Processus empiriques en sondages. Application à l'estimation du taux de pauvreté.

Hélène Boistard[(1)], Rik Lopuhaä[(2)] and Anne Ruiz-Gazen[(1)]

(1) TSE, Université Toulouse 1 Capitole,
contact: anne.ruiz-gazen@tse-fr.eu
(2) Delft University of Technology, The Netherlands.

9ème colloque francophone sur les sondages
Gatineau, Canada

Motivation

For a parameter θ and an estimator $\hat{\theta}$,

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

In survey sampling, confidence intervals for finite population or infinite population parameters are making use of the asymptotic normality of the point estimators.

Even when estimating a total with the Horvitz-Thompson estimator, proving the asymptotic normality is not an easy task.

Even more difficult for complex parameters :

- For some regular functions g, possible to use the Delta method to derive the asymptotic normality of $g(\hat{\theta})$ given that $\hat{\theta}$ is asymptotically normal.
- For parameters like the poverty rate, more difficult.

Poverty rate

For a distribution function F,

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

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).

Functional definition.

In classical statistics, use empirical process theory to derive the asymptotic normality of functionals that are regular in the sense of Hadamard differentiable using the functional Delta method (van der Vaart, 1998).

Motivation

Aims :

- Use the empirical processes theory in order to prove functional limit theorems for relevant empirical processes in survey sampling.
- Derive some 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)

Example : derive asymptotic normality of estimators of complex parameters such as the poverty rate.

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 $y_i \in \mathbb{R}$.

Simplified context (no design variable)

Super-population : we assume that the values in each finite population are realizations of the independent 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.

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

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

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

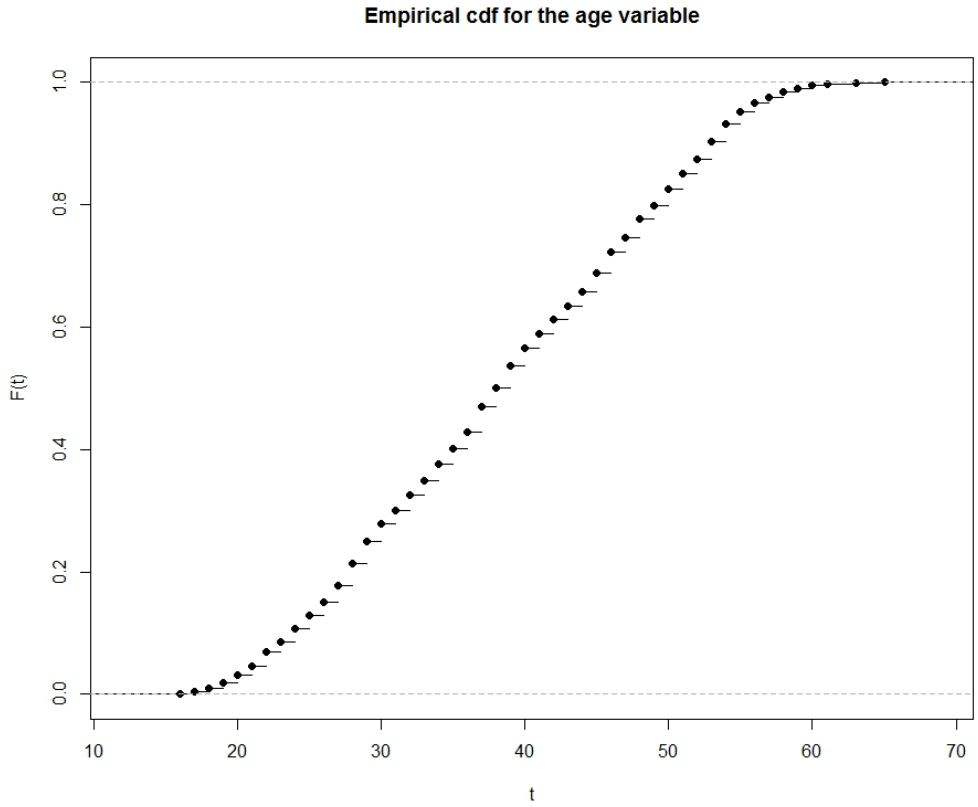
Note that n and the inclusion probabilities are considered as fixed in the present talk but could be random (dependent on the design variables).

Horvitz-Thompson processes

Let F be the c.d.f. of Y_i and

$$\mathbb{F}_N(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{Y_i \leq t\}}, t \in \mathbb{R}$$

the empirical (finite population) c.d.f. This function is cadlag.



Horvitz-Thompson processes

The Horvitz-Thompson (HT) empirical processes

obtained from the HT empirical c.d.f. :

$$\mathbb{F}_N^{\mathrm{HT}}(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{\xi_i \mathbb{1}_{\{Y_i \leq t\}}}{\pi_i}, \quad t \in \mathbb{R}. \tag{2}$$

where ξ_i is the indicator $\mathbb{1}_{\{s \ni i\}}$.

- $\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$, centered around the empirical c.d.f. $\mathbb{F}_N$ of the Y_i's,
- $\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - F)$, centered around c.d.f. F of the Y_i's.

Hájek processes

The Hájek empirical processes

obtained from the Hájek empirical c.d.f. :

$$\mathbb{F}_N^{\mathrm{HJ}}(t) = \frac{1}{\widehat{N}} \sum_{i=1}^{N} \frac{\xi_i \mathbb{1}_{\{Y_i \leq t\}}}{\pi_i}, \quad t \in \mathbb{R}, \tag{3}$$

where $\widehat{N} = \sum_{i=1}^{N} \xi_i / \pi_i$ is the HT estimator for the population total N.

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

In this presentation, we will only consider the Horvitz-Thompson processes.

Poverty rate

For a distribution function F,

$$\phi(F) = F\left(\beta F^{-1}(\alpha)\right) \tag{4}$$

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

The finite population parameter is $\phi(\mathbb{F}_N)$.

- $\phi(\mathbb{F}_N^{\mathrm{HT}})$ is the HT estimator,
- $\phi(\mathbb{F}_N^{\mathrm{HJ}})$ is the Hájek estimator.

Roughly speaking, if we are able to derive the limiting distribution of the empirical process

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

then for any functional ϕ Hadamard differentiable,
we can get the normality of

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

And the same for

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

A functional central limit theorem for $\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$ and $\sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - F)$ is obtained 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).

Assumptions on the design

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

$$D_{\nu,N} = \Big\{(i_1, i_2, \ldots, i_\nu) \in \{1, 2, \ldots, N\}^\nu : i_1, i_2, \ldots, i_\nu \text{ all different}\Big\}, \quad (5)$$

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 design

These conditions on higher order correlations are commonly used in the literature on survey sampling in order to derive asymptotic properties of estimators, e.g., Breidt and Opsomer (2000), and Cardot *et al.* (2010).

Breidt and Opsomer (2000) proved that they hold for simple random sampling without replacement and stratified simple random sampling without replacement, whereas Boistard *et al.* (2012) proved that they hold also for rejective sampling.

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.,

$$S_N^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} V_i V_j. \tag{6}$$

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

(HT3) $n/N \to \lambda$, where $\lambda \in [0, 1]$ ω-a.s.

(HT4) $\mu_{\pi 1} > 0$.

HT3 and HT4 only needed for the process centered by the super-population c.d.f.

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.

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

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

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

HT process centered by F

Theorem 2

Suppose that conditions (C1)-(C4) and (HT1)-(HT4) hold.

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

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

$\mathbb{E}_{d,m}\mathbb{G}_F^{\mathrm{HT}}(s)\mathbb{G}_F^{\mathrm{HT}}(t) = (\mu_{\pi_1} + \lambda)F(s \wedge t) + (\mu_{\pi 2} - \lambda)F(s)F(t)$, for $s, t \in \mathbb{R}$.

Process centered by $\mathbb{F}_N$

Tightness condition

Lemma 1

Let $\mathbb{X}_N = \sqrt{n}\,(\mathbb{F}_N^{\mathrm{HT}} - \mathbb{F}_N)$ and suppose that (C1)-(C4) hold.
Then there exists a constant $K > 0$ independent of N, such that for any t_1, t_2 and $-\infty < t_1 \leq t \leq t_2 < \infty$,

$$\mathbb{E}_{d,m}\left[(\mathbb{X}_N(t) - \mathbb{X}_N(t_1))^2\,(\mathbb{X}_N(t_2) - \mathbb{X}_N(t))^2\right] \leq K\Big(F(t_2) - F(t_1)\Big)^2.$$

Use of Billingsley (1999) theorem 13.5

- First for the Y_is' uniformly distributed on $[0; 1]$,
- Then using the proof of Theorem 14.3 (Billingsley, 1999) to generalize to any distribution.

Process centered by F

Remarks

- More terms in the tightness condition
- Finite dimensional convergence

$$\sqrt{n}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i V_i}{\pi_i} - \mu_V\right)$$
$$= \sqrt{n}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i V_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N} V_i\right) + \frac{\sqrt{n}}{\sqrt{N}} \times \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} V_i - \mu_V\right).$$

+ Theorem 5.1(iii) from Rubin-Bleuer and Schiopu Kratina (2005) with (HT4) in order to have the design-based variance converging to a strictly positive constant.

Definition

Let $D(\mathbb{R})$ be the space of càdlàg functions on $\mathbb{R}$ equipped with the Skorohod topology. Let $\mathbb{D}_\phi \subset D(\mathbb{R})$ consist of $F \in D(\mathbb{R})$ that are non-decreasing. Then for $F \in \mathbb{D}_\phi$, the poverty rate is defined as

$$\phi(F) = F\left(\beta F^{-1}(\alpha)\right) \tag{7}$$

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). Its Hadamard derivative is given by

$$\phi'_F(h) = -\beta \frac{f(\beta F^{-1}(\alpha))}{f(F^{-1}(\alpha))} h(F^{-1}(\alpha)) + h(\beta F^{-1}(\alpha)). \tag{8}$$

Result

Corollary

Under conditions defined previously, if F is differentiable at $F^{-1}(\alpha)$ with positive derivative $f(F^{-1}(\alpha))$, the random variables $\sqrt{n}(\phi(\mathbb{F}_N^{\mathrm{HT}}) - \phi(\mathbb{F}_N))$ and $\sqrt{n}(\phi(\mathbb{F}_N^{\mathrm{HT}}) - \phi(F))$ converge in distribution to a mean zero normal random variable with variance

$$\sigma^2_{\mathrm{HT},\alpha,\beta} = \beta^2 \frac{f(\beta F^{-1}(\alpha))^2}{f(F^{-1}(\alpha))^2} \left(\gamma_{\pi 1}\alpha + \gamma_{\pi 2}\alpha^2\right) + \gamma_{\pi 1}\phi(F) + \gamma_{\pi 2}\phi(F)^2 - 2\beta \frac{f(\beta F^{-1}(\alpha))}{f(F^{-1}(\alpha))}\phi(F)\left(\gamma_{\pi 1} + \gamma_{\pi 2}\alpha\right), \tag{9}$$

where $\gamma_{\pi 1} = \mu_{\pi 1} + \lambda$ and $\gamma_{\pi 2} = \mu_{\pi 2} - \lambda$.

Comparison

- Février and Ragache (2001) : similar to us.
- Breslow and Wellner (2007) and Saegusa and Wellner (2013) : equal probabilities.
- Wang (2012) : similar to us when centered by F but assumptions missing for tightness.
- Conti (2014) : similar to us.
- Bertail, Chautru and Clémençon (2016) : more general empirical processes, results for Poisson and high entropy designs.

Conclusions :

- Results also for Hájek.
- Results for random n and inclusion probabilities.
- Results for high entropy designs with conditions on the rate at which $d_N = \sum_{i \in U} p_i(1 - p_i)$ tends to infinity, compared to N and n.

Perspectives :

- Take into account auxiliary information (at the design stage or at the estimation stage)
- Consider stratified sampling designs,. . .

Thank you for your attention !

Some papers from the literature

- Berger, Y. G. and Skinner, C. J. (2003). Variance estimation for a low income proportion. J. Roy. Statist. Soc. Ser. C 52 457-468.
- Bertail, P. , Chautru, E. and Clémençon, S. (2014). Empirical processes in survey sampling. Accepted in Scand. J. Statist.
- Billingsley, P. (1999). Convergence of probability measures, second ed. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons, Inc., New York A Wiley-Interscience Publication.
- Boistard, H., Lopuhaä, H. P. and Ruiz-Gazen, A. (2016). Functional central limit theorems for one-stage sampling designs. Accepted in Ann. Statist.
- Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase strati ?ed samples, with application to Cox regression. Scand. J. Statist. 34 86-102.
- Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. Sankhya B 76 234-259.
- Dell, F. and d'Haultfœuille, X. (2008). Measuring the evolution of complex indicators : Theory and application to the poverty rate in France. Ann. Econom. Statist. 90 259-290.
- Rubin-Bleuer, S. and Schiopu Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. Ann. Statist. 33 2789-2810.
- van der Vaart, A. W. (1998). Asymptotic statistics. Cambridge Series in Sta- tistical and Probabilistic Mathematics 3. Cambridge University Press, Cambridge.
- van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York With applications to statistics.
- Wang, J. C. (2012). Sample distribution function based goodness-of-fit test for complex surveys. Comput. Statist. Data Anal. 56 664-679.