S. Pac. J. Nat. Sci., (2003), Vol. 21, pp. 91 – 95
Section B – Short Communication

A note on optimum allocation in multivariate stratified sampling

M. G. M. Khan[1] and **M. J. Ahsan**[2]
Department of Mathematics and Computing Science,
University of the South Pacific, Suva, FIJI,
[1]*E-mail: khan_mg@usp.ac.fj*

[2]*Aligarh Muslim University, Aligarh, INDIA*

ABSTRACT

In stratified random sampling when several characteristics are to be estimated simultaneously, an allocation that is optimum for one characteristic may be far away from optimum for others. To resolve this conflict the authors formulate the problem of determining optimum compromise allocation as a nonlinear programming problem ($NLPP$). The allocation obtained is optimum in the sense that it minimizes the sum of weighted variances of the estimated population means of the characteristics subject to a fixed sampling cost. The formulated NLPP is treated as multistage decision problem and solved using dynamic programming technique. A numerical example is presented to illustrate the computational details.

Keywords: Sample allocation, multivariate stratified random sampling, nonlinear programming problem, dynamic programming technique.

I. INTRODUCTION

Stratified sampling is the most popular among various sampling designs that are extensively used in sample survey. When a stratified sampling is to be used a sampler has to deal with three basic problems such as (i) the problem of determining the number of strata, (ii) the problem of cutting the stratum boundaries and (iii) the problem of optimum allocation of sample sizes to various strata. In this present paper the problem (iii) when more than one characteristics are under study is discussed.

The problem of allocation with more than one characteristics in stratified sampling is conflicting in nature, as the best allocation for one characteristic will not in general be best for others. Some compromise must be reached to obtain an allocation that is efficient for all characteristics. The problem was first considered by Neyman (1934). He pointed put that an allocation would be reasonably efficient for all characteristics if the characteristics themselves are positively correlated. However, in the absence of a strong positive correlation between characteristics when individual optimum allocation may differ a lot and there may be no obvious compromise, many authors such as Neyman (1934), Geary (1949), Dalenius (1957), Ghosh (1958), Aoyama (1963), Chatterjee (1967), Kokan and Khan (1967), Bethel (1989), Jahan, Khan and Ahsan (1994), Khan, Jahan and Ahsan (1997), etceteras, have made attempts for an acceptable allocation by either suggesting new criteria or exploring existing criteria further.

In this paper a more general problem of obtaining optimum allocation, when the cost of survey is fixed, is formulated as a nonlinear programming problem ($NLPP$) to minimize the sum of weighted variances of estimated population means. Since the functions involved are separable with respect to stratum sample size, the $NLPP$ is treated as a multistage decision problem and an explicit solution procedure using dynamic programming technique is presented.

II. THE PROBLEM

Let p independent characteristics are under study in a survey of a population with L strata. The variance of the stratified sample mean $\overline{y}_{jst}$, an unbiased estimate of population mean $\overline{Y}_j$, for jth characteristic is

$$V(\overline{y}_{jst}) = \sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{N_h};\ j = 1, 2, ..., p.$$

In a problem of stratification the loss in precision in the estimate of a characteristic increases, if the characteristic in a stratum is not internally homogeneous. To refrain from this increase in loss of precision the authors conjecture the following. If the jth characteristic in hth stratum ($h = 1, 2, ..., L$) is more heterogeneous, it produces more loss in precision in the estimate of stratum mean, as the value of stratum variance S_{jh}^2 for that characteristic is expected to be high. This results a high sampling variance $V(\overline{y}_{jst})$. A way to restrain this increase in the loss of precision is to assign a maximum weight w_j to jth characteristic as

$$w_j = \max(a_{j1}, a_{j2}, ..., a_{jL}). \qquad (1)$$

Where a_{jh} are the weights for jth ($j = 1, 2, ..., p$) characteristic in hth ($h = 1, 2, ..., L$) stratum and are obtained in proportion to their stratum variances S_{jh}^2, that is, $a_{jh} \, \alpha \, S_{jh}^2$. Letting $\sum_{j=1}^{p} a_{jh} = 1$, a_{jh} are worked out as

$$a_{jh} = \frac{S_{jh}^2}{\sum_{j=1}^{p} S_{jh}^2}; \; j = 1, 2, ..., p \;\text{ and }\; h = 1, 2, ..., L. \quad (2)$$

When the cost of survey is prefixed, it may be a reasonable criterion for determining an optimum allocation is to maximize the sum of weighted variances of the estimated population means, that is,

$$\text{Minimize } \sum_{j=1}^{p} w_j V(\overline{y}_{jst}). \quad (3)$$

Note that (3) is unlike the criterion due to Yates (1960) where the weights, w_j, are specified according to the importance of jth characteristics and are then used to form a linear combination of the variances $V(\overline{y}_{jst})$. The weakness of this compromise allocation is the arbitrariness in the choice of the importance weights, w_j.

For a fixed cost C, when n_h ($h = 1, 2, ..., L$) is the required allocation, c_0 is the overhead cost and c_h is the cost of measuring of all characteristics in hth stratum, the problem of determining an optimum allocation may be expressed as the following $NLPP$:

$$\text{Minimize } \sum_{j=1}^{p} w_j V(\overline{y}_{jst}) = \sum_{j=1}^{p} w_j \sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{n_h} - \sum_{j=1}^{p} w_j \sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{N_h}$$

$$\text{subject to } \; c_0 + \sum_{h=1}^{L} c_h n_h \le C \;\text{ and }\; 2 \le n_h \le N_h; \; h = 1, 2, ..., L \quad (4)$$

The bounded variable restrictions $2 \le n_h \le N_h$; $h = 1, 2, ..., L$ are imposed in the $NLPP$ (4) to meet the problem of estimating stratum variances and over sampling.

For the purpose of minimization the second term of objective function in (4) could be ignored, as it is independent of n_h. Further taking w_j inside the summation $\sum_{h=1}^{L}$, interchanging the order of summations and letting $A_h^2 = \sum_{j=1}^{p} w_j S_{jh}^2$, the $NLPP$ (4) may be rewritten as:

$$\text{Minimize } \sum_{h=1}^{L} \frac{W_h^2 A_h^2}{n_h}$$

$$\text{subject to } \; \sum_{h=1}^{L} c_h n_h \le C_0$$

$$\text{and } \; 2 \le n_h \le N_h; \; h = 1, 2, ..., L \quad (5)$$

where $C_0 = C - c_0$. Note that if $c_h = 1$ ($h = 1, 2, ..., L$), the $NLPP$ (5) reduces to the problem of minimizing the sum of weighted variances of the estimated population means subject to fixed sample size.

III. THE SOLUTION

It is observed that the objective function and the constraints of the $NLPP$ (5) are separable functions of n_h. It allows us to treat (5) as a sequence of interrelated multistage decision making problem that takes place in L stages. The dynamic programming technique may be used to solve (5) by dividing the L-stage and L-variable problem into L-stage single variable problems. The kth ($k = 1, 2, ..., L$) stage provides the optimum allocation, n_k^*, for kth stratum.

Consider the following sub problem of first k strata:

$$\text{Minimize } \sum_{h=1}^{k} \frac{W_h^2 A_h^2}{n_h}$$

$$\text{subject to } \; \sum_{h=1}^{k} c_h n_h \le C_k$$

$$\text{and } \; 2 \le n_h \le N_h; \; h = 1, 2, ..., k \quad (6)$$

where C_k is the available budget for the first k strata satisfying $C_k \le C_0$ and $k \le L$. Let $f(C_k)$ denotes the minimum value of the objective function of (6), that is,

$$f(C_k) \;=\; \min \Big[\sum_{h=1}^{k} \frac{W_h^2 A_h^2}{n_h} \;\Big|\; \sum_{h=1}^{k} c_h n_h \le C_k \;\text{ and }\; 2 \le n_h \le N_h; \; h = 1, 2, ..., k\Big].$$

With the above definition the problem (5) is equivalent to find $f(C_L)$ recursively by finding $f(C_k)$ for $k = 1, 2, ..., L$ and for all feasible C_k satisfying

$$2 \sum_{h=1}^{k} c_h \le C_k \le C_0 - 2 \sum_{h=k+1}^{L} c_h. \quad (7)$$

We may write

$$f(C_k) = \min [\frac{W_k^2 A_k^2}{n_k} + \sum_{h=1}^{k-1} \frac{W_h^2 A_h^2}{n_h} \mid \sum_{h=1}^{k-1} c_h n_h \leq C_k - c_k n_k \text{ and } 2 \leq n_h \leq N_h;\ h = 1, 2, ..., k].$$

For a fixed value of n_k over

$$2 \leq n_k \leq \min\ (C_k^{'}, N_k) \tag{8}$$

where $C_k^{'}$ is the maximum possible sample size that can be drawn from kth stratum within the available budget C_k, that is,

$$C_k^{'} = \frac{C_k - 2\sum_{h=1}^{k-1} c_h}{c_k}. \tag{9}$$

The function $f(C_k)$ is given by

$$f(C_k) = \frac{W_k^2 A_k^2}{n_k} + \min[\sum_{h=1}^{k-1} \frac{W_h^2 A_h^2}{n_h} \mid \sum_{h=1}^{k-1} c_h n_h \leq C_k - c_k n_k \text{ and } 2 \leq n_h \leq N_h;\ h = 1, 2, ..., k-1] \tag{10}$$

By the definition of $f(C_k)$, the term inside [ ] in (10) is the value of $f(C_{k-1})$. Consequently, if $f(C_{k-1})$ is known for all feasible C_{k-1} satisfying (7), the recursive relationship relating the functions $f(C_1),\ f(C_2), ..., f(C_k)$ for the problem (6) is

$$f(C_k) = \min_{2 \leq n_k \leq \min(C_k^{'}, N_k)} \left[\frac{W_k^2 A_k^2}{n_k} + f(C_{k-1})\right] \tag{11}$$

Initially we set $f(C_0) = 0$. The $NLPP$ (5) is equivalent to find $f(C_L)$. If (11) is solved recursively for each $k = 1, 2, ..., L,\ f(C_L)$ is solved. The optimum allocation n_L^* is obtained from $f(C_L)$, n_{L-1}^* is obtained from $f(C_{L-1})$ and so on until finally n_1^* is obtained.

IV. NUMERICAL EXAMPLE

To illustrate the suggested procedure discussed in earlier sections, the authors present the following example. For this purpose, data from Sukhatme et al (1984) have been used. The survey was conducted on a population of size 4190. The data are reproduced in Table I. It is assumed that the costs of measurement c_h in various strata for each unit are same and $c_h = 1$ unit and the total cost (excluding overhead cost c_0) available for measurements, $C_0 = 1000$ units. If the estimated s_{jh}^2 are used as the true values of S_{jh}^2, the weights a_{jh}; $j = 1, 2$ and $h = 1, 2, 3, 4$ are worked out from (2) as:

TABLE I: Data for four strata and two characteristics

h	N_h	W_h	s_{1h}^2	s_{2h}^2
1	1419	0.3387	4817.72	130121.15
2	619	0.1477	6251.26	7613.52
3	1253	0.2990	3066.16	1456.40
4	899	0.2146	56207.25	66977.72

$$a_{jh} = \begin{pmatrix} 0.0357 & 0.9643 \\ 0.4509 & 0.5491 \\ 0.6780 & 0.3220 \\ 0.4563 & 0.5437 \end{pmatrix}$$

From (1) the weights to be assigned to the characteristics are $w_1 = 0.6780$ and $w_2 = 0.9643$. The values of A_h^2; $h = 1, 2, 3, 4$ given by $A_h^2 = \sum_{j=1}^{p} w_j S_{jh}^2$ are obtained as $A_1^2 = 128742.2391$, $A_2^2 = 11580.0716$, $A_3^2 = 3483.2630$, and $A_4^2 = 102695.1309$.

The $NLPP$ (5) for the given example could be expressed as:

$$\begin{aligned} &\text{Minimize } Z(n_1, n_2, n_3, n_4) = \frac{14769.0123}{n_1} + \frac{252.6226}{n_2} + \frac{311.4072}{n_3} + \frac{4729.4353}{n_4} \\ &\text{subject to } \quad n_1 + n_2 + n_3 + n_4 \leq 1000 \\ &2 \leq n_1 \leq 1419, \\ &2 \leq n_2 \leq 619, \\ &2 \leq n_3 \leq 1253, \\ &2 \leq n_4 \leq 899, \end{aligned} \tag{12}$$

To solve the $NLPP$ (12) using the procedure discussed in Section 3 by dynamic programming technique we have

C_k: $k = 1, 2, 3, 4$ and their limits defined earlier are:

$$C_4 = n_1 + n_2 + n_3 + n_4 = 1000\,,$$

$$C_3 = C_4 - n_4; \quad 6 \le C_3 \le 998\,,$$

$$C_2 = C_3 - n_3; \quad 4 \le C_2 \le 996\,,$$

$$C_1 = C_2 - n_2; \quad 2 \le C_1 \le 994\,.$$

For the first-stage problem ($k = 1$)

$$\begin{aligned} f(C_1) &= \min_{2 \le n_1 \le \min(C_1', N_1)} \left[\frac{14769.0123}{n_1} + f(C_0)\right] \\ &= \min_{2 \le n_1 \le \min(\frac{C_1 - 0}{1}, 1419)} \left[\frac{14769.0123}{n_1}\right], \\ &\qquad \text{because } f(C_0) = 0 \\ &= \min_{2 \le n_1 \le C_1} \left[\frac{14769.0123}{n_1}\right] \end{aligned}$$

$$\Rightarrow \quad f(C_1) = \frac{14769.0123}{C_1}, \ \text{ at } n_1^* = C_1 \qquad (13)$$

For the second-stage problem ($k = 2$)

$$\begin{aligned} f(C_2) &= \min_{2 \le n_2 \le \min(C_2', N_2)} \left[\frac{252.6226}{n_2} + f(C_1)\right] \\ &= \min_{2 \le n_2 \le \min(\frac{C_2 - 2c_1}{c_2}, N_2)} \left[\frac{252.6226}{n_2} + \frac{14769.0123}{C_1}\right] \\ &= \min_{2 \le n_2 \le \min(\frac{C_2 - 2}{1}, 619)} \left[\frac{252.6226}{n_2} + \frac{14769.0123}{C_2 - n_2}\right] \end{aligned}$$

The optimal decision at this stage is obtained by using classical method of optimization for minimizing the quantity inside [ ] with respect to n_2 satisfying the conditions $2 \le n_2 \le \min(\frac{C_2 - 2}{1}, 619)$ and $4 \le C_2 \le 996$. Therefore,

$$f(C_2) = \frac{18884.78713}{C_2}, \ \text{ at } n_2^* = 0.1156591652 C_2 \quad (14)$$

Similarly, for the third-stage of problem ($k = 3$),

$$f(C_3) = \frac{24046.29069}{C_3}, \ \text{ at } n_3^* = 0.1137994802 C_3 \quad (15)$$

satisfying $2 \le n_3 \le \min(\frac{C_3 - 4}{1}, 1253)$ and $6 \le C_3 \le 998$.

The optimal decision for the fourth and final-stage of problem ($k = 4$) is

$$f(C_4) = 50.1041461, \ \text{ at } n_4^* = 307.232964762 \qquad (16)$$

Using this result, we obtain $C_3 = C_4 - n_4 = 1000 - 307.232964762 = 692.7670352381$. Substituting this value of C_3 in (15), we have $n_3^* = 78.836528509$. Proceeding in this manner we obtain $n_2^* = 71.006689909$ and $n_1^* = 542.9238168207$. Hence the optimum allocation for the problem (12) rounding off to the nearest integer value is $n_1^* = 543$, $n_2^* = 71$, $n_3^* = 79$ and $n_4^* = 307$ with $Z^* = 50.1042$.

TABLE II: The sum of the weighted variances under different allocations (ignoring f.p.c.)

Allocations	n_1	n_2	n_3	n_4	$SWV(\underline{n})$
Compromise					
(i) minimizing trace	524	73	85	317	50.23
(ii) average	417	89	109	385	53.40
(iii) Chatterjee's	427	86	113	374	52.93
(iv) proposed	542	71	79	307	50.10
Proportional	339	148	299	215	68.31

TABLE III: Variances of characteristics under different allocations

Allocations	$V(\overline{y}_{st}^1)$	$V(\overline{y}_{st}^2)$	Trace	R.E. w.r.t. proportional allocation
Compromise				
(i) minimizing trace	14.3	42.0	56.3	1.34
(ii) average	12.1	46.9	59.0	1.28
(iii) Chatterjee's	12.2	46.3	58.5	1.29
(iv) proposed	14.8	41.5	56.3	1.34
Proportional	15.5	59.9	75.4	1.00

V. DISCUSSION

In the following section a comparison study of the compromise allocation discussed in this article to other available compromise allocations is made. The Table II summarizes the results of various allocations. The compromise allocations to be compared are:

1. Minimizing the trace of the variance-covariance matrix (see Sukhatme et al., 1984).

2. Averaging the individual optimum allocation over characteristics (see Cochran, 1977).

3. Minimizing the total relative increase in the variances (see Chatterjee, 1967).

4. Minimizing the sum of weighted variances (proposed).

Note that in Table II $SWV(\underline{n})$ denotes the sum of the weighted variances of the estimated population means given by the objective function of $NLPP$ (5) obtained for allocation $\underline{n} = (n_1, n_2, n_3, n_4)$ under different criteria stated in first column of the table.

The variances of each characteristics (ignoring f.p.c) under different allocations, the trace of the variance-covariance matrix (or the total variance of the independent characteristics) and the relative efficiencies (R.E.) are given in Table III. The relative efficiencies of various compromise allocations are obtained over compromise allocation, which does not require the knowledge of S_{jh}. Though the criteria (i) and (iv) are identical in terms of

relative efficiency, the table reveals that the proposed criterion gives least variance at least for the characteristic that could produce high variance.

[1] Aoyama, H., (1963). Stratified Random Sampling with Optimum Allocation for Multivariate Populations. *Ann. Inst. Stat. Math*,**14**, 251-258.

[2] Bethel, J., 1989: Sample Allocation in Multivariate Surveys. *Survey Methodology*, **15**, 47-57.

[3] Chatterjee, S., (1967). A Note on Optimum Allocation. *Skand. Akt*, **50**, 40-44.

[4] Cochran, W.G., **Sampling Techniques, 3rd edition**, John Wiley and Sons, Inc., New York, 1997

[5] Dalenius, T., **Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice**, Almqvist Och Wiksell, Stockholm, 1957.

[6] Geary, R.C., (1949). *Sampling Methods Applied to Irish Agricultural Statistics*, Technical Series September.

[7] Ghosh, S.P., (1958). A Note on Stratified Random Sampling with Multiple Characters. *Cal. Stat. Bull.*, **8**, 81-89.

[8] Jahan, N., Khan, M.G.M., and Ahsan, M.J., (1994). A Generalized Compromise Allocation. *J. Ind. Stat. Assoc.*, **32**, 95-101.

[9] Khan, M.G.M., Ahsan, M.J., and Jahan, N., (1997). Compromise Allocation in Multivariate Stratified Sampling: An Integer Solution. *Naval Res. Logist.*, **44**, 69-79.

[10] Kokan, A.R., and Khan, S.U., (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution. *J. Roy. Stat. Soc., Ser. B*, **29**, 115-125.

[11] Neyman, J., (1934). On the Two Different Aspects of the Representative Methods: The Method Stratified Sampling and the Method of Purposive Selection. *J. Roy. Stat. Soc.*, **97**, 558-606.

[12] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C., **Sampling Theory of Surveys with Applications**, Iowa State University Press, Ames, IA, 1984.

[13] Yates, F., **Sampling Methods for Census and Surveys, 2nd edition**, Charles Griffin and Co. Ltd., London, 1960.