Empirical processes in survey sampling

Patrice Bertail, Emilie Chautru, Stéphan Clémençon

Empirical processes in survey sampling

Patrice Bertail[a,b], Emilie Chautru[c], and Stéphan Clémençon[d]

[a]MODAL'X - Université Paris Ouest

[b]Laboratoire de Statistique - CREST

[c]Laboratoire AGM - Université de Cergy-Pontoise

[d]Institut Télécom - LTCI UMR Télécom ParisTech/CNRS No. 5141

Abstract

It is the main purpose of this paper to study the asymptotics of variants of the empirical process in the context of survey data. Precisely, a functional central limit theorem is established when the sample is picked by means of a Poisson survey scheme. This preliminary result is then extended to the case of the rejective or conditional Poisson sampling case, and in turn to high entropy survey sampling plans which are close to the rejective design in the sense of the Bounded-Lipschitz distance. The framework we develop encompasses survey sampling designs with non-uniform first order inclusion probabilities, which can be defined so as to optimize estimation accuracy. Applications to Hadamard- and Fréchet-differentiable functionals are also considered together with the construction of uniform confidence bands of the cumulative distribution function. Related simulation results are displayed for illustration purpose.

Keywords: empirical processes; survey sampling; rejective scheme; Poisson scheme; functional central limit theorem; coupling

1 Introduction

This paper is devoted to the study of the limit behavior of extensions of the empirical process based on independent and identically distributed (iid) samples, when the data available have been collected through an explicit survey sampling scheme. Indeed, in many situations, statisticians have at their disposal not only data but also weights arising from some survey sampling plans. They 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, for instance). In most cases, the survey scheme is ignored, potentially yielding a significant sampling bias. When considering some functional of the empirical process, this may cause severe drawbacks and completely jeopardize the estimation, as can be revealed by simulation experiments (Bonnéry et al., 2011). From another point of view, when the available data is so voluminous that a single computer cannot treat the entire dataset, survey sampling appears as a natural remedy (Cardot et al., 2013). In particular, as opposed to simple sub-sampling, it permits to control the efficiency of estimators via the strategic definition of unequal survey weights.

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 $\mathbb{P}$ on a measurable space (viewed as a linear operator acting on a certain class of functions $\mathcal{F}$), in order to guarantee its asymptotic normality. This problem has been addressed by Breslow and Wellner (2007) and Gill et al. (1988) in the particular case of a stratified survey sampling, where the individuals are selected at random (without replacement) in each stratum, by means of bootstrap limit results. Our approach is different and follows (and extends) that of Hàjek (1964), considered next by Berger (1998, 2011), and is applicable to more general sampling surveys, namely those with unequal first order inclusion probabilities which are of the Poisson type or sequential/rejective. Such sampling designs have the advantage of allowing a fine control over the variance of

estimators via the preliminary definition of survey weights. In the context of big data, they can even be optimized to guarantee an efficiency almost as good as if the entire dataset was reachable.

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. The key argument involved in this asymptotic analysis consists in approximating the distribution of the extended empirical process by that related to a much simpler sampling plan. In order to illustrate the reach of this result, statistical applications are considered, where the extensions of the empirical process are used to construct confidence bands around the Horvitz-Thompson estimator of the cumulative distribution function.

The paper is organized as follows. In Section 2, the statistical framework is described at length, notations are set out and some basics on survey sampling theory are recalled, together with important examples of survey schemes to which the subsequent asymptotic analysis can be applied. The main result of the paper, a FCLT for the Horvitz-Thompson empirical process is stated in Section 4, after a thorough description of the processes of interest in Section 3. Supplementary Materials are also available for this article. They include in particular applications to nonparametric functional estimation, numerical experiments on confidence band interval construction for the cumulative distribution function and technical details.

2 Background and Preliminaries

We start off with recalling some crucial notions in survey sampling and in modern empirical process theory, which shall be extensively used in the subsequent analysis. Throughout the article, the Dirac mass at $\mathbf{x}$ in some vector space $\mathcal{X}$ is denoted by $\delta_{\mathbf{x}}$ and the indicator function of any event E by $\mathbb{I}\{\mathsf{E}\}$. We also

denote by $\#\mathcal{E}$ the cardinality of any finite set $\mathcal{E}$, and by $\mathcal{P}(\mathcal{E})$ its power set.

2.1 Survey sampling: some basics

The purpose of survey sampling is to study some characteristics of a population $\mathcal{U}_N$ of $N \geq 1$ units (or individuals) identified by an arbitrary collection of labels: $\mathcal{U}_N := \{1, \ldots, N\}$. When it is not possible to reach the whole population (*e.g.* with big data), the features of interest can be estimated from a finite, relatively small number of its elements, namely a *sample* $s := \{i_1, \ldots, i_{n(s)}\} \subset \mathcal{U}_N$ of size $n(s) \leq N$, selected at *random* within $\mathcal{U}_N$ (see for instance Tillé, 2006, Chapter 1, Tillé, 1999 or Gourieroux, 1981 for an introduction to random sampling). Equipped with this representation, a *sampling scheme* (design/plan) is determined by a discrete probability measure R_N on $\mathcal{P}(\mathcal{U}_N)$, the set of all possible samples in $\mathcal{U}_N$. Depending on the adopted point of view, like in superpopulation models, the characteristics of the population can be considered random too. In the next paragraphs, crucial concepts and notations are introduced concerning both sources of hazard.

2.1.1 Survey schemes without replacement

Consider a sampling scheme R_N *without replacement*; our analysis is restricted to this popular family of survey plans. By definition, we always have

$$\forall\, s \in \mathcal{P}(\mathcal{U}_N),\, R_N(s) \geq 0 \quad \text{and} \quad \sum_{s \in \mathcal{P}(\mathcal{U}_N)} R_N(s) = 1,$$

and the mean survey sample size is given by

$$\mathbb{E}_{R_N}(n(S)) = \sum_{s \in \mathcal{P}(\mathcal{U}_N)} n(s)\, R_N(s).$$

Here, the notation $\mathbb{E}_{R_N}(.)$ denotes the expectation taken with respect to the random sample S with distribution R_N. In a similar fashion, $\mathbb{P}_{R_N}(S \in \mathcal{S})$ refers to the probability of the event $\{S \in \mathcal{S}\}$ with $\mathcal{S} \subset \mathcal{P}(\mathcal{U}_N)$, when S is drawn from R_N. In particular, $R_N(s) = \mathbb{P}_{R_N}(S = s)$. Such distributions are entirely characterized by the concepts listed below.

Inclusion probabilities 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 on the sampling design, notations will be simplified and π_i will be used instead of $\pi_i(R_N)$. In the subsequent analysis, first order inclusion probabilities are assumed to be strictly positive: $\forall\, i \in \mathcal{U}_N,\ \pi_i(R_N) > 0$. We shall even require the stronger hypothesis that they never get either too small or too large, as formally stated below.

Assumption 2.1 There exist $\pi_\star > 0$ and $N_0 \in \mathbb{N}^*$ such that for all $N \geq N_0$ and $i \in \mathcal{U}_N$, $\pi_i(R_N) > \pi_\star$. In addition, $\limsup_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} \pi_i(R_N) < 1$.

When the first condition holds, the rate of convergence of the estimators considered in Section 3 and Section 4 will be shown to be typically of order $1/\sqrt{N}$. One could possibly relax it and allow $\pi_\star$ to depend on N, with $\pi_\star = \pi_\star(N)$ decaying to zero as N tends to infinity at a specific rate, and still be able to establish limit results. The analysis would be however much more technical; this is left for further research.

Conditions involving the *second order inclusion probabilities* shall also be used in our asymptotic analysis. They are denoted by

$$\pi_{i,j}(R_N) := \mathbb{P}_{R_N}\left((i,j) \in S^2\right) = \sum_{s \in \mathcal{P}(\mathcal{U}_N)} R_N(s)\, \mathbb{I}\{\{i,j\} \subset s\},$$

for all $(i,j) \in \mathcal{U}_N^2$. In other words, $\pi_{i,j}(R_N)$ is the probability that two distinct individuals labeled i and j are jointly selected under design R_N. Again, $\pi_{i,j}$ may eventually be used when there is no need to emphasize the dependency on the sampling plan R_N. Notice that higher order inclusion probabilities may be defined in a similar way, up to the maximal order for which the entire population is selected.

Inclusion indicators The information related to the observed sample S is encapsulated by the random vector $\boldsymbol{\epsilon}_{(N)} := (\epsilon_1, \ldots, \epsilon_N)$, where

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

Notice indeed that the set $\mathcal{P}(\mathcal{U}_N)$ of all possible samples is in one-to-one correspondence with $\{0,1\}^N$, which provides a handy alternative representation of sampling schemes. Again, for simplicity, the subscript (N) shall be omitted when no ambiguity is possible. By definition, the distribution of $\boldsymbol{\epsilon} := \boldsymbol{\epsilon}_{(N)}$ has univariate marginals that correspond to the Bernoulli distributions $\mathcal{B}(\pi_i)$, $i \in \mathcal{U}_N$, and covariance matrix given by

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

Incidentally we have $\sum_{i=1}^N \epsilon_i = n(S)$ and thus $\sum_{i=1}^N \pi_i = \mathbb{E}_{R_N}(n(S))$.

Before considering the issue of extending the concept of empirical process in the context of survey sampling, we recall a few important classes of survey schemes, to which the results established in Section 3 and Section 4 can be applied. One may refer to Deville (1987) for instance for an excellent account of survey theory, including many more examples of sampling designs.

Example 2.1 – *Simple Random Sampling Without Replacement.* A simple random sampling without replacement (SRSWOR in abbreviated form) is a sampling design of fixed size $n(S) = n$, according to which all samples with cardinality n in the population $\mathcal{U}_N$ are equally likely to be chosen, with probability $(N-n)!/n!$. It follows that all units of $\mathcal{U}_N$ have the same chance of being selected, n/N namely, and all second order probabilities are equal to $n(n-1)/(N(N-1))$.

Example 2.2 – *Poisson survey sampling.* The Poisson sampling plan without replacement (POISSWOR), denoted here by T_N, is one of the simplest survey

schemes. In this case, the N elements of $\boldsymbol{\epsilon}$ are *independent* Bernoulli random variables with respective parameters $\pi_i(T_N) =: p_i$, $i \in \{1, \ldots, N\}$ so that for any sample $s \in \mathcal{P}(\mathcal{U}_N)$,

$$T_N(s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)\,.$$

Notice that the size $n(S)$ of sample S with distribution T_N is random (except in the sole situation where $p_i \in \{0, 1\}$ for $i = 1, \ldots, N$) and that the corresponding survey plan is fully characterized by the first order inclusion probabilities. In the specific situation where they are all equal, *i.e.* $p_1 = \cdots = p_N = p$, the design is called Bernoulli.

Example 2.3 – *Stratified sampling.* A stratified sampling design permits to draw a sample S of fixed size $n(S) = n \leq N$ within a population $\mathcal{U}_N$ that can be partitioned into $K \geq 1$ distinct strata $\mathcal{U}_{N_1}, \ldots, \mathcal{U}_{N_K}$ (known a priori) of respective sizes $N_1, \ldots, N_K$ adding up to N. Let $n_1, \ldots, n_K$ be non-negative integers such that $n_1 + \cdots + n_K = n$, then the drawing procedure is implemented in K steps: within each stratum $\mathcal{U}_{N_k}$, $k \in \{1, \ldots, K\}$, perform a SRSWOR of size $n_k \leq N_k$ yielding a sample S_k. The final sample is obtained by assembling these sub-samples: $S = \bigcup_{k=1}^{K} S_k$. The probability of drawing a specific sample s by means of this survey scheme is

$$R_N^{\mathrm{str}}(s) = \sum_{k=1}^{K} \binom{N_k}{n_k}^{-1}.$$

Naturally, first and second order inclusion probabilities depend on the stratum to which each unit belong: for all $i \neq j$ in $\mathcal{U}_N$,

$$\pi_i(R_N^{\mathrm{str}}) = \sum_{k=1}^{K} \frac{n_k}{N_k} \mathbb{I}\{i \in \mathcal{U}_{N_k}\}$$

$$\text{and} \quad \pi_{i,j}(R_N^{\mathrm{str}}) = \sum_{k=1}^{K} \frac{n_k(n_k - 1)}{N_k(N_k - 1)} \mathbb{I}\left\{(i, j) \in \mathcal{U}_{N_k}^2\right\}.$$

Example 2.4 – *Canonical Rejective Sampling.* Let $n \leq N$ and consider a vector $\boldsymbol{\pi}^R := (\pi_1^R, \ldots, \pi_N^R)$ of first order inclusion probabilities. Further define $\mathcal{S}_n :=$

$\{s \in \mathcal{P}(\mathcal{U}_N) : \#s = n\}$, the set of all samples in population $\mathcal{U}_N$ with cardinality n. The rejective sampling (Hàjek, 1964; Berger, 1998), sometimes called conditional Poisson sampling (CPS), exponential design without replacement or maximum entropy design (Tillé, 2006, Section 5.6), is the sampling design R_N^R that selects samples of fixed size $n(s) = n$ so as to maximize the entropy measure

$$H(R_N) = -\sum_{s \in \mathcal{S}_n} R_N(s) \log R_N(s),$$

subject to the constraint that its vector of first order inclusion probabilities coincides with $\boldsymbol{\pi}^R$. It is easily implemented in two steps:

1. draw a sample S with a POISSWOR plan $T_N = T_N^P$, with properly chosen first order inclusion probabilities vector $\mathbf{p} := (p_1, \ldots, p_N)$. The representation is called canonical if $\sum_{i=1}^N p_i = n$. In that case, relationships between each p_i and π_i^R, $1 \leq i \leq N$, are established in Hàjek (1964).

2. If $n(S) \neq n$, then reject sample S and go back to step one, otherwise stop.

Vector $\mathbf{p}$ must be chosen in a way that the resulting first order inclusion probabilities coincide with $\boldsymbol{\pi}^R$, by means of a dedicated optimization algorithm (Tillé, 2006, Algorithms 5.5 to 5.9). The corresponding probability distribution is given for all $s \in \mathcal{P}(\mathcal{U}_N)$ by

$$R_N^R(s) = \frac{T_N^P(s)\,\mathbb{I}\{\#s = n\}}{\sum_{s' \in \mathcal{S}_n} T_N^P(s')} \propto \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i) \times \mathbb{I}\{\#s = n\},$$

where $\propto$ denotes the proportionality. We refer to Hàjek (1964, p.1496) for more details on the links between rejective and Poisson sampling plans.

Example 2.5 – *Rao-Sampford Sampling.* The Rao-Sampford sampling design generates samples $s \in \mathcal{P}(\mathcal{U}_N)$ of fixed size $n(s) = n$ with respect to some given first order inclusion probabilities $\boldsymbol{\pi}^{RS} := (\pi_1^{RS}, \ldots, \pi_N^{RS})$, fulfilling the condition $\sum_{i=1}^N \pi_i^{RS} = n$, with probability

$$R_N^{RS}(s) = \eta \sum_{i \in s} \pi_i^{RS} \prod_{j \notin s} \frac{\pi_j^{RS}}{1 - \pi_j^{RS}}.$$

Here, $\eta > 0$ is chosen such that $\sum_{s \in \mathcal{P}(\mathcal{U}_N)} R_N^{RS}(s) = 1$. In practice, the following algorithm is often used to implement such a design (Berger, 1998):

1. select the first unit i with probability π_i^{RS}/n,
2. select the remaining $n-1$ units j with drawing probabilities proportional to $\pi_j^{RS}/(1-\pi_j^{RS})$, $j = 1, \ldots, N$,
3. accept the sample if the units drawn are all distinct, otherwise reject it and go back to step one.

2.1.2 Superpopulation models

The characteristics of interest in population $\mathcal{U}_N$ are modeled as follows. We consider the probability space $(\mathcal{U}_N, \mathcal{P}(\mathcal{U}_N), \mathbb{P})$ and a random variable/vector $\mathbf{X}$ defined on the latter, taking its values in a *Banach* space $(\mathcal{X}, \|.\|)$, with probability measure $\mathbb{P}$. We set

$$\mathbf{X} : \begin{pmatrix} \mathcal{U}_N & \longrightarrow & \mathcal{X} \\ i & \longmapsto & \mathbf{X}(i) =: \mathbf{X}_i \end{pmatrix},$$

and the σ-algebra induced by the normed vector space topology structure of $\mathcal{X}$ is denoted by $\mathcal{A}$. Then, the studied features correspond to some synoptic mapping $(\mathbf{X}_1, \ldots, \mathbf{X}_N) \mapsto f(\mathbf{X}_1, \ldots, \mathbf{X}_N)$.

In survey sampling, a *superpopulation* is basically an imaginary infinite population, $\mathcal{U}_\infty$ say, from which $\mathcal{U}_N$ is supposed to be issued. In a model-based approach, it is assumed that the random vectors of interest $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are in fact realizations of N random vectors $\widetilde{\mathbf{X}}_j : \mathcal{U}_\infty \to \mathcal{X}$, $1 \leq j \leq N$, with joint distribution $\mathbb{Q}$. Then, a *superpopulation model* is simply a set of conditions that characterize $\mathbb{Q}$ (Droesbeke et al., 1987, Chapter 4). The main advantage of such a framework is that it often facilitates statistical inference; in particular, it permits the development of an asymptotic theory, when sample and population sizes grow conjointly to infinity. The superpopulation model we consider here

stipulates that all N random vectors $\mathbf{X}_i$, $i \in \mathcal{U}_N$, are independent identically distributed (iid) with common distribution $\mathbb{P}$, *i.e.* $\mathbb{Q} = \mathbb{P}^{\otimes N}$, where $\otimes$ denotes the tensor product of measures.

Remark 2.1 The most celebrated iid superpopulation model that we adopt here establishes a setting very similar to that of weighted bootstrap (Arcones and Giné, 1992; Barbe and Bertail, 1995): the original iid N-sample there would correspond to the complete vector $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$, from which sub-samples are drawn according to some procedure likened to the survey scheme. Actually, both approaches are completely equivalent if the survey weights $(\epsilon_1/\pi_1, \ldots, \epsilon_N/\pi_N)$ are exchangeable (*i.e.* the N-variate distribution of this vector is invariant to the order of its elements). For instance, in the specific case of stratified sampling, drawing units with equal probabilities in each stratum (with a finite and given stratum-size) amounts to bootstrapping (without replacement) in some given cell. It is not surprising then that both Breslow and Wellner (2007) and Saegusa and Wellner (2011) construct a general asymptotic theory in two-phase sampling by using bootstrap type results.

2.1.3 Auxiliary information

In practice, sampling from a population $\mathcal{U}_N$ is only possible if all individuals are listed somehow, and can be identified once selected. Such documents are called survey frames; in the case of social surveys, they are collected by government institutions and often provide some minimal information about its components. These *auxiliary variables*, supposedly known for all $i \in \mathcal{U}_N$, can sometimes be used to optimize in some sense the survey scheme. In a superpopulation framework, we denote by $\mathbf{W}$ the auxiliary random vector, valued in some measurable space $\mathcal{W}$, and set $\mathbf{W}_{(N)} := (\mathbf{W}_1, \ldots, \mathbf{W}_N)$. As soon as $\mathbf{W}$ is correlated with $\mathbf{X}$, the vector of interest, it becomes possible to boost the efficiency of estimators by defining inclusion probabilities as a function of $\mathbf{W}_{(N)}$ (Droesbeke et al., 1987).

In the present analysis, we denote by $\mathbb{P}_{\mathbf{X},\mathbf{W}}$ the joint distribution of $(\mathbf{X}, \mathbf{W})$ and by $\mathbb{P}_{\mathbf{W}}$ the marginal distribution of $\mathbf{W}$. Like in most applications, we as-

sume that the $\mathbf{W}_i$'s are independent (or exchangeable) random variables/vectors, linked to the variable of interest $\mathbf{X}$ through a *linear* model (notice that $\mathbf{W}$ may be constant over the population). It is required though that $\mathbf{W}$ is not proportional to $\mathbf{X}$ (in a deterministic sense) to avoid degenerate situations; in such a case, knowing $\mathbf{W}$ on the whole population would mean knowing the empirical process without any error. For the sake of simplicity, the dependence of survey weights in $\mathbf{W}$ will only be emphasized when it is necessary, starting in Section 4.

2.2 Empirical process indexed by classes of functions

In the context of iid realizations $\mathbf{X}_1, \ldots, \mathbf{X}_N$ of a probability measure $\mathbb{P}$, empirical process theory (Ledoux and Talagrand, 1991) consists in the study of the fluctuations of random processes of the type $\{\mathbb{G}_N f,\ f \in \mathcal{F}\}$, where $\mathbb{G}_N := \mathbb{P}_N - \mathbb{P}$. There, class $\mathcal{F}$ designates a certain set of $\mathbb{P}$-integrable real-valued functions,

$$\mathbb{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{X}_i}$$

is the "classical" empirical measure, and for any signed measure $\mathbb{Q}$ on a measurable space $(\mathcal{X}, \mathcal{A})$, $\mathbb{Q}f := \int_{\mathcal{X}} f(\mathbf{x})\, \mathbb{Q}(d\mathbf{x})$ when the integral is well-defined. We assume that class $\mathcal{F}$ admits a square integrable envelope H as defined below.

Assumption 2.2 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 $f \in \mathcal{F}$.

As a consequence, $\mathcal{F}$ is a subset of the space

$$L_2(\mathbb{P}) := \left\{ h : \mathcal{X} \to \mathbb{R},\ h \text{ measurable and } \|h\|_{2,\mathbb{P}}^2 := \mathbb{E}_{\mathbb{P}}\left(h^2(\mathbf{X})\right) < +\infty \right\}.$$

Notice that we may assume without loss of generality that there exists $\eta > 0$ such that $H(\mathbf{x}) > \eta$ for every $\mathbf{x} \in \mathcal{X}$, even if it entails replacing H by $H + \eta$ in the condition above.

2.2.1 Donsker classes

When viewed as a linear operator acting on $\mathcal{F}$, a probability measure $\mathbb{P}$ satisfying Assumption 2.2 may be considered as an element of $\ell^\infty(\mathcal{F})$, *i.e.* 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

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathcal{F}} := d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{h \in \mathcal{F}} \left| \int h \, d\mathbb{P} - \int h \, d\mathbb{Q} \right|,$$

for any couple of probability measures $\mathbb{P}$ and $\mathbb{Q}$. The main purpose of empirical process theory is to find conditions on the class of functions $\mathcal{F}$ guaranteeing that the distribution of $\sqrt{N}\,\mathbb{G}_N$ converges, as $N \to +\infty$, to that of a Gaussian, Banach space valued process in $\ell^\infty(\mathcal{F})$. Such collections of functions are called *Donsker classes* by analogy to the classical results on the empirical distribution function that analyze $\sqrt{N}\,(F_N - F)$, where

$$F_N(\mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{\mathbf{X}_i \in (-\infty\,, x_1] \times \cdots \times (-\infty\,, x_d]\}$$

and

$$F(\mathbf{x}) := \mathbb{P}\left(\mathbf{X} \in (-\infty\,, x_1] \times \cdots \times (-\infty\,, x_d]\right)$$

for $\mathbf{x} := (x_1, \ldots, x_d) \in \mathbb{R}^d$ (see Example 3.1). In particular, the study of the uniform deviations over $\mathcal{F}$

$$\sqrt{N}\,\|\mathbb{P}_N - \mathbb{P}\|_{\mathcal{F}}$$

is of great interest, with a variety of applications in statistics, see Shorack and Wellner (1986). A nearly exhaustive review of asymptotic results ensuring that $\mathcal{F}$ is a Donsker class of functions is available in van der Vaart and Wellner (1996). The purpose of this paper is to extend typical empirical processes results obtained for iid data to the framework of survey sampling.

2.2.2 On measurability issues

Recall that the normed vector space $(\ell^\infty(\mathcal{F}), \|.\|_{\mathcal{F}})$ is (generally) a non-separable Banach space. The major problem one faces when dealing with sums of random variables taking their values in such an infinite-dimensional non-separable

space concerns the measurability of events. For instance, the "classical" empirical process $\sqrt{N}\,(F_N - F)$, which can be viewed as a random sequence in the Skorokhod space $\mathbb{D}([0,1])$ of càd-làg functions endowed with the supremum norm, is not Borel/measurable. In this specific case, the topology induced by the sup-norm on $\mathbb{D}([0,1])$ can be classically replaced by the Skorokhod metric in order to overcome this technical difficulty. Alternative approaches can be found in Pollard (1984). The ideas developed in Hoffmann-Jørgensen (1991) have led to a general solution, based on the concept of *outer probability*, extending the original probability measure $\mathbb{P}$ to non-measurable events by setting $\mathbb{P}^*(A) := \inf\{\mathbb{P}(B) : A \subset B, B \text{ measurable}\}$. Then, the related concept of Hoffman-Jørgensen weak convergence permits somehow to forget the measurability assumptions. Hence, expectations and probabilities must now be understood as outer expectations and probabilities for non-measurable events. For simplicity, the same notations are kept to denote original and outer probabilities (resp. expectations). Here, weak convergence is metrized through the bounded Lipchitz metric on the space $\ell^\infty(\mathcal{F})$: for all random functions $\mathbf{X}$ and $\mathbf{Y}$ valued in $\ell^\infty(\mathcal{F})$,

$$d_{BL}(\mathbf{X},\mathbf{Y}) = \sup_{b \in BL_1(\ell^\infty(\mathcal{F}))} \left|\mathbb{E}\left(b(\mathbf{X})\right) - \mathbb{E}\left(b(\mathbf{Y})\right)\right|,$$

where $BL_1(\ell^\infty(\mathcal{F}))$ is the set of all 1-Lipchitz functions on $\ell^\infty(\mathcal{F})$ bounded by 1. In the following we define the $\rho_{\mathbb{P}}$ semi-metric under $\mathbb{P}$ as

$$\rho_{\mathbb{P}}(f,g) := \mathbb{E}_{\mathbb{P}}\left((f(\mathbf{X}) - g(\mathbf{X}))^2\right) =: \|f-g\|^2_{2,\mathbb{P}}.$$

We refer to van der Vaart and Wellner (1996) for technical details and general results.

2.2.3 Uniform covering numbers

A key concept in the study of empirical process is the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, |.|)$, which corresponds to the minimal number of balls of radius $\varepsilon > 0$ for a given semi metric $|.|$ needed to cover $\mathcal{F}$. Donsker classes of functions are often char-

acterized by some integrability conditions of the form

$$\int_0^1 \sqrt{\mathcal{N}(\varepsilon, \mathcal{F}, |.|)}\, d\varepsilon < \infty,$$

arising from maximal inequalities. Such a condition essentially ensures that the size of class $\mathcal{F}$ is not too big and that one may be able to approximate any of its elements (up to ε) by functions in a set of finite cardinality. In our non-iid setting, we will essentially consider the $L_2(\mathbb{P})$ norm for $|.|$ and use uniform covering numbers

$$\sup_{\mathbb{Q}\in\mathcal{D}} \mathcal{N}\left(\varepsilon \|H\|_{2,\mathbb{Q}}, \mathcal{F}, \|.\|_{2,\mathbb{Q}}\right),$$

where $\mathcal{D}$ is the set of all discrete probability measures $\mathbb{Q}$ such that $\int H^2 d\mathbb{Q}$ is in $(0, +\infty)$. Explicit calculus of (uniform) covering numbers for general classes of functions may be found in several textbooks, see van der Vaart and Wellner (1996) or van de Geer (2000).

3 Empirical process in survey sampling

We now introduce two different empirical processes built from survey data, whose asymptotic behaviors shall be investigated at length in Section 4.

3.1 The Horvitz-Thompson empirical process

In the context of survey data drawn through a general survey plan R_N, the empirical measure $\mathbb{P}_N$ cannot be computed since the whole statistical population is not observable. Hence, a variant based on the observations must be naturally considered. For any measurable set $\mathcal{M} \subset \mathcal{X}$, the Horvitz-Thompson estimator of the empirical probability $\mathbb{P}_N(\mathcal{M}) = N^{-1} \sum_{i=1}^N \delta_{\mathbf{X}_i}(\mathcal{M})$ based on the survey data described above is defined as follows, see Horvitz and Thompson (1951):

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

We highlight the fact that the measure $\mathbb{P}_{R_N}^{\boldsymbol{\pi}(R_N)}$ is an unbiased estimator of $\mathbb{P}$ (resp. $\mathbb{P}_N$, when conditioned upon $(\mathbf{X}_1, \ldots, \mathbf{X}_N)$) although it is not a probabil-

ity measure. For a fixed subset $\mathcal{M}$, the consistency and asymptotic normality of the estimator in Eq. (1) are established in Robinson (1982) 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

$$\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} f := \sqrt{N}\left(\mathbb{P}_{R_N}^{\boldsymbol{\pi}(R_N)} - \mathbb{P}_N\right) f = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\epsilon_i}{\pi_i(R_N)} - 1\right) f(\mathbf{X}_i), \qquad (2)$$

which shall be referred to as the $\mathcal{F}$-indexed *Horvitz-Thomson empirical process* (HT-empirical process, in short). The seemingly redundant notation $\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)}$ is motivated by the fact that extensions involving first order probabilities related to a different sampling scheme T_N will be considered in the sequel. Precisely, $\mathbb{G}_{R_N}^{\boldsymbol{\pi}(T_N)}$ shall denote the process obtained when replacing all $\pi_i(R_N)$ by $\pi_i(T_N)$, $1 \le i \le N$, in Eq. (2).

The main purpose of this chapter is to establish the convergence of the re-weighted empirical process $(\mathbb{G}_{R_N}^{\boldsymbol{\pi}} f)_{f\in\mathcal{F}}$ under adequate hypotheses involving some properties of measure $\mathbb{P}$, certain characteristics of the sequence of sampling plans (R_N), and the "complexity" of class $\mathcal{F}$ (in the classical metric entropy sense) as well. In particular, such a result would permit to describe the asymptotic behavior of the quantity below (assumed to be almost-surely finite, see Assumption 2.2):

$$\left\|\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)}\right\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)} f\right|.$$

By virtue of Cauchy-Schwarz inequality combined with Assumptions 2.1 and 2.2, we almost-surely have, $\forall N \ge 1$,

$$\begin{aligned}\left\|\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)}\right\|_{\mathcal{F}}^2 &\le \frac{1}{N}\left(\sum_{i=1}^{N}\left(\frac{\epsilon_i - \pi_i}{\pi_i}\right)^2\right)\left(\sum_{i=1}^{N} H^2(\mathbf{X}_i)\right)\\ &\le \frac{1}{\pi_\star^2}\sum_{i=1}^{N} H^2(\mathbf{X}_i) < +\infty.\end{aligned}$$

Under Assumptions 2.1 and 2.2, the $\mathcal{F}$-indexed HT-empirical process in Eq. (2) may thus be seen as a sequence of random elements of $\ell^\infty(\mathcal{F})$.

Example 3.1 – *Empirical cumulative distribution function.* In the case where $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$ for instance, a situation of particular interest is that where $\mathcal{F}$ is the class of indicator functions of rectangles of the type

$$\left\{ (-\infty, \mathbf{x}] := \prod_{j=1}^{d} (-\infty, x_j] \ , \ \mathbf{x} = (x_1, \ldots, x_d) \right\}.$$

Then, the empirical process can be identified with the Horvitz-Thompson version of the empirical cumulative distribution function (cdf), namely

$$F_{R_N}^{\boldsymbol{\pi}(R_N)}(\mathbf{x}) := \mathbb{P}_{R_N}^{\boldsymbol{\pi}(R_N)}(-\infty, \mathbf{x}],$$

$\mathbf{x} \in \mathbb{R}^d$, and the goal pursued boils down to investigating conditions under which uniform versions of the Law of Large Numbers (LLN) and of the Central Limit Theorem (CLT) hold for $F_{R_N}^{\boldsymbol{\pi}(R_N)}(\mathbf{x}) - F_N(\mathbf{x})$, where $F_N(\mathbf{x}) := \mathbb{P}_N(-\infty, \mathbf{x}]$. As shall be seen later, the study of the asymptotic behavior of this empirical process lies at the center of the validity of the confidence band construction considered in the Supplementary Materials.

3.2 Alternative estimate in the Poisson sampling case

The Poisson sampling scheme T_N (see Example 2.2) has been the subject of much attention, especially in Hàjek (1964), where asymptotic normality of (pointwise) Horvitz-Thompson estimators have been established in this specific case as well as that of a rejective sampling design with fixed sample size. Indeed, the sample size resulting from a Poisson sampling plan has a large variance equal to $\sum_{i=1}^{N} p_i(1 - p_i) =: d_N$. Conditional sampling plans intend to reduce this volatility. So as to account for the variations of the sample size we consider the following Poisson version of the empirical process rather than the original process :

$$\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\epsilon_i - p_i) \left(\frac{f(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(f) \right), \quad f \in \mathcal{F}, \tag{3}$$

where for all $f \in \mathcal{F}$, $\theta_{N,\mathbf{p}}(f) := d_N^{-1} \sum_{i=1}^N (1-p_i) f(\mathbf{X}_i)$. Notice that $\theta_{N,\mathbf{p}}(f)$ is the coefficient in the regression

$$\frac{1}{N} \sum_{i=1}^N \frac{\epsilon_i}{p_i} f(\mathbf{X}_i) = \theta_{N,\mathbf{p}}(f) \times \frac{1}{N} \sum_{i=1}^N \epsilon_i + \eta_N,$$

which fulfills $\mathrm{cov}_{T_N}\left(\eta_N\,,\,\sum_{i=1}^N \epsilon_i\right) = 0$. Provided that $d_N \to +\infty$ as N tends to $+\infty$, it has been established in Hàjek (1964, Lemma 3.2) that conditioned upon $(\mathbf{X}_1,\ldots,\mathbf{X}_N)$, for fixed $f \in \mathcal{F}$, when N tends to infinity and under a Lindeberg-Feller type condition, the weighted sum of independent random variables in Eq. (3) can be approximated by a centered Gaussian random variable with (conditional) variance

$$V_N^2(f) = \frac{1}{N} \sum_{i=1}^N \left(\frac{f(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(f)\right)^2 p_i(1-p_i).$$

To comprehend why the same result holds true for rejective (or conditional Poisson) sampling with fixed sample size n in the canonical case ($\sum_{i=1}^N p_i = n$, see Hàjek, 1964), simply notice that for a given f, the distribution of the Horvitz-Thompson mean in the rejective design computed with the Poisson inclusion probabilities is given by

$$\begin{aligned}
\mathbb{P}_{R_N}\left(\mathbb{G}_{R_N}^{\mathbf{p}} f \le x\right) &= \mathbb{P}_{R_N}\left(\widetilde{\mathbb{G}}_{R_N}^{\mathbf{p}} f \le x\right) \\
&= \mathbb{P}_{T_N}\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f \le x \;\middle|\; \sum_{i=1}^N \epsilon_i = n\right) \\
&= \frac{\mathbb{P}_{T_N}\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f \le x\,,\,\sum_{i=1}^N \epsilon_i = n\right)}{\mathbb{P}_{T_N}\left(\sum_{i=1}^N \epsilon_i = n\right)}.
\end{aligned}$$

By using the asymptotic normality of the joint distribution of $\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f, \sum_{i=1}^N \epsilon_i\right)$ under the Poisson sampling and the fact that the two components are asymptotically uncorrelated, this immediately ensures that as N tends to infinity, $\mathbb{P}_{R_N}\left(\mathbb{G}_{R_N}^{\mathbf{p}} f \le x\right)$ is equivalent to $\mathbb{P}_{T_N}\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f \le x\right)$. It is thus sufficient to study the behavior of the correctly recentered process $(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}} f)_{f\in\mathcal{F}}$.

As claimed by Theorem 4.2 in the next section, this result can be extended to a functional framework under adequate hypotheses. Combined with an ap-

proximation result, it will serve as the main tool for proving next a similar result in the context of rejective sampling. In the subsequent analysis, we start off by establishing that the process $\widetilde{\mathbb{G}}^{\mathbf{p}}_{T_N}$ can be asymptotically approximated by a Gaussian process.

4 Asymptotic results

The main results of the paper are stated in the present section. As a first go, we establish a FCLT for the empirical process variant of Eq. (3) in the Poisson survey scheme case, before extending it to the rejective design.

4.1 Limit of the empirical process for the Poisson survey scheme

The purpose of this section is to obtain a Gaussian approximation of the empirical process $\widetilde{\mathbb{G}}^{\mathbf{p}}_{T_N}$ related to a Poisson survey plan T_N with first order inclusion probabilities $\mathbf{p} = (p_1, \ldots, p_N)$ depending on some auxiliary variable $\mathbf{W}$ (see Section 2.1.3). The proof relies on Theorem 2.11.1 in van der Vaart and Wellner (1996), applied to the triangular collection of independent variables defined for all $f \in \mathcal{F}$ by

$$Z_{N,i}(f) := Z_{N,i}(f, \boldsymbol{\epsilon}) := \frac{1}{\sqrt{N}}(\epsilon_i - p_i)\left(\frac{f(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(f)\right) \text{ for } i \in \{1, \ldots, N\},$$

conditionally on the full vector $(\mathbf{X}_i, \mathbf{W}_i)_{1 \leq i \leq N}$ and for almost every such sequences. For clarity, the result is recalled below.

Theorem 4.1 – Triangular arrays (van der Vaart and Wellner, 1996). *Let $Z_{N,i}(f)$, $1 \leq i \leq N$ be independent $\mathcal{F}$-indexed stochastic processes defined on the product probability space $\prod_{i=1}^{N}(\{0,1\}, \mathcal{P}(\{0,1\}), \mathcal{B}(\pi_i(R_N))$ where the process $Z_{N,i}(f) := Z_{N,i}(f, \boldsymbol{\epsilon})$ only depends on the ith coordinate of $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_N)$. Assume that the maps*

$$(\epsilon_1, \ldots, \epsilon_N) \mapsto \sup_{\rho_{\mathbb{P}}(f,g)<\delta} \left| \sum_{i=1}^{N} e_i \left(Z_{N,i}(f, \boldsymbol{\epsilon}) - Z_{N,i}(g, \boldsymbol{\epsilon}) \right) \right|$$

and

$$(\epsilon_1, \ldots, \epsilon_N) \mapsto \sup_{\rho_{\mathbb{P}}(f,g)<\delta} \left| \sum_{i=1}^{N} e_i \left(Z_{N,i}(f) - Z_{N,i}(g) \right)^2 \right|$$

are measurable for every $\delta > 0$*, every* $(e_1, \ldots, e_N) \in \{-1, 0, 1\}^N$ *and every* $N \in \mathbb{N}$*. Further define the random semi-metric*

$$d_N^2(f, g) := \sum_{i=1}^{N} \left(Z_{N,i}(f) - Z_{N,i}(g) \right)^2,$$

and suppose that the following conditions are fulfilled, conditionally on the full vector $(\mathbf{X}_i, \mathbf{W}_i)_{1 \le i \le N}$ *and for almost every such sequences.*

i) $\sum_{i=1}^{N} \mathbb{E}\left(\|Z_{N,i}(f)\|_{\mathcal{F}}^2 \cdot \mathbb{I}\{\|Z_{N,i}(f)\|_{\mathcal{F}} > \eta\} \right) \underset{N\to\infty}{\longrightarrow} 0$ *for every* $\eta > 0$.

ii) $\sup_{\rho_{\mathbb{P}}(f,g)<\delta} \sum_{i=1}^{N} \mathbb{E}\left(\left(Z_{N,i}(f) - Z_{N,i}(g) \right)^2 \right) \underset{N\to\infty}{\longrightarrow} 0$ *as* $\delta \to 0$.

iii) $\int_0^{\delta} \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}, d_N)}\, d\varepsilon \underset{N\to\infty}{\longrightarrow} 0$ *as* $\delta \to 0$.

iv) *The sequence of covariance functions* $\mathrm{cov}\left(Z_{N,i}(f), Z_{N,i}(g) \right)$ *converges pointwise on* $\mathcal{F} \times \mathcal{F}$ *as* $N \to \infty$ *to a non degenerate limit* $\Sigma(f, g)$.

Then the sequence $\sum_{i=1}^{N} \left(Z_{N,i}(f) - \mathbb{E}\left(Z_{N,i}(f) \right) \right)$ *is* $\rho_{\mathbb{P}}$*-equicontinuous and converges in* $\ell^\infty(\mathcal{F})$ *to a Gaussian process with covariance function* $\Sigma(f, g)$.

4.1.1 Convergence of the covariance operator

The following intermediary results show that condition iv) in Theorem 4.1 is fulfilled in the particular case of Poisson survey plans. For $(f, g) \in \mathcal{F}^2$, set

$$\mathrm{cov}_{N,\mathbf{p}}(f, g) := \frac{1}{N} \sum_{i=1}^{N} \left(\frac{f(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(f) \right) \left(\frac{g(\mathbf{X}_i)}{p_i} - \theta_{N,\mathbf{p}}(g) \right) p_i(1 - p_i).$$

Due to the independence of the ϵ_i's, it is clear that

$$\begin{aligned} \mathrm{cov}_{T_N}\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}}(f), \widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}}(g) \right) &:= \mathrm{cov}\left(\widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}}(f), \widetilde{\mathbb{G}}_{T_N}^{\mathbf{p}}(g) \;\middle|\; (\mathbf{X}_i, \mathbf{W}_i)_{1 \le i \le N} \right) \\ &= \mathrm{cov}_{N,\mathbf{p}}(f, g). \end{aligned}$$

We thus essentially have to determine conditions ensuring that $\mathsf{cov}_{N,\mathbf{p}}(f,g)$ has a non-degenerate limit. The following assumptions are by no means necessary but provide a useful framework to derive such conditions. Similar types of assumptions may be found in Bonnéry et al. (2011) or Hàjek (1964) for instance.

Recall that inclusion probabilities were defined relative to some auxiliary variable $\mathbf{W}$. An additional assumption on the latter is required in the subsequent result.

Assumption 4.1 The couples of random vectors $(\mathbf{X}_1,\mathbf{W}_1),\ldots,(\mathbf{X}_N,\mathbf{W}_N)$ are iid (exchangeable at least) with distribution $\mathbb{P}_{\mathbf{X},\mathbf{W}}$. Moreover, the conditional inclusion probabilities $\mathbf{p} := (p_1,\ldots,p_N)$ are given for all $i \in \{1,\ldots,N\}$ and $\mathbf{W}_{(N)} \in \mathcal{W}^N$ by

$$p_i := p(\mathbf{W}_i) := \mathbb{E}\left(\epsilon_i \mid \mathbf{W}_{(N)}\right).$$

Remark 4.1 It can happen that p_i not only depends on $\mathbf{W}_i$, but on the entire vector $\mathbf{W}_{(N)}$. It is the case, for instance, when there is a unique auxiliary variable W to which weights are proportional:

$$p_i := n\,\frac{W_i}{\sum_{j=1}^N W_j}.$$

In such situations the $\mathtt{iid}$ property of the vectors $(\mathbf{X}_i,W_i)$, $1 \le i \le N$, can be used to bypass the part involving all $(W_1,\ldots,W_N)$ in the subsequent asymptotic analysis.

Under this supplementary condition, we have the following result, the proof of which can be found in the Supplementary Materials.

Lemma 4.1 – Limit of the covariance operator. *Suppose that Assumptions 2.1, 2.2 and 4.1 are fulfilled. Then we almost-surely have*

$$\frac{1}{N}\,d_N \underset{N\to\infty}{\longrightarrow} D_p := \int_{\mathcal{W}} (1-p(\mathbf{w}))\,p(\mathbf{w})\,\mathbb{P}_{\mathbf{W}}(d\mathbf{w}) > 0$$

and

$$\mathsf{cov}_{N,\mathbf{p}}(f,g) \underset{N\to\infty}{\longrightarrow} \Sigma(f,g),$$

where for all $(f, g) \in \mathcal{F}^2$,

$$\Sigma(f,g) := \int_{\mathcal{X}\times\mathcal{W}} f(\mathbf{x})g(\mathbf{x})\left(\frac{1}{p(\mathbf{w})} - 1\right) \mathbb{P}_{\mathbf{X},\mathbf{W}}(d\mathbf{x}, d\mathbf{w}) - \theta_p(f)\theta_p(g)\, D_p, \quad (4)$$

with

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

4.1.2 Functional Central Limit Theorem

Applying Theorem 4.1 to the empirical process $\widetilde{\mathbb{G}}^{\mathbf{p}}_{T_N} f$ defined in Eq. (3) thus leads to the theorem below, proved in the Supplementary Materials.

Theorem 4.2 – FCLT in the Poisson survey case. *Suppose that Assumptions 2.1, 2.2 and 4.1 hold, as well as the following conditions.*

i) *Lindeberg-Feller type condition:* $\forall \eta > 0$,

$$\mathbb{E}\left((\mathcal{Z}_{N,i})^2\, \mathbb{I}\left\{\mathcal{Z}_{N,i} > \eta\sqrt{N}\right\}\right) \underset{N\to\infty}{\longrightarrow} 0,$$

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 in Section 2.2.3, and assume*

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

Then there exists a $\rho_{\mathbb{P}}$*-equicontinuous Gaussian process* $\mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *with covariance operator* Σ *given by Eq. (4) such that*

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

Remark 4.2 – *On the Lindeberg-Feller condition.* Observe that, as can be proved using Hölder's inequality, condition i) in Theorem 4.2 can be replaced by the simpler condition: $\exists\, \delta > 0$ such that

$$i^*) \quad \mathbb{E}_{\mathbb{P}_{\mathbf{X},\mathbf{W}}}\left(\left|\frac{H(\mathbf{X}_i)}{p(\mathbf{W}_i)}\right|^{2+\delta} \mathbb{E}_{T_N}\left((\epsilon_i - p(\mathbf{W}_i))^{2+\delta} \mid (\mathbf{X}_i, \mathbf{W}_i)\right)\right) < +\infty.$$

4.2 The case of rejective sampling and its variants

As shall be shown herein-after, the result obtained above in the case of a Poisson sampling scheme may carry over to more general survey plans, as originally proposed in the seminal contribution of Hàjek (1964).

4.2.1 Empirical process for the rejective sampling

The Central Limit Theorem for rejective sampling and some variants of this survey scheme has been studied at length in Hàjek (1964). Consider the rejective sampling scheme defined in Example 2.4 from a given vector $\boldsymbol{\pi}^R$ corresponding to the vector $\mathbf{p} := (p_1, \ldots, p_N) = (p(\mathbf{W}_1), \ldots, p(\mathbf{W}_N)) =: p(\mathbf{W})$. Assume in addition that the representation is *canonical*, *i.e.* is such that $\sum_{i=1}^N p(\mathbf{W}_i) = n$. The key argument in Hàjek (1964) for proving a CLT in the rejective sampling case consists in exhibiting a certain coupling $((\epsilon_1, \ldots, \epsilon_N), (\epsilon_1^*, \ldots, \epsilon_N^*))$ of the Poisson sampling scheme with inclusion probabilities $p(\mathbf{W}_1), \ldots, p(\mathbf{W}_N)$ and the rejective sampling scheme with corresponding inclusion probabilities $\boldsymbol{\pi}^R$, see Hàjek (1964, p. 1503-1504) for further details. We point out that, under the rejective sampling scheme, the survey size is fixed, so that

$$\sum_{i=1}^N (\epsilon_i - p(\mathbf{W}_i)) = n - n = 0.$$

Thus, we have:

$$\begin{aligned}
\widetilde{\mathbb{G}}_{R_N}^{\mathbf{p}} f &:= \frac{1}{\sqrt{N}} \sum_{i=1}^N (\epsilon_i - p(\mathbf{W}_i)) \left(\frac{f(\mathbf{X}_i)}{p(\mathbf{W}_i)} - \theta_{N,\mathbf{p}}(f) \right) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\epsilon_i}{p(\mathbf{W}_i)} - 1 \right) f(\mathbf{X}_i) \\
&=: \mathbb{G}_{R_N}^{p(\mathbf{W})} f.
\end{aligned}$$

Hence, the Poisson-like empirical process coincides, in that case, with the original HT-empirical process where the weights $p(\mathbf{W})$ are involved instead of the true inclusion probabilities $\boldsymbol{\pi}^R$, the latter being however asymptotically equivalent to the former, see Hàjek (1964).

The next theorem is obtained by noticing that rejective sampling is a Poisson sampling conditioned on a fixed sample size.

Theorem 4.3 – FCLT in the rejective survey with Poisson weights case. *Suppose that Assumptions 2.1, 2.2, 4.1 and conditions* i) *and* ii) *of Theorem 4.2 are satisfied. Then, there exists a* $\rho_{\mathbb{P}}$*-equicontinuous Gaussian process* $\mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *with covariance operator* Σ *given by Eq. (4) such that*

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

Going back to the original HT-empirical process in Eq. (2) related to the plan R_N, the corollary below reveals that the asymptotic result still holds true for the latter (see the proof in the Supplementary Materials). This essentially follows from the fact that the weights $\mathsf{p}(\mathbf{W})$ and the inclusion probabilities corresponding to the rejective sampling are asymptotically equivalent.

Corollary 4.1 – FCLT in the rejective survey case. *Suppose that Assumptions 2.1, 2.2, 4.1 and conditions* i) *and* ii) *of Theorem 4.2 are satisfied. Then, there exists a* $\rho_{\mathbb{P}}$*-equicontinuous Gaussian process* $\mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *with covariance operator* Σ *given by Eq. (4) such that*

$$\mathbb{G}_{\mathsf{R}_\mathsf{N}}^{\boldsymbol{\pi}(\mathsf{R}_\mathsf{N})} \Rightarrow \mathbb{G} \text{ weakly in } \ell^\infty(\mathcal{F}), \text{ as } \mathsf{N} \to \infty.$$

This result generates many applications such as those introduced in the Supplementary Materials. In particular, one may deduce from Corollary 4.1 the asymptotic Normality of Hadamard- or Fréchet-differentiable functionals. In order to illustrate the practical assets of our theoretical results, numerical experiments were performed, the outcomes of which are also displayed in the Supplementary Materials. They show how confidence bands of the cumulative distribution function may be constructed in the context of a Rejective sampling scheme.

4.2.2 Extension to other sampling designs

The lemma stated below, following in the footsteps of Berger (1998), shows that the study of the empirical process related to a general sampling design $\widetilde{R}_N$ may be reduced to that related to a simpler sampling design, R_N say, which is close to $\widetilde{R}_N$ with respect to some metric and entirely characterized by its first order inclusion probabilities. The only "drawback" is that the estimator involved in this approximation result is not the Horvitz-Thompson estimator, since it does not involves the inclusion probabilities of the sampling plan of interest but those related to $\widetilde{R}_N$ (Hàjek, 1964). However, as will be shown next, the two estimators may asymptotically coincide, as N tends to $+\infty$.

In order to formulate the approximation result needed in the sequel, we introduce, for two sampling designs $\widetilde{R}_N$ and R_N, the *total variation metric*

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

as well as the *entropy*

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

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

Lemma 4.2 – Approximation result. *Let $\widetilde{R}_N$ and R_N be two sampling designs, then the empirical processes $\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)}$ and $\mathbb{G}_{\widetilde{R}_N}^{\boldsymbol{\pi}(R_N)}$ valued in $\ell^\infty(\mathcal{F})$ satisfy the relationships:*

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

Consequently, if the sequences $(\widetilde{R}_N)_{N \geq 1}$ and $(R_N)_{N \geq 1}$ are such that $\|\widetilde{R}_N - R_N\|_1$ tends to 0 or $D(R_N, \widetilde{R}_N) \to 0$ as $N \to \infty$ and if there exists a Gaussian process $\mathbb{G}$ such that

$$d_{BL}\left(\mathbb{G}_{R_N}^{\boldsymbol{\pi}(R_N)}, \mathbb{G} \right) \underset{N \to \infty}{\longrightarrow} 0,$$

then we also have

$$d_{BL}\left(\mathbb{G}_{\widetilde{R}_N}^{\boldsymbol{\pi}(R_N)}, \mathbb{G}\right) \underset{N\to\infty}{\longrightarrow} 0.$$

This result, proved in the Supplementary Materials, reveals that as soon as a possibly complicated survey design $\widetilde{R}_N$ can be approximated by a simpler one R_N through some coupling argument ensuring that the $\|.\|_1$ distance between them decays to zero (as in Berger, 1998), then an asymptotic approximation result possibly holding true for the empirical process related to R_N immediately extends to that related to $\widetilde{R}_N$, when built with the inclusion probabilities $\boldsymbol{\pi}(R_N)$. If in addition $\boldsymbol{\pi}(\widetilde{R}_N)$ and $\boldsymbol{\pi}(R_N)$ are asymptotically uniformly close (as is the case for the inclusion probabilities of the rejective and the Poisson survey schemes, see Hàjek, 1964), then the result also extends to the empirical process related to $\widetilde{R}_N$ involving the inclusion probabilities $\boldsymbol{\pi}(\widetilde{R}_N)$. A typical situation where this result applies corresponds to the case where $\widetilde{R}_N$ is a Rao-Sampford or successive sampling design, while R_N is a rejective sampling design, as in Berger (1998, 2011).

5 Discussion

Generalizing the seminal work of Breslow and Wellner (2007) and Saegusa and Wellner (2011) to the case of Poisson-like survey schemes with unequal first order inclusion probabilities depending on some appropriate auxiliary variable, we introduced in Section 3 a Horvitz-Thompson version of the empirical process the asymptotic properties of which were analyzed at length. The exhibited rate of convergence appeared to be the same as that of the standard empirical process, namely $\sqrt{N}$. Natural applications of these results to Hadamard- and Fréchet-differentiable functionals were also considered in the Supplementary Materials, in which simulations were performed to illustrate their utility in the construction of uniform confidence band intervals around the empirical cumulative distribution function in the entire population. Many improvements may be brought to these first results, for instance situations where the true inclusion

probabilities are not available and replaced by an estimated version issued from post-calibration methods could be inspected. The assumptions made on the inclusion probabilities are also quite restrictive. Following in the lines of Boistard et al. (2012), higher order conditions could permit to get rid of Assumption 2.1. Other sampling designs that can be written as a conditional Poisson scheme may be considered as well, like stratified sampling introduced in Example 2.3.

In view of the empirical results presented in the Supplementary Materials, it is clear that the definition of the inclusion probabilities could be handled so as to minimize the variance of the estimators of interest. So proceeding would be of great interest in the context of big data management. Indeed, it is more and more frequent to meet databases that increase regularly (in finance, information about the markets is stocked every hour at least) and cannot be saved, thus analyzed, on a single computer. When accessing such huge files becomes a challenge, sampling is a natural solution, as was already underlined by Cardot et al. (2013). In this context, the superpopulation model and the asymptotic nature of our results are perfectly relevant. Moreover, the analyst has then complete control over the survey scheme they desire to adopt, which is typically rarely the case with institutional data. Hence, the Poisson and rejective schemes, which are not of frequent use in practice, are revealed as especially convenient for such types of analyses.

Supporting Information

Additional information for this article is available online. It contains Appendix S1 - *Some applications to nonparametric statistics*, including Tables S1, S2, S3 and Figure S1, and Appendix S2 - *Technical details*.

Acknowledgments

This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

References

M.A. Arcones and E. Giné. On the bootstrap of M-estimators and other statistical functionals. *Exploring the Limits of Bootstrap, ed. by R. LePage and L. Billard, Wiley*, pages 13–47, 1992.

P. Barbe and P. Bertail. *The weighted bootstrap*, volume 98. Springer Verlag, 1995.

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

Y.G. Berger. Pak. J. Statist. 2011 Vol. 27 (4), 407-426 Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pak. J. Statist*, 27(4):407–426, 2011.

H. Boistard, H.P. Lopuhaä, and A. Ruiz-Gazen. Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics*, 6:1967–1983, 2012.

D. Bonnéry, J. Breidt, and F. Coquet. Propriétés asymptotiques de l'échantillon dans le cas d'un plan de sondage informatif. *Submitted for publication*, 2011.

N.E. Breslow and J.A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 35:186–192, 2007.

H. Cardot, C. Goga, and P. Lardin. Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *To appear in the Scandinavian J. of Statistics*, 2013.

J.C. Deville. *Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans les sondages*. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

J.J. Droesbeke, B. Fichet, and P. Tassi. *Les sondages*. Economica, 1987.

R.M. Dudley. Nonlinear functionals of empirical measures and the bootstrap. In *Probability in Banach Spaces 7*, pages 63–82. Springer, 1990.

R.D. Gill. Non- and semiparametric maximum likelihood estimators and the von Mises method. *Scand. J. Statistics*, 22:205–214, 1989.

R.D. Gill, Y. Vardi, and J.A. Wellner. Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, 16(3):1069–1112, 1988.

C. Gourieroux. *Théorie des sondages*. Economica, 1981.

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

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*. 1986.

J. Hoffmann-Jørgensen. *Stochastic processes on Polish spaces*. Aarhus Universitet, Denmark, 1991.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *JASA*, 47:663–685, 1951.

D.P. Kroese, T. Taimre, and Z.I. Botev. *Handbook of Monte Carlo methods*. Wiley, 2011.

M Ledoux and M. Talagrand. *Probability in Banach spaces: isoperimetry and processes* . Springer-Verlag, 1991.

D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, New-York, 1984.

O. Pons and E. de Turkheim. Von mises method, bootstrap and Hadamard differentiability for nonparametric general models. *Statistics= A Journal of Theoretical and Applied Statistics*, 22(2):205–214, 1991.

S.I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Verlag, 2007. ISBN 0387242724.

P.M. Robinson. On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2):234–238, 1982.

M. Rueda, S. Martínez, H. Martínez, and A. Arcos. Estimation of the distribution function with calibration methods. *Journal of statistical planning and inference*, 137(2):435–448, 2007.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. *Preprint available at http://arxiv.org/abs/1112.4951v1*, 2011.

G. Shorack and J.A. Wellner. *Empirical processes with applications to statistics*. Wiley, 1986.

Y. Tillé. Utilisation d'informations auxiliaires dans les enquêtes par sondage. *Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa*, 23(3):491–505, 1999.

Y. Tillé. *Sampling algorithms*. Springer Series in Statistics, 2006.

S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

A.W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

Corresponding author

Emilie Chautru,
Université de Cergy-Pontoise
Département de Mathématiques - AGM
2, avenue Adolphe Chauvin
95302 Cergy-Pontoise Cedex, France
Email : emilie.chautru@gmail.com