UNIVERSITI PUTRA MALAYSIA

ESTIMATION OF THE RAYLEIGH PARAMETERS BASED ON INTERVAL GROUPED DATA

HATIM SOLAYMAN AHMAD MIGDADI

FS 2008 35

ESTIMATION OF THE RAYLEIGH PARAMETERS BASED ON INTERVAL GROUPED DATA

BY

HATIM SOLAYMAN AHMAD MIGDADI

Thesis Submitted to the School of Graduate Studies, University Putra Malaysia, in Fulfillment of the Requirements for the degree of Doctor of Philosophy

June 2008

Dedication

TO THE MEMORY OF MY FATHER

Abstract of thesis presented to the Senate of University Putra Malaysia in fulfillment of the requirement for the degree of Doctor of philosophy

ESTIMATION OF THE RAYLEIGH PARAMETERS BASED ON INTERVAL GROUPED DATA

By

HATIM SOLAYMAN

June 2008

Chairman: Assoc Professor Isa Bin Daud, PhD

Faculty : Science

In this thesis the performance, the efficiency, the accuracy and the validity of the statistical estimation using the interval grouped data derived from the intermittent inspection life testing experiment are tested, improved and modified. To achieve these objectives several estimation methods are investigated employing Rayleigh as the underlying survival model.

Based on the interval grouped data the likelihood functions of the unknown Rayleigh parameters are constructed using the unconditional probability and the conditional probability(in case of censoring) of failure in the corresponding intervals. The existence and the uniqueness of the MLE's are proved. Using the equidistance case partitioning the MLE's of the scale parameter are bounded and hence bisection and secant numerical methods can be applied to arrive at faster solution. The intervals end points and the cumulative number of failures at these ends are used to derive the mid interval and the compound grouped estimators .These estimators are in explicit forms and evaluated in terms of their bias and consistency. The results of applying the maximum likelihood estimation to real life time data relatively show better

estimates of the survival and the hazard functions, as compared to the classical non parametric estimates. In the least square estimation and based on the multinomial distribution of failures the resulting estimators are compared to the corresponding estimators obtained by fitting regression models based on the nonparametric estimates of both the survival and the hazard functions at a pre given time.

In the Bayesian estimation approach the conjugate priors are derived using both the complete and the interval grouped data. High posterior credible intervals are obtained and mathematical improvements of the Bayesian estimators obtained by the interval grouped are made to increase their relative efficiency and performance. Applying the modified Bayesian estimation procedures to a generated Rayleigh lifetimes data show a significance efficiency of the Bayesian estimation method.

Despite the fact that there is a considerable loss of information in the exact unobservable lifetimes, simulation studies at different settings of the life testing experiment show a high relative efficiency of the estimators obtained using the interval grouped data in comparison with the estimators obtained using type I and right censored data.

To measure the loss of information due to the intermittent inspection life testing experiment Shannon information and distance divergence measures are considered. Modifications in the Shannon information measure and derivation of a new information measure based on the sufficient statistics are investigated to reflect the actual loss of information. A criterion for minimizing the loss of information, selecting the suitable number of intervals, the inspection times and the sample size is extracted.

The performance of the estimation procedures is also tested on some known survival analysis issues with application to a real lifetimes data. Hence, modifications in the conventional methods and formulating of alternative models are devoted to guarantee existence of the solution, improve the performance and reduce computations. Finally general conclusions on the overall thesis are given together with highlights for further researches.

Abstrak tesis yang dikemukakan Kepada Senat Universiti Putra Malaysia
Sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENGANGGARAN PARAMETER RAYLEIGH BERASASKAN DATA TERKUMPUL SELANG

Oleh

HATIM SOLAYMAN

Jun 2008

Pengerusi : Profesor Madya Isa Bin Daud, PhD

Fakulti : Sains

Dalam tesis ini prestasi, kecekapan, kejituan dan kesahihan dari penganggaran statistik menggunakan data terkumpul selang yang diperolehi dari ujikaji ujian hayat pemeriksaan terjadual diuji, diperbaiki dan dipinda. Untuk mencapai matlamat ini beberapa kaedah penganggaran diselidiki menggunakan Rayleigh sebagai model mandirian pokok.

Berdasarkan data terkumpul selang fungsi kebolehjadian dari parameter Rayleigh yang tidak diketahui dibina menggunakan kebarangkalian takbersyarat dan kebarangkalian bersyarat (dalam kes penapisan) dari kegagalan dalam selang yang selaras. Kewujudan dan keunikan dari MLE dibuktikan. Menggunakan kes pembahagian samajarak MLE daripada parameter skala adalah terbatas dan oleh yang demikian kaedah berangka pembahagian dua sama dan sekan dapat digunakan untuk mendapatkan selesaian yang lebih cepat. Titik hujung selang dan longgokan bilangan kegagalan pada titik hujung ini digunakan untuk memperolehi titik tengah

selang dan penganggar terkumpul majmuk. Penganggar ini adalah mempunyai rupa tak tersirat dan dinilai dalam hal pincang dan konsistennya. Keputusan dari penggunaan penganggaran kebolehjadian maksimum pada data mandirian nyata menunjukkan anggaran yang lebih baik bagi fungsi mandirian dan fungsi bahaya berbanding anggaran tak berparameter klasik. Dalam penganggaran kuasa dua terkecil dan berdasarkan taburan multinomial dari kegagalan penganggar yang dihasilkan dibandingkan dengan penganggar yang selaras yang diperolehi melalui penyuaian model regresi berdasarkan anggaran tak berparameter bagi kedua dua fungsi mandirian dan bahaya pada suatu masa yang ditetapkan pada awalnya.

Dalam pendekatan penganggaran Bayesan konjugat prior diperolehi menggunakan kedua dua data lengkap dan data terkumpul selang. Selang kredibel posterior yang lebih tinggi diperolehi dan pembaikan secara bermatematik penganggar Bayesan yang diperolehi melalui terkumpul selang dibuat untuk menambahkan kecekapan relatif dan prestasinya. Dengan menggunakan kaedah penganggaran Bayesan yang diubahsuai ke atas data masa hayat Rayleigh yang dibangkitkan menunjukkan kecekapan yang ketara bagi kaedah penganggaran Bayesan.

Walaupun hakikatnya ada maklumat yang hilang yang cukup penting dalam masa hayat tak tercerap yang tepat, kajian simulasi pada ujikaji ujian masa hayat dengan keadaan yang berbeza menujukkan kecekapan relatif tinggi untuk pengangar yang diperolehi menggunakan data terkumpul selang berbanding penganggar yang diperolehi menggunakan jenis 1 dan data tertapis kanan.

Untuk mengukur kehilangan maklumat kerana ujikaji ujian masa hayat pemeriksaan berjadwal maklumat Shannon dan ukuran kecapahan jarak diambil kira. Pengubahsuaian ukuran maklumat Shannon dan penerbitan ukuran maklumat baru berdasarkan statistik cukup diselidiki untuk mencerminkan kehilangan maklumat yang sebenarnya. Kriteria untuk meminimumkan kehilangan maklumat, pemilihan jumlah selang yang layak, masa pemeriksaan dan saiz sample dibina.

Prestasi dari kaedah penganggaran juga diuji pada beberapa isu analisis mandirian yang dikenal pasti dengan penerapan pada data masa hayat nyata. Oleh yang demikian pengubahsuaian dari kaedah yang lazim dipakai dan perumusan model alternatif ditumpukan untuk menjamin kewujudan dari solusi, menambah prestasi dan mengurangi pengiraan. Akhirnya kesimpulan am untuk tesis ini diberikan dengan sorotan untuk penyelidikan lebih lanjut.

ACKNOWLEDGEMENTS

First of all, prais is for ***Allah Subhanahu WA Talala*** for giving me the strength, guidance and patience to complete this study. May blessing and peace be upon our Prophet **Muhammad *Sallalhu Alaihi Wasallam***, the first preacher who was sent for mercy to the entire world.

I am particularly grateful to Assoc. Prof. Dr. Isa bin Daud, chairman of the supervisory committee, for his endless guidance, supervision, invaluable advises and fruitful discussions. His boundless assistance during this study is highly appreciated. Similarly I would like to express my gratitude to other members of supervisory committee namely (Assoc. Prof. Dr. Rizam abu Bakar, and Assoc. Prof. Dr. Noor binti Ibrahim) for their invaluable discussions, comments and help.

Likewise, I would like to convey my deepest appreciation to my beloved mother who has always been praying for my success. The same goes to my beloved wife who has always been a caring partner. I would like to express my sincere love to all my children for the patience that they have shown. I sincerely thank them all.

I also wish to express my thanks to all of my friends and colleagues during my study in Universiti Putra Malaysia, particularly Mr.Abedullrageeb Al Eryani for his continuous help, patience and for being a good friend.

Special acknowledgements to all the staff in the Department of Mathematics for their continuous support, help and encouragement.

I certify that an Examination Committee has met on **10/6/2008** to conduct the final examination of **Hatim Solayman Migdadi** on his **Doctor of Philosophy** thesis entitled "**Estimation the Parameters of the Rayleigh Distribution Based on Interval Grouped Data**" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the student be awarded the Phd degree.

Members of the Examination Committee are as follows:

Chairman, PhD
Professor.Dr.Malik Abu Hassan
Department of Mathematics
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Examiner 1, PhD
Associate Professor.Kassim Haron
Department of Mathematics
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Examiner 2, PhD
Dr.Jayanthi Arasan
Department of Mathematics
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

External Examiner, PhD
Professor.Dr. Ataharul Islam
Faculty of Science
Universiti of Scinece Malaysia
(External Examiner)

HASANAH MOHD. GHAZALI, PhD
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of **Doctor of Philosophy**. The members of the Supervisory Committee were as follows:

Isa bin DAUD, **PhD**
Associate Professor
Department of Mathematics
Faculty of Science
University Putra Malaysia
(Chairman)

Noor Akma Ibrahim, PhD
Associate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Member)

Mohd Rizam abu Bakar, PhD
Associate Professor
Department of Mathematics
Faculty of Science
Universiti Putra Malaysia
(Member)

AINI IDERIS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 11 September 2008

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

____

HATIM SOLAYMAN MIGDADI

Date: 12 August 2008

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LEST OF ABBREVATIONS

MLE: the maximum likelihood estimator

a_j : The interval end point

I_j : The interval j

m_j : The mid interval point of the interval I_j

d_j : The number of failures in I_j

N_j : The number of units at risk at the beginning of I_j

W_j : The number of units censored or withdrawals in I_j

E : Expectation

AV : Asymptotic variance

δ : The fixed length in the equidistance case partition

Δ : The fixed length in the third case partition with endpoints $a_j^2 = a_{j-1}^2 + \Delta$

$\Rightarrow$: implies

P : Probability

$(A \mid B)$: A conditional B

k : The number of intervals

n : The sample size

AMB : The absolute main bias

MSE : The mean squared error

EF : The estimated relative efficiencies

Π : Prior function

iid : Independent identically distributed

SELF: The Squared error loss function

ELELF: The Exponential linear error loss function

$|a|$: The absolute value of a

a' : The transpose of a

CHAPTER 1

INTRODUCTION

1.1 Introduction

The statistical analysis of survival time data is an important topic in many fields of study such as medicine, biology, engineering, public health and epidemiology.

Data that measure "the length of time" until the occurrence of an event are called lifetimes, failure times or survival data. Such data may consist together a set of variables referred to as covariates. Basically the analysis of survival times is to study the lifetime in association with these covariates.

Another characteristic of the survival data is the possibility of censoring. Censoring occurs when we are unable to observe completely the response variable of interest. When the available information is lower bounds of their lifetimes it is referred as right-censored. When the failure time is only known to be smaller than a given value then it is left-censored.

Survival data are often right censored. For example in clinical trials monitoring the remission of patients for a specific disease some patients may still be monitored at the trial clusters. So, for these patients we only know that their true period of remissions were longer than the duration of the trail. On the other hand patients may be withdrawn and their progress can no longer be followed up. Their status is described as “lost to follow up”. Such patients are called right censored.

In some situations it is difficult or even impossible to obtain exact lifetimes because of the limitations of the measuring instruments (as an example the microprocessor data provided by Nelson (1982)) or because of ethical and physical restrictions in research design which allow subjects in the follow up studies to be monitored only periodically. This type of study only provides the grouped information, i.e. the exact failure time is unknown and the only available information is whether the event of interest occurred between two inspection times. This procedure is frequently used because it requires less testing effort than the continuous inspection. The data resulting from the “intermittent inspection” is fall into the categorical of grouped data and referred to as interval grouped data and consist of the number of failures in each inspection interval.

In survival data analysis there is a growing interest in developing statistical methods for analyzing interval grouped data. In the recent years the amount of statistical research devoted to grouped data has considerably increased. Most of this research concentrate on the nonparametric estimation of the hazard and survival functions at a given inspection time.

One of the primary reasons for grouping can be found in studies involving large sample sizes. For example in epidemiological studies concerning follow up of large population groups over certain time periods to study the cause and the rate of deaths or to compare these rates among different population groups. This grouping into tabular representations (life tables) often provides a convenient format for presenting and summarizing life information. Details can be found in Heom (1997).

Some large data sets are publicly released only in grouped form as the American cancer society study of million men and women by Hammond (1966) and the life span study of over 100,000 Japanese atom bomb survivors in Hiroshima and Nagasaki by Beebe (1981), so grouping protect their individual records.

Various parametric models are used in the analysis of life times. Among these models there are only few distributions occupying a central role because of their demonstrated usefulness in wide range of situations.

1.2 Basic Definitions

The three major functions in the survival data analysis are the probability density function $f(t)$, the survival function $S(t)$ and the hazard function $h(t)$. The hazard function is the instantaneous rate of failure at time t and is defined by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{1.1}$$

This implies that the functions $f(t), F(t), S(t)$ and $h(t)$ mathematically give an equivalent specifications of the distribution of a random life time T. The cumulative hazard function $H(t)$ is defined as

$$H(t) = \int_0^t h(u)du$$

and is related to the survivor function by $S(t) = e^{-H(t)}$.

The main objective in many survival studies is to understand the relationship between lifetime and covariates. This can be formulated assuming that each individual in a population has a lifetime T and a column vector of covariates $\underline{X} = (x_1, x_2, ..., x_p)'$ then the survival function expressed as $S(t \mid \underline{X}) = S(t \mid \theta(\underline{X}))$ and the hazard function of T given $\underline{X}$ expressed in a proportional hazard model given by Cox (1972) as:

$$h(t \mid \underline{X}) = h_0(t)e^{\beta' \underline{X}} \quad (1.2)$$

Where $h_0(t)$ is the underlying hazard function. Hence, the estimation of the regression coefficients is necessarily.

Other part of the survival data analysis is related with situations in which the failure process comes from distinct modes denoted as (competing risks). Failure may occur due to one of the modes. The distinguishing feature of multiple failure modes setting is that each individual has lifetime T and a mod of failure C. the joint model of T and C can be approached by specifying model for $P(T \leq t, C = j)$ by the hazard function