A note on optimum allocation in multivariate stratified sampling

M. G. M. Khan\textsuperscript{1} and M. J. Ahsan\textsuperscript{2}

Department of Mathematics and Computing Science,
University of the South Pacific, Suva, FIJI,
\textsuperscript{1}E-mail: khan\_mg@usp.ac.fj
\textsuperscript{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 $\bar{y}_{jst}$, an unbiased estimate of population mean $\bar{Y}_j$, for $j$th characteristic is

$$V(\bar{y}_{jst}) = \sum_{h=1}^{L} \frac{W_h S^2_{jh}}{n_h} - \sum_{h=1}^{L} \frac{W^2_h S^2_{jh}}{N_h}; \quad j = 1, 2, \ldots, p.$$  

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 $j$th characteristic in $h$th stratum ($h = 1, 2, \ldots, L$) is more heterogeneous, it produces more loss in precision in the estimate of stratum mean, as the value of stratum variance $S^2_{jh}$ for that characteristic is expected to be high. This results a high sampling variance $V(\bar{y}_{jsh})$. A way to restrain this increase in the loss of precision is to assign a maximum weight $w_j$ to $j$th characteristic as

$$w_j = \max(a_{j1}, a_{j2}, \ldots, a_{jL}).$$  \hfill (1)
Where \( a_{jh} \) are the weights for \( j \)th \( (j = 1, 2, ..., p) \) characteristic in \( h \)th \( (h = 1, 2, ..., L) \) stratum and are obtained in proportion to their stratum variances \( S_{jh}^2 \), that is, \( a_{jh} \propto S_{jh}^2 \). Letting \( \sum_{j=1}^{p} a_{jh} = 1 \), \( a_{jh} \) are worked out as

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

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,

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

The bounded variable restrictions \( 2 \leq n_h \leq N_h; \quad h = 1, 2, ..., L \) are imposed in the \textit{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 \textit{NLPP} \((4)\) may be rewritten as:

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

subject to \( \sum_{h=1}^{L} c_h n_h \leq C \) and \( 2 \leq n_h \leq N_h; \quad h = 1, 2, ..., L \quad (5)\)

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

\section{III. THE SOLUTION}

It is observed that the objective function and the constraints of the \textit{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 \( k \)th \((k = 1, 2, ..., L)\) stage provides the optimum allocation, \( n_h^k \), for \( k \)th stratum.

Consider the following sub problem of first \( k \) strata:

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

subject to \( \sum_{h=1}^{k} c_h n_h \leq C_k \) and \( 2 \leq n_h \leq N_h; \quad h = 1, 2, ..., k \quad (6)\)

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

\[
f(C_k) = \min \left[ \sum_{h=1}^{k} \frac{W_h^2 A_h^2}{n_h} \right] \sum_{h=1}^{k} c_h n_h \leq C_k \]

and \( 2 \leq n_h \leq N_h; \quad h = 1, 2, ..., k \). With the above definition the problem \((5)\) is equivalent to find \( f(C_0^L) \) recursively by finding \( f(C_k) \) for \( k = 1, 2, ..., L \) and for all feasible \( C_k \) satisfying

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

S. Pac. J. Nat. Sci., 2003, 21, 91–95
We may write

\[ f(C_k) = \min \left[ \frac{W_k^2 A_k^2}{n_k} + \sum_{h=1}^{k-1} \frac{W_h^2 A_h^2}{n_h} \right] \]

\[ \sum_{h=1}^{k-1} c_h n_h \leq C_k - c_k n_k \]

and \( 2 \leq n_h \leq N_h; \ h = 1, 2, \ldots, k \).

For a fixed value of \( n_k \) over

\[ 2 \leq n_k \leq \min (C_k', N_k) \]

where \( C_k' \) is the maximum possible sample size that can be drawn from \( k \)th stratum within the available budget \( C_k \), that is,

\[ C_k' = \frac{C_k - 2 \sum_{h=1}^{k-1} c_h}{c_k} \]

The function \( f(C_k) \) is given by

\[ f(C_k) = \frac{W_k^2 A_k^2}{n_k} + \min \left[ \sum_{h=1}^{k-1} \frac{W_h^2 A_h^2}{n_h} \right] \]

\[ \sum_{h=1}^{k-1} c_h n_h \leq C_k - c_k n_k \]

and \( 2 \leq n_h \leq N_h; \ h = 1, 2, \ldots, k - 1 \]

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 \( f(C_1), f(C_2), \ldots, f(C_k) \) for the problem (6) is

\[ f(C_k) = \min_{2 \leq n_k \leq \min(C_k', N_k)} \left[ \frac{W_k^2 A_k^2}{n_k} + f(C_{k-1}) \right] \] (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 \( k = 1, 2, \ldots, L \), \( f(C_L) \) is solved. The optimum allocation \( n_k^* \) 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:

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

From (1) the weights to be assigned to the characteristics are \( w_1 = 0.6780 \) and \( w_2 = 0.9643 \). The values of \( A_k^2; \ h = 1, 2, 3, 4 \) given by \( A_h^2 = \sum_{j=1}^{p} w_j^2 s_{jh}^2 \) are obtained as \( A_1^2 = 12874.22391, A_2^2 = 11580.0716, A_3^2 = 3483.2630 \), and \( A_4^2 = 102695.1309 \).

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

\[
\text{Minimize } Z(n_1, n_2, n_3, n_4) = \frac{14769.0123}{n_1} + \frac{252.6226}{n_2} + \frac{311.4072}{n_3} + \frac{4729.4353}{n_4} \\
\text{subject to } n_1 + n_2 + n_3 + n_4 \leq 1000 \\
2 \leq n_1 \leq 1419, \\
2 \leq n_2 \leq 619, \\
2 \leq n_3 \leq 1253, \\
2 \leq n_4 \leq 899,
\]

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

S. Pac. J. Nat. Sci., 2003, 21, 91–95
$C_k$: $k = 1, 2, 3, 4$ and their limits defined earlier are:

\[
\begin{align*}
C_4 &= n_1 + n_2 + n_3 + n_4 = 1000, \\
C_3 &= C_4 - n_4; \quad 6 \leq C_3 \leq 998, \\
C_2 &= C_3 - n_3; \quad 4 \leq C_2 \leq 996, \\
C_1 &= C_2 - n_2; \quad 2 \leq C_1 \leq 994.
\end{align*}
\]

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

\[
f(C_1) = \min_{2 \leq n_1 \leq \min(C_1', N_1)} \left[ \frac{14769.0123}{n_1} + f(C_0) \right]
\]

\[
= \min_{2 \leq n_1 \leq \min(C_1', 0.1419)} \left[ \frac{14769.0123}{n_1} \right],
\]

because $f(C_0) = 0$

\[
f(C_1) = \frac{14769.0123}{C_1}, \quad {\text{at \ } n_1^* = C_1} \quad (13)
\]

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

\[
f(C_2) = \min_{2 \leq n_2 \leq \min(C_2', N_2)} \left[ \frac{252.6226}{n_2} + f(C_1) \right]
\]

\[
= \min_{2 \leq n_2 \leq \min(\frac{C_2-241}{2}, N_3)} \left[ \frac{252.6226}{n_2} + \frac{14769.0123}{C_1} \right]
\]

\[
= \min_{2 \leq n_2 \leq \min(\frac{C_2-241}{2}, 619)} \left[ \frac{252.6226}{n_2} + \frac{14769.0123}{C_2 - n_2} \right]
\]

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 $2 \leq n_2 \leq \min(\frac{C_2-241}{2}, 619)$ and $4 \leq C_2 \leq 996$. Therefore,

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

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

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

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

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

\[
f(C_4) = 50.1041461, \quad \text{at \ } n_4^* = 307.232964762 \quad (16)
\]

Using this result, we obtain $C_4 = C_4 - n_4 = 1000 - 307.232964762 = 692.7670352381$. Substituting this value of $C_4$ 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 $Z^* = 50.1042$.

**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 $SWV(n)$ denotes the sum of the weighted variances of the estimated population means given by the objective function of $NLP$ $P \text{ (5)}$ obtained for allocation $n = (n_1, n_2, n_3, n_4)$ 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 $S_j^*$. Though the criteria (i) and (iv) are identical in terms of

<table>
<thead>
<tr>
<th>TABLE II: The sum of the weighted variances under different allocations (ignoring f.p.c.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allocations</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Compromise</td>
</tr>
<tr>
<td>(i) minimizing trace</td>
</tr>
<tr>
<td>(ii) average</td>
</tr>
<tr>
<td>(iii) Chatterjee's</td>
</tr>
<tr>
<td>(iv) proposed</td>
</tr>
<tr>
<td>Proportional</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III: Variances of characteristics under different allocations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allocations</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Compromise</td>
</tr>
<tr>
<td>(i) minimizing trace</td>
</tr>
<tr>
<td>(ii) average</td>
</tr>
<tr>
<td>(iii) Chatterjee's</td>
</tr>
<tr>
<td>(iv) proposed</td>
</tr>
<tr>
<td>Proportional</td>
</tr>
</tbody>
</table>

S. Pac. J. Nat. Sci., 2003, 21, 91–95
relative efficiency, the table reveals that the proposed criterion gives least variance at least for the characteristic that could produce high variance.