Research Journal of Mathematical and Statistical Sciences ISSN 2320–6047
Vol. **3(1),** 10-14, January **(2015)** *Res. J. Mathematical and Statistical Sci.*

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 find a single usable allocation which is suits all the characteristics involved in a multivariate stratified random sampling. The idea is to minimize all the sampling variances of the estimates of the population means of the characteristics under study simultaneously. The problem when formulated mathematically terms out to be a Multi-objective Integer Nonlinear Programming Problem (MOINLPP). Two different approaches viz. '$D_1 -$ Distance and 'Goal Programming' are used to transformed 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, Cochran[1]. 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 Neyman[2], Dalenius[3], Aoyama[4], Gren[5], Hartley[6], Kokan and Khan[7], Chatterjee[8], Ahsan and Khan[9,10], Chromy[11], Wywial[12], Bethel[13], Jahan, Khan and Ahsan[14], Khan, Ahsan and Jahan[15], Ansari, Najmussehar and Ahsan[16], Kozak[17].

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. $D_1 -$ 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 $D_1 -$ 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 N_h denote the sizes of the h^{th} stratum and n_h units be drawn without replacement from it, $h = 1,2, \dots, L$.

For j^{th} character, an unbiased estimate of the population mean $\bar{Y}_j$ is given by

$$\bar{y}_{jst} = \sum_{h=1}^{L} W_h \, \bar{y}_{jh} \; ; j = 1,2, \dots, p \tag{1}$$

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

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

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

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

$$v(\bar{y}_{jst}) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 s_{jh}^2 \tag{3}$$

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

Ross[18] gave the sampling variance of s_{jh}^2 in terms of the fourth moment M_{4jh} about meanis given by

$$V(s_{jh}^2) = \frac{M_{4jh}}{n_h} - \frac{n_h - 3}{n_h(n_h - 1)} S_{jh}^4 \tag{4}$$

Letting $\beta_{2jh} = \frac{M_{4jh}}{S_{jh}^4}$, for largeN_h, $V(s_{jh}^2)$ may be approximated as

$$V(s_{jh}^2) \cong \frac{S_{jh}^4}{n_h} (\beta_{2jh} - 1) \tag{5}$$

where β_{2jh} is the coefficient of kurtosis of $\bar{y}_{jst}$ for the j^{th} characteristics in the h^{th} stratum.

Now

$V\left(v(\bar{y}_{jst})\right) = V\left(\sum_{h=1}^{L} \frac{W_h^2 s_{jh}^2}{n_h}\right) = \sum_{h=1}^{L} \frac{W_h^4}{n_h^2} V\left(s_{jh}^2\right) =$

$$\sum_{h=1}^{L} \frac{W_h^4 S_{jh}^4}{n_h^3}\left(\beta_{2jh} - 1\right) \quad (6)$$

$$= V_j \text{ (say)} \quad (7)$$

Letting the total cost C be expressed as

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \quad (8)$$

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 $\bar{y}_{jst}$ 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 $V_j \; ; j = 1,2,\ldots,p$ Simulteneously (9)

Subject to $\sum_{h=1}^{L} c_h n_h \leq C - c_0$ (10)

$2 \leq n_h \leq N_h$ (11)

and n_h *integers* (12)

Constraints $2 \leq n_h \leq N_h \; ; h = 1,2,\ldots,L$ are added to take care of over sampling and to provide an estimate of strata variances S_{jh}^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 (MOINLPP) (9)-(12).

The Goal Programming Approach

Let V_j^* be the optimum value of V_j at the optimal p points $n_j^* = \left(n_{1j}^*, n_{2j}^*, \ldots, n_{Lj}^*\right)$ of the integer nonlinear programming problems (INLPP).

Minimize V_j (13)

Subject to $\sum_{h=1}^{L} c_h n_h \leq C - c_0$ (14)

$2 \leq n_h \leq N_h$ (15)

n_h *integers* (16)

for $j = 1,2,\ldots,p$.

Further let $\tilde{V}_j = \tilde{V}_j\left(n_{1j}^c, n_{2j}^c, \ldots, n_{Lj}^c\right) = \sum_{h=1}^{L} \frac{W_h^4 S_{jh}^4}{\left(n_{hj}^c\right)^3}\left(\beta_{2jh} - 1\right)$ (17)

is the value of V_j at the compromise allocation $n_j^c = \left(n_{1j}^c, n_{2j}^c, \ldots, n_{Lj}^c\right)$.

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 jth characteristics.

The 'goal' may now be defined as: "Find $n_j^c = \left(n_{1j}^c, n_{2j}^c, \ldots, n_{Lj}^c\right)$ such the $\left(\tilde{V}_j - V_j^*\right) \leq x_j; j = 1,2,\ldots,p$". Where x_j, is the tolerance limit for the increase in V_j^* fixed in advance.

These tolerance limit impose the following restrictions.

$\tilde{V}_j - V_j^* \leq x_j$ or $\tilde{V}_j - x_j \leq V_j^*$

Substituting the value of $\tilde{V}_j$ from (17) we get

$$\sum_{h=1}^{L} \frac{W_h^4 S_{jh}^4}{\left(n_{hj}^c\right)^3}\left(\beta_{2jh} - 1\right) - x_j \leq V_j^* \quad (18)$$

A suitable compromise criterion will then be to minimize the quantity $\sum_{j=1}^{p} x_j$, which gives the total increase in $V_j^* s$.

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

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

Subject to $\sum_{h=1}^{L} \frac{W_h^4 S_{jh}^4}{n_h^3}\left(\beta_{2jh} - 1\right) - x_j \leq V_j^*$ (20)

$\sum_{h=1}^{L} c_h n_h \leq C - c_0$ (21)

$2 \leq n_h \leq N_h$ (22)

n_h *integers* (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 $n^{(r)} = \left\{n_1^{(r)}, n_2^{(r)}, \ldots, n_L^{(r)}\right\}$, $r = 1,2,\ldots,K!$ be the rth 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 $\sum_{k=1}^{2} x_j$ (24)

Subject to $\sum_{h=1}^{L} \frac{W_h^4 S_{1h}^4}{n_h^3}\left(\beta_{21h} - 1\right) - x_1 \leq V_1^*$ (25)

$\sum_{h=1}^{L} \frac{W_h^4 S_{2h}^4}{n_h^3}\left(\beta_{22h} - 1\right) - x_2 \leq V_2^*$ (26)

$\sum_{h=1}^{L} c_h n_h \leq C - c_0$ (27)

$x_j \geq 0$ (28)

$n_h \geq 0$ *integers* (29)

Let $n^{(1)^*} = \left(n_1^{(1)^*}, n_2^{(1)^*}, \ldots, n_L^{(1)^*}\right)$ the solution to the MINLPP problem (24)-(29).

When second characteristic is more important we have the problem as

Lex minimize $\sum_{j=1}^{2} x_j$ (30)

Subject to $\sum_{h=1}^{L} \frac{W_h^4 S_{1h}^4}{n_h^3}(\beta_{22h}-1) - x_1 \le V_2^*$ (31)

$\sum_{h=1}^{L} \frac{W_h^4 S_{1h}^4}{n_h^3}(\beta_{21h}-1) - x_2 \le V_1^*$ (32)

$\sum_{h=1}^{L} c_h n_h \le C - c_0$ (33)

$x_j \ge 0$ (34)

$n_h \ge 0$ *integers* (35)

Let the solution to the problem (30)-(35) is denoted by $n^{(2)^*} = \left(n_1^{(2)^*}, n_2^{(2)^*}, \ldots, n_L^{(2)^*}\right)$

Then,

$$n^* = \left\{\max\left(n_1^{(1)^*}, n_1^{(2)^*}\right), \max\left(n_1^{(1)^*}, n_1^{(2)^*}\right), \ldots, \max\left(n_1^{(1)^*}, n_1^{(2)^*}\right)\right\} = (n_1^*, n_2^*, \ldots, n_L^*) \text{, say} \quad (36)$$

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 $_{1\le r\le K!}$ $\sum_{h=i}^{L} \delta_{hr}$ (37)

Subject to $n_h^* - n_h^{(r)^*} - \delta_{hr} = 0$ (38)

$\delta_{hr} \ge 0$ (39)

$n_h \ge 0$ *integers* (40)

$r = 1,2, \ldots, K!$ (41)

where δ_{hr} are the deviational variable.

Now define the $D_1 -$ distance for the rth solution $n^{(r)^*}$ as

$$(D_1)^r = \sum_{h=1}^{L} \left|n_h^* - n_h^{(r)^*}\right| \quad (42)$$

This gives

$(D_1)_{\text{optimum}} = \text{Minimize}_{1\le r\le K!}\ (D_1)^r$ (43)

$= \underset{1\le r\le K!}{\text{Minimize}} \sum_{h=1}^{L} \left|n_h^* - n_h^{(r)^*}\right|$ (44)

$= \underset{1\le r\le K!}{\text{Minimize}} \sum_{h=1}^{L} \delta_{hr}$ (45)

$= \sum_{h=1}^{L} \delta_{hk}$ (46)

$= (D_1)^k$, say (47)

where it is assumed that the minimum is attained for $r = k$.

Hence, $\left(n_1^{(k)^*}, n_2^{(k)^*}, \ldots, n_L^{(k)^*}\right)$ will be the best compromise solution.

For notations and details of the formulation see Ali, Raghav and Bari[19].

A Practical Application

The data given in table 1 are from Sukhatme[2]. The values of β_{21h} and β_{22h} are assumed by authors.

The cost C, c_0 and $c_h; h = 1,2,3,4$ are assumed as $C = 1500$, $c_0 = 3000, c_1 = 3$, $c_2 = 4$, $c_3 = 5$ and $c_4 = 7$, units.

Table -1
Values of N_h, W_h true strata variances and coefficient of kurtosis

Stratum	N_h	W_h	S_{1h}^2	S_{2h}^2	β_{21h}	β_{22h}
1	1419	0.3387	4817.72	130121.15	1.5	5.5
2	619	0.1477	6251.26	7613.52	2.5	3.5
3	1253	0.2990	3066.16	1456.40	3.5	2.5
4	899	0.2146	56207.25	66977.72	5.5	1.5

Solution using Goal Programming Approach: With the values given in Table 1, the INLPP (9)-(12) takes the following form for $j = 1,2$.

For $j = 1$

$$\text{Minimize } V_1 = \frac{152726.3253}{n_1^3} + \frac{27894.09259}{n_2^3} + \frac{187851.3798}{n_3^3} + \frac{30151995.48}{n_4^3}$$

Subject to $3n_1 + 4n_2 + 5n_3 + 7n_4 \le 1200$

$2 \le n_1 \le 1419$

$2 \le n_2 \le 619$

$2 \le n_3 \le 1253$

$2 \le n_4 \le 899$

n_h *integers* (48)

The solution to INLPP (48) obtained by LINGO-13[20] is

$n_{11}^* = 38,\ n_{12}^* = 23,\ n_{13}^* = 35, n_{14}^* = 117$

with the optimal value of variance $V_1^* = 28.28331$

For $j = 2$

$$\text{Minimize } V_2 = \frac{1002695547}{n_1^3} + \frac{68956.25222}{n_2^3} + \frac{25429.47259}{n_3^3} + \frac{4757180.105}{n_4^3}$$

Subject to $3n_1 + 4n_2 + 5n_3 + 7n_4 \le 1200$

$2 \le n_1 \le 1419$

$2 \le n_2 \le 619$

$2 \le n_3 \le 1253$

$2 \le n_4 \le 899$

n_h *integers* (49)

This gives

$n_{21}^* = 234,\ n_{22}^* = 20,\ n_{23}^* = 15, n_{24}^* = 49$

with the optimal value of variance $V_2^* = 134.8463$

Using the values of V_1^* and V_2^* the Goal programming problem (19)-(23) becomes

Subject to

$$\text{Minimize } x_1 + x_2$$
$$\frac{152726.3253}{n_1^3} + \frac{27894.09259}{n_2^3} + \frac{187851.3798}{n_3^3} + \frac{30151995.48}{n_4^3} - x_1 \le 28.2833$$
$$\frac{1002695547}{n_1^3} + \frac{68956.25222}{n_2^3} + \frac{25429.47259}{n_3^3} + \frac{4757180.105}{n_4^3} - x_2 \le 134.846$$
$$3n_1 + 4n_2 + 5n_3 + 7n_4 \le 1200$$
$$2 \le n_1 \le 1419$$
$$2 \le n_2 \le 619$$
$$2 \le n_3 \le 1253$$
$$2 \le n_4 \le 899$$
$$x_j \ge 0; j = 1,2,\ldots,p$$
$$n_h \textit{ integers}$$
(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-2
D_1 Distances from the ideal solutions

Priorities of Variances	D_1 Distance
(V_1, V_2)	47
(V_2, V_1)	23

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

1. Cochran W.G., Sampling Techniques, *3rd edition. John Wileyand Sons, New York*, **(1977)**
2. Neyman J., On the Two Different Aspects of the Representative Methods: The Method of Stratified Sampling and The Method of Purposive Selection, *J. Roy. Statist. Soc*, **97(4)**, 558–625 **(1934)**
3. Dalenius T., Sampling in Sweden: Contributions to The Methods and Theories of Sample Survey Practice, *Almqvist and Wiksell, Stockholm*, **(1957)**
4. Aoyama H., Stratified Random Sampling with Optimum Allocation for Multivariate Populations, *Ann. Inst. Statist. Math*, **14**, 251–258 **(1963)**
5. Gren J., Some Application of Non-linear Programming in Sampling Methods, *Przeglad Statystyczny,* **13**, 203–217 (in Polish) **(1966)**
6. Hartley H.O., Multiple purpose optimum allocation in stratified sampling, *Proc. Amer. Statist. Assoc., Social Statist.,* 258-261**(1965)**
7. Kokan A.R. and Khan S.U., Optimum allocation in multivariate surveys: An analytical solution, *J. Roy. Statist. Soc. B*, **29**, 115–125 **(1967)**
8. Chatterjee S., A note on optimum allocation. *Scand. Actuar. J.*, **50**, 40–44 **(1967)**
9. Ahsan M.J. and Khan S.U., Optimum Allocation in Multivariate Stratified Random Sampling using Prior Information, *J. Indian Statist. Assoc.,* **15**, 57–67 **(1977)**
10. Ahsan M.J. and Khan S.U., Optimum allocation in multivariate stratified random sampling with overhead cost. *Metrika,* **29(1)**, 71–78 **(1982)**
11. Chromy R., Design Optimization with Multiple Objectives. Proceedings of the Survey Research Methods section, *American Statistical Association*, 194–199 **(1987)**

Table-3
Summary of the result

Approach	Allocation				Variance V_j		Trace	Cost
	n_1	n_2	n_3	n_4	V_1	V_2	$V = V_1 + V_2$	incurred
Goal Programming Approach	190	17	20	66	176.60	153.91	330.51	1200
D_1-Distance Approach	206	18	18	60	134.06	179.95	314.01	1200

12. Wywial J., Minimizing the Spectral Radius of Means Vector from Sample Variance–covariance Matrix Sample Allocation between Strata, *Prace Naukowe Akademii Ekonomicznej we Wroclawiu*. Wroclaw, Poland, **404**, 223–235 (in Polish) **(1988)**

13. Bethel J., Sample Allocation in Multivariate Surveys. *Survey Methodology*, **15**, 40-57 **(1989)**

14. Jahan N., Khan M.G.M. and Ahsan M.J., A Generalized Compromise Allocation, *J. Indian Statist. Assoc.*, **32**, 95–10 **(1994)**

15. Khan M.G.M., Ahsan M.J. and Jahan N., Compromise Allocation in Multivariate Stratified Sampling: An Integer Solution. *Naval Research Logistics,* **44**, 69–79 **(1997)**

16. Ansari A.H., Najmussehar and Ahsan M.J., On Multiple Response Stratified Random Sampling Design, *International Journal of Statistical Sciences*, **1(1)**, 1–11 **(2009)**

17. Kozok M., On sample allocation in multivariate surveys, *Communication in Statistics-Simulation and Computation*, **35**, 901-910 **(2006)**

18. Ross A., Variance Estimates in Optimum Sample Designs, *J. Amer. Stat. Assoc*., **56**, 135-142 **(1961)**

19. Ali I., Raghav Y.S. and Bari A., Compromise Allocation in Multivariate Stratified Surveys with Stochastic Quadrartic Cost Function, *Journal of Statistical Computation and Simulation*, 1-15 **(2011)**

20. LINGO, *LINGO User's Guide, LINDO Systems Inc., 1415, North Dayton Street, Chicago, Illinois, 60622, USA* **(2013)**