PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
Ministry of Higher Education and Scientific Research
SAAD DAHLAB UNIVERSITY, BLIDA 1
Faculty of Sciences
Department of Mathematics

THESIS

In Mathematics

Contribution to the estimation of risk measures for heavy-tailed distributions under censored data

By
BARI Amina

Examination Committee Members :

Maamar BENBACHIR	Chairman	Prof.	Univ. Blida 1
Abdelaziz RASSOUL	Advisor	Prof.	ENSH, Blida
Hamid OULD ROUIS	Co-Advisor	Prof.	Univ. Blida 1
Diffalah LAISSAOUI	Examinator	MCA	Univ. Médéa
Ghania SAIDI	Examinator	Prof.	ENSSEA Koléa
Redouane BOUDJEMAA	Examinator	MCA	Univ. Blida 1

January, 2022

DEDICATION

This work is dedicated

To my mother and to the memory of my father,

To my daughter and son Najia and Touhami,

To my brothers, sisters and all my family,

To all those who are dear to me.

ACKNOWLEDGMENTS

In the Name of Allah, the Most Gracious, the Most Merciful. Blessings and peace be upon our leader Muhammad and also upon his family and his companions.
I have completed this thesis with invaluable advice, support and encouragement from a variety of people whose names may not be fully acknowledged within this limited text.
First and foremost, I would like to express my deepest sense of gratitude to my supervisor Prof. Rassoul Abdelaziz, who offered his continuous advices and encouragement and his valuable support throughout the course of this thesis. Without his constructive comments, patience and knowledge this thesis would have not been possible.
I would be remiss if I did not thank him for engaging me in new ideas. Working with he is a great pleasure and I don't have enough words to express my deep and sincere appreciation. I honestly could have not wished for a better supervisor.
I would like to express my greatest gratitude to my Co-Advisor Pro. Hamid OULD ROUIS, thank you and the best of luck in your future endeavors.
Besides my advisors, I would like to thank the members of the examination committee: starting with its Chairman who honored me for the second time by sharing my scientific achievements Prof. Maamar Benbachir, as well as the examinees: Prof. Diffalah LAISSAOUI, Prof. Ghania SAIDI, Prof. Redouane BOUDJEMAA for accepting the verdict of my thesis.
Special thanks also to all my friends everywhere.
Finally, my special gratitude is saved for myall my family, and particularly my mother, for teaching me the value of hard work, my children, for ther infinite patience and support over the last years while I have been working on this research.

الملخص

النتيجة الرئيسية التي حصل عليها Fisher وTippett في عام 1928 المتعلقة بشأن قوانين الحد الممكنة للعينة القصوى أظهرت و أثبتت فكرة أن نظرية القيمة المتطرفة كانت شيئًا خاصًا إلى حد ما ، ومختلف تمامًا عن نظرية الحد المركزي الكلاسيكية.

في هذه الأطروحة ، قمنا بتعريف ودراسة أحد المؤشرات الأكثر شيوعا لقياس عدم المساواة في دخول رأس المال ، والمعروف بمؤشر Gini ، كما حاولنا بناء مُقدِّرًا لمؤشر جيني في حالة التوزيعات ذات الذيل الثقيل ، خصوصا عندما تكون البيانات خاضعة للرقابة.

الكلمات المفتاحية: المؤشر المتطرف ، مؤشر Gini ، توزيعات الذيل الثقيل ، مقاييس عدم المساواة ، مقدر Kaplan-Meier ، توزيعات الخسارة ، الرقابة العشوائية.

Le résumé

Le principal résultat obtenu par Fisher et Tippett en 1928 sur les lois limites possibles du maximum d'échantillon a apparemment créé l'idée que la théorie des valeurs extrêmes était quelque chose d'assez spécial, très différent de la théorie de la limite centrale classique.

Dans cette thèse, nous définissons et étudions l'un des indices les plus populaires qui mesure l'inégalité des revenus du capital, connu par l'indice de Gini, nous construisons un estimateur de l'indice de Gini dans le cas des distributions à queue lourde, surtout lorsque les données sont censurées.

Mots-clés : Indice extrême, indice de Gini, distributions à queue lourde, mesures d'inégalité, estimateur de Kaplan-Meier, distributions de pertes, censure aléatoire

Abstract

The main result obtained by Fisher and Tippett in 1928 on the possible limit laws of the sample maximum apparently created the idea that the extreme value theory was something rather special, very different from the classical central limit theory.

In this thesis, we define and study one of the most popular index which measure the inequality of capital incomes, known by Gini index, we construct an estimator for the Gini index in case of the heavy-tailed distributions, specially when data are censored.

Key-Words: Extreme index, Gini index, heavy tailed distributions, inequality measures, Kaplan-Meier estimator, loss distributions, random censoring.

Contents

List of Figures

List of Tables

Notations and Abbreviations

The following notation will be used throughout the thesis:

$r.v$: random variable.

$i.i.d$: Independent and identically distributed.

$i.e$: in other words.

$e.g$: for example.

$\overline{X_n}$: arithmitic mean.

S_n : arithmetic sum.

σ^2 : variance.

F : cumulative distribution function (cdf).

F^{-1} : the inverse function.

F_n : empirical distribution function.

f : probability distribution function (pdf).

Q : quantile function.

$\xrightarrow{p}$: convergence in probability.

$\xrightarrow{\mathcal{L}}$: convergence in law.

$\xrightarrow{a.s}$: almost sure convergence.

$(\Omega; A; \mathbb{P})$: probability space.

$(X_{1,n}, X_{2,n}, ..., X_{n,n})$: order statistics of n i.i.d observations from a r.v X.

$X_{1,n}$: minimum of $(X_1, ..., X_n)$.

$X_{n,n}$: maximum of $(X_1, ..., X_n)$.

$\Phi(x)$: the cdf of the standard Gaussian law.

CLT : Central Limit Theorem.

EVT : Extreme Value Theory.

EVI : Extreme Value Index.

$\mathcal{GEV}$: Generalized Extreme Value.

$\mathcal{GPD}$: Generalized Pareto distribution.

$G_{\gamma,\sigma}$: generalized Pareto law.

$\mathcal{H}_\gamma$: the cdf of generalized extreme value.

POT : Peaks Over Threshold.

ML : Maximum Likelihood.

MM : Method of moments.
$\mathcal{L}(\theta;(X_1,...,X_n))$: Maximum likelihood function.
$D(.)$: Domain of attraction.
Λ : Gumbel's distribution.
Φ_α : Frechet's distribution.
Ψ_α : Webull's distribution.
$L(.)$: slowly varying function.
$U(.)$: regular varying function.
x_F : The upper-end point.
$\hat{q}_{\alpha_n}$: quantile estimator.
$\hat{q}^W_{\alpha_n}$: Weissman's estimator.
$\hat{q}^{DH}_{\alpha_n}$: Dekkers and de Haan's estimator.
$\widehat{\gamma}^{(P)}_{n;k}$: Pickands estimator.
$\widehat{\gamma}^{(H)}_{n;k}$: Hill estimator.
$\mathbb{I}_{\{A\}}$: the indicator function of the set A.
$S(t) = \overline{F}(t)$: Survival function.
$h(t)$: Hazard function or "chance rate"
$H(t)$: Cumulative hazard function.
$\hat{S}_{KM}$: Kaplan-Meier estimator.
$\hat{S}_{GKM}$: Generalized Kaplan-Meier estimator.
$\hat{H}_{NA}$: Nealson-Aalen estimator.
VaR : Value-at-Risk.
$TVaR$: Tail Value-at-Risk
$\mathbb{CTE}$: Conditional Tail Expectation.
$\mathbb{ES}$: Expected shortfall.
MSE : Mean squared error.
$RMSE$: Root mean squared error
$CVaR$: conditional VaR.
GMD : Gini's mean difference.
I_A : Atkinson's measurement of inequality.

Generale Introduction

Problem description

In recent years the field of extreme value theory has been a very active research area. It is of relevance in many practical problems such as the flood frequency analysis, insurance premium, financial area, ... The theory of extreme values is a branch of statistics that tries to find a solution to these phenomena. It is mainly based on limit distributions of extremes and their domains of attraction. However, there are two models: generalized law of extremes (GEV: "Generalized Extreme Value") and generalized Pareto law (GPD: "Generalized Pareto Distribution"). Thus, it all started with the authors Fisher and Tippet (1928, [55]) when they studied the resistance of cotton threads and later Gnedenko (1943, [69]) became interested in these distributions. They stated a fundamental theorem with the creation of three domains of attraction: domain of attraction of Fréchet, Gumbel and Weibull. This interesting theorem refers to a parameter called the tail index which gives the shape of the tail of the distribution. For the literature concerning extreme value theory we refer to Reiss and Thomas (2007, [116]), Coles (2001, [22]) and Beirlant et al. (2007 [14]).

Extreme value analysis under random censoring is becoming more popular with applications for example in survival analysis, reliability and insurance. For instance, in certain long-tailed insurance products, such as car liability insurance, long developments of claims are encountered. One major departure from the unbiased-sample case is that where the sample has been censored. Censored data are commonly encountered in practical applications to income and wealth distributions, for several reasons. The modeling of censored extreme values emerged in 1997 in the literature of extremes with the publication of the book Reiss and Thomas (2007, [116]).

It is of great interest to guard against extreme risks, whether they result from a financial crisis, a nuclear accident or a natural disaster, taking into account the human, economic and financial repercussions that these can to have.

The statistical analysis of extremes is key to many of the risk management prob-

lems related to insurance, reinsurance, and finance. In statistics of extremes we deal essentially with the estimation of parameters of extreme or even rare events. Where the formulation of the possible limiting distributions of the affinely transformed maximum of a sample, shows that the parameter, i.e, the extreme value index is an important characteristic of the distribution.

The analysis of extreme values of random censorship is a new subject of research. The first aim of this thesis is to extend the results of the extreme value theory in the case where the sample consists of a censored data and estimate risk measures while making the necessary modifications.

The Gini concept or the mean difference of the Gini, initiated by Gini in 1912 [66], is a characteristic of widespread dispersion in the field of income distribution. The specificity of this indicator lies in its simple calculations. Monti (1991, 1993) showed that the Gini concentration ratio is very sensitive to extreme observations, it means that it records huge changes in the tail of the distribution, and this property may lead to problems in the presence of outliers. Nevertheless, the Gini index is robust against rounding errors, which is a good property especially in the case of grouped data.

Also a Bayesian nonparametric estimation of the Gini index has been proposed. Gigliarano and Muliere (2013) [65] suggested an alternative approach for dealing with contaminated observations and extreme income values: avoiding the common practice that removes these critical data, they instead treat them as censored observations and apply a Polya tree model for incomplete data.

The theme of this thesis revolves around three axes: extreme values, censored data and risk measures. In the remainder of this thesis we will mainly be concerned with the estimation of an inequality mesure when the data are censored. This thesis can be considered in its entirety as a contribution to the theory of extreme values and its statistical applications.

Methodology:

The treatment of this problem is established according to the following axes:

- Presentation of the extreme values theory and characterization of the domains of attraction for each extreme law.
- Review of the mains results about censored data for extremes distributions, Kaplan-Meier estimator of the cdf.
- Presentation of extreme income distributions and inequalities indices.

- Empirical estimation of the Gini index, presentation of the restriction of this estimator on distributions with moments of order two exist and finite.
- Proposed a Semi-parametric estimation of the Gini index for heavy-tailed income distributions, the study of its asymptotic behavior, and some results of simulation.
- Proposed a non parametric estimator of Gini index for heavy-tailed income distributions with censored data, the study of its asymptotic behavior, and some results of simulation.

Organization of the thesis:

In this work, we study the analysis of extreme values under random censorship from the theoretical development to its applications, we use it for the estimation of some risk index of certain classes of heavy-tailed distributions.

This thesis consists of two papers and is divided into four chapters:

- **Paper 1**: " Bari, A., Rassoul, A., Rouis, H. O. (2021). Estimating the Gini index for heavy-tailed income distributions. South African Statistical Journal, 55(1), 15-28.

-**Paper 2**: " Estimating the Gini index for income loss distributions under random censoring"

The chapters are organized as follows:

- **In Chapter 1**, the fundamentals of extreme value theory are provided which are necessary for a proper understanding of the following chapters. We provide an overview of the essential definitions and results of EVT. We start by the asymptotic properties of the sum of independent and identically distributed random variables, order statistics and distributions of upper order statistics. Afterwards, we are interested in the result, first discovered by Fisher and Tippett [55] and later proved in complete generality by Gnedenko; on the fluctuations and asymptotic behavior of the maximum $X_{n,n}$ of a series of independent and identically distributed random variables. A reminder on generalized extreme value $\mathcal{GEV}$ and generalized Pareto distribution $\mathcal{GPD}$ approximations, domains of attraction and regular variation functions is given as well.

- The **second chapter** presents the following two cases of incomplete data: censored and truncated. This chapter is divided into four section: an introduction, the second section present a foreword on survival data and censoring with these different types as well as estimating the survival function as well in a right-censoring model. Then we will present some definitions related to the statistics of survival times. Censorship is based on a few functions suchs as the distribution function,

the survival function, the risk function. Section 2.4 presents a summary of the main estimators of relevant quantities, of which the most famous non-parametric estimators are the Kaplan-Meier survival function estimator (Kaplan-Meier suggested a survival function estimator that Beran generalized in 1981 [15] in the conditional case called the generalized Kaplan-Meier estimator), and the Nelson Allen estimator for the cumulative risk function.

Chapter 3 is about measuring risk: we will start in Section 3.1 with an introduction and then provide some of concepts and definitions we need for the following sections. We also discuss risk measurement wich is a great part of an organization's overall risk management strategy. Risk measurment is a tool to used to assess the probability of a bad event happening. It can be done by businesses as part of disaster recovery planning and as part of the software development lifecycle. Then we'll recall the definition of inequality and some well-known measures of income inequality and the relationships between them. In the last section we discuss what the Lorenz curve and Gini coefficient are, and give some of their main properties.

- A simulation study is given in **chapter four**, this Chapter corresponds first to the article " Estimating the Gini index for heavy-tailed income distributions", we define and study one of the most popular indices which measure the inequality of capital incomes, known as "Gini index", we construct a semiparametric estimator for the Gini index in case of heavy-tailed income distributions, we establish its asymptotic distribution and derive bounds of confidence. We explore the performance of the confidence bounds in a simulation study and draw conclusions about capital incomes in some income distributions. Then we elaborate a non parametric estimation of the Gini index for income distributions when the data are randomly censored on the right. The estimator is constructed and its asymptotic normality established under appropriate conditions. Its performance is evaluated using simulated data sets, this corresponds to the article " Estimating the Gini index for income loss distributions under random censoring".

Finally, the conclusion is drawn from this work along with areas in which further research may be directed.

Chapter 1

Extreme value theory

Introduction

In this chapter we introduce the Extreme Value Theory, which is a branch of statistics that ful fils the modelling needs of extreme events and extreme probabilities in many disciplines, and wich emerged from the research of the limit distribution of the largest order statistics in a sample as the sample size increases to infinity (Fisher and Tippett, 1928) [55].

This chapter brings together some essential notions on the theory of extreme values which make it easier to read the thesis. After having introduced the behavior of the maximum, we will present the two main tools used to model the behavior of the extreme values of a sample: the law of extreme values, and the law of excess, we will then focus on the characterization of domains of attraction, functions with slow and regular variations and finally, we will recall the different methods of estimating extreme quantiles.

1.1 Brief history about EVT

Extreme Value Theory (EVT) is a broad theory aimed at studying rare events, that is, events with a low probability of occurrence. This theory has become one of the most important statistical disciplines for applied science in recent years. Extreme value techniques are also increasingly used in many other disciplines.

The origins of this theory go back to the pioneering work of Fréchet in 1927 and "Fisher and Tippett" in 1928 [55] on the convergence of the maximum of a sample of independent and identically distributed random variables.

At that time, the possible limits for the distributions of maximum samples of i.i.d random variables were derived. The complete proof was proposed by Gnedenko in 1943 [69], and later simplified by de Haan in 1976 [44]. These limit distributions

were unified by a parameterization given by Von Mises in 1954 [130] and Jenkinson in 1955 [84], leading to the distribution of generalized extreme values.

Gumbel in 1958 [75] was the first to be interested in the potential applications of the maximum value theory from a statistical point of view, and he exploited this finding in his famous book, Statistics of Outliers. Since then, great developments have taken place in theory and applications, and his book remains relevant today.

The theory of extreme values finds application in many fields such as reliability, metallurgy and astrophysics. It is also of interest to environmental sciences, with the modeling of large forest fires as well as climatology and meteorology. Two other main areas of application are: actuarial science for hedging against high impact claims, and finance.

Extreme value theory provides a rigorous probabilistic mathematical basis on which to build statistical models to predict the intensity and frequency of these extreme events.

Two important results from Extreme Values Theory are the limit distributions of a series of block maxima and of excesses over a threshold, called "Peaks Over Threshold" (POT), given that the distributions are non degenerate and the sample is iid.

1.2 Laws of extreme values

The extreme values theory (EVT) is mainly based on two findings. These two results give us the approximate behavior of the random variable X or the crosses of the threshold u. The benefit of these results comes from the fact that it is not necessary to know the law, F for the operation X we want to predict.

However, EVT is analogous to Central Limit Theorem (CLT) but for extreme events. Thus, where the CLT shows that the empirical mean of the variable X converges to a normal distribution (independently of the law of the variable of interest and when moments of order 1 and 2 exist); the EVT establishes similar results but for the extreme values of X.

1.2.1 Basic concepts

We begin our study with some definitions and results that we need throughout this thesis.

Definition 1.1 *(Sum and arithmetic mean)*
Let $X_1, X_2, ..., X_n$ be a sequence of random variables (rv's) that are independent and identically distributed (iid) defined on the same probability space. For an integer $n \geq 1$, we

define the sum and the corresponding arithmetic mean respectively by:

$$S_n = \sum_{i=1}^{n} X_i \quad and\ \overline{X}_n = \frac{S_n}{n} \tag{1.1}$$

$\overline{X}_n$ *is then called sample mean or empirical mean.*

1.2.1.1 Law of large numbers

The laws of large numbers indicate that a random draw is made from a series of large sizes, the more the sample size is increased, the more the statistical characteristics of the draw (the sample) approximate the statistical characteristics of the sample. They are of two types: weak laws bringing into play the convergence in probability P and strong laws relating to the convergence almost surely *a.s*.

Theorem 1.1 *Let* $(X_1, X_2, ..., X_n)$ *be a sample of independent rv of the same law admitting a mean* μ *and a variance* σ^2*. Then the sequence of empirical means* $\{\overline{X}_n\}$ *almost surely converges to* μ*, i.e.:*

$$\overline{X}_n \xrightarrow{a.s} \mu \ as\ n \to \infty$$

We state here the "Strong law" of large numbers. There are different versions of this law requiring more or less restrictive conditions than those used here, including the "Weak law" concerning convergence in probability.

From a practical point of view the law of large numbers ensures that the empirical mean increasingly approaches the mean of the law from which the sample is taken as the size of the sample is increased.

1.2.1.2 Central Limit Theorem

The study of sums of independent variables with the same law plays a crucial role in statistics. The following theorem is known under the name of Central Limit Theorem (CLT) establishes the convergence in law towards the normal distribution of a sum of rv i.i.d under very light assumptions. The first proof of this theorem, published in 1809, is due to Pierre-Simon de Laplace, but the particular case where the variables follow Bernoulli's law with parameter $\delta = 0.5$ was known since the work of De Moivre in 1733.

The CLT states that a sum of n rv's independently drawn from a common distribution function $F(x)$ with finite variance, converge to the normal distribution as n goes to infinity.

Theorem 1.2 *(CLT)*
If $X_1, X_2, \ldots$ is a sequence of rv's iid of mean μ and finite variance σ^2, then

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

The proof of this Theorem 1.2 could be found in any standard book of statistics (see example, Lejeune, M. page 83 [97]).

1.2.2 The laws of maximums

Historically, the study of the probability law of the maximum of a sample of n variables has been the first approach to describe extreme events. Fisher and Tippet in 1928 [55] were the first to heuristically deduce the possible limit laws for the maximum of a series of independent random variables with the same law, before Gnedenko in 1943 [69] rigorously obtained the convergence, the proof of which was simplified by de Haan in 1976 [44]. The work of von Mises in 1936 [130] and Jenkinson in 1955 [84] made it possible to give a unified form to this result. Applications began following the work of Gumbel in 1954 [75], particularly in hydrology.

Convergence of extremes for maximums is the equivalent of convergence of means: another law of large numbers. If the distribution of means around their expectation tends to be Gaussian when the variance is finite, the distribution of extremes also converges towards a particular limit.

1.2.2.1 Order statistics

Order statistics play an increasingly important role in extreme value theory.

Definition 1.2 *(Order statistics)*
If the random variables $X_1, X_2, \ldots, X_n$ are arranged in increasing order of magnitude and then written as

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

the random variable $X_{i,n}$ is called the ith order statistics ($i = 1, \ldots, n$).

In the following we will assume that X_i are independent and identically distributed (i.i.d.) random variables from a continuous population with cumulative distribution function (cdf) F, and probability density function (pdf) f:

Definition 1.3 *(Quantile function).*
Let F be a distribution function. The quantile function is

$$Q(s) = F^{\leftarrow}(s) = \inf\{x \in \mathbb{R} : F(x) \geq s\}, \quad 0 < s < 1 \tag{1.2}$$

For any cdf F, the quantile function is non-decreasing and right-continuous. If F is continuous, then Q is continuous. If F is strictly increasing, then Q is the inverse function F^{-1}.

We now present this limit distribution, let $\{X_1, X_2, ..., X_n\}$, n random variables i.i.d. having the distribution function:

$$F(x) = \mathbb{P}(X < x)$$

Consider a sample comprising n realizations $\{x_1, x_2, ..x_n\}$. We arrange them in ascending order, and present the convention:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

The largest of these realizations $x_{(n)}$ can be considered as the realization of a new random variable $X_{(n)}$. The same idea prevails for the other observations $x_{(k)}$. Thus, we present n new random variables with the convention:

$$X_{(n)} = \max(X_1, X_2, ..., X_n) \tag{1.3}$$

This represents the maximum of a sample of size n which is the random variable giving the greatest value. In the literature, this quantity is also noted by: $X_{n,n} = M_n$

The same:

$$X_{(1)} = X_{1,n} = \min(X_1, X_2, ..., X_n) \tag{1.4}$$

represents the minimum of a sample of size n which is the random variable giving the smallest observed value. More generally $X_{(k)}$ (or $X_{k;n}$) is the random variable attached to the kth value $x_{(k)}$ obtained among n realizations. These n random news items can be ordered as follows:

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

Two order statistics are particularly interesting for the study of extreme events. These are the extreme order statistics which are given by the following definition:

Definition 1.4 *(Extreme statistics).*
The extreme order statistics are defined as terms of the maximum and minimum of the n rv's $X_1, X_2, ..., X_n$. The variable $X_{1,n}$ is the smallest order statistic (or minimum statistic) and $X_{n,n}$ is the largest order statistic (or maximum statistic).

We want to determine the limiting behavior of the maximum, i.e., we want to characterize the probability law of maximum. All results for the minimum sample

can be obtained from those for the maximum using the following relation:

$$\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n) \tag{1.5}$$

The exact distribution of M_n can be obtained from the cdf F. We consider the cdf of $X_{(k)}$, which we denote by $F_{k,n}$ to indicate that the sample is of size n. We have:

$$F(x) = \mathbb{P}(X_{(k)} < x) = \sum_{p=k}^{n} C_n^p [F(x)]^p [1 - F(x)]^{n-p} \tag{1.6}$$

Let us use the fact that the random variables X_1, X_2,...., X_n are independent and have the same function cdf $F(x)$, we get the cdf of the maximum (and even the minimum) for the value $k = n$ (law of the maximum):

$$F_{n,n}(x) = \mathbb{P}(X_{(n)} < x) = \sum_{p=n}^{n} C_n^p [F(x)]^p [1 - F(x)]^{n-p}$$

so:

$$F_{n,n}(x) = [F(x)]^n \tag{1.7}$$

and for the value $k = 1$ (minimum law) we have:

$$F_{1,n}(x) = \mathbb{P}(X_{(1)} < x) = \sum_{p=1}^{n} C_n^p [F(x)]^p [1 - F(x)]^{n-p}$$

then:

$$F_{1,n}(x) = 1 - [1 - F(x)]^n \tag{1.8}$$

and their density functions are respectively:

$$\begin{cases} f_{n,n}(x) = n[F(x)]^{n-1} f(x) \\ f_{1,n}(x) = n[1 - F(x)]^{n-1} f(x) \end{cases} \tag{1.9}$$

We can also get these expressions directly. Consider for example the maximum. So:

$$\{M_n < x\} \Leftrightarrow \{\max(X_1, X_2, ..., X_n) < x\} \Leftrightarrow \cap_{k=1}^{n} \{X_k < x\}$$

and, using the independence of the initial random variables:

$$\mathbb{P}(M_n < x) = \mathbb{P}(\cap_{k=1}^{n} \{X_k < x\}) = \prod_{k=1}^{n} \mathbb{P}(X_k < x)$$

indeed:

$$F_{n,n}(x) = [F(x)]^n$$

In the rest of this thesis, we will focus on the study of the maximum.

We can segregate the values on equal time intervals and record the maximum value in each interval. We will end up with block maxima (see figure below).

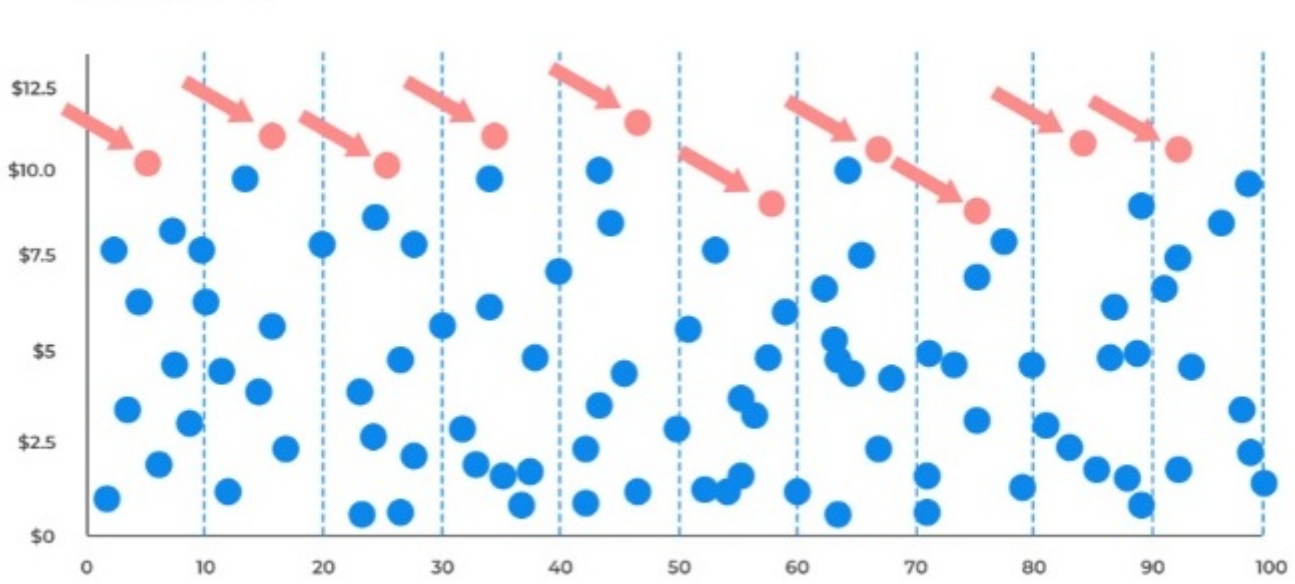

Figure 1.1: Illustrating Block Maxima

Hence, we see that, when the cdf F is known, the exact maximum cdf can be obtained easily. However, the cdf F is unknown in practice, it is reasonable to study the asymptotic behavior of M_n for $n \to \infty$, with the objective of approximating the distribution of M_n by a non-degenerate limiting distribution.

Denote by x_F to the upper endpoint[1] of cdf F with the convention $\sup\{\varnothing\} = \infty$. It represents the upper limit of the support of the law.

From relation (1.7), we can conclude as to the form of the limit law of M_n by making n tend towards infinity. We find:

$$\lim_{n\to\infty} F_{n,n}(x) = \lim_{n\to\infty} [F(x)]^n = \begin{cases} 1 & \text{if } x \geq x_F \\ 0 & \text{if } x < x_F \end{cases} \tag{1.10}$$

The result (1.10) indicates that the distribution of the maximum M_n is a degenerate law. This result provides a limited interest on the behavior of M_n.

[1] The upper (or right) endpoint of the cdf F is defined as follows:
$x_F = \sup\{x \in \mathbb{R}, F(x) < 1\} \leq \infty$

We want to find a distribution of some interest for the maximum. The idea is to apply a transformation to the maximum M_n so that the passage to the limit leads to a non-degenerate distribution. This outcome is well known in the context of the central limit theorem. The latter is concerned with the asymptotic behavior of the mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; when $n \to \infty$

Theorem 1.3 *Consider a sequence of random variables i.i.d* $\{X_1, X_2, ..., X_n\}$ *with* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2 < \infty$. *So,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x) \tag{1.11}$$

where $\Phi(x)$ *is the fdr of the standard Gaussian law,* $b_n = n\mu$ *and* $a_n = \sigma\sqrt{n}$.

The central limit theorem says that: when $n \to \infty$ the mean tends to be distributed according to a normal distribution if the variance is finite. If n is large enough then it seems relevant to use the normal distribution to model the mean. The extreme value theory follows the same logic, but studies the tail of the law instead of its mean. So, it's interested in a non-degenerate distribution for the maximum of the sample instead of the mean. For this, we need a similar theorem, i.e., we are looking for normalization sequences $a_n > 0$ and $b_n \in \mathbb{R}$, such as:

$$\frac{M_n - b_n}{a_n} \xrightarrow{\mathcal{L}} G, \quad \text{when } n \longrightarrow \infty \tag{1.12}$$

with G is non-degenerate distribution, i.e.

$$\lim \mathbb{P}(M_n < a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{1.13}$$

for any point x where G is continuous.

Consider the maximum of the sample rather than the average, also a double problem arises. On the one hand, we must identify all the possible non-degenerate distribution functions that can appear as a limit in (1.13), on the other hand, we must characterize the cdf F (in terms of necessary and sufficient conditions) for which there are sequences $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$, $a_n > 0$ and $b_n \in \mathbb{R}$, such that (1.13) is satisfied .

1.2.2.2 Law of excess

It is unrealistic to believe that only the maximum of the sample can model the behavior of extreme values. The other large values of the sample also contain information about the tail of the distribution. The approach by threshold overruns, or

"Peaks-Over-Threshold approach" denoted POT, is an alternative to the GEV law in the modeling of the behavior of the maximum of a sample based on the "large values" of the sample.

This approach relies on the use of higher order statistics from the sample. It consists in keeping only the observations exceeding a certain threshold. The excess beyond the threshold is defined as the difference between the observation and the threshold.

More precisely, let a sample of i.i.d $X_1, X_2, ..., X_n$, and let u be a fixed (non-random) threshold such that $u < x_F$. Consider the N_{ip} observations $X_{i1}, X_{i2}, ..., X_{ip}$ exceeding the threshold u. We call excess beyond the threshold u the $Y_j = X_{ij} - u$, where $j = 1, ..., p$.

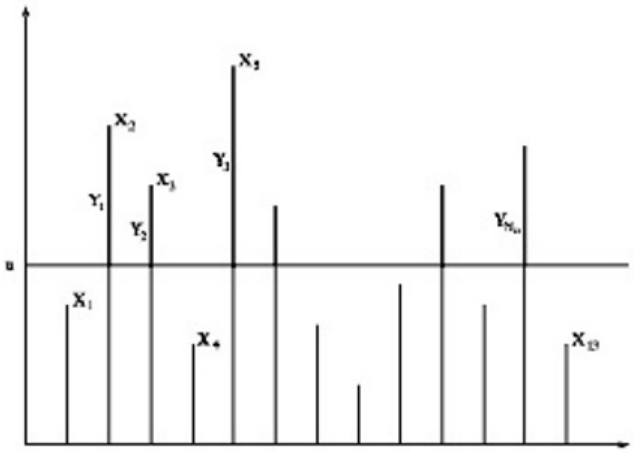

Figure 1.2: The $X_1, X_2, X_3, ..$ data, and the corresponding excesses $Y_1, ..., Y_N$ above threshold u.

We will denote by F_u the distribution function of the excess Y beyond the threshold u. The law of excesses is that of random variables i.i.d. admitting for distribution function $F_u(x) = \mathbb{P}(Y \leq x | X > u)$ representing the probability that the random variable Y does not exceed the threshold u by at least one quantity x knowing that it exceeds u. F_u thus describes the law of Y knowing that $X > u$. We can rewrite it as a function of F using the following result.

We have for $x \geq 0$:

$$F(x) = \mathbb{P}(Y \leq x | X > u) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(u+x) - F(u)}{1 - F(u)} \tag{1.14}$$

The Theorem below is fundamental in extreme value theory because it establishes the asymptotic law of the suitably normalized maximum $X_{n,n}$ of a sample.

Theorem 1.4 *(Fisher and Tippett)*
Let $(X_n)_{n \geq 1}$ be a sequence of independent and identically distributed random variables

of 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 distribution $H_\gamma(x)$ such that:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = H_\gamma(x) \tag{1.15}$$

then except for a translation and a change of scale, we have:

$$H_\gamma(x) = \exp\left\{-(1+\gamma x)_+^{-\frac{1}{\gamma}}\right\} \tag{1.16}$$

where $\gamma \in \mathbb{R}$ and $z_+ = \max(0, z)$

Using the result given in the previous theorem for n large enough we have,

$$F^n(u) \approx \exp\left\{-\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)\right)_+^{-\frac{1}{\gamma}}\right\} \tag{1.17}$$

with $a_n > 0$ and $(b_n, \gamma) \in \mathbb{R}^2$. So,

$$n\log(F(u)) \approx -\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)^{-\frac{1}{\gamma}}\right) \tag{1.18}$$

If u is large enough then a limited expansion gives:

$$\log(F(u)) \approx -(1-F(u))$$

By replacing in the expression (1.18) we obtain for u large enough:

$$1 - F(u) \approx \frac{1}{n}\left(1+\gamma\left(\frac{u-b_n}{a_n}\right)\right)^{-\frac{1}{\gamma}}$$

The same applies to $x > 0$ we have,

$$1 - F(u+x) \approx \frac{1}{n}\left(1+\gamma\left(\frac{u+x-b_n}{a_n}\right)\right)^{-\frac{1}{\gamma}}$$

by replacing in the expression (1.14) we obtain:

$$F(u) \approx 1 - \left(1+\gamma\frac{x}{\sigma}\right)^{-\frac{1}{\gamma}} \tag{1.19}$$

with: $\sigma = a_n + \gamma(u - b_n)$.

The works of Balkema, de Haan and Pickands give a precise result on the approximation of this distribution function when the threshold u is close to the terminal point x_F.

Theorem 1.5 *(Balkema and de Haan [9], and Pickands [111])*
F belongs to the domain of attraction of H_γ if and only if there exist $\sigma > 0$ and $\gamma \in \mathbb{R}$ such that the law of excesses F_u can be uniformly approximated by a generalized Pareto law denoted by $G_{\gamma,\sigma}$ i.e.

$$\lim_{u \to x_F} \sup_{x \in]0; x_F - u[} | F_u(x) - G_{\gamma,\sigma}(x) | = 0 \tag{1.20}$$

where:

$$G_{\gamma,\sigma}(x) = 1 - \left(1 + \gamma \frac{x}{\sigma}\right)^{-\frac{1}{\gamma}} \tag{1.21}$$

The proof of the **Theorem** (1.5) can be found in Embrechts et al. [53].

This approach is commonly called in the literature "the peaks beyond a threshold approach" (*POT* approach, "Peaks-Over-Threshold").

The case $\gamma = 0$ in the expression (1.21) can be seen as the limiting case when $\gamma \to 0$.

We then have:

$$G_{0,\sigma}(x) = 1 - \exp\left(-\frac{x}{\sigma}\right), \quad x \geq 0. \tag{1.22}$$

We find an exponential law of parameter $\frac{1}{\sigma}$. Note also that $G_{-1,\sigma}$ corresponds to the uniform law on $[0, \sigma]$.

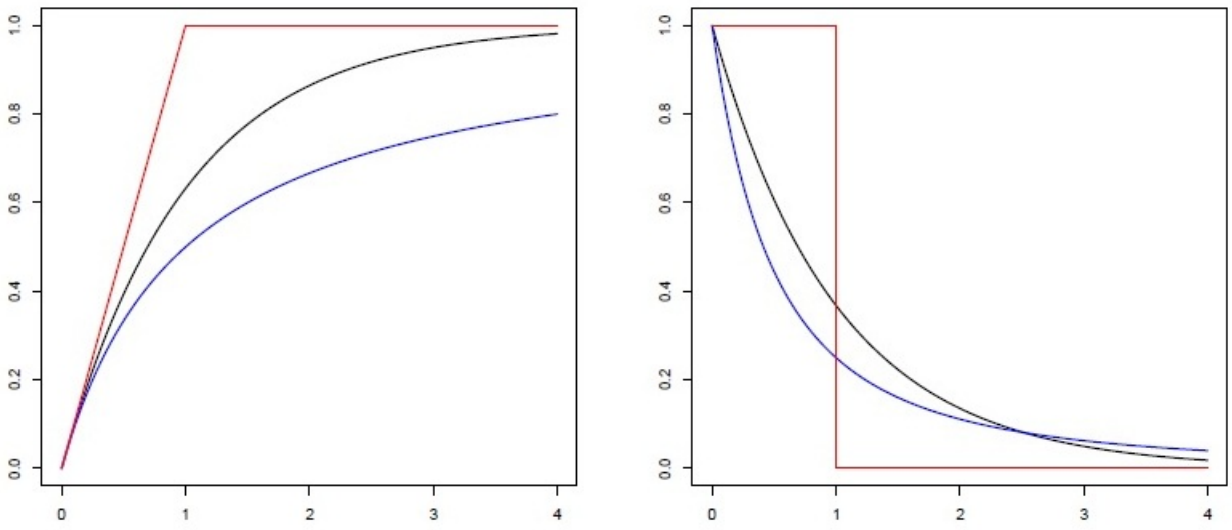

Figure 1.3: Left: $G_{\gamma,1}$. Right: the densities associated with the generalized Pareto law

1.2.3 GEV and GPD distribution

In Fisher and Tippet's theorem, the cdf of the limit is a type of the following three classes:

$$\begin{aligned}
\text{Gumbel} &: \Lambda(x) = \exp(-\exp(-x)), x \in \mathbb{R} \text{ and } \alpha = 0 \\
\text{Fréchet} &: \Phi_\alpha(x) = \begin{cases} 0 \text{ if } x < 0 \\ \exp(-x^{-\alpha}) \text{ if } x \geq 0 \end{cases} \text{ and } \alpha > 0 \\
\text{Weibull} &: \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\} \text{ if } x \leq 0 \\ 1 \text{ if } x > 0 \end{cases} \text{ and } \alpha > 0
\end{aligned}$$

These three distributions Λ, Φ_α et Ψ_α are called "the Standard extreme value distributions" and the corresponding values are "the extreme random variables". Index α is called sometimes the "extreme value index".

1.2.3.1 GEV distribution:

It is difficult to work with three families at the same time, Jenkinson in 1955 shows that these three families can be grouped together in a single form called the family of the laws of generalized extreme values ($\mathcal{GEV}$), Generalized Extreme Value distribution).

Definition 1.5 *The fdr of the family $\mathcal{H}_\gamma$ of generalized extreme values $\mathcal{GEV}$, is for $\gamma \in \mathbb{R}$ and $1 + \gamma x > 0$:*

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\{-(1+\gamma x)^{\frac{-1}{\gamma}}\} \text{ if } \gamma \neq 0 \\ \exp\{-\exp(-x)\} \text{ if } \gamma = 0 \end{cases} \tag{1.23}$$

The parameter γ which appears in the formula (**1.23**) is called "tail index", or "extreme values index, EVI".

For $\gamma = 0$ we must read $\mathcal{H}_0(x) = \exp\{-\exp(-x)\}$, $x \in \mathbb{R}$ which is obtained in the preceding formula by making $\gamma \to 0$. The laws of generalized extreme values correspond to a translation and a change of scale close to the laws of extreme values.

We have, where $\gamma = \frac{1}{\alpha}$, the following correspondances:

$$\begin{aligned}
\Lambda &= \mathcal{H}_0(x), x \in \mathbb{R} \\
\Phi_{\frac{1}{\gamma}} &= \mathcal{H}_\gamma((x-1)/\gamma),\ x > 0 \\
\Psi_{\frac{1}{\gamma}} &= \mathcal{H}_{-\gamma}((x+1)/\gamma),\ x < 0
\end{aligned} \tag{1.24}$$

For the non-centered and unreduced variables, we can write $\mathcal{H}_\gamma(x)$ in a more general form, denoted by $\mathcal{H}_{\mu,\sigma,\gamma}$ in which we reveal a localization parameter $\mu \in \mathbb{R}$

and a scale's parameter $\sigma > 0$. For $(1+\gamma(\frac{x-\mu}{\sigma}) > 0)$ the distribution $\mathcal{H}_{\mu,\sigma,\gamma}(x)$ is written as follows:

$$\mathcal{H}_{\mu,\sigma,\gamma}(x) = \begin{cases} \exp\left\{-\left(1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right)^{\frac{-1}{\gamma}}\right\} \text{ if } \gamma \neq 0 \\ \exp\left\{-\exp\left[-(\frac{x-\mu}{\sigma})\right]\right\} \text{ if } \gamma = 0 \end{cases} \tag{1.25}$$

We can easily show that the density fonction corresponding to $\mathcal{H}_{\mu,\sigma,\gamma}$ for $1 + \gamma\left(\frac{x-\mu}{\sigma}\right) > 0$, is**:**

$$h_{\mu,\sigma,\gamma}(x) = \begin{cases} \frac{1}{\sigma}\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1}{\gamma}}\} \text{ if } \gamma \neq 0 \\ \frac{1}{\sigma}\exp\{-\exp(-x)\} \text{ if } \gamma = 0 \end{cases} \tag{1.26}$$

Remark 1.1 *The quantile $Q(p)$ of the distribution $\mathcal{H}_{\mu,\sigma,\gamma}$ is given by the following formula:*

$$Q(p) = \mathcal{H}^{-1}_{\mu,\sigma,\gamma}(x) = \begin{cases} \mu - \sigma\gamma^{-1}[1-(-\log p)^{-\gamma}] & \text{if } \gamma \neq 0 \\ \mu - \sigma\log(-\log p) & \text{if } \gamma = 0 \end{cases} \tag{1.27}$$

This quantile is therefore strongly influenced by the two parameters σ and γ. Intuitively, we understand that the larger γ, the higher the quantile.

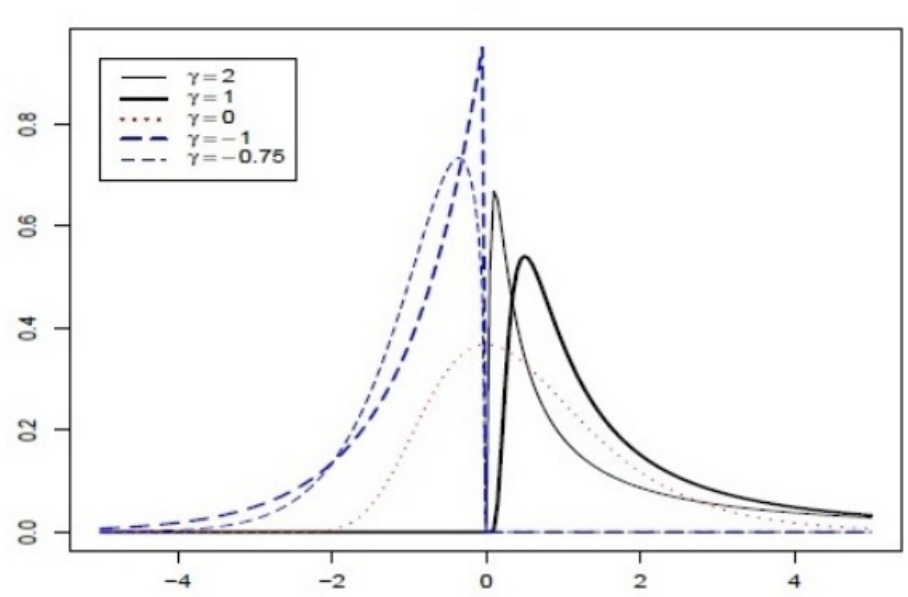

Figure 1.4: Density and Distributions of extreme value distributions

1.2.3.2 GPD distribution

The approach based on the $\mathcal{GEV}$ distribution can be reductive because the use of a single maxima leads to a continuous loss of information in the other large values of the sample.

Definition 1.6 *(Standard generalized **Pareto** distribution)*
The standard generalized Pareto cdf (GPD), denoted by $\mathcal{G}_\gamma$, is defined for $\gamma \in \mathbb{R}$ as follows:

$$\mathcal{G}_\gamma(x) = \begin{cases} \exp\left\{-(1+\gamma x)^{\frac{-1}{\gamma}}\right\} & \text{if } \gamma \neq 0 \\ \exp\{-\exp(-x)\} & \text{if } \gamma = 0 \end{cases} \tag{1.28}$$

with the support:

$$\begin{aligned} x &\geq 0 \qquad \text{if } \gamma \geq 0 \\ 0 &\leq x \leq \frac{-1}{\gamma} \quad \text{if } \gamma < 0 \end{aligned}$$

A general form of $\mathcal{GPD}$, denoted by: $\mathcal{G}_{\gamma,\mu,\sigma}(x) = \mathcal{G}_\gamma(\frac{x-\mu}{\sigma})$ is obtained by replacing the argument x by $(\frac{x-\mu}{\sigma})$ in (1.28) with a support must be, which adjusted accordingly, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively.

Note that the standard $\mathcal{GPD}$ is the case where $\mu = 0$ et $\sigma = 1$. When the location parameter is zero ($\mu = 0$) and the scale parameter is arbitrary ($\sigma > 0$), this distribution plays an important role, in the statistical analysis of extreme events, by providing an appropriate approximation for the excess beyond a large threshold. This special family, denoted by $\mathcal{G}_{\gamma,\sigma}$, is defined as follows:

$$\mathcal{G}_{\gamma,\sigma}(x) = \begin{cases} \exp\left\{-\left(1+\gamma\frac{x}{\sigma}\right)^{\frac{-1}{\gamma}}\right\} & \text{if } \gamma \neq 0 \\ \exp\left\{-\exp\frac{-x}{\sigma}\right\}\} & \text{if } \gamma = 0 \end{cases} \tag{1.29}$$

where

$$\begin{aligned} x &\geq 0 \qquad \text{if } \gamma \geq 0 \\ 0 &\leq x \leq -\frac{\sigma}{\gamma} \quad \text{if } \gamma < 0 \end{aligned}$$

Remark 1.2 *The density of the distribution GPD ($\mathcal{G}_{\gamma,\sigma}$) is written as follows:*

$$g_{\gamma,\sigma}(x) = \begin{cases} \sigma^{-1}\left(1+\gamma\frac{x}{\sigma}\right)^{\frac{-1}{\gamma}-1} & \text{if } \gamma \neq 0 \\ \sigma^{-1}\exp\left\{-\frac{x}{\sigma}\right\} & \text{if } \gamma = 0 \end{cases} \tag{1.30}$$

The quantile $Q(s)$ of the distribution $\mathcal{G}_{\gamma,\sigma}$, which is also the Var at the high confidence level s, is given by:

$$Q(s) = Var(s) = u + \frac{\sigma}{\gamma}\left\{\left(\frac{n}{N_u}s\right)^{-\gamma} - 1\right\} \tag{1.31}$$

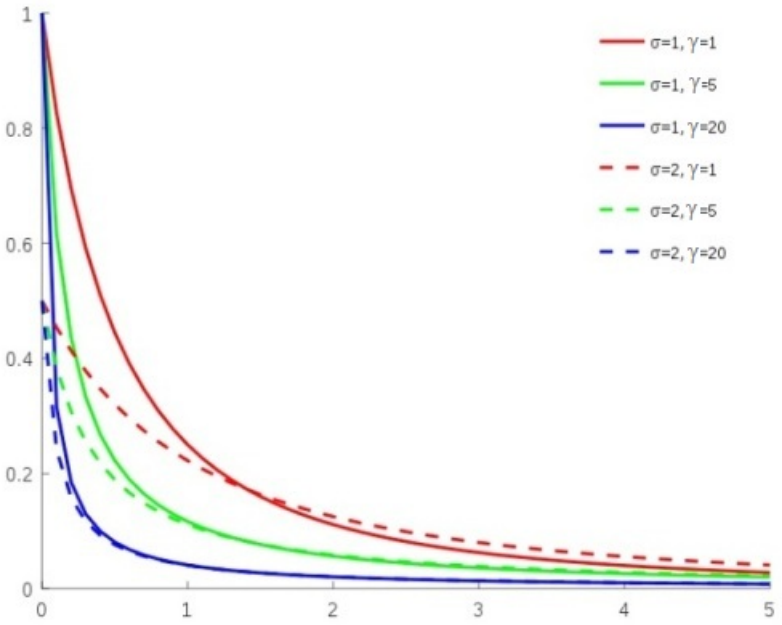

Figure 1.5: GPD distribution functions for $\mu = 0$ and different values of σ and γ.

1.3 Domain of attraction

Now in the following definition, we shall establish necessary and sufficient conditions for a distribution function F to belong to the domain of attraction of $\mathcal{H}_\gamma$.

Definition 1.7 *(Domain of d'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 $F \in D(\mathcal{H}_\gamma)$, if there are two normalizing sequences $(a_n)_{n\geq 1} > 0$ and $(b_n)_{n\geq 1} \in \mathbb{R}$ such that the condition **(1.15)** is verified.*

According to the sign of γ, there are three areas of attraction:

1. If $\gamma > 0$, we say that F belongs to **Fréchet's** D.A., and we note $F \in D(\Phi_\gamma)$; F has an upper-end point on the infinite right ($x_F = \infty$). This domain of attraction is that of heavy-tailed distributions, that is to say, which have a survival function with polynomial decay. Distributions from the Fréchet domain are widely used in mechanical reliability, in climatic phenomena such as meteorology, hydrology, the wind speed recorded continuously in airports and in finance in risk studies.

2. If $\gamma = 0$; we say that F belongs to **Gumbel's** D.A., and we note $F \in D(\Lambda)$, the upper-end point x_F can then be finite or not. This domain of attraction

is that of distributions with light tails, that is to say which have an exponentially decreasing survival function. These distributions are often used to make predictions in environmental events such as earthquake, hydrology (floods, destruction of dams), etc.

3. If $\gamma < 0$, we say that F belongs to **Weibull's** D.A., and we note $F \in D(\Psi_\gamma)$; F has a finite right endpoint ($x_F < \infty$): This domain of attraction is that of survival functions whose support is superiorly bounded. **Weibull** type distributions are often used to describe the mechanical strength of a material or the operating time of an electronic or mechanical device.

The following tables give different examples of standard distributions in these three attraction domains [as in Tables 1.1 to 1.3 in **Embrechts et al.**(1997 [53]).

Table 1.1: Some Distributions Associated With The Positive Index

Distributions	$\overline{F}(x)$	γ
Pareto (α), $\alpha > 0$	$x^{-\alpha}$, $x > 1$	$\frac{1}{\alpha}$
Burr (β, τ, λ) $\beta > 0$, $\tau > 0$, $\lambda > 0$	$(\frac{\beta}{\beta + x^\tau})^\lambda$	$\frac{1}{\lambda\tau}$
Fréchet $(\frac{1}{\alpha})$, $\alpha > 0$	$1 - \exp(-x^{-\alpha})$	$\frac{1}{\alpha}$
Log-gamma (m, λ), $m > 0$, $\lambda > 0$	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty \log(u)^{m-1} u^{-\lambda-1} du$	$\frac{1}{\lambda}$
Loglogistic (β, α), $\beta > 0$,$\alpha > 1$	$\frac{1}{1+\beta x^\alpha}$	$\frac{1}{\alpha}$

Table 1.2: Some Distributions Associated With The Negative Index

Distributions	$\overline{F}(x)$	γ
Uniform $(0, 1)$	$1 - x$	-1
Reverse Burr $(\beta, \tau, \lambda, x_\tau)$ $\beta, \tau, \lambda > 0$	$(\frac{\beta}{\beta + (x_F + x)^{-\tau}})^\lambda$	$-\frac{1}{\lambda\tau}$

Table 1.3: Some Distributions Associated With a Zero Index

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

These domains of attraction are characterized by functions with regular variations, so we have to define the notions of functions with regular variations and functions with slow variations which will be use later.

1.3.1 Regular variations

The concept of regular variation is widely used in extreme value theory to describe the deviation from pure power laws. Regular variation of the tails of a distribution appears as a condition in various theoretical results of probability theory, so in domain of attraction. In this section, we summarize some of the main results of regular variation theory, we present the basic properties of the aforementioned functions which will be used in the following. In general, regularly varying functions are functions which behave asymptotically like power functions.

An encyclopedic treatment of regular variation can be found in Bingham et al(1987 [17]). To describe the functions with regular variations in more detail, it is necessary to start with a definition of the functions with slow variations.

Definition 1.8 *We say that a function $L(.)$ is slowly varying at infinity if $L(x) > 0$ for x large enough and if for all $\lambda > 0$, we have:*

$$\lim_{x\to\infty} \frac{L(\lambda x)}{L(x)} = 1 \tag{1.32}$$

Among the slowly varying functions, we can cite:
- constant functions,
- the functions having a strictly positive limit at infinity,
- functions l such as:

$$\exists M > 0, \forall x \geq M \; l(x) = c + dx^{-\beta}(1 + 0(1))$$

where $c, \beta > 0$ and $d \in \mathbb{R}$. The set of these functions l is called "Hall class".

Theorem 1.6 ***(Karamata representation)***
All slowly varying $L(.)$ functions are written in the form :

$$L(x) = c(x)\exp\left(\int_1^x \frac{\Delta(u)}{u}du\right), \quad (x \geq 1) \tag{1.33}$$

where $c(x) \to c > 0$ and $\Delta(x) \to 0$ when $x \to \infty$. This formula of slowly varying functions is called "Karamata representation".

Proof. For a demonstration see **Resnick** (2007 [117]) Corollaire 2:1; page 29. ■

In the case where the function $c(.)$ is constant, the corresponding $L(.)$ fonction is said to be **normalized**. If the function L is **normalized** then it is differentiable from derivative $\grave{L}$ defined for all x by:

$$\grave{L}(x) = \frac{\Delta(x)L(x)}{x}$$

In particular, we have:

$$\lim_{x\to\infty} x\frac{\grave{L}(x)}{L(x)} = 0$$

Definition 1.9 *We say that a function $U(.)$ has regular variations of index $\rho \in \mathbb{R}$ at infinity, which we will denote by $U(.) \in \mathcal{RV}_\rho$, if U is positive at infinity (i.e. if there exists A such that for all $x > A$, $U(x) > 0$) and if for all $\lambda > 0$,*

$$\lim_{x\to\infty} \frac{U(\lambda x)}{U(x)} = \lambda^\rho \tag{1.34}$$

In the particular case where $\rho = 0$, $U(.)$ is a slowly varying function at infinity. We can easily show that any function with regular variations of index ρ can always be written in the form:

$$U(x) = x^\rho.L(x) \tag{1.35}$$

where L is a slowly varying function at infinity.

On the other hand, if $\rho = \infty$, we speak of a "**function with rapid variations at infinity**".

Lemma 1.1 *(Inverse of a function with regular variations)*
-If U has regular variations of index $\rho > 0$, then $U^{\leftarrow}(x)$ has regular variations of index $1/\rho$.
-If U has regular variations of index $\rho < 0$, then $U^{\leftarrow}(1/x)$ has regular variations of index $-1/\rho$.

For the proof of Lemma 1.1, we can refer to **Bingham et al.(1987 [17])**, Theorem 1.5.12 or Proposition 2.6 of **Resnick's** book (1987).

Lemma 1.2 *(**Resnick** (1987), **Proposition 0.5**)*
If U is a function with regular variations of index ρ at infinity then, for all $0 < a < b$,

$$\lim_{x\to\infty} \sup_{\lambda\in[a,b]} \left| \frac{U(\lambda x)}{U(x)} - \lambda^\rho \right| = 0 \tag{1.36}$$

Lemma 1.3 *(**de Haan and Ferreira** (2006), **Proposition B.1.9**)*
Let L be a slowly varying function at infinity and let a_n and b_n be two positive sequences which converge towards infinity as n approaches infinity. If $a_n \sim b_n$ (i.e. $\frac{a_n}{b_n} \to 1$ when $n \to \infty$), then:

$$L(a_n) \sim L(b_n) \text{ (i.e } \frac{a_n}{b_n} \to 1 \text{ when } n \to \infty) \tag{1.37}$$

The lemma (1.3) which is a consequence of the lemma (1.1) shows that the slowly varying function keep the equivalents. Consequently, this result implies that if U is regularly changing with index ρ and if $a_n \sim b_n$ then $U(a_n) \sim U(b_n)$ and we say that the functions with regular variations also keep the equivalent.

1.3.2 Potter's inequality

An important result of the extreme value analysis of infinitely variable functions is Potter's inequality.

Proposition 1.1 *Suppose that g be a regularly varying function with index ρ at infinity. Then for any real A strictly greater than* 1 *and for any ε strictly positive, there exists M such that:*

$$\forall x \geq M, \forall y \geq M : \ \frac{g(x)}{g(y)} \leq A \max\left\{\left(\frac{x}{y}\right)^{\rho+\varepsilon}, \left(\frac{x}{y}\right)^{\rho-\varepsilon}\right\} \tag{1.38}$$

This inequality can also be written:

$$\forall \varepsilon > 0, \exists t_0 : \forall x \geq 1, \forall t \geq t_0 : (1-\varepsilon)x^{\rho-\varepsilon} < \frac{g(tx)}{g(x)} < (1+\varepsilon)x^{\rho+\varepsilon} \tag{1.39}$$

Potter's bounds are very useful for the use of the dominated convergence theorem in the case of studies of integrations of functions with regular variations. For more other results on the theory of regularly varying functions, we referred the reader to Bingham et al. (1987 [17]).

With the help of the various results presented on functions with regular variations and slow variations to infinity, we will be able to characterize the different domains of attraction. Knowing the F distribution, we would like to know its domain of attraction and its normalization constants. We will indicate here the most widely used criteria, that is to say, the necessary and sufficient conditions on the fdr F for it to belong to one of the three domains of attraction that are defined above.

1.3.3 Fréchet's domain of attraction

Recall that the Fréchet domain of attraction contains the laws whose survival function is polynomial decat, i.e. the heavy-tailed laws or Pareto-type laws. The laws of this domain have an infinite upper endpoint x_F. Indeed, the result below stated by **Gnedenko[69]** and of which one will find a simple proof in the book of **Resnick** [Proposition 1.11]] ensures that any function belongs to the **Fréchet** domain of attraction is a function with regular variations and vice versa.

Theorem 1.7 *(characterization of $\mathcal{D}(\Phi_\gamma)$)*
A distribution function $F(.)$ belongs to the Fréchet domain of attraction, $F \in \mathcal{D}(\Phi_\gamma)$,with an index of the extreme values $\gamma > 0$ if and only if the upper-endpoint is finite (i.e. $x_F = +\infty$), and its survival function [2] $\overline{F}$ has regular variations of index $-1/\gamma$ ($\overline{F}(.) \in \mathcal{RV}_{-1/\gamma}$), that is to say:

$$\overline{F}(x) = x^{-1/\gamma} L(x)$$

[2] A brief study on the survival function will be given in the next chapter.

In this case, a possible choice of standardization sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ is:

$$a_n = F^{\leftarrow}\left(1-\frac{1}{n}\right) \text{ and } b_n = 0$$

where $F^{\leftarrow}$ is the generalized inverse of F.

From this theorem, we deduce that $F \in \mathcal{D}(\Phi_\gamma)$ if and only if the upper-endpoint x_F is infinite and $\overline{F}(x) = x^{-1/\gamma}L(x)$, where L is a slowly varying function infinite.

Fréchet's domain of attraction brings together a great diversity of laws including among them usual laws (Student's law, Chi-square law, Log-gamma law, Fréchet's law). It is therefore subject to numerous applications, in particular Gardes and Girard (2010 [62]) and Daouia et al. (2011 [33]).

1.3.4 Weibull's domain of attraction

All the laws belonging to the Weibull's domain of attraction have a finite upper endpoint x_F. The following result shows that we go from the Fréchet's domain of attraction to that of Weibull's by a simple change of variable in the distribution function.

Theorem 1.8 *(characterization of $\mathcal{D}(\Psi_\gamma)$)*
A distribution function $F(.)$ belongs to the Weibull's domain of attraction, $F \in \mathcal{D}(\Psi_\gamma)$, with $\gamma < 0$ if and only if $x_F < +\infty$ and in addition $1-F$ is a function with regular variations of index $-1/\gamma$, defined:

$$\overline{F}(x) = \left(x_F - \frac{1}{x}\right) = x^{-1/\gamma}L(x)$$

whith $L(.)$ is a slowly varying function at infinity.

In this domain of attraction the normalization sequences are determined as follows:

$$a_n = x - U(n) = x_F - F\left(1-\frac{1}{n}\right) \text{ and } b_n = x_F$$

the sequence $(a_n^{-1}(X_{n;n} - x_F))$ converges in law to a rv of cdf Ψ_γ when $n \to \infty$.

From Theorem 1.8, we deduce that $F \in \mathcal{D}(\Psi_\gamma)$ if and only if the upper-endpoint x_F is finite and

$$\overline{F}(x) = (x_F - x)^{-\frac{1}{\gamma}}L[(x_F - x)^{-1}],$$

with L is a slowly varying function at infinity and γ a strictly negative real.

This domain of attraction was considered by Gardes) (2010 [62]) to give an endpoint estimator of a distribution.

1.3.5 Gumbel's domain of attraction

The Gumbel's domain of attraction is more difficult to treat, since there is no direct linkage between the tail and the regular variation notion such as the domains of attraction of Fréchet and Weibull. We will find the extensions of the regular variation that take into account a complete characterization of $\mathcal{D}(\Lambda)$. The Gumbel class contains the exponential, normal, lognormal, gamma and classical Weibull distributions.

The Gumbel's domain of attraction contains the laws whose survival function is exponentially decreasing, i.e. the laws with light tails. Unlike the other two domains, there is no simple representation for the laws belonging to Gumbel's domain of attraction. It can be described from **Von Mises** functions type.

Definition 1.10 *(von Mises function).*
The df F is called a von Mises function with auxiliary function a if there exists some $z < x_F$ such that:

$$\overline{F}(x) = c.\exp\left(-\int_z^x \frac{dt}{a(t)}\right), \quad z < x < x_F < \infty \tag{1.40}$$

where $c > 0$ is some positive constant, and a is a positive absolutely continuous function (with respect to Lebesgue measure) with density $\grave{a}$ satisfying: $\lim_{x \to x_F} \grave{a}(x) = 0$.

As an example of the von Mises function, the exponential distribution function with parameter λ, $\overline{F}(x) = \exp\{-\lambda x\}$, the auxiliary function is $a(x) = 1/\lambda$.

Proposition 1.2 *(von Mises function's properties)*
Let F be a von Mises function with auxiliary function a. Then
- F is absolutely continuous on $(z; x_F)$ with positive pdf f. The auxiliary function can be chosen as $a(x) = \overline{F}(x)/f(x)$
- If $x_F = \infty$, then $\overline{F} \in \mathcal{RV}_{-\infty}$ and $\lim_{x \to x_F} \frac{xf(x)}{\overline{F}(x)} = \infty$
- If $x_F < \infty$, then $\overline{F}(x_F - x) \in \mathcal{RV}_{-\infty}$ and $\lim_{x \to x_F} \frac{(x_F - x)f(x)}{\overline{F}(x)} = \infty$

Theorem 1.9 *A distribution function $F(.)$ belongs to the Gumbel's domain of attraction, $F \in \mathcal{D}(\Lambda)$ if and only if:*

$$\overline{F}(x) = c(x)\exp\{-\int_z^x \frac{g(t)}{a(t)}dt\} \quad z < x < x_F$$

where g and c are some positive functions, such that: $c(x) \to c > 0$, $g(x) \to 1$ as $x \to x_F$, and $a(x)$ is a positive,absolutely continuous function (with respect to Lebesgue measure) with density $\grave{a}$ having $\lim_{x \to x_F} \grave{a}(x) = 0$. In this case, we can choose for the normalization

sequences $(a_n)_{n\geq 1}$ *and* $(b_n)_{n\geq 1}$:

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

Let us conclude on the behavior of the distribution tails:

• The Fréchet's domain of attraction $\mathcal{D}(\Phi_\gamma)$ contains all distributions characterized by a tail with polynomial decay at infinity, and an infinite x_F endpoint. They are also said to have heavy tails.

• The Gumbel's domain of attraction $\mathcal{D}(\Lambda)$ contains all distributions characterized by an infinitely exponentially decreasing tail. Examples of such distributions are exponential distributions with slightly heavy (light) tails and thicker than Gaussian.

• The Weibull's domain of attraction $\mathcal{D}(\Psi_\gamma)$ contains all distributions with right-bound supports i.e. $x_F < \infty$. They are also said to have bounded tails.

1.4 Estimating extreme quantiles

In what follows, we assume that F belongs to one of the attraction domains defined above. In order to summarize the estimation problem investigated in this work, we introduce the following result called the Poisson approximation.

Lemma 1.4 *If* $\alpha_n \to 0$ *and* $n\alpha_n \to c$ *(not necessarily finished) when* $n \to \infty$, *so:*

$$\mathbb{P}(X_{n,n} < q_{\alpha_n}) \to \exp(-c)$$

Thus, according to the previous Lemma, two situations can then be distinguished as a function of c when we want to estimate the quantiles of order α_n when $n \to \infty$.

First, if $c = \infty$ so, $\mathbb{P}(X_{n,n} < q_{\alpha_n}) = 0$. In such a context, a natural estimator of q_{α_n} is the empirical quantile which is nothing more than the $[n\alpha_n]$th largest observation of the sample $\{X_1, X_2, ..., X_n\}$ that is to say the statistic of order $X_{n-[n\alpha_n]+1,n}$.

Second, if $c = 0$ so, $\mathbb{P}(X_{n,n} < q_{\alpha_n}) = 1$. Therefore, we can't estimate the quantile empirically. To solve this problem, we have listed three main categories of methods:

- The extreme value theory presented by Guida and Longo (1988 [70]) and whose first bibliographic elements go back to Fisher and Tippet (1928 [55]) and Gnedenko (1943 [69]) consists in dividing the sample into m_0 disjoint subgroups of size $n_0 = n/m_0$ from which the maxima are determined. The law of these maxima is then approximated, for n_0 large enough, by a law of

extreme values. Using the relation

$$\mathbb{P}(\max(X_1,...,X_n < q_{\alpha_n})) = F^n(q_{\alpha_n})$$

we can then estimate the extreme quantile q_{α_n}.

- The excess method initially presented by Pickands (1975) [111]. It recommends to retain only the observations exceeding a fixed threshold u. The law of k_n observations thus retained, denoted by $\{Y_i,\ i = 1,...,k_n\}$ can be approximated, if u is large enough by a generalized Pareto distribution ($\mathcal{GPD}$). To estimate the extreme quantile q_{α_n}, it suffices to use the result of (Balkema and de Haan, 1974 [9]; Pickands, 1975 [111]) which establishes the equivalence between the convergence in the law of the maximum towards a law of extreme values and the convergence in law of an excess towards a $\mathcal{GPD}$.

- Semi-parametric methods where we assume that for all $\gamma > 0$ we have $\mathbb{P}(X > x) \sim x^{-\frac{1}{\gamma}}$ as x tends to infinity, that is to say that the survival function $\overline{F}(x)$ decreases in $x^{-\frac{1}{\gamma}}$. This assumption makes it possible to construct nonparametric estimators of the parameter γ, the most famous of which is the Hill (1975) [79] estimator. Based on this result, Weissman (1978) [132] proposed three years later an estimator of the extreme quantile q_{α_n}. Indeed, supposing that $\mathbb{P}(X > x) \sim x^{-\frac{1}{\gamma}}$ amounts to supposing that the quantile q_{α_n} decreases in α.

1.4.1 Extreme quantile approach by the law of extreme values

To estimate the quantile q_{α_n}, use the approximation $\mathbb{P}(X_{n,n} \leq a_n x + b_n) = F(a_n x + b_n) \simeq \mathcal{H}_\gamma(x)$. Indeed, according to Theorem 1.4 (Fisher and Tippet's theorem), we can write

$$\lim_{n\to\infty} n \log F(a_n x + b_n) = \log \mathcal{H}_\gamma(x)$$

still

$$\lim_{n\to\infty} n \log[1 - \overline{F}(a_n x + b_n)] = \log \mathcal{H}_\gamma(x)$$

when $n \to \infty$, we can show that $a_n x + b_n \to x_F$ and consequently $\overline{F}(a_n x + b_n)$ converges to 0. A first-order limited expansion of $\log(1 + u)$ therefore gives

$$\overline{F}(a_n x + b_n) \sim -\frac{1}{n} \log \mathcal{H}_\gamma(x)$$

for any γ, we can then approach the quantile q_{α_n} by:

$$q_{\alpha_n} \simeq a_n x_{\alpha_n} + b_n \text{ where } x_{\alpha_n} \text{ checks } -\log \mathcal{H}_\gamma(x_{\alpha_n}) = n\alpha_n$$

We then have an extreme quantile estimator of type

$$\begin{aligned} \widehat{q}_{\alpha_n} &= \hat{a}_n x_{\alpha_n} + \hat{b}_n \\ &= \begin{cases} \hat{a}_n (n\alpha_n)^{-\hat{\gamma}} + \hat{b}_n & \text{if } F \in \mathcal{D}(\Phi_\gamma) \\ -\hat{a}_n (n\alpha_n)^{-\hat{\gamma}} + \hat{b}_n & \text{if } F \in \mathcal{D}(\Psi_\gamma) \\ -\hat{a}_n \log(n\alpha_n) + \hat{b}_n & \text{if } F \in \mathcal{D}(\Lambda) \end{cases} \end{aligned} \tag{1.41}$$

where $(\hat{a}_n, \hat{b}_n)$ and $\hat{\gamma}$ are respectively estimators of the sequences (a_n, b_n) and the tail index γ.

In the particular case where $\gamma = 0$, the authors propose to use the approach based on the GEV law, the result of which is stated as follows.

Theorem 1.10 *(Weinstein [131])*
Let $F \in \mathcal{D}(\Lambda)$, there exist two sequences $(a_n) > 0$ and $(b_n) \in \mathbb{R}$ such that for all $x \in \mathbb{R}$ and $v > 0$,

$$\lim n\overline{F}[(b_n^v + c_n x)^{1/v}] = \exp(-x) \tag{1.42}$$

where $c_n = a_n v b_n^{v-1}$.

In such a situation, we approach the quantile by

$$\widehat{q}_{\alpha_n} \simeq (b_n^v + c_n x_{\alpha_n})^{1/v} \text{ where } x_{\alpha_n} \text{ checks } \exp(-x_{\alpha_n}) = n\alpha_n$$

and an estimator of the extreme quantile is obtained by replacing the sequences b_n and c_n respectively by their estimators $\hat{b}_n$ and $\hat{c}_n$, i.e.

$$\widehat{q}_{\alpha_n} \simeq (\hat{b}_n^v + \hat{c}_n \log(n\alpha_n))^{1/v} \tag{1.43}$$

the advantage of using the result (1.42) comes from the fact that there are values of v for which the convergence in (1.42) is faster than in the case $v = 1$. In this case, on simulation, the approximation of the quantile q_{α_n} is of better quality than the approximation based on the EVD approach, i.e with $v = 1$. The authors provide the optimal value of the parameter v.

The parameters γ, a_n, b_n and c_n of these distributions can be estimated by the maximum likelihood method, or the weighted moment method (Hosking et al., 1985 [82]). In the case of the EVD approach, Smith (1985 [122]) makes a detailed study of the asymptotic behavior of the estimators of the parameters γ, a_n, b_n obtained by the maximum likelihood method. However, it is advisable to use the weighted moment estimators because they are not only explicit and easy to calculate but also because they give better results than the maximum likelihood estimators when we have small medium samples. The main difficulty in estimating the

parameters γ, a_n, b_n and c_n is due to the fact that a sample of maxima is required, which is sometimes difficult to extract from the initial data.

1.4.2 Approaching extreme quantiles using the excess method

Before presenting this approach, it is useful to start with a definition.

Definition 1.11 *Generalized Pareto Distribution ($\mathcal{GPD}$).*
The Generalized Pareto Distribution is the law whose distribution function is given by:

$$\mathcal{G}_{\gamma,\sigma}(x) = \begin{cases} 1-(1+\gamma\frac{x}{\sigma})^{\frac{-1}{\gamma}}\} \text{ if } \gamma \neq 0 \text{ and } \sigma > 0 \\ 1-\exp(-\frac{x}{\sigma}) \text{ if } \gamma = 0 \text{ and } \sigma > 0 \end{cases}$$

with

$$\begin{array}{rcl} x & \geq & 0 \qquad \text{if } \gamma \geq 0 \\ 0 & \leq & x \leq \dfrac{-\sigma}{\gamma} \quad \text{if } \gamma < 0 \end{array}$$

In the previous expression, σ represents the scale parameter and γ the shape parameter: this is the same shape parameter introduced in section 1.2 and which is called the extreme value index.

The $\mathcal{GPD}$ distribution has some particularities. Here is a non-exhaustive list :

- If $\sigma = 1$, we are talking about the standard $\mathcal{GPD}$ distribution (see definition 12).
- If $\gamma = 0$, the $\mathcal{GPD}$ corresponds to an exponential law of expectation σ.
- If $\gamma = -1$, it corresponds to a uniform law on $[0, \sigma]$.
- If $\gamma > 0$, we find the decentered Pareto law.

In this approach to estimating extreme quantiles, only observations exceeding a fixed threshold $u < x_F$ are retained. We then define the excess Y of the variable X above the threshold u by $X - u$ knowing $X > u$. If we denote by F_u the distribution function of an excess above the threshold u, we have for all $y > 0$

$$\begin{array}{rcl} 1 - F_u(y) & = & \mathbb{P}(Y > y) \\ & = & \mathbb{P}(X - u > y | X > u) \\ & = & \dfrac{\mathbb{P}(X > u + y, X > u)}{\mathbb{P}(X > u)} \\ & = & \dfrac{1 - F(u + y)}{1 - F(u)} \end{array}$$

When the threshold u is large, we can approach this quantity by the survival function of a $\mathcal{GPD}$ law. In order to approach the quantile, it suffices to use the result of Balkema, de Haan (1974) [9] and Pickands (1975) [111] which establishes the equivalence between the convergence in the law of the maximum towards a law of extreme values $\mathcal{H}_\gamma$ and the convergence in law of an excess towards a $\mathcal{GPD}$, what was mentioned previously in Theorem 1.5.

From this result, if for an unknown distribution function F, the sample of normalized maxima converges in law towards a non-degenerate distribution, then it follows that the distribution of the excesses above a high threshold converges towards a $\mathcal{GPD}$ when the threshold tends towards the upper limit of the support of F. This characterization is the basis of the Peaks Over Threshold (POT) type estimation methods.

Like $1-F(u+y)=[1-F(u)][1-F_u(y)]$, if for all $y\geq 0$ we set $q_{\alpha_n}=u+y$, then

$$\begin{aligned}\alpha_n &= 1-F(q_{\alpha_n})=[1-F(u)][1-F_u(q_{\alpha_n}-u)]\\ &\simeq [1-F(u)](1-G_{\gamma,\sigma}(q_{\alpha_n}-u))\end{aligned}$$

for k_n excess above the threshold u, the approximation $1-F(u)\simeq k_n/n$ leads to

$$\frac{k_n}{n}(1-G_{\gamma,\sigma}(q_{\alpha_n}-u))\simeq\alpha_n$$

and if $\gamma\neq 0$, then we approach the quantile by

$$q_{\alpha_n}\simeq u+\frac{\sigma}{\gamma}((\frac{k}{n\alpha_n})^\gamma-1)$$

We then have an estimator of the type

$$\widehat{q_{\alpha_n}}\simeq\frac{(\frac{k}{n\alpha})^{\hat\gamma}-1}{\hat\gamma}\hat\sigma+u \tag{1.44}$$

where $\hat\gamma$ and $\hat\sigma$ are respectively estimators of the shape and scale parameters. We can note the similarity between the quantile estimator (1.44) and the expression of the quantile (1.41) with $\hat\sigma=\hat a_n$ and $u=\hat b_n$.

The parameters γ and σ of the $\mathcal{GPD}$ can be estimated by the method of moments, the method of weighted moments (Hosking and Wallis, 1987 [83]) or the method of maximum likelihood (Smith, 1987 [123]; Davison and Smith, 1990 [37]).

This method has an advantage over the previous one in that it is easier to have a sample of excess than of maxima. In practice, we replace u with $X_{n-k_n+1,n}$ that is the k_n largest observation of the sample $\{X_i, i=1,...,n\}$.

Two variants of this method have been presented by Breiman et al. (1990 [19]) under the names "Exponential tail" (**ET**) and "Quadratic tail" (**QT**).

1.4.3 Extreme quantile approach by semi-parametric method

We restrict to the functions $F \in \mathcal{D}(\Phi_\gamma)$ for which we have the following characterization

$$\bar{F}(x) = x^{-1/\gamma}\ell(x)$$

with ℓ a slowly varying function at infinity and $\gamma > 0$. According to "Lemma 18", this characterization implies that

$$\begin{aligned} q_{\alpha_n} &= \bar{F}^{\leftarrow}(\alpha_n) = \alpha_n^{-\gamma} L(1/\alpha_n) \text{ with } \alpha_n \leq 1/n \\ q_{\beta_n} &= \bar{F}^{\leftarrow}(\beta_n) = \beta_n^{-\gamma} L(1/\beta_n) \text{with } \beta_n \geq 1/n \end{aligned}$$

where L is a slowly varying function at infinity. Regarding the L and ℓ functions, it seems important to point out here that it is not the same slowly varying function.

Given the definition of a slowly varying function (refer to Definition 15), for β_n small enough, we have

$$\bar{F}^{\leftarrow}(\alpha_n) = \bar{F}^{\leftarrow}(\beta_n)(\frac{\beta_n}{\alpha_n})^{\gamma}$$

By replacing $\bar{F}^{\leftarrow}(\beta_n)$ and γ by estimators, we obtain the Weissman (1978) [132] estimator defined by:

$$\hat{q}^{W}_{\alpha_n} = X_{n-[n\beta_n]+1,n}(\frac{\beta_n}{\alpha_n})^{\hat{\gamma}}$$

For the properties of the Weissman estimator, one can refer to the work by Embrechts et al. (1997) [53].

As another estimator of the extreme quantiles, we can cite the one obtained by the approximation

$$\bar{F}^{\leftarrow}(\alpha_n) \simeq \frac{(\beta_n/\alpha_n)^{\gamma} - 1}{1 - 2^{-\gamma}}(\bar{F}^{\leftarrow}(\beta_n) - \bar{F}^{\leftarrow}(2\beta_n)) + \bar{F}^{\leftarrow}(\beta_n),$$

and valid regardless of the domain of attraction of the function F. The asymptotic normality of the resulting extreme quantile estimator, i.e.

$$\hat{q}^{DH}_{\alpha_n} = \frac{(\beta_n/\alpha_n)^{\hat{\gamma}} - 1}{1 - 2^{-\hat{\gamma}}}(X_{n-[n\beta_n]+1,n} - X_{n-[2n\beta_n]+1,n}) + X_{n-[n\beta_n]+1,n},$$

was established by Dekkers and de Haan (1989 [38]). It clearly appears that this

extreme quantile estimator can be put in the form (1.44) and therefore (1.41) with

$$\hat{a}_n = \frac{\hat{\gamma}}{1-2^{-\hat{\gamma}}}(X_{n-[n\beta_n]+1,n} - X_{n-[2n\beta_n]+1,n})$$

and

$$\hat{b}_n = X_{n-[n\beta_n]+1,n}.$$

1.5 Tail index estimators

Estimating parameters constitutes an important task in extreme values theory, since it is a starting point for statistical inference about extreme values of a population. In particular, the extreme value index (EVI) or tail index, measures the right tail's weight of the df F, allowing us to describe the behavior of the extreme values of a population. With the estimated EVI, it is possible to estimate other parameters of extreme events like the extreme quantile, the return period and the probability of exceedance of a hight threshold. There are two approaches : a parametric approach and a semi-parametric approach. In the parametric case, the most widely used estimation methods are: the maximum likelihood (ML) method and the method of moments (MM). For the semi-parametric approach we find several techniques, we cite the most famous: the Hill estimator (1975)[79] and the Pickands estimator (1975) [111].

Let $X_1, X_2, ..., X_n$ be a sequence of random variables i.i.d. whose distribution function F and denote by $X_{1;n}, X_{2;n}, ..., X_{n;n}$ the order statistics associated with this sample. We see that the asymptotic behavior of the maxima $X_{n,n}$ (after renormalization) is well known and can be modeled by the distribution of extreme values having for parameter γ. This key parameter is called "index of extreme values" or "index tail". It measures the weight of the right tail of the underlying fdr F, which allows us to understand and describe the behavior of extreme values of a population. According to its sign, three domains of attraction are possible for F: Fréchet ($\gamma > 0$), Gumbel ($\gamma = 0$) and Weibull ($\gamma < 0$).

Therefore, estimating this parameter is an important task in extreme value theory (tail index estimation is important for many aspects), since it is a starting point for statistical inference on the extreme values of a population. With its estimation, it is possible to estimate other extreme event parameters like the right end point x_F of the underlying fdr F, the extreme quantiles, the return period and the probability of exceeding a high level as well, provided that the existence of moments.

Depending on the importance, there is a large literature on tail index estimation. Knowledge of γ is therefore necessary to solve a number of problems in extreme value analysis, such as estimating extreme quantiles of X, which has made its esti-

mation a central topic in the literature. The first two estimators of this parameter were proposed in 1975 by Hill [79] and Pickands [111]. Then other estimators were suggested, such as the maximum likelihood estimator or the moments estimator (de Haan and Ferreira, 2006 [45]). Drees and Kaufmann (1998 [51]) showed that the estimators of γ are, in general, regular functions of large order statistics. These estimators are called semi-parametric estimators. In the following, we present some classic estimators for the extreme value index γ.

1.5.1 Maximum likelihood estimator

A widely used and flexible approach for parameter estimation is maximum likelihood, it is the first method that remains the most popular and which under certain conditions is the most effective. The aim of this approach is to obtain the set of parameter estimates for which the joint probability density of the observed data is maximised. In practice, the loglikelihood is maximised with respect to θ to obtain the maximum likelihood estimate (MLE) $\widehat{\theta}$.

The maximum likelihood estimator is built from observations of the maxima, it is about estimating the index of extreme values as well as the two normalizing sequences a_n and b_n:

Let $X_1, X_2, ..., X_n$ be a n-sample, the X_i are i.i.d. of density h_θ where $\theta = (\mu, \sigma, \gamma)$. The expression of the likelihood function is given by:

$$\mathcal{L}(\theta = (\mu, \sigma, \gamma); (X_1, ..., X_n)) = \prod_{i=1}^{n} h_\theta(x_i) \tag{1.45}$$

The estimator $\widehat{\theta}$ is given by solving the following system:

$$\begin{cases} \frac{\partial \log \mathcal{L}}{\partial \theta} = 0 \\ \frac{\partial^2 \log \mathcal{L}}{\partial^2 \theta} < 0 \end{cases} \tag{1.46}$$

Example, in the case $\gamma = 0$ (Gumbel's law), the log likelihood function is given:

$$\log \mathcal{L}(\theta = (\mu, \sigma, \gamma); (X_1, ..., X_n)) = -n \log \sigma - \sum_{i=1}^{n} \exp(-\frac{x_i - \mu}{\sigma}) - \sum_{i=1}^{n} \exp(\frac{x_i - \mu}{\sigma})$$

We derive this function relative to the two parameters, we obtain the following system of equations to be solved:

$$\begin{cases} \frac{\partial \log \mathcal{L}}{\partial \sigma} = 0 \Leftrightarrow n + \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} [\exp(-\frac{x_i - \mu}{\sigma} - 1)] = 0 \\ \frac{\partial \log \mathcal{L}}{\partial \mu} = 0 \Leftrightarrow n - \sum_{i=1}^{n} \exp(\frac{x_i - \mu}{\sigma}) = 0 \end{cases}$$

solving this system is relatively difficult and does not generally admit explicit solutions.

The maximum likelihood estimator is not straightforward. Jenkinson proposed an iterative algorithm for maximizing the likelihood function. The corresponding Newton-Raphson algorithm is given in Hosking[81] and improved in Macleod[99].

The asymptotic properties of the maximum likelihood estimator were studied by Smith in 1985 [122]. He shows that if $\gamma > -1/2$, we have the consistency, efficiency and asymptotic normality of these estimators. Zhou(2009 [136]) and Dombry (2013 [48]) proved that the maximum likelihood estimator exists and is consistent for $\gamma > -1$. Then in 2010, Zhou also obtained the asymptotic normality for $-1 < \gamma < -1/2$ and proved that it is not consistent for $\gamma < -1$.

1.5.2 Pickands estimator

The Pickands estimator (Pickands, 1975 [111]) is the first suggested estimator for the parameter γ. The interest of this estimator is to be defined for $\gamma \in \mathbb{R}$. This estimator is constructed using three order statistics, it has the advantage of being valid regardless of the domain of attraction of the distribution and therefore of the domain of definition of the index of extreme values.

Definition 1.12 *(Pickands estimator)*
The Pickands estimator is defined by:

$$\widehat{\gamma}_{n;k}^{(P)} = \frac{1}{\log 2} \log\left(\frac{X_{n-k+1;n} - X_{n-2k+1;n}}{X_{n-2k+1;n} - X_{n-4k+1;n}}\right). \tag{1.47}$$

We shall give weak consistency and asymptotic properties of $\widehat{\gamma}_{n;k}^{(P)}$:

Theorem 1.11 *(Weak consistency of $\widehat{\gamma}_{n;k}^{(P)}$)*
Let $(X_n)_{n\geq 1}$ be a sequence of iid rv's with df $F \in \mathcal{D}(H_\gamma)$ with $\gamma \in \mathbb{R}$. Then as $k \to \infty$ and $k/n \to 0$

$$\widehat{\gamma}_{n;k}^{(P)} \xrightarrow{p} \gamma \text{ as } n \to \infty$$

The properties of the Pickands estimator $\widehat{\gamma}_{n;k}^{(P)}$ have been studied by Pickands (1975 [111]) and extended by Dekkers and De Haan (1989 [38]). Pickands demonstrates the weak consistency of his estimator and Dekkers and De Haan proved that this estimator is strongly consistent (and therefore also weakly consistent) and asymptotically Normal-distributed.

Theorem 1.12 *(Asymptotic properties of $\widehat{\gamma}_{n;k}^{(P)}$)*
Suppose that $F \in \mathcal{D}(H_\gamma)$, $\gamma \in \mathbb{R}$, $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

(a) Strong consistency: If $k/\log\log n \to \infty$ *as* $n \to \infty$, *then*

$$\widehat{\gamma}_{n;k}^{(P)} \xrightarrow{a.s} \gamma \text{ as } n \to \infty$$

(b) Asymptotic normality: We suppose that U *admits positive derivatives* U' *and that* $\pm t^{1-\gamma}U'(t)$ *and that (with either choice of sign) is* Π*–varying at infinity with auxiliary function* a. *If* $k = o(n/g^{-1}(n))$, *where* $g(t) = t^{3-2\gamma}(U'(t)/a(t))^2$, *then*

$$\sqrt{k}\left(\widehat{\gamma}_{n;k}^{(P)} - \gamma\right) \xrightarrow{d} \mathcal{N}(0, \eta^2) \text{ as } n \to \infty,$$

where

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

An improvement of the Pickands estimator is proposed by Drees[50]. It is a convex combination of the Pickands estimators obtained for different values of k. This estimator, called the Drees-Pickands estimator, which was generalized later by Johan[85].

1.5.3 Hill estimator

After the Pickands'estimator, Hill (1975[79]) introduced another estimator for γ, but is restricted to the case of heavy tails df which belong to Fréchet maximum domain of attraction.

Let $X_1, X_2, ..., X_n$, n rv's i.i.d. from cdf $F \in \mathcal{D}(\Phi_\gamma)$; where $\gamma < 0$. Let $k = k(n)$ be a sequence of integers with $1 < k < n$.

Definition 1.13 *(Hill's estimator)*
Hill's estimator, denoted $\widehat{\gamma}_{n;k}^{(H)}$, *constructed from the k largest order statistics, is defined by:*

$$\widehat{\gamma}_{n;k}^{(H)} = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1;n} - \log X_{n-k;n} \tag{1.48}$$

The construction of this estimator is based on the Maximum Likelihood method where one uses statistics of order higher than a certain threshold u, to keep only the largest observations, so that they follow approximately a distribution of Pareto. After its construction, several researchers tried to determine its asymptotic properties. Thus, Mason (1982 [100]) proved the weak consistency of the Hill estimator for any sequence $k = k(n)$ satisfying $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ called an intermediate sequence of integers. The condition $k \to \infty$ ensures that the size of k-order statistics is large enough to obtain stable estimators. On the other hand, the $k/n \to 0$

condition makes it possible to stay in the tail of the distribution. Davis and Resnick (1984[36]) proposed its asymptotic normality under the conditions of Von Mises; Csörgő and Mason (1985 [29]) presented its asymptotic normality by introducing the approximation of empirical processes by Brownian bridges. In the same vein, Resnick and de Haan (1998 [47]) have shown this asymptotic property.

Theorem 1.13 *(Asymptotic properties of $\widehat{\gamma}_{n;k}^{(H)}$)*
Suppose that $F \in \mathcal{D}(\Phi_\gamma)$, $\gamma > 0$, $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

(a) *Weak consistency:*

$$\widehat{\gamma}_{n;k}^{(H)} \xrightarrow{p} \gamma \text{ as } n \to \infty$$

(b) *Strong normality: if $k/\log\log n \to \infty$ as $n \to \infty$, then*

$$\widehat{\gamma}_{n;k}^{(H)} \xrightarrow{a.s} \gamma \text{ as } n \to \infty.$$

(c) *Asymptotic normality : Suppose that the df F satisfies the second order condition. If $\sqrt{k}A(n/k) \to \lambda$ as $n \to \infty$, then*

$$\sqrt{k}\left(\widehat{\gamma}_{n;k}^{(H)} - \gamma\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right), \text{ as } n \to \infty.$$

Chapter 2

Censored data

This chapter is dedicated to reminder of the essential concepts on censorship and survival data. It will present a foreword on survival data and censoring with these different types as well as estimating the survival function as well in a right-censoring model. We will present some definitions related to the statistics of survival times. Censorship is based on a few functions suchs as the distribution function, the survival function, the risk function... Many authors have been interested in the notion, in particular Kaplan and Meier (1958 ,[88]), who have proposed an estimator of the survival function which Beran generalized it (1981,[15]) in the conditional case called the generalized or conditional Kaplan-Meier estimator.

2.1 Introduction

The problem of missing, incomplete or erroneous data is very vast and very rich by the multitude of works which have been devoted to it, it has aroused a lot of interest among statisticians in recent years.

Historically it was the British demographers John Graunt and William Petty who established the first statistics on the survival times of the population in the middle of the 17th century. It was not until the 19th century that tables linked to statistical variables began to appear and the modeling of the risk function was started.

Survival analysis underwent significant development in the second half of the twentieth century after Kaplan and Meier (1958 [88]) introduced their famous survival function estimator for right-censored data.

Survival analysis involves the modeling of time to event data. It has been a very active research for several decades. An important contribution that stimulated the entire eld was the counting process formulation given by Aalen (1975 [1]). The exibility of a counting process is that it allows modeling multiple (or recurrent) events. Since then a large number of textbooks have been written on survival

analysis and counting processes, with some key references being Andersen et al. (1993 [3], Fleming and Harrington (1991,[56]), and Lawless (2003,[96]). Excellent texts aimed at the biostatistical community with biomedical application as the motivating factor include Klein (1997 [91]) and Moeschberger (1997), Therneau and Grambsch (2000).

2.2 Survival times

Statistical life span analysis, or survival analysis, is the study of how long it takes for an event to occur. It is a generic term for any analysis of the occurrence over time of an event. The study of duration data (lifetime, failure time, re-employment time...) subject to random censoring is a major topic in statistics, which finds applications in many areas.

Survival times measured from an appropriate origin have two characteristics. The first is that they are non-negative and such that an assumption of normality is usually not reasonable due to pronounced asymmetry. The second is structural and stems from the fact that for some individuals the event studied does not occur during the observation period and consequently certain data are censored.

2.2.1 Definitions and Notations:

Analysis of survival data studies the appearance of an event over time.

In order to define a failure time random variable, we need some definitions:

2.2.1.1 Origin date:

It corresponds to the origin of the duration studied. It can be the date of birth, the onset of exposure to a risk factor, the date of onset of the disease or the date of surgery. Each individual can therefore have a different date of origin (not important because it is the duration that interests us), e.g. diagnosis of a disease, start of treatment (randomization)

2.2.1.2 Point date:

It's the date or the end of the study and we will no longer take subject information into account.

2.2.1.3 Date of the latest news:

This is the most recent date that information on a topic was collected.

2.2.1.4 Domains of Applications

Survival data analysis is an area of statistics that finds its place in all fields of application where the occurrence of an event, it is used in many fields;

In medicine, survival analysis is used to assess the effectiveness of a treatment. For example, we want to estimate the probable survival time of a patient. We use for this a sample of patients of which we know, for each of them,

- either the real survival time (uncensored data or detection),

-or a lower limit of this duration (censored data).

The second case occurs when a patient is lost, for example due to moving, or when he or she dies for an unrelated cause.

In demography, survival analysis is used to construct life tables. These are used by actuaries to determine the amount of life insurance and annuities, among others; we speak of actuarial tables when the data is grouped into intervals.

In engineering, survival analysis makes it possible to estimate the reliability of machines, electronic components...

2.2.2 Functions of Survival Time:

Now that we have introduced the main notions in survival analysis, let's define the variables and functions that we will be using.

Suppose that the survival time X is a positive or zero random variable, and absolutely continuous defined on a probability space $(\Omega; A; P)$; there are several equivalent ways to characterize the probability distribution of a survival random variable X. Some of these are familiar; others are special to survival analysis.

The functions most used in survival analysis and which best characterize the distribution of X are the survival function S, the instantaneous risk function h and the cumulative risk function H.

2.2.2.1 The distribution function

Definition 2.1 *Distribution function $F(t)$ describes the probability the time to event X is smaller or equal compared to a fixed time (t) and is given as:*

$$F(t) = \mathbb{P}(X \leq t) \tag{2.1}$$

Remark 2.1 *$F(t)$ is an increasing monotonic function, continuous to the right such that:*

$$\lim_{t \to 0} F(t) = 0 \text{ and } \lim_{t \to \infty} F(t) = 1$$

2.2.2.2 Survivorship Function (or Survival Function)

The survival function of X, also called the tail of the distribution, is defined as the probability that X is greater than a certain time t, and is of considerable interest in failure time analysis. Let $S(t)$ denote the survival function of X, then,

$$S(t) = \overline{F}(t) = 1 - F(t) = \mathbb{P}(X > t) \tag{2.2}$$

Survival function $S(t)$ can be also interpreted as the probability that a certain object of interest will survive beyond a certain time t.

Remark 2.2 *$S(t)$ is a decreasing monotonic function, continuous to the left such that:*

$$\lim_{t\to 0} S(t) = 1 \text{ and } \lim_{t\to\infty} S(t) = 0$$

2.2.2.3 Empirical distribution and survival functions

Let $X_1,...,X_n$ be a sample of size $n \geq 1$ of a positive r.v X of cdf F and survival function $\overline{F}$. The empirical functions of distribution and survival, F_n and $\overline{F_n}$ are respectively defined by:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq t\}} \ , \forall t \geq 0 \tag{2.3}$$

and

$$\overline{F_n}(t) = 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i > t\}} \ , \forall t \geq 0 \tag{2.4}$$

where $\mathbb{I}_{\{A\}}$ denotes the indicator function of the set A.

Remark 2.3 *$F_n(t)$ is the proportion of the n variables that are less than or equal to t.*
$\overline{F}_n(t)$ is the proportion of observations that exceeds t.
For $1 \leq i \leq n$, the functions F_n and $\overline{F}_n$ are written in terms of the values of the order statistics as follows:

$$F_n(t) = \begin{cases} 0 \text{ if } t < X_{1,n} \\ \frac{i}{n} \text{ if } X_{i;n} \leq t < X_{i+1,n} \\ 1 \text{ if } t \geq X_{n,n} \end{cases}$$

and

$$\overline{F}_n(t) = \begin{cases} 1 \text{ if } \ t < X_{1,n} \\ 1 - \frac{i}{n} \text{ if } \ X_{i,n} \leq t < X_{i+1,n} \\ 0 \text{ if } t \geq X_{n;n} \end{cases}$$

The application of the strong law of large numbers on $F_n(t)$ gives the following result.

Corollary 2.1

$$F_n(t) \xrightarrow{a.s} F(t) \quad as\ n \to \infty$$

The result of this corollary can be strengthened in the following fundamental result in nonparametric statistics, known under the name of theorem Glivenko-Cantelli.

Theorem 2.1 *(Glivenko-Cantelli)*
The convergence of F_n to F is almost surely uniform, i.e.

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s} 0 \quad as\ n \to \infty$$

The proof of this Theorem could be found in any standard textbook of probability theory such as (Billing [16], chapter 4)

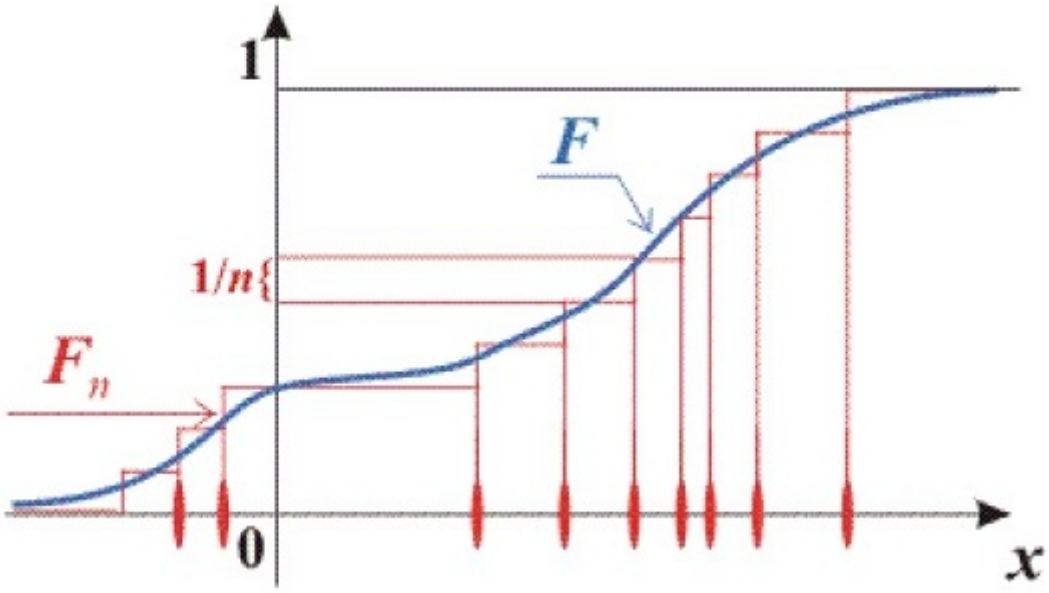

Figure 2.1: Empirical and theoretical distribution function

2.2.2.4 Density function

The probability density function represents the probability of the event occurring at time t, If F admits a derivative with respect to Lebesgue's measure on $\mathbb{R}_+$, the probability density function exists, and it is defined for all $t \geq 0$:

$$f(t) = \lim_{h\to 0} \frac{\mathbb{P}(t \leq X < t+h)}{h} = \frac{dF(t)}{dt} \quad , h > 0 \tag{2.5}$$

2.2.2.5 Hazard Function

An important function useful in survival is that of *hazard function* or *"chance rate"*, sometimes called *an instantaneous failure rate*.

Definition 2.2 *the hazard function of X at time t, Noted $h(t)$ and is defined by:*

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq X < t + \Delta t / X \geq t)}{\Delta t} \tag{2.6}$$

it represents the instantaneous probability that a subject fails at time t given that the subject has not failed before t.

Remark 2.4 *The hazard function h; can also be defined in terms of cdf $F(t)$ and the probability density function $f(t)$:*

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \tag{2.7}$$

2.2.2.6 Cumulative hazard function

There is another quantity that is also common in survival analysis, the cumulative hazard function, this is the integral of the hazard function from 0 to t:

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{dF(x)}{\overline{F}(x)} = -\log(S(t)) \tag{2.8}$$

The previous five functions are mathematically equivalent, so they are related to each other. It is sufficient to give any one of them, the others can be derived*: the knowledge of $S(t)$ allows that of $f(t)$ and therefore that of $h(t)$ then $H(t)$. Likewise, the knowledge of $h(t)$ allows that of $H(t)$ therefore of $S(t)$ and finally of $f(t)$. In other words, if we give ourselves only one of these functions, then the others are at the same time also defined. In particular, a choice of specification on the hazard function involves the selection of a certain distribution of the survival data.*

Example:

Suppose that the survival time of a population has the following density function:

$$f(t) = \exp(-t), \quad t \geq 0,$$

using the definition of the cumulative distribution function,

$$F(t) = \int_0^t f(u)du = 1 - \exp^{-t} \quad t \geq 0$$

from 2.2, we obtain the survivorship function

$$S(t) = 1 - F(t) = \exp(-t)$$

the hazard function can then be obtained from 2.7:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\exp^{-t}}{\exp^{-t}} = 1$$

2.2.2.7 Mean and variance of survival time

The mean survival time $\mathbb{E}(X)$ and the variance of the survival time $\mathbb{V}(X)$ are defined by the following quantities

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty t dF(t) \\ &= -\int_0^\infty t d(1 - F(t)) \\ &= \int_0^\infty S(t)dt \end{aligned} \tag{2.9}$$

We can also show that, if $\mathbb{V}(X)$ exists, then:

$$\mathbb{V}(X) = 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2 \tag{2.10}$$

2.2.2.8 Quantiles of survival time

The median of the survival time is the time t for which the probability of survival $S(t)$ is equal to 0.5, that is, the value t_k which satisfies $S(t_k) = 0.5$.

Sometimes it is possible to get a confidence interval of the median time. Let $[B_i; B_s]$ be a level α confidence interval of $S(t_k)$; then a level α confidence interval of the median time t_k is

$$[S(B_s), S(B_i)]$$

The quantile function of survival time is defined by:

$$\begin{aligned} q(p) &= \inf(t; F(t) \geq p), \quad 0 < p < 1 \\ &= \inf(t; S(t) \leq 1 - p) \end{aligned}$$

When the distribution function F is strictly increasing and continuous then:

$$\begin{aligned} q(p) &= F^{-1}(p), \quad 0 < p < 1, \\ &= S^{-1}(1-p). \end{aligned}$$

The quantile $q(p)$ is the time when a proportion p of the population has disappeared.

Remark 2.5 *Because the distribution of a failure time r.v. is often not symmetric (eg. Exponential), we often use median survival. Also, median survival is usually better estimated than mean survival.*

2.3 Censorship and truncation

Many researchers consider that the analysis of survival data is only the application of two conventional statistical methods to a particular type of problem: parametric if the distribution of survival times is known to be normal and nonparametric if the distribution is unknown. This hypothesis would be true if the survival times for all subjects were accurate and known; however, some survival times are not. In addition, the survival distribution is often biased or far from normal. There is therefore a need for new statistical techniques. One of the most important developments is due to a special feature of survival data in the life sciences that occurs when some study subjects have not experienced the event of interest at the end of the study or at the time of analysis. For example, some patients may still be alive or disease free at the end of the study period. The exact survival times for these subjects are unknown. These are called censored observations or censored times and can also occur when people are lost to follow-up after a period of study.

The phenomenon of censorship is linked to disruptive events that can occur in the time needed to collect data. It is therefore frequently involved in measurements that relate to variables modeling the time elapsed between two events: length of life of an individual, time between the onset of an illness and recovery, duration of an unemployment episode, ... etc. These disturbances prevent the observer from accessing all of the information concerning the phenomenon he is studying and leads to the appearance of incomplete so-called censored observations.

2.3.1 Censorship concept

An observation is censored when it is partially known. Censorship is the most common phenomenon encountered when collecting survival data.

Definition 2.3 *The censoring variable C is defined as the possible non-observation of the event. If instead of observing X, we observe C, and we know that $X > C$ (respectively $X < C$, $C_1 < X < C_2$), we say that there is right censorship (respectively left censorship , interval censorship).*

In reliability, censorship is the consideration of systems no breaches of the establishment of the law of reliability. More generally, the term applies when the exact date of the failure is not known, either because the failure has not yet occurred or has not been recorded with precision.

Example 2.1 *In a medical context, censorship can occur when:*

- *the patient leaves the study,*
- *the patient does not present any symptoms / evolution before the end of the study,*
- *the patient file is lost.*

2.3.2 Types of censoring:

Data censorship is done according to several mechanisms, observations can present different types of censorship such as right censorship, left censorship, double (or mixed) censorship.

For a given individual i, we will consider

- its survival time X_i, of distribution function F.
- its censorship variable C_i, with distribution function G.
- its actually observed variable Z_i with distribution function H.

2.3.2.1 Right censoring

The most common form of censoring is "Right censoring", occurs when a time-to-event is only known to be greater than a censoring time due to end of study, loss to follow-up, or patient's withdrawal. It is convenient to use the following notation: for a specific individual under study, we assume that there is a lifetime X and a fixed censoring time, C_r (C_r for "right" censoring time). The X's are assumed to be independent and identically distributed: The exact lifetime X of an individual will be known if, and only if, X is less than or equal to C_r. If X is greater than C_r, the individual is a survivor, and his or her event time is censored at C_r. The data from this experiment can be conveniently represented by pairs of random variables

$(Z;\delta)$, where δ indicates whether the lifetime X corresponds to an event ($\delta = 1$) or is censored ($\delta = 0$), and Z is equal to X, if the lifetime is observed, and to C_r if it is censored, i.e., $Z = \min(X;C_r)$.

We call this right-censoring because the true unobserved event is to the right of our censoring time; i.e., all we know is that the event has not happened at the end of follow-up.

2.3.2.2 Left censoring

Left censoring is much rare. There is "left censorship" when the individual has already experienced the event before it is observed. We only know that the variable of interest X is less than a censoring time C_l (C_l for "left" censoring time), that is, the event of interest has already occurred for the individual before that person is observed in the study at time C_l. For such individuals, we know that they have experienced the event sometime before time C_l, 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 C_l. The data from a left-censored sampling scheme can be represented by pairs of random variables $(Z;\delta)$, as in the previous kind, where Z is equal to X if the lifetime is observed and δ indicates whether the exact lifetime X is observed ($\delta = 1$) or not ($\delta = 0$). Note that, for left censoring as contrast with right censoring, $Z = \max(X;C_l)$.

For example if we want to reliably study a certain electronic component that is connected in parallel with one or more other components: the system can continue to operate, albeit in an aberrant fashion, until this failure is detected (for example during control or in case of system shutdown). Thus, the duration observed for this component is censored on the left.

In everyday life there are several phenomena which present both right and left censored data.

2.3.2.3 Double or mixed censorship

We say that we have "mixed or double censorship" in a sample if we have left-censored data and right-censored data in the same sample. Several non-parametric models have been presented for the study of double censorship. For example, the model of Turnbull (1974 [129]) is the most used, and several works are based on this model.

2.3.2.4 Interval censoring

Another type of censoring occurs when the lifetime is known only to lie in an interval, instead of being observed exactly. In this case, as the name suggests, we observe both a lower bound and an upper bound of the variable of interest. This pattern is typically found in follow-up studies where patients are checked periodically if a patient does not show up for one or more checks and then presents after the event of interest has occurred. We also have this kind of data which is censored on the right or, more rarely, on the left.

One advantage of this type is that it allows data to be presented censored to the right or to the left by intervals of the type $[a;+\infty[$and $[0;a]$ respectively.

Example 2.2 *(**right-censorship**) :*
A classic example of right-censorship is where the study examines the survival time X of patients with a certain disease. For patients lost to follow-up after time C while they were still alive, C censors X to the right since, for them, X is unknown but greater than C: $X > C$.

Example 2.3 *(**left censorship**) :*
An ethnologist studies the learning time of a task. This duration is a random variable X and C is the age of the child. For children who already know how to accomplish the task, C censors X to the left because for them X is unknown but less than C: $X < C$.

These four categories of censorship described above can arise depending on the mode or mechanism of censorship. Thus, in the literature we find the following types:

2.3.2.5 Type I Censoring

The experimenter sets a value (a non-random end date, for example). For example, in epidemiology, the maximum duration of participation is fixed and the difference between the date of end of the experiment and the date of entry of the patient into the study is valid for each observation, eg. animal's studies; all animals sacrificed after 2 years. The number of events observed is random.

Let C be a fixed value. For example in right-censorship, instead of observing the variables $X_1,...,X_n$ which interests us, we observe X_i only when it is less than or equal to the fixed duration C: We therefore observe a variable Z_i such that $Z_i = \min(X_i, C)$; $i = 1,...,n$.

This censorship mechanism is frequently encountered in industrial applications.

2.3.2.6 Type II Censoring

The experimenter fixes a priori the number of events to be observed. The end date of the experiment then becomes random, the number of events being non-random. This model is often used in studies of reliability, epidemiology, e.g. in engineering reliability experiments, the stop of the experiment after the failure of machine parts.

Let $X_{(i)}$ and $Z_{(i)}$ be the order statistics of the variables X_i and Z_i. The censorship date is therefore $X_{(r)}$ and we only observe the following variables:

$$\begin{aligned} Z_{1,n} &= X_{1,n} \\ &\vdots \\ Z_{r,n} &= X_{r,n} \\ Z_{r+1,n} &= X_{r,n} \\ &\vdots \\ Z_{n,n} &= X_{r,n} \end{aligned}$$

2.3.2.7 Type III Censoring

This is the random version of type I. Typically this model is used for therapeutic trials. In this type of experiment, the date of inclusion of the patient in the study is fixed, but the end date is unknown (this corresponds, for example, to the length of the patient's hospital stay).

Let $X_1,...,X_n$ a sample of a non-negative r.v X, we say that there is type III censorship of this sample if there is another positive r.v Y of sample $Y_1,...,Y_n$ in this case instead of observing the X_i's, we observe a couple of r.v's (Z_i,δ_i) with:

$$Z_i = \min(X_i, Y_i) \text{ and } \delta_i = \mathbb{I}\{X_i \leq Y_i\}_{i=1,...n}$$

Type I and type II censored observations are also called singly censored data, and type III, progressively censored data.

2.3.3 Truncated data

Truncation is another part of missing data of which left truncation is the most common. It occurs when the loans have been at risk before entering the study. Truncation is a condition other than the event of interest that is, for example, used to

screen respondents or patients, see e.g. Klein and Moeschberger (1997 [91]). This is very common in datasets, for example facilities enter the portfolio at a certain point in time because loans are bought. In that case loans are at risk before entering the portfolio and data is available.

2.3.3.1 Right truncation:

We say that there is right truncation when the variable of interest X_i (lifetime of the i^e individual) is not observable when it is greater than a fixed threshold $R > 0$, i.e, all observations greater than an R value are ignored. As example, if you ask a group of smoking school pupils at what age they started smoking, you necessarily have truncated data, as individuals who start smoking after leaving school are not included in the study.

2.3.3.2 Left truncation:

We say that there is left truncation when the variable of interest X_i (lifetime of the i^e individual) is not observable when it is less than a fixed threshold $r > 0$,i.e, all observations less than an r value are ignored. Example if you wish to study how long people who have been hospitalized for a heart attack survive taking some treatment at home, the start time is taken to be the time of the heart attack, only those individuals who survive their stay in hospital are able to be included in the study.

Remark 2.6 *We have left and right truncation if all observations less than an r value or greater than an R value are ignored. If the part is not connected, we say that the truncation is interval.*

Note that there are also models where truncation and censorship are exploited simultaneously to study practical cases.

2.4 Estimation of survival function

In the literature, several authors have been interested in estimating the survival function in the case where the data are censored.

The main estimators playing an essential role in the framework of censored data are:

• The Kaplan-Meier estimator (1958) for the survival function $S(t)$: It is also called the product-limit estimator.

• The Nelson-Aalen estimator for the cumulative hazard function $H(t)$.

The Kaplan–Meier is the most commonly used estimator of the survival function, while the Nelson-Aalen is an alternative estimator for the same function.

Kaplan and Meier proposed a very efficient estimator of the survival function when observations are right-censored, this estimator was generalized by Beran (1981 [15]).

2.4.1 Kaplan-Meier Estimator

The standard non-parametric estimator of the survival function is the Kaplan-Meier estimator introduced by Kaplan and Meier (1958). It is also called "Product-Limit" because it is obtained as the limit of a product, and has been extensively studied in the literature.

Definition 2.4 *Let $(Z_i, \delta_i)_{1\leq i\leq n}$ the actually observed sample and let $(Z_{(i)}, \delta_{(i)})$ its increasing order statistic, the Kaplan-Meier estimator of the survival function S, denoted $\hat{S}_{KM}$, is defined by:*

$$\hat{S}_{KM}(t) = \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)} \mathbb{I}_{\{Z_i \leq t\}}}$$

The great advantage of the Kaplan-Meier (K-M) estimator is that it is computable for right-censored data. The idea of the K-M estimator is given by the conditional probability. Let $t_i \leq t_{i+1}$:

$$\begin{aligned} S(t_i) &= \mathbb{P}(T > t_i) \\ &= \mathbb{P}(T > t_i, T > t_{i-1}) \\ &= \mathbb{P}(T > t_i | T > t_{i-1}).\mathbb{P}(T > t_{i-1}) \\ &= \mathbb{P}(T > t_i | T > t_{i-1}).\mathbb{P}(T > t_{i-1} | T > t_{i-2}).... \mathbb{P}(T > t_0 = 0) \end{aligned}$$

We assume that at the start of the study all subjects were alive, so $\mathbb{P}(\mathbb{T} > t_0 = 0) = 1$. The conditional probability is

$$\mathbb{P}(T > t_i | T > t_{i-1}) = \frac{n_i - d_i}{n_i}$$

where n_i is the number of subjects at risk in the study at the time t_i, and d_i is the number of subject dying at time t_i.The Kaplan-Meier estimator is :

$$\hat{S}_{KM}(t) = \prod_{i; t_i \leq t} \frac{n_i - d_i}{n_i}$$

• This Kaplan-Meier estimator is a staged function with jumps only at uncensored observations.

• The height of the jumps of this estimator is random.

• When all the observations are uncensored then we get the empirical distribution function.

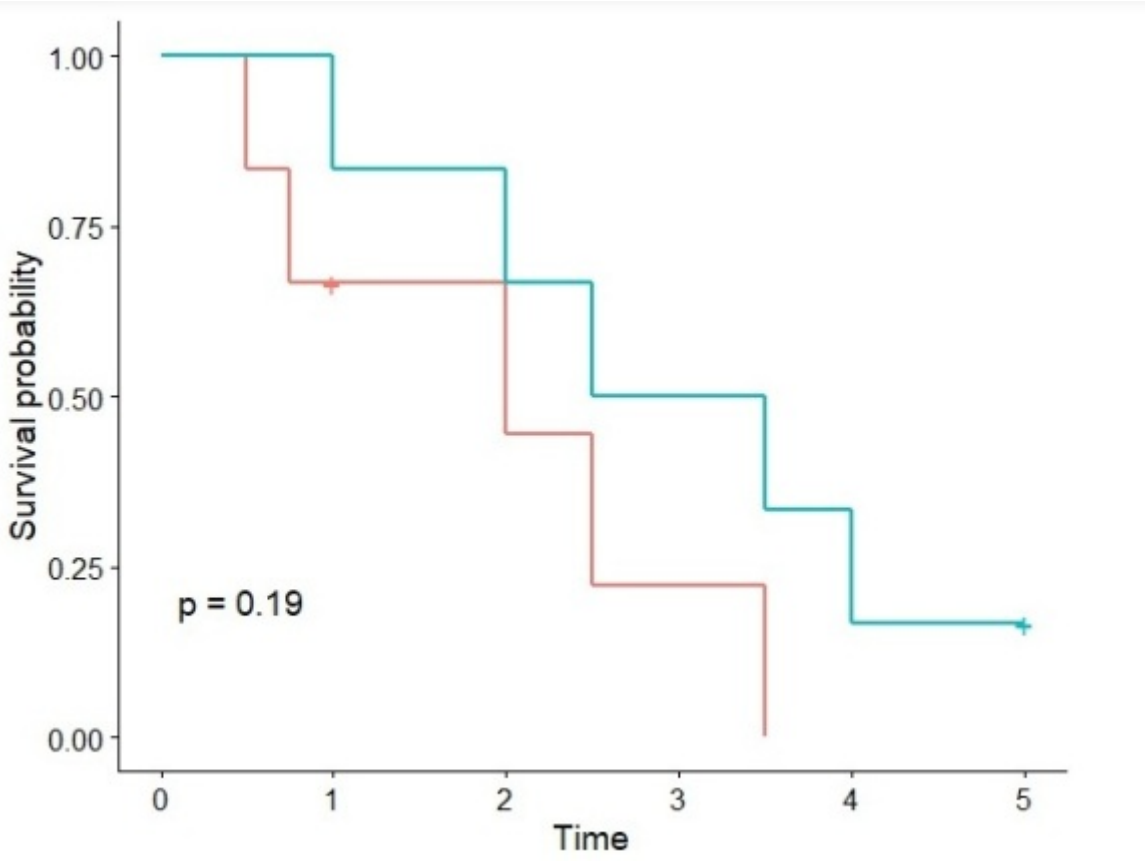

Figure 2.2: Kaplan-Meier curve

The Kaplan-Meier estimator converges almost surely and uniformly to S (FÖLDES et al. 1980 [57]). Under certain conditions of regularity, it converges in law to the Gaussian process (see Breslow and Crowley 1974 [20]). The mathematical properties of this estimator can also be found in Chapter 7 of Shorack and Wellner (1986 [120]).

2.4.1.1 Variance of Kaplan Meier estimator

The estimate of the variance is given by Greenwood's formula:

$$VAR(\hat{S}_{KM}(t)) = \hat{S}^2_{KM}(t) \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

This formula had been given in 1926 by Greenwood[1] before Kaplan and Meier published their estimator in 1958.

2.4.1.2 The Kaplan-Meier estimator for left-censored data

If we have left-censored data, we have to estimate the cumulative distribution function instead of the survival function. We could use an estimator derived from the idea of the Kaplan-Meier estimator. However, here we are interested in the infection time instead of the dead time. We have the following statement (assuming $t_i \leq t_{i+1}$):

$$\begin{aligned} F(t_i) &= \mathbb{P}(T \leq t_i) \\ &= \mathbb{P}(T \leq t_i | T \leq t_{i+1}).\mathbb{P}(T \leq t_{i-1}) \\ &= \mathbb{P}(T \leq t_i | T > t_{i-1}).\mathbb{P}(T \leq t_{i+1} | T \leq t_{i+2})....\mathbb{P}(T \leq t_n) \end{aligned}$$

We assume that we have only non-censored or left-censored data. Then we have $\mathbb{P}(T \leq t_n) = 1$, as t_n is the greatest time of realisation of all random variables. This suggests the following estimator:

$$\hat{F}_L(t) = \sum_{i;t_i>t} \frac{n_i - d_i}{n_i}$$

where d_i is the number of subjects getting infected at time t_i and n_i is the number of data in the study at time t_i. If a data is left-censored, it would enter in the study at the left-censor bound time, and may be count among n_i. If a data t_i is not left-censored, it would enter in the study at time t_i and count among d_i.

For the variance of this estimator, we may adapt the Greenwood's formula changing $t_i \leq t$ by $t_i > t$.

The Kaplan-Meier estimator can be adapted for both left-truncated and right-censored data or left-censored data. Other developments have been proposed to consider data truncated on the left, truncated on the right or even censored by interval (Peto, 1973 [110]; Turnbull, 1976 [129]; Lagakos et al., 1988; Andersen et al., 1993 [3]).

2.4.2 The Generalised K-M estimator

Beran (1981, [15]) added a local aspect to the Kaplan-Meier estimator using smoothing with Nadaraya-Watson weights. Thus, he was studying regression problems

[1] M. Greenwood, The natural duration of cancer, Reports on Public Health and Medical Subjects 33, 1926

with censored data in a completely nonparametric context. We refer to [103] where this estimator is presented with its properties.

The estimator thus proposed is defined as follows:

$$1-\hat{F}_n(z|x)=\hat{S}_{GKM}(z|x)=\begin{cases}\prod_{i=1}^{n}\left(1-\frac{B_i(x)}{\sum_{j=1}^{n}\mathbb{I}\{z_j\geq z_i\}B_j(x)}\right)^{\mathbb{I}\{z_{(i)}\leq z,\,\delta_{(i)}=1\}} & \text{if } z\leq Z_{(n)}\\ 0 & \text{if not}\end{cases}$$

where

$$B_i(x)=\frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right)}$$

represent the Nadaraya-Watson weights, $h\to 0$, the window and K, the kernel.

This estimator is also called "the conditional Kaplan-Meier estimator" or "the generalized Kaplan-Meier estimator". Note that this estimator in the absence of censorship corresponds to the estimator of the empirical distribution function, studied by Stone (1977,[125]).

2.4.3 Nealson-Aalen estimator

Kaplan and Meier introduced the limit product estimator for the survival function. The cumulative hazard function estimator is the Nelson-Aalen estimator introduced by Nelson in 1972 and generalized by Aalen in 1978 [1].

First of all, we observe that, under the general hypothesis of independence between X and Y, we can decompose $H(t)$ as follows:

$$\begin{aligned}H(t) &= 1-(1-F(t))(1-G(t))\\ &= H^{(0)}(t)+H^{(1)}(t)\end{aligned}$$

where

$$H^{(0)}(t)=\mathbb{P}(z\leq t,\delta=0)=\int_0^t\overline{F}(x)dG(x)$$

and

$$H^{(1)}(t)=\mathbb{P}(z\leq t,\delta=1)=\int_0^t\overline{G}(x)dF(x)$$

For $t\geq 0$; the cumulative hazard function (1.8) can be expressed as follows

$$H(t)=\int_0^t\frac{\overline{G}(x)dF(x)}{\overline{H}(x)}=\int_0^t\frac{dH^{(1)}(x)}{\overline{H}(x)}$$

Definition 2.5 *The Nelson-Aalen non-parametric estimator $\hat{H}_{NA}$ of H based on the sam-*

ple $\{(Z_i, \delta_i),\ 1 \leq i \leq n\}$ *defined by:*

$$\hat{H}_{NA}(t) = \int_0^t \frac{dH_n^{(1)}(x)}{\overline{H}_n(x)} = \begin{cases} \sum\limits_{z_{i:n} \leq t} \frac{\delta_{[i:n]}}{n-i+1} & \text{if } t < z_{i:n} \\ 1 & \text{if } t \geq z_{i:n} \end{cases}$$

where:

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{z_i \leq t\} \text{ and } H_n^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{I}\{z_i \leq t\}$$

respectively represent the empirical fdr of $H(t)$ *and the empirical version of* $H_n^{(1)}(t)$ *of sample* $Z_1, ..., Z_n$.

Asymptotically, Kaplan-Meier and Nelson-Aalen estimators are equivalent. On small samples, Kaplan-Meier would be better when chance decreases over time, while Nelson-Aalen would be better when chance increases over time (see Colosimo et al.[23]).

Remark 2.7 *We can notice that the Nelson-Aalen estimator of the cumulative chance rate leads to a natural estimator of the survival function, by exploiting the relation* $S(t) = \exp(-H(t))$*; we can thus propose as a survival function estimator*

$$\hat{S}_{HF} = \exp(-\hat{H}_{NA}(t))$$

this estimator is the Harrington and Fleming estimator (see Fleming [56]).

2.4.3.1 Variance of the Nelson-Aalen estimator

Using the theory of counting processes and approximating by a Poisson distribution, we show that the variance of the Nelson-Aalen estimator is,

$$Var(\hat{H}_{NA}(t)) = \sum_{i/z_{i:n} \leq t} \frac{\delta_{[i:n]}}{(n-i+1)^2}$$

Remark 2.8 *We show that, under certain conditions, the Nelson-Aalen estimator is uniformly consistent, asymptotically normal and asymptotically unbiased (when* $\mathbb{P}(Y = 0) \to 0$ *).*

Another often used estimator is the Breslow or Peterson estimator. It is obtained from the survival estimator KM and uses the equation linking the two functions,

$$\tilde{H}(t) = -\log S(t)$$

We can show that $\hat{H}(t) < \tilde{H}(t)$: the Nelson-Aalen estimator is always lower than the Breslow estimator [2]. There is no reason, however, to favor one over the other.

[2]This comes from the fact that the log function being concave it is located under its tangent and therefore, if we consider a Taylor expansion at order 1, it comes $\log(1 - x^+) < x^+$. As on the one hand $\tilde{H}(t) = -\sum_{i \backslash t_i \leq t} \log(1 - \frac{d_i}{n_i})$ and $\hat{H}(t) = \sum_{i \backslash t_i \leq t} \frac{d_i}{n_i}$, we immediately obtain the announced property.

Chapter 3

Risk Measurments and measures of income inequality

Lord Kelvin once said « Anything that exists, exists in some quantity and can therefore be measured»

Risk is a difficult notion to pin down. But generally speaking, it can be said that this is an undesirable and relatively innocuous. The risk is generally harmless, but still harmful enough to be undesirable. In this sense it is distinguished in particular from the danger which supposes the possibility of serious or even lethal damage. For example, someone who goes out bareheaded in cold weather will say that he runs the risk of catching a cold, but we will say that he is in danger if he crosses a highway.

A risk is an unlikely contingency, which is another difference from danger. We speak of danger when the probability of occurrence and the consequences are significant, while the risk exists when its probability of occurrence is not zero. The assessment of these different criteria is highly subjective, which may justify the search in scientific and technical fields for a quantifiable and rigorous definition of risk.

This chapter deals with risk measurement: in section 3.2 we present some notations and definitions for the following sections. Well-known measures of income inequality and the relationships between them are presented in section 3.3. In Section 3.4 we discuss what Lorenz curve and Gini coefficient, we give some main properties that it may have.

3.1 Introduction

In recent years, increasing attention has been paid to methods allowing a quantified measure to be associated with a risk, this risk possibly arising from a financial po-

sition or contingencies insured by an insurance company, the latter including risks. major varied such as natural disasters, pandemics, or industrial risks.

In general, a risk is an event which may or may not occur (*i.e.*, a random event) and has some adverse consequences. It is natural to model risk as a random variable that represents the random amount of loss a company may experience. It can be assumed that the random variables modeling risk losses are non-negative (similar to insurance risk).

Since 1997 the paper of Artzner et al[5] risk measurement, and hence risk measures, have gained enormously in interest under economist, bank regulators and mathematicians, giving rise to a new theory. A good reference for the Risk theory is the book of Denuit et al [42] and Kaas et al [87].

3.2 Definitions and notations

At the beginning risk measurement was mainly focussed on the mathematical properties which reflect the underlying economical meaning, however in the last years the statistical properties have become of increasing interest. Nowadays it is obvious to all working with risk, be it in practice or theory, that the procedure of risk measurement in fact involves two steps:

1) Estimating the loss distribution of the position.

2) Constructing a risk measure that summarizes the risk of the position.

The position's loss distribution in practice is generally unknown, and therefore must be estimated from data. The estimation is essentially done by backtesting[1]. Each one of the steps above should be regarded as equally important.

We are now ready to state the definition of a risk measure.

3.2.1 Risk measure

In recent years, increasing attention has been paid to methods allowing a quantified measure to be associated with a risk, this risk possibly arising from a financial position or contingencies insured by an insurance company, the latter including risks. major varied such as natural disasters, pandemics, or industrial risks. In general, risk can be defined as a random variable representing a future value, but we focus on the risk of loss , no profits.

Since risks are modelled as non-negative rv's, measuring risk is equivalent to establishing a correspondence between the space of rv's and non-negative real numbers $\mathbb{R}^+$: The real number denoting a general risk measure associated with the risk

[1] the backtesting is the procedure of periodically comparing the forecasted risk measure with realized values in the financial market.

X will henceforth be denoted as $\rho[X]$: Thus, a risk measure is nothing but a functional that assigns a non-negative real number to a risk. See Szegö (2004) for an overview. It is essential to understand which aspect of the riskiness associated with the uncertain outcome the risk measure attempts to quantify.

Consider a probability space $(\Omega, A, \mathbb{P})$ where:

- Ω represents the set of all possible scenarios;
- A is a tribe;
- $\mathbb{P}$ is a measure of probability.

The future value of a scenario is uncertain and can be represented by a r.v X. This is a function of all possible scenarios to the real numbers, $X : \Omega \to \mathbb{R}$.

Definition 3.1 *(risk)*
Let (Ω, A) *be a measurable space where* Ω *is the results space and A is the tribe defined above. A risk is a random variable defined on* (Ω, A).

When $X > 0$, we call it a loss, whereas when $X < 0$, we call it a gain.

The class of all random variables on (Ω, A) is denoted by $\mathcal{X}$. $\mathcal{X}$ contains the constants and is stable by addition and multiplication by a scalar.

As its name suggests, a risk measure quantifies the danger inherent in a risk represented by a random value X (measuring risks means establishing a correspondence between the random variable representing the risk and a non-negative real number).

Here we take up the definition of a risk measure as formalized in Denuit and Charpentier (2005) [**?**, **?**]

Definition 3.2 *(risk measure)*
A risk measure is a functional ρ *mapping a risk X to a non-negative real number* $\rho[X]$ *; possibly infinite, representing the extra cash which has to be added to X to make it acceptable.*

The idea is that ρ quantifies the riskiness of X: large values of $\rho[X]$ tell us that X is 'dangerous'. Specifically, if X is a possible loss of some financial portfolio over a time horizon, we interpret $\rho[X]$ as the amount of capital that should be added as a buffer to this portfolio so that it becomes acceptable to an internal or external risk controller. In such a case, $\rho[X]$ is the risk capital of the portfolio.

Such risk measures are used for determining provisions and capital requirements in order to avoid insolvency.

Another function that is useful in analysing the thickness of tails is the "mean-excess loss".

Definition 3.3 *Given a non-negative rv X; the associated mean-excess function (mef) e_X is defined as*

$$e_X(x) = \mathbb{E}[X - x | X > x] \, , x > 0 \tag{3.1}$$

The **mef** corresponds to the well-known expected remaining lifetime in life insurance. In reliability theory, when X is a non-negative rv, X can be thought of as the lifetime of a device and $e_X(x)$ then expresses the conditional expected residual life of the device at time x given that the device is still alive at time x.

There are many risk measures introduced in literature and practice, and choosing a risk measure can be difficult. One approach to dealing with the issue of risk measurement is to start with a list of properties that a risk measurement must satisfy.

To meet the need for theoretical and practical principles, it is customary for a risk measure to verify a number of properties. A list of axiomatic properties for a good measure of risk was introduced in the seminal article by Artzner et al. The verification of these properties leads to the notion of a coherent risk measurement.

Definition 3.4 *Two random variables with real values X and Y on (Ω, A) are comonotonic if*

$$(X(w) - X(\acute{w}))(Y(w) - Y(\acute{w})) \geq 0, \ \forall (w, \acute{w}) \in \Omega \times \Omega \tag{3.2}$$

Proposition 3.1 *For two random variables with real values X and Y on (Ω, A) the following conditions are equivalent:*

i) X and Y are comonotonic,

ii) They exist a r.v. Z on (Ω, A) and two non-decreasing functions f and g on $\mathbb{R}$ such that $X = f(Z)$ and $Y = g(Z)$.

Definition 3.5 *A risk measure $\rho : X \to \mathbb{R}$ which satisfies $\rho(X) = \rho(Y)$ for all $X, Y \in \mathcal{X}$ such that X and Y have the same distribution below ρ is called a law invariance risk measure.*

Definition 3.6 *A function ψ, defined over an interval I, is convex over I if the part of the plane located above the curve is convex, i.e., if any arc of its representative curve is located below the corresponding chord. This definition translates into:*

$$\psi(kx_1 + (1-k)x_2) \leq k\psi(x_1) + (1-k)\psi(x_2) \tag{3.3}$$

$\forall k \in [0,1]$ and $\forall x_1, x_2 \in I$. If $-\psi$ is convex, ψ is said to be concave.

For the definitions of all axioms, X and Y are random variables representing loss, $c \in \mathbb{R}$ is a scalar representing loss, and ρ is a risk measure.

Axiom 3.1 *Translation invariance:* $\rho(X+c) = \rho(X)+c$, *for any* $X \in \mathcal{X}$ *and for any* $c \geq 0$.

Axiom 3.2 *Monotonicity: If* X *and* Y *are two losses such that* $X < Y$ *, then* $\rho(X) \leq \rho(Y)$.

Axiom 3.3 *Positive homogeneity:* $\rho(cX) = c\rho(X)$ *, for all* $X \in \mathcal{X}$ *and all* $c \geq 0$.

Axiom 3.4 *Subadditivity:* $\rho(X+Y) \leq \rho(X)+\rho(Y)$ *for all* $X, Y \in \mathcal{X}$.

Axiom 3.5 *Additivity for comonotonic risks:* $\rho(X+Y) = \rho(X)+\rho(Y)$ *for all* $X, Y \in \mathcal{X}$ *such that* X *and* Y *are comonotonic.*

Artzner et al. analyzed measures of risk and stated a set of axioms that should be desirable for any measure of risk. Any measure of risk that satisfies these axioms is said to be coherent.

3.2.1.1 Coherent risk measures

Several authors have selected some of the precedent conditions to form a set of requirements that any risk measure should satisfy. The first class of risk measures which was introduced by Artzner et al (1999 [5]) is the coherent risk measures. And was constructed to possess all mathematical properties to properly reflect the economy. And hence it takes the second step within the risk measurement procedure into account. A risk measure is called coherent if it satisfies axioms presented in the following definition.

Definition 3.7 *A risk measure that is translative, positive homogeneous, subadditive and monotone is called coherent.*

Theorem 3.6 *A risk measure is coherent if and only if there is a family* $\mathcal{P}$ *of probability measure over all the states of nature such as:*

$$\rho(X) = \sup\{\mathbb{E}_P(X)/P \in \mathcal{P}\} \tag{3.4}$$

Proof. See Artenzer et al.(1999 [5]) for a demonstration. ■

This axiomatic definition is the cornerstone of a very rich theory which draws its modules from functional analysis and has interesting economic interpretations.

It is worth mentioning that coherence is defined with respect to a set of axioms, and no set is universally accepted. Modifying the set of axioms regarded as desirable leads to other 'coherent'risk measures.

3.2.2 Ways of measuring risk

Several families of risk measures are presented in the risk theory literature. The usual measures most used by practitioners are:

3.2.2.1 Value-at-Risk

One of the most popular risk measures is Value-at-Risk (VaR), it was developed in the 1990's as a response to financial disasters. Although developed in the 1990's, the methodology behind VaR is not new, it can be traced back to 1952 to the basic mean-variance framework

Definition 3.8 *Given a risk X and a probability level $p \in [0;1]$; the corresponding VaR; denoted by $VaR(X;p)$; is defined as*

$$VaR(X;p) = F_x^{-1}(p) \tag{3.5}$$

Note that the VaR risk measure reduces to the percentile principle of Goovaerts et al. (1984 [74]).

Value-at-Risk remains one of the main risk indicators for the management of financial portfolios. It is worth mentioning that VaR's always exist and are expressed in the proper unit of measure, namely in lost money. Since VaR is defined with the help of the quantile function F, all their properties immediately apply to VaR. We will often resort to the following equivalence relation, which holds for all

$$VaR(X;p) \leq x \Leftrightarrow p \leq F_x(x)$$

VaR fails to be subadditive (except in some very special cases, such as when the X_i are multivariate normal). Thus, in general, VaR has the surprising property that the VaR of a sum may be higher than the sum of the VaR's. In such a case, diversification will lead to more risk being reported. Consider two independent Pareto risks of parametre $1; X$ and Y : Show that the inequality

$$VaR(X;p) + VaR(Y;p) < VaR(X+Y;p)$$

holds for any p; so that VaR cannot be subadditive in this simple case. A possible harmful aspect of the lack of subadditivity is that a decentralized risk management system may fail because VaR's calculated for individual portfolios may not be summed to produce an upper bound for the VaR of the combined portfolio.

3.2.2.2 Tail Value at Risk (TVaR)

In addition to highlighting the inconsistency of the VaR, Artzner et al., member of coherent risk measures has also been proposed as an alternative risk measure: the Tail Conditional Expectation or Tail Value-at-Risk (denoted TCE or $TVaR$).

Definition 3.9 *Given a risk X and a probability level p; the corresponding $TVaR$, denoted by $TVaR(X,p)$; is defined as*

$$TVaR(X,p) = \frac{1}{1-p}\int_p^1 Var[x,\zeta]d\zeta, \quad 0 < p < 1 \tag{3.6}$$

We thus see that $TVaR(X,p)$ can be viewed as the arithmetic average of the VaR's of X; from p on $[0;1]$.

3.2.2.3 Some related risk measures

3.2.2.3.1 Conditional Tail Expectation The conditional tail expectation (CTE) represents the conditional expected loss given that the loss exceeds its VaR.

Definition 3.10 *For a risk X, the Conditional Tail Expectation ($\mathbb{CTE}$) at probability level $p \in (0;1)$ is defined as:*

$$\mathbb{CTE}(X,p) = E(X \backslash X > VaR(X,p)) \tag{3.7}$$

So the $\mathbb{CTE}$ is the 'average loss in the worst $100(1-p)\%$ cases'. Writing $d = VaR(X,p)$ we have a critical loss threshold corresponding to some confidence level p, $\mathbb{CTE}(X,p)$ provides a cushion against the mean value of losses exceeding the critical threshold d.

3.2.2.3.2 Conditional VaR An alternative to $\mathbb{CTE}$ is the conditional VaR (or $CVaR$). The $\mathbb{CTE}(X,p)$ is the expected value of the losses exceeding VaR.

$$\begin{aligned} CVaR(X) &= \mathbb{E}((X - VaR(X,p) \backslash X > VaR(X,p)) \\ &= \mathbb{CTE}(X,p) - VaR(X,p) \end{aligned} \tag{3.8}$$

It's easy to see that $CVaR$ is related to the mean-excess function through

$$CVaR = e_X(VaR(X,p)) \tag{3.9}$$

Therefore, evaluating the mef at quantiles yields $CVaR$.

3.2.2.3.3 Expected Shortfall As the VaR at a fixed level only gives local information about the underlying distribution, a promising way to escape from this shortcoming is to consider the so-called expected shortfall over some quantile. Expected shortfall at probability level p is the stop-loss premium with retention $VaR(X;p)$: Specifically,

$$\begin{aligned} \mathbb{ES}(X,p) &= \mathbb{E}((X - VaR_p)_+) \\ &= \pi_x(VaR(X,p)) \end{aligned} \tag{3.10}$$

3.2.2.4 Relationships between risk measures

The following relation holds between the risk measures defined above.

Proposition 3.2 *For any $p \in [0,1]$, The following identities are valid:*

$$TVaR(X,p) = VaR(X,p) + \frac{1}{1-p}\mathbb{ES}(X,p) \tag{3.11}$$

$$\mathbb{CTE}(X,p) = VaR(X,p) + \frac{1}{\overline{F_x}(VaR(X,p))}\mathbb{ES}(X,p) \tag{3.12}$$

$$CVaR(X,p) = \frac{\mathbb{ES}(X,p)}{\overline{F_x}(VaR(X,p))} \tag{3.13}$$

Proof. See Denuit et al (2005 [42]). ■

Corollary 3.1 *Note that if F_x is continuous then by combining (3.11) and (3.13) we find:*

$$\mathbb{CTE}(X,p) = TVaR(X,p),\ p \in [0,1] \tag{3.14}$$

so that $\mathbb{CTE}$ and $TVaR$ coincide for all p in thisin this special case. In general, however, we only have

$$TVaR(X,p) = \mathbb{CTE}(X,p) + (\frac{1}{1-p} - \frac{1}{\overline{F_x}(VaR(X;p))})\mathbb{ES}(X;p) \tag{3.15}$$

Since the quantity between the brackets can be different from 0 for some values of p; $TVaR$ and $\mathbb{CTE}$ are not always equal.

Remark 3.1 *Value-at-Risk, although widely used in finance, is not a coherent risk measure because it is not sub-additive; the same goes for the variance. On the other hand, the CVaR is a coherent risk measure.*

3.2.3 Income distribution

The discussion on income distribution and economic growth has gained importance since the emergence of the Great Recession. Central banks and Treasuries of developed countries avoided the collapse of the banking systems and the depression of these economies. The profitability of corporations has increased (*Richard Baldwin* 2011 [8]), but the economic growth of these economies has been low.

Income distribution is extremely important for development, since it influences the cohesion of society, determines the extent of poverty for any given average per capita income and the poverty-reducing effects of growth, and even affects people's health.

It finds that the Kuznets hypothesis that income distribution worsens as levels of income increase is not at all strongly supported by the evidence, while growth rates of income are not systematically related to changes in income distribution. However, evidence is accumulating that more equal income distribution raises economic growth.

A renewed interest in income distribution has developed because of recent history of the personal income distribution. After several decades of apparent stasis from the late 1970s onwards there has been a remarkable increase in the dispersion of incomes in many countries.

In economic analysis income distributionîis interpreted in two principal ways: the functional distribution of income - i.e. the distribution of income among factors- and the size distribution of income (or distribution of income among persons).

"Why the focus on income rather than some other measurable quantity?"

In many treatments of the subject income plays one of two roles, sometimes both:

- Income as a proxy for economic welfare. If one adopts an individualistic, welfarist approach to social economics then it is reasonable to be concerned with individual well-being or utility. In some respects the low of income captures this, but it has been argued that consumption expenditure may be a more appropriate economic indicator. It should also be acknowledged that individual well-being may be determined not only by the level of one's own income but also by its relation to the incomes of others

- Income as command over resources. This role of income can be interpreted in more than one way. If one has in mind spending power then perhaps disposable income (income after taxes and compulsory deductions) may be an appropriate concept. But if inequality is associated with economic power and status then a measure of wealth may be more appropriate.

3.2.4 Definition of inequality

There are many reasons why policymakers and researchers alike are concerned with a country's degree of economic inequality. Recent studies show that persistent income disparities among individuals are associated with poverty and deprivation, mental illness, social unrest, and crime, as well as lower levels of education, employment, and life expectancy (Stiglitz, 2012[124]). Many public policies such as taxes, welfare benefits, provision of education and health services, price, and competition regulations have distributional implications for income.

In the Oxford Dictionary of Economics, inequality is the differences in the distribution of economic stocks or flows among economic agents. For example, wealth inequality refers to the distribution of the stock of wealth, whereas income inequality refers to the distribution of the flow of income' (Black et al. 2012). Inequality is broadly defined as the 'unequal rewards or opportunities for different individuals within a group or groups within a society' (Scott and Marshall, 2009 [118]). This definition mentions two aspects of inequality:

- Unequal rewards,
- Unequal opportunities.

Researchers in the area of income distribution agree that a higher mean income increases social welfare while higher inequality decreases it. The question is how to measure inequality. In his seminal paper, Atkinson (1970 [6]) proved that if one Lorenz curve is always not lower than the other, then the income inequality in the distribution with the upper Lorenz curve is smaller than the inequality in the lower curve for every additive concave social welfare function, provided that the two distributions have equal means. If, on the other hand, the Lorenz curves of two distributions intersect, then one can always find two additive concave social welfare functions that will rank inequality in the two distributions in a reversed order. Shorrocks (1983 [121]) extended the above result to comparisons of distributions with different means. He showed that having a distribution with an absolute Lorenz curve that is always not lower than the other forms a necessary and sufficient condition for the expected value of every concave social welfare function to be greater than the expected welfare of the distribution with the lower absolute Lorenz curve.

The term inequality can be explained from a mathematical point of view as a distinction between two or more particular characters, provided that these characters can be quantified. From another perspective, the notion of inequality by many people is perceived as a failure to achieve equal opportunities, which is determined by different circumstances that people can not influence.

Inequality can have many dimensions. Economists are concerned specifically with the economics or monetarily-measurable dimension related to individual or household income and consumption. However, this is just one perspective and inequality can be linked to inequality in skills, education, opportunities, happiness, health, life expectancy, welfare, assets and social mobility.

An explanation for an automatic fall in inequality in rich countries from 1914 to 1945 could be simply because of the World Wars, the great economic depression, and political shocks (Piketty, 2014 [113]). Many critiques (for instance, Anand and Kanbur 1993 [2]) point out that there has been not an apparent and significant 'bell curve' relationship between economic growth and inequality within a country. Another evidence is an expansion in income gap in the Europe in the 1980s and 1990s (Doerrenberg and Peichl 2014 [43]). Even when supporting Kuznets's hypothesis in the case of early stage of the transformation from agricultural to industrial economy, economists (e.g. Ahluwalia 1976, Barro 2000) cannot predict a turning point where income inequality stops accelerating, or when it starts to decline.

"Income inequality" is the extent to which income is distributed unevenly across people or across households. Income encompasses labor earnings (such as wages, salaries, and bonuses), capital income derived from dividends, interest on savings accounts, rent from real estate, as well as welfare benefits, state pensions, and other government transfers. In addition, it is possible to distinguish between individual versus family income, pre-tax versus after-tax (disposable) income, and labor earnings versus capital income.

3.3 Measuring income inequality

The aim of this section is to summarise the major statistical measurements of inequality. This era has witnessed a considerable evolution of inequality economics with a variety of measurements of inequality. Applications of mathematical techniques for research in inequality have had an advantage since data indicating inequality has been better recorded from the beginning of the twentieth century (Piketty 2014 [113]). Furthermore, the exclusion of political elements in theoretical measurements of inequality allows the economics of inequality to focus entirely on economic issues. The majority of economists no longer classify society into three classes (i.e. capitalists, landlords and proletarians). Grouping individuals is instead subjected to different criteria which are based on particular research contexts such as income, educational background, gender, age and ethnicity. Measurements of inequality have been widely applied to many countries in the final third of the twentieth century. Thus, the literature on economic inequality should reflect these

empirical results of inequality measures.

It is well known that different indicators of income inequality can send conflicting messages about the evolution of inequality, both within countries and across time (Cowell, 2011 [28]; and World Bank, 2016 [2]). But even in a given country at one point in time, the same income inequality indicator can suggest significantly different levels of inequality.

There is a wide variety of inequality indices in common use, below we will discuss the most important indicators and the most common and used:

3.3.1 The GMD and GINI coefficient.

A century later, the Gini concentration ratio is still of great interest for the international scientific community. In fact, the large scientific production on the Gini concentration ratio in the last decades seems to confirm the long wave of its topical interest. In particular, new extensions, interpretations, and uses have continued to keep the interest for this index alive. The Gini concentration ratio is typically computed using data coming from sample surveys. Therefore, it should not be used only as a descriptive measure, but a formal statistical inference is rather necessary.

The Gini concept or the mean difference of the Gini, initiated by Gini in 1912, is a characteristic of widespread dispersion in the field of income distribution. The specificity of this indicator lies in its simple calculations. The Gini index uses the Euclidean distance between all pairs in the sample

Gini's mean difference (**GMD**) was first introduced by Corrado Gini in 1912 as an alternative measure of variability. **GMD** and the parameters which are derived from it (such as the Gini coefficient, also referred to as the concentration ratio) have been in use in the area of income distribution for almost a century, and there is evidence that the GMD was introduced even earlier (Harter, 1978 [78]). In other areas it seems to make sporadic appearances and to be "rediscovered" again and again under different names.

Gini's mean difference is defined as

$$GMD = 2\int F(t)[1-F(t)]dt \tag{3.16}$$

where $F(t)$ represent the cumulative distribution function. See Dorfman (1979[49]).

The most traditional member of the income inequality family is the Gini coefficient. It is widely used to measure income inequality, mainly because of its intuitive geometric interpretation.

[2] World Bank, 2016, "Poverty and Shared Prosperity 2016: Taking on Inequality," Washington, D.C: World Bank.

The Gini Coefficient is well established as a conventional, ad hoc measure of income inequality. Recently there has been a flurry of interest in it, stirred up by a debate about its significance as a measure of economic welfare (Atkinson, 1970[6]; Rothschild and Stiglitz, 1973 [124]; Sen, 1973 [119]) in the course of which a confusing variety of formulas for the coefficient have been published, some of them quite complicated (Atkinson, 1970 [6]; Sen, 1973 [119]; Theil, 1967[127], for example). The Gini coefficient was developed independently of the GMD, directly from the Lorenz curve and for a while it was called "the concentration ratio." Gini (1914) has shown the connection between the GMD and the concentration ratio. Ignoring the differences in definitions, the relationship between the GMD and the Gini coefficient is similar to the one between the variance and the coefficient of variation, $CV = \frac{\sigma}{\mu}$ a property that was already known in 1914. That is, the Gini coefficient is a normalized version of the GMD and it is unit-free (measured in percent). In order to calculate it one only needs to derive the GMD, and then easily convert the representation into a Gini coefficient by dividing by twice the mean.

We will propose a simple formula for the Gini Coefficient that will apply to both discrete and continuous distributions of income and will be well defined and valid whether or not there is a finite upper limit to the income that can be received by anyone, provided the mean of the distribution is finite.

Definition 3.11 *The Gini index is defined as twice the area between the egalitarian line and the Lorenz curve.*

Thus, if X is a random variable in $\mathcal{L}$ with Lorenz curve L_X, a formula for its Gini index, $G(X)$ or simply G if the random variable is known from the context, is

$$G(X) = 2\int_0^1 [u - L_X(u)]du = 1 - 2\int_0^1 L_X(u)du \tag{3.17}$$

Theorem 3.7 *The Gini index can be expressed as*

$$G(X) = 2\int_0^1 u\grave{L}_X(u)du - 1 \tag{3.18}$$

Theorem 3.8 *The Gini index can be written as*

$$G(X) = 1 - \frac{E(X_{1:2})}{\mu} = 1 - \frac{1}{\mu}\int_0^\infty [1 - F_X(x)]^2 dx \tag{3.19}$$

where $X_{1:2}$ is the smaller of a sample of size 2 with the same distribution as X.

If we denote by F the cumulative distribution function (CDF) of the incomes

under study, the Lorenz curve is defined implicitly by the equation

$$L_X(F(x)) = \frac{1}{\mu}\int_0^x y dF(y), \tag{3.20}$$

where $\mu = \int_0^\infty y dF(y)$ is expected income. It is assumed that there are nonegative incomes. The function L_X is increasing and convex, and maps the $[0,1]$ interval into itself. Twice the area between the graph of L and the 45^o-line is then

$$G = 1 - 2\int_0^1 L(y)dy \tag{3.21}$$

Using the definition (3.20) in (3.21), we find that:

$$\begin{aligned} G &= 1 - 2\int_0^\infty L(F(x))dF(x) \\ &= 1 - \frac{2}{\mu}\int_0^\infty \int_0^x y dF(y)dF(x) \end{aligned}$$

Then,on interchanging the order of integration and simplifying, we obtain:

$$\begin{aligned} G &= 1 - \frac{2}{\mu}\int_0^\infty y\int_y^\infty dF(x)dF(y) \\ &= 1 - \frac{2}{\mu}\int_0^\infty y(1-F(y))dF(y) \\ &= 1 + \frac{2}{\mu}\int_0^\infty yF(y)dF(y) - 2 \\ &= \frac{2}{\mu}\int_0^\infty yF(y)dF(y) - 1 \end{aligned} \tag{3.22}$$

The last expression above corresponds to a result cited in Modarres and Gastwirth (2006) [102] according to which G is $2/\mu$ times the covariance of Y and $F(Y)$, where Y denotes the random variable "income" of which the CDF is F. There are of course numerous other ways of expressing the index G, but 3.22 is most convenient for present purposes.

The Gini index is a complex and synthetic indicator of inequality. It provides condensed information on the distribution of income, but not on its characteristics, such as location and form. It is also an indicator associated with the descriptive approach to measuring inequality.

Morrison points out that: "The most important and the most common is the Gini coefficient." It must be recognized, however, that the scope of the results of this calculation also has limits. The Gini coefficient is too global and does not clearly distinguish the three social categories (rich, middle, poor). Chauvel raises that "theoreti-

cally and practically, Gini is a measure far too crude to provide a reliable diagnosis of inequalities".

The intersection of Lorenz curves is also a major limitation to Gini's results. "The Gini coefficients are only supposed to offer a valid instrument of comparison between two or more societies if the associated Lorenz curves don't intersect", Chauvel. When these curves intersect, a lot of information is hidden.

According to Morrisson, "the downside of the Gini coefficient is that very different Lorenz curves can correspond to the same value of the Gini index." We can see this in the next exemple:

Example 3.1 *Two income distributions with the same Gini index*

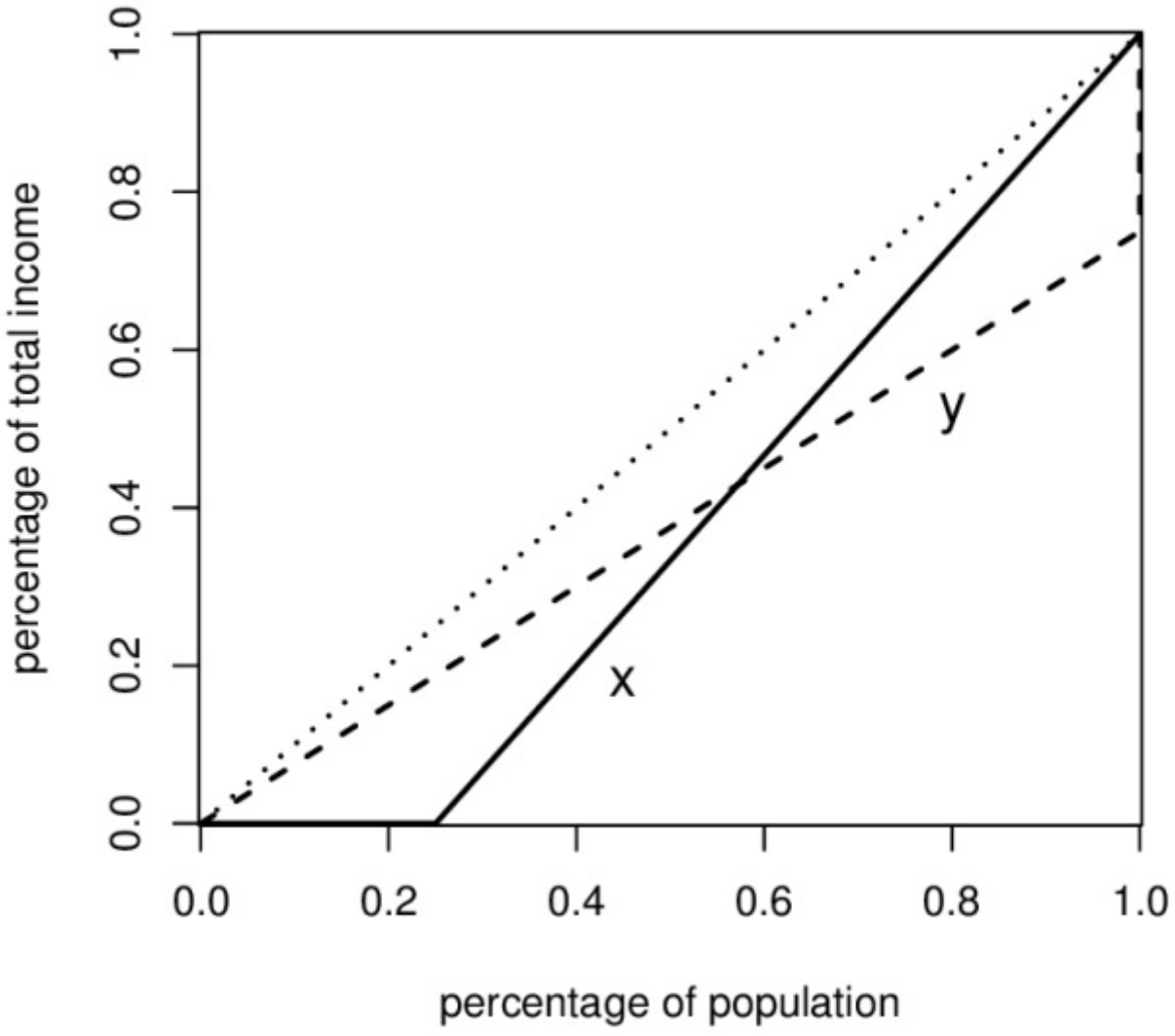

Figure 3.1: Lorenz Curve

These two distributions are represented by the Lorenz curves in Figure 3.3.1. Since they intersect, they cannot be used to classify them in terms of income inequality. But the way they intersect gives areas of equal value before and after the intersection. It can be concluded that the Gini index is the same, despite the large income differences.

3.3.2 Atkinson's measurement of inequality

Atkinson (1970 [6]) illustrates an alternative measurement of inequality calculated as follows:

$$I_A = 1 - \frac{Y_{EDE}}{\mu}$$

where Y_{EDE} is defined as "the equally distributed equivalent income"; and μ is the average real income.

His distinguishing idea is to emphasize the relationship between inequality and social welfare based on the aggregation of individual utilities. Equal distribution only occurs when "the equally distributed equivalent income", Y_{EDE}, is equal to the mean income. An absence of this ideal condition implies that Y_{EDE} deviates from the mean, μ; the larger the difference between Y_{EDE} and μ, the higher the inequality level. The result of this is that social wealth loss is proportionate to the level of inequality.

Alternatively, using the social welfare function (SWF)[3], inequality can be measured as follows:

$$I_A = 1 - \left[\sum_i \left(\frac{y_i}{\mu}\right)^{1-\varepsilon} f(y_i\right]^{\frac{1}{1-\varepsilon}} \tag{3.23}$$

In this equation, the level of inequality is clearly subject to changes in the inequality aversion degree $-\varepsilon$. The greater the ε, the greater the weight dedicated to the lower end of the distribution. Using equation (**??**), Atkinson resolves the problem of the crossing of the Lorenz curve and estimates inequality with a partial ordering solution. In these cases, measuring inequality with the Lorenz curve could not produce sensible results. However, by choosing ε in the range of 1.5 to 2, the number of controversial comparisons was reduced to five cases. However, this approach depends heavily on the choice of a value for ε.

The Atkinson index is advantageous in terms of evaluation of lost value in economies due to inequity. This approach also provides a series of results depending upon the social attitude to inequality. The more a community is concerned about inequal-

[3]The social welfare function is a statistical 'aggregator' that turns a distribution into a single number that provides an overall judgement on that distribution and that forces us to think coherently about welfare and its distribution.

ity, the higher the inequality aversion parameter (ε). Subsequently, the index will be greater, irrespective of the distribution being the same. However, when compared with the Gini coefficient and Theil indices, Atkinson's measurement is unable to analyse inequality attributions to different subgroups; thus, it cannot be used as a decomposition technique for understanding within-inequality and between-inequality (Gisbert et al. 2009 [68])

3.3.3 The standard deviation method

A simple measurement of inequality is an estimate of the deviations of every member of the population from the standard deviation. Given a population having n individuals (i), with a semi-infinite income distribution (range from 0 to $+\infty$) and the mean income ($\overline{y}$), the variance of this distribution (var) is defined as the second moment about the mean or 'the mean of squares of the deviations from the mean' (Kendall and Stuart 1977 [89], pp.42- 47); it is computed as follows:

$$var = \int_0^{+\infty} (\overline{y} - y)^2 dF = \frac{1}{n} \sum_{i=1}^{n} (\overline{y} - y_i)^2 \tag{3.24}$$

Then the positive square root of variance is called standard deviation ($\sigma = \left|\sqrt{var}\right|$), which is also the root-mean-square.

Another measurement of inequality that can avoid the 'arbitrariness of the units' uses the standard deviation in the logarithmic form (SDL) (Sen 1973 [119]):

$$SDL = \sqrt{\frac{\sum_{i=1}^{n} (\ln \overline{y} - \ln y_i)^2}{n}} \tag{3.25}$$

Yet, Sen (1973) [119] finds that the measurement is not concave at the high income levels and only considers the distances between each of individuals' income and the mean income. This could be a reason for the absence of applications of SDL for inequality analysis, which is also analogous to the case of the mean log deviation.

3.4 The Lorenz curve and Gini coefficient

The contents of this section is built on the contributions of two scholars who lived more than a century ago: the Italian statistician Corrado Gini, and the American economist Max Otto Lorenz [4], begining with the definition of classical index of Gini.

The "Lorenz curve" is a common graphical method of representing the degree of

[4] Max Otto LORENZ (1880 - 1962) is the American economist who invented the graph representing the curve that bears his name in 1905.

income inequality in a country [64]. It plots the cumulative share of income (y axis) earned by the poorest x% of the population, for all possible values of x. The 45-degree line represents the line of equality, when income is shared equally among all individuals. If, however, income is not shared equally, then the bottom x% of individuals earn less than x% of total income in the country, implying that Lorenz curves typically lie below the 45-degree line. Moreover, the further away the Lorenz curve is from the equality line, the more unequal the income distribution.

Several inequality indices can be derived from the Lorenz diagram. The Lorenz Curve construction also gives us a rough measure of the amount of inequality in the income distribution. This measure is called the Gini Coefficient. The Gini coefficient is a standard measure of inequality defined as the area between the Lorenz curve and the line of perfect equality divided by the area below the perfect equality line.

The most common measure that economists and sociologists use is the Gini index mainly because of clear economic interpretation. The Gini concentration index has been estimated in di erent ways to obtain valid variance.

3.4.1 Lorenz curve

The Lorenz or the concentration curve play important roles in the areas of GMD and the related measures such as Gini covariance, Gini correlation, Gini regression, and more.

Historically, Lorenz (1905 [98]) presented the "Lorenz curve" as based on the relationship between the cumulative distribution of the variable (the horizontal axis) and the cumulative value of the percentage of the variate (the vertical axis). The Lorenz curve is a pivotal tool in the study of economic inequality and the distribution of wealth in the society[90]. Consider a positive continuous random variable Y, belonging to the $\mathcal{L}^{-1}$ class, i.e. $\mu = E(Y) < \infty$, and let $F(Y) = \mathbb{P}(Y \leq y)$ be its cumulative distribution function. Define the quantile function of Y as $Q(\alpha) = F^{-1}(\alpha)$ (where $F^{-1}(\alpha) = \inf\{y : F(y) \geq \alpha\}$ with $0 \leq \alpha \leq 1$).

Consider a non-negative random variable (rv) X with a distribution function (df) F, quantile function $Q(p)$, and finite mean $E(X) = \mu$. The Lorenz curve $L(x)$ is formally given by:

$$L(x) = \frac{\int_0^x Q(\alpha)d\alpha}{\int_0^1 Q(\alpha)d\alpha} \, , \, 0 \leq x \, , \, \alpha \leq 1 \qquad . \tag{3.26}$$

In terms of wealth, the Lorenz curve reads as follows: for a given $x \in [0,1]$, $L(x)$ tells us that $x \times 100\%$ of the population owns $L(x) \times 100\%$ of the total wealth. Such an interpretation tells that the Lorenz curve is scale-free: the total amount of wealth is not taken into consideration, whereas the way it is distributed among the individu-

als is the key information.

Given its strong relation with the quantile function Q, the Lorenz curve can recover the cumulative distribution of Y up to a constant. However, despite the Lorenz curve is theoretically a one-to one mapping with a given distribution, discriminate among distributions just looking at their Lorenz curves it is not an easy task to perform by hand.

Mathematically, the Lorenz curve $L : [0,1] \to [0,1]$ defined in equation 3.26 is a continuous, non-decreasing, convex function, almost everywhere differentiable in $[0,1]$,such that $L(0) = 0$ and $L(1) = 1$. The curve $L(x)$ is bounded from above by the so-called perfect equality curve, i.e. $L_{pe}(x) = x$, and from below by the perfect inequality curve, i.e.

$$L_{pi}(x) = \begin{cases} 0 \text{ si } 0 \leq x < 1 \\ 1 \text{ si } x = 1 \end{cases} \tag{3.27}$$

The perfect equality line L_{pe} indicates the theoretical situation in which everyone possesses the same amount of wealth in the economy, while the perfect inequality line L_{pi}, reachable only as limiting case for continuous random variables, states that only one individual owns all the wealth in the society.

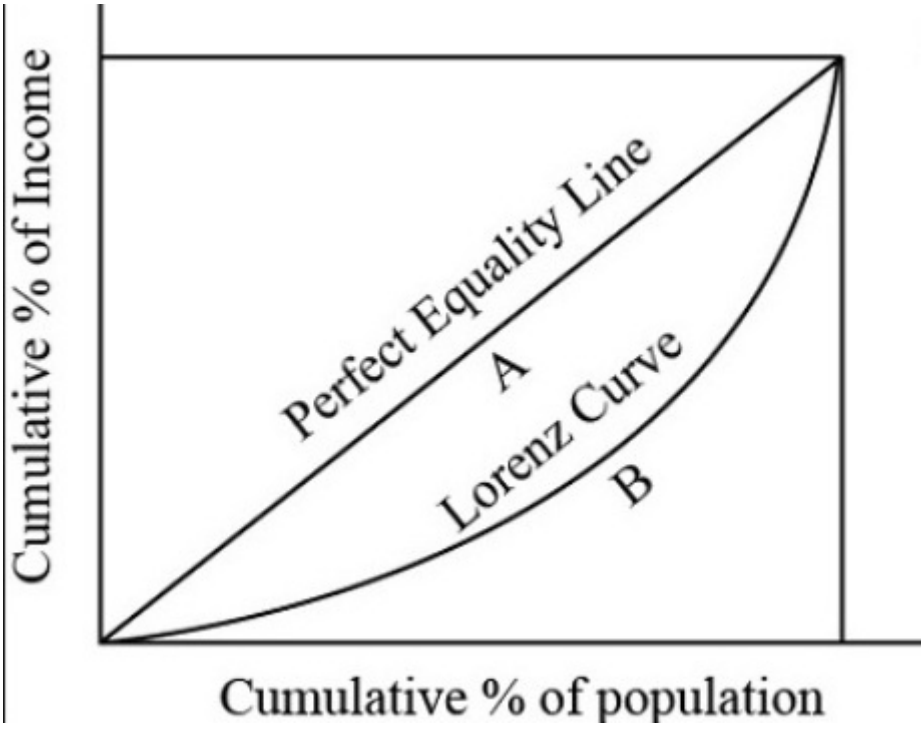

Figure 3.2: Illustration of Lorenz curve.

The very first mathematical definition of the Lorenz curve goes to **Kendall** and **Stuart** (1977 [89], who expressed it as two equations assuming an absolutely continuous distribution of income. Two years later, **Gastwirth** provided a general def-

inition of the Lorenz curve, applying to both continuous and discrete laws, in the form of a single formula **??**

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt \quad , 0 \leq p \leq 1 \tag{3.28}$$

where the income distribution and its inverse function are denoted by F and F^{-1}, respectively, and μ denotes the expectation. Due to its numerous applications in various fields such as economics (**Gastwirth)** and the Gini index, **Hart**, **Gail** and **Gastwirth** and tests, fishing (**Thompson,**1976) or even bibliometrics, the Lorenz curve has given rise to numerous works in nonparametric estimation. **Gastwirth** proposed a natural estimator of the Lorenz curve, defined by:

$$L_n(p) = \frac{1}{\overline{X}_n}\int_0^p F_n^{-1}(u)du \quad , 0 \leq p \leq 1 \tag{3.29}$$

where $\overline{X}_n$ represents the empirical mean of a sample of n independent observations $X_1, ..., X_n$ and with the same distribution F, F_n the empirical distribution function of this sample.

From this estimator, he was able to construct an estimator of the Gini index.

3.4.1.1 Properties

The Lorenz curve has several interesting mathematical properties.

1. It is entirely contained into a square, because p is defined over $[0,1]$ and $L(p)$ is at value also in $[0,1]$. Both the x-axis and the y-axis are percentages.

2. The Lorenz curve is not defined if μ is either 0 or ∞.

3. If the underlying variable is positive and has a density, the Lorenz curve is a continuous function. It is always below the 45^o line or equal to it.

4. $L(p)$ is an increasing convex function of p. Its first derivative:

$$\frac{dL(p)}{dp} = \frac{q(p)}{\mu} = \frac{x}{\mu} \text{ with } x = F^{-1}(p)$$

is always positive as incomes are positive. And so is its second order derivative (convexity). The Lorenz curve is convex in p, since as p increases, the new incomes that are being added up are greater than those that have already been counted. (Mathematically, a curve is convex when its second derivative is positive).

5. The Lorenz curve is invariant with positive scaling. X and cX have the same Lorenz curve

6. The mean income in the population is found at that percentile at which the

slope of $L(p)$ equals 1, that is, where $q(p) = \mu$ and thus at percentile $F(\mu)$. This can be shown easily because the first derivative of the Lorenz curve is equal to x/μ.

7. The median as a percentage of the mean is given by the slope of the Lorenz curve at $p = 0.5$. Since many distributions of incomes are skewed to the right, the mean often exceeds the median and $q(p = 0.5)/\mu$ will typically be less than one.

3.4.2 The Gini coefficient revisited

The most well-known member of the income inequality family is the Gini coefficient. The Gini mean difference and its normalized version, known as the Gini index, have aided decision makers since their introduction by Corrado Gini more than a hundred years ago.

The Gini index, which we denote by G_F, is usually interpreted as twice the area between the actual population Lorenz function, it is define by:

$$G_F = 1 - \frac{2}{\mu_F}\int_0^1 \left(\frac{1}{p}\int_0^p F(t)dt\right)dp \tag{3.30}$$

Several other equivalent ways to define the Gini index exist. An alternative expression is given by:

$$G = \frac{\eta}{\mu} - 1 \ = \frac{\int_0^\infty 2F(t)tdF(t)}{\mu} - 1 \tag{3.31}$$

(see **David** 1968) [34].

The Gini coefficient can be written in many different forms. We shall see how to pass from the standard definition of the Gini as a surface to its various expressions (covariance, mean of absolute difference). We shall suppose that the mean of F exists. As a consequence:

$$\lim_{t\to 0} tF(t) = \lim_{t\to\infty} t(1 - F(t)) = 0, \tag{3.32}$$

which means that both limits exists, which simplifies greatly the computation of some integrals when considering an infinite bound.

3.4.2.1 Gini coefficient as a surface

If everybody had the same income, the cumulative percentage of total income held by any bottom proportion p of the population would also be p. The Lorenz curve would then be $L(p) = p$: population shares and shares of total income would be identical. A useful informational content of a Lorenz curve is thus its distance, $p - L(p)$, from the line of perfect equality in income. Compared to perfect equality,

inequality removes a proportion $p - L(p)$ of total income from the bottom $100.p\%$ of the population. The larger that "deficit", the larger the inequality of income. There is thus an interest in computing the average distance between these two curves or the surface between the diagonal p and the Lorenz curve $L(p)$. We know that the Lorenz curve is contained in the unit square having a normalized surface of 1. The surface of the lower triangle is 1/2. If we want to obtain a coefficient at values between 0 and 1, we must take twice the integral of $p - L(p)$, i.e.:

$$G = 2\int_0^1 (p - L(p))dp = 1 - 2\int_0^1 L(p)dp \tag{3.33}$$

which is nothing but the usual Gini coefficient. Xu (2003 [133]) gives a good account of the algebra of the Gini index. We have given above an interpretation of the Gini index as a surface.

3.4.2.2 Gini as a covariance

Let us start from the above definition of the Gini coefficient and use integration by parts with $\grave{u} = 1$ and $v = L(p)$. Then

$$\begin{aligned} G &= 1 - 2\int_0^1 L(p)dp \\ &= 1 - 2[pL(p)]_0^1 + 2\int_0^1 p\grave{L}(p)dp \\ &= -1 + 2\int_0^1 p\grave{L}(p)dp \end{aligned} \tag{3.34}$$

We are then going to apply a change of variable $p = F(y)$ and use the fact proved above that $\grave{L}(p) = y/\mu$. We have

$$\begin{aligned} G &= \frac{2}{\mu}\int_0^\infty yF(y)f(y)dy - 1 \\ &= \frac{2}{\mu}[\int_0^\infty yF(y)f(y)dy - \frac{\mu}{2}] \end{aligned} \tag{3.35}$$

This formula opens the way to an interpretation of the Gini coefficient in term of covariance as

$$cov(y, F(y)) = \mathbb{E}(yF(y)) - \mathbb{E}(y).\mathbb{E}(F(y))$$

Using this definition, we have immediately that

$$G = \frac{2}{\mu}cov(y, F(y)) \tag{3.36}$$

which means that the Gini coefficient is proportional to the covariance between a variable and its rank. The covariance interpretation of the Gini coefficient open the way to numerical evaluation using a regression.

Meanwhile, noting that $cov(y, F(y)) = \int y[F(y) - \frac{1}{2}]dF(y)$, using integration by parts, we get

$$cov(y, F(y)) = \frac{1}{2}\int F(x)[1 - F(x)]dx$$

so that we arrive at the integral form

$$G = \frac{1}{\mu}\int F(x)[1 - F(x)]dx \tag{3.37}$$

We can remark that $F(x)(1 - F(x))$ is largest at $F(x) = 0.5$, which explains why the Gini index is often said to be most sensitive to changes in incomes occurring around the median income.

The above integral form can also be written as

$$G = 1 - \frac{1}{\mu}\int [1 - F(x)]^2 dx$$

We shall prove this equivalence by considering the last interpretation of the Gini which is the scaled mean of absolute differences.

3.4.2.3 Gini as mean of absolute differences

The initial definition of the Gini coefficient is the mean of the absolute differences divided by twice the mean. If y and x are two random variables of the same distribution F, this definition implies

$$I_G = \frac{1}{2\mu}\int_0^\infty \int_0^\infty \left|x - y\right| dF(x)dF(y) \tag{3.38}$$

As $F(x)$ and $1-F(x)$ are simply the proportions of individuals with incomes below and above x, integrating the product of these proportions across all possible values of x gives again the Gini coefficient, in its form $\frac{1}{\mu}\int F(x)[1 - F(x)dx]$. If we decide to proceed step by step, we first note that $\left|x - y\right| = (x + y) - 2\min(x, y)$, so that the expectation of this absolute difference is

$$\Delta = \mathbb{E}\left|x - y\right| = 2\mu - 2\mathbb{E}(\min(x, y))$$

To compute the last expectation, we need the distribution of the Min of two random variables having the same distribution. We know or we can show that it is

equal to $1-(1-F(y))^2$, while its derivative is $-d(1-F(y)$. So that

$$\Delta = 2\mu + 2\int_0^\infty y d(1-F(y))^2 \tag{3.39}$$

The last integral can be transformed using integration by parts with $u = y$, and $v = (1-F(y))^2$

$$\int_0^\infty y d(1-F(y))^2 = [y(1-F(y))^2]_0^\infty - \int [1-F(y)]^2 dy$$

So that we get the integral form of the Gini

$$I_G = \frac{\Delta}{2\mu} = 1 - \frac{1}{\mu}\int [1-F(x)]^2 dx, \tag{3.40}$$

because the first right hand term is zero.

3.4.2.4 The main properties of the Gini index

We will describe the main properties of the Gini Index in terms of the axioms it respects.

The main properties of the Gini Index are:

- G **has zero as lower limit for any** v. When all incomes are equal, the covariance between income levels and the cumulative distribution function is zero. The Gini Index is therefore zero. With regard to the geometrical interpretation of the standard Gini Index, note that when all incomes are equal, the Lorenz Curve is equal to the equidistribution line. Therefore, the sum of areas of the polygons (Z) is equal to $\frac{1}{2}$, i.e. the sum of the triangle under the Lorenz Curve. Therefore, the Gini Index ($1-2Z$) is equal to zero.

- The standard Gini Index G **has** $\frac{n-1}{n}$ **as upper limit**. The limit of this value, for very large populations, is 1. When all incomes are zero except for the last, the last income is also equal to total income, $y = Y$. It means that there is only one area to calculate, i.e., the last trapezium. However, for very large populations, this area tends to be smaller. In the limit (i.e. in a continuous framework) the value of the area Z tends towards zero. Therefore, the Gini Index tends towards 1. As a generalisation, $G(v)$ has $\frac{2}{v}\frac{n-1}{n}$ as upper limit. Remember, that the standard Gini Index is one in which $v = 2$.

- The Gini Index is **scale invariant**. By multiplying all incomes by a factor α, the value of the Gini Index G does not change. Intuitively, when all incomes

are scaled by a common factor, the cumulative distribution of income does not change, as a given fraction of population still holds the same fraction of total income. The areas under the Lorenz Curve, therefore, do not change. With regard to the covariance formula, the application of a common factor to all incomes makes the covariance and the average income increase by the same factor. The Gini Index does not change. The same is true for $G(v)$.

- On the other hand, the Gini Index G is **not translation invariant**. By adding/subtracting the same amount of money to all incomes, the Gini Index would increase (decrease) accordingly. The same is true for $G(v)$.

- The Gini Index **satisfies the principle of transfers for any** v. If income is redistributed from relatively richer individuals to relatively poorer individuals, both G and $G(v)$ decrease. The opposite holds true if income is redistributed from relatively poorer to relatively richer individuals. With regard to the standard Gini Index, we note that the size of its change, following a change in any income, depends on the rank of the individuals involved in redistribution and on the sample size. It does not depend on the level of individual incomes involved in redistribution, but it depends on total income. In particular, the Gini Index reacts more to redistribution occurring among individuals who have a greater difference in ranks. The same amount of redistribution, indeed, generates a much lower effect if the two individuals have a close rank.

3.4.3 The extended Gini family of measures

The GMD has many alternative presentations. Some of these alternative presentations can be extended into families of variability measures and the GMD can be viewed as one member of such a family.

A generalization of the Gini coefficient, called the extended Gini coefficient, was introduced by Yitzhaki (1983 [135]). The extended Gini family (EG) is a family of variability and inequality measures that depends on one parameter, the extended Gini parameter. The investigator can choose a member of the family by assigning a value to the parameter.

One advantage of having a family lies in the fact that one can perform a sensitivity analysis and evaluate the robustness of the conclusions by changing the EG parameter.

The basic definitions of the members of the EG family used in this part are based on the covariance. In order to simplify the presentation we will use $cov(X, F(X))$ as the Gini, ignoring the constant (4) that is needed to adjust the definition to the

GMD. Even with this simplification, the fact that the extended Gini is being used in different areas resulted in two alternative definitions. Let

$$\Delta(\theta, X) = -\theta cov(X, [1 - F(X)]^{\theta-1}) \ , \ \theta > 0, \ \theta \neq 1$$

Then the two definitions refer to $\theta = \upsilon$ and $\theta = \upsilon + 1$, where υ is the extended Gini parameter. More explicitly, the first definition is

$$\Delta(\upsilon, X) = -\upsilon cov(X, [1 - F(X)]^{\upsilon-1}) \ , \ \upsilon > 0, \ \upsilon \neq 1$$

This definition is mainly used in the areas of income distribution and finance, due to the need to adjust the definition to the theory of stochastic dominance. The other definition is

$$\Delta(\upsilon, X) = -(\upsilon + 1) cov(X, [1 - F(X)]^{\upsilon}) \ , \ \upsilon > (-1), \ \upsilon \neq 0$$

The definition (..) is mainly used in the area of econometrics, in which case the term $(\upsilon + 1)$ cancels because the parameters are expressed as ratios. The motivation for the different definitions is the need for a simple representation, relevant to the specific application.

The extended Gini index can be written as a covariance between the variate and a power function of its cumulative distribution. Specifically:

$$EG(X, \upsilon) = \upsilon cov(X, -[1 - F_X(X)]^{\upsilon-1})$$

where $F_X(x)$ is the cumulative distribution. The advantage of the covariance formula is that one can extend it quite easily to define the Gini covariance and Gini correlation. The latter can be used to decompose the EG of a sum of random variables to the contributions of the EG of each variable and the correlations, while other properties of the covariance may enable the imitation of ANOVA-like analysis.

Chapter 4

Estimating of the Gini index

Introduction

In this chapter, we discuss the current relevance of an index proposed by the Italian scientist more than a century ago. In 1914, Corrado Gini introduced his well-known concentration index for measuring the degree of inequality in the distribution of income and wealth. A century later, this index is still extremely relevant and widely used in several fields of research and application, such as economics, statistics, medicine, biology, ecology, and so on.

Due the simplicity and ease of interpretation, and thanks to its intuitive graphical relation with the Lorenz curve, the Gini index succeed. In addition, several sources have contributed to highlight its applicability in different aspect.

In fact, as stressed in Forcina and Giorgi (2005 [60]), "the political and economic debate on the way to reach a more equal distribution of income and wealth was particularly alive at the beginning of the last century." Gini first handled this issue in 1909 by proposing the index δ for describing the relations between social classes and distribution of wealth. He later introduced the mean difference (1912). Finally, in 1914 Gini developed the well-known index of concentration (1914), showing also the relation with the Lorenz curve and the mean difference. A few months later, Pietra (1915 [112]) proposed a simple geometrical interpretation of the Gini index of concentration, but after Gini's death, the Italian statistical academy did not follow nor develop his ideas immediately. Only in the last decades (starting from the 1970s) we have witnessed the rediscovery, extension, and reinterpretation of the Gini index of concentration. Important stimulus for the proliferation of studies on the Gini index have been provided, in particular, by Atkinson (1970 [6]), and Sen (1973 [119]).

The first section of this chapter is devoted to the construction of a semi-parametric estimator of the Gini index in the case of a heavy-tailed income distribution, to es-

tablish its asymptotic distribution and to derivate the confidence limits. In the second section, we construct an other estimator of the Gini index when the data are censored, then we study the asymptotic normality property. We show the performance of the estimator proposed by some simulation results.

4.1 Estimating the Gini index for heavy-tailed income distributions

This part was the subject of an article published in the paper [10].

Abstract

In the present section, we define and study one of the most popular indices which measure the inequality of capital incomes, known as "Gini index", we construct a semiparametric estimator for the Gini index in case of heavy-tailed income distributions, we establish its asymptotic distribution and derive bounds of confidence. We explore the performance of the confidence bounds in a simulation study and draw conclusions about capital incomes in some income distributions.

KeyWords: Heavy-tailed incomes, extreme quantile, income inequality, Gini index.

4.1.1 Introduction and motivation

The last decade has seen considerable use and development of statistical theory for inferring the dominance of one distribution (of income, wealth, wages, etc.) over another. The results thus provide the statistical framework within which to assess the progressivity of taxes and benefits, and the changes, in the inequality of income, or in the ranking of individuals with respect to income, which they may cause. The results can also be applied to the impact on poverty indices of a tax and benefit system, or of other socio-economic phenomena, when such poverty indices depend on estimated population quantiles. They furthermore encompass as special cases most of the previous statistical inference results for the measurement of inequality and social welfare.

There are many ways of measuring inequality, all of which have some intuitive or mathematical appeal (Cowel, 1985 [25]). However, many apparently sensible measures behave in perverse fashions. Numerous indices exist for measuring the degree of inequality in the distribution of income and wealth. They range from simple measures like the share of aggregate earnings received by each quintile to more complex measures such as the Gini, Theil (1967 [127]), Atkinson and generalized entropy indices (see Atkinson, 1970 [6]). All have different mathematical constructions, which can lead to different assessments concerning the degree of inequality. In our study, the main measure of inequality used as a proxy to show the

distribution of income is the Gini coefficient.
The Gini index is the most popular and important inequality measure. This index has a long history, dating back to Gini (1914 [**?**]), if not earlier. In particular, the Gini index has been widely used by economists and sociologists to measure economic inequality. Measures inspired by the index have been employed to assess the equality of opportunity and estimate income mobility. Naturally, numerous modifications and extensions of the classical Gini index have been proposed during the past 100 years, depending on one's needs and/or point of view.
The Gini index is based on the area between the egalitarian line and the Lorenz curve. This quantity is multiplied by 2, in order to have a range of values in the interval $[0,1]$. The Italian statistician, demographer and sociologist Corrado Gini (Gini,1914[**?**]).

Note that the Lorenz curve can be considered to be a cumulative distribution function on $[0,1]$ (Lorenz, 1905 [98]; Gastwirth, 1972[64]; Kovacevic and Binder, 1997 [92]; Cowell, 1977 [24]; and Langel and Tillé, 2013 [95]). We can exploit this fact and employ the moments of the Lorenz curve to develop new measures of inequality.

The Gini index has several possible interpretations and alternative ways in which it can be expressed. Perhaps, the most popular description of this measure is one related to the area between the population Lorenz curve and the egalitarian line.

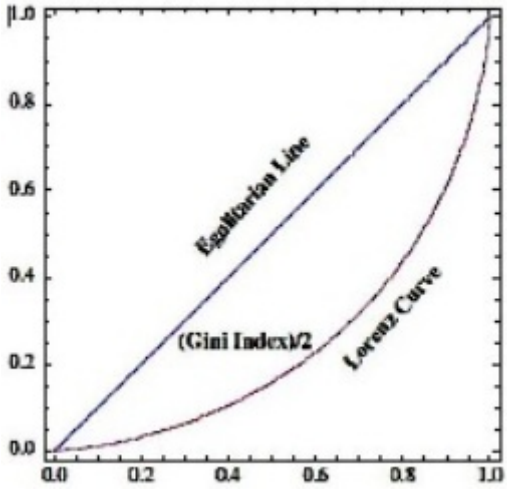

Figure 4.1: Egalitarian line $y = u$, Lorenz curve $y = L(u)$, and Gini index

Figure 4.1 represents the egalitarian line $y = u$, the Lorenz curve $y = L(u)$, and the Gini index for a hypothetical distribution. Consequently, if the Gini index is $G = 0$ we have perfect equality (all incomes identical), and $G = 1$ corresponds to perfect inequality.

The existing literature has intensively studied various estimators of G and the

associate inference theory, we cite here (Davidson, 2009 [35]; Qin et al., 2010 [114]; Yitzhaki, 1983 [134]; Kpanzou et al., 2013, 2017[93] , [94]).

More specifically, let $X \geq 0$ denote the income variable with distribution function $F(x)$, and the corresponding quantile function $\mathbb{Q}(t)$ for $0 < t < 1$, with Lorenz curve L_X, a formula for its Gini index, $G(X)$ or simply G if the random variable is known from the context, is

$$G = 2\int_0^1 [u - L_X(u)]\,du = 1 - 2\int_0^1 L_X(u)du, \tag{4.1}$$

where $u = F(x)$ is a cumulative distribution function (CDF) of a non-negative income with positive expectation $\mu = E(X)$, $L_X(p)$ is Lorenz function defined by

$$L_X(p) := \frac{1}{\mu}\int_0^p \mathbb{Q}(t)\,dt. \tag{4.2}$$

Using equation (4.1) and equation (4.2), it follows that we can also rewrite the Gini index as:

$$G = 1 - \frac{2}{\mu}\int_0^1 \int_0^p \mathbb{Q}(t)\,dt\,dp. \tag{4.3}$$

Inequality measures are often underestimated using sample data. It has been noted that the sample Lorenz curve often exhibits less inequality than does the population Lorenz curve. This fact suggests that the sample curve is a positively biased estimate of the population curve. If we have a sample $X_1, X_2, ..., X_n$ of size n from a distribution $F_X(x)$, recall that the corresponding sample Lorenz curve is defined to be a linear interpolation of the points $(0,0)$ and $(j/n, \sum_{i=1}^{j} X_i / \sum_{i=1}^{n} X_i)$, $j = 1, 2, ..., n$. As usual denote the sample Lorenz curve by $L_n(u)$.

Replacing the population quantile function $\mathbb{Q}$ by its empirical counterpart $\mathbb{Q}_n$, which is equal to the ith-order statistic $X_{i,n}$ for all $s \in ((i-1)/n, i/n]$, and for all $i = 1, \ldots, n$, where $X_{1,n} \leq X_{2,n} \leq .... \leq X_{n,n}$ are the order statistics based on the sample $X_1, X_2, ..., X_n$. Also the empirical estimator of the mean, μ_n, where $\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We arrive at the 'traditional' Gini estimator

$$\hat{G}_n = \frac{2}{n^2\mu_n}\sum_{i=1}^{n}\left(i - \frac{1}{2}\right)X_{i,n} - 1. \tag{4.4}$$

Of course, the empirical Gini index $\hat{G}_n$ can be rewritten in many other ways, such as the ratio of two L-statistics or the ratio of two U-statistics, which are perhaps more familiar to the reader, but formula (4.4) is best suited in the context of the present discussion.

The asymptotic theory for the empirical Gini index has been known at least since Hoeffding's paper (Hoeffding, 1948 [80]) on U-statistics. Indeed, the Gini index has been one of the most popular examples for illustrating the classes of L- and U-statistics. For this reason, Beach and Davidson (1983) [11] have developed an asymptotic theory for the traditional Gini estimator, assuming that the underlying i.i.d. random variables $X_1, X_2, ..., X_n$ have finite $(2+\epsilon)$ moments for some $\epsilon > 0$ as small as desired.

The latter moment assumption plays a crucial role. To illustrate the performance of G_n, we draw samples from the Pareto distribution

$$1 - F(x) = x^{-1/\gamma}, x > 1, \tag{4.5}$$

for some $\gamma > 0$, which is called the tail index. When $\gamma > 1$, then G is not defined. When $\gamma < 0.5$, then $E[X^{2+\epsilon}] < \infty$ for some $\epsilon > 0$, and so we can use the available estimator of G.

Therefore, from now on, we restrict ourselves to only those γ that are in the interval $(0.5, 1)$.
The present research has been motivated by the need for better understanding the distribution and inequality of capital incomes, which in many cases appear to be heavy-tailed. Since there are many individuals with no capital income, we restrict our attention to only those with positive capital incomes.
In mathematical terms, a heavy-tailed income distribution of a random variable X is regularly varying at infinity with index $(-1/\gamma) < 0$, where the parameter γ is referred to as the tail index of F. Its estimation is of fundamental importance to the applications of extreme value theory (see for example the monographs: see for example the monographs by Hill, 1975 [79]; Beirlant and Teugels, 1989 [12], Matthys and Beirlant, 2003 [101]; Beirlant et al. (2004 [13]); de Haan and Ferreira (2006 [45]), and the references therein). This class includes a number of popular income distributions such as Pareto, generalized Pareto, Burr, Fréchet, and Student t, etc., which are known to be appropriate models for fitting large incomes. In the remainder of this section, we restrict ourselves to this class of distributions. Moreover, we focus in our study on the case where $\gamma \in (1/2, 1)$ to ensure that the Gini index is finite, and in that case the results of Beach and Davidson (1983 [11]) cannot be applied, the second moment of X being infinite.

The present work is organized as follows, first 4.1.2, we construct an alternative estimator of the Gini index and we construct the bounds of confidences of this estimator, then 4.1.3 we illustrate the performance of the new estimator and the comparison with empirical estimator for some heavy-tailed models, the proof of the main results postponed until Section 4.1.4.

4.1.2 Main results

The idea behind the new estimator of G is to estimate the quantile function $\mathbb{Q}$ in the definition of the Gini index by the empirical quantile function for $s < 1 - k/n$, and by an extrapolated quantile function from the heavy-tail assumption for $s \geq 1 - k/n$. We next define an alternative estimator for the mean for a heavy-tailed distribution. Indeed, recall that, the mean μ can be rewritten as

$$\begin{aligned} \mu &= \int_0^1 \mathbb{Q}(s)ds = \int_0^{1-k/n} \mathbb{Q}(s)ds + \int_0^{k/n} \mathbb{Q}(1-s)ds \\ &= \mu_1 + \mu_2. \end{aligned}$$

We formulate the mean estimator for a heavy-tailed income distribution satisfying (**??**) as follows:

$$\hat{\mu}_{n,k} = \int_0^{1-k/n} \mathbb{Q}_n(s)ds + \left(\frac{k}{n}\right) \frac{X_{n-k,n}}{1-\widehat{\gamma}^H_{n,k}}, \tag{4.6}$$

where $\widehat{\gamma}^H_{n,k}$ is the Hill estimator of the tail index γ, (Hill, 1975[79]):

$$\widehat{\gamma}^H_{n,k} = \frac{1}{k}\sum_{i=1}^{k} i\,(\log X_{n-i+1,n} - \log X_{n-i,n}). \tag{4.7}$$

Note that to estimate μ_2 we use a Weissman-type estimator for $\mathbb{Q}$, (Weissman, 1978 [132]):

$$\widehat{\mathbb{Q}}(1-s) := X_{n-k,n}\,(k/n)^{\widehat{\gamma}^H_{n,k}}\, s^{-\widehat{\gamma}^H_{n,k}},\ s \to 0. \tag{4.8}$$

The estimation of Hill has been extensively studied in the literature for an intermediate sequence k, i.e. a sequence such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

Finaly, we obtain a semi-parametric estimator for the Gini index for heavy-tailed income distribution as follow:

$$\hat{G}_{n,k} = 1 - \frac{2}{\hat{\mu}_{n,k}} \int_0^1 \int_0^t \mathbb{Q}_n(s)dsdt. \tag{4.9}$$

Asymptotic normality for $\hat{G}_{n,k}$ is obviously related to that of $\widehat{\gamma}^H_{n,k}$. As usual in the extreme value framework, to prove such a type of result, we need a second-order condition on the tail quantile function $\mathbb{U}$, defined as

$$\mathbb{U}(z) = \inf\{y : F(y) \geq 1 - 1/z\}, z > 1. \tag{4.10}$$

We say that the function $\mathbb{U}$ satisfies the second-order regular variation condition

with second-order parameter $\rho \leq 0$ if there exists a function $A(t)$ which does not change its sign in a neighborhood of infinity, and is such that, for every $x > 0$,

$$\lim_{t\to\infty} \frac{\log \mathbb{U}(tx) - \log \mathbb{U}(t) - \gamma \log(x)}{A(t)} = \frac{x^\rho - 1}{\rho}, \tag{4.11}$$

when $\rho = 0$, the ratio on the right-hand side of equation (4.11) should be interpreted as $\log(x)$. As an example of heavy-tailed income distributions satisfying the second-order condition, we have the so-called and frequently used Hall's model (see Hall, 1982[76]; Hall and Welsh, 1985 [77]), which is a class of cdfs, such that

$$\mathbb{U}(t) = ct^\gamma (1 + dA(t)/\rho + o(t^\rho)) \text{ as } t \to \infty. \tag{4.12}$$

where $\gamma > 0$, $\rho \leq 0$, $c > 0$, and $d \in \mathbb{R}^*$. For statistical inference concerning the second-order parameter ρ we refer, for example, to de Haan and Stadtmüller (1996 [46]), Peng and Qi (2004 [108]), Gomes et al. (2005 [71]), and Gomes and Pestana (2007 [72]).

First, the family includes many of the most popular distributions used in the analysis of income, see for example Arnold and Sarabia (2018) [4], wealth and risk analysis, as special or limiting cases. This subclass of heavy-tailed distributions contains the Pareto, Burr, Fréchet, and t–Student. This family has several advantages for practical use.

Theorem 4.1 *Assume that the cdf F satisfies condition (4.11) with $\gamma \in (1/2, 1)$. Then for any sequence of integers $k = k_n \to \infty$ such that $k/n \to 0$ and $k^{1/2}A(n/k) \to 0$ when $n \to \infty$, on a suitable probability space, and with Brownian bridges appropriately constructed $\mathcal{B}_n(s)$, we have that*

$$\begin{aligned}\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} &= -\int_0^{1-k/n} \frac{v(s)\mathcal{B}_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) \\ &+ \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right) \\ &- \frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s} ds + o_p(1)\end{aligned} \tag{4.13}$$

when $n \to \infty$, where,

$$v(s) = \frac{2}{\mu^2}\int_0^s \int_0^t \mathbb{Q}(s) ds.$$

The proof of Theorem 4.1 is complex and relegated to Section 4.1.4. From the statistical inference point of view, the following corollary is our main result.

Corollary 4.1 *With the same hypothesis of Theorem 4.1, we have*

$$\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sigma(\gamma)\sqrt{k/n}\mathbb{Q}(1-k/n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \text{ as } n \longrightarrow \infty$$

where

$$\sigma^2(\gamma) = \frac{v^2(1)\gamma^4}{(1-\gamma)^4(2\gamma-1)}.$$

4.1.3 Simulation study

To discuss practical implementation of Theorem 4.1, we first fix a significance level $\alpha \in (0,1)$ and use the classical notation $z_{\alpha/2}$ for the $(1-\alpha/2)$-quantile of the standard normal distribution $\mathcal{N}(0,1)$. Given a realization of the random variables $X_1,\ldots,X_n$ (e.g., claim amounts), which follow a cdf F satisfying the conditions of Theorem 4.1, we construct a $(1-\alpha)$-level confidence interval for $\mathbb{G}$ as follows. First, we need to choose an appropriate number k of extreme values. Since Hill's estimator has in general a substantial variance for small k and a considerable bias for large k, we search for a k that balances between the two shortcomings, which is indeed a well-known hurdle when estimating the tail index.

To resolve this issue, several procedures have been suggested in the literature, and we refer to, e.g., Dekkers and de Haan (1993 [**?**]), Drees and Kaufmann (1998 [51]), Danielsson et al. (2001 [32]), Cheng and Peng (2001 [21]), Neves and Fraga Alves (2004 [106]), Gomes et al. (2009 [73]), and references therein.

In our current study we employ the method of Cheng-Peng for deciding on an appropriate value k^* of k. We note that, the optimal value of k that minimizes the absolute value of the leading coverage error term of Hill estimator, this fraction k depends on the sign of second-order regular variation, for more detail, see Cheng and Peng (2001 [21]). Having computed Hill's estimator and consequently determined $X_{n-k^*:n}$, we then compute the corresponding values of $\hat{G}_{n,k}$ and $\sigma^2(\widehat{\gamma}_n)$, and denote them by $\hat{G}_{n,k^*}$ and $\sigma^{2*}(\widehat{\gamma}_n)$, respectively. Finally, using Theorem 4.1 we arrive at the following $(1-\alpha)$-confidence interval for G:

$$\hat{G}_{n,k^*} \pm z_{\alpha/2}\frac{(k^*/n)^{1/2}\, X_{n-k^*:n}\sigma^*(\widehat{\gamma}_n)}{\sqrt{n}}. \tag{4.14}$$

To illustrate the performance of this confidence interval, we have carried out a small scale simulation study based on the Pareto cdf $F(x) = 1-x^{-1/\gamma}, x \geq 1$, and the Fréchet cdf $F(x) = \exp(-x^{-1/\gamma}), x \geq 0$ with the tail index γ set to 2/3 and 3/4, in which case we have fewer than two finite moments.

For the first part, we have generated 500 independent replicates from the selected

parent distribution of three samples of sizes n = 500, 1000, and 2000. For every simulated sample, we have obtained estimates $\hat{G}_{n,k}$. Then we have calculated the arithmetic averages over the values from the 500 repetitions, with the absolute error and root mean squared error ($RMSE$) of the new estimator $\hat{G}_{n,k}$ reported in Table 4.1 for the Pareto model, and Table 4.2 for the Frechet model.

In the tables, we have also reported 95%-confidence intervals with their lower (lcb) and upper bounds (ucb), coverage probabilities (covpr), and lengths.

Table 4.1: Simulation and confidence bounds of the estimator of the Gini index for Pareto distribution

$\gamma = 2/3$				$G = 0.500003$				
n	k^*	$\hat{G}_{n,k}$	error	rmse	lcb	ucb	covpr	length
500	26	0.47928	0.02074	0.01927	0.23462	0.76538	0.91782	0.53076
1000	70	0.48718	0.01288	0.01542	0.25881	0.74116	0.93526	0.48231
2000	103	0.50969	0.00969	0.01002	0.29499	0.70501	0.95236	0.41002
$\gamma = 3/4$				$G = 0.6000003$				
n	k^*	$\hat{G}_{n,k}$	error	rmse	lcb	ucb	covpr	length
500	27	0.58429	0.01572	0.01723	0.20304	0.99696	0.92671	0.79392
1000	51	0.58975	0.01025	0.01358	0.26944	0.92145	0.93056	0.66112
2000	102	0.59183	0.00817	0.000904	0.336381	0.83621	0.94748	0.47241

Table 4.2: Simulation and confidence bounds of the estimator of the Gini index for Frechet distribution

$\gamma = 2/3$				G =0.58693				
n	k^*	$\hat{G}_{n,k}$	error	rmse	lcb	ucb	covpr	length
500	26	0.5711	0.01481	0.11025	0.21351	0.97412	0.84811	0.76061
1000	52	0.59892	0.01198	0.07581	0.23069	0.96715	0.90124	0.73646
2000	103	0.58028	0.00665	0.03159	0.26062	0.89995	0.94201	0.63934
$\gamma = 3/4$				$G = 0.67979$				
n	k^*	$\hat{G}_{n,k}$	error	rmse	lcb	ucb	covpr	length
500	26	0.66812	0.01167	0.10231	0.35501	0.98124	0.86220	0.65623
1000	55	0.68541	0.00892	0.07814	0.37479	0.99603	0.89532	0.62124
2000	104	0.68101	0.00128	0.04215	0.37875	0.98327	0.91202	0.60452

The major observations from our simulations results presented in Table 4.1 and Table 4.2 are summarized as follows: (1) The error and $RMSE$ are decreasing when the sample size are increasing for all cases. (2) In terms of coverage probability, we find acceptable results, these results show that the coverage probability is increasing when the sample size is increasing. (3) In terms of average lengths of confidence intervals, our interval estimators decrease when the sample size is increasing.
The second part of our simulation study consists of a numerical comparison between the absolute bias and the mean squarre error (MSE) of $\hat{G}_n$ and $\hat{G}_{n,k}$. For two

models (Pareto and Frechet) with two values of tail index ($\gamma = 2/3$, and $\gamma = 3/4$). We vary the common size n of the sample, for each size, we generate 500 independent replicates. Our overall results are taken as the empirical means of the results obtained through the 500 repetitions. To determine the optimal number of upper order statistics (that we denote by k^*) used in the computation of $\hat{\gamma}^H_{n,k}$, we apply the algorithm of Cheng-Peng (2001). The simulation results are summarized in Table 4.3 for the model of Pareto and in Table 4.4 for the model of Frechet (where abs bias and MSE respectively stand for the absolute value of the bias and the mean squared error of the estimation).

Table 4.3: Results of comparison bias and mse between $\hat{G}_n$ and $\hat{G}_{n,k}$ for Pareto model

γ	2/3				3/4			
G	$\hat{G}_n$		$\hat{G}_{n,k}$		$\hat{G}$		$\hat{G}_{n,k}$	
n	bias	mse	bias	mse	bias	mse	bias	mse
500	0.2370	0.0642	0.0556	0.0263	0.2066	0.0509	0.0670	0.00141
1000	0.1898	0.0368	0.0356	0.0123	0.1668	0.0291	0.0049	0.00077
2000	0.1257	0.0225	0.0328	0.0016	0.1393	0.0199	0.0026	0.00043

Table 4.4: Results of comparison bias and rmse between $\hat{G}_n$ and $\hat{G}_{n,k}$ for Frechet model

γ	2/3				3/4			
G	$\hat{G}_n$		$\hat{G}_{n,k}$		$\hat{G}$		$\hat{G}_{n,k}$	
n	bias	mse	bias	mse	bias	mse	bias	mse
500	0.0387	0.0154	0.0197	0.00165	0.0455	0.0072	0.0111	0.00061
1000	0.0295	0.0148	0.0106	0.00138	0.0448	0.0028	0.0041	0.00014
2000	0.0168	0.0129	0.0102	0.00108	0.0331	0.0014	0.0027	0.000049

The results presented in Table 4.3 and Table 4.4 which represents the comparison between our proposed estimator $\hat{G}_{n,k}$ and the traditional estimator $\hat{G}_n$ in terms of bias and MSE, show the performance of our estimator, where the bias and MSE of our estimator are smaller in all cases in comparison with the bias and MSE of the traditional estimator, then, the values of the bias and MSE are decreasing when the size of the sample is increasing. In light of these results, we see that, from the point of view of the bias and the MSE, the estimation accuracy increases when the size of the sample is increased.

4.1.4 Proofs

Proof of Theorem 4.1. Denote $U_i = F(X_i)$ for $i = 1, 2, ..., n$. Then $U_1, U_2, ..., U_n$ is a sequence of i.i.d. random variables following the uniform distribution on $[0, 1]$. The following result shows that the empirical and quantile processes based on the

sequence $U_1, U_2, ..., U_n$ can be approximated by a series of Brownian bridges; see Csö650rgő et al. (1986) [30]. These Brownian bridges are the same as on the right-hand side of equation (4.13) in Theorem 4.1. Let $\alpha_n(s)$ be the uniform empirical process defined by

$$\alpha_n(s) = \sqrt{n}\,(H_n(s) - s), 0 \le s \le 1,$$

where $H_n(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \le s\}}$, and let $\beta_n(s)$ be the uniform quantile process defined by

$$\beta_n(s) = \sqrt{n}\left(H_n^{-1}(s) - s\right), 0 \le s \le 1.$$

Under a Skorokhod-type construction, there exists a sequence of Brownian bridges $\mathcal{B}_1, \mathcal{B}_2, ...$ such that, when $n \to \infty$, we have (cf. Csö650rgő et al., 1986 [30])

$$\sup_{0 \le s \le 1-1/n} n^{v_1} \frac{\left|\alpha_n(s) - \beta_n(s)\right|}{(1-s)^{1/2-v_1}} = O_P(1) \text{ for any } 0 \le v_1 \le \frac{1}{4}, \tag{4.15}$$

and

$$\sup_{0 \le s \le 1-1/n} n^{v_2} \frac{\left|\mathcal{B}_n(s) + \beta_n(s)\right|}{(1-s)^{1/2-v_2}} = O_P(1) \text{ for any } 0 \le v_1 \le \frac{1}{2}.$$

We start the proof of Theorem 4.1, by the calculation of the following difference

$$\begin{aligned}
\hat{G}_{n,k} - G &= \left(1 - \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t \mathbb{Q}_n(s)dsdt\right) - \left(1 - \frac{2}{\mu}\int_0^1\int_0^t \mathbb{Q}(s)dsdt\right) \qquad (4.16)\\
&= -\frac{2}{\hat{\mu}_n}\int_0^1\int_0^t \mathbb{Q}_n(s)dsdt + \frac{2}{\mu}\int_0^1\int_0^t \mathbb{Q}(s)dsdt\\
&= -\frac{2}{\hat{\mu}_n}\int_0^1\int_0^t \mathbb{Q}_n(s)dsdt + \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t \mathbb{Q}(s)dsdt\\
&\quad + \frac{2}{\mu}\int_0^1\int_0^t \mathbb{Q}(s)dsdt - \frac{2}{\hat{\mu}_n}\int_0^1\int_0^t \mathbb{Q}(s)dsdt.
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} &= -\frac{2}{\hat{\mu}_n}\left(\int_0^1 \sqrt{n}\frac{\int_0^t [\mathbb{Q}_n(s) - \mathbb{Q}(s)]\,dsdt}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\right) \qquad (4.17)\\
&\quad + \frac{2}{\mu\hat{\mu}_n}\int_0^1 \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\int_0^t \mathbb{Q}(s)dsdt \qquad (4.18)\\
&= I_1 + I_2. \qquad (4.19)
\end{aligned}$$

Since I_1 is an integral over $[0,1]$, we split it into the sum of two terms, I_{11} and I_{12}, which are the same integrals but over the intervals $[0, 1-k/n]$ and $[1-k/n, 1]$, respectively. A similar split is applied on I_2, which results in $I_2 = I_{21} + I_{22}$. We shall prove that $I_{12} = o_P(1)$ and $I_{22} = o_P(1)$ when $n \to \infty$. We shall next show in

several steps that $I_{11} = T_{n,1} + o_P(1)$ and $I_{21} = T_{n,2} + T_{n,3} + o_P(1)$ when $n \to \infty$. This will conclude the proof of Theorem 4.1. Hence, from now on we deal with the process A_n, which may be rewritten as follow

$$A_n(t) = \int_t^{1-k/n} [\mathbb{Q}_n(s) - \mathbb{Q}(s)]\, ds \tag{4.20}$$

which is an integral of the general quantile process $\mathbb{Q}_n - \mathbb{Q}$. To reduce it to an integral of the general empirical process $F_n - F$, we employ the (general) Vervaat process (see e.g., Zitikis, 1998 [137])

$$V_n(t) = \int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds + \int_{-\infty}^{\mathbb{Q}(t)} (F_n(x) - F(x))dx \tag{4.21}$$

The process $V_n(t)$ satisfies the boundary conditions $V_n(0) = 0$ and $V_n(1) = 0$, is non-negative for all $t \in [0,1]$, and such that

$$\sqrt{n} V_n(t) \leq |e_n(t)||\mathbb{Q}_n(t) - \mathbb{Q}(t)|. \tag{4.22}$$

Hence, upon recalling that $e_n(t) = \sqrt{n}(F_n(\mathbb{Q}(t)) - t)$, we conclude from (4.22) that the difference between the quantities

$$\sqrt{n} \int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds \tag{4.23}$$

and

$$-\sqrt{n} \int_{-\infty}^{\mathbb{Q}(t)} (F_n(x) - F(x))dx \tag{4.24}$$

tends to zero when $n \to \infty$ whenever $\mathbb{Q}_n(t)$ converges to $\mathbb{Q}(t)$, which holds because F is continuous and strictly increasing. This is the main idea of employing the Vervaat process in the present proof, as it allows us to replace quantity (4.23) by (4.24), which is much easier to tackle. We have the following equation

$$A_n(t) = -\int_{Q(t)}^{Q(1-k/n)} (F_n(x) - F(x))\, dx + V_n(1 - k/n) - V_n(t)$$

which we apply on the right-hand sides of equation 4.20 and 4.21. By changing the variable of integration, we get

$$A_n(t) = -\int_t^{1-k/n} \frac{e_n(s)}{\sqrt{n}} d\mathbb{Q}(s) + V_n(1 - k/n) - V_n(t) \tag{4.25}$$

and

$$\int_0^t (\mathbb{Q}_n(s) - \mathbb{Q}(s))ds = -\int_0^t \frac{e_n(s)}{\sqrt{n}} d\mathbb{Q}(s) + V_n(t). \tag{4.26}$$

Then

$$I_{11} = \frac{2}{\mu}\int_0^{1-k/n} \frac{\int_0^t e_n(s)d\mathbb{Q}(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt - \frac{2}{\mu}\int_0^{1-k/n} \frac{\sqrt{n}V_n(t)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt.$$

Taking into account that

$$\int_0^{1-k/n} \frac{\sqrt{n}V_n(t)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt = o_p(1),$$

when $n \to \infty$, we have

$$I_{11} = \frac{2}{\mu}\int_0^{1-k/n} \frac{\int_0^t e_n(s)d\mathbb{Q}(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}dt + o_p(1). \tag{4.27}$$

Here we replace e_n by $\mathcal{B}_n$ in the expressions for (4.27). Namely, when $n \to \infty$, by the use of the Fubini theorem, we obtain

$$\begin{aligned} I_{11} &= \int_0^{1-k/n} \frac{\mathcal{B}_n(s)v(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}d\mathbb{Q}(s) + o_p(1) \\ &\quad T_{n,1} + o_p(1). \end{aligned}$$

In a similar way, firstly writing I_{21} in terms of the empirical and Vervaat processes

$$I_{21} = \frac{2}{\mu\hat{\mu}_n}\int_0^{1-k/n} \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}\int_0^t \mathbb{Q}(s)dsdt$$

With the results of Peng (2001 [107]), Necir et al. (2010[104]), there exist a sequence of Brownien bridge $\{\mathcal{B}_n(s), 0 \le s \le 1\}_{n\ge 1}$ such that, for any n large enough, we have:h

$$\begin{aligned} \frac{\sqrt{n}(\hat{\mu}_n-\mu)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} \overset{d}{=} & -\int_0^{1-k/n} \frac{e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}d\mathbb{Q}(s) \\ & + \frac{\gamma^2}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} \\ & - \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds + o_P(1), \end{aligned}$$

Then

$$\begin{aligned} I_{21} \overset{d}{=} & \int_0^{1-k/n} \frac{v(s)e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)}d\mathbb{Q}(s) \\ & + \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} \\ & - \frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s}ds + o_P(1), \end{aligned}$$

we can easily show that

$$\int_0^{1-k/n} \frac{v(s)e_n(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s) = o_P(1).$$

Then

$$\begin{aligned} I_{21} \stackrel{d}{=} & \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\left\{\sqrt{\frac{n}{k}}\mathcal{B}_n\left(1-\frac{k}{n}\right)\right\} \\ & -\frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{n}{k}}\int_{1-k/n}^1 \frac{B_n(s)}{1-s} ds + o_P(1) \\ & = T_{n,2} + T_{n,3} \end{aligned}$$

Finally

$$\frac{\sqrt{n}(\hat{G}_{n,k} - G)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} = \sum_{i=1}^3 T_{n,i} + o_p(1)$$

where

$$T_{n,1} = -\int_0^{1-k/n} \frac{\mathcal{B}_n(s)v(s)}{\sqrt{k/n}\mathbb{Q}(1-k/n)} d\mathbb{Q}(s)$$

$$T_{n,2} = \frac{\gamma^2 v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\mathcal{B}_n(1-\frac{k}{n})$$

$$T_{n,3} = -\frac{\gamma v(1-k/n)}{(1-\gamma)^2}\sqrt{\frac{k}{n}}\int_{1-k/n}^1 \frac{\mathcal{B}_n(s)}{1-s} ds.$$

■

Proof of corollary 4.1. Without the remainder term $o_P(1)$, the right-hand side of equation (4.13) is a mean-zero normal random variable, whose variance converges to $\sigma^2(\gamma)$ when $n \to \infty$, as the following

$$\begin{aligned} E[T_{n,1}^2] &\to \frac{2\gamma v^2(1)}{2\gamma - 1}, \quad E[T_{n,2}^2] \to \frac{\gamma^4 v^2(1)}{(1-\gamma)^4} \\ E[T_{n,3}^2] &\to \frac{\gamma^2 v^2(1)}{(1-\gamma)^4}, \quad E[T_{n,1}T_{n,2}] \to \frac{\gamma^2 v^2(1)}{(1-\gamma)^2} \\ E[T_{n,1}T_{n,3}] &\to \frac{\gamma v^2(1)}{(1-\gamma)^2}, E[T_{n,2}T_{n,3}] \to \frac{\gamma^3 v^2(1)}{(1-\gamma)^4}. \end{aligned}$$

■

4.2 Estimating the Gini index for income loss distributions under random censo

Abstract:

The Gini index is one of the most widely used inequality indices, it has the distinction of being derived from the Lorenz curve, but generally it is estimated assuming that complete and unbiased samples are available. In this part, we make use of

the extreme value theory and the Kaplan-Meyer estimator to construct a new estimator of the Gini index when the data are censored, and we study the asymptotic normality property. We show the performance of our proposed estimator by some results of simulations.

KeyWords: Gini index, random censoring, loss distributions, Kaplan-Meier estimator.

4.2.1 Introduction

The Gini index, named in honor of the Italian statistician **Corrado Gini** (1884 – 1965), is one of the most commonly used statistical indices in the social sciences to measure the concentration in the distribution of a positive random variable, it is mainly used in economics as a measure of income or wealth inequality between individuals or households because of clear economic interpretation. Recently, the Gini coeffcient has been used to describe concentration in levels of mortality, or in length of life, among different socio-economic groups, and to evaluate inequality in health and in life expectancy (see, e.g., **Andreev and Begun** 2003).

Max Lorenz introduced the Lorenz curve corresponding to a non-negative random variable (rv) X with a distribution function (df) F, quantile function $Q(p)$, and finite mean $E(X) = \mu$ as:

$$L_F(t) = \frac{1}{\mu}\int_0^t Q(s)ds \quad 0 \leq t \leq 1 \tag{4.28}$$

In econometrics, with X representing income, $L(t)$ gives the fraction of total income that the holders of the lowest tth fraction of income possesses. Most of the measures of income inequality are derived from the Lorenz curve. An important example is the Gini index associated with F defined by:

$$G_F = \frac{\int_0^1 [u - L_F(u)]du}{\int_0^1 udu} = 1 - 2(CL)_F \tag{4.29}$$

where $(CL)_F = \int_0^1 L_F(u)du$ is the cumulative Lorenz curve corresponding to F. This is a ratio of the area between the diagonal and the Lorenz curve and the area of the whole triangle under the diagonal. The numerator is usually called the area of concentration. **Kendall** and **Stuart** (1963) showed that this is equivalent to a ratio of a measure of dispersion to the mean. In general, these notions are useful for measuring concentration and inequality in distributions of resources, and in size distributions. The Gini index has also been studied in its ability to detect requests in distributions (see, e.g., **Nygard** and **Sandröm** 1981; **Muliere** and **Scarsini** 1989).

It is now commonly recognised that income and wealth data may have been censored or trimmed for reasons of confidentiality or convenience (**Fichtenbaum** and **Shahidi** 1988), but several tests based on the sample Gini index have been proposed in literature for noncensored data. The goodness-of-fit test was proposed on the basis of the Gini index specified in spacing by **Rao** and **Goria** (2004), and they showed, by simulation, that such a test has a higher resistance than all the competitors considered for certain common alternatives.

Tse (2006) and **Bonetti** et **al.** (2009) considered Gini estimation under independent censoring. **Gigliarano** and **Muliere** (2013) required prior distribution information. Compared with the Gini estimation under complete data, these methods incorporated censoring into the Gini estimation. However, the independent censoring and prior distribution requirement may be limitations when applying these methods to real data.

As we are interested in applying the Gini index to censored data, the purpose of this part is to construct a new estimator of the Gini index for loss distribution when the data are censored, using the extreme value theory and the Kaplan-Meyer estimator.

The Gini index, which we denote by G_F, is usually interpreted as twice the area between the actual population Lorenz function (Lorenz 1905)

$$L_F(p) = \frac{1}{\mu_F} \int_0^p F^{-1}(t)dt \tag{4.30}$$

and the egalitarian Lorenz function $L_E(p) = p$, $0 \leq p \leq 1$, which is the hypotenuse of the right triangle into which every Lorenz curve falls.

Using the Lorenz curves, this Gini index is defined as the ratio of the area between the diagonal and the Lorenz curve and the whole area under the diagonal. The formula is

$$G = 1 - 2 \int_0^1 L(p)dp \tag{4.31}$$

This definition yields Gini coefficients satisfying the inequalities $0 < G < 1$, the higher the G value, the lower the Lorenz curve and the stronger the inequality. The reason for the popularity of the Gini coefficient is that it is easy to compute, being a ratio of two areas in Lorenz curve diagrams. The Gini coefficient allows direct comparison of the income of two income distributions, regardless of their sizes or patterns. This index doesn't capture where in the distribution the inequality occurs.

The Lorenz curve introduced by Max Otto Lorenz in 1905, it is a pivotal tool in the study of economic inequality and the distribution of wealth in the society.

Consider a non-negative random variable (rv) X with a distribution function (df) F, quantile function $Q(p)$, and finite mean $E(X) = \mu$. The Lorenz curve $L(x)$ is

formally given by:

$$L(x) = \frac{\int_0^x Q(\alpha)d\alpha}{\int_0^1 Q(\alpha)d\alpha} \ , \ 0 \leq x \ , \ \alpha \leq 1. \tag{4.32}$$

In terms of wealth, the Lorenz curve reads as follows: for a given $x \in [0,1]$, $L(x)$ tells us that $x \times 100\%$ of the population owns $L(x) \times 100\%$ of the total wealth. Such an interpretation tells that the Lorenz curve is scale-free: the total amount of wealth is not taken into consideration, whereas the way it is distributed among the individuals is the key information.

Given its strong relation with the quantile function Q, the Lorenz curve can recover the cumulative distribution of Y up to a constant. However, despite the Lorenz curve is theoretically a one-to one mapping with a given distribution, discriminate among distributions just looking at their Lorenz curves it is not an easy task to perform by hand.

Mathematically, the Lorenz curve $L : [0,1] \rightarrow [0,1]$ defined in Equation (2;3) is a continuous, non-decreasing, convex function, almost everywhere differentiable in $[0,1]$,such that $L(0) = 0$ and $L(1) = 1$. The curve $L(x)$ is bounded from above by the so-called perfect equality curve, i.e. $L_{pe}(x) = x$, and from below by the perfect inequality curve, i.e.

$$L_{pi}(x) = \begin{cases} 0 \ \text{ si } 0 \leq x < 1 \\ 1 \ \text{ si } \ x = 1 \end{cases} \tag{4.33}$$

The perfect equality line L_{pe} indicates the theoretical situation in which everyone possesses the same amount of wealth in the economy, while the perfect inequality line L_{pi}, reachable only as limiting case for continuous random variables, states that only one individual owns all the wealth in the society.

The very first mathematical definition of the Lorenz curve goes to **Kendall** and **Stuart**, who expressed it as two equations assuming an absolutely continuous distribution of income. Two years later, **Gastwirth** provided a general definition of the Lorenz curve, applying to both continuous and discrete laws, in the form of a single formula (4.29):

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt \ \ , 0 \leq p \leq 1$$

where the income distribution and its inverse function are denoted by F and F^{-1}, respectively, and μ denotes the expectation. Due to its numerous applications in various fields such as economics (**Gastwirth)** and the Gini index, **Hart**, **Gail** and **Gastwirth** and tests, fishing (**Thompson,**1976) or even bibliometrics (**Burell**, 1991), the Lorenz curve has given rise to numerous works in nonparametric estimation.

Gastwirth proposed a natural estimator of the Lorenz curve, defined by:

$$L_n(p) = \frac{1}{\overline{X}_n} \int_0^p F_n^{-1}(u)du \quad , 0 \leq p \leq 1$$

where $\overline{X}_n$ represents the empirical mean of a sample of n independent observations $X_1, ..., X_n$ and with the same distribution F, F_n the empirical distribution function of this sample, that is to say

$$F_n(x) = \frac{1}{n} \sum 1_{[X_i \leq x]} \quad , x > 0$$

and

$$F_n^{-1}(t) = \inf_{x>0} \{x : F_n(x) \geq t\} \quad , 0 \leq t \leq 1$$

the inverse of the empirical distribution function. From this estimator, he was able to construct an estimator of the Gini index.

The most well-known member of the income inequality family is the Gini coefficient. The Gini mean difference and its normalized version, known as the Gini index, have aided decision makers since their introduction by Corrado Gini more than a hundred years ago.

The Gini index, which we denote by G_F, is usually interpreted as twice the area between the actual population Lorenz function, it is define by:

$$G_F = 1 - \frac{2}{\mu_F} \int_0^1 \left(\frac{1}{p} \int_0^p F(t)dt \right) dp$$

Several other equivalent ways to define the Gini index exist. An alternative expression is given by:

$$G = \frac{\eta}{\mu} - 1 \ = \frac{\int_0^\infty 2F(t)tdF(t)}{\mu} - 1$$

(see **David** 1968).

Most research on the Gini coefficient, as well as the majority of applications, have focused on complete data. However, one often has to deal with censored data in applications.With respect to lifetime data, the data are often right censored.

The plan of this parts is organized as follows: In Section 2, and using the Kaplan-Meier estimator we propose an estimator of Gini index in the case of the presence of censored data, we give the properties of the new estimator, in section 3, we state our main results, some simulations are given in section 4. For the sake of completeness, some appendices contain the more technical details of this work.

4.3 Estimation from censored data

In the uncensored case, a wide variety of nonparametric estimators of Gini index have been proposed in the literature, in the case where the response variable is censored, some estimators which generalize the Kaplan-Meier estimator of the survival function (Kaplan and Meier 1958) have been proposed to estimate this index.

Csörgő and **Horváth** (1987)[31] first constructed an estimator for right-censored data and proved its convergence in law, as well as its almost safe uniform convergence. **Tse** (2006) then considered left-truncated and right-censored observations and then demonstrated the law of the iterated logarithm of the Lorenz process and the resulting convergence in law of the Gini index estimator.

In this section, we turn to the problem of estimating G from sample censored data. Let $X_1, ..., X_n$ be $n \geq 1$ independent copies of a non-negative random variable (rv) X, defined over some probability space (Ω, A, P), with continuous cumulative distribution fonction (cdf) F. An independent sequence of independent rv's $Y_1, ..., Y_n$ with continuous cdf G censor them to the right, so that at each stage j we only can observe $Z_j = \min(X_j, Y_j)$ and the variable $\delta_j = 1_{\{X_j \leq Y_j\}}$(with $1_{\{.\}}$ denoting the indicator function) informing whether or not there has been censorship.

We will use the method of Maximum Likelihood (ML) to estimate the parameters of the selected loss distribution. This method can be applied in a very wide variety of situations and the estimated obtained using ML generally have very good properties compared to estimates obtained by other methods (e. g. method of moments, method of quantile).

Notice that, under right random censoring, the well-known empirical estimator of the distribution function F is the nonparametric maximum likelihood estimator, given by Kaplan-Meier (1958) defined by:

$$1 - \hat{F}_n(x) = \hat{S}_n(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i;n]}}{n - i + 1}\right)^{1_{\{Z_{i,n} \leq x\}}}$$

where $Z_{1,n} \leq Z_{2,n} \leq ... \leq Z_{n,n}$ are the ordered Z-values, where ties within lifetimes or within censoring times are ordered arbitrarily and ties among lifetimes and censoring times are treated as if the former precedes the latter. $\delta_{[i;n]}$ is the concomitant of the i th order statistic, that is, $\delta_{[i;n]} = \delta_j$ if $Z_{i,n} = Z_j$.

Suppose indeed that we are interested in the lifetimes of n individuals or items, which are subject to K different and exclusive causes of death or failure, and to random censorship (from the right) as well.

The Kaplan-Meier estimator converges almost surely and uniformly to S (**Földes** et al. 1980). Under certain conditions of regularity, it converges in law towards a

Gaussian process (see **Breslow** and **Crowley** 1974).The mathematical properties of the Kaplan-Meier estimator can also be found in Chapter 7 of **Shorack** and **Wellner** (1986).

Let us assume that both F and G are heavy-tailed, that is there exist two constants $\gamma_1 > 0$ and $\gamma_2 > 0$, called tail index or extreme value index (EVI's), such that

$$\overline{F}(z) = z^{-1/\gamma_1}\ell_1(z) \text{ and } \overline{G}(z) = z^{-1/\gamma_2}\ell_2(z), \text{ as } z \to \infty, \tag{4.34}$$

where ℓ_1 and ℓ_2 are slowly varying functions at infinity, i.e. $\lim_{z\to\infty} \frac{\ell_i(tz)}{\ell_i(z)} = 1$ for every $t > 0, i = 1, 2$. If relations (4.34) hold, then we have, for any $x > 0$

$$\lim_{z\to\infty} \frac{\overline{F}(xz)}{\overline{F}(z)} = x^{-1/\gamma_1} \text{ and } \lim_{z\to\infty} \frac{\overline{G}(xz)}{\overline{G}(z)} = x^{-1/\gamma_2}$$

and we say that F and G are regularly varying at infinity as well, with respective tail indices $-1/\gamma_1$ and $-1/\gamma_2$ which we denote by $\overline{F} \in \mathcal{RV}_{-1/\gamma_1}$ and $\overline{G} \in \mathcal{RV}_{-1/\gamma 2}$. Note that, in virtue of the independence of X and Y, the cdf of the observed Z's, that we denote by H; is also heavy-tailed and we have $H \in \mathcal{RV}_{-1/\gamma}$ where $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

This class of distributions, which includes models such as Pareto, Burr, Fréchet, Lévy-stable and log-gamma, plays a prominent role in extreme value theory.

Let $\{(Z_i, \delta_i), 1 \le i \le n\}$ be a sample from the couple of rv's (Z, δ) and $Z_{1:n}, ..., Z_{n:n}$ the order statistics pertaining to $Z_1, ..., Z_n$. If we denote the concomitant of the ith order statistic by $\delta_{[i:n]}$ (i.e. $\delta_{[i:n]} = j$ if $Z_{i:n} = Z_j$), then Hill's estimator of γ_1 adapted to censored data is defined as $\widehat{\gamma}_1^{(H,c)} = \widehat{\gamma}^H/\widehat{p}$, where

$$\widehat{\gamma}^H = \frac{1}{k}\sum_{i=1}^{k} \log(Z_{n-i+1:n}/Z_{n-k:n}) \tag{4.35}$$

represents Hill's estimator (**Hill**, 1975) of γ with $k = k_n$ being an integer sequence satisfying

$$1 < k < n, k \to \infty \text{ and } k/n \to 0 \text{ as } n \to \infty, \tag{4.36}$$

and $\widehat{p} = \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]}$ being the proportion of upper non-censored observations. **Einmahl et al**. (2008) established the asymptotic normality of $\widehat{\gamma}_1^{(H,c)}$ by assuming that cdf's are absolutely continuous.

This leads us to derive a Weissman-type estimator (see Weissman, 1978 [132]) for the distribution tail F for censored data as follows:

$$\widehat{\overline{F}}(x) = \left(\frac{x}{Z_{n-k:n}}\right)^{-1/\widehat{\gamma}_1^{(H,c)}} \overline{F}_n(Z_{n-k:n})$$

In the context of randomly right censored observations, the nonparametric maximum likelihood estimator of F is given by Kaplan and Meier (1958) as the product limit estimator

$$\overline{F}_n(x) = \prod_{Z_{i:n}\leq x}\left(1-\frac{\delta_{[i:n]}}{n-i+1}\right) = \prod_{Z_{i:n}\leq x}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}}, \text{ for } x < Z_{i:n},$$

which gives

$$\overline{F}_n(Z_{n-k:n}) = \prod_{i=1}^{n-k}\left(1-\frac{\delta_{[i:n]}}{n-i+1}\right).$$

Thus, the distribution tail estimator is of the form

$$\widehat{\overline{F}}(x) = \left(\frac{x}{Z_{n-k:n}}\right)^{-1/\widehat{\gamma}_1^{(H,c)}} \prod_{i=1}^{n-k}\left(1-\frac{\delta_{[i:n]}}{n-i+1}\right)$$

and consequently, we define the Gini estimator as follows:

$$\hat{G}_n^c = 1-\frac{2}{\hat{\mu}_n^c}\left(\frac{1}{Z_{[nt],n}}\sum_{j=2}^{[nt]}\frac{\delta_{[j:n]}}{n-j+1}\prod_{i=1}^{j-1}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} Z_{i;n}\right),$$

$$\hat{\mu}_n^c = \hat{\mu}_{1,n}^c + \hat{\mu}_{2,n}^c$$

where:

$$\hat{\mu}_{1,n}^c = \prod\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} Z_{n-k,n} + \sum_{j=2}^{n-k}\frac{\delta_{[j:n]}}{n-j+1}\prod_{i=1}^{j-1}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}} Z_{i;n}$$

and

$$\hat{\mu}_{2,n}^c = \frac{\gamma_1^{(H,c)}}{1-\gamma_1^{(H,c)}}\left(\prod_{i=1}^{n-k}\left(\frac{n-i}{n-i+1}\right)^{\delta_{[i:n]}}\right) Z_{n-k;n}$$

4.3.1 Main results

We assume that, the cdf F and G satisfies the generalized second-order regular variation condition with second-order parameter ρ_1 and ρ_2 are negatives, if there exists two functions $a_1(t)$ and $a_2(t)$ which does not changes its sign in a neighbourhood of

infinity and such that, for every $x > 0$,

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{a_1(t)}\left(\frac{1-F(tx)}{1-F(t)}-x^{-1/\gamma_1}\right) &= x^{-1/\gamma_1}\frac{x^{\rho_1/\gamma_1}-1}{\rho_1/\gamma_1}\\ \lim_{t\to\infty}\frac{1}{a_2(t)}\left(\frac{1-G(tx)}{1-G(t)}-x^{-1/\gamma_2}\right) &= x^{-1/\gamma_2}\frac{x^{\rho_2/\gamma_2}-1}{\rho_2/\gamma_2}\end{aligned}$$

when $\delta_1 = 0$ and $\delta_2 = 0$ then the ratio on the right-hand side of equation (ref6.1) should be interpreted as $\log x$.

Assume that the second order condition (ref6.1) hold with $\gamma_2/(1+2\gamma_2) < \gamma_1 < 1$. Then for any sequence of integers $k = k_n \to \infty$ such that $k/n \to 0$ and $k^{1/2}a_1(n/k) \to 0$ when $n \to \infty$, we have that:

$$\frac{\sqrt{n}(\hat{G}_n^c - G)}{(k/n)^{1/2}Z_{n-k:n}\overline{F_n}(Z_{n-k:n})} \to_d \mathcal{N}(0,\sigma_\gamma^2) \quad \text{as } n\to\infty$$

where σ_γ^2 is the asymptotic variance.

4.3.2 Simulation study

Now, we carry out a simulation study (by means of the statistical software **R**) to illustrate the performance of our estimator, through three sets of censored and censoring data, all drawn, in the first part, from:

- Pareto model

$$F(x) = 1 - x^{-\gamma_1}, G(x) = 1 - x^{-\gamma_2} \qquad , x \geq 1,$$

- Fréchet model

$$F(x) = exp\{-x^{-\gamma_1}\}, G(x) = exp\{-x^{-\gamma_2}\} \qquad , x \geq 0,$$

- Burr model

$$F(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_1}, G(x) = 1 - \left(1 + x^{1/\eta}\right)^{-\eta/\gamma_2}, x \geq 0,$$

where η, γ_1, $\gamma_2 > 0$. We fix $\eta = 0.3$ and choose the values 0.3, 0.4 and 0.5 for γ_1.

For the proportion of the really observed extreme values, we take $p = 0.15$, 0.30, 0.40 and 0.50. 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 $(X_1,...,X_n)$ and $(Y_1,...,Y_n)$, then for each size, we generate 1000 independent replicates. Our overall results are taken as the empirical means of the results obtained through the 1000 repetitions.

To determine the optimal number (that we denote by k^*) of upper order statistics used in the computation of $\hat{\gamma}_1^{(H,c)}$, we apply the algorithm of Cheng and Peng (2003).

Pareto model:

Table 4.5: Results of simulation for Pareto model with index $\gamma_1 = 0.3$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.0215	0.0453	0.0570	0.0887	0.1170
500	0.0234	0.0445	0.0648	0.0784	0.0987
1000	0.0232	0.0477	0.0637	0.0762	0.1000
2000	0.0231	0.0464	0.0645	0.0790	0.1028
5000	0.0233	0.0463	0.0634	0.0798	0.1017

Table 4.6: Results of simulation for Pareto model with index $\gamma_1 = 0.4$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.0308	0.0621	0.0882	0.1121	0.1304
500	0.0309	0.0681	0.0772	0.1138	0.1444
1000	0.0295	0.0622	0.0862	0.1223	0.1359
2000	0.0324	0.0650	0.0892	0.1116	0.1365
5000	0.0313	0.0639	0.0862	0.1094	0.1330

Table 4.7: Results of simulation for Pareto model with index $\gamma_1 = 0.5$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.0351	0.0971	0.1094	0.1349	0.1982
500	0.0384	0.0770	0.1114	0.1221	0.1870
1000	0.0401	0.0806	0.1089	0.1359	0.1841
2000	0.0389	0.0823	0.1127	0.1407	0.1561
5000	0.0381	0.0807	0.1128	0.1410	0.1781

Fréchet Model:

Table 4.8: Results of simulation for Fréchet model with index $\gamma_1 = 0.3$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.0721	0.0959	0.1243	0.1302	0.1572
500	0.0649	0.0927	0.1221	0.1436	0.1778
1000	0.0660	0.0972	0.1125	0.1425	0.1716
2000	0.0701	0.0952	0.1168	0.1456	0.1760
5000	0.0672	0.0953	0.1157	0.1418	0.1737

Table 4.9: Results of simulation for Fréchet model with index $\gamma_1 = 0.4$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.0865	0.1284	0.1618	0.1736	0.2288
500	0.0927	0.1222	0.1597	0.1908	0.2156
1000	0.0886	0.1327	0.1577	0.1903	0.2431
2000	0.0905	0.1289	0.1629	0.1904	0.2411
5000	0.0905	0.1272	0.1613	0.1934	0.2333

Table 4.10: Results of simulation for Fréchet model with index $\gamma_1 = 0.5$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.1105	0.1517	0.2261	0.2578	0.3029
500	0.1097	0.1615	0.1899	0.2294	0.3109
1000	0.1107	0.1579	0.1864	0.2296	0.3118
2000	0.1112	0.1528	0.1962	0.2396	0.2994
5000	0.1092	0.1569	0.1928	0.2380	0.3022

Burr model:

Table 4.11: Results of simulation for Burr model with index $\gamma_1 = 0.3$ and $\eta = 1$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.1423	0.1824	0.2186	0.2463	0.2664
500	0.1337	0.1724	0.1936	0.2205	0.2909
1000	0.1382	0.1653	0.1960	0.2377	0.2878
2000	0.1398	0.1649	0.1998	0.2319	0.2841
5000	0.1404	0.1671	0.1939	0.2355	0.2934

Table 4.12: Results of simulation for Burr model with index $\gamma_1 = 0.4$ and $\eta = 1$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.1504	0.2142	0.2695	0.3275	0.3649
500	0.1698	0.2232	0.2555	0.2985	0.3483
1000	0.1739	0.2210	0.2586	0.3112	0.3530
2000	0.1784	0.2223	0.2572	0.3013	0.3702
5000	0.1762	0.2201	0.2528	0.3165	0.3685

Table 4.13: Results of simulation for Burr model with index $\gamma_1 = 0.5$ and $\eta = 1$

$n\|p$	0.15	0.30	0.40	0.5	0.6
250	0.2110	0.2896	0.3092	0.3954	0.4154
500	0.2100	0.2548	0.3137	0.3739	0.4746
1000	0.2204	0.2605	0.3184	0.3819	0.4478
2000	0.2070	0.2669	0.3126	0.3637	0.4481
5000	0.2119	0.2637	0.3148	0.3797	0.4436

Conclusion

in this thesis, we looked at a recent problem in extreme value theory, namely the presence of random censorship. This problem is very common in several areas of socio-economic life where data is often randomly censored on the right, such as medicine, finance, insurance, reliability,...

The objective of this thesis was twofold: initially, the goal was to clearly broaden the concepts of extremes values theories and its applications. Secondly, the goal was to estimate the income risk measures when the data are censored.

To facilitate the reading of the work, we recalled in Chapter 1 some fundamental concepts on the statistics of extreme values with its rich literature, that's constituted the starting point of the thesis.

In the same vein, we also recalled in Chapter 2 some fundamental notions on the statistics of censored data to better understand the field. heavy-tailed data based on the same number of extreme observations from both truncated and truncation variables.

Chapter 3 is dedicated to the study of risk measures and income inequality measures.

Finally, in Chapter 4, we are interested in the index estimation of risk measures for extreme values for incompletely observed data, with a particular interest in the case of right-censored data. We start by exploiting the first work on this subject, which is due to Gardes and Stupfler (2015), to obtain a simple tail index estimator based on a single sample fraction of extreme values.

Another objective of this thesis, is to combine the two problems of extreme values and censored data which poses a problem of double complexity, the first is that the data is too scarce giving statistics on small sizes, and the second is to reduce the sizes study of statistics due to censorship. This problem remains open in practice, and it is very interesting in this work to apply this knowledge to risk measurements.

Also, in the long term, it would be interesting to extend this theory of extreme values in the presence of data randomly right-censored in all domains (Weibull, Gumbel) and in other types of censorship (left-censored, interval).

This thesis offers interesting perspectives from a theoretical as well as a practical

point of view. In fact, in addition to the new lines of research it opens up, this work can contribute in many real situations to solving certain statistical problems such as insurance for example.

Bibliography

[1] Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. Annals of Statistics 6, 701–726.

[2] Anand, S., Kanbur, S. R. (1993). Inequality and development A critique. Journal of Development economics, 41(1), 19-43.

[3] Andersen, P., Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. Management science, 39(10), 1261-1264.

[4] Arnold,B. C and Sarabia, J. M. (2018), Majorization and the Lorenz Order with Applications in Applied Mathematics and Economics, Statistics for Social and Behavioral Sciences, Springer.

[5] Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999). Coherent measures of risk. Mathematical Finance 9(3), 203-228.

[6] Atkinson, Anthony. B ., "On the Measurement of Inequality," Journal of Economic Theory 2 (1970), 244-263.

[7] Atkinson, A. B. and A. Brandolini (2001). Promise and pitfalls in the use of secondary data-sets: Income inequality in oecd countries as a case study. Journal of Economic Literature 39, 771- 799.

[8] Baldwin, R., Forslid, R., Martin, P., Ottaviano, G., Robert-Nicoud, F. (2011). Economic geography and public policy. Princeton University Press.

[9] Balkema, A. A., De Haan, L. (1974). Limit laws for order statistics (No. 2099-2018-3082).

[10] Bari, A., Rassoul, A., Rouis, H. O. (2021). Estimating the Gini index for heavy-tailed income distributions. South African Statistical Journal, 55(1), 15-28.

[11] Beach, C.M. and R. Davidson (1983). "Distribution-Free Statistical Inference with Lorenz Curves and Income Shares." Review of Economic Studies, 50(3), 723–735.

[12] Beirlant, J., Teugels, J. (1989). Asymptotic normality of Hill's estimator. In J. Hüsler, R.-D. Reiss (Eds.) *Extreme Value Theory* (pp. 148–155). New York: Springer.

[13] Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J. (2004). *Statistics of Extremes: Theory and applications*. Chichester: Wiley.

[14] Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A., 2007. Estimation of the extreme value index and extreme quantiles under random censoring.Extremes 10, 151-174.

[15] Beran, R. Nonparametric regression with randomly censored data. Technical report. Univ. California, Berkeley, 1981.

[16] Billingsley, P. (1995). Probability and Measure, 3rd edition. Wiley, New York.

[17] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). Regular Variation. Cambridge University Press, Cambridge.

[18] Bonetti, M., Gigliarano, C., Muliere, P. (2009), The Gini concentration test for survival data. Lifetime Data Anal 15:493–518.

[19] Breiman, L., Stone, C. J. et Kooperberg, C. (1990). Robust confidence bounds for extreme upper quantiles. Journal of Statistical Computation and Simulation, 37(3- 4):127–149.

[20] Breslow, N., Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. The Annals of statistics, 437-453.

[21] Cheng, S. and Peng, L. (2001). Confidence intervals for the tail index. *Bernoulli*, 7, 751–760.

[22] Coles, S.G. (2001). An Introduction to Statistical Modelling of Extreme Values, Springer Series in Statistics.

[23] Colosimo E, Ferreira F, Oliveira M, Sousa C (2002) Empirical Comparisons Between Kaplan-Meier and Nelson-Aalen Survival Function Estimators. Journal of Statistical Computation and Simulation, 72 : 299-308.

[24] Cowell, F. A. (1977). Measuring inequality. Oxford: Phillip Allan.

[25] Cowell, F. A. (1985). Measures of distributional change: An axiomatic approach. Review of Economic Studies 52, 135–151.

[26] Cowell, F. A. and Flachaire, E. (2007). "Income Distribution and Inequality Measurement: The Problem of Extreme Values," Journal of Econometrics, vol. 141, pp. 1044-1072.

[27] Cowell, F. A. and Maria, E. (2007). "Statistical Inference for Lorenz Curves with Censored Data"

[28] Cowell, Frank, 2011, Measuring Inequality, Oxford: Oxford University Press, 3rd edition.

[29] Csörgő, S., Mason, D. Central limit theorems for sums of extreme values. Mathematical Proceedings of the Cambridge Philosophical Society, 98:547–558, 1985.

[30] Csörgő, M., Csörgő, S., Horváth, L., Mason, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.*, 14, 31–85.

[31] Csörgő,M., Csörgő,S., Horváth, L. (1987), Estimation of total time on test transforms and Lorenz curves under random censorship, Statistics, 18, 77-97.

[32] Danielsson, J., de Haan, L. Peng, L., de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.*, 76, 226–248.

[33] Daouia, A., Gardes, L., Girard, S. et Lekina, A. (2011). Kernel estimators of extreme level curves. Test, 20(2):311- 333.

[34] David, H. A. (1968). Miscellanea: Gini's mean difference rediscovered. Biometrika, 55(3), 573-575.

[35] Davidson, R. (2009). Reliable inference for the Gini index. Journal of Econometrics 150: 30-40.

[36] Davis, R., Resnick, S. (1984). Tail estimates motivated by extreme value theory. Annals of Statistics, 12:1467–1487.

[37] Davison, A. C. et Smith, R. L. (1990). Models for exceedances over high thresholds. Journal of the Royal Statistical Society, B, 52(3):393–442.

[38] Dekkers A. L. M. and de Haan L., On the estimation of the extreme value index and large quantile estimation, Ann. Statist. 17, (1989), 1795-1832.

[39] Dekkers A. L. M., Einmahl H. J. and de Haan L., A moment estimator for the index of an extreme value distribution, Ann. Statist. 17" (1989), 1833-55.

[40] Dekkers, A. L. M. and de Haan, L. (1993). Optimal choice of sample fraction in extreme-value estimation. Journal of Multivariate Analysis, 47, 173–195.

[41] Denuit, M. and A. Charpentier (2005). Mathématiques de l'assurance non-vie : Tarification et provisionnement . Economica, Paris.

[42] Denuit, M., Dhaene, J., Goovaerts, M. and Kaas, R. (2005). Actuarial theory for dependent risks measures, orders and models. Wiley.

[43] Doerrenberg, P., Peichl, A. (2014). The impact of redistributive policies on inequality in OECD countries. Applied Economics, 46(17), 2066-2086.

[44] de Haan, L. (1976). Sample extremes : an elementary introduction. Statistica Neerlandica, 30(4):161–172.

[45] de Haan, L., Ferreira, A. Extreme value theory: An introduction. New York, Springer, 2006.

[46] de Haan, L., Stadtmüller, U. (1996). Generalized regular variation of second order. *J. Austral. Math. Soc. Ser. A*, 61, 381–395.

[47] de Haan, L., Resnick, S.I. (1998), On asymptotic normality of the hill estimator. Stochastic Models, 4:849–867.

[48] Dombry, C. (2013). Maximum likelihood estimators for the extreme value index based on the block maxima method, arXiv:1301.5611.

[49] Dorfman, R. (1979). A formula for the Gini coefficient. The review of economics and statistics, 146-149.

[50] Drees, H. (1995). Refined pickands estimators of the extreme value index. The Annals of Statistics, 23(6):2059–2080.

[51] Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Processes and their Applications, 75(2):149–172.

[52] Einmahl, J. H. J., Fils-Villetard, A., Guillou, A. Statistics of extremes under random censoring. Bernoulli, 14:207–227, 2008. (Cité en pages xx, 36, 42, 43, 44, 45, 60, 61, 69 et 82.)

[53] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance, Springer-Verlag, Berlin.

[54] Fakoor, V., Ghalibaf, M.B., Azarnoosh, H. A., Asymptotic behaviors of the Lorenz curve and Gini index in sampling from a length-biased distribution, Statistics & probability letters 81 (9), 1425-1435 (2011).

[55] Fisher, R. A., and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Math. Proc. Cambridge Philos. Soc., 24(02), 180- 190.

[56] Fleming, T. R., and Harrington, D. P. (1984). Nonparametric estimation of the survival distribution in censored data. Comm. Statist. Theory Methods, 13(20), 2469- 2486.

[57] Földes, A., Rejtő, L., Winter, B. B. (1980). Strong consistency properties of nonparametric estimators for randomly censored data, I: The product-limit estimator. Periodica Mathematica Hungarica, 11(3), 233-250.

[58] Föllmer, H. and Schied, A. (2002a). Convex measures of risk and trading constraints. Finance and Stochastics, 6(4):429- 447.

[59] Föllmer, H. and Schied, A. (2002b). Robust preferences and convex measures of risk. In Advances in finance and stochastics, pages 39- 56. Springer.

[60] Forcina, A., & Giorgi, G. M. (2005). Early Gini's contributions to inequality measurement and statistical inference. Electronic journal for history of probability and statistics, 1(1), 1-15.

[61] Furman Edward, Ruodu Wang, and Ricardas Zitikis. (2017). Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. Journal of Business and Finance 83: 70-84.

[62] Gardes, L., Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. Extremes, 13(2), 177-204.

[63] Gastwirth, J. L. (1971). A general definition of the Lorenz curve, Econometrica, 39, 1037-1039.

[64] Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index, Rev. Economics and Statistics., 54, 306-316.

[65] Gigliarano, C., Muliere, P. (2013). Estimating the Lorenz curve and Gini index with right censored data: a Polya tree approach. Metron, 71(2), 105-122.

[66] Gini, C. (1914) On the measurement of concentration and variability of characters (Translated in 2005 from the 1914 Italian original by Fulvio De Santis.), Metron 63, pp. 3–38.

[67] Gini, Corrado. (1921). Measurement of inequality of incomes. Economic Journal 31: 124-26.

[68] Gisbert, J. P., Panés, J. (2009). Loss of response and requirement of infliximab dose intensification in Crohn's disease: a review. Official journal of the American College of Gastroenterology— ACG, 104(3), 760-767.

[69] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Annals of Mathematics 44 (3), 423- 453.

[70] Guida, M., Longo, M. (1988). Estimation of probability tails based on generalized extreme value distributions. Reliability Engineering System Safety, 20(3), 219-242.

[71] Gomes, M. I., Figueiredo, F., Mendonça, S. (2005). Asymptotically best linear unbiased tail estimators under a second-order regular variation condition. *J. Statist. Plann. Inference*, 134, 409–433.

[72] Gomes and Pestana (2007). A simple second-order reduced bias' tail index estimator. *J. Stat. Comput. Simul.*, 77, 487–504.

[73] Gomes, M. I., Pestana, D., Caeiro, F. (2009) A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator. *Statist. Probab. Lett.*, 79, 295–303.

[74] Goovaerts, M.J., De Vijlder, F.E., Haezendonck, J., 1984. Insurance Premiums: Theory and Applications. North-Holland, Amsterdam.

[75] Gumbel, E.J. (1958). Statistics of Extremes. Columbia University Press, New York.

[76] Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. R. Statist. Soc.*, 44, 37–42.

[77] Hall, P. and Welsh, A. H. (1985). Adaptative estimates of parameters of regular variation. *Ann. Statist.*, 13, 331–341.

[78] Harter, S. (1978). Effectance motivation reconsidered. Toward a developmental model. Human development, 21(1), 34-64.

[79] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. Ann. Statist., 3(5), 1163 - 1174.

[80] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19: 293–325.

[81] Hosking, J. (1985). Algorithm as 215 : Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. Journal of the Royal Statistical Society. Series C (Applied Statistics), 34(3):301- 310.

[82] Hosking, J. R. M., Wallis, J. R. et Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted comments. Technometrics, 27:251–261.

[83] Hosking, J. R. M., Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29(3):339–1349.

[84] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. The Quarterly Journal of the RoyalMeteorological Society, 81(384):158- 171.

[85] Johan, S. (2005). Generalized pickands estimators for the extreme value index. Journal of Statistical Planning and Inference, 128(2):381–396.

[86] Jorion, P. (2007). Value at risk : the new benchmark for managing financial risk. McGraw-Hill New York.

[87] Kaas, R. Goovaerts, M. Dhaene, J. and Denuit, M. (2009).Modern Actuarial Risk Theory Using R. Springer.

[88] Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc., 53(282), 457- 481.

[89] Kendall, M. and A. Stuart (1977). The advanced theory of statistics 4th Ed. London, C. Griffin.

[90] Kleiber ,C. and Kotz,S., Statistical size distributions in economics and actuarial sciences.,Vol. 470 (JohnWiley & Sons, 2003).

[91] Klein J.P., Moeschberger, M.L. (1997). Survival analysis : techniques for censored and truncated data. Springer-Verlag, New York.

[92] Kovacevic, M. and Binder, D. (1997). Variance estimation for measures of income inequality and polarization. The estimation equations approach. Journal of Official Statistics 13: 41–58.

[93] Kpanzou, T.A., de Wet, T. and Neethling, A. (2013). Semiparametric Estimation of Inequality Measures. South African Statistical Journal 47: 33-48.

[94] Kpanzou T.A., DeWet T. and Lo G.S. (2017). Measuring inequality: application of semi-parametric methods to real life data. African Journal of Applied Statistics 4: 157-164.

[95] Langel, M. and Tillé,Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. Journal of the Royal Statistical Society series A 62: 521–540.

[96] Lawless, Jerald F. (2003). Statistical Models and Methods for Lifetime Data (2nd ed.). Hoboken: John Wiley and Sons.

[97] Lejeune, M. (2010) Statistique : La théorie et ses applications. Springer, 2nd edition.

[98] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth, Publications of the American statistical association 9, 209.

[99] Macleod, A. (1989). A remark on algorithm as 215 : Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. Applied Statistics, 38(1):198- 199.

[100] Mason, D.M. Laws of large numbers for sums of extreme values. Ann. Probab., 10:754–764, 1982.

[101] Matthys, G., Beirlant, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models. *Statist. Sinica*, 13, 853–880.

[102] Modarres,R.and J.L. Gastwirth (2006). "A cautionary note on estimating the standard error of the Gini index of inequality", Oxford Bulletin of Economics and Statistics, Vol.68, pp.3: 85- 90.

[103] Ndao, P. (2015) Modélisation de valeurs extrêmes modélisation de valeurs extrêmes conditionnelles en présence de censure. Doctoral thesis, University of Gaston Berger.

[104] Necir, A., Rassoul, A. and Zitikis, R.,(2010). Estimating the conditional tail expectation in the case of heavy-tailed losses. J. Probab. Statist., doi:10.1155/2010/596839.

[105] Nelson, W. (1972). Theory and application of hazard plotting for censored survival data. Biometrics, 14, 945–966.

[106] Neves, C., Fraga Alves, M. I. (2004). Reiss and Thomas' automatic selection of the number of extremes. *Comput. Statist. Data Anal.*, 47, 689–704.

[107] Peng, L. (2001). Estimating the mean of a heavy tailed distribution. *Statist. Probab. Lett.*, 52, 255–264.

[108] Peng, L., Qi, Y. (2004). Estimating the first and second-order parameters of a heavy-tailed distribution. *Aust. N. Z. J. Stat.*, 46, 305–312.

[109] Peng, L. (2011) Empirical likelihood methods for the Gini index. Aust Nz J Stat 53(2):131-139.

[110] Peto, R. (1973). Experimental survival curves for intervalcensored data. Journal of the Royal Statistical Society: series C (Applied Statistics), 22(1), 86-91.

[111] Pickands, J. (1975). Statistical inference using extreme order statistics. Ann. Stat. 119- 131.

[112] Pietra, G. (1915) Delle relazioni tra gli indici di variabilità. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, 74, 775–804. English translation in Metron (2014), 72, 5–16.

[113] Piketty, T. (2014). trans. Arthur Goldhammer. Capital in the Twenty-first Century.

[114] Qin, Y., Rao, J.N.K., Wu, C., (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. Econ Model 27, 1429–1435.

[115] Reiss, R.D. (1989). Approximate distributions of order statistics. Springer, New York.

[116] Reiss, R.D. and Thomas, M. (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel.

[117] Resnick, S.I. (2007). Heavy-Tail Phenomena, probabilistic and statistical modeling. Springer.

[118] Scott, J., Marshall, G. (Eds.). (2009). A dictionary of sociology. Oxford University Press, USA.

[119] Sen, A., Ed. (1973). On Economic Inequality, Oxford University Press.

[120] Shorack, G. R. and Wellner, J. A. (1986). Empirical Processes with Applications to Statistics. Wiley, New York.

[121] Shorrocks, A. F. (1983). Ranking income distributions. Economica 50, 3- 17.

[122] Smith, R. (1985). Maximum likelihood estimation in a class of nonregular cases. Biometrika, 72(1):67- 90.

[123] Smith, R. (1987). Estimating tails of probability distributions. Annals of Statistics, 15(3): 1174- 1207.

[124] Stiglitz, J. E. (2012) The Price of Inequality: How Today's Divided Society Endangers Our Future. New York: W.W. Norton Co.

[125] Stone, C.J. (1977). Consistent nonparametric regression (with discussion). The Annals of Statistics, 5:595- 645.

[126] Stute, W. (1995). The central limit theorem under random censorship. Ann. Statist., 422- 439.

[127] Theil H., (1967). Economics and Information Theory, North-Holland, Amsterdam, The Netherlands.

[128] Tse S.M. (2006) Lorenz curve for truncated and censored data, Annals of the Institute of Statistical Mathematics, 58, 675–686.

[129] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. Journal of the American statistical association, 69(345), 169-173.

[130] Von Mises, R. (1936). La Distribution de la plus grande des n valeurs. Selected papers, American Mathematical Society, 271-294.

[131] Weinstein, S. B. (1973). Theory and application of some classical and generalized asymptotic distributions of extreme values. IEEE Transactions on Information Theory, 19(2):148–154.

[132] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. J. Amer. Statist. Assoc., 73 ,812-815.

[133] Xu, K. (2003). How has the literature on gini's index evolved in the past 80 years? Economics working paper, Dalhousie University.

[134] Yitzhaki S., (1983). On the Extension of the Gini Index, International Economic Review, 24, 617-628.

[135] Yitzhaki, S. (1998), "More than a dozen alternative ways of spelling Gini", Research on economic inequality 8, 13- 30.

[136] Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index. Journal of Multivariate Analysis, 100(4):794- 815.- R.T. Rockafellar, S. Uryasev. Optimization of conditional value-at-risk. Journal of Risk, 2:21–42, 2000.

[137] Zitikis, R. (1998). The Vervaat process. In B. Szyszkowicz (Ed.) Asymptotic Methods in Probability and Statistics (pp.667–694). Amsterdam: North-Holland.