Bootstrap algorithms for variance estimation in complex survey sampling

Alessandro Barbiero[1] and Fulvia Mecatti[2]

[1] Department of Economics, Business and Statistics, University of Milano
(e-mail: alessandro.barbiero@unimi.it)
[2] Department of Statistics, University of Milano-Bicocca
(e-mail: fulvia.mecatti@unimib.it)

Abstract. The problem of estimating the variance of the Horvitz-Thompson estimator under a probability proportional to size design is concerned. Some *IPPS*-bootstrap algorithms are proposed with the purpose of both simplifying available procedures and of improving efficiency. Results from a simulation study using both natural and artificial data are presented in order to empirically study the bias and stability of the bootstrap variance estimators proposed.

1 Introduction

In complex survey sampling every population unit $i \in U$ is assigned a specific probability $\pi_i, (i = 1 \ldots N)$ to be included in the sample and the random mechanism providing sample data further violates the classical *iid* hypothesis for instance with cluster, multistage and without replacement selection. We assume the total $Y = \sum_{i=1}^{N} y_i$ of a quantitative study variable y as the parameter to be estimated. A sampling design without replacement and with inclusion probability proportional to an auxiliary variable x (usually referred as *IPPS* sampling or πPS sampling) paired with the well-known unbiased Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i=1}^{n} y_i/\pi_i$ devises a strategy methodologically appealing since the estimator variance $V(\hat{Y}_{HT})$ tends to zero as the relationship between x and y approaches proportionality. From a practical prospective a variance estimator is essential for assessing estimate's accuracy and for providing confidence intervals. For a fixed sample size n, the Sen-Yates-Grundy estimator

$$\hat{v}_{SYG} = \sum_{i<j} \sum_{j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \qquad (1)$$

has a closed analytic form and is unbiased for $V(\hat{Y}_{HT})$ under non restrictive conditions. However it presents some drawbacks which limit the applications: $\hat{v}_{SYG}$ depends on the joint inclusion probability π_{ij} of pair of sampled units $i \neq j \in U$, $(i, j = 1 \ldots N)$ which can not be computed for sample sizes greater than 2 for the greatest part of the collection of *IPPS* designs available in literature, it is not uniformly positive for any *IPPS* design and it is often stated as highly instable in practical applications. A bootstrap estimate, although numeric, is a natural alternative for addressing those issues since it is positive by construction, can be computed for any sample size and does not require the explicit knowledge of joint inclusion probabilities. Since the original Efron's bootstrap applies in the classical *iid* setup (Efron, 1979; Shao and Tu, 1995) suitable modified bootstrap algorithms are needed in order to handle the complexity of the sampling. In section 2 *IPPS*-bootstrap algorithms previously appeared in literature are briefly discussed. Three modified *IPPS*-bootstrap algorithms aiming at both simplifying the procedure and improving efficiency are introduced in section 3 and 4. Results from an extended simulation study using both natural and artificial data are presented in section 5.

2 *IPPS* BOOTSTRAP ALGORITHMS

Several proposals to adapt the original Efron's bootstrap to handle with non-*iid* situations have been introduced, particularly for the without replacement selection. Among the other, the *with-replacement* bootstrap (Mc Carthy and Snowden, 1985), the *rescaling* bootstrap (Rao and Wu, 1988), the *mirror-match* bootstrap (Sitter, 1992) and the *without-replacement* bootstrap (Gross, 1980 and Chao and Lo, 1985). The latter is based on a pseudo-population U^* termed *bootstrap population*, formed by sampled units only each replicated N/n times, and on a without replacement resampling from U^* with the same sample size n as in the original sample. Holmberg (1998) generalized this approach for a general *IPPS* sampling design. For each unit $i = 1 \ldots n$ included in the original sample s, the inverse of the inclusion probability is decomposed as $1/\pi_i = c_i + r_i$ where $c_i = \lfloor 1/\pi_i \rfloor$, i.e. the integer part, and $0 \leq r_i < 1$. Let ε_i be the realization of n independent Bernoulli random variables with parameter r_i. Then define $d_i = c_i + \varepsilon_i$. The Holmberg's *IPPS*-bootstrap algorithm consists of the following steps:

1. Construct the bootstrap population $U^* = \{1^*, \ldots, i^*, \ldots, N^*\}$ replicating d_i times each unit in s . Thus $N^* = \sum_{i=1}^{n} d_i$ and $X^* = \sum_{i=1}^{n} d_i x_i$.
2. Select from U^* a sample s^* of size n according to the original *IPPS* sampling design with (resampling) inclusion probabilities $\pi_{i^*} = n x_{i^*} / X^*$.
3. Calculate the replication $\hat{Y}^* = \sum_{i \in s^*} y_i / \pi_{i^*}$ of $\hat{Y}_{HT}$.
4. Repeat steps 3 to 4 B times (B chosen sufficiently large) providing the bootstrap distribution $\left\{\hat{Y}_b^*, b = 1, \ldots, B\right\}$
5. Let $\hat{v}_{boot}$ be the variance of the bootstrap distribution; the Holmberg's bootstrap estimate of $V\left(\hat{Y}_{HT}\right)$ is given by

$$\hat{v}_{bH} = \frac{n}{n-1} \cdot \hat{v}_{boot} \tag{2}$$

Note that in the Holmberg's method a further step in the bootstrap algorithm is needed for constructing the bootstrap population U^*. Particularly, in step 1. n random variables have to be simulated in order to compute the weights d_i. Then, if r_i does not equal zero for some i, an entire class $\mathcal{U} = \left\{U_h^*, h = 1 \ldots 2^n\right\}$ of 2^n possible bootstrap populations remains defined. The further step is actually performed to select a unique bootstrap population by randomization into $\mathcal{U}$. As a consequence the Holmberg's *IPPS*-bootstrap results computationally heavy and resource consuming.

3 0.5 *IPPS* BOOTSTRAP

Our first proposal aims at simplifying the original Holmberg's algorithm by skipping the randomization step discussed above. We will call this modified algorithm "0.5 *IPPS*-bootstrap" since it is based on the following trivial approximation in order to compute weights d_i

$$d_i = \begin{cases} c_i & \text{if } r_i < 0.5 \\ c_i + 1 & \text{if } r_i \geq 0.5 \end{cases} \tag{3}$$

Hence a unique bootstrap population U^* is readily derived. Moreover, it is the maximum probability bootstrap population in the class $\mathcal{U}$. In fact it maximizes the joint probability function of the n independent Bernoulli trials required by the original Holberg's algorithm. Except for this slight modification, the bootstrap steps remain as described in section 2.

4 x-BALANCED *IPPS* BOOTSTRAP

With the last two proposals efficiency gains in the bootstrap variance estimator are fostered by a more complete use of the auxiliary information. We suggest to balance (Tillé, 2006) with respect to the known population total $X = \sum_{i=1}^{N} x_i$ when constructing U^* i.e. under the restriction $X^* \approx X$. Let U_0^* be the basic bootstrap population formed by sampled units each replicated c_i times. Starting from U_0^*, iteratively add sampled units $i \in s$ previously sorted in a decreasing order according to r_i. The process ends when the bootstrap population ensuring the best approximation to X is detected in $\mathcal{U}$, i.e. when $|X^* - X|$ reaches its minimum in $\mathcal{U}$. The algorithm consists of the following steps:

```
1. Start with $U^*_{(0)} = U^*_0$; $s_{(0)} \leftarrow s$
2. Iteration step:
   $t \leftarrow 1$
   Select unit $k_t$ in $s_{(t)}$ so that $r_{k_t} \geq r_j \quad \forall j \in s_{(t)}$
   Add unit $k_t$ to $U^*_{(t-1)}$ thus producing $U^*_{(t)}$
   If $X^*_{(t)} > X$, exit the loop
   next $t$
3. If $|X^*_{(t)} - X| < |X^*_{(t-1)} - X|$ then $U^* = U^*_{(t)}$, otherwise $U^* = U^*_{(t-1)}$
4. Perform steps 2 to 4 of the original Holmberg's algorithm as described in section 2.
```

We will denote the resulting bootstrap estimate of $V(\hat{Y}_{HT})$ with $\hat{v}_{bHx1}$. The last proposal consists of the previous algorithm except for the fact that an additional unit $i \in s$ is inserted into $U^*_{(t-1)}$ by considering the values $q_i = \pi_i^{-1}/(c_i + 1)$ instead of r_i. By using q_i an advantage to units with higher c_i for equal r_i is given, i.e. to units appearing with larger frequency in U_0^*. The resulting bootstrap estimate will be denoted by $\hat{v}_{bHx2}$. Notice that both the x-balanced algorithms ensure the construction of U^* in a number of steps less or equal to n leading to a bootstrap population included in $\mathcal{U}$. Hence a potential computational advantage with respect to the original Holmberg's algorithm is given while efficiency improvements are expected from using a bootstrap population closer the the actual population according to a basic bootstrap principle.

5 SIMULATION

In order to check the performance of the algorithms proposed a simulation study has been carried out. Several variance estimators have been compared: the customary $\hat{v}_{SYG}$, the naïve bootstrap estimator provided by the classical Efron's bootstrap, Holmberg's bootstrap estimator $\hat{v}_{bH}$ as given in section 2 and the three variance estimators provided by the algorithms proposed in section 3 and 4: $\hat{v}_{bH0.5}$, $\hat{v}_{bHx1}$, $\hat{v}_{bHx2}$. Two approximate estimators following by approximating the joint inclusion probabilities π_{ij} in terms of π_i only as recommended from previous simulation studies (Haziza, Mecatti and Rao, 2004; 2008) have been also considered: $\hat{v}_{HR}$ (Hartley and Rao, 1962) and $\hat{v}_{BM}$. Samples were simulated under the Rao-Sampford *IPPS* design and standard Monte Carlo performance indicators have been computed: the MC Relative Bias: $RB = \left(E_{MC}(\hat{v}) - V(\hat{Y}_{HT})\right)/V(\hat{Y}_{HT})$, the MC Relative Efficiency of a bootstrap estimator $\hat{v}$ with respect to $\hat{v}_{SYG}$: $Eff = MSE_{MC}(\hat{v}_{SYG})/MSE_{MC}(\hat{v})$ and the MC coverage of 95% bootstrap confidence intervals according to the percentile method *cove*.

5.1 SIMULATION DESIGN

Both natural and artificial populations have been considered. Two natural populations from the MU281 dataset of 281 Swedish municipalities (Särndal *et al.*) consisting of $N = 100$ units randomly selected from MU281 and three artificial populations produced as follows. The auxiliary variable x was generated according to a random variable Gamma with parameters α and β giving chosen levels of variability of X as measured by cv_x. The study variable y was generated conditionally to x under the model $y_i|x_i = ax_i + n_i$, where n_i are N independent random variables normally distributed with zero mean and variance $\sigma^2 x_i$. The values of a and σ^2 were chosen to garantee the correlation between x and y close to 0.9 since high correlation suggests the use of a *IPPS* sampling design. The simulation set up is described in Table 1.

Population	N	cv_y	cv_x	ρ_{xy}
MU100	100	1.107	1.015	0.9931
MU100CS	100	0.325	0.527	0.2829
GN1	100	0.529	0.598	0.897
GN2	100	0.981	1.122	0.916
GN3	100	1.419	1.692	0.928

Table 1. Characteristics of natural and artificial populations simulated

0.05, 0.10 and 0.15 were used for the sampling fraction $f = n/N$ under the restriction $\pi_i < 1$ for all population units. The number of simulation steps (between 1000 and 10000) has been used to control the Monte Carlo error according to the rule: $\left|\left[E_{MC}\left(\hat{Y}_{HT}\right) - Y\right]/Y\right| < 1\%$ and $\left|\left[E_{MC}(\hat{v}_{SYG}) - V(\hat{Y}_{HT})\right]/V(\hat{Y}_{HT})\right| < 3\%$.

5.2 SIMULATION RESULTS

A synthesis of the simulation results is displayed in Tables 2, 3 and 4. It clearly appears the poor performance of the naïve bootstrap algorithm when applied to a non-*iid* situation as in the *IPPS* sampling design. Simulation results also show the good performance of the modified algorithms proposed in section 3 and 4 in terms of bias and relative efficiency as compared with all the other estimators considered both bootstrap and analytic. In some cases they allows for efficiency gains greater or equal 7% with respect to the customary estimator. As a conclusion the bootstrap approach which is more general than the analytic approach for applying to any *IPPS* sampling design with any sample size n, can be improved as suggested in section 3 and 4 under both respect of computational simplification and statistical properties of the resulting variance estimator. Future research will concern the estimation of other population characteristics such as the median for which there is no analytic standard variance estimator.

estimator	MU100 $f = 0.05$			MU100 $f = 0.10$			MU100CS $f = 0.05$			MU100CS $f = 0.10$		
	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove
SYG	-1.62	1.0000	94.1	0.85	1.0000	94.5	-2.46	1.0000	88.2	-1.18	1.0000	87.3
bH	0.23	0.9671	79.3	1.76	1.0182	76.9	-2.42	1.0191	78.3	-1.24	1.0178	83.6
bH0.5	0.41	1.0031	79.9	5.58	1.0093	81.2	-2.20	1.0154	78.0	-1.82	1.0300	82.4
bHx1	0.53	0.9939	82.4	1.12	1.0740	89.2	-1.43	0.9842	78.1	-0.94	1.0191	83.4
bHx2	0.13	0.9872	82.3	3.00	1.0108	87.3	-2.12	0.9960	78.5	0.37	0.9995	83.5
bnaïve	8.22	0.8747	85.1	23.46	0.6538	92.3	1.15	0.9859	79.0	7.00	0.9217	85.1
BM	0.40	0.9957	94.2	3.45	1.0287	94.7	-2.14	1.0102	88.8	-0.83	1.0143	87.7
HR	-1.48	0.9991	94.1	1.53	0.9937	94.5	-2.44	0.9998	88.3	-1.06	0.9983	87.4

Table 2. Simulation results for natural populations.

estimator	GN1 $f = 0.05$			GN1 $f = 0.10$			GN1 $f = 0.15$		
	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove
SYG	2.01	1.0000	96.0	2.27	1.0000	94.6	0.46	1.0000	94.4
bH	3.21	0.9760	84.1	3.39	0.9828	89.6	1.50	0.9718	90.0
bH0.5	2.84	0.9960	83.6	5.00	0.9249	89.3	5.44	0.8680	90.4
bHx1	3.16	0.9784	84.7	3.59	0.9698	90.2	2.97	0.9587	91.4
bHx2	3.35	0.9785	85.0	3.66	0.9663	89.8	2.17	0.9557	91.3
bnaïve	8.63	0.8813	85.2	16.36	0.7302	91.4	22.64	0.5556	94.0
BM	3.24	0.9834	96.2	3.66	0.9851	94.8	1.83	0.9993	94.7
HR	1.99	1.0009	96.0	2.23	1.0020	94.6	0.38	1.0030	94.4

Table 3. Simulation results for artificial population GN1.

estimator	GN2 $f = 0.05$			GN2 $f = 0.10$			GN2 $f = 0.15$			GN3 $f = 0.05$			GN3 $f = 0.10$		
	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove	RB	Eff	cove
SYG	2.84	1.0000	95.8	0.98	1.0000	95.9	0.64	1.0000	95.4	2.04	1.0000	95.7	-2.29	1.0000	93.8
bH	4.57	0.9963	82.1	2.19	1.0055	89.7	1.87	0.9915	86.8	4.45	1.0080	84.0	0.20	0.9012	84.4
bH0.5	4.58	0.9925	82.3	5.61	0.9394	89.2	-2.44	1.1301	90.6	0.78	1.0769	85.8	-1.27	1.0421	86.6
bHx1	4.45	0.9996	82.3	3.02	0.9914	90.9	-0.43	1.0804	91.6	2.50	1.0350	84.9	-0.55	1.0140	88.0
bHx2	4.28	1.0011	82.3	2.49	0.9756	91.3	3.40	0.9424	91.1	4.34	0.9802	84.0	1.50	0.9298	88.6
bnaïve	13.76	0.8556	83.1	24.00	0.6554	94.1	40.33	0.4302	95.6	19.13	0.8212	86.8	40.57	0.4331	95.0
BM	4.89	0.9997	95.8	3.07	1.0064	94.8	3.18	1.0374	95.5	5.26	1.0279	95.8	2.46	1.0090	95.6
HR	2.81	1.0033	95.8	0.91	1.0061	94.6	0.49	1.0125	95.4	1.92	1.0184	95.7	-2.43	1.0330	94.4

Table 4. Simulation results for artificial populations GN2 and GN3.

REFERENCES

CHAO, M. T., LO, A. Y. (1985) A bootstrap method for finite population. *Sankhya: The Indian Journal of Statistics*, *47(A)*, 399–405

EFRON, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *7*, 1–26

GROSS, S. (1980) Median estimation in sample surveys. In *Proceedings of Section on Survey Research Methods*, *American Statistical Association*, 181–184

HARTLEY, H. O., RAO, J. N. K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, *32*, 350–374

HAZIZA, D., MECATTI, F., RAO, J.N.K. (2004). Comparison of variance estimators under Rao-Sampford method: a simulation study. In *Proceedings of the ASA Joint Statistical Meeting, Section on Survey Research methods*, 3638–3643

HAZIZA, D., MECATTI, F., RAO, J.N.K. (2008) Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, *1*, 89–106

HOLMBERG, A. (1998) A bootstrap approach to probability proportional to size sampling. In *Proceedings of Section on Survey Research Methods*, *American Statistical Association*, 378–383.

McCARTHY, P.J. and SNOWDEN, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, *2(95)*, U.S. Government Printing Office, Washington

RAO, J. N. K. and WU, C. F. J. (1988) Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83(401)*, 231–241

SÄRNDAL, C.E., SWENSSON B., WRETMAN J. (1992) Model Assisted Survey Sampling. Springer, New York

SHAO, J., TU, D. (1995) The jackknife and bootstrap. Springer, New York

SITTER, R. R. (1992) A resampling procedure for complex survey data. *Journal of the American Statistical Association*, *87(419)*, 755–765

TILLÉ, Y. (2006) Sampling algorithms. Springer, New York