Bootstrap for multistage sampling and unequal probability sampling of primary sampling units

Guillaume CHAUVET(*)

(*)ENSAI (CREST)

Résumé

Le tirage à plusieurs degrés est fréquemment utilisé quand on ne dispose pas d'une base de sondage, ou que la population est dispersée géographiquement. Le tirage à plusieurs degrés introduit une dépendance complexe dans la sélection des unités finales, ce qui rend les propriétés des estimateurs plus difficiles à prouver. Nous considérons ici un échantillonnage à plusieurs degrés avec tirage à probabilités inégales pour les unités primaires, et un plan de sondage arbitraire pour les tirages réalisés aux degrés suivants. Nous proposons une méthode de couplage entre le tirage réjectif et le tirage multinomial, afin de relier le premier plan de sondage à un autre plan de sondage où les unités primaires sont sélectionnées avec remise. Quand la fraction de sondage du 1er degré est faible, cette méthode permet de prouver que les estimateurs de variance par Bootstrap sont consistants pour des fonctions lisses de totaux.

Abstract

Multistage sampling is commonly used for household surveys when there exists no sampling frame, or when the population is scattered over a wide area. Multistage sampling usually introduces a complex dependence in the selection of the final units, which makes asymptotic results quite difficult to prove. In this work, we consider multistage sampling with unequal probability sampling at the first stage, and with an arbitrary sampling design for further stages. We introduce a new coupling method between rejective sampling and multinomial sampling. When the first-stage sampling fraction is small, this method is used to prove the consistency of with-replacement Bootstrap variance estimators for smooth functions of totals.

Keywords

Bootstrap; Coupling algorithm; Rejective sampling; With-replacement sampling; Without-replacement sampling.

1 Introduction

Multistage sampling is widely used for household and health surveys when there exists no sampling frame, or when the population is scattered over a wide area. Three or more stages of sampling may be commonly used. For example, the third National Health and Nutrition Survey (NHANES III) conducted in the United States involved four stages of sampling, with the selection of counties as Primary Sampling Units (PSUs), of segments as Secondary Sampling Units (SSUs) inside the selected counties, of households as Tertiary Sampling Units (TSUs) inside the selected segments, and of individuals inside the selected households (see Ezzati et al., 1992). A detailed treatment of multistage sampling may be found in textbooks like Ardilly (2006), Cochran (1977), Särndal et al. (1992) or Fuller (2009).

The use of bootstrap techniques in survey sampling has been widely studied in the literature. Most of them may be thought as particular cases of the weighted bootstrap (Bertail and Combris, 1997; Antal and Tillé, 2011; Beaumont and Patak, 2012); see also Shao and Tu (1995, chap. 6), Davison and Hinkley (1997, section 3.7), Lahiri (2003) and Davison and Sardy (2007) for detailed reviews. Bootstrap for multistage sampling under without-replacement sampling of PSUs has been considered for example in Rao and Wu (1988), Rao, Wu and Yue (1992), Nigam and Rao (1996), Funaoka et al. (2006), Preston (2009) and Lin et al. (2013), among others. Testing the validity of a bootstrap procedure has primarily consisted in showing that it led to the correct variance estimator in the linear case, and then in evaluating empirically the behavior of the method for complex parameters through simulations.

In this paper, we consider the so-called with-replacement Bootstrap of PSUs (see Rao and Wu, 1988). Extending the work in Chauvet (2014), We prove that this Bootstrap method is suitable for multistage sampling with rejective sampling of PSUs (see Hajek, 1964) and a small first-stage sampling fraction, and yields consistent variance estimators for smooth functions of means. Our framework is defined in Section 2. Multistage sampling with multinomial sampling of PSUs is presented in Section 3, and the principles of the with-replacement Bootstrap are briefly reminded in Section 4. Multistage sampling with rejective sampling of PSUs is presented in Section 5, and we describe in Section 6 a coupling procedure for a joint selection of a multinomial sample and of a rejective sample. This procedure is used to prove that the with-replacement Bootstrap leads to consistent variance estimators for smooth functions of means, by comparison with the multinomial case, which is the purpose of Section 7. The properties of the Bootstrap variance estimators for three parameters are evaluated in Section 8 through a small simulation study.

2 Framework

We consider a finite population U consisting of N sampling units that may be represented by their labels, so that we may simply write $U = \{1, \ldots, N\}$. 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

$$Y = \sum_{k \in U} y_k = \sum_{u_i \in U_I} Y_i \tag{1}$$

for some q-vector of interest y, where $Y_i = \sum_{k \in u_i} y_k$ is the sub-total of y on the PSU u_i. We are also interested in estimating some smooth function of the population total

$$\theta \quad = \quad f(Y) \tag{2}$$

where $f : \mathbb{R}^q \longrightarrow \mathbb{R}$ is a known function.

We note $E(\cdot)$ and $V(\cdot)$ for the expectation and the variance of some estimator. Also, we note $E_{\{X\}}(\cdot)$ and $V_{\{X\}}(\cdot)$ for the expectation and variance conditionally on some random variable X. Throughout the paper, we denote by $\hat{Y}_i$ an unbiased estimator of Y_i, and by $V_i = V(\hat{Y}_i)$ its variance-covariance matrix. Also, we denote by $\hat{V}_i$ an unbiased estimator of V_i.

In order to study the asymptotic properties of the sampling designs and estimators that we treat below, we consider the asymptotic framework of Isaki and Fuller (1982). We assume that the population U belongs to a nested sequence $\{U_t\}$ of finite populations with increasing sizes N_t, and that the population vector of values $y_{Ut} = (y_{1t}, \ldots, y_{Nt})^\top$ belongs to a sequence $\{y_{Ut}\}$ of N_t-vectors. For simplicity, the index t will be suppressed in what follows but all limiting processes will be taken as $t \to \infty$. We assume that:

H1: $n_I \underset{t \to \infty}{\longrightarrow} \infty$ and there exists some constant $\bar{N} \geq 1$ such that $\dfrac{N}{N_I} \underset{t \to \infty}{\longrightarrow} \bar{N}$.

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 $p_I(\cdot)$. For clarity of exposition, we consider non-stratified sampling designs for $p_I(\cdot)$, but the results may be easily extended to the case of stratified first-stage sampling designs with a finite number of strata. If the PSU u_i is selected in S_I, a second-stage sample S_i is selected in u_i by means of some sampling design $p_i(\cdot|S_I)$. We assume invariance of the second-stage designs: that is, the second stage of sampling is independent of S_I and we may simply write $p_i(\cdot|S_I) = p_i(\cdot)$. Also, we assume that the second-stage designs are independent from one PSU to another, conditionally on S_I. This implies that

$$\begin{aligned} Pr\left(\bigcup_{u_i \in S_I} \{S_i = s_i\} \middle| S_I \right) \quad &= \quad \prod_{u_i \in S_I} p_i(s_i|S_I) \\ &= \quad \prod_{u_i \in S_I} p_i(s_i) \end{aligned} \tag{3}$$

for any set of samples $s_i \subset u_i,\ i = 1, \ldots, N_I$, where the second line in (3) follows from the invariance assumption; see Särndal et al (1992, chapter 4) for further details. The second-stage sampling designs $p_i(\cdot)$ are left arbitrary. For example, they may involve censuses inside some PSUs (which means cluster sampling), or additional stages of sampling.

3 Multinomial sampling of PSUs

We first consider the case when a first-stage sample S_I^{WR} is selected in U_I by means of multinomial sampling (Tillé, 2006). That is, a sample S_I^{WR} is obtained from n_I independent draws, some unit u_i being selected with probability α_{Ii} at each draw. This will be noted as

$$S_I^{WR} \sim MULT(U_I; n_I; \alpha_I) \quad \text{with} \quad \alpha_I = (\alpha_{I1}, \ldots, \alpha_{IN_I})^\top \tag{4}$$

and $\sum_{u_i \in U_I} \alpha_{Ii} = 1$. We denote by W_i the number of selections of the PSU u_i in S_I^{WR}. The expected number of draws for the PSU u_i is

$$E(W_i) \quad = \quad n_I \alpha_{Ii}.$$

Each time $j = 1, \ldots, W_i$ that the PSU u_i is selected in S_I^{WR}, a second-stage sample $S_i^{[j]}$ is selected in u_i.

The population total Y is unbiasedly estimated by the Hansen-Hurwitz (1942) (HH)-estimator

$$\hat{Y}_{WR} \quad = \quad \frac{1}{n_I} \sum_{j=1}^{n_I} \frac{\hat{Y}_{i(j)}^{[j]}}{\alpha_{Ii(j)}} \tag{5}$$

where we denote by $i(j)$ the PSU selected at the j-th draw, and where $\hat{Y}_{i(j)}^{[j]}$ stands for an unbiased estimator of $Y_{i(j)}$ computed on $S_{i(j)}^{[j]}$. The HH-estimator may be rewritten as

$$\hat{Y}_{WR} = \bar{X} \equiv \frac{1}{n_I} \sum_{j=1}^{n_I} X_j \quad \text{where} \quad X_j = \frac{\hat{Y}_{i(j)}^{[j]}}{\alpha_{Ii(j)}}. \tag{6}$$

Note that the $X_j,\ j = 1, \ldots, n_I$ are independent and identically distributed random q-vectors, with $E(X_j) = Y$.

The variance of the HH estimator is

$$V\left(\hat{Y}_{WR}\right) \quad = \quad V_{PSU}\left(\hat{Y}_{WR}\right) + V_{SSU}\left(\hat{Y}_{WR}\right), \tag{7}$$

where

$$V_{PSU}\left(\hat{Y}_{WR}\right) \quad = \quad \frac{1}{n_I} \sum_{u_i \in U_I} \alpha_{Ii} \left(\frac{Y_i}{\alpha_{Ii}} - Y\right) \left(\frac{Y_i}{\alpha_{Ii}} - Y\right)^\top \tag{8}$$

is the variance due to the first stage of sampling, and

$$V_{SSU}\left(\hat{Y}_{WR}\right) \quad = \quad \frac{1}{n_I} \sum_{u_i \in U_I} \frac{V_i}{\alpha_{Ii}} \tag{9}$$

is the variance due to further stages. The variance due to the first-stage of sampling may be rewritten as

$$V_{PSU}\left(\hat{Y}_{WR}\right) \quad = \quad \frac{1}{2n_I} \sum_{u_i \neq u_j \in U_I} \alpha_{Ii}\alpha_{Ij} \left(\frac{Y_i}{\alpha_{Ii}} - \frac{Y_j}{\alpha_{Ij}}\right) \left(\frac{Y_i}{\alpha_{Ii}} - \frac{Y_j}{\alpha_{Ij}}\right)^\top. \tag{10}$$

An unbiased estimator for $V\left(\hat{Y}_{WR}\right)$ is

$$v_{WR}\left(\hat{Y}_{WR}\right) = \frac{s_X^2}{n_I} \quad \text{with} \quad s_X^2 = \frac{1}{n_I - 1} \sum_{j=1}^{n_I} (X_j - \bar{X})(X_j - \bar{X})^\top. \tag{11}$$

The simple form of the variance estimator in (11) is primarily due to (6), where $\hat{Y}_{WR}$ is written as a mean of independent and identically distributed random variables (see also Särndal et al, 1992, p. 151). In particular, unbiased variance estimators $\hat{V}_i$ inside PSUs are not needed to compute this variance estimator. This appealing property led to consider $v_{WR}(\cdot)$ as a possible simplified variance estimator when the PSUs are selected without replacement with a small sampling fraction at the first stage. In case of without-replacement sampling of PSUs, it can be shown that $v_{WR}(\cdot)$ succeeds in accounting for the variance due to further stages of sampling, but is usually biased for the variance due to the first-stage of sampling.

The parameter $\theta = f(Y)$ is approximately unbiasedly estimated by the plug-in estimator

$$\hat{\theta}_{WR} \quad = \quad f(\bar{X}). \tag{12}$$

An approximately unbiased variance estimator for $\hat{\theta}_{WR}$ may be obtained from (11) through the linearization technique, by substituting the variable y with the estimated linearized variable of the parameter θ (Deville, 1999; Goga, Deville and Ruiz-Gazen, 2009). In this paper, we rather resort to Bootstrap for variance estimation.

4 With-replacement bootstrap for multinomial sampling of PSUs

We consider the with-replacement Bootstrap of PSUs described for example in Rao and Wu (1988). Using the notation introduced in equation (6), we note

$$(X_1, \ldots, X_{n_I})^\top \tag{13}$$

the original sample of estimators under multinomial sampling of PSUs, whose mean $\bar{X}$ corresponds to the HH-estimator. We define

$$(X_1^*, \ldots, X_m^*)^\top \tag{14}$$

as a resample of estimators, obtained by sampling m times independently and with equal probabilities in the original set of estimators $(X_1, \ldots, X_{n_I})^\top$. The resample mean is denoted as

$$\bar{X}^* \quad = \quad \frac{1}{m} \sum_{i=1}^{m} X_i^*. \tag{15}$$

This is the Bootstrap estimator for the population total. The Bootstrap estimator for the parameter $\theta = f(Y)$ is

$$\hat{\theta}_{WR}^* \quad = \quad f(\bar{X}^*). \tag{16}$$

The resampling in the set $(X_1, \ldots, X_{n_I})^\top$ is repeated B times independently. The bootstrap variance estimator for Y is simply obtained by computing the dispersion of the Bootstrap estimators for the B resamples, which leads to

$$v_{boot}\left(\hat{Y}_{WR}\right) \quad = \quad \frac{1}{B-1} \sum_{b=1}^{B} \left(\bar{X}^{b*} - \frac{1}{B} \sum_{c=1}^{B} \bar{X}^{c*} \right) \left(\bar{X}^{b*} - \frac{1}{B} \sum_{c=1}^{B} \bar{X}^{c*} \right)^\top \tag{17}$$

with $\bar{X}^{b*}$ the mean computed on the b-th resample. Using $m = n_I - 1$ enables to match the usual unbiased variance estimator in (11) when $B \to \infty$. The bootstrap variance estimator for $\theta = f(Y)$ is also obtained by computing the dispersion of the Bootstrap estimators, which leads to

$$v_{boot}\left(\hat{\theta}_{WR}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{WR}^{b*} - \frac{1}{B}\sum_{c=1}^{B}\hat{\theta}_{WR}^{c*}\right)^2 \tag{18}$$

with $\hat{\theta}_{WR}^{b*}$ the estimator computed on the b-th resample.

5 Rejective sampling of PSUs

We now consider the case when a first-stage sample S_I^R is selected in U_I by means of rejective sampling (Hajek, 1964) with an inclusion probability π_{Ii} for the PSU u_i. This will be noted as

$$S_I^R \sim REJ(U_I; n_I; \pi_I) \quad \text{with} \quad \pi_I = (\pi_{I1}, \ldots, \pi_{IN_I})^\top \tag{19}$$

and $\sum_{u_i \in U_I} \pi_{Ii} = n_I$. If the PSU u_i is selected in S_I^R, a second-stage sample S_i is selected in u_i.

The population total Y is unbiasedly estimated by the Narain-Horvitz-Thompson (NHT)-estimator

$$\hat{Y}_R = \sum_{u_i \in S_I^R} \frac{\hat{Y}_i}{\pi_{Ii}}, \tag{20}$$

where $\hat{Y}_i$ stands for an unbiased estimator of Y_i computed on S_i.

The sample S_I^R can be obtained through a draw by draw procedure (see Chen, Dempster and Liu, 1994), which enables to rewrite the NHT-estimator as

$$\hat{Y}_R = \bar{Z} \equiv \frac{1}{n_I}\sum_{j=1}^{n_I} Z_j \quad \text{where} \quad Z_j = \frac{\hat{Y}_{i(j)}}{n_I^{-1}\pi_{Ii(j)}}, \tag{21}$$

with $i(j)$ the PSU selected at the j-th draw. Note that the draws are not performed independently and that the Z_j, $j = 1, \ldots, n_I$ are thus not i.i.d., unlike multinomial sampling. The parameter $\theta = f(Y)$ is approximately unbiasedly estimated by the plug-in estimator

$$\hat{\theta}_R = f(\bar{Z}). \tag{22}$$

The variance of the NHT- estimator is

$$V\left(\hat{Y}_R\right) = V_{PSU}\left(\hat{Y}_R\right) + V_{SSU}\left(\hat{Y}_R\right). \tag{23}$$

The variance due to the first stage of sampling is

$$V_{PSU}\left(\hat{Y}_R\right) = \sum_{u_i, u_j \in U_I} \Delta_{Iij}\left(\frac{Y_i}{\pi_{Ii}}\right)\left(\frac{Y_j}{\pi_{Ij}}\right)^\top \tag{24}$$

where $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$, and where π_{Iij} is the probability that the PSUs u_i and u_j are selected jointly in S_I^R. The variance due to further stages of sampling is

$$V_{SSU}\left(\hat{Y}_R\right) = \sum_{u_i \in U_I} \frac{V_i}{\pi_{Ii}}, \tag{25}$$

and is identical to the variance due to further stages of sampling obtained for multinomial sampling with drawing probabilities $\alpha_{Ii} = \pi_{Ii}/n_I$.

An unbiased estimator for $V\left(\hat{Y}_R\right)$ is

$$\hat{V}\left(\hat{Y}_R\right) = \hat{V}_1\left(\hat{Y}_R\right) + \hat{V}_2\left(\hat{Y}_R\right), \tag{26}$$

where

$$\hat{V}_1\left(\hat{Y}_R\right) = \sum_{u_i, u_j \in S_I^R} \frac{\Delta_{Iij}}{\pi_{Iij}} \left(\frac{\hat{Y}_i}{\pi_{Ii}}\right) \left(\frac{\hat{Y}_j}{\pi_{Ij}}\right)^\top, \tag{27}$$

$$\hat{V}_2\left(\hat{Y}_R\right) = \sum_{u_i \in S_I^R} \frac{\hat{V}_i}{\pi_{Ii}}. \tag{28}$$

It is well-known that $\hat{V}\left(\hat{Y}_R\right)$ is not a term by term unbiased estimator for (23), see for example Särndal, Swensson and Wretman (1992). Also, unbiased variance estimators $\hat{V}_i$ inside PSUs are needed, which may be cumbersome it the sampling strategy inside PSUs is somewhat complex. It is therefore desirable to exhibit simplified variance estimators with limited bias.

Making use of Theorem 6.1 in Hajek (1964), we note that the variance of $\hat{Y}_R$ may be approximately written as

$$V_{PSU}\left(\hat{Y}_R\right) = \frac{1}{2d_I} \sum_{u_i \neq u_j \in U_I} \pi_{Ii}(1-\pi_{Ii})\pi_{Ij}(1-\pi_{Ij}) \left(\frac{Y_i}{\pi_{Ii}} - \frac{Y_j}{\pi_{Ij}}\right) \left(\frac{Y_i}{\pi_{Ii}} - \frac{Y_j}{\pi_{Ij}}\right)^\top [1 + o(1)] \tag{29}$$

with $d_I = \sum_{u_i \in U_I} \pi_{Ii}(1-\pi_{Ii})$. As stated in Proposition 1 below, it follows from equation (10) that the variance of the NHT-estimator under rejective sampling of PSUs is close to that obtained for the HH-estimator under multinomial sampling of PSUs with drawing probabilities $\alpha_{Ii} = \pi_{Ii}/n_I$, if the inclusion probabilities π_{Ii} are small.

Proposition 1. *We take $\alpha_{Ii} = \pi_{Ii}/n_I$. Suppose that $H1$ holds. Suppose that:*

H2: There exists some constants $\lambda_1 \geq \lambda_0 > 0$ such that $\lambda_0 \leq \left\| N_I^{-2} n_I V_{PSU}\left(\hat{Y}_{WR}\right) \right\| \leq \lambda_1$ where $\|\cdot\|$ is the spectral norm.

H3: There exists some constants C_0 and C_1 such that for any $u_i \in U_I$ $\frac{n_I}{N_I} \leq \pi_{Ii} \leq C_1 \frac{n_I}{N_I}$.

If in addition: $\frac{n_I}{N_I} \underset{t\to\infty}{\to} 0$, we have

$$\left\| \left\{ V\left(\hat{Y}_{WR}\right) \right\}^{-1} \left\{ V\left(\hat{Y}_R\right) - V\left(\hat{Y}_{WR}\right) \right\} \right\| \underset{t\to\infty}{\longrightarrow} 0 \tag{30}$$

6 A coupling procedure between multinomial sampling of PSUs and rejective sampling of PSUs

Rejective sampling can be performed as conditional multinomial sampling (Hajek, 1981, chapter 7). A sample S_I^{WR} is first selected by means of multinomial sampling with drawing probabilities α_I. This sample is kept if all the selected units are distinct, otherwise it is discarded and a new sample is selected. This sampling process is repeated until we obtain a sample with distinct units, which is the final sample.

In what follows, we suppose that the drawing probabilities α_I are chosen so that the final inclusion probabilities π_I are matched; see Hajek (1981), Chen, Dempster and Liu (1994) and Deville (2000) for the computation of α_I. A coupling procedure leading to the joint selection of a multinomial sample with drawing probabilities α_I and of a rejective sample with inclusion probabilities π_I is proposed in Algorithm 1.

Algorithm 1 A coupling procedure for multinomial sampling of PSUs and rejective sampling of PSUs

1. Draw the sample $S_I^{WR} \sim MULT(U_I; n_I; \alpha_I)$. Each time $j = 1, \ldots, W_i$ that the PSU u_i is selected in S_I^{WR}, a second-stage sample $S_i^{[j]}$ is selected in u_i.
2. If all the PSUs inside S_I^{WR} are distinct, take $S_I^R = S_I^{WR}$, and for any $u_i \in S_I^R$ take $S_i = S_i^{[1]}$.
3. Otherwise, select a new sample $\sim MULT(U_I; n_I; \alpha_I)$ until all the selected units are distinct. The final sample is S_I^R. For any $u_i \in S_I^R$, select a second-stage sample S_i.

Proposition 2. *Assume that the samples S_I^{WR} and S_I^R are selected according to Algorithm 1. Assume that assumptions (H1)-(H3) hold. Suppose that:*

H4: f is homogeneous of degree α, and is a differentiable function on $\mathbb{R}^q$ with bounded partial derivatives and with $f'(\mu_y) \neq 0$.

If in addition: $\dfrac{n_I}{\sqrt{N_I}} \underset{t\to\infty}{\to} 0$, *then:*

$$E(\|\hat{Y}_{WR} - \hat{Y}_R\|^2) = o(N_I^2 n_I^{-1}), \tag{31}$$

$$E(\hat{\theta}_{WR} - \hat{\theta}_R)^2 = o(N_I^{2\alpha} n_I^{-1}). \tag{32}$$

7 With-replacement bootstrap for rejective sampling of PSUs

We still consider the with-replacement Bootstrap of PSUs described in Section (4). Recall that

$$(Z_1, \ldots, Z_{n_I})^\top \tag{33}$$

is the original sample of estimators under rejective sampling of PSUs, whose mean $\bar{X}$ corresponds to the NHT-estimator. The plug-in estimator of the parameter $\theta = f(Y)$ is

$$\hat{\theta}_R \;=\; f(\bar{Z}). \tag{34}$$

A resample of estimators

$$(Z_1^*, \ldots, Z_m^*)^\top \tag{35}$$

is obtained by sampling m times independently and with equal probabilities in the original set of estimators $(Z_1, \ldots, Z_{n_I})^\top$. The resample mean is denoted as

$$\bar{Z}^* \;=\; \frac{1}{m}\sum_{i=1}^{m} Z_i^*. \tag{36}$$

This is the Bootstrap estimator for the population total. The Bootstrap estimator for the parameter $\theta = f(Y)$ is

$$\hat{\theta}_R^* \;=\; f(\bar{Z}^*). \tag{37}$$

The resampling in the set $(X_1, \ldots, X_{n_I})^\top$ is repeated B times independently. The bootstrap variance estimator for Y is obtained by computing the dispersion of the Bootstrap estimators for the B resamples, which leads to

$$v_{boot}\left(\hat{Y}_R\right) \;=\; \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{Z}^{b*} - \frac{1}{B}\sum_{c=1}^{B}\bar{Z}^{c*}\right)\left(\bar{Z}^{b*} - \frac{1}{B}\sum_{c=1}^{B}\bar{Z}^{c*}\right)^\top \tag{38}$$

with $\bar{Z}^{b*}$ the mean computed on the b-th resample. The bootstrap variance estimator for $\theta = f(Y)$ is also obtained by computing the dispersion of the Bootstrap estimators, which leads to

$$v_{boot}\left(\hat{\theta}_R\right) \;=\; \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_R^{b*} - \frac{1}{B}\sum_{c=1}^{B}\hat{\theta}_R^{c*}\right)^2 \tag{39}$$

with $\hat{\theta}_R^{b*}$ the estimator computed on the b-th resample.

Proposition 3. *Assume that the samples S_I^{WR} and S_I^R are selected according to Algorithm 1. Assume that assumptions (H1)-(H4) hold.*
If in addition: $\dfrac{n_I}{\sqrt{N_I}} \underset{t\to\infty}{\to} 0$ and $m \underset{t\to\infty}{\longrightarrow} \infty$, then:

$$E(\|\bar{Z}^* - \bar{X}^*\|^2) = o(N_I^2 m^{-1}) + o(N_I^2 n_I^{-1}), \tag{40}$$

$$E(\hat{\theta}^* - \hat{\theta}_{WR}^*)^2 = o(N_I^{2\alpha} m^{-1}) + o(N_I^{2\alpha} n_I^{-1}). \tag{41}$$

Proposition 3 implies that

$$\frac{V_{\{Z_1,\ldots,Z_{n_I}\}}(\hat{Y}_R)}{V_{\{X_1,\ldots,X_{n_I}\}}(\hat{Y}_{WR})} \;\longrightarrow_{Pr}\; 1. \tag{42}$$

Therefore, the consistency of the Bootstrap variance estimator for rejective sampling of PSUs follows from that of the Bootstrap variance estimator for multinomial sampling of PSUs.

8 A simulation study

We conducted a limited simulation study to investigate on the performance of the Bootstrap variance estimator. We generated 1 finite population with $N_I = 2,000$ PSUs. The number of SSUs inside PSUs was generated so that the average number of SSUs per PSU was approximately equal to $\bar{N} = 40$, and so that the coefficient of variation for the sizes N_i of PSUs was equal to 0.06. For any PSU u_i, we generated:

$$\lambda_i \quad = \quad \lambda + \sigma \; v_i \tag{43}$$

with $\lambda = 20$ and $\sigma = 2$, and the v_i's were generated according to a normal distribution with mean 0 and variance 1. For each SSU $k \in u_i$, we generated three couples of values $(y_{1,k}, y_{2,k})$, $(y_{3,k}, y_{4,k})$ and $(y_{5,k}, y_{6,k})$ according to the model

$$y_{2h-1,k} \quad = \quad \lambda_i + \{\rho_h^{-1}(1-\rho_h)\}^{0.5}\sigma \; (\alpha \; \epsilon_k + \eta_k), \tag{44}$$

$$y_{2h,k} \quad = \quad \lambda_i + \{\rho_h^{-1}(1-\rho_h)\}^{0.5}\sigma \; (\alpha \; \epsilon_k + \nu_k), \tag{45}$$

for $h = 1, \ldots, 3$, where the values ϵ_k, η_k and ν_k were generated according to a normal distribution with mean 0 and variance 1. The parameter ρ_h was chosen so that the intra-cluster correlation coefficient was approximately equal to 0.1 for both variables y_1 and y_2, 0.2 for both variables y_3 and y_4, and 0.3 for both variables y_5 and y_6. Also, the parameter α was chosen so that the coefficient of correlation between variables y_{2h-1} and y_{2h}, $h = 1, \ldots, 3$, was approximately equal to 0.60.

We selected $B = 1,000$ samples in the population by means of a two-stage self-weighting sampling design. The sample S_I of PSUs was selected by means of rejective sampling of size $n_I = 20, 50, 100, 200$ or 500 with probabilities proportional to the size N_i of the PSUs. Inside each $u_i \in S_I$, the sample S_i of SSUs was selected by means of systematic sampling of size $n_0 = 5$ or 20. Note that, due to the systematic sampling at the second stage, the variance may not be unbiasedly estimated. Our objective is to use the with-replacement Bootstrap variance estimator in (39) for the NHT estimator of the total of the variables y_1, y_3 and y_5. Also, our objective is to use the with-replacement Bootstrap variance estimator in (39) for the substitution estimator of the ratios

$$R_h \quad = \quad \frac{\mu_{y,2h-1}}{\mu_{y,2h}} \tag{46}$$

with $\mu_{y,2h-1} = N^{-1}\sum_{k\in U} y_{2h-1,k}$ and $\mu_{y,2h} = N^{-1}\sum_{k\in U} y_{2h,k}$, and for the substitution estimator of the coefficient of correlations

$$r_h \quad = \quad \frac{\sum_{k\in U}(y_{2h-1,k} - \mu_{y,2h-1})(y_{2h,k} - \mu_{y,2h})}{\sqrt{\sum_{k\in U}(y_{2h-1,k} - \mu_{y,2h-1})^2}\sqrt{\sum_{k\in U}(y_{2h,k} - \mu_{y,2h})^2}}, \tag{47}$$

for $h = 1, \ldots, 3$. We used $B = 1,000$ Bootstrap replications and $m = n_I - 1$ in the simulation study. The true variance was approximated from a separate simulation run of $C = 20,000$ samples.

As a measure of bias of a point estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo percent relative bias (RB) given by

$$RB_{MC}(\hat{\theta}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{\theta}_{(b)} - \theta}{\theta}$$

where $\hat{\theta}_{(b)}$ gives the value of the estimator for the b^{th} sample. As a measure of variance of an estimator $\hat{\theta}$ we used the Monte Carlo percent relative stability (RS) given by

$$RS_{MC}(\hat{\theta}) = 100 \times \frac{\sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \theta)^2}}{\theta}.$$

Using the with-replacement bootstrap of PSUs, we assess the coverage of confidence intervals obtained by means of the percentile method. We used a nominal one-tailed error rate of 2.5 % in each tail.

The results obtained for the Bootstrap variance estimator are given in Tables 1 and 2. When the first-stage sampling fraction is small (less than 5 %), the bias of the Bootstrap variance estimator is small for all parameters (less than 6 %), and the coverage rates are approximately respected. When the first-stage sampling fraction is moderate ($f_I = 10$ %), the Bootstrap variance estimator is weakly biased for the ratio and the coefficient of correlation; as for the total, the bias of the Bootstrap variance estimator is small if the variance due to the second stage is appreciable ($n_0 = 5$), but can be as high as 10 % otherwise. When the first-stage sampling fraction is larger ($f_I = 25$ %), the variance estimators are positively biased in all cases. The bias tends to increase when the intra-cluster correlation coefficient increases, that is, when the variance due to the first-stage of sampling increases. Again, the bias is larger with $n_0 = 20$, that is, when the variance due to the second stage is small as compared to the variance due to the first-stage of sampling. As could be expected, the confidence intervals are too conservative when the biases are larger.

References

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, **106**, 534-543.

Ardilly, P. (2006). Les techniques de sondage. *Paris, éditions Technip.*

Beaumont, J.F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, **80**, 127-148.

Bertail, P. and Combris, P. (1994). Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique*, **46**, 49-83.

Chauvet, G. (2014). Coupling methods for multistage sampling Soumis.

Chen, X.H., and Dempster, A.P., and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy *Biometrika*, **81**, 457-469.

Cochran, W.G. (1977). Sampling Techniques. *New-York, Wiley.*

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and their Application, *Cambridge University Press.*

Davison, A.C. and Sardy, S. (2007). Resampling variance estimation in Surveys with missing data. *Journal of Official Statistics*, **23**, 371-386.

Deville, J-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes: linéarisation et techniques des résidus. *Techniques d'enquête*, **25**, 219-230.

Deville, J-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. *Rapport technique*, CREST-ENSAI.

Ezzati, T.M., Hoffman, K., Judkins, D.R., Massey, J.T., and Moore, T.F. (1992). Sample design: Third National Health and Nutrition Examination Survey. *Vital and Health Statistics*, **2**, 113, National Center for Health Statistics.

Fuller, W.A. (2009). Sampling Statistics. *New-York, Wiley.*

Funaoka, F. and Saigo, H. and Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, **32**, 151-156.

Goga, C, and Deville, J-C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, **96**, 691-709.

Gordon, L. (1983). Successive sampling in large finite populations. *The Annals of Statistics*, **11**, 702-706.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 1491-1523.

Hajek, J. (1981). Sampling from a finite population. *New-York, Marcel Dekker.*

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, **14**, 333-362.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, **18**, 199-210.

Lin, C.D. and Lu, W.W. and Rust, K. and Sitter, R.R. (2013). Replication variance estimation in unequal probability sampling without replacement: One stage and two stage. *Canadian Journal of Statistics*, **41**, 696-716.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169-175.

Nigam, A. K. and Rao, J.N.K. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica*, **6**, 199-214.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, **35**, 227-234.

Rao, J.N.K and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the american statistical association*, **83**, 231-241.

Rao, J.N.K. and Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey methodology*, **18**, 209-217.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. *New-York, Springer-Verlag.*

Shao, J. and Tu, D. (1995). The Jackknife and the Bootstrap, *New-York, Springer.*

Tillé (2011). Sampling Algorithms, *New-York, Springer.*

Table 1: Relative Bias and Relative Stability of a Bootstrap variance estimator and Nominal One-Tailed Error Rates of the Percentile Method for the estimation of a total, of a ratio and of a coefficient of correlation with a sample size of $n_0 = 5$ for the second stage

		$\rho = 0.1$					$\rho = 0.2$					$\rho = 0.3$				
		RB	RS	L	U	L+U	RB	RS	L	U	L+U	RB	RS	L	U	L+U
Total	$f_I = 1\%$	0.02	0.34	3.0	3.9	6.9	0.02	0.35	3.9	3.2	7.1	0.01	0.31	2.4	3.8	6.2
	$f_I = 2.5\%$	-0.01	0.21	3.8	2.0	5.8	0.01	0.21	3.1	3.2	6.3	0.01	0.22	2.9	2.4	5.3
	$f_I = 5\%$	0.01	0.15	2.4	2.9	5.3	0.02	0.15	2.0	2.3	4.3	0.02	0.15	1.8	2.5	4.3
	$f_I = 10\%$	0.04	0.12	3.3	1.9	5.2	0.04	0.12	3.3	2.3	5.6	0.04	0.12	2.3	2.2	4.5
	$f_I = 25\%$	0.08	0.12	2.0	2.0	4.0	0.13	0.16	2.0	2.6	4.6	0.17	0.19	2.0	2.0	4.0
Ratio	$f_I = 1\%$	0.03	0.33	3.3	3.2	6.5	0.00	0.33	3.5	3.0	6.5	0.02	0.33	3.1	3.7	6.8
	$f_I = 2.5\%$	-0.01	0.21	3.4	3.4	6.8	0.01	0.21	2.6	2.9	5.5	0.02	0.22	2.6	3.2	5.8
	$f_I = 5\%$	0.02	0.15	2.0	2.4	4.4	0.00	0.15	3.5	2.6	6.1	0.00	0.15	2.0	2.2	4.2
	$f_I = 10\%$	0.02	0.11	1.9	3.3	5.2	0.01	0.11	2.7	2.1	4.8	0.00	0.11	2.1	2.6	4.7
	$f_I = 25\%$	0.02	0.08	2.1	2.2	4.3	0.04	0.09	2.2	3.2	5.4	0.02	0.08	2.8	2.8	5.6
Coef. of correlation	$f_I = 1\%$	-0.03	0.41	3.5	3.7	7.2	0.00	0.42	3.9	4.1	8.0	0.01	0.44	3.3	3.3	6.6
	$f_I = 2.5\%$	0.01	0.27	3.8	2.8	6.6	0.02	0.27	3.6	2.8	6.4	0.04	0.29	3.1	4.4	7.5
	$f_I = 5\%$	0.01	0.19	3.4	2.9	6.3	0.01	0.20	2.8	3.5	6.3	-0.03	0.20	2.9	3.0	5.9
	$f_I = 10\%$	0.01	0.15	2.2	1.9	4.1	0.01	0.14	2.5	2.6	5.1	0.03	0.15	2.6	2.7	5.3
	$f_I = 25\%$	0.03	0.10	2.6	1.7	4.3	0.06	0.12	2.0	2.5	4.5	0.04	0.11	2.9	1.8	4.7

Table 2: Relative Bias and Relative Stability of a Bootstrap variance estimator and Nominal One-Tailed Error Rates of the Percentile Method for the estimation of a total, of a ratio and of a coefficient of correlation with a sample size of $n_0 = 20$ for the second stage

		$\rho = 0.1$					$\rho = 0.2$					$\rho = 0.3$				
		RB	RS	L	U	L+U	RB	RS	L	U	L+U	RB	RS	L	U	L+U
Total	$f_I = 1\%$	-0.01	0.33	2.6	3.1	5.7	0.00	0.33	2.8	3.3	6.1	0.01	0.33	2.4	3.2	5.6
	$f_I = 2.5\%$	0.02	0.21	3.2	2.3	5.5	0.04	0.21	2.2	2.1	4.3	0.04	0.22	3.0	2.9	5.9
	$f_I = 5\%$	0.03	0.16	2.8	1.9	4.7	0.06	0.17	2.6	2.0	4.6	0.06	0.17	2.3	1.3	3.6
	$f_I = 10\%$	0.10	0.15	1.9	2.3	4.2	0.10	0.15	2.3	2.4	4.7	0.13	0.18	2.8	2.1	4.9
	$f_I = 25\%$	0.23	0.24	1.4	1.9	3.3	0.25	0.27	1.0	2.0	3.0	0.29	0.30	1.2	2.1	3.3
Ratio	$f_I = 1\%$	-0.01	0.33	3.5	3.5	7.0	-0.01	0.32	2.9	3.5	6.4	0.01	0.33	2.3	3.8	6.1
	$f_I = 2.5\%$	0.03	0.22	3.1	3.1	6.2	0.02	0.21	3.0	2.2	5.2	0.01	0.21	3.1	3.3	6.4
	$f_I = 5\%$	0.01	0.15	2.6	2.8	5.4	0.03	0.16	2.3	3.3	5.6	0.02	0.15	2.1	2.1	4.2
	$f_I = 10\%$	0.04	0.13	1.9	2.6	4.5	0.03	0.12	3.2	1.7	4.9	0.04	0.12	3.5	1.9	5.4
	$f_I = 25\%$	0.14	0.17	2.4	1.8	4.2	0.15	0.17	1.4	1.7	3.1	0.13	0.16	2.0	2.0	4.0
Coef. of correlation	$f_I = 1\%$	0.00	0.36	2.3	2.6	4.9	0.00	0.35	1.7	4.6	6.3	-0.02	0.36	1.9	4.8	6.7
	$f_I = 2.5\%$	0.02	0.23	2.8	2.9	5.7	0.02	0.23	1.9	3.4	5.3	0.01	0.25	2.1	3.7	5.8
	$f_I = 5\%$	0.01	0.16	2.9	2.0	4.9	0.04	0.17	3.6	3.2	6.8	0.03	0.19	2.1	3.1	5.2
	$f_I = 10\%$	0.06	0.13	2.8	3.1	5.9	0.07	0.15	1.6	2.9	4.5	0.05	0.15	1.6	2.6	4.2
	$f_I = 25\%$	0.14	0.17	1.8	0.9	2.7	0.15	0.18	2.0	1.7	3.7	0.19	0.22	1.5	1.9	3.4