Bootstrap for multistage sampling and without replacement sampling at the first stage

Guillaume Chauvet

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

Séminaire de Statistique
Université de Besançon
17/11/2014
1. Multistage sampling
2. With replacement sampling of PSUs
3. Without replacement sampling of PSUs
4. A coupling procedure between SI/SIR sampling of PSUs
5. A simulation study
Multistage sampling
**Principle of multistage sampling**

The population $U$ of individuals is partitioned into $M$ big units called Primary Sampling Units (PSUs); the small units in $U$ are called the Secondary Sampling Units (SSUs).

- First stage: a sample $S_I$ of PSUs is selected.
- Second stage: a sample of SSUs is drawn in the selected PSUs $u_i$.

Multistage sampling consists in three stages of sampling, or more. In case of household surveys, a customary sampling design consists in

- selecting a sample of municipalities (PSUs),
- selecting a sample of districts inside the selected municipalities (SSUs),
- selecting a sample of households inside the selected districts (TSUs).
Motivation

Multistage sampling is mainly used for practical purposes:

- **Reducing the survey costs** when direct sampling would lead to a scattered sample. Using several stages of sampling enables to group the selected units.

- **Building of the sampling frame.** We only need a list of the final units inside the selected PSUs.
Examples

1. Household surveys: selection of a sample of municipalities (PSUs), of districts (SSUS) within, and of households (TSUs) inside (e.g., Ardilly, 2006).

2. Epidemiologic surveys: estimation of lead contamination by the selection of a sample of hospitals (PSUs), and then of children (SSUs) whose dwellings were investigated (Lucas, 2013).

3. PISA survey: in France, selection of a sample of schools (PSUs), and of a sample of students aged 15 within (SSUs).
Framework

We consider a finite population $U = \{1, \ldots, N\}$ of $N$ sampling units. The units are grouped inside $N_I$ non-overlapping subpopulations $u_1, \ldots, u_{N_I}$ called primary sampling units (PSUs). 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 denote by:

- $\hat{Y}_i$ an unbiased estimator of $Y_i$, with design variance $V_i = V(\hat{Y}_i)$,
- $\hat{V}_i$ an unbiased estimator of $V_i$. 

Framework

We consider the asymptotic framework of Isaki and Fuller (1982):

- The population $U$ belongs to a nested sequence $\{U_t\}$ of finite populations with increasing sizes $N_t$.
- The vector of values $y_{Ut} = (y_{1t}, \ldots, y_{Nt})^\top$ belongs to a sequence $\{y_{Ut}\}$ 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 $C_1$ such that $N_I^{-1} \sum_{u_i \in U_I} E|\hat{Y}_i|^4 < C_1$.

**H3**: There exists a constant $C_2$ such that $N_I^{-1} \sum_{u_i \in U_I} E(\hat{V}_i^2) < C_2$. 
With replacement sampling of PSUs
With replacement simple random sampling of PSUs

The first-stage sample $S_{WR}^{I}$ is selected by means of simple random sampling with replacement (SIR). The Hansen-Hurwitz estimator is

$$\hat{Y}_{WR} = \frac{N_{I}}{n_{I}} \sum_{j=1}^{n_{I}} \hat{Y}(j),$$

where

- $S_{WR}^{I}$ is obtained in $j = 1, \ldots, n_{I}$ independent draws,
- at each draw, a PSU $u_{(j)}$ with associated estimator $X_{j} \equiv \hat{Y}(j)$.

The variance of $\hat{Y}_{WR}$ and an unbiased variance estimator are

$$V\left(\hat{Y}_{WR}\right) = \frac{N_{I}^{2}}{n_{I}} \left\{ \frac{N_{I} - 1}{N_{I}} S_{Y,U_{I}}^{2} + \frac{1}{N_{I}} \sum_{u_{i} \in U_{I}} V_{i} \right\}$$

$$v_{WR}\left(\hat{Y}_{WR}\right) = \frac{N_{I}^{2}}{n_{I}} s_{X}^{2} \text{ with } s_{X}^{2} = \frac{1}{n_{I} - 1} \sum_{j=1}^{n_{I}} (X_{j} - \bar{X}_{n})^{2}.$$
With replacement sampling of PSUs

With replacement simple random sampling of PSUs

The simple form of the variance estimator is primarily due to the writing of $\hat{Y}_{WR}$ as a sum of independent random variables.

Under the assumptions:

H1: $N_I \xrightarrow{t \to \infty} \infty$ and $n_I \xrightarrow{t \to \infty} \infty$,

H2: there exists a constant $C_1$ such that $N_I^{-1} \sum_{u_i \in U_I} E|\hat{Y}_i|^4 < C_1$, we have

$$E \left| \frac{n_I}{N_I^2} \left\{ v_{WR} \left( \hat{Y}_{WR} \right) - V \left( \hat{Y}_{WR} \right) \right\} \right|^2 \xrightarrow{t \to \infty} 0.$$  

A variance estimator for further stages inside the selected PSUs is not needed.
Bootstrap for SIR of PSUs

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

\[
\bar{X}_m^* = \frac{1}{m} \sum_{j=1}^{m} X_j^* \quad \text{and} \quad s_X^2 = \frac{1}{m-1} \sum_{j=1}^{m} (X_j^* - \bar{X}_m^*)^2.
\]

Assume that (H1)-(H2) hold, and that \(m \rightarrow \infty\). Then (Bickel and Freedman, 1981):

\[
\frac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s_X^*} \xrightarrow{L} \mathcal{N}(0, 1).
\]

Using the BWR with \(m = n_I - 1\) enables to match the unbiased variance estimator \(v_{WR}(\hat{Y}_{WR})\) when estimating the total \(Y\).
Without replacement sampling of PSUs
Without replacement simple random sampling of PSUs

The first-stage sample $S_I$ is selected by means of simple random sampling without replacement (Sl). The Horvitz-Thompson estimator is

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

where

- $S_I$ is obtained in $j = 1, \ldots, n_I$ without-replacement draws,
- at each draw, a PSU $u(j)$ with associated estimator $Z_j \equiv \hat{Y}(j)$.

The variance of $\hat{Y}$ and an unbiased variance estimator are

$$V(\hat{Y}) = \frac{N_I^2}{n_I} \left\{ (1 - f_I)S_{Y,I}^2 + \frac{1}{N_I} \sum_{u_i \in U_I} V_i \right\}$$

$$\nu(\hat{Y}) = \frac{N_I^2}{n_I} \left\{ (1 - f_I)s_Z^2 + \frac{1}{N_I} \sum_{u_i \in S_I} \hat{V}_i \right\} \text{ with } f_I = n_I/N_I.$$
Without replacement simple random sampling of PSUs

Since $\hat{Y}$ is a sum of dependent random variables, there is no such simple unbiased variance estimator as for SIR sampling of PSUs.

Under the assumptions:

**H1:** $N_I \xrightarrow{t \to \infty} \infty$ and $n_I \xrightarrow{t \to \infty} \infty$,

**H2:** there exists a constant $C_1$ such that $N_I^{-1} \sum_{u_i \in U_I} E|\hat{Y}_i|^4 < C_1$,

**H3:** There exists a constant $C_2$ such that $N_I^{-1} \sum_{u_i \in U_I} E(\hat{V}_i^2) < C_2$.

we have

$$E \left| \frac{n_I}{N_I^2} \left\{ v(\hat{Y}) - V(\hat{Y}) \right\} \right|^2 \xrightarrow{t \to \infty} 0.$$ 

A variance estimator for further stages inside the PSUs is needed.
A coupling procedure between SI/SIR sampling of PSUs
Motivation

We would like to prove that, when the first-stage sampling fraction $f_I$ is small:

- the simplified variance estimator $v_{WR}(\hat{Y}) = \frac{N_I^2}{n_I} \frac{s_Z^2}{I}$ is also consistent in case of SI sampling of PSUs,
- the BWR of PSUs is suitable for SI sampling of PSUs.

We propose a coupling method (Hajek, 1960; Thorisson, 1980) to select jointly a with/without replacement sample of PSUs, in such a way that:

- $\bar{X}_n \sim \bar{Z}_n$ and $s^2_X \sim s^2_Z$,

\[
\frac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s^*_X} \sim \frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s^*_Z}.
\]
The coupling procedure

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

![Diagram showing the coupling procedure](image)
The coupling procedure

Step 2: each time \( u_i \in S_{WR}^I \), select a second-stage sample \( S_{i[j]}^I \).
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 $u_i \in S_I^C$. 
The coupling procedure

Suppose that the samples $S^{WR}_I$ and $S_I$ are selected according to the coupling procedure. Then

$$\frac{E(\hat{Y}_{WR} - \hat{Y})^2}{V(\hat{Y}_{WR})} \leq \frac{n_I - 1}{N_I - 1} \left( \leq \frac{n_I}{N_I} \right).$$

(1)

Suppose that (H1)-(H2) hold, and that $f_I \xrightarrow{t \to \infty} 0$. Then

$$E(\bar{Z} - \bar{X})^2 = o(n_I^{-1}) \quad \text{and} \quad \frac{V(\bar{Z})}{V(\bar{X})} \xrightarrow{t \to \infty} 1.$$

Also, the simplified variance estimator $v_{WR}(\hat{Y}) = \frac{N_I^2}{n_I} s^2_Z$ is such that:

$$E \left| \frac{n_I}{N_I^2} \left\{ v_{WR}(\hat{Y}) - v_{WR}(\hat{Y}_{WR}) \right\} \right| \xrightarrow{t \to \infty} 0.$$
With-replacement Bootstrap

We consider the same BWR of PSUs. Denote by

\[(Z_1^*, \ldots, Z_m^*)^{\top}\]

the resample obtained by sampling \(m\) times independently in \((Z_1, \ldots, Z_{n_I})\).

Let

\[\bar{Z}_m^* = \frac{1}{m} \sum_{j=1}^{m} Z_j^* \quad \text{and} \quad s^*_Z = \frac{1}{m-1} \sum_{j=1}^{m} (Z_j^* - \bar{Z}_m^*)^2\]
With-replacement Bootstrap

Mallows (1972) metric: let \( 1 \leq q < \infty \) and \( d_q(\alpha, \beta) = \inf \{E\|X - Z\|^q\}^{1/q} \), where the infimum is taken over all couples \((X, Z)\) with marginal distributions \(\alpha\) and \(\beta\).

Suppose that (H1) and (H2) hold, and that \( m \to \infty \). Then:

\[
d_2 [\sqrt{m}(\bar{Z}_m^* - \bar{Z}), \sqrt{m}(\bar{X}_m^* - \bar{X})] \to 0, \quad t \to \infty \tag{2}
\]

\[
d_1 [s_{\bar{Z}}^*, s_{\bar{X}}^*] \to 0, \quad t \to \infty \tag{3}
\]

\[
\frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_{\bar{Z}}^*} \xrightarrow{L} \mathcal{N}(0, 1). \tag{4}
\]

Using the BWR with \( m = n_I - 1 \) enables to match the simplified variance estimator \( \nu_{WR}(\hat{Y}) \) when estimating the total \( Y \).
Variance estimation

Suppose that $y_k$ is a $q$-vector of interest. We are interested in a parameter

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

where $f : \mathbb{R}^q \to \mathbb{R}$ is differentiable with bounded partial derivatives and $f'(\mu_Y) \neq 0$. The plug-in estimator of $\theta$ is:

- $\hat{\theta} = f(\bar{Z})$ under SI sampling of PSUs,
- $\hat{\theta}_{WR} = f(\bar{X})$ under SIR sampling of PSUs.

Suppose that $S_{WR}^I$ and $S_I^I$ are selected according to the coupling procedure + assumptions (H1)-(H2) hold + $f_I \to 0$. Then:

$$E(\|\bar{Z} - \bar{X}\|^2) = o(n_I^{-1}),$$
$$E(\hat{\theta} - \hat{\theta}_{WR})^2 = o(n_I^{-1}).$$

with $\| \cdot \|$ the Euclidean norm.
Variance estimation

Suppose that the samples $S_{I}^{WR}$ and $S_{I}$ are selected according to the coupling procedure. Suppose that assumptions (H1)-(H2) hold, $f_{I} \to 0$ and $m \to \infty$. Then:

$$E(\|\bar{Z}^{*} - \bar{X}^{*}\|^{2}) = o(m^{-1}) + o(n_{I}^{-1}),$$

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

This implies that

$$\frac{V(\hat{\theta}^{*}|Z_{i})}{V(\hat{\theta}_{WR}^{*}|X_{i})} \toPr 1.$$  

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

G. Chauvet (ENSAI)
A simulation study
Simulation study

We generated 2 finite populations, each with $N_I = 2,000$ PSUs, so that the CV for the sizes $N_i$ of PSUs was equal to 0 and 0.03. In each population, we generated for any PSU $u_i$:

$$\lambda_i = \lambda + \sigma v_i$$

where the $v_i$'s were generated according to a standardized normal distribution. For each SSU $k \in u_i$, we generated a couple of values according to the model

$$y_{1k} = \lambda_i + \left\{\rho^{-1}(1 - \rho)\right\}^{0.5} \sigma \left(\alpha \epsilon_k + \eta_k\right),$$

$$y_{2k} = \lambda_i + \left\{\rho^{-1}(1 - \rho)\right\}^{0.5} \sigma \left(\alpha \epsilon_k + \nu_k\right),$$

so as to have

- a coefficient of correlation approximately equal to 0.60,
- an intra-cluster correlation coefficient equal to 0.1 (similar results for 0.2 and 0.3).
Simulation study

From each population, we selected $B = 1,000$ two-stage samples by:
- Simple sampling of size $n_I = 20, 40, 100$ or $200$ at the first stage,
- Systematic sampling of size $n_0 = 5$ or $10$ at the second stage.

We want to estimate the variance of the substitution estimator for the parameters

$$R = \frac{\mu_{Y_1}}{\mu_{Y_2}}$$

$$r = \frac{\sum_{k \in U} (y_{1k} - \mu_{Y_1})(y_{2k} - \mu_{Y_2})}{\sqrt{\sum_{k \in U} (y_{1k} - \mu_{Y_1})^2} \sqrt{\sum_{k \in U} (y_{2k} - \mu_{Y_2})^2}},$$

by using the BWR of PSUs. The true variance was approximated from a separate simulation run of $C = 20,000$ samples.
## Estimation of the ratio

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop. 1</td>
<td>$n_0 = 5$</td>
<td>$n_I = 20$</td>
<td>0.02</td>
<td>0.34</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>$n_I = 40$</td>
<td>0.02</td>
<td>0.24</td>
<td>2.8</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>$n_I = 100$</td>
<td>0.01</td>
<td>0.15</td>
<td>2.8</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>$n_I = 200$</td>
<td>0.01</td>
<td>0.11</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>$n_0 = 10$</td>
<td>$n_I = 20$</td>
<td>0.00</td>
<td>0.33</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>$n_I = 40$</td>
<td>0.03</td>
<td>0.24</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>$n_I = 100$</td>
<td>0.00</td>
<td>0.16</td>
<td>3.3</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>$n_I = 200$</td>
<td>0.04</td>
<td>0.12</td>
<td>2.3</td>
<td>2.7</td>
</tr>
<tr>
<td>Pop. 2</td>
<td>$n_0 = 5$</td>
<td>$n_I = 20$</td>
<td>0.00</td>
<td>0.34</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>$n_I = 40$</td>
<td>0.00</td>
<td>0.22</td>
<td>2.1</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>$n_I = 100$</td>
<td>0.00</td>
<td>0.15</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>$n_I = 200$</td>
<td>0.02</td>
<td>0.11</td>
<td>3.4</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>$n_0 = 10$</td>
<td>$n_I = 20$</td>
<td>-0.01</td>
<td>0.33</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>$n_I = 40$</td>
<td>0.00</td>
<td>0.24</td>
<td>3.2</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>$n_I = 100$</td>
<td>0.02</td>
<td>0.16</td>
<td>3.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>$n_I = 200$</td>
<td>0.02</td>
<td>0.11</td>
<td>2.6</td>
<td>2.6</td>
</tr>
</tbody>
</table>
Estimation of the coefficient of correlation

<table>
<thead>
<tr>
<th></th>
<th>RB</th>
<th>RS</th>
<th>L</th>
<th>U</th>
<th>L+U</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pop. 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_0 = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_I = 20$</td>
<td>0.01</td>
<td>0.41</td>
<td>3.8</td>
<td>3.2</td>
<td>7.0</td>
</tr>
<tr>
<td>$n_I = 40$</td>
<td>0.00</td>
<td>0.29</td>
<td>2.9</td>
<td>2.8</td>
<td>5.7</td>
</tr>
<tr>
<td>$n_I = 100$</td>
<td>0.02</td>
<td>0.19</td>
<td>3.2</td>
<td>2.6</td>
<td>5.8</td>
</tr>
<tr>
<td>$n_I = 200$</td>
<td>0.01</td>
<td>0.14</td>
<td>2.8</td>
<td>2.1</td>
<td>4.9</td>
</tr>
<tr>
<td>$n_0 = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_I = 20$</td>
<td>-0.01</td>
<td>0.37</td>
<td>3.3</td>
<td>3.2</td>
<td>6.5</td>
</tr>
<tr>
<td>$n_I = 40$</td>
<td>0.01</td>
<td>0.27</td>
<td>2.5</td>
<td>3.0</td>
<td>5.5</td>
</tr>
<tr>
<td>$n_I = 100$</td>
<td>0.05</td>
<td>0.19</td>
<td>2.0</td>
<td>2.6</td>
<td>4.6</td>
</tr>
<tr>
<td>$n_I = 200$</td>
<td>0.03</td>
<td>0.13</td>
<td>2.2</td>
<td>2.4</td>
<td>4.6</td>
</tr>
<tr>
<td><strong>Pop. 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_0 = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_I = 20$</td>
<td>-0.01</td>
<td>0.41</td>
<td>4.2</td>
<td>3.2</td>
<td>7.4</td>
</tr>
<tr>
<td>$n_I = 40$</td>
<td>0.02</td>
<td>0.31</td>
<td>2.6</td>
<td>2.9</td>
<td>5.5</td>
</tr>
<tr>
<td>$n_I = 100$</td>
<td>0.02</td>
<td>0.19</td>
<td>3.0</td>
<td>2.9</td>
<td>5.9</td>
</tr>
<tr>
<td>$n_I = 200$</td>
<td>0.01</td>
<td>0.14</td>
<td>2.2</td>
<td>2.7</td>
<td>4.9</td>
</tr>
<tr>
<td>$n_0 = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_I = 20$</td>
<td>0.01</td>
<td>0.40</td>
<td>2.9</td>
<td>3.7</td>
<td>6.6</td>
</tr>
<tr>
<td>$n_I = 40$</td>
<td>0.00</td>
<td>0.28</td>
<td>4.1</td>
<td>2.8</td>
<td>6.9</td>
</tr>
<tr>
<td>$n_I = 100$</td>
<td>0.02</td>
<td>0.17</td>
<td>2.9</td>
<td>2.4</td>
<td>5.3</td>
</tr>
<tr>
<td>$n_I = 200$</td>
<td>0.04</td>
<td>0.13</td>
<td>2.5</td>
<td>3.4</td>
<td>5.9</td>
</tr>
</tbody>
</table>
References