PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

MOHAMED KHIDER UNIVERSITY, BISKRA

Laboratory of Applied Mathematics

Department of Mathematics

Submitted in partial fulfillment of the requirements for Doctorate degree in MATHEMATICS

Option: NONPARAMETRIC STATISTICS

On estimation of the hazard function for doubly truncated data

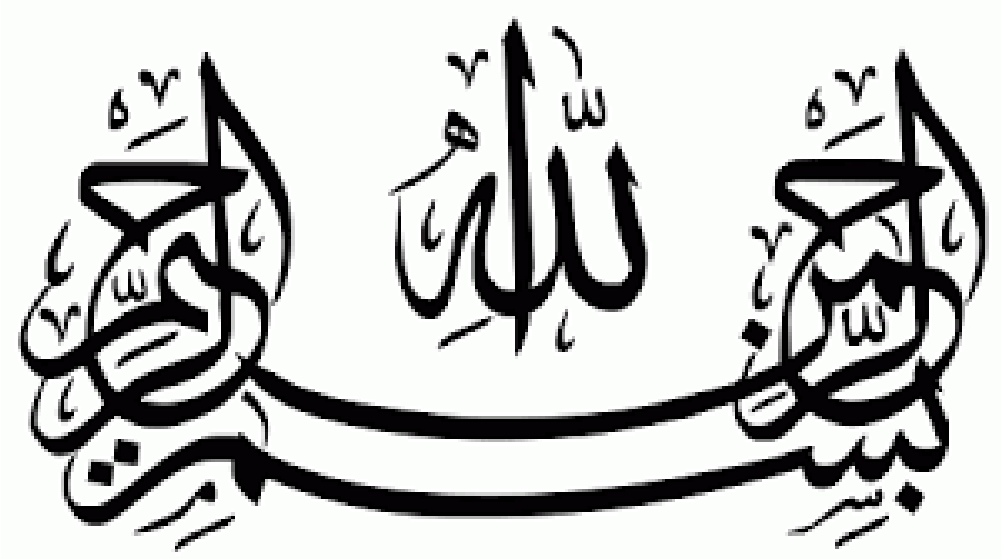
---

Presented by *ELBAY Roumaissa*

Examination Committee :

Mr.BENATIA Fatah	Professor	Biskra University	President
Mr.YAHIA Djabrane	Professor	Biskra University	Supervisor
Mr.SAYAH Abdallah	Professor	Biskra University	Examiner
Mr.DJENAIHI Youcef	MCA	Sétif 1 University	Examiner

defended publicly the


بسم الله الرحمن الرحيم

ACKNOWLEDGEMENTS

Praise be to the Almighty Alla'h
who has given me faith, courage, health, and patience to carry out this work.
I want to express my deep gratitude to my supervisor Prof. **YAHIA Djabrane** from Biskra university, for the confidence he has placed in me, through his presence always with me, by his direction, his modesty, his advice, and constructive remarks for the good progress of this work.
I would like to thank the members of examination committee: Prof. **BENATIA Fatah**; Prof. **SAYAH Abdallah**; Prof. **DJENAIHI Youcef** who they accepted the evaluation and discussion of my thesis to improve this work.
I express my deep gratitude to my mother and my father soul, my sister, my brothers, and my whole family for their encouragement and prayers that allowed me to achieve this modest work. I am very grateful for the confidence they have placed in me.
Finally, I express my gratitude to all my freinds .
O Alla'h, send your blessings on your noble messenger, his family, and companions, and bless us in our life.

Dedication

I dedicate this work to the memory of my dear father ammar :

who was always passionate about knowledge and learning

To my mother who give me love and liveliness.

To my sister and brothers and my brother's wife Khadidja and my sister's husband Amine and their families especially grandchildren: Mohamed, Sirine, Chahed, Takwa, Hind, Cham .

To all those who have helped me and those who shared with me the emotional moments during the accomplishment of this work and who warmly supported and encouraged throughout my journey.

To all my friends who have always encouraged me, including all the members of the LMA laboratory.

Thanks!

Roumaissa ELBAY

Abstract

In this thesis, we investigate the problem of incomplete data, specifically the phenomenon of double truncation, which make working with classical methods very hard, as truncation mean the loss of samples during the statistical analysis, and leads to negative results of the study and wrong decisions. Specifically, we focused in this thesis on estimating of the hazard function in the case of double truncation, where estimating the hazard function estimator is defined in many previous work, hence we make comparison with the hazard functions known in this case and our proposed estimator, and thus we result that the proposed hazard function estimator is more accurate through applied and theoretical comparison. In addition, a smoother cumulative distribution function estimator was proposed, as the previously proposed estimator of the distribution function it was not smooth and not continuous. In this context, several methods were also proposed to obtain the smoothing parameter for the cumulative distribution function within the data subject to double truncation.

Résumé

Dans cette thèse, nous avons étudié le problème des données incomplètes, en particulier le phénomène de double troncature, car la troncature signifie la perte d'échantillons au cours de l'analyse statistique, ce qui affecte négativement les résultats de l'étude. Nous nous sommes concentrés sur l'estimation de la fonction de risque dans le cas de double troncature, où un estimateur de fonction de risque plus précis a été donné, et cela apparaît en s'appuyant sur la comparaison avec les fonctions de risque définies dans ce cas, arrivant ainsi à la conclusion que nous estimateur de la fonction de risque est plus précis grâce à une comparaison appliquée et théorique. Grâce à cette recherche, un estimateur de fonction de répartition plus lisse a également été proposé, car l'estimateur de la fonction de répartition proposé précédemment signifiait qu'il n'était ni lisse ni continu. Dans ce contexte, plusieurs méthodes ont également été proposées pour obtenir le paramètre de lissage de la fonction de répartition dans le cas au les données soumises à double troncature.

CONTENTS

List of Figures

List of Tables

Symbols and Acronyms

General introduction **1**

1 *Censoring and truncation* **4**

1.1 *Preliminary definitions* 5

1.2 *Censoring* 7

1.2.1 *Types of censoring* 7

1.2.2 *A guide to defining likelihood functions with censored data* 12

1.2.3 *Techniques for estimation in the presence of right censored data* . . . 13

1.3 *Truncation* 15

1.3.1 *Types of truncation* 15

1.3.2 *The definition of the likelihood functions with truncation data* 18

1.3.3 *Techniques for estimation in the presence of right truncated data* . . 19

2 *Estimation under double truncation* **20**

2.1 *The definition of probability under double truncation* 21

2.2 *The construction of the likelihood function under double truncation data* . . . 22

2.2.1 *Nonparametric consideration in estimation* 22

2.2.2 *Semiparametric consideration in estimation* 26

2.2.3 Bootstrap method . . . 28
2.2.4 Particular case of double truncation: Fixed-Length . . . 28
2.3 Kernel density estimator . . . 29
2.3.1 *Asymptotic properties* . . . 29
2.3.2 *Selection of optimal bandwidth for kernel density estimator* . . . 31
2.4 *Kernel estimation of the cumulative distribution function* . . . 35
2.4.1 *Asymptotic properties* . . . 36

3 Hazard function for doubly truncated data **38**
3.1 *The NPMLE of Hazard function* . . . 39
3.2 *The smooth estimator of hazard function* . . . 42
3.3 *The proposed estimator* . . . 44
3.4 *Asymptotic properties of the proposed estimator* . . . 46

4 *Simulation* **48**
4.1 *Simulation data* . . . 49
4.2 Analyze the results . . . 51

5 On optimal bandwidth selection **56**
5.1 Kernel smoothing estimation of the distribution function . . . 57
5.1.1 *Normal reference bandwidth for the cumulative kernel distribution function* . . . 58
5.1.2 *Plug in method* . . . 59
5.1.3 *Cross-validation method for define the optimal bandwidth* . . . 61
5.1.4 *Bootstrap bandwidth selection* . . . 62

Bibliography **3**

LIST OF FIGURES

1.1 Right censoring data 9
1.2 Left censoring data 10
1.3 Interval censoring data 11
1.4 Left truncation data 16
1.5 Right truncation data 17
1.6 Example of double truncation data: The data on German companies from . 18

LIST OF TABLES

4.1 b_{AMISE} is by normal reference for the distribution function for model 1.1 . . 50
4.2 The bias and RMSE for the distribution function for model 1.1 50
4.3 b_{AMISE} is by normal reference for the distribution function for model1.2 . . 50
4.4 The bias and RMSE for the distribution function for model 1.2 51
4.5 b_{AMISE} is by normal reference for the distribution function for model1.3 . . 51
4.6 The bias and RMSE for the distribution function for model 1.3 51
4.7 b_{AMISE} is by normal reference for the distribution function for model2.1 . . 52
4.8 The bias and RMSE for the distribution function for model 2.1 52
4.9 b_{AMISE} is by normal reference for the distribution function for model2.2 . . 52
4.10 The bias and RMSE for the distribution function for model 2.2 53
4.11 b_{AMISE} is by normal reference for the distribution function for model3 . . . 53
4.12 The bias and RMSE for the distribution function for model 3 53
4.13 b_{AMISE} is by normal reference for the distribution function for model4 . . . 54
4.14 The bias and RMSE for the distribution function for model 4 54
4.15 b_{AMISE} is by normal reference for the distribution function for model5 . . . 54
4.16 The bias and RMSE for the distribution function for model 5 55

LIST OF SYMBOLS AND ACRONYMS

a.s.,:	almost surely.
e.g.,:	for example.
i.e.,:	that is.
NPMLE:	nonparametric maximum likelihood estimator.
rv:	random variable.
df:	distribution function.
RMSE:	root mean square error.
pdf:	probability density function.
cv:	Cross-validation.
DPI:	Direct plug-in.
iid:	Independent and identically distributed.
MSE:	Mean square error.
MISE:	Mean integrated square error.
NR:	Normal reference.
LNO:	Leave non out estimator.
ISE:	The integrated squared error.
ASE:	Approximate squared error.
b:	The bandwidth.
K:	Kernel function(or weight function).
W:	The cumulative Kernel function.
$(\Omega, \mathcal{A}, \mathcal{F})$:	Probability space.

X:	rv.
$\mathbf{E}(X)$:	Expectation of (or mean of) rv.
$\mathbf{Var}(X)$:	variance of rv.
f_b:	kernel density estimator.
F_b:	Kernel estimator of the distribution function.
F_n:	The NPMLE estimator of the distribution function.
F_n^*:	The ordinal empirical distribution function.
$\mathbb{R}$:	Set of real numbers.
$o(.)$:	$f(x) = o(g(x)) i.e., \lvert f(x)/g(x) \rvert \xrightarrow[x \longrightarrow x_{x_0}]{} 0.$
$O(.)$:	$f(x) = O(g(x)) i.e., f(x)/g(x) \xrightarrow[x \longrightarrow x_{x_0}]{} 1.$
$\mathcal{N}(\mu, \sigma)$:	Normal distribution.
$\forall$:	for all.
A^T:	The transpose of a matrix.

GENERAL INTRODUCTION

Statistics is the science of collecting, analysing and describing data and is one of the most important branches of mathematics used in various fields, including industry, technology, economics and even astronomy. Although precise methods and tools are used to collect data, we are sometimes exposed to problems that lead to partial or total loss of data, resulting in wrong decisions. Moreover, censoring and truncation are two types of incomplete data, as the main difference between these two types is the amount of information of the observation that is lost when it is partially lost (is censoring) or completely lost (is truncation). Where both these types categories in three parts. For censoring is considered for example in survival analysis when the observation can't be noted from the beginning to the end moreover when the observation can be noted from the beginning but in some part of the trail the observation is lost this kind is know as right censoring in contrast when the observation can't be considered in the data from the beginning but it include before the end of the trail and this is the second type of censoring which know as left censoring for the last kind of censoring is considered when both the previous kind are presented in the same trail we said that we have double censoring. Now, from the definition of truncation it can be also divided into three categories, where the observation is included in the study is only if it satisfies certain condition and this condition are known as the condition of truncation limit therefore when observation can't be included in the study because of it short amount is consider as left truncation and when observation has big amount make it out of the trail study is know as right truncation finally when both kinds of truncation exist in the same

study we said that we have double truncation, and throughout this definition we can say that these two kinds of incomplete data make the use of classical technique of statistics more difficult.

In the last few years, a number of researchers have developed new methods of working with incomplete data, where this kind of data need specials mechanism. We refer researchers to these books [[52]–[22]–[9]], which summarise many of the problems in these cases.
Throughout this thesis we focus in double truncation data, where one side truncation data is well defined in [22][19] [25] [26] [20] [44]. In addition, double truncation data has been considered in many papers in recent years where it is used when the data is known just within interval, for this reason Efron and Petrosian [10] establish the nonparametric likelihood in case of double truncation, following this work many research have considered the non parametric likelihood as a solution for solve the problem of working with double truncation data especially when Shen [38] proof the asymptotic properties of the nonparametric likelihood and after that both Xiao and Hudgens [49] and de Uña-Álvarez et al., [8] gives addition information about the NPMLE from the existing and uniqueness. After that many research investigate the problem of define a smooth estimator where the estimator of density is given in Moreira and Uña-Álvarez [29], additionally Moreira and Van Keilegom [32] given the selector of bandwidth when the data are sampling under double truncation, hence from this work Moreira et al., [31] provide an estimator of the hazard function for double truncation data which is defined as convolution between the estimator giving in Efron and Petrosian [10] and the ordinarily kernel function.
Estimating the hazard function under incomplete data has been an attractive topic and many literature reviews have dealt with it and this show in the work of Watson and Leadbetter [47], Tanner and Wong [42], Kim et al., [21], Efron and Petrosian [10], Shen [38], Patil et al., [35], recently [31] where all the research provide an estimator of the hazard function in both cases in censoring and truncation.

In this thesis, we investigate the problem of incomplete data, specifically the phenomenon of double truncation, which make working with classical methods very hard, as truncation mean the loss of samples during the statistical analysis, and leads to negative results of the study and wrong decisions. Specifically, we focused in this thesis on estimating of the hazard function in the case of double truncation, where estimating the hazard function estimator is defined in many previous work, hence we make comparison with the hazard functions

known in this case and our proposed estimator, and thus we result that the proposed hazard function estimator is more accurate through applied and theoretical comparison. In addition, a smoother cumulative distribution function estimator was proposed, as the previously proposed estimator of the distribution function it was not smooth and not continuous. In this context, several methods were also proposed to obtain the smoothing parameter for the cumulative distribution function within the data subject to double truncation.

The structure of this thesis is as follows

- **Chapter1** This chapter provides some concepts and definitions to make this thesis easier to read, we give the definition of incomplete data and its categories (i.e, censoring and truncation), in addition we define the estimator of the distribution function when the data are censoring or truncation from one side and in the last we give definition of the likelihood function in case of incomplete data.

- **Chapter2** We focus in this chapter in double truncation data problem where we defined the nonparametric likelihood estimator and we give the principal results and we defined the kernel estimator of density in this case, in addition we provide a new kernel estimator of the distribution function where we derive its asymptotic properties and also we derive the semiparametric estimator of the distribution function.

- **Chapter3** This chapter is the goal of our work in this thesis, we present in this part the estimators of the hazard function which have defined in case of double truncation, we defined our proposed estimator of the hazard function where we give the proprieties of the proposed estimator and we derive its asymptotic properties.

- **Chapter4** This chapter is related to the previous chapter where we present the finite sample behavior in order to make the comparison of the existing estimators and our proposed estimator and this by the program R.

- **Chapter5** In this chapter we proposed the bandwidth selector of the distribution function when the data are sampling under double truncation.

CHAPTER 1

CENSORING AND TRUNCATION

In many scientific fields, collecting and analysing data is an essential part of knowing the results of the studies. However, sometimes we suffer with problems of the data, because sometimes it's impossible to collect data, and some data can't be follow or lost. For this reason, many research deals with these problems by discovering new novel methods to solve these problems. In this these, we focus on a particular kind of incomplete data, thus we initially established censoring and truncation as the two categories of incomplete data.

1.1 *Preliminary definitions*

This section provides some definitions and basics, in order to facilitate the reading of this thesis, we assume X be continues random variable (rv) defined over the probability space $(\Omega, \mathcal{A}, \mathcal{F})$.

Definition 1 The distribution function (df) of X is defined on $\mathbb{R}$ by

$$F(x) := P(X \leq x) = \int_{a_X}^{x} f(z)dz,$$

where we have $a_X = \inf\{x : F(x) > 0\}$ and $b_X = \inf\{x : F(x) = 1\}$ are the left and right ends points of the df F respectively.

Definition 2 The survival function (or the tail distribution) of X is defined on $\mathbb{R}$, is also can be consider as the opposite of the df, thus is given by

$$S(x) := P(X > x) = \int_{x}^{b_X} f(z)dz.$$

Definition 3 The probability density function of X is a non-negative integration function defined by

$$f(x) := \frac{dF(x)}{dx} = -\frac{dS(x)}{dx}.$$

Definition 4 The hazard function of X is also know as instantaneous failure rate or instantaneous hazard rate in survival analysis is defined by

$$h(x) := \lim_{\delta t \to 0} \frac{P(X \in [x, x + \delta t[|X \geq x)}{\delta t} = \frac{f(x)}{S(x)} = -\frac{d \ln S(x)}{dx}.$$

from this definition we note that the hazard function is non negative function and doesn't have an upper bound, beside the rv X can take negative values.

Definition 5 The cumulative hazard function which defined as the accumulation hazard rate of X is defined by

$$\Delta(x) := \int_{a_X}^{x} h(t)dt = -\ln S(x).$$

Remark 1 We note that

$$S(x) := \exp\{-\Delta(x)\} \tag{1.1}$$

$$= \exp\{-\int_{a_X}^{x} h(z)dz\}, \tag{1.2}$$

hence, from this equation the density function is given by

$$f(x) := (\Delta(x))' \exp\{-\Delta(x)\} \tag{1.3}$$

$$= h(x) \exp\{-\int_{a_X}^{x} h(z)dz\}. \tag{1.4}$$

Definition 6 Let $X_1, X_2..., X_n$ be n identically distributed rv's with common distribution function F. An estimator of the df F which called the empirical distribution function (edf) is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}. \tag{1.5}$$

Definition 7 The different kinds of convergence Let $X_1, X_2..., X_n$ be n sequences of rv's with df noted by F_{X_n}, we have

- We said that X_n converge to X en law if

$$\lim_{n\to\infty} F_{X_n}(x) = F(x) \Longleftrightarrow X_n \xrightarrow[n\to\infty]{\mathcal{L}} X. \tag{1.6}$$

- We said that X_n converge to X en probability if

$$\forall \epsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0 \Longleftrightarrow X_n \xrightarrow[n\to\infty]{p} X. \tag{1.7}$$

- We said that X_n converges almost surely (or converges with probability 1, or converges strongly) at the point x, if

$$P(\lim_{n\to\infty} X_n = X) = 1 \Longleftrightarrow X_n \xrightarrow[n\to\infty]{as} X. \tag{1.8}$$

Theorem 1 (Central Limit Theorem) *Let $X_1, X_2..., X_n$ be n sequences of rv's, and assumed that is iid, hence we have*

$$\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

where $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$ is the empirical mean, and (μ, σ) is the mean and the variance the two parameters of the normal df, of course this convergence is obtained under the finite of the variance σ^2.

1.2 *Censoring*

Definition 8 In the field of statistics, there are situations in which observations cannot be followed from the beginning, the end, or both. This phenomenon is known as censoring, and it results in incomplete data, and that creates challenges for data analysis. This section focuses on survival analysis which is one area of statistics suffers from incomplete data, as incomplete data make it harder to take decisions. For technical reasons, we make the assumption that the variable of interest and the censoring variable are independent to avoid the problem in using classical analytical techniques. Time to event is the variable of interest in survival analysis which is a random positive variable that shows how long it will take for the event to occur (e.g., death of patient, broken machine, etc.).
There are three types of censorship: left, right, and interval censoring.

1.2.1 *Types of censoring*

1.2.1.1 *Right censoring*

First, we define right censoring, which comes in three kinds.

1. **Type I censoring (Fixed censoring)**
   Which is defined as time of the event is observed only if it happen before a specific time. (e.g., in a medical trial of covid 19 in hospital, the researcher uses a fixed number of patients to know the relationship between the covid 19 and increase in injury pressure before a medication or therapies are administered, but the researcher may decide to stop the study or announce the findings before the end of the study due to time. In these case, all censored observations have times equal to the duration of the study period, assuming no losses or persons withdrawals).

   Now, We suppose that there is a lifetime X assumed to be iid, and a fixed censoring time C^R, we assume there is independent between X and C^R. The time X be known if and only if X is less than or equal to C^R (i.e., $X \leq C^R$). When X is less then C^R, we said that the person is considered a survivor and their event time is censored at C^R, we have a sample of survival times $X_1, ..., X_n$ the censoring variable C^R here is fixed, so the observations defined by the pairs $(T_1, \delta_1), ..., (T_n, \delta_n)$ with :

$$T_i = min(X_i, C^R), \delta_i = \mathbf{1}_{\{X_i \leq C^R\}}, \forall i = 1, ..., n.$$

2. **Type II censoring (Censorship waiting)** is a second kind of right censoring where the study is continues up until the first k individual fail, where k is a predefined integer ($k < n$). This type is found in testing equipment life in factories where the test is run simultaneously for each item and it ends when k fail of n, so this kind of experiment could save time and money. It's also true that is has an easier statistical treatment because it consists of the k values less then n values.

   Now, let a sample of survival times $X_1, ..., X_n$ and $k > 0$ be fixed. Type II censoring is said to exist for this sample if, instead of directly observing $X_1, ..., X_n$ we observe $(T_1, \delta_1), ..., (T_n, \delta_n)$, which defined by

   $$T_i = min(X_i, X_{(k)}), \delta_i = \mathbf{1}_{\{X_i \leq X_{(r)}\}}, \forall i = 1, ..., n,$$

   where $X_{(k)}$ is the kth–order statistics associated to $X_1, ..., X_n$, note that if we take $X_{(k)} = C^R$ is equivalent to **censoring type I**.

3. **Type III censoring (Random censoring.)**
   This form of censorship, where C is a random variable, is can be consider as generalization of type I. We assume that we have sample of survival times $X_1, ..., X_n$ and anther sample which is independent sample and contained a positive random variables $C_1^R, ..., C_n^R$.

   We said that we have censored Type III if instead of observing $X_1, ..., X_n$ we observe the pairs $(T_1, \delta_1), ..., (T_n, \delta_n)$, which given by

   $$T_i = min(X_i, C_i^R), \delta_i = \mathbf{1}_{\{X_i \leq C_i^R\}}, \forall i = 1, ..., n.$$

   where C is censored variable.

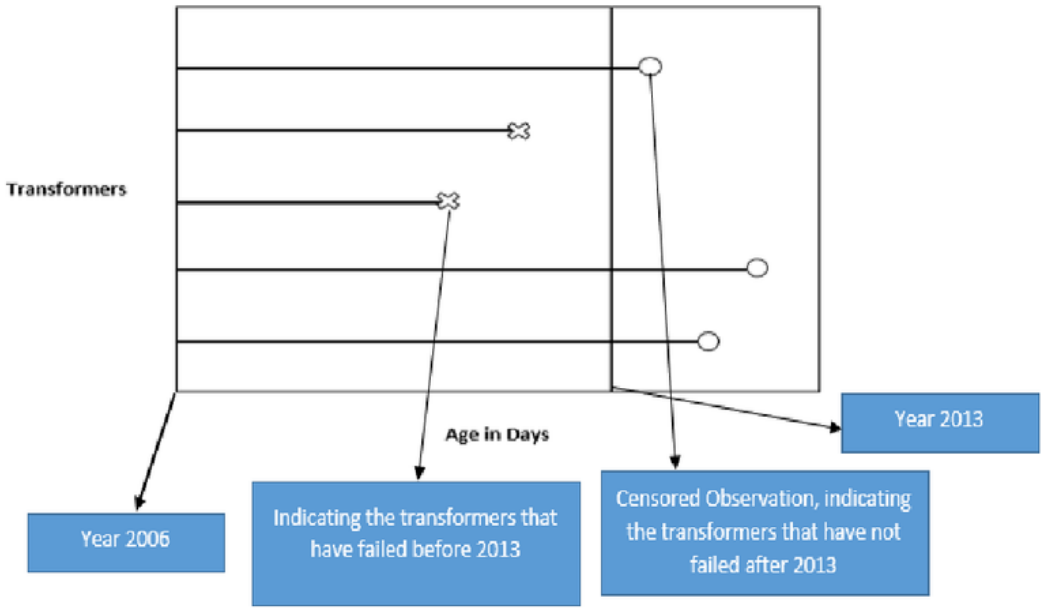


Figure 1.1: Right censoring data

1.2.1.2 *Left censoring*

The variable of interest X is considered to be left censored if it is less than a censoring variable C^L, meaning that the event of interest has already happened for the individual before that person is observed in the study at C^L. For such individuals, we know that they experienced the event at some point prior to variable C^L, but we are unsure of their precise event value. The time X will be known if and only if X is greater than or equal to C^L (i.e., $X \geq C^L$).

The data from a left-censored sampling scheme can be represented by pairs $(T_1, \delta_1), ..., (T_n, \delta_n)$ instead of observing $X_1, ..., X_n$, where this pairs are defined by

$$T_i = max(X_i, C_i^L), \delta_i = \mathbf{1}_{\{C_i^L \leq X_i\}}, \forall i = 1, ..., n.$$

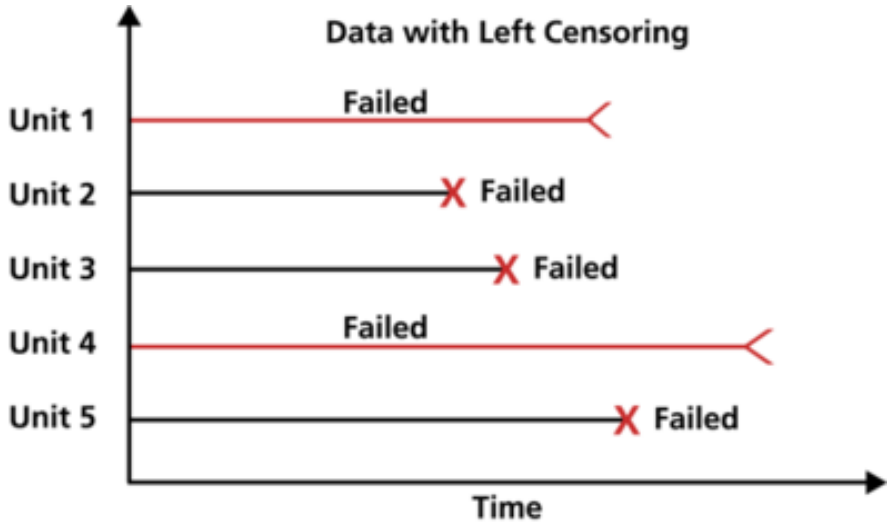


Figure 1.2: Left censoring data

1.2.1.3 *Mixed censored (double censored)*

Now, in some study we can find both left and right censoring, in this case the lifetimes are considered as doubly censored. The variable of interest X is said to be double censored if we observe pairs $(T_1, \delta_1), ..., (T_n, \delta_n)$ instead of observing $X_1, ..., X_n$, where the pairs giving by

$$T_i = max(min(X_i, C_i^R), C_i^L), \forall i = 1, ..., n,$$

where

- $\delta = 1$ the individual is survival.
- $\delta = 0$ the individual is left censored.
- $\delta = -1$ the individual is right censored.

Note that X will be observed if it less or equal to C^R and is great or equal to C^L.

1.2.1.4 *Interval censoring*

Interval censoring is widely used, which happens when we observe the individual just inside the interval (C^L, C^R), where C^L (or C^R) is the left endpoints (or the rights endpoints) of the censoring interval. This kind of censoring can happen in industrial experiments where equipment components are routinely inspected to ensure correct operation. Interval censoring can also be seen as a generalisation of one-sided censoring data, as left (or right) censoring is when the left end point is 0 (or C^R) and the right end point is C^R (or ∞).

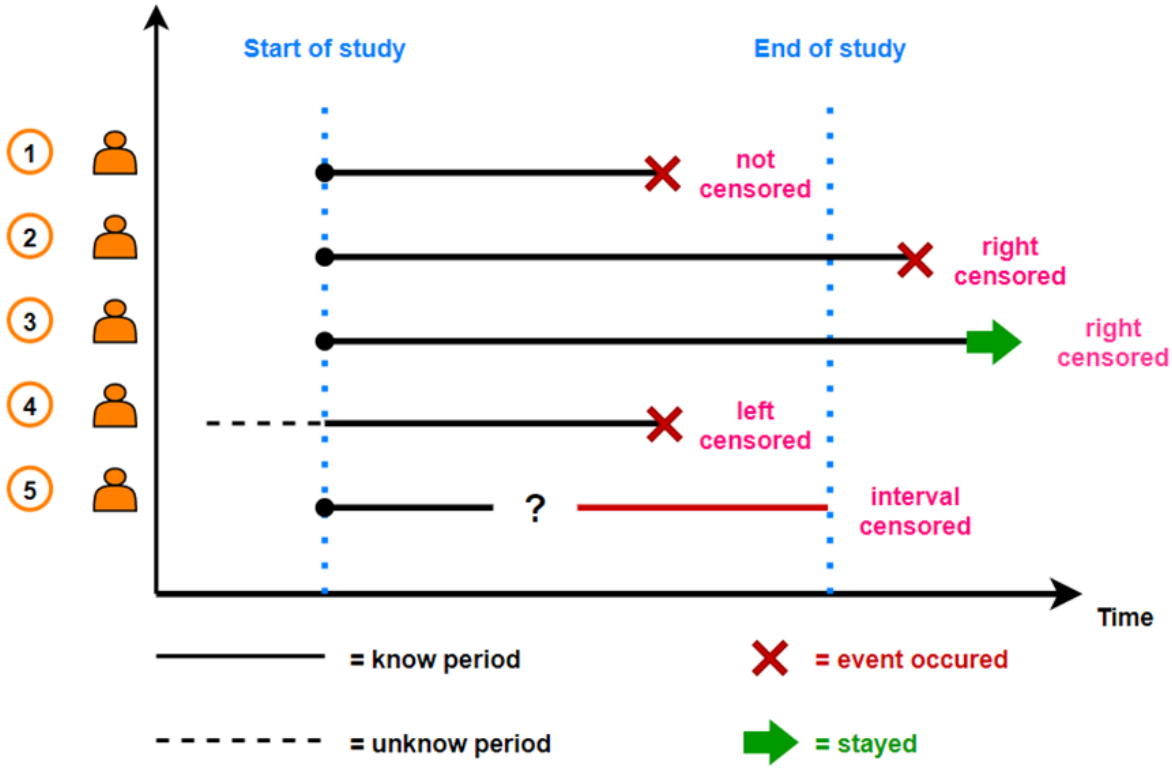


Figure 1.3: Interval censoring data

1.2.2 *A guide to defining likelihood functions with censored data*

It is typical that the construction of the likelihood under censored data need special mechanism. Initially, we make the assumption that there are independent between the lifetime and the censoring time. Additionally, we must take into account any knowledge came from the observed data and provide a mathematical definition for it.

- The probability in observed time is approximately equal to the density function $f(x)$ of X at this time.
- From the definition of right censored observation it approximately equal to survival function $S(C^R)$.
- From the definition of left censored observation it approximately equal to cumulative distribution function $1 - S(C^l) = F(C^l)$.
- From the definition of interval censored observation it similarly to probability in the interval $S(L) - S(R)$.

The likelihood function is given by this equation

$$L \propto \prod_{i \in d} f(x_i) \prod_{i \in r} S(C_i^R) \prod_{i \in l} (1 - S(C_i^L)) \prod_{i \in j} (S(C_i^L) - S(C_i^R)), \tag{1.9}$$

where d is set of death times, r (or l) is set of right (or left) censored observations, and j is set of interval censored observations.

1.2.3 *Techniques for estimation in the presence of right censored data*

In this part, we'll deal into right-censored data estimation as this kinds is well-known topic, where we'll defined the widely used Kaplan-Meier and Nelson-Aalen estimators.
Typically, the ordinary edf $F_n^*(x) = 1/n \sum_{i=1}^n I_{\{X_i \leq x\}}$ is commonly used when we need to estimate the df. However this estimator can't hold under censoring data because certain observations cannot be observed. For this reason, Kaplan and Meier (1958) discovered the first estimator of the df in the case of right censored data. The principle of this estimator is around the idea that if you survive to time x, you are sure that you was been survived before this time. For this reason, we assume that we have D distinct times $x_1 < x_2, ..., < x_D$, hence for $x_1 < x_2 < x$ we gate

$$\begin{aligned} S(x) &:= P(X > x) \\ &= P(X > x, X > x_2) \\ &= P(X > x | X > x_2) P(X > x_2) \\ &= P(X > x | X > x_2) P(X > x_2 | X > x_1) S(x_1), \end{aligned}$$

Now, consider

$$p_i := P(X > T_i | X > T_{i-1}), \tag{1.10}$$

which is the conditional probability that the individual survives during the interval $]T_{i-1}, T_i]$, where is known to have survived in the beginning of the interval.
Let r_i be the number of individual who are at risk (have not yet had an event) at time t_i, and d_i is the number of events (e.g., deaths, disease, relapse,etc) at the same time. We assume $q_i := 1 - p_i$ which represent the probability of individual had the event in the interval $]T_{i-1}, T_i]$, where is known to have survived in T_{i-1}. Hence the estimator of q_i is given by

$$\hat{q}_i = d_i / r_i.$$

Where

$$d_i = \begin{cases} 0 & , if \delta_i = 0 \\ 1 & , if \delta_i = 1. \end{cases}$$

Similar reasoning that lead to the construction of $\hat{q}_i$ estimator above, we arrive at the estimator of $\hat{p}_i$

$$1-\hat{p}_i=\begin{cases}\frac{1}{n-(i-1)} & ,if\delta_i=1\\ 1 & ,if\delta_i=0.\end{cases}$$

The Kaplan–Meier estimator (or **the product limit estimator**) for survival function of the variable of interest is then given by

$$\hat{S}(x):=1-\hat{F}(x)=\begin{cases}\prod_{X_{(i)}\leq x}(\frac{n-i}{n-(i-1)})^{\delta_{(i)}} & ,ifx<X_{(n)}\\ 0 & ,ifx\geq X_{(n)}.\end{cases}$$

The Kaplan–Meier estimator for survival function of the variable of censoring is then given by

$$\bar{G}(x):=1-\hat{G}(x)=\begin{cases}\prod_{X_{(i)}\leq x}(\frac{n-i}{n-(i-1)})^{1-\delta_{(i)}} & ,ifx<X_{(n)}\\ 0 & ,ifx\geq X_{(n)}.\end{cases}$$

Thus, by defining the relationship between the cumulative hazard function and the survival function $\Delta(x):=-\ln S(x)$, the product limit estimators can provide us with an estimator of the cumulative hazard function which define by $\hat{\Delta}(x):=-\ln\hat{S}(x)$.
The Nelson–Aalen estimator of the cumulative hazard, is given by

$$\hat{\Delta}(t):=\sum_{T_i\leq t}\frac{d_i}{y_i}.$$

This both estimators can be consider as the nonparametric likelihood, in addition the asymptotically normality of this both estimators is well defined in the book [52].

1.3 *Truncation*

Definition 9 As explained earlier, the data problems occurs when we collect observations and therefore some observations cannot be followed from the beginning to the end, and this is called censoring. Moreover, if some observations are lost, so that the data observation becomes conditional, which means that the observation is only known within conditional limits, and this leads to a change in the size of the sample, unlike the first case where the sample size was fixed, we are in the case of truncation, which is the second part of incomplete data.

The truncation can be divided into three parts: left, right and double truncation.

1.3.1 *Types of truncation*

1.3.1.1 *Left truncation*

This type of truncation is the most used, here we consider the left truncation limit U^* as a condition where the variable of interest X^* is observed if it exceeds U^* (i.e., $U^* \leq X^*$). Therefore, this type of truncation does not include individuals who don't meet the conditional limit. (e.g., In biology, researchers measure the diameter of a bacterium so that if a bacterium is too small, it is not taken into account and is excluded from the study or in economics when we classifying the people's income).

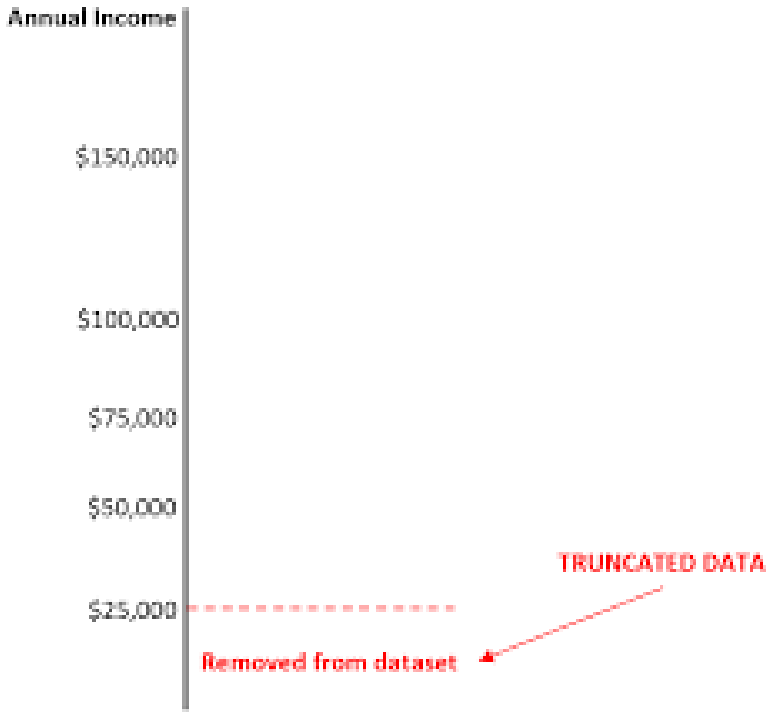


Figure 1.4: Left truncation data

1.3.1.2 *Right truncation*

Now the right truncation data which is the second type of truncation. In this type we define V^* as the right truncation limit where the condition of observed the data is when the observation X^* is less than the right truncation limit (i.e., $X^* \leq V^*$). Furthermore, this type of truncation takes into account the observations that are less then the threshold V. (e.g., in astronomy, when we are interested in the study of stars, only the stars that are closed from the earth are included in the study).

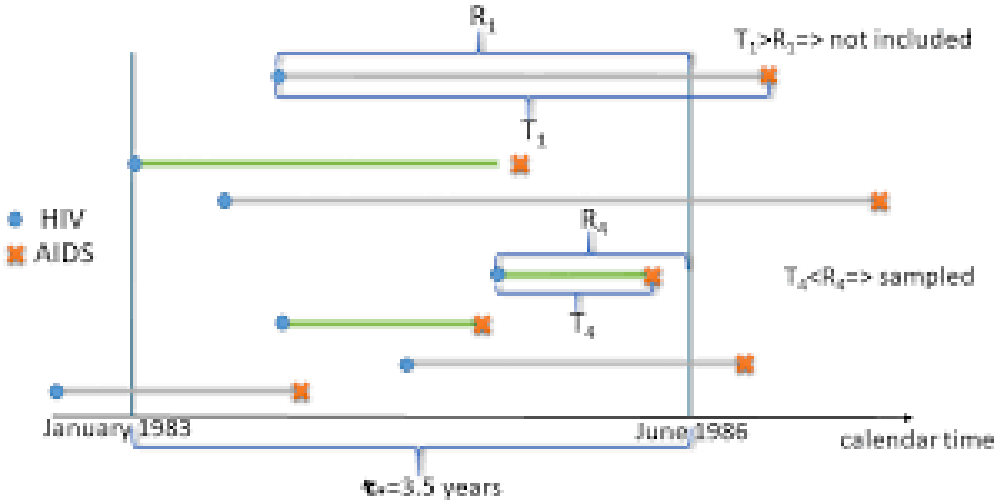


Figure 1.5: Right truncation data

1.3.1.3 *Double truncation*

In the context of collected data, deleting some observations is happening under certain limit conditions from above or below, but there is some situation where this eliminate is occurs from the both side which we called it double truncation data. The concept of double truncation is that the variable of interest is known in side interval (i.e., $U^* \leq X^* \leq V^*$). (e.g., In medicine studies can be so difficult or impossible that to examine every data point because there is time factor or the trial involve measurement error, hence some values may be invalid, or not all the information of interest might be available. For these reasons, a decision may be made and we may not get a full result.).

Example For example, we look at this example in the report in [9], which includes companies that failed from $U^* = 1/09/2013$ to $V^* = 31/03/2014$ and this example is a special case of double truncation data which will be defined in next chapter. Now, we assume that X^* is the rv represent the lifetime which has been chosen randomly in Germany, obviously if any companies fails outside this period it will not be included in this study and this is illustrated in the figure below, therefore the sample is formulae by $U^* \leq X^* \leq V^*$, where

- U^* is the age of the companies on 09/2013, which is the left truncated limit.
- $V^* = U^* + 7/12$ is the age of the companies, which is the right truncated limit.

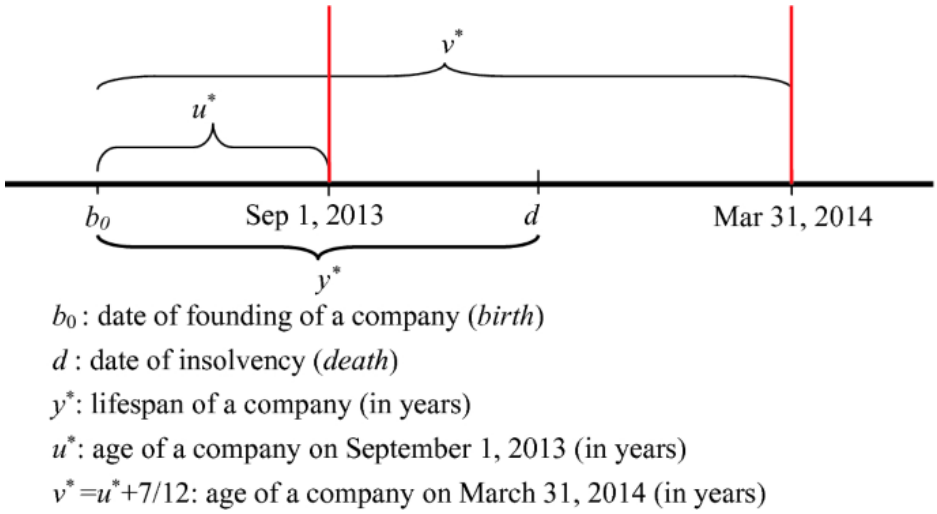


Figure 1.6: Example of double truncation data: The data on German companies from

1.3.2 *The definition of the likelihood functions with truncation data*

The concept of truncation refers to the deletion of some observations under certain conditions. For this reason, this part of the available data is often used by a conditional distribution to make inference on the data. Now, we define in this parts the likelihood function in the case of truncated data, under the same assumption in the case of censoring, by

- The density function $f(x)$ is consider as the probability observed value.
- Left truncation observations is defined by $f(x)/S(U)$.
- Right truncation observations is defined by $f(x)/(1 - S(V))$.
- Interval truncation observations is defined by $f(x)/(S(U) - S(V))$.

Now, the likelihood function in the case of truncation data is defined by

1. For left-truncated data we define the likelihood by make this change from the likelihood function equation (1.9)

$$f(x_i) \longrightarrow f(x_i)/S(U_i). \tag{1.11}$$

$$S(C_i^R) \longrightarrow S(C_i^R)/S(U_i). \tag{1.12}$$

2. For Right-truncated data, in this case only deaths are observed hence the likelihood function is defined by

$$L \propto \prod_i f(x_i)/(1 - S(V_i)). \tag{1.13}$$

1.3.3 *Techniques for estimation in the presence of right truncated data*

Now, let consider X as the random variable (rv) of interest defined over probability space $(\Omega, \mathcal{A}, \mathcal{F})$ with an unknown distribution function (df) F. Let Y be rv of truncation with an unknown df G^*. Due to the effect of sample selection under truncation, we observe only pairs $\{(X_i, Y_i)/1 \leq i \leq n\}$ which satisfying the condition $X_i \leq Y_i$. Therefore the observed sample size be n which is a subset of N (i.e., $n \leq N$) defined as Binomial rv by $n := \sum_{i=1}^{N} \mathbf{1}_{\{X_i \leq Y_i\}}$, before that we assume $\alpha := P(X \leq Y)$ is the probability of truncation, hence by the weak law of large numbers we have $n/N \xrightarrow[n\to\infty]{p} \alpha$. First, we define the joint df of the pairs (X^*, Y^*)

$$\begin{aligned} M(x,y) &:= P(X^* \leq x, Y^* \leq y | X^* \leq Y^*) \\ &= \alpha^{-1} P(X^* \leq min(x, Y^*), Y^* \leq y) \\ &= \alpha^{-1} \int_{a_Y}^{y} F(\min(x,z)) dG(z). \end{aligned}$$

The marginal df's of the observed pairs are given by

$$F^*(x) := M(x, \infty) = \alpha^{-1} \int_{a_Y}^{\infty} F(\min(x,z)) dG(z) = \alpha^{-1} \int_{a_X}^{x} (1 - G(z)) dF(z).$$
$$G^*(y) := M(\infty, y) = \alpha^{-1} \int_{a_Y}^{y} F(z) dG(z) = \alpha^{-1} \int_{a_X}^{y} F(z) dG(z).$$

Note that F^* and G^* can be estimated by the empirical distribution functions as in complete data. Now, we define the Woodroofe's nonparametric estimator of the df by

$$F_n^{(W)}(x) := 1 - \prod_{k: X_k^* \leq x} \exp(-(nC_n(X_k^*))^{-1}),$$
$$G_n^{(W)}(y) := 1 - \prod_{k: X_k^* \leq x} \exp(-(nC_n(Y_k^*))^{-1}),$$

where $C_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x \leq Y_i\}}$. From this approximation $\exp(x) \sim 1 + x$, Lynden [25] discovers a different estimator of the distribution function under right truncation, which is widely used. And is defined by

$$F_n^{(W)}(x) := \prod_{k: X_k^* > x} (1 - (nC_n(X_k^*))^{-1}),$$
$$G_n^{(W)}(y) := \prod_{k: X_k^* > x} (1 - (nC_n(Y_k^*))^{-1}).$$

CHAPTER 2

ESTIMATION UNDER DOUBLE TRUNCATION

Slleced data in order for make decision and give result about study can be hard as the data can be incomplete. For this reason, many researchers have focused on solving this problem, since the classical technique which is widely used cannot hold up, this has led researchers to make modifications to the classical technique to deal with these problems, including the truncation problems, where the data are only observed in a known interval. Where the fundamental change arises in this case by using conditional probability.

2.1 *The definition of probability under double truncation*

First, let X^* be the variable of interest which assume to be truncated by the random variables U^* (i.e., left truncation limit) and V^* (i.e., right truncation limit). Furthermore, in all our work we assume the independent between the variable of interest and the variables of truncation limit. Hence the observed double truncation data is given by $\{(U_i^*, X_i^*, V_i^*), U_i^* \leq X_i^* \leq V_i^*, i = 1, 2, ..., n\}$. Let consider $F(x) := P(X^* \leq x)$ and $G(u, v) := P(U^* \leq u, V^* \leq v)$ which is the distribution and the joint distribution of X and (U, V) respectively. Now we consider the probability of non truncation which defined by

$$\begin{aligned} \alpha &:= P(U^* \leq X^* \leq V^*) && (2.1)\\ &= \iint_{u\leq v} \int_u^v dF(x)dG(u, v) && (2.2)\\ &= \int \iint_{u\leq x\leq v} dG(u, v)dF(x). && (2.3) \end{aligned}$$

Therefore, we define the probability of non-truncation conditional on the observed $X^* = x$ by this formula

$$\begin{aligned} H(x) &:= P(U^* \leq X^* \leq V^*|X^* = x)\\ &= P(U^* \leq x \leq V^*). \end{aligned}$$

The distribution function of X_i is given by

$$\begin{aligned} F^*(x) &:= P(X^* \leq x|U^* \leq X^* \leq V^*)\\ &= \alpha^{-1}P(X^* \leq x, U^* \leq X^* \leq V^*)\\ &= \alpha^{-1} \int_{a_X}^x P(U^* \leq z \leq V^*)dF(z), \end{aligned}$$

by derivation we find the density function $f(x) = \alpha P^{-1}(U^* \leq x \leq V^*)f^*(x)$, and the joint distribution function of the truncation limit is defined by

$$\begin{aligned} G^*(u, v) &:= P(U^* \leq u, V^* \leq v|U^* \leq X^* \leq V^*)\\ &= \alpha^{-1}P(U^* \leq u, V^* \leq v, U^* \leq X^* \leq V^*)\\ &= \alpha^{-1} \int_{a_V}^v \int_{a_U}^{\min(z_2,u)} (F(z_2) - F(z_1^-))dG(z_1, z_2), \end{aligned}$$

let G_1 and G_1 be the marginal distribution of the truncation limits, hence the joint density of (U^*, V^*) is given by $g(u, v) = \alpha(F(v) - F(u^-))^{-1}g^*(u, v)$.

The survival function is defined by

$$\begin{aligned} S(x) &:= P(X^* \geq x | U^* \leq X^* \leq V^*) \\ &= \alpha^{-1} S(x^-) P(U^* \leq x^* \leq V^*), \end{aligned}$$

where $S(x^-) := P(X^* \geq x)$.

2.2 *The construction of the likelihood function under double truncation data*

2.2.1 *Nonparametric consideration in estimation*

The likelihood function based on the joint density of the observed data is given by

$$L = \prod_{i=1}^{n} P(U^* = u_i, X^* = x_i, V^* = v_i | U^* \leq X^* \leq V^*) = \prod_{i=1}^{n} \frac{P(U^* = u_i, X^* = x_i, V^* = v_i)}{P(U^* \leq X^* \leq V^*)}, \tag{2.4}$$

therefore, the likelihood function can be described by the formula

$$L = \prod_{i=1}^{n} \frac{f(x_i) g(u_i, v_i)}{\iiint\limits_{u \leq x \leq v} dF(x) dG(u, v)}. \tag{2.5}$$

In addition, the likelihood function may be decomposed as follows

$$L = \prod_{i=1}^{n} \frac{f(x_i)}{\int_{u_i}^{v_i} F(dx)} \times \prod_{i=1}^{n} \frac{\int_{u_i}^{v_i} dF(x) g(u_i, v_i)}{\iiint\limits_{u \leq x \leq v} dF(x) dG(u, v)} = \mathbf{L}_1(f) \times \mathbf{L}_2(f, g). \tag{2.6}$$

Now, in order to make an inference in density, Efron and Petrosian (1999) had to consider the first part

$$\mathbf{L}_1(f) = \prod_{i=1}^{n} P(X^* = x_i | u_i \leq X \leq v_i) = \prod_{i=1}^{n} \frac{f(x_i)}{\int_{u_i}^{v_i} dF(x)}. \tag{2.7}$$

In addition, [22]show that the conditional likelihood can be treated as a classical likelihood function, beside Shen (2010) establishes that the both parts of likelihood function $\mathbf{L}_1(f)$ and $\mathbf{L}_2(f, g)$ yields the same estimator of density.

Let us now define the probability distribution function according to the truncated interval $R_i := [u_i, v_i]$

$$F_i := \int_{R_i} f(x) dx. \tag{2.8}$$

Due to truncation effects, we only regard probability mass on observed vector $(x_1, ..., x_n)$, for this reason the distribution of X^* under the condition of regularity $\sum_{i=1}^n f_i = 1$ is given by $\mathbf{f} := (f_1, ..., f_n)$. Hence the probability of observed individual x_i is defined by

$$F_i := \sum_{j=1}^{n} J_{ij} f_j, \tag{2.9}$$

where $J_{ij} := \mathbf{1}_{\{u_i \leq X_j \leq v_i\}}$, thus we have $\mathbf{F} = \mathbf{Jf}$ where $\mathbf{J} = (J_{i,j})_{1 \leq i \leq n, 1 \leq j \leq n}$.
The log-likelihood function is given by

$$\mathcal{L}_1(f) = \ln L_1(f) = \sum_{i=1}^{n} \ln f_i - \sum_{i=1}^{n} \ln(\sum_{k=1}^{n} J_{ik} f_k).$$

According to the classical method of the maximum likelihood solution we derive $\mathcal{L}_1(f)$

$$\frac{\partial \mathcal{L}_1(f)}{\partial f_i} = \frac{1}{f_i} - \sum_{j=1}^{n} \frac{J_{ji}}{\sum_{k=1}^{n} J_{jk} f_k} = \frac{1}{f_i} - \sum_{j=1}^{n} \frac{J_{ji}}{F_j}. \tag{2.10}$$

Thus,

$$\frac{\partial \mathcal{L}_1(f)}{\partial f_i} = 0 \Longleftrightarrow \frac{1}{f_i} = \sum_{j=1}^{n} \frac{J_{ji}}{F_j}, \forall i = 1, ..., n. \tag{2.11}$$

The solution is iterative, hence Efron and Petrosian (1999) give EM algorithm for solve the equation

$$\frac{1}{\hat{f}_i} = \sum_{j=1}^{n} J_{ji} \frac{1}{\hat{F}_j}, \forall i = 1, ..., n \tag{2.12}$$

where $\hat{F}_j = \sum_{k=1}^{n} \hat{f}_k J_{ik}$. The EM algorithm is defined by

Step EP 1 : Introduce the first estimator of $\mathbf{f}^{(0)} = (1/n, ..., 1/n)$ according to $\mathbf{F}^{(0)} = \mathbf{Jf}^{(0)}$;

Step EP 2 : Calculate $\mathbf{F}^{(k)} = \mathbf{Jf}^{(k)}$;

Step EP 3 : Repeat Step EP 2 and make $k \to k+1$, calculate

$$\hat{f}_i^{(k)} = \sum_{j=1}^{n} J_{ji} \hat{F}_j^{(k-1)}, \tag{2.13}$$

with respect to $\sum_{i=1}^{n} \hat{f}_i^{(k)} = 1$;

Step EP 4 : Repeat Step EP 2 and Step EP 3 until the convergence;

The likelihood function also has another representation based on the joint truncation limit distribution.

$$L = \prod_{j=1}^{n} \frac{g_j}{G_j} \times \prod_{i=1}^{n} \frac{G_j f_j}{\sum_{i=1}^{n} G_j f_j} = \mathbf{L}_1(g) \times \mathbf{L}_2(g, f), \tag{2.14}$$

where $G_i = \sum_{k=1}^{n} g_k J_{ki}$, now we consider the first part $\mathbf{L}_1(g)$ then the maximum likelihood is given by

$$\frac{1}{\hat{g}_j} = \sum_{i=1}^{n} J_{ji} \frac{1}{\hat{G}_i}, \forall j = 1, ..., n, \tag{2.15}$$

where $\hat{G}_i = \sum_{k=1}^{n} \hat{g}_k J_{ki}$. Shen (2010) proof that both solution of equations 2.12 and 2.15 are the NPMLE's of the full likelihood defined in 2.4. In addition, he show that the estimators of $F(x)$ and $G(u, v)$ are given by solving the two equations

$$\hat{F}(x) = [\sum_{i=1}^{n} \frac{1}{\hat{G}(X_i, \infty) - \hat{G}(X_i, X_i)}]^{-1} \sum_{i=1}^{n} \frac{\mathbf{1}_{\{X_i \leq x\}}}{\hat{G}(X_i, \infty) - \hat{G}(X_i, X_i)} \tag{2.16}$$

$$\hat{G}(u, v) = [\sum_{i=1}^{n} \frac{1}{\hat{F}(V_i) - \hat{F}(U_i^-)}]^{-1} \sum_{i=1}^{n} \frac{\mathbf{1}_{\{V_i \leq v, U_i \leq u\}}}{\hat{F}(V_i) - \hat{F}(U_i^-)}. \tag{2.17}$$

Theorem 2 *Let $\hat{F}_{NP}(x) = \sum_{n=1}^{n} \hat{f}_i \mathbf{1}_{\{X_i \leq x\}}$ and $\hat{G}_{NP}(u, v) = \sum_{n=1}^{n} \hat{g}_i \mathbf{1}_{\{U_i \leq u, V_i \leq v\}}$ are according to likelihood function of $L_1(\boldsymbol{f})$ and $L_1(\boldsymbol{g})$ respectively, then*

1. *$\hat{F}(x) = \hat{F}_{NP}(x)$ and $\hat{G}(u, v) = \hat{G}_{NP}(u, v)$ are the NPMLE's F and G respectively.*
2. *$\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$ are the NPMLE's of the full likelihood L.*

Proof See Shen (2010) [38].

Shen [38] show that the NPMLE can obtained be use iterative algorithm based in the two system of equations

$$\hat{f}_j = [\sum_{i=1}^{n} \frac{1}{G_i}]^{-1} \frac{1}{\hat{G}_j}, \forall j = 1, ..., n. \tag{2.18}$$

$$\hat{g}_j = [\sum_{i=1}^{n} \frac{1}{\hat{F}_i}]^{-1} \frac{1}{\hat{F}_j}, \forall j = 1, ..., n. \tag{2.19}$$

Step EP 1 : Introduce the first estimator of $\mathbf{f}^{(0)} = (1/n, ..., 1/n)$ according to $\mathbf{F}^{(0)} = \mathbf{J}\mathbf{f}^{(0)}$;

Step EP 2 : Calculate $\mathbf{F}^{(k)} = \mathbf{J}\mathbf{f}^{(k)}$ and

$$\hat{g}_j^{(t+1)} = [\sum_{i=1}^{n} \frac{1}{\hat{F}_i^{(t)}}]^{-1} \frac{1}{\hat{F}_j^{(t)}}, \forall j = 1, ..., n. \tag{2.20}$$

Step EP 3 : Calculate $\mathbf{G}^{(k+1)} = \mathbf{J}^T \mathbf{g}^{(k+1)}$ and

$$\hat{f}_j^{(t+1)} = [\sum_{i=1}^{n} \frac{1}{\hat{G}_i^{(t)}}]^{-1} \frac{1}{\hat{G}_j^{(t)}}, \forall j = 1, ..., n. \tag{2.21}$$

and make $k \to k + 1$;

Step EP 4 : Repeat Step EP 2 and Step EP 3 until the convergence;

The NPMLE of the distribution function can be written as the following formula

$$F_n(x) := \alpha_n^{-1} \int_{a_F}^{x} \frac{dF_n^*(z)}{H_n(z)}, \tag{2.22}$$

where $H_n(x) = \int_{u \le x \le v} G_n(du, dv)$, and $\alpha_n^{-1} = \int_{a_F}^{\infty} (H_n(z))^{-1} dF_n^*(z)$ this both estimators are defined in Shen (2010).

2.2.1.1 *Asymptotic Properties of the NPMLE*

Under the stable conditions of woodroofe (1985) and Shen (2010), the following two theorems prove that F is consist estimator and is asymptotically normal.

Theorem 3 *Let $a_X \in [0, \infty)$ be such that $F(v) - F(u) > \delta > 0$ for $[u, v] \subseteq [a_X, \tau]$. Moreover, assume that*

1. $\int_{a_X}^{\tau} dF(x)/G(x, \infty) - G(x, x) < \infty$.
2. $dG(x, \infty) - dG(x, x)/dF(x)$ *is uniformly bounded on* $[a_X, \tau]$. *Then the NPMLE* $\hat{F}$ *is uniformly consistent on* $[a_X, \tau]$.

Proof See Shen (2010) [38].

Theorem 4 *Let $D(x) = G(x, \infty) - G(x, x)$, $\tilde{D}(x) = (x, \infty) - (x, x)$ and $\tilde{D}_n(x) = \tilde{G}_n(x, \infty) - \tilde{G}_n(x, x)$. Under the assumptions (a) and (b) of Theorem 2, we assume that (c) the class of functions F, where F consists of functions with envelop $1/D(s)$ is a $\tilde{F}(s)$-Donsker class, and*

$$(d) \int_{u}^{v} \frac{d\tilde{F}(x)}{D(x)\tilde{D}_n(x)} \le G(u, v) \tag{2.23}$$

with probability tending to 1, where $M(.,.)$ is such that the class of functions with envelope $M(.,.)$ is $\tilde{G}(u, v)$-Donsker. Then $\sqrt{n}(\hat{F}_n(x)) - F(x))$ is asymptotically normal for every $x \in [a_F, \tau]$.

Proof See Shen (2010).

2.2.2 *Semiparametric consideration in estimation*

In the previous section we dealt with estimation without condition on the distribution limit $G(u,v)$, however sometimes there is situation where the distribution of truncation limit is assumed to be follow parametric family $\{G(.,.;\theta), \theta \in \Theta)\}$, where θ is a vector parameters and Θ stands for the parametric space. Now, let us consider this definition of the probability distributions of the truncation limit in this case

$$H(x;\theta) := P(U^* \leq x \leq V^*;\theta) \tag{2.24}$$

$$= \int_{\{u\leq x\leq v\}} dG(u,v;\theta). \tag{2.25}$$

Hence, the probability of distribution function is defined by

$$F^*(x;\theta) := P(X^* \leq x | U^* \leq X^* \leq V^*) \tag{2.26}$$

$$= \alpha(\theta)^{-1} \int_{a_X}^{x} H(z;\theta) dF(z), \tag{2.27}$$

where

$$\alpha(\theta) := P(U^* \leq X^* \leq V^*;\theta) \tag{2.28}$$

$$= \int_{a_X}^{b_x} H(z;\theta) dF(z). \tag{2.29}$$

The conditional likelihood function can be represented as the previous case in two parts by

$$\mathcal{L}(F;\theta) := \mathcal{L}_m(F|X;\theta) \times \mathcal{L}_c(\theta|U,V,X) \tag{2.30}$$

$$= \prod_{i=1}^{n} \frac{H(X_i;\theta)dF(X_i)}{\alpha(\theta)} \times \prod_{i=1}^{n} \frac{g(U_i,V_i;\theta)}{H(X_i;\theta)}, \tag{2.31}$$

where $g(u,v;\theta)$ is the joint density of truncation limit. In addition [39] show that an estimator of the parameter θ is given by maximizing the conditional likelihood function $\mathcal{L}_c(\theta) = \prod_{i=1}^{n} \frac{g(U_i,V_i;\theta)}{H(X_i;\theta)}$, and $\mathcal{L}_m(F|X;\theta)$ is treated as a multinomial likelihood. Now, we define the estimator of $F(x)$ by

$$\hat{F}(x;\hat{\theta}) := \hat{\alpha}(\hat{\theta}) \int_{a_F}^{z} H(z;\hat{\theta})^{-1} dF_n^*(z) \tag{2.32}$$

$$:= n^{-1}\hat{\alpha}(\hat{\theta}) \sum_{i=1}^{n} H(X_i;\hat{\theta})^{-1} \mathbf{1}_{\{X_i\leq x\}}, \tag{2.33}$$

where $\hat{\alpha}(\hat{\theta})^{-1} = n^{-1}\sum_{i=1}^{n} H(X_i;\hat{\theta})^{-1}$.

2.2.2.1 Asymptotic properties

Consider the assumption given in [39], We define this theorem

Theorem 5

- *We have* $\sup_{a_X \leq x \leq b_X} |\hat{F}(x;\hat{\theta}) - F(x)| \xrightarrow[n\to\infty]{as} 0$.
- $\sqrt{n}(\hat{F}(.;\hat{\theta}) - F(.))$ *converges weakly to a mean zero Gaussian process with covariance* $A(x,y) = W^T(x)I{-}1(\theta)W(y) + \Sigma(x,y)$ *for* $x \leq y$. *Where* $I(\theta)$ *is the Fisher information matrix given by*

$$I(\theta) = E[(\frac{\partial \log G(U,V;\theta)/H(X;\theta)}{\partial\theta})(\frac{\partial \log G(U,V;\theta)/H(X;\theta)}{\partial\theta})^T], \tag{2.34}$$

and

$$W(s) = \int_{a_X}^{b_X} \frac{\partial H(z;\theta)/\partial\theta}{H(z;\theta)}(F(s) - \mathbf{1}_{\{z\leq s\}})dF(z), \tag{2.35}$$

where

$$\Sigma(x,y) = \omega_\theta \alpha(\theta)[S(y)\int_{a_X}^{x}(\omega_\theta H(z;\theta))^{-1}dF(z) \tag{2.36}$$

$$- F(x)(S(y))\int_{a_X}^{y}(\omega_\theta H(z;\theta))^{-1}dF(z)], \tag{2.37}$$

and $\omega = \int_{a_X}^{b_X} H(z;\theta)^{-1}dF(z)$.

Proof See [39].

Theorem 6 *Now consider* $\hat{\theta}$ *the solution of the maximum likelihood given by*

$$\frac{\partial log\mathcal{L}_c(\hat{\theta})}{\partial\theta} = 0. \tag{2.38}$$

We find

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, I(\theta)^{-1}), \tag{2.39}$$

in law, where $I(\theta) = E[(\frac{\partial^2 \log g(U_i,V_i;\theta)/H(X_i;\theta)}{\partial^2\theta})]$.

Proof See [28].

2.2.3 Bootstrap method

Now, in this part we consider method which addressed for the finite sample for define the NPMLE which is define in [27], hence we follow the same process as in completed data in order for small sample in order to provide an approximation of the NPMLE F_n, we look at the simple bootstrap, which is more consistent than the obvious bootstrap. The bootstrap procedure is defined as in completed data for $b = 1...B$ we defined $U_i^{boot}, X_i^{boot}, V_i^{boot}$ is consider as simple resample where in each observation $U_i \leq X_i \leq V_i$ we putting weight $1/n$, and we repeat this procedure for B, thus we define $\hat{F}^{boot}$ which is computed for $b = 1, ..., n$, hence $\hat{F}_1^{boot}(x), \hat{F}_2^{boot}(x), ..., \hat{F}_b^{boot}(x)$ can consider empirical distribution.

2.2.4 Particular case of double truncation: Fixed-Length

In this part, we discuss special case when the rv of right truncation limit is given by this formula $V = U + d$, where $d > 0$ is deterministic(ie., is not random) this situation is find in many situation in this case the likelihood function is given by

$$L = \prod_{i=1}^{n} \frac{f(x_i)g_1(u_i)}{\int (\int_u^{u+d} dF(x))g_1(u)du} \tag{2.40}$$

where g_1 is the density function of the rv of left truncation limit. Following the same procedure as before, the likelihood function can now be decomposed into

$$L = \prod_{i=1}^{n} \frac{f(x_i)}{\int_{u_i}^{v_i-d} dF(x)} \times \prod_{i=1}^{n} \frac{(\int_{u_i}^{v_i-d} dF(x))g(u_i)}{\int (\int_u^{u+d} dF(x))g_1(u)du} = \mathbf{L}_1(f) \times \mathbf{L}_2(f, g). \tag{2.41}$$

Hence, we find

$$\mathbf{L}_1(f) = \prod_{i=1}^{n} P(X^* = x_i | u_i \leq X \leq U_i + d) = \prod_{i=1}^{n} \frac{f(x_i)}{\int_{u_i}^{v_i+d} dF(x)}. \tag{2.42}$$

In order to define the likelihood in this case we use the last formula as before without modeling the right truncation limit.

2.3 Kernel density estimator

Let us now consider the smooth way to define the density estimators for double truncated data given in [29], which is given by the following equation

$$f_b(x) := (K_b * F_n)(x) = \alpha_n \frac{1}{n} \sum_{i=1}^{n} H_n(X_i)^{-1} K_b(x - X_i). \tag{2.43}$$

And the semiparametric kernel density estimator is given by

$$f_{b,\hat{\theta}}(x) := (K_b * F_{\hat{\theta}})(x) = \alpha_{\hat{\theta}} \frac{1}{n} \sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1} K_b(x - X_i), \tag{2.44}$$

where $K_b(x) = (1/b)K(x/b)$ which is the kernel or weight function, and b is the bandwidth.

2.3.1 *Asymptotic properties*

In addition, in order to study the asymptotic properties of this estimators we define artificial estimators based on the true distribution of α_n and $H_n(x)$ and this hold true under the conditions of convergence of Shen (2010) of α_n and $H_n(x)$, hence both estimators given in 2.43 and 2.44 have the same artificial estimator given by

$$\tilde{f}_b(x) := \alpha \frac{1}{n} \sum_{i=1}^{n} H(X_i)^{-1} K_b(x - X_i). \tag{2.45}$$

Theorem 7 *1. If K is bounded on a compact support, b is such that $\sum_{i=1}^{\infty} \exp(-\nu b n) < \infty$ for each $\nu > 0$, H is continuous at x, and x is a Lebesgue point of f, then $\tilde{f}_b(x) \xrightarrow[n\to\infty]{as} f(x)$.*

2. If, in addition to the conditions in part 1, K is an even function, $b = o(n^{-1/5})$, $H^{-1}f$ has a second derivative which is bounded in a neighbourhood of x, and $f(x) > 0$, then

$$(nb)^{-1/2}(\tilde{f}_b(x) - f(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \alpha H(x)^{-1} f(x) R(k)), \tag{2.46}$$

where $R(k) := \int K^2(x)dx$.

Proof For 1, let $f^*_{0,b}(x) = \alpha H(x)^{-1} f_{0,b}(x)$ and $f_{0,b}(x)$ is the kernel estimator of density which define for completed data and hence we have $f^*_{0,b}(x) \xrightarrow[n\to\infty]{p} \alpha H(x)^{-1} f(x)$.
Under the condition that K is contained in $[-a, a]$, we gate

$$|\tilde{f}_b(x) - f^*_{0,b}(x)| \leq \alpha f^*_{0,b}(x) \sup_{x-ab\leq y\leq x+ab} |H(y)^{-1} - H(x)^{-1}|,$$

hence $\sup_{x-ab\leq y\leq x+ab}|H(y)^{-1}-H(x)^{-1}|\to 0$ and this by the continuity of G at x.
For 2 we follow this same procedure in [12] we gate

$$(nb)^{-1/2}(\tilde{f}_b(x)-E(\tilde{f}_b)(x))\xrightarrow[n\to\infty]{\mathcal{N}}\mathcal{N}(0,\alpha H(x)^{-1}f(x)R(k)), \tag{2.47}$$

hence, we use the Taylor expansion we find $E(\tilde{f}_b(x))-f(x)=O(b^2)$. For $nb^5\to 0$, then we complete the proof.

Now, in order to determine the equation for the bias and the variance, we will based on the assumption of the classical approach as follows

H 1 The kernel function satisfies: $K(t)>0$, $\int K(t)=1$, $\int tK(t)=0$, $\int t^2K(t)<\infty$, and $\int K(t)^2dt<0$.

H 2 The bandwidth $b=b_n$ satisfies: $b\to 0$ and $nb\to\infty$ when $n\to\infty$.

H 3 The functions f and $H^{-1}f$ are twice continuously differentiable around x.

Through the conditions listed above, we find

$$E(\tilde{f}(x))=f(x)+\frac{1}{2}b^2f"(x)\mu_2(K)+o(b^2) \tag{2.48}$$

$$Var(\tilde{f}(x))=\frac{1}{nb}\alpha H(x)^{-1}f(x)R(K)+o((nb)^{-1}). \tag{2.49}$$

Now, we look at the asymptotic formula of MSE of the estimator, which define by

$$AMSE(\tilde{f}(x)):=\frac{1}{4}b^4f"^2(x)\mu_2^2(K)+\frac{1}{nb}\alpha H(x)^{-1}f(x)R(K). \tag{2.50}$$

Thus, the asymptotic formula of $MISE$ is given by

$$AMISE(\tilde{f}):=\int MSE(\tilde{f})(x)dx=\frac{1}{4}b^4R(f")\mu_2^2(K)+\frac{1}{nb}\alpha R(K)\int H(x)^{-1}f(x)dx. \tag{2.51}$$

In accordance with the classical method, we minimize the formula of $AMISE(\tilde{f})$ in order to get the asymptotically optimal bandwidth which define by

$$b_{AMISE}:=[\frac{\alpha R(K)\int H(x)^{-1}f(x)dx}{R(f")\mu_2^2(K)}]^{-1/5}n^{-1/5}. \tag{2.52}$$

2.3.2 *Selection of optimal bandwidth for kernel density estimator*

In this section, we consider the method of obtaining the optimal bandwidth for density when the data is sampling under double truncation which is defined in [32]. The proposed methods can be readily adapted to the nonparametric case, although in this section we restrict our attention to the semiparametric estimator.

2.3.2.1 *Normal reference as method for define bandwidth*

Now, in order for define the optimal bandwidth which give in 2.52 we need to estimate the unknown values in this formula, and as in the classical way given in [34], we assume that the distribution follows the normal distribution $\mathcal{N}(\mu, \sigma)$, therefore by this assumption we find $R(f') = 0.375/(\sigma^5\sqrt{\pi})$, and we take Gaussian kernel we find

$$b_{AMISE} = (0.375\alpha \int H^{-1}(z)f(z)dz)^{1/5}\sigma n^{-1/5}. \tag{2.53}$$

However, this estimator suffers from the problem of smoothing and is therefore reduced to

$$b_{AMISE} = (0.375\alpha \int H^{-1}(z)f(z)dz)^{1/5}IQRn^{-1/5}, \tag{2.54}$$

where IQR is the interquartile range of normal distribution, thus the bandwidth of type normal reference is given by

$$b_{NR} = (0.375\alpha_{\hat{\theta}} \int H_{\hat{\theta}}^{-1}(z)dF_{\hat{\theta}}(z))^{1/5} \min(\hat{\sigma}, 0.795IQR)n^{-1/5}, \tag{2.55}$$

for the parameter σ is can be estimated by

$$\hat{\sigma} = \alpha_{\hat{\theta}} \int (z - \hat{\mu}_{\hat{\theta}})^2 H_{\hat{\theta}}^{-1}(z)dF^*(z), \tag{2.56}$$

and $\hat{\mu}_{\hat{\theta}}$ is defined by

$$\hat{\sigma} = \alpha_{\hat{\theta}} \int z H_{\hat{\theta}}^{-1}(z)dF^*(z). \tag{2.57}$$

For the interquartile range is given by $IQR = F_{\hat{\theta}}^{-1}(0.75) - F_{\hat{\theta}}^{-1}(0.25)$.

2.3.2.2 *Plug in method*

In the previous part, we defined an easy way to find bandwidth. However, this method works well under the assumption of normality, so we need to use a more precise and more flexible method. We note that the estimation of the formula define in 2.52 need to estimate $R(f^{(r)}(x)) = \int f^{(r)}(z)^2 dz = (-1)^r \int f^{(2r)}(z) f(z) dz$, for this reason let study function of the form

$$\psi_r = \int f^{(r)}(x) f(x) dx \tag{2.58}$$

$$= E(f^{(r)}(x)), \tag{2.59}$$

then, we estimated this expression by

$$\begin{aligned}\hat{\psi}_r(g) &= \hat{\alpha} n^{-1} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i)(H(X_i))^{-1} \\ &= \hat{\alpha}^2 n^{-2} g^{-r-1} \sum_{i=1}^{n}\sum_{j=1}^{n} L^{(r)}(\frac{X_i - X_j}{g})(H(X_i))^{-1}(H(X_j))^{-1},\end{aligned}$$

where g and L are the bandwidth and kernel function respectively, and are defined under the same condition in [34] and [45] for completed data case. Hence by simply calculations we find the bandwidth given by

$$g_{AMISE} = [-\frac{\alpha k! L^r(0) \int (H(z))^{-1} dF(z)}{\psi_{r+k} \mu_k(L) n}]^{1/r+k+1} \tag{2.60}$$

Now, rewrite the formula of optimal bandwidth the definition of ψ_4 as

$$b_{AMISE} = [\frac{\alpha R(k) \int (H(z))^{-1} dF(z)}{\psi_4 \mu_2(K)^2}]^{1/5} n^{-1/5}. \tag{2.61}$$

Hence, we estimate $\alpha \int (H(z))^{-1} F(dz)$ and ψ_2. we defined the direct plug-in by this formula

$$\hat{b}_{DPI} = [\frac{\hat{\alpha} \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_4(g) \mu_2(K)^2}]^{1/3} n^{-1/3}. \tag{2.62}$$

However this formula keep depend to the bandwidth g. For this cause we use g by making use of formula 2.60 with $r = 4$ we gate

$$g_{AMISE} = [-\frac{\alpha 2 L^4(0) \int (H(z))^{-1} dF(z)}{\psi_6 \mu_2(L) n}]^{1/7}$$

Now, we need estimation of this bandwidth formula necessitates an estimator of ψ_4, which need again the selection of an appropriate bandwidth. Then we use anther use the normal reference rule for estimate ψ as in completed data we fined

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1} (r/2)! (\pi)^{1/2}}, \tag{2.63}$$

we estimated then we gate

$$\hat{\psi}_r^{NR} = \frac{(-1)^{r/2} r!}{(2\hat{\sigma})^{r+1}(r/2)!(\pi)^{1/2}}. \tag{2.64}$$

In actuality, the number of steps that will be included in this iterative process must be decided. Hence, we will work with $l = 0, 1, 2$. To summarize, for $l = 1$ the procedure contain this steps.

1. Calculate $\hat{\psi}_6^{NR} = \frac{(-1)^3 6!}{(2\hat{\sigma})^7 (3)! (\pi)^{1/2}}$.

2. Calculate $\hat{\psi}_4(g_1)$

$$g_1 = [-\frac{\hat{\alpha} 2L^{(4)}(0) \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_6^{NR} \mu_2(L) n}]^{1/7}. \tag{2.65}$$

3. The bandwidth is

$$\hat{b}_{DPI} = [-\frac{\hat{\alpha} R(k) \int H(z)^{-1} d\hat{F}(z)}{\hat{\psi}_4(g_1) \mu_2(K)^2}]^{1/5} n^{-1/5}. \tag{2.66}$$

Now, the two-stage plug-in bandwidth selector is given by

1. Calculate $\hat{\psi}_8^{NR} = \frac{(-1)^4 8!}{(2\hat{\sigma})^9 (4)! (\pi)^{1/2}}$.

2. Calculate $\hat{\psi}_6(g_1)$

$$g_1 = [-\frac{\hat{\alpha} 2L^{(4)}(0) \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_8^{NR} \mu_2(L) n}]^{1/9}. \tag{2.67}$$

3. Calculate $\hat{\psi}_4(g_1)$

$$g_2 = [-\frac{\hat{\alpha} 2L^{(4)}(0) \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_6(g_1) \mu_2(L) n}]^{1/7}. \tag{2.68}$$

4. The bandwidth is

$$\hat{b}_{DPI;2} = [-\frac{\hat{\alpha} R(k) \int H(z)^{-1} d\hat{F}(z)}{\hat{\psi}_4(g_2) \mu_2(K)^2}]^{1/5} n^{-1/5}. \tag{2.69}$$

2.3.2.3 *Cross-validation method for selecting optimal bandwidth*

Now, we are going to look at another method, which is different from the previous two methods in that it is based on the exact form of $MISE = \int E(f_b(z) - f(z))dz$. Hence we have

$$MISE(f_b) := E(ISE(f_b)) = E(\int (f_b(z) - f(z))^2 dz), \tag{2.70}$$

by simplification as in completed data where minimize $MISE(f_b)$ with respect to b is equivalent to minimize

$$MISE(f_b) - \int f(z)^2 dz = E(\int f_b^2(z)dz - 2\int f_b(z)f(z)dz). \tag{2.71}$$

Let us now consider the estimator of the density that covers all the data excluding the observation X_i by this formula

$$f_{b;-i}(x) = (k_b * F_{n;-i})(x) = \alpha_{n;-i}\frac{1}{n-1}\sum_{i\neq j} H_{n;-i}(X_j)^{-1}K_b(x - X_j), \tag{2.72}$$

where $H_{n;-i}(.)$ is an estimator of $H(.)$, since observation X_i is not included. Hence we have

$$LSCV(b) = \int f_b^2(z)dz - 2\alpha_{n;-i}\frac{1}{n}\sum_{i=1}^{n} f_b(X_i)H_{n;-i}(X_i)^{-1}. \tag{2.73}$$

Therefore, we minimise this formula with respect to b in order for define the optimal bandwidth, we find

$$\hat{b}_{LSCV} = argmin_b\{\int f_b^2(z)dz - 2\alpha_{n;-i}\frac{1}{n}\sum_{i=1}^{n} f_b(X_i)H_{n;-i}(X_i)^{-1}\}. \tag{2.74}$$

2.3.2.4 *Bootstrap method for selection an optimal bandwidth*

Now we look at the smoothed bootstrap which provide consist bandwidth. The bootstrap procedure is defined as in the completed data for $b = 1...B$ we gate

1. Let $X_{b,i}^{boot}$ be an iid sample from f_θ, g, where g is chosen to be well $\hat{b}_{DPI,2}$, and let $U_{b,i}^{boot}, V_{b,i}^{boot}$, $i = 1, .., n$, be an i.i.d. sample from G_n. We repeated this step until the condition of observed data $U_{b,i}^{boot} \leq X_{b,i}^{boot} \leq V_{b,i}^{boot}$.

2. Now, we define $\hat{\theta}$ and $f_{b,\hat{\theta}}^{boot}(x)$ be the estimators based on the bootstrap sample defined in step 1.

Now we define

$$BMISE(b) = B^{-1}\sum_{b=1}^{B}\int (f_{b,\hat{\theta}^{boot}}^{boot}(x) - f_{g,\hat{\theta}}(x)d)^2 x, \tag{2.75}$$

which as B is big it close to

$$MISE^{boot}(b) = E^{boot}[\int (f^{boot}_{b,\hat{\theta}^{boot}}(x) - f_{g,\hat{\theta}}(x)d)^2 x]. \tag{2.76}$$

Hence

$$\hat{b}^{boot} = argmin_b BMISE(b). \tag{2.77}$$

2.4 *Kernel estimation of the cumulative distribution function*

In the previous section, the definition of nonparametric kernel density of [29] is given by

$$\begin{aligned} f_b(x) &:= \int k_b(x-t)dF_n(t) \\ &= \alpha_n \frac{1}{n}\sum_{i=1}^{n} k_b(x-X_i)H_n(X_i)^{-1}. \end{aligned}$$

And the semiparametric kernel density estimator is defined by

$$f_{b,\hat{\theta}}(x) := (K_b * F_{\hat{\theta}})(x) = \alpha_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1} K_b(x-X_i), \tag{2.78}$$

since, now we want a smoother estimate of the distribution, as we have defined the cured estimates of the df in 2.2, we should integrate the density estimators in both cases, for define the nonparametric estimator of df which given by

$$\begin{aligned} F_b(x) &:= \int_{a_X}^{x} f_b(z)dz \\ &= \alpha_n \frac{1}{n}\sum_{i=1}^{n} H_n(X_i)^{-1} \int_{a_X}^{x} k_b(t-X_i)dt \\ &= \alpha_n \frac{1}{n}\sum_{i=1}^{n} H_n(X_i)^{-1} W(\frac{x-X_i}{b}), \end{aligned}$$

and the semiparametric estimator of df by

$$\begin{aligned} F_{b;\hat{\theta}}(x) &:= \int_{a_X}^{x} f_{b;\hat{\theta}}(z)dz \\ &= \alpha_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1} \int_{a_X}^{x} k_b(t-X_i)dt \\ &= \alpha_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1} W(\frac{x-X_i}{b}), \end{aligned}$$

where $W(t) = \int_{a_X}^{t} k(x)dx$ is the cumulative kernel function.
Now, in order to define the asymptotic properties of our estimators, we provide the pseudo-estimator $\tilde{F}_b(t) = \frac{\alpha}{n}\sum_{i=1}^{n} H(X_i)^{-1}W(\frac{x-X_i}{b})$, where this estimator based on the true value of α and $H(X_i)$ and this hold true under the theorem define in 2 for both kinds of estimators nonparametric and the semiparametric. Moreover, we assume that the kernel and bandwidth satisfied the regularity assumptions as in [29] and [33], hence based on the true value of α and H we find this results in next section.

2.4.1 *Asymptotic properties*

In the next part we give the asymptotic mean and variance, in addition to the consistency and the asymptotic normality of the estimator of the kernel distribution function.

H 1 The kernel function satisfies: $K(t) > 0$, $\int K(t) = 1$, $\int tK(t) = 0$, $\int t^2K(t) < \infty$, and $\int K(t)^2dt < 0$.

H 2 The bandwidth $b = b_n$ satisfies: $b \to 0$ and $nb \to \infty$ when $n \to \infty$.

H 3 The functions F and $H^{-1}F$ are twice continuously differentiable around x.

Theorem 8 *Now, under the assumption given before and assuming that H is have symmetric support on $[-1, 1]$, we have*

$$E(\tilde{F}_b(x)) = F(x) + \frac{b^2}{2}F''(x)\mu_2(K) + o(b^2). \tag{2.79}$$

$$Var(\tilde{F}_b(x)) = n^{-1}\alpha H(x)^{-1}\{F(x)[1 - F(x)] + bf(x)[J(k)] + o(b)\}, \tag{2.80}$$

where $J(k) = \int_{-1}^{1} W^2(z)dz - 1$.

Proof The proof follows as in completed data and this in consider of the condition of regularity see [15] for more details.

The asymptotic mean integrated squared error of the estimator of distribution function is defined by

$$\begin{aligned} AMISE(\tilde{F}_b) &= \int AMSE(\tilde{F}_b(x))dx \\ &= \frac{b^4}{4}\int f'(x)^2dx\mu_2(K)^2 + n^{-1}\alpha\int H(x)^{-1}\{F(x)[1 - F(x)] \\ &+ bf(x)[J(k)]dx\}. \end{aligned}$$

Then the asymptotically optimal bandwidth can be obtained by

$$b_{AMISE} = [\frac{(1 - \int_{-1}^{1} W^2(z)dz)\alpha \int H(x)^{-1} f(x)dx}{\int f'(x)^2 dx \mu_2(K)^2}]^{1/3} n^{-1/3}. \tag{2.81}$$

Since, there is unknown quantities in this expression it cannot be used in practise for this reason we provide the bandwidth selector for the df under double truncation in the last chapter.

Theorem 9 *Let K be a kernel function whose satisfied $0 < k(x)$ and bounded with $\int k(x) = 1$, $\lim_{x\to\infty} |xk(x) = 0|$, and $b \to 0$ with increasing n. Then our estimators is asymptotic unbiased and consists.*

Proof Based on the theorem of density under double truncation defined in [29] and the inequality instantly which based on fubini theorem lead to the first part of the theorem.

$$\begin{aligned} |E\tilde{F}_b(x) - F(x)| &= |E \int \tilde{f}_b(x) - f(x)dx| \\ &\leq \int E|\tilde{f}_b(x) - f(x)|dx \end{aligned}$$

Hence from [29] we can find that $E|\tilde{f}_b(x) - f(x)| \xrightarrow[n\to\infty]{} 0$ which leads to result of the part 1. Now for proof the consist of our estimators it results from the the first part of the theorem and the mean square error (MSE)

$$MSE(\tilde{F}_b(x)) = Var(\tilde{F}_b(x)) + (E\tilde{F}_b(x) - F(x))^2$$

hence as $n \to \infty$ the proof is completed.

Theorem 10 *If K satisfies $K(x) < M < \infty$ on a compact support, h is such that $\sum exp(-\nu hn) < +\infty$ for each $\nu > 0$ H is continuous at x, x is a Lebesgue point of F if K is an even function, $b = o(n^{-1/3})$, $H^{-1}F$ has a second derivative which is bounded in a neighbourhood of x, and $F(x) > 0$, then*

$$(n)^{1/2}(\tilde{F}_b(x) - F(x)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, \alpha H^{-1}(x)F(x)S(x)).$$

Proof Let start by define,

$$F_b^*(y) = \frac{1}{n}\sum_{i=1}^{n} W(x - X_i), \tag{2.82}$$

which is the kernel estimator of the df and based on the true value of α and H and under the stable regularity conditions given in [38]. Therefore the proof is follow the theorem 6 of Watson and Leadbetter(1964) [48] who proof the asymptomatic normality of the kernel estimator of the df.

CHAPTER 3

HAZARD FUNCTION FOR DOUBLY TRUNCATED DATA

Estimation under double truncation has been an attractive topic and many scientific papers and researches dealt with this topic. For this reason, in this chapter we try to identify the special mechanisms that should be used in the estimation in order to minimize the impact of truncation, and we focus in particular on estimation of the hazard function.

3.1 *The NPMLE of Hazard function*

Now, to define the hazard function based on defining of the NPMLE given in 2.2, first let consider X is the variable of interest which is suppose to be observed just inside known interval, then the observed data is defined by $\{(U_i, X_i, V_i)i = 1, ..., n/U_i \leq X_i \leq V_i\}$. Let f_i is the distribution probability on $X_i, i = 1, ..., n$, and Let G_i is the joint distribution probability on $(U_i, V_i), i = 1, ..., n$ respectively. Hence the NPMLE in this case is given by

$$L = \prod_{j=1}^{n} f_j g_j / \sum_{i=1}^{n} F_i g_i = L_1(f) * L_2(f, g), \tag{3.1}$$

where $F_i = \sum_{k=1}^{n} f_k \pi_{i,k}$, and $\pi_{i,k} = \mathbf{1}_{\{U_i \leq\ X_k \leq\ V_i\}}$ is the indicator function, Efron and Patrson (1999) consider the first parts which is given by $L_1(f) = \prod_{j=1}^{n} f_j / F_i$, thus the NPMLE of $(f_1, f_2, ..., f_n)$ is defined by

$$\hat{f}_j = (\sum_{i=1}^{n} \pi_{i,e}(1/\hat{F}_i))^{-1}, \forall j = 1, ..., n, \tag{3.2}$$

as $\hat{F}_i = \sum_{k=1}^{n} \hat{f}_k \pi_{i,k}$, thus the estimator of the hazard function in this case is defined by $\hat{h}_j = \hat{f}_j / 1 - \hat{F}_j$ is given by this formula

$$\hat{h}_j = (\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_j \leq\ X_i\}} + \sum_{i=1}^{n} \pi_{i,j} \frac{\hat{S}(V_i)}{\hat{F}_i})^{-1}, \forall j = 1, ..., n, \tag{3.3}$$

where $\hat{S}(V_i) = \sum_{j=1}^{n} \hat{f}_j \mathbf{1}_{\{X_j > V_i\}}$ is the curde estimator of the survival function when $X = V_i$.

Lemma 1 *We have*

$$\hat{h}(X_j) = \frac{\hat{f}_j}{\sum_{k=1}^{n} \hat{f}_k \boldsymbol{1}_{\{X_k \geq X_j\}}}. \tag{3.4}$$

Proof Let consider $\tilde{h}(X_{(j)}) = (\hat{F}(X_{(j)}) - \hat{F}(X_{(j-1)}))/(1 - \hat{F}(X_{(j-1)})), j = 1, ..., n$, we can write

$$\hat{F}(X_{(j)}) - \hat{F}(X_{(j-1)}) = (\sum_{i=1}^{n} \frac{\mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ V_i\}}}{\hat{F}(V_i) - \hat{F}(U_i^-)})^{-1}. \tag{3.5}$$

Then

$$\tilde{h}(X_{(j)}) = ((\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ V_i\}})\{\frac{1 - \hat{F}(X_{(j-1)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\})^{-1}. \tag{3.6}$$

And

$$\hat{h}(X_{(j)}) = ((\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ X_i\}} + \mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ V_i\}})\{\frac{1 - \hat{F}(V_i)}{\hat{F}(V_i) - \hat{F}(U_i^-)}\})^{-1}. \tag{3.7}$$

Note that to proof this lemma we proof $\tilde{h}(X_{(j)}) = \hat{h}(X_{(j)})$ for $j = 1, ..., n$ in more simplification we need to proof this equation

$$\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ X_i\}} = ((\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(j)} \leq\ V_i\}})\{\frac{\hat{F}(V_i) - \hat{F}(X_{(j-1)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}). \tag{3.8}$$

Now, to continue proving, we need to use induction on i.

- For $j = 1$, we have $\hat{F}(X_{(j-1)}) = \hat{F}(X_{(0)}) = 0$ and $\mathbf{1}_{\{U_i \leq\ X_{(1)} \leq\ X_i\}} = 1$ for $i = 1, ..., n$ we note that $\hat{F}(U_i^-) = 0$. We can now see that $\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(1)} \leq\ X_i\}} = \sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(1)}\}} = \sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(1)} \leq\ V_i\}}$, therefore, our conclusion is that the assertion is true.

- For $j = k$, we have

$$\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i\}} = ((\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i\}})\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k-1)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}). \tag{3.9}$$

  Now, we need to proof this equation i.e.,

$$\overbrace{\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k+1)} \leq\ X_i\}}}^{\text{Part 1}} = \underbrace{((\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k+1)} \leq\ V_i\}})\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\})}_{\text{Part 2}}. \tag{3.10}$$

  Note that part 2 is equivalent to

$$((\sum_{i=1}^{n} (\mathbf{1}_{\{U_i \leq\ X_{(k)} < X_{(k+1)} \leq\ V_i\}} + \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}))\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}). \tag{3.11}$$

$$= ((\sum_{i=1}^{n} (\mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i\}} - \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i \leq X_{(k+1)}\}} + \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}))\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}). \tag{3.12}$$

  Hence

$$= \overbrace{((\sum_{i=1}^{n} (\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i\}}\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k-1)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}}^{A} - \underbrace{\sum_{i=1}^{n} (\mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i\}}\{\frac{\hat{F}(X_{(k)}) - \hat{F}(X_{(k-1)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}}_{B}$$

$$- \overbrace{\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i \leq X_{(k+1)}\}}\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}}^{C} + \underbrace{\sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}))\{\frac{\hat{F}(V_i) - \hat{F}(X_{(k)})}{\hat{F}(V_i) - \hat{F}(U_i^-)}\}))}_{D}.$$

  Since the assertion holds for $j = k$, we gate

$$(A) = \sum_{i=1}^{n} (\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i\}}). \tag{3.13}$$

Now, we note that $\hat{F}(X_{(k)}) - \hat{F}(X_{(k-1)}) = (\sum_{i=1}^{n} \frac{\mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ V_i\}}}{\hat{F}(V_i) - \hat{F}(U_i^-)})^{-1}$. We find $B = 1$. We note that $\hat{F}(V_i) = \hat{F}(X_{(k)})$, as $X_{(k)} \leq\ V_i < X_{(k+1)}$, then $C = 0$. And $\hat{F}(X_{(k)}) = \hat{F}(U_i^-) =$, as $X_{(k)} < U_i \leq\ X_{(k+1)}$, then $D = \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}$.
By simplification of part 2 of equation 3.10 we find is equal to

$$\sum_{i=1}^{n}(\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i\}}) - 1 + \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}. \tag{3.14}$$

Hence the part 1 of equation 3.10 is given by

$$\sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq\ X_{(k+1)} \leq\ X_i\}} = \sum_{i=1}^{n}(\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i\}}) + \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ X_i\}} \tag{3.15}$$

$$- \sum_{i=1}^{n} \mathbf{1}_{\{U_i \leq X_{(k)} \leq\ X_i \leq\ X_{(k+1)}\}}. \tag{3.16}$$

By equivalent of this two equations we find

$$\sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ X_i\}} - \sum_{i=1}^{n}(\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i \leq\ X_{(k+1)}\}}) = \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}^{-1}}. \tag{3.17}$$

We have

$$\sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}^{-1}} = \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} \leq\ U_i \leq\ X_{(k+1)} \leq\ X_i\}} + \sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}}. \tag{3.18}$$

We have $\mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}} = 0, \forall i = 1, ..., n.$, we result that $\sum_{i=1}^{n} \mathbf{1}_{\{X_{(k)} < U_i \leq\ X_{(k+1)} \leq\ V_i\}} = 0$, and this give us $\sum_{i=1}^{n}(\ \mathbf{1}_{\{U_i \leq\ X_{(k)} \leq\ X_i \leq\ X_{(k+1)}\}}) = 1$ thus the proof is finished.

3.2 *The smooth estimator of hazard function*

In this section, we consider the smooth estimator of the hazard function which is defined in [31] by this formula

$$h_b(x) := \int k_b(x-z)\Delta_n(dz) = \int k_b(x-z)\frac{dF_n(z)}{1-F_n(z^-)},$$

where F_n is Efron and Patrson estimator's of the distribution function given in 2.2, and $k_b(z) = (1/b)k(z/b)$ is the kernel function and b is the ordinal bandwidth with classical condition of regularity.

Now, for the evaluation of the asymptotic behavior of the this estimator we define an asymptotically equivalent estimator by

$$\tilde{h}_b(x) := \alpha\sum_{i=1}^{n} k_b(x-X_i)\frac{H(X_i)^{-1}}{1-F(X_i)}. \tag{3.19}$$

Theorem 11 *1. If K is bounded on a compact support, b is such that $\sum_{i=1}^{\infty}\exp(-\nu bn) < \infty$ for each $\nu > 0$, H is continuous at x, and x is a Lebesgue point of f, then $\tilde{h}_b(x) \xrightarrow[n\to\infty]{as} h(x)$.*

2. If, in addition to the conditions in part 1, K is an even function, $b = o(n^{-1/5})$, $H^{-1}/(1-F)h$ has a second derivative which is bounded in a neighborhood of x, and $h(x) > 0$, then

$$(nb)^{-1/2}(\tilde{h}_b(x) - h(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \alpha H(x)^{-1}(1-F(x))^{-1}h(x)R(k)), \tag{3.20}$$

where $R(k) := \int K^2(x)dx$.

Proof For 1, let $\tilde{h}^*_{0,b}(x) = \alpha\frac{H(x)^{-1}}{1-F(x)}f_{0,b}(x)$ and $f_{0,b}(x)$ is the kernel estimator of density which define for completed data and is consist estimator i.e., $f^*_{0,b}(x) \xrightarrow[n\to\infty]{p} f^*(x)$.
under the condition that K is contained in $[-a,a]$, we have

$$|\tilde{h}_b(x) - h^*_{0,b}(x)| \le \alpha f^*_{0,b}(x)\sup_{x-ab\le y\le x+ab}|\frac{H(y)^{-1}}{1-F(y)} - \frac{H(x)^{-1}}{1-F(x)}|,$$

hence $\sup_{x-ab\le y\le x+ab}|H(y)^{-1} - H(x)^{-1}|$ is converge to zero and this by the continuity of G at x.

For 2 we follow this same procedure in [12] we gate

$$(nb)^{-1/2}(\tilde{h}_b(x) - E(\tilde{h}_b(x)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \alpha\frac{H(x)^{-1}}{1-F(x)}h(x)R(k)), \tag{3.21}$$

hence, we use the Taylor expansion we find $E(h^*_b(x)) - h(x) = o(b^2)$. For $nb^5 \to 0$, then we complete the proof.

Now, in order to define the bias and variance we need the list of assumptions below

H 1 The kernel function satisfies: $K(t) > 0$, $\int K(t) = 1$, $\int tK(t) = 0$, $\int t^2K(t) < \infty$, and $\int K(t)^2dt < 0$.

H 2 The bandwidth $b = b_n$ satisfies: $b \to 0$ and $nb \to \infty$ when $n \to \infty$.

H 3 The functions h and $H^{-1}/(1-F)h$ are continuously differentiable twice around x.

Assuming regularity of the kernel and bandwidth, the mean and variance, we gate

$$E(\tilde{h}(x)) = h(x) + \frac{1}{2}b^2h"(x)\mu_2(K) + o(b^2) \tag{3.22}$$

$$Var(\tilde{h}(x)) = \frac{1}{nb}\alpha\frac{H(x)^{-1}}{1-F(x)}h(x)R(K) + o((nb)^{-1}). \tag{3.23}$$

Now, the formula of MSE of the estimator, is given by

$$AMSE(\tilde{h}(x)) := \frac{1}{4}b^4h"^2(x)\mu_2^2(K) + \frac{1}{nb}\alpha\frac{H(x)^{-1}}{1-F(x)}h(x)R(K). \tag{3.24}$$

Hence, the asymptotic formula of $MISE$ is given by

$$AMISE(\tilde{h}) := \int MSE(\tilde{h})(x)dx = \frac{1}{4}b^4R(h")\mu_2^2(K) + \frac{1}{nb}\alpha R(K)\int\frac{H(x)^{-1}}{1-F(x)}h(x)dx. \tag{3.25}$$

Thus, the asymptotically optimal bandwidth is defined by

$$b_{AMISE} := [\frac{\alpha R(K)\int H(x)^{-1}(1-F(x))^{-1}h(x)dx}{R(h")\mu_2^2(K)}]^{-1/5}n^{-1/5}. \tag{3.26}$$

3.3 *The proposed estimator*

In this section, we are going to provide a new estimator of hazard function. Of course, as we know the hazard function is defined by

$$h(x) = \frac{f(x)}{1 - F(x)}. \tag{3.27}$$

Thus, our proposed estimator is based on the estimation of the denominator and numerator of 3.27. Where an estimator of $f(x)$ is given in [29] by

$$\begin{aligned} f_n(x) &= \int k_b(x-t)dF_n(t) \\ &= \frac{\alpha_n}{n}\sum_{i=1}^{n} k_b(x - X_i)H(X_i)^{-1}, \end{aligned}$$

and F_n is defined in [10] by

$$F_n(x) = \alpha_n \int_{a_X}^{x} H(z)^{-1} d\tilde{F}_n(z). \tag{3.28}$$

Now, our proposed estimator is given by this formula

$$\hat{h}(x) = \frac{\hat{f}(x)}{1 - \hat{F}(x^-)}. \tag{3.29}$$

We note that we have this approximation $\frac{1}{1-x} \sim 1 + x$ is for small $x > 0$, so by simplification we find an alternative estimator of the hazard function

$$\hat{h}(x) = \hat{f}(x)(2 - \hat{S}(x^-)). \tag{3.30}$$

From Shen [38] we have

$$\hat{S}(x^-) - S(x) = F(x) - \hat{F}(x^-) = o_p(n)^{-2}, \tag{3.31}$$

hence, for this reason we can say that this both estimators $\hat{h}(x) = \hat{f}(x)(2 - \hat{S}(x))$ and $\tilde{h}(x) = \hat{f}(x)(2 - S(x^-))$ are equivalent. Thus, for given the asymptotic properties of the alternative estimator we use the asymptotic equivalent estimator

$$E(\tilde{h}(x)) = f(x)(2 - S(x)) + \frac{1}{2}b^2 f"(x)(2 - S(x))\mu_2(K) + o(b^2). \tag{3.32}$$

For the bias we note that $f(x)(2 - S(x)) \sim h(x)$, then we find

$$bias(\tilde{h}(x)) = \frac{1}{2}b^2 f"(x)(2 - S(x))\mu_2(K) + o(b^2). \tag{3.33}$$

And

$$Var(\tilde{h}(x)) = (nb)^{-1}(2 - S(x))\alpha H(x)^{-1}h(x)R(k) + o((nb)^{-1}). \tag{3.34}$$

Now, the $AMSE$ is defined by

$$AMSE(h^*(x)) = \frac{1}{4}b^4 f"^2(x)(2-S(x))^2\mu_2^2(K) \tag{3.35}$$

$$+ (nb)^{-1}\alpha R(k)(2-S(x))H(x)^{-1}h(x), \tag{3.36}$$

and $AMISE = \int AMSE(h^*(x))dx$ which is given by

$$AMISE(h^*) = \frac{1}{4}b^4\mu_2^2(K)R(f"(2-S)) \tag{3.37}$$

$$+ \frac{1}{nb}\alpha R(k)\int(2-S(x))H(x)^{-1}h(x)dx. \tag{3.38}$$

The optimal bandwidth is defined by

$$b_{AMISE} = \{\frac{\alpha R(k)\int(2-S(x))H(x)^{-1}h(x)dx}{R(f"(2-S))\mu_2^2(K)}\}^{1/5}n^{-1/5}. \tag{3.39}$$

3.4 *Asymptotic properties of the proposed estimator*

In this part, we investigate the strongly consist and the asymptotic normality of our proposed estimator

Theorem 12 *1. We assume that regularity condition given in [29] and [38] are satisfied, hence we have*

$$(nb)^{-1}(\hat{h}(x)-h(x)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, S(x)^{-2}\alpha H(x)^{-1}f(x)R(k)). \tag{3.40}$$

2. We assume that the condition still hold, then we gate

$$\hat{h}(x) \xrightarrow[n\to\infty]{ps} h(x). \tag{3.41}$$

Proof 1. We note that from Moreira [29] in theorem 2, we can say that

$$(nb)^{-2}(\tilde{f}(x)-f(x)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, (\alpha H(x)^{-1}f(x)R(k)). \tag{3.42}$$

In addition, we have this simplifaction

$$\begin{aligned}
\hat{h}(x)-h(x) &= \{\hat{h}(x)-h(x)\}+\{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)}{S(x)}-\frac{\hat{S}(x^-)}{S(x)}) \\
&= \{\hat{h}(x)-h(x)\}+\{\frac{\hat{f}(x)}{S(x)}-\frac{h(x)\hat{S}(x^-)}{S(x)}\} \\
&- \{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)}{S(x)}-1+1) \\
&= \frac{\hat{f}(x)-h(x)\hat{S}(x^-)}{S(x)}-\{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)-S(x)}{S(x)}) \\
&= \frac{1}{S(x)}\{\hat{f}(x)-h(x)\hat{S}(x^-)+h(x)S(x)-h(x)S(x)\} \\
&- \{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)-S(x)}{S(x)}) \\
&= \frac{1}{S(x)}\{\hat{f}(x)-h(x)\{\hat{S}(x^-)-S(x)\}-h(x)S(x)\} \\
&- \{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)-S(x)}{S(x)}) \\
&= \frac{1}{S(x)}\{\hat{f}(x)-h(x)\{\hat{S}(x^-)-S(x)\}-f(x)\} \\
&- \{\hat{h}(x)-h(x)\}(\frac{\hat{S}(x^-)-S(x)}{S(x)}).
\end{aligned} \tag{3.43}$$

Now, we note that $\hat{S}(x^-) - S(x) = F(x) - \hat{F}(x^-) = o_p(n)^{-2}$ and this given by [38] who proof the uniform consistency and the asymptotically normality in Theorem 3 and 2 of the NPMLE given in section 2.2, hence we gate

$$\hat{h}(x) - h(x) = \frac{1}{S(x)}\{\hat{f}(x) + o_p(n)^{-2} - f(x)\} + o_p(n)^{-2}. \tag{3.44}$$

Finally $\tilde{f}$ is asymptotically equivalent with$f$, hence we can find the result.

2. For the proof of the strongly consist we follow the same simplification as in part 1 of the proof and we confide our results by the consist of density given in [29].

The regularity assumptions we make are as follows. In order to define the asymptotic variances and mean of the proposed estimator:

H 1 The kernel k is symmetric, positive function and satisfies $\int k(t)dt = 1$, $\mu_2(k) = \int t^2 k(t)dt < \infty$, and $R(k) = \int k^2(t)dt < \infty$.

H 2 The bandwidth sequence $b = b_n$ satisfied $b \to 0$, $bn \to \infty$ as $n \to \infty$.

H 3 The functions $h(x)$ and $h(x)H(x)^{-1}S^{-1}(x)$ are twice continuously differentiable around x.

Theorem 13 *Under the assumption of regularity list defined before, the mean and variance are given by*

$$E(\hat{h}(x)) = h(t) + \frac{1}{2}b^2\frac{f"(x)}{S(x)}\mu_2(K) + o(b^2), \tag{3.45}$$

and

$$Var(\hat{h}(x)) = (nb)^{-1}S(x)^{-1}\alpha H(x)^{-1}h(x)R(k) + o((nb)^{-1}). \tag{3.46}$$

Proof We follow the same preceding simplification for the proof as before, and we confine the result to the bias and variance of density which is defined in[29].

The $AMSE$ is given by

$$AMSE(\hat{h}(x)) = \frac{1}{4}b^4\frac{f"^2(x)}{S^2(x)}\mu_2^2(K) + nb^{-1}S(x)^{-1}\alpha H(x)^{-1}h(x)R(k), \tag{3.47}$$

then we have $AMISE(\hat{h}) = \int AMSE(\hat{h}(x))dx$

$$AMISE(\hat{h}) = \frac{1}{4}b^4(R(f"/S))\mu_2^2(K) + (nb)^{-1}\alpha R(k)\int S(x)^{-1}H(x)^{-1}h(x)dx. \tag{3.48}$$

The optimal bandwidth is given by

$$b_{AMISE} = \{\frac{\alpha R(k)\int S(x)^{-1}H(x)^{-1}h(x)dx}{R(f"/S)\mu_2^2(K)}\}^{1/5}n^{-1/5}. \tag{3.49}$$

CHAPTER 4

SIMULATION

Double truncation is one of the major challenges in data analysis, which happing when the observations are lost. For this reason, many researchers have focused on solving this problem, in this thesis we have introduced many technical and works in order to solve the problems that connected to truncation especially estimation of the hazard function. Now, to evaluate our solution we introduce simulation with the program R.

4.1 *Simulation data*

To investigate the estimator's behavior over a finite sample, we run five models with varying truncation percentages. We perform 1000 Monte Carlo trials using five different types of estimators, in order to define the bias and root mean square error RMSE of each estimator. And from the strong law of large numbers we have $n/N \to \alpha$.

- $\hat{h}_1(x) = \frac{\hat{f}(x)}{1-\hat{F}(x^-)}$ is our proposed estimator where $\hat{F}$ is Efron and Patrson estimators'.
- $\hat{h}_2(x) = \frac{\hat{f}(x)}{1-F_b(x)}$ is our proposed estimator where F is estimated by the kernel function and we take the cumulative kernel is Tukey integer kernel.
- $\hat{h}_3(x)$ is estimator of hazard function given in [31].
- $\hat{h}_4(x)$ is estimator of hazard function which is cured estimator given in section 2.2.
- $\hat{h}_5(x) = \hat{f}(x)(2 - \hat{S}(x^-))$ where $\hat{S}$ is the survival function.

Model 1 The tables are simulation for model where we assume X follow Weibull distribution with $(2, 1)$ and the truncation limit is

1. we assume that and U from $U(0, 1)$ and $V = U + 0.25$
2. we assume that and U from $U(-1, 1)$ and $V = U + 0.75$
3. we assume that and U from $U(-1/3, 1)$ and $V = U + 1.5$

Model 2 The tables are simulation for model where we assume X follow exponential distribution with $\lambda = 2$ and the truncation limit is

1. we assume that and U from $U(-1/3, 1)$ and $V = U + 1.25$
2. we assume that and U from $U(-1, 1)$ and $V = U + 0.75$

Model 3 The table is simulation for model where we assume X follow exponential distribution with $\lambda = 1/2$ and the truncation limit is

1. we assume that and U from $U(-1, 1)$ and $V = U + 0.75$

Model 4 The table is simulation for model where we assume X follow Pareto distribution with $(0.5, 1)$ and the truncation limit is

1. we assume that and U from $U(0, 1)$ and $V = U + 0.75$

Model 5 The table is simulation for model where we assume X follow normal distribution with $(0, 1)$ and the truncation limit is

1. we assume that and U from $U(0, 1)$ and $V = U + 0.75$.

$\hat{\alpha} \approx 0.2$				
n	50	150	250	500
b_{AMISE}	9.590787e-05	1.643863e-05	1.217965e-05	4.411124e-06

Table 4.1: b_{AMISE} is by normal reference for the distribution function for model 1.1

$\hat{\alpha} \approx 0.2$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-0.4987343	-1.551871	-1.404262	-1.400402	-0.4948653
RMSE		0.6883872	1.993602	1.514685	1.507872	0.6794141
bias	150	-0.4103777	-1.592309	-1.282785	-1.276606	-0.4038327
RMSE		0.7480329	2.073556	1.433679	1.423864	0.7239803
bias	250	-0.4155057	-1.603536	-1.295201	-1.288407	-0.429192
RMSE		0.8166817	2.139318	1.447967	1.437386	0.7909527
bias	500	-0.3864767	-1.615672	-1.295543	-1.288407	-0.4337807
RMSE		0.6112144	2.036588	1.448258	1.437386	0.6413404

Table 4.2: The bias and RMSE for the distribution function for model 1.1

$\hat{\alpha} \approx 0.3$				
n	50	150	250	500
b_{AMISE}	0.0006845909	2.615753e-05	5.406701e-05	1.735596e-06

Table 4.3: b_{AMISE} is by normal reference for the distribution function for model1.2

$\hat{\alpha} \approx 0.3$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-1.000838	-1.81778	-1.637239	-1.633883	-0.9674744
RMSE		1.316571	2.346835	1.814636	1.805315	1.261369
bias	150	-1.02246	-1.84628	-1.643964	-1.636491	-1.005916
RMSE		1.423244	2.382401	1.87008	1.857571	1.39499
bias	250	-1.152154	-1.892539	-1.73594	-1.730323	-1.149207
RMSE		1.01155	2.399303	2.39929	1.966733	1.516165
bias	500	-1.08678	-1.890453	-1.696052	-1.689128	-1.107132
RMSE		1.012587	2.398503	1.540688	1.941393	1.501335

Table 4.4: The bias and RMSE for the distribution function for model 1.2

$\hat{\alpha} \approx 0.8$				
n	50	150	250	500
b_{AMISE}	0.002905115	0.0002536418	0.0001298252	3.82446e-05

Table 4.5: b_{AMISE} is by normal reference for the distribution function for model1.3

$\hat{\alpha} \approx 0.8$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-1.306149	-1.891511	-1.865144	-1.851915	-1.199085
RMSE		1.650882	2.45611	2.05168	2.033896	1.55546
bias	150	-1.790735	-1.979969	-2.288248	-2.278835	-1.750873
RMSE		2.224652	2.561692	2.541543	2.526473	2.185148
bias	250	-1.709965	-1.987548	-2.221494	-2.210388	-1.692169
RMSE		2.164187	2.547219	2.498999	2.482699	2.139623
bias	500	-1.782657	-2.020467	-2.289813	-2.279496	-1.800552
RMSE		2.314576	2.593274	2.620787	2.605093	2.329907

Table 4.6: The bias and RMSE for the distribution function for model 1.3

4.2 Analyze the results

Through our simulation we can see that the percentage of truncation is fixed in each model where it is not affected by the sample size in each model, but the effect of changing the

$\hat{\alpha} \approx 0.5$				
n	50	150	250	500
b_{AMISE}	0.001905903	0.0002235539	6.899066e-05	1.750927e-05

Table 4.7: b_{AMISE} is by normal reference for the distribution function for model2.1

$\hat{\alpha} \approx 0.5$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-1.554278	-1.573753	-1.996503	-1.991573	-1.545449
RMSE		1.67521	1.681753	1.996565	1.996565	1.646042
bias	150	-1.521073	-1.566264	-1.998557	-1.991696	-1.530431
RMSE		1.620257	1.650866	1.998561	1.992434	1.613236
bias	250	-1.490832	-1.543618	-1.999164	-1.99171	-1.513274
RMSE		1.62075	1.659437	1.999165	1.99245	1.622201
bias	500	-1.477657	-1.539427	-1.999556	-1.990368	-1.507373
RMSE		1.57804	1.630695	1.999556	1.991127	1.594828

Table 4.8: The bias and RMSE for the distribution function for model 2.1

$\hat{\alpha} \approx 0.4$				
n	50	150	250	500
b_{AMISE}	0.0002263451	2.89504e-05	9.324716e-06	2.075316e-06

Table 4.9: b_{AMISE} is by normal reference for the distribution function for model2.2

truncation limit model make the percentage of truncation different. Moreover, from the analysis of each estimator we find that our proposed estimator is more stable where is give the smallest value of rmse then the other estimators, and of course we can see that the three estimators h_1, h_2, h_5 are so close from the results then the two others. In addition, as n increases, all the estimators behave well and the optimal bandwidth of the distribution function becomes more stable.

$\hat{\alpha} \approx 0.4$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-1.158746	-1.185233	-1.999556	-1.987973	-1.14011
RMSE		1.396928	1.405667	1.999556	1.988755	1.37798
bias	150	-1.117092	-1.170076	-1.997897	-1.987973	-1.127265
RMSE		1.33942	1.359977	1.997898	1.988755	1.322894
bias	250	-1.284083	-1.337228	-1.999287	-1.990167	-1.305629
RMSE		1.455126	1.486857	1.999287	1.990947	1.45848
bias	500	-1.31128	-1.364975	-1.999494	-1.99062	-1.337663
RMSE		1.439125	1.480825	1.999494	1.991394	1.454306

Table 4.10: The bias and RMSE for the distribution function for model 2.2

$\hat{\alpha} \approx 0.2$				
n	50	150	250	500
b_{AMISE}	0.001828573	4.353783e-05	1.149797e-05	3.043753e-06

Table 4.11: b_{AMISE} is by normal reference for the distribution function for model3

$\hat{\alpha} \approx 0.2$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	0.00965931	-0.1761755	0.1865518	-0.4972997	-0.4925952
RMSE		0.265063	0.3202854	0.4905645	0.4973049	0.4956071
bias	150	0.05419744	0.1921511	0.09718899	-0.4985292	-0.4917938
RMSE		0.1435797	0.3831946	0.1625938	0.4985293	0.4948693
bias	250	0.1323443	0.07666908	0.1164615	-0.4989891	-0.4904571
RMSE		0.1674689	0.3454693	0.1598307	0.4989891	0.4935912
bias	500	0.07045756	0.06168726	0.05378868	-0.4996259	-0.4904571
RMSE		0.2337028	0.2516658	0.2339528	0.4996259	0.4935912

Table 4.12: The bias and RMSE for the distribution function for model 3

$\hat{\alpha} \approx 0.4$				
n	50	150	250	500
b_{AMISE}	0.0005144181	6.356409e-05	2.598838e-05	5.136123e-06

Table 4.13: b_{AMISE} is by normal reference for the distribution function for model4

$\hat{\alpha} \approx 0.4$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-0.18512351	0.1086684	-1.126538	-0.5658987	-0.6027465
RMSE		0.9881321	0.4914275	1.183836	0.6129755	0.643658
bias	150	-0.5880179	0.2376235	-1.066797	-1.058288	-0.5762382
RMSE		0.6233867	0.7298765	1.125536	1.123233	0.6127834
bias	250	-0.5916553	0.2107076	-1.053169	-1.043707	-0.591513
RMSE		0.6147742	0.80411293	1.113373	1.109953	0.6138889
bias	500	-0.5788424	0.2200688	-1.030481	-1.021248	-0.5829258
RMSE		0.61113	0.8277529	1.094053	1.09041	0.6148102

Table 4.14: The bias and RMSE for the distribution function for model 4

$\hat{\alpha} \approx 0.2$				
n	50	150	250	500
b_{AMISE}	0.0004492293	0.0007347708	0.000527055	0.0002630326

Table 4.15: b_{AMISE} is by normal reference for the distribution function for model5

$\hat{\alpha} \approx 0.2$						
	n	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$
bias	50	-0.5532868	-1.016369	-0.9659516	-0.944781	-0.5392192
RMSE		0.735375	1.100765	1.046705	1.024815	0.6881205
bias	150	-0.5667239	-1.13382	-0.9734247	-0.9491034	-0.5622836
RMSE		0.7487492	1.351328	1.0666091	1.043343	0.7183722
bias	250	-0.5929445	-1.14138	-1.000867	-0.9764933	-0.6070306
RMSE		0.7898498	1.375191	1.098716	1.076008	0.771149
bias	500	-0.587167	-1.157865	-1.049023	-1.030998	-0.663146
RMSE		0.7892744	1.390444	1.156076	1.1383	0.8429926

Table 4.16: The bias and RMSE for the distribution function for model 5

CHAPTER 5

ON OPTIMAL BANDWIDTH SELECTION

Kernel estimation is a method to obtain a smooth estimator, however, this method has the problem of the selected nice bandwidth, as this bandwidth controls the degree of smoothness. For this reason in this chapter we investigate the problem of define bandwidth for the distribution function in the presence of double truncation data. Where we had defined both nonparametric and semiparametric estimators in the previous chapter.

5.1 Kernel smoothing estimation of the distribution function

Now, we will concentrate on the definition of the method of the chosen bandwidth of the distribution function, as the data are collected under double truncation. Let us therefore start by giving the definition of the nonparametric and the semiparametric of the df, the nonparametric estimator is given by

$$\begin{aligned} F_b(x) &:= \int_{a_X}^{x} f_b(z)dz \\ &= \alpha_n \frac{1}{n}\sum_{i=1}^{n} H_n(X_i)^{-1}\int_{a_X}^{x} k_b(t-X_i)dt \\ &= \alpha_n \frac{1}{n}\sum_{i=1}^{n} H_n(X_i)^{-1} W(\frac{x-X_i}{b}), \end{aligned}$$

and the semiparametric estimator of df is defined by

$$\begin{aligned} F_{b;\hat{\theta}}(x) &:= \int_{a_X}^{x} f_{b;\hat{\theta}}(z)dz \\ &= \alpha_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1}\int_{a_X}^{x} k_b(t-X_i)dt \\ &= \alpha_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} H_{\hat{\theta}}(X_i)^{-1} W(\frac{x-X_i}{b}), \end{aligned}$$

Now, under the assumption given in the previous chapter we have

$$E(\tilde{F}_b(x)) = F(x) + \frac{b^2}{2}F''(x)\mu_2(K) + o(b^2). \tag{5.1}$$

$$Var(\tilde{F}_b(x)) = n^{-1}\alpha H(x)^{-1}\{F(x)[1-F(x)] + bf(x)[J(k)] + o(b)\}, \tag{5.2}$$

where $J(k) = \int_{-1}^{1} W^2(u)du - 1$.

The asymptotic mean integrated squared error of this estimator is given under the regularity conditions by

$$\begin{aligned} AMISE(\tilde{F}_b) &= \int AMSE(\tilde{F}_b(x))dx \\ &= \frac{b^4}{4}\int f'(x)^2 dx\mu_2(K)^2 + n^{-1}\alpha\int H(x)^{-1}\{F(x)[1-F(x)] \\ &+ bf(x)[J(k)]dx\}. \end{aligned}$$

Then the asymptotically optimal bandwidth can be obtained by

$$b_{AMISE} = [\frac{(1-\int_{-1}^{1} W^2(u)du)\alpha\int[H(x)]^{-1}f(x)dx}{\int f'(x)^2 dx\mu_2(K)^2}]^{1/3} n^{-1/3}. \tag{5.3}$$

5.1.1 *Normal reference bandwidth for the cumulative kernel distribution function*

Now, as in the previous chapter on density function estimation, in order to facilitate the calculation of the undefined values in 5.3, we assume that the distribution of the variable of interest follows a normal distribution, so that this method is one of the simplest, although it suffers from some problems, especially when the variable of interest is far from normal distribution. Hence by this assumption we find $R(f') = 1/4\sigma^3\sqrt{\pi}$

$$\begin{aligned} b_{NR} &= [\frac{4\sqrt{\pi}(1-\int_{-1}^{1} W^2(u)du)\alpha \int H(x)^{-1}dF((x))}{\mu_2(K)^2}]^{1/3} n^{-1/3}\sigma \\ &= [\frac{4\sqrt{\pi}(1-\int_{-1}^{1} W^2(u)du)\alpha^2 \int H(x)^{-2}dF^*(x)}{\mu_2(K)^2}]^{1/3} n^{-1/3}\hat{\sigma}, \end{aligned}$$

where F^* can by estimated by the ordinary empirical distribution function. In addition $\hat{\sigma}$ is estimated as in density case and α and $H(x)$ as in [38].

5.1.2 *Plug in method*

We note that

$$\int F"^2(x)dx = \int f'^2(x)dx,$$

hence by integration by parties we find

$$\int F"^2(x)dx = -\int f"(x)f(x)dx.$$

Now, through the same steps as the kernel estimation we first define

$$\begin{aligned}\psi_r &= \int f^{(r)}(x)f(x)dx \\ &= E(f^{(r)}(x)),\end{aligned}$$

then we estimate this expression by

$$\begin{aligned}\hat{\psi}_r(g) &= \hat{\alpha}n^{-1}\sum_{i=1}^{n}\hat{f}^{(r)}(X_i)H(X_i)^{-1} \\ &= \hat{\alpha}^2 n^{-2}g^{-r-1}\sum_{i=1}^{n}\sum_{j=1}^{n}L^{(r)}(\frac{X_i - X_j}{g})H(X_i)^{-1}H(X_j)^{-1}.\end{aligned}$$

Now, by similar calculations as done [32], we find

$$g_{AMISE} = [-\frac{\alpha k! L^r(0)\int H(x)^{-1}f(x)dx}{\psi_{r+k}\mu_k(L)n}]^{1/r+k+1}.$$

Thus, rewrite the formula of optimal bandwidth the definition of ψ_2 as

$$b_{AMISE} = [-\frac{(1-\int_{-1}^{1}W^2(u)du)\alpha\int H(x)^{-1}f(x)dx}{\psi_2\mu_2(K)^2}]^{1/3}n^{-1/3}. \tag{5.4}$$

Then we estimate α, $\int H(x)^{-1}f(x)dx$ and ψ_2. In order for find the direct plug-in

$$\hat{b}_{DPI} = [-\frac{(1-\int_{-1}^{1}W^2(u)du)\hat{\alpha}^2\ \int H(x)^{-2}dF^*(x)}{\hat{\psi}_2(g)\mu_2(K)^2}]^{1/3}n^{-1/3}. \tag{5.5}$$

We follow the same procedure in [32] we find g with $r = 2$

$$g_{AMISE} = [-\frac{\alpha 2L^2(0)\int H(x)^{-1}f(x)dx}{\psi_4\mu_2(L)n}]^{1/5}$$

We find that the process is regressive in the sense that we keep needing unknown values and for this reason we use normal reference rule for estimate ψ as in completed data we gate

$$\psi_r = \frac{(-1)^{r/2}r!}{(2\sigma)^{r+1}(r/2)!(\pi)^{1/2}}, \tag{5.6}$$

which is estimated by

$$\hat{\psi}_r^{NR} = \frac{(-1)^{r/2}r!}{(2\hat{\sigma})^{r+1}(r/2)!(\pi)^{1/2}} \tag{5.7}$$

We work with the same steps $l = 0, 1, 2$ e.g., $l = 1$ the procedure is defined by

1. Calculate $\hat{\psi}_4^{NR}$.

2. Calculate $\hat{\psi}_2(g_1)$

$$g_1 = [-\frac{\hat{\alpha} 2L^{(2)}(0) \int H(x)^{-1} \hat{f}(x) dx}{\hat{\psi}_4^{NR} \mu_2(L) n}]^{1/5} \tag{5.8}$$

3. The selected bandwidth is

$$\hat{b}_{DPI} = [-\frac{(1 - \int_{-1}^{1} W^2(u) du) \hat{\alpha}^2 \int H(x)^{-2} dF^*(x)}{\hat{\psi}_2(g_1) \mu_2(K)^2}]^{1/3} n^{-1/3}. \tag{5.9}$$

In addition, and follow the density procedure we define the two-stage plug-in bandwidth selector by

1. Calculate $\hat{\psi}_6^{NR}$.

2. Calculate $\hat{\psi}_4(g_1)$

$$g_1 = [-\frac{\hat{\alpha} 2L^{(4)}(0) \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_6^{NR} \mu_2(L) n}]^{1/7}. \tag{5.10}$$

3. Calculate $\hat{\psi}_2(g_2)$

$$g_2 = [-\frac{\hat{\alpha} 2L^{(2)}(0) \int \hat{H}(z)^{-1} d\hat{F}(z)}{\hat{\psi}_4(g_1) \mu_2(L) n}]^{1/5}. \tag{5.11}$$

4. The bandwidth is

$$\hat{b}_{DPI;2} = [-\frac{(1 - \int_{-1}^{1} W^2(u) du) \hat{\alpha}^2 \int H(x)^{-2} dF^*(x)}{\hat{\psi}_2(g_2) \mu_2(K)^2}]^{1/3} n^{-1/3}. \tag{5.12}$$

5.1.3 *Cross-validation method for define the optimal bandwidth*

In completed data, cross-validation is widely used as a method to define an interesting bandwidth. Now in this part we define the method for gate bandwidth for df when the data are sampling under double truncation, for this reason let consider the integrated squared error ISE which define by

$$ISE(F_b) = \int (F_b(t) - F(x))^2 S(x) dF(x), \tag{5.13}$$

where S is a non negative weight function, thus the mean integrated squared error $MISE$ is defined by as in [1] for complete data by

$$MISE(F_b) = E(\alpha \int (F_b(t) - F(x))^2 S(x) H(x)^{-1} dF^*(x)). \tag{5.14}$$

Now, we note that this formula can be approximate by

$$ASE(F_b) = n^{-1} \alpha \sum ((F_b(X_i) - F(X_i))^2 S(X_i) H(X_i)^{-1}. \tag{5.15}$$

However, this formula is not useful in practice because it contains unknown values and, in order to solve this problem, we need to estimate the unknown values F, α and $H(x)$. Hence we gate the formula of leave-none-out estimator given by

$$LNO(b) = n^{-1} \alpha_n \sum ((F_b(X_i) - F_n(X_i))^2 S(X_i) H_n(X_i)^{-1}, \tag{5.16}$$

and the cross-validation formula define by

$$CV(b) = n^{-1} \alpha_n \sum ((\hat{F}_{b;-i}(X_i) - F_n(X_i))^2 S(X_i) H_n(X_i)^{-1}, \tag{5.17}$$

where $F_{b;-i}(x) = \alpha_{n;-i} \frac{1}{n-1} \sum_{i \neq j} H_{n;-i}(X_j)^{-1} W((x - X_j/b))$ is kernel estimator of the df where X_i is excluded. We have CV and LNO are asymptotic equivalent. The bandwidth that minimizes the criterion is chosen in either case.

5.1.4 *Bootstrap bandwidth selection*

The bootstrap procedure is defined as in completed data for $b = 1...B$ we defined

1. Let $X_{b,i}^{boot}$be an i.i.d. sample from F_θ, g , where g is chosen to be $\hat{b}_{DPI}$ well), and let $U_{b,i}^{boot}, V_{b,i}^{boot}$, $i = 1, ..., n$, be an i.i.d. sample from G_n. We repeated this step until the condition of observed data $U_{b,i}^{boot} \leq X_{b,i}^{boot} \leq V_{b,i}^{boot}$.

2. we define θ and $F_{b,\hat{\theta}}^{boot}(x)$ be the estimators based on the bootstrap sample defined in step 1.

Now we define

$$BMISE(b) = B^{-1} \sum_{b=1}^{B} \int (F_{b,\hat{\theta}^{boot}}^{boot}(x) - dF_{g,\hat{\theta}}(x))^2 \tag{5.18}$$

which as B big it close to

$$MISE^{boot}(b) = E^{boot} \int (F_{b,\hat{\theta}^{boot}}^{boot}(x) - dF_{g,\hat{\theta}}(x)d)^2 \tag{5.19}$$

Hence $\hat{b}^{boot} = argminBMISE(b)$.

GENERAL CONCLUSION

Throughout this thesis, we have defined a new estimator of the hazard function when the data is sampled under double truncation. Estimating the hazard function has been an attractive topics in sense that this function plays an important role in medicine and economics where it is used to know the probability of the risk, therefore we showed that our proposed estimator of this function give better results by simulation and the asymptotic properties of the proposed estimator is well defined.

Our proposed estimator is based in estimating the denominator and numerator by estimators defined in previous work and showed good behavior and also give very good simulation results. In addition in this thesis, we introduced a new smooth estimator of the distribution function which is purely nonparametric beside a semiparametric estimator and we defined the asymptotic properties of this estimators and also we use it in estimating of the denominator of the hazard function.

In order to evaluate our proposed estimators we use Monte Carlo simulation with different kind of distribution as the well know heavy-tailed distributions Weibull beside the Parto distribution, the Gaussian and exponential distribution etc. The simulation proof that our estimator has good behavior in all cases for the small and big percentage of truncation then the other existing estimators.

In the last chapter, we finish our work by introduce bandwidth selector of the distribution function for double truncation, where we defined the well know methods for define the bandwidth for example the plug in, cross validation and normal reference etc.

In our future work, we will investigate estimation in the quntile function, where the estimation under double truncation need special techniques to reduce the effect of truncation. In addition, in all this work we have not taken into account the small sample size, where under double truncation this situation needs a more robust estimator to give an attractive result. Moreover, this work does not deal with the situation for estimating hazard function as the variables of the truncation limit assume to follow parametric family, in another word, the semiparametric estimator of the hazard function when the data are subject to double truncation data which is consider in our future work.

BIBLIOGRAPHY

[1] Altman, N., and Leger, C. Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference 46*, 2 (1995), 195–214.

[2] Baszczyńska, A. Kernel estimation of cumulative distribution function of a random variable with bounded support. *Statistics in Transition. New Series 17*, 3 (2016), 541–556.

[3] Benchaira, S. *Statistics of incomplete data.* Doctorat these, Université Mohamed Khider Biskra, (2017).

[4] Bouhadjera, F. *Estimation non paramétrique de la fonction de régression pour des données censurées : méthodes locale linéaire et erreur relative.* Doctorat these, Université du Littoral Côte d'Opale et de l'Université Badji Mokhtar Annaba, (2020).

[5] Bowman, A., Hall, P., and Prvan, T. Bandwidth selection for the smoothing of distribution functions. *Biometrika 85*, 4 (1998), 799–808.

[6] Dai, H., and Wang, H. *Analysis for Time-to-Event Data under Censoring and Truncation.* Academic Press, (2016).

[7] de Uña-Álvarez, J. R packages for the statistical analysis of doubly truncated data: a review. *arXiv preprint arXiv:2004.08978* (2020).

[8] de Uña-Álvarez, J., Crujeiras, R., and Moreira, C. The statistical analysis of doubly truncated data: with applications in r. *John Wiley & Sons 64* (2021).

[9] Dörre, A., and Emura, T. *Analysis of doubly truncated data: an introduction.* Springer, (2019).

[10] Efron, B., and Petrosian, V. Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association 94* (1999), 824–834.

[11] El Bay, R., and Yahia, D. Estimating hazard functions in the presence of double truncation in statistical data analysis. *Studies in Engineering and Exact Sciences 5*, 1 (2024), 2813–2830.

[12] Emanuel, P. On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33* (1962), 1065–1076.

[13] Emura, T., Hu, Y., and Huang, C. Double. truncation: analysis of doubly-truncated data, 2018.

[14] Emura, T., Konno, Y., and Michimae, H. Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation. *Lifetime data analysis 21* (2015), 397–418.

[15] Falk, M. Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica 37*, 2 (1983), 73–83.

[16] Ghettab, S. *Estimation non Paramétrique par La méthode du noyau.* Doctorat these, Universite des sciences et de la technologie Houari Boumediene, (2014).

[17] Guessoum, Z., and Tatachak, A. On kernel hazard rate function estimate for associated and left truncated data. *REVSTAT-Statistical Journal 18*, 3 (2020), 337–355.

[18] Horová, I., Koláček, J., and Zelinka, J. *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing.* World scientific, (2012).

[19] Kalbfleisch, J., and Lawless, J. Some useful statistical methods for truncated data. *Journal of Quality Technology 24*, 3 (1992), 145–152.

[20] Kaplan, E. L., and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53* (1958), 457–481.

[21] Kim, C., Bae, W., Choi, H., and Park, B. Non-parametric hazard function estimation using the kaplan–meier estimator. *Nonparametric Statistics 17*, 8 (2005), 937–948.

[22] Klein, J., and Moeschberger, M. Censoring and truncation. *Survival Analysis: Techniques for Censored and Truncated Data 10* (2003), 63–90.

[23] Lejeune, M. *Statistique: La théorie et ses applications*. Springer Science & Business Media, 2004.

[24] Lejeune, M., and Sarda, P. Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis 14*, 4 (1992), 457–471.

[25] Lynden, B. A method of allowing for known observational selection in small samples applied to 3cr quasars. *Monthly Notices of the Royal Astronomical Society 155* (1971), 95–118.

[26] Michael, W. Estimating a distribution function with truncated data. *Annals of Statistics 13* (1985), 163–177.

[27] Moreira, C., and de Una-Alvarez, J. Bootstrapping the npmle for doubly truncated data. *Journal of Nonparametric Statistics 22* (2010), 567 – 583.

[28] Moreira, C., and de Uña Álvarez, J. A semiparametric estimator of survival for doubly truncated data. *Statistics in Medicine 29*, 30 (2010), 3147–3159.

[29] Moreira, C., and de Uña Álvarez, J. Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics 6* (2012), 501–521.

[30] Moreira, C., de Una-Alvarez, J., and Crujeiras, R. M. Dtda: An r package to analyze randomly truncated data. *Journal of Statistical Software 37* (2010), 1–20.

[31] Moreira, C., de Uña-Álvarez, J., Santos, A., and Barros, H. Smoothing methods to estimate the hazard rate under double truncation. *arXiv preprint arXiv:2103.14153* (2021).

[32] Moreira, C., and Van Keilegom, I. Bandwidth selection for kernel density estimation with doubly truncated data. *Comput. Stat. Data Anal. 61* (2013), 107–123.

[33] Nadaraya, E. Some new estimates for distribution functions. *Theory of Probability & Its Applications 9*, 3 (1964), 497–500.

[34] Ones, M., and Sheather, S. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters 11*, 6 (1991), 511–514.

[35] Patil, P., Wells, M., and Marron, J. Some heuristics of kernel based estimators of ratio functions. *Journal of Nonparametric Statistics 4*, 2 (1994), 203–209.

[36] Peter, D. Kernel estimation of a distribution function. *Communications in Statistics-Theory and Methods 14*, 3 (1985), 605–620.

[37] Sarda, P. Estimating smooth distribution functions. *Nonparametric Functional Estimation and Related Topics* (1991), 261–270.

[38] Shen, P. Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics 62*, 5 (2010), 835–853.

[39] Shen, P. Semiparametric analysis of doubly truncated data. *Communications in Statistics—Theory and Methods 39*, 17 (2010), 3178–3190.

[40] Shen, P. Nonparametric analysis of doubly truncated and interval-censored data. *Statistical Methods in Medical Research 31* (2022), 1157 – 1170.

[41] Silverman, B., and Young, G. The bootstrap: to smooth or not to smooth? *Biometrika 74*, 3 (1987), 469–479.

[42] Tanner, M., and Wong, W. The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics 11*, 3 (1983), 989–993.

[43] Terrell, G., and Scott, David, W. Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association 80*, 389 (1985), 209–214.

[44] Turnbull, B. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological) 38*, 3 (1976), 290–295.

[45] Wand, M., and Jones, M. *Kernel smoothing*. CRC press, (1994).

[46] Wand, M., and Jones, M. Univariate kernel density estimation. *Kernel smoothing* (1995), 10–57.

[47] Watson, G., and Leadbetter, M. Hazard analysis. i. *Biometrika 51*, 1/2 (1964), 175–184.

[48] Watson, G. S., and Leadbetter, M. Hazard analysis ii. *Sankhyā: The Indian Journal of Statistics, Series A* (1964), 101–116.

[49] Xiao, J., and Hudgens, M. On nonparametric maximum likelihood estimation with double truncation. *Biometrika 106*, 4 (2019), 989–996.

[50] Zahnit, A. *On robust tail index estimation under incomplete data*. Doctorat these, Université Mohamed Khider Biskra, (2022).

[51] Zerfaoui, K. *Sur l'estimation non paramétrique pour les données doublement tronquées*. Doctorat these, Université Mohamed Khider Biskra, (2023).

[52] Zhou, M. *Empirical likelihood method in survival analysis*, vol. 79. CRC Press, 2015.

ملخص

في هذه الاطروحة قمنا بالتطرق الى مشكلة البيانات الغير مكتملة و بالتحديد الى ظاهرة الاقتطاع المزدوج الذي يعيق العملية الاحصائية حيث وجود الاقتطاع يعني فقدان العينات خلال العملية الاحصائية مما يؤثر سلبا على نتائج الدراسة و بالتحديد ركزنا في هذه الاطروحة على تقدير دالة الخطر في حالة الاقتطاع المزدوج اين تم اعطاء مقدر لدالة خطرا اكثر دقة و هذا ظاهر من خلال الاعتماد على المقارنة مع دوال الخطر المعرفة في هذه الحالة و بالتالي الخروج بنتيجة ان مقدر دالة الخطر المقترح اكثر دقة من خلال المقارنة التطبيقية و النظرية. ومن خلال هذا البحث كذلك تم اقتراح مقدر لدالة الكثافة اكثر نعومة حيث مقدر الكثافة المقترح سابقا كان يعني من انه غير ناعم و كذا غير مستمر وفي اطار هذا السياق تم كذلك اقتراح عدت طرق للحصول على معامل التنعيم الخاص بدالة الكثافة في اطار المعطيات الخاضعة للاقتطاع المزدوج.

Abstract

In this thesis, we investigate the problem of incomplete data, specifically the phenomenon of double truncation, which make working with classical methods very hard, as truncation mean the loss of samples during the statistical analysis, and leads to negative results of the study and wrong decisions. Specifically, we focused in this thesis on estimating of the hazard function in the case of double truncation, where estimating the hazard function estimator is defined in many previous work, hence we make comparison with the hazard functions known in this case and our proposed estimator, and thus we result that the proposed hazard function estimator is more accurate through applied and theoretical comparison. In addition, a smoother cumulative distribution function estimator was proposed, as the previously proposed estimator of the distribution function it was not smooth and not continuous. In this context, several methods were also proposed to obtain the smoothing parameter for the cumulative distribution function within the data subject to double truncation.

Résumé

Dans cette thèse, nous avons étudié le problème des données incomplètes, en particulier le phénomène de double troncature, car la troncature signifie la perte d'échantillons au cours de l'analyse statistique, ce qui affecte

négativement les résultats de l'étude. Nous nous sommes concentrés sur l'estimation de la fonction de risque dans le cas de double troncature, où un estimateur de fonction de risque plus précis a été donné, et cela apparaît en s'appuyant sur la comparaison avec les fonctions de risque définies dans ce cas, arrivant ainsi à la conclusion que nous estimateur de la fonction de risque est plus précis grâce à une comparaison appliquée et théorique. Grâce à cette recherche, un estimateur de fonction de répartition plus lisse a également été proposé, car l'estimateur de la fonction de répartition proposé précédemment signifiait qu'il n'était ni lisse ni continu. Dans ce contexte, plusieurs méthodes ont également été proposées pour obtenir le paramètre de lissage de la fonction de répartition dans le cas au les données soumises à double troncature.