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

## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVER</td>
<td></td>
<td>i</td>
</tr>
<tr>
<td>DECLARATION</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td></td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td></td>
<td>vi</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>xiii</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td></td>
<td>xiv</td>
</tr>
</tbody>
</table>

1 INTRODUCTION OF RESEARCH 1

1.1 Introduction 1
1.2 Background of the problem 1
1.3 Statement of the problem 2
1.4 Objective of the study 3
1.5 Scope of the study 3
1.6 Significance of the study 3
1.7 Thesis Organization 4
2 LITERATURE REVIEW 5
2.1 Introduction 5
2.2 Survival Analysis 5
2.3 Survival Time 8
2.4 Mean Survival Time 8
2.5 Survival Function 8
2.6 Censoring 10
   2.6.1 Right-Censored Data 11
   2.6.2 Interval-Censored Data 13
2.7 Parameter Estimation for Right Censored Data using Various Methods. 13

3 RESEARCH METHODOLOGY 16
3.1 Introduction 16
3.2 Nonparametric Approach 16
   3.2.1 Kaplan-Meier Estimator 16
3.3 Parametric Approach 19
   3.3.1 Mean Survival Time for Weibull Distribution. 20
   3.3.2 Mean Survival Time for Exponential Distribution. 21
   3.3.3 Mean Survival Time for Lognormal Distribution. 22
   3.3.4 Anderson-Darling Test 22
3.3.5 Mean Square Error (MSE) 23
3.3.6 Relative Efficiency (RE) 24

4 RESULT AND DISCUSSION 25
4.1 Introduction 25
4.2 Breast Cancer Data 25
4.3 Results for Nonparametric Approach 28
4.4 Result for Parametric Approach 30
  4.4.1 Weibull Distribution 30
  4.4.2 Exponential Distribution 31
  4.4.3 Lognormal Distribution 32
4.5 Goodness of Fit 35
4.6 Simulation Study 36
  4.6.1 Generating the Data 36
4.7 Exponential Data 37
4.8 Lognormal Data 37
4.9 Weibull Data 38

5 CONCLUSION AND RECOMMENDATIONS 40
5.1 Introduction 40
5.2 Conclusion 40
5.3 Recommendations 41

REFERENCES 42
APPENDICES 45
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Possible Choices of Time Scales</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Summary on Parameter Estimation for Right Censored Data using Various Methods.</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Construction of the Kaplan-Meier Estimator</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td>Breast Cancer Data</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>Parameter Estimates and Major Characteristics of Interest of Weibull Distribution.</td>
<td>30</td>
</tr>
<tr>
<td>4.3</td>
<td>Parameter Estimates and Major Characteristics of Interest of Exponential Distribution.</td>
<td>31</td>
</tr>
<tr>
<td>4.4</td>
<td>Parameter Estimates and Major Characteristics of Interest of Lognormal Distribution.</td>
<td>32</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary of the Mean Survival Time for Each Estimator.</td>
<td>35</td>
</tr>
<tr>
<td>4.6</td>
<td>Anderson-Darling (adjusted) value of Three Distribution.</td>
<td>35</td>
</tr>
<tr>
<td>4.7</td>
<td>Mean Square Error (MSE) for Simulation Study Using Exponential Data.</td>
<td>37</td>
</tr>
<tr>
<td>4.8</td>
<td>Relative Efficiency (RE) Using Exponential Data.</td>
<td>37</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.9</td>
<td>Mean Square Error (MSE) for Simulation Study Using Lognormal Data.</td>
<td>38</td>
</tr>
<tr>
<td>4.10</td>
<td>Relative Efficiency (RE) Using Lognormal Data.</td>
<td>38</td>
</tr>
<tr>
<td>4.11</td>
<td>Mean Square Error (MSE) for Simulation Study Using Weibull Data.</td>
<td>38</td>
</tr>
<tr>
<td>4.12</td>
<td>Relative Efficiency (RE) Using Weibull Data.</td>
<td>39</td>
</tr>
</tbody>
</table>
## LIST OF FIGURE

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Lifetime State Model</td>
<td>6</td>
</tr>
<tr>
<td>2.2</td>
<td>Disease State Model</td>
<td>6</td>
</tr>
<tr>
<td>2.3</td>
<td>Diagram of Types of Censoring for Survival Time.</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Kaplan-Meier Survival Function for Right-Censored Data.</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td>Survival Function and Hazard Function for Breast Cancer Patients.</td>
<td>28</td>
</tr>
<tr>
<td>4.2</td>
<td>Output for nonparametric Estimates.</td>
<td>29</td>
</tr>
<tr>
<td>4.3</td>
<td>Distribution Overview Plot for Weibull Distribution.</td>
<td>33</td>
</tr>
<tr>
<td>4.4</td>
<td>Distribution Overview Plot for Exponential Distribution.</td>
<td>33</td>
</tr>
<tr>
<td>4.5</td>
<td>Distribution Overview Plot for Normal Distribution.</td>
<td>34</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>MINITAB Output for Weibull Distribution</td>
<td>45</td>
</tr>
<tr>
<td>B</td>
<td>MINITAB Output for Exponential Distribution</td>
<td>47</td>
</tr>
<tr>
<td>C</td>
<td>MINITAB Output for Lognormal Distribution</td>
<td>49</td>
</tr>
<tr>
<td>D</td>
<td>Coding for Simulation Data from Weibull Distribution</td>
<td>51</td>
</tr>
<tr>
<td>E</td>
<td>Coding for Simulation Data from Exponential Distribution</td>
<td>52</td>
</tr>
<tr>
<td>F</td>
<td>Coding for Simulation Data from Lognormal Distribution</td>
<td>53</td>
</tr>
</tbody>
</table>
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


