An Optimum Multivariate Stratified Sampling Design

Sana Iftekhar, M.J. Ahsan and Qazi Mazhar Ali
Department of Statistics and O.R., A.M.U., Aligarh, INDIA

Available online at: www.isca.in, www.isca.me
Received 1st December 2014, revised 11th December 2014, accepted 8th January 2015

Abstract
This article deals with the problem of finding a single usable allocation which fulfills all the characteristics involved in a multivariate stratified random sampling. The idea is to minimize all the sampling variances of the estimated population means of the characteristics under study simultaneously. The problem when formulated mathematically turns out to be a Multi-objective Integer Nonlinear Programming Problem (MOINLPP). Two different approaches viz. ‘Distance and Goal Programming’ are used to transform the formulated MOINLPP into a single objective integer nonlinear than can be solved through the well known optimization software LINGO (2013).

Keywords: Optimum multivariate stratified sampling design.

Introduction
In multivariate case individual optimum allocations do not help much unless the characteristics strongly correlated, Cochran1. An allocation is thus need that suits well to all the characteristics. Since this allocation will be based on some compromise criterion it is called compromise allocation. Some of the author who addressed the problem of obtaining a compromise allocation are Neyman2,3, Dalenius3, Aoyama4, Gren5, Hartley6, Kokan and Khan7, Chatterjee8, Ahsan and Khan9,10, Chromy11, Wywial12, Bethel13, Jahan, Khan and Ahsan4, Khan, Ahsan and Jahan14, Ansari, Najmussehar and Ahsan15,16, Kozak17.

This manuscript discusses a procedure to obtain a common allocation in multivariate stratified surveys by minimizing the sampling variances of the estimated variances for all characteristics for a fixed cost. The resulting problem is expressed as a Multi-objective nonlinear integer programming problem and solved using two approaches viz. Dl – Distance approach and Goal programming approach. The two approaches are compared through a numerical example.

The organization of the paper is as follows. Section 2 describes the problem of optimum allocation in multivariate stratified sampling with the linear cost. Section 3 gives the goal programming formulation of the problem. Section 4 discusses the Dl – Distance Approach. Section 5 provides the practical application of the discussed approaches through numerical data.

Formulation of the problem
Let there be a multivariate stratified population having number of strata as L and p characteristics on each population unit. Let Nh denote the sizes of the hth stratum and nh units be drawn without replacement from it, h = 1, 2, ..., L.

For jth character, an unbiased estimate of the population mean " is given by

\[ \bar{y}_{jst} = \sum_{h=1}^{L} W_h \bar{y}_{jh} ; j = 1, 2, ..., p \]  

(1)

The sampling variance of \( \bar{y}_{jst} \) is

\[ V(\bar{y}_{jst}) = \sum_{h=1}^{L} \left( \frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 \sigma_{jh}^2 \]  

(2)

where \( W_h = \frac{N_h}{N} \) is the stratum weight, \( \sigma_{jh}^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{jhi} - \bar{y}_{jh})^2 \) is the true variance and \( \bar{y}_{jh} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{jhi} \) is the true mean for the characteristics j and stratum h.

The usual estimate of \( V(\bar{y}_{jst}) \) is given by

\[ v(\bar{y}_{jst}) = \sum_{h=1}^{L} \left( \frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 \sigma_{jh}^2 \]  

(3)

where \( \sigma_{jh}^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{jhi} - \bar{y}_{jh})^2 \) is the usual estimate of \( S_{jh}^2 \) from the sample, \( y_{jhi} \) denotes the observation on the jth unit of the hth stratum in the sample as well as in the population, for the jth characteristics and \( \bar{y}_{jh} \) denotes the sample mean.

Ross18 gave the sampling variance of \( S_{jh}^2 \) in terms of the fourth moment \( \beta_{2jh} \) about means given by

\[ V(S_{jh}^2) = \frac{M_{4jh}}{n_h} \left( - \frac{n_h-3}{n_h(n_h-1)} S_{jh}^4 \right) \]  

(4)

where \( \beta_{2jh} = \frac{M_{4jh}}{S_{jh}^2} \), for largeNh, \( V(S_{jh}^2) \) may be approximated as

\[ V(S_{jh}^2) \approx \frac{S_{jh}^4}{n_h} (\beta_{2jh} - 1) \]  

(5)

where \( \beta_{2jh} \) is the coefficient of kurtosis of \( \bar{y}_{jst} \) for the jth characteristics in the hth stratum.
Now
\[ V\left(\bar{v}(\tilde{y}_{js})\right) = V\left(\sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}}\right) = \sum_{h=1}^{L} \frac{w_{h}}{n_{h}} V\left(s_{j}^{2}\right) \]
\[ = \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{zj} - 1) \]
\[ = V_{j} \quad \text{(say)} \]  
(6)

Letting the total cost \( C \) be expressed as
\[ C = c_{0} + \sum_{h=1}^{L} c_{h} n_{h} \]  
(8)

Where \( c_{0} \) is the fixed cost and \( c_{h} \) denote the measurement cost of each and every selected unit in the \( h^{th} \) stratum.

If the survey is to be conducted in such a way that the variances of the estimated variances of \( \tilde{y}_{jst} \) for all the \( p \) characteristics are minimized simultaneously for a fixed cost then the problem of allocation with linear cost function can be expressed as

Minimize \( V_{j} ; j = 1, 2, ..., p \) Simultaneously
(9)

Subject to \( \sum_{h=1}^{L} c_{h} n_{h} \leq C - c_{0} \)
\[ \leq n_{h} \leq N_{h} \]
\( n_{h} \) integers
(10)

Constraints \( 2 \leq n_{h} \leq N_{h} ; h = 1, 2, ..., L \) are added to take care of over sampling and to provide an estimate of strata variances \( s_{j}^{2} \).

In the following sections the two approaches namely the Goal Programming approach and the \( D_{1} \)-distance approach are discuss to solve the formulated (MOILPP) (9)-(12).

**The Goal Programming Approach**

Let \( V_{j}^{*} \) be the optimum value of \( V_{j} \) at the optimal \( p \) points \( n_{j}^{*} = (n_{1j}^{*}, n_{2j}^{*}, ..., n_{pj}^{*}) \) of the integer nonlinear programming problems (INLPP).

Minimize \( V_{j} \)
(13)

Subject to \( \sum_{h=1}^{L} c_{h} n_{h} \leq C - c_{0} \)
\[ \leq n_{h} \leq N_{h} \]
\( n_{h} \) integers
(14)

for \( j = 1, 2, ..., p \).

Further let \( \tilde{V}_{j} = \tilde{V}_{j}(n_{1j}^{c}, n_{2j}^{c}, ..., n_{pj}^{c}) = \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{zj} - 1) \) \( (17) \)

is the value of \( V_{j} \) at the compromise allocation \( n_{j}^{c} = (n_{1j}^{c}, n_{2j}^{c}, ..., n_{pj}^{c}) \).

As \( \tilde{V}_{j} \geq V_{j}^{*} \), the quantity \( \tilde{V}_{j} - V_{j}^{*} \geq 0 \) denotes the increase in \( V_{j}^{*} \) due to not using the individual allocation for \( j^{th} \) characteristics.

The 'goal' may now be defined as: "Find \( n_{j}^{c} = (n_{1j}^{c}, n_{2j}^{c}, ..., n_{pj}^{c}) \) such the \( (\tilde{V}_{j} - V_{j}^{*}) \leq x_{j} ; j = 1, 2, ..., p \). Where \( x_{j} \) is the tolerance limit for the increase in \( V_{j}^{*} \) fixed in advance.

These tolerance limit impose the following restrictions.
\[ V_{j} - V_{j}^{*} \leq x_{j} \text{ or } V_{j} - x_{j} \leq V_{j}^{*} \]

Substituting the value of \( \tilde{V}_{j} \) from (17) we get
\[ \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{zj} - 1) - x_{j} \leq V_{j}^{*} \]  
(18)

A suitable compromise criterion will then be to minimize the quantity \( \sum_{j=1}^{p} x_{j} \), which gives the total increase in \( V_{j}^{*} \).

The goal programming problem for obtaining a compromise allocation is then given as

Minimize \( \sum_{j=1}^{p} x_{j} \)
(19)

Subject to \( \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{zj} - 1) - x_{j} \leq V_{j}^{*} \) 
(20)

\[ \sum_{h=1}^{L} c_{h} n_{h} \leq C - c_{0} \]
\[ 2 \leq n_{h} \leq N_{h} \]
\( n_{h} \) integers
(21)-(23)

When numerical values of the parameters are available ((19)-(23)) may be solved by using an appropriate mixed integer nonlinear programming technique.

The next section discusses the \( D_{1} \) Distance Approach.

**\( D_{1} \) Distance Approach**

Let the priority of \( K \) objective functions be considered. This will lead to \( K! \) different priority structures. Thus one has to solve \( K! \) problems to get \( K! \) solutions.

Let \( n^{(r)} = \left(n_{1}^{(r)}, n_{2}^{(r)}, ..., n_{L}^{(r)}\right), \ r = 1, 2, ..., K! \) be the \( r^{th} \) solution.

Consider the case when there are only two characteristics, that is, \( K=2=K! \). If the first characteristic is more important then the Lexicographic goal programming problem may have the following form.

Lex minimize \( \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{z1h} - 1) - x_{1} \leq V_{1}^{*} \)
(24)

Subject to \( \sum_{h=1}^{L} \frac{w_{h} s_{j}^{2}}{n_{h}} (\beta_{z2h} - 1) - x_{2} \leq V_{2}^{*} \)
\[ \sum_{h=1}^{L} c_{h} n_{h} \leq C - c_{0} \]
\[ 2 \leq n_{h} \leq N_{h} \]
\( n_{h} \) integers
(25)-(29)

Let \( n^{(1)} = \left(n_{1}^{(1)}, n_{2}^{(1)}, ..., n_{L}^{(1)}\right) \) the solution to the MINLPP problem (24)-(29).

When second characteristic is more important we have the problem as

Lex minimize \( \sum_{j=1}^{L} x_{j} \)
(30)
Let the solution to the problem (30)-(35) be denoted by
\[ n^* = \left( n_1^{(2)^*}, n_2^{(2)^*}, \ldots, n_L^{(2)^*} \right) \]

Then,
\[ n^* = \{ \max(n_1^{(1)^*}, n_2^{(1)^*}), \max(n_3^{(1)^*}, n_1^{(2)^*}), \ldots, \max(n_1^{(1)^*}, n_1^{(2)^*}) \} = (n_1^*, n_2^*, \ldots, n_L^*), \text{ say} \]

will provide the ideal solution.
In fact the ideal solution is hard to achieve. Thus the solution, which is nearest to the ideal solution, is accepted as the available compromise solution. The corresponding priority structure is identified as most appropriate priority structure for planning.

The best compromise solution will be the solution to the following problem
Minimize \[ \sum_{i=1}^{K!} \delta_{hr} \]
Subject to \[ n_h^* - n_h^{(r)^*} - \delta_{hr} = 0 \]
\[ \delta_{hr} \geq 0 \]
\[ n_h \geq 0 \text{ integers} \]
\[ r = 1, 2, \ldots, K! \]
where \( \delta_{hr} \) are the devotional variable.

Now define the \( D_k \) distance for the \( r \)th solution \( n^{(r)^*} \) as
\[ (D_k)^r = \sum_{h=1}^{L} |n_h - n_h^{(r)^*}| \]
This gives
\[ (D_k)^{optimal} = \text{Minimize}_{1 \leq r \leq K!} (D_k)^r \]
\[ = \text{Minimize} \sum_{h=1}^{L} |n_h - n_h^{(r)^*}| \]
\[ = \text{Minimize} \sum_{h=1}^{L} \delta_{hr} \]
\[ = \sum_{h=1}^{L} \delta_{hk} \]
\[ = (D_k)^k, \text{ say} \]
where it is assumed that the minimum is attained for \( r = k \).
Hence, \( (n_1^{(k)^*}, n_2^{(k)^*}, \ldots, n_L^{(k)^*}) \) will be the best compromise solution.

For notations and details of the formulation see Ali, Raghav and Bari\(^1\).
Subject to
Minimize \( x_{1} + x_{2} \)
\[
\begin{align*}
\frac{152726.3253}{n_{1}^2} + \frac{27894.09259}{n_{2}^2} + \frac{187851.3798}{n_{3}^2} + \frac{30151995.48}{n_{4}^2} - x_{1} & \leq 28.2833 \\
\frac{1002658547}{n_{1}^2} + \frac{68056.2522}{n_{2}^2} + \frac{25429.47259}{n_{3}^2} + \frac{4757180.105}{n_{4}^2} - x_{2} & \leq 134.846
\end{align*}
\]
\[\begin{align*}
3n_{1} + 4n_{2} + 5n_{3} + 7n_{4} & \leq 1200 \\
2 \leq n_{1} & \leq 1419 \\
2 \leq n_{2} & \leq 619 \\
2 \leq n_{3} & \leq 1253 \\
2 \leq n_{4} & \leq 899 \\
x_{j} & \geq 0; j = 1, 2, \ldots, p \\
n_{k} & \text{ integers}
\end{align*}\]

(50)

Using LINGO, the optimum compromise solution for GPP (50) is found to be
\( n_{1}^{*} = 190, n_{2}^{*} = 17, n_{3}^{*} = 20, n_{4}^{*} = 66, \)
\( x_{1}^{*} = 105.7760, x_{2}^{*} = 45.10164 \)
with \( V^{*} = 150.8776 \)

Solution using \( D_{1} \) Distance Approach: If priority is given to the first characteristics, then solution of the lexicographic goal programming problem (24)-(29) is obtained as
\( n^{(1)*} = (157, 17, 23, 78) \)
If priority is given to the second characteristics, then solution of the lexicographic goal programming problem (30)-(35) is obtained as
\( n^{(2)*} = (206, 18, 18, 60) \)
From expression (36) gives the ideal solution as
\( n^{*} = (206, 18, 23, 78) \)
Table 2 gives the \( D_{1} \) Distances

<table>
<thead>
<tr>
<th>Priorities of Variances</th>
<th>( D_{1} ) Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>((V_{1}, V_{2}))</td>
<td>47</td>
</tr>
<tr>
<td>((V_{2}, V_{1}))</td>
<td>23</td>
</tr>
</tbody>
</table>

The \( D_{1} \) Distances from the ideal solution is minimum corresponding to the second priority. The resulting best compromise solution is \( n^{*} = (206, 18, 18, 60) \) with variances \( V_{1} = 134.06 \) and \( V_{2} = 179.95 \)

Conclusion

From table 3, considering the trace values as the measure of performance, we can conclude that out of the two discussed approaches the \( D_{1} \)-Distance approach provides better result in comparison to the goal programming approach.

Acknowledgement

The author M.J. Ahsan is grateful to the University Grant Commission for its financial support in the form of Emeritus Fellowship for carrying out this work.

References

3. Dalenius T., Sampling in Sweden: Contributions to The Methods and Theories of Sample Survey Practice, Almqvist and Wiksell, Stockholm, (1957)


