ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 235, 2010

Wojciech Gamrot[*]

ESTIMATION OF A QUANTILE USING PARETO SAMPLING SCHEME

Abstract. One of most common methods of utilizing available auxiliary information to improve stochastic properties of estimates for simple population parameters such as population total or population mean relies on drawing population units to the sample with individual inclusion probabilities proportional to known values of auxiliary variable. This leads to the construction of various non-simple sampling schemes. This paper focuses on properties of population quantile estimates when Pareto sampling scheme is used. Simulation results are presented.

Key words: non-simple samples, Pareto scheme, quantile estimation.

I. INTRODUCTION

Consider a finite population U of size N. Let Y be some characteristic taking fixed values $y_1,...,y_N$ for population units. The aim of the survey is to estimate various population parameters including the population total $t_y = \sum_{i\in U} y_i$, the population mean $\overline{Y} = t_y / N$ or the population quantile of the order p (or p-quantile):

$$Q_p(U) = (1-g)y_{(k)}+gy_{(k+1)} \qquad (1)$$

where k=[r], g = r-[r], $r=N\cdot p+0.5$ while $y_{(1)},...,y_{(N)}$ denote values of Y in U arranged in increasing order and the symbol [.] represents rounding down to the nearest integer (Gilchrist 2000). A random sample s of size n is drawn from U according to some general sampling design p(s) determining inclusion probabilities of the first order π_i for $i\in U$ and inclusion probabilities of the second order π_{ij} for $i\neq j\in U$. The population total may then be estimated without bias by the well-known Horvitz-Thompson estimator $\hat{t}_y = \sum_{i\in s}\frac{y_i}{\pi_i}$. For fixed-size sampling designs its variance may be expressed by the formula (Yates and Grundy (1953)):

[*] Ph.D. Department of Statistics, University of Economics, Katowice.

$$V(\hat{t}_y) = -\frac{1}{2}\sum_{i,j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{2}$$

This obviously suggests that the variance will be reduced if first order inclusion probabilities are proportional (or approximately proportional) to Y. And hence the values of Y are unknown, it is often attempted to set π_i's proportional to some other, observed characteristic X taking fixed values $x_1,...,x_N$ for all population units. At the same time it is necessary to guarantee the fixed sample size in order for the expression (2) to be valid (otherwise another term associated with sample size variability would appear in the variance formula). Several sampling schemes were constructed to simultaneously satisfy both these requirements, including the scheme of Lahiri-Midzuno, Hartley-Rao, Rao-Hartley-Cochran or Sunter (Bracha (1996)). In this paper a Pareto sampling scheme is considered. It aims to achieve first-order inclusion probabilities equal to $\pi_i = nx_i\left(\sum_{j\in U} x_j\right)^{-1}$ for $i\in U$. If this desired inclusion probability is greater than one for some population unit then it is set to one and inclusion probabilities for remaining units are recomputed accordingly (this is repeated for other units if necessary). The sampling procedure is executed in two steps. In the fist step a pseudo-random variable u_i following the uniform distribution on the <0,1> interval is generated for each population unit. Then the following expression is evaluated for each population unit:

$$q_i = \frac{u_i(1-\pi_i)}{\pi_i(1-u_i)} \tag{3}$$

In the second step n population units having lowest values of q_i are selected to the sample s. This guarantees constant sample size. It has been shown by Rosen (1997) that such a procedure also provides true first order inclusion probabilities approximately equal to desired ones. Let us also note that exact inclusion probabilities of the first and second order may be computed according to the procedure considered by Aires (2000).

II. QUANTILE ESTIMATION FOR NON-SIMPLE SAMPLES

For a simple sample an often used estimator of the population quantile is its sample analog, computed according to the formula:

$$Q_p(s) = (1 - g')y'_{(k')} + g' \cdot y'_{(k'+1)} \tag{4}$$

where $k' = [r']$, $g' = r' - [r']$, $r' = n \cdot p + 0.5$ while $y'_{(1)},..., y'_{(n)}$ denote values of Y in s arranged in an increasing order (Hyndman and Fan (1996)). However, Pareto sampling leads to underrepresentation of units with relatively low Y-values in the sample which distorts the distribution of the sample quantile. A positive bias is introduced that does not diminish with growing sample size. Hence, for non-simple sample we consider another estimator of the p-quantile, constructed following the general re-weighting approach outlined by Särndal et al (1992). It takes the form:

$$\hat{Q}_{p*}(s) = y'_{(k)} + \frac{p - W(k)}{W(k+1) - W(k)}(y'_{(k+1)} - y'_{(k)}) \tag{5}$$

where

$$W(k) = \sum_{i=1}^{k-1} w_{(i)} + 0.5 \cdot w_{(k)} \tag{6}$$

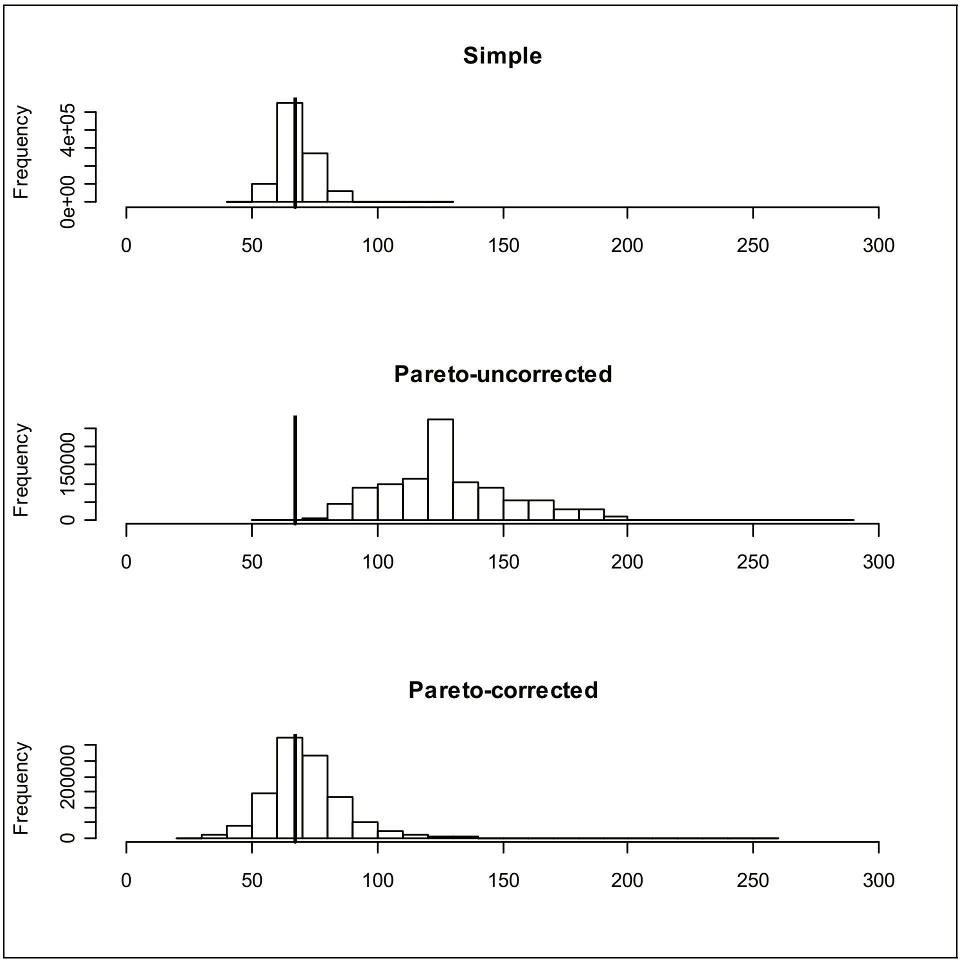
and W(k)≤p<W(k+1) while $w_{(1)},...,w_{(n)}$ represents an arranged in increasing order sequence of weights $w_1,..,w_n$ corresponding to respective sample units and computed as:

$$w_i = \left(\pi_i \sum_{j \in s} \pi_j^{-1} \right)^{-1}$$

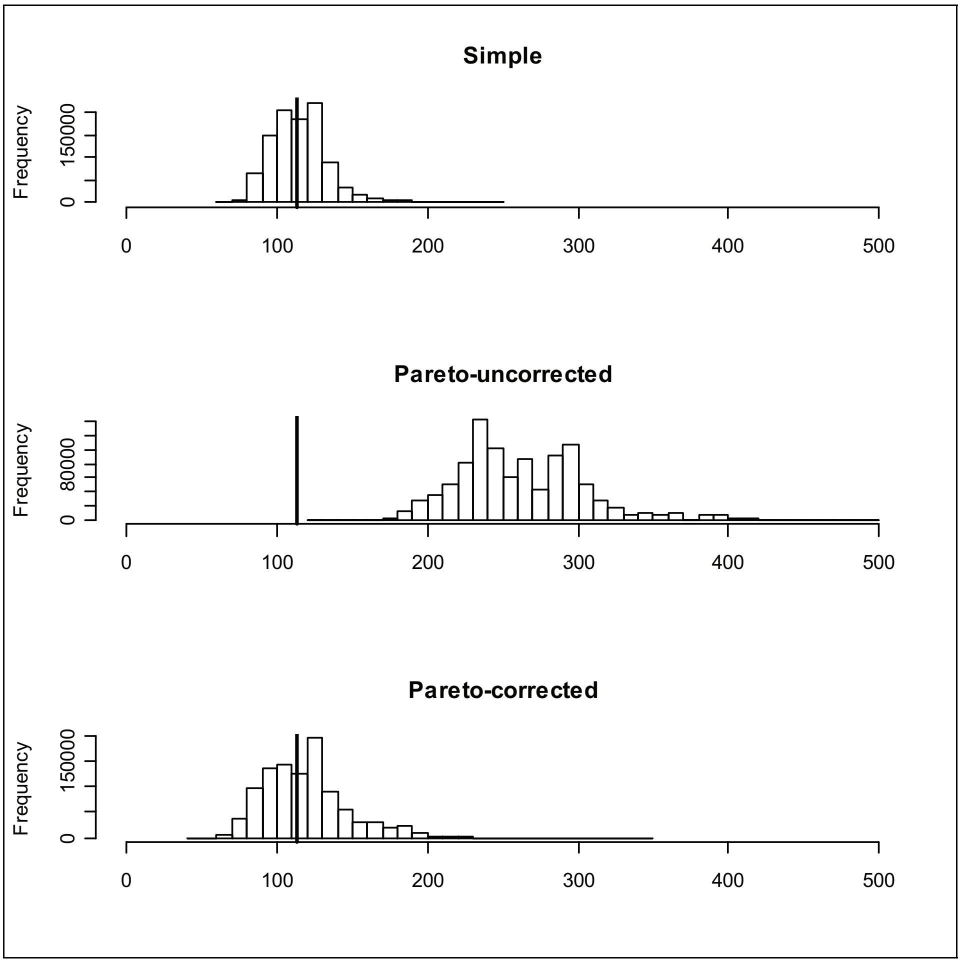
For simple sample we have $w_i=n^{-1}$ for $i \in U$ and this estimator is equivalent to (4). For non simple samples these estimators are generally not equivalent.

III. SIMULATION STUDY

A simulation study was carried out to compare empirical distributions of quantile estimators for simple random sampling without replacement and non simple samples drawn using Pareto scheme. The well-known RMT284 population published by Särndal et al (1992) and characterizing Swedish municipalities represented the population under study. The variable under study Y was the revenue from municipal taxation in 1985 (variable RMT85) while real estate values in the year 1984 (variable REV84) were used as an auxiliary variable X. Histograms illustrating the empirical distribution of 10^6 estimates of the p-quantile for p=0.25 and p=0.5 are respectively shown on pic. 1. and 2. Black vertical lines on all these graphs represent true values of the estimated quantile.



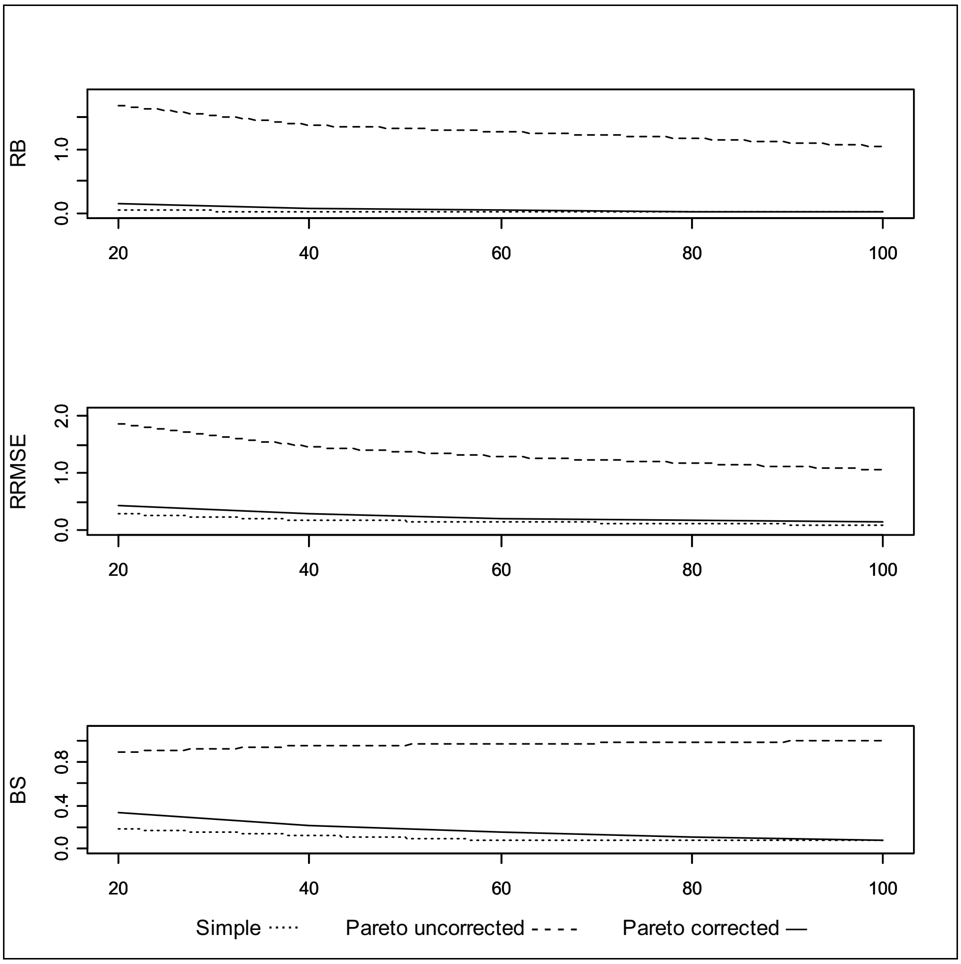
Pic.1 Empirical distributions of quantile estimators for p=0.25



Pic.2 Empirical distributions of quantile estimators for p=0.5

The graphs above show the extent of distortion in the quantile estimates when uncorrected estimator is used with Pareto sampling scheme. It is also evident, that the correction introduced in the estimator $\hat{Q}_{p*}(s)$ substantially reduces this unwelcome effect. It is worth noting that the empirical distribution of estimates is quite irregular despite large number of simulated sample replications – this is due to the irregular structure of the whole population.

The dependency of the relative bias (RB), relative root mean square error (RRMSE) and the share of bias in overall mean square error (BS) is illustrated on pic. 3.



Pic. 3. Relative bias (RB), relative root mean square error (RRMSE) and the share of bias in the total mean square error (BS) of estimators as a function of sample size n.

For the corrected estimator under Pareto sampling all the observed parameters: RB, RRMSE and BS decrease with growing n and take lower values than their equivalents for the uncorrected estimator. The latter is also characterized by increasing bias share when n grows. Anyway, the simple random sampling with replacement provided the most attractive properties of estimators for any value of n, although its advantage over the corrected estimator was modest or (for large samples) negligible.

IV. CONCLUSIONS

Simulation results clearly confirm that under Pareto sampling the corrected estimator of the p-quantile has better properties than the uncorrected one and should be preferred. The observed relative bias of the corrected estimator and its share in the MSE apparently tend to zero with growing sample size which suggests consistency. However, Pareto sampling did not lead to any gains in terms of precision when compared to simple random sampling. Instead, some modest loss of precision was observed. The considered estimator does not utilize explicitly any auxiliary variables. However, the application of such data could lead to construction of better quantile estimators based on model-based approach considered for example by Chambers and Dunstan (1986), Rao, Kovar and Mantel (1990), or Rueda et al. (2004)

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.

Bracha Cz. (1996) *Teoretyczne podstawy metody reprezentacyjnej* Wydawnictwo Naukowe PWN Warszawa.

Chambers, R., Dunstan, R., (1986). Estimating Distribution Functions from Survey Data. Biometrika, 73 (3): 597–604.

Gilchrist, W.G. (2000): Statistical Modelling with Quantile Functions. Chapman&Hall/CRC, Boca Raton.

Hyndman R.J., Fan Y, (1996) Sample Quantiles in Statistical Packages, *The American Statistician*, 50(4), 361–365.

Rao, J. N. K., Kovar J.G., Mantel, H. J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information, Biometrika 77(2), 365–375.

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

Rueda M. M., Arcos A., Martínez-Miranda M. D., Román Y. (2004), Some improved estimators of finite population quantile using auxiliary information in sample surveys, Computational Statistics & Data Analysis, 45(4), 825–848.

Särndal C.E., Swensson B., Wretman J.H. (1992) *Model Assisted Survey Sampling* Springer-Verlag New York.

Yates F. Grundy P.M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* B 15, 235–261.

Wojciech Gamrot

ESTYMACJA KWANTYLI Z WYKORZYSTANIEM SCHEMATU LOSOWANIA PARETO

Jedną z popularnych metod wykorzystywania dostępnych informacji o wartościach cech pomocniczych do poprawy dokładności oszacowań wartości globalnej lub średniej w populacji jest losowanie prób z prawdopodobieństwami inkluzji pierwszego rzędu proporcjonalnymi do wartości cechy pomocniczej. Podejście takie prowadzi do konstrukcji rozmaitych schematów losowania, taich jak schemat Lahiriego-Midzuno, Hartleya-Rao, Rao-Harleya-Cochrana, Suntera, czy też Pareto. W niniejszym artykule zbadano empirycznie, jak zastosowanie ostatniego z wymienionych schematów losowania próby wpłynie na własności stochastyczne uzyskiwanych oszacowań innego parametru, a mianowicie kwantyla.