Functional central limit theorems in Survey Sampling

Hélène Boistard\textsuperscript{(1)}, Rik Lopuhaā\textsuperscript{(2)} and Anne Ruiz-Gazen\textsuperscript{(1)}

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

6ème Journée Maasc, Ensai
Motivation

Aims :

- Use the empirical processes theory in order to prove functional limit theorems for relevant empirical processes in survey sampling.
- Using this theorems and the theory of empirical processes, derive some asymptotic properties of estimators in the survey sampling framework.
Motivation

Aims:

- Use the empirical processes theory in order to prove functional limit theorems for relevant empirical processes in survey sampling.
- Using this theorems and the theory of empirical processes, derive some asymptotic properties of estimators in the survey sampling framework.

Aims:

- Use the empirical processes theory in order to prove functional limit theorems for relevant empirical processes in survey sampling.
- Using this theorems and the theory of empirical processes, derive some asymptotic properties of estimators in the survey sampling framework.


Février and Ragache (2001)
Motivation

Context

Processes under study

Assumptions

Functional limit theorems

Proof sketch

- process centered by $F_N$
- process centered by $F$

Comparison with other results

Conclusion and perspectives
1 Motivation

2 Context

3 Processes under study

4 Assumptions

5 Functional limit theorems

6 Proof sketch
   - process centered by $\mathbb{F}_N$
   - process centered by $F$

7 Comparison with other results

8 Conclusion and perspectives
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.

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

Design:
for all $N = 1, 2, \ldots$, $S_N = \{s : s \subset U_N\}$: collection of subsets of $U_N$ and $\mathcal{A}_N = \sigma(S_N)$: $\sigma$-algebra generated by $S_N$.
We define a probability measure $P_d$ on the design space $(S_N, \mathcal{A}_N)$.

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

$$P_{d,m}(\{s\} \times E) = P_d(\{s\}) P_m(E).$$
Sample $s$ with $n$ denoting the expectation under the design of the size of the sample.

\[ \xi_i = 1_{\{i \in s\}}, \]

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

\[ \pi_{i_1i_2...i_k} = \mathbb{P}_d(\xi_{i_1} = 1, \xi_{i_2} = 1, \ldots, \xi_{i_k} = 1). \]
1 Motivation

2 Context

3 Processes under study

4 Assumptions

5 Functional limit theorems

6 Proof sketch
   - process centered by \( F_N \)
   - process centered by \( F \)

7 Comparison with other results

8 Conclusion and perspectives
Horvitz-Thompson processes

The Horvitz-Thompson (HT) empirical processes obtained from the HT empirical c.d.f.:

\[ F^\text{HT}_N(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i 1\{Y_i \leq t\}}{\pi_i}, \quad t \in \mathbb{R}. \]  \hspace{1cm} (1)
Horvitz-Thompson processes

The Horvitz-Thompson (HT) empirical processes obtained from the HT empirical c.d.f.:

\[ F_{N}^{HT}(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_{i} \mathbb{1}_{\{Y_{i} \leq t\}}}{\pi_{i}}, \quad t \in \mathbb{R}. \]  

- \( \sqrt{n}(F_{N}^{HT} - F_{N}) \), centered around the empirical c.d.f. \( F_{N} \) of the \( Y_{i} \)'s,
- \( \sqrt{n}(F_{N}^{HT} - F) \), centered around c.d.f. \( F \) of the \( Y_{i} \)'s.
The Hájek empirical processes

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

\[
F_{\hat{N}}^{HJ}(t) = \frac{1}{\hat{N}} \sum_{i=1}^{N} \frac{\xi_i 1\{Y_i \leq t\}}{\pi_i}, \quad t \in \mathbb{R}, \quad (2)
\]

where \( \hat{N} = \sum_{i=1}^{N} \xi_i / \pi_i \) is the HT estimator for the population total \( N \).
Hájek processes

The Hájek empirical processes

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

$$F^\text{HJ}_N(t) = \frac{1}{\hat{N}} \sum_{i=1}^{N} \frac{\xi_i \mathbb{1}\{Y_i \leq t\}}{\pi_i}, \quad t \in \mathbb{R},$$

(2)

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

- $\sqrt{n} \left( F^\text{HJ}_N - F_N \right)$
- $\sqrt{n} \left( F^\text{HJ}_N - F \right)$.

In this presentation, we will only consider the Horvitz-Thompson processes.
A functional central limit theorem for $\sqrt{n} \left( \mathbb{F}_{n}^{HT} - \mathbb{F}_{N} \right)$ and $\sqrt{n} \left( \mathbb{F}_{n}^{HT} - F \right)$ 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).
In order to prove the tightness condition, we make assumptions on the design.

Let

\[ D_{\nu,N} = \left\{ (i_1, i_2, \ldots, i_{\nu}) \in \{1, 2, \ldots, N\}^{\nu} : i_1, i_2, \ldots, i_{\nu} \text{ all different} \right\}, \quad (3) \]

for the integers \(1 \leq \nu \leq 4\).
Assumptions on the design

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

$$0 < K_1 \leq \frac{N\pi_i}{n} \leq K_2 < \infty,$$

(C2)

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

(C3)

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

(C4)

$$\limsup_{N \to \infty} \frac{N^4}{n^2} \max_{(i,j,k,l) \in D_{4,N}} \left| \mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)(\xi_l - \pi_l) \right| < \infty,$$
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).
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.
Small remark

\[
E \left( \prod_{j=1}^{k} (\xi_{ij} - \pi_{ij})^{n_{ij}} \right)
\]

\[
= \sum_{m=2}^{k} \sum_{(i_1,\ldots,i_m) \in D_{m,k}} (\pi_{i_1,\ldots,i_m} - \pi_{i_1} \cdots \pi_{i_m}) a_{i_1} \cdots a_{i_m} b_{i_{m+1}} \cdots b_{i_k}
\]

\[
+ \prod_{j=1}^{k} \left[ \pi_{ij} (1 - \pi_{ij})^{n_{ij}} + (-1)^{n_{ij}} \pi_{ij}^{n_{ij}} (1 - \pi_{ij}) \right],
\]

where \( a_i = (1 - \pi_i)^{n_i} - (-1)^{n_i} \pi_i^{n_i} \) and \( b_i = (-1)^{n_i} \pi_i^{n_i} \).

Note that \( \prod_{j=1}^{k} \left[ \pi_{ij} (1 - \pi_{ij})^{n_{ij}} + (-1)^{n_{ij}} \pi_{ij}^{n_{ij}} (1 - \pi_{ij}) \right] = 0 \) as soon as one of the powers \( n_{ij} \) is equal to 1.
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 \((\Omega, \mathcal{F}, \mathbb{P}_m)\), 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.,

\[
S_N^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} V_i V_j. \tag{4}
\]
Assumptions on the HT estimator

(HT1) For any sequence of bounded i.i.d. random variables $V_1, V_2, \ldots,$

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

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

(Boistard, Lopuhaä & Ruiz-Gazen)
Assumptions on the HT estimator

(HT1) For any sequence of bounded i.i.d. random variables $V_1, V_2, \ldots$,

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

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

(HT1) holds for simple random sampling without replacement if $n(N - n)/N$ tends to infinity when $N$ tends to infinity (see Thompson, 1997), as well as for Poisson sampling under some conditions on the first order inclusion probabilities (see Fuller, 2009). For rejective sampling, Hájek (1964) gives some sufficient conditions for (HT1) to hold.
Assumptions on the HT estimator

We also need that $nS_N^2$ converges in probability under $P_m$ to a constant and (HT2) and (HT3) are sufficient conditions:

(HT2) There exist constants $\gamma_{\pi_1}, \gamma_{\pi_2} \in \mathbb{R}$ such that

(i) $\lim_{N \to \infty} \frac{n}{N^2} \sum_{i=1}^{N} \frac{1}{\pi_i} = \gamma_{\pi_1}$,

(ii) $\lim_{N \to \infty} \frac{n}{N^2} \sum_{i \neq j} \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = \gamma_{\pi_2}$.

(HT3) $n/N \to \lambda$, where $\lambda \in [0, 1]$. 
Assumptions on the HT estimator

We also need that \( nS^2_N \) converges in probability under \( \mathbb{P}_m \) to a constant and (HT2) and (HT3) are sufficient conditions:

**(HT2)** There exist constants \( \gamma_{\pi 1}, \gamma_{\pi 2} \in \mathbb{R} \) such that

\[
(i) \quad \lim_{N \to \infty} \frac{n}{N^2} \sum_{i=1}^{N} \frac{1}{\pi_i} = \gamma_{\pi 1},
\]

\[
(ii) \quad \lim_{N \to \infty} \frac{n}{N^2} \sum_{i \neq j} \sum_{i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = \gamma_{\pi 2}.
\]

**(HT3)** \( n/N \to \lambda \), where \( \lambda \in [0, 1] \).

**(HT4)** \( \gamma_{\pi 1} \neq \lambda \) (only needed for the process centered by the super-population c.d.f.).
Assumptions on the HT estimator

(HT2) holds with $\gamma_{\pi 1} = 1$ and $\gamma_{\pi 2} = \lambda - 1$ for simple random sampling without replacement. For Poisson sampling, (HT2)(ii) holds because the trials are independent. For rejective sampling, (HT2) and (HT3) can be deduced from the associated Poisson sampling design. Indeed, suppose that (HT2)(i) holds for Poisson sampling with first order inclusion probabilities $p_1, \ldots, p_N$, such that $\sum_{i=1}^{N} p_i = n$. Then, from Theorem 1 in Boistard et al. (2012), if $d = \sum_{i=1}^{N} p_i (1 - p_i)$ tends to infinity, assumption (HT2)(i) holds for rejective sampling. Furthermore, if (HT3) holds and $N/d$ has a finite limit, then also (HT2)(ii) holds for rejective sampling.

(HT4) holds for simple random sampling without replacement as soon as $\lambda \neq 1$. 
1 Motivation

2 Context

3 Processes under study

4 Assumptions

5 Functional limit theorems

6 Proof sketch
   - process centered by $F_N$
   - process centered by $F$

7 Comparison with other results

8 Conclusion and perspectives
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)-(HT3) hold.

Then $\sqrt{n}(\mathbb{F}_{N}^{\text{HT}} - \mathbb{F}_{N})$ converges weakly in $D(\mathbb{R})$ to a mean zero Gaussian process $\mathbb{G}^{\text{HT}}$ with covariance kernel

$$\mathbb{E}_{d,m} \mathbb{G}^{\text{HT}}(s) \mathbb{G}^{\text{HT}}(t) = (\gamma_{\pi 1} - \lambda)F(s \wedge t) + \gamma_{\pi 2}F(s)F(t), \quad s, t \in \mathbb{R};$$
HT process centered by $F$

**Theorem 2**

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

Then $\sqrt{n}(\mathbb{F}_{HT}^N - F)$ converges weakly in $D(\mathbb{R})$ to a mean zero Gaussian process $\mathbb{G}_{HT}^F$ with covariance kernel

$$E_{d,m}\mathbb{G}_{HT}^F(s)\mathbb{G}_{HT}^F(t) = \gamma_{\pi 1}F(s \wedge t) + (\gamma_{\pi 2} - \lambda)F(s)F(t), \quad s, t \in \mathbb{R};$$
Motivation

Context

Processes under study

Assumptions

Functional limit theorems

Proof sketch
  - process centered by $F_N$
  - process centered by $F$

Comparison with other results

Conclusion and perspectives
Proof sketch process centered by $F_N$

**Tightness condition**

**Lemma 1**

Let $X_N = \sqrt{n}(F_N^{HT} - F_N)$ 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 $-\infty < t_1 \leq t \leq t_2 < \infty$,

$$
E_{d,m} \left[ (X_N(t) - X_N(t_1))^2 (X_N(t_2) - X_N(t))^2 \right] \leq K \left( F(t_2) - F(t_1) \right)^2.
$$
Lemma 2

Let $X_N = \sqrt{n}(F_N^{HT} - F_N)$ and suppose that (C1)-(C2),(HT1)-(HT3) hold. For any $k \in \{1, 2, \ldots\}$, and $t_1, \ldots, t_k \in \mathbb{R}$, $(X_N(t_1), \ldots, X_N(t_k))$ converges in distribution under $\mathbb{P}_{d,m}$ to a $k$-variate mean zero normal random vector with covariance matrix $\Sigma_{HT} = (\sigma_{HT,ij})^k_{i,j=1}$, where

$$\sigma_{HT,ij} = (\gamma_{\pi 1} - \lambda)F(t_i \wedge t_j) + \gamma_{\pi 2}F(t_i)F(t_j),$$
Use of Billingsley (1999) theorem 13.5

- First for the $Y_i$'s uniformly distributed on $[0; 1]$, 
- Then using the proof of Theorem 14.3 (Billingsley, 1999) to generalize to any distribution.
1 Motivation
2 Context
3 Processes under study
4 Assumptions
5 Functional limit theorems
6 Proof sketch
   - process centered by $F_N$
   - process centered by $F$
7 Comparison with other results
8 Conclusion and perspectives
Remarks

More terms in the tightness condition

\[
\sqrt{n} \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i V_i}{\pi_i} - \mu V \right) \\
= \sqrt{n} \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\xi_i V_i}{\pi_i} - \frac{1}{N} \sum_{i=1}^{N} V_i \right) + \frac{\sqrt{n}}{\sqrt{N}} \times \sqrt{N} \left( \frac{1}{N} \sum_{i=1}^{N} V_i - \mu V \right).
\]

+ 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.
Comparison with other results

1. Motivation
2. Context
3. Processes under study
4. Assumptions
5. Functional limit theorems
6. Proof sketch
   - process centered by $F_N$
   - process centered by $F$

7. Comparison with other results
8. Conclusion and perspectives
Comparison

- Wang (2012): similar to us when centered by $F$ but problem in the proof and assumptions missing.
- Lumley (2014): more general but assumptions not easy to interpret and check.
Conclusion and perspectives

1. Motivation
2. Context
3. Processes under study
4. Assumptions
5. Functional limit theorems
6. Proof sketch
   - process centered by $F_N$
   - process centered by $F$
7. Comparison with other results
8. Conclusion and perspectives
Possible to apply the theory by Billingsley quite directly.

Results also for Hájek.

Application to prove the asymptotic normality of the poverty rate.

Take into account auxiliary information (at the design stage or at the estimation stage). . .
Thank you for your attention!