruskal-Wallis test.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my supervisors Dr. M. Samanta and Dr. A. Thavaneswaran for their wise advice, valuable guidance, enthusiasm, patience and continuous encouragement. Their kindness and moral support have been a continuous source of inspiration. In fact, my gratitude to them cannot be expressed in a few words. I also acknowledge the partial financial support from NSERC.

I would like to thank Dr. Ken Mount and Dr. P. Shivakumar to serve on my committee and for their valuable suggestions to improve the presentation of this thesis. Additional thanks to Dr. Dennis Murphy for his help with “R” and his valuable comments on the thesis. Special thanks to Mr. K. Thaga for helping me when it was needed.

Finally, I wish to express my deep indebtedness to my parents, brother, my friends, and above all, my beloved husband, for their kind support, generous love and moral encouragement.

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	v
LIST OF FIGURES	vi
CHAPTER ONE: INTRODUCTION	1
CHAPTER TWO: SURVIVAL FUNCTIONS	5
2.1 Introduction	5
2.1.1 Survival function	6
2.1.2 Probability density function	7
2.1.3 Hazard function	9
2.2 Relationship among functions	11
CHAPTER THREE: NONPARAMETRIC ESTIMATORS	12
3.1 Introduction	12
3.2 Kaplan-Meier estimator	14
3.3 Nelson-Aalen estimator	16
3.4 Comparison of K-M estimator and Nelson-Aalen estimator	17
3.5 Modified Nelson-Aalen estimator for correlated data	19
3.5.1 Example	22
3.5.2 Modified product limit estimator	24
3.6 Delta Method	25

3.6.1 Examples	27
Example 3.6.1.1	27
Example 3.6.1.2	29
Example 3.6.1.3	30
3.6.2 Estimation of functionals of $S(t)$	33
3.7.1 Empirical distribution function	35
3.7.2 Nonparametric Bayes method	37
3.7.3 Comparison	42
CHAPTER FOUR: TESTING HYPOTHESES	43
4.1 Introduction	43
4.2 One-Sample Tests	43
4.2.1 Example	46
4.3 Comparison of two groups of survival data	49
4.3.1 The log-rank test	50
4.3.2 Example: Prognosis for women with breast cancer	52
4.3.3 The Wilcoxon test	56
4.3.4 Example: Prognosis for women with breast cancer	57
4.3.5 Comparison of the log-rank and Wilcoxon tests	60
4.4 Tests for more than two samples	64
CHAPTER FIVE: CONCLUSIONS	77
APPENDIX	80
REFERENCES	89

LIST OF TABLES

Table 3.1. Sum of squares of forecast errors for various α values.	24
Table 4.1. Computation of One-sample, Log-rank test.	48
Table 4.2. Number of deaths at t_i in each of two groups of individuals.	50
Table 4.3. Construction of two-sample, Log-rank test (following Klein).	54
Table 4.4. Calculation of the Log-rank and Wilcoxon statistic (following Collett).	58
Table 4.5. Comparison of two-sample tests, breast cancer data.	59
Table 4.6. Ranks and their corresponding expected order statistics.	73

LIST OF FIGURES

Figure 2.1 Two examples of survival curves.	6
Figure 2.2 Two examples of density curves.	8
Figure 2.3 Five different hazard functions.	9
Figure 3.1 Comparison of estimated survival functions.	18
Figure 3.2 Comparison of estimated variance of survival functions.	18
Figure 3.3 Empirical distribution function.	37
Figure 3.4(a): G_n for 100 random orderings (left) and the average-over-orderings $\bar{G}_n$ (right) with $\alpha = 1$. Dashed line represents the prior c.d.f. G_0 .	41
Figure 3.4(b): G_n for 100 random orderings (left) and the average-over-orderings $\bar{G}_n$ (right) with $\alpha = 5$. Dashed line represents the prior c.d.f. G_0 .	42
Figure 4.1(a): Survivor functions when $\delta > 1$.	61
Figure 4.1(b): Survivor functions when $\delta < 1$.	61

CHAPTER ONE

INTRODUCTION

The goal of survival analysis is to make inferences about a population with censored observations. In some sense, it describes incomplete information for the width of interval between an origin and an end point in time. Often, the end point corresponds to death and the length from the origin to the end is measured in number of days, months or years. The end point, generally called a failure, may also correspond to the occurrence of any type of event, such as response to a treatment, development of a disease, cancer recurrence, death of the patient or machine failure.

Imagine that a researcher in a hospital is studying the effectiveness of a new treatment for a generally terminal disease. The major variable of interest is the number of days that patients survive. At the end of the study there will often be patients who survive the entire study period, particularly those patients who entered the hospital (and the clinical study) late in the study; there may also be other patients with whom contact is lost. One would not want to exclude all of those patients from the study by declaring them to be missing data (since most of them are “survivors” and, therefore, they reflect on the success of the new treatment method). Those observations which contain only partial information are called censored observations (e.g., patient A survived at least 4 months before he moved away and contact was lost). In general,

censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time.

The analysis of survival experiments is complicated by issues of censoring, where an individual's life length is known to occur only in a certain period of time, and by truncation, where individuals enter the study only if they survive a sufficient length of time or individuals are included in the study only if the event has occurred by a given date. Possible censoring schemes are right censoring, where all that is known is that the individual is still alive at the end of the study; left censoring, when all that is known is that the individual has experienced the event of interest prior to the start of the study, or interval censoring, where the only information available is that the event occurs within some (time) interval. Truncation schemes are either left truncation or right truncation. Left truncation occurs, for example, when subjects enter a study at a particular age (not necessarily the origin for the event of interest), and are followed from this delayed entry time until the event occurs or until the subject is censored. Right truncation occurs when only individuals who have experienced the event of interest are observable.

Statistical methods for dealing with censored data might be divided into the following categories: parametric vs. nonparametric, and univariate vs. multivariate. Parametric methods assume that the censored survival times are drawn from a population with a known distribution function (e.g., Weibull, exponential, etc.). Nonparametric methods make no distributional assumption about the population and frequently rely on rank methods and density estimation techniques. Univariate and multivariate methods are

being used to make inference about one or more characteristics of the population from which the censored sample(samples) is(are) drawn and compare such characteristics for two or more populations using censored data. This thesis primarily concentrates on univariate nonparametric methods.

The problem of nonparametric estimation of the survival function with censored data has an elegant and efficient solution in the one-dimensional case: the Kaplan-Meier (K-M) estimator. In a recent paper, Cai (1998) developed the asymptotic properties of the K-M estimator for censored dependent data. However, in higher dimensions, with multiple, possibly correlated survival times, the task is much more difficult.

One of the basic assumptions in statistical inference is that we have a simple random sample from a population, which implies the observations are independently distributed. For correlated data (e.g., time series data), we can no longer assume independence; hence, the available estimation procedures for independent data cannot be applied.

In practice, the presence of censored correlated data is common, especially in long-term medical follow-up studies. In such settings, examples of censoring include loss to follow-up, dropout and termination of the study. There are not enough tools to handle such data and the field of study is open for research with various possibilities to extend existing results. One possibility is to develop a nonparametric estimator like the Nelson-Aalen (N-A) estimator to estimate the cumulative hazard function for censored correlated data. A possible choice is to use the exponentially weighted moving average (EWMA) smoothing technique to modify the existing estimator to

incorporate correlation. Alavi and Thavaneswaran (2002) applied such a modified estimator to a number of data sets, including both independent and correlated data with censored observations. In survival analysis, hypothesis testing procedures are also a topic of considerable interest to compare the effectiveness of one or more treatments.

Chapter one provides an introduction to the thesis. Chapter two describes some common notation dealing with different probability functions. Chapter three discusses nonparametric inference procedures with censored and censored correlated observations, including a discussion of the functional version of the delta method to derive the variance of the product limit estimator. Some hypothesis testing examples for one-sample and k ($k \geq 2$) samples situations are presented in Chapter four. Finally, the conclusion of the thesis is given in Chapter five. The data used for different examples are included in the Appendix. The computer package “R” has been used to develop the required algorithms, which are included in the Appendix.

CHAPTER TWO

SURVIVAL FUNCTIONS

2.1 Introduction

The terms “lifetime”, “survival time” or “failure time” are frequently used for data that measure the time to a certain event such as the actual length of life of an individual, the occurrence of some events, response, the development of a given disease, failure, death, relapse, parole, or divorce. These times are subject to random variations and, like any random variables, form a distribution. The distribution of survival times is most commonly described or characterized by three functions:

- 1) the survival function (also called survivorship function or cumulative survival rate),
- 2) the probability density function (or density function), and
- 3) the hazard function (also known as hazard rate, instantaneous failure rate, force of mortality, conditional mortality rate or age-specific failure rate).

These three functions are mathematically related. In the following sections, we will discuss these functions along with their interrelationships.

2.1.1 Survival function

Let T be a non-negative continuous random variable representing the lifetime of an individual drawn at random from some population. Then the survival function, denoted by $S(t)$, is defined as the probability that an individual survives longer than t :

$$S(t) = P(\text{an individual survives longer than } t)$$

$$= P(T \geq t)$$

From the definition of the cumulative distribution function $F(t)$ of T ,

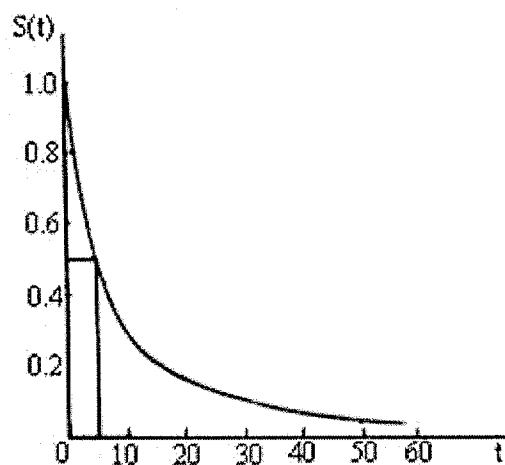
$$S(t) = 1 - P(\text{an individual fails before time } t)$$

$$= 1 - F(t)$$

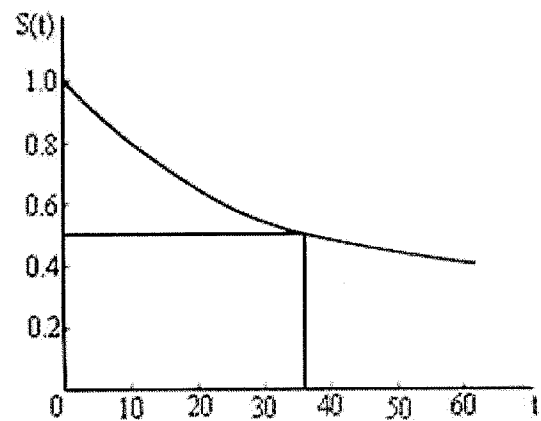
$S(t)$ is a non increasing function of time t with the following properties:

$$S(t) = 1 \quad \text{for } t = 0 \quad \text{and}$$

$$S(t) = 0 \quad \text{for } t = \infty$$



(a)



(b)

Figure 2.1 Two examples of survival curves.

A survival model in terms of the survival distribution function $S(t)$ can be defined and described. An important part of survival data analysis is to fit models. Berkson (1942) recommended a graphical presentation of $S(t)$, called the survival curve, for such a purpose.

A steep survival curve, such as the one in Figure 2.1(a) represents low survival rate or short survival time. A gradual or flat survival curve such as Figure 2.1(b) represents high survival rate or longer survival. One purpose of the survival curve is to find the 50th percentile (median) and other percentiles (e.g., 25th and 75th) of survival time and to compare survival distributions of two or more groups.

In survival distributions, the median is often a better estimator of the centre of a distribution than the mean, because a small number of individuals with exceptionally long or short lifetimes will cause the mean survival time to be disproportionately large or small. The median survival times in Figures 2.1(a) and 2.1(b) are approximately 5 and 36 units of time, respectively.

2.1.2 Probability density function

A probability density function of the survival time T is defined as the limit of the probability of failure in a small interval per unit time. It can be expressed as

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P\{\text{an individual dying in the time interval } (t, t + \Delta t)\}}{\Delta t}$$

The density function satisfies the following properties:

(1) $f(t)$ is a nonnegative function:

$$f(t) \geq 0 \text{ for all } t \geq 0$$

$$= 0 \text{ for } t < 0$$

(2) The area between the density curve and the t axis is equal to 1.

The density function $f(t)$ is also known as the unconditional failure rate at time t .

From this function, the proportion of individuals that fail in any time interval and the peaks of high frequency of failure can be found.

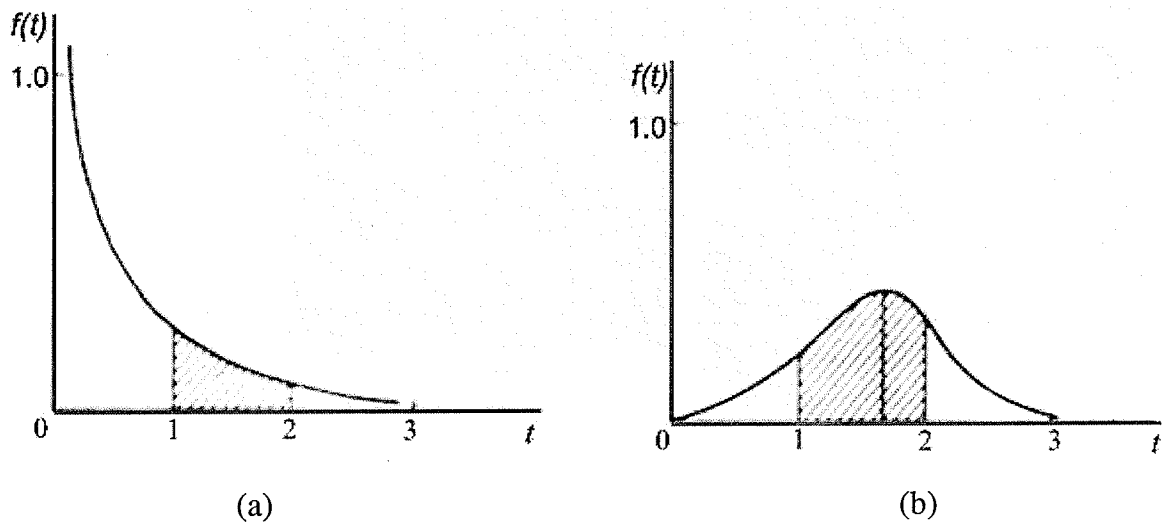


Figure 2.2 Two examples of density curves.

The density curve in Figure 2.2(a) gives a pattern of high failure rate at the beginning of the study and decreasing failure rate as time increases. In Figure 2.2(b), the peak of high failure frequency occurs at approximately 1.7 units of time. The proportion of individuals that fail between 1 and 2 units of time is equal to the shaded area between the density curve and the t axis.

2.1.3 Hazard function

The hazard function $h(t)$ of survival time T is defined as the probability of failure during a very small interval assuming that the individual has survived to the beginning of the interval. It gives the conditional failure rate and can be expressed as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P\{\text{an individual of age } t \text{ fails in the time interval } (t, t + \Delta t) | T \geq t\}}{\Delta t}$$

This function is particularly useful with lifetime distributions, since it describes the way in which the instantaneous probability of death for an individual changes with time. The cumulative hazard function is defined as

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

and can take any value between zero and infinity. The hazard function may increase, decrease, remain constant or indicate a more complicated process depending on different situations.

The following graph presents five different hazard functions, shown in Figure 2.3.

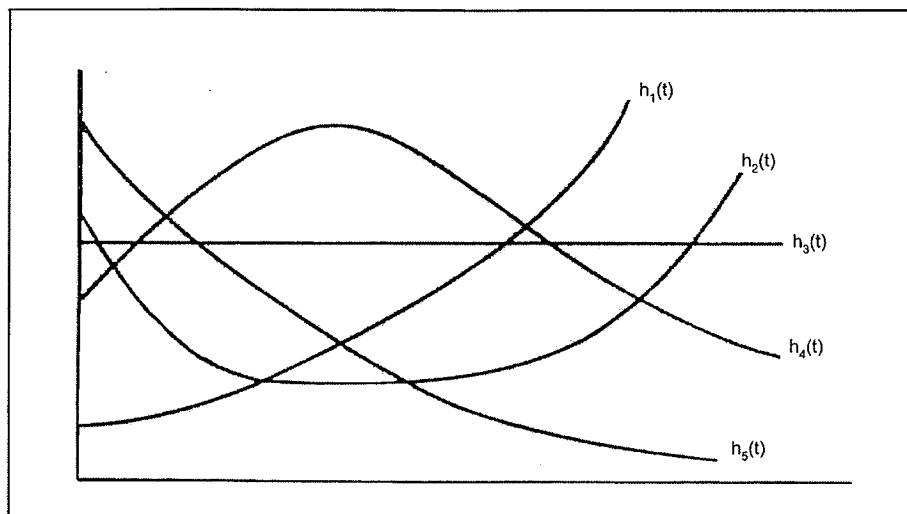


Figure 2.3 Five different hazard functions.

- (1) Models with increasing hazard functions are common in practice. For example, patients with acute cancer who do not respond to treatment have an increasing hazard rate $h_1(t)$.
- (2) $h_2(t)$ is the so-called “bathtub-shaped” or U-shaped hazard function. It describes a process whereby, after an initial period in which deaths occur primarily from defects or infant diseases, the death rate drops and is relatively constant until adulthood, after which it increases with age. The general survival experience of humans is compatible with this type of hazard function.
- (3) Models with a constant hazard function have a particularly simple structure. An example of a constant hazard function $h_3(t)$ is the risk of healthy individuals between 18 and 40 years of age whose main risks of death are accidents.
- (4) Patients with tuberculosis have risks that increase initially, and then decrease after treatment. Such an increasing, then decreasing, hazard function is described by $h_4(t)$.
- (5) Models with decreasing hazard functions are less common but are sometimes used. $h_5(t)$ is a decreasing hazard function that, for example, indicates the risk of soldiers wounded by bullets who undergo surgery. The main danger is the operation itself and this danger decreases if the surgery is successful.

2.2 Relationship among functions

The three functions $f(t)$, $S(t)$ and $h(t)$ are mathematically related. If any model is specified initially in terms of $h(t)$, then $S(t)$ and $f(t)$ can be obtained. The hazard function $h(t)$ is defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta(t)} = \frac{f(t)}{S(t)},$$

while the probability density function is the derivative of the cumulative distribution function; that is,

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

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

Integrating from 0 to t and using $S(0) = 1$, we have,

$$-\int_0^t h(x) dx = \ln S(t)$$

$$\Leftrightarrow H(t) = -\ln S(t)$$

$$\Leftrightarrow S(t) = \exp[-H(t)] = \exp \int_0^t h(x) dx$$

$$\Leftrightarrow f(t) = h(t) \exp[-H(t)].$$

CHAPTER THREE

NONPARAMETRIC ESTIMATORS

3.1 Introduction

In this chapter, different nonparametric methods are discussed for drawing inferences about the distribution of time to failure or some other event. In each case relevant examples are included. When there are no censored observations, the survivor function can be estimated by the empirical survivor function. With censored data, the well known nonparametric estimators are the product limit estimator (Kaplan-Meier estimator) for the survivor function and the Nelson-Aalen estimator for the cumulative hazard function. In section 3.4, these estimators are derived and compared considering right censored data. But for censored correlated data (e.g., time series data), we cannot use the commonly used nonparametric estimators. In this situation, the modified Nelson-Aalen estimator is more appropriate. Using the modified Nelson-Aalen estimator, an analogue of the product limit estimator is also introduced. In order to obtain the variance of different nonparametric estimators, we discuss the functional version of the delta method. In the last section of this chapter, a recursive algorithm is included to approximate nonparametric Bayes estimates using interval censored data.

We have also used the theory of counting processes to develop inference procedures with censored data. Counting processes are concerned with following a group of individuals who move among a small number of states over time. In survival analysis, the states are alive or dead. Analyses based on counting processes incorporate incomplete observations, of which right censoring is an important special case. Auxiliary information, such as fixed-time and/or time-varying covariates, is readily absorbed into the methodology.

In order to give an example of right censoring, suppose in a clinical trial, drug 6-MP was applied to a group of 12 acute leukemia patients. Let, $T_1, T_2, \dots, T_{12}$ (6, 6*, 7, 9*, 10, 10*, 16, 17*, 20*, 22*, 23, 32*- Starred quantities denote censored observations) represent the lengths of remission times (in weeks). These remission times are assumed to be independent and identically distributed (since those occurred to different individuals) with distribution function $F(t)$ and survivor function $S(t)$. Because of time and cost considerations, the investigator terminated the study after one year and reported the results before all individuals realized their events. For the i th individual under study, C_i represents a right censoring time. Here, the observation 6* refers to an individual who entered the trial 6 weeks before the end and whose remission was still in effect at the end of the trial. The exact remission time T_i for the i th individual will be known if and only if $T_i \leq C_i$. If $T_i > C_i$, the i th individual is a survivor and his remission time is censored at time C_i . To conveniently represent the data for this experiment, we introduce a pair of random variables (Z_i, δ_i) , where δ_i

indicates whether the remission time T_i corresponds to an event ($\delta_i = 1$) or is censored ($\delta_i = 0$).

Also, $Z_i = T_i$, if the remission time is observed

$= C_i$, if it is censored.

i.e. $Z_i = \min(T_i, C_i)$.

<u>Individual</u>	<u>T_i</u>	<u>C_i</u>	<u>δ_i</u>	<u>Z_i</u>
1	6	-	1	6
2	-	6	0	6
3	7	-	1	7
4	-	9	0	9
5	10	-	1	10
6	-	10	0	10
7	16	-	1	16
8	-	17	0	17
9	-	20	0	20
10	22	-	1	22
11	23	-	1	23
12	-	32	0	32

3.2 Kaplan-Meier estimator

The observations are given as (Z_i, δ_i) for $i = 1, 2, \dots, n$. Given a right-censored sample, the process

$$N_i(t) = I(Z_i \leq t, \delta_i = 1), \quad i = 1, 2, \dots, n$$

is a counting process. $N_i(t) = 0$ until the i th individual dies and then it jumps to one.

We define two stochastic processes on $[0, \infty)$. The first process deals with the number of uncensored observations less than or equal to t , given by

$$N(t) = \sum_{i=1}^n N_i(t),$$

The second process is defined by the number of censored or uncensored observations greater than or equal to t , defined as

$$Y(t) = \sum_{i=1}^n I(Z_i \geq t).$$

Heuristically, we define the term $dN(t)$ as the change in the process $N(t)$ over a short time interval $[t, t + dt)$,

$$dN(t) \equiv N[(t + dt)^-] - N(t^-),$$

where t^- is a time just prior to t .

The term $dN(t)$ can also be written in the following form.

$$dN(t) = Y(t)h(t)dt + dM(t)$$

This model is analogous to the regression model $Y_i = \beta X_i + \epsilon_i$,

where $N(t)$ is the number of deaths up to time t , $Y(t)$ is the number at risk just before time t , $h(t)$ is the hazard rate and $M(t)$ is the zero mean, square integrable martingale which corresponds to the error term in the regression model.

The K-M estimator of F is denoted by $\tilde{F}_n$ and is given by

$$\hat{S}(t) = 1 - \tilde{F}_n(t) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{Y(s)}\right) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right),$$

where $dN(s) = N(s) - N(s^-)$, n_j is the number of individuals who are alive and uncensored just prior to t_j and d_j is the number of deaths at time t_j .

For theoretical studies we define $\hat{S}(t)$ as $\prod_{s \leq t} \left(1 - \frac{dN(s)}{Y(s)}\right)$ and for practical purposes it

is defined as $\prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$.

For a continuous distribution function F on $[0, \infty)$, the cumulative hazard function is defined as

$$H(t) = \int_0^t \frac{dF(s)}{1 - F(s^-)} = -\ln[1 - F(t)] = -\ln[S(t)].$$

3.3 Nelson-Aalen estimator

Following Aalen (1978), we can write the counting process corresponding to the i th individual as,

$$N_i(t) = \int_0^t Y_i(s)h(s)ds + M_i(t),$$

or, equivalently

$$dN_i(t) = Y_i(t)h(t)dt + dM_i(t),$$

where $Y_i(t)$ is the number at risk (i.e. a process giving the number of individuals with a study time $T_i \geq t$, or, $Y_i(t) = I(Z_i > t)$), $h(t)$ is the hazard rate, $M_i(t)$ is a zero mean, square integrable martingale, which means for $i = 1, 2, \dots, n$,

$$E[M_i(t)] = 0 \text{ and } \sup_t E[M_i^2(t)] < \infty.$$

The empirical cumulative hazard function, or the estimator of $H(t)$ known as the Nelson-Aalen estimator, is given as

$$\tilde{H}_n(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{j:t_j \leq t} \frac{d_j}{n_j}, \quad (3.1)$$

where n_j is the number of individuals that are alive and uncensored just prior to t_j

and d_j is the number of deaths at time t_j .

Also, $\tilde{S}_n(t) = \exp[-\tilde{H}_n(t)]$.

$$\begin{aligned} \hat{Var}[\tilde{S}_n(t)] &= \langle \tilde{S}_n(t) \rangle_t \\ &= \sum_{j:t_j \leq t} \frac{d_j \exp[-2\tilde{H}_n(t_j)]}{n_j^2} \end{aligned}$$

Alavi and Thavaneswaran (2002) modified the Nelson-Aalen estimator for the cumulative hazard function of independent failure times for a homogeneous population, to incorporate dependence among the observations.

3.4 Comparison of K-M estimator and Nelson-Aalen estimator

This data set has been taken from Lawless (1982). The data consists of the remission times (in weeks) for a group of 21 acute leukemia patients. The group was given the drug 6-mercaptopurine (6-MP). Censored times are starred, which means that the disease was still in a state of remission at the end of the experiment.

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

With this data set, we first estimate the survival functions using the K-M and Nelson-Aalen estimators and plot them on the same axis. Figure 3.1 shows that the two

estimated survival functions are approximately the same. Next, we compare the estimated variance for those two survival functions. Figure 3.2 shows that the estimated variance based on K-M estimator is lower than that of the Nelson-Aalen estimator.

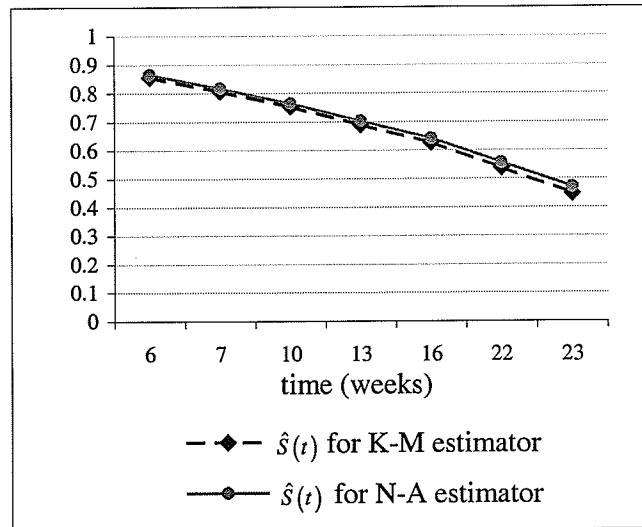


Figure 3.1 Comparison of estimated survival functions.

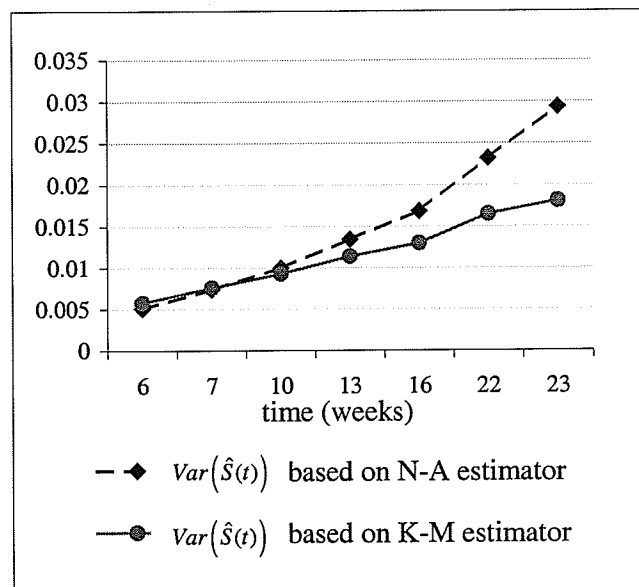


Figure 3.2 Comparison of estimated variance of survival functions.

3.5 Modified Nelson-Aalen estimator for correlated data

Let $T_1, T_2, \dots, T_n$ be a sequence of true survival times for n individuals in a life table.

These random variables are not mutually independent and have a common but unknown continuous marginal distribution function, defined as

$$F(x) = P(T_i \leq x) \text{ such that } F(0) = 0.$$

Let us assume the right-censoring defined by random variables C_i for the random variables T_i in such a way that one observes (Z_i, δ_i) where

$$Z_i = T_i \wedge C_i \text{ and } \delta_i = I(T_i \leq C_i), \quad i = 1, 2, \dots, n,$$

where $\wedge$ denotes the minimum and $I(A)$ is an indicator random variable of an event A . We also assume that the C_i 's are independent of the T_i 's and are independent, identically distributed random variables with distribution function

$$G(c) = P(C_i \leq c) \text{ such that } G(0) = 0.$$

Now we want to draw nonparametric inferences about F taking into account the censored observations along with their dependence.

The Nelson-Aalen estimator, as given in (3.1), is an estimator of an absolutely continuous function $H(t)$. The trajectories of $\tilde{H}(t)$ are right-continuous. An absolutely continuous estimator for $H(t)$ has been studied by Thavaneswaran (1988). Here, we introduce a weighted version of the Nelson-Aalen estimator for correlated data and demonstrate its superiority over the usual Nelson-Aalen estimator. The modified estimator has a form very similar to the single exponentially weighted

moving average with slight differences. First of all, the weights are not exponentially diminishing, rather oscillating between positive and negative values with most of the weight given to the recent value. For a set of correlated observations $z_1, z_2, \dots, z_n$, the exponentially smoothed statistic S_j based on j observations is given by

$$S_j = \alpha [z_j + (1-\alpha)z_{j-1} + (1-\alpha)^2 z_{j-2} + \dots],$$

or, equivalently

$$S_j = \alpha z_j + (1-\alpha)S_{j-1}$$

for $j=1, 2, \dots, n$, where $S_0 = \bar{z}$ is the sample mean.

Here, $z_j = \frac{d_j}{n_j}$, where d_j and n_j are defined in (3.1), and α is a weight which could

be greater than 1. Note that

$$\alpha = 0 \Leftrightarrow S_i = S_0 = \bar{z},$$

i.e., if the observations are independent we expect to get $\alpha = 0$.

For computing purposes an alternative form of the estimator of the cumulative hazard function at a given time t (not necessarily an integer) is given by

$$\hat{H}_n(t) = \alpha z_t + (\alpha - 1)\hat{H}_n(t-1),$$

where z_t is a simplified form of writing z_{t_j} , defined as $z_{t_j} = \frac{d_j}{n_j}$, where n_j is the number of individuals at risk of experiencing the event of interest at time t_j and d_j is the number of events that occur at time t_j . z_t can also be thought of as an estimate of

the conditional probability that an individual experiences the event at time t_j given that he has neither experienced the event nor been censored just prior to time t_j .

In the counting process notation the estimator has a form

$$\hat{H}(t) = \int_0^t w(s) \frac{dN(s)}{Y(s)},$$

where $w(s)$ is a continuous function.

The main drawback of using the counting process technique is that it only allows for a single event at a specified time, which is usually not the case in practice. Using the counting process approach, when $d_j = 1$ (i.e., no multiple events at t_j), the mean and conditional variance of the proposed estimator are given in the following theorem.

Theorem 3.1.

Let $\hat{H}(t) = \int_0^t w(s) \frac{dN(s)}{Y(s)}$, where $N(s) = \sum_{i=1}^n N_i(s)$, and $Y(s) = \sum_{i=1}^n Y_i(s)$, and $w(s)$ is a

smooth continuous function of s . Then,

$$(i) \ E[\hat{H}(t)] = \int_0^t w(s)h(s)ds,$$

$$(ii) \ \langle \hat{H} \rangle(t) = \int_0^t \frac{w^2(s)h(s)ds}{Y(s)}$$

where $\langle \rangle$ is the notation used for the predictable variation process, the conditional variance, of a martingale. It can easily be shown that part (i) of the theorem follows

from the properties of Itô's integrals and $E \int_0^t \frac{w(s)dN(s)}{Y(s)} = \int_0^t w(s)h(s)ds$.

Part (ii) follows from the fact that

$$\langle M_i \rangle(t) = \int_0^t h(s) Y_i(s) ds,$$

and from the properties of stochastic integrals,

$$\langle \hat{H} \rangle(t) = \int_0^t \frac{w^2(s)}{Y^2(s)} d\langle M \rangle_s = \int_0^t \frac{w^2(s)}{Y(s)} h(s) ds.$$

3.5.1 Example

The following data set is taken from Zeger and Brookmeyer (1986). This is a time series data and also a good example of censored correlated data. The data were collected on the chemical composition of atmospheric deposition. The Environmental Measurement Laboratory collected monthly data between 1977 and early 1980 at a number of sites in the United States on the composition of dry deposition of NH_4 . There are lower detection limits on the assays, which depend on the total quantity of deposition collected each month; smaller volumes collected give higher detection limits. Because of the detection limits, we have some censored observations. These data are also correlated in time.

The data are listed in the Appendix (Table 1). Here, the time t indicates to the monthly concentration of ammonia at the Lawrence Livermore, California, site. Of the 43 observations in the data set, 3 were missing and 6 were left censored.

The modified Nelson-Aalen estimator is used to calculate the sum of squares of the forecast errors. The aim is to find the value of α which minimizes the sum of squares of forecast errors. When the usual Nelson-Aalen estimator is used, the sum of squares of the forecast errors is 32.66, while the sum of squares of forecast errors for modified Nelson-Aalen estimator is 0.2089.

Table 3.1 shows various α values, along with their respective sum of squares of the forecast errors. From this table, it becomes clear that the α value which minimizes the sum of squares of forecast errors is 1.75. Here, the forecast errors are defined by

$$\text{forecast error} = z_t - \hat{H}(t-1),$$

and the sum of squares of forecast(prediction) errors (SSE) is given by

$$\text{SSE} = \sum [z_t - \hat{H}(t-1)]^2.$$

Another example from the same paper concerns an autoregressive process of order one (AR(1)). The data are listed in the Appendix (Table 2). Zeger and Brookmeyer (1986) generated a single realization of length 50 from a Gaussian autoregressive process of order $p=1$ with mean $\mu=0$, first-lag autocorrelation $\theta=0.5$ and white noise variance $\sigma^2=1$. The data were left censored at -1.2 and eight observations were censored.

Here, also, the α value which minimizes the sum of squares of forecast errors is 1.75. When the usual Nelson-Aalen estimator is used, the sum of squares of the forecast errors is 40.31, while the sum of squares of forecast errors for modified Nelson-Aalen estimator is 0.2233.

Table 3.1. Sum of squares of forecast errors for various α values.

α	SSE(NH_4)	SSE(AR(1))
0.3	0.6156	0.6157
0.5	0.4678	0.4689
0.7	0.3804	0.3825
0.9	0.3218	0.3250
1.0	0.2991	0.3031
1.1	0.2796	0.2846
1.2	0.2627	0.2690
1.25	0.2552	0.2623
1.3	0.2482	0.2561
1.5	0.2249	0.2383
1.7	0.2103	0.2368
1.75	0.2089 (min)	0.2233 (min)
1.8	0.2093	0.2501
1.85	0.2135	0.2640

3.5.2 Modified product limit estimator

In the literature, an analogue of the product limit estimator for censored correlated data has not been defined. For correlated data, we have modified the Nelson-Aalen estimator and determined the optimum α value by minimizing the sum of squares of forecast errors. This is a data driven method. Using the relationship $\hat{S}_n(t) = \exp[-\hat{H}_n(t)]$, we produce an estimate of the survivor function, which is defined as the modified product limit estimator. We can also denote this new estimator as

$$\hat{S}_M(t) = \exp[-\hat{H}_M(t)]. \quad (3.2)$$

3.6 Delta Method

The delta method has been used to obtain the variance of the MLE of differentiable functions of the parameters with independent observations. A functional version of the delta method is used to obtain the variance of different nonparametric estimators. A new estimate of the variance of the product limit estimator is proposed and it is shown that it has smaller variance than the one used in the literature.

Suppose a random variable Y has mean μ and variance σ^2 (denoted by $Y \sim (\mu, \sigma^2)$) and suppose we want to find the distribution of some function $g(Y)$. Expanding $g(Y)$ about μ in a Taylor series,

$$g(Y) = g(\mu) + (Y - \mu)g'(\mu) + \dots,$$

ignoring higher order terms, the first two moments are

$$E(g(Y)) \approx g(\mu),$$

$$\text{Var}(g(Y)) \approx \sigma^2 (g'(\mu))^2; \text{ hence,}$$

$$g(Y) \approx (g(\mu), \sigma^2 (g'(\mu))^2)$$

(where ' $\approx$ ' denotes "is approximately distributed as"). Moreover, if $Y \underset{\sim}{\sim} N(\mu, \sigma^2)$, i.e., Y has an asymptotically normal distribution, then

$$g(Y) \underset{\sim}{\sim} N(g(\mu), \sigma^2 (g'(\mu))^2).$$

For example, if $Y \sim (\mu, \sigma^2)$ and $g(Y) = Y^2$, then

$$g(Y) \approx (\mu^2, \sigma^2 (2\mu)^2).$$

If $g(Y) = e^Y$, then $g(Y) \approx (e^\mu, \sigma^2(e^{2\mu}))$.

The delta method also has a multivariate version; suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right)$$

and suppose we want the distribution of $g(X, Y)$. Then, since

$$g(X, Y) \approx g(\mu_X, \mu_Y) + (X - \mu_X) \frac{\partial g}{\partial X}(\mu_X, \mu_Y) + (Y - \mu_Y) \frac{\partial g}{\partial Y}(\mu_X, \mu_Y) + \dots,$$

the approximate distribution of $g(X, Y)$ is

$$g(X, Y) \approx \left[g(\mu_X, \mu_Y), \sigma_X^2 \left(\frac{\partial g}{\partial X} \right)^2 + 2\sigma_{XY} \frac{\partial g}{\partial X} \frac{\partial g}{\partial Y} + \sigma_Y^2 \left(\frac{\partial g}{\partial Y} \right)^2 \right].$$

If furthermore $(X, Y) \stackrel{a}{\sim} \text{Normal}$, then $g(X, Y) \stackrel{a}{\sim} \text{Normal}$.

The delta method is very useful in practice; for example, we can use it to get an approximate value for $\text{Var}\left(\frac{\bar{X}}{\bar{Y}}\right)$ or $\text{Var}(\bar{X}\bar{Y})$.

Following Thavaneswaran (1988), consider the semimartingale model $X(t)$, with

$$dX(t) = h(t)Y(t)dt + dM(t)$$

where $h(t)$ is a time varying parameter of interest, $Y(t)$ is a predictable process and $M(t)$ is a zero mean, square integrable martingale.

$$\text{Let } H(t) = \int_0^t \frac{h(s)}{Y(s)} ds,$$

$$J(s) = I[Y(s) > 0]$$

Then an estimate of $H(t)$ is

$$\hat{H}(t) = \int_0^t \frac{dX(s)}{Y(s)I[Y(s) > 0]},$$

which can be written as $\hat{H}(t) = H(t) + \tilde{M}(t)$.

Under some regularity conditions, it can be shown that (Thavaneswaran, 1985) for large values of t ,

$$\hat{H}_t \underset{d}{\sim} N\left(H_t, \int_0^t \frac{h(s)}{Y(s)J(s)} ds\right).$$

Then for large values of t , viewing $\hat{H}_t$ as a diffusion process of the form

$$d\hat{H}_t = dH_t + d\tilde{M}_t$$

and using Itô's formula (Elliot, 1982) for any function $g \in C^2$ (twice continuously differentiable), we have

$$\begin{aligned} dg(\hat{H}(t)) &\equiv g'(\hat{H}(t))d\hat{H}_t + \frac{1}{2}g''(\hat{H}(t))d\langle \tilde{M} \rangle_t, \\ \langle g(\hat{H}(t)) \rangle_t &\equiv \int_0^t \left(g'(\hat{H}(s))\right)^2 d\langle \tilde{M} \rangle_s. \end{aligned}$$

Here, the notation $\langle \tilde{M} \rangle_t$ represents the conditional variance depending on time t .

Then, from Itô's formula, it follows that

$$\hat{H}_t \underset{d}{\sim} N(\hat{H}(t), \langle \tilde{M} \rangle_t) \text{ implies}$$

$$g(\hat{H}_t) \underset{d}{\sim} N\left(\int_0^t g'(\hat{H}(s))dH_s + \frac{1}{2}\int_0^t g''(\hat{H}(s))d\langle \tilde{M} \rangle_s, \int_0^t \left(g'(\hat{H}(s))\right)^2 d\langle \tilde{M} \rangle_s\right).$$

3.6.1 Examples

We start with an example of the delta method in a stochastic process context, and in subsequent examples, we use the delta method in a survival analysis context.

Example 3.6.1.1: Hasminskii and Ibragimov (1985) have studied nonparametric estimation of a linear functional in a Gaussian white noise process. Here we consider estimation of nonlinear functionals under Gaussian white noise. Consider the model

$$dX(t) = \theta(t) dt + \epsilon dW_t, \quad 0 \leq t \leq T$$

where ϵ is a positive number, $\theta(t)$ is a time varying parameter and $W(t)$ is a standard Gaussian process having mean zero and variance t ; i.e., for fixed t , $W(t) \sim N(0, t)$.

Hasminskii and Ibragimov (1985) have studied the estimation of $F(\theta)$,

$$F(\theta) = \int_0^t f(u) \theta(u) du,$$

a linear functional of the unknown parameter θ .

They have proposed the estimator

$$\hat{F} = F(\hat{\theta}) = \int_0^t f(u) dX(u).$$

Then the variance process (quadratic characteristic) of the estimator is

$$\langle \hat{F} \rangle = \epsilon^2 \int_0^t f^2(u) du.$$

Now consider $F(\theta) = \exp \left[\int_0^t \theta(u) du \right]$, a nonlinear functional in $\theta(t)$.

We propose the estimator $\hat{F} = \exp \left[\int_0^t \hat{\theta}(u) du \right]$ for F , where

$$\hat{F} = \exp \left[\int_0^t dX(u) \right] = e^{x_t} - 1 \quad [X(0) = 0]$$

To find the variance process of $\hat{F}$, we use Itô's formula with $g(x) = e^x$:

$$de^{x_t} = e^{x_t} dX_t + \frac{\epsilon^2}{2} e^{x_t} dt$$

$$e^{x_t} = \int_0^t e^{x_u} \theta(u) du + \epsilon \int_0^t e^{x_u} dW_u + \frac{\epsilon^2}{2} \int_0^t e^{x_u} du$$

Hence, the variance process of the estimate $\hat{F} = e^{x_t}$ is $\epsilon^2 \int_0^t e^{2x_u} du$.

Example 3.6.1.2: Suppose we want to estimate a functional of $H(t)$, the cumulative hazard rate, in a censoring setup,

$$N(t) = h(t) Y(t) dt + dM_t,$$

$$\text{where } H(t) = \int_0^t h(s) ds,$$

$N(t)$ is the number of deaths up to time t , $Y(t)$ is the number at risk, $M(t)$ is a zero mean, square integrable martingale, $h(t) = \frac{f(t)}{1 - F(t)}$ and the survival function is

$$S(t) = e^{-H_t}, \quad \hat{H}_t \approx N \left(H_t, \int_0^t \frac{h(s)}{Y(s)J(s)} ds \right),$$

$$g(x) = e^{-x}, \quad g'(x) = -e^{-x},$$

$$\langle \hat{S}(t) \rangle_t = \int_0^t e^{-2\hat{H}(s)} \frac{h(s)}{Y(s)} ds. \quad (3.3)$$

Hence, the variance process of an estimator of $S(t) = e^{-\hat{H}(t)}$ is given by (3.3).

We propose to estimate $\langle \tilde{S}(t) \rangle_t$ by

$$\int_0^t \frac{e^{-2\hat{H}(s)} h(s)}{Y^2(s)} dN(s). \quad (3.4)$$

In the literature the estimator,

$$\int_0^t \frac{1}{Y^2(s)} dN(s) = \sum_{j: X_j \leq t} \frac{1}{n_j^2} \quad (3.5)$$

has been used to construct confidence intervals for $S(t)$.

Since $e^{-x} < 1$ for $x > 0$, we note that (3.4) < (3.5).

Hence, an approach based on $It\hat{o}$'s formula provides a smaller variance for the product limit estimator than the nonparametric estimator in (3.5).

Example 3.6.1.3: Robust Estimators.

In estimation problems with independent observations $X_1, X_2, \dots, X_n$, the parameter of interest can frequently be expressed as a functional $\theta = g(F)$ of the underlying distribution function. With no censoring present, the usual estimator is

$$\tilde{\theta} = g(F_n), \text{ where } F_n \text{ is the empirical distribution function.}$$

With censoring present, a reasonable estimator is

$$\hat{\theta} = g(\hat{F}) = g(1 - \hat{S}), \text{ where } \hat{S} = 1 - \hat{F} \text{ is the product limit estimator.}$$

That is, with censored observations, functionals of the survival function $S(t)$, e.g., the mean, etc., are estimated by functionals of the product limit estimator $\hat{S}$.

$$\begin{aligned}\text{The mean } \theta = g(F) &= \int_0^{\infty} x dF(x) = \int_0^{\infty} (1 - F(x)) dx \\ &= \int_0^{\infty} S(t) dt, \text{ a functional of } S(t).\end{aligned}$$

For any positive random variable the n th moment, $\theta = E[X^n]$ (see, for example, Shirayayev, 1984, p. 206) is

$$\theta = g(F) = \int_0^{\infty} x^n dF(x) = n \int_0^{\infty} x^{n-1} (1 - F(x)) dx = n \int_0^{\infty} t^{n-1} S(t) dt.$$

When there is no censoring a nonparametric estimate of mean $\theta = E[X]$ is

$$\hat{\theta} = g(F_n) = \int_0^{\infty} x dF_n(x) = \bar{X} = \int_0^{\infty} (1 - F_n(x)) dx,$$

where $F_n(x)$ is the empirical cumulative distribution function.

For any nonnegative random variable X , the mean is

$$\theta = E[X] = \int_0^{\infty} S(t) dt.$$

With censoring, an estimator of the mean θ is

$$\hat{\theta} = \int_0^{\infty} \hat{S}(t) dt$$

and the n th moment, $\theta = g(S) = n \int_0^{\infty} t^{n-1} S(t) dt$ can be estimated by

$$\hat{\theta} = g(\hat{S}) = n \int_0^{\infty} t^{n-1} \hat{S}(t) dt.$$

For example, here we have a situation where $\hat{S}$ is a semimartingale; since $\hat{\theta}$ is a function of $\hat{S}$, we can use Itô's formula to obtain variance process of $\hat{\theta}$. We now obtain a semimartingale form for the product limit estimate $\hat{S}(t)$.

By definition, $-\frac{dF(t)}{1-F(t)} = h(t)dt = dH(t)$.

where $F(t)$ is the cumulative distribution function, $h(t)$ is the hazard rate and $H(t)$ is the cumulative hazard function. Hence, the survival function $S(t) = 1 - F(t)$ can be written as

$$\begin{aligned} dS(t) &= S(t-)d(-H(t)) \\ S(t) &= 1 + \int_0^t S(u-)d(-H(u)). \end{aligned} \quad (3.6)$$

We note that, for every semimartingale X_t (Elliot 1982, p. 156), the equation

$$Z_t = 1 + \int_0^t Z_{s-}dX_s$$

has a unique solution

$$Z_t = e^{X_t^c} \prod_{s \leq t} (1 + \Delta X_s), \text{ where } \Delta X_s = X_s - X_{s-}.$$

Hence, from (3.3), the solution for $S(t)$ can be written as $S(t) = e^{-H(t)}$.

Moreover, if we try to estimate $S(t)$ using $\hat{H}(t) = \int_0^t \frac{dN(s)}{J(s)Y(s)}$ which is a discontinuous

function, then, $\hat{S}(t) = 1 + \int_0^t \hat{S}(u)d(-\hat{H}(u))$. (3.7)

and the solution $\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{Y(s)}\right) \quad [\because \hat{H}^c(t) = 0]$

turns out to be the product limit estimator

$$\hat{S}(t) = 1 - \int_0^t \hat{S}(u-) d(\hat{H}(u)), \quad (3.8)$$

where $\hat{H}(t) = H(t) + \tilde{M}(t)$.

$$\text{Hence, } \langle \hat{S}(t) \rangle_t = \int_0^t \hat{S}(u-)^2 d\langle \tilde{M} \rangle_u = \int_0^t [\hat{S}(u-)]^2 \frac{h(s)}{Y(s)} ds.$$

An estimator of $\langle \hat{S}(t) \rangle_t$ proposed in Ordoukhani, Thavaneswaran and Samanta (1991) is

$$\int_0^t [\hat{S}(u-)]^2 \frac{dN(s)}{Y(s)} = \sum_{j: t_j \leq t} \hat{S}(t_j) \frac{d_j}{n_j}.$$

3.6.2 Estimation of functionals of $S(t)$

For any functional $g(\cdot)$ of $S(t)$,

$$\theta = g(S(t)) = n \int_0^\infty t^{n-1} S(t) dt \quad \text{can be estimated by}$$

$$\hat{\theta} = g(\hat{S}) = n \int_0^\infty t^{n-1} \hat{S}(t) dt.$$

Using the martingale central limit theorem (Liptser and Shinyayev, 1980), it can be shown that asymptotically, as $n \rightarrow \infty$, $\hat{S}(t)$ becomes a diffusion process. Now we try to find the variance process of

$$\hat{\theta}_t = n \int_0^t u^{n-1} \hat{S}(u) du.$$

Using Itô's formula, i.e., if

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t$$

then for any twice continuously differentiable function $f(\cdot)$

$$df(X_t) = f'(X_t)dX_t + \frac{1}{2}f''(X_t)b^2(t, X_t)dt$$

In other words, integration by parts in ordinary calculus may be related to the following formula in Itô's calculus,

$$\frac{1}{2} \int_0^t f''(X_u)du = f(X_t) - f(X_0) - \int_0^t f'(X_u)dX_u. \quad (3.9)$$

When $X_t = W_t$, $f(x) = x^2$, $f'(x) = 2x$, $f''(x) = 2$, (3.9) becomes

$$t = W_t^2 - W_0^2 - 2 \int_0^t W_u dW_u,$$

which shows that rules of ordinary calculus do not necessarily apply in stochastic calculus.

We have noticed that $\hat{S}(t)$ can be written as

$$d\hat{S}(t) = -\hat{S}(t)h(t)dt - \hat{S}(t-)\tilde{dM}(t). \quad (3.10)$$

Hence, to find $\int_0^t \hat{S}(u)du$ or $n \int_0^t u^{n-1} \hat{S}(u)du$, we have to express the variance process

$\int_0^t \hat{S}(u)du$ in terms of stochastic integrals. Using Itô's formula with

$f(X) = X^3$, $f'(X) = 3X^2$, $f''(X) = 6X$, we have

$$3 \int_0^t X_u du = X_t^3 - X_0^3 - 3 \int_0^t X_u^2 dX_u$$

$$3 \int_0^t \hat{S}(u) du = \hat{S}^3(t) - N^3 - 3 \int_0^t \hat{S}(u)^2 d\hat{S}(u), \quad (3.11)$$

where $X_0 = N$ is the initial value

$$\langle \hat{\theta} \rangle_t \equiv 3 \int_0^t [\hat{S}(u-)]^4 \frac{h(u)}{Y(u)} du. \quad (3.12)$$

Hence, for the parameter $\theta = n \int_0^\infty u^{n-1} S(u) du$, expressed as a functional of $S(t)$,

we propose $\hat{\theta}_t = n \int_0^\infty u^{n-1} \hat{S}(u) du$ as an estimator of θ , which has the variance process

(for $n = 1$)

$$\langle \hat{\theta} \rangle_t = 3 \int_0^t [\hat{S}(u-)]^4 \frac{h(u)}{Y(u)} du. \quad (3.13)$$

An unbiased estimate of the variance of $\hat{\theta}_t$ is

$$3 \int_0^t [\hat{S}(u)]^4 \frac{dN(u)}{Y^2(u)} = 3 \sum_{j: t_j < t} \frac{[\hat{S}(t_j)]^4}{n_j^2},$$

where n_j is the number at risk at time t_j .

3.7.1 Empirical distribution function

If there are no censored observations in a sample of size n , the empirical survivor function (ESF) is defined as

$$\tilde{S}_n(t) = \frac{\text{number of individuals with survival times } \geq t}{\text{total number of individuals under study}}; \quad t \geq 0$$

$\tilde{S}_n(t)$ is assumed to be constant between two adjacent death times. That is, $\tilde{S}_n(t)$ is a step function that decreases by $\frac{1}{n}$ just after each observed survival time. In particular,

$$\begin{aligned}\tilde{S}_n(t) &= 1, \text{ for values of } t \text{ before the first death time.} \\ &= 0, \text{ for values of } t \text{ after the final death time.}\end{aligned}$$

equivalently, $\tilde{S}_n(t) = 1 - \tilde{F}_n(t)$.

where $\tilde{F}_n(t)$ is the empirical distribution function (e.d.f.), i.e.

$$\begin{aligned}\tilde{F}_n(t) &= \frac{1}{n} \sum_{i=1}^n I(T_i \leq t) \\ \tilde{F}_{n+1}(t) &= \frac{1}{n+1} \sum_{i=1}^{n+1} I(T_i \leq t) = \frac{1}{n+1} \left[\sum_{i=1}^n I(T_i \leq t) + I(T_{n+1} \leq t) \right] = \frac{1}{n+1} [n\tilde{F}_n(t) + I(T_{n+1} \leq t)] \\ \Leftrightarrow (n+1)\tilde{F}_{n+1}(t) &= n\tilde{F}_n(t) + I(T_{n+1} \leq t) \\ \Leftrightarrow \tilde{F}_{n+1}(t) &= \frac{n}{(n+1)} \tilde{F}_n(t) + \frac{1}{(n+1)} I(T_{n+1} \leq t) \\ \Leftrightarrow \tilde{F}_{n+1}(t) &= (1-w)\tilde{F}_n(t) + wI(T_{n+1} \leq t) \tag{3.14}\end{aligned}$$

where $w = \frac{1}{(n+1)}$.

The relation in (3.14) gives the recursive form of the e.d.f.; that is, if we have an empirical distribution function based on n observations, we can easily update the empirical distribution function based on $(n+1)$ observations and so on. However, the resulting estimator need not be optimal.

Suppose we have the following survival times (in weeks) of seventeen patients:

$$5, 17, 21, 12, 45, 17, 28, 23, 27, 54, 39, 15, 57, 68, 74, 86, 32.$$

With this data set, the plot of the empirical distribution function is shown in Figure 3.3. The relation (3.14) can also be generated using a recursive algorithm given in the Appendix.

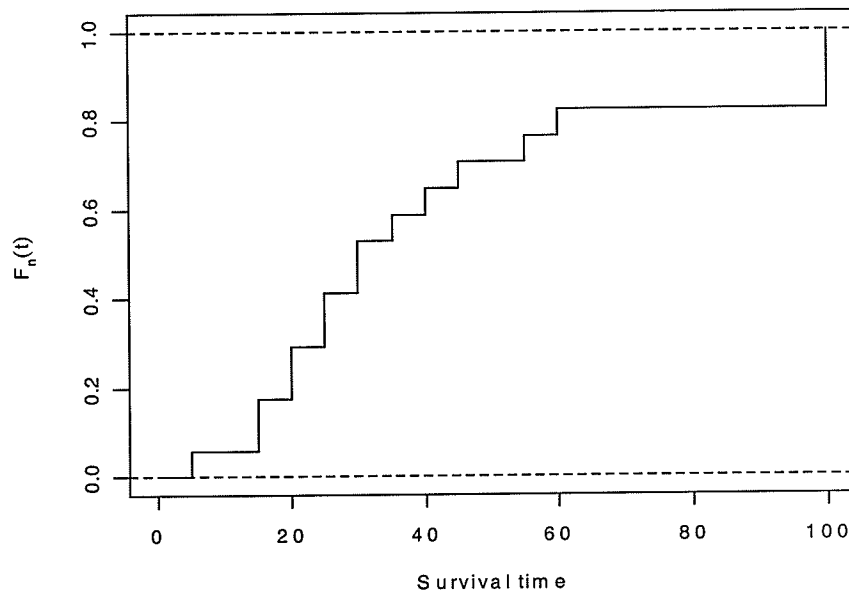


Figure 3.3 Empirical distribution function.

3.7.2 Nonparametric Bayes method

Newton, Quintana and Zhang (1998) developed a recursive algorithm using elementary facts about nonparametric predictive distributions to approximate nonparametric Bayes estimates.

Suppose that an unobserved random sample $\theta_1, \theta_2, \dots, \theta_n$ from an unknown distribution G determines the conditional probability structure of the observations

$Y_1, \dots, Y_n$. We can consider θ_i as real or vector-valued and observations can come in various types. The unobserved random sample may be missing data or it may represent parameters that fluctuate across experimental units. This recursive algorithm is applied to an interval censoring problem considering θ_i as a survival time and Y_i an interval known to contain θ_i .

The prior on the unknown distribution G is indexed by a positive finite measure defined on the range χ (a finite set) of θ_i . We denote the prior measure by m_0 , which contains the Bayesian information about G and may be factored into a probability measure G_0 representing a prior guess and a prior mass (predetermined weight) $\alpha = m_0(\chi)$. The prior mass α measures the extent to which new information will change the Bayesian opinion.

Conditional on data $D = \{Y_i\}$, for a measurable subset $B \subset \chi$, the Bayes estimate $E\{G(B)|D\}$ coincides with the posterior predictive distribution $G_n(B) = P(\theta_{n+1} \in B|D)$ and we want to calculate this distribution. The most simple case motivates the general recursive approximation. The object of interest is the posterior predictive distribution on $n=1$,

$$\begin{aligned} G_1(B) &= P(\theta_2 \in B|Y_1) \\ &= E\{P(\theta_2 \in B|\theta_1)|Y_1\} \\ &= E\left\{\frac{\alpha}{1+\alpha}G_0(B) + \frac{1}{1+\alpha}I_B(\theta_1)|Y_1\right\} \\ &= \frac{\alpha}{1+\alpha}G_0(B) + \frac{1}{1+\alpha}P(\theta_1 \in B|Y_1) \end{aligned}$$

Whereas the measure $m_0 = \alpha G_0$ represents uncertainty prior to observing data, the measure $m_1 = (\alpha+1)G_1$ encodes updated uncertainties. By accumulating the measures, we get:

$$m_1(B) = m_0(B) + P_0(\theta_1 \in B | Y_1) \quad (3.15)$$

It is interesting to note that Bayes rule enters the second term in (3.15). The corresponding posterior calculations are driven by the prior G_0 on θ (hence the subscripting in P_0). If no information is lost going from θ_1 to Y_1 , then the added measure is simply a point mass at θ_1 .

The idea that predictive uncertainty can be encoded in single measure and that learning occurs by adding measures lead us to the following recursion generalizing (3.15):

$$m_i(B) = m_{i-1}(B) + P_{i-1}(\theta_i \in B | Y_i) \quad (3.16)$$

for $i \geq 1$.

In terms of probability distributions

$$G_i(B) = (1 - w_i)G_{i-1}(B) + w_i P_{i-1}(\theta_i \in B | Y_i) \quad (3.17)$$

where again the Bayes rule enters the second term taking G_{i-1} as the updated prior distribution for θ_i . The nominal weights are $w_i = 1/(\alpha + i)$. It is interesting to note that relation (3.17) has a similar form as relation (3.14) for the empirical distribution function.

As constructed, G_i is not the exact posterior predictive distribution for θ_{i+1} in general, even though it is so when $i=1$. It will be only in the case of no information loss,

that is, when $Y_i = \theta_i$. Thus the value of G_i depends on the order in which $Y_1, \dots, Y_i$ are processed. What gives some credence to the recursion is that the dependence on order can be relatively weak. The suggested algorithm then is to arrange in some order $Y_1, \dots, Y_n$ and to process them through (3.17) to produce an approximate Bayes estimate G_n . Calculations being $O(n)$, we can easily re-evaluate G_n over a random sample of orderings and average the results.

The form of information loss that we have studied most extensively is interval censoring. Rather than observing $\theta_i \in \mathcal{X}$, we observe an interval $Y_i \subset \mathcal{X}$ that is known to contain θ_i . The recursive formula (3.17) is particularly simple for interval censored data:

$$G_i(B) = (1 - w_i) G_{i-1}(B) + w_i \frac{G_{i-1}(B \cap Y_i)}{G_{i-1}(Y_i)}. \quad (3.18)$$

As an illustrative example, we consider a much studied data set on cosmetic deterioration after radiotherapy of the breast. Each patient is monitored periodically and a time to deterioration θ_i is known either to occur either in an interval between hospital visits or is right-censored and known to occur only beyond some maximum observed time. This example has been considered by a number of authors. Our goal is to illustrate how the recursive approximations work in a typical problem. The data are reported in Table 1 of Gentleman and Geyer (1994). There are $n = 46$ intervals in the data set. Associated with each woman in the study is a time θ_i (days) indicating the time after radiotherapy treatment that a defined amount of change occurs in the treated tissue, and the i th interval is known to contain θ_i .

Figures 3.4(a) and 3.4(b) show the results of the recursive approximation (3.18) using the nominal weighting scheme. We take G_0 to be an exponential prior with mean equal to one half a year and we consider two values of α : 1 and 5. The left figures show G_{46} for 100 random orderings of the data. We use sets $B = [0, \theta]$ to drive the recursion for θ in the grid $\{0, 1, 2, \dots, 100, 1000\}$. The outlying grid point at 1000 days is included to account for mass beyond the range of data. The average-over-orderings $\bar{G}_n$ shown in the right-hand plots, is a very accurate approximation to the correct posterior predictive distribution in this example. The dashed line represents the prior c.d.f. G_0 for $\theta \in [0, 100]$. The Appendix provides some code to implement the recursion using built-in functions in “R”.

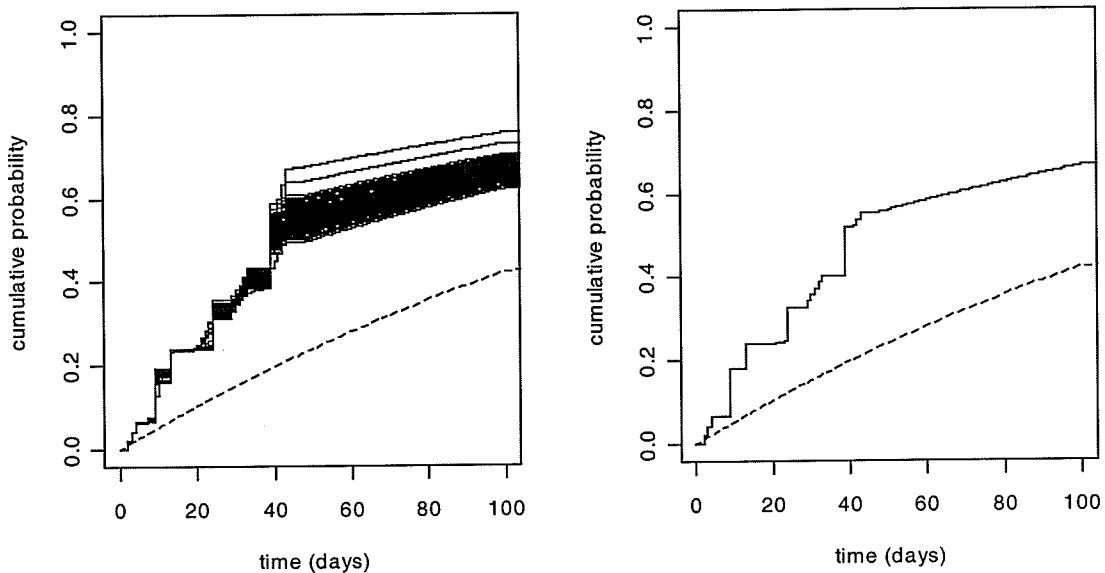


Figure 3.4(a): G_n for 100 random orderings (left) and the average-over-orderings $\bar{G}_n$ (right) with $\alpha = 1$. Dashed line represents the prior c.d.f. G_0 .

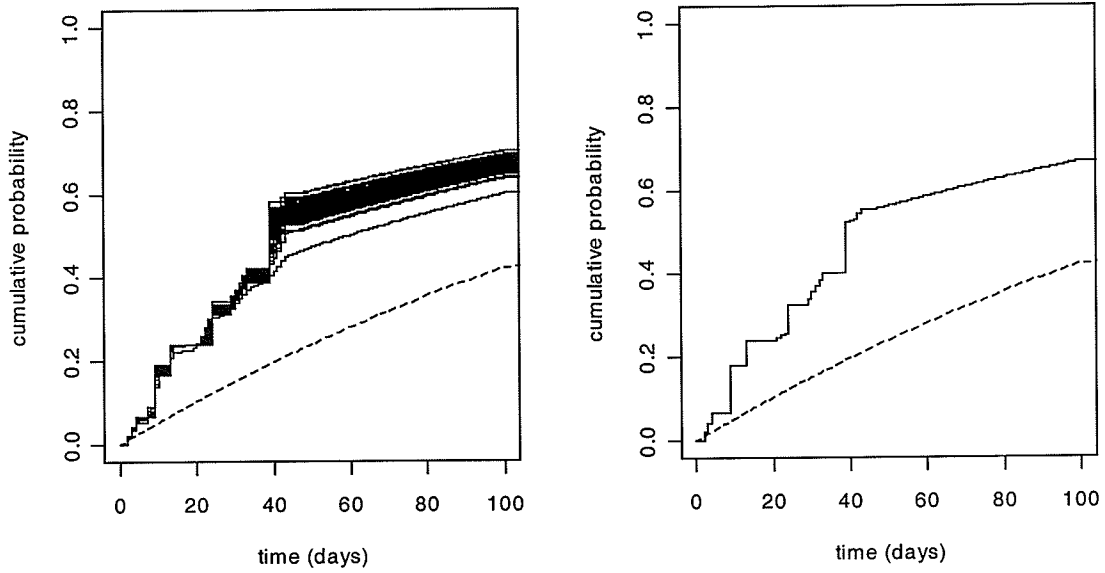


Figure 3.4(b): G_n for 100 random orderings (left) and the average-over-orderings $\bar{G}_n$ (right) with $\alpha = 5$. Dashed line represents the prior c.d.f. G_0 .

3.7.3 Comparison

Newton, Quintana and Zhang (1998) considered estimation of posterior distribution functions using a Bayesian method, with predetermined weights (i.e., prior mass) $\alpha = 1$ and $\alpha = 5$. This method is applied to independent, interval censored data. For censored correlated data we have modified the Nelson-Aalen estimator and determined the optimum α value as follows: using the EWMA method, we choose α by minimizing the mean squared forecast errors, and obtain the estimate of the distribution function with censored correlated observations by using the relationship (3.2).

CHAPTER FOUR

TESTING HYPOTHESES

4.1 Introduction

In biomedical research, the comparison of populations based on sample observations is very common. In this chapter, we discuss one-sample and k ($k \geq 2$) sample tests that compare the integrated difference between the observed and expected hazard rate. Some tests use ranks instead of original observations and are easy to apply. In each situation, we include examples to illustrate the testing procedures. Examples are based on censored, truncated and complete data sets. In section 4.2 and 4.3, examples using censored data sets are considered, and in section 4.4 both complete and truncated data sets are used. All tests discussed herein are based on large sample approximations to the distribution of chi-square statistics.

4.2 One-Sample Tests

Suppose $Y(t_i)$ is the number of individuals under study prior to an observed event time t_i and d_i is the number of deaths (event) at time t_i , where $t_1 < t_2 < \dots < t_r \leq T$.

Here, t_r is the time when the last death occurs and T is the largest of the observed

study times.

On the basis of the above sample observations, we wish to test the null hypothesis

$$H_0 : h(t) = h_0(t) \text{ for all } t \leq T$$

against the alternative hypothesis

$$H_a : h(t) > h_0(t) \text{ for all } t \leq T, \text{ with strict inequality for some } t.$$

where, $h(t)$ is the population hazard rate and $h_0(t)$ is completely specified.

The Nelson-Aalen estimator of the cumulative hazard function $H(t)$ is given by

$$\hat{H}(t) = \sum_{i: t_i \leq t} \frac{d_i}{Y(t_i)}.$$

We note that the ratio $\frac{d_i}{Y(t_i)}$ gives an estimate of $h(t_i)$, the hazard rate at time t_i .

In the present study, the difference between a weighted version of the estimate of $h(t_i)$ and the hypothesized rate is used as a test statistic. Let $W(t)$ be a weight function with the property that $W(t) = 0$ whenever $Y(t) = 0$. The test statistic for testing H_0 is

$$\begin{aligned} Z(T) &= O(T) - E(T) \\ &= \sum_{i=1}^r W(t_i) \frac{d_i}{Y(t_i)} - \int_0^T W(s) h_0(s) ds \\ &= \int_0^T W(s) \frac{dN(s)}{Y(s)} - \int_0^T W(s) h_0(s) ds \end{aligned}$$

where, $dN(s) = h_0(s)Y(s)ds + dM(s)$.

Under the null hypothesis, the variance of $Z(T)$ is given by

$$V[Z(T)] = \int_0^T W^2(s) \frac{h_0(s)}{Y(s)} ds.$$

For large samples, the statistic $\frac{[Z(T)]^2}{V[Z(T)]}$ has approximately a central chi-square distribution with one degree of freedom when the null hypothesis H_0 is true. We can also use $\frac{Z(T)}{\sqrt{V[Z(T)]}}$ as a statistic for testing the null hypothesis H_0 against the alternative hypothesis H_a . Clearly under H_0 , $\frac{Z(T)}{\sqrt{V[Z(T)]}}$ has approximately a standard normal distribution provided the sample is large enough. Large values of the statistic leads to the rejection of the null hypothesis H_0 .

The most popular choice of weight function is $W(t) = Y(t)$ (the number of individuals prior to the observed event time t), which yields the one-sample log-rank test. To allow for possible left truncation, let T_j be the time on study and L_j be the delayed entry time for the j th patient. When T is equal to the largest time on study,

$$O(T) = \text{observed number of events at or prior to time } T$$

and with $W(t) = Y(t)$,

$$V[Z(T)] = \sum_{j=1}^n \int_0^T Y_j^2(s) \frac{h_0(s)}{Y_j(s)} ds = \sum_{j=1}^n \int_{L_j}^{T_j} Y_j(s) h_0(s) ds = \sum_{j=1}^n \int_{L_j}^{T_j} h_0(s) ds,$$

since for the j th patient $Y_j(s) = 1$; therefore

$$V[Z(T)] = \sum_{j=1}^n [H_0(s)]_{L_j}^{T_j} = \sum_{j=1}^n [H_0(T_j) - H_0(L_j)]$$

where $H_0(t)$ is the cumulative hazard under the null hypothesis.

Other weight functions proposed in the literature include the Harrington and Fleming (1982) family $W_{HF}(t) = Y(t) S_0(t)^p [1 - S_0(t)]^q$, $p \geq 0, q \geq 0$, where $S_0(t) = \exp[-H_0(t)]$ is the hypothesized survival function. By choice of p and q , one can put more weight on early departures from the null hypothesis (p much larger than q), late departures from the null hypothesis (p much smaller than q), or on departures in the mid-range ($p = q > 0$). The log-rank weight is a special case of this model with $p = q = 0$.

4.2.1 Example

In a study of the effectiveness of a combined drug regimen for the treatment of rheumatoid arthritis, 40 white patients were followed for a period ranging from 1 to 18 years. During the course of the study, 9 patients died, so 31 patients were still alive at the end of the study. The ages at entry into the study and at death for the 9 patients and also the ages at entry and last follow-up for the 31 patients were as follows:

Female deaths	Male deaths	Female survivors	Male survivors
66, 74	50, 59	50, 68	53, 68
60, 76	60, 66	55, 72	55, 62
70, 77	51, 69	56, 60	56, 63
71, 81	69, 71	45, 55	45, 51
	58, 71	48, 51	48, 61
		44, 55	49, 55
		33, 51	43, 51
		44, 50	44, 54
		60, 70	61, 70
		55, 60	45, 60
		60, 72	63, 72
		77, 80	74, 80
		70, 75	70, 76
		66, 70	66, 72
		59, 63	54, 70
		62, 63	

Using the all cause U.S. mortality table for 1989 presented in the Appendix (Table 3), we test the hypothesis that the death rate of these rheumatoid arthritis patients is not different from that in the general population using the log-rank test.

To perform this test, we use the sex specific survival rates. Define

T_j = time which shows the j th individual's age at death or the end of the study .

L_j = left-truncation time which shows the j th individual's age at entry into the study.

Table 4.1. Computation of One-sample, Log-rank test.

Individual		Status	Age at entry	Age at exit			
j	Sex	d_i	L_j	T_j	$H_0(L_j)$	$H_0(T_j)$	$H_0(T_j) - H_0(L_j)$
1	f	1	66	74	0.16053	0.31861	0.15808
2	f	1	60	76	0.09700	0.37999	0.28299
3	f	1	70	77	0.22516	0.41530	0.19015
4	f	1	71	81	0.24309	0.59789	0.35480
5	m	1	50	59	0.09269	0.17080	0.07811
6	m	1	60	66	0.18470	0.29934	0.11463
7	m	1	51	69	0.09829	0.38069	0.28240
8	m	1	69	71	0.38069	0.44840	0.06771
9	m	1	58	71	0.15818	0.44840	0.29022
10	f	0	50	68	0.04526	0.18993	0.14467
11	f	0	55	72	0.06493	0.26759	0.20266
12	f	0	56	60	0.07019	0.09700	0.02681
13	f	0	45	55	0.03362	0.06493	0.03131
14	f	0	48	51	0.03987	0.04842	0.00855
15	f	0	44	55	0.03188	0.06493	0.03305
16	f	0	33	51	0.02015	0.04842	0.02827
17	f	0	44	50	0.03188	0.04526	0.01338
18	f	0	60	70	0.09700	0.22516	0.12816
19	f	0	55	60	0.06493	0.09700	0.03206
20	f	0	60	72	0.09700	0.26759	0.17060
21	f	0	77	80	0.41530	0.54488	0.12958
22	f	0	70	75	0.23081	0.34787	0.11706
23	f	0	66	70	0.16053	0.22516	0.06463
24	f	0	59	63	0.08932	0.12471	0.03539
25	f	0	62	63	0.11464	0.12471	0.01008
26	m	0	53	68	0.11116	0.35122	0.24006
27	m	0	55	62	0.12706	0.21678	0.08972
28	m	0	56	63	0.13639	0.23506	0.09867
29	m	0	45	51	0.07084	0.09829	0.02744
30	m	0	48	61	0.08284	0.20000	0.11716
31	m	0	49	55	0.08755	0.12706	0.03951
32	m	0	43	51	0.06431	0.09829	0.03397
33	m	0	44	54	0.06745	0.11867	0.05122
34	m	0	61	70	0.20000	0.41300	0.21300
35	m	0	45	60	0.07084	0.18470	0.11386
36	m	0	63	72	0.23506	1.15706	0.92199
37	m	0	74	80	0.57558	0.95829	0.38271
38	m	0	70	76	0.41300	0.68120	0.26820
39	m	0	66	72	0.29934	0.48709	0.18776
40	m	0	54	70	0.11867	0.41300	0.29433
Total		9					6.07495

The observed value of the test statistic is $\chi^2 = (9 - 6.07495)^2 / 6.07495 = 1.408$ with one degree of freedom. The p -value of the test is 0.234. So, we conclude that the death rates of the rheumatoid arthritis patients do not differ from those in the general population.

4.3 Comparison of two groups of survival data

In the comparison of two groups of survival data, there are several nonparametric methods which can be used to quantify the extent of between-group differences. Here, we consider two nonparametric procedures:

- the log-rank test,
- the Wilcoxon test.

Suppose we have two groups of survival data, namely Group I and Group II. Across the two groups (for the pooled data) there are r distinct death times, $t_1 < t_2 < \dots < t_r$. At time t_i , d_{i1} individuals in Group I and d_{i2} individuals in Group II die, for $i = 1, 2, \dots, r$. If there are no ties, that is, unless two or more individuals in a group have the same recorded death time, the values of d_{i1} and d_{i2} will either be zero or unity. We further suppose that just before time t_i , there are Y_{i1} and Y_{i2} individuals at risk (i.e., alive and uncensored) in Group I and Group II, respectively. The situation is summarized in Table 4.2.

Table 4.2. Number of deaths at t_i in each of two groups of individuals.

Group	Number of deaths at t_i	Number surviving beyond t_i	Number at risk just before t_i
I	d_{i1}	$Y_{i1} - d_{i1}$	Y_{i1}
II	d_{i2}	$Y_{i2} - d_{i2}$	Y_{i2}
Total	d_i	$Y_i - d_i$	Y_i

4.3.1 The log-rank test

Let $h_j(t)$ be the hazard rate for group j ($j = 1, 2$). Then our hypothesis set is

$$H_0 : h_1(t) = h_2(t) \text{ on } [0, T]$$

against $H_a : h_1(t) \neq h_2(t) \text{ on } [0, T]$.

The null hypothesis considers that there is no difference in the survival experience of individuals in the two groups. One way of assessing the validity of this hypothesis is to consider the extent of the difference between the observed number of individuals in the two groups who die at each of the death times and the numbers expected under the null hypothesis. Information about the extent of these differences can then be combined over each of the death times.

In Table 4.2, if the marginal totals are regarded as fixed and the null hypothesis is true, the four entries in the table are solely determined by the value of d_{i1} . We can therefore regard d_{i1} as a random variable, which can take any value in the range from 0 to the minimum of d_i and Y_{i1} . In fact, d_{i1} has a hypergeometric distribution, according to which the probability that the random variable associated with the number of deaths

in Group I takes the value d_{i1} is

$$\frac{\binom{d_i}{d_{i1}} \binom{Y_i - d_i}{Y_{i1} - d_{i1}}}{\binom{Y_i}{Y_{i1}}}$$

The mean of the hypergeometric random variable d_{i1} is $e_{i1} = \frac{Y_{i1} d_i}{Y_i}$;

i.e., e_{i1} is the expected number of individuals who die at time t_i in Group I.

$$\text{So, } E(d_{i1} | d_i) = e_{i1} = \frac{Y_{i1} d_i}{Y_i} \quad (4.1)$$

$$\text{and } \text{Var}(d_{i1} | d_i) = v_{i1} = \frac{Y_{i1} Y_{i2} d_i (Y_i - d_i)}{Y_i^2 (Y_i - 1)}. \quad (4.2)$$

In the next step, we combine the information from the individual 2×2 tables for each death time to give an overall measure of the deviation of the observed values of d_{i1} from their expected values. The most straightforward way of doing this is to sum the differences $d_{i1} - e_{i1}$ over the total number of death times, r , in the two groups. The resulting statistic is given by $U_L = \sum_{i=1}^r (d_{i1} - e_{i1})$.

The mean and variance of U_L are:

$$E(U_L) = 0$$

$$\text{Var}(U_L) = \sum_{i=1}^r v_{i1} = V_L$$

When the number of death times is not too small, U_L has an approximate normal

distribution. Also, $Z = \frac{U_L}{\sqrt{V_L}}$ is approximately a standard normal random variable.

$$\text{i.e. } Z = \frac{\sum_{i=1}^r (d_{i1} - e_{i1})}{\sqrt{\sum_{i=1}^r \frac{Y_{i1}Y_{i2}d_i(Y_i - d_i)}{Y_i^2(Y_i - 1)}}$$

Clearly, $W_L = Z^2$ is approximately distributed as χ_1^2 , a chi-square random variable with one degree of freedom.

The statistic W_L summarizes the extent to which the observed survival times in the two groups of data deviate from those expected under the null hypothesis of no group differences. The larger the value of this statistic, the greater the evidence against the null hypothesis. Since $\sqrt{W_L}$ has approximately a standard normal distribution when the null hypothesis is true, the p -value can be easily calculated using a table of the standard normal distribution.

This method of combining information over a number of 2×2 tables was proposed by Mantel and Haenszel (1959), and is known as the Mantel-Haenszel procedure. In fact, the test based on this statistic has various names, including Mantel-Cox and Peto-Mantel-Haenszel, but it is probably best known as the log-rank test.

4.3.2 Example: Prognosis for women with breast cancer

In this example, the data are the survival times of women with breast cancer, grouped according to whether a section of the tumour was positively or negatively stained.

This data set has been taken from Collett (1994) and are presented in the Appendix (Table 4). We test the null hypothesis that there is no difference in the survival experience of the two groups of women using the log-rank test. The required calculations are shown in Table 4.3.

We begin by ordering the observed death times across the two groups of women. The numbers of women in each group who die at each death time and the numbers who are at risk at each time are then calculated. These values are d_{i1} , Y_{i1} , d_{i2} and Y_{i2} . Also, d_i and Y_i contain the total numbers of deaths and the total numbers of women at risk over the two groups, at each death time. Next, we compute the values of e_{i1} and v_{i1} from equations (4.1) and (4.2), respectively. Then, the log-rank statistic is calculated from $U_L = \sum d_{i1} - \sum e_{i1} = 5 - 9.565 = -4.565$ and the value of $Var(U_L)$ is $V_L = \sum v_{i1} = 5.929$. Finally, the value of the log-rank test statistic is $W_L = (-4.565)^2 / 5.929 = 3.515$.

The corresponding p -value is calculated from the probability that a chi-squared variate with one degree of freedom is greater than or equal to 3.515. This probability is equal to $2\Pr(Z > \sqrt{3.515}) = 0.061$. Since the p -value is greater than 0.05, we do not reject the null hypothesis of the equality of two survivor functions at the 5% level of significance. We also note that we have a different conclusion if the test is performed at the 6.5% level of significance.

Table 4.3. Construction of two-sample, Log-rank test (following Klein).

t_i	Y_{i1}	d_{i1}	Y_{i2}	d_{i2}	Y_i	d_i	$Y_{i1} \left(\frac{d_i}{Y_i} \right)$	$d_{i1} - Y_{i1} \left(\frac{d_i}{Y_i} \right)$	$\frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i$
5	13	0	32	1	45	1	0.288889	-0.28889	0.20543
8	13	0	31	1	44	1	0.295455	-0.29545	0.20816
10	13	0	30	1	43	1	0.302326	-0.30233	0.21092
13	13	0	29	1	42	1	0.309524	-0.30952	0.21372
18	13	0	28	1	41	1	0.317073	-0.31707	0.21654
23	13	1	27	0	40	1	0.3250	0.6750	0.21938
24	12	0	27	1	39	1	0.307692	-0.30769	0.21302
26	12	0	26	2	38	2	0.631579	-0.63158	0.42045
31	12	0	24	1	36	1	0.333333	-0.33333	0.22222
35	12	0	23	1	35	1	0.342857	-0.34286	0.22531
40	12	0	22	1	34	1	0.352941	-0.35294	0.22837
41	12	0	21	1	33	1	0.363636	-0.36364	0.23140
47	12	1	20	0	32	1	0.3750	0.6250	0.23438
48	11	0	20	1	31	1	0.354839	-0.35484	0.22893
50	11	0	19	1	30	1	0.366667	-0.36667	0.23222
59	11	0	18	1	29	1	0.37931	-0.37931	0.23543
61	11	0	17	1	28	1	0.392857	-0.39286	0.23852
68	11	0	16	1	27	1	0.407407	-0.40741	0.24143
69	11	1	15	0	26	1	0.423077	0.57692	0.24408
71	9	0	15	1	24	1	0.3750	-0.3750	0.23438
113	6	0	10	1	16	1	0.3750	-0.3750	0.23438
118	6	0	8	1	14	1	0.428571	-0.42857	0.24490
143	6	0	7	1	13	1	0.461538	-0.46154	0.24852
148	6	1	6	0	12	1	0.5000	0.5000	0.25000
181	5	1	4	0	9	1	0.555556	0.44444	0.24691
Total		5		21			9.565128	- 4.565127	5.929

$$\sum_{i=1}^r e_{i1}$$

$$U_L$$

$$V_L$$

Moreover, it is shown by Anscombe (1948) that for moderately large values of Y_{i1} and Y_{i2} , the statistic

$$Z^* = \frac{\sin^{-1} \sqrt{\frac{d_{.1} + (3/8)}{Y_{11} + (3/4)}} - \sin^{-1} \sqrt{\frac{d_{.2} + (3/8)}{Y_{12} + (3/4)}}}{\sqrt{\frac{1}{4Y_{11} + 2} + \frac{1}{4Y_{12} + 2}}}$$

has approximately a standard normal distribution, where $d_{.1}$ and $d_{.2}$ are the total number of individuals who died in Group I and Group II, respectively, and Y_{11} and Y_{12} are the total number of individuals at risk at the beginning of the study over the two groups.

We can simplify Z^* as $Z^* = \frac{a-b}{\sqrt{c}}$, where

$$a = \sin^{-1} \sqrt{\frac{d_{.1} + (3/8)}{Y_{11} + (3/4)}}, \quad b = \sin^{-1} \sqrt{\frac{d_{.2} + (3/8)}{Y_{12} + (3/4)}}$$

$$\text{and } c = \frac{1}{4Y_{11} + 2} + \frac{1}{4Y_{12} + 2}$$

Taking values $d_{.1} = 5$, $d_{.2} = 21$, $Y_{11} = 13$ and $Y_{12} = 32$, from the data on the survival times of women with breast cancer, we calculate

$$a = 0.675$$

$$b = 0.941$$

$$\sqrt{c} = 0.162.$$

So, we get $Z^* = -1.638$, with

$$p\text{-value} = 2 P(Z > 1.638) = 0.5 - 0.4495 = 0.101.$$

Again, we do not reject the null hypothesis of the equality of two survivor functions at

the 5% level of significance. Anscombe's $\sin^{-1}$ transformation of the binomial proportions is useful when the sample size is not large.

4.3.3 The Wilcoxon test

The Wilcoxon test of the null hypothesis that there is no difference in the survivor functions for two groups of survival data is very similar to the log-rank test. The Wilcoxon test is based on the statistic

$$U_w = \sum_{i=1}^r Y_i (d_{i1} - e_{i1}),$$

where d_{i1} is the number of deaths at time t_i in Group I and e_{i1} is as defined in equation (4.1). The difference between U_w and U_L is that, in the Wilcoxon test, each difference $d_{i1} - e_{i1}$ is weighted by Y_i , the total number of individuals at risk at time t_i . The effect of this is to give less weight to differences between d_{i1} and e_{i1} at those times when the total number of individuals who are still alive is small, that is, at the longest survival times. This statistic is therefore less sensitive than the log-rank statistic to deviations of d_{i1} from e_{i1} in the right tail of the distribution of survival times.

The variance of the Wilcoxon statistic U_w is given by

$$V_w = \sum_{i=1}^r Y_i^2 v_{i1},$$

where v_{i1} is given in equation (4.2), and so the Wilcoxon test statistic is

$$W_w = \frac{U_w^2}{V_w},$$

which has an asymptotic chi-square distribution with one degree of freedom when the null hypothesis is true. The Wilcoxon test is therefore conducted in the same manner as the log-rank test.

4.3.4 Example: Prognosis for women with breast cancer

For the data on the survival times of women with tumours which were positively or negatively stained, the value of the Wilcoxon statistic is $U_w = -159$, and the variance of the statistic is $V_w = 6048.014$. The value of the approximate chi-squared statistic W_w is 4.180, and the corresponding p -value is 0.041. The required calculations are shown in the Table 4.4.

On the basis of this calculation, we reject the null hypothesis of the equality of two survivor functions at the 5% level of significance. Unlike the log-rank test, the Wilcoxon test leads to a different conclusion.

Table 4.4. Calculation of the Log-rank and Wilcoxon statistic (following Collett).

Death time	d_{i1}	Y_{i1}	d_{i2}	Y_{i2}	d_i	Y_i	e_{i1}	v_{i1}	$(d_{i1} - e_{i1})/\sqrt{v_{i1}}$	$Y_i(d_{i1} - e_{i1})$	$Y_i^2 v_{i1}$
5	0	13	1	32	1	45	0.2889	0.2054	-0.637	-13.001	415.94
8	0	13	1	31	1	44	0.2955	0.2082	-0.648	-13.002	403.08
10	0	13	1	30	1	43	0.3023	0.2109	-0.658	-12.999	389.95
13	0	13	1	29	1	42	0.3095	0.2137	-0.67	-12.999	376.97
18	0	13	1	28	1	41	0.3171	0.2165	-0.682	-13.001	363.94
23	1	13	0	27	1	40	0.3250	0.2194	1.441	27.00	351.04
24	0	12	1	27	1	39	0.3077	0.2130	-0.667	-12.00	323.97
26	0	12	2	26	2	38	0.6316	0.4205	-0.974	-24.001	607.2
31	0	12	1	24	1	36	0.3333	0.2222	-0.707	-11.999	287.97
35	0	12	1	23	1	35	0.3429	0.2253	-0.722	-12.002	275.99
40	0	12	1	22	1	34	0.3529	0.2284	-0.738	-11.999	264.03
41	0	12	1	21	1	33	0.3636	0.2314	-0.756	-11.999	251.99
47	1	12	0	20	1	32	0.3750	0.2344	1.291	20.00	240.03
48	0	11	1	20	1	31	0.3548	0.2289	-0.742	-10.999	219.97
50	0	11	1	19	1	30	0.3667	0.2322	-0.761	-11.001	208.98
59	0	11	1	18	1	29	0.3793	0.2354	-0.782	-11.00	197.97
61	0	11	1	17	1	28	0.3929	0.2385	-0.805	-11.001	186.98
68	0	11	1	16	1	27	0.4074	0.2414	-0.829	-11.00	175.98
69	1	11	0	15	1	26	0.4231	0.2441	1.168	14.999	165.01
71	0	9	1	15	1	24	0.3750	0.2344	-0.775	-9.00	135.01
113	0	6	1	10	1	16	0.3750	0.2344	-0.775	-6.00	60.006
118	0	6	1	8	1	14	0.4286	0.2449	-0.866	-6.0004	48.00
143	0	6	1	7	1	13	0.4615	0.2485	-0.926	-5.9995	41.997
148	1	6	0	6	1	12	0.5000	0.2500	1.00	6.00	36.00
181	1	5	0	4	1	9	0.5556	0.2469	0.894	3.9996	19.999
Total	5						9.5652	5.9289	-9.324	-159	6048.01
							$\sum_{i=1}^r e_{i1}$	V_L	U_w		

Table 4.5. Comparison of two-sample tests, breast cancer data.

Test	$W(t_i)$	$Z_1(T)$	σ_{11}^2	χ^2	p - value
Log-Rank	1.0	-4.565	5.929	3.515	0.061
Wilcoxon(Gehan)	Y_i	-159.0	6048.135	4.180	0.041
Tarone-Ware	$Y_i^{1/2}$	-26.922	178.865	4.052	0.044
Peto-Peto	$\tilde{S}(t_i)$	-3.509	2.992	4.115	0.042
Modified Peto-Peto	$\tilde{S}(t_i)Y_i/(Y_i + 1)$	- 3.412	2.811	4.141	0.041
Fleming-Harrington $p = 0, q = 1$	$[1 - \hat{S}(t_{i-1})]$	-0.962	0.687	1.347	0.246
Fleming-Harrington $p = 1, q = 0$	$\hat{S}(t_{i-1})$	-3.603	3.155	4.115	0.042
Fleming-Harrington $p = 1, q = 1$	$\hat{S}(t_{i-1})[1 - \hat{S}(t_{i-1})]$	-0.7099	0.218	2.312	0.129
Fleming-Harrington $p = 0.5, q = 0.5$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^{0.5}$	-1.677	1.043	2.696	0.101
Fleming-Harrington $p = 0.5, q = 2$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^2$	-0.227	0.074	0.696	0.407

From the above comparisons, we can see that some of the tests are statistically significant but some others are not, depending on the weight functions. All tests whose weight functions are monotonically nonincreasing turn out to be statistically significant at the 0.05 level; all other tests are nonsignificant at the 0.05 level. This means that group differences of hazard rates occurred early in the survival experience.

An important consideration in applying the tests is the choice of weight function to be used. A variety of weight functions have been proposed in the literature. Different weights allow us to present tests that can be more sensitive to early or late departures from the hypothesized relationship between samples as specified by the null hypothesis. In some applications, other weight functions $W(\cdot)$ may be more appropriate based on the investigator's desire to emphasize either late or early departures between hazard rates. For the two-sample tests, the log-rank weights have optimal local power to detect differences in the hazard rates when the hazard rates are proportional.

4.3.5 Comparison of the log-rank and Wilcoxon tests

Of the two tests, the log-rank test is more suitable when the alternative to the null hypothesis of no difference between the survivor functions of the two groups is that the hazard rate at any given time in one group is proportional to the hazard rate at that time in the other group. This is the assumption of proportional hazards, which underlies a number of methods for analyzing survival data. For other types of departure from the null hypothesis, the Wilcoxon test is more appropriate than the log-rank test for comparing two survivor functions.

In order to help decide which test is more suitable in any given situation, we make use of the result that if the hazard functions are proportional, the survivor functions for the two groups of survival data do not cross one another. To show this, suppose that $h_1(t)$ is the hazard of death at time t for an individual in Group I, and $h_2(t)$ is the

hazard at that same time for an individual in Group II. If these two hazards are proportional, then we can write $h_1(t) = \delta h_2(t)$, where, $\delta = \exp \theta \neq 1$ is a constant that does not depend on the time t . Integrating both sides of this expression, multiplying by -1 and exponentiating gives

$$e^{-\int_0^t h_1(u) du} = e^{-\int_0^t \delta h_2(u) du} \quad (4.3)$$

Now, we know, $S(t) = e^{-\int_0^t h(u) du}$, so if $S_1(t)$ and $S_2(t)$ are the survivor functions of the two groups, then from equation (4.3), we have

$$S_1(t) = [S_2(t)]^\delta.$$

Since the survivor function takes values between zero and unity, this result shows that $S_1(t)$ is greater than or less than $S_2(t)$, according to whether δ is less than or greater than unity, at any time t . Thus, if two hazard functions are proportional, their survivor functions do not cross. This situation is shown in the following figure. This is a necessary, but not a sufficient condition for proportional hazards.

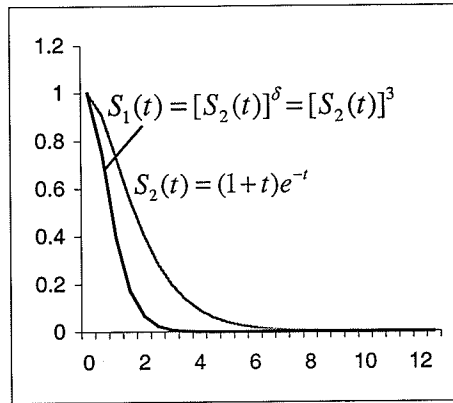
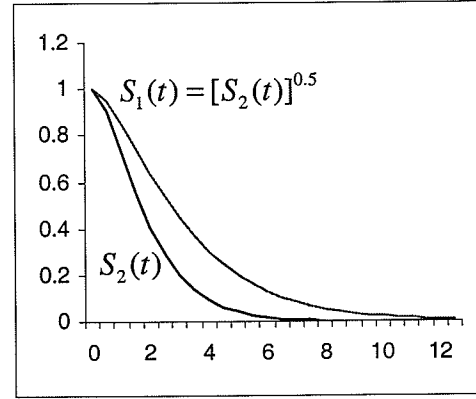


Figure 4.1(a): Survivor functions when $\delta > 1$.



(b): Survivor functions when $\delta < 1$.

Specifically, a test will be developed for the hypothesis

$$H_0 : S_1(t) = S_2(t)$$

that the two distributions are the same. This can be handled by treating observations from both distributions as coming from a single population and by defining a dummy regressor variable x that takes on the value 0 or 1 according to whether an observation comes from the first or second distribution. The hazard functions for the two distributions are then $h_2(t) = h_0(t)$ and $h_1(t) = h_0(t)e^\theta$, where $h_0(t)$ is the baseline hazard function. These are called the proportional hazards alternatives and the two distributions are identical if and only if $\theta = 0$. This is equivalent to assuming that $S_1(t)$ and $S_2(t)$ are related by

$$S_1(t) = [S_2(t)]^{\exp \theta}.$$

so that by testing that, $\theta = 0$, one is testing the hypothesis

$$H_0 : S_1(t) = S_2(t) \quad \text{vs.} \quad H_a : S_1(t) = [S_2(t)]^\delta \quad (4.4)$$

where $\delta \neq 1$.

The family of hypotheses represented by H_a in (4.4) is sometimes referred to as the Lehmann family of alternatives; it specifies that the two distributions have proportional hazard functions. The test of H_0 obtained below should consequently be powerful against alternatives in H_a .

Suppose that random observations are taken on Y_{i1} individuals from the first population having a survivor function $S_1(t)$ and on Y_{i2} individuals from the second

population having a survivor function $S_2(t)$. Some observations may be censored. We regard the two samples together as a single sample. Then the survivor function for T , given x , can be written as

$$S(t|x) = \exp\left(-\int_0^t h(u|x) du\right) \\ = [S_0(t)]^{\exp(x\theta)}$$

where the dummy variable x signifies whether an observation is from that first ($x=1$) or second ($x=0$) population. Let $t_1 < t_2 < \dots < t_r$ denote the r distinct observed lifetimes in the combined sample and let d_i represent the number of deaths at t_i . Further, let Y_{i1} and Y_{i2} be the numbers of individuals in the risk set R_i at t_i who are from the first and second populations, respectively, and let d_{i1} and d_{i2} denote the number of deaths at t_i among individuals from populations I and II. $Y_{i1} + Y_{i2} = Y_i$ and $d_{i1} + d_{i2} = d_i$. Under this setup, the log likelihood for θ becomes

$$\ln L(\theta) = D_1\theta - \sum_{i=1}^r d_i \ln(Y_{i1} + Y_{i2}e^\theta) \quad [\text{Lawless (1982), p. 349}]$$

where, $D_1 = \sum_{i=1}^r d_{i1}$ is the total number of deaths among individuals from population I.

The score function and the negative of the second derivative of $\ln L(\theta)$ are

$$U(\theta) = \frac{\partial \ln L}{\partial \theta} = D_1 - \sum_{i=1}^r \frac{d_i Y_{i1} e^\theta}{Y_{i1} e^\theta + Y_{i2}} \quad (4.5)$$

and

$$I(\theta) = -\frac{\partial^2 \ln L}{\partial \theta^2} = \sum_{i=1}^r \frac{d_i Y_{i1} Y_{i2} e^\theta}{(Y_{i1} e^\theta + Y_{i2})^2} \quad (4.6)$$

The likelihood equation $U(\theta)=0$ is readily solved by iteration to get the MLE $\hat{\theta}$.

Inferences for θ can be obtained by the likelihood ratio method, by treating $\hat{\theta}$ as being approximately $N(\theta, I^{-1}(\hat{\theta}))$, or by treating $U(\theta)$ as $N(0, I(\theta))$.

The score function provides a particularly simple test of the equality of $S_1(t)$ and $S_2(t)$ without the need to compute $\hat{\theta}$. In particular, under $H_0 : \theta = 0$, the statistic

$$Z = \frac{U(0)}{[I(0)]^{1/2}} \quad (4.7)$$

is approximately distributed as $N(0,1)$. Large absolute values of Z provide evidence against equality of the two distributions. From (4.5) and (4.6) the expressions for $U(0)$ and $I(0)$ are

$$U(0) = D_1 - \sum_{i=1}^r \frac{d_i Y_{i1}}{Y_i}$$

$$I(0) = \sum_{i=1}^r \frac{d_i Y_{i1} Y_{i2}}{Y_i^2}.$$

4.4 Tests for more than two samples

The comparison of k ($k > 2$) samples has been considered by many statisticians; for example, Kruskal and Wallis (1952), Mantel and Haenszel (1959), Breslow (1970) and Peto and Peto (1972). k -sample problems are a generalization of two-sample problems. For k samples, we test whether the k independent samples can be regarded as coming from the same population; i.e., we test the hypothesis set

$$H_0: S_1(t) = S_2(t) = \dots = S_k(t) \text{ for all } t$$

against $H_a: S_i(t) \neq S_j(t) \text{ for some } i \neq j.$

We discuss these tests in the context of a data set on 137 bone marrow transplant patients. The data are included in the Appendix (Table 5). Patients were grouped into risk categories based on their status at the time of transplantation. The three groups are as follows:

Group 1, consisting of 38 ALL patients; Group 2, consisting of 54 AML low-risk patients; and Group 3, consisting of 45 AML high-risk patients.

Without any truncation, we perform the Kruskal –Wallis test and an extension of the Median test. We define, for $i = 1, 2, 3$

$n_i = \text{ith sample size};$

$$N = \sum_{i=1}^3 n_i ;$$

$R_i = \text{sum of the ranks for the } \text{ith} \text{ sample in the combined ranking of all samples.}$

The Kruskal–Wallis statistic H is given as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^3 \frac{R_i^2}{n_i} - 3(N+1)$$

For moderately large samples, the null distribution of the statistic H is approximately chi-square distribution with $(k - 1)$ degrees of freedom.

For the given data,

$$\begin{aligned} H &= \frac{12}{137(137+1)} \left(\frac{2403^2}{38} + \frac{4572.5^2}{54} + \frac{2477.5^2}{45} \right) - 3(137+1) \\ &= 14.777 \end{aligned}$$

At the 5% level, the critical value of a chi-square variable with 2 d.f. is 5.99.

We therefore conclude that k survivor functions are different at the 5% significance level.

Next, we want to know which population(s) differ(s) from which others. Several methods of simultaneous multiple comparisons between the locations of all pairs of the k populations have been proposed. We include here a procedure discussed by Dunn (1964). It gives conservative results but is the simplest method available.

The treatment pairs that do not differ significantly are those for which the sample means $(\bar{R}_i, \bar{R}_j)$ satisfy the inequality

$$|\bar{R}_i - \bar{R}_j| \leq z_{\left(1 - \frac{\alpha}{k(k-1)}\right)} \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

where $\bar{R}_i = \frac{R_i}{n_i}$.

A systematic method of generating all differences $|\bar{R}_i - \bar{R}_j|$ is shown in the following table.

Group i \ Group j	$\bar{R}_i$	1	2	3
		$ \bar{R}_i - 63.24 $	$ \bar{R}_i - 84.68 $	$ \bar{R}_i - 55.06 $
1	63.24	0		
2	84.68	21.44	0	
3	55.06	8.18	29.62	0

With $k = 3$ and $\alpha = 0.05$, we get the value of

$$z_{\left(1-\frac{\alpha}{k(k-1)}\right)} \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = 2.394 \sqrt{\frac{137(137+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

for 1st and 2nd sample as 20.12, for 1st and 3rd sample as 20.94 and for 2nd and 3rd sample as 19.18. Comparing these values with the differences $|\bar{R}_i - \bar{R}_j|$, we find that group 2 differs significantly from both groups 1 and 3. We conclude that the survivor function for group 2 differs from the survivor functions of groups 1 and 3.

For the extension of the Median test, we calculate the median of the three samples combined. This pooled median is $M = 538$. We classify observations according to whether they are less than or greater than or equal to the pooled median. We then construct the following contingency table:

Sample	< M	≥ M	Total
1	23 (18.86)	15 (19.139)	38
2	15 (26.803)	39 (27.197)	54
3	30 (22.336)	15 (22.664)	45
Total	68	69	137

The numbers in brackets represent the expected values under the null hypothesis.

The test statistic is $\chi^2 = \frac{(O - E)^2}{E}$ with O = observed frequency and E = expected frequency. This test statistic has approximately a chi-square distribution with 2 d.f. The calculated value of chi-square is 17.345. Since this observed value is greater than the critical value 5.991, we again reject the null hypothesis of the equality of survivor functions.

Using the $\sin^{-1}$ transformation of the binomial proportions suggested by Anscombe (1948), we construct the following table.

Sample	Sample size n_i	n_{i1}	U_i	Reciprocal of Variance $4n_i + 2$	$(4n_i + 2)U_i$	$(4n_i + 2)U_i^2$
1	38	23	0.889	154	136.963	121.8112
2	54	15	0.559	218	121.756	68.0024
3	45	30	0.952	182	173.341	165.0935
Total	137			554 = V	432.06 = T_1	354.907 = T_2

Here,

$$U_i = \sin^{-1} \sqrt{\frac{n_{i1} + \frac{3}{8}}{n_i + \frac{3}{4}}}$$

The test statistic $\chi^2 = T_2 - \frac{T_1^2}{V}$ has approximately a chi-square distribution with 2 d.f.

From the data,

$$\begin{aligned}\chi^2 &= 354.907 - \frac{(432.06)^2}{554} \\ &= 17.947\end{aligned}$$

Since, this calculated value is greater than the critical value 5.991, we again reject the null hypothesis of the equality of survivor functions.

We have used the Median test to test the equality of three population medians. We have also used Anscombe's $\sin^{-1}$ transformation of the binomial proportions in this situation, because in the Median test we have in fact tested the equality of three binomial proportions. Each binomial proportion is the population proportion of observations that are less than the common population median under the null

hypothesis. Using truncation, we can also test the equality of other types of proportions created in any arbitrary manner.

Next, we truncate the observations for each group of patients for time periods over 1500.

We define, for $i = 1, 2, 3$:

$p_i = \text{Pr}(\text{the lifetime of an individual chosen at random from the } i\text{th population is less than 1500}).$

Our hypothesis becomes

$$H_0: p_1 = p_2 = p_3$$

against $H_a: p_i \neq p_j$ for some $i \neq j$.

We construct the following table using the notations given below.

Sample	Sample size Y_i	# of Death D_i	U_i	Reciprocal of Variance $4Y_i+2$	$(4Y_i+2)U_i$	$(4Y_i+2)U_i^2$
1	38	36	1.321	154	203.376	268.584
2	54	36	0.953	218	207.732	197.947
3	45	38	1.158	182	210.691	243.905
Total	137	110		$554 = V$	$621.799 = T_1$	$710.436 = T_2$

Here, $U_i = \sin^{-1} \sqrt{\frac{D_i + \frac{3}{8}}{Y_i + \frac{3}{4}}}$, $i = 1, 2, 3$. Now,

$$T_1 = \sum_{i=1}^3 (4Y_i + 2)U_i \text{ and } T_2 = \sum_{i=1}^3 (4Y_i + 2)U_i^2, \text{ so that}$$

$$\chi^2 = T_2 - \frac{T_1^2}{V}.$$

The value of this chi-square is 12.5406. This result also leads to rejection of the above null hypothesis.

With the truncated observations, we calculate the value of the chi-square statistic using the following contingency table. The value of the corresponding test statistic is 11.84 and we again reject the above null hypothesis.

Sample	Sample size	Number of Deaths D_i	Number Surviving
1	38	36 (30.51)	2 (7.49)
2	54	36 (43.36)	18 (10.64)
3	45	38 (36.13)	7 (8.87)
Total	137	110	27

The numbers in brackets represent the expected values under the null hypothesis.

Now, consider another example of a randomized double-blind clinical trial comparing four headache relief medications (active compound # 1, active compound # 2, rival product, and control). We want to test whether or not there are statistically significant differences among the four groups. The data set are included in the Appendix (Table 6).

First, we calculate the Kruskal –Wallis statistic H as

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \sum_{i=1}^4 \frac{R_i^2}{n_i} - 3(N+1) \\
 &= \frac{12}{65(65+1)} \left(\frac{324.5^2}{10} + \frac{513.5^2}{12} + \frac{464.5^2}{14} + \frac{777.5^2}{29} \right) - 3(65+1) \\
 &= 10.88
 \end{aligned}$$

The critical value of chi-square with 3 d.f. at the 5% level of significance is 7.8147.

We conclude that there is a significant difference among the four medications on their effect of treating headache at the 5% level of significance.

For this data set, we have one control group and three different treatment groups.

Now, we are interested in which treatment groups differ significantly from the control group. So, a total of $k - 1 = 3$ comparisons would be made with $p = 3$ sets of contrasts.

Treatment Group	Sample size (n)	R_i	$\bar{R}_i$
Active compound # 1	10	324.5	32.45
Active compound # 2	12	578.5	48.21
Rival product	14	464.5	33.18
Control	29	777.5	26.81

The treatment pairs that do not differ significantly are those for which the sample means $(\bar{R}_i, \bar{R}_j)$ satisfy the inequality

$$|\bar{R}_i - \bar{R}_j| \leq z \sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (4.8)$$

The critical value of z in (4.8) for $p = 3$ contrasts with overall significance level α is always that quantile of the standard normal curve corresponding to the right-tail probability $\alpha/2p$. With $\alpha = 0.05$, the value of z is 2.394.

Contrast	Right-hand side of (4.8)	Observed mean difference $ \bar{R}_i - \bar{R}_j $
Active #1 vs. Control	16.5995	5.64
Active #2 vs. Control	15.5369	21.40
Rival vs. Control	14.7310	6.37

We find that only the second contrast is statistically significant which implies that there is a significant difference between active compound #2 and the control medication. But there is no significance difference between active compound #1 and the control medication. Similarly, there is no difference between rival product and the control medication.

Following Puri (1964), we perform another test. We transform the data by replacing each rank in the combined sample by the expected value of the observation having that rank in a random sample of N observations from a standard normal distribution. For this transformation we have used a special table given by Harter (1961). In the Appendix (Table 7) we have included the expected values for $N = 65$.

In table 4.6, we have calculated the ranks and their corresponding expected order statistics for the observations of four groups. Define,

$$\mu = \text{average of all expected values} = 0$$

$$\sigma^2 = \text{variance of all expected values}$$

$$= \frac{1}{N} (\text{sum of squares of all expected values})$$

$$= \frac{65.36}{65} = 0.959$$

T_{Nj} = average of the expected values of order statistics for the j th group,

$$j = 1, 2, 3, 4.$$

Table 4.6. Ranks and their corresponding expected order statistics.

Active compound # 1		Active compound # 2		Rival product		Control	
Rank	Expected order statistics	Rank	Expected order statistics	Rank	Expected order statistics	Rank	Expected order statistics
5	-1.47241	21	-0.47894	8	-1.19252	2	-2.024287
14	-0.81123	24	-0.35312	12	-0.92329	2	-2.024287
23	-0.39435	35	0.07694	16	-0.70856	2	-2.024287
27.5	-0.21303	40	0.27237	17	-0.65996	4	-1.59732
34	0.03844	46	0.52252	22	-0.43625	6	-1.36698
38	0.19337	50	0.70856	26	-0.27237	7	-1.2749
39	0.23269	53	0.8659	27.5	-0.21303	9	-1.11754
41	0.31249	55.5	1.01615	32	-0.03844	10	-1.04838
49	0.65996	62	1.59732	43	0.39435	11	-0.98391
54	0.92329	63	1.75363	44	0.43625	13	-0.8659
		64	1.96965	45	0.47894	15	-0.75889
		65	2.34958	52	0.81123	18	-0.61288
				59	1.2749	19.5	-0.54482
				61	1.4724	19.5	-0.54482
						25	-0.31249
						29	-0.15435
						30	-0.11556
						31	-0.07694
						33	0.0
						36	0.11556
						37	0.15435
						42	0.35312
						47	0.56712
						48	0.61288
						51	0.75889
						55.5	1.01615
						57	1.11754
						58	1.19252
						60	1.36698
324.5	-0.53078	578.5	10.30056	464.5	0.42366	777.5	-10.193436

It is known that

$$E(T_{Nj}) = 0 \text{ and } Var(T_{Nj}) = \frac{\sigma^2(N-n_j)}{n_j(N-1)}, \quad j = 1, 2, 3, 4.$$

$$\therefore Z_j = \frac{T_{Nj} - 0}{\sqrt{\frac{\sigma^2(N-n_j)}{n_j(N-1)}}}$$

Puri's test statistic is defined as

$$\begin{aligned} \mathcal{L} &= \sum_{j=1}^4 \frac{(N-n_j)}{N} Z_j^2 \\ &= \sum_{j=1}^4 \frac{(N-n_j)}{N} \frac{T_{Nj}^2}{\frac{\sigma^2(N-n_j)}{n_j(N-1)}} \\ &\equiv \sum_{j=1}^4 n_j \frac{T_{Nj}^2}{\sigma^2} \end{aligned}$$

The calculation of the statistic $\mathcal{L}$ is shown below.

n_j	T_{Nj}	$Var(T_{Nj})$	Z_j	$n_j T_{Nj} / \sigma^2$
10	-0.05308	0.08241	-0.18490	0.02938
12	0.85838	0.06618	3.33667	9.21981
14	0.03026	0.05459	0.12952	0.01337
29	-0.35150	0.01860	-2.57723	3.73620
Total				12.9988 = $\mathcal{L}$

Puri (1964) has shown that the distribution of the test statistic $\mathcal{L}$ under the null hypothesis of identical distributions tends to approach the chi-square distribution with $(k-1)$ d.f. as the sample sizes become large.

The observed value of $\mathcal{L}$ is greater than the tabulated value of chi-square with 3 d.f. at 5% level of significance. We conclude that there is a difference among the four medications on their effect of treating a headache.

Here, we can also test which treatment groups differ significantly from the control group. The treatment pairs that do not differ significantly are those for which the averages of the expected values of order statistics (T_{Ni}, T_{Nj}) satisfy the inequality

$$|T_{Ni} - T_{Nj}| \leq z\sigma \sqrt{\frac{N}{N-1} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (4.9)$$

where T_{Ni} is the average of the expected values of order statistics for the control group. So, a total of only $k-1=3$ comparisons would be made. The critical value of z in (4.9) for $p=3$ sets of contrasts with overall significance level α is always that quantile of the standard normal curve that corresponds to a right-tail probability of $\alpha/2p$. With $\alpha=0.05$, the value of z is 2.394.

Contrast	Right-hand side of (4.9)	Observed mean difference $ T_{Ni} - T_{Nj} $
Active #1 vs. Control	0.8664	0.29842
Active #2 vs. Control	0.8110	1.20988
Rival vs. Control	0.7689	0.38176

We find that only the second contrast is statistically significant, which implies that there is a significant difference between active compound #2 and the control medication. But there is no difference between active compound #1 and the control medication, nor is there a difference between the rival product and control medication. This test produces the same conclusions as the test of comparing all treatments with a control using the ranks.

The Kruskal-Wallis test is a natural extension of the Wilcoxon rank sum test for two samples. Similarly, Puri's normal-scores test is a natural extension of the Terry-Hoeffding test for two samples.

It is known that the asymptotic efficiency of Puri's test relative to the parametric F -test for normal samples is 1 and the asymptotic efficiency of the Kruskal-Wallis test relative to the F -test for normal samples is $3/\pi$. Puri's test is as efficient as the parametric F -test and more efficient than the Kruskal-Wallis test.

CHAPTER FIVE

CONCLUSIONS

In survival analysis, parametric procedures are more efficient when the underlying distributional assumptions are valid. However, when we use the parametric approach, we end up having non-linear estimating functions with no closed form expressions available, in general, for the estimates. For certain inference problems this difficulty is successfully overcome by using nonparametric inference procedures.

In this thesis, nonparametric estimation techniques are discussed and compared for different nonparametric estimators (e.g., K-M estimator, N-A estimator, modified N-A estimator) and illustrated with relevant data sets. We have also discussed a nonparametric Bayesian method. We have found that the modified N-A estimator performs better than any other estimators for censored correlated data in the sense that this estimator minimizes the sum of squares of forecast errors.

Nonparametric tests are specially useful when:

- the sample sizes are quite small.
- data are measured on an ordinal scale.
- sample medians seem to be more reliable than sample means.
- little or nothing can be assumed about the probability distribution.

- homogeneity of the variances seems highly questionable.

Their power is more robust against gross errors than that of the t - and F -tests and their efficiency loss is quite small even in the rare case in which the suspicion of the possibility of gross errors is unfounded. For k -sample problems, the asymptotic relative efficiency of each of the discussed nonparametric tests in comparison to the corresponding parametric test is always greater than or equal to 0.864 and can be large. The only assumption required for the validity of these tests is that the unknown common probability distribution is continuous.

The Median test uses less information than any of the others and is therefore expected to be less powerful. Its asymptotic relative efficiency to the F -test in the case of normal populations with equal variance is $2/\pi = 0.637$. Anscombe's $\sin^{-1}$ transformation of the binomial proportions can also be used in conjunction with the Median test. The Kruskal-Wallis test is a nonparametric analogue of the one-way analysis of variance test (F -test). For normal samples, this test has an ARE of $3/\pi = 0.955$. Puri's normal score test is an improvement of the Kruskal-Wallis test and as efficient as the parametric F -test. All the tests are much quicker and easier to apply than the F -test and may perform better than the F -test if the normality assumption is not satisfied.

This thesis is mainly concerned with applications of nonparametric estimation and hypothesis testing procedures dealing with survival analysis. These procedures include both well-known methods and some recently developed techniques. We have

focused on the applications and comparisons of these techniques through several data sets. The inference for censored correlated data can be extended in future research.

APPENDIX

Table 1. Data List of Monthly Dry Deposition of NH_4 as Measured at Lawrence Livermore, California.

	1977	1978	1979	1980
January		NA	34	117
February		53	128	99
March		546	79	< 41
April		332	< 31	29
May	27	51	174	138
June	< 16	203	< 36	168
July	< 19	171	< 42	107
August	796	321	169	253
September	268	402	223	445
October	97	460	750	446
November	356	2080	NA	1260
December	94	487	NA	

Note: NA = not available.

Table 2. Data list of 50 observations from an AR(1) process with mean 0, white noise variance 1, and first lag correlation, 0.5; left censored at -1.2 .

0.3	2.02	-0.2	-0.39	0.75
-0.19	0.74	0.18	-0.28	1.18
-0.75	-0.29	-0.17	-0.58	-0.2
-1	-1.08	-0.15	-1.05	-0.04
<-1.2	<-1.2	1.4	<-1.2	<-1.2
<-1.2	-0.94	1.68	<-1.2	-0.17
0.54	<-1.2	0	<-1.2	-0.59
0.64	-0.1	0.73	-0.73	-0.57
0.42	1.29	2.55	0.23	-0.48
0.06	2.01	0.8	1.43	-1.15

Table 3. Survival Functions of U.S. Population By Race and Sex in 1989.

Age	White Male	White Female	Age	White Male	White Female
0	1.00000	1.00000	43	0.93771	0.97016
1	0.99092	0.99285	44	0.93477	0.96862
2	0.99024	0.99232	45	0.93161	0.96694
3	0.98975	0.99192	46	0.92820	0.96511
4	0.98937	0.99160	47	0.92450	0.96311
5	0.98905	0.99134	48	0.92050	0.96091
6	0.98898	0.99111	49	0.91617	0.95847
7	0.98850	0.99091	50	0.91148	0.95575
8	0.98825	0.99073	51	0.90639	0.95273
9	0.98802	0.99056	52	0.90086	0.94938
10	0.98782	0.99041	53	0.89480	0.94568
11	0.98765	0.99028	54	0.88810	0.94161
12	0.98748	0.99015	55	0.88068	0.93713
13	0.98724	0.98999	56	0.87250	0.93222
14	0.98686	0.98977	57	0.86352	0.92684
15	0.98628	0.98948	58	0.85370	0.92096
16	0.98547	0.98909	59	0.84299	0.91455
17	0.98445	0.98862	60	0.83135	0.91756
18	0.98326	0.98809	61	0.81275	0.89995
19	0.98197	0.98755	62	0.80511	0.89169
20	0.98063	0.98703	63	0.79052	0.88275
21	0.97924	0.98654	64	0.77501	0.87312
22	0.97780	0.98607	65	0.75860	0.86278
23	0.97633	0.98561	66	0.74131	0.85169
24	0.97483	0.98514	67	0.72309	0.83980
25	0.97332	0.98466	68	0.70383	0.82702
26	0.97181	0.98416	69	0.68339	0.81324
27	0.97029	0.98365	70	0.66166	0.79839
28	0.96876	0.98312	71	0.63865	0.78420
29	0.96719	0.98257	72	0.61441	0.76522
30	0.96557	0.98199	73	0.58897	0.74682
31	0.96390	0.98138	74	0.56238	0.72716
32	0.96217	0.98073	75	0.53470	0.70619
33	0.96038	0.98005	76	0.50601	0.68387
34	0.95852	0.97933	77	0.47641	0.66014
35	0.95659	0.97858	78	0.44604	0.63494
36	0.95457	0.97779	79	0.41503	0.60822
37	0.95245	0.97696	80	0.38355	0.57991
38	0.95024	0.97607	81	0.35178	0.54997
39	0.94794	0.97510	82	0.31991	0.51835
40	0.94555	0.97404	83	0.28816	0.48502
41	0.94307	0.97287	84	0.25677	0.44993
42	0.94047	0.97158	85	0.22599	0.41306

Table 4. Survival times of women grouped according to whether a section of the tumour was positively or negatively stained.

Negative staining	Positive staining	
23	5	68
47	8	71
69	10	76*
70*	13	105*
71*	18	107*
100*	24	109*
101*	26	113
148	26	116*
181	31	118
198*	35	143
208*	40	154*
212*	41	162*
224*	48	188*
	50	212*
	59	217*
	61	225*

Table 5. Data on 137 Bone Marrow Transplant Patients.

T for Group-1	T for Group-2	T for Group-2	T for Group-3
2081	2569	53	2640
1602	2506	80	2430
1496	2409	35	2252
1462	2218	1499	2140
1433	1857	704	2133
1377	1829	653	1238
1330	1562	222	1631
996	1470	1356	2024
226	1363	53	1345
1199	1030		1136
1111	860		845
530	1258		491
1182	2246		162
1167	1870		1298
418	1799		121
417	1709		2
276	1674		62
156	1568		265
781	1527		547
172	1324		341
487	957		318
716	932		195
194	847		469
371	848		93
526	1850		515
122	1843		183
1279	1535		105
110	1447		128
243	1384		164
86	414		129
466	2204		122
262	1063		80
162	481		677
262	105		73
1	641		168
107	390		74
269	288		16
350	522		248
	79		732
	1156		105
	538		392
	48		63
	431		97
	1074		153
	393		363
	10		

T – Time (in Days) to
death or on study time

Disease Group:
Group 1- ALL
Group 2 – ALL low risk
Group 3 - AML high risk

Table 6. Clinical trial for comparing four headache relief medications.

Active 1	Active 2	Rival	Control
16.1	31.9	11.6	7.8
15.1	26.4	3.3	9.7
8.2	27.7	9	20.7
1.3	18	24.7	10.8
10.3	21.7	8.8	13.1
22.3	28.1	17.9	12
11.6	15.5	21.6	9.7
19.4	14.6	8	5.3
14.4	19.5	13.3	14.7
15	22.9	11	8.1
	10.6	10.2	23.9
	10	17.7	0.5
		17.2	0.5
		25.2	14.9
			22.9
			0.5
			25.1
			1
			5.8
			2.6
			13.6
			9.3
			19.1
			24.5
			16.3
			1.5
			11.7
			18.4
			8.4

Table 7. Expected values of normal order statistics.

k	$n = 65$
1	2.34958
2	1.96965
3	1.75363
4	1.59732
5	1.47241
6	1.36698
7	1.2749
8	1.19252
9	1.11754
10	1.04838
11	0.98391
12	0.92329
13	0.8659
14	0.81123
15	0.75889
16	0.70856
17	0.65996
18	0.61288
19	0.56712
20	0.52252
21	0.47894
22	0.43625
23	0.39435
24	0.35312
25	0.31249
26	0.27237
27	0.23269
28	0.19337
29	0.15435
30	0.11556
31	0.07694
32	0.03844
33	0.00000

Source: Harter (1961), p.160.

Note: Tabular values are the expected values of the k th largest normal variate for a random sample of size $n = 65$ from $N(0,1)$; when preceded by a minus sign, they are expected values of the k th smallest normal variate.

“R” code for censoring example:

Have left endpoints of intervals in a vector.

“lefts” and right endpoints in “rights” and

set “N” equal to the common length.

```
lefts <- c(45, 6, 0, 46, 46, 7, 17, 7, 37, 0, 4, 15, 11, 22, 46, 46, 25, 46, 26, 46, 27, 36,
46, 36, 37, 40, 17, 46, 11, 38, 5, 37, 0, 18, 24, 36, 5, 19, 17, 24, 32, 33, 19, 37, 34, 36)
```

```
rights <- c(Inf, 10, 7, Inf, Inf, 16, Inf, 14, 44, 8, 11, Inf, 15, Inf, Inf, Inf, 37, Inf, 40,
Inf, 34, 44, Inf, 48, Inf, Inf, 25, Inf, 18, Inf, 12, Inf, 5, Inf, Inf, Inf, 11, 35, 25, Inf, Inf,
Inf, 26, Inf, Inf, Inf)
```

```
npcens <- function(lefts, rights, alpha = 1, K = 100)
```

```
{
  if(length(lefts) != length(rights))
    stop(“Lengths of left and right endpoints differ”)
```

```
  N <- length(lefts)
```

```
  # Partition support (0, infty) into sets B.
```

```
  grid <- c( seq(0,100,by = 1), 1000 )
```

```
  ngrid <- length(grid)
```

```
  grt <- grid[2:ngrid]; glt <- grid[1:(ngrid-1)]
```

```
  cdfs <- matrix(0, K, length(grid))
```

```
  # Exponential prior guess and prior sample size
```

```
  gg <- exp( -glt/(365/2) ) - exp( -grt/(365/2) )
```

```
  g0 <- gg
```

```
  # alpha <- 1
```

```
  # Identify partition elements that may contain each survival time.
```

```
  less <- function(x,y){ return( x < y ) }
```

```
  ma <- t( outer(grt, rights, FUN = “less”) )
```

```
  mb <- outer(lefts, glt, FUN= “less”)
```

```
  dmat <- ma & mb
```

```

# Recursion yields approximate Bayes estimate gg
weight <- 1/(alpha + (1:N)) # Nominal weight sequence
for( j in 1:K )
{
# weight <- 1/(alpha+1:N) # Nominal weight sequence
ord <- sample( 1:N ) # Permute censoring intervals randomly
for( i in 1:N )
{
ok <- dmat[ord[i],] # A_i
numer <- rep(0, ngrid-1)
numer[ok] <- gg[ok] # G(B and A_i)
denom <- sum( numer ) # G(A_i)
gg <- gg*( 1-weight[i] ) + weight[i]*numer/denom
}
cdfs[j,] <- c(0, cumsum(gg))
}
# Repeat loop to see variation over orderings.
# cdf2 <- cbind(rep(0, K), cdfs)
gnavg <- apply(cdfs, 2, mean) # produces average cdf
# return(dim(cdfs), dim(cdf2), length(gnavg), length(grid))
# Approximate Bayes estimate of distribution function.
par(mfrow = (c(1, 2)))
matplot(grid, t(cdfs), type = "s", xlim = c(0,100), col = 1, lty = 1,
xlab = "time (days)", ylab = "cumulative probability")
lines(grid, c(0, cumsum(g0)), lty = 2)
plot(grid, gnavg, type = "s", xlim = c(0,100),
xlab = "time (days)", ylab = "cumulative probability")
lines(grid, c(0, cumsum(g0)), lty = 2)
}
# Standard call: npcens(lefts, rights)
# with alpha=5: npcens(lefts, rights, alpha = 5)

```

“R” code for Empirical distribution function example:

```
t<- c(5, 17, 21, 12, 45, 17, 28, 23, 27, 54, 39, 15, 57, 68, 74, 86, 32)
brks <- c(seq(0, 65, 5), 100)
# mycdf <- function(t, brks)
{
# EDF function for interval survival time.
# t is the vector of event times.
# brks is the vector of end points (including first left and first right end point).
# if (brks[1] > y[1] | brks[length(brks)] < y[length(y)])
# stop("Data out of range")
y <- sort(t)
z<- cut(y,breaks=brks, labels=NULL, include.lowest=FALSE,right=TRUE,dig.lab=3)
tab <- table(z)
edf <- cumsum(tab)/length(y)
plot (brks, c(0, edf), type = "s", xlab = "Survival time", ylab = expression (F[n](t)))
abline (0, 0, lty = 2)
abline (1, 0, lty = 2)
mat <- cbind(table = tab, edf = edf)
rownames(mat) <- names(tab)
mat
}
{
# Suppose (n +1)th observation is 40.
# return(edf)
n <- length(t)
weight <- 1/(n+1)
brks2 <- brks[2:length(brks)]
I <- as.numeric(40 <= brks2)
edf2 <- ((1-weight)*edf) + (weight*I)
edf2
}
```

REFERENCES:

1. Alavi, A. and Thavaneswaran, A. (2002). "Nonparametric Estimators For Censored Correlated Data." *Communication in Statistics - Theory and Method*, Vol. 31, p. 977-985.
2. Anscombe, F. J. (1948). "The Transformation of Poisson, Binomial and Negative-Binomial Data." *Biometrika*, Vol. 35, p. 246-254.
3. Cai, Z. (1998). "Asymptotic Properties of Kaplan-Meier Estimator for Censored Dependent Data." *Statistics and Probability Letters*, Vol. 37, p. 381-389.
4. Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London, UK.
5. Daniel, W. W. (1990), *Applied Nonparametric Statistics*. 2nd ed., PWS-KENT Publishing Company, Boston, USA.
6. Dunn, O. J. (1964), "Multiple Comparisons Using Rank Sums." *Technometrics*, Vol. 6, p. 241-252.
7. Elliot, R. J. (1982). *Stochastic Calculus and Applications*. Springer-Verlag, New York, USA.
8. Gentleman, R. and Geyer, C. J. (1994). "Maximum Likelihood Estimation for Interval Censored Data: Consistency and Computation." *Biometrika*, Vol. 81, p. 618-623.

9. Gibbons, J. D. (1985). *Nonparametric Statistical Inference*. 2nd ed., Marcel Dekker Inc., New York, USA.
10. Gibbons, J. D. (1995). *Nonparametric Methods for Quantitative Analysis*. 3rd ed., American Sciences Press Inc., Ohio, USA.
11. Harter, H. L. (1961). "Expected Values of Normal Order Statistics." *Biometrika*, Vol. 48, p. 151-165.
12. Hasminskii R. Z. and Ibragimos I. A. (1985). "On Nonparametric Estimation of the Value of a Linear Functional in a Gaussian White Noise". *Theory of Probability and its Applications*, Vol. 29, p. 18-32.
13. Klein, J. P. and Moeschberger, M. L. (1997), *Survival Analysis Techniques for Censored and Truncated Data*. Springer – Verlag New York, USA.
14. Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*. John Willey & Sons Inc., New York, USA.
15. Lee, E. T. (1992), *Statistical Methods for Survival Data Analysis*. 2nd ed., John Wiley & Sons Inc. New York, USA.
16. Lipster, R. S. and Shirayayev A. N. (1980). "A Functional Central Limit Theorem for Semimartingales." *Theory of Probability and its Applications*, Vol. 25, p. 667-688.
17. Newton, M.A., Quintana, F.A. and Zhang, Y. (1998). "Nonparametric Bayes Methods Using Predictive Updating." In *Practical Nonparametric and Semiparametric Bayesian Statistics* (eds D. Dey, P. Muller, and D. Sinha), p. 45-61. New York: Springer.

18. Ordoukhani, N., Thavaneswaran, A. and Samanta, M. (1991), "Functional Version of the Delta Method and its Application." *Communication in Statistics - Theory and Method*, Vol. 20, p. 373-388.
19. Puri, M. L. (1964). "Asymptotic Efficiency of a Class of c-Sample Tests." *Annals of Mathematical Statistics*, Vol. 35, p.102-121.
20. Reid N. M. (1981). "Influence Functions for Censored Data." *Annals of Statistics*, Vol. 9, p. 78.
21. Shirayayev A. N. (1984). *Probability*. Springer-Verlag, New York, USA.
22. Thavaneswaran A. (1988). "Smoothing Signals for Semimartingales." *Stochastic Processes and Their Applications*, Vol. 28, p. 81-89.
23. Thavaneswaran A. (1985). "Estimation for Semimartingales." Ph.D. Thesis.
24. Zeger, L. and Brookmeyer, R. (1986). "Regression Analysis With Censored Autocorreletad Data." *Journal of American Statistical Association*, Vol. 81, p. 722-729.
25. Ihaka, R. and Gentleman, R. (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, Vol. 5, No. 3, p. 299-314.