QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

WEIBULL DISTRIBUTION BASED ON EDUCATION PARTLY INTERVAL CENSORED DATA

BY

NOORA MOHSIN SAEED

A Project Submitted to

the Faculty of the College of Arts and Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Masters of Science in

Applied Statistics

June 2018

COMMITTEE

The members of the Committee approve the Project of Noora Mohsin Saeed defended on 23/05/2018.

____

Dr. Faiz Ahmed Elfaki
Thesis/Dissertation Supervisor

____

Dr. Ayman Suleiman Bakleezi
Committee Member

____

Dr. Saddam Akber Khan Abbasi
Committee Member

ABSTRACT

SAEED, NOORA, MOHSIN, Masters: June : [2018:], Applied Statistics

Title:_Weibull distribution based on education partly interval censored data

Supervisor of Project: Dr. Faiz Ahmed Elfaki.

The work in this project is concerned with the applying of techniques for the assessment of survival analysis in data that include censored observations. Survival analysis has a lot of achievement in the medical, engineering, economic, education and other fields and it also known as failure time analysis. Partly Interval Censoring (PIC) is one of the techniques of the censoring that used in the survival analysis and it can help to treat many types of data especially the incomplete data. One of the most commonly lifetime distribution used in the reliability applications is Weibull distribution. In this project we use Weibull model based on modified education partly interval censored data as well as medical data and simulation data. Based on the medical data, we found that our model is comparable with Turnbull method. From the education data and simulation study for this particular case, we can conclude that our proposed distribution describes well the nature of the model as compared to the Turnbull method in terms of the value of scale and shape parameter estimates. Plots of survival distribution function against failure time are used to examine the predicted survival patterns for the two types of failures.

DEDICATION

I would like to dedicate this work to my family. To my parents, who have always encouraged me to pursue my goals. To my husband, whom without his support, patience, and understanding, the completion of this work would have never been possible.

ACKNOWLEDGMENTS

First and foremost, I would like to express my heartfelt gratitude towards my supervisor Dr. Faiz Elfaki for granting me the chance to work under his supervision and also for his academic guidance, support, and encouragement during my research work. Also many thanks to committee members Dr. Ayman Suleiman Bakleezi, Dr. Saddam Akber Khan Abbasi and Dr. Ameen Jameel Deeb Alawneh for their help, value comments and support. I also would like expand my thanks to all the members of statistics program and Mathematics, Statistics and Physics Department, Qatar University, for their kind assistance during my studies, and making my stay a memorable one.

Last but not the least, my heartfelt thanks should go to my husband for his patience, support & understanding and to my family and friends for all their support and encouragement.

I would like to thank my colleagues Rasha Adnan and Mariam Altamimi for all their support and encouragement.

Special thanks go to my manager in work Dr.Aziza Alsaadi for encouragement, continuous help and support. Also, I would like to thank my colleagues in work for all their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

CHAPTER OVERVIEW

In this chapter we introduce, the survival analysis, censored data and Weibull distribution model. The chapter also, present the background of the research, the problem statement, the objective and the scope of the research.

1.1 INTRODUCTION

1.1.1 SURVIVAL ANALYSIS

The analysis of data in various applications is quite important as it can discover many useful information, options and conclusion in decision making. Statistical method is one way that is widely used by researchers as it is provides much kind of methods in dealing with data. One of that methods used in the data analysis is the survival analysis method.

Survival analysis or failure time analysis in the lifetime it described as one of the most significant advancements method of mathematical statistics in the last quarter of the 20th century (Sam and Krongs, 2008). In fact, Singh and Totawattage (2013) mention that it's a major focus of statistics because it is involved with death and failures of components. Kleinbaum and Klein, (2005) described the survival analysis as the procedures of analysis data in statistic and the outcome is time until an event occurs.

There are many applications in survival analysis for examples in; medical, engineering, education, economic and other areas. Mostly the survival analysis method is has been widely used in the engineering applications as well as in the biomedical application. As mentioned by Xian Liu (2012), one of the examples of engineering application that deals with the survival analysis method is in the life of testing the durability of a mechanical or an electrical component. The scientist applies this technique to track the products and material's life span for predicting the product reliability. The probability of the patients / participant will survive this the main estimate of the duration or computing a survival function in survival analysis.

The duration of certain life for humans this are interested by demographers and social scientists. As an example, marriage and, in particular, the marriages formed during the year 1980 in a particular country. Lawless, (2003) described that the duration would be lifetime of a marriage; a marriage may end due to annulment, divorce, or death. Other example from education scope, as mentioned by Eagle and Barnes (2014) used survival analysis approach for measuring time until an event occurs and account for teacher's attrition.

Yin et al., (2012) described the variable that measures the time from a starting time to a particular endpoint of interest is the survival time. There are some participants in any study who may incomplete a survival time due to censoring and some time we do not know exactly their survival time. In the following section will be introduce the censored data.

1.1.2 CENSORED DATA

Censored data is one of the important factors in survival analysis. Also, censored data is a key issue for the analysis of survival data and is the reasons why survival analysis is a special topic within statistics. The presence of censored data make the survival analyses difference with other statistical analysis (Thamrin, 2013). The reasons for censoring is due to; a participants does not experience the event before the study ends; a participant is lost to follow-up during the study period and; might withdraw from the study because of death or some other reasons. Right, left, and interval censoring, this are the main three types of censoring data.

If the study ends before the event has occurred when a subject leaves the study before an event occurs in this situation we described the right censored (Kleinbaum and Klein, 2005). In this right censoring, the data that is known is in minimum value. This right censoring is a common type that many researchers used in their study. Left censoring will happened when the subject is already failed before the study. The data that is known from this study is maximum value only. However, the use of left censored data is rarely used compared to the use of right censored. Moreover, when the survival time of a subject is occur with interval for example [a,b] then this type we called interval censoring. When censored consist interval censored and exact data then we have Partly Interval Censored (PIC) (Kim, 2003).

In this study, Weibull distribution model will be used based on partly interval censoring data. In next section will be introduce the Weibull distribution model.

1.1.3 WEIBULL DISTRIBUTION MODEL

In lifetime data one of the most useful distributions for analyzing & modelling is the Weibull distribution in various fields such as; the medical, biological, engineering and other fields. It was applicable to various failure situations and was proposed by Weibull (1939). Lee and Wang, (2003) proposed that Weibull distribution it used in many mortality of the human disease studies and reliability studies. It is described by two parameters that is; shape parameter that is determine the distribution curve and the other parameter determine the scaling.

Rinne, (2008) described that the data used by Weibull distribution have been modelled that originate from such different areas such as the biological, environmental, health, physical and social sciences.

Lee and Wang (2003) proposed that the probability density function of the two Weibull distribution parameters as;

$$f(t;\alpha,\beta)=\begin{cases}\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}e^{-\left(\frac{t}{\alpha}\right)^{\beta}}, & if\ t\geq 0\\ 0, & if\ t<0.\end{cases} \qquad (1.1)$$

where β and α are represented the shape and scale parameters respectively. For which both parameters are positive.

The cumulative Weibull distribution function is give as;

$$F(t;\alpha,\beta)=\begin{cases}1-e^{-\left(\frac{t}{\alpha}\right)^{\beta}}, & if\ t\geq 0\\ 0, & if\ t<0.\end{cases} \qquad (1.2)$$

In the next chapters, we will derive the model under survivorship and the estimation of the curve of the survival probability function will be calculated by using censoring data based on Weibull distribution model.

1.2 BACKGROUND

In this study, the general aim is to apply the survival analysis concept in the prediction of the survival of students in schools from one grade to another during an educational stage in a specific school year. The high survival rate in school indicates a high retention level and a low dropout rate. As mention (UIC ,2009), the survival rate in the last grade of an educational level is a matter of interest for observation. For example, the rate of survival in the last grade of primary education, which is the main objective of Education for All (EFA) and the Millennium Development Goals (MDGs), is of interest to monitor and follow the primary education program. The calculation of this indicator is based on student flow rates whereas the reliability of the retention rate to the last grade in an educational stage depends on the constancy of the enrollment data and grade repeat from their comprehensiveness over time and grades. This indicator is predicted by using Cohort analysis models based on a number of assumptions. For example; the survival of flow rates is constant throughout the study life of the cohort.

The data set which will be used in this study refers to students in a specific school year who are in Grade 7 and follow up each student until the graduate from Grade 12. The sample of students there are in Grade 7 for 2010-2011 school year and follow up each student until graduate Grade 12 at the end of 2015-2016 school year. In this study we will estimate the curve of the survival probability function based on Weibull distribution via partly interval censoring data.

1.3 PROBLEM STATEMENT

Several researchers used PIC data via Weibull distribution. For instance, Alharpy and Ibrahim (2013) used parametric Weibull distribution for score test and likelihood ratio test. Similarly, Elfaki et al. (2012) used Cox model with Weibull distribution and applied it to AIDS studies. Seguro and Lambert (2000) used Weibull distribution to estimate the parameters by three methods that is; the graphical method, maximum likelihood method, and the modified maximum likelihood method and applied it to wind speed study.

However, there are few studies that focus on the partly interval censored data and even fewer applied it to education related applications. This includes Singer and Willett (1993) who study duration and the timing of events based on discrete time survival analysis. Moreover, Plank et. al., (2008) estimated the duration of student surviving in high school.

In this research we will tackle partly interval censored data for survival analysis and apply a model that is significantly applicable to be used in education data. Consequently, the Weibull distribution model based on simple imputation techniques will be used to simplify the procedure.

1.4 OBJECTIVE

This study aims to predict of survival ability of students in schools for education partly interval censored based on Weibull model with different imputation techniques. However, the parameters in the model will be estimated using MLE method. Therefore, the main objectives of this study are:

- To use the Weibull model that is suitable for education (PIC) data based on simple imputation techniques.
- To compare the performance of our Weibull model with Turnbull model.
- To use secondary medical data, real education data and simulation data.

1.5 SCOPE OF THE RESEARCH

The research study is limited to use education partly interval censored data based on the Weibull model to predict of the survival-ability of students in Qatar school. This model is described in chapter III and the MLE will be used to estimate the parameters in the model. Simple imputation techniques will also be used to modify the real data set into PIC data.

The literature review of the survival analysis, partly interval censoring and Weibull distribution model are presented in chapter 2. The Weibull distribution model based on survival analysis and derivation of maximum likelihood estimator for parameters will be described in chapter 3. In chapter3 also, the-likelihood ratio test will be presented. At the end of chapter III, the real education data set and the process to be treat as survival time data will be described. In addition, the simple imputation techniques that used to modify the education data to be right, interval and PIC.

Weibull model that is suitable for our modified data sets based on simple imputation methods will be presented in Chapter IV. At the end of this chapter, an illustrate based on our model secondary medical data, real data set and simulated data will be given. Finally, chapter V summarizes the conclusions arrived in previous chapters and present some suggestions for future research.

CHAPTER 2

LITERATURE REVIEW

CHAPTER OVERVIEW

In this chapter, we will review some existing related literature for survival analysis which have been applied in various areas such as medicine, engineering and education. Then we will focus on one type of censoring data that used in this project that is partly interval censoring. Finally, some existing related literature for Weibull distribution model will be provided.

2.1 SURVIVAL ANALYSIS

Many researchers have applied survival analysis in their research area such as; medical and engineering application and other fields. Sam and Krongs, (2008) described that the survival analysis is one of the most significant advancements of mathematical statistics in the lifetime. There have been only a few research in survival analysis that applied it in the education application. However, in this research we will present some related literature for general survival analysis by considering different situations.

Giolo (2004) presented the nonparametric method for estimation of the survival function using interval censored data. Giolo used the Turnbull's algorithm to obtain the estimation survival functions and implemented the procedure by the software R. It has been observed that the analysis based on this type of censored data and then applying this method to standard time to event data can lead to worthless inferences. Therefore, Giolo advises that the analysts to be more careful when using the latest methods for analysing interval censored data.

Singh and Totawattage (2013) discussed the applications of survival analysis via interval failure time data. Five different techniques were used based on parametric and non-parametric methods. The techniques which were used in this study were Kaplan-Meier estimator, Turnbull method, Logspline of the survival curve, Weibull model and piecewise exponential model to estimate the parameters of survival function. They used different data sets, which were AIDS, Hemophilia, and Breast Cancer to illustrate the methods in their study. From this data, they showed that the parametric method can be more satisfying in the performance particularly when the log-normal family or Weibull is chosen to estimate the parametric because it provides a wide range of distributive forms. So, they suggest using a piecewise constant hazard model to allow additional flexible modelling with weak parametric assumptions.

From the scope of education application, Singer and Willett (1993) studied duration and the timing of events based on discrete-time survival analysis. Moreover, they have observed that the discrete-time survival analysis provides an easily applicable framework for analysing a type of event occurrence data which is frequently collected in educational research. The interpretation of the parameters of the discrete-time hazard model that can easily be fitted based on analysis of the standard logistic regression. In addition, the discrete-time approach facilitates examination the hazard function shape which is in contrast to Cox regression model, where the shape of the hazard function is ignored in support of estimating only the shift parameters associated with covariates under the assumption of proportionality.

Plank et al. (2008) provides the surviving of student in high school. Their study objective was explore to combined of Career & Technical Education (CTE) and core academic courses that influence the likelihood of leaving school. In their study, they used the hazard model for dropout of youths by one of the most common methods in estimation, which is Cox Regression model. It has been observed that the hazard model indicate that to significant curvilinear association between the CTE to academic course that taking ratio and the risk of reducing out for youths who were aged 14 and younger when they entered the ninth grade.

Similarly, Eagle and Barnes (2014) used survival analysis for measuring time to event data and account for participant attrition by Cox regression proportional hazards model and accelerated failure time (AFT) model. They demonstrated that the duration data collected from intelligent tutors are applicable for survival analysis and useful when a study experiences participant attrition.

Weybright et al. (2017) identified the risk of dropping out of secondary school for male and female teenagers based on survival analysis approach. Based on secondary longitudinal data, they examine the influence of substance and leisure experience predictors while controlling for demographic. They used Kaplan-Meier (KM) with SAS PROC LIFETEST to estimate the survival and hazard functions. Also, they used SAS

PROC PHREG to estimate the parameters of Cox Regression model with discrete time survival analysis based on demographic, substance use, and leisure experience variables to predict dropout.

While the reviewed studies have similarity with our research project in the sense that they are also focused on survival analysis. However, they are different from this research in many ways. For example, they are different based on the type of censoring data and methods used. In this research we will focus on Weibull model based on partly interval censored.

2.2 PARTLY INTERVAL CENSORING

There are several researchers that used partly interval censored in their studies. In addition, partly interval censoring data appear a lot in practice and in follow-up studies. Kim (2003) investigated the partly interval censored data using the MLE for the Cox model. In their study they used two methods to estimate the variance–covariance matrix of the MLE of the regression parameter that generalized missing information principle and generalized profile information procedure. The simulation studies indicate that both methods work well in terms of variance for samples of moderate size and the bias. In addition, the researcher illustrated this method using an application to diabetes data in Denmark. Zhao et al. (2008) studied the partly interval censored failure data using generalized log-rank test as that discussed by Peto and Peto (1972). They evaluated the method using a set of real data from a diabetes study and simulation studies.

Guure et al. (2012) used partly interval censored data based on Weibull distribution model with several estimations of parameter methods which were MLE, Least Square (LS) estimators of one variable on other variable to determine the survival estimator among these methods for estimating the parameters and to show that the bias the estimators' of the parameters are to the true values. Their used MSE bias to compared their mentioned methods Bias based on simulation study. They observed that the MLE was better for estimating the scale parameter. On the other hand, the least square for first variable was more reliable for estimating the shape parameter with relatively small samples, but with larger samples the least square on other variable was the preferred method.

Elfaki et al. (2012) examined the parametric proportional hazard Weibull model based on Expectation Maximization (EM) algorithm for PIC data of the application for an AIDS study. In their research also, partly interval censored will be mainly used in order to estimate the survivability of failure rate. They investigated the treated of HIV/AIDS of hemophiliacs in two hospitals in Sudan. In their study they showed that there are no differences between the two treatment in the real data set. But they are strongly support the generalized missing information principle in a parametric context in simulation studies and suggested to use the generalized profile information for non-identically distributed samples.

Moreover, Alharpy and Ibrahim (2013) used parametric Weibull distribution for two testing that Likelihood Ratio Test (LRT) and Score Test (ST) based on PIC data. They observed that the LRT is better than score test to test the parametric for partly interval censored under Weibull distribution.

In another study, Elfaki et al. (2013) proposed the semiparametric Cox's proportional hazards regression for PIC a competing risks model based on EM algorithm to estimate the parameters. They used two competing risks models that is the Censoring Complete (CC) model and a Weighting Technique (WT) model. They investigated the possible association among the treatment and the anti D in Rhesus time to study the effect of covariates on the development of complications which have been applied to a set of time data arising from anti D in Rhesus D negative pregnant women in Sudan. The study conclude that the covariates do not have a significant difference due to the negative group had a significantly higher risk of the onset of anti D rhesus after infection.

Yousif et al. (2016) also presented the Cox Model to estimate the regression coefficients for partly interval-censored data using Bayesian method. They were simulated the data to verify the model which was developed and which worked well as well as because easy to implement. From simulation data they showed that the developed model performs well and is demonstrated that it is applicable.

Zyoud et al. (2016) studied the partly interval-censored data using nonparametric analysis based imputation methods to estimate the survival function. The simple imputation includes right-point, left-point and mid-point imputation. On the other hand, the probability based imputation methods that includes mean imputation, median imputation, conditional mode, multiple imputation and random imputation. They implemented their proposal for estimating the survival function using R software. They observed that the random, mean and median imputations are better compared with other imputation techniques.

Wu et al. (2017) proposed the semiparametric sieve MLE method to analyse partly interval censored data using Cox Regression. In their study, they considered the non-mixture Cox Regression cure rate model and adopt the semiparametric spline-based sieve maximum likelihood approach to analyse such data. Also they are illustrated the methods using modern empirical process theory for both the parametric and the nonparametric parts of the sieve estimator. Then, they observed that the sieve estimator was consistent. They simulated the data to show the performance for the proposed method and they observed that the proposed sieve MLE is satisfactory. Following that, they have applied the proposed method on spontaneous abortion studies and have applied it successfully.

As can see in this chapter, even though many studies have used distribution method based the partly interval censored data. This is because this method is flexible when being used to estimate the parameter. The purpose of our study is to apply the Weibull model that is suitable for education PIC data. Also, to investigate the performance of Weibull model and to ascertain its effectiveness by using suitable methods.

2.3 WEIBULL DISTRIBUTION MODEL

There are many researchers who used the Weibull distribution method in survival and human disease mortality, among other; Murthy, Xie, and Jiang (2004) they promoted the usefulness of this distribution by modelling data sets from various disciplines.

Harter and Moore (1965) presented Weibull and Gamma distribution to estimate the parameters using the complete and censored data by maximum likelihood estimates (MLEs). They used numeric examples to estimate the parameters from the first failure times simulation. They show that the MLE have shown good results in cases when the estimation is non-regular.

Cohen (1965) used MLE to estimate the two parameters of Weibull distribution for complete, singly censored and progressively (multiple) censored data. In addition, he computed the asymptotic variance-covariance matrices for each of these sample.

Moreover, Seguro and Lambert (2000) used MLE, the proposed modified MLE method and graphical method to estimated the parameters in Weibull distribution and applied it to wind speed study. They have demonstrated that the MLE is the more appropriate computer based method and recommended method for prediction of the parameters of the Weibull distribution for wind energy analysis.

Odell et al. (1992) presented accelerated failure time regression model based on Weibull distribution to find the parameters for interval censored data by two methods maximum likelihood estimates (MLEs) and midpoint estimator (MDE). They have observed that the maximum likelihood estimates better than midpoint estimator for decreasing hazards. While midpoint estimator can be used in others situations if the study is based on the covariate factor. Thus, the percentage of missing data, the size of the sample, and the width of the intervals should be considered.

Farnum and Booth (1997) also used Weibull distribution to estimate the two parameters for complete failure data and right censored data by maximum likelihood estimates (MLEs). They have observed that the maximum likelihood estimates have simple lower bounds on the parameter, quick approximation for parameter estimates and they can be used to show that the MLE for the parameters are unique.

Balakrishnan and Kateri (2008) proposed alternative approach for Weibull distribution to estimate the parameters using graphical method to determinate the MLE of the shape parameter. They have taken samples from simple censored and progressive censored data (Type I & II). They showed that the MLE have existence and uniqueness.

As can be shown in foregoing to discussion the nature of the review of material are many of the reviewed materials have used the syntax method to demonstrate how important to employ PIC data in application by using the Weibull distribution model. Accordingly, this is the reason why in our present research we trying to adopt this approach by using a data from education. The following chapter will provide a methodology for this study.

CHAPTER 3

METHODOLOGY

CHAPTER OVERVIEW

This chapter presents the estimation of the parameters of Weibull distribution using maximum likelihood estimator in general and under censored data. The chapter also, present the likelihood ratio test and describes the real data set that is used in this study and the process to treat the data to survival time data. At the end of the chapter, the simple imputation technique will be presented.

3.1 WEIBULL MODEL

The Weibull distribution model is one of the continuous probability distributions and failure time model. In the past few decades Weibull distribution has had importance in survival analysis and reliability application. After the evolution of scientific research, Weibull distribution has become an accompaniment to these developments in the lifetime analysis. Many researchers have contributed to studying the characteristics of this distribution and its application in various areas such as the analysis of wind speed, rainfall and flood data, and other analysis in health sciences. However, the most extensive use of this distribution remains in reliability and survival analysis.

As mentioned early the Weibull distribution is commonly used in analysis of lifetime data. Moreover, it can be used as the underlying survival process and the process that leads to censored observation as well. Now, let T be a random variable that follows the probability density function of the two parameters in Weibull distribution presented in early Chapter. Lee and Wang (2003) give the survivorship function as;

$$S(t) = \mathrm{P}\{T > t\}$$

$$S(t) = 1 - \mathrm{P}\{T \leq t\}$$

$$S(t) = 1 - \mathrm{F(t)}$$

$$S(t) = 1 - \int_0^t f(t) = 1 - \int_0^t \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}} dt$$

Let *u*= $-\left(\frac{t}{\alpha}\right)^{\beta}$ then *du*= $-\frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} dt$. Hence,

$$S(t) = 1 - \int_0^t e^u du = 1 - \left(1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}}\right) = e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$$

$$S(t) = e^{-\left(\frac{t}{\alpha}\right)^{\beta}} \qquad (3.1)$$

And the hazard function as

$$h(t) = \frac{f(t)}{S(t)} = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \qquad (3.2)$$

In the next section, the two parameters of Weibull distribution will be estimated by using one of the most common method in estimation, that is a maximum likelihood estimator.

3.2 MAXIMUM LIKELIHOOD ESTIMATORS

The method of maximum likelihood is the most popular technique for deriving estimators and has a very wide application. Maximum likelihood estimation (MLE) is the parameter value for which the observed sample is most likely similar. There are several advantages when this method is used. One advantage of using it is presents a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations. Another advantage of the maximum likelihood estimators is having desirable mathematical and optimality properties, especially when the sample size increased the minimum variance will be unbiased estimators.

The likelihood function for Weibull distribution is given as (Lee and Wang, 2003):

$$L(\alpha,\beta)=\prod_{i=1}^{n} f(t_i;\alpha,\beta)$$

$$=\prod_{i=1}^{n}\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}$$

$$=\left(\frac{\beta}{\alpha}\right)^{n}\prod_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}$$

$$=\left(\frac{\beta}{\alpha}\right)^{n}\left[\left(\frac{1}{\alpha}\right)^{\beta-1}\right]^{n}\sum_{i=1}^{n} t_i{}^{\beta-1}\, e^{-\sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}} \qquad (3.3)$$

Then take the log-likelihood function, we can have;

$$l(\alpha,\beta)=\ln L(\alpha,\beta)$$

$$=n\ln\beta-n\ln\alpha-n(\beta-1)\ln\alpha+\sum_{i=1}^{n}(\beta-1)\ln t_i-\sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}$$

$$=n\ln\beta-n\ln\alpha-n\beta\ln\alpha+n\ln\alpha+\sum_{i=1}^{n}(\beta-1)\ln t_i-\sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}$$

$$= n\ln\beta - n\beta\ln\alpha + \sum_{i=1}^{n}(\beta - 1)\ln t_i - \sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}$$

Differentiating the above log-likelihood with respect to α and β, we have

$$\frac{\partial l}{\partial \alpha} = 0 - \frac{n\beta}{\alpha} + 0 + \beta\sum_{i=1}^{n}\left({t_i}^{\beta}\alpha^{-\beta-1}\right)$$

$$\frac{\partial l}{\partial \beta} = \frac{n}{\beta} - n\ln\alpha + \sum_{i=1}^{n}\ln t_i - \frac{\sum_{i=1}^{n}\left[{t_i}^{\beta}\ln(t_i)\right] - \ln(\alpha)\sum_{i=1}^{n}{t_i}^{\beta}}{\alpha^{\beta}}$$

Setting the above equations equal to zero to maximize the function, α and β can be obtained as follows:

$$\alpha = \frac{\left(\sum_{i=1}^{n}{t_i}^{\beta}\right)^{\frac{1}{\beta}}}{n^{\frac{1}{\beta}}} \qquad (3.4)$$

and

$$0 = \frac{n}{\beta} - n\ln\alpha + \sum_{i=1}^{n}\ln t_i - \frac{\sum_{i=1}^{n}\left[{t_i}^{\beta}\ln(t_i)\right] - \ln(\alpha)\sum_{i=1}^{n}{t_i}^{\beta}}{\alpha^{\beta}} \qquad (3.5)$$

Substituting (3.4) into (3.5) and solving for β, we obtain $\hat{\beta}$, which is the MLE. Then we can derive α easily.

The maximum likelihood estimates can be used to estimate the Weibull distribution with censored case. In this case let $t_1, t_2, \dots, t_r, t_1^+, t_2^+, \dots, t_n^+$ be the survival times observed from the *n* individuals, with *r* is an exact times and (n-r) is censored times.

The likelihood function for Weibull distribution with data involve failure, right censored and interval-censored is given as (Guure et al. 2012):

$$L\big(t_i,u_j,v_j,\alpha,\beta\big)=\prod_{i=1}^{k}f(t_i)\prod_{j=k+1}^{r}(1-F(t_i))\prod_{j=r+1}^{n}(F(v_i,\alpha,\beta)-F(u_i,\alpha,\beta))$$

This implies that,

$$L\big(t_i,u_j,v_j,\alpha,\beta\big)=\prod_{i=1}^{k}\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta-1}e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}\times\prod_{j=k+1}^{r}e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}\times\prod_{j=r+1}^{n}(e^{-\left(\frac{u_j}{\alpha}\right)^{\beta}}-e^{-\left(\frac{v_j}{\alpha}\right)^{\beta}})$$

Let $L\big(t_i,u_j,v_j,\alpha,\beta\big)=l$

Then

$$l=\left(\frac{\beta}{\alpha}\right)^{k}\prod_{i=1}^{k}\left(\frac{t_k}{\alpha}\right)^{\beta-1}e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}\times\prod_{j=k+1}^{r}e^{-\left(\frac{t_k}{\alpha}\right)^{\beta}}\times\prod_{j=r+1}^{n}(e^{-\left(\frac{u_j}{\alpha}\right)^{\beta}}-e^{-\left(\frac{v_j}{\alpha}\right)^{\beta}})$$

Implying

$$l=\left(\frac{\beta}{\alpha}\right)^{k}\prod_{i=1}^{k}\left[\left(\frac{t_k}{\alpha}\right)^{\beta-1}\right]exp\left\{\prod_{j=k+1}^{r}\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\prod_{j=k+1}^{r}\left[-\left(\frac{t_k}{\alpha}\right)^{\beta}\right]^{r-k}\right\}$$

$$\times\prod_{j=r+1}^{n}(e^{-\left(\frac{u_j}{\alpha}\right)^{\beta}}-e^{-\left(\frac{v_j}{\alpha}\right)^{\beta}})$$

Then the log-likelihood will be

$$\ln(l)=k[ln\beta-\beta ln\alpha]+(\beta-1)\sum_{i=1}^{k}\ln(t_i)-\frac{1}{\alpha^{\beta}}\left[\sum_{i=1}^{k}(t_i)^{\beta}+(r-k)\,(t_k)^{\beta}\right]$$

$$+\sum_{j=r+1}^{n}\ln\left(e^{-\left(\frac{u_j}{\alpha}\right)^{\beta}}-e^{-\left(\frac{v_j}{\alpha}\right)^{\beta}}\right)\qquad(3.6)$$

Then differentiating with respect to α and β give the following:

$$\frac{\partial ln(l)}{\partial \alpha} = \Bigg(-\frac{k\beta}{\alpha} + \frac{\beta}{\alpha}\left[\sum_{i=1}^{k}\left(\frac{t_i}{\alpha}\right)^{\beta} + (r-k)\left(\frac{t_k}{\alpha}\right)^{\beta}\right] + \sum_{i=r+1}^{n}\left[\frac{\left(\frac{u_j}{\alpha}\right)^{\beta}\left(\frac{u_j}{\alpha}\right) exp\left\{-\left(\frac{u_j}{\alpha}\right)^{\beta}\right\} - \left(\frac{v_j}{\alpha}\right)^{\beta}\left(\frac{v_j}{\alpha}\right) exp\left\{-\left(\frac{v_j}{\alpha}\right)^{\beta}\right\}}{exp\left\{-\left(\frac{u_j}{\alpha}\right)^{\beta}\right\} - exp\left\{-\left(\frac{v_j}{\alpha}\right)^{\beta}\right\}}\right]\Bigg) \tag{3.7}$$

and

$$\frac{\partial ln(l)}{\partial \beta} = \Bigg(\frac{k}{\beta} - kln\alpha + \sum_{i=1}^{k}\ln(t_i) - \frac{1}{\alpha^{\beta}}\left[\sum_{i=1}^{k}(t_i)^{\beta} ln\left(\frac{t_i}{\alpha}\right) + (r-k)(t_k)^{\beta} ln\left(\frac{t_k}{\alpha}\right)\right] + \sum_{i=r+1}^{n}\left[\frac{\left(\frac{u_j}{\alpha}\right)^{\beta}\left(\frac{u_j}{\alpha}\right) exp\left\{-\left(\frac{u_j}{\alpha}\right)^{\beta}\right\} - \left(\frac{v_j}{\alpha}\right)^{\beta}\left(\frac{v_j}{\alpha}\right) exp\left\{-\left(\frac{v_j}{\alpha}\right)^{\beta}\right\}}{exp\left\{-\left(\frac{u_j}{\alpha}\right)^{\beta}\right\} - exp\left\{-\left(\frac{v_j}{\alpha}\right)^{\beta}\right\}}\right]\Bigg) \tag{3.8}$$

Setting (3.7) and (3.8) equal to zero and use the numerical method, such as Newton Rapson method we can obtained β and α.

The Likelihood Ratio Test (LRT) will be present in the following section.

3.3 LIKELIHOOD RATIO TEST

Likelihood Ratio Test (LRT) is a hypothesis test that help to choose the best model from two nested models. In this research project, the likelihood ratio test will be used to perform tests of hypotheses about parameters that have been estimated by MLE in two situations. One of the test statistics is for testing weather all parameters in the distribution are equal to certain values and the other test statistics is for testing whether some of the parameters in the distribution are equal to certain values. To test a subset of parameter in a distribution, let $\beta = (\beta_1, \beta_2)$ denote all the parameters in a parametric distribution, where β_1 and β_2 are subsets of parameters. Then the hypothesis will be

$$H_0: \beta_2 = \beta_0 \qquad (3.9)$$

where β_0 is a vector of specific numbers. Let $\hat{\beta}$ be MLE of b , $\hat{\beta}_1(\beta_0)$ the MLE of β_1 given $\beta_2 = \beta_0$, and $\widehat{V}_2(\beta)$ the submatrix of the covariance matrix. Under H_0 the statistic test has chi-square distribution with degrees of freedom equal to the dimension of β_2 or the number of parameters in β_2. Then the likelihood ratio test statistic is given as;

$$X_L = 2\left[l(\hat{\beta}) - l\left(\hat{\beta}_1(\beta_0), \beta_0\right)\right] \qquad (3.10)$$

If the number of parameters in β_2 is equal to q, for a given significant level α. Then H_0 is rejected if $X_L > X^2_{q,\alpha}$.

To test whether all of the parameters in β equal a given set of known values β_0, the null hypothesis will be

$$H_0: \beta = \beta_0 \qquad (3.11)$$

Then the LRT is given as;

$$X_L = 2\left[l(\hat{\beta}) - l(\beta_0)\right] \qquad (3.12)$$

Then H_0 will be rejected if $X_L > X^2_{p,\alpha}$ for a given significant level α (Lee and Wang, 2003).

3.4 DATA SET

The data set which is used in this study was collected from the Ministry of Education and Higher Education in Qatar by following up the students in National Student Information System (NSIS) from grade 7 until graduation from grade 12. In this data set, the sample of students was followed up from September-2010/June-2011 school academic year to September-2015/June -2016. Two schools were chosen randomly, one of the school for boys and the other for girls. However, all the students were started from grade 7 and were followed up until graduation from grade 12.

As mentioned earlier, the data will be treated as survival time data. In this data, the event is graduation and the outcome is time in years until the students graduate. Censoring occurs when the information about student graduate do not know exactly. In this case, there are two reasons why censoring occurs;

- A student does not graduate before the study ends
- A student is lost to follow-up during the study period.

The data set contains five column which are student ID, Gender, Nationality, Year (event) and Censored. The sample size in the first group was 390 students for which were (208) male among them (135) Qatari and (73) Non-Qatari and they were (182) female (among them (105) Qatari, (77) Non-Qatari).

3.5 IMPUTATION

Imputation approaches are sometimes used to transform the problem of analyzing data. In this research project, we will modified the data based on imputation technique to be as right censored data, interval censored data and partly interval censored (PIC) data. The motivation behind that is the imputation process is very simple and there are numerous methods to deal with the data. There are two diverse types of imputation; simple imputation and multiple Imputation. The simple imputation technique will used in this research.

3.5.1 SIMPLE IMPUTATION

Simple imputation technique is one of most common used to treat the missing data. Because the simple imputation is conjectural and appealing often utilized this technique in the simple cases of observations. As mentioned by Zyoud, et al. (2016), the simple imputation methods have three main types that is;

1. The right limit of the interval R_i which represent right point.

2.The left limit of the interval L_i that represent the left point.

3. The midpoint of the interval $[L_i, R_i]$ as $(L_i + R_i)/2$ for which represent the midpoint.

CHAPTER 4

RESULTS AND DISCUSSION

CHAPTER OVERVIEW

We will illustrate the implementation of the methods discussed in the earlier chapters using three data sets. The first one is breast cancer data, the second one education real data, and the third one is from generated data. All calculations were computed using R software.

4.1 BREAST CANCER DATA

Several researchers used this data in their study such as; Zyoud, et al. (2016) modified to partly interval censored and compared with the Turnbull method. They were two failure times that is Radiation (R) and Radiation + Chemotherapy (R+C). In the first failure time (R) there were 66 patients and they were 68 patients for the second failure time (R+C). The breast cancer data are shown in Table 4.1. Our objective is to compare the cosmetic effects of the first failure time against the second failure time on women with early breast cancer and the event of interest was represent by the time to first occurrence of breast retraction. The actual dates will be recorded exactly if available, when the patients visits the clinic every 4 to 6 months.

To set up the data as the PIC we follow the same way that used by Alharphy & Ibrahim (2013) and Zyoud, et al. (2016). For more details of this data set reader can refer to Zyoud, et al. (2016) and Alharphy & Ibrahim (2013).

Table 4.1: The breast cancer data for patients with two treatments.

Radiotherapy	Radiotherapy + Chemotherapy
(0,7]; (0,8]; (0,5]; (4,11]; (5,12]; (5,11]; (6,10]; (7,16]; (7,14]; (11,15]; (11,18]; ≥ 15; ≥17; (17,25]; (17,25]; ≥18; (19,35]; (18,26]; ≥22; ≥24; ≥24; (25,37]; (26,40]; (27,34]; ≥32; ≥33; ≥34; (36,44]; (36,48]; ≥36; ≥36; (37,44]; ≥ 37; ≥37; ≥37; ≥38; ≥40; ≥45; ≥46; ≥ 46; ≥46;≥46; ≥46; ≥46; 4≥6; ≥46; 10; 23; 20; 37; 36; 20; 30; 20; 18; 30; 44; 23; 29; 15; 20; 22; 15; 45; 41; 38	(0,22]; (0,5]; (4,9]; (4,8]; (5,8]; (8,12]; (8,21]; (10,35]; (10,17]; (11,13]; ≥11; (11,17]; ≥11; (11,20]; (12,20]; ≥13; (13,39]; ≥13; ≥13; (14,17]; (14,19]; (15,22]; (16,24]; (16,20]; (16,24]; (16,60]; (17,27]; (17,23]; (17,26]; (18,25]; (18,24]; (19,32]; ≥21; (22,32]; ≥23; (24,31]; (24,30]; (30,34]; (30,36]; ≥31; ≥32; (32,40]; ≥34; ≥34; ≥35; (35,39]; (44,48]; ≥48; 16; 25; 14; 12; 24; 28; 26; 18; 40; 13; 21; 17; 27; 21; 22; 27; 9; 20; 40; 14

Figure 4.1 shows the survival curve for Radiotherapy and Radiation + Chemotherapy using Weibull distribution and Turnbull method. It is clear from the figure that the estimated survival curves obtained by our Weibull model lies close to the one obtained by Turnbull. These results indicated that our model is similar to Turnbull method. Parameters estimates (shape and scale) of Weibull distribution and standard errors (se) of the two treatments are presented in Table 4.2. Moreover, the likelihood ratio test for this model is (12.86, with P-value is almost zero). In next section will be implement our model using education data set.

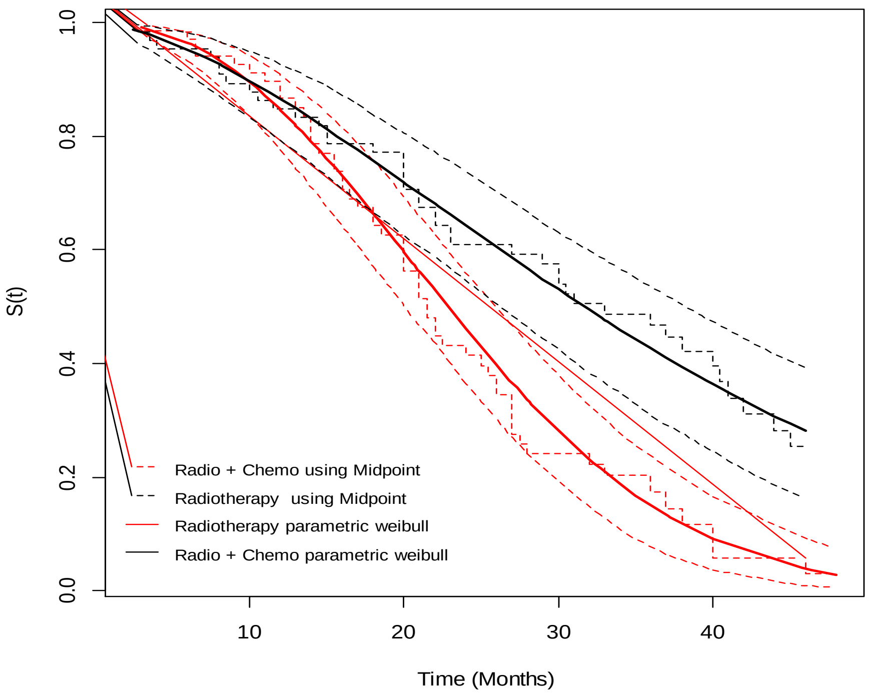

Figure 4.1: The survival function obtained by our Weibull model and Turnbull methods based on PIC data.

Table 4.2: The parameters estimated based on Weibull distribution for breast cancer PIC Data.

PICD		*est	L95%	U95%	**se
All patients	shape	1.858	1.574	2.193	0.157
	scale	32.402	29.095	36.085	1.780
Radiotherapy	shape	1.620	1.240	2.110	0.220
	scale	39.690	32.720	48.160	3.920
Radio + Chemo	shape	2.195	1.785	2.699	0.232
	scale	26.786	23.716	30.254	1.664

*est: estimation **se: standard error

4.2 EDUCATION DATA SET

The education data set will be used to implement the methods in this study that is the data set collected from the Ministry of Education and Higher Education in Qatar, reader referred to chapter 3 for more detail. In this study, we will compare the duration rate of students staying in school for the local and international students and the time represent the event of interest. All the students were followed up from grade 7 until graduated in grade 12 where the graduate year represent the event that recorded as exact value if available. We consider one year as interval and the exact value will be given from data set. In this situation, we consider the data as PIC data as well as interval data. In the next sections, we will discuss the result obtained for right censored and PIC.

Right Censored

We will implements the method that discussed in chapter 3 using the above education data set when the data is right censored. The survival curve of student using Weibull distribution show that the local students have slightly longer survival compared with international students as presented in Figure 4.2 which indicate that the local students have less drop from school than the others.

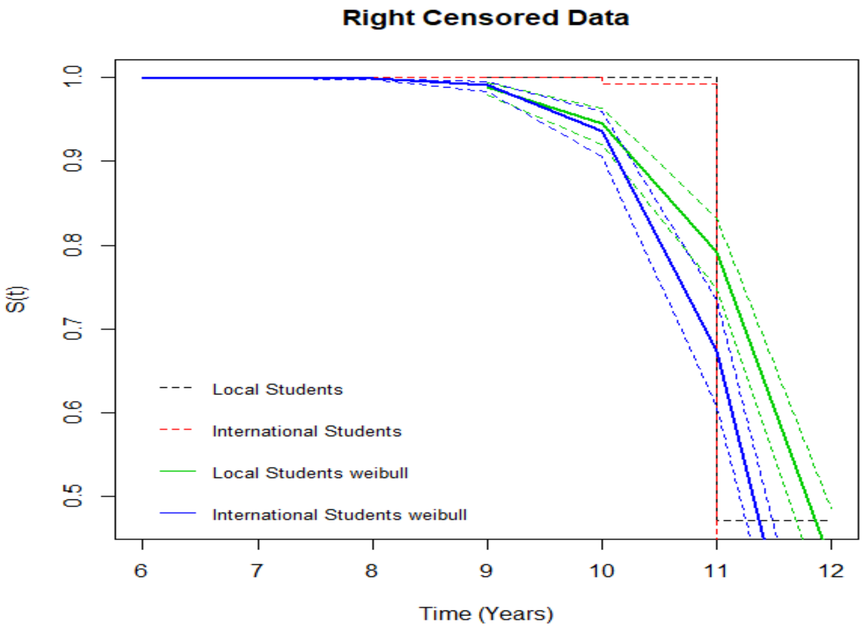

Figure 4.2: Estimated of survival function using Weibull distribution based on Right Censored Data

Table 4.3 shows the results from R software that present the shape parameter for all students is (16.1034) with standard deviation (0.8855). The shape parameter estimate for local students and international students is 15.0607 (se=1.1707) and 18.8465 (1.4057) respectively. Moreover, the scale parameter of all students, local students and international students are almost similar with small standard deviation.

The result shows that the parameters estimate of local students and international students are almost similar. Furthermore, the likelihood ratio test for this model is (19.95) and zero P-value.

Table 4.3: The parameters estimated using Weibull distribution based on Right Censored Data

Right Censored		est	L95%	U95%	se
All Students	shape	16.1034	14.4582	17.9358	0.8855
	scale	11.8879	11.7915	11.9851	0.0494
Local Students	shape	15.0607	12.9324	17.5394	1.1707
	scale	12.1143	11.9642	12.2662	0.0771
International Students	shape	18.8465	16.2833	21.8132	1.4057
	scale	11.5563	11.4411	11.6726	0.0591

Partly Interval Censored Data

In this section we will be implemented two scenarios for the PIC data. We have 50% exact observed and 50 interval observed data in the first scenario where by in the second scenario we have 70 exact and 30 interval observed data. In contrast, we estimate the survival function for the two failures time that is the local and international students based on right point and left point imputations based on PIC data.

Figure 4.3 and Figure 4.4 show similar survival curve for local and international students based on 50% exact observed data obtained by right and left point, respectively. Furthermore, Figure 4.5 and Figure 4.6 show the survival curve of local and international students when observed 70% of exact observed data that obtained also by right and left point, respectively. The survival curve on those Figures represent that the local students have slightly longer survival compared with international students especially on the right point. Which indicate that the international student have more drop from the school than the local students.

The shape and scale parameter for all students, local students and international students look similar that are as shown in Tables 4.4, 4.5, 4.6 and 4.7. The results also show that the likelihood ratio test obtained by right point for two scenarios are (27.07) and (30.33), respectively. Also, the likelihood ratio test obtained by left point for two scenarios is (19.95) in both scenario. Moreover, all likelihood ratio test for this model PIC is small P-value which implies significant fit the model. This result indicate that the Weibull model is significant enough based on the P-value obtained.

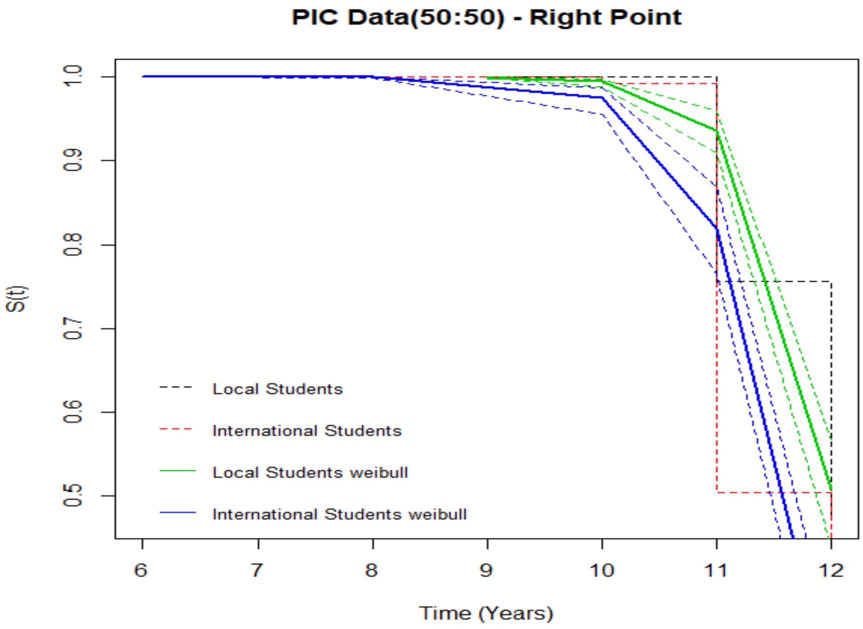

Figure 4.3: The survival function based on PIC data (50:50) _ Right point imputation.

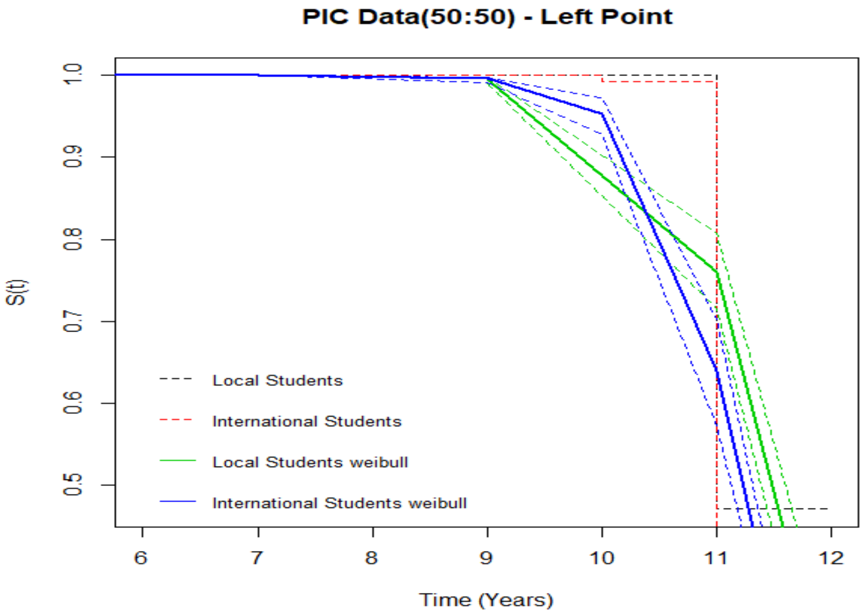

Figure 4.4: The survival function based on PIC data (50:50) _ Left point imputation.

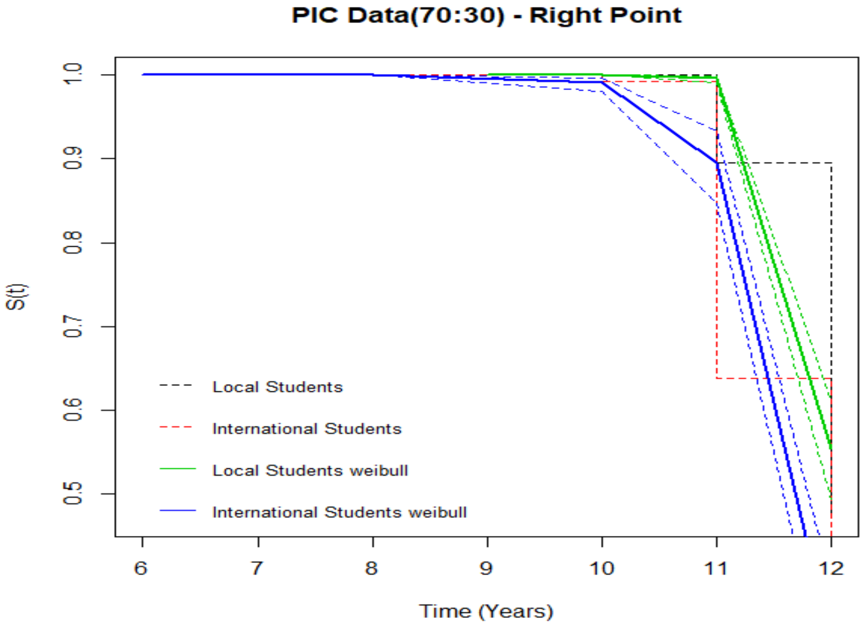

Figure 4.5: The survival function based on PIC data 70:30 (70% exact)_ Right point imputation.

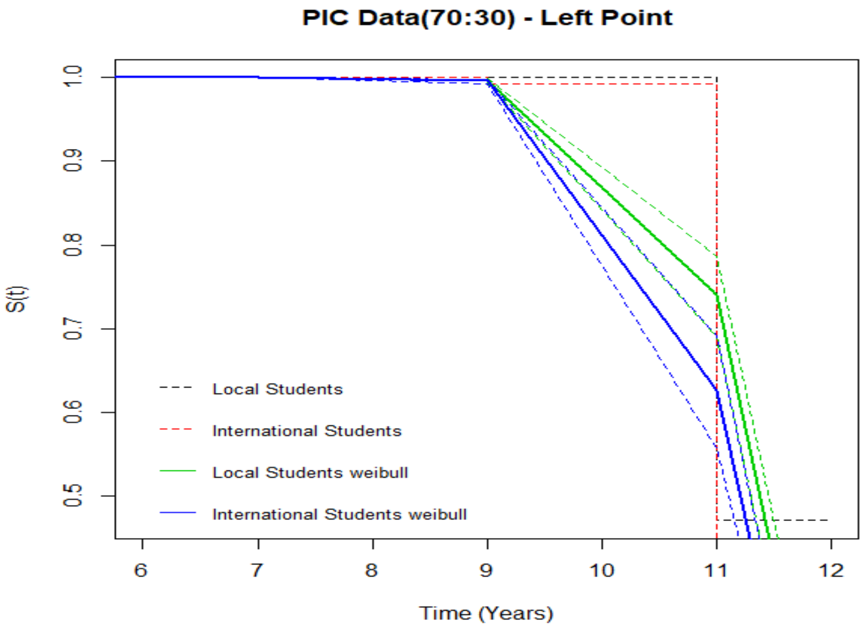

Figure 4.6: The survival function based on PIC data 70:30 (70% exact) _ Left point imputation.

Table 4.4: The parameters estimated based on Weibull distribution from PIC Data (50:50) - Right Point

Partly Interval Censored Data (50:50) - Right Point		est	L95%	U95%	se
All Students	shape	23.4833	20.9383	26.3377	1.3744
	scale	12.0687	12.0012	12.1365	0.0345
Local Students	shape	26.7436	22.7432	31.4476	2.2109
	scale	12.1752	12.0891	12.2619	0.0441
International Students	shape	21.7677	18.5794	25.5030	1.7589
	scale	11.8458	11.7435	11.9489	0.0524

Table 4.5: The parameters estimated based on Weibull distribution from PIC Data (50:50) - Left Point

Partly Interval Censored Data (50:50) - Left Point		est	L95%	U95%	se
All Students	shape	20.6814	18.8014	22.7494	1.0056
	scale	11.5996	11.5278	11.6719	0.0368
Local Students	shape	19.5738	17.1239	22.3742	1.3354
	scale	11.7499	11.6430	11.8577	0.0548
International Students	shape	23.4900	20.5950	26.7930	1.5760
	scale	11.3820	11.2910	11.4750	0.0470

Table 4.6: The parameters estimated based on Weibull distribution from PIC Data (70:30) - Right Point

Partly Interval Censored Data (70:30) - Right Point		est	L95%	U95%	se
All Students	shape	35.6397	31.5554	40.2526	2.2133
	scale	12.0900	12.0449	12.1353	0.0231
Local Students	shape	58.1418	48.9106	69.1153	5.1288
	scale	12.1090	12.0680	12.1501	0.0209
International Students	shape	26.3643	22.3207	31.1405	2.2396
	scale	11.9379	11.8528	12.0236	0.0436

Table 4.7: The parameters estimated based on Weibull distribution from PIC Data (70:30) - Left Point

Partly Interval Censored Data (70:30) - Left Point		est	L95%	U95%	se
All Students	shape	23.5433	21.5283	25.7469	1.0748
	scale	11.4866	11.4242	11.5493	0.0319
Local Students	shape	23.0743	20.4000	26.0991	1.4502
	scale	11.5874	11.4992	11.6763	0.0452
International Students	shape	24.6682	21.6870	28.0592	1.6211
	scale	11.3437	11.2565	11.4315	0.0446

4.3 SIMULATION DATA

Rubinstein, (1981) described the simulation study as a technique that performing experiments on a computer that involve certain types of mathematical models for which explain the behavior of a certain system. To evaluate and study the behavior of statistical procedures in statistic we mostly used the simulation especially for the situation when a problem cannot be solved analytically (Elfaki, et al. (2017)).

The technique consists of setting up a large number of samples. The samples are then individually reckoned in terms of statistics of interest, and the overall statistics of interest is used to study distribution properties. The simulations can also be used to generate estimates of the mean, variance, coverage probability of confidence intervals. The objective of this simulation study is to compare the survival function for local and international students as well as different type for partly interval censored and right censored based on Weibull model.

The simulated data were generated based on the education data set with two failure times that the local and international students. The Weibull distribution is used to generate the data (used because the Weibull was found to be fit the data well based on early results). To generate the data we used the mean and standard deviation of the shape and scale parameters that were presented in Tables 4.3, 4.4, 4.5, 4.6 and 4.7). The data generated was for 500 for local students and international students. In the next section, we will discuss the result obtained for right censored and PIC.

Right Censored Data

Figure 4.7 and Table 4.8 showed the estimated survival function and the parameters from simulation with right censored. The results show that the survival curves data lies very close to Weibull curves. The local students showed to be have along survival compare on to international students which is indicate that the local students more stable. In additional that the shape and scale parameters for local and international student are almost similar. Moreover, the likelihood ratio test show to be 30.29, with P-value equal zero which is indicate that the model was fit well.

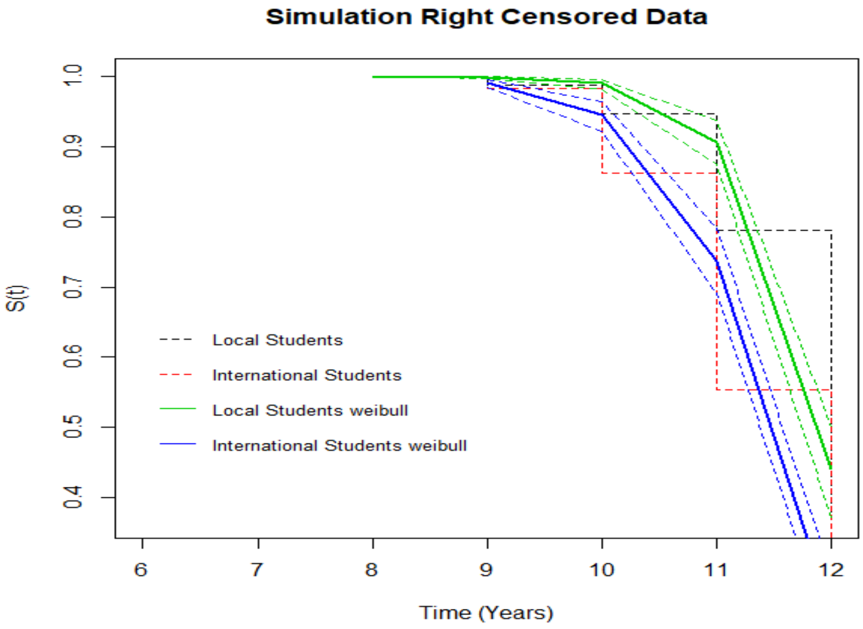

Figure 4.7: The survival function based on Weibull distribution from right censored data.

Table 4.8: The parameters estimated based on Weibull distribution from simulation data with right censored.

Right Censored		est	L95%	U95%	se
All Students	shape	19.5370	17.7535	21.4996	0.9542
	scale	11.9522	11.8836	12.0212	0.0351
Local Students	shape	24.4585	21.0071	28.4770	1.8983
	scale	12.0969	12.0103	12.1840	0.0443
International Students	shape	17.8576	15.8080	20.1728	1.1107
	scale	11.7570	11.6596	11.8553	0.0499

Partly Interval Censored Data

We following similar situation as in the real data in early section. There are two scenarios in for which the first scenario will take 50% exact observed and 50% the observed are interval. The second scenario will take the exact observed of 70% and 30% for the observed as interval. The result of this partly interval censored shows in Figure 4.8, 4.9, 4.10 and Figure 4.11. There figures show that the survival curves obtaining by Weibull for the local students and international students, respectively. The figures indicate the that the result from the simulation study are similar as in the education data set.

Table 4.9 and Table 4.11 show the shape and scale parameter obtained by our model and the results look similar for all students, local students and international students that obtained by right point in two scenarios. Likewise, the estimated parameters are similar that obtained by left point imputation for two scenarios with respect standard error (Tables 4.10 & 4.12).

Moreover, the likelihood ratio test obtained by right point for two scenarios are 38.11 (0) and 35.6 (0), respectively. Similarity, the likelihood ratio test obtained by left point for two scenarios are 38.48 (0.) and 40.38 (0), respectively, which implement the significant of the model.

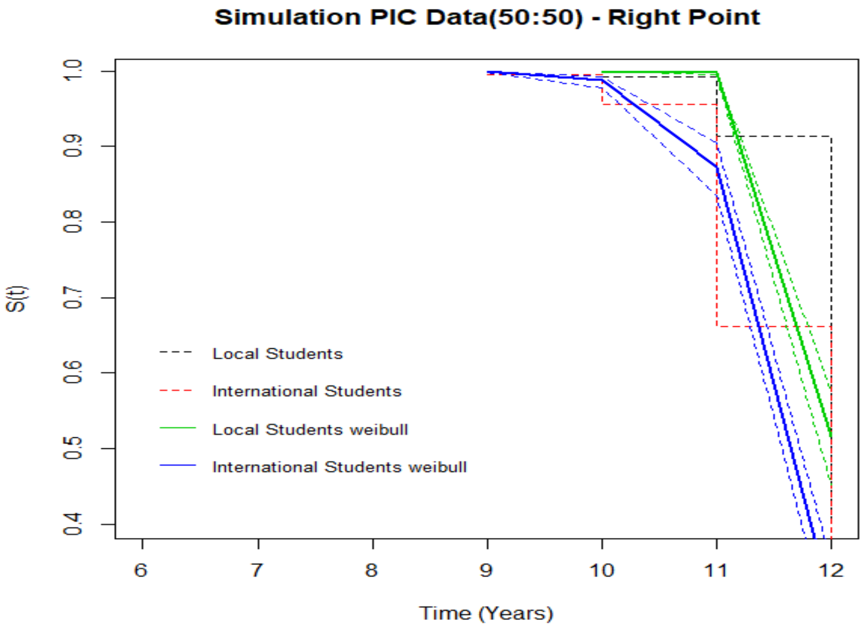

Figure 4.8: The survival function based on PIC data (50:50) _ Right point imputation.

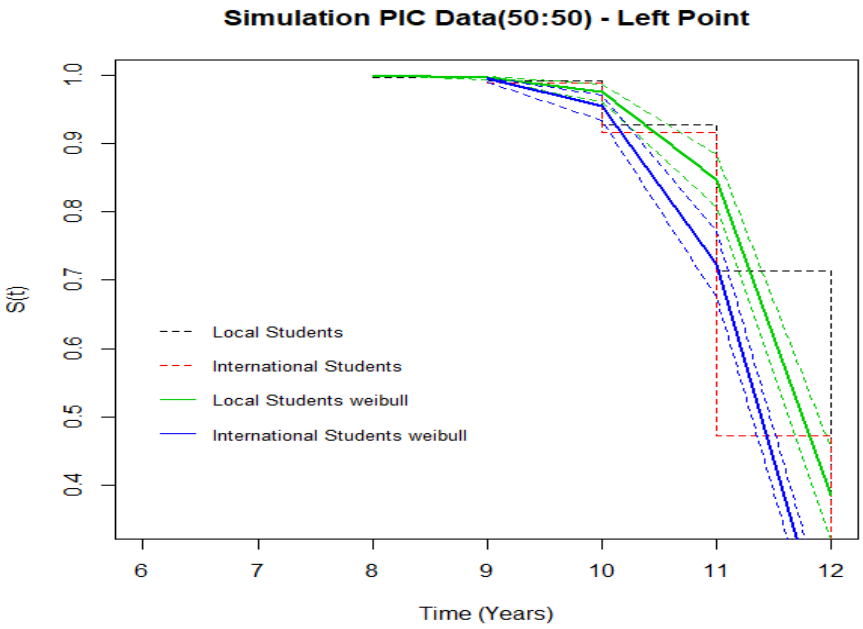

Figure 4.9: The survival function based on PIC (50:50) _ Left point imputation.

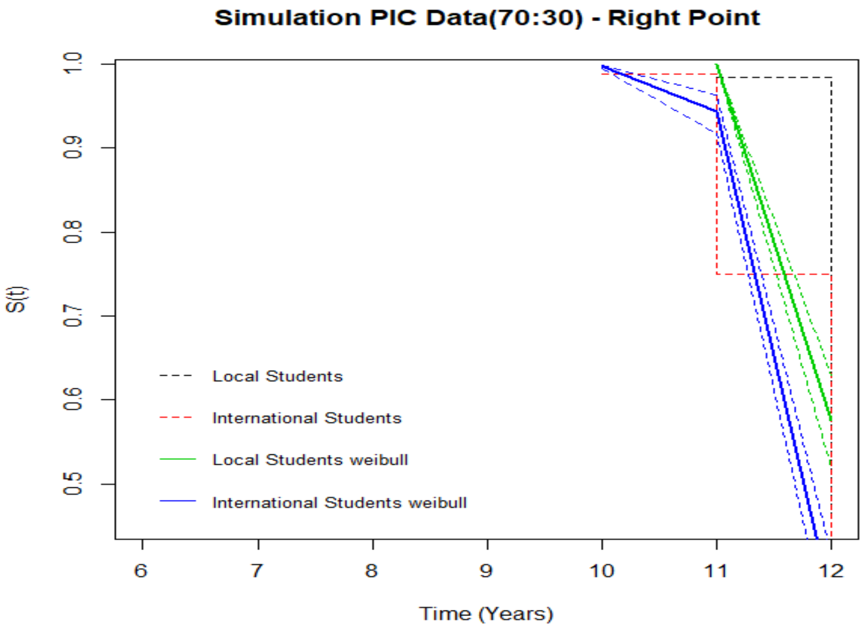

Figure 4.10: The survival function based on PIC data 70:30 (70% exact) _ Right point imputation.

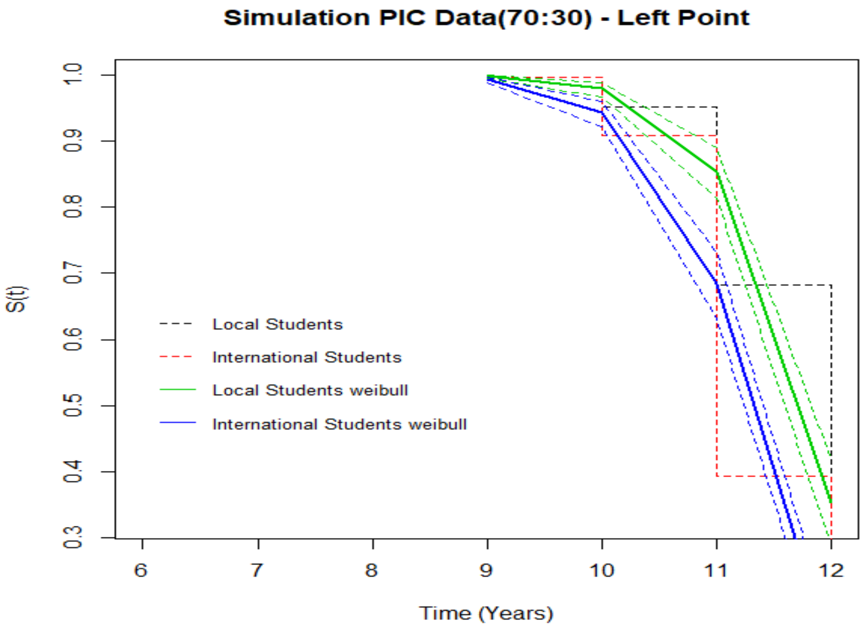

Figure 4.11*:* The survival function based on PIC data 70:30 (70% exact) _ Left point imputation.

Table 4.9: The parameters estimated based on Weibull distribution from simulation with PIC Data (50:50) - Right Point

Partly Interval Censored Data (50:50) - Right Point		est	L95%	U95%	se
All Students	shape	32.2031	29.1061	35.6297	1.6614
	scale	12.0375	11.9959	12.0793	0.0213
Local Students	shape	66.1562	55.9205	78.2654	5.6736
	scale	12.0753	12.0418	12.1088	0.0171
International Students	shape	25.1130	22.1534	28.4679	1.6067
	scale	11.9086	11.8401	11.9775	0.0351

Table 4.10:The parameters estimated based on Weibull distribution from simulation with PIC Data (50:50) - Left Point

Partly Interval Censored Data (50:50) - Left Point		est	L95%	U95%	se
All Students	shape	19.6628	17.9674	21.5183	0.9046
	scale	11.8267	11.7595	11.8943	0.0344
Local Students	shape	20.0321	17.3228	23.1652	1.4852
	scale	12.0285	11.9216	12.1364	0.0548
International Students	shape	20.3040	18.1430	22.7220	1.1650
	scale	11.6260	11.5420	11.7110	0.0430

Table 4.11: The parameters estimated based on Weibull distribution from simulation with PIC Data (70:30) - Right Point

Partly Interval Censored Data (70:30) - Right Point		est	L95%	U95%	se
All Students	shape	51.7831	46.4364	57.7454	2.8793
	scale	12.0663	12.0392	12.0933	0.0138
Local Students	shape	387.8800	328.8260	457.5400	32.6870
	scale	12.0180	12.0120	12.0240	0.0030
International Students	shape	32.9567	28.7980	37.7159	2.2681
	scale	11.9926	11.9386	12.0469	0.0276

Table 4.12:The parameters estimated based on Weibull distribution from using simulation with PIC Data (70:30) - Left Point

Partly Interval Censored Data (70:30) - Left Point		est	L95%	U95%	se
All Students	shape	19.7145	18.0668	21.5124	0.8779
	scale	11.7745	11.7091	11.8402	0.0335
Local Students	shape	21.6267	18.8002	24.878	1.5454
	scale	11.9787	11.8836	12.0745	0.0487
International Students	shape	19.7644	17.7192	22.0458	1.1016
	Scale	11.5507	11.4662	11.6359	0.0433

CHAPTER 5

CONCLUSION AND SUGGESTION FOR FUTURE RESEARCH

CHAPTER OVERVIEW

Two sections will be present in this chapter. The conclusion, which summaries the result obtained in the previous chapters will be presents in first section. Suggestion for future studies presented in second section.

5.1 CONCLUSION

In this study, we used Weibull model based on simple imputation technique to simplify the procedure for partly interval censored data. Weibull distribution model have been mostly applied in engineering and medical application. In this research project we used it for medical data and education data.

In this research project, the estimated survival function was obtained based on the maximum likelihood estimation and comparisons were made with existing one under the Turnbull (chapter three).

The first step of this study is to look for secondary data set to conform our model well fit to used. Thus, we used the medical data that was modified by Zyoud, et al. (2016) to partly interval censored. Based on the result from this data set, we found that our model fit well and easy to implement compared with the one obtained by Turnbull method.

In the second step, the education data set was used to implement our methods. The data was collected from the Ministry of Education and Higher Education in Qatar. The data was followed up the students in National Student Information System (NSIS) from grade 7 until graduation from grade 12 in period from September-2010/June-2011 school academic year to September-2015/June -2016. Moreover, the time in years represent the event of interest until the students graduate from grade 12. In addition, the interval of one year period and the exact value will be given from data set. Then, the data set become partly interval censored data as well as interval data.

Overall the result from education data has shown that, from survival curves for the two failure time that the local students and the international students. The local students showed to have longer survival compare with international students. The scale parameter is almost similar for the different type of censored. On the contrary, we note that the estimated shapes are slight difference for the different types of censored. The results indicate that the local students are more stable in school and they less drop from school compare with international.

For the short period of interval data the simple imputation is better to used, because it is easy to used and always the result reliable enough. However, education partly interval censored data is preferable to used compared with medical data, because in the education data can easy control the exact data and interval data but for medical data dealing with observation that has a lot of missing observation.

The simulation data was used based on the education data. The data generated for 500 times from international and local students. We conclude that the simulation results is similar to the results obtained from education data for the both scenarios. Moreover, the education data is suitable for partly interval censored.

Finally, the result observed that when the observed have more exact in the data the model is better fit which is same line with other results obtained some researchers such as Kim (2003), Zyoud et al. (2016) and Alharpy and Ibrahim (2013). The simulation study strongly supports if the data is partly interval censored then the Weibull model is better compared with interval data.

5.2 SUGGESTION FOR FUTURE RESEARCH

This study can be extending to look into the properties of more than two parameters in the model. Also, can used different sample sizes with more factors in the data such as age, gender and others. Moreover, the type of data will be more accurate if the data is available from grade 1 to grade 12.

REFERENCES

Alharphy and Ibrahim (2013), Parametric Test for Partly Interval-censored Failure Time Data under Weibull Distribution via Multiple Imputation. *Journal of Applied Science*, 13(4), pp 621-626.

Balakrishnan and Kateri (2008), On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data, *Statistics and Probability Letters, 78, pp 2971_2975*

Eagle and Barnes (2014), Survival Analysis on Duration Data in Intelligent Tutors, *Intelligent Tutoring Systems, 8474,* pp 178-187

Elfaki, Abobakar, Azram and Usman (2013), Survival Model for Partly Interval-Censored Data with Application to Anti D in Rhesus D Negative Studies*, International Scholarly and Scientific Research & Innovation, 7, pp 347-350.*

Elfaki, Azram and Usman (2012) Parametric Cox's Model for Partly Interval-Censored Data with Application to AIDS Studies. *International Journal of Applied Physics and Mathematics*, 2(5), pp 352-354.

Elfaki, Bin Daud, Ibrahim, Abdullah and Usman, (2007), Competing risks for reliability analysis using Cox's model, Engineering Computations, Vol. 24 Issue: 4, pp.373-383

Farnum and Booth (1997), Uniqueness of Maximum Likelihood Estimators of the 2-Parameter Weibull Distribution, *IEEE Transactions on Reliability, 46, pp 523-525.*

Giolo (2004), Turnbull's Nonparametric Estimator for Interval-Censored Data. *Department of Statistics, Federal University of Paraná, researchgate.net.*

Guure, Ibrahim and Adam (2012), On partly censored data with the Weibull distribution. *ScienceAsia*, 39S: 75–79.

Harter and Moore (1965), Maximum-Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and from Censored Samples, *Taylor and Francis, 7, pp 639-643.*

Kim (2003), Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *J R. Statist. Soc*., Series B65, pp 489-502.

Kleinbaum and Klein (2005). Survival analysis: A self-learning text. *Springer, NewYork, NY.*

Lawless (2003), Statistical Models and Methods for Lifetime Data. *John Wiley & Sons, Canada*

Lee and Wang (2003), Statistical Methods for Survival Data Analysis, *John Wiley & Sons, Canada.*

Murthy, D. P., Xie, M., and Jiang, R. (2004), Weibull Models, *John Wiley and Sons, Inc*.

Odell, Anderson and D'Agostino, (1992), Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model, *International Biometric Society, 48, pp 951-959.*

Peto R. and Peto, J. (1972), Asymptotically Efficient Rank Invariant Test Procedures".*J R. Statist. Soc*., Series A135, pp. 187-220.

Plank, DeLuca and Estacion (2008), High School Dropout and the Role of Career and Technical Education A Survival Analysis of Surviving High School, *Sociology of Education, 81, pp 345-370.*

Rinne (2008), The Weibull Distribution. *New York: Chapman and Hall/CRC.*

Rubinstein, Reuven Y.,(1981) Simulation and the Monte Carlo Method, John Wiley & Sons, New York-Chichester-Brisbane Toronto, pp. 278.

Sam and Krongs (2008), Survival Analysis Approach to Reliability, Survivability and Prognostics and Health Management (PHM). *IEEE Aerospace Conference.*

Seguro and Lambert (2000) Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics, 85, pp 75-84.*

Singh and Totawattage (2013), The Statistical Analysis of Interval-Censored Failure Time Data with Applications. *Journal of Statistics, 3, pp 155-166.*

Singer and Willett (1993), It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics, 18(2), pp 155-195.*

Thamrin (2013), Bayesian Survival Analysis Using Gene Expression, *eprints.qut.edu.au*

UIS (2009), Education indicators: technical guidelines, *UNESCO Institute for Statistics, UIS, Montreal. (http://uis.unesco.org/sites/default/files/documents/education-indicators-technical-guidelines-en_0.pdf)*

Weybright, Caldwell, Xie, Wegner and Smith, (2017), Predicting secondary school dropout among South African adolescents: A survival analysis approach. *South African Journal of Education, 37, pp 1-11.*

Wu, Chambers and Xu, (2017), arXiv: 1708.06838*, https://arxiv.org/abs/1708.06838*

Xian Liu (2012), Survival Analysis Models and Applications. *John Wiley & Sons, United Kingdom*

Yin, Huang, Li and Zhou (2013), A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia, *IEEE Transactions on Knowledge and Data Engineering, 25, pp 1201-1212.*

Yousif, Elfaki and Hrairi (2016), Analysis of Partly Interval-Censored Data under Competing Risks Framework, *International Journal of Computer Science and Network Security, 16, pp* 25-28.

Zhao, Zhao, Sun and Kim, (2008), Generalized Log-Rank Tests for Partly Interval-Censored Failure Time Data. *Biometrical Journal, 50, pp 375–385.*

Zyoud, Elfaki and Hrairi (2016), Non-parametric estimate based in imputations techniques for interval and partly interval censored data, *Science International (Lahore), 28, pp 879-884.*

APPENDICES

R CODE FOR OBTAINING THE SURVIVAL CURVES

APPENDIX A: RIGHT CENSORED DATA

```
require(survival)
require(MASS)
require("flexsurv")
year <- dat$year - 2005
eve <- year
cens <- dat$Censored
ident <- dat$Nationality

for (i in 1:length(ident)){if(ident[i]==2) {ident[i]= 0}}
data = data.frame(eve, cens, ident)
dat1 <- data[data$ident==1,]   #Local Students
dat2 <- data[data$ident==0,]   #International Students

y=Surv(data$eve,data$cens==1)
kmfit1=survfit(y~1)
summary(kmfit1)

survdiff(Surv(eve,cens)~ident,data=data)
x=coxph(y~ident,data=data)
```

```
summary(x)
y1=Surv(dat1$eve,dat1$cens==1)
kmfit2=survfit(y1~1)
summary(kmfit2)

y2=Surv(dat2$eve,dat2$cens==1)
kmfit3=survfit(y2~1)
summary(kmfit3)

plot(kmfit2$time,kmfit2$surv,type="s",main="Right Censored Data",col=1,lty=2, xlab="Time (Years)",ylab="S(t)",xlim=range(c(0,12)))
lines(kmfit3$time,kmfit3$surv,type="s",col=2,lty=2)
legend(1,0.65,lty=2,col=1,"Local Students", bty="n",cex=0.8)
legend(1,0.6,lty=2,col=2,"International Students", bty="n",cex=0.8)

#All Students
est = flexsurvreg(Surv(data$eve,data$cens)~1, dist="weibull")
est

#Local Student
est1 = flexsurvreg(Surv(dat1$eve,dat1$cens)~1, dist="weibull")
est1

#International Student
est2 = flexsurvreg(Surv(dat2$eve,dat2$cens)~1, dist="weibull")
```

```
est2


lines(est1 , col=3, lty=1, xlab="Time (Years)",ylab="S(t)")

lines(est2 , col=4, lty=1, xlab="Time (Years)",ylab="S(t)")

legend(1,0.55,lty=1,col=3, "Local Students weibull", bty="n",cex=0.8)

legend(1,0.5,lty=1,col=4, "International Students weibull", bty="n",cex=0.8)
```

APPENDIX B: PARTLY INTERVAL CENSORED DATA

```
require(survival)

require("flexsurv")


year <- dat$year - 2005

eve <- year

left = double(length(eve))

right = double(length(eve))


for (i in 1:length(eve)){

if (i<196) {

left[i]=(ceiling(eve[i]/2))*2-1

right[i] =(ceiling(eve[i]/2))*2

} else {

left[i] =eve[i]

right[i] =eve[i]

}}

cens <- dat$Censored

ident <- dat$Nationality

for (i in 1:length(ident)){if(ident[i]==2) {ident[i]= 0}}

data = data.frame(left, right, cens, ident)

eve <- data$right
```

```
data = data.frame(eve, cens, ident)
dat1 <- data[data$ident==1,]     #Local Students
dat2 <- data[data$ident==0,]     #International Students

y=Surv(data$eve,data$cens==1)


kmfit1=survfit(y~1)
summary(kmfit1)
survdiff(Surv(eve,cens)~ident,data=data)
x=coxph(y~ident,data=data)
summary(x)

y1=Surv(dat1$eve,dat1$cens==1)
kmfit2=survfit(y1~1)
summary(kmfit2)

y2=Surv(dat2$eve,dat2$cens==1)
kmfit3=survfit(y2~1)
summary(kmfit3)

plot(kmfit2$time,kmfit2$surv,type="s",main="PIC Data(50:50) - Right Point",col=1,lty=2,
xlab="Time (Years)",ylab="S(t)",xlim=range(c(0,12)))
lines(kmfit3$time,kmfit3$surv,type="s",col=2,lty=2)
legend(1,0.65,lty=2,col=1,"Local Students", bty="n",cex=0.8)
legend(1,0.6,lty=2,col=2,"International Students", bty="n",cex=0.8)
```

```
#All Students

est = flexsurvreg(Surv(data$eve,data$cens)~1, dist="weibull")

est


#Local Student

est1 = flexsurvreg(Surv(dat1$eve,dat1$cens)~1, dist="weibull")

est1


#International Student

est2 = flexsurvreg(Surv(dat2$eve,dat2$cens)~1, dist="weibull")

est2


lines(est1 , col=3, lty=1, xlab="Time (Years)",ylab="S(t)")

lines(est2 , col=4, lty=1, xlab="Time (Years)",ylab="S(t)")

legend(1,0.55,lty=1,col=3, "Local Students weibull", bty="n",cex=0.8)

legend(1,0.5,lty=1,col=4, "International Students weibull", bty="n",cex=0.8)
```