Grim, J. (1990). Probabilistic Expert Systems and Distribution Mixtures. Computers and Artificial Intelligence, 9, 241–256.

Grim, J. (1992). A Dialog Presentation of Census Results by Means of the Probabilistic Expert System PES. In Proceedings of the Eleventh European Meeting on Cybernetics and Systems Research, R. Trappl (ed.). Singapore: World Scientific, Vienna 21–24 April, 997–1005.

Grim, J. (1994). Knowledge Representation and Uncertainty Processing in the Probabilistic Expert System PES. International Journal of General Systems, 22, 103–111.

Grim, J. and Boček, P. (1996). Statistical Model of Prague Households for Interactive Presentation of Census Data. In SoftStat'95. Advances in Statistical Software 5. Stuttgart: Lucius & Lucius, 271–278.

Grim, J., Boček, P., and Pudil, P. (2001). Safe Dissemination of Census Results by Means of Interactive Probabilistic Models. In Proceedings of the ETK-NTTS 2001 Conference, P. Nanopoulos and D. Wilkinson (eds). Rome: European Communities, 849–856.

Grim, J., Hora, J., Boček, P., Somol, P., and Pudil, P. (2004a). Information Analysis of Census Data by Using Statistical Models. In Proceedings: Statistics – Investment in the Future. Prague.

Grim, J., Hora, J., and Pudil, P. (2004b). Statistical Model for Interactive Presentation of Census Results under Protection of Confidentiality. Statistika, 40(5), 400–414, [In Czech]

McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models. New York, Toronto: John Wiley & Sons.

Schlesinger, M.I. (1968). Relation between Learning and Self-learning in Pattern Recognition. Kibernetika, (Kiev), (2), 81–88, [In Russian]

Willenborg, L.C.R.J. and de Waal, A.G. (2001). Elements of Statistical Disclosure Control. New York: Springer Verlag.

Williams, T. (2005). The Development of Truth Decks for the 2010 Census Count Imputation Research. In Proceedings of the American Statistical Association, Section on Survey Research Methods.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Research in Official Statistics, 2, 87–104.

Received April 2009
Revised December 2009

Journal of Official Statistics, Vol. 26, No. 4, 2010, pp. 695–708

An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach

M. G. M. Khan[1,4], *T. Maiti*[2], *and M. J. Ahsan*[3]

In multivariate stratified random sampling, for practical purposes we need an allocation which is optimum in some sense for all characteristics because the individual optimum allocations usually differ widely unless the characteristics are highly correlated. Cochran (1977) suggested the use of characteristicwise average of individual optimum allocations as a compromise allocation for correlated characteristics. For uncorrelated ones, many authors have suggested various criteria to work out an optimum compromise allocation. For example a compromise criterion may be to minimize the total loss in efficiency of the estimate due to not using the individual optimum allocations. When auxiliary information is also available, it is customary to use it to increase the precision of the estimates. Moreover, for practical implementation of an allocation, we need integer values of the sample sizes. The present article addresses the problem of determining the integer optimum compromise allocation when the population means of various characteristics are of interest and auxiliary information is available for the separate and combined ratio and regression estimates. The problem is formulated as a multiobjective nonlinear programming problem and a solution procedure is developed using goal programming technique. The goal is to minimize the weighted sum of the increases in the variances due to not using the individual optimum allocations subject to budgetary and other constraints.

Key words: Multivariate stratified sampling; optimum allocation; optimum compromise allocation; all-integer multi-objective nonlinear programming problem; goal programming; auxiliary information; ratio and regression methods of estimation.

1. Introduction

The problem of optimum allocation in stratified random sampling for a univariate population is well-known in sampling literature (see Cochran 1977 and Sukhatme et al. 1984). In multivariate stratified sampling where more than one characteristic is to be measured on each sampled unit, the optimum allocations for the individual characteristics are not of much practical use unless the various characteristics under study are highly correlated, because an allocation that is optimum for one characteristic will generally be

[1] School of Computing Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji. Email: khan_mg@usp.ac.fj
[2] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, U.S.A.
[3] Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh 202002, India.
[4] M.G.M. Khan was on leave at the Iowa State University in Ames, U.S.A. when the research was undertaken.
Acknowledgment: The authors are grateful to Professor Lars Lyberg, the Associate Editor and the learned referees for their valuable comments and suggestions that helped a lot in presenting the manuscript in its present form.

far from optimum for others. To resolve this problem, a compromise criterion is needed to determine a useable allocation which is optimum, in some sense, for all characteristics. Such an allocation may be called a compromise allocation. Many authors have discussed various criteria for obtaining a useable compromise allocation among them Neyman (1934), Dalenius (1953, 1957), Ghosh (1958), Yates (1960), Aoyama (1963), Folks and Antle (1965), Kokan and Khan (1967), Chatterjee (1967, 1968), Ahsan and Khan (1977, 1982), Bethel (1985), Chromy (1987), Bethel (1989), Jahan, Khan, and Ahsan (1994), Khan, Ahsan, and Jahan (1997), and Khan, Khan, and Ahsan (2003). Obviously, the compromise allocation always results in a loss of precision as compared to the individual optimum allocations.

In surveys where auxiliary information on a variable that is highly correlated with the main study variable is available it is desirable to consider the estimators that use the auxiliary information to increase the precision of estimates. Dayal (1985) discussed the optimum allocation of sample sizes for univariate stratified sample surveys using auxiliary information. For multivariate stratified surveys, Ahsan and Khan (1982) discussed the optimum allocation when the auxiliary information is available in the form of a joint multivariate normal distribution.

For practical purposes, integer values of the sample sizes are required. They can be obtained by rounding off the noninteger sample sizes to their nearest integer values. When the sample sizes are large enough and/or the measurement costs in various strata are not too high, the rounded-off sample allocations may work well. However, in many situations the rounded-off sample allocations may violate the cost constraint and/or there may exist other sets of integer sample allocations with relatively smaller values of the objective function. There may also be other goals that must be met in order to satisfy constraints on cost, and minimum and maximum sample sizes from different strata to permit variance estimation and to avoid over-sampling. In such circumstances the problem of allocation becomes a problem of mathematical programming with a nonlinear objective function and linear as well as nonlinear constraints.

The present article treats the problem of obtaining a compromise allocation in multivariate stratified sampling using auxiliary information as an all integer multi-objective nonlinear programming problem (AIMNLPP). A goal is set that the increase in the variance for characteristic due to the use of compromise allocation instead of its individual optimum allocation must not exceed a certain quantity, called the goal variable, and a solution procedure is developed using goal programming technique. The optimum compromise allocation is then obtained by executing a program coded in LINGO, a computer software package for solving linear, nonlinear and integer optimization problems developed by LINDO Systems, Inc. This software is user-friendly and does not require much knowledge of computer programming or computer languages.

The proposed method deals with the problem of determining an optimum compromise allocation when the population means of the characteristics are of interest and the combined ratio estimator is to be used. The method can also be applied to other separate ratio and separate and combined regression estimators. Two numerical examples are given to demonstrate the practical application of the algorithm and its computational details. Comparison of the proposed method has been made with several other available methods

of obtaining a compromise allocation. The comparison shows that the proposed method provides more precise estimates than other methods.

2. The Problem

Let the population be divided into L nonoverlapping strata and n_h be the number of units drawn by simple random sampling without replacement (SRSWOR) from the hth stratum consisting of N_h units, and $n = \sum_{h=1}^{L} n_h$ and $N = \sum_{h=1}^{L} N_h$ give the total sample size and the population size. Let p characteristics be defined on each population unit and the estimation of unknown population means $\bar{Y}_j$, $j = 1, 2, \ldots, p$ be of interest. Let y_{jhi} and x_{jhi} denote the values of the ith population (sampled) unit of the jth main variable (Y_j) and the jth auxiliary variable (X_j) respectively, in the hth stratum. Also let $\bar{y}_{jh} = n_h^{-1}\sum_{i=1}^{n_h} y_{jhi}$ and $\bar{x}_{jh} = n_h^{-1}\sum_{i=1}^{n_h} x_{jhi}$ denote the sample means of the jth characteristic in the hth stratum for the main and auxiliary variables, respectively.

In stratified sampling, in the presence of auxiliary information, the estimate of the population mean $\bar{Y}_j$ may be given as

$$\bar{y}_{j,st} = \sum_{h=1}^{L} W_h \bar{r}_{jh}, \quad j = 1, 2, \ldots, p \tag{1}$$

with the sampling variance

$$\mathrm{Var}(\bar{y}_{j,st}) = \sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{N_h} \tag{2}$$

where $W_h = N_h/N, \quad h = 1, 2, \ldots, L.$

The quantities $\bar{r}_{jh}$ and A_{jh}^2 define the method of estimation using auxiliary information. In this article we are using the combined ratio estimate for which

$$\bar{r}_{jh} = (\hat{R}_{st})_j \bar{X}_j \tag{3}$$

and

$$A_{jh}^2 = S_{y_jh}^2 + R_j^2 S_{x_jh}^2 - 2R_j S_{y_jx_jh} \tag{4}$$

Where

$$(\hat{R}_{st})_j = \frac{(\bar{y}_{st})_j}{(\bar{x}_{st})_j}; \quad (\bar{y}_{st})_j = \sum_{h=1}^{L} W_h \bar{y}_{jh}, \quad (\bar{x}_{st})_j = \sum_{h=1}^{L} W_h \bar{y}_{jh}$$

$$\bar{X}_j = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} x_{jhi}$$

$$S_{y_jh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} \left(y_{jhi} - \bar{Y}_{jh}\right)^2$$

$$S_{x_jh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} \left(x_{jhi} - \bar{X}_{jh}\right)^2$$

$$S_{y_jx_jh} = \text{cov}(y_j, x_j)_h = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{jhi} - \bar{Y}_{jh})(x_{jhi} - \bar{X}_{jh})$$

$$R_j = \frac{\bar{Y}_j}{\bar{X}_j}$$

$$\bar{X}_{jh} = \frac{\sum_{i=1}^{N_h} x_{jhi}}{N_h}$$

and

$$\bar{Y}_{jh} = \frac{\sum_{i=1}^{N_h} y_{jhi}}{N_h}$$

Assume a linear cost function $C = c_0 + \sum_{h=1}^{L} c_h n_h$, where $c_h = \sum_{j=1}^{p} c_{jh}$, $h = 1, 2, \ldots, L$ denote the cost of measuring all the p characteristics on a sampled unit from the hth stratum and c_{jh} denote the per unit cost of measuring the jth characteristic in the hth stratum.

For estimating of the population means $\bar{Y}_j$, Chatterjee (1967) proposed a compromised allocation by minimizing the total relative increase in variances due to the use of a nonoptimum allocation for a fixed budget C_0. Chatterjee formulated the problem as:

$$\text{Minimize} \quad \sum_{j=1}^{p}\frac{1}{C}\sum_{h=1}^{L}\frac{C_h\left(n'_{jh} - n_h\right)^2}{n_h}$$

$$\text{subject to} \quad \sum_{h=1}^{L} c_h n_h = C,$$

where $C = C_0 - c_0$ and n'_{jh} is the individual optimum allocation for the jth characteristic in the hth stratum for a fixed value of C.

Chatterjee used Lagrange multiplier technique to work out the compromise allocation as

$$n_h = \frac{C\sqrt{\sum_{j=1}^{p} n'_{jh}}}{\sum_{h=1}^{L} c_h\sqrt{\sum_{j=1}^{p} n'_{jh}}}; \quad h = 1, 2, \ldots, L \tag{5}$$

The values of n_h given by (5) are rounded off to the nearest integer for practical use.

Cochran (1977) suggested a compromise allocation by averaging the individual optimum allocation $n^{\dagger}_{jh}$ over the p characteristics, that is,

$$n_h = \frac{1}{p}\sum_{j=1}^{p} n^{\dagger}_{jh}, \quad \text{where} \quad n^{\dagger}_{jh} = \frac{CW_h A_{jh}/\sqrt{c_h}}{\sum_{h=1}^{L} W_h A_{jh}\sqrt{c_h}} \tag{6}$$

Another alternative compromise allocation indicated by Sukhatme et al. (1984) that

minimizes the trace of the variance-covariance matrix is given by

$$n_h = \frac{CW_h\sqrt{\sum_{j=1}^{p} A^2_{jh}/c_h}}{\sum_{h=1}^{L} W_h\sqrt{c_h\sum_{j=1}^{p} A^2_{jh}}} \tag{7}$$

If the strata sizes are not large enough, then the above compromise allocations of Chatterjee, Cochran, and Sukhatme may give us an allocation in which some $n_h \geq N_h$; that is, the problem of over-sampling occurs. Furthermore, to estimate the parameters $S^2_{y_jh}$, $S^2_{x_jh}$ and $S_{y_jx_jh}$, we must have $n_h \geq 2$, and in order to implement the allocation, we need integer values of n_h, for the reasons discussed in Section 1.

When n_h are restricted to be integers the equality sign in the cost constraint may give a distorted optimum or even result in an empty feasible region. Interested readers may find the related discussion in Khan et al. (1997).

After taking care of all the factors discussed above, the problem of finding the optimum allocation $(n_1, \ldots, n_L)$ that minimizes the $\text{var}(\bar{y}_{j,st})$ in (2) simultaneously for all $j = 1, 2, \ldots, p$ may be formulated as the following AIMNLPP:

$$\begin{aligned} &\text{Minimize} && [\text{var}(\bar{y}_{1,st}),\ \text{var}(\bar{y}_{2,st}),\ \ldots,\ \text{var}(\bar{y}_{p,st})] \\ &\text{subject to} && \sum_{h=1}^{L} c_h n_h \leq C \\ & && 2 \leq n_h \leq N_h \\ &\text{and} && n_h \text{ are integers}, \quad (h = 1, 2, \ldots, L) \end{aligned} \tag{8}$$

In Expression (2) for $\text{var}(\bar{y}_{j,st})$ the second summation is independent of n_h and therefore may be ignored for the purpose of minimization, giving the AIMNLPP as:

$$\begin{aligned} &\text{Minimize} && \left[\sum_{h=1}^{L}\frac{W_h^2 A^2_{1h}}{n_h}, \sum_{h=1}^{L}\frac{W_h^2 A^2_{2h}}{n_h}, \ldots, \sum_{h=1}^{L}\frac{W_h^2 A^2_{ph}}{n_h}\right] \\ &\text{subject to} && \sum_{h=1}^{L} c_h n_h \leq C \\ & && 2 \leq n_h \leq N_h \\ &\text{and} && n_h \text{ are integers}, \quad (h = 1, 2, \ldots, L) \end{aligned} \tag{9}$$

The next section deals with the solution procedure using goal programming technique.

3. The Solution: Using Goal Programming

The AIMNLPP (9) can be solved by using some scalarization methods in which the multiobjective functions are converted into a single objective function (see Khan et al. 1997, 2003). In this section, the multiobjective programming problem (9) is solved using the goal programming technique.

Let n^*_{jh} denote the individual optimum allocation for the jth characteristic, that is, n^*_{jh} is the solution to the following all-integer nonlinear programming problem (AINLPP):

$$\begin{aligned} &\text{Minimize} && \sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n_h}; \quad j = 1, 2, \ldots, p \\ &\text{subject to} && \sum_{h=1}^{L} c_h n_h \le C \\ & && 2 \le n_h \le N_h \\ &\text{and} && n_h \text{ are integers}; \quad h = 1, 2, \ldots, L \end{aligned} \tag{10}$$

and can be obtained by using dynamic programming technique (see Arthanari and Dodge, 1981).

Let $V^*_j = \sum_{h=1}^{L} W_h^2 A_{jh}^2 / n^*_{jh}$ be the optimum value of the objective function at n^*_{jh}.

Further, let n^*_h be the optimum compromise allocation, that is, n^*_h is the solution to the AIMNLPP (9) with V_j as the value of the jth objective function.

Obviously,

$$V_j = \sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n^*_h} \ge V^*_j$$

and the quantity $V_j - V^*_j \ge 0$; $j = 1, 2, \ldots, p$ will denote the increase in the variance of the jth characteristics for using n^*_h instead of n^*_{jh}.

The solution procedure of AIMNLPP (9) starts with setting the following "goal":

"For the jth characteristic, the increase in the value of the objective function due to the use of a compromise allocation, n^*_h, should not exceed v_j".

Therefore, n^*_h must satisfy:

$$\sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n^*_h} - V^*_j \le v_j$$

or

$$\sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n^*_h} - v_j \le V^*_j \tag{11}$$

A reasonable compromise criterion to determine the optimum compromise allocation n^*_h can be worked out as to minimize the weighted sum of the increase in the variances of all the p characteristics for using n^*_h instead of n^*_{jh}, that is, to

$$\text{minimize} \quad \sum_{j=1}^{p} a_j v_j \tag{12}$$

where $a_j \ge 0; j = 1, 2, \ldots, p$ are the weights assigned to v_j according to the relative importance of the characteristics. As the characteristics are expected to be homogeneous within strata, a_j are assumed to be equal, in general, and then (12) reduces to

$$\text{minimize} \quad \sum_{j=1}^{p} v_j$$

From (11) and (12), one should note that the purpose of the proposed formulation in (12) is to minimize the total increase in the variances. However, it is possible that if the jth characteristic in the hth stratum $(h = 1, 2, \ldots, L)$ is heterogeneous, it may produce more loss in precision in the estimate of the stratum mean, as the value of A^2_{jh} for that characteristic is expected to be high, which may result in an increase in the value of v_j. If one has enough evidence, a_j may be predetermined as considered by Geary (1949) or may be determined by some other method, for example by use of the money value of the different variates as proposed by Mahalanobis (1944) (see Dalenius 1957, Chapter 9). Khan and Ahsan (2003) also studied this case and proposed a procedure of determining the unequal weights of the characteristics to control the increase in variances. This conjecture gives the choice of weights a_j such that:

$$a_j = \max(a_{j1}, a_{j2}, \ldots, a_{jL}); \quad (j = 1, 2, \ldots, p) \tag{13}$$

where a_{jh} are the weights for jth $(j = 1, 2, \ldots, p)$ characteristic in the hth $(h = 1, 2, \ldots, L)$ stratum and are obtained in proportion to A^2_{jh}, that is, $a_{jh} \propto A^2_{jh}$.

Letting $\sum_{j=1}^{p} a_{jh} = 1$, a_{jh} are worked out as:

$$a_{jh} = \frac{A^2_{jh}}{\sum_{j=1}^{p} A^2_{jh}}; \quad j = 1, 2, \ldots, p \text{ and } h = 1, 2, \ldots, L \tag{14}$$

Therefore, incorporating the goal and the objective given in (11) and (12), respectively, the equivalent problem to the AIMNLPP (9) may be expressed as the following goal programming problem (GPP):

$$\begin{aligned} &\text{Minimize} && \sum_{j=1}^{p} a_j v_j \\ &\text{subject to} && \sum_{h=1}^{L} \frac{W_h^2 A_{jh}^2}{n^*_h} - v_j \le V^*_j \\ & && \sum_{h=1}^{L} c_h n_h \le C \\ & && 2 \le n_h \le N_h \\ & && n_h \ge 0 \text{ are integers}, \quad (h = 1, 2, \ldots, L) \\ &\text{and} && v_j \ge 0, \quad (j = 1, 2, \ldots, p) \end{aligned} \tag{15}$$

The GPP in (15) may be solved by executing a program in the LINGO software package. For a numerical example, a program is coded in LINGO programming language with functions "MIN" and "@GIN" for the solution procedure as discussed above, and is available from the authors. For more information one may visit the site http://www.lindo.com. The documentation and the trial version of the software that has all the features and functionality of the standard version but with limited capacities may be downloaded from this site. The trial version enable users to become familiar with the software and its features and to solve sample problems.

Table 1. Data for four Strata, two Main and two Auxiliary Characteristics

h	N_h	W_h	$S^2_{y_1h}$	$S^2_{x_1h}$	$S^2_{y_2h}$	$S^2_{x_2h}$	$S_{x_1y_1h}$	$S_{x_2y_2h}$
1	8	0.0808	29,267,524,195.5	21,601,503,189.8	777,174.1	1,154,134.2	24,360,422,802.3	904,170.6
2	34	0.3434	26,079,256,582.8	19,734,615,816.7	4,987,812.9	7,056,074.8	22,003,466,630.3	5,813,439.5
3	45	0.4545	42,362,842,460.8	27,129,658,750.0	1,074,510.6	2,082,871.3	33,367,597,192.0	1,285,355.6
4	12	0.1212	30,728,265,336.9	17,258,237,358.5	388,378.5	732,004.9	21,033,769,867.3	456,991.5

4. Numerical Examples

The following two numerical examples are presented to illustrate the practical use and the computational details of the proposed solution procedure. The data of these examples are from the 2002 and 1997 Agricultural Censuses in Iowa State conducted by the National Agricultural Statistics Service, USDA, Washington D.C. (source: http://www.agcensus.usda.gov/). For the purpose of the illustration, the data of $N = 99$ counties in Iowa State are divided into four strata. The relevant data in respect to two characteristics, (i) the quantity of corn harvested and (ii) the quantity of oats harvested are given in Table 1 where Y_1 = The quantity of corn harvested in 2002, Y_2 = the quantity of oats harvested in 2002, X_1 = The quantity of corn harvested in 1997, and X_2 = the quantity of oats harvested in 1997. The values of X_1 and X_2 are used as the auxiliary information on the main variables Y_1 and Y_2, respectively, with $\bar{X}_1 = 405{,}654.19$, $\bar{Y}_1 = 474{,}973.90$, $\bar{X}_2 = 2{,}116.70$, and $\bar{Y}_2 = 1{,}576.25$. Therefore, $R_1 = \bar{Y}_1/\bar{X}_1 = 1.1709$ and $R_2 = \bar{Y}_2/\bar{X}_2 = 0.7447$.

Example 1. For the combined ratio estimate the numerical values of A^2_{jh}; $j = 1, 2$ and $h = 1, 2, \ldots, 4$ are worked out for the data in Table 1 using (4) and presented in Table 2.

We assume that the costs of measurement, c_h, in the four strata are $c_1 = 15$, $c_2 = 7$, $c_3 = 5$ and $c_4 = 9$ units, respectively, and the total amount available for conducting the survey is $C_0 = 250$ units, which includes an expected overhead cost $c_0 = 50$ units. The total amount available for the measurements is $C = 200$ units.

Using the values in Tables 1 and 2, the rounded-off compromise allocations of Chatterjee, Cochran, and Sukhatme given by (5), (6) and (7), respectively, are shown in Table 3. Note that these allocations are infeasible to the problem (9) because they violate the constraint $n_1 \geq 2$.

To work out the compromise allocation using the goal programming approach as discussed in Section 3, the AINLPP (10) is solved for each individual characteristic.

Using the LINGO program, the individual optimum allocations for $j = 1$ and 2 are found to be:

$$n^*_{11} = 2, n^*_{12} = 7, n^*_{13} = 17, n^*_{14} = 4 \text{ with minimum variance } V^*_1 = 96,754,589.11.$$

and

$$n^*_{21} = 2, n^*_{22} = 11, n^*_{23} = 15, n^*_{24} = 2 \text{ with minimum variance } V^*_2 = 27,061.62.$$

Table 2. Values of A^2_{jh}

h	A^2_{1h}	A^2_{2h}
1	2,230,755,317.1	242,098.5
2	1,875,241,116.7	1,047,747.4
3	2,939,641,936.7	920,058.2
4	6,029,490,311.7	321,766.5
Total	13,075,128,682.2	2,531,670.6

Table 3. Individual sampling variances (ignoring fpc) and relative efficiency (RE) of different allocations with respect to the proportional allocation

	Allocation				
	Proportional	Chatterjee	Cochran's average	Minimizing trace (Sukhatme)	Proposed
n_1	2	1	1	1	2
n_2	10	10	10	8	7
n_3	13	16	17	16	17
n_4	3	3	3	5	4
$\sum_{h=1}^{4} c_h n_h$	192	192	197	196	200
var $(\bar{y}_{1,st})$ (% increased[a])	105,650,784.8 (9.19)	104,174,082.3 (1.18)	101,941,126.3 (5.36)	97,891,892.2 (7.67)	96,754,589.1 (0.00)
var $(\bar{y}_{2,st})$ (% increased[a])	29,346.8 (8.44)	27,395.5 (10.32)	26,696.7 (0.00)	29,854.7 (1.23)	30,808.5 (13.84)
Trace (Sum of Variances)	105,680,131.6	104,201,477.9	101,967,823.0	97,921,746.8	96,785,397.6
R.E. w. r. t. Proportional Allocation (%)	100	101.42	103.64	107.92	109.19

[a] Percentage increase in the variance due to not using individual optimum allocation $\left(n_{jh}^{*}\right)$

From (13) and (14), the weights to be assigned to the first and second characteristics are

$$a_1 = \max \begin{pmatrix} 0.1706 \\ 0.1434 \\ 0.2248 \\ 0.4611 \end{pmatrix} = 0.4611 \text{ and } a_2 = \max \begin{pmatrix} 0.0956 \\ 0.4139 \\ 0.3634 \\ 0.1271 \end{pmatrix} = 0.4139, \text{ respectively,}$$

which shows that the weights are approximately equal.

For the purpose of comparison of the proposed allocations with other allocations, we may assume that the weights are approximately equal – that is, both the characteristics are equally important and $a_1 = a_2 = 0.5$. The GPP (15), which is equivalent to AIMNLPP (9), may be expressed as:

$$\text{Minimize} \quad 0.5v_1 + 0.5v_2$$

$$\text{subject to} \quad \frac{14{,}566{,}711.59}{n_1} + \frac{221{,}179{,}342.00}{n_2} + \frac{607{,}364{,}036.51}{n_3} + \frac{88{,}587{,}552.79}{n_4} - v_1 \leq 96{,}754{,}589.11,$$

$$\frac{1{,}580.89}{n_1} + \frac{123{,}578.82}{n_2} + \frac{190{,}094.68}{n_3} + \frac{4{,}727.52}{n_4} - v_2 \leq 27{,}061.62,$$

$$15n_1 + 7n_2 + 5n_3 + 9n_4 \leq 200,$$

$$2 \leq n_1 \leq 8,$$

$$2 \leq n_2 \leq 34,$$

$$2 \leq n_3 \leq 45,$$

$$2 \leq n_4 \leq 12,$$

$$n_1, n_2, n_3, \text{ and } n_4 \text{ are integers}$$

$$\text{and} \quad v_1, v_2 \geq 0.$$

Using the LINGO program, the optimum compromise allocation is obtained as:

$$n_1 = 2, \quad n_2 = 7, \quad n_3 = 17, \quad n_4 = 4.$$

The results of compromise allocations using different criteria are summarized in Table 3. All the four compromise allocations, Chatterjee's, Cochran's, Sukhatme's and the allocation proposed by authors, are compared with the proportional allocation $\left(n_h = CW_h / \sum_{h=1}^{L} W_h c_h\right)$. The Relative Efficiency (R.E.) of an allocation $(\underset{\sim}{n})$ with respect to another allocation $(\underset{\sim}{n}^*)$, is defined as R.E. $=$ Trace $(\underset{\sim}{n}^*)$/Trace $(\underset{\sim}{n})$.

Example 2. In the data of Example 1 we change only the cost of measurement c_h to equal cost in all four strata, as $c_1 = c_2 = c_3 = c_4 = 7.5$ units, which is common in practice and motivates surveyors to use the proportional allocation as a kind of noninformative prior choice. Keeping the remaining values the same as in Example 1, we arrive at the results given in Table 4.

Table 4 shows that only the proposed and the proportional allocations are feasible while other allocations are infeasible because they violate the cost constraint $\sum_{h=1}^{4} c_h n_h \leq 200$.

The proportional allocation, although feasible, is less efficient. Furthermore, the proposed allocation has a relative efficiency of 104.46% as compared to the proportional allocation.

5. Discussion

In this section, we compare the proposed compromise allocation with the other available compromise allocations discussed in Section 2 and 4, when the true values of the parameters A_{jh}^2; $(j = 1, 2 \text{ and } h = 1, 2, \ldots, 4)$ are available. For the data in Example 1, if the restrictions $2 \leq n_h \leq N_h$ are relaxed to $1 \leq n_h \leq N_h$, all the allocations given in Table 3 become feasible and can be considered for the comparison.

Rows 6 and 7 in Table 3 show the individual sampling variances of $\bar{y}_{j,st}$ for the different allocations, and their relative efficiencies with respect to the proportional allocation are given in Row 9. It reveals that different allocations differ considerably from each other, and the maximum relative efficiency is attained for the proposed allocation. Also, the proposed allocation, which gives least variances for the estimates, is most efficient among all allocations, and the gain in efficiency of the proposed allocation over the proportional allocation is 9.19%.

The percentage increase in the variance of estimates for different characteristics due to not using the individual optimum allocations $\left(n_{jh}^*\right)$ are also presented in the parentheses of Rows 6 and 7. It is evident that the compromise criteria differ little from each other with respect to the percentage increase in the variance. The proposed compromise allocation gives 0% increase in variance for $\bar{y}_1$ and slightly higher percentage increase for $\bar{y}_2$ as compared to other allocations. However, one should remember that the other compromise allocations are infeasible and therefore are of no practical use as they violate the constraint $n_1 \geq 2$.

Thus we conclude that the proposed allocation provides an answer to the problem of determining integer optimum compromise allocation in the situation where other allocations do not guarantee even a feasible solution. The proposed compromise allocation is also more precise than other compromise allocations discussed here. Furthermore, the proposed method could easily be implemented by survey designers using LINGO without much knowledge of computer programming.

Table 4. Allocation with equal measurement cost between strata

Allocations	n_1	n_2	n_3	n_4	$\sum_{h=1}^{4} c_h n_h$	RE as compared to proportional allocation %
Proportional	2	9	12	3	195	100.00
Chatterjee	2	9	13	4	210	111.19
Cochran	2	9	13	3	202.5	103.60
Sukhatme	2	8	12	5	202.5	108.46
Proposed	2	7	12	5	195	104.46

6. References

Ahsan, M.J. and Khan, S.U. (1977). Optimum Allocation in Multivariate Stratified Random Sampling Using Prior Information. Journal of the Indian Statistical Association, 15, 57–67.

Ahsan, M.J. and Khan, S.U. (1982). Optimum Allocation in Multivariate Stratified Random Sampling with Overhead Cost. Metrika, 29, 71–78.

Arthanari, T.S. and Dodge, Y. (1981). Mathematical Programming in Statistics. New York: John Wiley.

Aoyama, H. (1963). Stratified Random Sampling with Optimum Allocation for Multivariate Populations. Annals of the Institute of Statistical Mathematics, 14, 251–258.

Bethel, J. (1985). An Optimum Allocation Algorithm for Multivariate Surveys. Proceedings of the American Statistical Association, Section on Survey Methods Research, 209–212.

Bethel, J. (1989). Sample Allocation in Multivariate Surveys. Survey Methodology, 15, 47–57.

Chatterjee, S. (1967). A Note on Optimum Allocation. Skandinavisk Aktuarietidskrift, 50, 40–44.

Chatterjee, S. (1968). Multivariate Stratified Surveys. Journal of the American Statistical Association, 63, 530–534.

Chromy, J. (1987). Design Optimization with Multiple Objectives. Proceedings of the American Statistical Association, Section on Survey Methods Research, 194–199.

Cochran, W.G. (1977). Sampling Techniques (Third Edition). New York: John Wiley and Sons, Inc.

Dalenius, T. (1953). The Multivariate Sampling Problem. Skandinavisk Aktuarietidskrift, 36, 92–102.

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

Dayal, S. (1985). Allocation of Sample Using Values of Auxiliary Characteristic. Journal of Statistical Planning and Inference, 11, 321–328.

Folks, J.L. and Antle, C.E. (1965). Optimum Allocation of Sampling Units to the Strata when there are R Responses of sampling Units of Interest. Journal of the American Statistical Association, 60, 225–233.

Geary, R.C. (1949). Sampling Methods Applied to Irish Agricultural Statistics. Technical Series, Central Statistical Office, Dublin, September.

Ghosh, S.P. (1958). A Note on Stratified Random Sampling with Multiple Characters. Calcutta Statistical Association Bulletin, 8, 81–89.

Jahan, N., Khan, M.G.M., and Ahsan, M.J. (1994). A Generalized Compromise Allocation. Journal of the Indian Statistical Association, 32, 95–101.

Khan, M.G.M. and Ahsan, M.J. (2003). A Note on Optimum Allocation in Multivariate Stratified Sampling. The South Pacific Journal of Natural Science, 21, 91–95.

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

Khan, M.G.M., Khan, E.A., and Ahsan, M.J. (2003). An Optimal Multivariate Stratified Sampling Design Using Dynamic Programming. Australian & New Zealand Journal of Statistics, 45, 107–113.

Kokan, A.R. and Khan, S.U. (1967). Optimum Allocation in Multivariate Surveys: An Analytical Solution. Journal of the Royal Statistical Society, Series B, 29, 115–125.

Mahalanobis, P.C. (1944). On Large-scale Sample Surveys. Philosophical Transactions of the Royal Society of London, Series B, 231, 329–451.

Neyman, J. (1934). On the Two Different Aspects of the Representative Methods: The Method Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, 97, 558–606.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984). Sampling Theory of Surveys with Applications. Ames, IA: Iowa State University Press.

Yates, F. (1960). Sampling Methods for Census and Surveys (Second Edition). London: Charles Griffin and Co. Ltd.

Received January 2008
Revised June 2010