Simulation-assisted Horvitz-Thompson statistic and isotonic regression

Wojciech Gamrot[1]

Abstract. Under some finite-population sampling schemes the calculation o exact inclusion probabilities may be prohibitively complex even for modest population sizes. This is especially true for various sequential procedures used in spatial sampling and for fixed-cost (or sum-quota) schemes. Such a phenomenon presents a significant challenge for constructing estimates of finite population totals based on the Horvitz-Thompson approach. Such a challenge may be overcome by replacing unknown first-order inclusion probabilities with estimates computed in a simulation study which is enabled by the knowledge of the sampling scheme. Such estimates may be calculated in several ways, which influences stochastic properties of the Horvitz-Thompson statistic. Available auxiliary information may also be used to improve their accuracy. In this paper isotonic regression algorithms are applied to capitalize on limited auxiliary information and to improve the accuracy of simulation-assisted design-based estimates for finite population totals.

Keywords: empirical inclusion probability, Horvitz-Thompson estimator, simulation, population total

JEL classification: C83
AMS classification: 62D05

1 Introduction: empirical Horvitz-Thompson estimation

Consider a finite population represented by a set of indices $U = \{1, ..., N\}$. Values $y_1, ..., y_N$ of a fixed characteristic correspond to each population unit. The parameter under study is the population total:

$$t = \sum_{i \in U} y_i \tag{1}$$

In order to estimate it, an unordered sample s is drawn from U through some sampling scheme characterized by inclusion probabilities of the first-order: $p_1, ..., p_N$ where $p_i = Pr\{i \in s\}$ for $i \in U$. If inclusion probabilities are known then the population total may be estimated without design bias using the well-known Horvitz-Thompson (H-T) statistic [9]:

$$\widehat{t} = \sum_{i \in s} \frac{y_i}{p_i} \tag{2}$$

However, for some sampling schemes exact calculation of first-order inclusion probabilities $p_1, ..., p_N$ may be impossible because of prohibitive computational complexity. This is particularly true for various sequential sampling schemes like those described in [2] and [4] where the combinatorial explosion prevents the computation of inclusion probabilities even for very modest population sizes. At first sight, this effect seems to make the H-T estimation impossible. Fortunately, as noted in [5],[15], another potent source of information still remains in the hands of a statistician. Namely, the sampling procedure itself. This may be used to generate a large number of independent sample replications. By examining sample membership indicator values corresponding to any particular population unit within all replications one may arrive at the estimate of associated inclusion probability. This estimate may take the form of sample proportion of ones or some other function of membership indicators. Let R be the number of replications and let f_i be a number of times a certain i-th population unit is drawn to a sample replication. A classic approach

[1]University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland, e-mail: wojciech.gamrot@ue.katowice.pl

proposed in [6] would rely on estimating the inclusion probability p_i through the statistic:

$$\hat{p}_{iF} = \frac{f_i + 1}{R + 1} \tag{3}$$

for $i \in s$ and inserting these estimates instead of p_i into (2). However, as noted in [6], the number of sample replications needed to guarantee a desired level of accuracy for population total estimates may still be uncomfortably large, even with respect to contemporary computing capabilities. One possibility of overcoming such a difficulty is the adoption of sequential methods at the simulation stage as proposed in [6]. In this paper another approach to empirical H-T estimation based on the use of external information is investigated.

Let us assume that available auxiliary information takes form of an ordering constraint on first-order inclusion probabilities so that these probabilities are known to behave monotonically. Such a situation arises for many sampling schemes such as Pareto sampling discussed in [13] where inclusion probability grows with increasing values of some known auxiliary variate. It also appears in the case of fixed-cost sum-quota sampling proposed in [11], where inclusion probabilities decrease with growing cost of observing the variable. It is worth noting, that when subsequent replications are generated, the values of sample membership indicators may be recorded for all population units, as opposed to observing only units in the sample s. The additional computational effort associated with recording all of them is negligible. Then various isotonic regression algorithms may be applied to enhance the accuracy of inclusion probability estimates within the sample by forcing the ordering constraints to be satisfied by all the estimates in the population. Hence such approach may be viewed as a special indirect case of "strength-borrowing" technique discussed in [10]. In the next section one such algorithm is presented.

2 The PAVA procedure

The well-known Pool-Adjacent-Violators Algorithm (PAVA) works in the following way (see [1], [8], [3]). Let $p_1, p_2, ..., p_N$ be unknown probabilities satisfying a simple order:

$$p_1 \leq p_2 \leq ... \leq p_N \tag{4}$$

Let R_i independent trials be made of an event with probability p_i and let f_i denote the number of successes in these trials ($i = 1, ..., N$). Constraint-preserving estimates $\hat{p}_1, ..., \hat{p}_N$ of $p_1, ..., p_N$ satisfying (1) are calculated by iteratively grouping (merging) initial unconstrained estimates $f_1/R_1, ..., f_N/R_N$ (sample proportions of ones) into groups and repeatedly averaging them within each group. The procedure works through following steps .

Step one: assign the index of each probability p_i for $i = 1, ..., N$ to a separate group so that initial groups are $A_1^{(0)} = \{1\}, ..., A_N^{(0)} = \{N\}$ and initial number of groups is $a^{(0)} = N$. Set an initial estimate of mean probability in each i-th group to $q_i^{(0)} = f_i/R_i$ for $i = 1, ..., N$.

Step two: in each subsequent step of the procedure (numbered $m = 1, 2, ...$) whenever mean probability estimates in some neighboring groups are found to breach the order constraint, a maximum-length sequence $A_i^{(m-1)}, ..., A_{i+z}^{(m-1)}$ (where $1 \leq i < i + z \leq a^{(m-1)}$) of such groups is merged together so that

$$A_i^{(m)} = A_i^{(m-1)} \cup ... \cup A_{i+z}^{(m-1)}$$

and

$$A_j^{(m)} = A_{j+z}^{(m-1)}$$

for $j = i + 1, ..., a^{(m)}$ while $a^{(m)} = a^{(m-1)} - z$. Then a new within-group mean probability estimate:

$$q_i^{(m)} = \frac{\sum_{j \in A_i^{(m)}} f_j}{\sum_{j \in A_i^{(m)}} R_j}$$

is assigned to the group $A_i^{(m)}$ while $q_j^{(m)} = q_{j+z}^{(m-1)}$ for $j = i + 1, ..., a^{(m)}$. This step is repeated until all estimates $q_1^{(m)}, ..., q_{a^{(m)}}^{(m)}$ of mean within-group probabilities satisfy ordering constraints or there is just one group left.

Step three: when the iteration stops after the last - say M-th - step ($M \in \{1, 2, ...\}$) a mean probability estimate computed for a group is assigned to each of its member components so that the final estimate for the probability p_i is $\hat{p}_i = q_j^{(M)}$ for $i \in A_j$, $j = 1, ..., a^{(M)}$.

If $f_1, ..., f_N$ are independent, this procedure leads to a vector of restricted maximum likelihood estimates for probabilities $p_1, p_2, ..., p_N$. Such estimates may find non-trivial applications in various fields of study including medicine, toxicology or calculating insurance premiums ([14],[16]). At the same time, for most sampling schemes a significant dependence between some sample membership indicators may appear. Resulting lack of independence among $f_1, ..., f_N$ seems to prevent the use of PAVA to estimate ordered inclusion probabilities. However, new results obtained by Gamrot [7] indicate that PAVA-based estimates remain consistent even in the case of strong correlation between sample membership indicators. Hence the empirical Horvitz-Thompson estimator constructed upon them should also remain consistent. A more detailed investigation of its properties for a specific sampling scheme is presented in the next section.

3 Simulation results

The sequential fixed-cost sampling scheme of Pathak [11] features varying inclusion probabilities. Their exact evaluation is very demanding computationally even for modest sample sizes. Nevertheless, despite the existence of sufficiency-based unbiased estimators that do not rely on inclusion probabilities, empirical H-T estimation may be of interest when nonresponse corrections need to be incorporated or when some modifications are introduced to the original scheme. In this section the Pathak procedure in its original form serves as an illustration of PAVA-based empirical H-T estimation. The selection procedure works in the following way. Let $c_1, ..., c_N$ be costs of observing the value of characteristic under study for corresponding population units, known in advance. Without a loss of generality one may assume that units are pre-ordered by decreasing value of this cost so that $c_1 \geq c_2 \geq ... \geq c_N$. Individual units are drawn to the sample one-by-one with equal probabilities until the sum of costs corresponding to drawn units exceeds some pre-determined survey budget C. More specifically, the procedure is stopped when the cumulative cost of the sample becomes greater or equal C and the population unit for which it occurs is not included in the sample. As a result, inclusion probabilities satisfy the simple order expressed by multiple inequality (4). Hence their estimates $\hat{p}_1, ..., \hat{p}_N$ may be obtained through PAVA. By inserting these estimates into formula (2) an estimator of the population total for the Pathak sampling scheme is obtained.

A simulation study was carried out in order to assess the properties of resulting H-T estimator for the population total and to compare it to the classic empirical H-T statistic involving inclusion probability estimates computed through Fattorini's formula (3). In simulation experiments, a sampling frame corresponding to the finite population under study was represented by the data set obtained during agricultural census carried out by Polish Central Statistical Office (GUS) in 1996. The dataset described population of 695 farms in the Gręboszów municipality of the Dąbrowa Tarnowska district. Total yearly sales of a farm represented the variable under study for which the population total was to be estimated. It was also assumed that the cost of observing this variable for individual farms was proportional to the farm area, assumed to be known. It was assumed that $C = 0.05 \cdot (c_1 + ... + c_N)$ so the survey budget was equal to five percent of the census cost. As a result sample size could vary in the range between 10 and 102 units depending on the cost of sampled units. The simulation study was designed to jointly capture the variability of estimates resulting from both sampling of finite population and a random simulation study. It was carried out by sequentially generating independent sets of replications for each sample drawn from the finite population and computing corresponding empirical H-T estimates. All computations were made in the R computing environment [12].

In the first experiment $R = 300$ replications were drawn for each of 15000 Pathak samples. Histograms of empirical distributions for both empirical H-T estimates are shown in the Figure 1. The dashed vertical line represents true value of the population total while the solid vertical line represents observed average of estimates. Moreover, observed characteristics of empirical distributions for both empirical H-T estimators are listed in the Table 1. They include their biases, relative biases (Relbias) computed as a ratio of bias to the true estimated value, variances, mean square errors (MSE), relative root mean square errors (RRMSE) and ratios of squared bias to the total mean square error (Bias share). Distributions of both estimators feature a very similar shape, with slight positive skew and nearly the same variance. However the distribution of Fattorini's estimator is substantially shifted to the left which is reflected by its strong

negative bias. The bias of the PAVA-based estimator is positive, but its absolute value is several times lower. This advantage in terms of absolute bias also influences the overall accuracy of estimates. The mean square error of the PAVA-based estimator is 17 % lower than that of Fattorini's statistic.

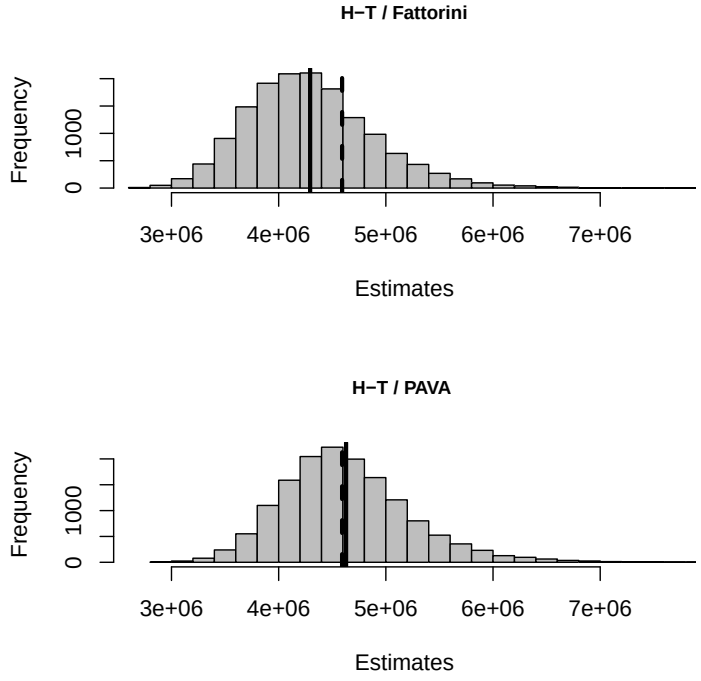

Figure 1: Distribution of estimates for Fattorini's and PAVA-based empirical H-T statistic

	Bias	**Relbias**	**Variance**	**MSE**	**RRMSE**	**Bias share**
Fattorini	-298242.8	-0.0649	$3.592 \cdot 10^{11}$	$4.482 \cdot 10^{11}$	0.1457	0.1984
PAVA-based	37207.4	0.0081	$3.702 \cdot 10^{11}$	$3.716 \cdot 10^{11}$	0.1327	0.0037

Table 1: Selected distribution characteristics of both empirical H-T estimators

In the second experiment, the investigation was extended to compare the behavior of estimators for $R = 100, 200, ..., 1000$ replications. For each value of R a total of 10000 samples were drawn using Pathak scheme, with corresponding set of R replications again generated independently for each sample. The absolute bias, relative bias, relative root mean square error and the share of bias in the MSE for varying values of R are shown in the Figure 2. It turns out that the bias of the proposed PAVA-based estimator is very stable for small numbers of replications, while it slowly tends to zero when R grows. For Fattorini's estimator this tendency was more pronounced, but absolute values of bias were 13 to 3.46 times higher reaching nearly 20 percent for $R = 100$. The relative root mean square error of both estimators exhibits similar behavior. For Fattorini's statistic it is always higher than for the PAVA-based one, with the relative difference reaching 58 percent for $R = 100$ but quickly diminishing when R grows. For any value of R the share of bias in the mean square error did not exceed one percent for the PAVA-based estimator while it reached over 70 percent for the Fattorini's statistic and $R = 100$.

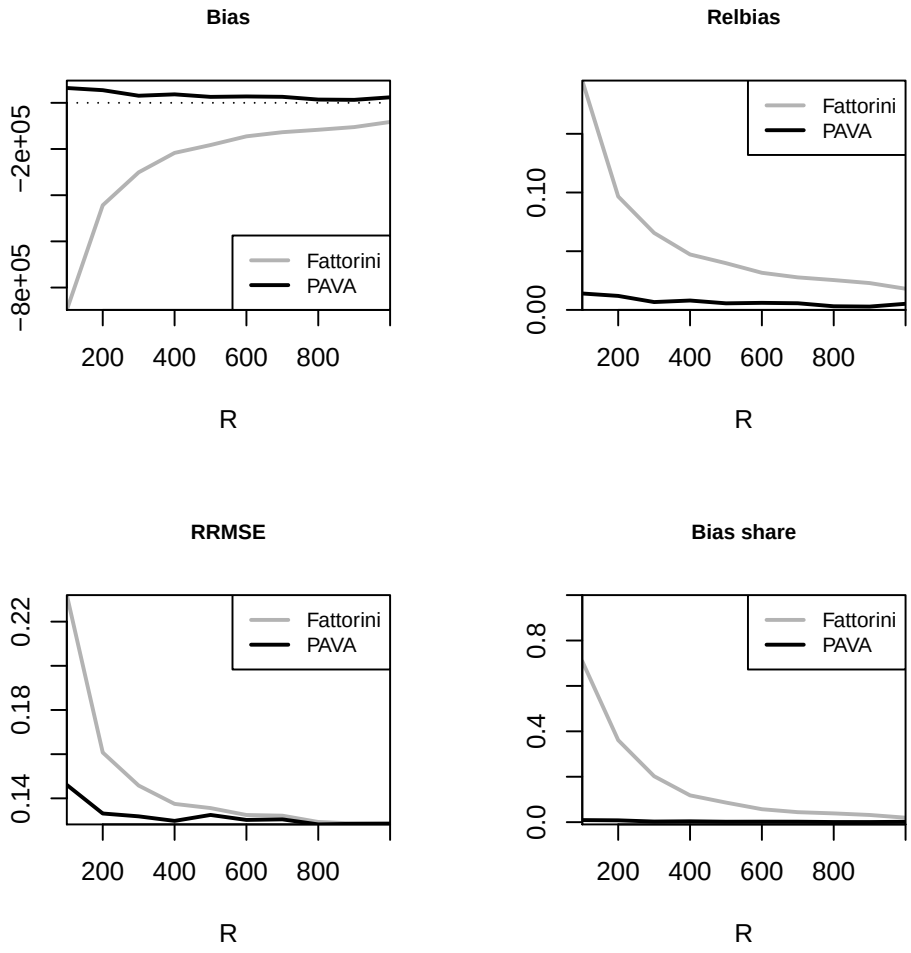

Figure 2: Stochastic properties of Fattorini's and PAVA-based empirical H-T statistic for varying R

4 Conclusions

The proposed estimator indirectly utilizes observations of sample membership indicators associated with all population units to increase the accuracy of first-order inclusion probability estimates corresponding to sampled units. Presented simulation study suggests, that this in turn significantly reduces the bias and improves the accuracy of empirical H-T estimator itself, even for quite large numbers of replications. One may reasonably expect that the strength borrowing effect should be particularly beneficial when differences between true individual inclusion probabilities are small, the population size is large and when generation of sample replications is time-consuming.

Acknowledgements

The work was supported by the grant No N N111 558540 from the Ministry of Science and Higher Education

References

[1] Ayer, M., and Brunk, H.D., and Ewing, G.M., and Reid, W.T., and Silverman, E.,: An empirical Distribution function for Sampling with Incomplete Information, *The Annals of Mathematical Statistics* **6** (1955), 641-647.

[2] Barabesi, L., and Fattorini, L., and Ridolfi, G.: Two-phase surveys of elusive populations. In *Proceedings of the Statistics Canada Symposium 97: New Direction in Surveys and Censuses*, Statistics Canada, Ottawa 1997, 285-287.

[3] de Leeuw, J., and Hornik, K., and Mair, P.: Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods, *Journal of Statistical Software* **32** (2009), 1-24.

[4] Fattorini, L., and Ridolfi, G.: A sampling design for areal units based on spatial variability, *Metron* **55** (1997): 59-72.

[5] Fattorini, L.: Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities, *Biometrika* **93** (2006), 269-278.

[6] Fattorini, L.: An adaptive algorithm for estimating inclusion probabilities and performing the Horvitz-Thompson criterion in complex designs, *Computational Statistics* **24** (2009), 623-639.

[7] Gamrot, W.: On Pool-Adjacent-Violators Algorithm and its Performance for Non-Independent Variables, To appear in *Studia Ekonomiczne* (2012).

[8] Härdle, W.,: *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1992.

[9] Horvitz, D.G., and Thompson, D.J.,: A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association* **47** (1952), 663-685.

[10] Myrskylä, M.: Generalized Regression Estimation for Domain Class Frequencies. Research report, Statistics Finland, Helsinki 2007.

[11] Pathak, P.K.,: Unbiased estimation in fixed-cost sequential sampling schemes, *Annals of Statistics* **4** (1976), 1012-1017.

[12] R Development Core Team: A language and environment for statistical computing (2011), R Foundation for Statistical Computing, Vienna.

[13] Rosén, B.: On sampling with probability proportional to size, *Journal of Statistical Planning and Inference*, **62** (1997), 159-191.

[14] Stylianou, M., and Proschan, M., and Flournoy, N.,: Estimating the probability of toxicity at the target dose following an up-and-down design, *Statistics in Medicine* **22** (2003), 535-543.

[15] Thompson, M.E., and Wu, C.: Simulation-based randomized systematic PPS sampling under substitution of units, *Survey Methodology* **34** (2008), 3-10.

[16] Wolny-Dominiak, A.: The multi-level factors in insurance rating technique. In: *Proceedings of 27th international conference "Mathematical Methods in Economics 2009"* (Brožová, H. ed.), Czech University of Life Sciences, Prague 2009, 346-351.