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

Balanced sampling

Treatment of item non-response

Variance estimation: 3+1 papers

F. J. Breidt, G. Chauvet (2011). Improved variance estimation for balanced samples drawn via the Cube method. JSPI.


coupling methods

G. Chauvet (ENSAI)

Sampling and Estimation in Surveys

HDR Defense
Overview of the manuscript

Balanced sampling

Treatment of item non-response

Variance estimation

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.
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 $\theta$, such as:

- A total: $t_y = \sum_{k \in U} y_k$
- A pop c.d.f.: $F_N(t) = \frac{1}{N} \sum_{k \in U} 1(y_k \leq t)$.

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

The Horvitz-Thompson (HT) estimator

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

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

Contributions: 6+2 papers

Balanced sampling

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

\[
\sum_{k \in U} \frac{x_k}{\pi_k} (I_k - \pi_k) = 0 \quad \iff \quad A (I - \pi) = 0 \tag{3}
\]

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 \( \pi \) to \( I \) so that (3) is approximately satisfied.
General procedure for the cube method

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

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

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

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

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

Minimizing a variance approximation of [DT05], [CBD11] propose a choice of the $\pi_k$'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 \( \pi_{kl} \), 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

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
  \[ \Pr(r_k = r_l = 1) = \Pr(r_k = 1) \times \Pr(r_l = 1); \]
- 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

$$
\hat{t}_{yI} = \sum_{k \in S} d_k r_k y_k + \sum_{k \in S} d_k (1 - r_k) y_k^*,
$$

$$
\hat{F}_I(t) = \frac{1}{\hat{N}} \sum_{k \in S} d_k r_k 1(y_k \leq t) + \frac{1}{\hat{N}} \sum_{k \in S} d_k (1 - r_k) 1(y_k^* \leq t).
$$

Many imputation mechanisms can be motivated by some imputation model

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

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

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

The vector of parameters $\beta$ is estimated by

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

where $\omega_k$ is an imputation weight attached to unit $k$.

In case of random regression imputation (RRI), it is natural to select the $\epsilon^*_k$’s from the observed residuals with prob. $Pr(\epsilon^*_k = e_l) = \frac{\omega_l}{\sum_{j \in s} \omega_j r_j}$.

**Theorem (CDH09)**

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

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

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

**Theorem (CDH09)**

Assume that the random residuals $\epsilon_i^*$ are selected by means of the Cube method s.t. (8) holds. Then under mild assumptions:

$$E_{mpqI} |\hat{F}_I(t) - F_N(t)| \rightarrow_{n \to \infty} 0.$$
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 $U_1, \ldots, U_G$:

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

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

**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 $\alpha$. Then under mild assumptions:

$$E_{mpqI}\left| \hat{F}_I(t) - F_N(t) \right| \longrightarrow_{n \to \infty} 0 \text{ under the IM approach},$$  

$$E_{pqI}\left| \hat{F}_I(t) - F_N(t) \right| \longrightarrow_{n \to \infty} 0 \text{ under the NM approach}.$$
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

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

[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, J.C. Deville (201X). *Asymptotic Results for Deville’s Systematic 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 \((X_t, Z_t)^	op\) 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 \(X_t\).
Lemma

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

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

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

$$\frac{V(Z_t)}{V(X_t)} \xrightarrow{t \to \infty} 1. \quad (9)$$

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

$$\sqrt{a_t}\{X_t - E(X_t)\} \xrightarrow{L} X_0.$$
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 $y_{U_t} = (y_{1t}, \ldots, y_{N_t})^\top$ 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:** $N_I \xrightarrow{t \to \infty} \infty$ and $n_I \xrightarrow{t \to \infty} \infty$.

**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
Coupling Methods

Bernoulli sampling of PSUs

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

$$\hat{Y}^{BE} = \frac{N_I}{n_I} \sum_{i \in S_{BE}^{I}} \hat{Y}_i.$$  

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

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

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 $S_I \sim SI(S_I^{BE}; n_I)$. 

\[ S_I \subseteq S_I^{BE} \]
\[ n_I^{BE} > n_I \]
The coupling procedure

Step 2: If $n_{I}^{BE} < n_{I}$,
The coupling procedure

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

If $S_{I}^{BE}$ and $S_{I}$ are selected with the coupling procedure:

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

**Hint for the proof:**

$$
N_{I}^{-1} \sum_{u_i \in S_{I}} (\hat{Y}_i - \mu_Y) - N_{I}^{-1} \sum_{u_i \in S_{I}^{BE}} (\hat{Y}_i - \mu_Y) = \epsilon n_{I}^{-1} \sum_{u_i \in S_{I}^+} (\hat{Y}_i - \mu_Y),
$$

with $S_{I}^+$ the surplus/complementary sample, and $\epsilon = \text{Sign}(n_{I} - n_{I}^{BE})$.

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 $S_{WR}^I$ 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 $S_{WR}^I$ 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 $S^{WR}_I$. Denote by $S^d_I$ the set of distinct PSUs in $S^{WR}_I$. 
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 $S^d_I$, and $S_i = S_i[1]$ for $u_i \in S^d_I$. 

$S^W_R$ $S_I$ $S_I$
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

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

$$
\hat{\mu}_{YWR} = \frac{1}{n_I} \sum_{j=1}^{n_I} X_j \equiv \bar{X} \quad \text{and} \quad \hat{\theta}_{WR} = f(\bar{X}).
$$
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 \(S^{WR}_I\) and \(S_I\) are selected according to the coupling procedure + assumptions (H1)-(H2) + \(f_I \rightarrow 0 + m \xrightarrow{t \to \infty} \infty\). Then:

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

This implies that

\[
\frac{V(\hat{\theta}^* | Z_1, \ldots, Z_{n_I})}{V(\hat{\theta}^{WR}_* | X_1, \ldots, X_{n_I})} \xrightarrow{Pr} 1.
\]

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

3. Coupling methods
   - G. Chauvet, J.C. Deville: *Asymptotic Results for Deville’s Systematic Sampling.*
   - Extension of the results presented to unequal probability sampling of PSUs.
References