# PARAMETER ESTIMATION OF MEAN SURVIVAL TIME USING PARAMETRIC AND NONPARAMETRIC APPROACHES

HASNAH BINTI ISMAIL

UNIVERSITI TEKNOLOGI MALAYSIA

# PARAMETER ESTIMATION OF MEAN SURVIVAL TIME USING PARAMETRIC AND NONPARAMETRIC APPROACHES

HASNAH BINTI ISMAIL

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

AUGUST 2011

To my beloved parents…..ayah & mak

May Allah bless you all

## ACKNOWLEDGEMENT

In the name of Allah, Most Merciful, Most Graceful. First and foremost, I would like to express my full gratitude and gratefulness to Allah for His blessings, I finally can finish my project.

I would like to express my deepest appreciation and gratitude to my supervisor, Dr. Ani bin Shabri and Dr. Zarina Bt Mohd Khalid for their guidance, invaluable advice and encouragement throughout the process of doing this project.

I also would like to thank to my family of their patience and support.

Finally, I also would like to thank to my friends for their moral and emotional support as well as fruitful ideas, comments and help in completing my project and making it a success.

## ABSTRACT

Exploring health related quality of life is usually the focus of survival studies. Using the data of breast cancer, an investigation about the mean survival time of cancer patients was explored, using the nonparametric and parametric modeling approaches. The Kaplan-Meier method and three of the distribution were considered in this study which is Weibull distribution, exponential distribution and lognormal distribution. Other than that, the Anderson Darling test is used to test if a sample data came from a population with a specific distribution. Based on the result, the data came from a Weibull distribution because the distribution has the minimum Anderson-Darling (adjusted) value. The simulation study has been done to see the efficiency of parametric and nonparametric estimator by observing the Relative Efficiency (RE) values. The results show that parametric estimator provide better estimates than the Kaplan-Meier estimator if the correct distribution is assumed.

# ABSTRAK

Kajian berkaitan hubungan kualiti kesihatan dalam kehidupan biasanya tertumpu kepada kajian *survival.* Dengan menggunakan data kanser payudara, satu penyelidikan tentang min tempoh *survival* pesakit telah dijalankan dengan menggunakan pendekatan *nonparametric* dan *parametric.* Kaedah Kaplan-Meier dan tiga taburan lain telah digunakan iaitu taburan Weibull, taburan exponential dan taburan lognormal. Selain itu, ujian Anderson-Darling telah digunakan untuk menguji sama ada data yang digunakan berasal daripada taburan tertentu ataupun tidak. Berdasarkan keputusan, data yang digunakan adalah daripada taburan Weibull disebabkan taburan Weibull mempunyai nilai Anderson-Darling paling minimum berbanding taburan exponential dan taburan lognormal. Kajian simulasi telah dibuat untuk melihat tahap kejituan penganggar *parametric* dan *nonparametric* berdasarkan nilai hubungan kejituan. Keputusan menunjukkan bahawa pengganggar *parametric* memberi anggaran yang lebih baik berbanding pengganggar Kaplan-Meier sekiranya andaian dibuat pada taburan yang betul.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURE

# LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION OF RESEARCH

## 1.1 Introduction

This study discusses the survival analysis in general followed by statement of the problem, the objectives of the study, scope of the study as well as the significance of the study. Lastly, we included the thesis organization to review the overall of the study.

## 1.2 Background of the problem

In logistic regression, interest lies in studying how risk factors were associated with presence or absence of disease. Sometimes, we are interested in how a risk factor or treatment affects time to disease or some other event. In these cases, logistic regression is not appropriate.

Survival analysis is commonly applied in many fields such as medicine, biology, public health and epidemiology. A typical analysis of survival data involves the

modeling of time-to-event data, such as the time until death. The time to the event of interest is called either survival time or failure time. The survival function is a basic quantity employed to describe the probability that an individual survives beyond a specified time. In other words, this is the amount of time until the event of interest occurs. In survival analysis, a data set can be exact or censored, and it may also be truncated. In this study, only right censored data are considered.

In the presence of right censoring, the usual estimate of the mean survival time is not appropriate. In the absence of censoring, this is equivalent to the usual estimate of the mean. When the largest observed time is censored, the Kaplan-Meier estimator is undefined beyond the largest observed time. Thus, this estimator is only appropriate when the largest observed time is a death time. One approach to overcome the limitation is to change the largest observation to a death time if it is censored. A simulation study was conducted using parametric lifetime distribution to assess the behavior of this estimator of the mean survival time in the presence of right censoring. Common parametric lifetime distributions were exponential, uniform, log-logistic, log-normal, gamma and Weibull distribution. In this study, only three distributions will be considered, that is exponential, log-normal and Weibull distribution.

## 1.3 Statement of the problem

A non-parametric estimate of the mean survival time can be obtained as the area under the Kaplan-Meier estimate of the survival curve in the absence of censoring. A common modification is to change the largest observation to a death time if it is censored. A simulation study was conducted to assess the behavior of this estimator of the mean survival time in the presence of right censoring using parametric lifetime distribution.

## 1.4 Objectives of the study

The objectives of this study are as follows:

(a) To estimate the mean survival time using standard Kaplan-Meier estimator and three other distributions, that is Weibull, Exponential and Log-Normal distribution.

(b) To fit an appropriate parametric lifetime distribution in order to test if a sample data come from a population with a specific distribution using the Anderson- Darling goodness of fit test.

(c) To compare the efficiency between Kaplan-Meier and three distributions (Weibull, Exponential and Log-Normal) of the mean survival time.

## 1.5 Scope of the study

This study discusses both parametric and nonparametric approach. This study only considered three specific distributions in simulations which are Weibull distribution, Exponential distribution and Log-Normal distribution. This study will use the right censored data. The analysis is be performed by using MINITAB and Microsoft Excel while simulation data is be performed by using MATLAB.

## 1.6 Significance of the study

The contribution of this study is to investigate different techniques to estimate the mean survival time with right censored data. In this study, both parametric and nonparametric estimators were considered. Besides focusing on different techniques of estimation, this study also helps to test if a sample data come from a population with a specific distribution using the Anderson Darling goodness of fit test. This study will also help to examine results when incorrect distribution is assumed.

## 1.7 Thesis Organization

This dissertation consists of 5 chapters.

Chapter 1 discusses the survival analysis in general followed by statement of the problem, the objectives of the study, scope of the study as well as the significance of the study.

Chapter 2 introduces the survival analysis including the definition of survival time, mean survival time and survival function. In addition, it includes the types of censoring of survival times. Furthermore, we discuss in more detail on right-censored data and interval-censored data.

Chapter 3 discusses about the nonparametric and parametric approach to estimate the mean survival time. Also, we discuss the Anderson-Darling goodness of fit test to fit an appropriate parametric lifetime model as well as the Relative Efficiency (RE) method to measure the efficiency of one estimator to another.

Chapter 4 discusses the estimation of mean survival time on a set of breast cancer data. The simulation study involves right-censored data are introduced. The results show that the parametric estimator provides goods estimates than the nonparametric if the correct distribution is assumed.

Chapter 5 discusses the conclusion of the whole study and some recommendations for those who interested to pursue the study based on survival analysis.

**REFERENCES**


Akram, M., Aman Ullah, M., Taj, R. 2007. Survival Analysis of Cancer Patients using Parametric and Nonparametric Approaches. 27(4): 194-198

Anderson, T. W. and Darling, D. A. 1954. A test of goodness of fit. *Journal of the American Statistical Association* , 49: 765-769.

Barker, C. 2009. The Mean, Median, and Confidence Interval of the Kaplan-Meier Survival Estimate. Computational and Applications. *The American Statistician*; 63:78-80

Cantor and Alan,B. 2003. *SAS Survival Analysis Techniques for Medical Research*. Cary,NC: SAS Publishing,

Cox, D. R. and D. Oakes, 1984. *Analysis of Survival Data*. Chapman and Hall, New York, USA.

Datta S. 2005. Estimating the Mean Lifetime using Right Censored Data. *Journal of Statistical Methodology*; 2:65-69.

Dobson, A. J. 2002. *An Introduction to Generalized Linear Models.* 2nd Edition. New York: Chapman & Hall.

Efron, B. 1967. The Two Sample Problem with Censored Data. *In Preceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* New York. Prentice Hall; 4:831-853

Hougaard, P. 1999. Fundamentals of Survival Data. *Biometrics.* 55: 13-22

Kalbfleish, J. D and Prentice, R.L. 2002. *The Statistical Analysis of Failure Time Data.* John Wiley and Sons Inc, New York, USA.

Kaplan ,E.L., Meier ,P. 1958. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*; 53: 457-481.

Klein, J.P and Moeschberger, L. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd Edition. New York: Springer.

Klein, J.P, Moeschberger M.L. 1997. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Lawless, J. F. 2003. *Statistical models and methods for lifetime data*. 2nd Edition. New Jersey. John Wiley & Sons

.

Lehmann, E. L.; Casella, George .1998. *Theory of Point Estimation* (2nd ed.). New York: Springer.

Ross, S.M. 2007. *Introduction to Probability Models.* 9th Edition. California. Elsevier Inc.

Zhoa Guolin, M. A. 2008. *Nonparametric and Parametric Survival Analysis of Censored Data with Violation of Method Assumptions*. Greensboro.