STRATIFIED RANDOM SAMPLING WITH OPTIMUM ALLOCATION FOR MULTIVARIATE POPULATION

BY HIROJIRO AOYAMA

(Received Dec. 18, 1962)

1. Introduction

We want to consider the optimum allocation for the estimation of population means in stratified random sampling when a multivariate problem is treated. In ordinary sampling survey we must consider many variables at a time, but we treat the estimation problems of these variables independently, and the optimum allocation method by Neyman is obliged to be modified so that a certain compromised size is taken from among different sample sizes for each variable. In these circumstances we must know the correlation coefficients among the variables in each stratum, but if it is possible from previous experience we can get an optimum allocation procedure for the estimation of means. Even if we stratify the sampling units by each variable separately as used in the ordinary case, we can also get an optimum size of allocation for the multivariate case.

2. Allocation method for multivariate case

Let us now consider the estimation problem of the means of a certain multivariate population by stratified random sampling.

At first we treat the case of two variables x and y. Let R denote the number of strata, N and n the sizes of the population and sample, respectively. Further, let N_i and n_i denote the sizes of the ith population and the corresponding sample, respectively. Let $\bar{x}$ and $\bar{y}$ be the sample estimates of the population means $\bar{X}$ and $\bar{Y}$, respectively, and the variances and covariances of $\bar{x}$ and $\bar{y}$ be $\sigma^2_{\bar{x}}$, $\sigma^2_{\bar{y}}$, and $\sigma_{\bar{x}\bar{y}}$, respectively. Then the ellipse of concentration is given by the variance-covariance matrix as follows.

$$\begin{aligned}&(\bar{x}-\bar{X}, \bar{y}-\bar{Y})\begin{pmatrix}\sigma^2_{\bar{x}} & \sigma_{\bar{x}\bar{y}}\\ \sigma_{\bar{x}\bar{y}} & \sigma^2_{\bar{y}}\end{pmatrix}^{-1}(\bar{x}-\bar{X}, \bar{y}-\bar{Y})'\\ &\quad=(\xi, \eta)\Lambda^{-1}(\xi, \eta)' \qquad (1)\end{aligned}$$

where the prime shows the column vector. If we make use of the expressions

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{R}N_i\bar{x}_i\,,\qquad \bar{y}=\frac{1}{N}\sum_{i=1}^{R}N_i\bar{y}_i \tag{2}$$

for the estimation of means $\bar{X}$ and $\bar{Y}$ in stratified random sampling with

$$\bar{x}_i=\frac{1}{n_i}\sum_{l=1}^{n_i}x_{il}\,,\qquad \bar{y}_j=\frac{1}{n_j}\sum_{k=1}^{n_j}y_{jk}\,, \tag{3}$$

we obtain

$$E(\bar{x}\bar{y})=\frac{1}{N^2}E\left(\sum_{i=1}^{R}\sum_{j=1}^{R}N_iN_j\bar{x}_i\bar{y}_j\right)=\frac{1}{N^2}\sum_{i=1}^{R}\sum_{j=1}^{R}N_iN_jE(\bar{x}_i\bar{y}_j) \tag{4}$$

$$E(\bar{x}_i\bar{y}_i)=\frac{N_i-n_i}{N_i-1}\frac{1}{n_i}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}X_{il}Y_{il}\right)+\frac{n_i-1}{n_i}\frac{N_i}{N_i-1}\bar{X}_i\bar{Y}_i \tag{5}$$

and for $i\neq j$

$$E(\bar{x}_i\bar{y}_j)=\frac{1}{N_iN_j}\sum_{l=1}^{N_i}\sum_{k=1}^{N_j}X_{il}Y_{jk}\,. \tag{6}$$

Therefore, we have from (4), (5) and (6)

$$\begin{aligned}\sigma_{\bar{x}\bar{y}}=E(\bar{x}\bar{y})-E(\bar{x})E(\bar{y})&=\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\left\{\frac{N_i-n_i}{N_i-1}\frac{1}{n_i}\left(\frac{1}{N_i}\sum_{l=1}^{N_i}X_{il}Y_{il}\right)\right.\\&\quad\left.+\frac{n_i-1}{n_i}\frac{N_i}{N_i-1}\bar{X}_i\bar{Y}_i\right\}+\frac{1}{N^2}\sum_{i\neq j}^{R}\sum^{R}N_iN_j\cdot\frac{1}{N_iN_j}\sum_{l=1}^{N_i}\sum_{k=1}^{N_j}X_{il}Y_{jk}\\&\quad-\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\bar{X}_i\bar{Y}_i-\frac{1}{N^2}\sum_{i\neq j}^{R}\sum^{R}N_iN_j\bar{X}_i\bar{Y}_j\\&=\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_{x_iy_i}}{n_i}\,.\end{aligned} \tag{7}$$

For $\sigma_{\bar{x}}^2$ and $\sigma_{\bar{y}}^2$, we can get the similar equations to (7). Thus the equation (1) becomes by (7)

$$\begin{aligned}(\xi,\eta)\Lambda^{-1}(\xi,\eta)'&=(\xi^2\sigma_{\bar{y}}^2-2\xi\eta\sigma_{\bar{x}\bar{y}}+\eta^2\sigma_{\bar{x}}^2)/(\sigma_{\bar{x}}^2\sigma_{\bar{y}}^2-\sigma_{\bar{x}\bar{y}}^2)\\&=\left\{\eta^2\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_{y_i}^2}{n_i}-2\xi\eta\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_{x_iy_i}}{n_i}\right.\\&\quad\left.+\eta^2\frac{1}{N^2}\sum_{i=1}^{R}N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_{x_i}^2}{n_i}\right\}\Big/(\sigma_{\bar{x}}^2\sigma_{\bar{y}}^2-\sigma_{\bar{x}\bar{y}}^2)\end{aligned} \tag{1}'$$

and the area of this ellipse of concentration is given by

$$\pi\sqrt{\sigma^2_{\bar x}\sigma^2_{\bar y}-\sigma^2_{\bar x\bar y}}\quad .$$

Hence, to decide the optimum size n_i for the minimum error in the simultaneous estimation we minimize the value $(\sigma^2_{\bar x}\sigma^2_{\bar y}-\sigma^2_{\bar x\bar y})$ under the condition $n=\sum_{i=1}^{R} n_i$.

Let us consider

$$F=\left(\sum_{i=1}^{R} N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma^2_{x_i}}{n_i}\right)\left(\sum_{i=1}^{R} N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma^2_{y_i}}{n_i}\right)$$
$$-\left(\sum_{i=1}^{R} N_i^2\frac{N_i-n_i}{N_i-1}\frac{\sigma_{x_iy_i}}{n_i}\right)^2+\lambda\sum_{i=1}^{R} n_i \tag{8}$$

where λ is a Lagrange's multiplier. Differentiating by n_i and equating to zero we have

$$\frac{N_i^2}{n_i^2}(\sigma^2_{x_i}\sigma^2_{\bar y}+\sigma^2_{y_i}\sigma^2_{\bar x}-2\sigma_{x_iy_i}\sigma_{\bar x\bar y})=\lambda\ . \tag{9}$$

At first we get

$$n_i=N_i\sqrt{\sigma^2_{x_i}\sigma^2_{\bar y}+\sigma^2_{y_i}\sigma^2_{\bar x}-2\sigma_{x_iy_i}\sigma_{\bar x\bar y}}\ /\sqrt{\lambda}\ , \tag{10}$$

hence n_i is apparently proportional to N_i, and also depends on σ_{x_i}, σ_{y_i} and $\sigma_{x_iy_i}$. Putting

$$\sigma_{x_iy_i}=\rho_i\sigma_{x_i}\sigma_{y_i}\qquad\text{and}\qquad N_i-n_i\doteqdot N_i-1$$

we have from (9) after some calculation

$$\frac{a_{i1}}{n_1}+\frac{a_{i2}}{n_2}+\cdots+\frac{a_{iR}}{n_R}=\lambda N^2n_i^2\qquad i=1,2,\cdots,R \tag{11}$$

where

$$a_{ij}=N_i^2N_j^2(\sigma^2_{x_i}\sigma^2_{y_j}+\sigma^2_{x_j}\sigma^2_{y_i}-2\sigma_{x_iy_i}\sigma_{x_iy_i})>0\ . \tag{12}$$

Thus the optimum size of allocation is obtained by solving the simultaneous equations (11). For this purpose put $n_i=1/u_i$. Then

$$a_{i1}u_1+a_{i2}u_2+\cdots+a_{iR}u_R=\frac{\lambda N^2}{u_i^2}\ ,\qquad i=1,2,\cdots,R\ . \tag{13}$$

At first we insert the first approximate value $\boldsymbol{u}^{(1)}=(u_1^{(1)}, u_2^{(1)}, \cdots, u_R^{(1)})=(1, 1, \cdots, 1)$ into the left side of (13) and get the second approximate value $\boldsymbol{u}^{(2)}=(u_1^{(2)}, u_2^{(2)}, \cdots, u_R^{(2)})$ with

$$u_i^{(2)}=\mu\left(\sum_{j=1}^{R} a_{ij}u_j^{(1)}\right)^{-1/2} \tag{14}$$

where $\mu=\sqrt{\lambda}\,N>0$. Inserting this $\boldsymbol{u}^{(2)}$ into the left side of (13) we have the third approximate value $\boldsymbol{u}^{(3)}$ and we iterate this procedure until we can get a limit. As to the existence of this limit we prove it in the following.

From the above assumption we get

$$u_i^{(2)}=\mu\left(\sum_{j=1}^{R} a_{ij}\right)^{-1/2}$$

$$u_i^{(3)}=\mu\left(\sum_{j=1}^{R} a_{ij}u_j^{(2)}\right)^{-1/2}$$

$$\begin{aligned}u_i^{(4)}-u_i^{(2)}&=\mu\left(\sum_{j=1}^{R} a_{ij}u_j^{(3)}\right)^{-1/2}-\mu\left(\sum_{j=1}^{R} a_{ij}\right)^{-1/2}\\&=\mu\left(\sum_{j=1}^{R} a_{ij}u_j^{(3)}\right)^{-1/2}\left(\sum_{j=1}^{R} a_{ij}\right)^{-1/2}\left\{\left(\sum_{j=1}^{R} a_{ij}u_j^{(3)}\right)^{1/2}+\left(\sum_{j=1}^{R} a_{ij}\right)^{1/2}\right\}^{-1}\\&\quad\times\sum_{j=1}^{R} a_{ij}(1-u_j^{(3)})\ .\end{aligned}$$

We can assume $0<u_j^{(3)}\leqq 1$ and we can conclude

$$u_i^{(4)}\geqq u_i^{(2)}\ .$$

In a similar manner we can prove

$$u_i^{(5)}\leqq u_i^{(3)}$$

by the above inequality. Thus, we have generally

$$0<u_i^{(2)}\leqq u_i^{(4)}\leqq {}_i^{(6)}\leqq \cdots <1\ , \tag{15}$$

$$0< \cdots \leqq u_i^{(7)}\leqq u_i^{(5)}\leqq u_i^{(3)}<1\ , \tag{16}$$

hence we can see $\lim u_i^{(2k)}$ and $\lim u_i^{(2k+1)}$ exist. Moreover

$$u_i^{(2k+1)}\geqq u_i^{(2k)} \tag{17}$$

$$u_i^{(2k)}\leqq u_i^{(2k-1)} \tag{18}$$

so we get

$$\begin{aligned}&(u_i^{(2k+3)}-u_i^{(2k+2)})-(u_i^{(2k+1)}-u_i^{(2k)})\\&\qquad=(u_i^{(2k+3)}-u_i^{(2k+1)})+(u_i^{(2k)}-u_i^{(2k+2)})\leqq 0\ .\end{aligned}$$

Hence we have

$$|u_i^{(2k+3)}-u_i^{(2k+2)}|\leqq |u_i^{(2k+1)}-u_i^{(2k)}|$$

and in general (except the case when equalities hold in all stages)

$$\lim_{k\to\infty} |u_i^{(2k+1)} - u_i^{(2k)}| = 0$$

that is, there exists $\lim u_i^{(k)}$.

Example: Let us put

$$\begin{array}{lll} R=2\ , & N_1=N/3\ , & N_2=(2/3)N\ , \\ \sigma_{x_1}^2=1\ , & \sigma_{x_2}^2=3\ , & \rho_1^2=0.3\ , \\ \sigma_{y_1}^2=2\ , & \sigma_{y_2}^2=5\ , & \rho_2^2=0.7\ . \end{array}$$

Then the equations (13) are given as follows:

$$\begin{cases} 2.8u_1+24u_2=\dfrac{\mu^2}{u_1^2} \\ 24u_1+144u_2=\dfrac{\mu^2}{u_2^2}\ . \end{cases}$$

By the above mentioned procedure we obtain the following table:

k	1	2	3	4	5	6	7	8	9	10	11	12
$u_1^{(k)}$	1	0.194	0.645	0.356	0.481	0.414	0.447	0.429	0.438	0.433	0.435	0.434
$u_2^{(k)}$	1	0.077	0.253	0.139	0.187	0.161	0.174	0.167	0.171	0.169	0.170	0.169

Thus we have

$$\frac{1}{n_1}\propto 0.434\ , \qquad \frac{1}{n_2}\propto 0.169$$

that is,

$$\frac{n_1}{n_2}=0.389\ .$$

If we use the ordinary procedure of optimum allocation, we shall have

$$\frac{n_1}{n_2}=\frac{1/3}{2\sqrt{3}/3}=0.289 \qquad \text{for } x$$

and

$$\frac{n_1}{n_2}=\frac{\sqrt{2}/3}{2\sqrt{5}/3}=0.316 \qquad \text{for } y\ .$$

Therefore, when we take a compromised procedure, we have

$$\frac{n_1}{n_2}=\frac{1}{2}(0.289+0.316)\doteqdot 0.303\ .$$

If we can put $\rho_1=\rho_2=0$ in the above example, we have the following equations

$$\begin{cases} 4u_1+44u_2=\dfrac{\mu^2}{u_1^2} \\ 44u_1+480u_2=\dfrac{\mu^2}{u_2^2} \end{cases}$$

from which we obtain

$$\frac{n_1}{n_2}\doteqdot 0.303$$

which is very near to the value in the above compromised case.

3. Method of stratification

As was seen in the preceding section we may treat the equation (8) in order to stratify the sampling units to get better estimates of means. The essential part of the method for a proper stratification is to minimize the expression

$$\begin{aligned} G&=\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{y_i}^2\right)-\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_iy_i}\right)^2 \\ &=\sum_{i=1}^{R}\frac{N_i^4}{n_i^2}\sigma_{x_i}^2\sigma_{y_i}^2(1-\rho_i^2)+\sum_{i\neq j}^{R}\sum^{R}\frac{N_i^2N_j^2}{n_in_j}\sigma_{x_i}\sigma_{y_j}(\sigma_{x_i}\sigma_{y_j}-\rho_i\rho_j\sigma_{x_j}\sigma_{y_i}) \end{aligned} \tag{19}$$

where ρ_i is the correlation coefficient between x and y in the ith stratum. Let us take the second term of G in the last expression. To minimize this term we stratify the sampling units so that

$$\sigma_{x_i}\sigma_{y_j}-\rho_i\rho_j\sigma_{x_j}\sigma_{y_i}$$

approaches to zero. That means the gradient of the regression line of y on x in the ith stratum equals that of x on y in the jth stratum for any i and j. Accordingly, the two regression lines must coincide in each stratum and have the same gradient with correlation coefficient ± 1. In this case the first term of G equals 0 and the value of G also equals 0. In practical problems it is better to use a stratification method which makes ρ_i near to 1. Generally it is more difficult to make ρ_i near to -1 than to 1.

On the other hand, when the sampling units are stratified so that ρ_i becomes 0 in each stratum, which is used in many practical cases, the value of G becomes larger than that in the above-mentioned better stratification.

4. General method for multivariate stratification

We can develop a similar argument in the case of 3 or more variables.

In the case of 3 variables we put

$$F=N^6\begin{vmatrix}\sigma_{\bar x}^2 & \sigma_{\bar x\bar y} & \sigma_{\bar x\bar z}\\ \sigma_{\bar x\bar y} & \sigma_{\bar y}^2 & \sigma_{\bar y\bar z}\\ \sigma_{\bar x\bar z} & \sigma_{\bar y\bar z} & \sigma_{\bar z}^2\end{vmatrix}+\lambda\sum_{i=1}^{R}n_i\,. \tag{20}$$

Differentiating F by n_i and equating to zero we have

$$\sum_{j=1}^{R}\sum_{k=1}^{R}\frac{a_{ijk}}{n_jn_k}=\lambda N^2n_i^2\,,\qquad i=1,2,\cdots,R\,, \tag{21}$$

or putting $u_i=1/n_i$, we obtain

$$\sum_{j=1}^{R}\sum_{k=1}^{R}a_{ijk}u_ju_k=\frac{\lambda N^2}{u_i^2}\,,\qquad i=1,2,\cdots,R \tag{21'}$$

where

$$a_{ijk}=N_i^2N_j^2N_k^2\sum_{(i,j,k)}\begin{vmatrix}\sigma_{x_i}^2 & \sigma_{x_iy_i} & \sigma_{x_iz_i}\\ \sigma_{x_jy_j} & \sigma_{y_j}^2 & \sigma_{y_jz_j}\\ \sigma_{x_kz_k} & \sigma_{y_kz_k} & \sigma_{z_k}^2\end{vmatrix} \tag{22}$$

and the summation covers all permutation (i,j,k). The method to solve this simultaneous equations (21)′ is carried out by the successive approximation in a similar way as in two variables.

As for a better stratification we can proceed in the same way as in two variate case.

Let us put

$$\begin{aligned}G=&\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{y_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{z_i}^2\right)\\ &+2\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_iy_i}\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{y_iz_i}\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_iz_i}\right)\\ &-\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{y_iz_i}\right)^2-\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{y_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_iz_i}\right)^2\\ &-\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{z_i}^2\right)\left(\sum_{i=1}^{R}\frac{N_i^2}{n_i}\sigma_{x_iy_i}\right)^2\\ =&\sum_{(i,j,k)}\frac{N_i^2N_j^2N_k}{n_in_jn_k}\begin{vmatrix}\sigma_{x_i}^2 & \sigma_{x_iy_i} & \sigma_{x_iz_i}\\ \sigma_{x_jy_j} & \sigma_{y_j}^2 & \sigma_{y_jz_j}\\ \sigma_{x_kz_k} & \sigma_{y_kz_k} & \sigma_{z_k}^2\end{vmatrix}\,.\end{aligned} \tag{23}$$

In order to minimize the value of this expression G we try to minimize these values of determinants for any i, j and k. This is accomplished when the regression planes in all strata have the same direction cosines with multiple correlation coefficient 1. Because if the value of the determinant for $j=k$, for example, is zero, then the direction cosines $R_{11}^{(j)}/\sigma_{x_j}$: $R_{12}^{(j)}/\sigma_{y_j}$: $R_{13}^{(j)}/\sigma_{z_j}$ of the normal line of regression plane of x on y and z in the jth stratum satisfy the following equation:

$$\frac{R_{11}^{(j)}}{\sigma_{x_j}}\sigma_{x_i}+\frac{R_{12}^{(j)}}{\sigma_{y_j}}\rho_{x_iy_i}\sigma_{y_i}+\frac{R_{13}^{(j)}}{\sigma_{z_j}}\rho_{x_iz_i}\sigma_{z_i}=0 \tag{24}$$

where $R_{lm}^{(j)}$ is the cofactor of ρ_{lm} in the determinant

$$R^{(j)}=\begin{vmatrix} 1 & \rho_{x_jy_j} & \rho_{x_iz_i} \\ \rho_{x_jy_j} & 1 & \rho_{y_jz_j} \\ \rho_{x_jz_j} & \rho_{y_jz_j} & 1 \end{vmatrix}$$

in the jth stratum.

On the other hand if the multiple correlation coefficient equals 1 in the ith stratum, then $R^{(i)}=0$, hence

$$R_{11}^{(i)}+R_{12}^{(i)}\rho_{x_iy_i}+R_{13}^{(i)}\rho_{x_iz_i}=R^{(i)}=0$$

that is,

$$\frac{R_{11}^{(i)}}{\sigma_{x_i}}\sigma_{x_i}+\frac{R_{12}^{(i)}}{\sigma_{y_i}}\rho_{x_iy_i}\sigma_{y_i}+\frac{R_{13}^{(i)}}{\sigma_{z_i}}\rho_{x_iz_i}\sigma_{z_i}=0\ . \tag{25}$$

As from (24) and (25) we obtain that two direction cosines are proportional to each other, we can stratify the sampling units so that the regression planes in all strata become parallel to each other with multiple correlation coefficient 1.

THE INSTITUTE OF STATISTICAL MATHEMATICS