# Tail Index Estimation Based on Survey Data

Patrice Bertail*
MODAL'X Université Paris-Ouest, Nanterre, France - patrice.bertail@gmail.com

Emilie Chautru
Mines de Paris, Fontainebleau, France - emilie.chatru@gmail.com

Stephan Clémencon
LTCI UMR Telecom ParisTech/CNRS, Paris, France - stephan.clemencon@gmail.com

## Abstract

This paper is devoted to tail index estimation in the context of survey data. Assuming that the population of interest is described by a heavy-tailed statistical model, we prove that the survey scheme plays a crucial role in the design of consistent inference methods for extremes. Focus is here on the celebrated Hill method for tail index estimation; it is shown how to modify it in order to take into account the survey design. Precisely, under specific conditions on the inclusion probabilities of first and second orders, we establish the consistency of the variant of the Hill estimator we propose. Additionally, its asymptotic normality is proved for specific sampling schemes (Poisson and Rejective sampling or sampling plans close to rejective. Applications are thoroughly discussed and illustrated by numerical results.

**Keywords**: Survey sampling, tail index estimation, Hill estimator, Poisson survey scheme, rejective sampling.

**1. Introduction** It is the main purpose of this paper to study the impact of a survey sampling scheme on tail index estimation. Indeed, in many situations, statisticians have at their disposal not only data but also weights arising from some survey sampling stratification. These weights correspond either to true inclusion probabilities, as is often the case for institutional data, or to some calibrated or post-stratification weights (minimizing some discrepancy with the inclusion probabilities subject to some margin constraints). In most cases, the survey design is ignored, the statistics thus computed possibly exhibiting then a significant sampling bias. When considering statistics of extremes in particular, this may cause severe drawbacks and completely jeopardize the estimation, as can be revealed by simulation experiments.

Our goal is here to show how to incorporate the survey scheme into extreme value statistical techniques, in order to guarantee their asymptotic validity. Our approach is illustrated through the tail index estimation problem in the context of heavy-tailed survey data. We propose a specific modification of the Hill estimator, accounting for the survey plan by means of which data have been collected, and establish its consistency and asymptotic normality under adequate assumptions on the probabilities of inclusion of first and second orders.

## 2. Hill estimator in survey sampling

Here and throughout, we consider a finite population of size $N \geq 1$, $\mathcal{U}_N := \{1, \ldots, N\}$ say. We call a *sample* of (possibly random) size $n \leq N$, any subset $s := \{i_1, \ldots, i_{n(s)}\} \in \mathcal{P}(\mathcal{U}_N)$ with cardinality $n =: n(s)$ less that $N$. A sampling scheme (design/plan) without replacement is determined by a probability distribution $R_N$ on the set of all possible samples $s \in \mathcal{P}(\mathcal{U}_N)$. For any $i \in \{1, \ldots, N\}$, the following quantity, generally called (first order) *inclusion probability*,

$$\pi_i(R_N) := P_{R_N}\{i \in S\},$$

is the probability that the unit $i$ belongs to a random sample $S$ drawn from distribution $R_N$. In vectorial form, we shall write $\pi(R_N) := (\pi_1(R_N), \ldots, \pi_N(R_N))$. First order inclusion probabilities are assumed to be strictly positive in the subsequent analysis: $\forall i \in \{1,\dots,n\}$, $\pi_i(R_N) > 0$. Additionally, the second order inclusion probabilities are denoted by

$$\pi_{i,j}(R_N) := P_{R_N}\{(i,j) \in S^2\},$$

for any $i \neq j$ in $\{1,\dots,N\}^2$. When no confusion is possible, we shall fail to mention the dependence in $R_N$ when writing the first/second order probabilities of inclusion. The information related to the observed sample $S \subset \{1,\dots,N\}$ is encapsulated by the random vector $(\epsilon_1,\dots,\epsilon_N)$, where

$$\epsilon_i = \begin{cases} 1 & if i \in S \\ 0 & otherwise. \end{cases}$$

The distribution of the sampling scheme $\epsilon$ has 1-d marginals that correspond to the Bernoulli distributions $\mathcal{B}(\pi_i)$, $1 \leq i \leq N$, and covariance matrix given by

$$\Gamma_N := \{\pi_{i,j} - \pi_i\pi_j\}_{1\leq i,j\leq N}.$$

Notice incidentally that, equipped with these notations, we have $\sum_{i=1}^N \epsilon_i = n(S)$.

The *superpopulation model* we consider here stipulates that :

**Assumption** $H_1$ a real-valued random variable $X$ with cdf $F$, supposed to be absolutely continuous, with density $f$, observable on the population $\mathcal{U}_N$, *i.e.* $X_1,\dots,X_N$ are iid realizations drawn from $f$.

In practice, it is customary to determine the first order inclusion probabilities as a function of an *auxiliary variable*, which is observed on the entire population. Here, it is denoted by $\mathbf{W}$. Hence, for all $i \in \{1,\dots,N\}$ we can write $\pi_i = \pi(\mathbf{W}_i)$ for some link function $\pi(.)$. In the following we will assume that
**Assumption** $H_2$ : The random vectors $W_1,\dots,W_N$ are iid with continuous distribution $P_W$ on $\mathcal{W} \subset R^d$, cdf $F_{\mathbf{W}}$ and density $f_{\mathbf{W}}$. The joint distribution of the entailed iid sequence $\{(X_i,\mathbf{W}_i),\ 1 \leq i \leq N\}$ is denoted by $P_{X,\mathbf{W}}$ with corresponding cdf $F_{X,\mathbf{W}}$ Moreover we assume that the joint cdf $F_{X,\mathbf{W}}$ is absolutely continuous with Lebesgue-integrable density $f_{X,\mathbf{W}}$ such that for all $(x,\mathbf{w}) \in (0,+\infty] \times \mathcal{W}$,

$$f_{X,\mathbf{W}}(x,\mathbf{w}) := c\,(F(x), F_W(\mathbf{w}))\ f(x)\, f_{\mathbf{W}}(\mathbf{w}),$$

for some copula density $c : R_+^\star \times R^d \to R$.

We shall denote by $X_{1,n} \leq \dots \leq X_{n,n}$ the order statistics related to the survey *sample* $(X_{i_1},\dots,X_{i_n})$, where $n = n(S)$ may be random. When unit $j$ is such that $X_j = X_{i,N}$, the $i$-the largest observation in the population, $1 \leq i,j \leq N$, its inclusion indicator and probability are denoted by $\epsilon_{i,N} = \epsilon_j$ and $\pi_{i,N} = \pi_j$ respectively. Similarly, we write $\pi_{i,n} := \pi_j$ when $X_{i,n} = X_j$, $1 \leq i,j \leq n$.

In a wide variety of situations, it is appropriate to assume that a statistical population is described by a heavy-tailed probability distribution (the field of heavy-tail analysis is well depicted in [6]). A distribution with Pareto-like right tail is any probability measure $P$ on $R$ with cdf $F$ such that for all $x \in R$,

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

where $\gamma > 0$ is the *extreme value index* (EVI) of distribution $P$ and $L(x)$ is a *slowly varying function*, *i.e.* a function such that $L(t\,x)/L(x) \to 1$ as $x \to +\infty$ for all $t > 0$. Notice that instead of the EVI, focus is often on $\alpha := 1/\gamma$, the *tail index* of the distribution $P$. Functions of the form introduced in eq:HSSRV are said to be *regularly varying* with index $-1/\gamma$; the set of such functions is denoted by $\mathcal{R}_{-1/\gamma}$. Notice first that, under the heavy-tail assumption above, we have:

$$\gamma = \lim_{x\to\infty} \int_x^{+\infty} \frac{\overline{F}(u)}{\overline{F}(x)} \frac{du}{u},$$

The empirical estimator based on the Horvitz-Thompson estimator of this quantity is given by (based on the $k$ largest values in the sample)

$$\begin{aligned} H^{\pi}_{k,n}. &= \left(\sum_{j=1}^{k} \frac{1}{\pi_{n-j+1,n}}\right)^{-1} \sum_{i=1}^{k} \frac{1}{\pi_{n-i+1,n}} \log\left(\frac{X_{n-i+1,n}}{X_{n-k,n}}\right) \\ &= \left(\sum_{j=1}^{K} \frac{\epsilon_{N-j+1,N}}{\pi_{N-j+1,N}}\right)^{-1} \sum_{i=1}^{K} \frac{\varepsilon_{N-i+1,N}}{\pi_{N-i+1,N}} \log\left(\frac{X_{N-i+1,N}}{X_{N-K,N}}\right) \\ &= H^{\pi}_{K,N} \end{aligned}$$

where $K$ is the chosen number of largest observations in the population corresponding to $k$ in the sample (we may choose indiferrent K or k).

**3. Asymptotic results** Here we investigate the limit properties of the estimator $H^{\pi}_{K,N}$ as $N$ and $n$ simultaneously go to infinity, with $n \leq N$. The following assumptions, related to the sample design, shall be involved in the asymptotic analysis. In the following we will assume that
**Assumption** $H_3$ There exist $\pi_\star > 0$ and $N_0 \in N^*$ such that for all $N \geq N_0$ and $i \in \mathcal{U}_N$,

$$\pi_\star < \pi_i < 1 - \pi_\star.$$

**Assumption** $H_4$ There exists $c < +\infty$ such that we have $\forall\, N \geq 1$,

$$\max_{1\leq i,\, j\leq N} |\pi_{i,j} - \pi_i\pi_j| \leq \frac{c}{n}.$$

We then have the following result

**Theorem** [Consistency] Let $K = K(N)$ be a sequence of integers such that $K \to +\infty$ and $K/N \to 0$ as $N, n \to +\infty$. Provided that Assumptions $H_1 - H_4$ hold then , in $P$-probability, we have

$$H^{\pi}_{K,N} \longrightarrow \gamma.$$

Whereas the consistency of the standard Hill estimator can be proved for any sequence $K$ going to infinity at a reasonable rate, asymptotic normality cannot be guaranteed at such a level of generality. Higher-order regular variation properties of the heavy-tail model eq:HSSRV are required [2, 3]. More specifically, consider the hypothesis below, referred to as the *Von Mises condition* [4].
**Assumption** $H_5$ The regularly varying tail quantile function $U \in \mathcal{R}_\gamma$ with $\gamma > 0$ is such that there is a real parameter $\rho < 0$, referred to as the *second order parameter*, and a positive or negative function $A$ with $\lim_{x\to+\infty} A(x) = 0$ such that for any $t > 0$, as $x \to \infty$, we have

$$\frac{1}{A(x)}\left(\frac{U(tx)}{U(x)} - t^{\gamma}\right) \longrightarrow t^{\gamma}\,\frac{t^{\rho}-1}{\rho},$$

or equivalently

$$\frac{1}{A\left(\frac{1}{\overline{F}(x)}\right)}\left(\frac{\overline{F}(t\,x)}{\overline{F}(x)} - t^{-1/\gamma}\right) \longrightarrow t^{-1/\gamma}\,\frac{t^{\rho/\gamma}-1}{\gamma\,\rho}.$$

**The Poisson survey sampling sheme** The following result reveals that under the Poisson survey scheme, when based on the $K$ largest values among the whole population $X_1, \ldots, X_N$, $H^{\mathbf{p}}_{K,N}$ converges at the same rate ($1/\sqrt{K}$ namely) to the same limit distribution as $H_{K,N}$, up to a multiplicative term in the asymptotic variance induced by the sampling scheme. Further details about the convergence of the classical Hill estimator

can be found *e.g.* in [2, Theorem 1] and [6, Section 9].
**Theorem** [Limit distribution in the Poisson survey case] Suppose that Assumptions $H_1$ to $H_5$ are satisfied. In addition, assume that

$$\int_{[0,1]^d} c(1, \mathbf{v})\, d\mathbf{v} < \infty.$$

Then, for

$$\sigma_p^2 := \int_{[0,1]^d} \frac{1}{p\left(F_{\mathbf{W}}^{\leftarrow}(\mathbf{v})\right)}\, c(1, \mathbf{v})\, d\mathbf{v}$$

and provided that $K \to +\infty$ as $N \to +\infty$ so that $\sqrt{K} A(N/K) \to \lambda$ for some constant $\lambda \in R$, we have the convergence in distribution as $N \to +\infty$:

$$\sqrt{K}\left(H^{\mathbf{p}}_{K,N} - \gamma\right) \Rightarrow \mathcal{N}\left(\frac{\lambda}{1-\rho}\, , \, \gamma^2\, \sigma_p^2\right).$$

### Extension to conditional Poisson sampling plans

We will show how the result stated in the above theorem can be extended to an important class of survey plans (with fixed sampling size), namely *rejective sampling schemes*. In opposition to what happens to the Horvitz-Thompson mean, the asymptotic variance is the same and need not to be recentered as in [5]. For clarity's sake, we provide a brief description of the latter, refer to see [5] and [1] for further details.

Fix $n \le N$ and consider a vector $(\pi_1^R,\ \ldots,\ \pi_N^R)$ of first order inclusion probability. The rejective sampling, sometimes referred to as *conditional Poisson sampling* (CPS in short), exponential design without replacement or maximum entropy design is the sampling plan $R_N$ which picks samples of fixed size $n(\mathcal{S}) = n$ in order to maximize the entropy measure

$$H(R_N) = -\sum_{\{\mathcal{S} \subset \mathcal{P}_N:\ \#\mathcal{S} = n\}} R_N(\mathcal{S}) \log R_N(\mathcal{S})$$

subject to the constraint stipulating that its vector of first order inclusion probabilities coincides with $(\pi_1^R,\ \ldots,\ \pi_N^R)$. It can be implemented in two steps, as follows.

($i$) Draw a sample $\mathcal{S}$ with a Poisson sampling plan (without replacement), with properly chosen first order inclusion probabilities $(p_1,\ \ldots,\ p_N)$. The representation is called canonical if $\sum p_i = n$. In that case relationship between $p_i$ and $\pi_1^R$ are established in [5].

($ii$) If $n(\mathcal{S}) \neq n$, then reject it and go back to step ($i$), otherwise stop.

The vector $(p_1,\ \ldots,\ p_N)$ must be chosen in a way that the resulting first order inclusion probabilities coincide with $\pi_1^R,\ \ldots,\ \pi_N^R$, by means of a dedicated optimization algorithm, see [**?**]. The corresponding probability distribution is given by: $\forall \mathcal{S} \subset \mathcal{P}_N$,

$$R_N^R(\mathcal{S}) = \frac{R_N^P(\mathcal{S}) I\left\{\#\mathcal{S} = n\right\}}{\sum_{\{\mathcal{S}' \subset \mathcal{P}_N:\ \#\mathcal{S}' = n\}} R_N^P(\mathcal{S}')} \propto \prod_{i \in \mathcal{S}} p_i \prod_{i \notin \mathcal{S}} (1 - p_i) \times I\left\{\#\mathcal{S} = n\right\}\}$$

Refer to [5] p. 1496 for more details on the $p_i's$.

In addition, as shown in [5] (see p.1510 therein), the decomposition below holds uniformly, for all $i \in \{1, \ldots, N\}$:

$$p_i - \pi_i^R = \left(\frac{\bar{p}_N - p_i}{d_N} + o(1/d_N)\right) p_i(1 - p_i) = O(1/N).$$

This roughly means that the inclusion probabilities of the rejective sampling scheme are very close to those of the underlying Poisson design from which it was built. Therefore, the main Theorem in the Poisson case also holds when the sample is constructed with a rejective plan (because the rate of convergence is much more smaller than $N$) as revealed by the detailed proof.

**5. Conclusions** Following in the footsteps of [5], the asymptotic normality of an Horvitz-Thompson type of estimator of the tail index is investigated in particular survey sampling schemes (for survey designs which are "close" to the Poisson scheme). Extensions to other sampling plans are also possible by following our approach. Based on this limit result, the issue of building Gaussian confidence intervals for the tail index will be discussed during the talk.

## References

[1] Y.G. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Stat. Plan. Inf*, 67(2):209–226, 1998.

[2] L. de Haan and L. Peng. Comparison of tail index estimators. *Statist. Neerlandica*, 52:60–70, 1998.

[3] L. de Haan and S. Stadtmüller. Generalized regular variation of second order. *J. Austral. Math. Soc. Ser. A*, 61:381–295, 1996.

[4] C.M. Goldie and R.L. Smith. Slow variation with remainder: theory and applications. *Quart. J. Math. Oxford*, 38(1):45–71, 1987.

[5] J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[6] S.I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling.* Springer Verlag, 2007.