THE ESTIMATION OF MULTIPLE CORRELATION COEFFICIENT IN STRATIFIED RANDOM SAMPLING

KAZUMASA WAKIMOTO

1. Introduction

The main object of stratification in the current sampling survey has been thought to improve the precision in estimating the population mean and the most of the current studies on the stratified random sampling seem so far to have been made within its limitation. But in practice we consider the problem of estimating several population characteristics including mean, variance, covariance, correlation coefficient, etc.

Yanagawa and the author [2] proposed the stratified random sampling theory for the estimation of some functional $\theta(F)$ of the population distribution E including the population mean, variance and covariance.

The author [3] proposed the stratified random sampling theory for the estimation of the correlation coefficient of the population having the bivariate distributions. In this paper, we deal with the problems concerning the estimation of the multiple correlation coefficient of the population based on the stratified random sample.

The principal purpose of this paper is to indicate:

(i) A consistent estimator of the population multiple correlation coefficient $\rho_{1(2\cdots k)}$ together with its mean sqare error (Theorem 1).

(ii) When the stratification is preassigned in a way, the precision of the estimator of $\rho_{1(2\cdots k)}$, in the proportional allocation, is better than one in the simple random sampling (Theorem 2).

(iii) When the allocation of the sample is preassigned as the proportional allocation, the optimum stratification is determined by a number of simple quadratic hypersurfaces (Theorem 3).

2. Notation and preliminary

Let $F(x_1, \cdots, x_k)$ be the distribution function. The number L of strata is preassigned and the population with F is classified into L strata, the i-th of which has the distribution function $F_i(x_1, \cdots, x_k)$, $i=1, \cdots, L$. Then we have

$$F=\sum_{i=1}^{L} w_i F_i, \quad \text{for all } (x_1, \cdots, x_k), \tag{2.1}$$

where w_i is a positive constant such that $\sum_{i=1}^{L} w_i=1$. The value w_i is considered as a weight or a ratio of the i-th stratum to the whole population.

Let $\begin{pmatrix}\mu_1\\ \vdots\\ \mu_k\end{pmatrix}$ and $M=(\sigma_{pq})$ be the population mean vector and the population covariance matrix given by

$$\mu_p=\int x_p\, dF,\quad p=1,\cdots,k \tag{2.2}$$

and

$$\sigma_{pq}=\int(x_p-\mu_p)\,(x_q-\mu_q)\, dF,\quad p,\ q=1,\cdots,k, \tag{2.3}$$

respectively. Then the multiple correlation coefficient of the population between x_1 and $(x_2,\cdots,x_k)$ is defined by

$$\rho_{1(2\cdots k)}=\left(1-\frac{|M|}{\sigma_{11}\tilde{\sigma}_{11}}\right)^{1/2}, \tag{2.4}$$

where $\tilde{\sigma}_{11}$ is the cofactor of σ_{11} in M.

Let $\begin{pmatrix}\mu_{i1}\\ \vdots\\ \mu_{ik}\end{pmatrix}$ and $M_i=(\sigma_{ipq})$ be the mean vector and the covariance matrix of the i-th stratum $(i=1,\cdots,L)$ given by

$$\mu_{ip}=\int x_p\, dF_i,\quad p=1,\cdots,k \tag{2.5}$$

and

$$\sigma_{ipq}=\int(x_p-\mu_{ip})\,(x_q-\mu_{iq})\, dF_i,\quad p,\ q=1,\cdots,k, \tag{2.6}$$

respectively.

From (2.1), (2.2), (2.3), (2.5) and (2.6) we have the following relations:

$$\mu_p=\sum_{i=1}^{L} w_i\mu_{ip},\quad p=1,\cdots,k \tag{2.7}$$

$$\sigma_{pq}=\sum_{i=1}^{L} w_i\sigma_{ipq}+\sum_{i=1}^{L} w_i(\mu_{ip}-\mu_p)\,(\mu_{iq}-\mu_q),\quad p,\ q=1,\cdots,k. \tag{2.8}$$

Throughout this paper, we shall denote the expected value of a random variable T by $E(T)$.

3. A consistent estimator h_i and its mean square error

Let $(X_{i11},\cdots,X_{i1k}),\cdots,(X_{in_i1},\cdots,X_{in_ik})$ be n_i independent and identically distributed random variables having a distribution function F_i, $i=1,\cdots,L$. Let n be total sample size such that $\sum_{i=1}^{L} n_i=n$. Let $U_i=(U_{ipq})$ be the covariance matrix given by

$$(3.1)\qquad U_{spq}=\sum_{i=1}^{L}\frac{w_i^2}{n_i(n_i-1)}\sum_{h<l}^{n_i}(X_{ihp}-X_{ilp})(X_{ihq}-X_{ilq})$$
$$+\sum_{i<j}^{L}\frac{w_iw_j}{n_in_j}\sum_{h=1}^{n_i}\sum_{l=1}^{n_j}(X_{ihp}-X_{jlp})(X_{ihq}-X_{jlq}),$$
$$p,\ q=1,\ \cdots,\ k.$$

Then we propose a consistent estimator of $\rho_{1(2\cdots k)}$ given by

$$(3.2)\qquad h_s=\left(1-\frac{|U_s|}{U_{s11}\widetilde{U}_{s11}}\right)^{1/2},$$

where $\widetilde{U}_{s11}$ is the cofactor of U_{s11} in U_s.

We define *the mean square error* (*MSE*) of h_s by

$$(3.3)\qquad MSE(h_s)=E\{(h_s-\rho_{1(2\cdots k)})^2\}.$$

Then we have the following theorem.

Theorem 1. *The MSE of the estimator h_s is given by*

$$(3.4)\qquad MSE(h_s)=\sum_{p\leq q}^{k}\sum_{p'\leq q'}^{k}\left(\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}\right)\left(\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{p'q'}}\right)\left[\sum_{i=1}^{L}\frac{w_i^2}{n_i}\right.$$
$$\left\{\int(x^p-\mu_{ip})(x_q-\mu_{iq})(x_{p'}-\mu_{ip'})(x_{q'}-\mu_{iq'})\,dF_i-\sigma_{ipq}\sigma_{ip'q'}\right.$$
$$+(\mu_{ip}-\mu_p)\int(x_q-\mu_{iq})(x_{p'}-\mu_{ip'})(x_{q'}-\mu_{iq'})\,dF_i$$
$$+(\mu_{iq}-\mu_q)\int(x_p-\mu_{ip})(x_{p'}-\mu_{ip'})(x_{q'}-\mu_{iq'})\,dF_i$$
$$+(\mu_{ip'}-\mu_{p'})\int(x_p-\mu_{ip})(x_q-\mu_{iq})(x_{q'}-\mu_{iq'})\,dF_i$$
$$+(\mu_{iq'}-\mu_{q'})\int(x_p-\mu_{ip})(x_q-\mu_{iq})(x_{p'}-\mu_{ip'})\,dF_i$$
$$+\sigma_{ipp'}(\mu_{iq}-\mu_q)(\mu_{iq'}-\mu_{q'})$$
$$+\sigma_{ipq'}(\mu_{ip'}-\mu_{p'})(\mu_{iq}-\mu_q)$$
$$+\sigma_{ip'q}(\mu_{ip}-\mu_p)(\mu_{iq'}-\mu_{q'})$$
$$\left.\left.+\sigma_{iqq'}(\mu_{ip}-\mu_p)(\mu_{ip'}-\mu_{p'})+O\left(\frac{1}{n_i}\right)\right\}\right].$$

Proof. From (3.2), we have the following equation using the Taylor expansion.

$$(3.5)\qquad MSE(h_s)=E\left\{\sum_{p\leq q}^{k}\sum_{p'\leq q'}^{k}\left(\frac{\partial h_s}{\partial U_{spq}}\right)_0\left(\frac{\partial h_s}{\partial U_{sp'q'}}\right)_0\right.$$
$$\left.(U_{spq}-\sigma_{pq})(U_{sp'q'}-\sigma_{p'q'})+R\right\},$$

where the suffix 0 denotes the value at the point $(\sigma_{11}, \cdots, \sigma_{1k}, \sigma_{22}, \cdots, \sigma_{2k}, \cdots, \sigma_{kk})$ and R denotes the term of higher order.

From (3. 1) and (3. 5), after some culculations we obtain the equation (3. 4).

4. Improvement of the precision of estimator due to proportinal allocation

In the case of *proportional allocation* $(n_i=w_in,\ i=1, \cdots, L)$, from (3. 2) we have the following estimator of $\rho_{1(2\cdots k)}$:

$$h_{\bar{s}}=\left(1-\frac{|U_{\bar{s}}|}{U_{\bar{s}11}\widetilde{U}_{\bar{s}11}}\right)^{1/2}, \tag{4. 1}$$

where $U_{\bar{s}}=(U_{\bar{s}pq})$ is the covariance matrix given by

$$U_{\bar{s}pq}=\frac{1}{n^2}\left\{\sum_{i=1}^{L}\frac{n_i}{n_i-1}\sum_{h<l}^{n_i}(X_{ihp}-X_{ilp})(X_{ihq}-X_{ilq})\right.$$
$$\left.+\sum_{i<j}^{L}\sum_{h=1}^{n_i}\sum_{l=1}^{n_j}(X_{ihp}-X_{jlp})(X_{ihq}-X_{jlq})\right\}$$
$$p,\ q=1, \cdots, k$$

and $\widetilde{U}_{\bar{s}11}$ is the cofactor of $U_{\bar{s}11}$ in $U_{\bar{s}}$.

On the other hand, we consider a consistent estimator of $\rho_{1(2\cdots k)}$ based on a simple random sample of size n drawn from the population.

Let $(X_{11}, \cdots, X_{1k}), \cdots, (X_{n1}, \cdots, X_{nk})$ be independent and identically distributed random varables having the distribution function F. Then a consistent estimator of $\rho_{1(2\cdots k)}$ is given by

$$h_r=\left(1-\frac{|U_r|}{U_{r11}\widetilde{U}_{r11}}\right)^{1/2}, \tag{4. 2}$$

where $U_r=(U_{rpq})$ is the covariance matrix given by

$$U_{rpq}=\frac{1}{n(n-1)}\sum_{i<j}^{n}(X_{ip}-X_{jp})(X_{iq}-X_{jq}),$$
$$p,\ q=1, \cdots, k$$

and $\widetilde{U}_{r11}$ is the cofactor of U_{r11} in U_r.

From (3. 4), (4. 1) and (4. 2), after some calculations we obtain the following theorem.

Theorem 2. *For any stratification of the population having the distribution function* $F(x_1, \cdots, x_k)$, *we obtain the following equation*:

$$(4.3) \qquad MSE(h_r)-MSE(h_s)$$
$$=\frac{1}{n}\sum_{i<j}^{L} w_i w_j\left[\sum_{p\leq q}^{k}\frac{\partial \rho_{1(2\cdots k)}}{\partial \sigma_{pq}}\{\sigma_{ipq}-\sigma_{jpq} +(\mu_{ip}-\mu_p)(\mu_{iq}-\mu_q)-(\mu_{jp}-\mu_p)(\mu_{jq}-\mu_q)\}\right]^2+O\left(\frac{1}{n^2}\right).$$

5. Optimum stratification in the case of proportional allocation

From (2.1) we have

$$F(x_1\cdots, x_k)=\sum_{i=1}^{L}\overline{F}_i(x_1, \cdots, x_k),$$

where $\overline{F}_i=w_iF_i$ for all $(x_1, \cdots, x_k)$, $i=1, \cdots, L$. Then we call $\{\overline{F}_i\}$ "an L-decomposition of F". Among all possible L-decompositions $\{\overline{F}_i\}$ of the distribution function F, the one which minimizes $MSE(h_{\bar{s}}/\{\overline{F}_i\})$ is called "*optimum stratification* (OS)" for the estimation of $\rho_{1(2\cdots k)}$. For any given n, L and F, to show the existence and the form of OS in the case of proportional allocation, we must show the existence and the form of OS $\{\overline{F}_i^*\}$ such that

$$MSE(h_{\bar{s}}/\{\overline{F}_i^*\})=\inf_{\{\overline{F}_i\}} MSE(h_{\bar{s}}/\{\overline{F}_i\}).$$

Let us consider a vector valued function $\phi=(\phi_1, \cdots, \phi_L)$ such that

$$(5.1) \qquad d\overline{F}_i=\phi_i(x_1, \cdots, x_k)\,dF, \quad i=1, \cdots, L$$

and

$$(5.2) \qquad \sum_{i=1}^{L}\phi_i=1, \quad 0\leqq\phi_i\leqq 1.$$

Let Φ and Ψ be the set of all vector valued function satisfying (5.2) and the set of all decompositions of F, respectively. Then, there is the one-to-one correspondence between Φ and Ψ. Therefore, we can consider such a vector valued function ϕ to be a stratification, and the MSE of $h_{\bar{s}}$ is given by $MSE(h_{\bar{s}}/\phi)$ instead of $MSE(h_{\bar{s}}/\{\overline{F}_i\})$. An OS ϕ^* in Φ is defined as the one at which attains

$$(5.3) \qquad \sup_{\phi\in\Phi} n\{MSE(h_r)-MSE(h_{\bar{s}}/\phi)\}.$$

For any given n, L and F, from (4.3) and (5.3) *an approximate* OS (AOS) ϕ_0 in Φ is defined as the one at which attains

$$\sup_{\phi\in\Phi}\sum_{i<j}^{L} w_i w_j\Big[\sum_{p\leq q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}\{\sigma_{ipq}-\sigma_{jpq}+(\mu_{ip}-\mu_p)(\mu_{iq}-\mu_q)-(\mu_{jp}-\mu_p)(\mu_{jq}-\mu_q)\}\Big]^2,$$

where

$$w_i=\int\phi_i\,dF,\quad \mu_{ip}=\frac{1}{w_i}\int x_p\phi_i\,dF,\quad p=1,\cdots,k$$

and

$$\sigma_{ipq}=\frac{1}{w_i}\int(x_p-\mu_{ip})(x_q-\mu_{iq})\,\phi_i\,dF,\quad p,q=1,\cdots,k,\quad \text{for } i=1,\cdots,L.$$

Put

$$v_{ipq}(\phi)=\int(x_p-\mu_p)(x_q-\mu_q)\,\phi_i\,dF,\quad p,q=1,\cdots,k,\quad \text{for } i=1,\cdots,L.$$

Then after some calculations we obtain

$$\sup_{\phi\in\Phi}\sum_{i<j}^{L} w_i w_j\Big[\sum_{p\leq q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}\{\sigma_{ipq}-\sigma_{jpq}+(\mu_{ip}-\mu_p)(\mu_{iq}-\mu_q)-(\mu_{jp}-\mu_p)(\mu_{jq}-\mu_q)\}\Big]^2$$
$$=\sup_{\phi\in\Phi}\sum_{i=1}^{L}\frac{1}{w_i(\phi)}\Big\{\sum_{p\leq q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}v_{ipq}(\phi)\Big\}^2.$$

Therfore the *AOS* ϕ_0 in Φ is defined by the one at which attaines

$$\text{(5. 4)}\qquad \sup_{\phi\in\Phi}\sum_{i=1}^{L}\frac{1}{w_i(\phi)}\sum_{p\leq q}^{k}\Big\{\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}v_{ipq}(\phi)\Big\}^2.$$

From [2, Lemma 1. 3], it is easy to verify that ϕ_0 in Φ at which attains the expression (5. 4) exists.

Now, to show the form of ϕ_0 in Φ at which attains the expression (5. 4), we use the useful result of Isii and Taga [1].

We state their result in a version suitable to our present situation.

Lemma. *Let* $\phi=(\phi_1,\cdots,\phi_L)$ *be a vector valued function such that* $\sum_{i=1}^{L}\phi_i=1$, $\phi_i\geqq 0$, *for* $i=1,\cdots,L$, *let* S *be a closed set in* R^m *and let* $y(\phi)=(y_1(\phi),\cdots,y_m(\phi))$ *be a linear mapping from* Φ *into* S. *Suppose that* $G(y)$ *is a continuous mapping from* S *into* R^1 *and that* $G(y)$ *is continuously differentiable on* S.

If there exists ϕ_0 *such that*

$$G\{y(\phi_0)\}=\sup_{\phi\in\Phi}G\{y(\phi)\},$$

then

$$\sup_{\phi\in\Phi} \sum_{j=1}^{m} y_j(\phi)\alpha_j$$

is attained at ϕ_0, *where*

$$\alpha_j=\frac{\partial G\{y(\phi)\}}{\partial y_j}\Big|_{\phi=\phi_0}, \quad j=1, \cdots, m.$$

Moreover in case $G(y)$ *is a convex function with respect to* y, *if there exists, for the point* ϕ_0 *in* Φ, *a point* ϕ' *in* Φ *satisfying*

$$\sum_{j=1}^{m} y_j(\phi')\alpha_j=\sum_{j=1}^{m} y_j(\phi_0)\alpha_j,$$

then we have

$$G\{y(\phi')\}=G\{y(\phi_0)\}.$$

By putting

$$G\{g_1(\phi), \cdots, g_L(\phi), w_1(\phi), \cdots, w_L(\phi)\}=\sum_{i=1}^{L}\frac{g_i^2(\phi)}{w_i(\phi)},$$

where $g_i(\phi)=\sum_{p\le q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}v_{ipq}(\phi)$, in the above lemma, we find that ϕ_0 in Φ at which attains the expression (5. 4) also attains

$$\sup_{\phi\in\Phi}\int\sum_{i=1}^{L}2a_i\left\{\sum_{p\le q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}(x_p-\mu_p)(x_q-\mu_q)-\frac{a_i}{2}\right\}\phi_i\,dF, \tag{5.5}$$

where

$$a_i=\frac{\sum_{p\le q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}v_{ipq}(\phi_0)}{w_i(\phi_0)}, \quad i=1, \cdots, L.$$

Since $G(g_1, \cdots, g_L, w_1, \cdots, w_L)$ is convex, from the lemma, ϕ^{**} in Φ at which attains the expression (5. 5) also attains the expression (5. 4). Therefore, from (5. 5) we obtain the following theorem.

Theorem 3. *In the case of proportional allocation, for any given* n, L *and* $F(x_1, \cdots, x_k)$, *an AOS* ϕ^{**} *in* Φ *is given by*

$$\phi_i^{**}(x_1, \cdots, x_k)=\begin{cases}1, & \Big\{(x_1, \cdots, x_k)\ \textit{such that}\\ & a_i\Big[\sum_{p\le q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}(x_p-\mu_p)(x_q-\mu_q)-\frac{a_i}{2}\Big]\\ & \le a_j\Big[\sum_{p\le q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}(x_p-\mu_p)(x_q-\mu_q)-\frac{a_j}{2}\Big]\\ & \qquad\qquad \textit{for all}\ j(\neq i)\Big\},\\ 0, & \textit{otherwise},\end{cases}$$

where

$$a_i=\frac{\sum_{p\leq q}^{k}\frac{\partial\rho_{1(2\cdots k)}}{\partial\sigma_{pq}}v_{ipq}(\phi_0)}{w_i(\phi_0)},\quad i=1,\cdots,L.$$

Aknowlegement. The author is grateful to Professors Yasushi Taga, Shizuoka University, Takashi Yanagawa, Kyushu University and the referee for their useful comments.

REFERENCES

[1] K. ISII and Y. TAGA: On optimal stratifications for multivariate distributions, Skand. Aktuarietidskr. **52** (1969), 24—38.

[2] T. YANAGAWA and K. WAKIMOTO: Estimation of some functional of the population distribution based on a stratified random sample, Ann. Inst. Statist. Math. **24** (1972), 137—151.

[3] K. WAKIMOTO: Stratified random sampling (III); Estimation of the correlation coefficient, Ann. Inst. Statist. Math. **23** (1971), 339—353.

DEPARTMENT OF MATHEMATICS
COLLEGE OF LIBERAL ARTS AND SCIENCE
OKAYAMA UNIVERSITY

(*Received August 24, 1973*)