SURVIVAL MODELS
AND
AN APPLICATION

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of
Dokuz Eylül University
in partial Fulfillment of the Requirements for
the Degree of Master of Science in Statistics

by
Ayşe Övgü TEKİN

December, 2001
İZMİR

Ms.Sc. THESIS EXAMINATION RESULT FORM

We certify that we have read the thesis, entitled **"SURVIVAL ANALYSIS AND AN APPLICATION"** completed by Ayşe Övgü TEKİN under supervision of Prof. Dr. Nilgün MORALI and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Nilgün MORALI

Supervisor

Doç. Dr. Gül ERGÖR

Committee Member

Prof. Dr. Serdar KURT

Committee Member

Approved by the
Graduate School of Natural and Applied Sciences

Prof.Dr.Cahit Helvacı
Director

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my supervisor Prof. Dr. Nilgün MORALI for her guidance throughout course of this work.

I am thankful to Prof. Dr. Serdar KURT for his encouragement in my studies.

I am grateful to Assoc. Prof. Serdar ÖZKÖK for his special help in obtaining the data and also grateful to Alper VAHAPLAR for all his assistance.

Finally, I wish to express my deepest gratitude to my family members and my friends for their encouragement.

Ayşe Övgü TEKİN

ABSTRACT

The objective of this study is to examine the survival models and describing the data obtained in a particular field by a suitable model.

In the first section, the goal of the study was described by comparing the methods used in survival analysis. The second section gives general information about the data types and the functions used in survival analysis. In the third section, Exponential, Gompertz, Makeham, Gamma, Weibull and Lognormal distributions which are the parametric methods used frequently in survival analysis, are examined in order. In the fourth section, the nonparametric methods, which are Kaplan-Meier product limit and Life table methods, were investigated.

In the fifth section of the study, the data obtained were grouped in a convenient way and examined by using Kaplan-Meier product limit method. The related data were taken from Ege University Faculty of Medicine, Branch of Radiation Oncology and contains information about survival data and type of tumor belonging to NSCLC patients. The differences in two groups were researched by logrank and Wilcoxon tests.

In the last section, the results of the application are discussed. Using Kaplan-Meier product limit method discovers the differences between the levels of variables measured in classification level and the survival times of the patients according to NSCLC are investigated.

ÖZET

Bu çalışmanın amacı, sağkalım modellerinin incelenmesi ve elde edilen özel bir alandaki verilerin uygun bir modelle açıklanmasıdır.

Birinci bölümde, sağkalım analizlerinde kullanılan yöntemler karşılaştırılarak çalışmanın amacı belirtilmiştir. İkinci bölümde, veri tipleri ve sağkalım analizlerinde kullanılan fonksiyonlarla ilgili genel bilgiler verilmiştir. Üçüncü bölümde, sağkalım analizlerinde sıkça kullanılan parametrik yöntemlerden sırasıyla Exponential, Gompertz, Makeham, Gamma, Weibull ve Lognormal dağılımları incelenmiştir. Dördüncü bölümde, parametrik olmayan yöntemlerden Kaplan-Meier product limit ve Life table metodları incelenmiştir.

Beşinci bölümde ise elde edilen veriler uygun bir şekilde gruplandırıldıktan sonra Kaplan-Meier product limit metoduyla incelenmiştir. Veriler, Ege Üniversitesi Tıp Fakültesi Radyasyon Onkolojisi Anabilim Dalı'ndan elde edilmiştir ve NSCLC hastalarına ait tümör cinsi ve sağkalım bilgilerinden oluşmaktadır. İki grup arasındaki farklılıklar logrank ve Wilcoxon testleriyle incelenmiştir.

Son bölümde, uygulamanın sonuçları tartışıldı. Sınıflama düzeyinde ölçülen değişkenlerin düzeyleri arasındaki farklar Kaplan-Meier product limit metoduyla ortaya çıkarıldı ve NSCLC'e göre hastaların sağkalım süreleri incelenmiştir.

CONTENTS

Chapter One
INTRODUCTION

Chapter Two
STRUCTURE OF SURVIVAL DATA

Chapter Three
PARAMETRIC MODELS

Chapter Four

NONPARAMETRIC MODELS

Chapter Five

APPLICATION

Chapter Six

CONCLUSION

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Survival analysis is used in analyzing the data that occurs at the time of a predefined event to happen. The main difficulty met in survival data analysis is the lack of observation in some of the units examined or the time failure of the individuals because of various reasons. These observations are named as censored observations and are constructed of units or individuals with longer failure times mostly. In order to use all the data and to reach at better results, censored observations must be used correctly as other observations.

There are different approaches about the solution of related problems in survival analysis. One of these approaches is estimating by using various parametric survival distributions and another one is estimating by using nonparametric methods, which does not depend on any distribution assumptions.

It causes a preference problem having both parametric and nonparametric methods in analyzing censored survival data. Kaplan-Meier, which is a nonparametric method, is one of the frequently used method because it can be computed easily and is understandable. Although parametric models are powerful, any corruption in their hypothesis causes biasedness in results.

If the distributions used in parametric analysis are convenient to the data examined, parametric modeling gives better results. But the existence of stopped observations especially, results some problems in researching the convenience of a parametric distribution.

Depending on this information, the objective of this study is to choose the most appropriate analysis method and to find the factors that effect the survival times. So, firstly, the structure of the survival data was examined, then parametric and nonparametric methods are introduced and lastly, an analysis was applied on the data obtained from Ege University Faculty of Medicine, Branch of radiation Oncology. In this analysis, the selection of the method is discussed and determining the factors that effect survival times is focused.

CHAPTER TWO
STRUCTURE OF SURVIVAL DATA

Survival time can be defined as the time to the occurance of a given event and survival data can include survival time, response to a given treatment, and patient characteristics related to response and survival.

In this section, the types of censored observations used in survival analysis, probability density, survival and hazard functions are mentioned.

2.1 Cencored Data

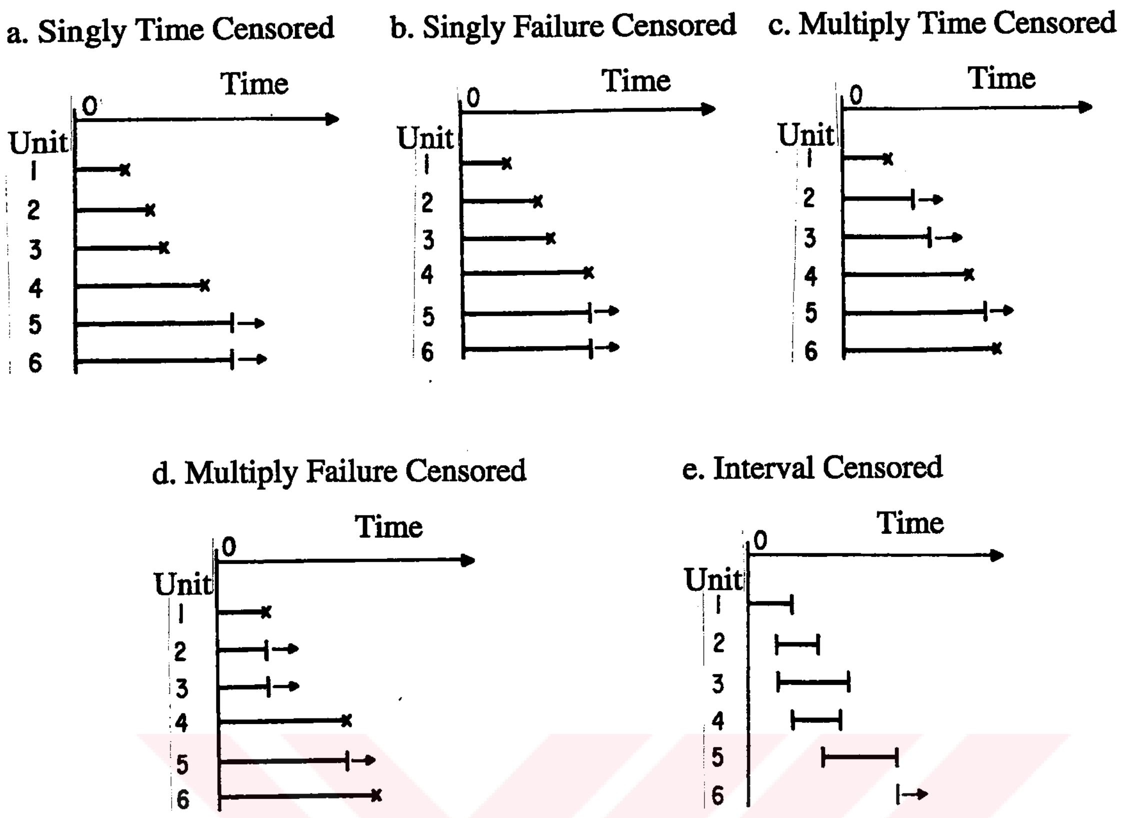
Many researchers consider survival data analysis to be merely the application of two conventional statistical methods to a special type of problem; parametric if the distribution of survival times is known to be normal and nonparametric if the distribution is unknown. This assumption would be true if the survival times of all the subjects were exact and known. However, some survival times are not (Lee, 1992).

Survival data are not appropriate to standard statistical procedures used in data analysis. The first reason is survival data are generally not symmetrically distributed. This difficulty could be resolved by first transforming the data to give a more symmetric distribution, for example by taking logarithms. However, a more satisfactory approach is to adopt an alternative distributional model for the original data (Collett,1994).

The second reason is that survival times are frequently censored. In an experiment in which subjects are followed over time until an event of intererest occurs, it is not always possible to follow every subject until the event is observed. Subjects may drop out of the study and be lost to follow-up, or be deliberately withdrawn, or the end of the data collection period may arrive before the event is observed to happen. For such a subject, all that is known is that the time to the event was at least as long as the time to when the subject was last observed. The observed time to the event under such circumstances is censored (PROPHET StatGuide).

Sometimes when survival data are analyzed, some subjects are unfailed, and their failure times are known only to be beyond their present survival times. Such data are said to be *censored on the right* or *right censoring*. Similarly, a failure time known only to be before a certain time is said to be *censored on the left* or *left censoring*. If all unfailed subjects have a common running time and all failure times are earlier, the data are said to be *singly censored* on the right. Singly censored data arise when subjects are started on the experiment together and the data are analyzed before all subjects fail. Such data are *singly time censored* if the censoring time is fixed; then the number of failures in that fixed time is random. Figure 2.1.*a* depicts such a sample. Time censored data are also called *Type I censored*. Data are *singly failure censored* if the experiment is stopped when a specified number of failures occurs, the time to that fixed number of failures being random. Figure 2.1.*b* depicts such a sample. Time censoring is more common in practise; failure censoring is more common in the literature, as it is mathematically more tractable.

Much data censored on the right have differing survival times intermixed with the failure times. Such data are called *multiply censored* (also progressively, hyper-, and arbitrarily censored). Multiply censored data usually come from the field, because subjects go into service at different survival times when the data are recorded. Such data may be time censored or failure censored. Figure 2.1.*c* and Figure 2.1.*d* depict such samples, respectively. And another type of censoring is *interval censoring*. Here, subjects are known to have experienced a failure within an interval of time. Figure 2.1.*e* depicts such a sample (Nelson, 1982).



a. Singly Time Censored

b. Singly Failure Censored

c. Multiply Time Censored

d. Multiply Failure Censored

e. Interval Censored

Figure 2.1 Types of data (failure time ×, running time ↦).

Analyses of such censored and interval data have much the same purposes as analyses of complete data, for example, estimation of model parameters and prediction of future observations.

2.2 Functions of Survival Time

Survival times are data that measure the time to a certain event such as failure, death, response, relapse, the development of a given disease etc. These times are subject to random variations and form a distribution.

The distribution of survival times described by three functions: the probability density function, the survival function (or survivorship function), and the hazard function. These are mathematically equivalent in that each can be derived from the others.

2.2.1 Probability Density Function

The actual survival time of an individual, *t*, can be regarded as the value of a variable *T*, which can take any non-negative value. The different values that *T* can take have a probability distribution, and we call *T* the random variable associated with the survival time. The survival time *T* has a probability distribution with underlying *probability density function f*(*t*). The density function is also known as the *unconditional failure rate*.

This function is defined as the limit of the probability that an individual fails in the short interval *t* to $t+\Delta t$ per unit width Δt, or simply the probability of failure in a small interval per unit time. It can be expressed as

$$f(t)=\lim_{\Delta t\to 0}\frac{P\{\text{an individual dying in the interval }(t,\, t+\Delta t)\}}{\Delta t} \qquad (2.1)$$

In practice, if there are no censored observations, the probability density function $f(t)$ is estimated as the proportion of patients dying in an interval per unit width.

$$\hat{f}(t)=\frac{\text{number of patients dying in the interval beginning at time } t}{\text{(total number of patients)(interval width)}} \qquad (2.2)$$

However, we cannot determine this function when censored observations are present. An appropriate method will be discussed in chapter four.

The distribution function of *T* is given by;

$$F(t)=P(T\le t)=\int_0^t f(u).du, \qquad (2.3)$$

and represents the probability that the survival time is less than some value *t*.

2.2.2 Survival (Survivorship) Function

Survival function is defined as the probability that an individual survives longer than t and denoted by $S(t)$. In mathematical terms:

$$S(t) = P(\text{an individual survives longer than } t) = P(T>t) \qquad (2.4)$$

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

$$S(t) = 1 - P(\text{an individual fails before time } t) = 1 - F(t) \qquad (2.5)$$

The probability of surviving at least at the time zero is 1, and $\lim_{t \to \infty} S(t) = 0$ means the probability of surviving an infinite time is zero. In accordance with these, the survival function always has a value between 0 and 1 inclusive, and is nonincreasing.

The function $S(t)$ is also known as the *cumulative survival rate*. To depict the course of survival, Berkson(1942) recommended a graphic presentation of $S(t)$. The graph of $S(t)$ is called the survival curve (Lee, 1992). Figure 2.2 shows a hypothetical survival function for a population.

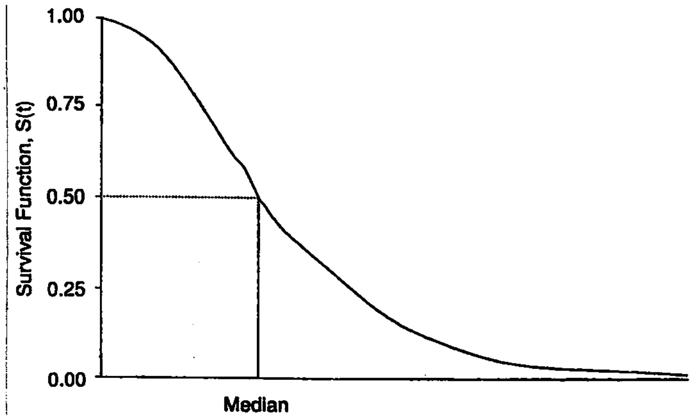


Figure 2.2 Survival function for a population.

The survival function is used to find percentiles for survival time, and to compare the survival experience of two or more groups. The mean is usually used to describe the central tendency of a distribution, but in survival distributions the median is often better because a small number of individuals with exceptionally long or short lifetimes will cause the mean survival time to be disproportionately large or small.

In practice, if there are no censored observations, the survival function is estimated as the proportion of patients surviving longer than t;

$$\hat{S}(t)=\frac{\text{number of patients surviving longer than } t}{\text{total number of patients}} \tag{2.6}$$

Similar to the estimation of *f(t)*, when censored observations are present, (2.6) is not applicable. Nonparametric methods of estimating $S(t)$ for censored data will be discussed in chapter four.

2.2.3 Hazard Function

The hazard function *h(t)* of survival time *T* gives the *conditional failure rate*. This function is a time to failure function that gives the instantaneous probability of the event (failure) given that it has not yet occured. That is, in a survival experiment where the event is death, the value of the hazard function at time *T* is the probability that an individual will die precisely at time *T*, given that the subject has survived to time *T* or the limit of the probability that an individual fails in a very short interval, *t* to $t+\Delta t$ per unit time, given that the individual has survived to time *t*:

$$h(t)=\lim_{\Delta t\to 0}\frac{P\{\text{an individual of age } t \text{ fails in the time interval } (t,t+\Delta t)\}}{\Delta t} \tag{2.7}$$

From the formula (2.7), the hazard function can also be defined in terms of the cumulative distribution function and the probability density function:

$$h(t)=f(t|T>t)=\frac{f(t)}{P(T>t)}=\frac{f(t)}{S(t)}$$

therefore

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.8}$$

In practice, when there are no censored observations, the hazard function is estimated as the proportion of patients dying in an interval per unit time, given that they have survived to the beginning of the interval:

$$\hat{h}(t) = \frac{\text{number of patients dying per unit time in the interval}}{\text{number of patients surviving at } t} \tag{2.9}$$

(Lee, 1992).

This function may increase with time, meaning that the longer subjects survive, the more likely it becomes that they will die shortly. It may decrease with time, meaning that the longer subjects survive, the more likely it is that they will survive into the near future. It may remain constant, as for a population with an exponential survival distribution. Or it may have a more complicated shape, like the well-known *bathtub* curve for human mortality, where the hazard is high for newborns, drops quickly, stays low through adulthood, and then rises again in old age.

Any function *h(t)* satisfying

1. $h(t) \geq 0$ for $-\infty < t < \infty$
2. $\int_0^\infty h(t).dt = \infty$

is a hazard function of a distribution.

The cumulative hazard function is defined as

$$H(t) = \int_0^t h(x).dx \tag{2.10}$$

and is estimated as the negative logarithm of the survival function.

Any continuous function *H(t)* satisfying

1. *H(t)* is an increasing function.
2. $\lim_{t \to \infty} H(t) = \infty$
3. *H(t)* is continuous on the right

is a cumulative hazard function of a continuous distribution (Nelson, 1982).

2.3 Relationships of the Functions

The probability density function, the survival function and the hazard function are mathematically equivalent. Given any one of them, the other two functions can be derived.

1. From the equations (2.5) and (2.8) the hazard function can be shown as

$$h(t) = \frac{f(t)}{S(t)} \tag{2.11}$$

2. Since the probability density function is the derivative of the cumulative distribution function,

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \tag{2.12}$$

3. Substituting (2.12) into (2.13) yields

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log_e S(t) \tag{2.13}$$

4. Integrating (2.13) from zero to t, we have

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

or

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

or

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(x).dx\right] \qquad (2.14)$$

5. From (2.11) and (2.14) we supply

$$f(t) = h(t).\exp[-H(t)] \qquad (2.15)$$

It is said that *f(t)* and *F(t)* are common representations of the distribution of a random variable. The hazard function *h(t)* is a more specialized characterization but is particularly useful in modeling survival data. In many instances, information is available as to how the failure rate will change with the amount of time on test. This information can be used to model *h(t)* and easily translated into something that is suggested for *F(t)* and *f(t)* using the above formulas.

CHAPTER THREE
PARAMETRIC MODELS

In this chapter, several theoretical distributions that have been used to describe survival time are explained and their characteristics summarized.

3.1 The Exponential Distribution

The simplest and most important distribution in survival studies is the exponential distribution. The exponential distribution plays a role in lifetime studies analogous to that of the normal distribution in other areas of statistics. Applications in human and animal studies of chronic and infectious diseases can be found.

The exponential distribution, with its property of a constant hazard rate, is frequently used in reliability engineering as a survival model for inanimate objects such as machine parts. The hazard of death at any time after the time origin of the study is the same, that is for an individual; death is a random event independent of time. Under this model, the hazard function may be written

$$h(t) = \lambda \qquad (3.1)$$

A large λ indicates high risk and short survival while a small λ indicates low risk and long survival. Figure 3.1 depicts the survival function, the probability density function, and the hazard function of the exponential distribution with parameter λ. When $\lambda = 1$, the distribution is often referred to as the unit exponential distribution.

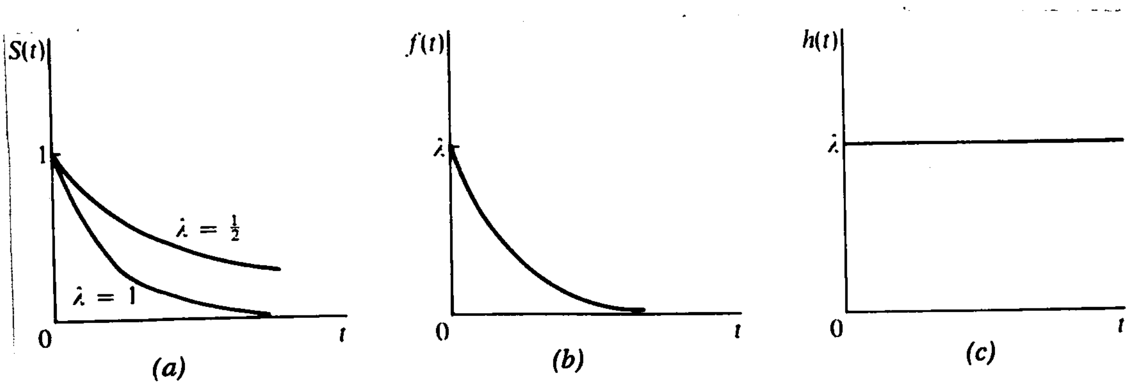


Figure 3.1 The exponential distribution: (a) survivorship function (b) probability density function (c) hazard function.

When the survival time T follows the exponential distribution with a parameter λ, the probability density function is defined as

$$f(t)=\begin{cases}\lambda e^{-\lambda t} & ,t\geq 0,\lambda>0\\ 0 & ,t<0\end{cases} \tag{3.2}$$

The cumulative distribution function is then

$$F(t)=1-e^{-\lambda t} \qquad ,t\geq 0 \tag{3.3}$$

and the survival function is

$$S(t)=e^{-\lambda t} \qquad ,t\geq 0 \tag{3.4}$$

When the natural logarithm of the survival function is taken, $\log_e S(t)=-\lambda t$, which is a linear function of t. Thus it is easy to determine whether data come from an exponential distribution by plotting $\log_e S(t)$ against t, where $\hat{S}(t)$ is an estimate of $S(t)$.

The mean and variance of the exponential distribution with parameter λ are, respectively, $1/\lambda$ and $1/\lambda^2$.

3.2 The Gompertz Distribution

This distribution was suggested as a model for human survival by Gompertz in 1825. The distribution is usually defined by its hazard rate as

$$h(t) = Bc^{t} \qquad , t \geq 0,\ B > 0,\ c > 1 \tag{3.5}$$

Then the survivorship function is

$$S(t) = \exp\left[\frac{B}{\ln c}\left(1 - c^{t}\right)\right] \tag{3.6}$$

The probability distribution function is given by $h(t) \cdot S(t)$, and is clearly not a very convenient mathematical form.

3.3 The Makeham Distribution

In 1860 Makeham modified the Gompertz distribution by taking the hazard rate function to be

$$h(t) = A + Bc^{t} \qquad , t \geq 0,\ B > 0,\ c > 1,\ A > -B \tag{3.7}$$

Makeham was suggesting that part of the hazard at any age is independent of the age itself, so a constant was added to the Gompertz hazard rate.

The survivorship function is

$$S(t) = \exp\left[\frac{B}{\ln c}\left(1 - c^{t}\right) - At\right] \tag{3.8}$$

Again it is clear that the probability distribution function for this distribution is not mathematically tractable, so the calculation of probabilities, moments, or other quantities is somewhat difficult.

3.4 The Weibull Distribution

This distribution was proposed by Weibull (1939) and its applicability to various failure situations discussed again by Weibull.

The Weibull distribution is a generalization of the exponential distribution. However, unlike the exponential distribution, it does not assume a constant hazard rate and therefore has broader application.

It is frequently used in industrial applications (Kao, 1959; Lieblein and Zelen, 1956; Nelson, 1972) and medical researches (Pike, 1966; Peto et al 1972, Williams 1978; Scott and Hahn 1980) (Başar, 1993).

The distribution is characterized by two parameters, γ and λ. The value of γ determines the shape of the distribution curve and the value of λ determines its scaling. Consequently, γ and λ are called shape and scale parameters, respectively.

The relationship between the value of λ and survival time can be seen from Figure 3.2, which shows the hazard rate of the Weibull distribution with γ=0.5, 1, 2, 4. When γ=1, the hazard rate remains constant as time increases; this is the exponential case. The hazard rate increases when $\gamma > 1$ and decreases when $\gamma <1$ as t increases. Thus, the Weibull distribution may be used to model the survival distribution of a population with increasing, decreasing, or constant risk.

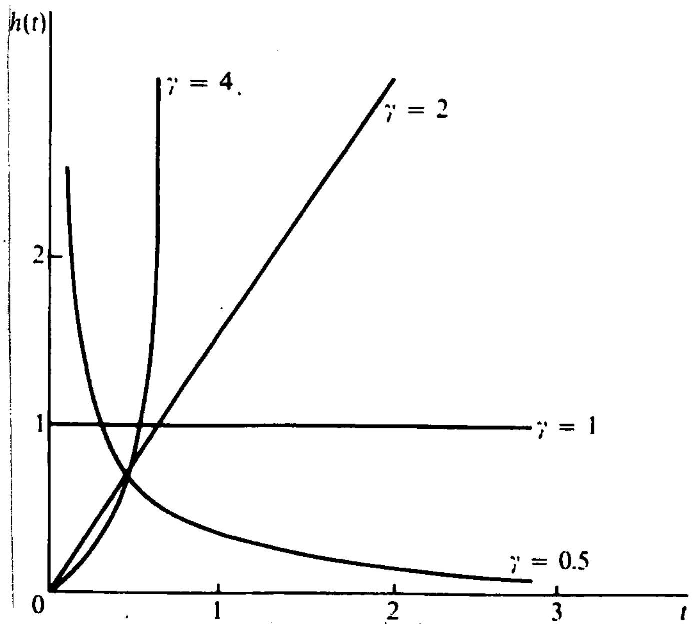


Figure 3.2 Hazard functions of Weibull distribution with $\lambda = 1$.

The hazard function is defined as

$$h(t) = \lambda\gamma\,(\lambda t)^{\gamma-1} \tag{3.9}$$

The probability density function and cumulative distribution functions are, respectively,

$$f(t) = \lambda\gamma\,(\lambda\gamma)^{\gamma-1} e^{-(\lambda t)^{\gamma}} \qquad t \geq 0, \lambda, \gamma > 0 \tag{3.10}$$

and

$$F(t) = 1 - e^{-(\lambda t)^{\gamma}} \tag{3.11}$$

The survivorship function is therefore

$$S(t) = e^{-(\lambda t)^{\gamma}} \tag{3.12}$$

The mean of the Weibull distribution is

$$\mu = \frac{\Gamma(1 + 1/\gamma\ )}{\lambda} \tag{3.13}$$

and the variance is

$$\sigma^2 = \frac{1}{\lambda^2}\left[\Gamma\left(1+\frac{2}{\gamma}\right)-\Gamma^2\left(1+\frac{1}{\gamma}\right)\right] \tag{3.14}$$

3.5 The Gamma Distribution

The gamma distribution is a natural extension of the exponential distribution as Weibull distribution and has sometimes been considered as a model in life test problems.

The disadvantage of the gamma distribution is that it suffers from its survival and hazard functions can only be expressed in terms of integrals. Especially the hazard function for the gamma distribution is

$$h(t) = \frac{\lambda^\gamma t^{\gamma-1} e^{-\lambda t}}{\Gamma(\gamma)\left[1-\Gamma_{\lambda t}(\gamma)\right]} \tag{3.15}$$

where $\Gamma(\gamma)$ is the well-known gamma function defined as

$$\Gamma(\gamma) = \int_0^\infty x^{\gamma-1} e^{-x}\,dx = (\gamma-1)! \qquad ,\gamma > 0$$

and $\Gamma_{\lambda t}(\gamma)$ is the incomplete gamma function given by

$$\Gamma(\gamma) = \frac{1}{\Gamma(\gamma)}\int_0^{\lambda t} u^{\gamma-1} e^{-u}\,du$$

This integral has to be evaluated by numerically.

The gamma distribution is characterized by shape and scale parameters, γ and λ, respectively. When 0<γ<1, there is negative aging and the hazard rate decreases monotonically from infinity to λ as time increases from zero to infinity. When γ>1, there is positive aging and the hazard rate increases monotonically from zero to λ as time increases from zero to infinity. When γ =1, the hazard rate equals λ, a constant, as in the exponential case. Figure 3.3 illustrates the gamma hazard function for λ=1 and γ<1, γ=1,2,4. Thus the gamma distribution describes a different type of survival pattern where the hazard rate is decreasing or increasing to a constant value as time approaches infinity.

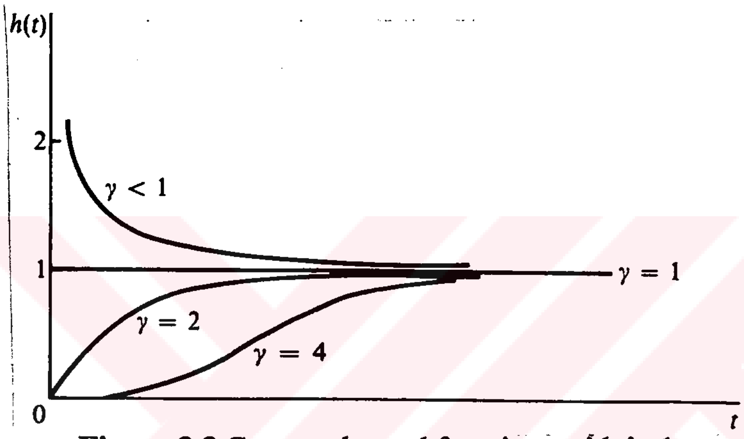


Figure 3.3 Gamma hazard functions with λ=1.

The probability density function of a gamma distribution is

$$f(t) = \frac{\lambda}{\Gamma(\gamma)}(\lambda x)^{\gamma-1}e^{-\lambda t} \qquad t > 0, \gamma > 0, \lambda > 0 \qquad (3.16)$$

and the survival function is

$$S(t) = \int_{t}^{\infty} \frac{\lambda}{\Gamma(\gamma)}(\lambda x)^{\gamma-1}e^{-\lambda x}dx \qquad (3.17)$$

for the gamma distribution.

The mean and variance of the gamma distribution are, γ / λ and γ / λ^2 respectively.

3.6 The Lognormal Distribution

This distribution is useful if the range of the data is several powers of 10. It is often used for economic data, data on response of biological material to stimulus, certain types of life data and also used for the distribution of repair times of equipment.

Consider the survival time *T* such that $\log_e T$ is normally distributed with mean μ and variance σ^2.

The lognormal probability density function and survival function are, respectively

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log_e t - \mu)^2\right] \qquad t > 0, \sigma > 0 \qquad (3.18)$$

and

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} \exp\left[-\frac{1}{2\sigma^2}(\log_e x - \mu)^2\right] dx \qquad (3.19)$$

Let $a = \exp(-\mu)$. Then $-\mu = \log_e a$, $f(t)$ and $S(t)$ can be written as

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log_e at)^2\right] \qquad (3.20)$$

and

$$S(t) = \frac{1}{\sigma\sqrt{2\sigma}} \int_t^\infty \exp\left[-\frac{1}{2\sigma^2}(\log_e ax)^2\right] \frac{dx}{x} = 1 - G(\log_e ax/\sigma) \qquad (3.21)$$

Where *G(y)* is the cumulative distribution function of a standard normal variable

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_0^y e^{-u^2/2} du$$

and the hazard function, from (3.16) and (3.17), has the form

$$h(t)=\frac{\dfrac{1}{t\sigma\sqrt{2\pi}}\exp\left[-\dfrac{(\log_e at)^2}{2\sigma^2}\right]}{1-G(\log_e at/\sigma)} \tag{3.22}$$

and is plotted in Figure 3.4.

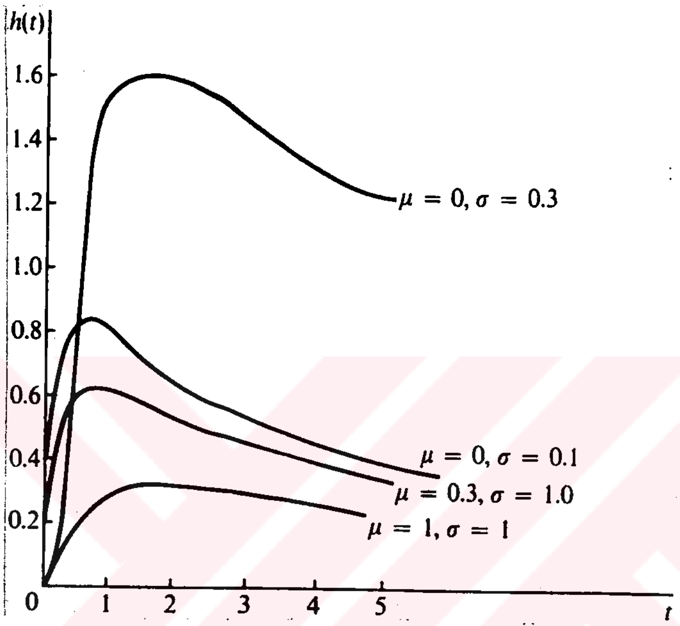


Figure 3.4 Hazard of the lognormal distribution with different parameters.

The hazard function increases initially to a maximum and then decreases to zero as time approaches infinity. Therefore, the lognormal distribution is suitable for survival patterns with an initially increasing and then decreasing hazard rate.

CHAPTER FOUR
NONPARAMETRIC MODELS

In this chapter, the methods of estimating the three survival functions (survival, density and hazard) for censored data will be discussed. These methods are said to be *non-parametric* or *distribution-free*, since they do not require specific assumptions to be made about the underlying distributions of the survival times.

Nonparametric methods are less efficient than parametric methods when survival times follow a theoretical distribution and more efficient when no suitable theoretical distributions are known. Therefore, before attempting to fit a theoretical distribution, nonparametric methods to analyze survival data are suggested. If the main objective is to find a model for the data, estimates obtained by nonparametric methods and graphs can be helpful in choosing a distribution.

In this chapter, Kaplan-Meier product limit method and life table method are discussed.

4.1 Kaplan-Meier Product Limit Method

Kaplan-Meier product limit method is the most commonly used technique for estimating the survival function for samples of small and moderate sizes and this method was developed by Kaplan and Meier in 1958.

The simple case where all of the patients are observed to death so that the survival times are exact and known. Let $t_1, t_2, ..., t_n$ be the exact survival times of the n individuals under study. We assume that this group of patients as a random sample from a much larger population of similar patients. The n survival times $t_1, t_2, ..., t_n$ ascending order such that $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(n)}$, the survival function at $t_{(i)}$ can be estimated as

$$\hat{S}(t_{(i)}) = \frac{n-i}{n} = 1 - \frac{i}{n} \tag{4.1}$$

where *n-i* is the number of individuals in the sample surviving lower than $t_{(i)}$. If two or more $t_{(i)}$ are equal (tied observations), the largest i value is used. For example, if $t_{(5)} = t_{(6)} = t_{(7)}$, then

$$\hat{S}(t_{(5)}) = \hat{S}(t_{(6)}) = \hat{S}(t_{(7)}) = \frac{n-7}{n}$$

This gives a conservative estimate for the tied observations.

This method can only be applied if all the individuals are followed to death. But some of these individuals may be censored, and there may also be more than one observation with the same survival time. Accordingly, there are r death times amongst the individuals, where $r \leq n$. These death times are arranged as $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(r)}$ and the i'th is denoted $t_{(i)}$, for $i = 1,2,...,r$. The number of individuals who are alive just before time $t_{(i)}$, including those who are about to die at this time, will be denoted n_i, for $i = 1,2,...,r$ and d_i will denote the number who die at this time.

The time interval from $t_i - \delta$ to $t_{(i)}$, where δ is an infinitesimal time interval, then includes one death time. Since there are n_i individuals who are alive just before $t_{(i)}$ and d_i deaths at $t_{(i)}$ the probability that an individual dies during the interval from $t_i - \delta$ to $t_{(i)}$ is estimated by d_i / n_i. The corresponding estimated probability of survival through that interval is then $(n_i - d_i)/n_i$ (Collett, 1994).

The deaths of the individuals in the sample occur independently of one another. Then, the estimated survival function at any time in the k'th constructed time interval from $t_{(k)}$ to $t_{(k+1)}$, for $k=1,2,...,r$, where $t_{(r+1)}$ is defined to be ∞, will be the estimated probability of surviving beyond $t_{(k)}$. This is the Kaplan-Meier estimate of the survival function, which is given by

$$\hat{S}(t)=\begin{cases}\prod_{i=1}^{k}\left(\frac{n_i-d_i}{n_i}\right) & ,t_{(k)}\leq t<t_{(k+1)}\ ,k=1,2,...r\\ 1 & ,t<t_{(1)}\end{cases} \tag{4.2}$$

Suppose that the following failure times are observed from 8 patients with small-cell lung cancer. Five patients dead at 8, 10.5, 12.5, 15, and 15 months and three patients are still alive at the end of the study after 11, 13.5, and 16 months.

The Kaplan-Meier estimate of the survival function $\hat{S}(t)$ is readily obtained using equation (4.2), and the required calculations are set out in Table 4.1.

Table 4.1 Kaplan-Meier estimate of the survival function for the eight cancer patients.

Time	n_i	d_i	$(n_i-d_i)/n_i$	$\hat{S}(t)$
8.0	8	1	0.875	0.875
10.5	7	1	0.857	$0.875\times0.857=0.750$
11.0+	---	---	---	---
12.5	5	1	0.800	$0.750\times0.800=0.600$
13.5+	---	---	---	---
15.0	3	1	0.667	$0.600\times0.667=0.400$*
15.0	2	1	0.500	$0.400\times0.500=0.200$*
16.0+	---	---	---	---

* 0,200 is used as $\hat{S}(15.0)$. It is conservative estimate.

If the largest observation is uncensored, the Kaplan-Meier estimate at that time equals zero. But if the largest observation is censored the Kaplan-Meier estimate can never equal zero and is undefined beyond the largest observation.

The Kaplan-Meier survival estimate is a step-function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death time.

The median survival time is the most commonly used summary statistic in survival analysis and a simple estimate of the median can be read from survival curves estimated by the Kaplan-Meier method as the time t at which $\hat{S}(t)=0.5$ (Lee, 1992, p77).

The Kaplan-Meier estimator can be viewed in any of the following ways;

1. *The maximum likelihood estimator (Kaplan and Meier, 1958),*
2. *The estimator obtained from a product of estimators of conditional probabilities (Kaplan and Meier, 1958),*
3. *The self-consistent estimator (Efron, 1967) and*
4. *The redistribute-to-the-right estimator (Efron, 1967; Peterson, 1975).* (Peterson, 1977).

4.1.1 Standard Error of the Kaplan-Meier Estimate

The Kaplan-Meier estimator is unbiased and consistent.

For the variance of the estimator, let us consider,

$$\hat{S}(t)=\prod_{i=1}^{k}\hat{p}_i \qquad ,k=1,2,\ldots,r$$

(London, 1988, p167) where $\hat{p}_i=(n_i-d_i)/n_i$ is the estimated probability that an individual survives longer through the time interval which begins at $t_{(i)}$, $i=1,2,\ldots,r$. Taking logarithms,

$$\log\hat{S}(t)=\sum_{i=1}^{k}\log\hat{p}_i \tag{4.3}$$

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

$$\operatorname{var}\left\{\log \hat{S}(t)\right\}=\sum_{i=1}^{k} \operatorname{var}(\log \hat{p}_i) \tag{4.4}$$

The number of individuals who survive through the interval beginning at $t_{(i)}$ can be assumed to have a binomial distribution with parameters n_i and p_i. The variance of $(n_i - d_i)$ is given by

$$\operatorname{var}(n_i - d_i)=n_i p_i (1-p_i) \tag{4.5}$$

therefore the variance of $\hat{p}_i$ is estimated by $\hat{p}_i(1-\hat{p}_i)/n_i$.

In order to obtain the variance of $\log \hat{p}_i$, *we make use of a general result for the approximate variance of a function of a random variable. According to this result, the variance of a function g(X) of the random variable X is given by*

$$\operatorname{var}\{g(X)\} \approx \left\{\frac{dg(X)}{dx}\right\}^2 \operatorname{var}(X) \tag{4.6}$$

This is known as the Taylor series approximation to the variance of a function of a random variable (Collett, 1994, p23).

Using equation (4.6), the approximate variance of $\log \hat{p}_i$ is $\operatorname{var}(\hat{p}_i)/\hat{p}_i^{\,2}$ and then the $\operatorname{var}(\log \hat{p}_i)$ equals to $(1-\hat{p}_i)/(n_i \hat{p}_i)$. If we denote this formula by d_i and n_i;

$$\frac{(1-\hat{p}_i)}{n_i \hat{p}_i}=\frac{d_i}{n_i(n_i - d_i)} \tag{4.7}$$

From equation (4.4)

$$\operatorname{var}\left\{\log \hat{S}(t)\right\} \approx \sum_{i=1}^{k} \frac{d_i}{n_i\left(n_i - d_i\right)} \tag{4.8}$$

and a further application of the result in equation (4.6) gives

$$\operatorname{var}\left\{\log \hat{S}(t)\right\} \approx \frac{1}{\left[\hat{S}(t)\right]^2} \operatorname{var}\left\{\hat{S}(t)\right\}$$

so that

$$\operatorname{var}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{i=1}^{k} \frac{d_i}{n_i\left(n_i - d_i\right)} \tag{4.9}$$

As a result, the standard error of the Kaplan-Meier estimate of the survival function is the square root of equation (4.9), is given by

$$s.e.\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right] \left\{\sum_{i=1}^{k} \frac{d_i}{n_i\left(n_i - d_i\right)}\right\}^{\frac{1}{2}} \qquad , t_{(k)} \leq t < t_{(k+1)} \tag{4.10}$$

This result is known as Greenwood's formula (Collett, 1994).

4.1.2 Linear (Greenwood) Confidence Interval for Survival Function

"A pointwise confidence interval for the survival probability *S(t)* at a specified *t* can be obtained by the usual normal-theory approximation using Greenwood's formula" (Oakes, 2001, p102).

The interval is computed from percentage points of the standard normal distribution. Thus $100(1-\alpha)\%$ confidence interval for the survival function at some specified time *t* is calculated from

$$\hat{S}(t) \pm z_{\alpha/2} \, s.e.\left\{\hat{S}(t)\right\} \tag{4.11}$$

But there is one difficulty with this procedure arises from the fact that the confidence intervals are symmetric. When the estimated survival function is close to zero or unity, symmetric intervals are inappropriate, since they can lead to confidence limits for the survival function that lie outside the interval (0,1). A pragmatic solution to this problem is to replace any limit that is greater than unity by 1.0, and any limit that is less than zero by 0.0.

4.1.3 Nelson-Aalen Hazard Estimator

The Nelson-Aalen estimator is recomended as the best estimator of the cumulative hazard function, *H(t)*. This estimator is give as

$$\tilde{H}(t) = \begin{cases} 0 & ,t_{\min} > t \\ \sum_{t_i \le t} \frac{d_i}{n_i} & ,t_{\min} \le t \end{cases} \tag{4.12}$$

The variance of this estimate is given by the formula

$$\sigma_H{}^2(t) = \sum_{t_i \le t} \frac{(n_i - d_i)d_i}{(n_i - 1)n_i{}^2} \tag{4.13}$$

4.2 Life Table Method

The most straightforward way to describe the survival in a sample is to compute the life table. The life table technique is one of the oldest methods for analyzing survival data. Berkson ang Gage (1950) and Cutler and Ederer (1958) give a life table method for estimating the survival function; Gehan (1969) provides methods for estimating all three functions (survival, density, hazard). This method is also sometimes referred to in the medical literature as the Cutler-Ederer method (1958).

If the data have been grouped into intervals or the sample size is very large or the interest is in a large population, it may be more convenient to perform a life table analysis (Lee, 1992).

The life table estimate of the survival function is obtained by first dividing the period of observation into a series of time intervals. These intervals are usually equal length, but need not necessarily to be equal. Suppose that the *i*'th of *m* such intervals, $i = 1,2,...,m$, extends from time t_i' to t_{i+1}', and let d_i and c_i denote the number of deaths and the number of censored survival times, respectively, in this time interval. And also let n_i be the number of individuals who are alive. We now make the assumption that the censoring process is such that the censored survival times occur uniformly throughout the i'th interval, so that the average number of individuals who are at risk during this interval is

$$n_i' = n_i - \frac{c_i}{2} \tag{4.14}$$

This assumption is sometimes known as the actuarial assumption (Collett, 1994).

In the *i*'th interval, the probability of death can be estimated by d_i / n_i', so that the corresponding survival probability is $(n_i' - d_i)/n_i'$. According to this probability that an individual survives beyond time t_k', $k = 1,2,...,m$, that is, until some time after the start of the *k*'th interval. This will be the product of the probabilities that an individual survives beyond the start of the *k*'th interval and through each of the *k*-1 preceding intervals, and so the life table estimate of the survival function is given by

$$\hat{S}^*(t) = \prod_{i=1}^{k} \left(\frac{n_i' - d_i}{n_i'} \right) \qquad , t_k' \le t < t_{k+1}' \,,\ k = 1,2,...,m \tag{4.15}$$

For example, we can compute Table 4.2, using 62 patients with small cell lung cancer, which are originally reported by Maksymiuk et al. (1993), were also included in Ying et al. (1995) paper.

Table 4.2 Life-table estimate of the survival function.

Interval	Time period	n_i	c_i	n_i'	d_i	$(n_i' - d_i)/n_i'$	$S^*(t)$
1	0-	62	0	62.0	2	0.967	1.000
2	180-	60	0	60.0	9	0.850	0.968
3	360-	51	0	51.0	16	0.686	0.823
4	540-	35	0	35.0	7	0.800	0.564
5	720-	28	2	27.0	5	0.814	0.452
6	900-	21	1	20.5	6	0.707	0.368
7	1080-	14	5	11.5	2	0.826	0.260
8	1260-	7	1	6.5	0	0.923	0.215
9	1440-	6	0	6.0	0	0.916	0.198
10	1620-	6	3	4.5	0	0.888	0.182
11	1800-	3	2	2.0	0	0.750	0.162
12	1980-	1	1	0.5	0	0.000	0.121

The estimated probability of surviving until the start of the first interval, t_1' is unity, while the estimated probability of surviving beyond t_{m+1}' is zero.

The life table method is primarily designed for situations in which actual failure and censoring times are unavailable and only d_i's and c_i's are given for the i'th interval (Kalbfleisch et al, 1980).

4.2.1 Standard Error of the Life Table Estimate

The standard error of the life table estimate can be found as the standard error of the Kaplan-Meier estimator. So, if the survival function of the life table estimate is $\hat{S}^*(t)=\prod_{i=1}^{k}\left(\frac{n_i'-d_i}{n_i'}\right)$, then the variance of $\hat{S}^*(t)$ is estimated by,

$$\mathrm{var}\left\{\hat{S}^*(t)\right\}=\left[\hat{S}^*(t)\right]^2\left\{\sum_{i=1}^{k}\frac{d_i}{n_i'(n_i'-d_i)}\right\} \qquad ,t_k'\leq t<t_{k+1}',k=1,2,\ldots,m \qquad (4.16)$$

As a result the standard error of the life table estimate of the survival function is the square root of equation (4.16), is given by

$$s.e.\left\{\hat{S}^*(t)\right\}=\left[\hat{S}^*(t)\right]\left\{\sum_{i=1}^{k}\frac{d_i}{n_i'(n_i'-d_i)}\right\}^{1/2} \qquad (4.17)$$

4.2.2 Confidence Interval for Survival Function of Life Table

This confidence interval is computed from percentage points of the standard normal distribution. Thus $100(1-\alpha)\%$ confidence interval for the survival function at some specified time t is calculated as similar as equation (4.11), that is

$$\hat{S}^*(t)\pm z_{\alpha/2}.s.e.\left\{\hat{S}^*(t)\right\} \qquad (4.18)$$

4.2.3 Hazard Function for Life Table Estimate

It is supposed that the observed survival times have been grouped into a series of m intervals, as in the construction of the life table estimate of the survival function. An appropriate estimate of the average hazard of death per unit time over each interval is the observed number of deaths in that interval divided by the average time survived in that interval. This latter quantity is the average number of persons at risk in the interval, multiplied by the length of the interval. Assuming that the death rate is constant during the i'th interval, the average time survived in that interval is $(n_i'-d_i/2)\tau_i$, where τ_i is the length of the i'th time interval. Thus the life table estimate of the hazard function in the i'th time interval is given by

$$h^*(t)=\frac{d_i}{(n_i'-d_i/2)\tau_i} \qquad ,t_k' \le t < t_{k+1}, k=1,2,\ldots,m \tag{4.19}$$

The asymptotic standard error of this estimate has been shown by Gehan (1969) to be given by

$$s.e.\{h^*(t)\}=\frac{h^*(t)\sqrt{1-[h^*(t)\tau_i/2]^2}}{\sqrt{d_i}} \tag{4.20}$$

(Collett, 1994).

CHAPTER FIVE
APPLICATION

In this study, methods of survival analysis are examined and it is objected to make an analysis on real data. In order to get the data of the application, the physicians who have made researches on cancer have been connected. As these researches are difficult works prepared by large teams and by the reason of importance of the research results to be published in the name of the team, the research data was used limited in the light of ethical reasons.

5.1 General Informations

When a patient is diagnosed as cancer, the survival time of the patient varies depending on the percentage location of the tumor, the characteristics, type of operation and treatment and the opportunities of operation.

In this research, the data about 58 patients operated for non-small cell lung cancer diagnose and applied to Ege University Faculty of Medicine Clinic of Radiation Oncology(*) for radiotherapy are evaluated. The data cover some information about the patients, which had medical treatment after the operation between the dates November 31st, 1994 and November 28th, 2001. Each record comprises of the following information:

1) Personal data: Name-surname
Sex
Age

(*) Data are obtained from the records kept by Dr. Serdar ÖZKÖK, Ege University.

2) Follow-up data :	Karnofsky performance status
	Number of hemoglobin
	Weight loss
	Time without far metastasis
	Time without local recurrence
	Survival time
	Censored / Failure
3) Tumor's feature :	Tumor stage
	Nodal stage
	Overall stage
	Local recurrence
	Far metastasis
	Histopathology
4) Related with the operation :	Type of Operation
	Date of operation
	Date of last follow-up
	Localization
5) Characteristics of the treatment:	Dose of Radiotherapy
	Delay in Radiotherapy

It is possible to obtain valuable and important results from the details of the existing records in sight of medicine. But this subject is outlined from the interest of this research.

5.2 Data

While limiting the study with the idea of finding a significant relationship between the properties of the tumor and the survival time, only the age, tumor stage, nodal stage, histopathology and survival time of the patient are considered. Details are presented in Table 5.1.

Table 5.1 Data for 58 patients of NSCLC.

Age	Histopathology	Tumor Stage	Nodal Stage	Survival Time (month)
35	large-cell carcinoma	T2	N1	10
35	large-cell carcinoma	T3	N0	74
36	adenocarcinoma	T3	N0	14
37	large-cell carcinoma	T3	N0	66+
39	adenocarcinoma	T2	N1	10
40	large-cell carcinoma	T2	N0	17
42	large-cell carcinoma	T3	N0	54+
42	large-cell carcinoma	T2	N1	18+
42	large-cell carcinoma	T2	N1	24+
45	other	T2	N0	7
46	epidermoid	T3	N0	24
47	epidermoid	T3	N0	72
47	epidermoid	T2	N0	31
47	large-cell carcinoma	T2	N2	21
48	epidermoid	T3	N1	21
50	epidermoid	T2	N1	40+
50	epidermoid	T2	N0	11
52	adenocarcinoma	T3	N1	76+
52	large-cell carcinoma	T3	N0	34+
53	epidermoid	T2	N0	75+
54	epidermoid	T3	N1	14
54	epidermoid	T2	N0	41+
55	epidermoid	T2	N1	84+
55	adenocarcinoma	T2	N0	54+
55	adenocarcinoma	T2	N2	23+
56	adenocarcinoma	T2	N1	92+
57	epidermoid	T3	N0	68+
57	epidermoid	T2	N0	60+
58	epidermoid	T3	N0	8
58	epidermoid	T3	N1	25
59	large-cell carcinoma	T3	N0	8
59	epidermoid	T3	N1	15
59	epidermoid	T3	N1	40+
59	epidermoid	T3	N1	24+
60	epidermoid	T3	N0	15
60	other	T2	N0	28
60	epidermoid	T3	N2	24
60	epidermoid	T3	N1	13
61	adenocarcinoma	T3	N1	52
61	other	T2	N1	54+
61	epidermoid	T3	N1	10
61	epidermoid	T2	N2	24+
62	adenocarcinoma	T2	N1	9
62	epidermoid	T3	N0	10
62	large-cell carcinoma	T2	N1	21+
64	epidermoid	T3	N1	9
64	other	T2	N1	41+
64	epidermoid	T2	N0	32+
64	large-cell carcinoma	T3	N0	27+
65	large-cell carcinoma	T2	N1	18
66	epidermoid	T3	N0	16
66	adenocarcinoma	T1	N1	33+
68	adenocarcinoma	T2	N1	61+
68	epidermoid	T3	N1	13
70	epidermoid	T2	N1	12
73	epidermoid	T3	N0	6
74	other	T3	N1	54
75	epidermoid	T2	N0	24

Histopathology is defining the kind of tumor by microscopic investigation. Epidermoid, adenocarcinoma, large-cell carcinoma and "other" (out of these tumor types) are the expressions used in our study.

Tumor stage is defined according to the tumor's size and its behavior in spreading to other tissues around it. It consists T1, T2 and T3 phases.

Nodal stage is determined according to the lymph nodes to be effected. It has N0, N1 and N2 phases. Data is summarized in Table 5.2.

Table 5.2 Number of patients and tumor characteristics.

	Number of patients	Percent	Censored	Failure
Histology				
epidermoid	30	51.7	10	20
adenocarcinoma	10	17.3	6	4
large-cell carcinoma	13	22.4	7	6
other	5	8.6	2	3
Tumor (T) stage				
T1	1	1.7	1	0
T2	28	48.3	16	12
T3	29	50.0	8	21
Nodal (N) stage				
N0	26	44.8	10	16
N1	28	48.3	13	15
N2	4	6.9	2	2

5.3 Methodology

As we have no information about the data of a parametric distribution and as there are censored observations, it is not convenient to use one of the parametric methods. While choosing one of the nonparametric methods, Kaplan-Meier analysis –which is more effective– is preferred instead of Life-Table because there is suitable data for Kaplan-Meier analysis.

5.4 Statistical Analysis

Statistical analysis is performed in three steps. In the first step, the survival function is obtained due to the survival times of 58 patients and it is investigated if it can be examined by one of the parametric models or not. Ensuring this is impossible, the research for the factors that effect survival began.

In the second stage, it was investigated if only one of the factors of age, histopathology, tumor stage and nodal stage effects the survival time or not. But it was seen this is also not possible. So that, it was investigated if it is possible to explain the survival times with two factors.

5.4.1 Overall Survival Function

Kaplan-Meier survival function about the 58 patients is obtained as presented by the graph in Figure 5.1.

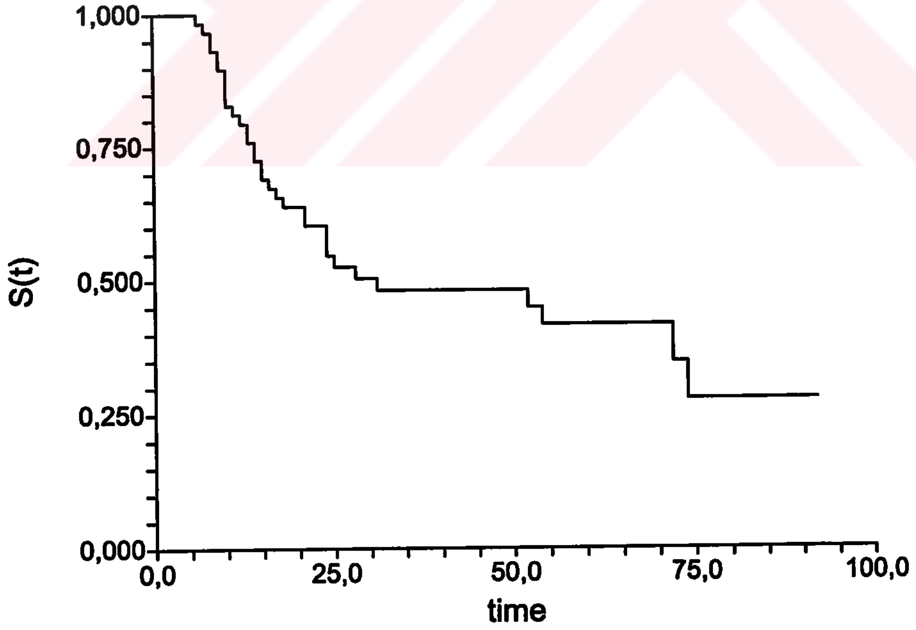


Figure 5.1 Survival function for all patients.

The alternative models Exponential, Weibull, Gamma and Lognormal are suggested to be suitable from the graph of survival function and the together graphs are given in Figure 5.2 - 4. It is observed from these graphs and the test statistics that none of these parametric distributions represent the data exists.

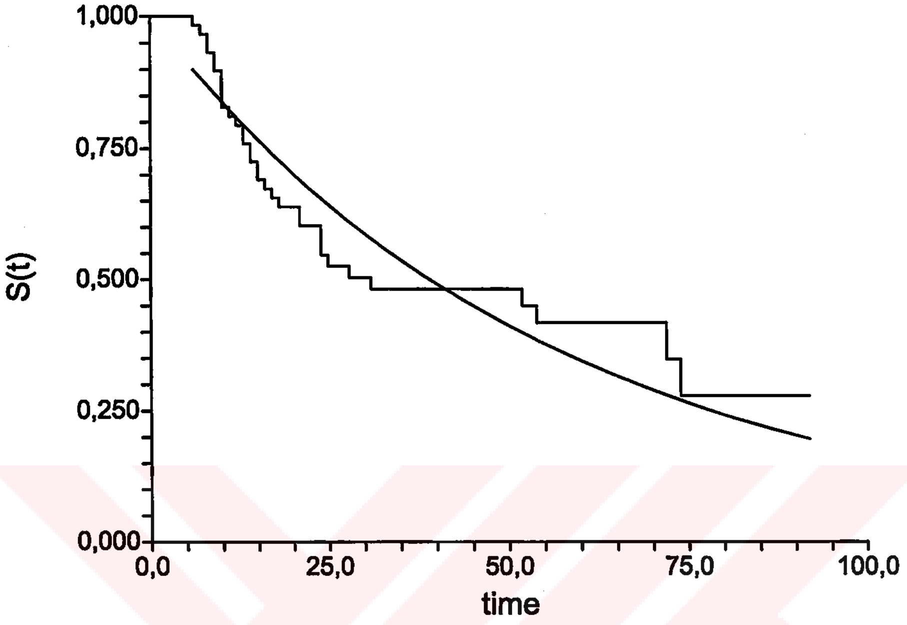


Figure 5.2 Survival function estimate for Exponential distribution.

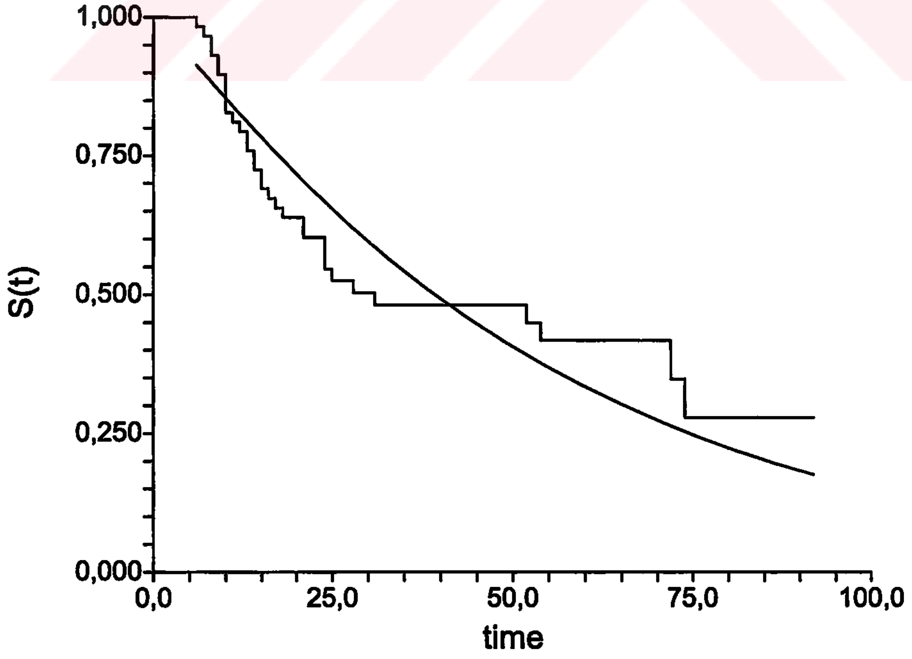


Figure 5.3 Survival function estimate for Weibull distribution.

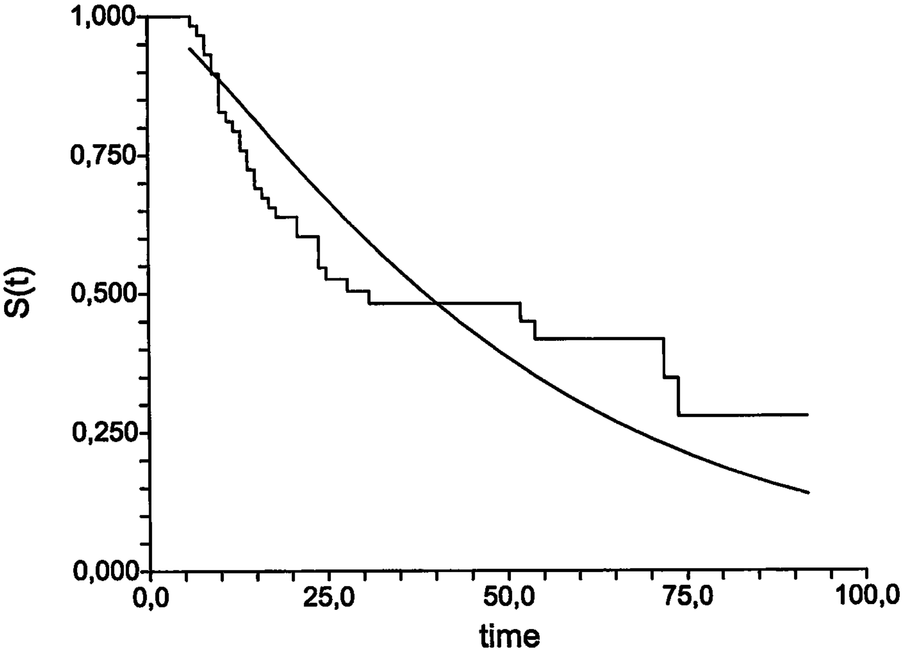


Figure 5.4 Survival function estimate for Gamma distribution.

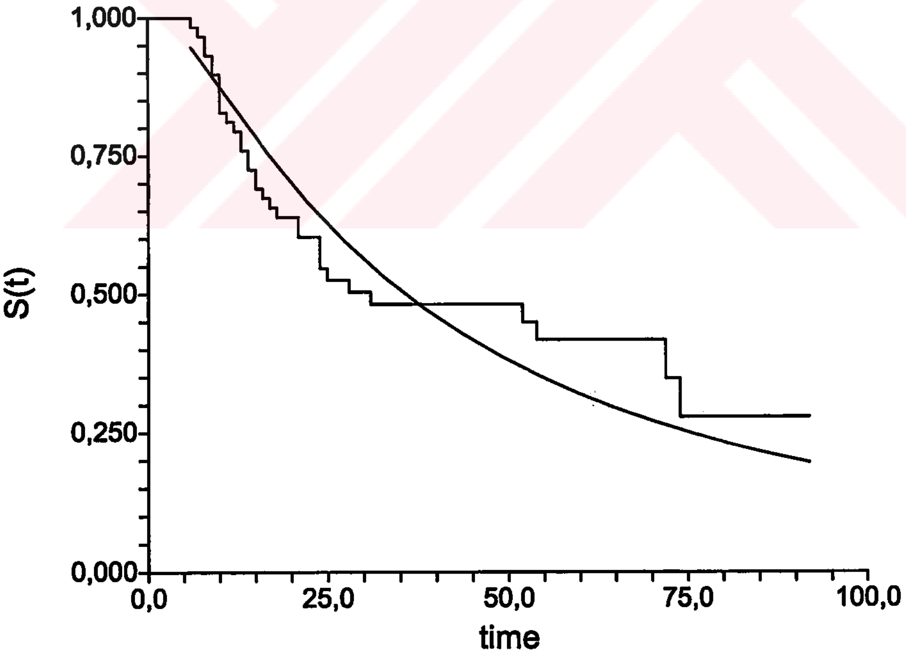


Figure 5.5 Survival function estimate for Lognormal distribution.

5.4.2 Kaplan-Meier Analysis with respect to Each Characteristic

Kaplan-Meier analysis is made in order to define if the survival varies depending on the properties age, tumor stage, nodal stage and histopathology that we considered.

While searching the effect of age, first 1/3 slice and the last 1/3 slice of the sorted records according to age are compared. 20 patients are considered as young in 58 patients and ages of 35-53 are considered as age group. In old category, again 20 of 58 patients are considered as old and the age group is accepted as 61-75. The analysis results for these two groups are given in Table 5.3, Table 5.4, Figure 5.6 and Figure 5.7 in order.

Table 5.3 Kaplan-Meier Product-Limit survival distribution for young group of age.

Rank	Sample Size	Time	Survivorship S(t)	Std Error of S(t)	Cumulative Hazard Fn H(t)=-Log(S(t))	Std Error of H(t)
1	20	7.0	0.950000	0.048734	0.051293	0.232377
2	19	10.0	0.900000	0.067082	0.105361	0.287780
3	18	10.0	0.850000	0.079844	0.162519	0.332431
4	17	11.0	0.800000	0.089443	0.223144	0.373837
5	16	14.0	0.750000	0.096825	0.287682	0.414889
6	15	17.0	0.700000	0.102470	0.356675	0.457298
7	14	18.0+				
8	13	21.0	0.646154	0.107811	0.436718	0.508153
9	12	21.0	0.592308	0.111465	0.523729	0.563666
10	11	24.0	0.538462	0.113596	0.619039	0.625930
11	10	24.0+				
12	9	31.0	0.478632	0.115661	0.736822	0.710545
13	8	34.0+				
14	7	40.0+				
15	6	54.0+				
16	5	66.0+				
17	4	72.0	0.358974	0.135142	1.024504	1.024076
18	3	74.0	0.239316	0.132900	1.429969	1.523318
19	2	75.0+				
20	1	76.0+				

Table 5.4 Kaplan-Meier Product-Limit survival distribution for old group of age.

					Cumulative	
	Sample		Survivorship	Std Error	Hazard Fn	Std Error
Rank	Size	Time	S(t)	of S(t)	H(t)=-Log(S(t))	of H(t)
1	20	6.0	0.950000	0.048734	0.051293	0.232377
2	19	9.0	0.900000	0.067082	0.105361	0.287780
3	18	9.0	0.850000	0.079844	0.162519	0.332431
4	17	10.0	0.800000	0.089443	0.223144	0.373837
5	16	10.0	0.750000	0.096825	0.287682	0.414889
6	15	12.0	0.700000	0.102470	0.356675	0.457298
7	14	13.0	0.650000	0.106654	0.430783	0.502429
8	13	16.0	0.600000	0.109545	0.510826	0.551625
9	12	18.0	0.550000	0.111243	0.597837	0.606420
10	11	21.0+				
11	10	24.0	0.495000	0.112899	0.703198	0.678798
12	9	24.0+				
13	8	27.0+				
14	7	32.0+				
15	6	33.0+				
16	5	41.0+				
17	4	52.0	0.371250	0.136584	0.990880	0.995484
18	3	54.0	0.247500	0.136017	1.396345	1.490121
19	2	54.0+				
20	1	61.0+				

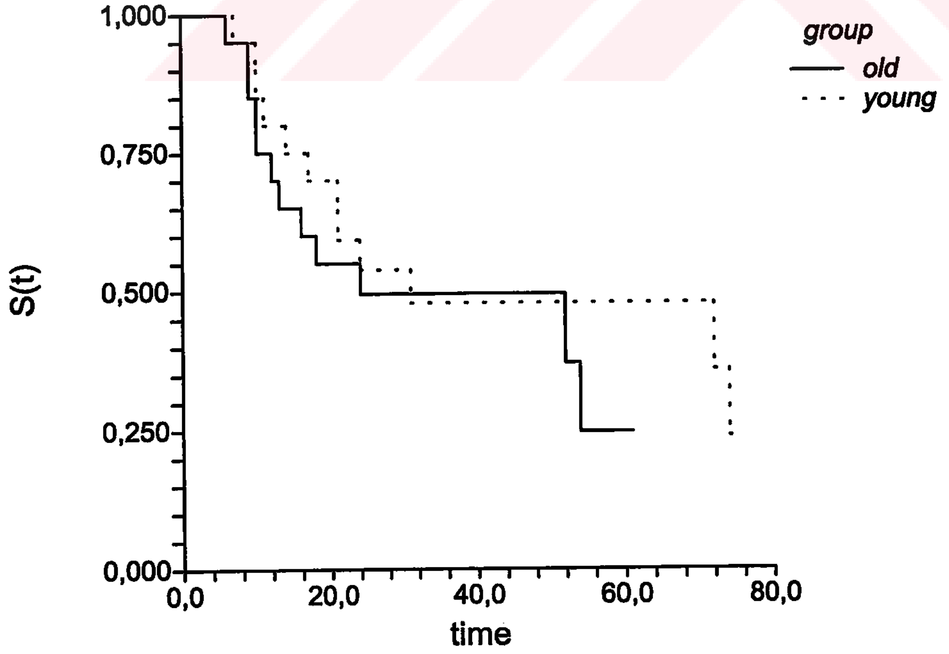


Figure 5.6 Survival functions plot of young and old group for age.

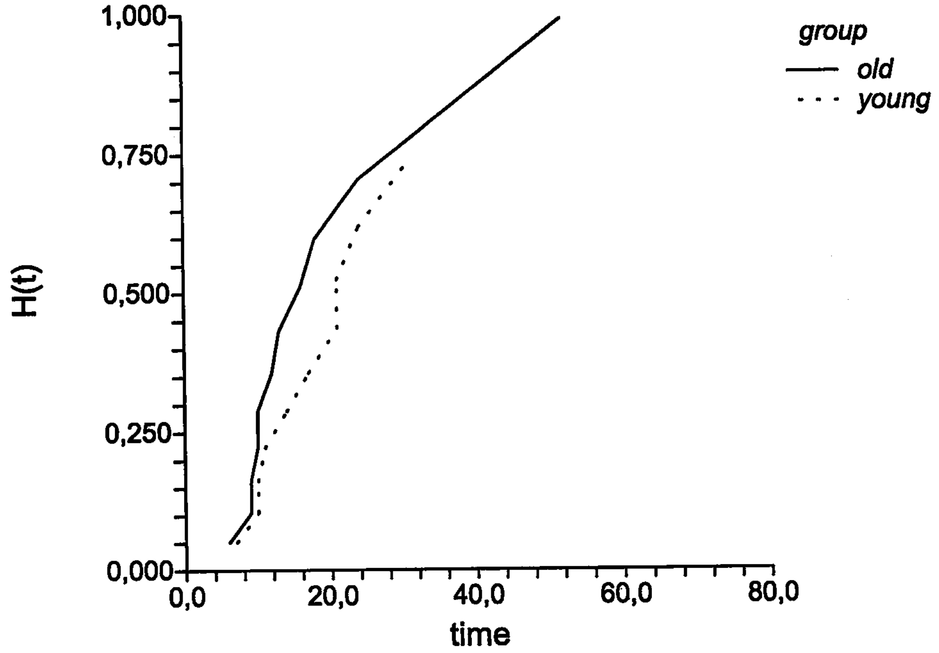


Figure 5.7 Cumulative hazard functions plot of young and old group for age.

Examining Figure 5.6, we can say that grouping by age does not make any difference in survival time of all patients. In order to test this result statistically, logrank and Wilcoxon tests are performed and chi-square and p-value results are obtained as in Table 5.5.

Table 5.5 Test statistics for age.

Test	Chi-Square	d.f.	p-value
Logrank Wilcoxon	0.61 0.49	1 1	0.4364 0.4843

Examining Table 5.5 we have no clue about the survival times of young group for age is longer than the survival times of old group for age because the p-values are greater than $\alpha = 0.05$. This means the null hypothesis cannot be declined and it can be said that there is no difference between the survival times.

While investigating how the histopathology of the tumor effects survival, it was decided as unreasonable to deal with the 4 groups in Table 5.2 separately even in medical sight. Instead, a group of 30 epidermoid patients – called epidermoid, – and another group of 23 patients considering adenocarcinoma and large-cell carcinoma – called non-epidermoid, – is taken, and a group of 5 patients – named as "other" – is excluded. The survival function seen in Figure 5.8 is obtained as a result of Kaplan-Meier analysis, which compares the survival times of epidermoid and non-epidermoid patients.

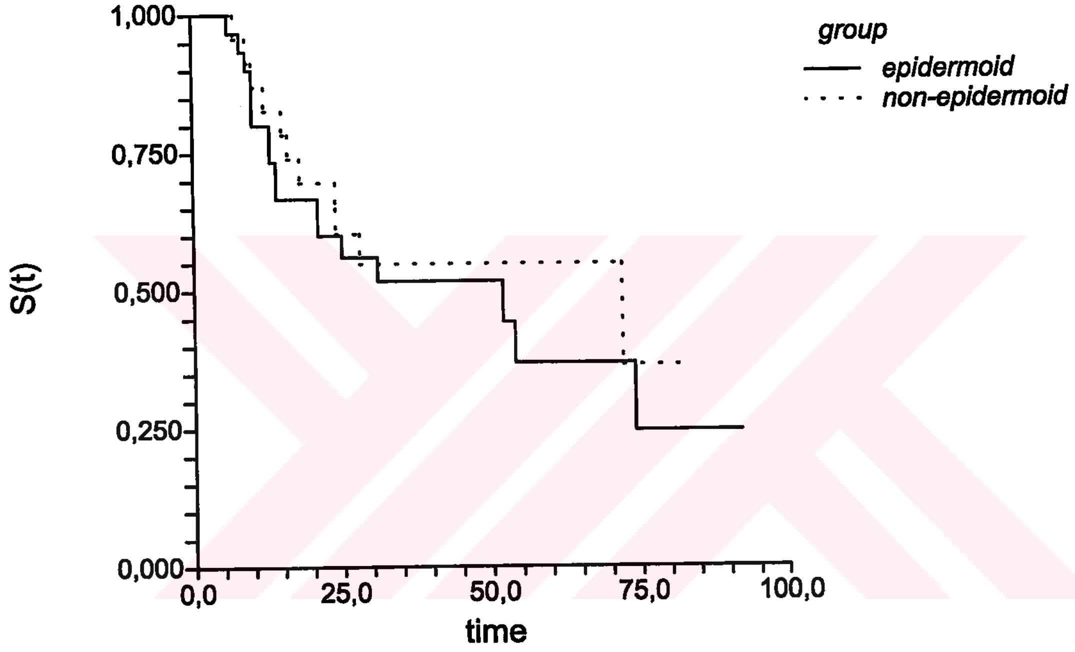


Figure 5.8 Survival functions plot of epidermoid and non-epidermoid group for histopathology.

Test statistics determines that there is not a significant difference between these survival functions. (Table 5.6)

Table 5.6 Test statistics for histopathology.

Test	Chi-Square	d.f.	p-value
Logrank Wilcoxon	0.47 0.35	1 1	0.4912 0.5569

While investigating the effect of tumor stage on survival, T2 and T3 phases are compared as there is only 1 patient in T1 phase. There are 28 patients in T2 and 29 patients in T3. The survival function plot seen in Figure 5.9 is obtained when Kaplan-Meier analysis is applied for T2 and T3 phases.

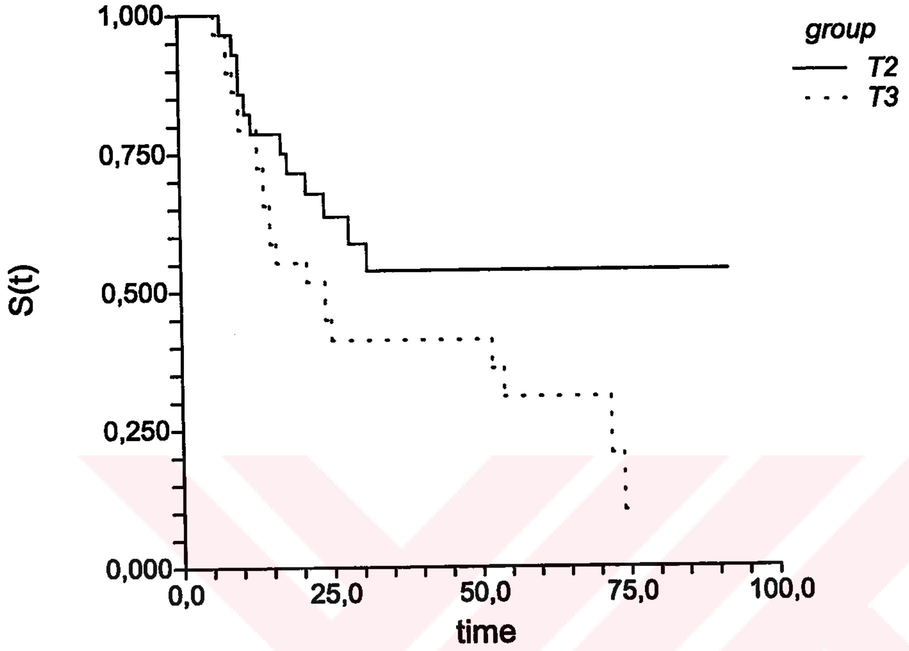


Figure 5.9 Survival functions plot of T2 and T3 group for tumor stage.

Test statistics show that there is not a significant difference between these survival functions. (Table 5.7)

Table 5.7 Test statistics for tumor stage.

Test	Chi-Square	d.f.	p-value
Logrank Wilcoxon	3.37 2.04	1 1	0.0665 0.1531

While investigating the effect on nodal stage to survival, the analyses are applied to N0 and N1 phases as there are only 4 patients in N0 phase. There are 26 patients in N0 phase and 28 patients in N1 phase. The survival function plot shown in Figure 5.10 is obtained by applying Kaplan-Meier analysis for N0 and N1 phases.

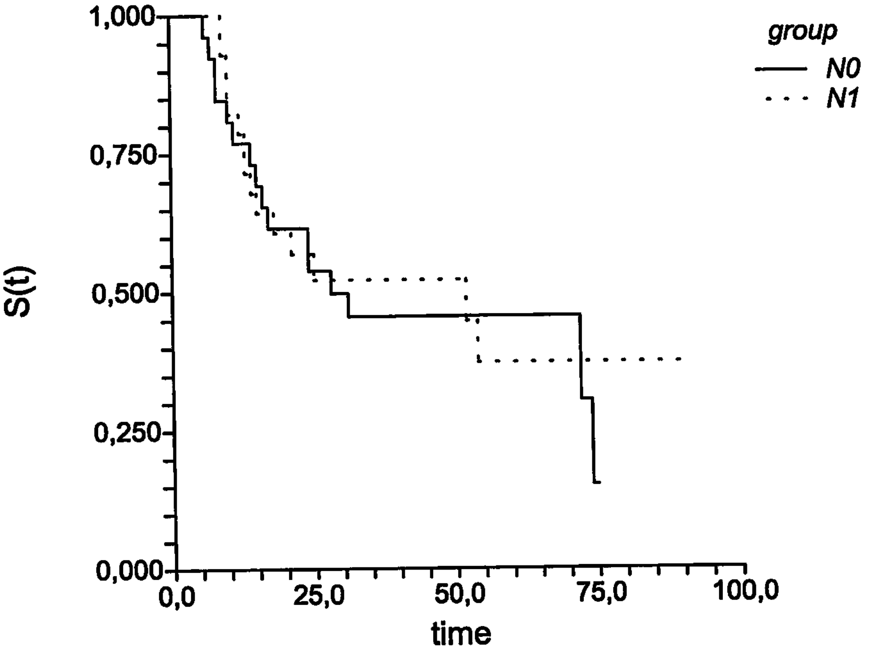


Figure 5.10 Survival functions plot of N0 and N1 group for nodal stage.

Test statistics show that there is not a significant difference between these survival functions. (Table 5.8)

Table 5.8 Test statistics for nodal stage.

Test	Chi-Square	d.f.	p-value
Logrank Wilcoxon	0.11 0.04	1 1	0.7453 0.8449

5.4.3 Kaplan-Meier Analysis for Pairwise Effect on Survival

It is seen that none of the factors affect survival alone but it is known that these characteristics have effect on survival. Consequently, a certain characteristic is considered as base and the effects of other factors on survival are examined. This investigation is started from the histopathology characteristic of the tumor. The information depending on the epidermoid or non-epidermoid type of cancer of patients are summarized in Table 5.9.

Table 5.9 Distribution of histopathology according to age, tumor stage and nodal stage.

	Histopathology			Total number of the patients
	Epidermoid	Non-epidermoid	Other	
Age				
-53	7	12	1	20
54-60	13	4	1	18
61-	10	7	3	20
Tumor stage				
T1	0	1	0	1
T2	11	13	4	28
T3	19	9	1	29
Nodal stage				
N0	15	9	2	26
N1	13	12	3	28
N2	2	2	0	4

As in the previous analysis, Kaplan-Meier analysis is performed on 2 groups for age and 2 groups for both tumor stage and nodal stage. P-values computed to test the difference are given in Table 5.10. In this table, it is seen that tumor stage effects survival on epidermoid patients and others have no effect.

Table 5.10 Test statistics.

	Test	Histopathology	
		Epidermoid	Non-epidermoid
Age	Logrank	0.1099	0.9935
	Wilcoxon	0.0595	0.8075
Tumor stage	Logrank	0.0067*	0.6546
	Wilcoxon	0.0176*	0.4883
Nodal stage	Logrank	0.6426	0.8961
	Wilcoxon	0.6195	0.9334

In the second comparison, analysis between tumor stage and other variables are performed. The data used in this comparison are given in Table 5.11.

Table 5.11 Distribution of tumor stage according to age, histopathology and nodal stage.

	Tumor Stage			Total number of the patients
	T1	T2	T3	
Age				
-53	0	11	9	20
54-60	0	7	11	18
61-	1	10	9	20
Histopathology				
Epidermoid	0	11	19	30
Non-epidermoid	1	13	9	23
Other	0	4	1	5
Nodal stage				
N0	0	11	15	26
N1	1	14	13	28
N2	0	3	1	4

It is tested if there is a difference in survival times for each variable and the p-values are given in Table 5.12. In this table it is seen that age groups and histopathology has effect on survival but no effect of others.

Table 5.12 Test statistics.

		Tumor Stage	
	Test	T2	T3
Age	Logrank	0.2491	0.0159*
	Wilcoxon	0.2677	0.0145*
Histopathology	Logrank	0.3794	0.0089*
	Wilcoxon	0.2698	0.0312*
Nodal stage	Logrank	0.5981	0.6322
	Wilcoxon	0.8284	0.8299

Lastly, nodal stage is compared with age, histopathology and tumor stage separately. The data are summarized in Table 5.13.

Table 5.13 Distribution of nodal stage according to age, histopathology and tumor stage.

	Nodal Stage			Total number of the patients
	N0	N1	N2	
Age				
-53	12	7	1	20
54-60	8	8	2	18
61-	6	13	1	20
Histopathology				
Epidermoid	15	13	2	30
Non-epidermoid	9	12	2	23
Other	2	3	0	5
Tumor stage				
T1	0	1	0	1
T2	11	14	3	28
T3	15	13	1	29

In order to test the difference between the groups, again logrank and Wilcoxon tests were used, and p-values have been obtained as in Table 5.14.

Table 5.14 Test statistics.

	Test	Nodal Stage	
		N0	N1
Age	Logrank	0.4089	0.6173
	Wilcoxon	0.3257	0.6963
Histopathology	Logrank	0.3415	0.1891
	Wilcoxon	0.3221	0.2939
Tumor stage	Logrank	0.4444	0.1061
	Wilcoxon	0.4927	0.3710

Examining Table 5.14, it can be said that none of the variables have effect on survival time according to the nodal stage.

CHAPTER SIX
CONCLUSION

6.1 Conclusions

In this study, it was objected to investigate the survival models with an application. Parametric and nonparametric methods were examined, respectively. In the application part, various statistical analyses were applied on the data, which was obtained from Ege University Faculty of Medicine, Branch of Radiation Oncology.

At the end of the analysis, it is observed that it is impossible to explain the data by any of the available parametric models. The effects of various factors on survival are investigated by using Kaplan-Meier analysis, log-rank and Wilcoxon tests. It was determined that none of the characteristics alone has effect on survival but tumor stage has effects on survivability with epidermoid type cancer patients and age and histopathological situation has effects on survivability with the patients in T3 phase.

REFERENCES

Başar, E. (1993), Yaşam tabloları analizinde kullanılan bazı istatistiksel tekniklerin böbrek nakli verilerine uygulanması, Hacettepe Üniversitesi F.B.E.

Chai, Z. (1998). Asymptotic properties of Kaplan-Meier estimator for censored dependent data. Statistics & Probability Letters, 37, 381-389.

Chen, Y.Y., Hollander, M.,& Langberg, N.A. (1982). Small-sample results for the Kaplan-Meier estimator. Journal of the American Statistical Association, 77, 141-144.

Collett, D. (1994). Modelling survival data in medical research. London: Chapman & Hall.

Crowder, M.J., Kimber, A.C., Smith, R.L.,& Sweeting, T.J. (Eds.) (1991). Statistical analysis of reliability data. London: Chapman & Hall.

Efron, B. (1981). Censored data and the bootstrap. Journal of the American Statistical Association, 76, 312-319.

Everitt, B.S.,& Dunn, G. (Eds).(1998). Statistical analysis of medical data. London: Arnold.

Gehan, E.A. (1969). Estimating survival functions from the life table. Journal of Choronic Diseases, 21, 629-644.

Gentleman, R.,& Crowley, J. (1991). Graphical methods for censored data. Journal of the American Statistical Association, 86, 678-683.

Glantz, S.A. (1977). Primer of bio-statistics.(4th ed.). McGrawhill.

Kalbfleisch, J.D.,& Prentice, R.L. (1980). The statistical analysis of failure time data. Canada: John Wiley & Sons.

Kleinbaum, D.G. (1996). Survival analysis-A self-learning text. Springer.

Kuo, C.W., Chen, Y.M., et al. (2000). Non-small cell lung cancer in very young and very old patients. Chest, 117, 354-357.

Lee, E.T. (1992). Statistical methods for survival data analysis. USA: John Wiley & Sons.

London, D. (1988). Survival models and their estimation (2nd ed.). USA: Actex Publications.

Nelson, W. (1982). Applied life data analysis. Canada: John Wiley & Sons.

Maguire, P.D., Marks, L.B., et al. (2001). 73.6 Gy and beyond: hyperfractionated, accelerated radiotherapy for non-small-cell lung cancer. Journal of Clinical Oncology, 19, 705-711.

Maksymiuk, A. W., Earle, J. R., et al. (1994). Sequencing and schedule effects of cisplatin plus etoposide in small cell lung cancer results of a North Central Cancer treatment group randomized clinical trial. Journal of Clinical Onkology, 12, 70-76.

Miller, R. G. (1981). Survival analysis. John Wiley & Sons.

Oakes, D. (2001). Biometrika centenary: survival analysis. Biometrika, 88, 99-142.

Peterson, A.V. (1997). Expressing the Kaplan-Meier estimator as function of empirical subsurvival functions. Journal of the American Statistical Association, 72, 854-858.

Pocock, S.J. (1991). Clinical trials. John Wiley & Sons.

PROPHET StatGuide: Glossary, //www.basic.nwu.edu/statguidefiles/sg_glos.html#independent

Ying, Z. Jung, S. H.,& Wei, L. J. (1995). Survival analysis with median regression models. Journal of the American Statistical Association, 90, 178-184.