Wojciech Gamrot
Uniwersytet Ekonomiczny w Katowicach

ON KERNEL SMOOTHING AND HORVITZ-THOMPSON ESTIMATION

Introduction

Design-based estimation of population parameters usually relies on the knowledge of inclusion probabilities characterizing the sampling scheme. These are needed to construct sampling weights that form the well-known Horvitz-Thompson estimator of the population total and estimates for other parameters of interest. Sometimes, the complexity of sampling scheme prevents the exact calculation of inclusion probabilities. Such a situation arises for example for spatial sampling schemes (Fattorini and Ridolfi, 1997) some order sampling schemes (Rosen, 1997; Aires, 2000) as well as in fixed-cost sequential sampling where the composition of the sample depends on individual costs of sampled units (Pathak, 1976; Kremers, 1985).

The lack of exact inclusion probabilities does not necessarily render the Horvitz-Thompson statistic useless, since the statistician still possesses the knowledge of the sampling procedure used to draw the sample. When all the information needed to carry out sampling is readily available (such as: auxiliary variable values, unit sampling costs, adjacency matrix in spatial sampling), Fattorini (2006) proposes to conduct a simulation study and to estimate unknown inclusion probabilities, by drawing large numbers of sample replications and then counting appearances of individual units. By replacing unknown inclusion probabilities with estimates an alternative statistic known as empirical Horvitz--Thompson estimator is obtained.

Estimation of inclusion probabilities by simple sample proportions (or some statistics functionally dependent on it) usually requires large numbers of sample replications to achieve desired accuracy of Horvitz-Thompson estimates. Hence it appears reasonable to employ some form of strength-borrowing to capitalize on available auxiliary information and to improve accuracy of the simulation-based Horvitz-Thompson statistic. In this paper a nonparametric strength-borrowing technique is proposed for sampling schemes where first order inclusion probabilities satisfy simple ordering constraints. The fixed-cost sequential sampling scheme of Pathak (1976) is used as an example.

1. Estimators

Let the finite population be represented as a set of unit indices U={1, ... N}. Also, let y_1 , ... , y_N represent fixed values of some characteristic of interest and let $t = \sum_{i \in U} y_i$ be the population total to be estimated. An unordered sample s is drawn from U using some sampling scheme characterized by a set of first-order inclusion probabilities π_1, ... ,π_N where $\pi_i = P(i \in s)$ for $i \in U$. If inclusion probabilities were known, a design-unbiased Horvitz-Thompson estimator for t would be easily calculated from s according to the formula:

$$\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1}$$

When inclusion probabilities are impossible to calculate exactly, one may use the known sampling scheme to generate M independent sample replications s_1, ... ,$s_M \subseteq U$. For $i \in U$ let

$$k_i = \#\{r \in 1, \ \ldots \ , M : i \in s_r\} \tag{2}$$

be the number of replications containing the i-th unit. A very simple estimate of π_i is the sample proportion:

$$\hat{\pi}_i = \frac{k_i}{M} \text{ for } i \in s \tag{3}$$

However, when plugged into the formula (1) in place of π_i it could lead to division by zero if $k_i = 0$ for some $i \in s$. Such an event would require the i-th unit not to be drawn at all to any replication and is extremely unlikely for large M, but formally it prevents moments of the Horvitz-Thompson statistic from being computed. Hence, Fattorini (2006) proposes to estimate the inclusion probability π_i by the statistic:

$$\hat{\pi}_{iF} = \frac{k_i + 1}{M + 1} \text{ for } i \in s \tag{4}$$

and to estimate the population total t through the estimator:

$$\hat{t}_F = \sum_{i \in s} \frac{y_i}{\hat{\pi}_{iF}} \tag{5}$$

He derives an exact formula for its bias and a tight upper bound for the mean square error. However, as noted by the same author, the number of replications needed to guarantee high accuracy of this statistic may still be very large. This justifies efforts aimed at finding an alternative method of estimating π_i. Let

us notice, that during the simulation experiment involving generation of M replications, one may calculate estimates of inclusion probabilities not only for units in the sample s, but in fact for all N population units at negligible additional cost. Hence, any known relationships between individual inclusion probabilities corresponding to units included in s and units not included in s may be utilized to improve accuracy of estimates. In particular, such relationships may take the form of multiple inequality:

$$\pi_1 \leq \pi_2 \leq \ ... \ \leq \pi_N \tag{6}$$

As a simple example one may consider the well-known Pareto sampling scheme of Rosén (1997). By arranging population units in non-decreasing order with respect to known auxiliary variable on which the Pareto sampling is based one may easily guarantee that first-order inclusion probabilities characterizing this scheme satisfy the multiple inequality above. Gamrot (2012) proposed to incorporate the ordering constraint into empirical Horvitz-Thompson framework by calculating restricted estimates of inclusion probabilities satisfying (6) using isotonic regression algorithms such as Pool-Adjacent-Violators Algorithm (PAVA) or active set methods (see: Ayer et al., 1955; Robertson et al., 1988; Best and Chakravarti, 1990) and then by replacing unknown probabilities in (1) with these restricted estimates. However, isotonic regression only corrects for the breaches of ordering constraint (6) but it produces estimates equivalent to respective sample proportion when ordering is not violated. Hence properties of PAVA-based estimates should differ only slightly from sample proportions for larger replication numbers where such violations are rare. We will now propose another method that may be less prone to this unwelcome effect.

Let us start by noting that by definition we have $\pi_i \in [0,1]$ for $i \in U$. When N is large the ordering constraint (6) implies that either for all pairs (π_i, π_{i+1}) the difference $\pi_{i+1}-\pi_i$ is relatively small, or at least that the number of pairs where this difference is relatively large is itself not large. This leads to the intuition that for large N a particular inclusion probability π_i corresponding to the i-th population unit is unlikely to differ much from inclusion probabilities for its closest neighbors. Hence, combining probability estimates for inclusion probabilities of neighboring units may lead to better precision than using simple sample proportion.

A kernel estimator originally proposed by Rosenblatt (1956) appears to be a convenient way of forming a combined estimate of any individual inclusion probability in the population. For our purposes it is constructed as a weighted mean of simple proportions using the formula (see: Kulczycki, 2005; Härdle, 1992):

$$\hat{\pi}_{iK} = \frac{\sum_{j=1...N} \hat{\pi}_j w_{ij}}{\sum_{j=1...N} w_{ij}} \tag{7}$$

with

$$w_{ij} = \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) \tag{8}$$

where K(.) represents a certain non-negative symmetric real function having weak global maximum at 0 (so that K(x)=K(-x) and K(0)≥K(x) for x∈R) which is usually called a *kernel function* while h is a positive real constant known as *smoothing factor* or *bandwidth*. The symbol x_i represents for i∈U the value of some auxiliary characteristic of the i-th population unit. It is natural to intuitively assume it to be the unit index so that $x_i = i$ for i∈U. Another more interesting possibility of choosing x_i is discussed in the next section. Ultimately, the non-parametric empirical Horvitz-Thompson estimator of the population total is calculated according to the formula:

$$\hat{t}_K = \sum_{i \in s} \frac{y_i}{\hat{\pi}_{iK}} \tag{9}$$

Kulczycki (2005) argues, that the choice of a particular kernel function influences the accuracy of the kernel estimator (7) much less than the choice of bandwidth. In applications associated with sample surveys the normal kernel given by the formula:

$$K(x) = \exp\left(-\frac{x^2}{2} \right) \tag{10}$$

seems to be particularly popular (Giommi, 1987). From our perspective it is important that (10) always takes strictly positive values. As a result, all the terms w_{ij} in the linear combination (7) are strictly positive. Meanwhile, if the sampling scheme never produces empty samples (which may be safely assumed to be true), then at least one population unit belongs to some replication and consequently at least one of simple proportions $\hat{\pi}_1, \ldots, \hat{\pi}_N$ is strictly positive. This means that all kernel estimators $\hat{\pi}_{1K}, \ldots, \hat{\pi}_{NK}$ always take strictly positive (although possibly very small) values. Such an effect guarantees the finiteness of the Horvitz-Thompson statistic itself, and hence may be considered an advantage. In the following discussion it will be assumed that the normal formula (10) is used as a kernel.

As a general side note, it should also be stated that the proposed nonparametric estimator does not guarantee the constraint (6) to be satisfied. Although the likelihood of violating this restriction by individual estimates is apparently lower than for simple proportions computed through (3), such violations may still happen relatively often. Having said that one should keep in mind that the constraint (6) was discussed here only in order to motivate and justify the use of kernel smoothing, and was not meant to be strictly imposed.

In the following sections the proposed estimator (9) will be compared to other alternatives for a specific sampling design.

2. Application to fixed-cost sampling

Let us consider the fixed-cost sequential sampling scheme of Pathak (1976). It is characterized by varying inclusion probabilities which are generally difficult to calculate for larger sample sizes due to the combinatorial explosion (Schuster, 2000). Despite the existence of some sufficiency-based design-unbiased estimators which do not utilize inclusion probabilities, the empirical Horvitz-Thompson estimators may be of interest when nonresponse corrections need to be incorporated or when some modifications are made to the original scheme. In this paper the Pathak's scheme in its original form illustrates the use of nonparametric empirical Horvitz-Thompson approach. The sampling procedure is carried out as follows. Let $c_1, \ldots, c_N$ denote known per-unit costs of observing the characteristic under study for individual population units. Population units are drawn to the sample one-by-one without replacement and with equal probabilities until the total cumulative cost of the sample is greater or equal to some budget constraint C fixed in advance. The element for which this happens is not appended to the sample. The sample size is random in general, but instead the variability of random sample cost is largely limited.

Meanwhile, it may be shown that inclusion probabilities of the first order – although hard to compute – constitute a non-increasing function of the per-unit cost, so that:

$$\underset{i,j\in U}{\forall} c_i < c_j \Rightarrow \pi_i \geq \pi_j \tag{11}$$

and

$$\underset{i,j\in U}{\forall} c_i = c_j \Rightarrow \pi_i = \pi_j \tag{12}$$

Consequently, by arranging population units in a non-increasing order with respect to individual unit cost one may easily guarantee that inclusion probabilities satisfy the ordering constraint (6). This suggests that for most population units their inclusion probabilities should not differ dramatically from those having similar cost. This in turn justifies the use of nonparametric empirical Horvitz-Thompson estimator (9) for the population total, with costs $c_1, \ldots, c_N$ treated as auxiliary variables $x_1, \ldots, x_N$ in (8).

3. A simulation study

A simulation study was carried out in order to compare performance of the proposed non-parametric empirical Horvitz-Thompson estimator (9), the PAVA-based estimator proposed by Gamrot (2012) and the classic Fattorini's (2006) statistic (5). In experiments, the finite population was represented by the data set describing 695 farms in the Gręboszów municipality of the Dąbrowa Tarnowska district obtained during the agricultural census conducted by Polish Central Statistical Office in 1996. It was assumed that the cost of sampling individual units is strictly proportional to the farm area, which featured high positive skew and that the budget constraint C is equal to five percent of the total cost of exhaustively enumerating the whole population.

The simulation experiment accounted for two sources of randomness, namely the randomness of the actual sample s, and the randomness of inclusion probability estimates. It was carried out by drawing 20000 samples and executing an independent simulation study involving 300 sample replications for each such sample to arrive at population total estimates. Figure 1 shows the observed relative bias (RBIAS) of kernel-based estimates for h = 0.2, 0.4, ... , 30. Figure 2 shows the observed relative root mean square error (RRMSE) of kernel-based population total estimates for h = 0.2, 0.4, ... , 30. The corresponding levels of RRMSE's for PAVA-based Horvitz-Thompson estimator and for Fattorini's statistic are also shown in the Figure 2.

The relative bias of the proposed estimator exhibits rather complex behavior. For very small h it takes values very close to zero, but quite unstably fluctuating between positive and negative values. With growing h at first it also quickly grows, reaching 0.00537 for h = 4.2 but then it steadily decreases to reach 0.00010 for h=17.6 to finally slowly increase again for h>17.6. The biases of PAVA-based estimator and Fattorini's statistic do not depend on h and they are respectively equal to 0.00801 and –0.06470 with the absolute value of the latter obviously the greatest of all for any h. Hence one may conclude that for any h = 0.2, 0.4, ... , 30 the proposed estimator clearly dominated the other two by a wide margin in terms of bias.

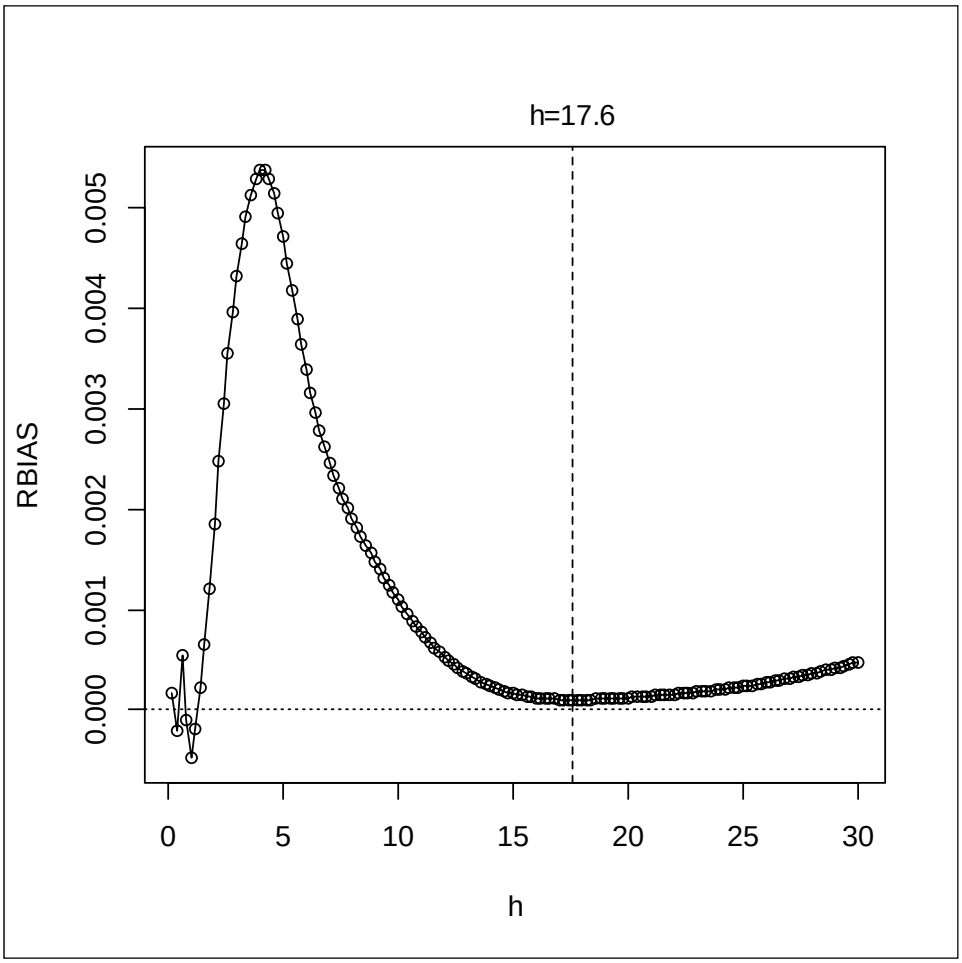

Figure 1. Relative bias of the proposed estimator for h = 0.2, 0.4, ... , 30.

The relative root mean square error of the proposed estimator also exhibited rather complicated behavior, reflecting to some extent the tendencies in the bias. It took the maximum value of 0.13877 for h = 0.1, but also featured two local minima around h = 1.2 and h = 15.8. For h = 15.8 it was equal to 0.12896 which is respectively about 12% and 3% lower than RRMSE's of PAVA-based estimator and Fattorini's statistic.

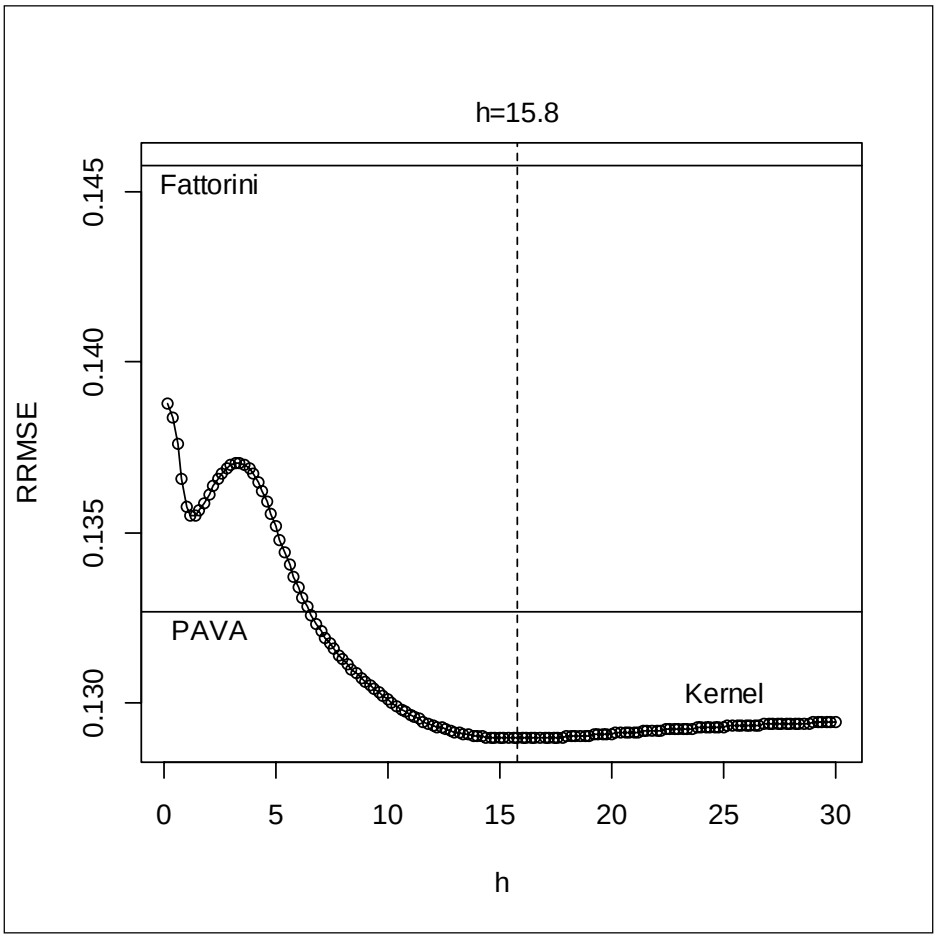

Figure 2. Relative root mean square errors of three population total estimators for h = 0.2, 0.4, ... , 30

Conclusion

Presented simulation results suggest that proposed nonparametric empirical Horvitz-Thompson estimator of the population total constitutes an attractive alternative to its two counterparts, especially in terms of bias reduction. The main challenge for it to gain a wider popularity most likely lies in choosing an optimal value for the smoothing factor h. In our study it could easily be chosen through simulation on the basis of known values of the characteristic under study in the whole population. In practice of the field work the statistician does not possess such information and would have to resort to using cross-validation or the plug-in method of Sheather and Jones (1991). Nevertheless the wide range of h-values for which the proposed estimator dominates its counterparts in terms of bias and mean square error seems to justify such approach.

References

Aires N. (2000): *Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs*, Phd thesis, Chalmers, Göteborg University, Göteborg.

Ayer M., Brunk H.D., Ewing G.M., Reid W.T., Silverman E. (1955): *An Empirical Distribution Function for Sampling with Incomplete Information*. *The Annals of Mathematical Statistics* 6(4), s. 641-647.

Best M.J., Chakravarti N. (1990): *Active Set Algorithms for Isotonic Regression*. A Unifying Framework. *Mathematical Programming* 47, s. 425-439.

Fattorini L., Ridolfi G. (1997): *A Sampling Design for Areal Units Based on Spatial Variability*. *Metron* 55, s. 59-72.

Fattorini L. (2006): *Appling the Horvitz-Thompson Criterion in Complex Designs: A Computer-Intensive Perspective for Estimating Inclusion Probabilities*. „Biometrica", 93(2), s. 269-278.

Gamrot W. (2012) Simulation-Assisted Horvitz-Thompson Statistic and Isotonic Regression. *Proceedings of the 30th International Conference on Mathematical Methods in Economics 2012* (accepted).

Giommi A. (1987): *Nonparametric Methods for Estimating Individual Response Probabilities*. „Survey Methodology", Vol. 13, No. 2, s. 127-134.

Härdle W. (1992): *Applied Nonparametric Regression*. Cambridge University Press.

Kulczycki P. (2005): *Estymatory jądrowe w analizie systemowej*. WNT, Warszawa.

Kremers W.K. (1985): *The Statistical Analysis of Sum-Quota Sampling*. Unpublished PHD thesis. Cornell University.

Pathak K. (1976): *Unbiased Estimation in Fixed-Cost Sequential Sampling Schemes*. „Annals of Statistics", 4 (5), s. 1012-1017.

Robertson T., Wright F.T., Dykstra R.L. (1988): *Order Restricted Statistical Inference*. Wiley, New York.

Rosén B. (1997): *On Sampling with Probability Proportional to Size*. „Journal of Statistical Planning and Inference", 62, s. 159-191.

Rosenblatt M. (1956): *Remarks on Some Nonparametric Estimates for the Density Function*. „Annals of Mathematical Statistics", No. 27, s. 832-837.

Schuster P. (2000): *Taming Combinatorial Explosion*. Proceedings of the National Academy of Sciences of the United States of America, 97 (14), s. 7678-7680.

Sheather S.J., Jones M.C. (1991): *A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation*. „Journal of the Royal Statistical Society", B, 53(3), s. 683-690.

ON KERNEL SMOOTHING AND HORVITZ-THOMPSON ESTIMATION

Summary

Estimation of the total value of fixed characteristic of interest in a finite population is considered for a complex sampling scheme featuring unknown inclusion probabilities. The general empirical Horvitz-Thompson statistic is adopted as an estimator for the unknown total. In the presence of additional knowledge on inclusion probabilities taking form of inequality constraints it is proposed to use the well-known kernel estimator for individual inclusion probabilities. For a fixed-cost sequential sampling scheme this leads to a new nonparametric empirical Horvitz-Thompson estimator of a total. Its properties are compared to known alternatives in a simulation study.