Far East Journal of Theoretical Statistics

https://pphmjopenaccess.com
https://doi.org/10.17654/0972086326002
Volume 70, Number 1, March 2026, Pages 17-37
P-ISSN: 0972-0863
E-ISSN: 3048-7358

PROBABILISTIC APPROACH TO RISKS ASSOCIATED WITH TRUNCATED DATA

Mathieu Tiene[1], Wendkouni Yaméogo[1], Bakary Compaoré[2] and Diakarya Barro[1]

[1]Université Thomas Sankara
Burkina Faso
e-mail: leprobabilistetiene@gmail.com

[2]LANIBIO
Université Joseph Ki Zerbo
Burkina Faso

Abstract

This article presents a contribution to probabilistic modeling risks linked to truncated data. We first model that the sample density of the last random sample is given by the truncation of the distribution function. Then we show that our estimator is unbiased, and also the inverse of the variance measures whatever the sample size, the precision of the estimator and if the distribution function is log-concave, then the precision is shown to be much lower than the value to be estimated.

Received: August 20, 2025; Revised: October 15, 2025; Accepted: October 30, 2025
2020 Mathematics Subject Classification: 93E03, 60-XX, 92D30, 62G32.
Keywords and phrases: probabilistic modeling, truncated data, random sample.

How to cite this article: Mathieu Tiene, Wendkouni Yaméogo, Bakary Compaoré and Diakarya Barro, Probabilistic approach to risks associated with truncated data, Far East Journal of Theoretical Statistics 70(1) (2026), 17-37. https://doi.org/10.17654/0972086326002

Published Online: January 3, 2026

1. Introduction

Censored data is not the only type of incomplete data. The other classic case of incomplete data is that of so-called truncated data.

Several authors have commented on censored or truncated data.

Reference [5] examines the connections between climate change and global inequalities. It first analyzes differences in exposure and vulnerability to the impacts of climate change among countries and individuals. It then investigates inequalities in contributions to greenhouse gas emissions. Finally, it demonstrates that understanding these inequalities helps to ensure a more equitable allocation of the actions required to mitigate climate change.

Reference [10] proposes a new method to estimate extreme precipitation at points where we do not have observations. Reference [1] approximates the extreme value index estimator of a heavy-tailed distribution under random censoring. Reference [7] gives a novel extension of the Lomax distribution, aiming to enhance its applicability in various contexts and emphasizes a pragmatic approach in deriving mathematical properties of the new distribution, prioritizing its practical implications.

Thus, the phenomenon of truncation is very different from censorship. Truncation, for its part, eliminates from the study, part of X_j. During a practical study on lifespans, it is not uncommon for the variable of interest X to be unobservable when it is lower than a random threshold Y, which will have the consequence that the analysis can only relate to the conditional law of X knowing $X > Y$.

The problem of significant, incomplete or erroneous data is very vast and has attracted a lot of interest from statisticians in recent years. Reference [6] gives a bias-reduced tail estimate for censored Pareto distributions, [2] proposes an adaptation of the method to an ungauged site while attempting to retain its strengths, namely, a spatial and probabilistic structuring of precipitation conditioned by weather types, and a cross-referencing of rainfall and basin saturation hazards using stochastic simulation.

Reference [3] analyzes the spatio-temporal variability of precipitation in the upper part of the Senegal River basin using data from ten reference stations. The stations were selected on the basis of data quality and their proximity to the catchment area. Homogeneity tests applied to the annual time series reveal breakpoints for all stations, eight of which occur between 1960 and 1970, with rainfall deficits ranging from 12% to 24%. At the monthly scale, a significant decrease in precipitation is observed between the two study periods. At the daily scale, heavy rainfall events (> 40mm) become less frequent from the breakpoint years onward. Finally, the southern part of the basin is the most contrasted area, experiencing both the largest surpluses during wet periods and the largest deficits during dry periods.

The attitude towards this type of data has long been either to eliminate them or to minimize the bad impact they could have on statistical procedures adapted to complete data. In the field of survival times, data are often incomplete due to two distinct phenomena: truncation and censoring, so our topic will focus on truncation.

Data play an important role in the statistical analysis of astronomical observations as well as in survival analysis [8]. The motivating example for this paper concerns a set of quasar measurements in which there is a double truncation. That is, quasars are observed only if their luminosity occurs within a certain finite interval, bounded at both ends, with the interval varying for different observations. Nonparametric methods for testing and estimating doubly truncated data are developed. These methods extend some known techniques for data that are truncated only on one side, in particular the Lynden-Bell estimator and the truncated version of Kendall's tau statistic. They derive asymptotic results and illustrate in the simulation study the performance and robustness of this estimator for both small and large sample sizes.

Let $X_1, ..., X_n$ be n copies of independent and identically distributed random variable X, with a common cumulative distribution function F assumed to be heavy-tailed. In other words, the distribution tail $\overline{F} = 1 - F$

is regularly varying, with index $(-\alpha_1)$ notation: $\overline{F} \in \mathcal{R}v_{(-\alpha_1)}$. That is,

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha_1}, \text{ for any } x > 0,$$

where $\alpha_1 > 0$ is called the *shape parameter*, *tail index* or *extreme value index*. It plays a very crucial role in the analysis of extremes as it governs the thickness of the distribution tails. Reference [9] shows that it would be important to control the number of observations in a given region, while the oscillation of the log-quantile function is small. More precisely, we have also shown that the extreme quantile is more stable, the further we are from the frontier. We now show that the random variable is observed if it belongs to the interval of the non-negative and absolutely continuous variable. Here τ is an integral functional of the bivariate survival function. We construct a natural estimator via the von Mises functional approach. This does not necessarily yield a consistent estimator since tail region information on the survival curve may not be identifiable due to right censoring. To assess the magnitude of the inconsistency, some estimable bounds on τ are proposed. It is shown that estimates of the bounds shrink to provide consistency if the largest observations on both marginal conditions are uncensored and satisfy certain regularity conditions,

$$\hat{\tau} = \binom{n}{2} \sum\nolimits_{1\le i,\, j\le n} a_{ij} b_{ij},$$

where $a_{ij}b_{ij} = 1$, if the $(i,\ j)$ pair is concordant and is –1 if discordant.

Thus,

$$\Gamma = \frac{\sum_{1\le i,\, j\le n} a_{ij} b_{ij}}{\left(\sum_{i,\, j} a_{ij}^2 \sum_{i,\, j} b_{ij}^2\right)^{1/2}}.$$

The objective of this article is to show that when we study truncated observations whose smallest limit is the first sample, with a zero probability density, then the frequency of observations is not zero.

Also, we illustrate that when the estimator is unbiased, the inverse of the variance measures, whatever the size of the sample be, the precision of the estimator. Also, if the distribution function is log-concave, then the precision is shown to be much lower than the value to be estimated.

The rest of the article is organized as follows: In Section 2, we recall the concepts essential to the study, in Section 3, we present the main results obtained and in Section 4, we give a conclusion and a discussion.

2. Preliminary

In this section, we bring together important definitions and theorems on modeling truncated data, the Fisher information that is necessary for our approach. We direct the reader to the references on detailed introductions [4, 7].

Another situation where incomplete data appears is truncated data.

An observation is said to be *truncated* if it is conditional on another event. The lifetime variable Y is said to be *truncated* if Y is only observable under a certain condition dependent on the value of Y.

There is another type of truncation:

Left truncation. There is a left truncation, when the variable of interest X is only observable if it is greater than T.

T is then the left truncation random variable, i.e.,

$$X \text{ is only observed if } X > T.$$

Right truncation. There is a right truncation, when X is only observable if it is less than T. T is then the right truncation random variable:

$$X \text{ is only observed if } X < T.$$

Interval truncation. When a duration is truncated on the right and left, we say that it is *interval truncated*.

Remark 1. In general, truncation arises when observation of the variable of interest X is restricted to those realizations for which an event B occurs. One can also distinguish the case of double truncation.

So, $\nabla F(\theta)$ designates the gradient of $F : \Theta \to \mathbb{R}$ evaluated at $\theta \in \Theta$. By convention, the gradient of a function is only calculated at θ if the function is of class C^1 over a neighborhood of θ. Furthermore, $\mathbb{V}_\theta$ denotes the variance matrix.

Definition 2.1. Suppose that Θ is open and $\nabla \ln L_n(x\,;\,\theta) \in \mathbb{L}^2(P_\theta)$.

For each $\theta \in \Theta$:

$$I_n(\theta) = \mathbb{V}_\theta(\nabla \ln L_n(x;\,\theta))$$

$$= \left(\operatorname{cov}_\theta\left(\frac{\partial}{\partial \theta_i} \ln L_n(x;\,\theta),\, \frac{\partial}{\partial \theta_j} \ln L_n(x;\,\theta) \right) \right)_{i,\,j=1,\,\ldots,\,d}.$$

Fisher information is a function with value in the set of positive matrices. As a measure of Kullback-Leibler information curvature, it specifies the model's power of discrimination between two values close to the model parameter. If $d = 1$, a large value for $I_n(\theta)$ reflects a significant variation in the nature of the model's probabilities in the vicinity of P_θ, hence makes it easy to determine the true value of the unknown parameter.

Conversely, if $I_n(\theta)$ is small, then the law is very steep and we are led to seek the maximum likelihood in a very vast region.

In our example, the statistical model $(\{0;\,1\}^n,\, \{\mathfrak{B}(\theta)^{\oplus n}\}_{\theta \in]\,0,\,1\,[})$ of the coin toss game, for which the likelihood L_n is, if $\theta \in]\,0,\,1\,[$ and $(x_1,\,\ldots,\,x_n) \in \{0,\,1\}^n$,

$$L_n(x_1,\,\ldots,\,x_n;\,\theta) = \theta^{n\hat{x}_n}(1-\theta)^{n(1-\hat{x}_n)},\ \text{with}\ \hat{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

According to the calculation made previously for the variance $\nabla \ln L_n(x; \theta)$, the I_n information of the model is given by

$$I_n = \mathbb{V}_\theta (\nabla \ln L_n(x; \theta)) = \frac{n}{\theta(1-\theta)}.$$

The uncertainty is low for θ close to 0 and 1, while it is even greater as θ is close to $1/2$, which results in information $I_n(\theta)$ maximum for θ close to 0 and 1, and minimum for $\theta = 1/2$.

In an independent sampling situation, Fisher information is proportional to the sample size.

We consider $(\mathcal{H}^n, \{P_\theta\}_{\theta \in \Theta})$ as a model dominated by a measure σ-finite μ, with $\mathcal{H} \in \mathbb{R}^K$ and $\theta \subset \mathbb{R}^d$.

Likelihood and maximum likelihood

When $\mathcal{H}^n$ is discrete, the probability that the sample $(X_1, ..., X_n) \sim P_\theta$ is equal to $(x_1, ..., x_n) \in \mathcal{H}^n$ represents the degree of likelihood of this observation for the law P_θ. So the definition is given below:

Definition 2.2. The likelihood of the model $(\mathcal{H}^n, \{P_\theta\}_{\theta \in \Theta})$ is the map

$$L_n : \mathcal{H}^n \times \theta \to \mathbb{R}_+$$

such that, for each $\theta \in \Theta$,

$$L_n(x; \theta) : \mathcal{H}^n \to \mathbb{R}_+$$

is an element of the equivalence class of the density of P_θ with respect to μ.

If L_n is the likelihood of the model, then a value of the parameter which fits the observation is in the set of maxima of the function

$$[0; 1] \to \mathbb{R}, \quad \theta \mapsto L_n(x_1, ..., x_n; \theta),$$

because $L_n(x_1, ..., x_n; \theta)$ represents the probability that a sample of the distribution $\mathcal{B}(\theta)^{\otimes n}$ is equal to $(x_1, ..., x_n)$.

This principle of constructing estimators is exported to the case of continuous models, for example $(\mathbb{R}^n, \{\mathcal{N}(\theta, 1)^{\otimes n}\}_{\theta\in\mathbb{R}})$. For the observation $(x_1, ..., x_n)$ generated according to the law $\mathcal{N}(\theta, 1)^{\otimes n}$, the curve of the function

$$\theta \mapsto \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right),$$

is then maximum at a point close to 0. This intuitive principle leads us to choose as the estimator of the model, a parameter which maximizes the likelihood. This is the concept of maximum likelihood estimator.

3. Main Results

In reliability analysis, failure data frequently contains individual failure times.

Proposition 1. *Let* $\hat{Y} = (y_1, ..., y_n)$ *be a sample of a real random variable y with distribution function density F and* $\hat{Y}^\star = (y_1^\star, ..., y_n^\star)$ *be the associated order statistic. Then the density of* $(y_1^\star, ..., y_{n-1}^\star)$ *is conditioned by* $y_n^\star$ *of a real random variable x whose distribution function is given by the truncation of F into* $\hat{Y}_n^\star$.

Proof. Let $(y_1, ..., y_n) \in \mathbb{R}^n$, $y_1 \leq \cdots \leq y_n$. Then the density of $(y_1^\star, ..., y_{n-1}^\star)$ conditioned by $y_n^\star$ is given by

$$f_{y_n}(y_1, ..., y_{n-1}) = \frac{\dfrac{\partial^n}{\partial y_1, ..., \partial y_n} P(y_1^\star < y_1, ..., y_{n-1}^\star < y_{n-1}, y_n^\star < y_n)}{\dfrac{d}{dy_n} P(y_n^\star < y_n)},$$

$$\frac{d}{dy_n} P(y_n^\star < y_n) = nf(y_n)F^{n-1}(y_n),$$

$$P(y_1^\star < y_1, ..., y_{n-1}^\star < y_{n-1}, y_n^\star < y_n)$$

$$= \sum_{p=0}^{n=1} \left\{ \frac{n!}{(n-p)!} [F(y_{p+1}) - F(y_p)]^{n-p} \prod_{i=1}^{p} [F(y_i) - F(y_{i-1})] \right\}, \quad (1)$$

with by convention $F(y) = 0$.

In the calculation of the derivative n^{iem}, only the term corresponds to $p = n - 1$ intervenes, which is the only one which contains $(y_1, ..., y_n)$. So

$$\frac{\partial^n}{\partial y_1, ..., \partial y_n} P(y_1^\star < y_1, ..., y_n^\star < y_n)$$

$$= n! \frac{\partial^n}{\partial y_1, ..., \partial y_n} \left\{ \prod_{i=1}^{n} (F(y_i) - F(y_{i-1})) \right\}.$$

Thus,

$$\frac{\partial^n}{\partial y_1, ..., \partial y_n} P(y_1^\star < y_1, ..., y_n^\star < y_n) = n! \prod_{i=1}^{n} f(y_i). \quad (2)$$

Consequently, according to equations (1) and (2), we have

$$f_{y_n}(y_1, ..., y_{n-1}) = \frac{n! \prod_{i=1}^{n} f(y_i)}{nf(y_n)F^{n-1}(y_n)}$$

$$= (n-1)! \prod_{i=1}^{n-1} \left[\frac{f(y_i)}{F(y_n)} \right].$$

Furthermore, let $\hat{X} = (x_1, ..., x_{n-1})$ be an $(n-1)$ sample of a real random variable by distribution function F_{y_n} (F is right truncated to y_n). Then the density of $\hat{X}$ is given by

$$g(y_1, ..., y_{n-1}) = \prod_{i=1}^{n-1} \left[\frac{f(y_i)}{F(y_n)} \right].$$

With $\hat{X}^\star$, the associated order statistic is

$$h(y_1, ..., y_{n-1}) = \frac{\partial^{n-1}}{\partial y_1, ..., \partial y_{n-1}} P(x_1^\star < y_1, ..., x_{n-1}^\star < y_{n-1}).$$

Thus

$$h(y_1, ..., y_{n-1}) = (n-1)! \prod_{i=1}^{n-1} \left[\frac{f(y_i)}{F(y_n)} \right].$$

There exists $i < j$ such that $y_i > y_j$, and hence

$$f_{y_n}(y_1, ..., (y_{n-1})) = h(y_1, ..., (y_{n-1})) = 0. \qquad \square$$

The density of the sample, excluding the last random observation, is the distribution function given by the truncation of F at $\hat{Y}_n^\star$.

Thus, if we study observations for which the lower bound is y_1, the frequency of observations less than or equal to y_1 is not zero, whereas the density of the probability law chosen a priori to perform a statistical adjustment on these data is zero. Moreover, between y_1 and y_h with $h < n-1$ (y_h being a threshold higher than the bound y_1), the number of observations that should have been made is unknown, as are the individual values of these observations. The threshold y_h can be determined by the sensitivity of the measuring instrument, see [9].

Consider the probability law of a random variable X varying from $-\infty$ to $+\infty$. Let $f(x)$ denote a probability density function, and $F(x)$ the distribution function. Then

$$F(x) = Pr(x \leq x) = \int_{-\infty}^{x} f(x)dx.$$

It is assumed that the n observed values of this variable belong only to a $[a; b]$ part of its domain of variation. The probability of the variable taking a value outside the $[a; b]$ domain is non-zero but unknown. The distribution of the random variable can be studied only on the $[a; b]$ part. This defines the truncated probability distribution function $\xi(x)$ such that

$$\xi(x) = Pr_{a \le x \le b}(X \le x) = \frac{F(x) - F(a)}{F(b) - F(a)},$$

$$\xi(a) = 0 \quad \text{and} \quad \xi(b) = 1$$

and probability density function:

$$\frac{\delta\xi(x)}{\delta x} = \frac{f(x)}{\int_a^b f(x)dx}$$

if an analytical expression is chosen as a function of the parameters θ_k for the distribution $f(x)$, an estimate of the parameters θ_k from the only n observed values available, belonging to the interval $[a; b]$ can be obtained using the maximum likelihood method. The maximum likelihood method consists in finding the parameters θ_k that maximize the probability of obtaining the sample of observed values with the envisaged law. The likelihood function to be maximized $L(x, \theta)$ is therefore written as

$$L(x, \theta) = \prod_{i=1}^{n} \frac{\delta\xi(x)}{\delta x}$$

$$= \frac{\prod_{i=1}^{n} f(x)}{\left[\int_a^b f(x)dx\right]^n}.$$

Let us take observations whose lower bound is x_0. Then the frequency of observations less than or equal to x_0 is not zero, whereas the density of the a priori probability distribution chosen to perform a statistical adjustment on these data is zero or indefinite at this point $[.]$.

What is more, between x_0 and x_h (x_h threshold is greater than the x_0 boundary), the number of observations that should have been made is unknown, as are the individual values of these observations.

The number x_h can be determined by the sensitivity of the measuring device, so

$$L(x > 0,\, \theta_h) = \frac{\prod_{i=1}^{n} g(x_0)}{\left[\int_{x_h}^{+\infty} g(x)\,dx\right]^n}.$$

If $G(x)$ denotes the distribution function of the probability distribution of non-zero values, and $F(x)$ the distribution function of the probability distribution of zero or non-zero rain data, then the relationship between these two functions is

$G(x) = \dfrac{F(x) - P_0}{1 - P_0}$, where P_0 is the probability of zero or non-zero rain data.

It is important to establish uniform asymptotic developments, so let us consider μ_y and $\partial\mu_y/\partial y$.

Lemma 1. *Suppose* $y \in \mathcal{C}$ *and* $h_y \in \mathcal{C}^{\infty}(\mathbb{R}^+)$. *Let*

$$i(y) = \min\{j \in \mathbb{N} / h_y^{(j)}(0) \neq 0\}.$$

If

$$\left|\, \sup_{y \in \mathcal{C}_{w \geq 0}}(w) \,\right| < \infty,$$

then

$$\lim_{t \to \infty} \sup_{y \in \mathcal{C}} \left| \, t^{i(y)+1}\, \hat{h}_y(t) - h_y^{(i(y))}(0) \,\right| = 0,$$

with $\hat{h}_y(t) = \int_0^{+\infty} \exp(-tu) h_y(u)\,du$ *the Laplace transform of* h_y.

Proof. Using Taylor expansion to order $i+1$, we have

$$h_y(u) = \frac{u^i}{i!} h_y^{(i)}(0) + \frac{u^{i+1}}{(i+1)!} h_y^{(i+1)}(\eta u),$$

with $\eta \in]0; 1[$ and consequently,

$$\hat{h}_y(t) = \int_0^\infty \exp(-tu) \frac{u^i}{i!} h_y^{(i)}(0) du + \int_0^\infty \exp(-tu) \frac{u^{i+1}}{(i+1)!} h_y^{(i+1)}(\eta u) du,$$

so

$$\hat{h}_y(t) = T_y^{(1)}(t) + T_y^{(2)}(t).$$

It follows that

$$T_y^{(1)}(t) = \frac{h_y^{(i)}(0)}{i!} \int_0^\infty u^i \exp(-tu) du,$$

which also gives

$$T_y^{(1)}(t) = \frac{h_y^{(i)}(0)}{t^{i+1}},$$

and the change of variable $v = tu$ implies

$$T_y^{(2)}(t) = \frac{1}{(i+1)!} \frac{1}{t^{i+2}} \int_0^\infty \exp(-v) v^{i+1} h_y^{(i+1)}\left(\eta \frac{v}{t}\right) dv.$$

Therefore, uniformly for $y \in \mathcal{C}$,

$$| T_y^{(2)}(t) | \leq \sup_{z \in \mathcal{C}_{w \geq 0}} | h_z^{(i+1)}(w) | \frac{1}{(i+1)!} \frac{1}{t^{i+2}} \int_0^\infty \exp(-v) v^{i+1} dv.$$

On the other hand,

$$\sup_{z \in \mathcal{C}_{w \geq 0}} | h_z^{(i+1)}(w) | \frac{1}{(i+1)!} \frac{1}{t^{i+2}} \int_0^\infty \exp(-v) v^{i+1} dv$$

$$= \frac{1}{t^{i+2}} \sup | h_z^{(i+1)}(w) | = O\left(\frac{1}{t^{i+2}}\right).$$

So

$$\hat{h}_y(t) = \frac{h_y^{(i)}(0)}{t^{i+1}} + O\left(\frac{1}{t^{i+2}}\right).$$

Hence

$$\hat{h}_y(t) = \frac{1}{t^{i+1}}\left(h^{(i)}(0) + O\left(\frac{1}{t}\right)\right),$$

which concludes the proof. □

Proposition 2. *Let F be a distribution function such that $\int_{\mathbb{R}} x^2 dF(x) < \infty$, a belonging to A and $a' = \inf\{x/F(x) = F(a)\}$. Then an estimator $X_n^\star$ of a' is asymptotically unbiased and convergent.*

The amount of Fisher information $I_{\hat{X}}(a)$ contributed by the n-sample $\hat{X}$ is proportional to n^2. For n fixed, $I_{\hat{X}}(a)$ is a decreasing (*respectively*, *increasing*) *function of a in the log-concavity* (*respectively*, *log-convexity*) *domain of F.*

Proof. Integration by parts shows that the expectation $M(a)$ of a distribution function F_a is written as

$$M(a) = \int_{\mathbb{R}} x dF_a(x) = a - \int_{-\infty}^{a} F_a(x) dx$$

$$= a - \int_{-\infty}^{a} \frac{F(x)}{F(a)} dx.$$

Putting $X_n^\star$, we get

$$E(X_n^\star) = a - \int_{-\infty}^{a} \left(\frac{F(x)}{F(a)}\right)^n dx,$$

and also

$$E(X_n^\star) = a - \int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx - \int_{a'}^{a} \left(\frac{F(x)}{F(a)}\right)^n dx.$$

Then

$$E(X_n^\star) = a' - \int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx.$$

$X_n^\star$ is therefore an unbiased estimator of a and a'. Then

$$\lim E(X_n^\star) = a'.$$

We next show that $X_n^\star$ is convergent and for this, $V(X_n^\star) \to 0$.

Using integration by parts, we have

$$V(X_n^\star) = 2a\int_{-\infty}^{a} \left(\frac{F(x)}{F(a)}\right)^n dx - \left(\int_{-\infty}^{a} \left(\frac{F(x)}{F(a)}\right)^n dx\right)^2 - 2\int_{-\infty}^{a} x\left(\frac{F(x)}{F(a)}\right)^n dx.$$

Taking into account that $\forall x \in]\, a', a\,],\ F(x) = F(a),$ we obtain

$$V(X_n^\star) = 2a\left[\int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx + a - a'\right] - \left[\int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx + a - a'\right]^2$$

$$- 2\left[\int_{-\infty}^{a'} x\left(\frac{F(x)}{F(a)}\right)^n dx + \frac{a^2 - a'^2}{2}\right]$$

so we have

$$V(X_n^\star) = 2a\int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx - \left[\int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx\right]^2$$

$$- 2(a - a')\int_{-\infty}^{a'} \left(\frac{F(x)}{F(a)}\right)^n dx - 2\int_{-\infty}^{a'} x\left(\frac{F(x)}{F(a)}\right)^n dx. \qquad \square$$

Applying Lebesgue's theorem for each of the integrals, we find that $\lim_{n\to\infty} V(X_n^\star) = 0,$ so for the last integral, we have

$$\left| x\left(\frac{F(x)}{F(a)}\right)^n \right| \leq \left| x\,\frac{F(x)}{F(a)} \right|,$$

since $\int_{\mathbb{R}} x^2 dF(x) < \infty,$ by hypothesis,

$$\forall x < a', \quad x\left(\frac{F(x)}{F(a)}\right)^n \to_{n\to\infty} 0.$$

As a reminder,

$$I_{\hat{X}}(a) = E_a\left(\frac{\partial}{\partial a}\log L_{\hat{X}}(a)\right)^2,$$

$$\log L_{\hat{X}}(a) = -n\log F(a) + \sum_{i=1}^{n} \log f(X_i), \text{ if } \max X_i \leq a.$$

So,

$$\frac{\partial}{\partial a}\log L_{\hat{X}}(a) = -n\,\frac{f(a)}{F(a)}, \text{ if } \max X_i \leq a.$$

Then

$$I_{\hat{X}}(a) = \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} \left(-n\,\frac{f(a)}{F(a)}\right)^2 \frac{1}{F^n(a)} \prod_{i=1}^{n} f(x_i) d\hat{X}.$$

As a result,

$$I_{\hat{X}}(a) = n^2\left(\frac{f(a)}{F(a)}\right)^2.$$

Hence

$$I_{\hat{X}}(a) = n^2\left(\frac{d}{da}\log F(a)\right)^2.$$

Bearing in mind that $X^\star$ is asymptotically unbiased, the inverse of the variance measures the precision of this estimator for large *n*. Thus, when *F* is log-concave, the greater the value to be estimated, the lower the precision.

Corollary 1. *When F is log-concave, the precision of* $X^\star$ *is a decreasing function of a.*

Proof. For *n* fixed, either of the following holds:

$$g(x) = E(X_n^\star - a)^2 = E[(X_n^\star - E(X_n^\star))] + (E(X_n^\star) - a)^2,$$

$$g(x) = E(X_n^\star - E(X_n^\star))^2 + (E(X_n^\star) - a)^2.$$

The distribution function of $X_n^\star$ being (F^n) is log-concave. According to the hypothesis on *F*, the first term is an increasing function of *a*.

On the other hand, $h(a) = a - E(X_n^\star)$ is positive, and

$$h(a) = \frac{(F^n)_{[1]}(a)}{F^n(a)},$$

where

$$(F^n)_{[1]}(a) = \int_{-\infty}^{a} F^n(x)dx.$$

Now, if a function is log-concave, then its antiderivative that vanishes at $-\infty$ is itself log-concave, which implies that $(F^n)_{[1]}$ is therefore log-concave. Consequently, *h* is increasing, as is h^2. The same holds for *g*, from which the corollary follows. □

Proposition 3. *Let* I_n *be Fisher information of the model* $(\mathcal{H}^n, \{P_\theta\}_{\theta\in\Theta})$.

Assume that for each $\theta \in \Theta$, *there is a neighborhood* $v \subset \Theta$ *of* θ *such that* $\sup_{\alpha\in v}\| \nabla L_n(x; \alpha) \| \in \mathbb{L}^1(\mu)$. *Then*

(i) $\mathbb{E}_\theta \nabla \ln L_n(x;\, \theta) = 0$;

(ii) *if furthermore*,

$$\sup_{\alpha \in v} \| \nabla^2 L_n(x;\, \alpha) \| \in \mathbb{L}^1(\mu), \quad I_n(\theta) = -\mathbb{E}_\theta \nabla^2 \ln L_n(x;\, \theta).$$

Proof. Under the condition $\sup_{\alpha \in v} \| \nabla L_n(x;\, \alpha) \| \in \mathbb{L}^1(\mu)$, according to the derivation theorem for the integral, we have

$$\int_{\mathcal{H}^n} \nabla L_n(x;\, \theta) d\mu = \nabla \int_{\mathcal{H}^n} L_n(x;\, \theta) d\mu.$$

Thus, since $dP_\theta = L_n(x;\, \theta) d\mu$ and $P_\theta(\mathcal{H}^n) = 1$,

$$\int_{\mathcal{H}^n} \nabla L_n(x;\, \theta) d\mu = \nabla P_\theta(\mathcal{H}^n) = 0.$$

Now, since

$$dP_\theta = L_n(x;\, \theta) d\mu$$

and

$$\nabla \ln L_n(x;\, \theta) = \nabla L_n(x;\, \theta) / L_n(x;\, \theta),$$

$$\mathbb{E}_\theta \nabla \ln L_n(x;\, \theta) = \int_{\mathcal{H}^n} \nabla \ln L_n(x;\, \theta) L_n(x;\, \theta) d\mu,$$

$$\mathbb{E}_\theta = \int_{\mathcal{H}^n} \nabla L_n(x;\, \theta) d\mu,$$

(i) follows.

We now show (ii).

If $F : \Theta \to \mathbb{R}$ is a class C^2, for all $i,\ j = 1, ..., d$ and $\theta \in \Theta$, then

$$\nabla_i F(\theta) = \frac{\partial F}{\partial \theta_i}(\theta) \quad \text{and} \quad \nabla^2_{i,\, j} F(\theta) = \frac{\partial^2 F}{\partial \theta_i \partial \theta_j}(\theta).$$

The derivation theorem under the integral shows that

$$\int_{\mathcal{H}^n} \nabla^2_{i,\, j} L_n(x;\, \theta) d\mu = \nabla^2_{i,\, j} \int_{\mathcal{H}^n} L_n(x;\, \theta) d\mu,$$

under the assumption $\sup_{\alpha \in \nu} \| \nabla^2_{i,j} L_n(x; \alpha) \| \in \mathbb{L}^1(\mu)$, and therefore

$$\int_{\mathcal{H}^n} \nabla^2_{i,j} L_n(x; \theta) d\mu = \nabla_{i,j} P_\theta(\mathcal{H}^n) = 0.$$

We also check that for any $x \in \mathcal{H}^n$,

$$\nabla^2_{i,j} \ln L_n(x; \theta) = \frac{\nabla^2_{i,j} L_n(x, \theta)}{L_n(x, \theta)} - \frac{\nabla_i L_n(x; \theta) \nabla_j L_n(x; \theta)}{L^2_n(x; \theta)}.$$

As a result,

$$\mathbb{E}_\theta[\nabla^2_{i,j} \ln L_n(x; \theta)] = \int_{\mathcal{H}^n} \nabla^2_{i,j} \ln L_n(x; \theta) L_n(x; \theta) d\mu$$

and likewise

$$\int_{\mathcal{H}^n} \nabla^2_{i,j} \ln L_n(x; \theta) L_n(x; \theta) d\mu = -\int_{\mathcal{H}^n} \frac{\nabla_i L_n(x; \theta) \nabla_j L_n(x; \theta)}{L_n(x; \theta)}.$$

So we get

$$\mathbb{E}_\theta[\nabla^2_{i,j} \ln L_n(x; \theta)] = -\mathbb{E}_\theta \nabla_i \ln L_n(x; \theta) \nabla_j \ln L_n(x; \theta),$$

because $\nabla \ln L_n(x; \theta) = \nabla L_n(x; \theta) / L_n(x; \theta)$. □

Now, by Fisher's definition of information:

$$I_n(\theta)_{i,j} = \operatorname{cov}_\theta(\nabla_i \ln L_n(x; \theta), \nabla_j \ln L_n(x; \theta))$$

$$= \mathbb{E}_\theta \nabla_i \ln L_n(x; \theta), \nabla_j \ln L_n(x; \theta).$$

Since $\mathbb{E}_\theta \nabla_i \ln L_n(x; \theta) = 0,$ hence follows (ii).

4. Conclusion and Discussion

The spatial accuracy of data is a crying phenomenon in the scientific world. In the present work, we have made a contribution to this theme. In particular, the results allow us to model and describe that the sample density

of the last random sample is given by the truncation of the distribution function.

We also managed to show that when the estimator is unbiased, the inverse of the variance measures the precision of the estimator, whatever the sample size, and if the distribution function is log-concave, then the precision is very low compared to the value to be estimated.

In the future, it would be interesting to build coherent estimators for predictive measures in the context of extreme values for possible applications to real data.

Acknowledgement

We thank the anonymous referees for their comments and feedback on earlier version of this document.

References

[1] Brunet-Moret Yves, Etude de quelques lois statistiques utilisées en hydrologie, Cahiers ORSTOM, Série Hydrologie 6(3) (1969), 3-100.

[2] David Penot, Catographie des AVA “nements” hydrologique extrêmes et estimation schadex en site non jaugés, Thèse de Doctorat, Université de Grenoble, 2014.

[3] Dacosta A. Ansoumana, Characterization of the rainfall regime in the upper Senegal River basin in a context of climatic variability, Physico-geo-Geographie Physique and Environment 5 (2011), 116-133. DOI: 10.4000/pysico-geo.1958.

[4] Cheikh Faye, Amadou Abdou Sow and Jean Baptiste Ndong, Study of rainfall and hydrological droughts in tropical Africa: characterization and mapping of drought by indices in the upper Senegal River basin, Physio-Gêo 9 (2015), 17-35. DOI: 10.4000/Physio-geo.4388.

[5] Celine Guivarch and Nicolas Taconet, Global inequality and climate change, Revue de l’OFCE 2020/1(No 165). https://doi.org/10.3917/reof.165.0035.

[6] J. Heckman and B. Singer, A method for minimizing the impact of distributional assumptions in econometric models for duration data, Econometrica 52(2) (1984), 271-320.

[7] Wahid A. M. Shehata, Hafida Goual, Talhi Hamida, Aiachi Hiba, G. G. Hamedani, Abdullah H. Al-Nefaie, Mohamed Ibrahim, Nadeem S. Butt, Rania M. A. Osman and Haitham M. Yousof, Censored and uncensored Nikulin-Rao-Robson distributional validation : characterizations, classical and Bayesian estimation with censored and uncensored application, Pakistan Journal of Statistics and Operation Research 20(1) (2024), 11-35.

[8] Bradley Efron and Vahé Petrosian, Nonparametric methods for doubly truncated data, J. Amer. Statist. Assoc. 94(447) (1999), 824-834.

[9] Mathieu Tiene, M. Dodo Natatou and Diakarya Barro, Probabilistic method for modeling the conditional extreme quantiles and censored data, Kongzhi yu Juece/Control and Decision 39(2) (2024).

[10] Bewentaore Sawadogo and Diakarya Barro, Space-time trend detection and dependence modeling in extreme event approaches by functional peaks-over thresholds : application to precipitation in Burkina Faso, Int. J. Math. Math. Sci. 2022, Art. ID 2608270, 12 pp. https://doi.org/10.1155/2022/2608270.