ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Marcin Skibicki**

ON MAXIMIZATION OF ESTIMATION ACCURACY IN MULTIPARAMETER TWO STAGE SAMPLING

Abstract. In the paper the problem of sample allocation for both stages in such a way, that the accuracy of an estimation of the means of many variables is maximal and the survey cost is restricted is considered. As the measure of accuracy, value of the spectral radius of covariance matrix of means estimators vector is taken. It was proved that the spectral radius is a convex function of sample sizes. That allows effective solving the problem using known methods adapted to this issue.

Key words: two stage sampling, sample allocation, covariance matrix, spectral radius.

1. INTRODUCTION

Let us consider the problem of sample allocation on both sampling stages in the case of estimation of many variables means. Assuming limited financial resources, we want to maximize accuracy of the estimation. As the measure of accuracy we take value of the spectral radius of covariance matrix of means estimators vector (see Wywiał 1995). The problem of minimizing of the spectral radius when the survey cost is restricted in the stratified sampling case was considered among others by J. Wywiał (1988), J. Wywiał and G. Kończak (1994). The solution proposed here is based on convexity of the maximum eigenvalue of the covariance matrix as a function of sample sizes.

2. NOTATIONS

Let $U = \{1, ..., N\}$ be a population of N units partitioned into G groups $U_1, ..., U_G$ $(G > 2)$, so that $U_j \cap U_k = \varphi$, for $j \neq k$. There sizes are adequately

* PhD, Department of Statistics, University of Economics, Katowice, e-mail: skibi@ae.katowice.pl.

$N_1, ..., N_G$ ($N_j \geqslant 2$ for $j = 1, ..., G$). Suppose that m variables are observed in the population ($m \geqslant 2$). Considering the two-stage sampling as the first stage sampling units let take groups $U_1, ..., U_G$. On the first sampling stage a sample S of size g is drawn without replacement from the groups. On the second sampling stage form each group, a sample S_h of size n_h is drawn without replacement for $h \in S$. The vector of means values in the population is of the form:

$$\overline{\mathbf{y}} = [\overline{y}_1, ..., \overline{y}_m]^T,$$

where

$$\overline{y} = \frac{1}{N}\sum_{k=1}^{N} y_{ik}, \quad i = 1, ..., m.$$

An unbiased estimator of the means vector $\overline{\mathbf{y}}$ is $\overline{\mathbf{y}}_s = [\overline{y}_{1S}, ..., \overline{y}_{mS}]^{\mathrm{T}}$ with its co-ordinates:

$$\overline{y}_{iS} = \frac{G}{Ng}\sum_{h\in S}\left[\frac{N_h}{n_h}\sum_{k\in S_h} y_{ik}\right].$$

The covariance matrix of vector $\overline{\mathbf{y}}_s$ is of the form:

$$\mathbf{V}(\overline{y}_s) = \frac{G}{N^2}\left[\frac{(G-g)}{g}\mathbf{C}_{\bullet} + \frac{1}{g}\sum_{h=1}^{G}\frac{N_h(N_h - n_h)}{n_h}\mathbf{C}_{\bullet h}\right].$$

$\mathbf{C}_{\bullet}$ is the between-groups covariance matrix with its elements:

$$\mathbf{C}_{\bullet h}(y_i, y_j) = \frac{1}{G-1}\sum_{h=1}^{G}(\overline{y}_{ih} - \overline{y}_i)(\overline{y}_{jh} - \overline{y}_j),$$

where

$$\overline{y}_{ih} = \frac{1}{N_h}\sum_{k\in U_h} y_{ih}, \quad i = 1, ..., m, \quad h = 1, ..., H.$$

$\mathbf{C}_{\bullet}$ is the within-groups covariance matrix with its elements:

$$\mathbf{C}_{\bullet h}(y_i, y_j) = \frac{1}{N_h - 1}\sum_{k\in U_h}(y_{ik} - \overline{y}_{ih})(y_{jk} - \overline{y}_{jh}).$$

3. FROM OF THE PROBLEM

Let us define the total cost of the survey as

$$k(g,n_1, \ldots, n_G) = k_0 g + \frac{g}{G}\sum_{h=1}^{G} k_h n_h \tag{3}$$

where k_0 is the unit cost of surveying the first stage elements (e.g. cost of organizing the survey of each group, area). However k_h is the unit cost of surveying the second stage elements from h-th group. Defined in this way cost of the survey is the expected costs of sample observation.

The problem consist in calculating sample sizes on both sampling stages in such a way that the spectral radius (the maximal absolute eigenvalue) of the covariance matrix $\mathbf{V}(\overline{\mathbf{y}}_S)$ is minimal and the survey cost $k(g,n_1, \ldots, n_G)$ do not exceed specified value. Let us denote the covariance matrix as the function of sample sizes $\mathbf{V}(g,n_1, \ldots, n_H) = \mathbf{V}(\overline{\mathbf{y}}_s)$. The problem may be written down as:

$$\begin{cases} \max|\lambda| : \lambda \in \operatorname{Spect}(\mathbf{V}(g,n_1, \ldots, n_G))\} = \min \\ k(g,n_1, \ldots, n_G) \leqslant K \\ 2 \leqslant g \leqslant G \\ 2 \leqslant n_h \leqslant N_h, \quad h = 1, \ldots, G \end{cases} \tag{4}$$

The value of objective function can not be calculated analytical, however if it is a convex function of sample sizes, searching for solution is simplified. In this case the problem is a non-linear constrained programming problem with convex objective function (see Wit 1986).

Theoreme 1. Maximum eigenvalue of the covariance matrix $\mathbf{V}(g,n_1, \ldots, n_H)$ is a convex function on set:

$$\begin{cases} 2 \leqslant g \leqslant G \\ 2 \leqslant n_h \leqslant N_h, \quad h = 1, \ldots, G \end{cases} \tag{5}$$

Proof: Let $\mathbf{n} = [g,nm_1, \ldots, n_H]^{\mathrm{T}}$ be the vector of sample sizes. The functions defined as follow:

$$f_0(\mathbf{n}) = \frac{G}{N^2}\left(\frac{G}{g} - 1\right),$$

$$f_h(\mathbf{n}) = \frac{GN_h}{N^2}\frac{1}{g}\left(\frac{N_h}{n_h} - 1\right), \quad h = 1, \ ..., \ G,$$

are convex on the set given by inequalities (5). For notation simplification let us write $\mathbf{C}_{\bullet 0} = \mathbf{C}_{\bullet}$. The covariance matrix $\mathbf{V}(g, n_1, \ ..., \ n_H)$ may be written in the form:

$$\mathbf{V}(\boldsymbol{n}) = \sum_{h=0}^{G} f_h(\boldsymbol{n})\mathbf{C}_{\bullet h}.$$

Symbol $\lambda_1(\mathbf{V}(\boldsymbol{n}))$ denotes the maximum eigenvalue of the matrix $\mathbf{V}(\boldsymbol{n})$. Maximum absolute value of the eigenvalues is a norm on the space of real matrices of fixed size. Moreover, the eigenvalues of the covariance matrix are non-negative, then for all covariance matrices $\mathbf{V}_1$ i $\mathbf{V}_2$ and every $t \in (0, 1)$ we have

$$\lambda_1(t\mathbf{V}_1 + (1-t)\mathbf{V}_2) \leqslant t\lambda_1(\mathbf{V}_1) + (1-t)\lambda_1(\mathbf{V}_2).$$

Because the functions $\lambda_h(\mathbf{n})$ are convex and the matrices $\mathbf{C}_{\bullet h}$ are non-negative definite we obtain:

$$\lambda_1(\mathbf{V}(t\mathbf{n}^{(1)} + (1-t)\mathbf{n}^{(2)})) = \sup_{\mathbf{x} \in \mathbf{R}^m} \frac{\mathbf{x}^T\left[\sum_{h=0}^{H} f_h(t\mathbf{n}^{(1)} + (1-t)\mathbf{n}^{(2)})\mathbf{C}_{\bullet h}\right]\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} =$$

$$= \sup_{\mathbf{x} \in \mathbf{R}^m} \frac{\sum_{h=0}^{H} f_h(t\mathbf{n}^{(1)} + (1-t)\mathbf{n}^{(2)})x^T\mathbf{C}_{\bullet h}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \leqslant \sup_{\mathbf{x} \in \mathbf{R}^m} \frac{\sum_{h=0}^{H} [tf_h(\mathbf{n}^{(1)} + (1-t)f_h\mathbf{n}^{(2)})]\mathbf{x}^{\mathrm{T}}\mathbf{C}_{\bullet h}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} =$$

$$= \sup_{\mathbf{x} \in \mathbf{R}^m} \frac{\mathbf{x}^T\left[t\sum_{h=0}^{H} f_h(\mathbf{n}^{(1)})\mathbf{C}_{\bullet h} + (1-t)\sum_{h=0}^{H} f_h(\mathbf{n}^{(2)})\mathbf{C}_{\bullet h}\right]\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \lambda_1(tV(\mathbf{n}^{(1)}) + (1-t)V(\mathbf{n}^{(2)}))$$

for all vectors $\mathbf{n}^{(1)}$ and $\mathbf{n}^{(2)}$ which fulfils (5) and every $t \in (0, 1)$. Finally we have

$$\lambda_1(\mathbf{V}(t\mathbf{n}^{(1)} + (1-t)\mathbf{n}^{(2)})) \leqslant t\lambda_1(\mathbf{V}(\mathbf{n}^{(1)})) + (1-t)\lambda_1(\mathbf{V}(\mathbf{n}^{(2)})).$$

which means convexity of the maximum eigenvalue as a function of sample sizes $g, n_1, \ ..., \ n_G$ which fulfils (5).

The problem (4) can be solved using known non-linear programming methods and the maximum eigenvalue may be calculated for example with the power method. Since we use a non-integer programming method, there is a problem of rounding obtained sample sizes. The simplest method is to round the sample sizes to lower integers, witch allows to fulfil the survey cost restriction but also increases the spectral radius value. In practice, rounding size g of the first stage sample may indeed affect the solution. Because of that, if non-integer value of g was obtained, we may for example solve the problem two more times for g rounded respectively to lower and higher integer. Finally better from these two solutions is selected.

4. EXAMPLE

We have the population of 7523 farms from administrative unit Dąbrowa Tarnowska. The data used for calculating optimal sample sizes come from the general agricultural census for 1996.

Four variables have been selected to survey:

1) arable land (in hectares),
2) stock of cattle (in heads),
3) stock of pigs (in heads),
4) value of sale (in thousands of zlotys).

The population was divided into 6 groups on the basis of territorial division. Group sizes, unit costs and mean values of the variables are presented in Tab. 1. The value 30 was taken as the unit cost of surveying first stage elements.

Table 1

Sizes of groups	Unite costs of survey	Mean values of variables			
		I	II	III	IV
1046	1.0	3.864	2.41	5.42	3.732
1312	1.0	4.574	1.97	3.44	2.247
695	1.5	5.141	3.66	9.87	6.607
1459	2.0	4.671	2.63	5.19	3.181
1163	2.0	4.847	1.98	3.98	2.559
1848	2.5	3.714	1.82	5.26	2.751

The between-groups covariance matrix is of the form:

$$\mathbf{C}_{\bullet} = \mathbf{c}_{\bullet}(y_i,y_j) = \begin{bmatrix} 0.326 & 0.239 & 0.499 & 0.425 \\ 0.239 & 0.491 & 1.478 & 1.092 \\ 0.499 & 1.478 & 5.315 & 3.699 \\ 0.425 & 1.092 & 3.699 & 2.675 \end{bmatrix}.$$

Restriction of the survey cost amounts $K = 800$. Solutions obtained for each value of the first stage sample size presents Tab. 2. The first row of results in the table contains solution optimal in the sense of considered problem.

Table 2

First stage, sample size	Second stage sample size	Spectral radius, value
2	236, 140, 131, 144, 87, 474	0.2300
3	150, 90, 84, 92, 56, 303	0.2617
4	106, 64, 60, 66, 41, 218	0.2850
5	87, 49, 46, 50, 30, 166	0.3052
6	69, 39, 36, 40, 24, 132	0.3249

REFERENCES

Greń J. (1966), *O pewnym zastosowaniu programowania nieliniowego do metody reprezentcyjnej*, "Przegląd Statystyczny", **13**, s. 203–217.

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

Wit R. (1986), *Metody programowania nieliniowego*, WNT, Warszawa.

Wywiał J. (1988), *Lokalizacja próby w warstwach minimalizująca promień spektralny macierzy wariancji i kowariancji wektora średnich z próby*, "Prace Naukowe Akademii Ekonomicznej we Wrocławiu", **404**, s. 195–200.

Wywiał J. (1995), *Wielowymiarowe aspekty metody reprezentacyjnej*, Ossolineum, Katowice.

Wywiał J., Kończak G. (1994), *O lokalizacji próby w warstwach minimalizującej promień spektralny macierzy wewnątrzwarstwowej macierzy wariancji i kowariancji*, [w:] *XI Seminarium Ekonometryczne im. Profesora Zbigniewa Pawłowskiego, Trzemieśnia 24–26 III*, Akademia Ekonomiczna, Kraków, 85–92.

Marcin Skibicki

O MAKSYMALIZACJI DOKŁADNOŚCI ESTYMACJI W WIELOPARAMETROWYM LOSOWANIU DWUSTOPNIOWYM

W pracy rozważano zadanie ustalenia liczebności prób losowanych na obydwu stopniach losowania tak, aby dokładność estymacji średnich wielu cech populacji była maksymalna przy kosztach obserwacji próby nieprzekraczających zadanego poziomu. Za miarę dokładności estymacji przyjęto wartość promienia spektralnego macierzy kowariancji wektora estymatorów wartości średnich cech. Wykazano, że promień spektralny tej macierzy dla przekształconego zadania jego minimalizacji jest wypukłą funkcją odpowiedniego wektora. Pozwala to na efektywne poszukiwanie optymalnego rozwiązania przy użyciu znanych metod adaptowanych do tego problemu.