



# A sequential design for extreme quantiles estimation under binary sampling.

Michel Broniatowski, Emilie Miranda

## ► To cite this version:

Michel Broniatowski, Emilie Miranda. A sequential design for extreme quantiles estimation under binary sampling.. 2019. hal-02190810

**HAL Id: hal-02190810**

**<https://hal.science/hal-02190810>**

Preprint submitted on 19 Jan 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# A sequential design for extreme quantiles estimation under binary sampling.

Michel Broniatowski and Emilie Miranda  
LPSM Sorbonne Universite, Paris

July 22, 2019

## Abstract

We propose a sequential design method aiming at the estimation of an extreme quantile based on a sample of dichotomic data corresponding to peaks over a given threshold. This study is motivated by an industrial challenge in material reliability and consists in estimating a failure quantile from trials whose outcomes are reduced to indicators of whether the specimen have failed at the tested stress levels. The solution proposed is a sequential design making use of a splitting approach, decomposing the target probability level into a product of probabilities of conditional events of higher order. The method consists in gradually targeting the tail of the distribution and sampling under truncated distributions. The model is GEV, and sequential estimation of its parameters involves an improved maximum likelihood procedure for binary data, due to the large uncertainty associated with such a restricted information.

Consider a non negative random variable  $X$  with distribution function  $G$ . Let  $X_1, \dots, X_n$  be  $n$  independent copies of  $X$ . The aim of this paper is to estimate  $q_{1-\alpha}$ , the  $(1 - \alpha)$ -quantile of  $G$  when  $\alpha$  is much smaller than  $1/n$ . We therefore aim at the estimation of so-called extreme quantiles. This question has been handled by various authors, and we will review their results somehow later. The approach which we develop is quite different since we do not assume that the  $X_i$ 's can be observed. For any threshold  $x$ , we define the r.v.

$$Y = \begin{cases} 1 & \text{if } X \leq x \\ 0 & \text{if } X > x \end{cases}$$

which therefore has a Bernoulli distribution with parameter  $G(x)$ . We may choose  $x$  but do not observe  $X$ , but merely  $Y$ . Therefore any inference on  $G$  suffers from a loss of information. In industrial statistics such a setting is quite common: when exploring the strength of a material, or of a bundle, we may set a constraint  $x$  on it, and observe whether the bundle has broken or not at this level of constraint. Denoting  $X$  the resistance of this bundle, we observe  $Y$ . Inference on  $G$  can be performed for large  $n$  making use of many thresholds  $x$ . Unfortunately such a procedure will not be of any help for extreme quantiles. It will be assumed that we are able to observe  $Y$  not only when  $X$  follows  $G$  but also when  $X$  follows the conditional distribution of  $X$  given  $X > x$ . In such a case we will be able to estimate  $q_{1-\alpha}$  even when  $\alpha < 1/n$  where  $n$  designates the total number of trials. In material sciences this amounts to consider trials based on artificially modified materials; in the case when we aim at estimation of extreme upper quantiles, this amounts to strengthen the material. Clearly we would consider a family of increasing thresholds  $x_1, \dots, x_m$  and for each of them realize  $n_1, \dots, n_m$  trials, each block of iid realizations  $Y$ 's being therefore functions of the corresponding unobserved  $X$ 's with distribution  $G$  conditioned upon  $X > x_l$ ,  $1 \leq l \leq m$ . Our procedure thus aims at the definition of the design which allows for the estimation of extreme quantiles.

The present setting is therefore quite different from that usually considered for similar problems under complete information. As sketched above it is specifically suited for industrial statistics and reliability studies in the science of materials.

From a strictly statistical standpoint, the above description may also be considered when the distribution  $G$  is of some special form, namely when the conditional distribution of  $X$  given  $X > x$  has a functional form which differs from that of  $G$  only through some changes of the parameters. In this case simulation under these conditional distributions can be performed for adaptive choice of the thresholds  $x_l$ 's, substituting the above sequence of trials. This sequential procedure allows to estimate iteratively the initial parameters of  $G$  and to obtain  $q_{1-\alpha}$  combining corresponding quantiles of the conditional distributions above thresholds, a method named splitting. In this method, we will choose sequentially the  $x_l$ 's in a way that  $q_{1-\alpha}$  will be obtained easily from the last distribution of  $x$  conditioned upon  $X > x_m$ .

In safety or in pharmaceutical control, the focus is usually set on the behavior of a variable of interest (strength, acceptable tolerable dose) for small (or even very small) levels. In these settings the above considerations

turn to be equivalently stated through a clear change of variable, considering the inverse of the variable of interest. As an example which is indeed at the core of the motivation for this paper, and in order to make this approach more intuitive, we first sketch briefly the industrial situation which motivated this work. In the next section, we look at a safety property, namely the supremum of all values of  $x$  for which  $X < x$  happens with very low probability  $\alpha$ ; think at  $(X < x)$  being the event that the material brakes when its strength  $X$  is less than  $x$  after  $n_0$  solicitations ( $n_0$  large), and  $x$  is small. So we look at values  $x$  which specify very rare events, typically failures under very small solicitation. The sequence of thresholds is decreasing, starting from a material with good properties with respect to failure, and downgrading it, in the sense that trials under the distribution of  $X$  given  $X < x$  are obtained on a downgraded material.

As stated above, the problem at hand is the estimation of very small quantiles. Classical techniques in risk theory pertain to large quantiles estimation. For example, the Generalized Pareto Distribution, to be referred to later on, is a basic tool in modeling extreme risks and exceedances over thresholds. Denoting  $R$  a variable of interest and  $\tilde{R} := 1/R$ , then obviously, for  $x > 0$ ,  $(R < x)$  is equivalent to  $(\tilde{R} > u)$  with  $u = 1/x$ . In this paper we will therefore make use of this simple duality, stating formulas for  $R$ , starting with classical results pertaining to  $\tilde{R}$  when necessary. Note that when  $q_\alpha$  designates the  $\alpha$ -quantile of  $R$  and respectively  $\tilde{q}_{1-\alpha}$  the  $(1 - \alpha)$ -quantile of  $\tilde{R}$ , it holds  $q_\alpha = 1/\tilde{q}_{1-\alpha}$ . The resulting notation may seem a bit cumbersome; However the reader accustomed to industrial statistics will find it more familiar.

## 1 Industrial challenge

In aircraft industry, one major challenge is the characterization of extreme behaviors of materials used to produce engine pieces. Especially, we will consider extreme risks associated with fatigue wear, which is a very classical type of damage suffered by engines during flights. It consists in the progressive weakening of a material due to cyclic loadings applied on it many times and can lead to the material failure. As shown in Figure 1, a loading cycle is defined by several quantities: the minimal and maximal stresses  $\sigma_{\min}$  et  $\sigma_{\max}$ , the stress amplitude  $\sigma_a = \frac{\sigma_{\max} - \sigma_{\min}}{2}$ , and other indicators such as the

stress ratio  $\frac{\sigma_{\min}}{\sigma_{\max}}$ .

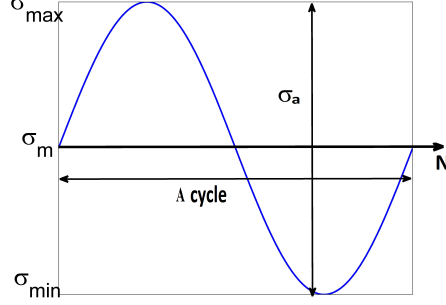


Figure 1: Loading cycle on a material

The fatigue strength of a given material is studied through experimental campaigns designed at fixed environmental covariates to reproduce flight conditions. The trials consist in loading at a given stress level a dimensioned sample of material up to its failure or the date of end of trial. The lifetime of a specimen is measured in terms of number of cycles to failure, usually subject to right censoring.

The campaign results are then used to study fatigue resistance and are represented graphically in an S-N scale (see figure 2). S-N curves highlight the existence of three fatigue regimes. Firstly, low cycle fatigue corresponds to short lives associated with high levels of stress. Secondly, during high cycle fatigue, the number of cycles to failure decreases log-linearly with respect to the loading. The last regime is the endurance limit, in which failure occurs at very high number of cycles or doesn't occur at all. We will focus in the following on the endurance limit, which is also the hardest regime to characterize since there is usually only few and scattered observations.

In this framework, we are focusing on extreme minimal risk. The critical quantities that are used to characterize minimal risk linked to fatigue damage are failure quantiles, called in this framework allowable stresses at a given number of cycles and for a fixed level of probability. Those quantiles are of great importance since they intervene in decisions to dimension engine pieces, in pricing decisions as well as in maintenance policies.

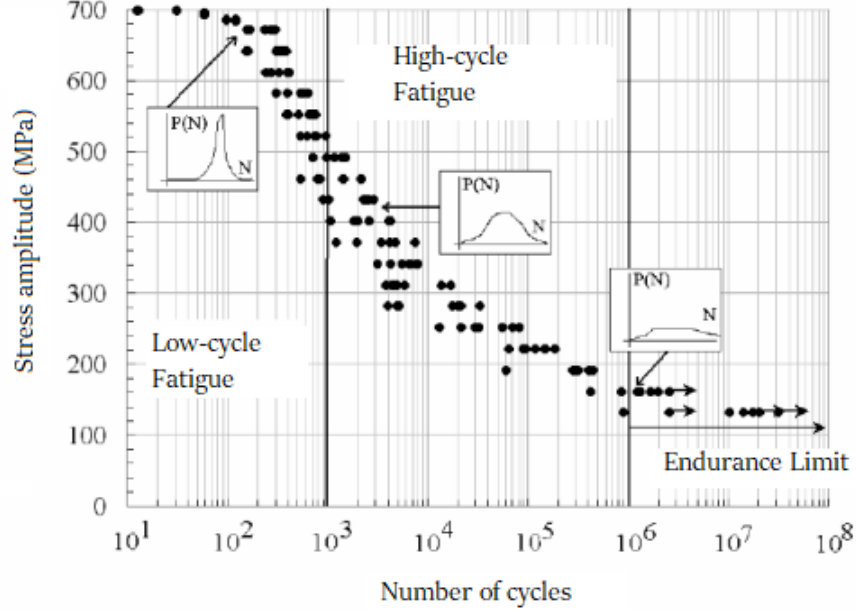


Figure 2: S-N curve

## 2 Objectives

The aim of this study is to propose a new design method for the characterization of allowable stress in very high cycle fatigue, for a very low risk  $\alpha$  of order  $10^{-3}$ . We are willing to obtain a precise estimation method of the  $\alpha$ -failure quantile based on a minimal number of trials.

Denote  $N$  the lifetime of a material in terms of number of cycles to failure and  $S$  the stress amplitude of the loading, in MPa. Let  $n_0$  be the targeted time span of order  $10^6 - 10^7$ .

Define the allowable stress  $s_\alpha$  at  $n_0$  cycles and level of probability  $\alpha = 10^{-3}$  the level of stress that guarantee that the risk of failure before  $n_0$  does not exceed  $\alpha$ :

$$s_\alpha = \sup \{s : \mathbb{P}(N \leq n_0 | S = s) \leq \alpha\} \quad (1)$$

where  $\mathbb{P}_s(N \leq n_0) = \mathbb{P}(N \leq n_0 | S = s)$ .

We will now introduce a positive r.v.  $R = R_{n_0}$  modeling the resistance of the material at  $n_0$  cycles and homogeneous to the stress.  $R$  is the variable of

interest in this study and its distribution  $\mathbb{P}$  is defined as:

$$\mathbb{P}_s(N \leq n_0) = \mathbb{P}(R \leq s) \quad (2)$$

Thus, the allowable stress can be rewritten as the  $\alpha$ -quantile of the distribution of  $R$ .

$$s_\alpha = \sup \{s : \mathbb{P}(R \leq s) \leq \alpha\} \quad (3)$$

However,  $R$  is not directly observed. Indeed, the usable data collected at the end of a test campaign consists in couples of censored fatigue life - stress levels  $(\min(N, n_0), s)$  where  $s$  is part of the design of the experiment. The relevant information that can be drawn from those observations to characterize  $R$  is restricted to indicators of whether or not the specimen tested has failed at  $s$  before  $n_0$ . Therefore, the relevant observations corresponding to a campaign of  $n$  trials are formed by a sample of variables  $Y_1, \dots, Y_n$  with for  $1 \leq i \leq n$ ,

$$Y_i = \begin{cases} 1 & \text{if } R_i \leq s_i \\ 0 & \text{if } R_i > s_i \end{cases}$$

where  $s_i$  is the stress applied on specimen  $i$ .

Note that the number of observations is constrained by industrial and financial considerations; Thus  $\alpha$  is way lower than  $1/n$  and we are considering a quantile lying outside the sample range.

While we motivate this paper with the above industrial application, note that this kind of problem is of interest in other domains, such as broader reliability issues or medical trials through the estimation of the maximal dose of tolerated toxicity for a given drug.

### 3 Extreme quantile estimation, a short survey

As seen above estimating the minimal admissible constraint raises two issues; on one hand the estimation of an extreme quantile, and on the other hand the need to proceed to inference based on exceedances under thresholds. We present a short exposition of these two areas, keeping in mind that the literature on extreme quantile estimation deals with complete data, or data under right censoring.

### 3.1 Extreme quantiles estimation methods

Extreme quantile estimation in the univariate setting is widely covered in the literature when the variable of interest  $X$  is either completely or partially observed.

The usual framework is the study of the  $(1 - \alpha)$ -quantile of a r.v  $X$ , with very small  $\alpha$ .

The most classical case corresponds to the setting where  $x_{1-\alpha}$  is drawn from a  $n$  sample of observations  $X_1, \dots, X_n$ . We can distinguish estimation of high quantile, where  $x_{1-\alpha}$  lies inside the sample range, see Weissman 1978[16] and Dekkers and al. 1989[4], and the estimation of an extreme quantile outside the boundary of the sample, see for instance De Haan and Rootzén 1993[3]. It is assumed that  $X$  belongs to the domain of attraction of an extreme value distribution. The tail index of the latter is then estimated through maximum likelihood (Weissman 1978[16]) or through an extension of Hill's estimator (see the moment estimator by Dekkers and al. 1989[4]). Lastly, the estimator of the quantile is deduced from the inverse function of the distribution of the  $k$  largest observations. Note that all the above references assume that the distribution has a Pareto tail. An alternative modeling has been proposed by De Valk 2016[5] and De Valk and Cai 2018[6], and consists in assuming a Weibull type tail, which enables to release some second order hypotheses on the tail. This last work deals with the estimation of extreme quantile lying way outside the sample range and will be used as a benchmark method in the following sections.

Recent studies have also tackled the issue of censoring. For instance, Beirlant and al. 2007[2] and Einmahl and al. 2008[9] proposed a generalization of the peak-over-threshold method when the data are subjected to random right censoring and an estimator for extreme quantiles. The idea is to consider a consistent estimator of the tail index on the censored data and divide it by the proportion of censored observations in the tail. Worms and Worms 2014[17] studied estimators of the extremal index based on Kaplan Meier integration and censored regression.

However the literature does not cover the case of complete truncation, i.e when only exceedances over given thresholds are observed. Indeed, all of the above are based on estimations of the tail index over weighed sums of the higher order statistics of the sample, which are not available in the problem of interest in this study. Classical estimation methods of extreme quantiles are thus not suited to the present issue.



In the following, we study designs of experiment at use in industrial contexts and their possible application to extreme quantiles estimation.

## 3.2 Sequential design based on dichotomous data

In this section we review two standard methods in the industry and in biostatistics, which are the closest to our purpose. Up to our knowledge, no technique specifically addresses inference for extreme quantiles.

We address the estimation of small quantiles, hence the events of interest are of the form  $(R < s)$  and the quantile is  $q_\alpha$  for small  $\alpha$ .

The first method is the *staircase*, which is the present tool used to characterize a material fatigue strength.

The second one is the *Continual Reassessment Method (CRM)* which is adapted for assessing the admissible toxicity level of a drug in Phase 1 clinical trials.

Both methods rely on a parametric model for the distribution of the strength variable  $R$ . We have considered two specifications, which allow for simple comparisons of performances, and do not aim at an accurate modelling in safety.

### 3.2.1 The Staircase method

Denote  $\mathbb{P}(R \leq s) = \phi(s, \theta_0)$ . Invented by Dixon and Mood (1948)[7], this technique aims at the estimation of the parameter  $\theta_0$  through sequential search based on data of exceedances under thresholds. The procedure is as follows.

#### Procedure

Input data:

- The initial value for the constraint, say  $S_{ini}$ . It is chosen as close as possible to the supposed expectation of  $R$ .
- The step  $\delta$  between two levels of stress during the trial. It is chosen as close as possible to the supposed standard deviation of  $R$ . In practice good results for the Staircase method are obtained for  $\delta$  between  $0.5\sigma$  and  $2\sigma$ .

- The number of cycles to perform before concluding on the issue of the trial on an item, denoted  $n_0$ .
- The total number of items to be tested,  $K$ .

The first item is tested at level  $s_{(1)} = S_{ini}$ . Under failure, the next item is tested at level  $s_{(2)} = S_{ini} - \delta$ . Under no failure  $s_{(2)} = S_{ini} + \delta$ . The process is illustrated in figure 3.2.1

Denote  $Y_i$  the variable associated to the issue of the trial on ,  $1 \leq i \leq K$ , where  $Y_i$  assumes value 1 under failure and 0 under no failure,  $Y_i = \mathbb{1}_{N_a \leq n_0} \sim \mathcal{B}(\phi(s_i, \theta_0))$ .

### Estimation

After the  $K$  trials, the parameter  $\theta_0$  is estimated through maximization of the likelihood, namely

$$\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^n \phi(s_i, \theta)^{y_i} (1 - \phi(s_i, \theta))^{(1-y_i)} \quad (4)$$

A major advantage of the Staircase method lies in the fact that the number of trials to be performed in order to get a reasonable estimator of  $\theta_0$  is small, as seen below.

### Numerical results

In this section numerical performances of this method is considered through simulation for the estimation of an admissible constraint at level  $10^{-3}$ . This method has been tested on the two above models, with  $K = 100$ .

#### Exponential case

We chose  $R \sim \mathcal{E}(0.2)$ . The input parameters are  $S_{ini} = 5$  and  $\delta = 15 \in [0.5 \times \frac{1}{\lambda^2}, 2 \times \frac{1}{\lambda^2}]$ .

The accuracy of the procedure are obtained on a batch of 1000 replications , each with  $K = 100$ . The relative error pertaining to the parameter  $\lambda$  is roughly 25%, although the input parameters are somehow optimal for the method. The resulting relative error on the  $10^{-3}$  quantile is 40%. Indeed the parameter  $\lambda$  is underestimated, which results in an overestimation of the variance  $1/\lambda^2$  , which induces an overestimation of the  $10^{-3}$  quantile.

Relative error			
On the parameter		On $s_\alpha$	
Mean	Std	Mean	Std
-0.252	0.178	0.4064874	0.304

Table 1: Results obtained using the *Staircase* method through simulations under the exponential model.

Relative error					
On $\mu$		On $\sigma$		On $s_\alpha$	
Mean	Std	Mean	Std	Mean	Std
-0.059	0.034	1.544	0.903	-1.753	0.983

Table 2: Results obtained using the *Staircase* method through simulations under the Gaussian model.

### Gaussian case

We now chose  $R \sim \mathcal{N}(60, 10)$ . The value of  $S_{\text{ini}}$  is set to the expectation and  $\delta = 7$  belongs to the interval  $[\frac{\sigma}{2}, 2\sigma]$ . The same procedure as above is performed and yields the following results:

The expectation of  $R$  is recovered rather accurately, whereas the estimation of the standard deviation suffers a loss in accuracy, which in turn yields a relative error of 180 % on the  $10^{-3}$  quantile.

### Drawback of the Staircase method

Both cases show that this method is not adequate for the estimation of extreme quantiles, which follows from an extrapolation based on estimated parameters, which furthermore may suffer of bias. Also, reparametrization of the distribution making use of the theoretical extreme quantile would not help, since the estimator would inherit of a large lack of accuracy.

### 3.2.2 The Continuous Reassessment Method (CRM)

The CRM (O’Quigley, Pepe and Fisher, 1990[12]) aims at the estimation of  $q_\alpha$  among  $K$  stress levels  $s_1, \dots, s_K$ .

$$\mathbb{P}(R \leq s) = \psi(s, \beta_0)$$

The estimator of  $q_\alpha$  is

$$s^* := \underset{s \in \{s_1, \dots, s_K\}}{\operatorname{arginf}} |\psi(s_k, \beta_0) - \alpha|.$$

This optimization is performed iteratively.

Start with an initial estimator  $\widehat{\beta}_1$  of  $\beta_0$ , for example through a Bayesian choice as proposed in [12]. Define

$$s_1^* := \underset{s \in \{s_1, \dots, s_K\}}{\operatorname{arginf}} |\psi(s_k, \widehat{\beta}_1) - \alpha|.$$

**Step 1.** Perform  $J$  trials under  $\psi(\cdot, \beta_0)$ , say  $R_{1,1}, \dots, R_{1,J}$  and observe only their value under threshold, say  $Y_{1,j} := 1_{R_{1,j} < s_1^*}, 1 \leq j \leq J$ .

**Step i.** Iteration  $i$  consists in two steps :

- Firstly an estimate  $\widehat{\beta}_i$  of  $\beta_0$  is produced on the basis of the information beared by the  $Y_{l,j}$ 's trials performed in all the preceding iterations , say  $1 \leq l \leq i - 1$ , through maximum likelihood under  $\psi(\cdot, \beta_0)$ .
- Secondly the current estimate of  $q_\alpha$  is defined by

$$s_{i+1}^* := \underset{s \in \{s_1, \dots, s_K\}}{\operatorname{arginf}} |\psi(s_k, \widehat{\beta}_i) - \alpha|;$$

This stress level  $s_{i+1}^*$  is the one under which the trials of iteration  $i + 1$  will be performed.

The stopping rule depends on the context (maximum number of trials or stabilization of the results).

Both the Staircase and the CRM have the same drawback in the context of extreme quantile estimation, since they aim at the estimation of quantiles of order 0.2 or so, far from the target  $\alpha = 10^{-3}$ .

Relative error			
On the 0.1-quantile		On the $10^{-3}$ -quantile	
Mean	Std	Mean	Std
0.129	0.48	-0.799	0.606

Table 3: Results obtained through *CRM* on simulations for the exponential model

### Numerical simulation for the CRM

Under the exponential model with parameter  $\lambda = 0.2$  and through  $N = 10$  iterations of the procedure, and  $K = 10$ , with equally distributed thresholds  $s_1, \dots, s_K$ , and performing  $m = 50$  trials at each iteration, the following results are obtained.

The  $10^{-3}$ -quantile is poorly estimated. Indeed for thresholds close to the expected quantile, nearly no failure is observed. So, for acceptable  $m$ , the method is not valid; Figure 3.2.2 shows the increase of accuracy with respect to  $m$ .

## 4 Estimation of extreme quantiles

### 4.1 Splitting

Denote  $\mathbb{P}$  the distribution of the r.v.  $R$ .

For  $s_\alpha = s_m < s_{m-1} < \dots < s_1$  it holds:

$$\{R \leq s_\alpha\} = \{R \leq s_m\} \subset \dots \subset \{R \leq s_1\}.$$

It follows that

$$\mathbb{P}(R \leq s_\alpha) = \mathbb{P}(R \leq s_1) \prod_{j=1}^{m-1} \mathbb{P}(R \leq s_{j+1} \mid R \leq s_j) \quad (5)$$

The thresholds  $(s_j)_{j=1, \dots, m}$  should be chosen such that all  $\mathbb{P}(R \leq s_{j+1} \mid R \leq s_j)_{j=1, \dots, m}$  be of order 0.2 or 0.3, in such a way that  $(R \leq s_{j+1})$  is observed in experiments performed under the conditional distribution of  $R$  given  $(R \leq s_j)$ , and in a way which makes  $\alpha$  recoverable by a rather small number of such probabilities  $\mathbb{P}(R \leq s_{j+1} \mid R \leq s_j)$  making use of (5).

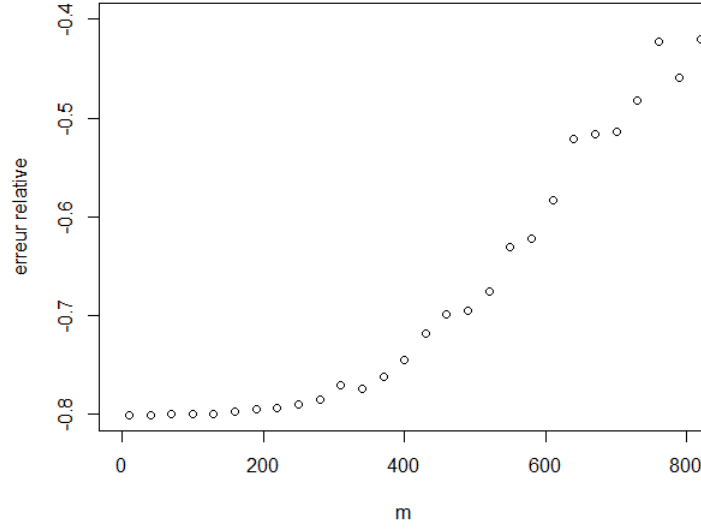


Figure 3: Relative error on the  $10^{-3}$ -quantile with respect to the number of trials for each stress level

## 4.2 Sampling under the conditional probability

In practice batches of specimen are put under trial, each of them with a decreasing strength; this allows to target the tail of the distribution  $\mathbb{P}$  iteratively.

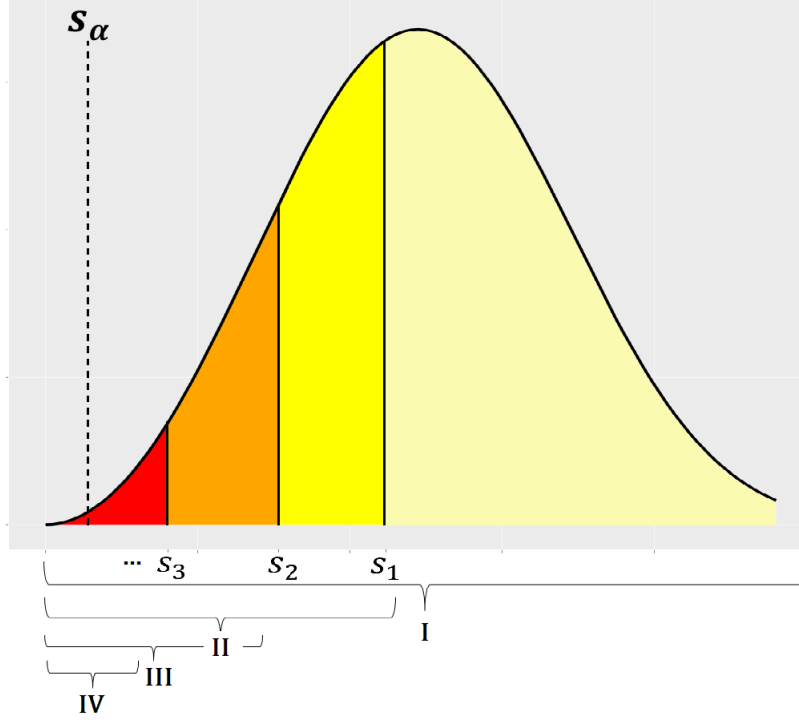


Figure 4: Sampling under the strength density at  $n_0$  cycles

- In the first step, points are sampled in zone (I)
- In the next step, only specimen with strength in zone II are considered, and so on.
- In the final step, the specimen are sampled in zone IV under test have a very small probability to fail under  $n_0$  cycles under  $\mathbb{P}_0$ , but still with a probability of order 0.2 to fail under their own law of failure, which is  $\mathbb{P}(\cdot \mid R \leq s_m)$ .

In practice flaws in the specimen are introduced during manufacturing, since the strength varies inversely with respect to their size. Preliminary trials require to assess that for a given stress constraint  $s$  there exists a size of flaw  $\theta$  such that, noting and noting  $R_\theta$  the strength of the specimen with flaw of size  $\theta$ , it holds  $\mathcal{L}(R_\theta) \approx \mathcal{L}(R \mid R \leq s)$ .

### 4.3 Modeling the distribution of the strength, Pareto model

The events under consideration have small probability under  $\mathbb{P}$ . By (5) we are led to consider the limit behavior of conditional distributions under smaller and smaller thresholds, for which we make use of classical approximations due to Balkema and de Haan (1974[1]) which stands as follows, firstly in the commonly known setting of exceedances over increasing thresholds.

Denote  $\tilde{R} := 1/R$ . Then

**Theorem 1.** *For  $\tilde{X}$  of distribution  $F$  belonging to the maximum domain of attraction of an extreme value distribution with tail index  $c$ , i.e.  $F \in MDA(c)$ , it holds that: There exists  $a = a(s)$ , such that:*

$$\lim_{s \rightarrow \infty} \mathbb{P} \left( \tilde{X} - s > x \mid \tilde{X} > s \right) = 1 - G_{(c,a)}(x)$$

where  $G_{(c,a)}$  is defined through

$$G_{(c,a)}(x) = 1 - \exp \left\{ - \int_0^{\frac{x}{a}} [(1 + ct)_+]^{-1} dt \right\}$$

where  $a > 0$  and  $c \in \mathbb{R}$ .

The distribution  $G$  is the Generalized Pareto distribution  $GPD(c, a)$  which is defined explicitly through

$$1 - G(x) = \begin{cases} (1 + \frac{c}{a}x)^{-1/c} & \text{when } c \neq 0 \\ \exp(-\frac{x}{a}) & \text{when } c = 0 \end{cases}$$

where  $x \geq 0$  for  $c \geq 0$  and  $0 \leq x \leq -\frac{a}{c}$  if  $c < 0$ .

Generalized Pareto distributions enjoy invariance through threshold conditioning, an important property for our sake. Indeed it holds, for  $\tilde{R} \sim GDP(c, a)$  and  $x > s$ ,

$$\mathbb{P} \left( \tilde{R} > x \mid \tilde{R} > s \right) = \left( 1 + \frac{c(x-s)}{a+cs} \right)^{-1/c} \quad (6)$$

We therefore state:



**Proposition 2.** *When  $\tilde{R} \sim GPD(c, a)$  then, given  $(\tilde{R} > s)$ , the r.v.  $\tilde{R} - s$  follows a  $GPD(c, a + cs)$ .*

The GPD's are on the one hand stable under thresholding and on the other appear as the limit distribution for thresholding operations. This chain of arguments is quite usual in statistics, motivating the recourse to the ubiquitous normal or stable laws for additive models. This plays in favor of GPD's as modelling the distribution of  $\tilde{R}$  for excess probability inference. Due to the lack of memory property, the exponential distribution which appears as a possible limit distribution for excess probabilities in Theorem 1 do not qualify for modelling.

Since we handle variables  $R$  which can approach 0 arbitrarily (i.e. unbounded  $\tilde{R}$ ) the parameter  $c$  is assumed positive.

Turning to the context of the minimal admissible constraint, we make use of the r.v.  $R = 1/\tilde{R}$  and proceed to the corresponding change of variable.

When  $c > 0$ , the density of the r.v.  $R$  is

$$f_{c,a}(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{ax^2} \left(1 + \frac{c}{ax}\right)^{-1/c-1} & \text{for } x > 0 \end{cases} \quad (7)$$

and its distribution function writes

$$F_{c,a}(x) = \left(1 + \frac{c}{ax}\right)^{-1/c}$$

for nonnegative  $x$ .

For  $0 < x < u$ , the conditional distribution of  $R$  given  $(R < u)$  is

$$\mathbb{P}(R < x \mid R < u) = \left(1 - \frac{c(\frac{1}{x} - \frac{1}{u})}{a + \frac{c}{u}}\right)^{-1/c}$$

which proves that the distribution of  $R$  is stable under threshold conditioning with parameter  $(a_u, c)$  with

$$a_u = a + \frac{c}{u}. \quad (8)$$

In practice at each step  $j$  in the procedure the stress level  $s_j$  equals the corresponding threshold  $1/\tilde{s}_j$ , a right quantile of the conditional distribution of  $\tilde{R}$  given  $(\tilde{R} > \tilde{s}_{j-1})$ . Therefore the observations take the form  $Y_i = \mathbb{1}_{R_i < s_{j-1}} = \mathbb{1}_{\tilde{R}_i > \tilde{s}_{j-1}}$ ,  $i = 1, \dots, n_j$ .

A convenient feature of model (7) lies in the fact that the conditional distributions are completely determined by the initial distribution of  $R$ , therefore by  $a$  and  $c$ . The parameters  $a_j$  of the conditional distributions are determined from these initial parameters and by the corresponding stress level  $s_j$ ; see (8).

#### 4.4 Notations

The distribution function of the r.v.  $\tilde{R}$  is a  $GPD(c_T, a_T)$  of distribution function  $G_{(c_T, a_T)}$ . Note  $\bar{G}_{(c_T, a_T)} = 1 - G_{(c_T, a_T)}$ .

Our proposal relies on iterations. We make use of a set of thresholds  $(\tilde{s}_1, \dots, \tilde{s}_m)$  and define for any  $j \in \{1, \dots, m\}$

$$G_{(c_j, a_j)}(x - \tilde{s}_j) = \mathbb{P}(\tilde{R} > x \mid \tilde{R} > \tilde{s}_j)$$

with  $c_j = c_T$  and  $a_j = a_T + c_T \tilde{s}_j$  where we used (6).

At iteration  $j$ , denote  $(\hat{c}, \hat{a})_j$  the estimators of  $(c_j, a_j)$ . Therefore  $1 - G_{(\hat{c}, \hat{a})_j}(x - \tilde{s}_j)$  estimates  $\mathbb{P}(\tilde{R} > x \mid \tilde{R} > \tilde{s}_j)$ . Clearly, estimators of  $(c_T, a_T)$  can be recovered from  $(\hat{c}, \hat{a})_j$  through  $\hat{c}_T = \hat{c}$  and  $\hat{a}_T = \hat{a} - \hat{c} \tilde{s}_j$ .

#### 4.5 Sequential design for the extreme quantile estimation

Fix  $m$  and  $p$ , where  $m$  denotes the number of stress levels under which the trials will be performed, and  $p$  is such that  $p^m = \alpha$ .

Set a first level of stress, say  $s_0$  large enough (i.e.  $\tilde{s}_0 = 1/s_0$  small enough) so that  $p_0 = \mathbb{P}(R < s_0)$  is large enough and perform trials at this level. The optimal value of  $s_0$  should satisfy  $p_0 = p$ , which cannot be secured. This choice is based on expert advice.

Turn to  $\tilde{R} := 1/R$ . Estimate  $c_T$  and  $a_T$ , for the GPD  $(c_T, a_T)$  model describing  $\tilde{R}$ , say  $(\hat{c}, \hat{a})$ , based on the observations above  $\tilde{s}_0$  (note that under  $s_0$  the outcomes of  $R$  are easy to obtain, since the specimen is tested under medium stress, say).

Define

$$\tilde{s}_1 := \sup \{s : \bar{G}_{(\hat{c}, \hat{a})_0}(s - \tilde{s}_0) < p\}$$

the  $(1 - p)$ -quantile of  $G_{(\hat{c}, \hat{a})_0}$ .

At step  $j = 1$  perform  $K$  trials under  $G_{(c_1, a_1)}$  say  $\tilde{R}_{1,1}, \dots, \tilde{R}_{1,K}$  and consider the observable variables  $Y_{1,i} := 1_{\tilde{R}_{1,i} > \tilde{s}_1}$ . Therefore the  $K$  iid replications  $Y_{1,1}, \dots, Y_{1,K}$  follow a Bernoulli  $\mathcal{B}(\overline{G}_{(c_1, a_1)}(\tilde{s}_1 - \tilde{s}_0))$ . Estimate  $(c_1, a_1)$  in the resulting Bernoulli scheme, say  $(\hat{c}, \hat{a})_1$ . Then define

$$\tilde{s}_2 := \sup \{s : \overline{G}_{(\hat{c}, \hat{a})_1}(s - \tilde{s}_1) < p\}$$

which is the  $(1 - p)$ -quantile of the estimated conditional distribution of  $\tilde{R}$  given  $\tilde{R} > \tilde{s}_1$ .

Iterate from step 1, sampling under  $G_{(c_j, a_j)}$ ; set

$$\tilde{s}_{j+1} := \sup \{s : \overline{G}_{(\hat{c}, \hat{a})_j}(s - \tilde{s}_j) < p\}.$$

Defining  $\tilde{R}_{j+1,1}, \dots, \tilde{R}_{j+1,K}$  iid with distribution  $G_{(\hat{c}, \hat{a})_j}$  and the corresponding  $Y_{j+1,i} := 1_{\tilde{R}_{j+1,i} > \tilde{s}_{j+1}}$ , estimate  $(\hat{c}, \hat{a})_{j+1}$  in the Bernoulli  $\mathcal{B}(\overline{G}_{(c_j, a_j)}(\tilde{s}_{j+1} - \tilde{s}_j))$  scheme, define  $\tilde{s}_{j+1}$  as the  $(1 - p)$ -quantile of  $G_{(\hat{c}, \hat{a})_{j+1}}$ , namely

$$\tilde{s}_{j+1} := G_{(\hat{c}, \hat{a})_j}^{-1}(1 - p) + \tilde{s}_j,$$

and so on.

In practice a conservative choice for  $m$  is given by  $m + 1 = \left\lceil \frac{\log \alpha}{\log p} \right\rceil$ , where  $\lceil \cdot \rceil$  denotes the ceiling function. This implies that the attained probability  $\tilde{\alpha}$  is less than or equal to  $\alpha$ .

The  $m + 1$  stress levels  $\tilde{s}_0 < \tilde{s}_1 < \dots < \tilde{s}_m = \tilde{s}_\alpha$  satisfy

$$\begin{aligned} \tilde{\alpha} &= \overline{G}(\tilde{s}_0) \prod_{j=0}^{m-1} \overline{G}_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j+1} - \tilde{s}_j) \\ &= p_0 p^m \end{aligned}$$

Finally by its very definition  $\tilde{s}_m$  is a proxy of  $\tilde{q}_{1-\alpha}$ .

Although quite simple in its definition, this method bears a number of drawbacks, mainly in the definition of  $(\hat{c}, \hat{a})_j$ . The next section addresses this question.

Minimum	Q25	Q50	Mean	Q75	Maximum
67.07	226.50	327.40	441.60	498.90	10 320.00

Table 4: Estimation of the  $(1-\alpha)$ -quantile,  $\tilde{s}_\alpha = 469.103$ , through procedure 4.5 with  $K = 50$

## 5 Sequential enhanced design in the Pareto model

In this section we focus on the estimation of the parameters  $(c_T, a_T)$  in the  $GPD(c_T, a_T)$  distribution of  $\tilde{R}$ . One of the main difficulty lies in the fact that the available information does not consist of replications of the r.v.  $\tilde{R}$  under the current conditional distribution  $G_{(c_j, a_j)}$  of  $\tilde{R}$  given  $(\tilde{R} > \tilde{s}_j)$  but merely on very downgraded functions of those. At step  $j$  we are given  $G_{(\hat{c}, \hat{a}_0)_j}$  and define  $\tilde{s}_{j+1}$  as its  $(1-p)$ -quantile. Simulating  $K$  r.v.  $\tilde{R}_{j,i}$  with distribution  $G_{(c_j, a_j)}$ , the observable outcomes are the Bernoulli  $(p)$  r.v.'s  $Y_{j,i} := 1_{\tilde{R}_{j,i} > \tilde{s}_{j+1}}$ . This loss of information with respect to the  $\tilde{R}_{j,i}$ 's makes the estimation step for the coefficients  $(\hat{c}, \hat{a})_{j+1}$  quite complex; indeed  $(\hat{c}, \hat{a})_{j+1}$  is obtained through the  $Y_{j,i}$ 's,  $1 \leq i \leq K$ .

It is of interest to analyze the results obtained through standard Maximum Likelihood Estimation of  $(\hat{c}, \hat{a})_{j+1}$ . The quantile  $\tilde{q}_{1-\alpha}$  is loosely estimated for small  $\alpha$ ; as measured on 1000 simulation runs, large standard deviation of  $\hat{\tilde{q}}_{1-\alpha}$  is due to poor estimation of the iterative parameters  $(\hat{c}, \hat{a})_{j+1}$ . We have simulated  $n = 200$  realizations of r.v.'s  $Y_i$  with common Bernoulli distribution with parameter  $\overline{G}_{(c_T, a_T)}(\tilde{s}_0)$ . Figure 5 shows the log likelihood function of this sample as the parameter of the Bernoulli  $\overline{G}_{(c', a')}(\tilde{s}_0)$  varies according to  $(c', a')$ . As expected this function is nearly flat in a very large range of  $(c', a')$ .

However we see in Table 5 the improvement of accuracy of the estimator of  $\tilde{q}_{1-\alpha}$  as the number of replications  $Y_{j,i}$ ,  $1 \leq i \leq K$  increases.

Changing the estimation criterion by some alternative method does not improve significantly; Figure 6 shows the distribution of the resulting estimators of  $\tilde{q}_{1-\alpha}$  for various estimation methods ( minimum Kullback Leibler, minimum Hellinger and minimum L1 distances) of  $(c_T, a_T)$ .

This motivates the need for an enhanced estimation procedure.

$\tilde{s}_\alpha$	$\tilde{s}_m$ for $K = 30$		$\tilde{s}_m$ for $K = 50$	
	Mean	Std	Mean	Std
469.103	1 276.00	12 576.98	441.643	562.757

Table 5: Estimation of the  $(1-\alpha)$ -quantile,  $\tilde{s}_\alpha = 469.103$ , through procedure 4.5 for different values of  $K$

## 5.1 An enhanced sequential criterion for estimation

We consider an additional criterion which makes a peculiar use of the iterative nature of the procedure. We will impose some control on the stability of the estimators of the conditional quantiles through the sequential procedure.

At iteration  $j - 1$ , the sample  $Y_{j-1,i}$ ,  $1 \leq i \leq n$  has been generated under  $G_{(\hat{c}, \hat{a})_{j-2}}$  and provides an estimate of  $p$  through

$$\hat{p}_{j-1} := \frac{1}{n} \sum_{i=1}^n Y_{j-1,i}. \quad (9)$$

The above  $\hat{p}_{j-1}$  estimates  $\mathbb{P}(\tilde{R} > \tilde{s}_{j-1} \mid \tilde{R} > \tilde{s}_{j-2})$  conditionally on  $\tilde{s}_{j-1}$  and  $\tilde{s}_{j-2}$ . We write this latter expression  $\mathbb{P}(\tilde{R} > \tilde{s}_{j-1} \mid \tilde{R} > \tilde{s}_{j-2})$  as a function of the parameters obtained at iteration  $j$ , namely  $(\hat{c}, \hat{a})_j$ . The above r.v's  $Y_{j-1,i}$  stem from variables  $\tilde{R}_{j-1,i}$  greater than  $\tilde{s}_{j-2}$ . At step  $j$ , estimate then  $\mathbb{P}(\tilde{R} > \tilde{s}_{j-1} \mid \tilde{R} > \tilde{s}_{j-2})$  making use of  $G_{(\hat{c}, \hat{a})_j}$ . This backward estimator writes

$$\frac{\overline{G}_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j-1})}{\overline{G}_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j-2})} = 1 - G_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j-1} - \tilde{s}_{j-2}).$$

The distance

$$\left| (\overline{G}_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j-1} - \tilde{s}_{j-2})) - \widehat{p}_{j-1} \right| \quad (10)$$

should be small, since both  $(\overline{G}_{(\hat{c}, \hat{a})_j}(\tilde{s}_{j-1} - \tilde{s}_{j-2}))$  and  $\hat{p}_{j-1}$  should approximate  $p$ .

Consider the distance between quantiles

$$\left| (\tilde{s}_{j-1} - \tilde{s}_{j-2}) - G_{(\hat{c}, \hat{a})_j}^{-1}(\hat{p}_{j-1}) \right|. \quad (11)$$

An estimate  $(\hat{c}, \hat{a})_j$  can be proposed as the minimizer of the above expression for  $(\tilde{s}_{j-1} - \tilde{s}_{j-2})$  for all  $j$ . This backward estimation provides coherence with respect to the unknown initial distribution  $G_{(c_T, a_T)}$ . Would

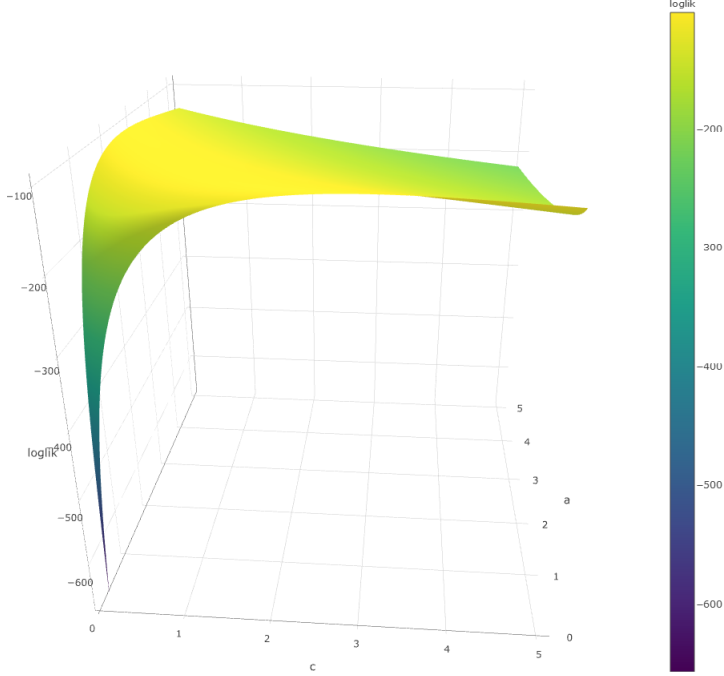


Figure 5: Log-likelihood of the Pareto model with binary data

we have started with a good guess  $(\hat{c}, \hat{a}) = (c_T, a_T)$  then the successive  $(\hat{c}, \hat{a})_j$ ,  $\tilde{s}_{j-1}$  etc would make (11) small, since  $\tilde{s}_{j-1}$  (resp.  $\tilde{s}_{j-2}$ ) would estimate the  $p$ -conditional quantile of  $\mathbb{P} \left( . | \tilde{R} > \tilde{s}_{j-2} \right)$  (resp.  $\mathbb{P} \left( . | \tilde{R} > \tilde{s}_{j-3} \right)$ ).

It remains to argue on the set of plausible values where the quantity in (11) should be minimized.

We suggest to consider a confidence region for the parameter  $(c_T, a_T)$ . With  $\widehat{p}_{j-1}$  defined in (9) and  $\gamma \in (0, 1)$  define the  $\gamma$ -confidence region for  $p$  by

$$I_\gamma = \left[ \widehat{p}_j - z_{1-\gamma/2} \sqrt{\frac{\widehat{p}_j(1-\widehat{p}_j)}{n-1}}; \widehat{p}_j + z_{1-\gamma/2} \sqrt{\frac{\widehat{p}_j(1-\widehat{p}_j)}{n-1}} \right]$$

where  $z_\tau$  is the  $\tau$ -quantile of the standard normal distribution. Define

$$\mathcal{S}_j = \left\{ (c, a) : (1 - G_{(c,a)}(\tilde{s}_j - \tilde{s}_{j-1})) \in I_\gamma \right\}.$$

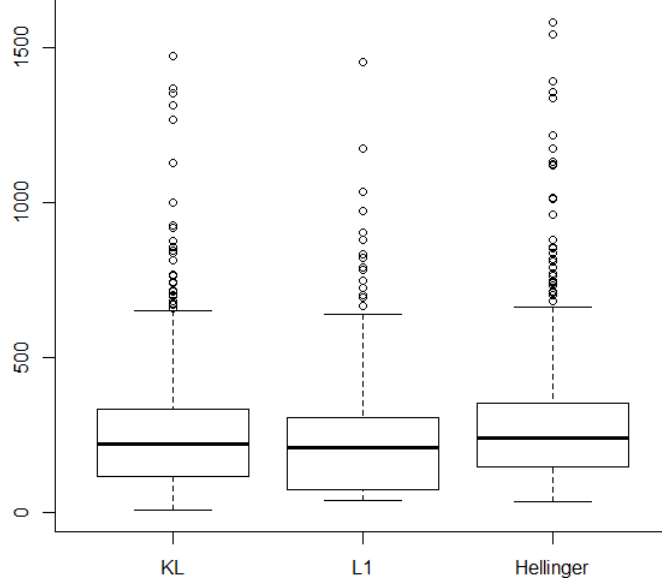


Figure 6: Estimations of the  $\alpha$ -quantile based on the Kullback-Leibler, L1 distance and Hellinger distance criterion

Therefore  $\mathcal{S}_j$  is a plausible set for  $(\hat{c}_T, \hat{a}_T)$ .

We summarize this discussion:

At iteration  $j$ , the estimator of  $(c_T, a_T)$  is a solution of the minimization problem

$$\min_{(c,a) \in \mathcal{S}_j} \left| (\tilde{s}_{j-1} - \tilde{s}_{j-2}) - G_{(c,a+c\tilde{s}_{j-2})}^{-1}(1 - \hat{p}_{j-1}) \right|.$$

As seen hereunder, this heuristics works.

## 5.2 Simulation based numerical results

This procedure has been applied in three cases. A case considered as reference is  $(c_T, a_T) = (1.5, 1.5)$ ; Secondly the case when  $(c_T, a_T) = (0.8, 1.5)$  describes a light tail with respect to the reference. Thirdly, a case  $(c_T, a_T) = (1.5, 3)$  defines a distribution with same tail index as the reference, but with a larger dispersion index.

Figure 6 shows that the estimation of  $\tilde{q}_{1-\alpha}$  deteriorates as the tail of the distribution gets heavier; Also the procedure underestimates  $\tilde{q}_{1-\alpha}$ .

Parameters	Relative error on $\tilde{s}_\alpha$	
	Mean	Std
$c = 0.8, a_0 = 1.5$ and $\tilde{s}_\alpha = 469.103$	-0.222	0.554
$c = 1.5, a_0 = 1.5$ and $\tilde{s}_\alpha = 31621.777$	-0.504	0.720
$c = 1.5, a_0 = 3$ and $\tilde{s}_\alpha = 63243.550$	0.310	0.590

Table 6: Mean and std of relative errors on the  $(1 - \alpha)$ -quantile of GPD calculated through 400 replicas of procedure 5.1.

Despite these drawbacks, we observe an improvement with respect to the simple Maximum Likelihood estimation; this is even more clear, when the tail of the distribution is heavy. Also, in contrast with the ML estimation, the sensitivity with respect to the number  $K$  of replications at each of the iterations plays in favor of this new method: As  $K$  decreases, the gain with respect to Maximum Likelihood estimation increases notably, see Figure 8.

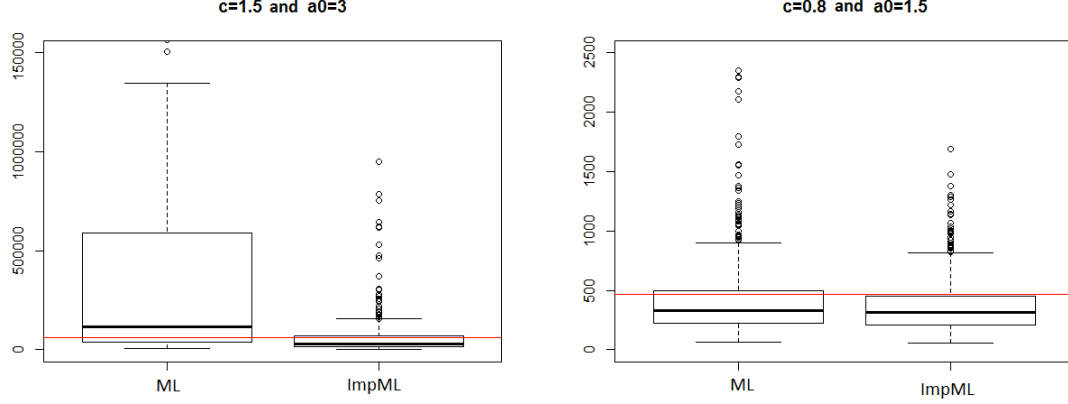
### 5.2.1 Performances of the sequential estimation

To our knowledge no method dealing with similar question is available in the literature; a vast literature is devoted to the estimation of extreme quantiles, but merely on complete data. We compare the results of our method, based on observed exceedances over thresholds, with the results which can be obtained by methods which make use of complete data; those may be seen as benchmarks for an upper bound of the performance of our method.

#### Estimation of an extreme quantile based on complete data, de Valk's estimator

In order to provide an upper bound for the performances of the estimator, we make use of the estimator proposed by C. De Valk (2016). This work aims at the estimation of a quantile of order  $p_n \in [n^{-\tau_1}; n^{-\tau_2}]$ , with  $\tau_2 > \tau_1 > 1$ , where  $n$  is the sample size. This question is in accordance with the industrial context which motivated the present paper. De Valk's proposal is a modified Hill estimator adapted to log-Weibull tailed models. De Valk's estimator is consistent, asymptotically normally distributed, but is biased for finite sample size. We briefly recall some of the hypotheses which set the context of de Valk's approach.





The red line stands for the real value of  $s\alpha$

Figure 7: Estimations of the  $(1 - \alpha)$ -quantile of two GPD obtained by Maximum Likelihood and by the improved Maximum Likelihood method

Let  $X_1, \dots, X_n$  be  $n$  iid r.v.'s with distribution  $F$ , and denote  $X_{k:n}$  the  $k$ -order statistics. A tail regularity assumption is needed in order to estimate a quantile with order greater than  $1 - 1/n$ .

Denote  $U(t) = F^{-1}(1 - 1/t)$ , and let the function  $q$  be defined by

$$q(y) = U(e^y) = F^{-1}(1 - e^{-y})$$

for  $y > 0$ .

Assume that

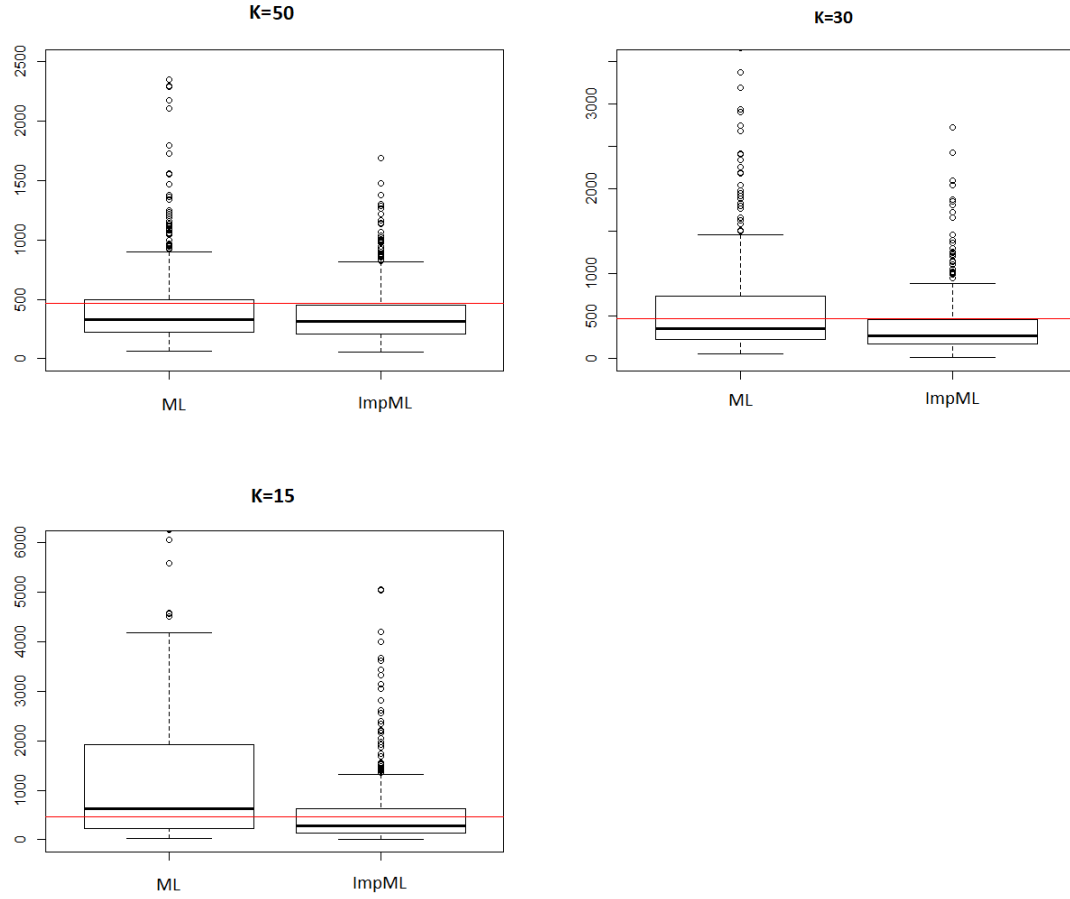
$$\lim_{y \rightarrow \infty} \frac{\log q(y\lambda) - \log q(y)}{g(y)} = h_\theta(\lambda) \quad \lambda > 0 \quad (12)$$

where  $g$  is a regularly varying function and

$$h_\theta(\lambda) = \begin{cases} \frac{\lambda^\theta - 1}{\theta} & \text{if } \theta \neq 0 \\ \log \lambda & \text{if } \theta = 0 \end{cases}$$

de Valk writes condition 12 as  $\log q \in ERV_\theta(g)$ .

*Remark :* Despite its naming of log-Generalized tails, this condition also holds for Pareto tailed distributions, as can be checked, providing  $\theta = 1$ .



The red line stands stands for the real value of  $s\alpha$

Figure 8: Estimations of the  $(1 - \alpha)$ -quantile of a  $GPD(0.8, 1.5)$  obtained by Maximum Likelihood and by the improved Maximum Likelihood method for different values of  $K$ .

We now introduce de Valk's extreme quantile estimator.

Let

$$\vartheta_{k,n} := \sum_{j=k}^n \frac{1}{j}.$$

Let  $q(z)$  be the quantile of order  $e^{-z} = p_n$  of the distribution  $F$ . The estimator makes use of  $X_{n-l_n:n}$ , an intermediate order statistics of  $X_1, \dots, X_n$ , where  $l_n$  tends to infinity as  $n \rightarrow \infty$  and  $l_n/n \rightarrow 0$ .

de Valk's estimator writes

$$\hat{q}(z) = X_{n-l_n:n} \exp \left\{ g(\vartheta_{l_n,n}) h_\theta \left( \frac{z}{\vartheta_{l_{n+1},n}} \right) \right\}. \quad (13)$$

When the support of  $F$  overlaps  $\mathbb{R}^-$  then the sample size  $n$  should be large; see de Valk ([6]) for details.

Note that, in the case of a  $GPD(c, a)$ , parameter  $\theta$  is known and equal to 1 and the normalizing function  $g$  is defined by  $g(x) = cx$  for  $x > 0$ .

### Loss in accuracy due to binary sampling

We compare the performance of de Valk's method with ours on the model, making use of complete data in de Valk's estimation, and of dichotomous ones in our approach. Clearly de Valk's results cannot be attained by the present sequential method, due to the loss of information induced by thresholding and dichotomy. Despite of this, the results can be compared, since the bias of the estimator does not exceed five times the corresponding bias of de Valk's, and its dispersion is of the same order of magnitude, when handling heavy tailed GPD models. We can assess that a large part of the volatility of the estimator produced by our sequential methodology is due to the nature of the GPD model as well as to the sample size.

## 6 Sequential design for the Weibull model

The main property which led to the GPD model is the stability through threshold conditioning. However the conditional distribution of  $\tilde{R}$  given  $(\tilde{R} > s)$  takes a rather simple form which allows for some variation of the sequential design method.

Denote  $\tilde{R} \sim W(\alpha, \beta)$ , with  $\alpha, \beta > 0$  a Weibull r.v. with scale parameter  $\alpha$  and shape parameter  $\beta$ . let  $G$  denote the distribution function of  $\tilde{R}$ ,  $g$

Parameters	Relative error on the $(1 - \alpha)$ -quantile			
	On complete data		On binary data	
	Mean	Std	Mean	Std
$c = 0.8, a_0 = 1.5$ and $s\alpha = 469.103$	0.052	0.257	-0.222	0.554
$c = 1.5, a_0 = 1.5$ and $s\alpha = 31621.777$	0.086	0.530	-0.504	0.720
$c = 1.5, a_0 = 3$ and $s\alpha = 63243.550$	0.116	0.625	0.310	0.590

Table 7: Mean and std of the relative errors on the  $1 - \alpha$ -quantile of GPD on complete and binary data for samples of size  $n = 250$  computed through 400 replicas of both estimation procedures.

Estimations on complete data are obtained with de Valk's method; Estimations on binary data are provided by the sequential design.

its density function and  $G^{-1}$  its quantile function. We thus write for non negative  $x$

$$G(x) = 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right)$$

$$\text{for } 0 < u < 1, \quad G^{-1}(u) = \alpha(-\log(1 - u))^{1/\beta}$$

The conditional distribution of  $\tilde{R}$  is a truncated Weibull distribution

$$\begin{aligned} \text{for } \tilde{s}_2 > \tilde{s}_1, \quad \mathbb{P}(\tilde{R} > \tilde{s}_2 \mid \tilde{R} > \tilde{s}_1) &= \frac{\mathbb{P}(\tilde{R} > \tilde{s}_2)}{\mathbb{P}(\tilde{R} > \tilde{s}_1)} \\ &= \exp\left\{\left(-\left(\frac{\tilde{s}_2}{\alpha}\right)^\beta + \left(\frac{\tilde{s}_1}{\alpha}\right)^\beta\right)\right\} \end{aligned}$$

Denote  $G_{s_2}$  the distribution function of  $\tilde{R}$  given  $(\tilde{R} > \tilde{s}_2)$ .

The following result helps. For  $\tilde{s}_2 > \tilde{s}_1$ ,

$$\log \mathbb{P}(\tilde{R} > \tilde{s}_2 \mid \tilde{R} > \tilde{s}_1) = \left[ \left( \frac{\tilde{s}_2}{\tilde{s}_1} \right)^\beta - 1 \right] \log \mathbb{P}(\tilde{R} > \tilde{s}_1) \quad (14)$$

Assuming  $\mathbb{P}(\tilde{R} > \tilde{s}_1) = p$ , and given  $\tilde{s}_1$  we may find  $\tilde{s}_2$  the conditional quantile of order  $1 - p$  of the distribution of  $\tilde{R}$  given  $(\tilde{R} > \tilde{s}_1)$ . This solves

the first iteration of the sequential estimation procedure through

$$\log p = \left[ \left( \frac{\tilde{s}_2}{\tilde{s}_1} \right)^\beta - 1 \right] \log p$$

where the parameter  $\beta$  has to be estimated on the first run of trials.

The same type of transitions holds for the iterative procedure; indeed for  $\tilde{s}_{j+1} > \tilde{s}_j > \tilde{s}_{j-1}$

$$\begin{aligned} \log \mathbb{P}(\tilde{R} > \tilde{s}_{j+1} \mid \tilde{R} > \tilde{s}_j) &= \left[ \frac{\log \mathbb{P}(\tilde{R} > \tilde{s}_{j+1} \mid \tilde{R} > \tilde{s}_{j-1})}{\log \mathbb{P}(\tilde{R} > \tilde{s}_j \mid \tilde{R} > \tilde{s}_{j-1})} - 1 \right] \log \mathbb{P}(\tilde{R} > \tilde{s}_j \mid \tilde{R} > \tilde{s}_{j-1}) \\ &= \left[ \frac{\tilde{s}_{j-1}^\beta - \tilde{s}_{j+1}^\beta}{\tilde{s}_{j-1}^\beta - \tilde{s}_j^\beta} - 1 \right] \log \mathbb{P}(\tilde{R} > \tilde{s}_j \mid \tilde{R} > \tilde{s}_{j-1}) \end{aligned} \quad (15)$$

At iteration  $j$  the thresholds  $\tilde{s}_j$  and  $\tilde{s}_{j-1}$  are known; the threshold  $\tilde{s}_{j+1}$  is the  $p$ -quantile of the conditional distribution,  $\mathbb{P}(\tilde{R} > \tilde{s}_{j+1} \mid \tilde{R} > \tilde{s}_j) = p$ , hence solving

$$\log p = \left[ \frac{\tilde{s}_{j-1}^\beta - \tilde{s}_{j+1}^\beta}{\tilde{s}_{j-1}^\beta - \tilde{s}_j^\beta} - 1 \right] \log p$$

where the estimate of  $\beta$  is updated from the data collected at iteration  $j$ .

## 6.1 Numerical results

Similarly as in Sections 5.2 and 5.2.1 we explore the performances of the sequential design estimation on the Weibull model. We estimate the  $(1 - \alpha)$ -quantile of the Weibull distribution in three cases. In the first one, the scale parameter  $a$  and the shape parameter  $b$  satisfy  $(a, b) = (3, 0.9)$ . This corresponds to a strictly decreasing density function, with heavy tail. In the second case, the distribution is skewed since  $(a, b) = (3, 1.5)$  and the third case is  $(a, b) = (2, 1.5)$  and describes a less dispersed distribution with lighter tail.

Table 8 shows that the performances of our procedure depend on the shape of the distribution. The estimators are less accurate in case 1, corresponding to a heavier tail. Those results are compared to the estimation errors on complete data through de Valk's methodology. As expected, the

Parameters	Relative error on the $(1 - \alpha)$ -quantile			
	On binary data		On complete data	
	Mean	Std	Mean	Std
$a_0 = 3, b_0 = 0.9$ et $s\alpha = 25.69$	0.282	0.520	0.127	0.197
$a_0 = 3, b_0 = 1.5$ et $s\alpha = 10.88$	-0.260	0.490	0.084	0.122
$a_0 = 2, b_0 = 1.5$ et $s\alpha = 7.25$	-0.241	0.450	0.088	0.140

Table 8: Mean and std of relative errors on the  $(1 - \alpha)$ -quantile of Weibull distributions on complete and binary data for samples of size  $n = 250$  computed through 400 replicas.

Estimations on complete data are obtained with de Valk's method; Estimations on binary data are provided by the sequential design.

loss of accuracy linked to data deterioration is similar to what was observed under the Pareto model, although a little more important. This can be explained by the fact that the Weibull distribution is less adapted to the splitting structure than the GPD.

## 7 Perspectives, variations around mixture forms

In this paper, we have consider two models for  $\tilde{R}$  that exploits the thresholding operations used in the splitting method. In the following, we present ideas of extensions and generalizations of those models, based on common properties of the GPD and Weibull models.

When the tail index is positive, the GPD is completely monotone, and thus can be written as the Laplace transform of a probability distribution. Thyron (1964[15]) and Thorin (1977[14]) established that a  $GPD(a_T, c_T)$ , with  $c_T > 0$ , can be written as the Laplace transform of a Gamma r.v  $V$  whose parameters are functions of  $a_T$  and  $c_T$ :  $V \sim \Gamma\left(\frac{1}{c_T}, \frac{a_T}{c_T}\right)$ . Denote  $v$  the density of  $V$ ,

$$\begin{aligned} \forall x \geq 0, \quad \bar{G}(x) &= \int_0^\infty \exp(-xy) v(y) dy \\ \text{where } v(y) &= \frac{(a_T/c_T)^{1/c}}{\Gamma(1/c_T)} y^{1/c_T-1} \exp\left(-\frac{a_T y}{c_T}\right). \end{aligned} \quad (16)$$

It follows that the conditional survival function of  $\tilde{R}$ ,  $\bar{G}_{s_j}$ , is given by:

$$\begin{aligned} \mathbb{P}(\tilde{R} > \tilde{s}_{j+1} \mid \tilde{R}_j > \tilde{s}_j) &= \bar{G}_{\tilde{s}_j}(\tilde{s}_{j+1} - \tilde{s}_j) \\ &= \int_0^\infty \exp\{-(\tilde{s}_{j+1} - \tilde{s}_j)y\} v_j(y) dy, \\ &\text{where } V_j \text{ is a r.v of distribution } \Gamma\left(\frac{1}{c_j}, \frac{a_j}{c_j}\right). \end{aligned}$$

with  $c_j = c_T$  and  $a_j = a_{j-1} + c_T(\tilde{s}_j - \tilde{s}_{j-1})$ .

Expression (16) gives room to an extension of the Pareto model. Indeed, we could consider distributions of  $\tilde{R}$  that share the same mixture form with a mixing variable  $W$  that possesses some common characteristics with the Gamma distributed r.v.  $V$ .

Similarly, the Weibull distribution  $W(\alpha, \beta)$  can also be written as the Laplace transform of a stable law of density  $g$  whenever  $\beta \leq 1$ . Indeed, it holds from Feller 1971[10]) (p. 450, Theorem 1) that:

$$\forall x \geq 0, \quad \exp\{-x^\beta\} = \int_0^\infty \exp(-xy) g(y) dy \quad (17)$$

where  $g$  is the density of an infinitely divisible probability distribution.

It follows, for  $s_j < s_{j+1}$

$$\begin{aligned} \mathbb{P}(\tilde{R} > \tilde{s}_{j+1} \mid \tilde{R}_j > \tilde{s}_j) &= \frac{\exp\{-(\tilde{s}_{j+1}/\alpha)^\beta\}}{\exp\{-(\tilde{s}_j/\alpha)^\beta\}} \\ &= \frac{\int_0^\infty \exp\{-(\tilde{s}_{j+1}/\alpha)y\} g(y) dy}{\int_0^\infty \exp\{-(\tilde{s}_j/\alpha)y\} g(y) dy} = \frac{\int_0^\infty \exp\{-(\tilde{s}_{j+1}/\alpha)y\} g(y) dy}{K(s_j)} \\ &= \frac{1}{K(s_j)} \int_0^\infty \exp\{-\tilde{s}_{j+1}u\} g_\alpha(u) du \\ &\quad \text{with } u = y/\alpha \text{ and } g_\alpha(u) = \alpha g(\alpha u) \end{aligned} \quad (18)$$

Thus an alternative modeling of  $\tilde{R}$  could consist in any distribution that can be written as a Laplace transform of a stable law of density  $w_{\alpha,\beta}$  defined on  $\mathbb{R}_+$  and parametrized by  $(\alpha, \beta)$ , that complies to the following condition: For any  $s > 0$ , the distribution function of the conditional distribution of  $\tilde{R}$  given  $\tilde{R} > s$  can be written as the Laplace transform of  $w_{\alpha,\beta}^{(\alpha,s)}(.)$  where

$$x > s, w_{\alpha,\beta}^{(\alpha,s)}(x) = \frac{\alpha w_{\alpha,\beta}(\alpha x)}{K(s)},$$

where  $K(.)$  is defined in (18).



## References

- [1] Balkema A. A. and De Haan L.: Residual Life Time at Great Age. *Ann. Prob.*, vol. 2 (5) pp. 762 - 804 (1974).
- [2] Beirlant J., Guillou A. and Dierckx G. and Fils-Villetard A.: Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*, vol. 10 (3) pp. 151 - 174 (2007).
- [3] De Haan L. and Rootzén H.: On the estimation of high quantiles. *J. Statist. Plann. Inference*, vol. 35 (1) pp. 1 - 13 (1993).
- [4] Dekkers A.L.M., Einmahl J. H. J. and De Haan L.: A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, vol. 17 (4) pp. 1833—1855 (1989).
- [5] De Valk C.: Approximation of high quantiles from intermediate quantiles. *Extremes*, vol. 4 pp. 661 - 684 (2016).
- [6] De Valk C. and Cai J.J: A high quantile estimator based on the log-Generalised Weibull tail limit. *Econometrics and Statistics*, vol. 6 pp. 107 - 128 (2018).
- [7] Dixon W. J. and Mood A. M.: A Method for Obtaining and Analyzing Sensitivity Data. *Journal of the American Statistical Association*, vol. 43 pp. 109 - 126 (1948).
- [8] Dixon W. J.: The Up-and-Down Method for Small Samples. *Journal of the American Statistical Association*, vol. 60 pp. 967 - 978 (1965).
- [9] Einmahl J.H.J, Fils-Villetard A. and Guillou A. Statistics of extremes under random censoring, *Bernoulli*, vol 14(1) pp. 207227 (2008)
- [10] Feller W.: An introduction to probability theory and its applications. *Wiley*, vol 2 (1971)
- [11] Fouchereau R.: Modélisation probabiliste des courbes S-N. *Université Paris Sud* (2014).
- [12] O’Quigley J. and Pepe M. and Fisher L.I.: Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer. *Biometrics*, vol. 46 pp. 33 - 48 (1990).

- [13] Pickands J.: Statistical Inference using extreme order statistics. *Ann. Prob.*, vol. 3 (1) pp. 119 - 131 (1975).
- [14] Thorin O.: On the infinite divisibility of the Pareto distribution. *Scand. Actuarial J.*, vol. 1 pp. 31 - 40 (1977).
- [15] Thyron P.: Les lois exponentielles composées. *Bulletin de l'Association Royale des Actuaires Belges*, vol. 62 pp. 35 - 44 (1964).
- [16] Weissman, I.: Estimation of parameters and large quantiles based on the  $k$  largest observations. *J. Amer. Statist. Assoc.*, vol. 73 pp. 812 - 815 (1978).
- [17] Worms J. and Worms R.: New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes*, vol. 17 pp. 337 - 358 (2014).