Bootstrap Algorithms for Probability Proportional to Size Sampling

Algoritmi bootstrap per campioni a probabilità variabili

Alessandro Barbiero, Fulvia Mecatti
Department of Statistics, University of Milan-Bicocca
email: fulvia.mecatti@unimib.it

Keywords: joint inclusion probability, model-assisted πPS bootstrap, variance estimation, simulations

1. The problem

Unequal probability sampling without replacement with inclusion probabilities π_i exactly proportional to a measure of size *x* known for each unit (denoted by πPS) is extensively used in large-scale surveys, especially for the selection of primary sampling units in multi-stage sampling. For simplicity, we focus on unistage sampling from a finite population U of size N. The design-unbiased Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$ is used to estimate the total Y of a characteristic of interest *y*, where *s* is a sample of fixed size *n*, $\pi_i = n\, x_i / X$ and $X = \sum_{i \in U} x_i$. The problem of evaluating the accuracy of $\hat{Y}_{HT}$ by estimating its variance $V\left(\hat{Y}_{HT}\right)$ is concerned. The customary Sen-Yates-Grundy variance estimator $v_{SYG}\left(\hat{Y}_{HT}\right)$ is exactly unbiased; on the other hand it is not uniformly positive under any πPS design and it involves the joint inclusion probabilities π_{ij} which are computationally cumbersome for $n > 2$. It is also often stated that $v_{SYG}\left(\hat{Y}_{HT}\right)$ can be very unstable. Consequently, several approximately unbiased variance estimators, based on approximating π_{ij} in terms of the π_i's only, have been proposed in recent literature and extensively analyzed via simulations (Haziza, Mecatti and Rao, 2004). A natural alternative is a bootstrap variance estimator.

2. Methods

Since Efron's original bootstrap applies to the classical *iid* framework, suitable modifications are needed in order to handle the πPS context. We focus on πPS bootstrap algorithms based upon the notion of bootstrap population, as a natural extension of the Gross-Chao-Lo bootstrap for equal probability sample without replacement from a finite population (Chao and Lo, 1985).

We refer to a bootstrap population as a set U^* formed by replicating w_i times every sampled unit $i \in s$ so that the bootstrap population includes data from the original sample *s* only, according to a basic bootstrap principle. Moreover U^* has size $N^* = \sum_{i \in s} w_i$ and total of the auxiliary variable $X^* = \sum_{i \in s} w_i\, x_i$. A class of πPS bootstrap algorithms originates for different choices of weights w_i and of the re-

sampling design. We suggest to evaluate w_i by calibrating with respect to known features of the actual population generating the data, namely N and X.

With the Holmberg πPS bootstrap algorithm (Holmberg, 1998) U^* follows by setting $w_i = \pi_i^{-1}$ and by mimicking the original sample design in the re-sampling step. Hence $E_*(N^*) = N$ and $E_*(X^*) = X$ where E_* denotes expectation under the re-sampling design. Since in the general case $\pi_i^{-1} = [\pi_i^{-1}] + r_i$ with $0 \le r_i < 1$, U^* actually arises via randomization by performing n independent bernoulli trials with probability r_i. Former empirical results (Manzi and Mecatti, 2007) suggest that the Holmberg algorithm is able to give encouraging results in terms of unbiasdness and stability of the variance estimator but is computational heavy, considerably resource-consuming and it allows efficiency improvements.

In this paper alternative πPS bootstrap algorithms are considered with the main purpose of a) simplifying the re-sampling step of the Holmberg algorithm according to Mecatti (2000) and Manzi and Mecatti (2007) in order to foster computational advantages; b) exploring alternatives to the randomization proposed by Holmberg to provide the bootstrap population U^* according to the calibration restrictions $N^* = N$ and $X^* = X$.

It is also known that the efficiency of a πPS sampling over a simple random sampling with equal probability increases as the relationship between y and x approaches proportionality. Hence a ratio model $y_i = \beta x_i + \varepsilon_i$ where the ε_i's are independent random variables with $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) \propto x_i^2$, is implicitly assumed with a πPS design. Alternative choices of the bootstrap weights w_i according to a model-assisted approach assuming the ratio model above are also explored.

A simulation study using artificial data will be performed in order to empirically study the bias and stability of the variance estimator supplied by the πPS bootstrap algorithms developed. Comparisons with the original Holmberg algorithm, with the classical Sen-Yates-Grundy variance estimator and with a selection of nearly unbiased variance estimators based on approximating the π_{ij}, will be also provided.

References

Haziza D., Mecatti F., Rao J.N.K. (2004) Comparison of variance estimators under Rao-Sampford method: a simulation study, *Proceedings of the ASA Joint Statistical Meeting, Section on Survey Research methods [CD-ROM]*, 3638-3643.

Chao M.T., Lo S.H. (1985) A bootstrap method for finite population, *Sankhyā*, 47, A, 399-405.

Holmberg A. (1998) A bootstrap approach to probability proportional-to-size sampling, *Proceedings of the ASA Section on Survey research Methods*, 378-383.

Mecatti F. (2000) Bootstrapping unequal probability samples, *Statistica Applicata*, 12, 67-77.

Manzi G.C., Mecatti F. (2007) Bootstrap algorithms for risk models with auxiliary variable and complex samples, *Proceedings of the 2007 intermediate conference SIS*, 555-556. http://sis2007.unive.it/sessioniposter-it.html#sessione6