Some contributions to Sampling and Estimation in Surveys

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

Defense of Habilitation à Diriger des Recherches
28/11/2014

Overview of the manuscript

Overview of the manuscript

1. Balanced sampling
2. Treatment of item non-response
3. Variance estimation
   - F.J. Breidt, G. Chauvet (2011). *Improved variance estimation for balanced samples drawn via the Cube method*. JSPI.
   - G. Chauvet (2011). *On variance estimation for the French Master Sample*. JOS.
   - G. Chauvet (201X). *Variance Estimation for the 2006 French Housing Survey*. To appear in Mathematical Population Studies (invited submission).
   - G. Chauvet, C. Goga (201X). *Gini coefficient and Gini coefficient change: linearization versus Bootstrap to estimate the variance*. In revision for Survey Methodology.
4. Coupling methods

Overview of the manuscript

1. Balanced sampling
2. Treatment of item non-response
3. Variance estimation: 3+1 papers
   - F.J. Breidt, G. Chauvet (2011). *Improved variance estimation for balanced samples drawn via the Cube method*. JSPI.
   - G. Chauvet (2011). *On variance estimation for the French Master Sample*. JOS.
   - G. Chauvet (201X). *Variance Estimation for the 2006 French Housing Survey*. To appear in Mathematical Population Studies (invited submission).
   - G. Chauvet, C. Goga (201X). *Gini coefficient and Gini coefficient change: linearization versus Bootstrap to estimate the variance*. In revision for Survey Methodology.
4. Coupling methods

Overview of the talk

1. Balanced sampling
2. Treatment of item non-response
3. Coupling Methods

Notation

We consider a finite labeled population $U = \{1, \ldots, N\}$ with some variable of interest y. We are interested in some parameter θ, such as:

$$\begin{aligned} \text{a total} : t_y &= \sum_{k \in U} y_k \\ \text{a pop c.d.f.} : F_N(t) &= \frac{1}{N} \sum_{k \in U} 1(y_k \leq t). \end{aligned}$$

A random sample S is selected in U by means of some sampling design $p(\cdot)$. We note $\pi = (\pi_1, \ldots, \pi_N)^\top$ the vector of first-order inclusion probabilities.

The Horvitz-Thompson (HT) estimator

$$\hat{t}_{y\pi} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \tag{1}$$

is design-unbiased for t_y, with $I = (I_1, \ldots, I_N)^\top$ the vector of sample membership indicators.

Balanced sampling

Contributions: 6+2 papers

- G. Chauvet, Y. Tillé (2006). *A fast algorithm of Balanced Sampling*. Computational Statistics.
- G. Chauvet, Y. Tillé (2007). *Application of Fast SAS Macros for Balancing Samples to the Selection of Addresses*. Case Studies in Business, Industry, and Government Statistics.
- G. Chauvet (2009). *Stratified Balanced Sampling*. Survey Methodology.
- G. Chauvet, D. Bonnery, J.C. Deville (2011). *Optimal inclusion probabilities for balanced sampling*. JSPI.
- G. Chauvet (2012). *On a characterization of ordered pivotal sampling*. Bernoulli.
- F.J. Breidt, G. Chauvet (2012). *Penalized Balanced Sampling*. Biometrika.
- G. Chauvet, D. Haziza et E. Lesage (201X). *Examining some aspects of balanced sampling in surveys*. In revision for Statistica Sinica.
- G. Chauvet, A. Ruiz-Gazen (201X). *A comparison of pivotal sampling and unequal probability sampling with replacement*. Submitted.

Principle

The accuracy of HT-estimators relies on auxiliary information, frequently incorporated by using some form of balanced sampling.

Suppose that a q-vector x_k is known at the design stage for any $k \in U$. A sampling design $p(\cdot)$ is balanced on x_k if

$$\forall s \subset U \quad p(s) > 0 \Rightarrow \hat{t}_{x\pi}(s) = t_x. \tag{2}$$

The balancing equation (2) is equivalent to

$$\sum_{k \in U} \frac{x_k}{\pi_k}(I_k - \pi_k) = 0 \quad \Leftrightarrow \quad A\,(I - \pi) = 0 \tag{3}$$

where $A = \left(\frac{x_1}{\pi_1}, \ldots, \frac{x_N}{\pi_N}\right)$. Balanced sampling may be performed by means of the cube method [DT04]: random walk from π to I so that (3) is approximately satisfied.

General procedure for the cube method

Initialize at $\pi(0) = \pi$. Next, at time $t = 0, \cdots, T$:

1. Flight phase: if there exists $u(t) \in Ker(A)$ s.t. $u(t) \neq 0$ and $u_k(t) = 0$ if $\pi_k(t)$ is an integer:
   1. take any such $u(t)$ and the largest values $\lambda_1^*(t)$ and $\lambda_2^*(t)$ s.t.

$$0 \leq \pi(t) + \lambda_1^*(t)u(t) \leq 1 \quad \text{and} \quad 0 \leq \pi(t) - \lambda_2^*(t)u(t) \leq 1.$$

   2. Take $\pi(t+1) = \pi(t) + \delta(t)$ with

$$\delta(t) = \begin{cases} \lambda_1^*(t)u(t) & \text{with proba. } \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}, \\ -\lambda_2^*(t)u(t) & \text{with proba. } \lambda_1^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}. \end{cases}$$

2. Landing phase: otherwise, drop the last column from A and go back to Step 1.

Alternatively, a rejective method can be used [H81; F09; CHL14].

Motivation

Suppose that the variable of interest y follows the linear model

$$y_k = \beta^\top x_k + \epsilon_k \quad \Rightarrow \quad \hat{t}_{y\pi} = \beta^\top \hat{t}_{x\pi} + \hat{t}_{\epsilon\pi}. \tag{4}$$

Balanced sampling withdraws the variability of the first term in (4).

Minimizing a variance approximation of [DT05], [CBD11] propose a choice of the π_k's which reduces the variability of the second term in (4).

[BC11] studied the case when y may be described by a linear mixed model. They proposed a penalized balanced sampling method, where a ranking of the balancing variables is used to limit the balancing error.

A fast procedure for balanced sampling

At any step t of the cube method, the search for a vector in the kernel of A may be time-consuming. A faster solution is:

- to extract from A the sub-matrix A_t whose columns are associated to the $q+1$ first units in U that are still at stake,
- to find a vector $v(t)$ in $Ker(A_t)$, which is complemented with zeros for the rest of the columns in A.

This led to the Macro Fastcube [CT06; CT07] and to the stratified balanced sampling procedure [C09]. Applications include:

- selection of the rotation groups of the New Census [B12],
- sampling the PSUs for the Master Sample [CF09],
- selection of areas in the Labour Force Survey [L09].

Pivotal sampling

When $x_k = \pi_k$ (fixed-size sampling), the fast procedure leads to pivotal sampling [DT98] based on duels between units. This sampling algorithm possesses some nice properties, including:

- computable second-order inclusion probabilities π_{kl}, obtained by [C12] from an exact coupling with Deville's systematic sampling [D88];
- better efficiency than multinomial sampling [CRG14], which entails that the HT-estimator is consistent in mean-square under some mild assumptions [C14];
- asymptotic normality for the HT-estimator [CD09].

Treatment of item non-response

Contributions: 3+3 papers

- G. Chauvet, J.C. Deville, D. Haziza (2011). *On balanced random imputation in surveys*. Biometrika.
- G. Chauvet, D. Haziza (2012). *Fully efficient estimation of coefficients of correlation in the presence of imputed data*. Canadian Journal of Statistics.
- D. Haziza, C-O. Nambeu, G. Chauvet (2014). *Doubly robust imputation procedures for populations containing a large amount of zeroes in surveys*. Canadian Journal of Statistics.
- H. Chaput, G. Chauvet, D. Haziza, L. Salembier, J. Solard (201X). *Joint imputation procedures for categorical variables with application to the French Wealth Survey*. Second revision for JRSS C.
- H. Boistard, G. Chauvet, D. Haziza (201X). *Consistency of the estimated distribution function with missing data under a non-response model*. In revision for Scandinavian Journal of Statistics.
- G. Chauvet, J.C. Deville, D. Haziza (201X). *Adapting the Cube algorithm for balanced random imputation in surveys*. Submitted.

Introduction

Item non-response occurs when some variables of interest (but not all) are missing for some unit $k \in S$. Imputation is typically used to compensate for item non-response.

We focus on simple imputation methods [H09] where some missing value y_k is replaced by some artificial value y_k^*. We will use the following assumptions:

- the units answer independently
  $\Rightarrow Pr(r_k = r_l = 1) = Pr(r_k = 1) \times Pr(r_l = 1)$;
- there exists $\kappa > 0$ such that $Pr(r_k = 1) > \kappa$ for any $k \in S$;
- the data are MAR: $E(y_k|z_k, r_k = 1) = E(y_k|z_k, r_k = 0)$
  for a vector of auxiliary variables z_k known for $k \in S$.

Imputed estimators

The imputed estimators of the total t_y and of the c.d.f. $F_N(t)$ are

$$\begin{aligned}\hat{t}_{yI} &= \sum_{k\in S} d_k r_k y_k + \sum_{k\in S} d_k(1-r_k)y_k^*, \\ \hat{F}_I(t) &= \frac{1}{\hat{N}}\sum_{k\in S} d_k r_k 1(y_k \leq t) + \frac{1}{\hat{N}}\sum_{k\in S} d_k(1-r_k)1(y_k^* \leq t).\end{aligned}$$

Many imputation mechanisms can be motivated by some imputation model

$$m: \; y_k = f(z_k;\beta) + \sigma v_k^{1/2}\epsilon_k, \tag{5}$$

$$\Rightarrow \; I: \; y_k^* = f(z_k;\hat{B}_r)(+\hat{\sigma} v_k^{1/2}\epsilon_k^*). \tag{6}$$

We take $f(z_k;\beta) = z_k^\top \beta$ to simplify. With/without the random residual ϵ_k^*, we obtain random/deterministic regression imputation.

Random regression imputation

The vector of parameters β is estimated by

$$\hat{B}_r = \left(\sum_{k \in S} \omega_k r_k v_k^{-1} z_k z_k^\top \right)^{-1} \sum_{k \in S} \omega_k r_k v_k^{-1} z_k y_k, \tag{7}$$

where ω_k is an imputation weight attached to unit k.

In case of random regression imputation (RRI), it is natural to select the ϵ_k^*'s from the observed residuals with prob. $Pr\left(\epsilon_k^* = e_l\right) = \dfrac{\omega_l}{\sum_{j \in s} \omega_j r_j}$.

THEOREM (CDH09)

Assume that the random residuals ϵ_i^* are selected independently with replacement from the set of observed residuals. Then under mild assumptions: $E_{mpqI} \left| \hat{F}_I(t) - F_N(t) \right| \longrightarrow_{n \to \infty} 0$.

Balanced random imputation

When the total t_y is estimated, the imputed estimator may be written as

$$\hat{t}_{yI} = \sum_{k\in S} d_k r_k y_k + \sum_{k\in S} d_k (1-r_k)(z_k^\top \hat{B}_r) + \hat{\sigma} \sum_{k\in S} d_k (1-r_k)(v_k^{1/2}\epsilon_k^*).$$

The imputation variance is eliminated if

$$\sum_{k\in S} d_k (1-r_k)(v_k^{1/2}\epsilon_k^*) = 0. \tag{8}$$

[CDH09] proposed an adaptation of the cube method to select the random residuals ϵ_k^* so that the balancing equation (8) is approximately satisfied.

THEOREM (CDH09)

Assume that the random residuals ϵ_i^* are selected by means of the Cube method s.t. (8) holds. Then under mild assumptions: $E_{mpqI}\left|\hat{F}_I(t) - F_N(t)\right| \longrightarrow_{n\to\infty} 0.$

Doubly robust imputation

Under the Non-Response Model approach (NM), the response probability $p_k \equiv p(z_k; \alpha)$ is modeled and estimated. [BCH14] considered the mean imputation model within classes, where U is divided into disjoint imputation cells $U_1, \ldots, U_G$:

$$m: \; y_k \sim (\mu_g, \sigma_g^2), \qquad k \in U_g.$$

$$I: \; y_k^* = y_l \text{ for } l \in S_r \cap U_g \quad \text{with} \quad \mathbb{P}(y_k^* = y_l) = \frac{\omega_l}{\sum_{j \in S_g} \omega_j r_j}.$$

THEOREM (BCH14)

Assume that $\omega_k = d_k \, \frac{1-\hat{p}_k}{\hat{p}_k}$, *where* $\hat{p}_k = p(z_k; \, \hat{\alpha})$ *and* $\hat{\alpha}$ *is a consistent estimator of* α. *Then under mild assumptions :*
$E_{mpqI}|\hat{F}_I(t) - F_N(t)| \longrightarrow_{n\to\infty} 0$ *under the IM approach,*
$E_{pqI}|\hat{F}_I(t) - F_N(t)| \longrightarrow_{n\to\infty} 0$ *under the NM approach.*

Taylor-made imputation methods

In practice, the imputation regression model may not be appropriate. For example, if the study variable contains a large number of zeroes, it seems natural to postulate

$$m : y_k = \begin{cases} z_k^\top \beta + \sigma_k \epsilon_k & \text{w.p. } \phi_k, \\ 0 & \text{w.p. } 1 - \phi_k, \end{cases} \Rightarrow I : y_k^* = \begin{cases} z_k^\top \hat{B}_{\phi r} & \text{w.p. } \hat{\phi}_k, \\ 0 & \text{w.p. } 1 - \hat{\phi}_k. \end{cases}$$

[HNC14] proposed doubly robust balanced imputation methods for estimating t_y under this imputation model.

[CH11] considered balanced imputation methods to preserve the correlation between continuous variables. [CCHSS11] considered balanced hot-deck methods to preserve the correlation between categorical variables, with application to the French Wealth Survey.

Coupling methods

Contributions: 1 paper + 2 works in progress

- G. Chauvet (201X). *Coupling Methods for multistage sampling*. Submitted.
- G. Chauvet, J.C. Deville (201X). *Asymptotic Results for Deville's Systematic Sampling*.
- G. Chauvet, J. Opsomer (201X). *Coupling Methods for two-phase sampling*.

Overview of the chapter

1. Introduction: what is a coupling?
2. Multistage sampling:
   - A coupling algorithm between SI/BE sampling
   - Asymptotic normality of the HT-estimator
3. Multistage sampling:
   - A coupling algorithm between SI/SIR sampling
   - Validity of a bootstrap method

Introduction

Introduction

The dependence in the selection of units may be complex, which makes limiting results quite difficult to prove. In some cases, we can resort to coupling methods [T00] to link a sampling design under study to a close, simpler sampling design.

We look for a random vector $(X_t, Z_t)^\top$ such that:

1. X_t has an appropriate marginal law (e.g., that of the HT estimator $N^{-1}\hat{t}_{y\pi}$ under the sampling design);
2. Z_t has a marginal law which is simpler to study;
3. X_t and Z_t are close: $E(X_t - Z_t)^2$ is smaller than the rate of convergence of X_t.

Introduction (2)

Lemma

Let X_t and Z_t denote two random variables such that $E(X_t) = E(Z_t)$. Assume that

$$V(X_t) = O(a_t) \quad and \quad E(X_t - Z_t)^2 = o(a_t),$$

where $a_t \underset{t\to\infty}{\longrightarrow} 0$. Then

$$\frac{V(Z_t)}{V(X_t)} \underset{t\to\infty}{\longrightarrow} 1. \tag{9}$$

Also, if $\sqrt{a_t}\{Z_t - E(Z_t)\} \underset{\mathcal{L}}{\longrightarrow} X_0$, then

$$\sqrt{a_t}\{X_t - E(X_t)\} \underset{\mathcal{L}}{\longrightarrow} X_0.$$

Framework for multistage sampling

We consider a finite population $U = \{1, \ldots, N\}$ of N sampling units. The units are grouped inside N_I Primary Sampling Units $u_1, \ldots, u_{N_I}$. We are interested in estimating the population total

$$Y = \sum_{k \in U} y_k = \sum_{u_i \in U_I} Y_i \quad \text{with} \quad Y_i = \sum_{k \in u_i} y_k,$$

for some variable of interest y. We note $\mu_Y = N_I^{-1} \sum_{u_i \in U_I} Y_i$.

We denote by $\hat{Y}_i$ an unbiased estimator of Y_i, with design variance

$$V_i = V(\hat{Y}_i).$$

Framework for multistage sampling (2)

We consider the asymptotic framework of [IF82]:

- The population U belongs to a nested sequence $\{U_t\}$ of finite populations with increasing sizes N_t.
- The vector of values $y_{U_t} = (y_{1_t}, \ldots, y_{N_t})^\top$ belongs to a sequence $\{y_{U_t}\}$ of N_t-vectors.

The subscript "t" is suppressed in the sequel.

In the population $U_I = \{u_1, \ldots, u_{N_I}\}$ of PSUs:

- a first-stage sample S_I is selected according to some sampling design $p_I(\cdot)$,
- if $u_i \in S_I$, a second-stage sample S_i is selected in u_i by means of any sampling design (census, stratified sampling, multistage sampling, ...).

Assumptions

We assume:

- **Invariance of the second-stage designs:** the second stage of sampling is independent of S_I,
- **Independence of the second-stage designs:** the second-stage designs are independent from one PSU to another, conditionally on S_I.

We will also make use of the following assumptions:

H1: $N_I \underset{t\to\infty}{\longrightarrow} \infty$ and $n_I \underset{t\to\infty}{\longrightarrow} \infty$.

H2: There exists a constant C_1 and $\delta > 0$ such that

$$N_I^{-1} \sum_{u_i \in U_I} E|\hat{Y}_i|^{2+\delta} < C_1.$$

Central limit theorem for multistage sampling

Bernoulli sampling of PSUs

Suppose that the first-stage sample S_I^{BE} is selected by Bernoulli sampling (BE) with N_I independent Bernoulli trials. The HT estimator is

$$\hat{Y}^{BE} = \frac{N_I}{n_I} \sum_{u_i \in S_I^{BE}} \hat{Y}_i.$$

Under assumptions (H1) and (H2), we have

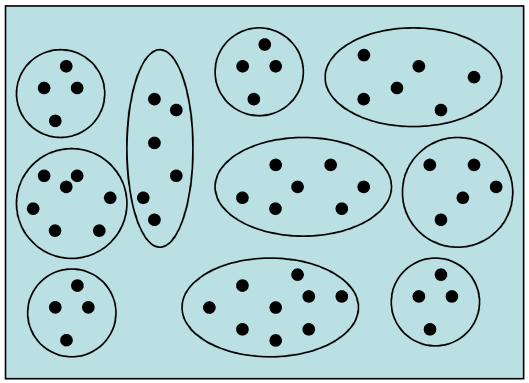
$$\frac{\sum_{u_i \in S_I^{BE}} \left(\hat{Y}_i - \mu_Y \right)}{\sqrt{V \left[\sum_{u_i \in S_I^{BE}} \left(\hat{Y}_i - \mu_Y \right) \right]}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1).$$

If the first-stage sample S_I is selected by means of simple random sampling without replacement (SI), the HT estimator is denoted as

$$\hat{Y} = \frac{N_I}{n_I} \sum_{u_i \in S_I} \hat{Y}_i.$$

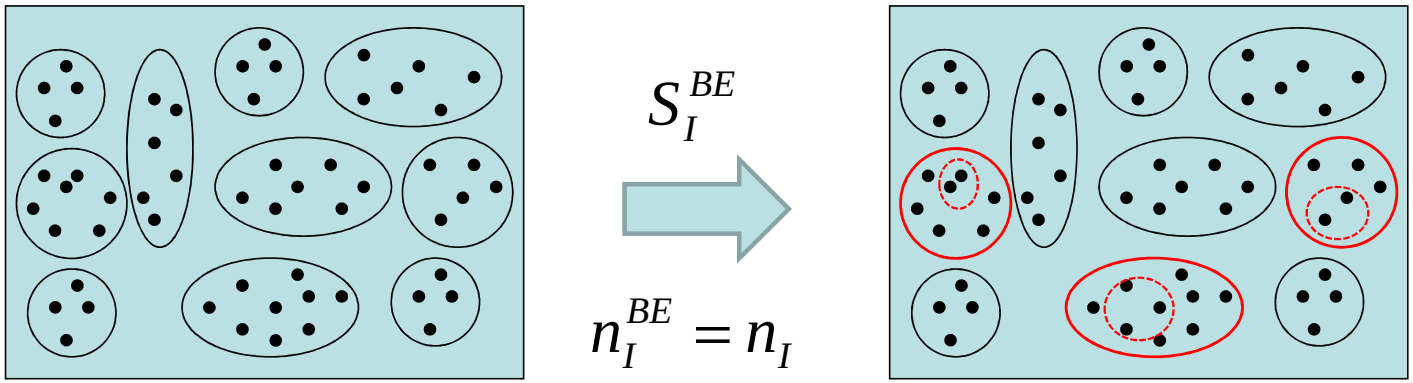
The coupling procedure

Step 1: Draw $S_I^{BE} \sim BE(U_I; n_I)$. Denote by n_I^{BE} its random size.



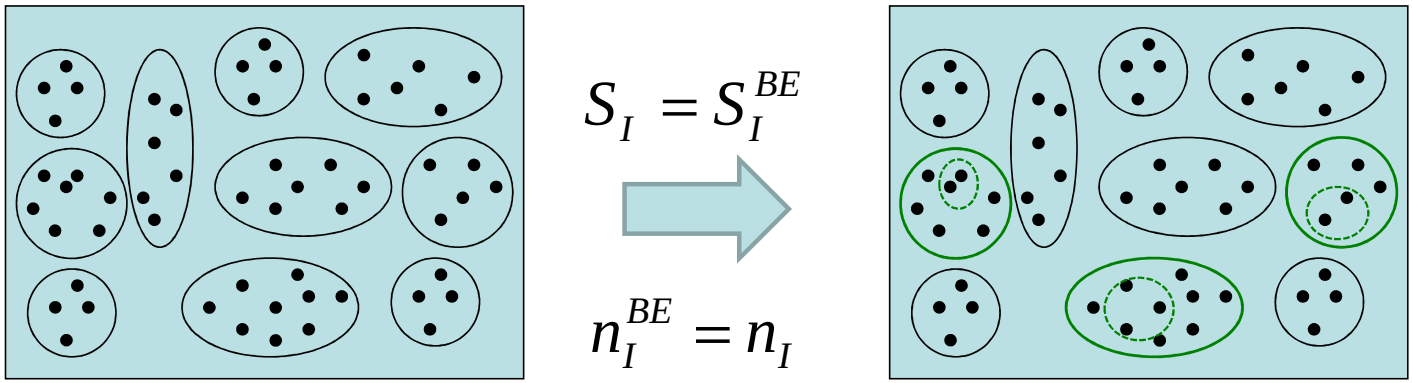
The coupling procedure

Step 2: If $n_I^{BE} = n_I$,



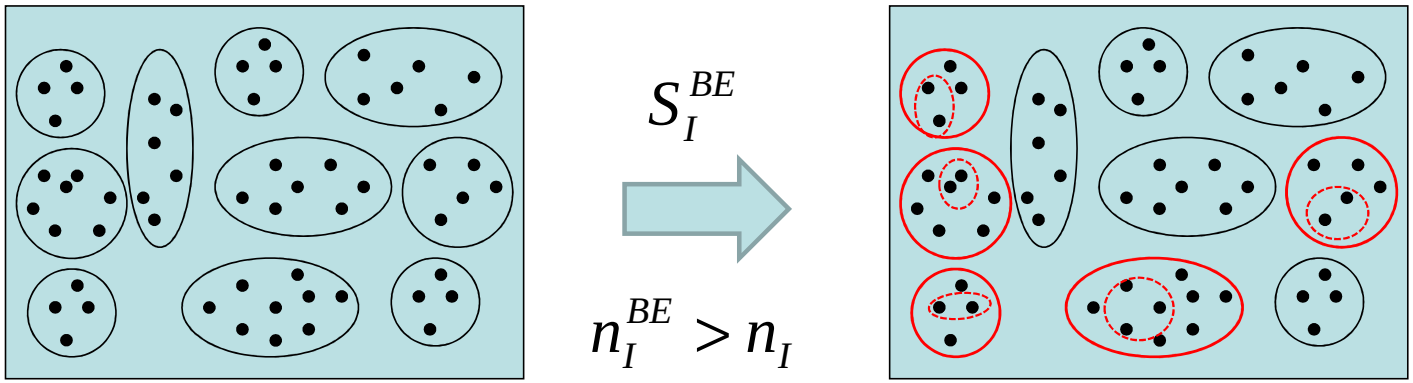
The coupling procedure

Step 2: If $n_I^{BE} = n_I$, take $S_I = S_I^{BE}$.



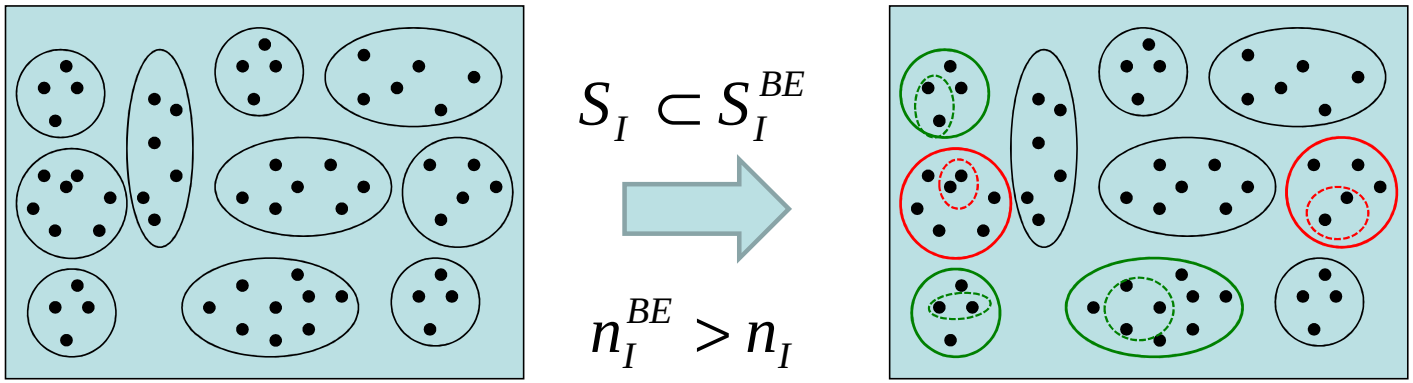
The coupling procedure

Step 2: If $n_I^{BE} > n_I$,



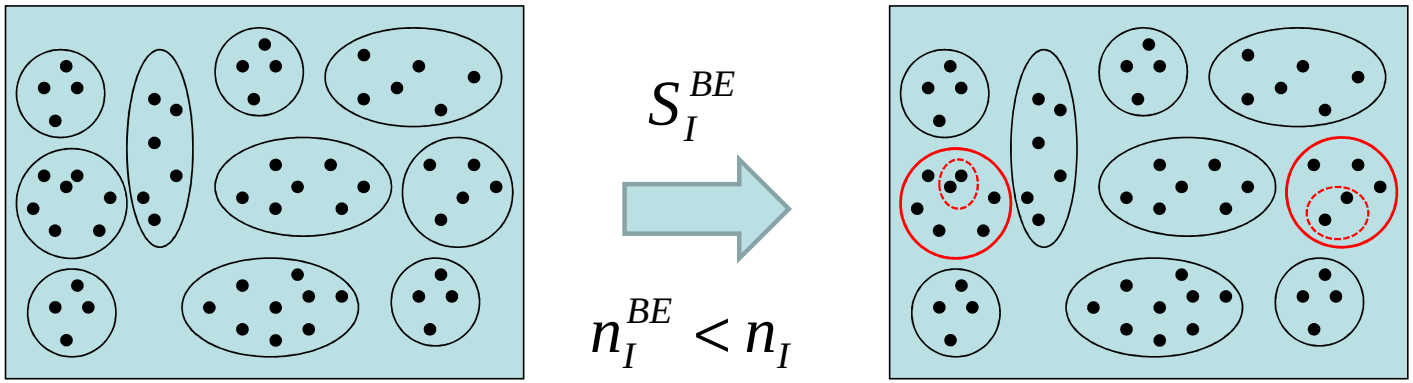
The coupling procedure

Step 2: If $n_I^{BE} > n_I$, draw $S_I \sim SI(S_I^{BE}; n_I)$.



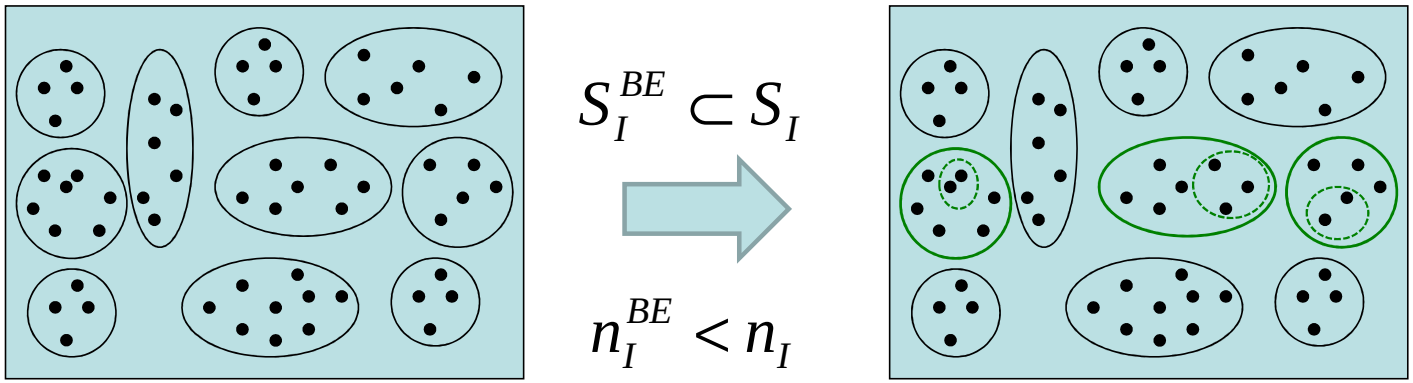
The coupling procedure

Step 2: If $n_I^{BE} < n_I$,



The coupling procedure

Step 2: If $n_I^{BE} < n_I$, take $S_I = S_I^{BE} \cup SI(U_I \setminus S_I^{BE}; n_I - n_I^{BE})$.



CLT for SI sampling of PSUs

PROPOSITION

If S_I^{BE} and S_I are selected with the coupling procedure:

$$\frac{E\left[\sum_{u_i \in S_I}\left(\hat{Y}_i - \mu_Y\right) - \sum_{u_i \in S_I^{BE}}\left(\hat{Y}_i - \mu_Y\right)\right]^2}{V\left[\sum_{u_i \in S_I^{BE}}\left(\hat{Y}_i - \mu_Y\right)\right]} \leq \sqrt{\frac{1}{n_I} + \frac{1}{N_I - n_I}}$$

Hint for the proof:

$$N_I^{-1}\sum_{u_i \in S_I}\left(\hat{Y}_i - \mu_Y\right) - N_I^{-1}\sum_{u_i \in S_I^{BE}}\left(\hat{Y}_i - \mu_Y\right) = \epsilon n_I^{-1}\sum_{u_i \in S_I^{+}}\left(\hat{Y}_i - \mu_Y\right),$$

with S_I^+ the surplus/complementary sample, and $\epsilon = Sign(n_I - n_I^{BE})$.

Under assumptions (H1) and (H2), we have

$$\frac{\hat{Y} - Y}{\sqrt{V(\hat{Y})}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1).$$

Bootstrap for multistage sampling

Simple random sampling of PSUs

If the first-stage sample S_I is selected by means of SI sampling, the HT estimator is

$$\hat{Y} = \frac{N_I}{n_I} \sum_{j=1}^{n_I} \hat{Y}_{(j)} \equiv \frac{N_I}{n_I} \sum_{j=1}^{n_I} Z_j,$$

where S_I is obtained in $j = 1, \ldots, n_I$ without-replacement draws.

If the first-stage sample S_I^{WR} is selected by means of simple random sampling with replacement (SIR), the Hansen-Hurwitz estimator is

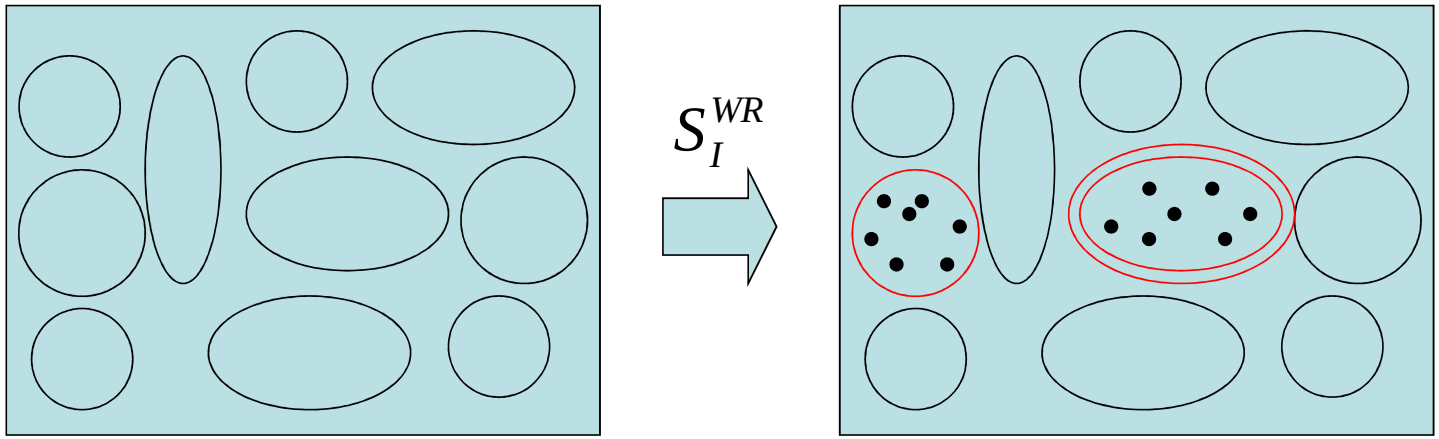
$$\hat{Y}_{WR} = \frac{N_I}{n_I} \sum_{j=1}^{n_I} \hat{Y}_{(j)} \equiv \frac{N_I}{n_I} \sum_{j=1}^{n_I} X_j,$$

where S_I^{WR} is obtained in $j = 1, \ldots, n_I$ independent draws.

The two estimators are expected to be close if the first stage sampling rate $f_I = n_I/N_I$ is small.

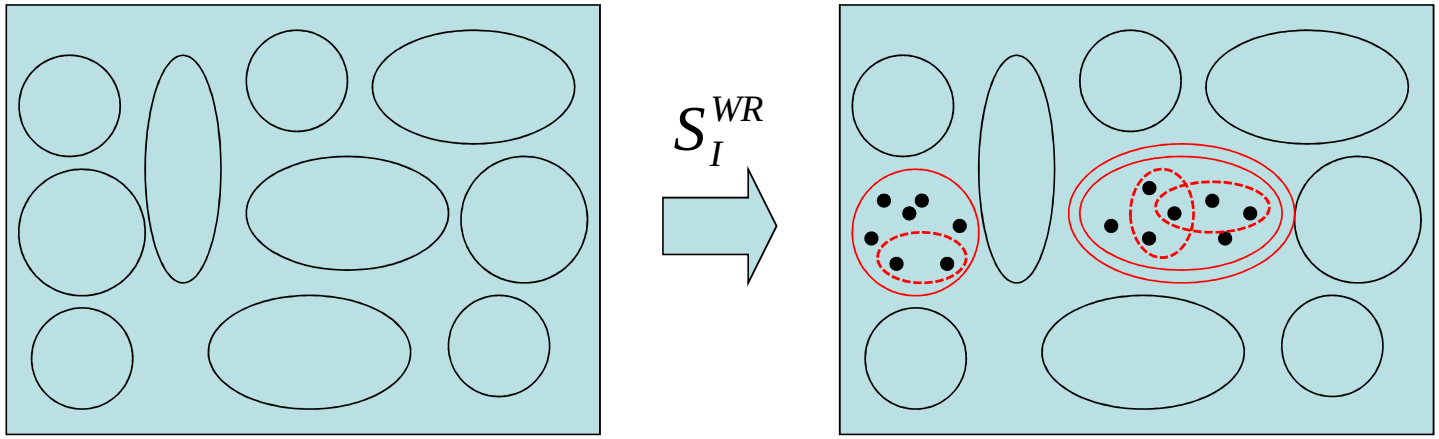
The coupling procedure

Step 1: draw S_I^{WR}. Denote by S_I^d the set of distinct PSUs in S_I^{WR}.



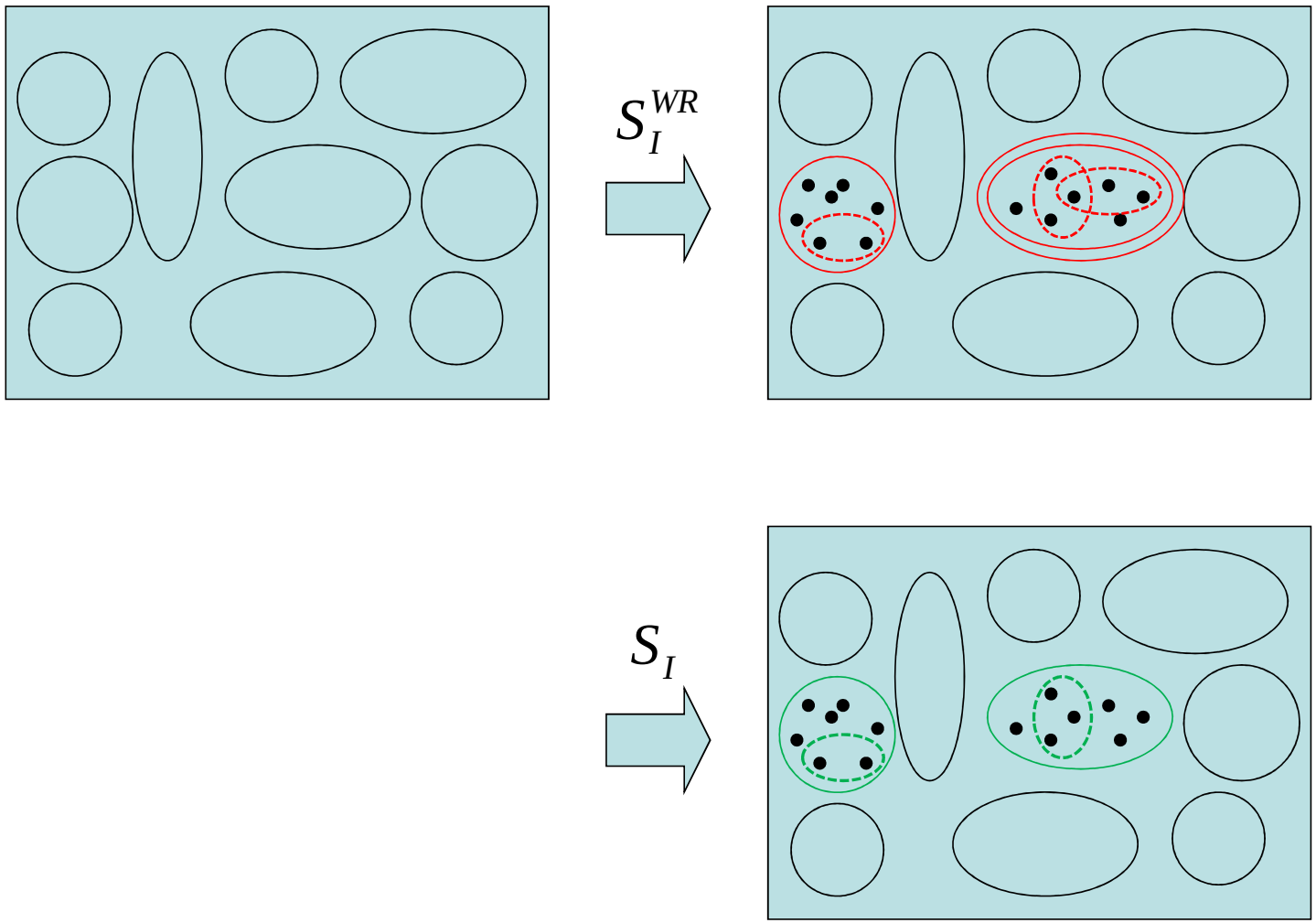
The coupling procedure

Step 2: each time $u_i \in S_I^{WR}$, select a second-stage sample $S_{i[j]}$.



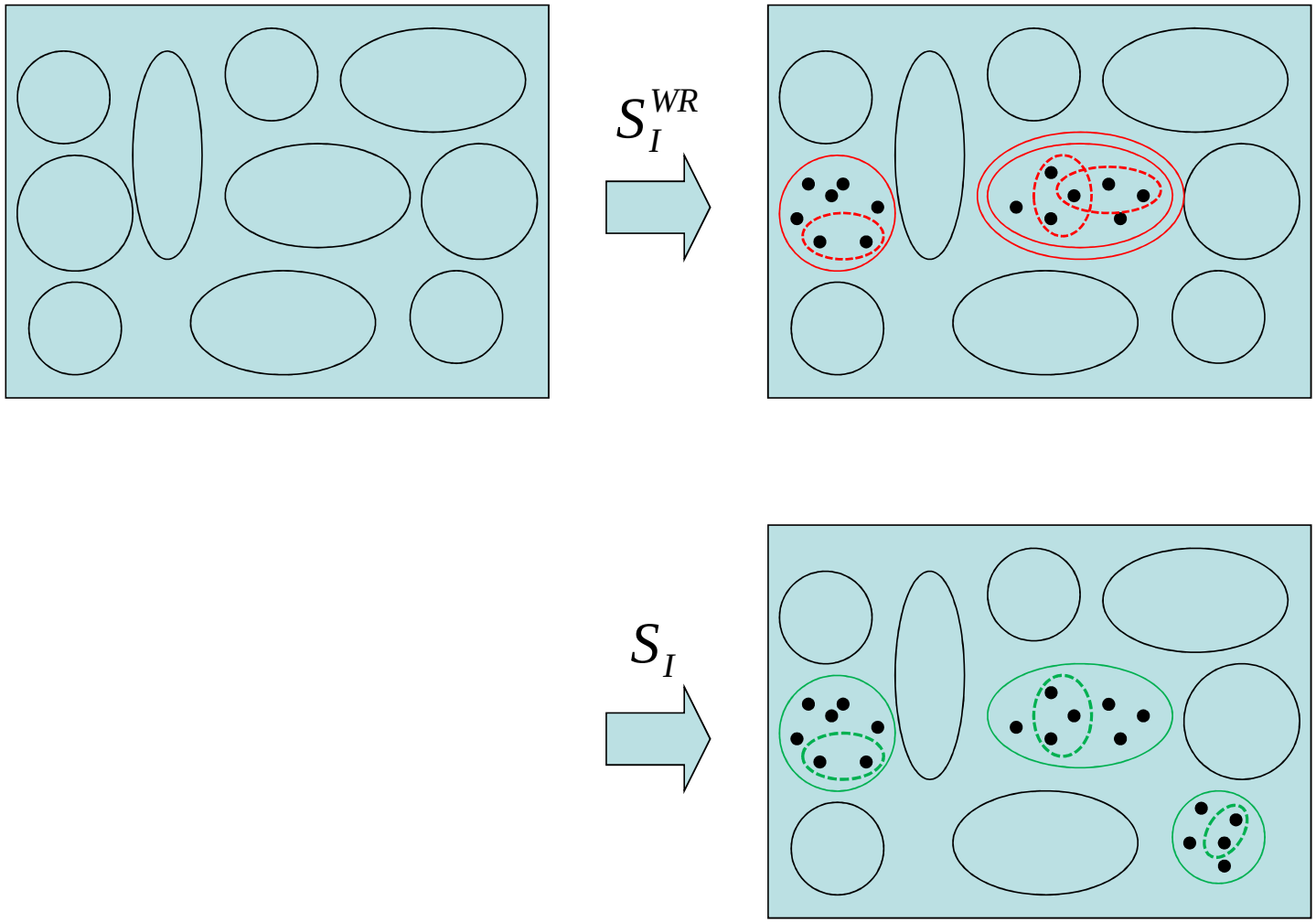
The coupling procedure

Step 3: initialize S_I with S_I^d, and $S_i = S_{i[1]}$ for $u_i \in S_I^d$.



The coupling procedure

Step 4: draw a complementary sample S_I^c, and S_i for $u_i \in S_I^c$.



Plug-in estimation

For some smooth function $f(\cdot)$, we consider the parameter

$$\theta = f(\mu_Y) \quad \text{with} \quad \mu_Y = \frac{1}{N_I} \sum_{u_i \in U_I} Y_i.$$

Under SI or SIR sampling of PSUs, we have

$$\hat{\mu}_Y = \frac{1}{n_I} \sum_{j=1}^{n_I} Z_j \equiv \bar{Z} \quad \text{and} \quad \hat{\theta} = f(\bar{Z}),$$

$$\hat{\mu}_Y^{WR} = \frac{1}{n_I} \sum_{j=1}^{n_I} X_j \equiv \bar{X} \quad \text{and} \quad \hat{\theta}_{WR} = f(\bar{X}).$$

Bootstrap of PSUs

We consider the with-replacement Bootstrap (BWR) of PSUs (Rao and Wu, 1988). The resample $(X_1^*, \ldots, X_m^*)^\top$ is obtained by sampling m times independently in $(X_1, \ldots, X_{n_I})$, and similarly for $(Z_1, \ldots, Z_{n_I})$.

Suppose that S_I^{WR} and S_I are selected according to the coupling procedure + assumptions (H1)-(H2) + $f_I \to 0$ + $m \underset{t\to\infty}{\longrightarrow} \infty$. Then :

$$E(\hat{\theta}^* - \hat{\theta}_{WR}^*)^2 = o(m^{-1}) + o(n_I^{-1}). \tag{10}$$

This implies that

$$\frac{V(\hat{\theta}^*|Z_1, \ldots, Z_{n_I})}{V(\hat{\theta}_{WR}^*|X_1, \ldots, X_{n_I})} \quad \underset{Pr}{\longrightarrow} \quad 1.$$

If the with-replacement Bootstrap provides consistent variance estimation for $\hat{\theta}_{WR}$, it is also consistent for $\hat{\theta}$.

Work in progress

1. Treatment of item non-response
   - G. Chauvet, Do Paco, W., Haziza, D: *Exact balanced imputation for sample survey data*.
2. Variance estimation
   - G. Chauvet, H. Juillard, A. Ruiz-Gazen: *Variance estimation for product sampling: an application to the ELFE survey*.
3. Coupling methods
   - G. Chauvet, J.C. Deville: *Asymptotic Results for Deville's Systematic Sampling*.
   - G. Chauvet, J. Opsomer: *Coupling methods for two-phase sampling*.
   - Extension of the results presented to unequal probability sampling of PSUs.

References

- Brilhaut, G. (2012). La précision du recensement français de la population. 7 ème colloque francophone sur les Sondages.
- Christine, M., and Faivre, S. (2009). Le projet OCTOPUSSE de nouvel Echantillon-Maître de l'INSEE. Journées de Méthodologie Statistique, 2009.
- Deville, J. C. (1998). Une nouvelle méthode de tirage à probabilités inégales. Technical Report 9804, Ensai, France.
- Deville, J. C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. Biometrika, 85(1), 89-101.
- Deville, J. C., and Tillé, Y. (2004). Efficient balanced sampling: the cube method. Biometrika, 91(4), 893-912.
- Deville, J. C., and Tillé, Y. (2005). Variance approximation under balanced sampling. Journal of Statistical Planning and Inference, 128(2), 569-591.
- Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. Biometrika, 96(4), 933-944.
- Hajek, J. (1981). Sampling from a finite population (Vol. 37). V. Dupac (Ed.). M. Dekker.
- Haziza, D. (2009). Imputation and inference in the presence of missing data. Handbook of Statistics, 29, 215-246.
- Isaki, C. T., and Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77(377), 89-96.
- Loonis, V. (2009). La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation. Journées de Méthodologie Statistique, 2009.
- Thorisson, H. (2000). Coupling, stationarity, and regeneration (pp. 90095-1555). New York: Springer.