PROCESSUS EMPIRIQUE DANS LES SONDAGES

Patrice Bertail [1], Emilie Chautru [2], Stéphan Clémençon[3]
[1] *Université Paris-Ouest, 200 ave de la République 92000 Nanterre ,*
patrice.bertail@gmail.com
[2] *Université de Cergy, Emilie.Chautru@gmail.com*
[3] *ENST, 37-39 rue Darreau, 75014 Paris Stephan.clemencon@gmail.com*

Résumé. Cet exposé présente une étude du comportement du processus empirique dans le cadre non i.i.d. des sondages et est motivé par des problèmes liés au traitement de grandes masses de données (big data). Dans de nombreuses situations les statisticiens ont à leur disposition non pas des données i.i.d. mais des données issues de sondages avec des poids déterminés par un plan de sondage spécifique. Dans certaines situations de données massives (big data), un plan de sondage peut être également un moyen efficace de réduire la taille du problème. Pour étudier des procédures de minimisation de risques empiriques issues d'algorithmes complexes, il est naturel d'étudier dans un premier temps le comportement asymptotique d'une version adéquate du processus empirique. Notre but est d'étudier comment incorporer le plan de sondage dans l'estimation uniforme d'une mesure de probabilité P sur un espace mesurable, vue comme un opérateur linéaire agissant sur une classe de fonctions $\mathcal{F}$ afin d'obtenir des résultats de normalité asymptotique. Notre but est d'étudier des plans de sondages plus généraux proches en un certain sens des plans de sondages Poissoniens et suit plus particulièrement les approches de Hajek(1964) et Berger(1998). Ceci inclue en particulier les plans rejectifs et les plans de type Rao-Sandford. Le principal résultat de ce travail est un théorème fonctionel donnant le comportement asymptotique du processus empirique repondéré (ou processus empirique de Horvitz-Thompson), indéxé par une classe de fonction dans un modèle de superpopulation.

Mots-clés. Sondages, probabilité d'inclusion, Horvitz-Thompson, processus empirique, Classe de fonctions.

Abstract. This paper is devoted to the study of the limit behavior of extensions of the empirical process, when the data available have been collected through an explicit survey sampling scheme and is motivated by some problems linked to practical exploitation of big datas. Indeed, in many situations, statisticians have at their disposal not only data but also weights arising from some survey sampling plans. On the other hand,for big data, survey sampling may be an efficent tool to reduce the size of computational costs involved by massive data. Our main goal is here to investigate how to incorporate the survey scheme into the inference procedure dedicated to the estimation of a probability measure Pon a measurable space (viewed as a linear operator acting on a certain class of functions $\mathcal{F}$), in order to guarantee its asymptotic normality. limit results. Our approach follows that of Hajek(1964), extended next by Berger(1998), and is applicable to general sampling surveys, namely those with unequal first order inclusion probabilities which are of the Poisson type or sequential/rejective. The main result of the paper is a

Functional Central Limit Theorem (FCLT) describing the limit behavior of an adequate version of the empirical process (referred to as the Horvitz-Thompson empirical process throughout the article) in a superpopulation statistical framework.

Keywords. Survey sampling, inclusion probability, Horvitz-Thompson, empirical process, Classes of functions.

1 EMPIRICAL PROCESS IN SURVEY SAMPLING

1.1 Survey sampling

We consider a finite population of size $N \geq 1$, $\mathcal{U}_N := \{1, \ldots, N\}$ say. A *sample* $\mathcal{S}$ of size $n(\mathcal{S}) \leq N$ is any subset $s := \{i_1, \ldots, i_{n(s)}\} \subset \mathcal{U}_N$ with cardinality $n(s) \leq N$. A sampling scheme (design/plan) is determined by a discrete probability distribution R_N on $\mathcal{P}(\mathcal{U}_N)$. For any $i \in \mathcal{U}_N$, the quantity usually referred to as the i-th (first order) *inclusion probability*

$$\pi_i(R_N) := \mathbb{P}_{R_N}\{i \in S\} = \sum_{s \in \mathcal{P}(\mathcal{U}_N)} R_N(s)\, \mathbb{I}\{i \in s\},$$

is the probability that the individual labeled i belongs to a random sample S under the survey scheme R_N. When there is no ambiguity, we will simplify notations and write π_i instead of $\pi_i(R_N)$.The information related to the observed sample S is encapsulated by the random vector $\boldsymbol{\epsilon}_{(N)} := (\epsilon_1, \ldots, \epsilon_N)$, where

$$\epsilon_i := \mathbb{I}\{i \in S\} = \begin{cases} 1 & \text{with probability } \pi_i, \\ 0 & \text{with probability } 1 - \pi_i. \end{cases}.$$

We will assume

H1 There exist $\pi_\star > 0$ and $N_0 \in \mathbb{N}^*$ such that for all $N \geq N_0$ and $i \in \mathcal{U}_N$, $\pi_i(R_N) > \pi_\star$. and $\limsup_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} \pi_i(R_N) < 1$.

1.2 The Horvitz-Thompson empirical process indexed by classes of function

Troughout the article, we assume that the class $\mathcal{F}$ admits a square integrable envelope H, as defined below.

H2 There exists a measurable function $H : \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} H^2(\mathbf{x})\, \mathbb{P}(d\mathbf{x}) < \infty$ and $|f(\mathbf{x})| \leq H(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and any $f \in \mathcal{F}$ and there exists $\eta > 0$ such that $H(\mathbf{x}) > \eta$.

When viewed as a linear operator acting on $\mathcal{F}$, a probability measure $\mathbb{P}$ may then be considered as an element of $\ell^{\infty}(\mathcal{F})$, the space of all maps $\Phi : \mathcal{F} \to \mathbb{R}$ such that

$$\|\Phi\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\Phi(f)| < +\infty,$$

equipped with the uniform convergence norm (or, equivalently, with Zolotarev metric), namely $\|\mathbb{P} - \mathbb{Q}\|_{\mathcal{F}} := \sup_{h \in \mathcal{F}} \left| \int h \, d\mathbb{P} - \int h \, d\mathbb{Q} \right|$,for any couple of probability measures $\mathbb{P}$ and $\mathbb{Q}$. See Van der Vaart and Wellner (2000) for details.

1.2.1 The Horvitz-Thompson empirical process

The Horvitz-Thompson estimator of the empirical probability $\mathbb{P}_N = N^{-1} \sum_{i=1}^{N} \delta_{\mathbf{X}_i}$ based on the survey data described above is defined as follows,

$$\mathbb{P}_{R_N}^{\boldsymbol{\pi}(R_N)} := \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i} \delta_{\mathbf{X}_i}.$$

Pointwise consistency and asymptotic normality of the estimator in can be deduce from Hajek(1964) and Berger(1998), as N tends to infinity. When considering the estimation of measure $\mathbb{P}_N$ (the measure of interest in survey sampling) over a class of functions $\mathcal{F}$, we are led to the asymptotic study of the collection of random processes $\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} := \left(\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} f \right)_{f \in \mathcal{F}}$, where

$$\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} f := \sqrt{N} \left(\mathbb{P}_{R_N}^{\boldsymbol{\pi}(R_N)} - \mathbb{P}_N \right) f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\epsilon_i}{\pi_i(R_N)} - 1 \right) f(\mathbf{X}_i),$$

which shall be referred to as the $\mathcal{F}$-indexed *Horvitz-Thomson empirical process.*

1.2.2 Poisson Horvitz-Thompson empirical process

The Poisson sampling scheme T_N has been the subject of much attention, especially in Hajek(1964), where asymptotic normality of (pointwise) Horvitz-Thompson estimators have been established in this specific case. Following in the footsteps of this seminal contribution, we consider the following Poisson like version of the empirical process rather than the original process : for all $f \in \mathcal{F}$,

$$\widetilde{\mathbb{G}}_{R_N}^{\mathbf{p}} f := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\epsilon_i - p_i) \left(\frac{f(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(f) \right), \quad f \in \mathcal{F},$$

$$\theta_{N,\mathbf{p}}(f) := \frac{1}{D_N} \sum_{i=1}^{N} (1 - p_i) f(\mathbf{X}_i) \quad \text{and} \quad D_N := \sum_{i=1}^{N} p_i (1 - p_i).$$

Assume that the inclusion probabilities are chosen according to some auxiliary random variable $\mathbf{W}$, valued in some measurable space $\mathcal{W}$ and observed over the whole population: $\mathbf{W}_{(N)} = (\mathbf{W}_1, \dots, \mathbf{W}_N)$.

H3 The pairs of random vectors $(\mathbf{X}_1, \mathbf{W}_1), \dots, (\mathbf{X}_N, \mathbf{W}_N)$ are iid (exchangeable at least) with distribution $\mathbb{P}_{\mathbf{X},\mathbf{W}}$. In addition, the conditional inclusion probabilities $\mathbf{p} := (p_1, \dots, p_N)$ are then given for all $i \in \{1, \dots, N\}$ and $\mathbf{W}_{(N)} \in \mathcal{W}^N$ by $p_i := p(\mathbf{W}_i) = E(\varepsilon_i|W_i)$ with $\int \frac{1}{p(w)^2}\,\mathbb{P}_{\mathbf{X},\mathbf{w}}(d\mathbf{x}, d\mathbf{w}) < \infty$.

Theorem Suppose that H1-H3 as well as technical measurability assumptions, hold, as well as the following conditions.

i) *Lindeberg-Feller type condition*: $\forall\,\eta > 0,\ \mathbb{E}\left((\mathcal{Z}_{N,i})^2\,\mathbb{I}\left\{\mathcal{Z}_{N,i} > \eta\sqrt{N}\right\}\right) \to 0$,with $\mathcal{Z}_{N,i} := (\epsilon_i - p(\mathbf{W}_i))\ \sup_{f\in\mathcal{F}}\left|\frac{f(\mathbf{X}_i)}{p(\mathbf{W}_i)} - \theta_{N,\mathbf{p}}(f)\right|$.

ii) *Uniform entropy condition*: let $\mathcal{D}$ be the set of all finitely discrete probability measures defined and assume that the covering numbers satify the uniform entropy condition

$$\int_0^\infty \sup_{\mathbb{Q}\in\mathcal{D}} \sqrt{\log(N(\varepsilon\|H\|_{2,Q}, \mathcal{F}, \|.\|_{2,\mathbb{Q}})}\, d\varepsilon < \infty.$$

Then there exists a $\rho_{\mathbb{P}}$-equicontinuous Gaussian process $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$ with covariance operator Σ given by

$$\Sigma(f,g) := \int_{\mathcal{X}\times\mathcal{W}} f(\mathbf{x})g(\mathbf{x})\left(\frac{1}{p(\mathbf{w})} - 1\right)\mathbb{P}_{\mathbf{X},\mathbf{W}}(d\mathbf{x}, d\mathbf{w}) - \theta_p(f)\theta_p(g)\,D_p, \tag{1}$$

with

$$\theta_p(f) := \frac{1}{\int_{\mathcal{W}}(1 - p(\mathbf{w}))\,p(\mathbf{w})\,\mathbb{P}_{\mathbf{W}}(d\mathbf{w})}\int_{\mathcal{X}\times\mathcal{W}}(1 - p(\mathbf{w}))\,f(\mathbf{x})\,\mathbb{P}_{\mathbf{X},\mathbf{W}}(d\mathbf{x}, d\mathbf{w}).$$

such that

$$\widetilde{\mathbb{G}}^{\mathbf{p}}_{T_N} \Rightarrow \mathbb{G} \text{ weakly in } \ell^\infty(\mathcal{F}),\ \text{ as } N \to \infty.$$

1.2.3 Generalization to other sampling plans including rejective sampling, Rao Sampford etc...

In order to formulate the approximation result needed in the sequel, we introduce, for two sampling designs R_N and T_N, the *total variation metric*

$$\|R_N - T_N\|_1 := \sum_{s\in\mathcal{P}(\mathcal{U}_N)} |R_N(s) - T_N(s)|\,,$$

as well as the *entropy*

$$D(T_N, R_N) := \sum_{s \in \mathcal{P}(\mathcal{U}_N)} T_N(s) \, \log \left(\frac{T_N(s)}{R_N(s)} \right).$$

In practice, T_N will typically be the Poisson sampling plan investigated in the previous subsection and $\widetilde{\mathbb{G}}_{T_N}^{\boldsymbol{\pi}(T_N)}$ the corresponding Poisson like empirical process.

Theorem Let R_N and T_N be two sampling designs and assume that T_N is entirely characterized by its first order inclusion probabilities, $\boldsymbol{\pi}(T_N)$. Then, the empirical processes $\widetilde{\mathbb{G}}_{T_N}^{\boldsymbol{\pi}(T_N)}$ and $\widetilde{\mathbb{G}}_{R_N}^{\boldsymbol{\pi}(T_N)}$ valued in $\ell^\infty(\mathcal{F})$ satisfy the relationships:

$$d_{BL} \left(\widetilde{\mathbb{G}}_{T_N}^{\boldsymbol{\pi}(T_N)}, \widetilde{\mathbb{G}}_{R_N}^{\boldsymbol{\pi}(T_N)} \right) \leq \|R_N - T_N\|_1 \leq \sqrt{2 \, D(T_N, R_N)}.$$

A consequence of this result is that for sampling plans R_N such that the L1 norm or the entropy is close to the Poisson sampling plan T_N $\|R_N - T_N\|_1 \to 0$ or $D(T_N, R_N) \to 0$ as $N \to \infty$ and for which it is possible to appoximate its inclusion probability by the one of a Poisson sampling plan, then, there exists a $\rho_{\mathbb{P}}$-equicontinuous Gaussian process $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$ with covariance operator Σ given by such that $\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} \Rightarrow \mathbb{G}$ weakly in $\ell^\infty(\mathcal{F})$, as $N \to \infty$. This include Rejective Sampling, Rao-Sampford sampling, sucessive sampling etc...

Bibliographie

[1] Hajek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. Ann. Math. Statist., 35, 1419-1880
[2] Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. Journal of Statistical Planning and Inference, 67:209–226.
[3] van der Vaart A.W., Wellner, J.A. (2000). Weak Convergence and Empirical Processes: With Applications to Statistics, Springer Series in Statistics.