IIPS SAMPLING VERSUS STRATIFIED SAMPLING — COMPARISON OF EFFICIENCY IN AGRICULTURAL SURVEYS

Marcin Kozak¹, Robert Wieczorkowski²

ABSTRACT

The paper contains considerations on efficiency of three sampling schemes, i.e. simple random sampling without replacement, stratified sampling, and πps sampling, in an application to large agricultural populations. A simulation study using data from Agricultural Census 2002 was carried out to make a comparison of the schemes. Stratification of the populations was carried out using a method of Rivest (2002), and Hartley-Rao scheme with randomly permutated units was chosen as the most popular πps sampling scheme.

Key words: agriculture surveys, πps sampling, stratified sampling.

1. Introduction

An optimal sampling design is the one in which the first order inclusion probabilities \( \pi_j \), i.e. the probability of sampling a \( j \)-th \( j = 1, ..., N \) unit, are proportional to values of a survey variable \( Y \), i.e. \( \pi_j \propto Y_j, j = 1, ..., N \) (Särndal et al., 1992, pp. 88—89, Bracha 1996, p. 39). In such a case a variance of a mean or total estimator would be zero. Certainly, in practice we cannot evaluate the inclusion probabilities \( \pi_j \), for we do not know the values of the variable under study \( Y \) before a survey. Although, if we know values of some auxiliary variable \( X \) which is strongly correlated with the survey one, we can evaluate the probabilities \( \pi_j \) which are proportional to the values of this auxiliary variable,
i.e. \( \pi_j \propto X_j, j = 1, \ldots, N \). It should lead to getting the small variance of the considered estimator; the variance depends on both the mentioned correlation coefficient and certainly on a sample.

The above procedure should be more efficient than other sampling designs, e.g. a stratified sampling which is the most often used scheme in agricultural surveys in Central Statistical Office of Poland. Unfortunately, the \( \pi ps \) schemes are quite difficult to realize in practice, for several reasons. First of all, they are often very difficult in implementation; although, in several packages, e.g. SAS (SAS Institute, 1999), R (R Development Core Team, 2003), etc., these schemes are available. Secondly, to have a possibility of estimation a variation of a considered estimator (see e.g. Särndal et al., 1992, section 2.8), we have to know the joint inclusion probabilities, i.e. \( \pi_{ij}, i, j = 1, \ldots, N, i \neq j \). In most \( \pi ps \) schemes we do not have such opportunity; just several of them enables calculating these probabilities, e.g. Hanurav-Vijayan (Hanurav, 1967, and Vijayan, 1968), or Sampford (1967) methods. Unfortunately there are no direct formulas on the joint inclusion probabilities, just the approximate ones, for the most popular \( \pi ps \) scheme, i.e. Hartley-Rao (1962). There is a possibility of evaluating the \( \pi_{ij} \) in the Sunter’s (1977, 1986) and Rao-Hartleya-Cochran (1962) schemes, but in these schemes the first order probabilities for the units are not strictly proportional to their value.

There are several alternative methods that enable approximate variance estimation in complex surveys; they might be applicable in a case of \( \pi ps \) sampling. Wolter (1985) lists such methods: a method of random groups, variance estimation based on balanced half-samples, a jackknife method, generalized variance functions, and a Taylor series method. Probably a bootstrap method gives the biggest future respects; unfortunately, its theory regarding the application in the \( \pi ps \) sampling is still in its infancy and needs improving.

Moreover, the \( \pi ps \) sampling is usually used in a two-stage sampling scheme; it is rather not used in a case of one-stage surveys. In the two-stage surveys the frame of primary sampling units is rather small; it can consist of hundreds, maybe thousands units. In agricultural surveys we use one-stage sampling; it is, certainly, more efficient, but a frame consists of even a few hundreds of thousands units. Is the \( \pi ps \) sampling efficient enough in such a case?

An additional question is as follows. Is the \( \pi ps \) sampling more efficient than other schemes, mostly stratified sampling? The stratified sampling scheme proved its usefulness in many surveys, in a contrary to the \( \pi ps \) sampling. Theory points that the latter should be more efficient; is so indeed? The answer on such question should be sought against a background of a different correlation between the auxiliary and survey variable.
An aim of the paper is to give the answer on such question. A comparison of the schemes is based on a simulation study based on data from the National Agricultural Census 2002.

2. Simulation study

Let us consider three populations of farms from Łódzkie, Mazowieckie, and Wielkopolskie voivodships. A cereals area is the variable under study $X$; we know values of it in all farms. Let frames (in voivodships) consist of farms which satisfies three conditions, i.e. their land area is bigger than 2 ha, and the cereals area is non-zero. Moreover, let us exclude the farms which do not satisfy a condition $nX_{j}\left(\sum_{j=1}^{N}X_{j}\right)^{-1}<1$, where $n$ is a fixed sample size, and $X_{j}$ is a value of the stratification variable in a $j$-th unit (farm) of the population; it is a condition required in $\pi\pi\pi$ sampling – one cannot consider the units which sampling probability is bigger than one; (in practice they are taken to the sample with probability one). Afterwards, the frames consisted of 62,570 farms satisfying the above conditions in the Łódzkie voivodship, 100,252 farms in Mazowieckie voivodship, and 89,321 farms in Wielkopolskie voivodship. A sample size from each voivodship was fixed on $n=1000$ farms.

Given the fixed sample size, the population in each voivodship has been stratified. Ten strata have been created, and one of them was a “take-all” stratum; (the populations under study were strongly right-skewed, for details see Lednicki and Wieczorkowski, 2003). It should make us certain the stratification almost optimal; the bigger number of strata would not make the efficiency markedly better. The strata boundaries obtained in that way have been used in following stages of a study, i.e. in a drawing a sample and estimation.

Hartely-Rao scheme (1962) with randomly permutated units has been chosen as the representation of $\pi\pi\pi$ sampling schemes, for it is the most popular and often used from a midst of them (because of its simplicity).

The below procedure have been carried out independently in each voivodship. Four artificial survey variables have been generated in each voivodship. They were correlated with the stratification variable as follows: $r\in\{0.80, 0.85, 0.90, 0.95\}$. The variables under study in each voivodship were generated using the stratification variable in such a way that all characteristics of the variables $Y_{i}, i=1,...,4$, were the same as of the stratification variable $X$, and correlation coefficients $r_{XY_{i}}$ were as above; (for details of the algorithm of the initialization of the populations see Appendix 1). Such procedure resulted from the fact that the comparison of the efficiency of the sampling schemes should be carried out against a background of the similar populations; it excludes an influence of the characteristics of the populations on the efficiency.
For each survey variable in each voivodship a simulation study has been carried out independently. In each of a 5,000 iterations of the particular study three artificial surveys were carried out. First one was a simple random sampling without replacement (SI), second — stratified sampling (see Lednicki and Wieczorkowski, 2003) and third one a πps Hartley-Rao sampling. In such a way we have obtained thousand estimators for each sampling scheme. A variation of these results shows an efficiency of the considering sampling scheme. Moreover, a variance of the results can be treated as an approximate variance of the real variance of the estimator, and a precision, i.e. a coefficient of variation, of the results as an approximate precision of the estimator. Therefore, a following approximate coefficient of variation of the estimator from the particular sampling scheme were estimated

\[
\hat{c}_{ijv} = \frac{\hat{D}^2(\hat{c}_{ijv})}{t_v},
\]

where \(\hat{c}_{ijv}\) is the approximate precision (coefficient of variation) of the \(i\)-th estimator, \((i = 1,...,4)\), in the \(v\)-th voivodship, under \(j\)-th sampling scheme, \(\hat{D}^2(\hat{c}_{ijv})\) is the sample estimator of the variance of the estimator \(\hat{c}_{ijv}\), and \(t_v\) is the real value of the considered parameter; index \(j\) regards three types of the sampling scheme, i.e. \(j=1\) for SI, \(j=2\) for STSI, and \(j=2\) for πps sampling.

The coefficients (1) obtained in such a way are comparable; a big value of (1) implicates bad precision in the particular combination \(i \times j \times v\). The scheme with the smallest values of (1) from a group of the coefficients from the particular voivodship can be seen as the most efficient one.

Comparing of the schemes should take into account all values of the correlation coefficient between the stratification and survey variable. The best sampling scheme would be the one that is the most efficient in a case of all considered values of the correlation coefficient. Of course we should consider such values that can be met in practice; therefore we have chosen four mentioned values of the correlation coefficient.

The results for three studied voivodships are presented in table 1. As it was mentioned, the survey variables in particular voivodships differed only with the value of correlation coefficient with the stratification variable, so the results of simple random sampling should be similar for them; indeed, it has been recorded. The results showed univocally that the stratified sampling were markedly the best in three studied voivodships.

The biggest differences between the precisions of estimation were recorded in Wielkopolskie voivodship. In a case of the small value of the correlation coefficient between the stratification and survey variable, i.e. \(r = \{0.80, 0.85\}\), the efficiency of πps sampling were the worst (even worse than the SI sampling).
Also in Mazowieckie voivodship, in a case of $r=0.80$, the SI sampling was better than the $\pi ps$ one. It shows the obvious property that decreasing value of this coefficient makes the sampling schemes, which are based on the values of the stratification variable, less efficient; although, the decrease of the precision of estimation under the stratified sampling were smaller than in a case of $\pi ps$ sampling. In the rest voivodships the efficiency of the $\pi ps$ sampling were better than the efficiency of the SI sampling, but worse than the stratified sampling.

An obvious conclusion of the study is that the stratified sampling led to the best efficiency from among the midst of the studied sampling schemes.

Table 1. Estimated precision of estimation of total value obtained by using three sampling schemes in Voivodships against a background of various correlations between a stratification and survey variable

<table>
<thead>
<tr>
<th>Survey variable</th>
<th>$r_{XY_i}$ (1)</th>
<th>SI (2)</th>
<th>STSI (3)</th>
<th>$\pi ps$ (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Łódzkie Voivodship</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_1$</td>
<td>0.80</td>
<td>0.0662</td>
<td>0.0398</td>
<td>0.0546</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>0.85</td>
<td>0.0671</td>
<td>0.0348</td>
<td>0.0500</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>0.90</td>
<td>0.0674</td>
<td>0.0281</td>
<td>0.0430</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.95</td>
<td>0.0660</td>
<td>0.0201</td>
<td>0.0298</td>
</tr>
<tr>
<td>Mazowieckie Voivodship</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_1$</td>
<td>0.80</td>
<td>0.0716</td>
<td>0.0431</td>
<td>0.0867</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>0.85</td>
<td>0.0703</td>
<td>0.0370</td>
<td>0.0617</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>0.90</td>
<td>0.0708</td>
<td>0.0297</td>
<td>0.0427</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.95</td>
<td>0.0706</td>
<td>0.0222</td>
<td>0.0287</td>
</tr>
<tr>
<td>Wielkopolskie Voivodship</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Y_1$</td>
<td>0.80</td>
<td>0.0608</td>
<td>0.0369</td>
<td>0.0672</td>
</tr>
<tr>
<td>$Y_2$</td>
<td>0.85</td>
<td>0.0598</td>
<td>0.0339</td>
<td>0.0662</td>
</tr>
<tr>
<td>$Y_3$</td>
<td>0.90</td>
<td>0.0606</td>
<td>0.0276</td>
<td>0.0530</td>
</tr>
<tr>
<td>$Y_4$</td>
<td>0.95</td>
<td>0.0606</td>
<td>0.0189</td>
<td>0.0308</td>
</tr>
</tbody>
</table>

Source: own calculations based on CSO data from the Agricultural Census 2002

(1) $r_{XY_i}$ – correlation coefficient between the stratification variable $X$ and the survey variable $Y_i, i = 1,...,4$.

(2) SI – Simple Random Sampling without Replacement,

(3) STSI – Stratified Sampling with SI in strata.

(4) $\pi ps$ – Hartley-Rao (1962) without Replacement Proportional-to-Size Sampling (with random permutation of the population units)
3. Discussion and Conclusions

The paper contains a comparison of the efficiency of three sampling schemes, i.e. a simple random sampling without replacement (SI), stratified sampling (STSI), and without replacement proportional-to-size (πps) sampling. The comparison regards the agriculture populations of farms in some Polish voivodships (experiments have been carried out for three of them); therefore the conclusions can be treated as regarding the agricultural populations of big size; (although we think they can be widen to large populations similar to agricultural, e.g. business populations). A case of surveys regarding small populations should be considered in independent investigations.

The results showed that the stratified sampling scheme is more efficient than πps and SI sampling schemes when designing an agricultural survey in a case of a large population. The stratified sampling was the most accurate in all cases; certainly its superiority to the πps sampling results from a smaller susceptibility on the value of the correlation coefficient between the stratified and survey variable. Moreover, another advantage of the stratified sampling scheme is that it is easier to design a survey and analyze its results, including estimation under situation of occurring of non-responses.

Afterwards, the main conclusion of the paper is that we can recommend the stratified sampling as the most efficient and appropriate sampling scheme in the agricultural surveys conducted by the Central Statistical Office of Poland; in the contrary, the results of the πps sampling are not satisfying, so it rather should not be entered to the surveys.
APPENDIX 1

Consider a population $U$ consisting of $N$ units. For each unit of the population $U$ we know values of a normal variable $X$ with a population mean $\mu_X$ and variance $\sigma_X^2$. We want to create a variable $Y$ with the same population mean and variance, i.e. $\mu_Y = \mu_X = \mu$, $\sigma_Y^2 = \sigma_X^2 = \sigma^2$, subject to a fixed correlation coefficient between variables $X$ and $Y$ $\rho$. To make it one has to create such variable $Y$ that after sorting it would be the same as the original one ($X$); therefore, it is enough to replace the values of $X$ between some of the randomly chosen units of the population $U$. The correlation coefficient $\rho$ depends, first, on a number of the units with replaced values and, secondly and certainly, which units are sampled; therefore $\rho$ is a random variable. We should only control a probability of the replacement of the units.

An algorithm is easy. Fix the replacement probability, say $p_r$, taking into account $\rho$; ($p_r$ should increase for small $\rho$, and vice versa – it has to be chosen experimentally). For each unit repeat a following procedure – draw a random number, say $u$, from an interval $[0,1]$. If $p_r > u$ for the $i$-th unit, take $Y_i = X_i$, else draw one unit, say $j$-th, from the following units of the population, i.e. $X_{i+1}, \ldots, X_N$, and take $Y_i = X_j, Y_j = X_i$. Do not include the $j$-th unit in further steps of the procedure.
REFERENCES


