Estimation de l'indice de valeurs extrêmes à partir de données de sondage

Emilie Chautru [1], Patrice Bertail [2], Stéphan Clémençon [3]

[1] *Université de Cergy-Pontoise, 2 av. Adolphe Chauvin, 95302 Cergy-Pontoise cedex, emilie.chautru@gmail.com*

[2] *Université Paris-Ouest, 200 av. de la République, 92000 Nanterre, patrice.bertail@gmail.com*

[3] *Télécom ParisTech, 37/39 rue Dareau, 75014 Paris, stephan.clemencon@telecom-paristech.fr*

Résumé. Dans de nombreux domaines d'application de la statistique, il peut arriver que les données disponibles ne soient pas indépendantes et identiquement distribuées, mais issues d'un plan de sondage. A l'ère des "Big Data", échantillonner peut aussi être une solution naturelle aux problèmes informatiques induits par les quantités phénoménales de données. Comme en ignorer le processus de collecte peut conduire à un biais non-négligeable des estimateurs, une solution classique consiste à pondérer les observations par l'inverse de leur probabilité d'inclusion dans l'échantillon. Si de nombreux travaux ont d'ores et déjà été réalisés pour ainsi estimer sans biais et efficacement des quantités moyennes, à notre connaissance tel n'est pas le cas de la théorie des valeurs extrêmes. Tentant de faire le pont entre ces deux pans de théorie statistique, nous proposons ici une version Horvitz-Thompson du classique estimateur de Hill, qui évalue l'indice de valeurs extrêmes dans le cadre de plans de sondage à forte entropie de type Poisson. Après avoir démontré sa consistance et sa normalité asymptotique sous des hypothèses portant sur les probabilités d'inclusion et le modèle de surpopulation sous-jacent, nous illustrons nos résultats théoriques à l'aide de données simulées. Il en ressort en particulier qu'une calibration astucieuse des probabilités d'inclusion peut permettre de neutraliser la perte d'efficacité due à la phase d'échantillonnage.

Mots-clés. Sondage, théorie des valeurs extrêmes, estimateur de Hill, indice de valeurs extrêmes, probabilités d'inclusion, Horvitz-Thompson.

Abstract. In many application fields of theoretical statistics, the available observations are not independent and identically distributed, but originate from a potentially complex survey scheme. Moreover, in the "Big Data" era, sampling can be viewed as a natural solution to the computational issues induced by the immoderate size of databases. Since ignoring the survey scheme can impede estimation by introducing a non-negligible bias, it is customary to weight the data with the inverse of their probability of inclusion in the sample. While a plethora of analyzes has already been conducted to provide unbiased and efficient estimators of average quantities, to our knowledge, such is not the

case for phenomenons involving tails of distributions in the framework of extreme value theory. The analysis of extreme events is yet of major importance for risk management in a plurality of fields, ranging from biology or climatology to finance. In an attempt to conciliate both branches of statistics, we propose here a Horvitz-Thompson variant of the Hill estimator, which assesses the extreme value index when the observations are drawn according to a large entropy survey plan like the Poisson design. After having proved its consistency and asymptotic normality under a set of hypotheses involving the calculation of inclusion probabilities and the underlying superpopulation model, we illustrate our results on numerical experiments. It appears in particular that an appropriate choice of inclusion probabilities can neutralize the loss of efficiency due to the sampling phase.

Keywords. Survey sampling, extreme value theory, Hill estimator, extreme value index, inclusion probabilities, Horvitz-Thompson.

1 Estimation de l'indice de valeurs extrêmes à partir de données de sondage

Nous nous intéressons à la queue de distribution d'une certaine variable aléatoire X à valeurs dans $]0, +\infty[$ et de distribution $\mathbb{P}$, dont la fonction de répartition est de la forme

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

où $\gamma > 0$ et $L(x)$ est une fonction à variation lente ($\forall\, t > 0,\ L(x\,t)/L(x) \to 1$ lorsque $x \to +\infty$). De telles distributions sont dites à variation régulière. Le paramètre γ n'est alors autre que l'*indice de valeurs extrêmes* (IVE) et détermine l'épaisseur de la queue de la distribution: plus il est grand, moins les événement extrêmes sont rares. Dans un tel contexte, de nombreuses quantités d'intérêt en gestion du risque (*eg.* les très grands quantiles ou la faible probabilité de dépasser un seuil élevé) dépendent directement de l'IVE, qu'il est donc souhaitable d'estimer.

1.1 Cadre statistique

Nous considérons ici une population de taille N, notée $\mathcal{U}_N := \{1, \ldots, N\}$, sur laquelle des réalisations de X peuvent potentiellement être observées. Dans le cadre des "Big Data", elle peut correspondre par exemple à l'ensemble des unités d'une base de données stockée sur un ou plusieurs serveurs. Lorsque N est si grand que pour des raisons de coût (économique ou informatique) $\mathcal{U}_N$ n'est plus accessible dans son intégralité, il reste néanmoins possible d'en extraire un échantillon S de taille n (potentiellement aléatoire) beaucoup plus petite que N, sur lequel reposeront ensuite les analyses statistiques. Cette phase de sélection est généralement réalisée selon un plan de sondage spécifique.

1.1.1 Plan de sondage

Un plan de sondage n'est autre qu'une distribution de probabilité, notée ici R_N, portant sur $\mathcal{P}(\mathcal{U}_N)$, l'ensemble des parties de la population. La probabilité qu'un échantillon $s \in \mathcal{P}(\mathcal{U}_N)$ soit tiré parmi tous les échantillons possibles lorsque le plan R_N est utilisé s'écrit alors $\mathbb{P}_{R_N}(S = s) = R_N(s)$. Une manière alternative de définir un plan de sondage repose sur le vecteur de *variables d'inclusion* $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_N)$, dont les marginales sont des lois de Bernoulli de paramètres respectifs $\pi_1, \ldots, \pi_N$. Ces derniers sont naturellement appelés *probabilités d'inclusion d'ordre 1*, et la probabilité que l'unité $i \in \mathcal{U}_N$ soit sélectionnée vaut $\mathbb{P}(\epsilon_i = 1) = \pi_i$. De manière récurrente, on peut définir les probabilités d'inclusion d'ordre supérieur jusqu'à l'ordre N. En particulier, les *probabilités d'inclusion d'ordre 2* correspondent aux probabilités que deux unités distinctes i et j de la population soient simultanément retenues dans l'échantillon: $\pi_{i,j} := \mathbb{P}(\epsilon_i = 1, \epsilon_j = 1)$. Il est clair que tout échantillon s peut être identifié par un unique vecteur $\boldsymbol{\epsilon}_s \in \{0,1\}^N$. Si l'on note $\boldsymbol{\epsilon}$ le vecteur associé à l'échantillon aléatoire S, alors choisir la loi multivariée du vecteur $\boldsymbol{\epsilon}$ revient à choisir un plan de sondage R_N. Notons que dans ce cadre nous avons $\sum_{i=1}^N \epsilon_i = n$ et $\sum_{i=1}^N \pi_i = \mathbb{E}(n)$. Ainsi, les probabilités d'inclusion affectent directement la taille de l'échantillon.

Exemple : le pan de Poisson Le plan de sondage de Poisson, noté ici T_N, est obtenu lorsque toutes les variables d'inclusion sont indépendantes. Il est ainsi entièrement caractérisé par les probabilités d'inclusion du premier ordre.

Nous formulons ici un certain nombre d'hypothèses concernant R_N, sous lesquelles il sera montré que l'estimateur étudié possède de bonnes propriétés asymptotiques.

($\mathcal{H}_1$) Les probabilités d'inclusion ne peuvent pas devenir trop petites:
$\exists\, \pi_\star > 0 \ : \ \forall i \in \mathcal{U}_N,\ \pi_i \geq \pi_\star.$

($\mathcal{H}_2$) La population ne peut être sélectionnée dans son intégralité:
$\limsup_{N \to +\infty} \sum_{i=1}^N \pi_i < N$

($\mathcal{H}_3$) Les probabilités d'inclusion du second ordre sont proches de celles obtenues avec un tirage indépendant des unités:
$\exists\, \ell < +\infty \ : \ \forall\, N \geq 1,\ \max_{1 \leq i,\, j \leq N} |\pi_{i,j} - \pi_i \pi_j| \leq \frac{\ell}{n}.$

En pratique, les probabilités d'inclusion sont souvent déterminées en fonction de variables dites auxiliaires, accessibles sur l'ensemble de la population. Lorsqu'elles sont fortement corrélées à la variable d'intérêt, la performance des estimateurs peut alors être optimisée.

1.1.2 Information auxiliaire et modèle de surpopulation

Nous supposons qu'il existe un vecteur $\mathbf{W} := (W_1, \ldots, W_d)$ de variables auxiliaires à valeurs dans un espace $\mathcal{W} \subset \mathbb{R}^d$, $d \geq 1$. Sa distribution (resp. fonction de répartition) est notée $\mathbb{P}_{\mathbf{W}}$ (resp. $F_{\mathbf{W}}$), de marginales $\mathbb{P}_{W_1}, \ldots, \mathbb{P}_{W_d}$ (resp. $F_{W_1}, \ldots, F_{W_d}$), et la loi jointe de $\mathbf{W}$ et de la variable d'intérêt X est notée $\mathbb{P}_{X,\mathbf{W}}$. Quelques hypothèses concernant ces mesures de probabilités sont nécessaires dans nos analyses. Elles forment ce qui est appelé le modèle de surpopulation. Ici, nous supposons que les observations $(X_1, \mathbf{W}_1)$, $\ldots, (X_N, \mathbf{W}_N)$ sont des copies *indépendantes* du vecteur aléatoire $(X, \mathbf{W})$.

1.2 Version Horvitz-Thompson de l'estimateur de Hill

Si l'on note $X_{1,N} \leq \cdots \leq X_{N,N}$ les statistiques d'ordre dans la population, alors l'estimateur de Hill [1] calculé sur tout $\mathcal{U}_N$ peut s'écrire

$$H_{K,N} := \frac{1}{K} \sum_{i=1}^{K} \log\left(\frac{X_i}{X_{N-K,N}}\right) \mathbb{I}\{X_i > X_{N-K,N}\}.$$

Au sein d'un échantillon aléatoire caractérisé par le vecteur d'inclusion $\boldsymbol{\epsilon}$ de probabilités d'inclusion $\pi_1, \ldots, \pi_N$, nous en proposons une contrepartie pondérée à la manière Horvitz-Thompson [2] :

$$H^{\boldsymbol{\pi}}_{K,N} := \left(\sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i} \mathbb{I}\{X_i > X_{N-K,N}\}\right)^{-1} \sum_{i=1}^{K} \frac{\epsilon_i}{\pi_i} \log\left(\frac{X_i}{X_{N-K,N}}\right) \mathbb{I}\{X_i > X_{N-K,N}\}.$$

1.2.1 Consistance

Quelque soit le plan de sondage considéré, sous les hypothèses $(\mathcal{H}_1)$ à $(\mathcal{H}_3)$, dans le cadre du présent modèle de surpopulation, la consistance de $H^{\boldsymbol{\pi}}_{K,N}$ est assurée au même titre que celle de $H_{K,N}$.

Théorème 1 Soit $K = K(N)$ une suite d'entiers naturels telle que $K \to +\infty$ et $K/N \to 0$ lorsque $N, n \to +\infty$. Si les hypothèses $(\mathcal{H}_1)$ à $(\mathcal{H}_3)$ sont satisfaites, alors quand N et n tendent vers $+\infty$, nous avons $H^{\boldsymbol{\pi}}_{K,N} \xrightarrow{\mathbb{P}} \gamma$.

1.2.2 Normalité asymptotique

La normalité asymptotique de notre estimateur est d'abord démontrée dans le cadre du plan de sondage de Poisson T_N de probabilités d'inclusion $p_1, \ldots, p_N$. Une hypothèse supplémentaire, déjà nécessaire pour garantir la normalité asymptotique de $H_{K,N}$, est requise:

($\mathcal{H}_4$) Soit $U(x) := \inf\{y \in]0, +\infty[: \ F(y) \geq 1 - 1/x\}$ la fonction quantile extrême de la variable d'intérêt X. Il existe $\rho < 0$, appelé *paramètre de second ordre*, ainsi qu'une fonction de signe constant A satisfaisant $\lim_{x\to+\infty} A(x) = 0$, tels que pour tout $t > 0$,

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

Théorème 2 Supposons que les hypothèses ($\mathcal{H}_1$), ($\mathcal{H}_2$) et ($\mathcal{H}_4$) sont satisfaites et que lorsque $N, K \to \infty$, $K/N \to 0$, nous avons

$$\mathbb{E}\left(\frac{1}{p(\mathbf{W})} \;\middle|\; X > U(N/K)\right) \to \sigma_p^2 < \infty,$$

et $\sqrt{K}A(N/K) \to \lambda$ pour une certaine constante $\lambda \in \mathbb{R}$. Alors nous avons la convergence en distribution lorsque $N \to +\infty$:

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

Ce théorème est ensuite étendu au cas du plan de sondage dit réjectif, en en contrôlant la distance au plan de Poisson comme dans les travaux de Hàjek [3] et Berger [4]. Sous quelques hypothèses supplémentaires concernant la loi jointe $\mathbb{P}_{X,\mathbf{W}}$, nous montrons finalement comment calibrer les probabilités d'inclusion de manière à minimiser l'impact de l'échantillonnage sur la variance, en résolvant le programme d'optimisation suivant:

$$\min_{p_1,\ldots,p_N} \sigma_p^2 \ \text{ sous contrainte que } \ \begin{cases} \lim_{N\to\infty} \frac{\mathbb{E}(n)}{N} < 1, \\ \forall\, w \in \mathcal{W},\ 0 < p_\star \leq p(w) \leq 1. \end{cases}$$

Bibliographie

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

[2] Horvitz, D.G. et Thompson, D.J. (1951), *A generalization of sampling without replacement from a finite universe*, JASA, 47, 663–685.

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

[4] Berger, Y.G. (1998), *Rate of convergence to normal distribution for the Horvitz-Thompson estimator*, Journal of Statistical Planning and Inference, 67, 209–226.