Functional Central Limit Theorems for Stratified Single-Stage Sampling Designs

Anne Ruiz-Gazen

TSE, Université Toulouse 1 Capitole,
anne.ruiz-gazen@tse-fr.eu

(collaboration with Rik Lopuhaä)

ITACOSM 2019
Firenze, Italy

Motivation

In survey sampling, when one wants to derive confidence intervals for estimators of complex parameters that are functions of the distribution function (e.g. the poverty rate), need to prove the asymptotic normality of the estimator $\Longrightarrow$ functional central limit theorem.

$$\left[\hat{\theta} - 1.96\sqrt{\widehat{Var}(\hat{\theta})}\,;\, \hat{\theta} + 1.96\sqrt{\widehat{Var}(\hat{\theta})}\right].$$

For a distribution function F, the poverty rate is:

$$\phi(F) = F\left(\beta F^{-1}(\alpha)\right)$$

for fixed $0 < \alpha, \beta < 1$, where $F^{-1}(\alpha) = \inf\{t : F(t) \geq \alpha\}$.

Typical choices are $\alpha = 0.5$ and $\beta = 0.5$ (INSEE) or $\beta = 0.6$ (EUROSTAT).

Motivation

The empirical process theory and the functional Delta method can be used to derive the asymptotic normality of regular functionals (in the sense of Hadamard differentiable).

Our aims:

- use the empirical process theory to prove functional limit theorems for relevant empirical processes in survey sampling for stratified designs.
- derive asymptotic properties of estimators in the survey sampling framework using the functional delta method for Hadamard differentiable functionals.

(van der Vaart, 1998, van der Vaart and Wellner, 1996)

In short

In Boistard, Lopuhaä and Ruiz-Gazen (BLRG 2017), we derive functional limit theorems but we do not consider the case of stratified sampling designs.

This presentation gives:

- a reminder of the results by BLRG (2017),
- some extensions in the case of a stratified sampling design with a fixed number of strata,
- some extensions when the number of strata depends on N and may grow to infinity (Krewski and Rao, 1981).

1. Introduction
2. BLRG (2017) reminder
   - Context
   - Empirical process under study
   - Assumptions and theorem
3. Stratified sampling designs
   - Fixed number of strata
   - Allowing the number of strata change with N

1. Introduction

2. BLRG (2017) reminder
   - Context
   - Empirical process under study
   - Assumptions and theorem

3. Stratified sampling designs
   - Fixed number of strata
   - Allowing the number of strata change with N

Context

We follow Rubin-Bleuer and Schiopu Kratina (2005) and consider a product probability space that includes the super-population and the design space, assuming that sample selection and model characteristic are independent given the design variables (non-informative sampling).

Consider a sequence of nested finite populations associated to a set of indices $U_N = \{1, 2, \ldots, N\}$ of sizes $N = 1, 2, \ldots$.

For each index $i \in U_N$, we have the variable of interest $y_i \in \mathbb{R}$.

Context

Super-population: we assume that the values y_i in each finite population are realizations of i.i.d. random variables $Y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, N$, on a common probability space $(\Omega, \mathfrak{F}, \mathbb{P}_m)$.

Design (without replacement):
for all $N = 1, 2, \ldots$, $\mathcal{S}_N = \{s : s \subset U_N\}$: collection of subsets of U_N and $\mathcal{A}_\mathcal{N} = \sigma(\mathcal{S}_\mathcal{N})$: σ-algebra generated by $\mathcal{S}_N$.
We define a probability measure $\mathbb{P}_d$ on the design space $(\mathcal{S}_N, \mathcal{A}_\mathcal{N})$.

Product:
let $(\mathcal{S}_N \times \Omega, \mathcal{A}_\mathcal{N} \times \mathfrak{F})$ be the product space with probability measure:

$$\mathbb{P}_{d,m}(\{s\} \times E) = \mathbb{P}_d(\{s\})\, \mathbb{P}_m(E).$$

Context

Sample s where n denotes the expectation of the size of the sample under the design.

$$\xi_i = \mathbb{1}_{\{i \in s\}},$$

$$\pi_i = \mathbb{P}_d(\xi_i = 1) > 0$$

$$\pi_{i_1 i_2 \ldots i_k} = \mathbb{P}_d(\xi_{i_1} = 1, \xi_{i_2} = 1, \ldots, \xi_{i_k} = 1).$$

Note that n and the inclusion probabilities are considered as fixed in the present talk but could be random (dependent on the design variables).

Horvitz-Thompson process

Let F be the c.d.f. of Y_i and $\mathbb{F}_N(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Y_i \le t\}}, t \in \mathbb{R}$

The Horvitz-Thompson (HT) empirical process centered with $\mathbb{F}_N$:

$$\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$$

with

$$\mathbb{F}_N^{\mathrm{HT}}(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i \mathbb{1}_{\{Y_i \le t\}}}{\pi_i}, \quad t \in \mathbb{R}.$$

In the paper, we also consider the HT process centered around F and the Hájek processes but in this presentation, we only focus on the Horvitz-Thompson process centered with $\mathbb{F}_N$.

We derive the limiting distribution of the empirical process

$$\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$$

using Theorem 13.5 in Billingsley, 1999.

This requires

- weak convergence of all finite dimensional distributions
- and a tightness condition (see (13.14) in Billingsley, 1999).

1. Introduction
2. BLRG (2017) reminder
   - Context
   - Empirical process under study
   - Assumptions and theorem
3. Stratified sampling designs
   - Fixed number of strata
   - Allowing the number of strata change with N

Assumptions on the design

In order to prove the tightness condition, we make assumptions on the design.
Let

$$D_{\nu,N} = \left\{ (i_1, i_2, \ldots, i_\nu) \in \{1, 2, \ldots, N\}^\nu : i_1, i_2, \ldots, i_\nu \text{ all different} \right\}, \quad (1)$$

for the integers $1 \leq \nu \leq 4$.

Assumptions on the design

(C1) there exist constants K_1, K_2, such that for all $i = 1, 2, \ldots, N$,

$$0 < K_1 \leq \frac{N\pi_i}{n} \leq K_2 < \infty, \quad \omega - \text{a.s.}$$

There exists a constant $K_3 > 0$, such that for all $N = 1, 2, \ldots$:

(C2) $\max_{(i,j)\in D_{2,N}} \left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)\right| < K_3 n/N^2$,

(C3) $\max_{(i,j,k)\in D_{3,N}} \left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)\right| < K_3 n^2/N^3$,

(C4) $\max_{(i,j,k,l)\in D_{4,N}} \left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)(\xi_l - \pi_l)\right| < K_3 n^2/N^4$,

ω-almost surely.

Assumptions on the HT estimator

To establish the convergence of finite dimensional distributions, for sequences of bounded i.i.d. random variables $V_1, V_2, \ldots$ on $(\Omega, \mathfrak{F}, \mathbb{P}_m)$, we need a Central Limit Theorem for the HT estimator in the design space, conditionally on the V_i's.

Let S_N^2 be the (design-based) variance of the HT estimator of the population mean, i.e.,

$$S_N^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} V_i V_j. \quad (2)$$

Assumptions on the HT estimator

(HT1) Let $V_1, V_2, \ldots$ be a sequence of bounded i.i.d. random variables, not identical to zero, and such there exists an $M > 0$, such that $|V_i| \leq M$ ω-almost surely, for all $i = 1, 2, \ldots$.
Suppose that for N sufficiently large, $S_N > 0$ and

$$\frac{1}{S_N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i V_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N} V_i\right) \to N(0,1), \qquad \omega - \text{a.s.},$$

in distribution under $\mathbb{P}_d$.

Assumptions on the HT estimator

We also need that nS_N^2 converges in probability under $\mathbb{P}_m$ to a constant:

(HT2) there exist constants $\mu_{\pi 1}$, $\mu_{\pi 2} \in \mathbb{R}$ such that

(i) $$\lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1 \right) = \mu_{\pi 1},$$

(ii) $$\lim_{N\to\infty} \frac{n}{N^2} \sum\sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = \mu_{\pi 2}.$$

Result for the HT process centered by F_N

Let $D(\mathbb{R})$ be the space of càdlàg functions on $\mathbb{R}$ equipped with the Skorohod topology.

Theorem 1

Suppose that conditions (C1)-(C4) and (HT1)-(HT2) hold.

Then $\sqrt{n}(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$ converges weakly in $D(\mathbb{R})$

to a mean zero Gaussian process $\mathbb{G}^{\mathrm{HT}}$ with covariance kernel

$\mathbb{E}_m \mathbb{G}^{\mathrm{HT}}(s)\mathbb{G}^{\mathrm{HT}}(t) = \mu_{\pi_1} F(s \wedge t) + \mu_{\pi 2} F(s)F(t)$, for $s, t \in \mathbb{R}$

Context

The population U_N is divided into H strata, $U_{N_1}, \ldots, U_{N_H}$ of sizes $N_1, N_2, \ldots, N_H$.

A sample s_h of size n_h is selected according to a design with inclusion probabilities π_{hi}, π_{hij}, and inclusion indicators ξ_{hi}. The H samples are selected independently on each stratum.

The overall sample $s = \cup_{h=1}^{H} s_h$ and has size $n_s = \sum_{h=1}^{H} n_{sh}$.

We have $Y_1, Y_2, \ldots, Y_N$ independent identically distributed with distribution function F in the super population setup. They can be divided into $Y_{h1}, \ldots, Y_{hN_h}$ per stratum U_{Nh}, for $h = 1, 2, \ldots, H$.

Process definition

The Horvitz-Thompson empirical process centered with $\mathbb{F}_N$:

$$\sqrt{n}\left(\mathbb{F}_N^{\mathrm{HT}}(t) - \mathbb{F}_N(t)\right) = \sum_{h=1}^{H} \frac{N_h}{N} \frac{\sqrt{n}}{\sqrt{n_h}} \sqrt{n_h}\left(\mathbb{F}_{N_h}^{\mathrm{HT}}(t) - \mathbb{F}_{N_h}(t)\right).$$

where

$$\mathbb{F}_{N_h}^{\mathrm{HT}}(t) = \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{\xi_{hi} \mathbb{1}_{\{Y_{hi} \le t\}}}{\pi_{hi}}, \quad \mathbb{F}_N^{\mathrm{HT}}(t) = \sum_{h=1}^{H} \frac{N_h}{N} \mathbb{F}_{N_h}^{\mathrm{HT}}(t),$$

$$\mathbb{F}_{N_h}(t) = \frac{1}{N_h} \sum_{i=1}^{N_h} \mathbb{1}_{\{Y_{hi} \le t\}}, \quad \mathbb{F}_N(t) = \sum_{h=1}^{H} \frac{N_h}{N} \mathbb{F}_{N_h}(t), \quad t \in \mathbb{R};$$

Proposed approach

Let us assume that $N_h \to \infty$ as $N \to \infty$.

We prove the weak convergence of the process $\sqrt{n}\left(\mathbb{F}_N^{\mathrm{HT}}(t) - \mathbb{F}_N(t)\right)$ in $D(\mathbb{R})$ by assuming:

- (Ch1)-(Ch4) and (HTh1) and (HTh2) for all $h = 1, \ldots, H$.
- and one extra condition:
  there exist constants α_h for $h = 1, \ldots, H$ such

$$\frac{N_h}{N}\frac{\sqrt{n}}{\sqrt{n_h}} \to \alpha_h \quad \mathbb{P}_{d,m}\text{-probability.} \tag{3}$$

Proposed approach

$$\sqrt{n}\left(\mathbb{F}_N^{\mathrm{HT}}(t) - \mathbb{F}_N(t)\right) = \sum_{h=1}^{H_N} \frac{N_h}{N} \frac{\sqrt{n}}{\sqrt{n_h}} \sqrt{n_h}\left(\mathbb{F}_{N_h}^{\mathrm{HT}}(t) - \mathbb{F}_{N_h}(t)\right).$$

When H_N changes with N, we propose to:

- assume (Ch1)-(Ch4) and (HTh1) and (HTh2) for all $h = 1, \ldots, H$,
- find conditions such that (C1)-(C4) and (HT1) and (HT2) are verified.

Example of the (C1) condition

(C1)

There exist constants K_1, K_2, such that for all $h = 1, 2, \ldots, H_N$ and $i = 1, 2, \ldots, N_h$,

$$0 < K_1 \leq \frac{N\pi_{hi}}{n} \leq K_2 < \infty, \quad \omega - \text{a.s.}$$

If we assume

$$0 < \liminf_{N\to\infty} \min_{1\leq h\leq H_N} \frac{N}{N_h}\frac{n_h}{n} \leq \limsup_{N\to\infty} \max_{1\leq h\leq H_N} \frac{N}{N_h}\frac{n_h}{n} < \infty,$$

and that (Ch1) holds for $h = 1, 2, \ldots, H_N$, with constants K_{h1} and K_{h2}.

then (C1) holds with:

$$K_1 = \liminf_{N\to\infty} \min_{1\leq h\leq H_N} K_{h1} > 0$$

$$K_2 = \limsup_{N\to\infty} \max_{1\leq h\leq H_N} K_{h2} < \infty,$$

Thank you for your attention !