ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Wojciech Gamrot**

A TWO-PHASE SAMPLING STRATEGY FOR ESTIMATING MULTIPLE MEAN VALUES IN THE PRESENCE OF NONRESPONSE

Abstract. The phenomenon of nonresponse in sample surveys usually results in biased estimates of population characteristics. One of the means to deal with nonresponse is the subsampling technique. It relies on re-contacting some subset of nonrespondents by using more expensive and more efficient tools (e.g. direct interview) than those used in the first attempt to collect data. This allows to increase response rate and to obtain unbiased estimates of population characteristics. In this paper, the problem of establishing the sample and subsample sizes minimizing the expected cost of the survey, while achieving desired precision of multiple mean value estimates, is considered. An algorithm is proposed that allows to establish the optimum initial sample and subsample sizes for two-phase sampling strategy.

Key words: nonresponse, callbacks, subsampling, two-phase sampling.

1. INTRODUCTION – TWO-PHASE SAMPLING

One of the measures taken to deal with nonresponse in surveys is the callback technique. It relies on re-contacting nonrespondents and allows to increase response rate and, under some conditions, to obtain unbiased estimates of population characteristics. To reduce the survey cost, usually only a small fraction of nonrespondents from the initial sample is questionned in the next phases of the survey. The subset of re-contacted nonrespondents is called a subsample. Obviously, this technique is a special case of multiphase sampling. In this paper a two-phase sampling strategy for estimating multiple mean values of population characteristics is considered. A deterministic nonresponse is assumed.

Let us assume that a population Ω of the size N is divided into two nonoverlaping strata Ω_1 and Ω_2 whose sizes are equal to N_1 and N_2 respectively. All the units belonging to stratum Ω_1 would respond to all the questions, if contacted. All the units from the stratum Ω_2 would not

* PhD, Department od Statistics, University of Economics, Katowice.

respond to any question[1]. Let us also denote: $W_1 = N_1/N$, $W_2 = N_2/N$. The aim of the survey is to estimate mean values of k population characteristics. In the first phase of the survey a simple random sample s of the size n is drawn without replacement from the population Ω, according to the sampling design:

$$P_1(s) = \binom{N}{n}^{-1} \tag{1}$$

The sample s is partitioned into two disjoint sets $s_1 \subset \Omega_1$ and $s_2 \subset \Omega_2$ of the sizes $0 \leqslant n_{s_1} \leqslant n$ and $0 \leqslant n_{s_2} \leqslant n$ respectively and $s_1 \cup s_2 = s$, $s_1 \cap s_2 = \emptyset$, $n_{s_1} + n_{s_2} = 1$. The sizes n_{s_1} and n_{s_2} are random variables having the following hypergeometrical ditribution function:

$$P^*(n_{s_1}) = P^*(n_{s_2}) = \frac{\binom{N_1}{n_{s_1}}\binom{N_2}{n_{s_2}}}{\binom{N}{n}} \tag{2}$$

where:

$\max\{0,\ n-N_2\} \leqslant n_{s_1} \leqslant \min(n,\ N_1\}$,

$\max\{0,\ n-N_1\} \leqslant n_{s_2} \leqslant \min\{n,\ N_2\}$.

All the sampling units from the set s1 would respond to questions concerning all the characteristics, and all the sampling units from the set s_2 would not respond to any question. In the second phase of the survey a subsample u of the size n_u is drawn from among n_{s_2} units of the set s_2 with the probability of selection equal to:

$$P_2(u|n_{s_2}) = \binom{n_{s_2}}{n_{s_2}}^{-1} \tag{3}$$

It is assumed that all units in the sample u respond in the second attempt of contact, so the data is collected for the $n_{s_1} + n_u$ units, and, according to W. G. Cochran (1977), the following statistic is an unbiased estimator of mean value of the i-th population characteristic:

$$\bar{x}_w^{(i)} = \frac{1}{2}(n_{s_1}\bar{x}_{s_1}^{(i)} + n_{s_2}\bar{x}_u^{(i)}) \tag{4}$$

[1] This kind of response mechanism is commonly called "unit nonresponse".

where:

$\bar{x}_{s_1}^{(i)}$ – the mean of the characteristic under study in the set s_1,

$\bar{x}_u^{(i)}$ – the mean of the characteristic under study in the subsample u.

For fixed n and n_{s_2} the variance of the estimator described above is given by expression[2]:

$$V(\bar{x}_w^{(i)}) = \left(\frac{N-n}{Nn}\right)S_i^2 + \frac{n_{s_2}^2}{n^2}\left(\frac{1}{n_u} - \frac{1}{n_{s_2}}\right)S_{2i}^2 \tag{5}$$

where:

S_i^2 – the variance of the i-th characteristic in the population Ω,

S_{2i}^2 – the variance of the i-th characteristic in the stratum Ω_2.

Assume that the cost of observing all the population characteristics is given by the expression:

$$K = C_0 n + C_1 n_{s_1} + C_2 n_u \tag{6}$$

where:

C_0 – per-unit cost of making the first contact attempt.

C_1 – per-unit cost of processing data obtained during the first contact attempt.

C_2 – per-unit cost of getting and processing data from the second stratum.

The quantities n_{s_1}, n_{s_2} and n_u are random variables, so we will consider expected value of the cost given by expression[3]:

$$E(K) = C_0 n + C_1 E(n_{s_1}) + C_2 E(n_u) \tag{7}$$

2. SPECIAL CASE: LIMITED VARIANCE FOR SINGLE CHARACTERISTIC

Let us assume that desired precision V_{0i} should be achieved for the i-th population characteristic[4]. For fixed n and n_{s_2}, the minimum subsample size needed to obtain variance $V(\bar{x}_w^{(i)})$ not exceeding V_{0i} is given by:

[2] See e.g. W. G. Cochran (1977) or C. E. Sarndal, B. Swensson, J. Wretman (1992) for a proof.

[3] To simplify the notation, the symbol $E(\cdot)$ is used to denote the expectation over both sampling designs P_1 and P_2, so it is equivalent to $E_{P_1}(E_{P_2}(\cdot))$.

[4] At this moment we assume, that variances of all other characteristics are not limited. However all the characteristics are observed and the cost is still given by the expression (6).

$$n_{ui} = \frac{NS_{2i}^2 n_{s_2}^2}{n^2(NV_{0i} + S_i^2) - NnS_i^2 + NS_{2i}^2 n_{s_2}} = \frac{n_{s_2}^2}{n(na_i - b_i) + n_{s_2}} = \frac{n_{s_2}^2}{n\gamma_i(n) + n_{s_2}} \tag{8}$$

where:

$$a_i = \frac{NV_{0i} + S_i^2}{NS_{2i}^2} \tag{9}$$

$$b_i = \frac{S_i^2}{S_{2i}^2} \tag{10}$$

$$\gamma_i(n) = na_i - b_i \tag{11}$$

and initial sample size satisfies the condition:

$$n \geqslant n_i^*, \quad \text{where} \quad n_i^* = \frac{NS_i^2}{NV_{0i} + S_i^2} \tag{12}$$

For fixed n and n_{s_2}, the cost given by (6) grows with the subsample size n_u. Thus, given the variance limit V_{0i}, it is minimized, when subsample size is established by using expression (8). Assuming that the subsample size is given by (8), and that $E(n_{s_1}) = nW_1$, the expected cost of achieving the desired precision for the i-th characteristic is approximately[5] equal to:

$$E(K_{(i)} = f_{(i)}(n) = nC_0 + nC_1W_1 + \frac{NnW_2^2S_{2i}^2C_2}{n(NV_{0i} + S_i^2) - N(S_i^2 - W_2S_{2i}^2)} \tag{13}$$

and it reaches minimum for initial sample size equal to:

$$n_x^{(i)} = n_i^*\left(W_2\frac{S_{2i}^2}{S_i^2}(\alpha_i - 1) + 1\right) \tag{14}$$

where:

$$\alpha_i = \min\left(\beta_i, \frac{\mathrm{NV}_{0i}}{\mathrm{W}_2\mathrm{S}_{2i}^2} + 1\right) \tag{15}$$

$$\beta_i = \max\left(1, \sqrt{\frac{\mathrm{C}_2(\mathrm{S}_i^2 - \mathrm{W}_2\mathrm{S}_{2i}^2}{\mathrm{S}_{2i}^2(\mathrm{C}_0 + \mathrm{C}_1\mathrm{W}_1)}}\right) \tag{16}$$

[5] Under assumption, that W_2 is a sufficiently accurate estimate of nonrespondent fraction (n_{s_2}/n) in the sample.

3. GENERAL CASE: LIMITED VARIANCES FOR ALL THE CHARACTERISTICS

For fixed n and n_{s_2}, the minimum subsample size needed to achieve desired precision levels for all the k characteristics is equal to:

$$n_u = \max_{i,\ldots,k} \{n_{ui}\} \tag{17}$$

so it is equal to the quantity n_{uj}, where j is the number of the characteristic for which the expression $\gamma_j(n)$ takes the minimum value. Hence the result of the comparison of expressions $\gamma_i(n)$ does not depend on n_{s_2}, we may write

$$E(K) = f_{(j)}(n), \quad \text{where} \quad \gamma_j(n) = \min_{i=1,k} \{\gamma_i(n)\} \tag{18}$$

The expected cost of estimating all the k characteristics is equal to the expected cost of estimating this one characteristic, for which $\gamma_i(n)$ is minimal. Let us formulate the problem of establishing initial sample size n, minimizing the expected cost without violating the precision limits V_{0i} for all k characteristics. This can be expressed as:

$$\begin{cases} E(K) \to \min \\ V(\bar{x}_w^{(i)}) \leqslant V_{0i} \quad \text{for} \quad i = 1, \ldots, k \\ 2 \leqslant n \leqslant N \end{cases} \tag{19}$$

To solve the problem stated above, we determine intervals of the initial sample size[6] n, inside which $\gamma_i(n)$ yields the minimum value for the same characteristic. Hence $\gamma_i(n)$ is a linear function of n, the bounds of intervals can easily be established by finding the values of n, for which some of the expressions $\gamma_i(n)$ are equal to each other. In each interval corresponding to some j-th characteristic we find the value n_{xj} minimizing the cost[7], under assumption that only the variance of this single characteristic is limited by V_{0j}. By comparing the values of minimum expected cost evaluated in each interval we choose the optimum initial sample size, and the corresponding value of expected cost $E(K)$.

[6] At this point we assume, that the initial sample size n is a real number, not an integer.

[7] The optimum value is obtained by evaluating the expression (14). If the result of evaluation falls outside the appropriate interval, it is assumed to be equal to the corresponding bound of this interval because $f_{(j)}(n)$ is a convex function of n.

The optimum initial sample size may be evaluated according to the algorithm presented below. The variables m_0 and m_1 are used to denote lower and upper bound of current interval.

1. Assign the value $n^*_{\min}$ to the variable m_0.

2. Denote by the index j the characteristic for which the expression $\gamma_i(m_0)$ yields the minimum value. If there exists more than one characteristic having this property, the one with minimal ai value should be chosen. The characteristics for which $\gamma_i(m_0) \geqslant \gamma_j(m_0)$ and $a_i > a_j$ may be eliminated from further considerations.

3. If all the characteristics except the j-th one were eliminated, assume $m_1 = N$ and go to the step 5.

4. For each not eliminated characteristic evaluate the expression:

$$n_i = \frac{b_i - b_j}{a_i - a_j} \tag{20}$$

Assume $m_1 = \min\{n_i\}$. If $m_1 > N$ assume $m_1 = N$.

5. Evaluate the initial sample size n_x minimizing the expected cost, under assumption that only the variance of j-th characteristic is limited.

$$n_x = n_x^{(j)} = n_j^* \left(W_2 \frac{S_{2j}^2}{S_j^2} (\alpha_j - 1) + 1 \right) \tag{21}$$

6. If $n_x < m_0$, assume $n_x = m_0$.

7. If $n_x > m_1$ assume $n_x = m_1$.

8. Evaluate minimum expected cost corresponding to n_x, according to expression:

$$E(K|n_x) = n_x(C_0 + C_1 W_1) + \frac{N n_x W_2^2 S_{2j}^2 C_2}{n_x(N V_{0j} + S_j^2) + N(W_2 S_{2j}^2 - S_j^2)} \tag{22}$$

9. In the first iteration, store the values n_x and $E(K|n_x)$ obtained in steps 7–8. In the successive iterations, compare the cost evaluated in step 8 with previously stored value of cost and if it is greater, then store the values n_x and $E(K|n_x)$ obtained in steps 7–8.

10. Eliminate the i-th characteristic from further considerations. If all the characteristics were eliminated, terminate execution: the stored values of n_x and $E(K|n_x)$ constitute the solution. In other case assign the value of m_1 to m_0 and go to the step 2.

The number of iterations to execute is not greater than the number k characteristics under study. In every iteration the expression (20) is

evaluated in step 4 for each of not eliminated characteristics, so if we assume, that evaluation of this expression is a dominating operation, the computational complexity of the algorithm is of the order $O(k^2)$.

4. EXAMPLE

In the population of size $N = 100\ 000$, values of 9 characteristics are observed. Per-unit costs are $C_0 = 0.1$, $C_1 = 0.4$, $C_2 = 4$ respectively. Table 1 shows the variances S_i^2, S_{2i}^2, the desired precisions V_{0i}, and the minimum initial sample sizes n_i^* corresponding to each characteristic.

Table 1

Example data

Characteristic	S_i^2	S_{2i}^2	V_{0i}	n_i^*
1	1 400	1 800	1	1 380.671
2	500	1 160	0.5	990.099
3	950	1 700	0.8	1 173.564
4	400	700	2	199.6008
5	1 450	1 550	1	1 429.276
6	3 000	1 800	2	1 477.833
7	1 000	1 340	1.2	826.4463
8	325	250	0.3	1 283.317
9	1 500	500	4	373.599

Table 2 shows the interval bounds m_0 and m_1, evaluated in successive iterations, optimum initial sample sizes for each iteration and corresponding expected costs.

Table 2

Computation results

Iteration	m_0	m_1	n_x	$E(K)$
1	1 477.833	1 544.986	1 544.986	2 603.747
2	1 544.986	1 729.56	1 729.56	2 443.612
3	1 729.56	2 512.186	2 167.548	2 382.019
4	2 512.186	3 129.657	2 512.186	2 403.948
5	3 129.657	100 000	3 129.657	2 568.449

The lowest value of cost $E(K) = 2382.019$ was achieved in the third interval, for $n_x = 2167.548$. As n_x is not an integer, the result was rounded to 2168. The expected cost for the rounded size is equal to 2382.020 so it does not differ significantly from the cost obtained for not-rounded initial sample size.

REFERENCES

Cochran W. G. (1977), *Sampling Techniques*, Wiley, New York.

Hansen M. H., Hurwitz W. N. (1946), *The Problem of Nonresponse in Sample Surveys*, "Journal of the American Statistical Association", **41**.

Srinath K. P. (1971), *Multiphase Sampling in Nonresponse Problems*, "Journal of the American Statistical Association", 335.

Sarndal C. E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, Berlin.

Wojciech Gamrot

STRATEGIA LOSOWANIA DWUFAZOWEGO DLA ESTYMACJI WARTOŚCI PRZECIĘTNEJ WIELU CECH W OBECNOŚCI BRAKÓW ODPOWIEDZI

W badaniach statystycznych występuje często zjawisko nieuzyskania danych od części badanych jednostek. Prowadzi to do obciążenia ocen badanego parametru populacji. Jedną z technik stosowanych dla przeciwdziałania temu zjawisku jest ponawianie badania (ang. *callback*) w grupie jednostek populacji, od których nie uzyskano danych. Często spotykanym rozwiązaniem, umożliwiającym ograniczenie kosztu badania jest wylosowanie jedynie pewnego podzbioru tych jednostek w celu ponowienia próby kontaktu. W niniejszym artykule rozważono strategię polegającą na jednokrotnym ponowieniu badania, dla jednoczesnej estymacji wartości przeciętnych wielu cech w populacji. Zaproponowano algorytm iteracyjny umożliwiający ustalenie optymalnej liczebności próby i podpróby, cechujący się wielomianową złożonością obliczeniową.