ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Janusz Wywial**

ESTIMATION OF POPULATION AVERAGES ON THE BASIS OF A VECTOR OF CLUSTER MEANS**

Abstract. The estimation of a vector of mean values is being considered. The vector estimator consists of simple cluster sample means. It is assumed that a population of a fixed size is divided into mutually disjoint clusters each of the same size. The variance-covariance matrix of the vector estimator is derived. It is a function of a homogeneity matrix of multidimensional variable which describes within-cluster spread of the multidimensional variable under research. The accuracy of estimation is measured by means of standard deviations of particular sample cluster means as well as by means of the trace or the determinant or the maximal eigenvalue of the variance-covariance matrix of the vector estimator. The accuracy of the vector of simple sample cluster means is compared with the accuracy of the vector of the simple sample means. The accuracy of the vector of simple sample cluster means increases when the degree of within-cluster spread of the distribution of a multidimensional variable increases. Hence, the population should be divided into such clusters that the within-cluster spread is as large as possible.

Key words: cluster sample, vector estimation, clustering methods, generalised variance, relative efficiency, homogeneity coefficient of multidimensional variable, eigenvalue of variance-covariance matrix.

1. THE BASIC PROPERTIES OF THE VECTOR OF CLUSTER MEANS

A fixed population of the size N is denoted by Ω. It is convenient to treat the population as a subset of the natural numbers: $\Omega = \{1, 2, ..., N\}$. Let us assume that the population Ω is divided into G such mutually disjoint clusters Ω_p $(p = 1, ..., G)$ that $\bigcup_{p=1}^{G} \Omega_p = \Omega$. If each cluster is of the same size denoted by M, the population Ω is of the size $N = GM$. Let S be the cluster sample of the size g. The random sample S is drawn according to the following design:

* Prof., Department of Statistics, University of Economics, Katowice, e-mail: wywia@lae.katowice.pl.

** The research was supported by the grant number 1 H02B 015 10 from the Polish Scientific Research Committee.

$$P_g(S) = \frac{1}{\binom{G}{g}}.$$

A k-th ($k = 1, ..., N$) outcome of an i-th ($i = 1, ..., m$) variable is denoted by y_{ki}. The sum of observations of an i-th variable in a p-th cluster is as follows:

$$z_{ip} = \sum_{k \in \Omega_p} y_{ki}.$$

The mean value of an i-th variable in a p-th cluster is:

$$\bar{y} = \frac{1}{M} z_{pi}.$$

The mean value of an i-th variable per cluster is:

$$\bar{z}_i = \frac{1}{G} \sum_{p=1}^{G} z_{pi}.$$

The population mean of an i-th variable takes the following form:

$$\bar{y}_i = \frac{1}{N} \sum_{p=1}^{G} z_{pi}.$$

The variance-covariance matrix is denoted by: $\mathbf{C} = [\mathrm{cov}(y_i, y_j)]$, where:

$$\mathrm{cov}(y_i, y_j) = \frac{1}{N-1} \sum_{p=1}^{G} \sum_{k \in \Omega_p} (y_{ki} - \bar{y}_i)(y_{ki} - \bar{y}_j).$$

The variance-covariance matrix of cluster sums is denoted by: $\mathbf{C}_z = [\mathrm{cov}(z_i, z_j)]$, where:

$$\mathrm{cov}(z_i, z_j) = \frac{1}{G-1} \sum_{p=1}^{G} (z_{pi} - \bar{z}_i)(z_{pi} - \bar{z}_k).$$

The estimator of the vector $\bar{\mathbf{y}} = [\bar{y}_1, ..., \bar{y}_m]$ is defined as the vector $\bar{\mathbf{y}}_{gS} = [\bar{y}_{1gS}, ..., \bar{y}_{mgS}]$, where:

$$\bar{y}_{igS} = \frac{1}{gM} \sum_{p \in S} \sum_{l \in \Omega_p} y_{ki} = \frac{1}{gM} \sum_{p \in S} z_{pi} \tag{1}$$

The vector $\bar{y}_{igS}$ is the unbiased estimator of the mean vector $\bar{y}$.

The covariance of the estimators $\bar{y}_{igS}$, $\bar{y}_{jgS}$ $(i \neq j = 1, ..., m)$ can be derived similarly as variance of $\bar{y}_{igS}$ $(i = 1, ..., m)$, see e.g. W. G. Cochran (1963) or C. E. Särndal, B. Swenson, J. Wretman (1992).

$$\operatorname{cov}(\bar{y}_{igS}, \bar{y}_{jgS}) = \frac{G-g}{GgM^2} \operatorname{cov}(z_i, z_j) \tag{2}$$

The variance-covariance matrix of the $\bar{y}_{gS}$ can be written down in the following way:

$$\mathbf{V}(\bar{y}_{gS}, P_g) = \frac{G-g}{GgM^2} \mathbf{C}(z) \tag{3}$$

where: $\mathbf{C}(z) = [\operatorname{cov}(z_i, z_j)]$.

The unbiased estimator of the covariance is obtained through substitution of the following statistic for the parameter $\operatorname{cov}(z_i, z_j)$:

$$\operatorname{cov}_S(z_i, z_j) = \frac{1}{g-1} \sum_{p \in S} (z_{pi} - \bar{z}_i)(z_{pj} - \bar{z}_j).$$

2. HOMOGENEITY COEFFICIENT OF MULTIDIMENSIONAL VARIABLE

Let $\mathbf{C}_b = [\operatorname{cov}_b(y_i, y_j)]$ be the between-cluster matrix of the variances and covariances, where:

$$\operatorname{cov}_b(y_i, y_j) = \frac{1}{G-1} \sum_{p=1}^{G} (\bar{y}_{ip} - \bar{y}_i)(\bar{y}_{jp} - \bar{y}_j).$$

The within-cluster matrix of the variances and covariances is denoted by $\mathbf{C}_w = [\operatorname{cov}_w(y_i, y_j)]$, where:

$$\operatorname{cov}_w(y_i, y_j) = \frac{1}{G(M-1)} \sum_{p=1}^{G} \sum_{k \in \Omega_p} (y_{ik} - \bar{y}_{ip})(y_{jk} - \bar{y}_{jp}).$$

Similarly to the one dimensional case (see e.g. Cochran 1963, p. 243) the variance-covariance matrix $\mathbf{C}$ can be decomposed in the following way:

$$(N-1)\mathbf{C} = (G-1)M\mathbf{C}_b + (N-G)\mathbf{C}_w \tag{4}$$

The matrix $\mathbf{C}(z)$ can be rewritten as follows:

$$\mathbf{C}(z) = M^2\mathbf{C}_b \tag{5}$$

This expression and the equation (4) lead to the following results:

$$\mathbf{C}(z) = \frac{M}{G-1}((N-1)\mathbf{C} - (N-G)\mathbf{C}_w)$$

$$\mathbf{C}(z) = M\mathbf{C}\left(\mathbf{I} + \frac{N-G}{G-1}\Delta\right) \tag{6}$$

where:

$$\Delta = \mathbf{I} - \mathbf{C}^{-1}\mathbf{C}_w \tag{7}$$

In the case of an one-dimensional variable y_i, when $\mathbf{C}$ reduces to the variance $\operatorname{var} y_i$ and $\mathbf{C}_w$ is the within-cluster variance var_w the matrix Δ reduces to the homogeneity coefficient (see Särndal, Swenson, Wretman 1992, p. 130):

$$\delta(y_i) = 1 - \frac{\operatorname{var}_w(y_i)}{\operatorname{var}(y_i)}, \quad -\frac{G-1}{N-G} \leqslant \delta(y_i) \leqslant 1 \tag{8}$$

where:

$$\operatorname{var}(y_i) = \frac{1}{N}\sum_{p=1}^{G}\sum_{k\in\Omega_p}(y_{ik} - \overline{y}_i)^2, \quad \overline{y}_i = \frac{1}{N}\sum_{p=1}^{G}\sum_{k\in\Omega_p} y_{ik} \tag{9}$$

$$\operatorname{var}_w(y_i) = \frac{1}{G(M-1)}\sum_{p=1}^{G}\sum_{k\in\Omega_p}(y_{ik} - \overline{y}_{ip})^2, \quad \overline{y}_{ip} = \frac{1}{M}\sum_{k\in\Omega_p} y_{ik} \tag{10}$$

Then, the matrix Δ can be treated as generalization of the homogeneity coefficient δ. That is why the matrix Δ can be named as homogeneity matrix of multidimensional variable.

Theorem 1. If the variance-covariance matrix $\mathbf{C}$ is non-singular then the eigenvalues λ_i $(i = 1, ..., m)$ of the matrix Δ fulfill the following inequalities:

$$-\frac{G-1}{N-G} \leqslant \lambda_i \leqslant 1, \quad \text{for each} \quad i = 1, ..., m \tag{11}$$

Proof. The characteristic equation for the matrix Δ can be transformed as follows:

$$|\Delta - \lambda \mathbf{I}| = 0 \tag{12}$$

$$|\mathbf{I} - \mathbf{C}^{-1}\mathbf{C}_w \lambda \mathbf{I}| = 0$$

$$|\mathbf{C}^{-1}\mathbf{C}_w - \kappa \mathbf{I}| = 0 \tag{13}$$

where $\kappa = (1 - \lambda)$. Since the matrix $\mathbf{C}^{-1}\mathbf{C}_w$ is positive semi-definite its eigenvalues $\kappa_i \geqslant 0$ for each $i = 1, ..., m$. Hence, the eigenvalues of the matrix Δ are: $\lambda_i \leqslant 1$ for each $i = 1, ..., m$.

Since the matrix $\mathbf{C}_b$ is positive semi-definite the equation (4) leads to the matrix

$$\mathbf{A}_1 = (N-1)\mathbf{C} - (N-G)\mathbf{C}_w$$

which is positive semi-definite, too. Because the matrix $\mathbf{C}$ is positively defined the following matrix is positive semi-definite:

$$\mathbf{A}_2 = \frac{1}{N-G}\mathbf{C}^{-1}\mathbf{A}_1 = \frac{N-1}{N-G}\mathbf{I} - \mathbf{C}^{-1}\mathbf{C}_w.$$

After simple algebraic transformations we have:

$$\mathbf{A}_2 = \Delta + \frac{G-1}{N-G}\mathbf{I} \tag{14}$$

Let us do the following transformations:

$$|\Delta - \lambda \mathbf{I}| = 0,$$

$$\left|\Delta + \frac{G-1}{N-G}\mathbf{I} - \frac{G-1}{N-G}\mathbf{I} - \lambda\right| = 0,$$

$$|\mathbf{A}_2 - \varsigma \mathbf{I}| = 0 \tag{15}$$

where:

$$\varsigma = \frac{G-1}{N-G}\lambda \tag{16}$$

Since the matrix $\mathbf{A}_2$ is positive semi-definite the eigenvalue $\varsigma_i \geqslant 0$ for each $i = 1, ..., m$. Hence, on the basis of the expression (16) we have: $\lambda_i \leqslant -\frac{G-1}{N-G}$ for $i = 1, ..., m$. This completes the proof.

We can say that the within-cluster spread of observations of a multidimensional variable is less than their population spread if the matrix Δ is positive definite. When Δ is negative definite, then we say that the population spread of values of a multidimensional variable is less than the within-cluster spread.

3. ACCURACY OF A CLUSTER SAMPLE MEAN VECTOR IN RELATION TO SIMPLE SAMPLE MEAN VECTOR

Let $\bar{\mathbf{y}}_S$ be the vector of the mean from the simple random sample of the size n, selected without replacement from a population of the size N. Its variance-covariance matrix is of the following form:

$$\mathbf{V}(\bar{\mathbf{y}}_S, P_S) = \frac{N-n}{Nn}\mathbf{C} \tag{17}$$

where:

$$P_S = \frac{1}{\binom{N}{n}}.$$

On the basis of the equations (3) and (6) we have:

$$\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g) = \frac{G-g}{GgM}\mathbf{C}\left(\mathbf{I} + \frac{N-G}{G-1}\Delta\right) \tag{18}$$

Under the assumption that $GM = N$ and $gM = n$:

$$\mathbf{V}(\bar{y}_{gS}, P_g) = \frac{N-n}{Nn}\mathbf{C}\left(\mathbf{I} + \frac{N-G}{G-1}\Delta\right)$$

Hence:

$$\mathbf{V}(\bar{\mathbf{y}}_S, P_s) - \mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g) = \frac{N-n}{Nn}\frac{N-G}{G-1}\mathbf{C}\Delta \tag{20}$$

or

$$\mathbf{V}(\bar{\mathbf{y}}_S, P_s) - \mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g) = \frac{N-n}{Nn}\frac{N-G}{G-1}(\mathbf{C}-\mathbf{C}_w) \tag{21}$$

This leads to the following property:

Theorem 2. If the matrix $(\mathbf{C}-\mathbf{C}_w)$ is non-positive definite (non-negative definite) then the strategy $\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_S)$ is not worse (not better) than the strategy $\mathbf{V}(\bar{\mathbf{y}}_S, P_S)$. Particularly, if the matrix C is nonsingular and the Δ is non-positive definite (non-negative definite) then the strategy $\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g)$ is not worse (not better) than the strategy $\mathbf{V}(\bar{\mathbf{y}}_S, P_S)$.

Hence, The strategy $\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g)$ is not worst than the strategy $\mathbf{V}(\bar{\mathbf{y}}_S, P_S)$, if the within-cluster spread of a multidimensional variable represented by the matrix $\mathbf{C}_w$ is larger than its population spread represented by the matrix $\mathbf{C}$.

Let us denote the variance of a strategy, the determinant, the trace and the maximal eigenvalue of a variance-covariance matrix of a vector strategy by $\mathrm{D}^2(.,.)$, $\det(.,.)$, $\mathrm{tr}(.,.)$ and $\lambda_1(.,.)$, respectively. The relative efficiency coefficients are defined as follows:

$$e_{0i} = \frac{\mathrm{D}^2(\bar{y}_{gS}, P_g)}{\mathrm{D}^2(\bar{\mathbf{y}}_S, P_s)} = 1 + \frac{N-G}{G-1}\delta(y_i), \quad i = 1, \ldots, m \tag{22}$$

where $\delta(y_i)$ expresses the formulas (8–10).

$$e_1 = \frac{\det \mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g)}{\det \mathbf{V}(\bar{\mathbf{y}}_S, P_S)} = \det\left(\mathbf{I} + \frac{N-G}{G-1}\Delta\right) \tag{23}$$

$$e_2 = \frac{\mathrm{tr}\,\mathbf{V}(\bar{\mathbf{y}}_{gS}, P_g)}{\mathrm{tr}\,\mathbf{V}(\bar{\mathbf{y}}_S, P_s)} = 1 + \frac{N-G}{G-1}\bar{\delta} \tag{24}$$

where:

$$\bar{\delta} = \sum_{i=1}^{m} \delta(y_i)a_i,$$

$$a_i = \frac{\operatorname{var}(y_i)}{\sum_{i=1}^{m} \operatorname{var}(y_i)},$$

$$e_3 = \frac{\lambda_1(\bar{\mathbf{y}}_{gS}, P_g)}{\lambda_1(\bar{\mathbf{y}}_S, P_3)} \tag{25}$$

Theorem 3. If the matrix $\mathbf{C}$ is positive definite and matrix Δ is non-positive (non-negative) definite, then $e_k \leqslant 1$ for $k = 1, 2, 3$ and $e_{0i} \leqslant 1$ for $i = 1, ..., m$. Particularly, if the matrix Δ is negative (positive) definite, $e_k < 1$ for $k = 1, 2, 3$ and $e_{0i} \leqslant 1$ for $i = 1, ..., m$ and $e_{0j} < 1$ for at least one index $j = 1, ..., m$.

C. R. R a o (1982, p. 89), showed: if $\mathbf{B}$ is positive definite and $(\mathbf{A} - \mathbf{B})$ is non-negative definite then $\det(\mathbf{A}) \geqslant \det(\mathbf{B})$. This and the expression (7) lead to inequality $e_1 \leqslant 1$. The properties of the trace of a sum of matrixes lead to the inequality $e_2 \leqslant 1$. If the matrix Δ is non-positive definite, the matrix $(\mathbf{C} - \mathbf{C}_w)$ is non-positive definite, too. Let $\lambda_1(\mathbf{A})$ be the maximal eigenvalue of a matrix $\mathbf{A}$. Hence:

$$\lambda_1(\mathbf{C}) = \max_{\alpha^T\alpha - 1} \{\alpha^T \mathbf{C} \alpha\},$$

$$\lambda_1(\mathbf{C}_w) = \max_{\alpha^T\alpha - 1} \{\beta^T \mathbf{C}_w \beta\}.$$

If $(\mathbf{C} - \mathbf{C}_w)$ is non-negative defined then for all non-zero vectors γ:

$$\gamma^T \mathbf{C} \gamma - \gamma^T \mathbf{C}_w \gamma \geqslant 0 \tag{26}$$

Hence:

$$\alpha^T \mathbf{C} \alpha - \alpha^T \mathbf{C}_w \alpha = \lambda_1(\mathbf{C}) - \alpha^T \mathbf{C}_w \alpha \geqslant 0,$$

$$\beta^T \mathbf{C} \beta - \beta^T \mathbf{C}_w \beta = \beta^T \mathbf{C} \beta - \lambda_1(\mathbf{C}_w) \geqslant 0,$$

$$\lambda_1(\mathbf{C}) - \lambda_1(\mathbf{C}_w) \geqslant \beta^T \mathbf{C} \beta - \lambda_1(\mathbf{C}_w) \geqslant 0,$$

$$\lambda_1(\mathbf{C}) \geqslant \lambda_1(\mathbf{C}_w).$$

This leads to inequality: $e_3 \leqslant 1$. The inequality (26) let us derive the inequalities $e_{0i} \leqslant 1$, $i = 1, ..., m$ when we assume that the elements of the vector γ are equal to zero except the i-th element equal to one.

The strategy $(\bar{y}_{gS}, P_g)$ can be better than the strategy $(\bar{y}_S, P_s)$ if the matrix $(\mathbf{C} - \mathbf{C}_w)$ is negative definite. It means that the within-cluster spread of values of the multidimensional variable (under research) should be bigger than the population spread of observations of those variables.

REFERENCE

Cochran W. G. (1963), *Sampling Techniques*, John Wiley, New York.
Rao C. R. (1982). *Modele liniowe statystyki matematycznej*, PWN, Warszawa.
Särndal C. E., Swenson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York–Berlin–Heidelberg–London–Paris–Tokyo–Hong Kong–Barcelona–Budapest.

Janusz Wywial

ESTYMACJA WARTOŚCI PRZECIĘTNYCH W POPULACJI NA PODSTAWIE WEKTORA ŚREDNICH Z PRÓBY GRUPOWEJ

Zakłada się, że skończona i ustalona populacja jest podzielona na równoliczne i rozłączne grupy. Na podstawie prostej próby grupowej jest wyznaczany wektor średnich, który daje oceny wektora przeciętnych w populacji. Wyprowadzono macierz wariancji i kowariancji wektora wartości średnich z próby grupowej. Jest ona zależna od macierzy wewnątrzgrupowej jednorodności rozkładu wielowymiarowej zmiennej. Precyzja estymacji jest oceniana za pomocą wariancji poszczególnych średnich z próby grupowej, śladu, wyznacznika lub maksymalnej wartości własnej macierzy wariancji i kowariancji. Precyzja wektora średnich z próby grupowej jest porównywana z precyzją wektora średniej z próby prostej. Okazuje się, że wektor średnich z próby grupowej jest precyzyjniejszy od wektora przeciętnych z próby prostej, gdy stopień wewnątrzgupowego zróżnicowania wartości zmiennych jest dostatecznie duży.