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.

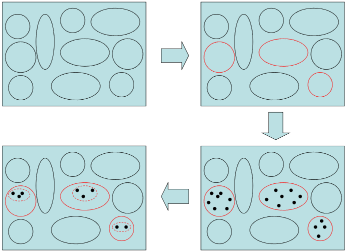
Multistage sampling consists in three stages of sampling, or more. In case of household surveys, a customary sampling design consists in

- 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 purpose:

- **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).

ENSAI École nationale de la statistique et de l'analyse de l'information

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.

École nationale de la statistique et de l'analyse de l'information

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, ...).

ENSAI École nationale de la statistique et de l'analyse de l'information

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 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_I^{WR} is selected by means of simple random sampling with replacement (SIR). The Hansen-Hurwitz estimator is

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

where

- S_I^{WR} 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

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

$$v_{WR}\left(\hat{Y}_{WR}\right) = \frac{N_I^2}{n_I} s_X^2 \text{ with } s_X^2 = \frac{1}{n_I - 1} \sum_{j=1}^{n_I} (X_j - \bar{X}_n)^2.$$

With replacement 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: $N_I \underset{t\to\infty}{\longrightarrow} \infty$ and $n_I \underset{t\to\infty}{\longrightarrow} \infty$,

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

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

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Bootstrap for SIR of PSUs

We consider the with-replacement Bootstrap (BWR) of PSUs described in Rao and Wu (1988). The resample $(X_1^*, \ldots, X_m^*)^\top$ 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 $m \underset{t \to \infty}{\longrightarrow} \infty$. Then (Bickel and Freedman, 1981) :

$$\frac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s_X^*} \underset{\mathcal{L}}{\longrightarrow} \mathcal{N}(0,1).$$

Using the BWR with $m = n_I - 1$ enables to match the unbiased variance estimator $v_{WR}\left(\hat{Y}_{WR}\right)$ when estimating the total Y.

ENSAI École nationale de la statistique et de l'analyse de l'information

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 (SI). 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

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

$$v(\hat{Y}) = \frac{N_I^2}{n_I} \left\{ (1 - f_I) s^2_Z + \frac{1}{N_I} \sum_{u_i \in S_I} \hat{V}_i \right\} \text{ with } f_I = n_I / N_I.$$

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

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

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

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:

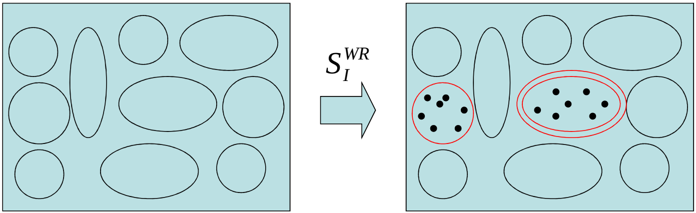
- the simplified variance estimator $v_{WR}(\hat{Y}) = \frac{N_I^2}{n_I} s_Z^2$ is also consistent in case of SI sampling of PSUs,
- the BWR of PSUs is suitable for SI sampling of PSUs.

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

- $\bar{X}_n \simeq \bar{Z}_n$ and $s_X^2 \simeq s_Z^2$,
- $\frac{\sqrt{m}(\bar{X}_m^* - \bar{X})}{s_X^*} \simeq \frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*}$.

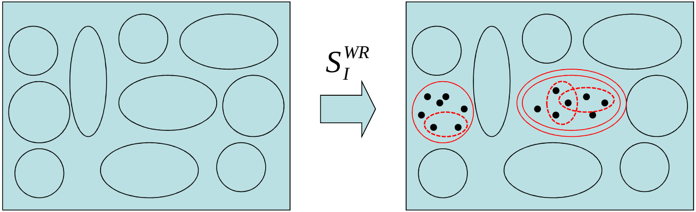
The coupling procedure

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



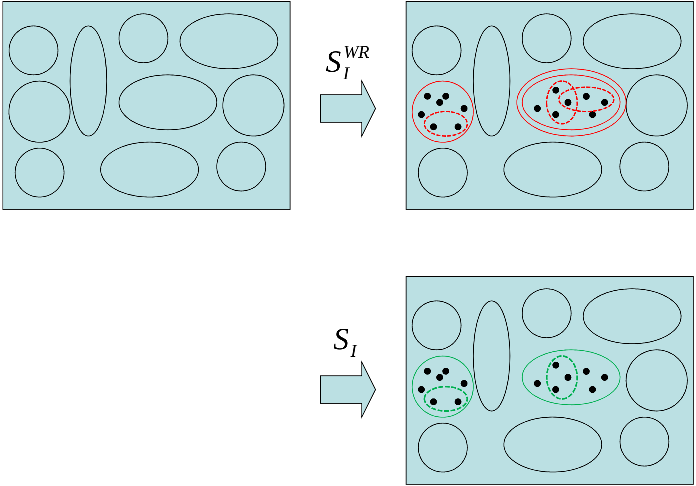
The coupling procedure

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



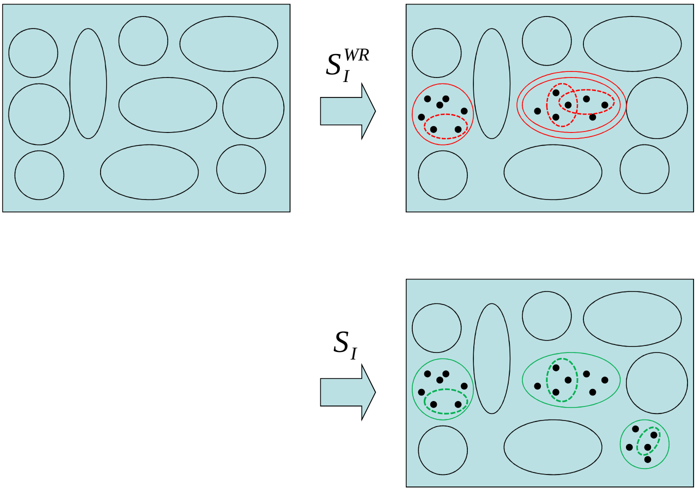
The coupling procedure

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 $u_i \in S_I^c$.



The coupling procedure

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

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

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

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

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

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

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

$$\bar{Z}_m^* = \frac{1}{m} \sum_{j=1}^{m} Z_j^* \quad \text{and} \quad s_Z^{*2} = \frac{1}{m-1} \sum_{j=1}^{m} \left(Z_j^* - \bar{Z}_m^*\right)^2$$

ENSAI École nationale de la statistique et de l'analyse de l'information

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 α and β.

Suppose that (H1) and (H2) hold, and that $m \underset{t\to\infty}{\longrightarrow} \infty$. Then :

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

$$d_1\left[s_Z^{*2}, s_X^{*2}\right] \underset{t\to\infty}{\longrightarrow} 0, \quad (3)$$

$$\frac{\sqrt{m}(\bar{Z}_m^* - \bar{Z})}{s_Z^*} \underset{\mathcal{L}}{\longrightarrow} \mathcal{N}(0,1). \quad (4)$$

Using the BWR with $m = n_I - 1$ enables to match the simplified variance estimator $v_{WR}\left(\hat{Y}\right)$ when estimating the total Y.

ENSAI École nationale de la statistique et de l'analyse de l'information

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 $f : \mathbb{R}^q \longrightarrow \mathbb{R}$ is differentiable with bounded partial derivatives and $f'(\mu_Y) \neq 0$. The plug-in estimator of θ 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_I^{WR} and S_I are selected according to the coupling procedure + assumptions (H1)-(H2) hold + $f_I \underset{t\to\infty}{\longrightarrow} 0$. 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, $f_I \underset{t\to\infty}{\longrightarrow} 0$ and $m \underset{t\to\infty}{\longrightarrow} \infty$. Then :

$$E(\|\bar{Z}^* - \bar{X}^*\|^2) = o(m^{-1}) + o(n_I^{-1}), \quad (5)$$

$$E(\hat{\theta}^* - \hat{\theta}_{WR}^*)^2 = o(m^{-1}) + o(n_I^{-1}). \quad (6)$$

This implies that

$$\frac{V(\hat{\theta}^*|Z_i)}{V(\hat{\theta}_{WR}^*|X_i)} \underset{Pr}{\longrightarrow} 1. \quad (7)$$

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

ENSAI École nationale de la statistique et de l'analyse de l'information

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:

$$\lambda_i = \lambda + \sigma\, v_i \qquad (8)$$

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

$$y_{1k} = \lambda_i + \{\rho^{-1}(1-\rho)\}^{0.5}\sigma\,(\alpha\,\epsilon_k + \eta_k), \qquad (9)$$
$$y_{2k} = \lambda_i + \{\rho^{-1}(1-\rho)\}^{0.5}\sigma\,(\alpha\,\epsilon_k + \nu_k), \qquad (10)$$

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:

- SI 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

$$
\begin{aligned}
R &= \frac{\mu_{y1}}{\mu_{y2}} \\
r &= \frac{\sum_{k\in U}(y_{1k}-\mu_{y1})(y_{2k}-\mu_{y2})}{\sqrt{\sum_{k\in U}(y_{1k}-\mu_{y1})^2}\sqrt{\sum_{k\in U}(y_{2k}-\mu_{y2})^2}},
\end{aligned}
$$

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

ENSAI École nationale de la statistique et de l'analyse de l'information

Estimation of the ratio

			RB	RS	L	U	L+U
Pop. 1	$n_0 = 5$	$n_I = 20$	0.02	0.34	3.6	2.9	6.5
		$n_I = 40$	0.02	0.24	2.8	3.3	6.1
		$n_I = 100$	0.01	0.15	2.8	2.2	5.0
		$n_I = 200$	0.01	0.11	3.0	3.0	6.0
	$n_0 = 10$	$n_I = 20$	0.00	0.33	3.9	3.1	7.0
		$n_I = 40$	0.03	0.24	3.2	2.8	6.0
		$n_I = 100$	0.00	0.16	3.3	2.4	5.7
		$n_I = 200$	0.04	0.12	2.3	2.7	5.0
Pop. 2	$n_0 = 5$	$n_I = 20$	0.00	0.34	3.8	3.6	7.4
		$n_I = 40$	0.00	0.22	2.1	3.0	5.1
		$n_I = 100$	0.00	0.15	2.5	2.5	5.0
		$n_I = 200$	0.02	0.11	3.4	2.9	6.3
	$n_0 = 10$	$n_I = 20$	-0.01	0.33	3.7	2.6	6.3
		$n_I = 40$	0.00	0.24	3.2	3.5	6.7
		$n_I = 100$	0.02	0.16	3.3	2.2	5.5
		$n_I = 200$	0.02	0.11	2.6	2.6	5.2

Estimation of the coefficient of correlation

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

References

- Ardilly, P. (2006). Les techniques de sondage. Editions Technip.
- Bickel, P.J., and Freedman, D.A. (1981). Some asymptotic theory for the Bootstrap. The Annals of Statistics, 9, 1196-1217.
- Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. Publications of the Mathematics Institute of the Hungarian Academy of Science, 5, 361-74.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. JASA, 77, 89-96.
- Lucas, J-P. (2013). Contamination des logements par le plomb : prévalence des logements à risque et identification des déterminants de la contamination. PhD dissertation, Université de Nantes.
- Mallows, C. (1972). A note on asymptotic joint normality. The Annals of Mathematical Statistics, 43, 508-515.
- Rao, J.N.K., Wu, C.F.J. (1988). Resampling inference with complex survey data. JASA, 83(401), 231-241.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New-York, Springer-Verlag.
- Thorisson, H. (2000). Coupling, stationarity, and regeneration. New York, Springer.

ENSAI École nationale de la statistique et de l'analyse de l'information