M A T H E M A T I C A L E C O N O M I C S

No. 7(14) 2011

ON SOME MODIFICATION OF THE SUM-QUOTA SAMPLING SCHEME

Wojciech Gamrot

Abstract. A simple extension of the well-known sequential sum-quota sampling scheme is proposed. The modification facilitates the simultaneous examination of several sampling units between checks of the cost limit. This may speed up the data gathering process while some degree of control over the variable sample cost is still retained. It is proposed to estimate the population total of the studied characteristic under such a sampling scheme using empirical estimates of inclusion probabilities evaluated in a simulation study. It appears that at least in some situations the empirical Horvitz-Thompson estimator is approximately unbiased.

Keywords: fixed-cost sampling, sum-quota sampling, empirical inclusion probabilities.

JEL Classification: C83.

1. Introduction

The costs of carrying out the sample survey are usually divided into two broad categories, namely: *fixed costs* and *variable costs* (Groves, 1989, pp. 51). Fixed costs do not depend on the particular sample which is actually drawn while variable costs do depend on it. Variable costs arise due to various circumstances associated with the survey, such as the geographical localization of sample units, the extent of effort needed to gather the data from individual units or the effect of interviewer training, and they are generally harder to control. As an example let us consider the situation where the population under study consists of 100 units u_1, ..., u_{100} such that per unit cost of gathering information is equal to 5 for u_1, ..., u_{10}, and it is equal to 2 for the remaining units. If the sample of ten elements is drawn from such a population using simple random sampling without replacement (SRSWOR), then the expected cost of examining all sample units is proportional to the average of per-unit cost in the whole population and so it equals $10 \cdot 2.3 = 23$. However, a much less lucky sample containing units u_1, ..., u_{10}

Wojciech Gamrot
Department of Statistics, University of Economics in Katowice, Bogucicka Street 14, 40-226 Katowice, Poland.
E-mail: wojciech.gamrot@ue.katowice.pl

yields the total cost of 50, which is over two times greater than the expected cost. Obviously, such an extreme event is unlikely. However, even the theoretical possibility of a dramatic increase in the cost may render the analysis based on the assumption of equal inclusion probabilities questionable when the institution carrying out the survey does not have the capacity to handle such unexpected costs and would be forced to modify the sample or cancel the survey completely.

The well known sum-quota sampling scheme (Pathak, 1976; Kremers, 1985, 1986; Lehmann, Cassella, 1998) is dedicated to controlling the survey cost in unfavorable situations as described above. The scheme does not rely on the requirement of fixed sample size. Instead it is attempted to keep the total sample cost constrained. This is achieved by sequentially drawing individual units one by one to the sample with equal probabilities until the total cost of the sample reaches some predefined limit. Hence the total cost of the sample is guaranteed not to exceed the fixed budget of the survey. A significant drawback of the sum-quota procedure manifests itself when per-unit sampling costs are not known in advance. Then the procedure requires examining all units sequentially and to test if the cost limit has not been breached before every examination. This practically prevents the simultaneous examination of more than one unit. Hence the survey may become very time-consuming and its organization (e.g. division of field work) problematic. In this paper a modification of the sum-quota sampling scheme is considered. It makes possible the simultaneous examination of more than one population unit while still maintaining some control over the total cost of the sample.

2. Sum-quota sampling scheme and its extension

Let the finite population of size N be represented by a set of unit indices $U = \{1, ..., N\}$. Let $c_1, ..., c_N$ represent individual per-unit costs of examining individual population units. Let some pre-determined maximum cost limit L be given. The original sum-quota sampling procedure of Pathak (1976) is carried out by drawing units randomly until for some – say M-th – unit the total cost of M units already drawn is greater or equals L. The M-th unit is discarded and preceding M-1 units form a sample. The joint cost of the sample is guaranteed not to exceed L. The sample size is random, and inclusion probabilities may in general vary, so the sample is non-simple. The inclusions of any two distinct units in the sample are not necessarily inde-

pendent events. Let us now propose the following extension of this original procedure:

- *K*-sets of units are sequentially drawn from *U* using simple random sampling without replacement.
- Each drawn *k*-set is removed from the population so that *i*-th *k*-set is drawn from among just $N–k$ (i-1) units (i = 1, 2, ...). If N-k (i-1) < k for some i, then the whole population is taken as a sample (a trivial special case).
- The procedure stops when the cumulative cost of all the elements in all drawn *k*-sets exceeds *L*.
- The last *k*-set for which the limit *L* is exceeded is included in the sample.

Denote the sample obtained this way by *s*. The proposed procedure differs from Pathak's algorithm in two ways. Firstly, it allows for the simultaneous examination of *k* units so that testing if the cost limit is breached may be done less frequently. Secondly, the last *k*-set for which the cost limit is breached is included in the sample, which eliminates the need to assess the cost of examining all units in the current *k*-set before they are sampled. The proposed procedure differs from the *k*-finite populations sampling scheme considered by Kremers, Robson (1987) because all the units within the *k*-set are examined and any particular unit may be drawn within many *k*-sets (sampling within *k*-sets is not independent).

As a result of the introduced modifications the guarantee not to exceed the cost limit exactly is lost. However some control over the random total cost of the sample *s* given by:

$$C = \sum_{i \in s} c_i$$

is retained. Let $c_{(1)}$, ..., $c_{(N)}$ be a sequence of individual costs sorted in decreasing order. The maximum excess $\Delta = C - L$ of sample cost over the limit *L* may be computed as:

$$\Delta_{\max} = \sum_{i=1,\ldots,k} c_{(i)}$$

Moreover, since *C* is not lower than *L*, and consequently $C \in [L, L + \Delta_{\max}]$, the variance of *C* is not greater than:

$$V_{\max}(C) = \frac{\Delta_{\max}^2}{4}.$$

When the population is large and its distribution not extremely skewed, both Δ_{max} and $V_{max}(C)$ may be reasonably small in comparison to the expected cost. It may also be possible to estimate both on the basis of external knowledge of the survey subject, and to set the limit L in such a way that sufficient funds are available to examine any possible sample. The first-order inclusion probabilities defined as:

$$\pi_i = P(i \in s) = \sum_{s \ni i} P(s)$$

for $i = 1, ..., N$ and characterizing the proposed procedure may vary. Their exact calculation from the definition is prevented by the combinatorial explosion effect even for rather modest sample sizes. The following example presents their estimates for a certain specific situation.

Example 1. Let $N = 40$ and $c_i = i$ for $i = 1, ..., 40$. Thus, per unit costs are distributed quite uniformly on the (1, 40) interval. Let $L = 80$. A total of 100000 samples were drawn independently using both sum-quota procedure and the modified sum-quota algorithm with $k = 2$. The observed frequencies of each population unit appearing in the sample that may be treated as estimates for first-order inclusion probabilities are shown in Figure 1. Somewhat surprisingly, inclusion probabilities for a modified scheme differ very modestly from each other and even tend to grow with increasing per-unit cost. This is in contrast with the original sum-quota sampling scheme where they clearly decrease when the per-unit cost increases. Let us also define a vector of sample membership indicators $I = [I_1, ..., I_N]$ in such a way that

$$I_i = \begin{cases} 0 & \text{for} \quad i \in s \\ 1 & \text{for} \quad i \notin s \end{cases}$$

for $i = 1, ..., N$. Its observed correlation matrix for the original scheme as well as for the modified scheme with $k = 2$ are shown in Figure 2. It appears that population units which are relatively cheap to examine corresponding sample membership indicators are positively correlated while for relatively expensive units correlation tends to be negative. This effect manifests for both schemes, and the correlation appears to be rather weak.

In general for both schemes the sample size may vary. However its expectation which is equal to the sum of all first order inclusion probabilities is clearly greater for the modified scheme. This is not surprising since the modified scheme allows for drawing between one and k extra units above the cost limit L. Distributions of sample size are compared in the following example involving real data.

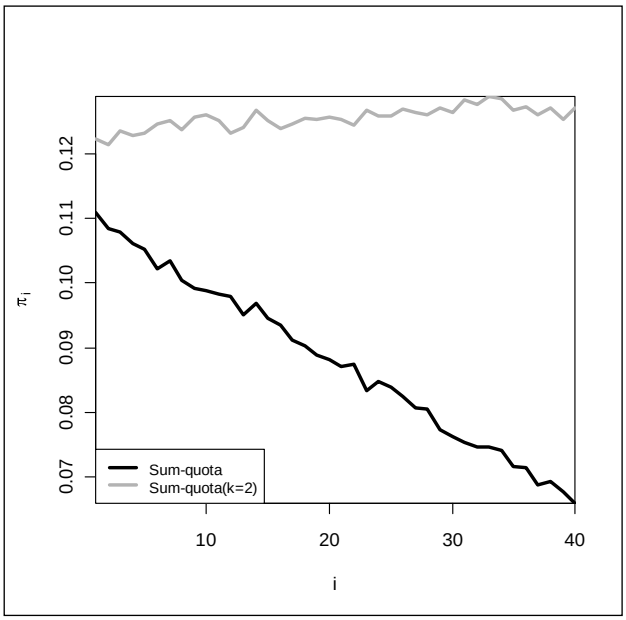

Fig. 1. First order inclusion probabilities characterizing sum-quota sampling and modified sum-quota sampling scheme with $k = 2$ for the same cost limit

Source: author's own work.

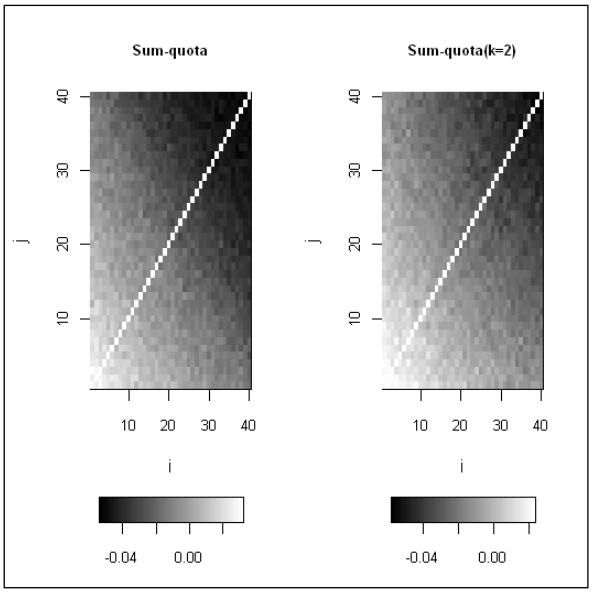

Fig. 2. Correlations $c_{ij} = \mathrm{cor}(I_i, I_j)$ between sample membership indicators for sum-quota sampling and modified sum-quota sampling scheme with $k = 2$ for the same cost limit

Source: author's own work.

Example 2. Consider the data obtained in the agricultural census carried out in 1996 by the Polish Central Statistical Office (GUS). The data describes a population of 695 farms in the Gręboszów borough. Assume that the cost for obtaining data from any individual farm is roughly proportional to its area (this might be justified in the case of a survey involving the assessment of geo-botanic assets, veterinary inspection of the cattle or some specialized examination of crops). The histogram of the farm area in the whole population is shown in Figure 3. The farm area exhibits a strong positive skew with a mean equal to 734.05 and a median equal to 651. Denote $G = c_1 + ... + c_N$. A total of 10000 samples were drawn for $L = 0.1 \cdot G$ using original sum quota sampling scheme and a modified sum-quota sampling with $k = 2$. The distribution of sample cost relative to G computed as $\lambda = C/G$ is shown in Figure 4.

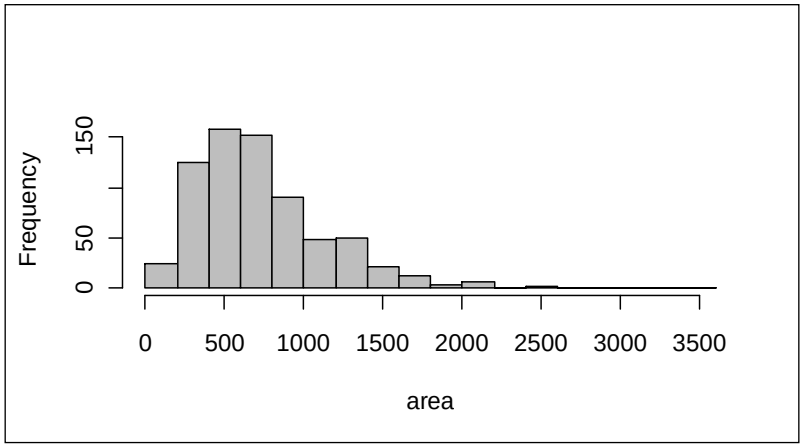

Fig. 3. Distribution of farm area

Source: author's own study.

As expected, the observed values of λ for the original sum-quota sampling scheme never exceed the desired value 0.1 corresponding to the limit L, while for the modified scheme they always do. However, absolute deviations (in plus or in minus) from L never exceed 10% and in most cases they do not exceed 5% (for original scheme 99.83% of absolute deviations did not exceed 5%, for the modified scheme 98.12% of absolute deviations did not exceed 5%). The maximum possible deviation of $\Delta_{max} = 57163$ corresponding to $\lambda = 0.1120$ was never observed for the modified scheme. Hence one might conclude that both sampling schemes provide a reasonable degree of control over sampling costs.

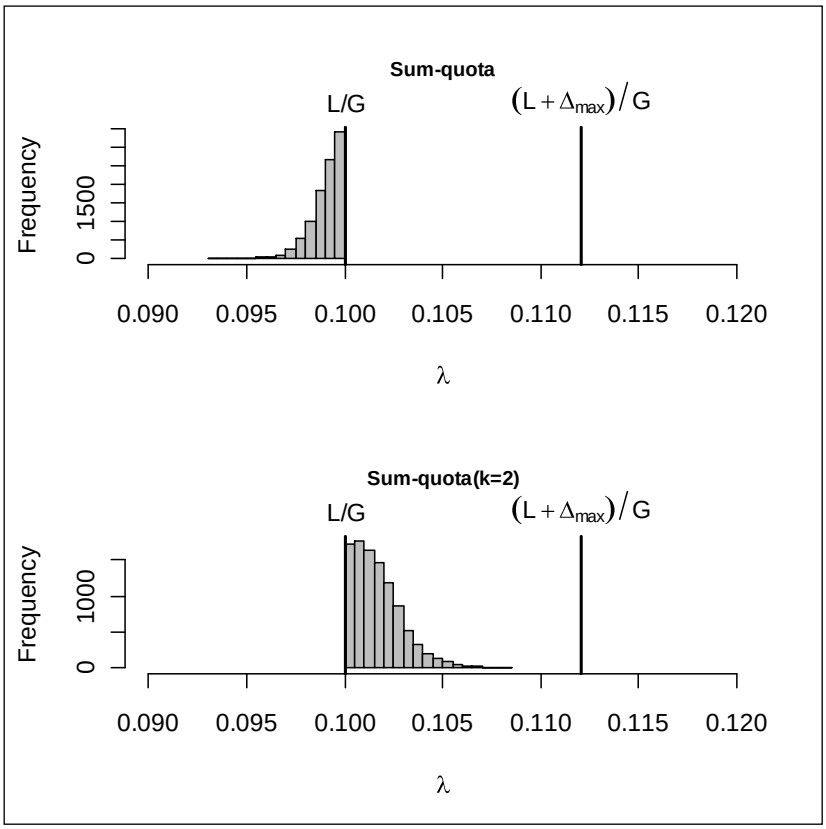

Fig. 4. Distribution of sample cost under sum-quota sampling and modified sum-quota sampling scheme with $k = 2$ for the same cost limit

Source: author's own study.

3. Estimation based on empirical inclusion probabilities

Let $y_1, ..., y_N$ be fixed values of some population characteristic of interest. If first-order inclusion probabilities were known, one would easily estimate the population total $t = y_1 + ... + y_N$ without design bias using the well-known Horvitz-Thompson statistic:

$$\hat{t}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

When unknown, these probabilities may be estimated from a simulation experiment involving massive numbers of independently drawn sample replications $s_1, ..., s_H$. Let X_i represent the number of times i-th unit is

included in H sample replications. It was shown by Fattorini (2006) that the statistic:

$$\hat{\pi}_i = \frac{X_i + 1}{H + 1}$$

is a consistent estimator for π_i. It may be used to form an empirical Horvitz--Thompson estimator for t in the form:

$$\hat{t}_{EHT} = \sum_{i \in s} \frac{y_i}{\hat{\pi}_i}.$$

The above statistic always takes finite values since $\hat{\pi}_i > 0$ for $i = 1, ..., N$ and it is asymptotically unbiased for the population total t (in terms of H growing to infinity for all inclusion probability estimates).

4. Simulation study

To shed some light on the properties of $\hat{t}_{EHT}$, a simulation study was carried out using the population from Example 2. A total of 180000 samples were drawn using modified sum-quota sampling scheme with $k = 2$, and $L = \lambda G$ with $\lambda = 0.1, 0.2, 0.3, 0.4$. For each sample simulation, experiments were independently executed for $r = 100, 200, ... ,500$ replications. Hence, $4 \cdot 5 \cdot 180000 = 36 \cdot 10^5$ sample replications were drawn in the whole study. Relative bias and relative root mean square error of empirical Horvitz--Thompson estimates for the variable representing total farm sales are shown in Figure 5.

The observed relative bias seems to behave in a very irregular way with no systematic tendency. This is because bias values are very small in comparison to the limited accuracy of this simulation study. The absolute value of the observed relative bias was never greater than $5 \cdot 10^{-4}$, which means that the bias was extremely small in comparison to the estimated parameter. The mean square error (MSE) of estimates decreases slowly when number H of the sample replications grows, and seems to stabilize around some positive value, dependent on the cost limit L. By itself, the cost limit seems to influence the MSE more strongly (the greater the cost limit, the lower the MSE). The share of squared bias in the MSE never exceeded $3.5 \cdot 10^{-5}$, which indicates that the bias was negligible in comparison to the MSE.

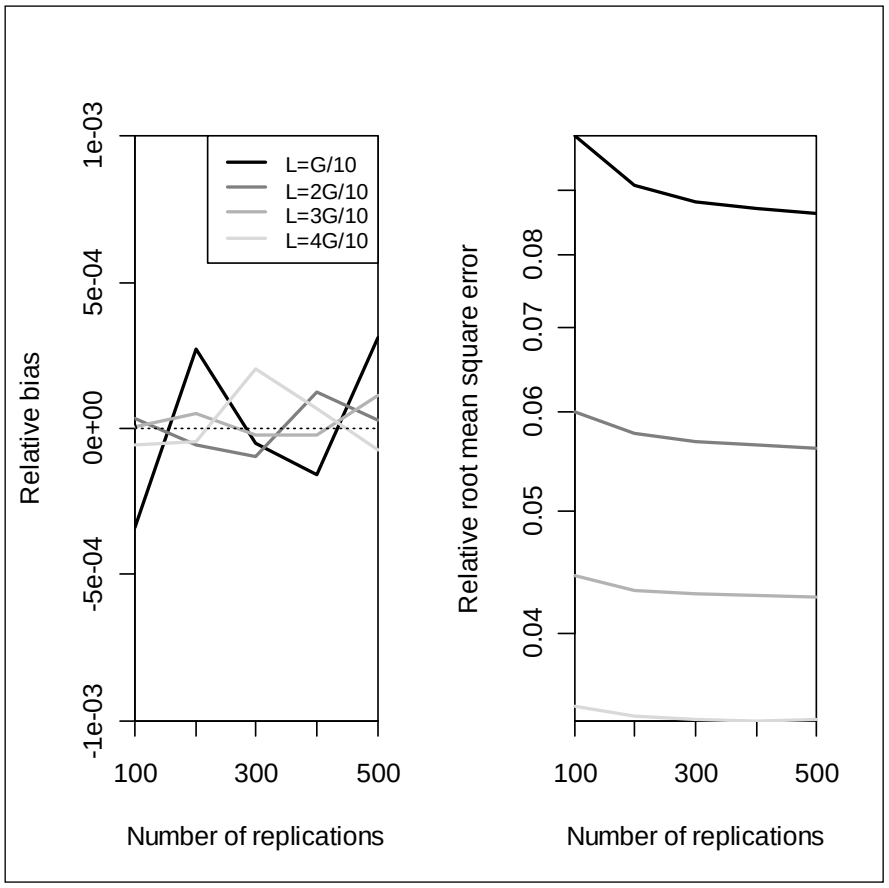

Fig. 5. Relative bias and relative root mean square error for empirical Horvitz-Thompson estimator under modified sum-quota sampling scheme with $k = 2$

Source: author's own study.

5. Conclusions

The proposed simple modification of the sum-quota sampling scheme facilitates a simultaneous examination of several units and may lead to a serious speeding-up of the data gathering process when individual per-unit costs of data acquisition are not known in advance. On the other hand, some degree of control over the total variable cost of the survey is still retained. The proposed method for estimating the finite population total through the Horvitz-Thompson formula using Fattorini's empirical estimates of inclusion probabilities seems to provide an approximate design-unbiasedness at least in some situations, apart from asymptotic considerations.

Acknowledgement

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

Literature

Fattorini L. (2006). *Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities*. Biometrika. Vol. 93(2). Pp. 269-278.

Groves R.M. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons. New York.

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

Kremers W.K. (1986). *Completeness and unbiased estimation for sum-quota sampling*. Journal of the American Statistical Association. Vol. 81(396). Pp. 1070-1073.

Kremers W.K., Robson D. (1987). *Unbiased estimation when sampling from renewal processes: The single sample and k-sample random means cases*. Biometrika. Vol. 74(2). Pp. 329-336.

Lehmann E.L., Casella G. (1998). *Theory of Point Estimation*. Springer. New York.

Pathak K. (1976). *Unbiased estimation in fixed-cost sequential sampling schemes*. Annals of Statistics. Vol. 4(5). Pp. 1012-1017.