People ’s Democratic Republic of Algeria

Ministry of Higher Education and Scientific

Research

THIRD CYCLE LMD FORMATION

A Thesis submitted in partial execution of the requirements of the degree of

DOCTOR IN MATHEMATICS

Suggested by

Mohamed Khidher University Biskra

Presented by

BEN DAHMANE Khanssa

Titled

Estimating the mean of heavy tailed distribution under random truncation

Supersivor: **Pr.BEANTIA Fatah**

Examination Committee :

Mr. BRAHIMI Brahim	Professor	University of Biskra	President
Mr. BEANTIA Fatah	Professor	University of Biskra	Supervisor
Mr. SAYEH Abdellah	MCA	University of Biskra	Examiner
Mr. AISSAOUI Adel	Professor	University of El-Oued	Examiner

Estimating the mean of a heavy tailed distribution under random truncation

Estimation de la moyenne d'une distribution à queues Lourdes pour les données tronquées

Presented by:
BEN DAHMANE Khanssa

MOHAMED KHIDER UNIVERSITY, BISKRA
FACULTY of EXACT SCIENCES, SIENCE of NATURE and LIFE
DEPARTMENT of MATHEMATICS

A thesis submitted for the fulfillment of the requirements of :
(The Doctorate Degree in Mathematics)

Option : Statistics

March 22, 2022

Supervisor: Prof. Dr. BEN ATIA fatah

Abstract

The main aim of this thesis is to deploy and develop a new estimator for the mean that is based on the famous paper by Peng, 2001. Our case focuses on dealing with data when it becomes incomplete with a particular interest in the case of right-truncated, an asymptotic estimator is proposed and its behavior examined in a simulation study. We treat throughout our study two branches: Survival Analysis and Extreme Value Theory which has emerged as one of the most important statistical disciplines for the applied sciences over the last 50 years.

The first objective of this thesis is to collect and simplify what has been done in the study of extreme values theory. This branch is interested in rare events and the causes of all disasters we know and of all economic crises. In addition, the second objective is to present an introduction that is devoted to the basic notions of survival analysis. Furthermore, we present two cases of incomplete data (censored and truncated) with giving the non-parametric estimator of the mean for each case.

Keywords: Asymptotically normality, Extreme value Theory, Extreme value index, Lynden-Bell estimator, Random variation, Heavy-tails, Random truncation

Dedication

"*To Mom and Dad, sisters: Raouia, Ikram and Lina and my brother Mohammed Charef Eddine. Who always picked me on time and encouraged me to go on every adventure, espicially this one; Thank you...*"

B.khanssa

Acknowledgments

“Words cannot thank them enough, I would just extend profound thanks to : Particularly my supervisor Pr. BENATIA Fatah and Pr. BRAHIMI Brahim for accompanying me throughout the academic course as well as the moral supports. The honorable examination committee members for accepting the evaluation and discussion of this dissertation. I would like to express my special appreciation to my friends Malika and Khaoula that helped me in variety of ways. As well as anyone who helped the cause of this work in any possible way....”

Thank you, B.khanssa

Appendix A: Abbreviations and Notations

The different abbreviations and symbols used throughout this thesis are explained below:

$(\Omega, \mathcal{A}, \mathcal{P})$	:Probability space
rv	:random variable
X	:rv defined on $(\Omega, \mathcal{A}, \mathcal{P})$, population
$(X_1, ..., X_n)$	:samples of size n from X
$\left(X_{1,n}, ...X_{n,n}\right)$	:order statistics pertaining to $(X_1, ..., X_n)$
$X_{i,n}$	:ith order statistics $(i = \overline{1, n})$
$X_{1,n}$	:minimum of $(X_1, ..., X_n)$
$X_{n,n}$	:maximum of $(X_1, ..., X_n)$
$E[X]$	:expectation of (or mean of X)
$Var(X)$	:variance of X
pdf	:probability density function
df	:distribution function
f	:pdf of X
F	:df of X
F_n	:empirical df
$F^{\leftarrow}$	:generalized inverse of F, quantile function

$a.s$	:almost sure
CLT	:Central limit theorem
$Cov(X,Y)$	:covariance between X and Y.
$\mathcal{D}(.)$	:domain of attraction
$e.g$	:for example
i.e.	:in other words
EVI	:extreme value index
EVT	:extreme value theory
$GEVD$	:extreme value distribution
GPD	:generalized Pareto distribution
$\mathbf{1}_A$	:indicator function of set A
iid	:independent identically distributed
$\inf A$	:infimum of set A
ML	:maximum likelihood
MLE	:maximum likelihood estimator
$RMSE$	:root mean squared error
n	:integer number greater than 1
$\mathbb{N}$	:set of non-negatives integers
$\mathbb{R}$	:set of real numbers
$\exists$	:exist
$\mathbb{R}^+$	set of positive real numbers

Q	:quantile function, generalized inverse of F
Q_n	:empirical quantile function
$\sup A$	:supermum of set A
$\mathcal{N}(\mu, \delta^2)$	:normal or Gaussian distribution
$\mathcal{N}(0, 1)$	:standard normal or standard Gaussian distribution
$o(.)$	:$f(x) = o(g(x))$ as $x \longrightarrow x_0 : f(x)/g(x) \longrightarrow 0$ as $x \longrightarrow x_0$
$O(.)$	:$f(x) = O(g(x))$ as $x \longrightarrow x_0 : \exists M > 0, \lvert f(x)/g(x) \rvert \leq M$ as $x \longrightarrow x_0$
$\overset{a.s}{\longrightarrow}$	:a.s converge
$\overset{d}{\longrightarrow}$	:convergence in distribution
$\overset{p}{\longrightarrow}$	:convergence in probability
$\wedge$	:$a \wedge b = \min(a, b)$
$\vee$	:$a \vee b = \max(a, b)$
$[x]$	:integer part of a real number x
S_n	:the partial sum X
$\overline{X}$	:arithmetic mean of X
$\forall$	:$\forall x$ i.e. for any x
$\mathcal{RV}_\alpha$	:regular variation at ∞ with index α
$\mathcal{RV}^0_\alpha$	:regular variation at 0 with index α
$\in$	:belongs
$\log$	:logarithm
$\exp$ or e	:exponential

List of Figures

List of Tables

Contents

1 Introduction

"The essence of mathematics is not to make simple things complicated but to make complicated things simple"

S.Gudder

In statistical analysis the main data or the overage behavior of any phenomenon was the only element of interest. But we could face in many cases extreme situations; For example, Since 12 July 2021 several European countries have been effected by floods, some were catastrophic causing deaths and wide spread damage, besides to the virus that has swept the world and represented as an extreme situation caused millions of deaths around the world Like flooding most of extreme events such as Fires [Algeria(Khenchla, Annaba,..); Turkey..],earthquakes, volcanoes, severe weather conditions (extremely high or low) in order to reduce severe damage and for facing the above extreme situations **E**xtreme **V**alue **T**heory could give a great help.

Figure 1.1: German Floods Kill at Least 133 in 15 july 2021.

the particularity of the extreme value theory is that it focuses on the tail of distribution that generate the studied various extreme phenomenon. it is developed for the estimation of the probability of rare events and make it possible to obtain reliable estimates of the extreme values for which there are few observations. EVT or EVA (Extreme Value Analysis) is mainly based on limit distributions of the extremes and their domain of attractions however there are two models:

- Generalized Extreme Value Distributions.

- Generalized Pareto Distributions.

Thus, it all started with the authors Fisher and Tippett, 1928 ,when they were studying resistance. They stated a fundamental theorem with the creation of three domains of attractions: Fréchet, Gumbell and Weibull. This interesting theorem refers to a parameter called the tail index which gives the shape of the distribution tail. Indeed, if the tail index is positive we are in the presence of Fréchet's domain of attraction; then if it's negative, domain of Weibull attraction on the other hand if the index is zero then Gumbel domain. von Mises, 1936, Jenkinson, n.d., gathered the distributions of these three domains in one writing. It is at this time that several authors have focused on estimating of the extreme value index. We can cite Hill, 1975, in the case where the index is positive. Pickands III, 1975 in the same year proposed an estimator of the index of extreme values in the general case. On the other hand, Dekkers et al., 1989. have generalized Hill's estimator, referred to as the Moments estimator, Beirlant et al., 2016. presented in turn, the estimator of the extreme value index generated at from the Hill estimator and the quantile function.

the second peculiarity of this analysis, and that it is very common to be found in faced with the problem of missing data, i.e. survival data are not fully observed, they are incomplete. Censorship and truncation are both the most common causes of this problem. Censorship is a mechanism that prevents the exact observation of the time of occurrence of interest. We know well that this deadline belongs to a certain time interval. Truncation of an object can be detected only if its value is greater or less than some number, and the value is completely known. In this case, the classic techniques do not adapt correctly to incomplete data. The literature is much richer in censorship than the truncation, which is more recent. For full details on censorship and survival analysis, the reader might check the books: Cox and Oakes, 1984, Kalbfleisch and Prentice, 2011 ,Lee and Wang, 2003 ,Klein and Moeschberger, 2005 , Wienke, 2010

Figure 1.2: Forest fires in Khenchela in 08 july 2021.

.In 1951, Weibull designed a parametric model in the field of reliability; at this effect, it provides a new probability distribution which will subsequently be frequently used in survival analysis: Weibull's law.In 1958, Kaplan and Meier, 1958 presented important results concerning the estimation non-parametric of the survival function, of the resulting estimator, they study expectation, variance and asymptotic properties. Asymptotic behavior of the Kaplan and Meier estimator Kaplan and Meier, 1958 are used the interest of a large number of authors including Breslow and Crowley, 1974 who are the first to deal with convergence and the asymptotic normality of the Kaplan and Meier, 1958 estimator.In this thesis we deal essentially with the case of right-truncation For full details on truncation the reader back to the books: Woodroofe, 1985, Benchaira et al., 2015,Gardes and Stupfler, 2015,and Lynden-Bell, 1971 etc...

Our thesis is organized into two parts We start as preliminary chapters: 2,3 and 4 then the second part present our main results.the content of each chapter is presented as follows:

Figure 1.3: Coronavirus disease (COVID-19).

Chapter 2

This chapter contains some mathematical preliminaries (the asymptotic properties of the sum of iid rv's; order statistics and distributions of upper order statistics), also contains a derivation of the three families of classical Gnedenko limit distributions for extremes of iid variables and an account of regular variation and its extensions and domains of attraction. So, this chapter gives you an introduction to the mathematical and statistical theory underlying EVT.

Chapter 3

This chapter is devoted to the basics of survival analysis, we begins with a few reminders on basic concepts such as fdr, the three survival functions and the equivalence relationship between these three functions is discussed.

The more we talk about the laws of large numbers and properties asymptotic of the sum of the iid values (TCL). with giving the different estimators for the mean in each case of finite and infinite second moment.

Chapter 4

Contains the essential definitions and results of incomplete data, with the main basic concepts on truncated data and some important and useful results existing in the literature for the random right truncation model. In this chapter, we start by censored data, which can be further classified into three categories: right censoring, left censoring, and interval censoring.it worth to mention that we present the different work in estimating the mean in this case. Afterward, we will be interested in the truncated data. Which in turn has three types as follows: right truncation, left truncation, and interval truncation, but in the present thesis, we are concerned with data that are right truncated.

Chapter 5

The chapter deals with the estimation of the mean under random truncation. The main objective of this chapter is to propose a method for estimating the mean of this type of distribution in the presence of random right truncation, its asymptotic normality established and its performance evaluated on simulated data.

Part I

Preliminary Theory

2 Extreme Value theory

*Much statistical analysis study the main body of data, and look at its behavior in terms of means in many cases however, the extreme value in the data is more interested in example; if we were studying river level over time then the only values we really care about are those that are really high or really low; if they are too low then the river could dry out. For this spot, an extreme value of some random variable is often either their maximum or their minimum; we could analogously come up with some results for their minimum as well will just assume it's their maximum, In this chapter, we have made a general overview on the theory of extreme values, mentioning the different characteristics and the basics which are very useful for the estimation of the extreme quantiles and the truncated data that we will discuss in the next chapter. A very good variety of textbooks and books is devoted to Extreme Values Theory (**EVT**) for example: Resnick, 1987 , De Haan et al., 2006, Embrechts et al., 1997b and Leadbetter et al., 2012 and we don't forget to mention Meraghni, 2008 etc.*

2.1 Order statistics

Order statistics or (OS) play an increasingly important role in the theory of extreme values because they provide information about the tail distribution (right). for a long time, we face OS when we study survival analysis (truncation or censored) but recently the order statistics appeared in the research for robust methods. We start in this section, by giving the definitions and some properties of the statistics order, then we study their exact and asymptotic distributions. For more detailed presentations in this area, we can cite, for example, the books of Reiss, 1987 and Coles et al., 2001, Arnold et al., 2008, and David and Nagaraja, 2004. Which covered virtually all the topics of order statistics.

▶ **Definition 2.1 (Order statistics).** Let $(X_1, ..., X_n)$ n iid random variable with a common distribution F and density f. We call statistics of order (increasing) the sequence of random variables $(X_1, ..., X_n)$ which are ordered by ascending order, either:

$$X_{1,n} \leq ... \leq X_{n,n}.$$

◀

▶ **Remark 2.2.** For $1 \leq k \leq n$ the variable $X_{k,n}$ is known under the name of the kth order statistic or k order statistic. Two order statistics are particular-interesting for the study of extreme events. These are the order statistics extremes which are given by the following definition. ◀

Extreme order statistics are defined as terms of the maximum and minimum of n random variables $(X_1, ..., X_n)$:

- the variable $X_{1,n}$ is the smallest statistic of order (or statistic of the minimum) and $X_{1,n} := \min(X_1, ..., X_n)$.
- the variable $X_{n,n}$ is the greatest statistic of order (or maximum statistic) $X_{n,n} := \max(X_1, ..., X_n)$.

▶ **Remark 2.3.** We can find in other books the next notation for extreme order statistics (minimum and maximun) as follow:

$$M_n := \max(X_1, ..., X_n) \quad \text{and} \quad m_n := \min(X_1, ..., X_n).$$

◀

2.1.1 Empirical distribution function and order statistics

- $F_n(x)$ is the proportion of the n variables which are less than or equal to x .
- $\overline{F}_n(x)$ is the proportion of the n variable which are greater or equal to x.
- the empirical df (or sample df) of the sample $(X_1, ..., X_n)$ is defined by:

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x) \quad , x \in \mathbb{R}. \tag{2.1}$$

- The functions $F_n(x)$ and $\overline{F}_n(x)$ are written by using order statistics as follows:

$$F_n(x) := \begin{cases} 0 & \text{if } x \leq X_{1,n} \\ \frac{i-1}{n} & \text{if } X_{i-1,n} \leq x \leq X_{i,n} \text{ for } i = \overline{2, n} \\ 1 & \text{if } x \geq X_{n,n} \end{cases}$$

$$\overline{F}_n(x) := \begin{cases} 1 & \text{if } x \leq X_{1,n} \\ 1 - \frac{i-1}{n} & \text{if } X_{i-1,n} \leq x \leq X_{i,n} \text{ for } i = \overline{2, n} \\ 0 & \text{if } x \geq X_{n,n} \end{cases}$$

2.1.2 Distribution function and density of the maximum

▶ **Proposition 2.4 (Balakrishnan & Nagaraja).** – The distribution function (df) $F_{X_{n,n}}$of $X_{n,n}$ is given by:

$$\forall x \in \mathbb{R}, \qquad F_{X_{n,n}}(x) = P(X_{n,n} \leq x) = F^n(x). \tag{2.2}$$

– If X is absolutely continuous of density f, then the density function $f_{X_{n,n}}$of $X_{n,n}$ is given by:

$$\forall x \in \mathbb{R}, \quad f_{X_{n,n}}(x) = nF^{n-1}(x)f(x). \tag{2.3}$$

◀

2.1.3 Upper end point

We denote by x_F (resp x_F^*) the upper extreme point (resp. Lower)of the distribution F (i.e. the greatest possible value for $X_{k,n}$ which can take the value $+\infty$(resp $-\infty$) in the sense that:

$$x_F := \sup\{x : F(x) < 1\} \leq \infty$$

and :

$$x_F^* := \inf\{x : F(x) > 0\}.$$

2.1.4 Quantile function

▶ **Definition 2.5 (Quantile function).** The quantile function of df F is generalized inverse function of F defined by: For all $0 < s < 1$

$$Q(s) := F^{\leftarrow}(s) := \inf\{x : F(x) \geq s\}. \tag{2.4}$$

with the convention that $\inf(\varnothing) = \infty$. ◀

▶ **Remark 2.6.** For all $0 < s < 1$ the distribution function F is strict-increasing and continues. ◀

2.1.5 Empirical quantile function

▶ **Definition 2.7 (Empirical quantile function).** The empirical quantile function of the sample $(X_1, ..., X_n)$ is defined by For all $0 < s < 1$:

$$Q_n(s) := \inf\{x : F_n(x) \geq s\}$$

Q_n can be expressed as a simple function of order statistics concerning the sample $(X_1, ..., X_n)$ So we have:

$$Q_n(s) = X_{i,n} \quad \text{for} \quad \frac{i-1}{n} < s \leq \frac{i}{n}, \quad i = \overline{1, n}$$

note that for $0 < p < 1$; $X_{[np]+1,n}$ is the sample quantile of order p, where $[np]$ denote the integer part of np, if $s = 1/2$ then one also speaks of the sample median. ◀

2.1.6 Tail quantile and emperical tail quantile function

▶ **Definition 2.8 (Tail quantile function).** denoted by U and called tail quantile function is used quite often; it is defined by:

$$\mathrm{U}(t) := Q(1 - 1/t) = (1/\overline{F})^{\leftarrow}(t), \quad 1 < t < \infty$$

the corresponding empirical function is:

$$\mathrm{U}_n(t) := Q_n(1 - 1/t), \quad 1 < t < \infty$$

◀

▶ **Proposition 2.9 (Quantile transformation).** Let $(\mathrm{U}_1, ..., \mathrm{U}_n)$ be a sample from the standard uniform rv U and $(\mathrm{U}_{1,n}, ..., \mathrm{U}_{n,n})$ be the corresponding ordered sample. ◀

(i) For any df F, we have:

$$X_{i,n} = F^{\leftarrow}(U_{i,n}), \ i = 1, .., n \tag{2.5}$$

(ii) When F is continuous, we have:

$$F(X_{i,n}) = U_{i,n}, \quad i = 1, .., n. \tag{2.6}$$

In this case the rv's $F(X_1), ..., F(X_n)$ are iid standard uniform.

2.1.7 Distributions of order statistics

▶ **Proposition 2.10 (Maximum and minimum distributions).** Let $(X_1, ..., X_n)$ be n rv independent identically distributed of distribution function F, the exact distribution of the maximum $X_{n,n}$ is simply given by the following formula:

$$\forall x \in \mathbb{R}, \quad F_{X_{n,n}}(x) = [F(x)]^n.$$

The exact distribution of the minimum is given by:

$$\forall x \in \mathbb{R}, \quad F_{X_{1,n}}(x) = 1 - [1 - F(x)]^n.$$

◀

▶ **Proposition 2.11 (Distribution function of the k th upper order statistic).** These are important special cases of the general result of $F_{k,n}$ denote the df of $X_{k,n}$ where $k = 1, ..., n$ which it is given by:

- $F_{k,n}(x) := \sum_{r=0}^{k-1} \begin{pmatrix} n \\ r \end{pmatrix} \overline{F}^r(x) F^{n-r}(x).$

– if F is continuous; then

$$F_{k,n}(x) := \int_{-\infty}^{x} f_{k,n}(z) dF(z),$$

Where

$$f_{k,n}(x) := \frac{n!}{(k-1)!(n-k)!} F^{n-k}(x) \overline{F}^{k-1}(x);$$

i.e. $f_{k,n}$ is a density of $F_{k,n}$ with respect to F.

Proof. see e.g Embrechts et al., 1997b page 183. ◀

◀

2.2 Distribution of extreme values

Now suppose there is a sequence $(a_n)_{n\in\mathbb{N}^*}$ srictly real numbers positive and a sequence $(b_n)_{n\in N}$ of real numbers such as the sequence of normal maxima $\left\{\frac{1}{a_n}(X_{n,n} - b_n),\ n \in \mathbb{N}^*\right\}$ converges in distribution to a random variable non degenerate of distribution function $\mathcal{H}$, i.e.

$$\forall x \in \mathbb{R}, \quad \lim_{n\to\infty} P\left(\frac{X_{n,n} - a_n}{b_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}(x). \tag{2.7}$$

▶ **Definition 2.12.** The sequences $\{a_n > 0, n \geq 1\}$ and $\{b_n, n \geq 1\}$are called sequences of normalization, the constants $a_n \in \mathbb{R}_+^*$ and $b_n \in \mathbb{R}$ are called constants of normalization and the random variable $\frac{1}{a_n}(X_{n,n} - b_n)$ is called the normalized maximum. ◀

2.3 Limit distributions

In the central limit theorem, we have defined only possible limit laws for the sequence of sums, normalized by independent and identically distributed random variables when n tend to ∞, we have a similar notion in extreme value theory called max-stable law.

▶ **Definition 2.13 (Embrechts & Mikosch).** The random variable, non-degenerate, X or the probability law of X or, again the distribution function F of X is said to be max-stable, if there are constant $a_n \in \mathbb{R}_+^*$ and $b_n \in \mathbb{R}$ such that for all $n \in \mathbb{N}^*$.

$$M_n \overset{d}{=} a_n X + b_n.$$

Or, which is equivalent, if there are constants for all $n \in \mathbb{N}^*$ and $x \in \mathbb{R}$

$$F^n(a_n x + b_n) = F(x).$$

◀

▶ **Theorem 2.14 (Fisher & Tippett).** Let $(X_i)_{i\geq 1}$ be a sequence of n independently distributed random variables with distribution function F. If there are two real normalizing sequences $(a_n)_{n\geq 1} > 0$ and $(b_n)_{n\geq 1} \in \mathbb{R}$and a non-degenerate law of distribution $\mathcal{H}$ such that

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = \mathcal{H}(x). \qquad (2.8)$$

$\mathcal{H}$ is the distribution of extreme values. So except for a translation and a change of scale, the distribution function of the limit is of the type of the following three classes:

$$\begin{aligned} &Fr\acute{e}chet: && \Phi_\alpha(x) = \exp(-x^{-\alpha})\mathbf{1}_{x>0}.\\ &Gumbel: && \Lambda(x) = \exp(-e^{-x}), \qquad x \in \mathbb{R}\\ &Weibull: && \Psi_\alpha(x) = \exp(-(-x)^{\alpha})\mathbf{1}_{x<0}.\end{aligned}$$

◀

Proof. For a proof of this theorem, the reader can refer to the following work: Resnick, 1987. ■

▶ **Remark 2.15.** To distinguish the three distributions, we generally use the notation:Λ for the Gumbel distribution ,Φ_α for the Fréchet distribution and $\Psi_\alpha$for the Weibull distribution ◀

We can choose the normalization constants according to the following theorem.

▶ **Theorem 2.16.** We have $(a_n)_{n\geq 1} > 0$ and $(b_n)_{n\geq 1} \in \mathbb{R}$ such that:

$$F_{X_{n,n}}(a_n x + b_n) \overset{n\to\infty}{\to} \mathcal{H}(x),$$

$$\begin{aligned} &b_n = 0, && a_n = F^{-1}(1-\tfrac{1}{n}) && \text{if } \mathcal{H} = \Phi\\ &b_n = F^{-1}(1), && a_n = F^{-1}(1) - F^{-1}(1-\tfrac{1}{n}) && \text{if } \mathcal{H} = \Psi\\ &b_n = F^{-1}(1-\tfrac{1}{n}), && a_n = F^{-1}(1) - F^{-1}(1-\tfrac{1}{n}) && \text{if } \mathcal{H} = \Lambda.\end{aligned}$$

◀

▶ **Proposition 2.17 (Density function of extreme values).** The density functions of the distribution of standard extreme values and the different types of extreme distribution, are as follows:

$$\begin{aligned} &Fr\acute{e}chet: && \phi_\alpha(x) = \alpha x^{-\alpha-1}\exp(-x^{-\alpha})\mathbf{1}_{x>0}.\\ &Gumbel: && \lambda(x) = \exp(-\{x + e^{-x}\}), \qquad x \in \mathbb{R}\\ &Weibull: && \psi_\alpha(x) = \alpha(-x)^{-\alpha-1}\exp(-(-x)^{\alpha})\mathbf{1}_{x<0}.\end{aligned}$$

◀

2.4 Generalized extreme values distributions (GEVD)

As we have just seen, the three types of extreme distributions Fréchet, Weibull and Gumbel have different behaviors that correspond to different behaviors of the function of tail F of the random variable X . This resulted in the first applications of extreme value theory, to adopt one of these three types for data analysis. But this method has drawbacks because,

- first, we must have a technique to choose which of the three distribution extremes are more appropriate to the data that we have.
- secondly, once such a decision is made, subsequent deductions confirm that our choice is correct and does not take into account the uncertainty that such a selection implies,

although this uncertainty attitude can be substantial. A better analysis is offered thanks to the work of von Mises, 1936 and Jenkinson, n.d. who showed that the three extreme types of distribution Fréchet, Weibull and Gumbel can be combined into a single type of distribution called a "type of generalized extreme values distribution (GEVD) ”or“ type of distribution of extreme values of Von Mises-Jenkinson ”

▶ **Definition 2.18 (Embrechts & Mikosch).** Let $\gamma \in \mathbb{R}$, We call the distribution of standard generalized extreme values any distribution function $\mathcal{H}_\gamma$ or any probability law which has $\mathcal{H}_\gamma$ as a function of distribution such that for $\gamma \in \mathbb{R}$ and $1 + \gamma x > 0$ where the parameter γ is called the index of extreme values. such that:

$$\mathcal{H}_\gamma(x) := \begin{cases} \exp\left\{-(1+\gamma x)^{-\frac{1}{\gamma}}\right\}, & \text{if } \gamma \neq 0. \\ \exp\{-\exp(-x)\}, & \text{if } \gamma = 0. \end{cases} \tag{2.9}$$

◀

▶ **Proposition 2.19 (Ferreira (2006)).** Let $\mathcal{H}_\gamma(\gamma \in \mathbb{R})$be the generalized extreme value distribution and Λ, Φ_α and $\Psi_\alpha$the distribution of standard extreme values with $\alpha > 0$ we have :

$$\mathcal{H}_\gamma(x) := \begin{cases} \Phi_{\frac{1}{\gamma}}(1+\gamma x) & , \text{ if } \gamma > 0 \\ \Psi_{-\frac{1}{\gamma}}\{-(1+\gamma x)\} & , \text{ if } \gamma < 0 \\ \Lambda(x) & , \text{ if } \gamma = 0 \end{cases} ; \tag{2.10}$$

$\forall x \in \mathbb{R}$ such that $1 + \gamma x > 0$. ◀

This proposition gives us a very important result, in the applications of the theory of extreme values, which make it possible to classify the three types of extreme distributions Fréchet, Weibull and Gumbel in a single type which is the type of generalized extreme value distribution. Indeed, we have the following proposition:

▶ **Proposition 2.20.** Let $\mathcal{H}_\gamma(\gamma \in \mathbb{R})$be the generalized extreme value distribution and Λ, Φ_α and $\Psi_\alpha$the distribution with $\alpha > 0$, the types of extreme value distribution which are,respectively, of Fréchet , of Weibull and of Gumbel , then we have:

$$\mathcal{H}_\gamma := \begin{cases} \Phi_{\frac{1}{\gamma}} & , \text{ if } \gamma > 0 \\ \Psi_{-\frac{1}{\gamma}} & , \text{ if } \gamma < 0 \\ \Lambda & , \text{ if } \gamma = 0 \end{cases} . \tag{2.11}$$

◀

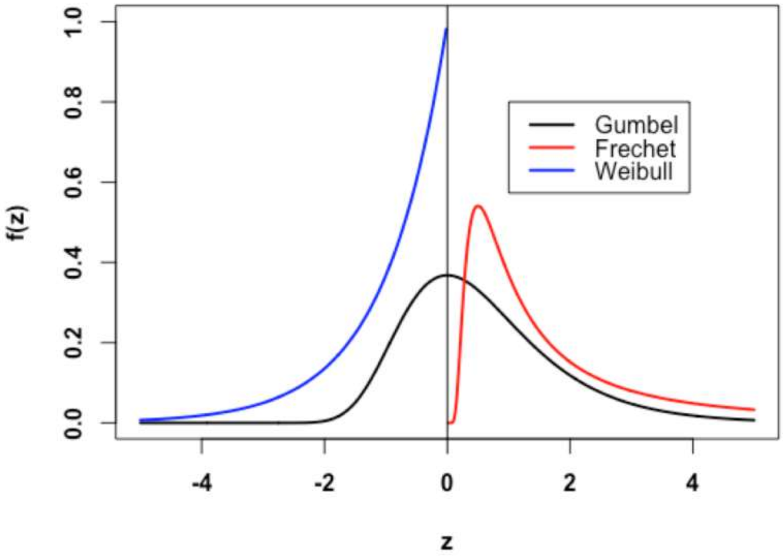

Figure 2.1: Densities of the standard extreme value distributions. we chose $\alpha = 1$ for the Fréchet and the Weibull distribution

▶ **Remark 2.21.** $H_{\gamma,\mu,\sigma}(x)$ is a general form for non-centered and unreduced variables, so for $\left(1+\gamma(\frac{x-\mu}{\sigma}) > 0\right)$ the distribution $H_{\gamma,\mu,\sigma}(x)$ is written as follows:

$$H_{\gamma,\mu,\sigma}(x) := \begin{cases} \exp\left\{-(1+\gamma(\frac{x-\mu}{\sigma}))^{-\frac{1}{\gamma}}\right\}, & \text{if } \gamma \neq 0. \\ \exp\left\{-\exp(-(\frac{x-\mu}{\sigma}))\right\}, & \text{if } \gamma = 0. \end{cases} \tag{2.12}$$

- in which the shown paramters are : for localization $\mu \in \mathbb{R}$ and a scale parameter $\sigma > 0$. ◀

2.5 Generalized Pareto Distribution (GPD)

As mentioned in the previous section, the distribution of generalized extremes is very useful in application of extreme value theory, because it is the one and only law of probability which models the behavior of the maximum of a sample. To estimate its parameters, statisticians often use methods which are:

- **The block maxima (BM) method:**

statistical inference about rare events is linked to observations which are extreme in some sense. Different ways to define such observations lead to different approaches to statistics of univariate extreme (SUE). Sometimes, only yearly or block maxima which consists in constructing a sample of maximums from a sample data in block format of the same size. This method has a major draw back which leads to a loss of certain information, in particular, some blocks may contain more than one extreme value, while other blocks may not contain any.

- **The peaks over threshold (POT) approach:**

The POT approach is also popular to SUE, in a certain sense parallel to the MEV model, and was introduced Smith, 1987. Our attention is restricted to the observations that exceed a certain high threshold u, this method allows to take into account much more data to ensure much more precision in the estimation of the parameters of the distribution of extreme values generalized, in particular, the index of extreme values $\gamma \in \mathbb{R}$. Indeed based on Generalized Pareto Distribution (GPD) Balkema de haan-Pickands theorem, consists in studying the behavior not of the maximum data that we have but all data that greater a high threshold u , and more precisely, the differences between these data and the threshold u , called "excess"i.e

$$P(X - u \leq y \mid X > u)$$

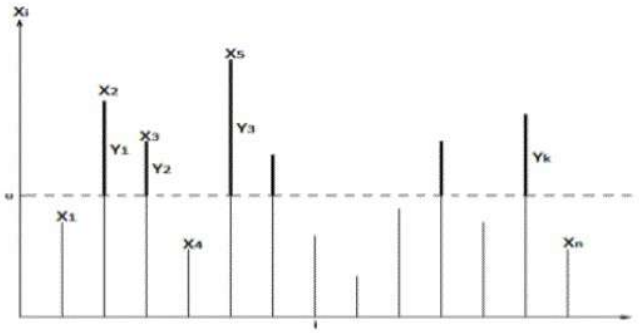

Figure 2.2: The data $X_1, X_2, ..., X_n$ and their corresponding k excess beyond the $u$threshold $Y_1, Y_2, ..., Y_n$ $(k \leq n)$.

▶ **Definition 2.22 (excess).** We call excess of the random variable X beyond a threshold $u < x_F$ the random variable Y, which takes its values on $]0, x_F - u[$, defined by :

$$Y = X - u \mid X > u, \qquad u < x_F.$$

We call the distribution of the excesses of the random variable X with respect to a threshold $u < x_F$ the probability law of the random variable Y excess of X beyond the threshold $u < x_F$, given by its distribution function F_u , which we call the distribution function excess, following:

$$F_u(y) = P(X - u \leq y \mid X > u) = \begin{cases} 0 & , \text{if } y \leq 0, \\ 1 - \frac{1 - F(u+y)}{1 - F(u)} & , \text{if } 0 < y < x_F - u, \\ 1 & , \text{if } y \geq x_F - u. \end{cases}$$

the mean function of the excesses of the random variable X with respect to the threshold $u < x_F$, and we denote it by $e(u)$, the expectation function of the random variable Y excess of X beyond the threshold $u < x_F$, defined by:

$$\forall u < x_F, \qquad e(u) = E(X - u \mid X > u) = \frac{1}{\overline{F}(u)} \int_u^{x_F} \overline{F}(t)dt.$$

◀

▶ **Definition 2.23 (Generalized Pareto Distribution).** Let $\gamma \in \mathbb{R}$. standard generalized Pareto distribution any function distribution G_γ or any probability law which has $\mathcal{G}_\gamma$ as a distribution function such as : $\forall x > 0$ and $1 + \gamma x > 0$;

$$\mathcal{G}_\gamma(x) := \begin{cases} 1 - (1+\gamma x)^{-\frac{1}{\gamma}} & , \text{if } \gamma \neq 0, \\ 1 - e^{-x} & , \text{if } \gamma = 0. \end{cases} \tag{2.13}$$

◀

▶ **Remark 2.24.** – We can give a more general form to the distribution function $\mathcal{G}_\gamma$ given in the definition above, that we denote by $\mathcal{G}_{\gamma,\mu,\sigma}(x)$ a parameter is shown of localization $\mu \in \mathbb{R}$ and a scale parameter $\sigma > 0$ for $\big(1 + \gamma(\frac{x-\mu}{\sigma}) > 0\big)$ and $\forall x > \mu$ as follows:

$$\mathcal{G}_{\gamma,\mu,\sigma}(x) := \begin{cases} 1 - \left[1 + \gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}} & , \text{if } \gamma \neq 0, \\ 1 - \exp\left(\frac{x-\mu}{\sigma}\right) & , \text{if } \gamma = 0. \end{cases} \tag{2.14}$$

– The parameter $\gamma \in \mathbb{R}$ that we can therefore see the generalized Pareto distribution is a shape parameter called "tail index".

◀

2.6 Regularly Varying distributions

In this section, we treat the class C of functions that appear in a large number of applications in the whole of mathematics. Here, we will define some generalities on these functions with some of their most important properties. Those who are interested in the theory of regular variation can consult for example: Teugels et al., 1987, Klüppelberg and Mikosch, 1997.

▶ **Definition 2.25 (Regularly varying and slowly varying functions).** A measurable function $V : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is regularly varying at ∞ with the index ρ, and we denote by $V \in \mathcal{RV}_\rho$, if:

$$\lim_{x\to\infty} \frac{V(tx)}{V(x)} = x^\rho \quad ,t > 0. \tag{2.15}$$

ρ is the variation exponent or the regular variation index.

A measurable function $l :]a, +\infty[\longrightarrow \mathbb{R}^+$ with $(t > 0)$ is said slowly varying at infinity, if:

$$\lim_{x\to\infty} \frac{l(tx)}{l(x)} = 1 \quad ,t > 0.$$

A function with regular variation of index $\rho \in \mathbb{R}$ can always be written under the following form: ◀

▶ **Example 2.26.** The following table content some examples for slowly varying functions and not:

regularly varying	not regularly varying
x^ρ	$\exp(x)$
$x^\rho \log(1+x)$	$\sin(x+2)$
$(x\log(1+x))^\rho$	$\exp(\log(1+x))$

◀

we give some elementary properties, functions with slow variations:

▶ **Proposition 2.27 (Slowly varying function properties).** – $\mathcal{RV}_0$ is closed under addition, multiplication and division.

- If l is slowly varying,

$$\lim_{x\to\infty}\frac{\log(l(x))}{\log(x)} = 0.$$

- If l is slowly varying, then the l^α is slowly varying for any $\alpha \in \mathbb{R}$.
- If l is slowly varying and $\rho > 0$.

$$\lim_{x\to\infty} x^\rho l(x) = \infty \text{ and } \lim_{x\to\infty} x^{-\rho} l(x) = 0.$$

◀

▶ **Lemma 2.28 (Inverse of regular variation function).** – if f is a regular variation at infinity with index $\alpha > 0$, then $f^{\leftarrow}$ is regular variation at infinity with index $\frac{1}{\alpha} > 0$.

- if f is regular variation at infinity with index $\alpha < 0$, then $f^{\leftarrow}$ is regular variation at infinity with index $\frac{1}{\alpha} < 0$.

◀

Proof. the proof of this lemma could be found in Bingham et al., 1989 . ■

▶ **Theorem 2.29 (Kramata representation).** – every slowly varying function l(i.e $l \in \mathcal{RV}_0$) if and only if can be represented as:

$$l(x) = c(x)\exp\left\{\int_1^x \frac{r(x)}{t}dt\right\}, \quad x > 0.$$

◀

where $c(.)$, $r(.)$ are two measurable functions, and

$$\lim_{x\longrightarrow\infty} c(x) = c_0 \in [0, \infty[\text{ and } \lim_{t\longrightarrow\infty} r(t) = 0.$$

if the function $c(.)$ is a constant, then we said l is normalised.l

- A function $V : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is regularly varying at ∞ with the index ρ if and only if V has the representation:

$$V(x) = c(x) \exp\left\{\int_1^x \frac{\rho(t)}{t} dt\right\}, \quad x > 0.$$

$\lim_{t\longrightarrow\infty} \rho(t) = \rho$

Proof. See Resnick, 1987, Corollary 2.1; page 29 ■

2.6.1 First Order Regular variation Assumption

For a df function F and **U** the tail quantile function, the following assumptions are equivalent:

- $\overline{F}$ is regularly varying at infinity with index $-1/\gamma$

$$\lim_{z\longrightarrow\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma}, x > 0.$$

- $Q(1 - s)$ is regularly varying at infinity with index $-\gamma$

$$\lim_{s\longrightarrow\infty} \frac{Q(1 - sx)}{Q(1 - s)} = x^{-\gamma}, x > 0.$$

- U is regularly varying at infinity with index γ

$$\lim_{z\longrightarrow\infty} \frac{U(xz)}{U(z)} = x^{\gamma}, x > 0.$$

- F is heavy-tailed.

2.6.2 Second Order Regular variation Assumption

We say that F is second order regularly varying at infinity if it satisfies one of the following conditions:

- there exist some parameter $\rho \leq 0$ and a function A^*, such that for all $x > 0$

$$\lim_{t\to\infty} \frac{\overline{F}(tx)/\overline{F}(t) - x^{-1/\gamma}}{A^*(t)} = x^{-1/\gamma}\frac{x^\rho - 1}{\rho}.$$

- there exist some parameter $\rho \leq 0$ and a function A^{**}, such that for all $x > 0$

$$\lim_{s\to\infty} \frac{Q(1-sx)/Q(1-s) - x^{-1/\gamma}}{A^{**}(s)} = x^{-\gamma}\frac{x^\rho - 1}{\rho}.$$

- there exist some parameter $\rho \leq 0$ and a function A, such that for all $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{-1/\gamma}\frac{x^\rho - 1}{\rho}.$$

where A^*, A^{**} and A are regularly varying functions with

$$A^*(t) = A(1/\overline{F}(t)) \text{ and } A^{**}(t) = A(1/t).$$

Their role is to control the speed of convergence in First order regular variation condition.

If $\rho = 0$; interpret $(x^\rho - 1)/\rho$ as $\log x$.

2.6.3 Third Order Regular variation Assumption

There exist a positive real parameter γ, negative real paramters ρ and β; functions b and $\widetilde{b}$ with $b(t) \longrightarrow 0$ and $\widetilde{b}(t) \longrightarrow 0$

for $t \longrightarrow \infty$, both of constant sign for large value of t; such that:

$$\lim_{t\longrightarrow\infty} \frac{\frac{\ln U(tx)-\ln U(t)-\gamma\ln x}{b(t)} - \frac{x^\rho-1}{\rho}}{\widetilde{b}(t)} = \frac{1}{\beta}\left(\frac{x^{\rho+\beta}}{\rho+\beta} - \frac{x^\rho - 1}{\rho}\right), \text{ for } x > 0$$

Where$\left|\widetilde{b}\right|$ is regularly varying of index β.

2.7 Domain of attraction

In this section, we shall determine a sufficient and necessary conditions on the distribution function F that ensure the membership of this distribution to a domain of attraction. Basically, these conditions due to von Mises, 1936, require the existence of one or two derivative of F. Next we present a definition of domain of attraction and the following theorem states a sufficient conditions for belonging to domain of attraction the conditions called von Mises condition.

▶ **Definition 2.30 (Domain of attraction).** We say that a distribution F belongs to the domain of attraction of the maximum of the distribution $\mathcal{H}_\gamma$, and we denote by $F \in \mathcal{D}(\mathcal{H}_\gamma)$, if there are two normalizing sequences $(a_n > 0)$and $(b_n \in \mathbb{R})$ such that the condition holds:

$$\lim_{n\longrightarrow\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\longrightarrow\infty} F^n(a_n x + b_n) = \mathcal{H}_\gamma(x), \quad \forall x \in \mathbb{R}. \tag{2.16}$$

◀

▶ **Theorem 2.31.** Let F be a distribution function and x_F its right endpoint. Suppose $\acute{F}(x)$ exists and $\acute{F}(x)$ is positive for all x in some left neighborhood of x_F. If

$$\lim_{t\to x_F}\left(\frac{1-F}{\acute{F}}\right)^{'}(t)=\gamma \tag{2.17}$$

or equivalently

$$\lim_{t\to x_F}\frac{(1-F(t))\acute{F}(t)}{\left(\acute{F}(t)\right)^2}(t)=-\gamma-1$$

then F is in the domain of attraction of $\mathcal{H}_\gamma$. ◀

▶ **Theorem 2.32.** ◀

1. $(\gamma>0)$ suppose $x_F=\infty$ and $\acute{F}$ exist. If

$$\lim_{t\to\infty}\frac{t\acute{F}(t)}{1-F(t)}=\frac{1}{\gamma},$$

for some positive γ, then F is in the domain of attracttion of $\mathcal{H}_\gamma$.

2. $(\gamma<0)$ suppose $x_F<\infty$ and $\acute{F}$ exist for $x<x_F$. If

$$\lim_{t\to x_F}\frac{(x_F-t)\acute{F}(t)}{1-F(t)}=-\frac{1}{\gamma},$$

for some negative γ, then F is in the domain of attracttion of $\mathcal{H}_\gamma$.

For the proofs and more details on this issue, one my consult De Haan et al., 2006, please check page 15.

▶ **Theorem 2.33.** the distribution function F is in the domain of attraction of the extreme value distribution $\mathcal{D}(\mathcal{H}_\gamma)$ if and only if

1. for $\gamma > 0 : x_F$ is infinite and
$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\frac{1}{\gamma}}. \tag{2.18}$$
for all $x > 0$. this means that the function $1 - F$ is regularly varying at infinity with index $-\frac{1}{\gamma}$.
2. for $\gamma < 0 : x_F < \infty$ and for all $x > 0$
$$\lim_{t\downarrow 0} \frac{1-F(x_F - tx)}{1-F(x_F - t)} = x^{-\frac{1}{\gamma}}. \tag{2.19}$$
3. for $\gamma = 0$: here the right endpoint x_F may be finite or infinite and
$$\lim_{t\uparrow x_F} \frac{(1-F(t+xf(t)))}{1-F(t)} = e^{-x}. \tag{2.20}$$
for all real x, where f is a suitable positive function.if equation (2.20) holds for some f then $\int_t^{x_F}(1-F(s))ds < \infty$ for $t < x_F$ and equation (2.20) holds with:
$$f(t) = \frac{\int_t^{x_F}(1-F(s))ds}{1-F(t)}.$$

◀

Characterization of domain of attraction

Different characterizations of three domain of attraction of Fréchet, Weibull and Gumbel have been proposed in Resnick, 1987, De Haan et al., 2006 and Embrechts et al., 1997b. So According to the sign of γ, we can distinguish three domain of attraction:

- If $\gamma > 0$, we say that $F \in \mathcal{D}(\Phi_\gamma)$, and F has an infinite right end point ($x_F = +\infty$), this domain of attraction of heavy-tailed distributions,that is, which have a polynomial decay survival function.

- If $\gamma < 0$, we say that $F \in \mathcal{D}(\Psi_\gamma)$, and F has a finite right endpoint ($x_F < +\infty$). This domain of attraction of survival functions whose support is bounded above.
- If $\gamma = 0$ we say that $F \in \mathcal{D}(\Lambda)$ the end point x_F can then be finite or not.This domain of attraction of distributions with light tails, that is to say which have an exponentially decaying survival function.

We indicate here the most used criteria, that is to say, the conditions on the fdr for which belongs to one of the three domains of attraction which are defined previously.

▶ **Theorem 2.34 (Characterization of $\mathcal{D}(\Phi_\alpha)$).** The fdr belongs to the domain of attraction of Fréchet's law with parameter $\alpha > 0$ if and only if :

$$\overline{F}(x) = x^{-\alpha} l(x). \tag{2.21}$$

where the function l is slowly varying. In particular $x_F = +\infty$ moreover if $F \in \mathcal{D}(\Phi_\alpha)$,with $a_n = F^{-1}(1 - \frac{1}{n})$ and $b_n = 0$, the sequence $\left(a_n^{-1} X_{n,n}\right)_{n\geq 1}$converges in law to goes from fdr Φ_α when $n \longrightarrow +\infty$. ◀

Proof. See Embrechts et al., 1997a , Theorem 3.3.7, page 13 ■

▶ **Theorem 2.35 (Characterization of $\mathcal{D}(\Psi_\alpha)$).** The fdr belongs to the domain of attraction of the Weibull law with parameter $\alpha > 0$ iff $x_F < +\infty$and:

$$\overline{F}(x) = \left(x_F - \frac{1}{x}\right) = x^{-\alpha} l(x) \tag{2.22}$$

where the function l is slowly varying. Moreover if $F \in \mathcal{D}(\Psi_\alpha)$ with $a_n = x_F - F^{-1}(1 - \frac{1}{n})$ and $b_n = x_F$ the sequence $\left(a_n^{-1}\left(X_{n,n} - x_F\right)\right)_{n\geq 1}$converges in law to goes from fdr Ψ_α when $n \longrightarrow +\infty$. ◀

Proof. The proof of this theorem is similar to that of the previous theorem, See Embrechts et al., 1997a , Theorem 3.3.7, page 131. for the converse. ■

The results concerning the domain of attraction of Gumbel's law are more delicate, since there is no simple representation for the laws belonging to the domain attraction of Gumbel

▶ **Theorem 2.36 (Characterization of $\mathcal{D}(\Lambda)$).** The fdr belongs to the domain of attraction of Gumbel's law if and only if :

$$\overline{F}(x) = c(x)\exp\left\{-\int_z^x \frac{g(t)}{a(t)}dt\right\}, z < x < x_F. \tag{2.23}$$

where c and g are two satisfying measurable functions $c(x) \longrightarrow c > 0$ and $g(x) \longrightarrow 1$when $x \longrightarrow x_F$ and a is a positive, absolutely continuous function (with respect to the Lebesgue measure) with the density $a^{'}$ having $\lim_{x\to x_F} a^{'}(x) = 0$. In this case, a choice possible for the standardization sequences is:

$$a_n = x_F - F^{-1}(1-\frac{1}{n}) \quad \text{and} \quad b_n = \frac{1}{\overline{F}(a)}\int_{a_n}^{x_F}\overline{F}(y)dy,$$

◀

Distributions	$\overline{F}(x)$	γ
$\mathcal{U}[0,1]$	$1-x$	-1
inverse Burr$(\beta,\tau,\lambda,x_\tau)\beta,\tau,\lambda>0$	$\left(\frac{\beta}{\beta+(x_\tau+x)^{-\tau}}\right)^{\lambda}$	$-\frac{1}{\lambda}$

Table 2.1: Some distributions associated witha negative index

Distributions	$\overline{F}(x)$	γ
Pareto$(\alpha),\alpha>0$	$x^{-\alpha},x>0$	$\frac{1}{\alpha}$
Burr$(\beta,\tau,\lambda),\beta>0,\tau>0,\lambda>0$	$\left(\frac{\beta}{\beta+x^{\tau}}\right)^{\lambda}$	$\frac{1}{\lambda^{\tau}}$
Fréchet $\left(\frac{1}{\alpha}\right),\alpha>0$	$1-\exp(-x^{-\alpha})$	$\frac{1}{\alpha}$
log gamma$(m,\lambda),\lambda>0,m>0$	$\frac{\lambda^m}{\Gamma(m)}\int_x^\infty(\log u)^{m-1}u^{-\lambda-1}du$	$\frac{1}{\lambda}$
loglogistic$(\alpha,\beta),\alpha>0,\beta>0$	$\frac{1}{1+\beta x^{\alpha}}$	$\frac{1}{\alpha}$

Table 2.2: Some distributions associated with a positive index

► **Example 2.37.** The following tables give different examples of standard distributions in these three domain of attraction: ◄

Distributions	$\overline{F}(x)$	γ
Gamma$(m,\lambda),\lambda>0,m\in\mathbb{N}$	$\frac{\lambda^m}{\Gamma(m)}\int_x^\infty u^{-m-1}\exp(-\lambda u)du$	0
Gumbel $(\mu,\beta),\beta>0,\mu\in\mathbb{R}$	$\exp\Big(-\exp(-\frac{x-\mu}{\beta})\Big)$	0
Logistic	$\frac{2}{1+\exp(x)}$	0
log gamma$(\mu,\sigma),\mu\in\mathbb{R},\sigma>0$	$\frac{1}{\sqrt{2\pi}}\int_1^\infty\frac{1}{\mu}\exp\Big(-\frac{1}{2\sigma^2}(\log u-u)^2\Big)du$	0
Weibull$(\lambda,\tau),\lambda>0,\tau>0$	$\exp(-\lambda x^{\tau})$	0

Table 2.3: Some Distributions Associated with a Null Index

2.8 Estimation of the extreme value index

We focus in this section on the tail index parameter, we give study for different estimators with some of their statistical properties Perhaps the two most popular estimators in the literature are the estimators from Hill 1975 and Pickands 1975.

we shall mention that the case of $\gamma > 0$ has got more interest because data sets in most real life applications exhibit heavy-tails.

We denote by $\left(X_{1,n}, ..., X_{n,n}\right)$ the order statistics associated with the sample $(X_1, ..., X_n)$, i.e. say that we classify $(X_1, ..., X_n)$ in ascending order so that:

$$X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}.$$

Consider the k largest (or smallest) values. k depends a priority on n, even if we will not mention it in the notation: the idea is to have $k \longrightarrow \infty$ when $n \longrightarrow \infty$, but without taking "too many" values from the sample, which leads to impose $\frac{k}{n} \longrightarrow 0$. Incidentally, this implies that we will ask ourselves the question of the optimal choice of k .Indeed, it is essential to calculate this estimator on the tails of the distribution. Choosing too high a k generates the risk of taking into account values which are not extreme my conversely, too small sub sample does not allow the estimators reach their level of stability. Finally, it will be noted that the non parametric approach is only possible if one has a large number of observations: if the samples are small,we will turn to the parametric approach.

2.8.1 Pickand's estimator

The Pickands estimator was introduced in 1975 by James Pickands III, 1975 for any $\gamma \in \mathbb{R}$.

▶ **Definition 2.38 (Pickand's estimator).** Either $(X_1,...,X_n)$ n iid random variable $F \in \mathcal{D}(\Phi_{\frac{1}{\gamma}})$, where $\gamma \in \mathbb{R}$. Let $k = k_n$a series of integers with $1 < k < n$, the Pickand estimator is defined by:

$$\widehat{\gamma}^p = \widehat{\gamma}^p(k) := \frac{1}{\log(2)} \log\left[\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n-}}\right], \tag{2.24}$$

◀

The author demonstrates the weak Convergence of his estimator. Strong convergence as well that asymptotic normality have been demonstrated by Dekkers and De Haan, 1989. Of improvements of this estimator were introduced in particular by Drees, 1995 .Under certain conditions on the entire sequence k and the fdr F, the estimator of γ has good asymptotic properties, they are grouped together in the following theorem

▶ **Theorem 2.39 (asymptotic properties of $\widehat{\gamma}^p$).** Let $F \in \mathcal{D}(\mathcal{H}_\gamma)$,$\gamma \in \mathbb{R}$, $k \longrightarrow \infty$, $\frac{k}{n} \longrightarrow 0$,when $n \longrightarrow \infty$

- Convergence in probability:

$$\widehat{\gamma}^p \xrightarrow{p} \gamma \quad \text{when } n \longrightarrow \infty.$$

- Strong convergence (almost sure): If $k/\log\log(n) \longrightarrow \infty$ when $n \longrightarrow \infty$, then

$$\widehat{\gamma}^p \xrightarrow{p.s} \gamma \quad \text{when } n \longrightarrow \infty.$$

- Asymptotic normality: We suppose that U admits positive derivatives $U^{'}$ and that $\pm t^{1-\gamma}U^{'}(t)$ (with one or the other choice of sign) is regularly varying at infinity with the auxiliary function a. If $k = 0(n/g^{-1}(n))(n \longrightarrow \infty)$, $g(t) = t^{3-2\gamma}(U^{'}(t)/a(t))^2$, then :

$$\sqrt{k}(\widehat{\gamma}^p - \gamma) \xrightarrow{L} \mathcal{N}(0, \sigma^2) \quad \text{when } n \longrightarrow \infty$$

$$\text{and } \sigma^2 = \frac{\gamma^2(2^{2\gamma+1}+1)}{(2(2\gamma-1)\log 2)^2}.$$

◀

2.8.2 Hill's estimator

Research has mainly focused on when the EVI is positive $\left(\gamma = \frac{1}{\alpha} > 0\right)$ because data sets in most real applications , which corresponds to the distributions belonging to the domain of attraction of Fréchet $F \in \mathcal{D}(\Phi_{\frac{1}{\gamma}})$, that is, when the distribution tail has a Pareto shape. the best known estimator of γ is the estimator proposed by Hill is given by the next definition:

▶ **Definition 2.40 (Hill estimator $\widehat{\gamma}^H$).** Let $X_1, ..., X_n$ be n iid random variable $F \in \mathcal{D}(\Phi_{\frac{1}{\gamma}})$, where $\gamma \in \mathbb{R}$. Let $k = k_n$ a series of integers with $1 < k < n$ the Hill estimator is defined by.

$$\widehat{\gamma}^H = \widehat{\gamma}^H(k) := \frac{1}{k}\sum_{i=1}^{k} \log\left[\frac{X_{n-i+1,n}}{X_{n-k,n}}\right]. \tag{2.25}$$

◀

The construction of this estimator is given in the book by De Haan et al., 2006 and in the book by Beirlant et al., 2016. Other TI estimators have been proposed in particular by Beirlant et al., 2016 who use an exponential regression model base to the Hill estimator and by Csorgo et al., 1985 who use a kernel in the Hill estimator. A large number of theoretical works have been devoted to the study of the properties of the Hill estimator. The weak consistency was established by Mason, 1982, and the strong consistency was established in 1988 by Deheuvels et al., 1988 and more recently by Necir, 2006.

▶ **Theorem 2.41 (asymptotic properties of $\widehat{\gamma}^H$).** Let $F \in \mathcal{D}(\mathcal{H}_\gamma)$,$\gamma \in \mathbb{R}, k \longrightarrow \infty, \frac{k}{n} \longrightarrow 0$,when $n \longrightarrow \infty$

- Convergence in probability:

$$\widehat{\gamma}^H \xrightarrow{p} \gamma \quad \text{when } n \longrightarrow \infty.$$

- Strong convergence (almost sure): If $k/\log\log(n) \longrightarrow \infty$ when $n \longrightarrow \infty$, then

$$\widehat{\gamma}^H \xrightarrow{p.s} \gamma \ \text{ when } n \longrightarrow \infty.$$

- Asymptotic normality: We suppose that F satisfying the second order Condition if $\sqrt{k}A(k/n) \longrightarrow \lambda$ when $n \longrightarrow \infty$. then :

$$\sqrt{k}\left(\widehat{\gamma}^H - \gamma\right) \xrightarrow{L} \mathcal{N}(\frac{\lambda}{1-\tau}, \gamma^2) \quad \text{when } n \longrightarrow \infty$$

◀

In the general case of the Fréchet domain, the survival function has the form $\overline{F}(x) = x^{-\frac{1}{\gamma}} l(x)$with l a slowly varying function. This induces a significant bias on the Hill estimator, which is therefore in practice a delicate handling in the general case.

2.8.3 Optimal sample fraction selection

The number k of the order statistic is difficult to choose. The results concerning the estimators of the extreme value index are asymptotic when $k \longrightarrow \infty$ and$\frac{k}{n} \longrightarrow 0$. As in practice, we only have a finite number of observations n , it is to choose k so that we have enough statistical data while remaining in the distribution queue.

Graphic method

It is the simplest method for the determination of k It consists in tracing the graph$\left(k, \widehat{\gamma}^{H}_{k_n,n}\right)$with $k_n = k$ a sequence of integers and $1 < k < n$ .and take the value where$\left(k, \widehat{\gamma}^{H}_{k_n,n}\right)$becomes horizontal. this estimator is valid only in the Fréchet domain of attraction, i.e.$\gamma > 0$. for generalize to other domain of attraction, different estimators have been proposed, among others the Pickands estimator.

Analytical method

It is necessary to give precision to the estimator $\left(k, \widehat{\gamma}^{H}_{k_n,n}\right)$calculate the root mean square error (RMSE), it is a function of k

$$\begin{aligned} RMSE\left(\widehat{\gamma}_{k_n,n}\right) &= RMSE\left(\widehat{\gamma}_{k_n,n} - \gamma\right)^2 \\ &= biais^2(\widehat{\gamma}_{k_n,n}) - Var(\widehat{\gamma}_{k_n,n}). \end{aligned}$$

The optimal choice of k , corresponds to minimize MSE. Regarding the Hill estimator for functions belonging to the domain of maximum attraction of Fréchet, of Haan and Peng in 1998 proposed to retain the number of observations k_{opt} which minimizes the root mean square error of the Hill estimator which is

$$k_{opt} = \begin{cases} 1 + 2^{\frac{2\gamma}{\left(\frac{2\gamma+1}{2\gamma}\right)}} \left[\frac{(\gamma+1)^2}{2\gamma}\right]^{\frac{1}{2\gamma+1}} & \text{, if } 0 < \gamma < 1 \\ 2n^{\frac{2}{s}} & \text{, if } \gamma > 1. \end{cases}$$

Numerical method

There are several algorithms to find an estimator $\widehat{k}_{opt}$of k_{opt}

$$\frac{\widehat{k}_{opt}}{k_{opt}} \longrightarrow 1 \quad \text{when } n \longrightarrow \infty.$$

then $\widehat{\gamma}_{\widehat{k}_{opt},n}$converges asymptotically to $\gamma_{k_{opt},n}$

▶ **Remark 2.42.** – If k is small, $\widehat{\gamma}_{k,n}$ uses little observation and has a large variance.

– if k is large, the bias is large, the variance is small

◀

3 Survival Analysis

The statistical analysis or what is variously referred to as lifetime survival time or failure, time is an important brunch that deals with analyzing the expected duration of time until one event occurs, for example in the biological organisms the event is death same event when it comes to mechanical systems.

This topic is very significant in many areas such as; biomedical, social sciences; which is called event history analysis, in engineering is named by reliability theory or reliability analysis and in the economy, it's known as duration analysis or duration modeling. Some methods of dealing with lifetime data are quite old, but starting from 1970 the field had known a rapid extend with respect to methodology and fields application. Since the importance of this brunch in our work; and in order to make this thesis easier to read, this chapter is concerned about giving some basic concepts and definitions, please check thesis of Soltane, 2017 who's deal with survival analysis for more information.

3.1 Basic concepts and definitions

*Consider a **probability space** or **probability triple*** $(\Omega, \mathrm{A}, \mathrm{P})$ *such that:*

(i) Ω *: a set of all possibles outcomes (**a sample space**).*

(ii) A: *a set of events* A.

(iii) P: *a probability function, which assigns each event in the event space a probability, which is a number between **0** and **1**.*

Let X be a random variable (rv), defined on some probability space $(\Omega, \mathrm{A}, \mathrm{P})$ *representing the survival time. there are* ***three*** *basic conditions for survival time must be defined precisely which are :*

- ***Time origin:*** *must be specified such that individuals are as much as possible on equal footing. For example if the survival time of patients with particular type of cancer is being studied the time origin could be chosen to be the time point of diagnosis of that type of cancer.*
- ***End point:*** *or event of interest should be appropriately specified, such that the times considered are well defined. In the above example this is could be death due to the cancer studied.*
- ***Length of time:*** *From the time origin to the end end point could be calculated.*

The distribution of X could be characterized by the following functions:

* *Survival function,*
* *Density function,*
* *Hazard function.*

Before talking about those functions; we define the distribution function (fdr or fd) of X:

▶ **Definition 3.1 (Distribution function).** The distribution function or df is an application F defined on $\mathbb{R}_+$ to $[0, 1]$ by:

$$F(t) := P(X \leq t) \tag{3.1}$$

◀

- *F also called the distribution function or cumulative distribution function.*

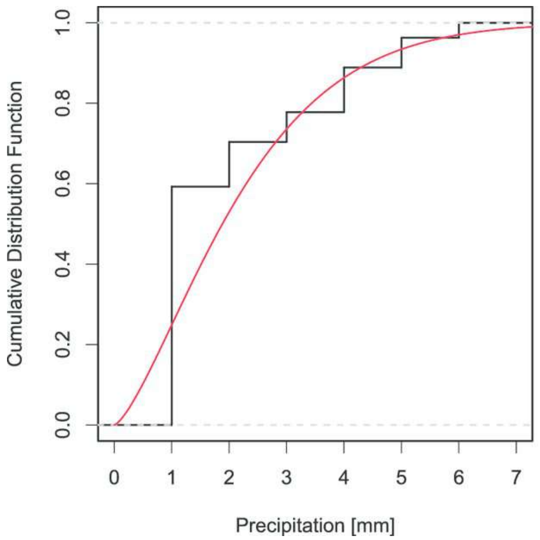

Figure 3.1: Cumulative distribution function

- *The function F is a right continuous monotonic increasing function such as:*

$$\lim_{t \longrightarrow 0} F(t) = 0 \qquad \mathit{and} \qquad \lim_{t \longrightarrow \infty} F(t) = 1.$$

▶ **Definition 3.2 (Survival function).** Mathematically a survival function is quite obviously a function of time. Survival function can be also interpreted as the probability that a certain object of interest will survive beyond a certain time so; then the survival function of rv Xcan be represented as:

$$S(t) := \overline{F}(t) = 1 - F(t) = P(X > t). \tag{3.2}$$

◀

- *The survival function of a rv X is left monotonic decreasing* *continuous such as:*

$$\lim_{t \to 0} S(t) = 1 \qquad \textit{and} \qquad \lim_{t \to \infty} S(t) = 0.$$

▶ **Definition 3.3 (Probability density function).** The probability density function $f(t)$ is the frequency of events per unit-time can be represented as:

$$f(t) := \frac{dF(t)}{dt} = \lim_{dx \to \infty} \frac{P(t < X < t + dx)}{dx}. \tag{3.3}$$

◀

▶ **Definition 3.4 (Hazard function).** The Hazard function is the instantaneous rate at which events occur for individual which are surviving at the time; so if X is a positive continuous variable representing a duration . The hazard function denoted by $h(t)$; is defined by:

$$h(t) := \lim_{dx \to \infty} \frac{P(t < X < t + dx \mid X > t)}{dx}, \tag{3.4}$$

Recall that cumulative Hazard function of distribution F is defined by:

$$\Lambda(t) := \int_0^t h(x)dx = \int_0^t \frac{f(x)}{1 - F(x)} dx; \tag{3.5}$$

it is clearly easy that the two different notions had a relation between them for example 3.5 implies that:

$$\Lambda(t) = -\log(\overline{F}(t)).$$

also we can write:

$$\overline{F}(t) = -\exp(-\Lambda(t)) = -\exp\left\{-\int_0^t \frac{f(x)}{1 - F(x)} dx\right\}.$$

◀

3.2 Laws of large numbers

The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. These Laws of two kinds describe the asymptotic behavior of the sample mean. the weak law is about the convergence in probability or consistency of $\overline{X}_n$ *while the strong low, due to Klongorov, concerns the a.s converge or the strong converge of* $\overline{X}_n$*; i.e converge with probability.*

▶ **Definition 3.5 (Sum and arithmetic mean).** Let $X_1, X_2, \ldots$ be a sequence of iid rv with common df F. for an integer $n \geq 1$, define the partial sum and the corresponding arithmetic mean respectively by:

$$\mathcal{S}_n := \sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{X}_n := \frac{\mathcal{S}_n}{n},$$

In what follows $(X_1, X_2, \ldots, X_n)$will considered as a sample from rv X defined on probability space(Ω, A, P), $\overline{X}_n$is called the sample mean or empirical mean. ◀

▶ **Definition 3.6 (Empirical distribution and survival functions).** Let $X_1, X_2, \ldots, X_n$ a sample of size $n \geq 1$ of a positive rv X fdr F and a function of survival S. the empirical distribution and survival functions; F_n and S_n are respectively defined by:

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq t) \quad \text{and} \quad S_n(t) = 1{-}F_n(t) = \frac{1}{n}\mathbf{1}(X_i \geq t). \quad \forall t \geq 0,$$

we can write F_n and S_n in terms of the value statistics order as follows:

$$F_n \begin{cases} 0 & if\ t < X_{1,n}, \\ \frac{i}{n} & if\ X_{i,n} < t < X_{i+1,n}, \\ 1 & if\ \ \text{t} \geq X_{n,n}. \end{cases}$$

$$S_n \begin{cases} 1 & if\ t < X_{1,n}, \\ 1 - \frac{i}{n} & if\ X_{i,n} < t < X_{i+1,n}, \\ 0 & if\ \ t \geq X_{n,n}. \end{cases}$$

◀

▶ **Theorem 3.7 (Laws of large numbers).** if $(X_1, X_2, ..., X_n)$ is a sample from a rv X such that $E(X) < \infty$, then

$$\overline{X}_n \overset{p}{\longrightarrow} \mu \quad \text{as} \quad n \longrightarrow \infty; \quad \text{Weak law,}$$
$$\overline{X}_n \overset{a.s}{\longrightarrow} \mu \quad \text{as} \quad n \longrightarrow \infty; \quad \text{Strong law.}$$

where $\mu := E(X)$. ◀

applying this Strong law of large numbers on $F_n(x)$ yield the following result:

▶ **Corollary 3.8.**

$$\overline{F}_n \overset{a.s}{\longrightarrow} F \quad \text{as} \quad n \longrightarrow \infty; \quad \text{for every } x \in \mathbb{R}.$$

◀

▶ **Theorem 3.9 (Glivenko-Contelli).**

$$\sup_{x\in\mathbb{R}} \left| \overline{F}_n(x) - F(x) \right| \overset{a.s}{\longrightarrow} 0 \quad \text{as} \quad n \longrightarrow \infty.$$

◀

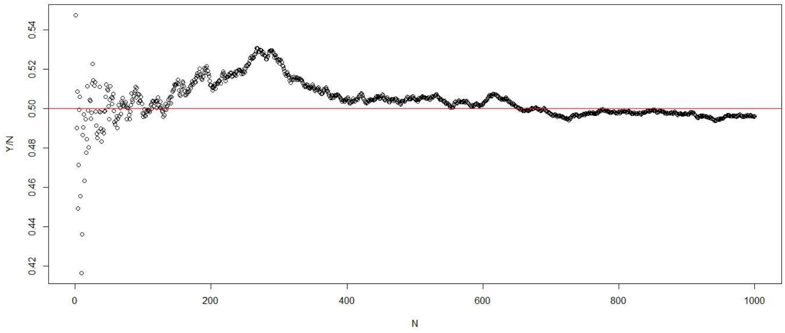

Figure 3.2: The empirical mean for a sample of the uniform distribution on $[0, 1]$ with $n = 1000$

For more details and proofs please check any standard textbook of probability theory such as Billingsley, 1995.

▶ **Theorem 3.10 (CLT).** Let $(X_n, n \in N^*)$ be a sequence of independent, random variables and identically distributed defined on the same probability space (Ω, A, P) with mean μ and finite variance σ^2 suppose that: $\forall i \in N^* \ \ ,E(X^2) < \infty$.

then

$$\sqrt{n}\frac{(S_n - n\mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad n \longrightarrow \infty.$$

3.3 Estimating the mean of a heavy-tailed distribution

3.3.1 Estimating the mean of a heavy-tailed distribution in the case of finite second moment

The most fundamental problem of statistics is that estimating the expected value μ of random variable X :

▶ **Definition 3.11.** Let X be a random variable with a distribution function F. we call the mathematical expectation of X, which we denote by: $E[X]$ or μ that it is defined by :

$$E[X] := \int_{\mathbb{R}} x dF(x) := \int_{\mathbb{R}} x f(x) dx, \tag{3.6}$$

where f is the density function of F. ◀

▶ **Definition 3.12.** we could give another formula of the mean by using the quantile function if we set $t = F(x)$ with change of limit condition of integral we find that $E[X]$ defined by:

$$E[X] := \int_0^1 Q(t) dt.$$

◀

▶ **Remark 3.13.** The moment of order k is $E[X^k]$ where $k > 0$. ◀

The obvious choice of an estimator for the mean is of course, the empirical mean:

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

There are two methods to estimate any parameters of a population: **empirical distribution** and the **quantile empirical**; even in this case there two method for estimating the mean which are :

- **Empirical distribution:** after definition 3.6 we have:

$$
\begin{aligned}
E[X] := \mu &:= \int_{\mathbb{R}} x dF(x); \\
\widehat{\mu} &:= \int_0^1 x dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i, \\
\widehat{\mu} &:= \overline{X}.
\end{aligned}
$$

- **Empirical quantile:** as we defined the other formula for the expectation of X by:

$$
\begin{aligned}
E[X] := \mu &:= \int_0^1 Q(t)dt, \\
\widehat{\mu} &:= \int_0^1 Q_n(t)dt, \\
\widehat{\mu} &:= \overline{X}.
\end{aligned}
$$

3.3.2 Estimating the mean of a heavy-tailed distribution in the case of infinite second moment

Peng, 2001 proposed an estimator for the mean of a heavy tailed distribution with the tail index $\alpha > 1$ the sample mean is not good estimator of the population mean. Peng, 2001 defined his estimator as the sum of mean estimates for the tail and non-tail regions.

Let $X_1, ..., X_n$ iid random variables with the common distribution function F, with regularly varying tails and index $\alpha > 1$ that satisfies following conditions

$$\lim_{t\to\infty} \frac{1 - F(tx) + F(-tx)}{1 - F(t) + F(-t)} = x^{-\alpha}; x > 0 \tag{3.7}$$
$$\lim_{t\to\infty} \frac{1 - F(t)}{1 - F(t) + F(-t)} = p \in [0, 1].$$

this is implies when:

$\alpha-$**value**	**The domain attraction of** F.
$1 < \alpha < 2$	domain attraction of a stable law.
$\alpha \geq 2$	domain attraction of a normal distribution.

To obtain a consistent estimator of the sample mean for any $\alpha > 1$ Peng, 2001 partitioned the population mean $E[X]$ into:

$$\begin{aligned} E[X] = \int_0^1 Q(t)dt = \int_0^{k/n} Q(t)dt + \int_{k/n}^{1-k/n} Q(t)dt + \int_{1-k/n}^1 Q(t)dt. \\ = \mu_n^{(1)}(k) + \mu_n^{(2)}(k) + \mu_n^{(3)}(k). \end{aligned}$$

where $Q(t) := \inf\{x : F(x) \geq t\}, 0 \leq t \leq 1$; denote the inverse function of F and the sample fraction extremes k, is equal to number of observation in the upper tail with $k := k(n)$ satisfying the following condition:

$$k \longrightarrow \infty \quad \text{and} \quad \frac{k}{n} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty;$$

Then the $\mu_n^{(1)}$, $\mu_n^{(2)}$and $\mu_n^{(3)}$are estimated separately; for more details please check Peng, 2001. and with uses of extreme value theory the mean for right tail is estimated as follow

$$\widehat{\mu}_n^{(3)}(k) := \frac{k}{n} X_{n-k,n} \frac{\widehat{\alpha}}{\widehat{\alpha}-1},$$

note that $\widehat{\gamma} = \frac{1}{\widehat{\alpha}}$. The new estimator is $\widehat{\mu}_n^{(1)}(k) + \widehat{\mu}_n^{(2)}(k) + \widehat{\mu}_n^{(3)}(k)$ where $\widehat{\mu}_n^{(1)}(k)$ is the simple average of all observation excluding observation in the right tail; This estimator has normal distribution presented on the following theorem.

▶ **Theorem 3.14 (Peng 2001).** Assume that conditions holds for $\alpha > 0$ and $\beta > 0$:

$$\lim_{t\to\infty} \frac{1-F(tx)+F(-tx)/1-F(t)+F(-t)}{\mathbf{A}(t)} = x^{-\alpha}\frac{x^{-\beta}-1}{-\beta}; x > 0$$
$$\lim_{t\to\infty} \frac{1-F(t)/1-F(t)+F(-t)-p}{\mathbf{A}(t)} = q.$$

where: • $\mathbf{A}(t)$:=function with constant sign.

• $q \in \mathbb{R}$.

Let $k = k(n)$ denote an intermediate integer sequence satisfying $k = o(n^{2\beta/(\alpha+2\beta)})$then:

$$\frac{\sqrt{n}}{\sigma(k/n)}\left(\widehat{\mu}_n^{(1)}(k) + \widehat{\mu}_n^{(2)}(k) + \widehat{\mu}_n^{(3)}(k) - E[X]\right) \xrightarrow{d} \mathcal{N}(0, 1 + \left\{\frac{(2-\alpha)(2\alpha^2-2\alpha+1)}{2(\alpha-1)^4} + \frac{2-\alpha}{\alpha-1}\right\} 1_{(\alpha<2}$$

◀

Proof. For those who interested in proof of the theorem please check the paper of Peng, 2001, pages from 259 to 264 ◀

3.3.3 Kernel-type estimator of the mean for a heavy tailed distribution

aforementioned in the previous section that the classical mean estimator introduced by Peng, 2001 which is based under the second order regular variation. in this section we introduce the work of Rassoul, 2015 who defined a kernel type estimator for the mean and proposed a reduced bias estimator with asymptotic distributional properties.

Let the non-negative and independent and identically distributed (iid) random variables $X_1, X_2, ..., X_n$ with the cdf F and let $X_{1,n} < X_{2,n} < ... < X_{n,n}$ be the corresponding order statistics in this case we shall mention that $F(x) = 0$ for $x < 0$ and $P = 1$ in Peng, 2001 condition 3.7.

To obtain the kernel type estimator Rassoul, 2015 work with some assumptions about the kernel:

Conditions ($\mathcal{K}$): let $\mathcal{K}$ be a function defined on $(0, 1]$

Condition 01: $\mathcal{K}(s) \geq 0$ whenever $0 < s < 1$ and $\mathcal{K}(1) = 0$;

Condition 02: $\mathcal{K}(.)$ is differentiable, non increasing and right continuous on $(0, 1]$;

Condition 03: $\mathcal{K}$ and $\acute{\mathcal{K}}$ are bounded;

Condition 04: $\int_0^1 \mathcal{K}(u)du = 1$;

Condition 05: $\int_0^1 u^{-1/2}\mathcal{K}(u)du < \infty$.

where $\mathcal{K}$ is a kernel integrating to one the proposed kernel-type estimator for the mean defined by:

$$\widehat{\mu}_n^{\mathcal{K}}(k) := \int_0^{1-k/n} Q_n(s)ds + \frac{(k/n)X_{n-k,n}}{(1-\widehat{\gamma}_n^{\mathcal{K}}(k))}; \tag{3.8}$$

Csorgo et al., 1985 extended Hill estimator into estimator into a kernel class of estimation $\widehat{\gamma}_n^{\mathcal{K}}(k)$:

$$\widehat{\gamma}_n^{\mathcal{K}}(k) := \frac{1}{k}\sum_{i=1}^{k}\mathcal{K}\left(\frac{i}{k+1}\right)Z_{i,k}; \tag{3.9}$$

where

$$Z_{i,k} = i\left(\log X_{n-i+1,n} - \log X_{n-i,n}\right), \quad 1 \le i \le k < n.$$

▶ **Remark 3.15.** the Hill estimator corresponds to the particular case where $\mathcal{K} = \underline{\mathcal{K}} := \mathbf{1}_{(0,1)}$ ◀

Asymptotic results for the mean estimator

Obviously the asymptotic normality if $\widehat{\mu}_n^{\mathcal{K}}(k)$is related to $\widehat{\gamma}_n^{\mathcal{K}}(k)$ to prove that type of results in the extreme value framework we need a second order condition on the tail quantile function U, with second order parameter $\rho \le 0$ if there exists a function $\mathbf{A}(t)$ which does not change its sign in a neighbourhood of infinity with $\lim_{t\to\infty}\mathbf{A}(t) = 0$ such that:

$$\lim_{t\to\infty}\frac{\log U(tx) - \log U(t) - \gamma\log(x)}{\mathbf{A}(t)} = \frac{x^\rho - 1}{\rho} \tag{3.10}$$

▶ **Theorem 3.16 (Asymptotic results for the mean estimator).** Assume that F satisfies 3.10 with $\gamma \in (1/2, 1)$ if further $(\mathcal{K})$ holds and the sequence k satisfies

$$k \to \infty, \quad k/n \to 0 \quad \text{and if} \sqrt{k}\mathbf{A}(n/k) \to \lambda \in \mathbb{R}, \ \textit{as } n \to \infty,$$

we have

$$\frac{\sqrt{k}}{(k/n)\mathbf{U}(n/k)}\left(\widehat{\mu}_n^{\mathcal{K}}(k) - \mu\right) \xrightarrow{d} \mathcal{N}(\lambda\mathcal{AB}_{\mathcal{K}}(\lambda,\rho), \mathcal{AC}_{\mathcal{K}}(\lambda,\rho)),$$

Where:

$$\mathcal{AB}_{\mathcal{K}}(\lambda,\rho) = \left(\frac{1}{(\gamma-1)(\gamma+\rho-1)} + \frac{1}{(1-\gamma)^2}\int_0^1 \frac{\mathcal{K}(s)}{s^{\rho}}ds\right);$$

and

$$\mathcal{AC}_{\mathcal{K}}(\lambda,\rho) = \left(\frac{\gamma^2}{(1-\gamma)^2(2\gamma-1)} + \frac{\gamma^2}{(1-\gamma)^4}\int_0^1 \mathcal{K}^2(s)ds\right).$$

◀

▶ **Remark 3.17.** In the case of $\lambda \neq 0$ and when we use the general kernel instead of $\mathcal{K}$; the result of this theorem will be generalizes theorem in Peng. ◀

▶ **Theorem 3.18 (Bias-correction for the mean estimator).** Under the same assumptions of Theorem 3.16, and if $\widehat{\rho}$ is a consistent estimator for $\rho < 0$, then we have :

$$\frac{\sqrt{k}}{(k/n)\mathbf{U}(n/k)}\left(\widehat{\mu}_n^{\mathcal{K},\widehat{\rho}}(k) - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \widetilde{\mathcal{AC}}_{\mathcal{K}}(\lambda,\rho)\right),$$

where :

$$\begin{aligned}\widetilde{\mathcal{AC}}_{\mathcal{K}}(\lambda,\rho) &= \mathcal{AC}_{\mathcal{K}}(\lambda,\rho) + \frac{\gamma^2}{\rho^2}(1-2\rho)(1-\rho)^2\mathcal{AB}^2_{\mathcal{K}}(\lambda,\rho)\\ &+ \frac{2\gamma^2(1-2\rho)(1-\rho)}{\rho^2(1-\gamma)^2}\\ &\times\left(1-(1-\rho)\int_0^1\frac{\mathcal{K}(s)}{s^{\rho}}ds\right)\mathcal{AB}_{\mathcal{K}}(\lambda,\rho).\end{aligned}$$

◀

Proof. For more details and proofs please check the paper of Rassoul, 2015. ◀

4 Taxonomy of incomplete Data

The key that distinguishes survival analysis from another area in statistics is that survival data are usually incomplete in some way. This Incomplete data are questions without answers or variables without observations. Even a small percentage of missing data can cause serious problems with the analysis leading to the drawing of wrong conclusions and imperfect knowledge. There are many techniques to overcome imperfect knowledge and manage data with incomplete items, but no one is absolutely better than the others. To handle such problems, researchers are trying to solve them in different directions and then propose to handle the information system. Since our work deals with incomplete data; we choose to spot a light on some definitions and examples of incomplete data i.e truncated or censored. the famous thesis deal with this issue are the following: Djabrane, 2010, Benchaira, 2017, HAOUAS, 2017, and Soltane, 2017.

4.1 Censoring

Since censoring is the most common phenomenon, encountered when collecting survival data and as we mentioned the statistical techniques for analyzing censored data sets are quite well studied In this section, we will concentrate on talking about censored such that for a specific individual i under study we assume that:

(a) its life time is: X_i,

(b) its censoring time is: Y_i,

(c) the time actually observed: Z_i.

▶ **Definition 4.1.** Censoring is when an observation is incomplete due to some random case. The cause of the censoring must be independent of the event of interest if we are going to use standard methods of analysis. So, When a data set contains observations within a restricted range of values but otherwise not measured. it is called censored data set. ◀

4.1.1 Types of Censoring Mechanisms

There are three categories of censoring: The right, the left and interval censoring we define each one as follows:

Right censoring

▶ **Definition 4.2 (Right censoring).** The variable of interest is said to be right-censored if the individual concerned has no information about his observation. Thus in the presence of the right censoring the variables of interest are not observed at all. ◀

Type I of censoring *Let Y be a fixed value, instead of observing the variable of interest $X_1, ..., X_n$ we observe X_i is less than or equal to Y $(X_i \leq Y)$, otherwise the only thing that we know is $(X_i > Y)$, we can use the next notation $Z_i = X_i \wedge Y = \min(X_i, Y)$ which means we observe the variable Z_i where $Z_i = \min(X_i, Y)$ $i = 1, .., n$. This mechanism of censorship is frequently encountered in industrial applications; For example, we can test the lifetime of n identical objects (Lampes) over a fixed interval of observation $[0, u]$. This type of censoring is called* ***Fixed censoring*** *.*

Type II of censoring *This model is often used in reliability and epidemiology studies it is present when we decide to observe the survival times of n patients up to that k of them died and stop the study* *at that time; Let $X_{i,n}$ and $Z_{i,n}$ the order statistics of the variable X_i and Z_i, the censoring date is therefore $X_{k,n}$ and we observe the following variables:*

$$\begin{aligned} Z_{1,n} &= X_{1,n} \\ &. \\ &. \\ &. \\ Z_{k,n} &= X_{k,n} \\ Z_{k+1,n} &= X_{k,n} \\ Z_{n,n} &= X_{k,n} \end{aligned}$$

For a general formula:

$$\begin{cases} Z_{i,n} = X_{i,n} & \text{for } i \leq k, \\ Z_{i,n} = X_{k,n} & \text{for } i \geq k. \end{cases}$$

This kind of censoring is known as ***Censorship waiting****.*

Type III of censoring *Let $X_1, ..., X_n$ be a sample of a positive rv X, we say that is a random censoring of this sample if there exists another random variable Y of a sample $Y_1, ..., Y_n$ we observe in this case the couple of rv(Z_i, δ_i) :*

$$Z_i = X_i \wedge Y_i,$$

we can summarize the information that could be available to:

- *the actual time observed Z_i,*
- *$\delta_i = \mathbf{1}_{(X_i \leq Y_i)}$ the indicator of censor, which determines when the X has been censored or not*

$$(i)\delta_i = 1 \text{ the variable of interest is observed } (Z_i = X_i),$$

$$(ii)\delta_i = 0 \ \textit{it is censored} \ (Z_i = Y_i).$$

This type of censoring could named by random censoring and it is the most common.

▶ **Example 4.3 (Right censoring).** Consider the following example where we have 3 patients (A, B, C) enrolled in a clinical study that runs for a period of time (Study end - Study start).

These 3 patients have three different trajectories:

- Patient A: Experience a death before the study ends. we count this as an event.
- Patient B: Survives past the end of the study.
- Patient C: withdraws from the study.

Patient A requires no censoring since we know their exact survival time is the time until death.

Patient B however needs to be censored (indicated with the + at the end to the follow-up time) since we don't know the exact survival time of the patient, we only know that they survived up to at least the end of the study.

Patient C also needs to be censored since they withdrew before the study ended. So we only know that they survived up to the time they withdrew; but again we don't exact the survival time of this patient. In right censoring; the true survival times will always equal to or greater than the observed survival time. ◀

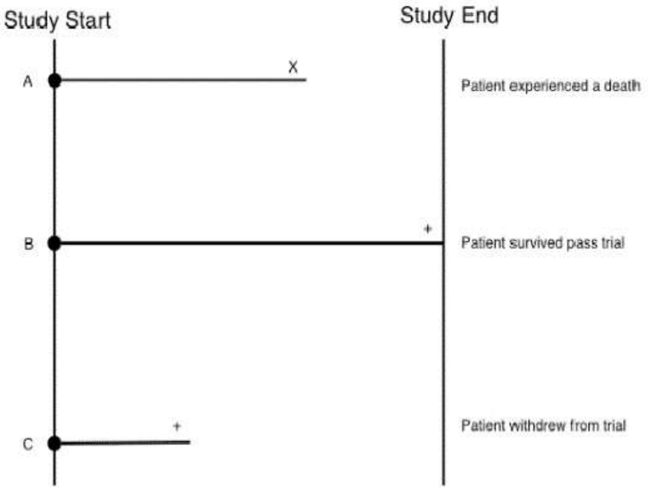
Study Start
Study End
A
X
Patient experienced a death
B
+
Patient survived pass trial
C
+
Patient withdrew from trial

Figure 4.1: Example of Right Censoring Mode

Left censoring

Left censoring is much rare. A lifetime X associated with a specific individual in a study is considered to be left censored if it is less than a censoring time Y, that is, the event of interest has already occurred for the individual before that person is observed in the study at time Y. For such individuals, we know that they have experienced the event sometime before time Y, but their exact event time is unknown. The exact lifetime X will be known if, and only if, X is greater than or equal to Y. The data from a left-censored sampling scheme can be represented by pairs of random variables (Z, δ_i), as in the previous kind, where Z is equal to X if the lifetime is observed and indicates whether the exact lifetime X is observed $(\delta = 1)$ or not $(\delta = 0)$. Note that, for left censoring as contrasted with right censoring, $Z = \max(X, Y_l)$.

▶ **Example 4.4 (Left censoring).** An example of a situation could be for virus testing. For instance, if we've been following an individual and

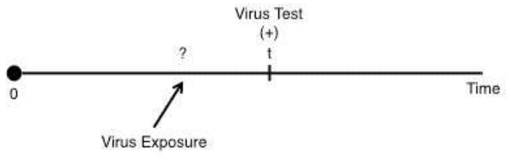

Figure 4.2: Example for Left Censoring Model

recorded an event when for instance the individual test's positive for a virus. But we don't know the exact time of when the individual was exposed to the disease. We only know that there was some exposure between 0 and the time they were tested.

◀

Interval censoring

As its name indicate and for more general type of censoring occurs when the lifetime is known to occur only within an interval i.e we observe both lower bound and upper bound of interest variable we found this model in general in medical followup studies such interval censoring occurs when patients in a clinical trial or longitudinal study have periodic follow-up and the patient's event time is only known to fall in an interval $(L_i, R_i]$ *where:*

- *L for left endpoint,*
- *R for right endpoint of the censoring interval.*

This type of censoring can be found in industrial experiments where there is periodic and animal tumorigenicity experiments where there is periodic insectation.

▶ **Remark 4.5.** Interval censoring is a generalization of left and right censoring because, when the left end point is 0 and the right end point is C we have left censoring by interval of type $[0, C]$and, when the left end point is C and the right end point is infinite, we have right censoring by interval of type $[C, \infty]$ ◀

▶ **Example 4.6.** Using the virus testing example, if we have the situation whether we've performed testing on the individual at some time point (t_1) and the individual was negative. But then at a time point further on (t_2), the individual tested positive. In this scenario, we know the individual was exposed to the virus sometime between $t_1$and t_2, but we do not know the exact timing of the exposure. ◀

4.1.2 Estimation under random right-censoring

In this section; we will place ourselves in the most frequent framework of a type I of a random right censoring (RRC) the main estimators:

- *The Kaplan-Meier estimator,*
- *The Nalson-Aalen cummulative risk estiamtor.*

Kaplan-Meier estimator

This section deals with the non parametric estimation of the df by means of the Kaplan–Meier estimator (also called the product–limit estimator) and with the estimator for the mean. We start with remarks about the statistics

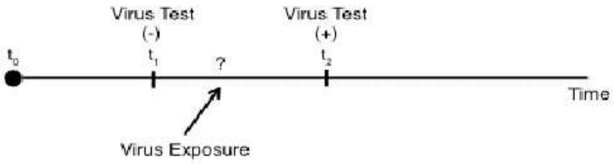

Figure 4.3: Example for Interval Censoring Model

of extremes of randomly censored data. The topic was mentioned in Reiss et al., 2007, Section 6.1, where an estimator of a positive extreme value index was introduced, but no (asymptotic) results were derived. In the last decade, several authors started to be interested in the estimation of the tail index along with large quantiles under random censoring as one can see in Worms and Worms, 2014, Beirlant et al., 2007, Einmahl et al., 2008 and Gomes and Neves, 2011 also made a contribution to this estimator by providing a detailed simulation study and applying the estimation procedures on some survival data sets.

Let $X_1, .., X_n$ be $n \geq 1$ independent copies of a positive random variable X; defined over some probability space(Ω, A, P) with continuous cumulative distribution function F.Rather then $X_1, .., X_n$, the variables of interest, one observes

$$Z_i = \min(X_i, Y_i) \quad \textit{and } \delta_i = \mathbf{1}_{X_i \leq Y_i}; \qquad 1 \leq i \leq n.$$

where $Y_1, .., Y_n$ is another i.i.d. sequence from some (censoring) d.f. G being also independent of the X's: This model is very useful in a variety of areas where random censoring is very likely to occur such as in bio-statistics, medical research, reliability analysis, actuarial science,...In the context of this randomly right censored observations, the non parametric maximum likelihood estimator of the survival distribution F is given by Kaplan and Meier, 1958 as the product limit estimator defined by:

$$F_n(x) = 1 - \prod_{Z_{i,n} \le x} (1 - \frac{\delta_{i,n}}{n-i+1}) \quad \textit{for} \;\; x < Z_{n;n};$$

where $Z_{i,n}$ denote the order statistics associated to $Z_1, .., Z_n$and $\delta_{i,n}$is the concomitant of the ith order statistics, that is, $\delta_{i,n} = \delta_j$ if $Z_{i,n} = Z_j$ This estimator may be expressed as follows:

$$F_n(x) = \sum_{i=2}^{n} W_{i,n} \mathbf{1}_{\{Z_{i,n} \le x\}} \qquad \textit{where } i = 2, ..., n$$

$$W_{i,n} = \frac{\delta_{i,n}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{j,n}} .$$

Estimating the mean under random censoring

In this section we are interested in estimating the mean of a distribution under random censoring; as we have presented before the different estimators for the expectation of the X but in case Stute, 1995 introduced an estimator called the Kaplein-Meire integral.

▶ **Definition 4.7.** the non-parametric estimator of the mean under random censoring is defined by:

$$\widetilde{\mu}_n = \sum_{i=2}^{n} W_{i,n} Z_{i,n,} \tag{4.1}$$

where

$$W_{i,n} = \frac{\delta_{i,n}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{j,n}}. \tag{4.2}$$

◀

Stute, 1995 showed that this estimator is asymptotically normal under the two following conditions:

$$\begin{aligned} I_1 &= \int_0^{\infty} x^2 \Gamma_0^2(x) dH^1(x) < \infty, \qquad (4.3) \\ I_2 &= \int_0^{\infty} x \left(\int_0^x \frac{dH^0(x)}{\left[\overline{H}(x)\right]^2} \right)^{1/2} dF(x) < \infty. \end{aligned}$$

where H^0 and H^1 two functions defined as :

$$\begin{aligned} H^0(t) &= P(Z \leq t, \sigma = 0) = \int_0^t \overline{F}(x) dG(x), \qquad (4.4) \\ H^1(t) &= P(Z \leq t, \sigma = 1) = \int_0^t \overline{G}(x) dF(x). \end{aligned}$$

with:

$$\Gamma_1(x) = \int_0^x \frac{s\Gamma_0(s)}{\overline{F}(s)} dH^1(s) \quad \textit{and} \quad \Gamma_2(x) = \int_0^x \frac{\int_s^{\infty} t\Gamma_0(t) dH^1(s)}{\left[\overline{F}(s)\right]^2} dH^0(t). \tag{4.5}$$

▶ **Theorem 4.8.** Suppose 4.3 holds we have :

$$\frac{\widetilde{\mu} - \mu}{\sqrt{n}} \to \mathcal{N}(0, \sigma^2).$$

where $\sigma^2 = Var[Z_1 \Gamma_0(Z_1)\delta_1 + \Gamma_1(Z_1)(1-\delta_1) - \Gamma_2(Z_1)]$. ◀

Proof. See Stute, 1995. ◀

4.2 Truncation

A second feature of many survival studies sometimes confused with censoring is truncation, *there are three categories of truncation are Right, left and interval:*

4.2.1 Right truncation

Right truncation occurs when only individuals with event timeless threshold included in the study, that is we observed the survival time X only when $X \leq Y$.

Right truncation arises, for example in estimating the distribution of stars from the earth in that stars too faraway are not visible and right truncated.

The second example of a right truncated sample is a morality study based on death records right-censored data is particularly relevant to studies of AIDS.

4.2.2 Left truncation

Here we only observe those individuals whose event time X exceeds the truncation time Y, i.e we observe X if and only if $X > Y$. famous example of left truncation is the problem of estimating the distribution of the diameters of microscopic particles. The only particles big enough to be seen based on the resolution of the microscope are observed and smaller particles do not come to the attention of the investigator. In survival studies the truncation event may be exposure to some disease, diagnosis of a disease, entry into a retirement home, occurrence of some intermediate event such as graft-versus-host disease after a bone marrow transplantation, etc. In this type of truncation any subjects who experience the event of interest prior to the *truncation time are not observed. The truncation time is often called a delayed entry time since we only observe subjects from this time until they die or are censored. Note that, as opposed to left censoring where we have partial information on individuals who experience the event of interest prior to age at entry, for left truncation these individuals were never considered for inclusion into the study.*

4.2.3 Interval truncation

Or doubly truncated failure-time arises if an individual is potentially observed and only if its failure-time falls within a certain interval, unique to that individual. Doubly truncated data play an important role in the statistical analysis of astronomical observations as well as in survival analysis.

▶ **Example 4.9.** data on the luminosity of quasars in astronomy: One of the main aims of astronomers interested in quasars is to understand the evolution of the luminosity of quasars see Efron and Petrosian, 1999. The motivating example presented in this paper concerns a set of measurements on quasars in which there is double truncation, because the quasars are observed only if their luminosity occurs within a certain finite interval, that is bounded at both ends, with the interval varying for different observations. ◀

4.3 Estimation under random truncated model

4.3.1 Estimation the distribution function under truncation model

In this section we will present the different estimators of the distribution function that will be presented in the case of incomplete data especially *truncated one because when the empirical data is incomplete empirical estimators will not produce good results. There are two famous estimators:*

- ***Woodroofe** estimator,*
- ***Lynden-Bell** estimator.*

Lynden-Bell estimator

Let $(\mathbf{X}_1, ..., \mathbf{X}_N)$ *be independent copies of a non-negative random variable (rv)* $\mathbf{X}$ *with cumulative distribution (cdf)* $\mathbf{F}$ *, defined over some probability space* $(\Omega, \mathcal{A}, \mathcal{P})$*, suppose that* $\mathbf{X}$ *is right truncated by sequences of independent copies* $(\mathbf{Y}_1, ..., \mathbf{Y}_N)$ *of (rv)* $\mathbf{Y}$ *with cdf* $\mathbf{G}$*, in the sense that* X_i *is only observed when* $X_i \leq Y_i$ *.*

Let denote now by (X_i, Y_i)*,* $i = 1, ..., n$ *to be observed data, as copies of a couple of rv's* (X, Y)*, corresponding to the truncated sample* $(\mathbf{X}_i, \mathbf{Y}_i), i = 1, .., N$*, where* $n = n_N$ *is a sequence of discrete rv's by the weak law of large numbers, we have*

$$\frac{n}{N} \longrightarrow p = \mathbf{P}(\mathbf{X} \leq \mathbf{Y}) \quad \textit{as} \ N \rightarrow \infty.$$

we shall assume that $p > 0$*, otherwise nothing will be observed. the joint* $\mathbf{P}$*-distribution of on observed* (X, Y) *is given by:*

$$\begin{aligned} H(x, y) &= \mathbf{P}(X \leq x, Y \leq y) \\ &= \mathbf{P}(\mathbf{X} \leq x, \mathbf{Y} \leq y \mid \mathbf{X} \leq \mathbf{Y}) = p^{-1} \int_0^y \mathbf{F}(\min(x, z)) d\mathbf{G}(z), \end{aligned}$$

The marginal distributions of the rv's X *and* Y *respectively denoted by* F *and* G *are defined by:*

$$F(x) = p^{-1} \int_0^x \overline{\mathbf{G}}(z) d\mathbf{F}(z) \quad \textit{and} \ G(y) = p^{-1} \int_0^y \mathbf{F}(z) d\mathbf{G}(z),$$
$$\overline{F}(x) = -p^{-1} \int_x^\infty \overline{\mathbf{G}}(z) d\overline{\mathbf{F}}(z) \ \ \textit{and} \ \ \overline{G}(y) = -p^{-1} \int_y^\infty \mathbf{F}(z) d\overline{\mathbf{G}}(z).$$

since right endpoint of F and G are infinite and thus they are equal.From Woodroofe, 1985 we may write :

$$\int_x^\infty d\mathbf{F}(y)/\mathbf{F}(y) = \int_x^\infty dF(y)/C(y)$$

where

$$\begin{aligned} C(z) &:= P(X \le z \le Y \mid X \le Y) \\ &= p^{-1}F(x)G(x). \end{aligned}$$

Differentiating the previous equation leads the following crucial equation

$$C(x)d\mathbf{F}(x) = \mathbf{F}(x)dF(x). \tag{4.6}$$

For randomly truncated data; the truncation product-limit estimate is the maximum likelihood estimate ($\mathcal{MLE}$) for non-parametric models the well-known non-parametric estimator of F in $\mathcal{RRT}$ model, proposed by Lynden-Bell, 1971 defined by:

$$\mathbf{F}_n^{(\mathbf{LB})}(x) = \prod_{i:X_i>x} \exp(1 - \frac{1}{nC_n(X_i)}).$$

where

$$C_n(x) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i \le x \le Y_i) \tag{4.7}$$

Woodroofe estimator

we can define the solution of 4.6 by:

$$\mathbf{F}(x) = \exp\left\{-\int_x^\infty \frac{dF(z)}{c(z)}\right\},$$

by replacing the df's F and C by their respective empirical counterparts:

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x) \text{ and } C_n(x) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x \leq Y_i),$$

this leads to non parametric estimator of df $\mathbf{F}$*;Wodroofe estimator given by Woodroofe, 1985:*

$$\mathbf{F}_n^{(\mathbf{W})}(x) := \prod_{i:X_i>} \exp\left\{-\frac{1}{nC_n(X_i)}\right\},$$

4.3.2 Estimation Tail-index under truncation model

We assume that $\mathbf{F}$ *and* $\mathbf{G}$ *are heavy-tailed in other words that* $\overline{\mathbf{F}} = 1 - \mathbf{F}$ *and* $\overline{\mathbf{G}} = 1 - \mathbf{G}$ *are regularly varying* $(\mathcal{RV})$ *at infinity with respective negative indices* $-1/\gamma_1$ *and* $-1/\gamma_2$*; we will use the notation:* $\overline{\mathbf{F}} \in \mathcal{RV}(-1/\gamma_1)$ *and* $\overline{\mathbf{G}} \in \mathcal{RV}(-1/\gamma_2)$ *that is for any* $x > 0$.

$$\lim_{t\to\infty} \frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)} = x^{-\frac{1}{\gamma_1}} \text{ and } \lim_{t\to\infty} \frac{\overline{\mathbf{G}}(tx)}{\overline{\mathbf{G}}(t)} = x^{-\frac{1}{\gamma_2}},$$

being characterized by their heavy tails, these distributions play a prominent role in extreme value theory. After making use of the proposition B.1.10 in De Haan et al., 2006 for regularly varying functions $\overline{\mathbf{F}}$ *and* $\overline{\mathbf{G}}$*, we may show that both* $\overline{F}$ *and* $\overline{G}$

are regularly at infinity as well, with respective indices γ_2 *and* $\gamma := \frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)}$*. For any* $x > 0$,

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\frac{1}{\gamma}} \text{ and } \lim_{t\to\infty} \frac{\overline{G}(tx)}{\overline{G}(t)} = x^{-\frac{1}{\gamma_2}},$$

The work of Gardes and Stupfler, 2015 addressed the estimation of extreme value index γ_1 *under random truncation. They used the definition of* γ *to* *derive the following consistent estimator:*

$$\widehat{\gamma}_1(k,k^{'}) := \frac{\widehat{\gamma}(k)\widehat{\gamma}_2(k^{'})}{\widehat{\gamma}_2(k^{'}) - \widehat{\gamma}(k)},$$

Where $\widehat{\gamma}$ *and* $\widehat{\gamma}_2$ *are the well-known Hill estimators of* γ *and* γ_2 *defined by:*

$$\widehat{\gamma}(k) := 1/k \sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{X_{n-k,n}} \quad \textit{and} \quad \widehat{\gamma}_2(k^{'}) := 1/k^{'} \sum_{i=1}^{k^{'}} \log \frac{Y_{n-i+1,n}}{Y_{n-k,n}},$$

with $X_{1,n} \leq ... \leq X_{n,n}$ *and* $Y_{1,n} \leq ... \leq Y_{n,n}$ *being the order statistics pertaining to the samples* $(X_{1,n}, ..., X_{n,n})$ *and* $(Y_{1,n}, ..., Y_{n,n})$ *respectively. The two sequences* $k = k_n$ *and* $k^{'} = k_n'$ *of integer rv's, respectively represent the numbers of top observations from truncated and truncation data satisfying the following conditions:*

$$1 < k, k^{'} < n, \ \min(k, k^{'}) \rightarrow \infty \quad \textit{and} \quad \max(k/n, k^{'}/n) \rightarrow 0 \textit{ as } n \rightarrow 0.$$

by exploiting the work of Gardes and Stupfler, 2015 and starting from the first-order condition of regular variation Benchaira et al., 2015 construct a new estimator with the situation $k = k^{'}$ *for the shape parameter of a right-truncated heavy-tailed distribution. and they prove the its asymptotic normality by use the tail Lynden-Bell process for which a weighted Gaussian approximation is provided:*

$$\widehat{\gamma}_1(k) := \widehat{\gamma}_1 := 1/k \frac{\sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{X_{n-k,n}} \sum_{i=1}^{k} \frac{Y_{n-i+1,n}}{Y_{n-k,n}}}{\sum_{i=1}^{k} \log \frac{X_{n-k,n}Y_{n-i+1,n}}{Y_{n-k,n}X_{n-i+1,n}}}.$$

Proof. See Benchaira et al., 2015.

Part II

Main results

5 Estimating the mean of a heavy tailed distribution under random truncation

The main aim of this chapter is to introduce an alternative estimator for the mean of heavy-tailed distribution when it comes to the right truncated and study its asymptotic normality this work inspired by L. Peng's work in the case of completed data. A simulation study is executed to evaluate the finite sample behavior on the proposed estimator.

5.1 Introduction

Let $(\mathbf{X}_1, ..., \mathbf{X}_N)$ *be independent copies of a non-negative random variable (rv)* $\mathbf{X}$ *with cumulative distribution (cdf)* $\mathbf{F}$ *, defined over some probability space* $(\Omega, \mathcal{A}, \mathcal{P})$*, suppose that* $\mathbf{X}$ *is right truncated by sequences of independent copies* $(\mathbf{Y}_1, ..., \mathbf{Y}_N)$ *of (rv)* $\mathbf{Y}$ *with cdf* $\mathbf{G}$*, throughout this chapter, we assume that* $\mathbf{F}$ *and* $\mathbf{G}$ *are heavy-tailed in other words that* $\overline{\mathbf{F}} = 1-\mathbf{F}$ *and* $\overline{\mathbf{G}} = 1-\mathbf{G}$ *are regularly varying (*$\mathcal{RV}$*) at infinity with respective negative indices* $-1/\gamma_1$ *and* $-1/\gamma_2$*; we will use the notation:* $\overline{\mathbf{F}} \in \mathcal{RV}(-1/\gamma_1)$ *and* $\overline{\mathbf{G}} \in \mathcal{RV}(-1/\gamma_2)$ *that is for any* $x > 0$.

$$\lim_{t\to\infty} \frac{\overline{\mathrm{F}}(tx)}{\overline{\mathrm{F}}(t)} = x^{-\frac{1}{\gamma_1}} \quad \textit{and} \quad \lim_{t\to\infty} \frac{\overline{\mathrm{G}}(tx)}{\overline{\mathrm{G}}(t)} = x^{-\frac{1}{\gamma_2}}, \tag{5.1}$$

The statistical literature on such problems of extremes events is very extensive, one of those problems is for the estimation of the mean $\mathbf{E}(X)$*, this problem was already treated by Peng, 2001 in the case of complete data, nevertheless in numerous survival practical applications, it happens that one is not able to observe a subject entire lifetime. the subject may leave the study may survive to the closing data, or may enter the study at some time after its lifetime has* *started, the most current forms of such incomplete data are censorship and truncation. As we mention our aim is to propose an asymptotically normal estimator for the mean of* X*:*

$$\mu = \mathbf{E}(X) = \int_0^\infty \overline{\mathrm{F}}(x)dx. \tag{5.2}$$

Whose existence requires that $\gamma_1 < 1$*, The sample mean for censored data is obtained and equal to:*

$$\widetilde{\mu}_n = \sum_{i=2}^{n} \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}} Z_{i,n}. \tag{5.3}$$

the asymptotic normality of $\widetilde{\mu}_n$ *is established by Stute, 1995. The model studied here is based on the random right truncated (*$\mathcal{RRT}$*) data, in the sense that the rv of interest* $\mathbf{X}_i$ *and the truncated rv* $\mathbf{Y}_i$ *are observable only when* $\mathbf{X}_i \leq \mathbf{Y}_i$*, whereas nothing is observed if* $\mathbf{X}_i > \mathbf{Y}_i$*.We denote* (X_i, Y_i)*,*$i = 1; n$ *to be observed data as copies of a couple of rv's* (X, Y) *corresponding to the truncated sample* $(\mathbf{X}_i, \mathbf{Y}_i)_{1 \leq i \leq N}$*, where* $n = n_N$ *is a sequence of discrete rv's by the weak law of large numbers, we have*

$$\frac{n}{N} \longrightarrow p = \mathbf{P}(\mathbf{X} \leq \mathbf{Y}) \quad as \ \ N \rightarrow \infty.$$

we shall assume that $p > 0$*, otherwise nothing will be observed. the joint* $\mathbf{P}$*-distribution of on observed* (X, Y) *is given by:*

$$\begin{aligned} H(x, y) &= \mathbf{P}(X \leq x, Y \leq y) \\ &= \mathbf{P}(\mathbf{X} \leq x, \mathbf{Y} \leq y \mid \mathbf{X} \leq \mathbf{Y}) = p^{-1} \int_0^y \mathrm{F}(\min(x, z))d\mathrm{G}(z), \end{aligned}$$

The marginal distributions of the rv's X *and* Y *respectively denoted by* F

and G are defined by:

$$F(x) = p^{-1} \int_0^x \overline{\mathrm{G}}(z) d\mathrm{F}(z) \quad \textit{and} \;\; G(y) = p^{-1} \int_0^y \mathrm{F}(z) d\mathrm{G}(z),$$
$$\overline{F}(x) = -p^{-1} \int_x^\infty \overline{\mathrm{G}}(z) d\overline{\mathrm{F}}(z) \;\; \textit{and} \;\; \overline{G}(y) = -p^{-1} \int_y^\infty \mathrm{F}(z) d\overline{\mathrm{G}}(z).$$

For randomly truncated data; the truncation product-limit estimate is the maximum likelihood estimate (MLE) for non-parametric models the well-known non-parametric estimator of F in $\mathcal{RRT}$ model, proposed by Lynden-Bell, 1971 :

$$\mathbf{F}_n^{(\mathbf{LB})}(x) = \prod_{i: X_i > x} \exp(1 - \frac{1}{n C_n(X_i)}). \tag{5.4}$$

Where $C_n(x) = n^{-1} \sum_{i=1}^n \mathbf{1}(X_i \leq x \leq Y_i)$ the empirical counterparts of $C(z) = P(X \leq z \leq Y)$.Since F and G are heavy-tailed their right endpoints are infinite and thus are equal.

As we mentioned this problem has been studied by Peng, 2001 in the case of sets of complete data from heavy-tailed distributions with a range of $\gamma_1 \in (1/2, 1)$ we restrict ourselves on the case where γ_1 belongs to the following range:

$$\mathcal{R} = \left\{ \gamma_1 \, , \gamma_2 > 0 : \frac{\gamma_2}{1 + 2\gamma_2} < \gamma_1 < 1 \right\}, \tag{5.5}$$

To ensure that the mean is finite and since we have applied both conditions of Stute and Wang, 2008 paper:

$$I_1 = \int_1^\infty \frac{\varphi^2(x)}{\mathrm{G}(x)} d\mathrm{F}(x) \qquad I_2 = \int_1^\infty \frac{d\mathrm{F}(x)}{\mathrm{G}(x)}, \tag{5.6}$$

We find those conditions may be infinite when we deal with heavy-tailed distributions. Assumed that both of X and Y are $Pareto(\gamma_1)$ and $Pareto(\gamma_2)$

respectively :

$$1-\mathrm{F}(x)=\overline{\mathrm{F}}(x)=x^{-\frac{1}{\gamma_1}} \quad 1-\mathrm{G}(x)=\overline{\mathrm{G}}(x)=x^{-\frac{1}{\gamma_2}}. \ \textit{with} \ \ \gamma_1>0\,,\gamma_2>0 \ \ \textit{and} \ \ x\geq 1.$$

we figure out that the central limit theorem (CTL) established by Stute and Wang, 2008 cannot be applied in the previous range when $I_1 = I_2 = \infty$. It is worth to mention that in the case of non truncation we have $\gamma_1 = \gamma$ and $\gamma_2 = \infty$ so $\mathcal{R}$ abbreviate to Peng's range. To define our new estimator we introduce an integer sequences $k = k_n$ representing a fraction of extreme order statistics satisfying the following conditions:

$$1<k<n,\ k\longrightarrow\infty \text{ and } k/n\longrightarrow 0 \text{ as } n\longrightarrow\infty. \tag{5.7}$$

So by decomposing μ as the sum of two terms

$$\begin{aligned}\mu &= \int_0^t \overline{\mathrm{F}}(x)dx+\int_t^\infty \overline{\mathrm{F}}(x)dx \\ &= \mu_1+\mu_2.\end{aligned} \tag{5.8}$$

Then we can estimate $\mu_i, i=\overline{1,2}$ separately, after integration μ_1 by parts and after changing variables in μ_2 we may write:

$$\mu_1 = t\overline{\mathrm{F}}(t)+\int_0^t xd\mathrm{F}(x) \qquad \textit{and} \qquad \mu_2 = t\overline{\mathrm{F}}(t)\int_1^\infty \frac{\overline{\mathrm{F}}(tx)}{\overline{\mathrm{F}}(t)}dx,$$

By replacing t by $X_{n-k,n}$ where $X_{1,n} < ... < X_{n,n}$ denote the order statistics pertaining to $X_1, ..., X_n$; and $\mathbf{F}$ by $\mathbf{F}_n^{(\mathbf{LB})}$ we get that:

$$\widehat{\mu}_1 = X_{n-k,n}\overline{\mathrm{F}_n}^{(\mathbf{LB})}(X_{n-k,n})+\int_0^{X_{n-k,n}} xd\mathrm{F}_n^{(\mathbf{LB})}(x),$$

Hence from Woodroofe, 1985 we may write:

$$\widehat{\mu_1} = X_{n-k,n}\overline{\mathbf{F}_n}^{(\mathbf{LB})}(X_{n-k,n}) + \frac{1}{n}\sum_{i=1}^{n-k}\frac{\mathbf{F}_n^{(\mathbf{LB})}(X_{i,n})}{C_n(X_{i,n})}X_{i,n}. \tag{5.9}$$

Back to μ_2 *building on the Karamata theorem (De Haan et al., 2006, page 363) we may write :*

$$\mu_2 \sim \frac{\gamma_1}{1-\gamma_1} t\overline{\mathbf{F}}(t) \ \textit{as } n \longrightarrow \infty \qquad 0 < \gamma_1 < 1, \tag{5.10}$$

Notice to estimate (5.10) it is based on estimator of tail index γ_1*, in view of the history of the estimation of* γ_1*.Gardes and Stupfler, 2015 introduced an estimator of* γ_1 *under random truncation. Benchaira et al., 2015 established the asymptotic normality of this estimator under the tail dependence and the second order conditions of regular variation . throughout this work we use the estimation of Benchaira et al., 2015 . So that yield us to an estimator to* μ_2 *:*

$$\widehat{\mu_2} = \frac{\widehat{\gamma_1}}{1-\widehat{\gamma_1}} X_{n-k,n}\overline{\mathbf{F}_n}^{(\mathbf{LB})}(X_{n-k,n}), \tag{5.11}$$

Finally with (5.9) and (5.11). we build our estimator $\widehat{\mu}$ *for the mean (5.2) as follow:*

$$\hat{\mu} = X_{n-k,n}\ \overline{\mathbf{F}}_n(X_{n-k,n})\frac{1}{1-\hat{\gamma}_1} + \frac{1}{n}\sum_{i=1}^{n-k}\frac{\mathbf{F}_n^{LB}(X_{i,n})}{C_n(X_{i,n})}X_{i,n}.$$

The rest of this chapter is organized as follows. In the second section, we state our main result. This is followed by a simulation study of our proposed estimator where we discuss its behavior with a finite sample.

5.2 The assumptions and the main results

In extreme value analysis and in the second-order frame work(see,.e.g De Haan et al., 2006), weak approximation are achieved. Consequently, it seems quite natural to suppose that df's $\mathbf{F}$ *and* $\mathbf{G}$ *satisfy the well-known second-order condition of regular variation we express in terms of the tail quantile functions. That is we assume that for* $x > 0.$ *we have*

$$\lim_{t\to\infty} \frac{U_{\mathbf{F}}(tx)/U_{\mathbf{F}}(t) - x^{\gamma_1}}{\mathbf{A}_{\mathbf{F}}(t)} = x^{\gamma_1}\frac{x^{\tau_1}-1}{\tau_1}, \tag{5.12}$$

and

$$\lim_{t\to\infty} \frac{U_{\mathbf{G}}(tx)/U_{\mathbf{G}}(t) - x^{\gamma_2}}{\mathbf{A}_{\mathbf{G}}(t)} = x^{\gamma_2}\frac{x^{\tau_2}-1}{\tau_2}, \tag{5.13}$$

where $\tau_1,\tau_2 < 0$ *are the second-order parameters and* $\mathbf{A}_{\mathbf{F}}, \mathbf{A}_{\mathbf{G}}$ *are functions tending to zero and not changing signs near infinity with regularly varying absolute values at infinity with indices* τ_1,τ_2 *respectively.*

▶ **Theorem 5.1.** Assume that (5.12 and 5.13) hold and $\sqrt{k}\mathbf{A}_\circ(n/k) = O(1)$ for

$\gamma_2/(1+2\gamma_2) < \gamma_1 < 1$. Let $k = k_n$ denote an intermediate integer sequences satisfying (5.7). then $\hat{\mu} \to \mu$ in probability:

$$\begin{aligned}\frac{\sqrt{k}(\widehat{\mu}-\mu)}{\overline{\mathbf{F}}(X_{n-k,n})X_{n-k,n}} &= \mathbf{c}_1\mathbf{W}(1)\\ &+ \int_0^1 \left\{\mathbf{c}_2 s^{-\frac{2\gamma_1}{\gamma}+\frac{\gamma}{\gamma_2}+1} + \mathbf{c}_3 s^{-\gamma_1+\frac{\gamma}{\gamma_2}+1} + \mathbf{c}_4\log(s) + \mathbf{c}_5\right\} s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds\\ &+ \frac{(\gamma_1+\tau_1-1)(1-\gamma_1)+(1-\tau_1)}{(1-\tau_1)(\gamma_1+\tau_1-1)(1-\gamma_1)}\sqrt{k}\mathbf{A}_\circ(n/k).\end{aligned}$$

◀

▶ **Corollary 5.2.** Under the assumptions of Theorem 5.1 we suppose that $\sqrt{k}\mathbf{A}_{\circ}(n/k) \longrightarrow \lambda$

$$\frac{\sqrt{k}(\widehat{\mu}-\mu)}{\overline{\mathrm{F}}(X_{n-k,n})X_{n-k,n}} \longrightarrow \mathcal{N}\left(\lambda\frac{(\gamma_1+\tau_1-1)(1-\gamma_1)+(1-\tau_1)}{(1-\tau_1)(\gamma_1+\tau_1-1)(1-\gamma_1)}, \sigma^2\right), \text{ as } n \longrightarrow \infty,$$

Where

$$\begin{aligned}\sigma^2 :=& \frac{p(1-p)\left[p(1-p)+2\gamma_1^2\right]}{(1-\gamma_1)^2} + \frac{p^3\gamma_1}{1-\gamma_1} + \frac{2p^2(1-p)}{(1-\gamma_1)(-\gamma_1+2)} \\ &+ \frac{-2p^4}{(-2+p)(-4+3p)} + \frac{3p^5\gamma_1}{(-2+p)(-2+\gamma_1 p+3p)} + \frac{-2\gamma_1 p^3(1-p)}{(-2+p)(-\gamma_1+2)} \\ &+ 3p^5\gamma_1^2(\frac{p}{2}-\frac{1}{4-p})^2 - 2p^3\gamma_1^2(1-p)\frac{3p-2}{6}(\frac{p}{1+p})^2 \\ &+\frac{p^2\gamma_1(p-1)(1-\gamma_1)-p^2\gamma_1^3}{(-1+p)(-2+p)(1-\gamma_1)}\left[\frac{\gamma_1(-p^3+4-6p)+p^2(\gamma_1-2)+2}{(-1-p)+\gamma_1(-p-2)}\right] \\ &+ \frac{1-2p}{p^2} + \frac{-2p^2(1-p)^2(1-\gamma_1)+\gamma_1^2 p}{(1-\gamma_1)(\gamma_1+2)(-\gamma_1+p+1)^2} \\ &+ \frac{2p^2(1-p)(1-\gamma_1)+\gamma_1^2 p}{(1-\gamma_1)^2}\left(\left(\frac{p}{p^2-1}\right)^2+\left(\frac{1}{1-p}\right)^2\right).\end{aligned}$$

and

$$p = \frac{\gamma_2}{\gamma_1+\gamma_2}.$$

◀

5.3 Simulation study

The main purpose of this section is to study the execution of our new estimator $\widehat{\mu}$ for that we generate the data as follows:

- *The interest and the truncated variable:*

we generate two sets of truncated and truncation data both pulled for the first hand from Fréchet model:

$$\overline{\mathrm{F}}(x) = 1 - \exp(-x^{\frac{1}{\gamma_1}}),\ \ \overline{\mathrm{G}}(x) = 1 - \exp(-x^{\frac{1}{\gamma_2}}) \qquad x \geq 0.$$

and the other hand from Burr model:

$$\overline{\mathrm{F}}(x) = (1 + x^{\frac{1}{\delta}})^{-\frac{\delta}{\gamma_1}},\ \ \overline{\mathrm{G}}(x) = (1 + x^{\frac{1}{\delta}})^{-\frac{\delta}{\gamma_2}} \quad x \geq 0 \ \ \textit{and}\ \delta, \gamma_1, \gamma_2 > 0.$$

- *The observed data :*

for the proportion of observed data is equal to $p = \gamma_2/\gamma_1 + \gamma_2$ *we take* $p = 70\%, 80\%$ *and* 90% *we fix* $\delta = 1/4$ *and choose the values* $0.6, 0.7$ *and* 0.8 *for* γ_1. *For each couple* (γ_1, p)*; we solve the equation* $p = \gamma_2/\gamma_1 + \gamma_2$ *to get the pertaining* γ_2*-value.*

- *We vary the common size N of both samples* $(\mathbf{X}_1, ..., \mathbf{X}_N)$ *and* $(\mathbf{Y}_1, ..., \mathbf{Y}_N)$.
- *We apply the algorithm of Reiss et al., 2007 page* 137. *to select the optimal numbers of upper order statistics* (k^*)*used in the computation of* $\hat{\gamma}_1$.

The performance of this new estimator named by $\hat{\mu}$ *is evaluated in terms of absolute bias (abs bias) root mean squared error (RMSE) which are summarized in tables for Burr model in Tables: 5.1 for* $\gamma_1 = 0.6$, *5.2 for* $\gamma_1 = 0.7$, *5.3 for* $\gamma_1 = 0.8$ *and for Fréchet models Tables: 5.4 for* $\gamma_1 = 0.6$, *5.5 for* $\gamma_1 = 0.7$, *5.6 for* $\gamma_1 = 0.8$.

After the inspection of all tables and as expected the sample size influences the estimation in the sense that the large N gets the better the estimation is.

It is noticeable that the estimation accuracy of estimator decreases when the truncation percentage increase and it is quite expected.

Moreover the estimator performs best for the larger value of the tail index larger than 0.5 *especially when truncation proportion is high.*

γ_1= 0.6 $\longrightarrow$ $\mu = 2.371$															
	$p = 0.7$					$p = 0.8$					$p = 0.9$				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n
300	0.002	0.130	27	2.374	198	0.008	0.180	10	2.380	244	0.005	0.040	4	2.406	268
400	0.069	0.858	31	2.440	278	0.008	0.119	16	2.379	318	0.006	0.028	7	2.406	361
500	0.072	0.257	39	2.300	355	0.001	0.174	27	2.372	399	0.003	0.067	8	2.374	445
1000	0.001	0.048	40	2.372	681	0.001	0.106	25	2.372	811	0.003	0.097	12	2.374	886

Table 5.1: Biases and RMSE's of the mean estimator based on samples of Burr models with $\gamma_1 = 0.6$

γ_1= 0.7 $\longrightarrow$ $\mu = 3.218$															
	$p = 0.7$					$p = 0.8$					$p = 0.9$				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	bias	RMSE	k^*	$\hat{\mu}$	n	bias	RMSE	k^*	$\hat{\mu}$	n
300	0.016	0.634	25	3.234	215	0.021	0.178	18	3.239	246	0.005	0.028	19	3.223	268
400	0.008	0.067	34	3.227	290	0.002	0.306	23	3.221	319	0.000	0.134	21	3.218	368
500	0.008	0.063	58	3.226	3362	0.002	0.367	39	3.220	403	0.008	0.246	25	3.226	458
1000	0.004	0.023	88	3.222	701	0.001	0.193	52	3.219	788	0.002	0.049	37	3.220	896

Table 5.2: Biases and RMSE's of the mean estimator based samples of Burr models with $\gamma_1 = 0.7$

γ_1= 0.8 $\longrightarrow$ $\mu = 4.896$															
	$p = 0.7$					$p = 0.8$					$p = 0.9$				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n
300	0.000	0.152	73	4.896	207	0.106	0.613	55	5.002	239	0.094	0.962	67	4.990	275
400	0.029	0.070	75.	4.925	278	0.014	0.446	14	4.910	315	0.058	0.240	86	4.954	359
500	0.065	0.631	147	4.961	348	0.001	0.321	146	4.897	404	0.029	0.171	67	4.925	451
1000	0.013	0.302	228	4.919	697	0.030	0.039	173	4.926	810	0.006	0.041	187	4.902	894

Table 5.3: Biases and RMSE's of the mean estimator based on samples of Burr models with $\gamma_1 = 0.8$

γ_1= 0.6 $\longrightarrow$ μ = 2.218															
	p = 0.7					p = 0.8					p = 0.9				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n
300	0.155	0.537	28	2.373	170	0.259	0.263	17	2.475	178	0.010	0.084	5	2.228	180
400	0.153	0.186	25	2.371	217	0.031	0.598	40	2.249	241	0.009	0.185	11	2.218	231
500	0.004	0.065	32	2.222	284	0.066	0.222	33	2.284	293	0.004	0.052	19	2.222	314
1000	0.002	0.010	43	2.220	568	0.074	0.076	31	2.307	569	0.008	0.106	23	2.227	594

Table 5.4: Biases and RMSE's of the mean estimator based on samples of Frechét models with $\gamma_1 = 0.6$

γ_1= 0.7 $\longrightarrow$ μ = 2.992															
	p = 0.7					p = 0.8					p = 0.9				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n
300	0.085	0.213	23	3.076	168	0.031	0.171	30	3.022	169	0.001	0.213	22	2.993	193
400	0.080	0.356	57	3.072	227	0.000	0.063	26	2.992	250	0.082	0.206	25	3.074	225
500	0.025	0.365	49	3.016	278	0.016	0.352	44	3.007	274	0.086	0.189	29	3.078	306
1000	0.020	0.385	58	3.011	564	0.001	0.122	48	2.993	598	0.000	0.257	40	2.992	584

Table 5.5: Biases and RMSE's of the mean estimator based on samples of Frechét models with $\gamma_1 = 0.7$

γ_1= 0.8 $\longrightarrow$ μ = 4.591															
	p = 0.7					p = 0.8					p = 0.9				
N	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n	abs bias	RMSE	k^*	$\hat{\mu}$	n
300	0.084	0.720	15	4.675	164	0.267	0.282	12	4.857	173	0.222	0.301	37	4.813	172
400	0.185	0.604	42	4.776	225	0.131	0.147	29	4.722	222	0.128	0.283	72	4.719	256
500	0.001	0.037	52	4.591	297	0.044	0.045	41	4.635	306	0.057	0.576	70	4.648	302
1000	0.063	0.674	109	4.654	540	0.011	0.331	68	4.690	597	0.001	0.382	133	4.592	604

Table 5.6: Biases and RMSE's of the mean estimator based on samples of Frechét models with $\gamma_1 = 0.8$

5.4 Proofs

5.4.1 Proof of Theorem 2.1

We begin by seting $U_i = \overline{F}(X_i)$ and define the corresponding uniform tail process by $\alpha_n(s) = \sqrt{k}(U_n(s) - s)$, for $0 \leq s \leq 1$ where $U_n(s) = 1/k \sum_{i=1}^{n} \mathbf{1}(\mathbf{U}_i \leq k\frac{s}{n})$. The weighted weak approximation to $\alpha_n(s)$ given in terms of either a sequence of wiener processes (see, eg., Einmahl, 1992 and Drees et al., 2006) or a single Wienner process as in proposition 3.1 of Einmahl, 1992, will be very crucial to our proof procedure.

In the sequel, we use the latter representation which says that: there exists a Wiener process $\mathbf{W}$, such that for every $0 \leq \eta \leq 1$

$$\sup_{0<s\leq 1} \mid \alpha_n(s) - \mathbf{W}(s) \mid \to \mathbf{0}, \text{ as } n \to \infty \tag{5.14}$$

Observe that $\widehat{\mu} - \mu = (\widehat{\mu}_1 - \mu_1) + (\widehat{\mu}_2 - \mu_2)$ and starting by:

$$\widehat{\mu}_1 - \mu_1 = \int_0^{X_{n-k;n}} \overline{\mathbf{F}_n}(x)dx - \int_0^t \overline{\mathbf{F}}(x)dx.$$

we consider the following decomposition:

$$\widehat{\mu}_1 - \mu_1 = T_{n_1}(x) + T_{n_2}(x)$$

Where:

$$T_{n_1}(x) = \int_0^{X_{n-k;n}} \Big(\overline{\mathbf{F}}_n(x) - \overline{\mathbf{F}}(x)\Big)dx$$
$$T_{n_2}(x) = \int_{X_{n-k;n}}^{t} \overline{\mathbf{F}}(x)dx.$$

it follows after changing variables that:

$$T_{n_1}(x) = X_{n-k,n} \int_0^1 \frac{\overline{\mathbf{F}}(a_k x)}{\overline{\mathbf{F}}(a_k x)} \overline{\mathbf{F}}_n(xX_{n-k,n}) - \overline{\mathbf{F}}(xX_{n-k,n}) dx$$

$$T_{n_2}(x) = -X_{n-k,n} \int_1^{\frac{t}{X_{n-k,n}}} \overline{\mathbf{F}}(xX_{n-k,n}) dx$$

In order to established the result of theorem we apply the results of Benchaira et al., 2016, we have :

$$\sqrt{k}\frac{\overline{\mathbf{F}}_n(xX_{n-k,n}) - \overline{\mathbf{F}}(xX_{n-k,n})}{\overline{\mathbf{F}}(a_k x)} = x^{\frac{1}{\gamma}} \frac{\gamma}{\gamma_1} W(x^{-\frac{1}{\gamma_1}}) + \frac{\gamma}{\gamma_1 + \gamma_2} x^{\frac{1}{\gamma_1}} \int_0^1 s^{-\frac{\gamma}{\gamma_2}-1} \mathbf{W}(x^{-\frac{1}{\gamma_1}} s) ds,$$

After some elementary but tedious manipulations of integral calculus (change of variables and integration by parts) and by making use of the uniform inequality of the second-order regularly varying functions $\overline{\mathbf{F}}$ *,*

to $T_{n_1}(x)$ *becomes :*

$$\sqrt{k}\frac{T_{n_1}(x)}{X_{n-k,n}\overline{\mathbf{F}}(a_k)} = \int_0^1 (-\gamma s^{-\frac{2\gamma_1}{\gamma}} + \frac{\gamma\gamma_1}{(\gamma_1+\gamma_2)(\gamma_1+1)} s^{-\frac{\gamma}{\gamma_2}-1} + \frac{\gamma\gamma_1}{(\gamma_1+\gamma_2)(\gamma_1+1)} s^{-\gamma_1}) \mathbf{W}(s) ds + o_{\mathbf{p}}(1) \tag{5.15}$$

Next we move to$T_{n_2}(x)$ *which we may write it as follow after changing variables :*

$$\frac{\sqrt{k}T_{n_2}(x)}{X_{n-k,n}\overline{\mathbf{F}}(X_{n-k,n})} = \int_1^{\frac{t}{X_{n-k,n}}} \sqrt{k}\frac{\overline{\mathbf{F}}(xX_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})} - x^{-\frac{1}{\gamma_1}} dx + \int_1^{\frac{t}{X_{n-k,n}}} x^{-\frac{1}{\gamma_1}} dx$$
$$= \mathbf{I}_1 + \mathbf{I}_2$$

For $\mathbf{I}_1$ we apply the results of Benchaira et al., 2016

$$\sqrt{k}\frac{\overline{\mathbf{F}}(xX_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})} - x^{-\frac{1}{\gamma_1}} = x^{-\frac{1}{\gamma_1}}\frac{x^{-\frac{\tau_1}{\gamma_1}}-1}{\gamma_1\tau_1}\sqrt{k}\mathbf{A}_\circ(n/k) + o_p(x^{-\frac{1}{\gamma_1}+(1-\eta)/\gamma\pm\varepsilon})$$

This implies, almost surely, that

$$\int_1^{\frac{t}{X_{n-k,n}}} \sqrt{k}\frac{\overline{\mathbf{F}}(xX_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})} - x^{-\frac{1}{\gamma_1}}dx = \int_1^{\frac{t}{X_{n-k,n}}} x^{-\frac{1}{\gamma_1}}\frac{x^{-\frac{\tau_1}{\gamma_1}}-1}{\gamma_1\tau_1}\sqrt{k}\mathbf{A}_\circ(n/k)dx,$$

Which is equal after simple calcul and by using the mean value theorem we get $\mathbf{I}_1 = o_{\mathbf{p}}(1)$, for the second step by similar argument and using the fact that from Theorem 2.1 of Benchaira et al., 2015; we have $\sqrt{k}\left(\frac{X_{n-k,n}}{t}-1\right) - \gamma\mathbf{W}(1) = o_{\mathbf{P}}(1)$ we get $\mathbf{I}_2 = -\gamma\mathbf{W}(1) + o_{\mathbf{p}}(1)$, that yield to :

$$\frac{\sqrt{k}T_{n_2}(x)}{X_{n-k,n}\overline{\mathbf{F}}(X_{n-k,n})} = -\gamma\mathbf{W}(1) + o_{\mathbf{p}}(1) \tag{5.16}$$

The two approximation 5.15 and 5.16 together give:

$$\begin{aligned}\sqrt{k}\frac{\widehat{\mu}_1-\mu_1}{X_{n-k,n}\overline{\mathbf{F}}(X_{n-k,n})} &= \int_0^1 (-\gamma s^{-\frac{2\gamma_1}{\gamma}} + \frac{\gamma\gamma_1}{(\gamma_1+\gamma_2)(\gamma_1+1)}s^{-\frac{\gamma}{\gamma_2}-1} \\ &+ \frac{\gamma\gamma_1}{(\gamma_1+\gamma_2)(\gamma_1+1)}s^{-\gamma_1})\mathbf{W}(s)ds \\ &- \gamma\mathbf{W}(1) + o_{\mathbf{p}}(1)\end{aligned} \tag{5.17}$$

Let us now treat term $\frac{\sqrt{k}(\widehat{\mu}_2-\mu_2)}{t\overline{\mathbf{F}}(t)}$ Consider the following forms of μ_2 and $\widehat{\mu}_2$:

$$\widehat{\mu}_2 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1}X_{n-k,n}\overline{\mathbf{F}}_n(X_{n-k,n}) \quad \text{and } \mu_2 = \int_t^\infty \overline{\mathbf{F}}(x)dx$$

$$\widehat{\mu}_2 - \mu_2 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} X_{n-k,n}\overline{\mathbf{F}}_n(X_{n-k,n}) - \int_t^\infty \overline{\mathbf{F}}(x)dx$$

After changing variables we can obtain:

$$\begin{aligned}\mu_2 &= \int_1^\infty t\overline{\mathbf{F}}(tx)dx \\ &= t\overline{\mathbf{F}}(t)\int_1^\infty \frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)}dx\end{aligned}$$

and

$$\widehat{\mu}_2 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} X_{n-k,n}\overline{\mathbf{F}}_n(X_{n-k,n})\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}$$

so the previous equation leads to

$$\widehat{\mu}_2 - \mu_2 = \frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} X_{n-k,n}\overline{\mathbf{F}}_n(X_{n-k,n})\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})} - t\overline{\mathbf{F}}(t)\int_1^\infty \frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)}dx$$

if we devise this equation by $t\overline{\mathbf{F}}(t)$ we can get:

$$\frac{\sqrt{k}\widehat{\mu}_2 - \mu_2}{t\overline{\mathbf{F}}(t)} = \sqrt{k}\frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} X_{n-k,n}\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{t\overline{\mathbf{F}}(t)}\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})} - \sqrt{k}\int_1^\infty \frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)}dx$$

So after adding and Subtract some terms we can decompose $\frac{\sqrt{k}(\widehat{\mu}_2-\mu_2)}{t\overline{\mathbf{F}}(t)}$ into the sum of:

$$\begin{aligned}\mathbf{I}_1 &:= \sqrt{k}\frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1}\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{\overline{\mathbf{F}}(t)}\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}\left[\frac{X_{n-k,n}}{t} - 1\right] \\ \mathbf{I}_2 &:= \sqrt{k}\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{\overline{\mathbf{F}}(t)}\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}\left[\frac{\widehat{\gamma}_1}{1-\widehat{\gamma}_1} - \frac{\gamma_1}{1-\gamma_1}\right]\end{aligned}$$

$$\mathbf{I}_3 := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(t)}\left[\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}-1\right]$$

$$\mathbf{I}_4 := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\left[\frac{\overline{\mathbf{F}}(X_{n-k,n})}{\overline{\mathbf{F}}(t)}-\left(\frac{X_{n-k,n}}{t}\right)^{-\frac{1}{\gamma_1}}\right]$$

$$\mathbf{I}_5 := \sqrt{k}\frac{\gamma_1}{1-\gamma_1}\left[\left(\frac{X_{n-k,n}}{t}\right)^{-\frac{1}{\gamma_1}}-1\right]$$

$$\mathbf{I}_6 := \sqrt{k}\left[\frac{\gamma_1}{1-\gamma_1}-\int_1^\infty\frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)}dx\right]$$

For, we have $\mathbf{I}_1$, $\widehat{\gamma}_1 \longrightarrow \gamma_1$ *and* $X_{n-k,n}/t \longrightarrow 1$. *Since* $\overline{\mathbf{F}}$ *is regular variation we obtain* $\overline{\mathbf{F}}(X_{n-k,n}) = (1+o_{\mathbf{P}}(1))\overline{\mathbf{F}}(t)$. *From remark 4.1 ofBenchaira et al., 2015, we have* $\overline{\mathbf{F}}_n(X_{n-k,n})/\overline{\mathbf{F}}(X_{n-k,n}) \longrightarrow 1$.*So,*

$$\sqrt{k}\mathbf{I}_1 = (1+o_{\mathbf{P}}(1))\sqrt{k}\left(\frac{X_{n-k,n}}{t}-1\right)$$

From Theorem 2.1 of Benchaira et al., 2015; we have

$$\sqrt{k}\left(\frac{X_{n-k,n}}{t}-1\right)-\gamma\mathbf{W}(1) = o_{\mathbf{P}}(1).$$

Then

$$\sqrt{k}\mathbf{I}_1 = (1+o_{\mathbf{P}}(1))\frac{\gamma_1\gamma}{1-\gamma_1}\mathbf{W}(1). \tag{5.18}$$

For $\mathbf{I}_2$, *by using a similar way of* $\mathbf{I}_1$, *we prove that:*

$$\sqrt{k}\mathbf{I}_2 = (1+o_{\mathbf{P}}(1))\frac{1}{(1-\gamma_1)^2}\sqrt{k}(\widehat{\gamma}_1-\gamma_1). \tag{5.19}$$

From Theorem 3.1 of Benchaira et al., 2016, we have

$$\sqrt{k}(\widehat{\gamma}_1-\gamma_1)=\frac{\sqrt{k}\mathbf{A}_\circ(n/k)}{1-\tau_1}-\gamma\mathbf{W}(1)+\frac{\gamma}{\gamma_1+\gamma_2}\int_0^1(\gamma_2-\gamma_1-\gamma\log s)s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds$$
$$+o_{\mathbf{P}}(1)$$

For $\mathbf{I}_3$, *we have*

$$\sqrt{k}\mathbf{I}_3=(1+o_{\mathbf{P}}(1))\frac{\gamma_1\gamma}{1-\gamma_1}\sqrt{k}\left(\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}-1\right)$$

From Theorem 4.1 of Benchaira et al., 2015, we have

$$\sqrt{k}\left(\frac{\overline{\mathbf{F}}_n(X_{n-k,n})}{\overline{\mathbf{F}}(X_{n-k,n})}-1\right)=\frac{\gamma_2}{\gamma_1+\gamma_2}\mathbf{W}(1)+\frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2}\int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds+o_{\mathbf{P}}(1).$$

So,

$$\sqrt{k}\mathbf{I}_3=(1+o_{\mathbf{P}}(1))\frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)}\mathbf{W}(1)+(1+o_{\mathbf{P}}(1))\frac{\gamma_1\gamma_2^2}{(\gamma_1+\gamma_2)^2(1-\gamma_1)}\int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds \tag{5.20}$$
$$+o_{\mathbf{P}}(1).$$

For $\mathbf{I}_4$, *after the second-order condition of regular variation*

$$\sqrt{k}\mathbf{I}_4=o_{\mathbf{P}}(1). \tag{5.21}$$

For $\mathbf{I}_5$, *using the mean value theorem with* $X_{n-k,n}/t\longrightarrow 1$, *we get*

$$\sqrt{k}\mathbf{I}_5=-(1+o_{\mathbf{P}}(1))\frac{1}{1-\gamma_1}\sqrt{k}\left(\frac{X_{n-k,n}}{t}-1\right). \tag{5.22}$$

From Theorem 2.1 of Benchaira et al., 2015; we have

$$\sqrt{k}\left(\frac{X_{n-k,n}}{t}-1\right)-\gamma\mathbf{W}(1)=o_{\mathbf{P}}(1).$$

then

$$\sqrt{k}\mathbf{I}_5=-(1+o_{\mathbf{P}}(1))\frac{\gamma}{1-\gamma_1}\mathbf{W}(1).$$

For $\mathbf{I}_6$, *we have*

$$\int_1^\infty x^{-1/\gamma_1}dx=\frac{\gamma_1}{1-\gamma_1},$$

then

$$\mathbf{I}_6=\int_1^\infty x^{-1/\gamma_1}dx-\int_1^\infty\frac{\overline{\mathbf{F}}(tx)}{\overline{\mathbf{F}}(t)}dx.$$

Then, by applying the uniform inequality of regularly varying functions(see, e.g., Theorem 2.3.9 in De Haan et al., 2006); page 48) together with the regular variation of $|\mathbf{A}_\circ|$, *we show that*

$$\sqrt{k}\mathbf{I}_6\sim\frac{\sqrt{k}\mathbf{A}_\circ(t)}{(\gamma_1+\tau_1-1)(1-\gamma_1)}. \tag{5.23}$$

Summing up above equations, we get

$$\begin{aligned}\frac{\sqrt{k}(\widehat{\mu}_2-\mu_2)}{t\overline{\mathbf{F}}(t)}&=\left(\frac{\gamma_1\gamma_2-2\gamma(\gamma_1+\gamma_2)}{(1-\gamma_1)(\gamma_1+\gamma_2)}\right)\mathbf{W}(1)-\frac{\gamma^2}{\gamma_1+\gamma_2}\int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}W(s)\log s ds\\&+\frac{\gamma_1^2\gamma_2(\gamma_2-\gamma_1)}{(\gamma_1+\gamma_2)^2(1-\gamma_1)}\int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds+\frac{\sqrt{k}\mathbf{A}_\circ(n/k)}{1-\tau_1}\\&+\frac{\sqrt{k}\mathbf{A}_\circ(t)}{(\gamma_1+\tau_1-1)(1-\gamma_1)}.\end{aligned} \tag{5.24}$$

Finally, Summing up equations 5.17 and 5.24 achieves the proof.

5.4.2 Proof of Corollary 2.1

We set :

$$\frac{\sqrt{k}(\widehat{\mu}-\mu)}{\overline{\overline{\mathbf{F}}}(X_{n-k,n})X_{n-k,n}} = \Delta + \frac{(\gamma_1+\tau_1-1)(1-\gamma_1)+(1-\tau_1)}{(1-\tau_1)(\gamma_1+\tau_1-1)(1-\gamma_1)}\sqrt{k}\mathbf{A}_\circ(n/k),$$

Where $\Delta = c_1\Delta_1 + c_2\Delta_2 + c_3\Delta_3 + c_4\Delta_4 + c_5\Delta_5$ *with*

$$\Delta_1 = \mathbf{W}(1),\ \Delta_2 = \int_0^1 s^{-\frac{2\gamma_1}{\gamma}}\mathbf{W}(s)ds,\ \Delta_3 = \int_0^1 s^{-\gamma_1}\mathbf{W}(s)ds$$
$$\Delta_4 = \int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}\log(s)\mathbf{W}(s)ds,\ \Delta_5 = \int_0^1 s^{-\frac{\gamma}{\gamma_2}-1}\mathbf{W}(s)ds$$

After elementary but tedious computations, we find the following covariance as asymptotic variance: $\mathbf{\Gamma\Sigma\Gamma}^t$

Where $\mathbf{\Gamma} = (\frac{p(1-p)}{1-\gamma_1}, -p\gamma_1, p(1-p), \gamma_1 p^2(1-p), p(1-p)+\frac{\gamma_1^2 p}{1-\gamma_1})$ *and* $\mathbf{\Gamma}^t$ *is the transpose of* $\mathbf{\Gamma}$, Σ *is the variance-covariance matrix:*

$$\Sigma = \begin{bmatrix} \mathbf{1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} & \alpha_{1,5} \\ \alpha_{1,2} & \alpha_2 & \alpha_{2,3} & \alpha_{2,4} & \alpha_{2,5} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_3 & \alpha_{3,4} & \alpha_{3,5} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_4 & \alpha_{4,5} \\ \alpha_{1,5} & \alpha_{2,5} & \alpha_{3,5} & \alpha_{4,5} & \alpha_5 \end{bmatrix}$$

$$\mathbf{E}(\Delta_1^2) = 1,\ \alpha_2 := \mathbf{E}(\Delta_2^2) = \frac{2p^2}{(-2+p)(-4+3p)}$$

$$\alpha_3 := \mathbf{E}(\Delta_3^2) = \frac{(1-2p)}{p^4(1-p)},$$

$$\alpha_4 :=\ \mathbf{E}(\Delta_4^2){=}\frac{1-2p}{p^4(1-p)^2} - \frac{2\gamma_1 p}{(1-p)^3} - \frac{2(1-p)^{-2}}{(-1-p)} + \frac{1}{(1-p)^2(2p-1)^2}$$

$$\alpha_5 := \mathbf{E}(\Delta_5^2) = \frac{4p-3}{-p(1-p)^2(2p-1)},$$

$$\alpha_{1,2} := \mathbf{E}(\Delta_1\Delta_2)=\frac{p}{-2(1-p)},$$

$$\alpha_{1,3} := \mathbf{E}(\Delta_1\Delta_3) =\frac{1}{-\gamma_1+2},$$

$$\alpha_{1,4} := \mathbf{E}(\Delta_1\Delta_4) = -\frac{1}{p^2},$$

$$\alpha_{1,5} := \mathbf{E}(\Delta_1\Delta_5) =\frac{1}{p},$$

$$\alpha_{2,3} := \mathbf{E}(\Delta_2\Delta_3)=\frac{3p^3}{2(-2+p)(p-1)(-2+\gamma_1 p+3p)} + \frac{p}{(-2+p)(-\gamma_1+2)},$$

$$\alpha_{2,4} := \mathbf{E}(\Delta_2\Delta_4)=\frac{3p^2}{2(p-1)}\left(\frac{p}{2} - \frac{1}{4-p}\right)^2 + \frac{3p-2}{6}\left(\frac{p}{1+p}\right)^2,$$

$$\alpha_{2,5} := \mathbf{E}(\Delta_2\Delta_5)=\frac{-p^3\gamma_1}{2(-1+p)(-2+p)(-1-p+\gamma_1(-2+p))} + \frac{1}{-2+p},$$

$$\alpha_{3,4} := \mathbf{E}(\Delta_3\Delta_4) =\frac{-1}{(\gamma_1+2)(-\gamma_1+p+1)^2}+\frac{1}{(-\gamma_1+1)}\left[\left(\frac{p}{-1+p^2}\right)^2 + \left(\frac{1}{1-p}\right)^2\right],$$

$$\alpha_{3,5} := \mathbf{E}(\Delta_3\Delta_5) =\frac{1}{(-\gamma_1+2)(-\gamma_1+p+1)}+\frac{p^3\gamma_1^3}{(-\gamma_1+1)(-p\gamma_1-p\gamma_1^2-p^2\gamma_1^2-p+1)},$$

$$\alpha_{4,5} := \mathbf{E}(\varDelta_4 \varDelta_5) = \frac{(1-p)^2}{p\gamma_1(-\gamma_1 - 1)(2p-1)} + \frac{1-p}{p^2}.$$

Conclusions & Outlook

In this thesis, we are interested in a recent problem in the theory of their extremes, namely the presence of random truncation. This problem is very common in several areas of socio-economic life where data are often randomly censored on the right, such as medical insurance...

This thesis is broken down into two distinct parts to which is added an introduction. In the introduction, we recalled the areas where we meet the incomplete data (censored-truncated) with particular attention to truncated data. To facilitate the reading of the document, we recalled the high light points:

- *Giving you an introduction to the mathematical and statistical theory underlying EVT. contains some mathematical preliminaries also contains a derivation of the three families of classical Gnedenko limit distributions for extremes of iid variables and an account of regular variation and its extensions and domains of attraction.*
- *Beginning with a few reminders on basic concepts such as fdr, the three survival functions, and the equivalence relationship between these three functions is discussed. The more we talk about the laws of large numbers and asymptotic properties of the sum of the iid values (TCL).*
- *Moving to talk about incomplete data, with the main basic concepts on truncated data and some important and useful results existing in the literature for the random right truncation model. In this chapter we start with censored data, which can be further classified Afterwards, we will be interested in the truncated data. but in the present thesis, we are concerned with data that are right truncated.*

- *Last but not least we propose a method for estimating the mean of this type of distribution in the presence of random right truncation, its asymptotic normality established and its performance evaluated on simulated data; Our outlook in this subject is presented in the following question what is the kernel type of our estimator !.*

Bibliography

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (2008). ***A first course in order statistics****. SIAM. (see page 8).*

Beirlant, J., Bardoutsos, A., de Wet, T., & Gijbels, I. ***Bias reduced tail estimation for censored pareto type distributions****.* Statistics & Probability Letters, 109, *78–88 (see pages 2, 34).*

Beirlant, J., Guillou, A., Dierckx, G., & Fils-Villetard, A. ***Estimation of the extreme value index and extreme quantiles under random censoring****.* Extremes, 10*(3), 151–174 (see page 62).*

Benchaira, S. ***Statistics of incomplete data*** (Doctoral dissertation)*. Doctoral dissertation. Université Mohamed Khider-Biskra, 2017 (see page 55).*

Benchaira, S., Meraghni, D., & Necir, A. ***On the asymptotic normality of the extreme value index for right-truncated data****.* Statistics & Probability Letters, 107, *378–384 (see pages 3, 70, 77, 85, 87–89).*

Benchaira, S., Meraghni, D., & Necir, A. ***Tail product-limit process for truncated data with application to extreme value index estimation****.* Extremes, 19*(2), 219–251 (see pages 84, 85, 88).*

Billingsley, P. ***Probability and measure. 3rd wiley****.* New York *(see page 45).*

Bingham, N. H., Goldie, C. M., Teugels, J. L., & Teugels, J. (1989). ***Regular variation****. Cambridge university press. (see page 23).*

Breslow, N., & Crowley, J. ***A large sample study of the life table and product limit estimates under random censorship****.* The Annals of statistics, *437–453 (see page 3).*

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). ***An introduction to statistical modeling of extreme values*** (Vol. 208)*. Springer. (see page 7).*

Cox, D. R., & Oakes, D. (1984). Analysis of survival data. vol. vol. 21. (see page 2).

Csorgo, S., Deheuvels, P., & Mason, D. ***Kernel estimates of the tail index of a distribution.*** The Annals of Statistics, *1050–1077 (see pages 34, 50).*

David, H. A., & Nagaraja, H. N. (2004). ***Order statistics.*** *John Wiley & Sons. (see page 8).*

De Haan, L., Ferreira, A., & Ferreira, A. (2006). ***Extreme value theory: An introduction*** *(Vol. 21). Springer. (see pages 7, 27, 28, 34, 69, 77, 78, 89).*

Deheuvels, P., Haeusler, E., & Mason, D. M. ***Almost sure convergence of the hill estimator.*** Mathematical Proceedings of the Cambridge Philosophical Society, 104*(2), 371–381 (see page 34).*

Dekkers, A. L., & De Haan, L. ***On the estimation of the extreme-value index and large quantile estimation.*** The annals of statistics, *1795–1832 (see page 33).*

Dekkers, A. L., Einmahl, J. H., & De Haan, L. ***A moment estimator for the index of an extreme-value distribution.*** The Annals of Statistics, *1833–1855 (see page 2).*

Djabrane, Y. ***Conditional quantile for truncated dependent data*** *(Doctoral dissertation). Doctoral dissertation. Univesity of Biskra, 2010 (see page 55).*

Drees, H. ***Refined pickands estimators of the extreme value index.*** The Annals of Statistics, *2059–2080 (see page 33).*

Drees, H., de Haan, L., & Li, D. ***Approximations to the tail empirical distribution function with application to testing extreme value conditions.*** Journal of Statistical Planning and Inference, 136*(10), 3498–3538 (see page 83).*

Efron, B., & Petrosian, V. ***Nonparametric methods for doubly truncated data.*** Journal of the American Statistical Association, 94*(447), 824–834 (see page 66).*

Einmahl, J. H. ***Limit theorems for tail processes with application to intermediate quantile estimation.*** Journal of Statistical Planning and Inference, 32*(1), 137–145 (see page 83).*

Einmahl, J. H., Fils-Villetard, A., & Guillou, A. ***Statistics of extremes under random censoring.*** Bernoulli, 14*(1), 207–227 (see page 62).*

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997a). Modeling extremal events for insurance and financespringer. (see pages 29, 30).

Embrechts, P., Klüppelberg, C., & Mikosch, T. ***Modelling extremal events. for insurance and finance, volume 33 of.*** Applications of Mathematics (New York) *(see pages 7, 13, 28).*

Fisher, R. A., & Tippett, L. H. C. ***Limiting forms of the frequency distribution of the largest or smallest member of a sample.*** Mathematical proceedings of the Cambridge philosophical society, 24*(2), 180–190 (see page 2).*

Gardes, L., & Stupfler, G. ***Estimating extreme quantiles under random truncation.*** Test, 24*(2), 207–227 (see pages 3, 69, 70, 77).*

Gomes, M. I., & Neves, M. M. ***Estimation of the extreme value index for randomly censored data.*** Biometrical Letters, 48*(1), 1–22 (see page 62).*

HAOUAS, N. ***Contribution to statistics of rare events of incomplete data*** *(Doctoral dissertation). Doctoral dissertation. UNIVERSITE DE MOHAMED KHIDER BISKRA, 2017 (see page 55).*

Hill, B. M. ***A simple general approach to inference about the tail of a distribution.*** The annals of statistics, *1163–1174 (see page 2).*

Jenkinson, A. F. ***The frequency distribution of the annual maximum (or minimum) values of meteorological elements.*** Quarterly Journal of the Royal Meteorological Society, 81*(348), 158–171 (see pages 2, 16).*

Kalbfleisch, J. D., & Prentice, R. L. (2011). ***The statistical analysis of failure time data*** *(Vol. 360). John Wiley & Sons. (see page 2).*

Kaplan, E. L., & Meier, P. ***Nonparametric estimation from incomplete observations.*** Journal of the American statistical association, 53*(282), 457–481 (see pages 3, 63).*

Klein, J., & Moeschberger, M. ***Survival analysis: Techniques for censored and truncated data: Springer science & business media.[google scholar]*** *(see page 2).*

Klüppelberg, C., & Mikosch, T. (1997). ***Modelling extremal events for insurance and finance.*** *Springer. (see page 22).*

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (2012). ***Extremes and related properties of random sequences and processes.*** *Springer Science & Business Media. (see page 7).*

Lee, E. T., & Wang, J. (2003). ***Statistical methods for survival data analysis*** *(Vol. 476). John Wiley & Sons. (see page 2).*

Lynden-Bell, D. ***A method of allowing for known observational selection in small samples applied to 3cr quasars***. Monthly Notices of the Royal Astronomical Society, 155*(1), 95–118 (see pages 3, 68, 75).*

Mason, D. M. ***Laws of large numbers for sums of extreme values***. The Annals of Probability*, 754–764 (see page 34).*

Meraghni, D. ***Modelling distribution tails*** (Doctoral dissertation)*. Doctoral dissertation. Biskra, Université Mohamed Khider. Faculté des Sciences et des Sciences de l . . . , 2008 (see page 7).*

Necir, A. ***A functional law of the iterated logarithm for kernel-type estimators of the tail index***. Journal of statistical planning and inference, 136*(3), 780–802 (see page 34).*

Peng, L. ***Estimating the mean of a heavy tailed distribution***. Statistics & Probability Letters, 52*(3), 255–264 (see pages iii, 47–50, 73, 75).*

Pickands III, J. ***Statistical inference using extreme order statistics***. the Annals of Statistics*, 119–131 (see pages 2, 32).*

Rassoul, A. ***Kernel-type estimator of the mean for a heavy tailed distribution***. Statistics and Its Interface, 8*(1), 85–91 (see pages 50, 53).*

Reiss, R.-D. (1987). ***Approximate distributions of order statistics: With applications to nonparametric statistics***. *Springer science & business media. (see page 7).*

Reiss, R.-D., Thomas, M., & Reiss, R. (2007). ***Statistical analysis of extreme values*** (Vol. 2)*. Springer. (see pages 62, 80).*

Resnick, S. I. (1987). ***Extreme values, regular variation and point processes***. *Springer. (see pages 7, 15, 24, 28).*

Smith, R. L. ***Estimating tails of probability distributions***. The annals of Statistics*, 1174–1207 (see page 19).*

Soltane, L. ***Analyse des valeurs extrêmes en présence de censure*** (Doctoral dissertation)*. Doctoral dissertation. Université Mohamed Khider-Biskra, 2017 (see pages 39, 55).*

Stute, W. ***The central limit theorem under random censorship***. The Annals of Statistics*, 422–439 (see pages 63–65, 74).*

Stute, W., & Wang, J.-L. ***The central limit theorem under random truncation***. Bernoulli: official journal of the Bernoulli Society for

Mathematical Statistics and Probability, 14*(3), 604 (see pages 75, 76).*

Teugels, J., Bingham, N., & Goldie, C. (1987). ***Regular variations****. Cambridge University Press. (see page 22).*

von Mises, R. ***La distribution de la plus grande de n valeurs, vol. 2 of selected papers of richard von mises****.* Providence, RI: American Mathematical Society, pp. 271r*, 294 (see pages 2, 16, 26).*

Wienke, A. (2010). ***Frailty models in survival analysis****. CRC press. (see page 2).*

Woodroofe, M. ***Estimating a distribution function with truncated data****.* The Annals of Statistics, 13*(1), 163–177 (see pages 3, 68, 69, 77).*

Worms, J., & Worms, R. ***New estimators of the extreme value index under random right censoring, for heavy-tailed distributions****.* Extremes, 17*(2), 337–358 (see page 62).*

Notations

Appendix B: Software R

R is a system, commonly known as language and software, which allows statistical analyzes to be carried out. More particularly, it comprises means which make possible the manipulation of the data, the calculations and the graphical representations. R also has the ability to run programs stored in text files and includes a large number of statistical procedures called packets. The latter make it possible to deal fairly quickly with subjects as varied as linear models (simple and generalized), regression (linear and non-linear), time series, classic parametric and non-parametric tests, the various methods of data analysis. , ... Several packages, such ade4, FactoMineR, MASS, multivariate, scatterplot3d and rgl among others are intended for the analysis of multidimensional statistical data.

It was originally created in 1996 by Robert Gentleman and Ross Ihaka of the Department of Statistics at the University of Auckland in New Zealand. Since 1997, an "R Core Team" has been formed which develops R. It is designed to be used with Unix, Linux, Windows and MacOS operating systems.

A key element in R's development mission is the Comprehensive R Archive Network (CRAN) which is a collection of sites that provides everything needed for the distribution of R, its extensions, documentation, source files and files. binaries. The master site of CRAN is located in Austria in Vienna, it can be accessed by the URL: "http://cran.r-project.org/". The other CRAN sites, called mirror sites, are spread all over the world.

R is free software distributed under the terms of the "GNU Public License". It is an integral part of the GNU Project and has an official site at "http: //www.R-project.org". It is often presented as a clone of S which is a high level language developed by AT&T Bell Laboratories and more specifically by Rick Becker, John Chambers and Allan Wilks. S can be used through *the S-Plus software which is marketed by the company Insightful (http://www.splus.com/).*

List of Publications and Communication

5.1 Articles in Refereed Journals

- *Estimating the mean of heavy tailed distribution under random truncation.Journal of Siberian Federal University Mathematics and Physics 14:3 (2021), 273–286. Joint work with BEN DAHMANE Khanssa, BENATIA Fateh, and BRAHIMI Brahim.*

5.2 Communications

- *International Workshop on Perspectives On High dimensional Data Analysis (HDDA-2018), Marakesh, Morocco;09-13 April 2018, poster communication poster : Estimating the mean of heavy tailed distribution under random truncation.*
- *Applied Mathematics Days in Biskra, a poster communication entitled: The t-hill estimator of the extreme value index for right truncated data.*
- *Congress of Algerian Mathematicians, 12, 13 May 2018, University of Boumerdes, a poster communication entitled: The t-hill estimator of the extreme value index for right truncated data.*
- *International conference on computational methods in applied sciences;12-16 july,2019 , Istanbul; Turkey, oral presenter : Nonparametric estimation for heavy tail distributions with incomplete data.*