USE OF DISTANCE FUNCTION TO OPTIMIZE SAMPLE ALLOCATION IN MULTIVARIATE SURVEYS: A NEW PERSPECTIVE

M.A. Rahim, Retired
44 Birchview Road, Ottawa, K2G 3G6

KEY WORDS: *Optimum allocation, stratified design.*

1. INTRODUCTION

In a stratified random sampling design involving multiple response variables, optimizing sample allocation to different strata is an old problem. In the literature this problem has been dealt with partially. One approach is based on averaging the stratum variances of the variables and using it to optimize the allocations. Since the choice of weights in averaging is arbitrary, the optimality property remains unclear. Notable works are those of Dalenius (1953), Yates (1960), Hartley (1965), and Kish (1976). The other approach is concerned with minimizing the cost of survey while the sampling errors of the estimates do not exceed certain preassigned upper bounds. This is accomplished by convex programming. Notable works are those of Yates (1960), Hartley (1965), Kokan and Khan (1967), Chatterjee (1972) and Bethel (1989). In this approach, however, no optimality criterion is offered when the cost of survey is preassigned which is usually the case. Further, convex programming is suitable when the number of variables and strata are small. When the number of variables is large, say over hundred, convex programming is impractical. The cost involved to ensure that all the sampling error constraints are satisfied becomes unacceptable.

To obviate these difficulties, particularly from a practioner's point of view, it is necessary to deal with the optimization problem in a more comprehensive manner. To that end we have to take into account all of the following considerations.

(i) optimality criterion should be based on an aggregate measure of sampling errors i.e. joint sampling error of all the estimates.

(ii) when cost of survey is preassigned we should term the allocation as optimum for which the joint sampling error of estimates is minimum.

(iii) when an upper bound to the joint sampling error of estimates is preassigned we should term the allocation as optimum for which the cost of survey is minimum.

(iv) when the number of response variables is small and we wish to set upper bounds to each individual sampling errors of estimates, we should term the allocation as optimum for which the cost is minimum while satisfying each individual sampling error constraints.

(v) keeping the above considerations as the basis we should be able to develop allocation formula such that it becomes identical with the Neyman (1934) allocation in the single variable case.

In this paper a simple weighted Eucledean distance function is proposed as a measure of joint sampling error of all the estimates. Formulae for sample allocation are then derived which meets the above requirements.

2. MATHEMATICAL PRELIMINARIES

In a stratified random sampling design n_i denotes the number of units sampled from a

population of N_i units in the i-th stratum; $i = 1, 2, \ldots, I$. We assume that we have records of values of J independent variables, Y_j; $j = 1, 2, \ldots, J$; on these units. $\bar{Y}_{ij}$, $\bar{y}_{ij}$; S_{ij}^2, s_{ij}^2; denote the mean and variance of the j-th variable in i-th stratum of the population and sample respectively. The population mean $\bar{Y}_j$ is estimated by a function $\bar{y}_j = \sum_i (N_i \bar{y}_{ij})/N$; $N = \sum_i N_i$, $E(\bar{y}_j) = \bar{Y}_j$. Variance of the estimate is given by $V(\bar{y}_j) = \sum_i (N_i^2 S_{ij}^2)/(N^2 n_i) - \sum_i (N_i^2 S_{ij}^2)/N^2$. Following Bethel (1989) – to make our subsequent results comparable – we will neglect the second term and write the variance approximately as $V(\bar{y}_j) = \sum_i (N_i^2 S_{ij}^2)/(N^2 n_i)$. In that case the coefficient of variation (cv) of $\bar{y}_j$ is $cv(\bar{y}_j) = \left\{ \sum_i (N_i^2 S_{ij}^2)/(\bar{Y}_j^2 N^2 n_i) \right\}^{\frac{1}{2}}$ which is used as a measure of the sampling error of estimate. The cost function is written as $g(\mathbf{x}) = \sum_i C_i n_i$; C_i being the cost of enumeration per unit in the i-th stratum. Writing $n_i = 1/x_i$, $A_{ij} = (N_i^2 S_{ij}^2)/(\bar{Y}_j^2 N^2)$, we can write $cv^2(\bar{y}_j) = \sum_i A_{ij} x_i$. An inequality relation $cv(\bar{y}_j) \leq \nu_j$; $\nu_j > 0$, can also be written as $\{cv^2(\bar{y}_j)/\nu_j^2\} \leq 1$ or $\sum_i a_{ij}\, x_i \leq 1$ where $a_{ij} = A_{ij}/\nu_j^2$. For the sake of brevity we shall also use the following notations and symbols:

(i) $\mathbf{A}$ will denote the matrix $[A_{ij}]_{I \times J}$: $A_{ij} = (N_i^2 S_{ij}^2)/(\bar{Y}_j^2 N^2)$.

(ii) $\mathbf{A_j} = (A_{1j}, A_{2j}, \ldots A_{IJ})'$ will denote the j-th column vector of $\mathbf{A}$.

(iii) $\mathbf{a}$ will denote the matrix $[a_{ij}]_{I \times J}$: $a_{ij} = A_{ij}/\nu_j^2$.

(iv) $\mathbf{a_j} = (a_{1j}, a_{2j}, \ldots a_{IJ})'$ will denote the j-th column vector of $\mathbf{a}$.

(v) $\mathbf{x}$ will denote the vector $\mathbf{x} = (x_1, x_2, \ldots, x_I)'$.

(vi) Notation $g(\mathbf{x})_{\mathbf{a_j'x} \leq 1}$ will denote a constrained function $g(\mathbf{x})$ under the constraint $a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{IJ}x_I \leq 1$.

(vii) Notation $g(\mathbf{x})_{\mathbf{a_j'x} \leq 1}$; $j = 1, 2, \ldots, J$ will denote a constrained function $g(\mathbf{x})$ under J different constraints $\mathbf{a_1'x} \leq 1$, $\mathbf{a_2'x} \leq 1, \ldots, \mathbf{a_J'x} \leq 1$.

(viii) Notation $g(\mathbf{x})_{\sum_j (\mathbf{a_j'x} - 1)} \leq 0$ will mean a constrained function $g(\mathbf{x})$ under the single constraint $(\mathbf{a_1'x} - 1) + (\mathbf{a_2'x} - 1) + \ldots + (\mathbf{a_J'x} - 1) \leq 0$. This single constraint will also be referred as an aggregate constraint as opposed to an individual constraint like $(\mathbf{a_k'x} - 1) \leq 0$.

3. DEFINITION OF A DISTANCE FUNCTION $D_{(R)}$ AND OPTIMALITY CRITERIA

If $cv(\bar{y}_j)$ is plotted along the j-th axis of a J dimensional space then $\sum_j cv^2(\bar{y}_j)$ is simply the square of the Eucledean distance that can be used as a measure of the joint sampling error of the J independent estimates $\bar{y}_j$; $j = 1, 2, \ldots, J$. In an actual survey we may be more concerned about the sampling error of $\bar{y}_j$ as compared to that of $\bar{y}_k$. To accommodate this situation we assign certain weights to each of the sampling errors of estimates and define, in general, a weighted distance function $D_{(R)}$ as

$$D_{(R)} = W_1 cv^2(\bar{y}_1) + W_2 cv^2(\bar{y}_2) + \ldots, + W_J cv^2(\bar{y}_J) \qquad (3.1)$$

where $W_j > 0$ is an arbitrary weight. Based on this $D_{(R)}$ we define the optimality criteria as follows:

(i) Under a preassigned upper bound C_0 to the cost of survey, the allocation will be termed as optimum for which $D_{(R)}$ is minimum subject to the condition $g(\mathbf{x}) \leq C_0$.

(ii) Under a preassigned upper bound ν^2 to the joint sampling error of estimate, the allocation will be termed as optimum for which the cost of survey is minimum subject to the condition $D_{(R)} \leq \nu^2$.

(iii) When the number of response variables is small we may wish to assign upper bounds ν_j, $j = 1, 2, \ldots, J$, to the individual sampling errors of estimates $\bar{y}_j$. In that case the allocation will be termed as optimum for which the cost $g(\mathbf{x})$ is minimum subject to the conditions $cv^2(\bar{y}_j) \leq \nu_j^2$; $j = 1, 2, \ldots, J$.

4. ALLOCATION UNDER PREASSIGNED UPPER BOUND TO THE COST OF SURVEY

In this case, under the optimality criterion laid above we have to minimize the joint sampling error $D_{(R)}$ with respect to $x_1, x_2, \ldots, x_I$ under the constraint $g(\mathbf{x}) \leq c_0$. In other words we have to find the minimum of a constrained function $D_{(R)g(\mathbf{x})\leq c_0}$. This is done by finding the free minimum of a function $F(\mathbf{x}, \gamma)$ of $(I+1)$ variables x_i; $i = 1, 2, \ldots, I$ and γ, $\gamma > 0$, where

$$F(\mathbf{x}, \gamma) = D_{(R)} + \gamma(g(\mathbf{x}) - c_0)$$
$$= \sum_j W_j \mathbf{A}_j' \mathbf{x} + \gamma\Big(\sum_i \frac{c_i}{x_i} - c_0\Big) \quad (4.1)$$

To find this minimum we solve the equations

$$\frac{\partial F(\mathbf{x}, \gamma)}{\partial x_i} = \sum_j W_j A_{ij} - \gamma(c_i/x_i^2) = 0$$
$$i = 1, 2, \ldots, I \quad (4.2)$$

$$\frac{\partial F(\mathbf{x}, \gamma)}{\partial \gamma} = \sum_i (c_i/x_i) - c_0 = 0 \quad (4.3)$$

and obtain the allocation in terms of x_i as

$$x_i = \Big\{\sum_i \sqrt{\Big(\sum_j W_j A_{ij}\Big) c_i}\Big\} \Big/ \Big\{c_0 \sqrt{\Big(\sum_j W_j A_{ij}\Big)/c_i}\Big\}; \quad (4.4)$$
$$i = 1, 2, \ldots, I$$

With large number of variables we treat all as of equal importance. In that case we can write $W_j = 1$ for all j. Then, assuming $c_i = c$ (4.4) reduces to the simple form

$$x_i = \Big\{\sum_i \sqrt{\sum_j A_{ij}}\Big\} \Big/ \Big\{n \sqrt{\sum_j A_{ij}}\Big\} \quad (4.5)$$

Remembering that $A_{ij} = (N_i^2 S_{ij}^2)/(\bar{Y}_j^2 N^2)$ and dropping the subscript j, (4.5) further reduces to $n_i = \{N_i S_i / \sum_i N_i S_i\} n$ which is the well known Neyman allocation in the single variable case.

5. ALLOCATION UNDER PREASSIGNED UPPER BOUND TO THE JOINT SAMPLING ERROR OF ESTIMATES

In this case our problem is to minimize the cost of survey $g(\mathbf{x})$ with respect to $x_1, x_2, \ldots, x_I$ under the constraint $D_{(R)} \leq \nu^2$. In other words we have to find the minimum of a constrained function $g(\mathbf{x})_{D_{(R)}\leq \nu^2}$. This is done by finding the free minimum of a function $F(\mathbf{x}, \gamma)$ of $(I+1)$ variables x_i; $i = 1, 2, \ldots, I$ and $\gamma, \gamma > 0$, where

$$F(\mathbf{x}, \gamma) = g(\mathbf{x}) + \gamma(D_{(R)} - \nu^2) \quad (5.1)$$
$$= g(\mathbf{x}) + \gamma\Big(\sum_j W_j \mathbf{A}_j' \mathbf{x} - \nu^2\Big)$$

To find this minimum we solve the equations

$$\frac{\partial F(\mathbf{x}, \gamma)}{\partial x_i} = -\frac{c_i}{x_i^2} + \gamma \sum_j W_j A_{ij} = 0 \quad (5.2)$$

$$\frac{\partial F(\mathbf{x}, \gamma)}{\partial \gamma} = \sum_j W_j \mathbf{A}_j' \mathbf{x} - \nu^2 = 0 \quad (5.3)$$

and obtain the allocation in terms of x_i as

$$x_i = \frac{\nu^2\sqrt{c_i}}{\sqrt{\sum_j W_j A_{ij}}\sum_i \sqrt{c_i \sum_j W_j A_{ij}}}; \quad i = 1, 2, \ldots, I \qquad (5.4)$$

With large number of variables we can, as before, assume the weights $W_j = 1$. Then (5.4) reduces to the simple form

$$x_i = \left\{\nu^2\sqrt{c_i}\right\} \Big/ \left\{\sqrt{\sum_j A_{ij}}\sum_i \sqrt{c_i \sum_j A_{ij}}\right\}. \qquad (5.5)$$

For $j = 1$, dropping the subscript j and remembering that $a_i = A_i/\nu^2$, (5.5) further reduces to $x_i = \sqrt{c_i}/\left\{\sqrt{a_i}\sum_i \sqrt{c_i a_i}\right\}$ which again is the Neyman allocation in the single variable case.

6. ALLOCATION UNDER PREASSIGNED UPPER BOUNDS TO THE INDIVIDUAL SAMPLING ERRORS OF ESTIMATES

If the number of variables is small we would like to preassign upper bounds to each individual sampling errors of estimates. In that case our problem is to find the minimum of the constrained function $g(\mathbf{x})_{cv^2(\bar{y}_j)\leq\nu_j^2}$ i.e. $g(\mathbf{x})_{\mathbf{a}_j'\mathbf{x}\leq 1}$; $j = 1, 2, \ldots, J$ under J individual constraints. For an exact mathematical solution we would have to find the free minimum of a function $F(\mathbf{x}, \boldsymbol{\lambda})$ of $(I + J)$ variables x_i and λ_j; $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$ where

$$F(\mathbf{x}, \boldsymbol{\lambda}) = g(\mathbf{x}) + \sum_j \lambda_j(\mathbf{a}_j'\mathbf{x} - 1) \qquad (6.1)$$

It has been shown (Kokan and Khan, 1967) that there exist a unique set of values $(x_1^*, x_2^*, \ldots, x_I^*, \lambda_1, \lambda_2, \ldots, \lambda_J)$ for which $F(\mathbf{x}, \boldsymbol{\lambda})$ is minimum although a mathematical solution by solving the $(I + J)$ equations $\partial F(\mathbf{x}, \boldsymbol{\lambda})/\partial x_i = 0$, $\partial F(\mathbf{x}, \boldsymbol{\lambda})/\partial \lambda_j = 0$, is not possible because of nonlinearity. We therefore take an alternative route using the distance function $D_{(R)}$.

Since the individual constraints $cv^2(\bar{y}_i) \leq \nu_j^2$ can also be written as $W_j cv^2(\bar{y}_j) \leq W_j\nu_j^2$, by summing over j we can write an aggregate constraint as $\sum_j W_j cv^2(\bar{y}_j) \leq \sum_j W_j\nu_j^2$ or $D_{(R)} \leq \sum_j W_j\nu_j^2$. In the formula at (5.4) putting $\nu^2 = \sum_j W_j\nu_j^2$ we immediately get an expression for x_i as

$$x_i = \frac{\sqrt{c_i}\sum_j W_j\nu_j^2}{\sqrt{\sum_j W_j A_{ij}}\sum_i \sqrt{c_i \sum_j W_j A_{ij}}}; \quad i = 1, 2, \ldots, I \qquad (6.2)$$

Writing $k_j = W_j\nu_j^2$ and remembering that $cv^2(\bar{y}_j) = \nu_j^2\mathbf{a}_j'\mathbf{x}$ the individual constraint $W_j cv^2(\bar{y}_j) \leq W_j\nu_j^2$ can be written as $k_j\mathbf{a}_j'\mathbf{x} \leq k_j$. We can then rewrite our problem in an alternative form namely, minimize $g(\mathbf{x})_{k_j\mathbf{a}_j'\mathbf{x}\leq k_j}$; $j = 1, 2, \ldots, J$. This in turn means we have to find the free minimum of a function $F(\mathbf{x}, \mathbf{k})$ of $(I+J)$ variables x_i and k_j; $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$, where

$$F(\mathbf{x}, \mathbf{k}) = g(\mathbf{x}) + \sum k_j(\mathbf{a}_j'\mathbf{x} - 1) \qquad (6.3)$$

Since λ_j is unique, comparing (6.1) and (6.3) we conclude that we must have $k_j = \lambda_j$ or $W_j\nu_j^2 = \lambda_j$ or $W_j = \lambda_j/\nu_j^2$. Substituting the value of W_j in (6.2) and remembering $A_{ij} = \nu_j^2 a_{ij}$ we can directly write an expression for x_i as

$$x_i = \frac{\sqrt{c_i}}{\sqrt{\sum_j \alpha_j^* a_{ij}}\sum_i \sqrt{c_i \sum_j \alpha_j^* a_{ij}}}; \quad i = 1, 2, \ldots, I \qquad (6.4)$$

where $\alpha_j^* = \lambda_j/\sum_j \lambda_j$.

Note that this result is identical with what was earlier obtained by Bethel (1989). He used Khun-Tucker Theorem (1951) and wrote the expression (6.4) following an analogical expression in the univariate case, where as using the distance function $D_{(R)}$ the result follows directly and easily.

Also note that x_i is expressed in terms of a function $(\lambda_j / \sum_j \lambda_j)$, not in terms of individual values of λ_j that would result from an exact mathematical solution. Hence x_i, as expressed in (6.4), is still not the optimum solution x_i^*. However the true value of the function $\alpha_j^* = (\lambda_j / \sum_j \lambda_j)$ can be successively approximated by computer programming for which several software package – including one outlined by Bethel (1989) – is available. And therefore x_i can be brought sufficiently closer to x_i^* so as to satisfy all the individual sampling error constraints.

7. CONCLUSION

In this paper we have provided a comprehensive, unified treatment to the problem of optimizing sample allocation in the multivariate case using a distance function $D_{(R)}$ as a measure of joint sampling error of all the estimates. Particularly, from a practioner's point of view this will be helpful to calculate optimum sample allocation to different strata under any of the following situations, namely:

(i) When cost of survey is preassigned and we want to minimize the joint sampling error of all the estimates.

(ii) When an upper bound to the joint sampling error is preassigned and we want to minimize the cost of survey.

(iii) When upper bounds to each individual sampling errors of estimates are preassigned and we want to minimize the cost of survey.

The formulae provided are easily applicable particularly when the number of variables are very large and also in the special case when the number is small.

REFERENCES

Bethel, J.W. (1989). "Sample Allocation in Multivariate Surveys," *Survey Methodology*, 15, No. 1, 46-57.

Chatterjee, S. (1972), "A Study of Optimum Allocation in Multivariate Stratified Surveys," *Skandinavisk Actuarietidskrift*, 55, 73-80.

Dalenius, T. (1953), "The Multivariate Sampling Problem," *Skandinavisk Actuarietidskrift*, 36, 92-102.

Hartley, H.O. (1965), "Multiple purpose optimum allocation in stratified sampling," Proceedings of the Social Statistics Section, American Statistical Association, 258-261.

Kish, L. (1976), "Optima and proxima in linear sample designs," *Journal of the Royal Statistical Society A*, 139, 80-95.

Kokan, A.R. and Khan, S. (1967), "Optimum allocation in multivariate surveys: an analytical solution," *Journal of the Royal Statistical Society B*, 29, 115-125.

Kuhn, H.W. and Tucker, A.W. (1951), "Nonlinear Programming," Proceedings of the 2nd Berkeley Symposium Mathematical Statistics and Probability.

Yates, F. (1960). Sampling Methods for census and surveys. London: Charles Griffin & Co.