A Goodness-of-fit Test for the Distribution Tail

J. Diebolt[(1)], M. Garrido[(2)] and S. Girard[(3)]

[(1)] CNRS, Université de Marne-la-Vallée,

5 bd Descartes, 77454 Marne-la-Vallée Cedex 2, France.

[(2)] ENAC Dept. MI,

7, av. Belin, BP 4005, 31055 Toulouse cedex 04 , France.

[(3)] SMS/LMC-IMAG, Université Grenoble 1,

BP 53, 38041 Grenoble cedex 9, France.

Abstract

In order to check that a parametric model provides acceptable tail approximations, we present a test which compares the parametric estimate of an extreme upper quantile with its semiparametric estimate obtained by extreme value theory. To build this test, the sampling variations of these estimates are approximated through parametric bootstrap. Numerical Monte Carlo simulations explore the covering probability and power of the test. A real-data study illustrates these results.

Keywords : Goodness-of-fit Test, Extremes, Structural Reliability.

1 Introduction.

In many fields (financial analysis, climatology, decision making, structural reliability and safety engineering) special attention is devoted to the modelling of distribution tails, in particular upper tails, and the estimation of occurrence probabilities of rare events.

The present work is primarily motivated by questions related to extreme value estimation arising in the field of structural reliability. Typically, engineers working in this field have to evaluate failure probabilities of critical components, based on failure criteria and random input variables. The latter are modelled by probability distributions fitted to data. Such components can fail when some of the input variables take large values. Hence, in general failure probabilities are sensitive to the upper tail of some of the distributions used in the modelling process. Therefore, finding distributions that closely fit the largest values of input variables and provide reasonable extrapolations above often is of crucial importance.

Extreme Value Theory (see the monographs [14, 15]) has been developed to estimate probabilities of occurrence of rare events. It enables us to extrapolate the distribution tail behavior from the largest observed data (the extreme values of the sample). Unfortunately, in the small or moderate sample situations we have to deal with in industrial context (see Section 5) extreme models are helpless to make statistical inference. Since semiparametric Extreme Value models and methods take only a small part of the sample into account, it may remain too little information for estimation.

We are then led to use parametric models which take the information contained in the whole data set into account. These models have the advantage to make full use of available data in case of small data sets. In addition, these parametric models are easily interpretable for engineers: they have meaningful interpretations for each parameter, and use these parameters to make sensitivity analysis. Furthermore, in our context the rare events we want to model can be quantities entering complex

physical systems. Our modelisation then have to be easily re-introduced in these systems, wich would not be the case for semiparametric Extreme Values models. Only parametric models are available in current software tools, in particular those dedicated to risk evaluation for complex systems.

Usual goodness-of-fit tests help selecting some of these models. For exemple, we used here the Anderson-Darling and Cramér-von Mises tests [3]. However, such procedures essentially test the adequacy of each model to the central range of data. Dangers of extrapolating in the upper tail the results of such tests are detailed, *e.g.*, in [13] and [17].

Consider data issued from n independent identically distributed random variables with common distribution function F. Suppose that usual goodness-of-fit tests have not rejected the null hypothesis $\mathcal{H}_0$ that F belongs to one of those parametric models, $\{F_\theta : \theta \in \Theta\}$. The purpose of the GPD (Generalized Pareto Distribution) test is to check the fit of the tail of $F_{\widehat{\theta}_n}$, where $\widehat{\theta}_n$ is typically the maximum likelihood estimate of θ, to the largest data values and to make sure that this tail also provides reasonable extrapolations above the maximal observation. Anderson-Darling and associated tests can check the adequation to the largest sample values, but they give no indication concerning extrapolation quality outside the sample. Therefore, we essentially wish to test

$$\mathcal{H}_0 : F \in \{F_\theta : \theta \in \Theta\} \quad \text{against} \quad \mathcal{H}_1 : F \notin \{F_\theta : \theta \in \Theta\}$$

in the upper tail.

The principle of the GPD test is to compare two different estimates of some extreme upper quantile under $\mathcal{H}_0$. Let us recall that a $(1 - p_n)$ th upper quantile x_{p_n} is said to be extreme when $p_n \leq 1/n$ since it is generally larger than the maximal observation. The first estimate is the parametric estimate of the quantile, $\widehat{x}_{\text{param};\, n} = F^{-1}_{\widehat{\theta}_n}(1 - p_n)$. The second one is $\widehat{x}_{\text{GPD};\, n}$ a semiparametric estimate deduced form Extreme Value Theory.

The sampling variations of both parametric and semiparametric estimates of

x_{p_n} are approximated through parametric bootstrap, *i.e.* by resampling from $F_{\widehat{\theta}_n}$. Also, in order to deal with distributions whose parameter estimates and/or related quantile estimates are computationally demanding (e.g., mixtures), we introduce a simplified version of the parametric bootstrap version. In some situations, it is also possible to introduce an asymptotic version of the test based on the asymptotic distributions of the semiparametric estimate. Nevertheless, this version turns out to be disappointing in finite sample situations.

In Section 2, the main issues of Extreme Value Theory are recalled. Section 3 details the different versions of the test. Section 4 summarizes the results of intensive numerical simulation experiments. In each case, we recommend the best ranges of values for the number of excesses and order of the quantile, and give the corresponding values of the power. Along with various warnings, this will help practitioners wishing to check tails with the help of the GPD test. Finally, in Section 5 we apply both classical and GPD tests to a real-data set of size $n = 121$. These data were provided to us by the French electricity company *Electricité de France* (EDF). For both data sets, and even the small sample of size 24, our test helps in selecting a model.

2 Theoretical background.

Denote by $X_{n,n}$ the largest observation of the sample $\{X_1, \ldots, X_n\}$. The common distribution function F is said to belong to a *domain of attraction (for the maximum)* if we can find two deterministic sequences t_n and $s_n > 0$ such that $(X_{n,n} - t_n)/s_n$ converges in distribution as $n \to \infty$ to some nondegenerate random variable. Conditions ensuring that F belongs to a domain of attraction can be found, e.g., in [14, 15]. Up to a scale parameter, the only possible limiting distribution functions are of the form

$$H_\gamma(x) = \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right),$$

with the notation $u_+ = \max(u, 0)$, and for some real number γ, which reduces to $H_0(x) = \exp(-\exp(-x))$ when $\gamma = 0$. In the latter case, we say that F belongs to *Gumbel's domain of attraction*, DA(Gumbel).

Let $\bar{F} = 1 - F$ be the survival function associated to F. An extreme upper quantile is a $(1 - p_n)$ th quantile x_{p_n} of F, *i.e.* such that $\bar{F}(x_{p_n}) = p_n$ with $p_n \leq 1/n$. It is usually larger than the maximal observation. Estimation of extreme quantiles requires specific methods. The POT (Peaks Over Threshold) method has become the cornerstone of tail estimation and statistical inference for extreme quantiles. It relies on an approximation of the distribution of excesses over a given threshold ([23]). More precisely, let us introduce deterministic thresholds u_n such that $\bar{F}(u_n) = k_n/n$ with

$$1 \leq k_n \leq n, \quad k_n \to \infty, \quad k_n/n \to 0 \text{ as } n \to \infty. \tag{1}$$

The excesses above the threshold u_n of the n-sample $\{X_1, \ldots, X_n\}$ from F are defined on the basis of the $X_i > u_n$'s by $Y_i = X_i - u_n$. The common survival function of the excesses is

$$\bar{F}_{u_n}(y) = P\left(X - u_n > y \,|X > u_n\right) \;=\; \bar{F}(u_n + y)/\bar{F}(u_n), \quad y \geq 0\,. \tag{2}$$

Pickands' theorem [22] implies that if F belongs to a domain of attraction, then under (1), $\bar{F}_{u_n}$ can be approximated by a generalized Pareto distribution (GPD) with survival distribution function given by

$$\bar{G}_{\gamma,\,\sigma}(x) \;=\; \Big(1 \;+\; \frac{x\gamma}{\sigma}\Big)^{-1/\gamma} \quad \text{for} \begin{cases} x \in \mathbb{R}_+ \text{ if } \gamma > 0. \\ x \in [0,\, -\sigma/\gamma[\text{ if } \gamma < 0, \end{cases}$$

if $\gamma \neq 0$, and

$$\bar{G}_{0,\,\sigma}(x) \;=\; \exp\left(-x/\sigma\right) \quad \text{ for } x \in \mathbb{R}_+, \tag{3}$$

in the case $\gamma = 0$. As a consequence, the extreme quantile x_{p_n} can be approximated by the deterministic term

$$x_{\mathrm{GPD};\,n} = u_n + \frac{\sigma(u_n)}{\gamma}\left[\left(\frac{k_n}{np_n}\right)^{\gamma} - 1\right], \tag{4}$$

where $\sigma(u_n)$ and γ are respectively the scale and shape parameters of the GPD. Then, the POT method consists of estimating these two unknown parameters and replacing u_n by its empirical counterpart $X_{n-k_n,n}$. This leads to:

$$\widehat{x}_{\mathrm{GPD};\,n} = X_{n-k_n,n} + \frac{\widehat{\sigma}_n}{\widehat{\gamma}_n}\left[\left(\frac{k_n}{np_n}\right)^{\widehat{\gamma}_n} - 1\right], \tag{5}$$

where $\widehat{\sigma}_n$ and $\widehat{\gamma}_n$ are some estimates of respectively the scale and shape parameters of the GPD. There exists several methods to estimate these parameters such as maximum likelihood method [23], Bayesian methods [8], weighted moments method [19], Hill's method [18], and the generalisation of Hill's method proposed by [5]. Let us note that (5) relates the estimator $X_{n-k_n,n}$ of a quantile u_n within the data range to estimators $\widehat{x}_{\mathrm{GPD};\,n}$ of quantiles x_{p_n} larger than the maximum value of the sample.

In the particular case where F belongs to DA(Gumbel), $\gamma = 0$ and the GPD reduces to an exponential distribution (3) with scale parameter $\sigma(u_n)$. Thus, approximation (4) can then be rewritten as

$$x_{\mathrm{ET};\,n} = u_n + \sigma(u_n)\ln\left(\frac{k_n}{np_n}\right).$$

The corresponding estimator [1] is

$$\widehat{x}_{\mathrm{ET};\,n} = X_{n-k_n,n} + \widehat{\sigma}_n \ln\left(\frac{k_n}{np_n}\right), \tag{6}$$

where $\widehat{\sigma}_n$ is the estimation of $\sigma(u_n)$, the scale parameter of the asymptotic exponential distribution, given by the empirical mean of the k_n excesses: $\widehat{\sigma}_n = k_n^{-1}\sum_{i=1}^{k_n} Y_i$. Finally, we define the parametric estimate under $\mathcal{H}_0$ of x_{p_n} by

$$\widehat{x}_{\mathrm{param};\,n} = F^{-1}_{\widehat{\theta}_n}(1-p_n)\,. \tag{7}$$

3 The GPD test.

The principle of the GPD test is to compute under $\mathcal{H}_0$ a $(1-\alpha)$-confidence interval $CI_{\alpha;n}$ for the difference $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n}$ of semiparametric and parametric estimates of an extreme upper quantile. The test rejects $\mathcal{H}_0$ hypothesis at level α

when $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n}$ does not belong to $CI_{\alpha;n}$. Three versions of the test can be derived, depending on the method used to compute $CI_{\alpha;n}$.

3.1 Full parametric bootstrap version of the GPD test.

In this section, a parametric bootstrap evaluation of the distribution of $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n}$ is used. Since under $\mathcal{H}_0$, F is approximated by $F_{\widehat{\theta}_n}$, first we independently generate N independent samples of size n from $F_{\widehat{\theta}_n}$. Up to the approximation of F by $F_{\widehat{\theta}_n}$ (under $\mathcal{H}_0$), each of these N samples is a replication of the initial one. Then, for each of these samples, we compute the estimates $\widehat{x}^*_{\mathrm{GPD};\,n}$ and $\widehat{x}^*_{\mathrm{param};\,n} = F^{-1}_{\widehat{\theta}^*_n}(1-p_n)$. This provides N independent values of $\delta^*_n = \widehat{x}^*_{\mathrm{GPD};\,n} - \widehat{x}^*_{\mathrm{param};\,n}$. Let us note $\delta^*_{j,N}$, $j = 1, \ldots, N$ the corresponding ordered sample. Sorting these values and eliminating the $[N\alpha/2]$ largest and smallest ones, we deduce a Monte Carlo based $(1-\alpha)$-confidence interval: $FPB.CI_{\alpha;n} = [\delta^*_{[N\alpha/2],N}, \delta^*_{[N(1-\alpha/2)],N}]$. This leads to the

FPB.GPD Test (Full Parametric Boostrap GPD test):
Reject $\mathcal{H}_0$ when $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n} \notin FPB.CI_{\alpha;n}$.

In the particular case where F_θ belongs to DA(Gumbel), since $\gamma = 0$, no estimation of this parameter is needed. Thus $\widehat{x}_{\mathrm{GPD};\,n}$ and $\widehat{x}^*_{\mathrm{GPD};\,n}$ are replaced by $\widehat{x}_{\mathrm{ET};\,n}$ and $\widehat{x}^*_{\mathrm{ET};\,n}$ in the previous procedure. In the following, this particular version of the test is referred to as the FPB.ET test. Let us emphasize that the computation of $\widehat{x}_{\mathrm{ET};\,n}$ only requires the estimation of one parameter from the largest observations (see Section 2) instead of two in the general case for $\widehat{x}_{\mathrm{GPD};\,n}$. This is an important point in the case of small data sets, that is to say when the information contained in the extreme values of the sample is very poor.

3.2 Simplified parametric bootstrap version of the GPD test.

When computations of F_θ^{-1} and $\widehat{\theta}_n$ are heavy (e.g., for mixtures) and if n is large enough, then it is possible to neglect the sampling fluctuations of $\widehat{x}_{\mathrm{param};\,n}$, as k_n is much smaller than n. We then construct, similarly to the previous subsection, a Monte Carlo based $(1-\alpha)$-confidence interval for $\widehat{x}_{\mathrm{GPD};\,n}$, $SPB.CI_{\alpha;n}$ from the N independent values of $\widehat{x}^*_{\mathrm{GPD};\,n}$, leading to the

SPB.GPD Test (Simplified Parametric Bootstrap GPD test):
Reject $\mathcal{H}_0$ when $\widehat{x}_{\mathrm{GPD};\,n} \notin SPB.CI_{\alpha;n}$.

When F_θ belongs to DA(Gumbel), GPD estimates can be replaced by ET estimates as in Subsection 3.1 to obtain the so-called SPB.ET test.

3.3 Asymptotic version of the GPD test.

The computational cost of the test can be reduced even more by approximating the sampling fluctuations of $\widehat{x}_{\mathrm{GPD};\,n} - x_{\mathrm{GPD};\,n}$ by its asymptotic distribution. This yields an $(1-\alpha)$-confidence interval $A.CI_{\alpha;\,n}$ for $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n}$. Thus, the

A-GPD test (Asymptotic GPD test):
Reject $\mathcal{H}_0$ when $\widehat{x}_{\mathrm{GPD};\,n} - \widehat{x}_{\mathrm{param};\,n} \notin A.CI_{\alpha;n}$.

Of course, the asymptotic distribution of $\widehat{x}_{\mathrm{GPD};\,n} - x_{\mathrm{GPD};\,n}$, and thus $A.CI_{\alpha;n}$, strongly depend on the method used to estimate γ and $\sigma(u_n)$. For instance, when F belongs to DA(Gumbel), [11] takes advantage of the asymptotic distribution of $\widehat{x}_{\mathrm{ET};\,n} - x_{\mathrm{ET};\,n}$ (established in [12]) to propose the asymptotic $(1-\alpha)$-confidence interval,

$$A.CI_{\alpha;\,n} = \left[b_n \pm \hat{\sigma}_n \ln\left(\frac{k_n}{np_n}\right) k_n^{-1/2} z_\alpha \right], \tag{8}$$

where $P(|\xi| > z_\alpha) = \alpha$ with $\xi \sim N(0,1)$ and b_n is a bias correction term which is

not detailed here.

However, this asymptotic version is not satisfactory in the general case. For instance, the test based on (8) reveals a rather low power [10] because the normal approximation to the distribution of $\widehat{x}_{\mathrm{ET};\,n} - x_{\mathrm{ET};\,n}$ is not appropriate for realistic values of n and k_n.

4 Simulations.

Our purpose in the present section is to experiment performance of the GPD test through Monte-Carlo simulations. Let us note that several intensive numerical studies about the estimation of extreme quantiles or the parameters of generalized Pareto distributions have already been made (see, *e.g.*, [19, 20, 21, 2, 9]).

4.1 Experimental design.

We focus on the SPB.GPD version which offers the best compromise in terms of power and computational cost. The performance of the test clearly depends on the choice of the estimation method for $\hat{\sigma}_n$ and $\hat{\gamma}_n$ and on the choice of the parameters k_n and p_n.

Four estimators of $\hat{\sigma}_n$ and $\hat{\gamma}_n$ have been considered: maximum likelihood (ML) method [23], weighted moments (WM) method [19] Hill's method [18], and the generalisation of Hill's method proposed by [5]. We focus in ML and WM estimators since they benefit of scale and location invariance. It is thus possible to prove (see Appendix) that the result of the GPD test built on these methods does not depend on the location and scale parameters of the data. In the following we limit ourselves to WM estimates, since ML estimates revealed a poor computational behaviour (convergence problems ...) [16, 4].

The choice of the parameters involves two steps. First, for each of the previous methods, we have selected the numbers k_n of excesses to be used to guarantee that the actual simulated levels are reasonably close to the theoretical one $(1 - \alpha)$. Let

us note that the choice of the parameter k_n is a recurrent problem in practical extreme value analysis, see for instance [6]. Then, we have studied the power of the booststrap versions of the test in a number of typical cases and try to make precise the values of k_n and p_n achieving the best powers.

As a conclusion, we have conducted Monte Carlo simulations by varying the following parameters:

- Sample size: $n \in \{100, 200, 500\}$.
- Number of excesses: $k_n \in \{5, 10, 20, 40\}$.
- Order of the extreme quantile: $p_n \in \{10^{-2}, 5.10^{-3}, 10^{-3}, 5.10^{-4}\}$.
- Simulation distributions: Normal $N(0,1)$, Lognormal $LN(0,1)$, Exponential $Exp(1/2)$, Gamma $\Gamma(3,3)$, Weibull $W(1/2,2)$, Khi2 χ^2_4, Pareto $Pa(2,3)$, Student $T(10)$, GPD$(1/5,5)$, Uniform $U[0,1]$.
- Null hypothesis: Normal, Lognormal, Exponential, Gamma, Weibull, Khi2, Pareto, Student, GPD, Uniform.

The number of bootstrap replications is fixed to $N = 200$ and for each simulated data set, the test is applied 100 times to estimate its level and power. Finally, the complete experiment involves $3 \times 4 \times 4 \times 10 \times 10 \times 200 \times 100 = 96,000,000$ basic statistical operations. In order to make these experiments easier, a software [7] dedicated to that task is available at

`http://www.inrialpes.fr/is2/pub/software/EXTREMES/accueil.html`.

4.2 Experimental results.

The results are collected in Table 1. In the first column, the distribution considered in the null hypothesis $\mathcal{H}_0$ is recalled. In the second one, the sample size n is given. The next two columns present the number of excesses k_n and the order of quantile p_n to use preferably in each situation. Column 5 provides the range of the power we can expect with such choices. The last column presents the exceptions to this

rule. For the simulated distributions quoted here, the wrong null hypothesis has been usually accepted with the advised parameters k_n and p_n. In some cases, we could reject the null hypothesis with different values of k_n outside the test set $\{5, 10, 20, 40\}$. Such situations are highlighted by an asterisk. For instance, when testing Gaussian hypothesis on samples of size $n = 500$, one can choose $k_n \in [20, 40]$ and $p_n \in [5.10^{-4}, 10^{-3}]$. The resulting power is at least 92% except if the true data distribution is Weibull.

In this latter case, the GPD test cannot discriminate between Gaussian and Weibull with shape parameter $\beta = 2$ since their survival distribution functions are very close in the upper tails. They essentially behave as $\exp(-x^2)$ when x is large. Similar observations can be made to explain the confusions between lognormal/gamma, exponential/lognormal distributions (when $n \leq 200$) and gamma/weibull distributions. Let us also note that, since the Student $T(m)$ distribution converges to a standard Gaussian distribution when $m \to \infty$, the Student assumption is accepted on Gaussian data sets. Finally, observing that the GPD distribution includes the exponential distribution, the confusion between GPD and lognormal distribution is a consequence of the confusion between exponential and lognormal distributions.

5 Real-data set.

Our real-data set consists of amounts of chromium (Cr) measured on $n = 121$ steel blades. These steel blades are samples from steel used in sensible components of EDF nuclear plants. The amounts of chromium, as well as carbon (C), manganese (Mn), molybden (Mo), nickel (Ni) and silicon (Si), correspond to residual dirtiness that may deteriorate the steel quality and therefore alter reliability characteristics of the components. When the amount of chromium (or other chemical substance) increases, the steel quality is deteriorated and failure risks become larger. [6] found that the amounts of C, Mo, Ni and Si measured on the same steel blades contained

outliers. Further investigations with the data collectors led them to remove one to three largest observations for the C, Mo, Ni and Si data. However, they found no evidence of the existence of outliers for the Cr and Mn data.

The Cramér-von Mises test did not reject the normal, lognormal and gamma models at the level $\alpha = 5\%$. The Anderson-Darling test did not reject the lognormal and gamma models at the 5%-level. Figure 1 depicts the histogram and densities of the estimated normal, lognormal and gamma distributions (which are accepted by the Cramér-von Mises test) and Figure 2 depicts the corresponding failure rate functions.

Let us note that the three remaining models are in DA(Gumbel). Besides, usual goodness-of-fit tests do not reject the exponentiality of excess distributions corresponding to proper values of $k_n \in \{2, \ldots, 40\}$. Thus, we focus on the FPB.ET and SPB.ET versions of the test. First, the SPB.ET test is used with $k_n = 5$ and $p_n = 0.01$, according to the advised value of Table 1. The test rejected tail normal and gamma models and did not reject tail lognormal distributions (see Table 2) at the 5%-level. However, in each case, the test statistic $\widehat{x}_{\mathrm{ET};\,n}$ is very close to the boundary of the confidence interval. To confirm these results, we used the FPB.ET test with the same parameters. The conclusions remain the same (see Table 2) and the test statistic $\widehat{x}_{\mathrm{ET};\,n} - \widehat{x}_{\mathrm{param};\,n}$ is clearly outside the confidence interval in the normal and gamma cases.

Therefore, as far as there is no unsuspected outlier in these data, the lognormal distribution is appropriate to model both the central part and the tail of the data.

6 Conclusion.

Motivated by questions arising in the field of structural reliability, we have introduced the GPD test to check the goodness-of-fit of distribution tails. The GPD test can be used in conjunction with usual goodness-of-fit tests, in order to also test the adequacy of models to the central range of data. We approximate the sampling fluc-

tuations of the difference between GPD and parametric estimates of some extreme quantile through resampling from an estimate of the model distribution under the null hypothesis.

In order to deal with distributions whose parameter and quantile estimates are difficult to compute, we have introduced a simplified version of the parametric bootstrap version of our test.

We have made tables displaying the values for k_n that we recommend for $n =$ 100, 200 and 500; normal, lognormal, exponential, Weibull, gamma, Khi2, Pareto, Student, GPD and uniform distributions; quantiles of order $1 - p_n$ with $p_n = 10^{-2}$, 5.10^{-3}, 10^{-3} and 5.10^{-4}; and significance level $\alpha = 5\%$. We have also indicated the values of k_n and p_n achieving the best experimental powers in a number of typical situations. Our tables, examples, software and recommendations will help practitioners using the GPD test. In cases where central range goodness-of-fit tests reject all models but one, the GPD test performed at several values of k_n and p_n gives insight into tail adequacy of the only plausible distribution.

We have treated a real-data set issued from structural reliability questions arising in nuclear industry. It turns out that GPD tests help deciding which model is best fitted.

Acknowledgements.

This work has been partially supported by INRIA and Électricité de France. The authors are grateful to V. Durbec and B. Villain for stimulating and helpful discussions and for providing the real-data set.

References

[1] L. Breiman, C.J. Stone, and C. Kooperberg. Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37:127–149, 1990.

[2] S. G. Coles and M. J. Dixon. Likelihood-based inference for extreme value models. *Extremes*, 2:5–23, 1999.

[3] R.B. D'Agostino and M.A. Stephens. *Goodness-of-fit Techniques*, volume 68 of *Statistics textbooks and monographs*. 1986.

[4] A.C. Davison and R.L. Smith. Models for exceedances over high tresholds. *J. R. Statist. Soc. B*, 52(3):393–442, 1990.

[5] A.L.M. Dekkers, J.H.J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

[6] J. Diebolt, V. Durbec, M. A. El Aroui, and B. Villain. Estimation of extreme quantiles: empirical tools for method assessment and comparison. *International Journal of Reliability, Quality and Safety Engineering*, 7(1):75–94, 2000.

[7] J. Diebolt, J. Ecarnot, M. Garrido, S. Girard, and D. Lagrange. Le logiciel Extremes, un outil pour l'étude des queues de distribution. *La revue de Modulad*, 30:55–60, 2003.

[8] J. Diebolt, M. El-Aroui, M. Garrido, and S. Girard. Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling. *Extremes*, 8:57–78, 2005.

[9] J. Diebolt and M. A. El Aroui. On the use of the peaks over thresholds method for estimating out-of-sample quantiles. *Computational Statistics and Data Analysis*, 39(4):453–475, 2002.

[10] J. Diebolt, M. Garrido, and S. Girard. Le test ET : test d'adéquation d'un modèle central à une queue de distribution. Technical report RR-4170, INRIA, 2001.

[11] J. Diebolt, M. Garrido, and S. Girard. Asymptotic normality of the ET method for extreme quantile estimation. application to the ET test. *Comptes-Rendus de l'Académie des Sciences*, t. 337, Série I:213–218, 2003.

[12] J. Diebolt and S. Girard. A Note on the asymptotic normality of the ET method for extreme quantile estimation. *Statistics and Probability Letters*, 62(4):397–406, 2003.

[13] O. Ditlevsen. Distribution arbitrariness in Structural Reliability. In Shinozuka Schuller and Yao, editors, *Structural Safety and Reliability*, Proc. of ICOSSAR'93: 6th International Conference on Structural Safety and Reliability, pages 1241–1247, Balkema, Rotterdam, 1994.

[14] P. Embrechts, C. Klüppelberg, and Mikosh T. *Modelling Extremal Events*, volume 33 of *Applications of Mathematics*. Springer-Verlag, 1997.

[15] J. Galambos. *The asymptotic theory of extreme order statistics*. R.E. Krieger publishing compagny, 1987.

[16] S.D. Grimshaw. Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics*, 35:185–191, 1993.

[17] G. Hahn and W. Meeker. Pitfalls and practical considerations in product life analysis, part 1: Basic concepts and dangers of extrapolation. *Journal of Quality Technology*, 14:144–152, 1982.

[18] B.M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[19] J. Hosking and J. Wallis. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 3:339–349, 1987.

[20] R. Hughey. A survey and comparison of methods for estimating extreme right tail-area quantiles. *Communication in statistics - Theory and Methods*, 20(4):1463–1496, 1991.

[21] Y. Moon, U. Lall, and K. Bosworth. Comparison of tail probability estimators for flood frequency analysis. *Journal of Hydrology*, 151(2-4):343–363, 1993.

[22] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131, 1975.

[23] R.L. Smith. Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207, 1987.

null hyp.	n	k_n	p_n	power	no result for true
Normal	100	5	10^{-2}	$[60, 100]$	Weibull
	200	5, 10	5.10^{-3}, 10^{-3}	$[67, 100]$	Weibull
	500	20, 40	10^{-3}, 5.10^{-4}	$[92, 100]$	Weibull
Lognormal	100	5, 10	10^{-2}	$[71, 100]$	Gamma, GPD
	200	20, 40	5.10^{-3}	$[77, 100]$	Gamma
	500	20, 40	10^{-3}, 5.10^{-4}	$[85, 100]$	
Exponential	100	40	10^{-2}, 10^{-3}	$[60, 100]$	Lognormal
	200	40	5.10^{-3}, 10^{-3}	$[91, 100]$	Lognormal
	500	40	10^{-3}, 5.10^{-4}	$[85, 100]$	
Gamma	100	5	10^{-2}, 10^{-3}	$[62, 100]$	Weibull
	200	5, 10, 20, 40	10^{-3}	$[77, 100]$	Weibull
	500	20, 40	10^{-3}, 5.10^{-4}	$[90, 100]$	
Weibull	100	10	10^{-3}	$[52, 100]$	Gamma
	200	5, 10, 20	10^{-3}, 5.10^{-3}	$[71, 100]$	Gamma
	500	40	10^{-3}	$[97, 100]$	Gamma
Khi2	100	10	10^{-2}, 10^{-3}	$[58, 100]$	Lognormal
	200	20	10^{-3}	$[67, 100]$	
	500	40	10^{-3}	$[85, 100]$	Weibull*
Pareto	100	5	10^{-2}, 10^{-3}	$[46, 99]$	Weibull
	200	10	10^{-3}, 5.10^{-3}	$[60, 100]$	
	500	10, 20	10^{-3}, 5.10^{-4}	$[90, 100]$	
Student	100	10	10^{-3}	$[60, 100]$	Normal, Exponential
	200	10	10^{-3}	$[51, 100]$	Normal, Exponential
	500	10	10^{-3}, 5.10^{-4}	$[85, 100]$	Normal, Exponential
GPD	100	40	10^{-2}, 10^{-3}	$[57, 100]$	Lognormal
	200	40	10^{-3}, 5.10^{-3}	$[81, 100]$	Lognormal
	500	20, 40	10^{-3}, 5.10^{-4}	$[93, 100]$	Lognormal, Gamma*
Uniform	100	20	10^{-2}	$[77, 100]$	
	200	40	10^{-3}, 5.10^{-3}	$[90, 100]$	
	500	20, 40	10^{-3}, 5.10^{-4}	100	

Table 1: Best values of k_n and p_n to use for different null hypothesis.

* for these true distributions, k_n has to be greater to obtain a satisfactory power (for instance $k_n = 80$).

SPB.ET test					
distribution	k_n	p_n	result	confidence interval	$\widehat{x}_{\mathrm{ET};\,n}$
normal	5	10^{-2}	rejected	[21.984 , 22.995]	23.083
lognormal	5	10^{-2}	**accepted**	[21.937 , 23.083]	23.083
gamma	5	10^{-2}	rejected	[21.910 , 22.997]	23.083
FPB.ET test					
distribution	k_n	p_n	result	confidence interval	$\widehat{x}_{\mathrm{ET};\,n} - \widehat{x}_{\mathrm{param};\,n}$
normal	5	10^{-2}	rejected	[-0.420 , 0.420]	0.484
lognormal	5	10^{-2}	**accepted**	[-0.422 , 0.488]	0.421
gamma	5	10^{-2}	rejected	[-0.373 , 0.352]	0.453

Table 2: Results of the GPD test for amounts of chromium measured on $n = 121$ steel blades, with $\alpha = 0.05$.

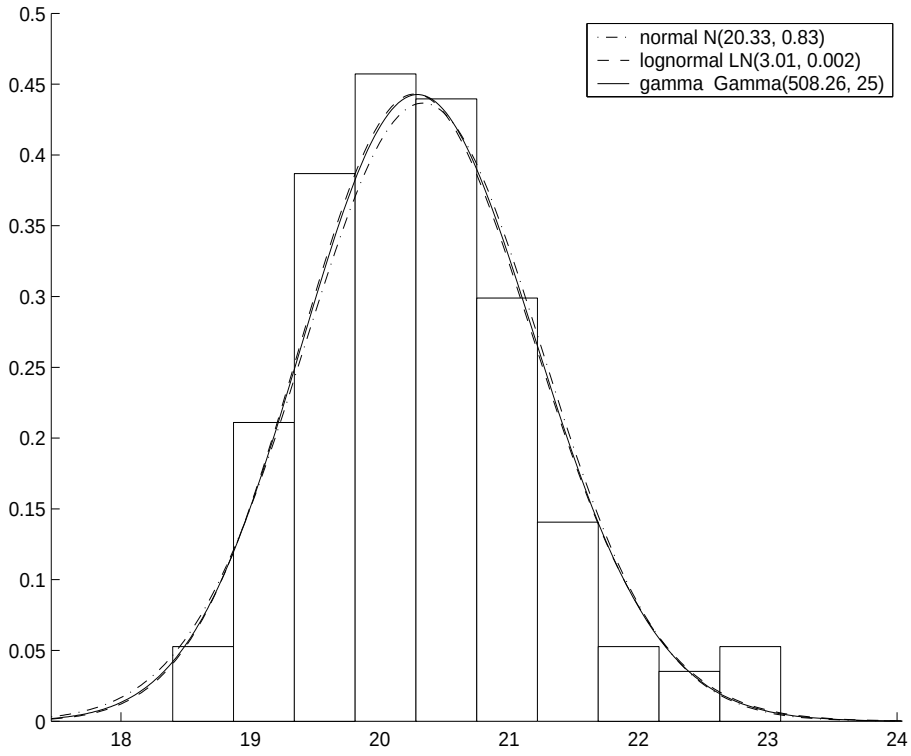

Figure 1: Histogram and densities of estimated distributions for amounts of chromium measured on $n = 121$ steel blades.

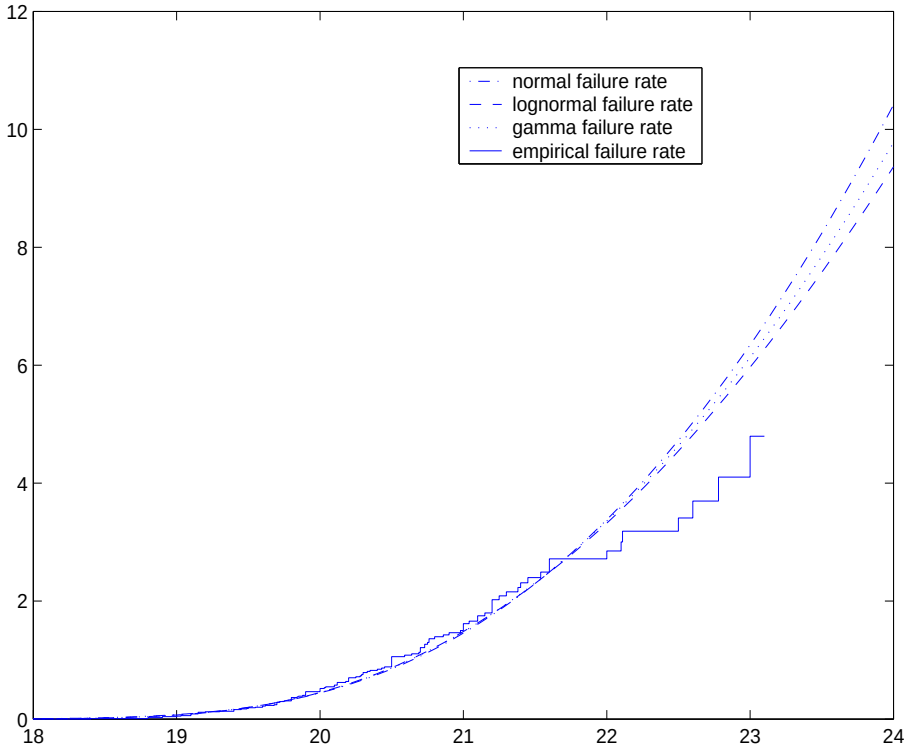

Figure 2: Failure rate functions (empirical and estimated) for amounts of chromium measured on $n = 121$ steel blades.