Hacettepe Journal of Mathematics and Statistics
Volume 33 (2004), 91 – 96

DETERMINATION OF COMPROMISE INTEGER STRATA SAMPLE SIZES USING GOAL PROGRAMMING

Mustafa Semiz*

Received 15 : 09 : 2003 : Accepted 03 : 11 : 2004

Abstract

This article deals with the determination of compromise integer strata sample sizes using goal programming in multivariate stratified sampling. Firstly, the problem of determining optimum integer strata sample sizes is formulated for the univariate case, and then based on these individual optimal solutions, individual goal variances are calculated. A new compromise criteria is defined for the goal programming approach based on predetermined or calculated goal variances. It is shown that the proposed approach provides relatively more efficient and feasible compromise integer strata sample sizes for multivariate surveys.

Keywords: Stratified sampling, Compromise allocation, Goal programming, Relative efficiency.

1. Introduction

Several alternative compromise criteria and methods have been suggested in order to determine strata sample sizes for multivariate surveys by authors such as Neyman [6], Cochran [2], Chatterjee [1], Kokan and Khan [5], Sukhatme, Sukhatme, Sukhatme, and Asok [8], Jahan, Khan, and Ahsan [3], Khan, Ahsan, and Jahan [4], etc. Determining the compromise strata sample sizes in multivariate stratified sampling has been commonly called compromise allocation. If the total sample size is known and this sample size is divided among stratum, it is called an allocation procedure. However, this study is intended to determine strata sample sizes directly, and the proposed goal programming approach does not involve any allocation techniques. The problem of determining compromise strata sample sizes may be defined as a goal programming problem, since it consists of multiple objectives. In this study, the compromise criteria is the sum of the

*Selçuk University, Faculty of Art and Sciences, Department of Statistics, Konya, Turkey.
E-mail : msemiz@selcuk.edu.tr

proportional increase in variances resulting from absolute deviations from the individual desired variances over all k characteristics. The criterion is formulated as

$$\text{minimize} \sum_{j=1}^{k} \frac{\left|V_{\text{comp}}(\bar{y}_j) - V_d(\bar{y}_j)\right|}{V_d(\bar{y}_j)}, \tag{1.1}$$

where $V_{\text{comp}}(\bar{y}_j)$ is the variance of the sample mean of the jth characteristic under optimum compromise integer strata sample sizes (n_h^*), and $V_d(\bar{y}_j)$ is the desired variance of the sample mean of the jth characteristic under optimum individual strata sample sizes (n_{jh}) in the hth strata. The desired individual variance $V_d(\bar{y}_j)$ can be either predetermined or calculated. If one has no idea how to predetermine $V_d(\bar{y}_j)$, the minimum value of the individual variances $V_{\min}(\bar{y}_j)$ can be used instead of the desired variance $V_d(\bar{y}_j)$. In the first step, the desired individual optimal variances should be predetermined or calculated as $V_{\min}(\bar{y}_j)$ for every characteristic.

2. The individual optimal integer strata sample sizes

The most popular way of calculating the individual optimal strata sample sizes for the jth characteristic in the hth strata is to use the equation

$$n_{jh} = \frac{CW_hS_{jh}/\sqrt{c_h}}{\sum_{h=1}^{L} W_hS_{jh}\sqrt{c_h}}, \tag{2.1}$$

as indicated by Cochran [2], where c_h is the cost of a sample taken from the hth strata, W_h is the weight of the size of the hth strata, $(N_h/\sum_h N_h)$ and S_{jh} is the standard deviation of the hth strata for the jth characteristic. The solution of equation (2.1) depends on the total sampling cost function $f = \sum_{h=1}^{L} c_h n_{jh}$ and a fixed budget C. It is known that equation (2.1) provides non-integer solutions, and Khan, Ahsan, and Jahan [4] showed that it sometimes provides unfeasible solutions, too. However, they used these solutions as an initial point of their algorithms for determining the optimum compromise integer strata sample sizes in multivariate surveys.

For the univariate case, the goal is to minimize the jth individual variance, $V(\bar{y}_j)$, subject to $f \le C$, $n_{jh} \le N_h$, where n_{jh} are integers $(h = 1, 2, \ldots, L)$. This problem can also be presented as a non-linear integer programming (NIP) problem, as proposed by Semiz and Esin [7]. This problem for every jth characteristic is formulated by the following model:

$$\begin{aligned} &\text{minimize } V_{\min}(\bar{y}) = \sum_{h=1}^{L} \frac{W_h^2 S_{jh}^2}{n_{jh}} \\ &\text{subject to } f \le C \\ &\qquad 0 \le n_{jh} \le N_h,\ h = 1, 2, \ldots, L, \\ &\qquad n_{jh} \text{ are integers},\ h = 1, 2, \ldots, L. \end{aligned} \tag{2.2}$$

Individual optimum integer values n_{jh} can be determined by solving the problem (2.2) using the *Lingo* package program [9]. The NIP solution of the problem (2.2) has advantages over the solution of equation (2.1) since one can add different constraints to problem (2.2), and obtain optimal integer results.

2.1. Example. The data, exhibited in Table 1, of the example reviewed by Khan, Ahsan, and Jahan [4], is reconsidered here for the comparison of alternative methods.

Table 1. Data for five strata and three characteristics.

	c_h	N_h	W_h	S_{1h}	S_{2h}	S_{3h}	$W_h^2 S_{1h}^2$	$W_h^2 S_{2h}^2$	$W_h^2 S_{3h}^2$
1	3	39,552	0.197	4.6	11.7	332	0.82119844	5.31256401	4277.683216
2	4	38,347	0.191	3.4	9.8	357	0.42172036	3.50363524	4649.466969
3	5	43,969	0.219	3.3	7.0	246	0.52229529	2.35008900	2902.407876
4	6	36,942	0.184	2.8	6.5	173	0.26543104	1.43041600	1013.276224
5	7	41,760	0.208	3.7	9.8	279	0.59228416	4.15507456	3367.713024

The data includes three characteristics:

(i) The number of cows milked per day,
(ii) The number of gallons of milk yielded per day,
(iii) The total annual cash receipts from dairy products.

The fixed budget for this sampling design is $C = 5,000\,\$$. The individual optimum solutions of Cochran's equation (2.1), and the NIP problem defined in (2.2), are illustrated in Table 2.

Table 2. The individual optimal strata sample sizes, cost and variances obtained from the solution of the Cochran (2.1) and NIP (2.2) methods.

	Cochran (2.1)			NIP (2.2)		
h/j	1	2	3	1	2	3
1	336	341	314	335	340	314
2	209	240	283	210	239	284
3	208	176	200	207	175	200
4	135	125	108	135	126	108
5	187	198	182	187	199	182
Total cost (\$)	**5,003**	4,999	4,996	4,999	5,000	5,000
$V_{\min}(\bar{y}_j)$	0.01210	0.07595	72.45055	0.01211	0.07594	72.39270

The NIP (2.2) solutions are feasible solutions which do not violate any constraints at all. However, sometimes Cochran's solutions may violate some of the constraints due to rounding off. In this example, for the first characteristic the sampling cost is over the fixed budget of 5,000 \$. The individual variances are smaller with the NIP (2.2) solutions. These optimum integer individual strata sample sizes determined by NIP can be considered as a starting point for algorithms such as Dynamic programming used by Khan, Ahsan, and Jahan [4], and the related individual minimum variances $V_{\min}(\bar{y}_j)$ are considered as the individual desired variances $V_d(\bar{y}_j)$.

3. Compromise integer solution via goal programming

Goal programming aims to attain predetermined goals for multiple objectives. In multivariate surveys, there are k predetermined goal variances $V_d(\bar{y}_j)$. Therefore, there

are k absolute deviations between the compromise variances $V_{\text{comp}}(\bar{y}_j)$ and the minimum individual, or desired known variances $V_d(\bar{y}_j)$. The absolute positive deviations are formulated as

$$V_{\text{comp}}(\bar{y}_j) = d_j^- - d_j^+ = V_d(\bar{y}_j),\ j = 1, 2, \ldots, k,$$

where

$$\begin{aligned} V_{\text{comp}}(\bar{y}_j) < V_d(\bar{y}_j) &\implies d_j^- > 0,\ d_j^+ = 0, \\ V_{\text{comp}}(\bar{y}_j) > V_d(\bar{y}_j) &\implies d_j^- = 0,\ d_j^+ > 0, \\ V_{\text{comp}}(\bar{y}_j) = V_d(\bar{y}_j) &\implies d_j^- = 0,\ d_j^+ = 0. \end{aligned}$$

For the jth characteristic, if the variances are not equal, one of these positive deviations d_j^+ or d_j^- come into existence. Therefore, the decision criteria is to minimize the sum of the deviations d_j^+ and d_j^-. However, the deviations of different characteristics may have different units. For each characteristic, the deviation d_j^+ or d_j^- becomes unit free by applying the transformation

$$\frac{d_j^+}{V_d(\bar{y}_j)} \text{ or } \frac{d_j^-}{V_d(\bar{y}_j)},\ j = 1, 2, \ldots, k, \tag{3.1}$$

respectively. As seen in Equation (3.1), the jth unit free standardized deviation is equal to the jth proportional increase in variance resulting from the absolute deviation between $V_{\text{comp}}(\bar{y}_j)$ and $V_d(\bar{y}_j)$ in (1.1). Consequently, by using goal programming, the compromise integer strata sample sizes in the multivariate case may be presented as the following nonlinear integer programming problem:

$$\begin{aligned} &\text{minimize} \sum_{j=1}^{k} w_j \frac{d_j^+ + d_j^-}{V_d(\bar{y}_j)} \equiv \text{minimize} \sum_{j=1}^{k} w_j \frac{|V_{\text{comp}}(\bar{y}_j) - V_d(\bar{y}_j)|}{V_d(\bar{y}_j)} \\ &\text{subject to } V_{\text{comp}}(\bar{y}_j) + d_j^- - d_j^+ = V_d(\bar{y}),\ j = 1, 2, \ldots, k, \\ &\qquad f_c \le C, \\ &\qquad 1 \le n_h^* \le N_h,\ h = 1, 2, \ldots, L, \\ &\qquad n_h^* \text{ are integers},\ h = 1, 2, \ldots, L, \end{aligned} \tag{3.2}$$

where $f_c = \sum_{h=1}^{L} c_h n_h^*$ can be of any form. The problem (3.2) may accept many constraints, and w_j can be added as the weight of the jth characteristic according to its importance. Therefore, this approach is much more flexible than the other algorithms. This problem can be solved by the Lingo package program [9].

Taking NIP (2.2) individual optimal solutions as the desired variances, and assuming the importance of all characteristics to be equal ($w_j = 1,\ j = 1, 2, 3$), the solution of the compromise integer problem defined in (3.2) gives the compromise integer strata sample sizes as

$$n_1^* = 329,\ n_2^* = 246,\ n_3^* = 195,\ n_4^* = 123,\ n_5^* = 188,$$

and the compromise variances of characteristics as

$$V_{\text{comp}}(\bar{y}_1) = 0.012197215,\ V_{\text{comp}}(\bar{y}_2) = 0.076172628,\ V_{\text{comp}}(\bar{y}_3) = 72.93787706.$$

4. Comparisons and Conclusions

The mean sum of relative efficiencies of variances is used for the comparison of the proposed goal programming approach (3.2) with other compromise methods. The compared methods are:

i) Minimizing the trace of the covariance matrix, as proposed by Sukhatme, Sukhatme, Sukhatme and Asok [8],
ii) Averaging the individual strata sample sizes over the characteristics calculated using (2.1),
iii) Minimizing the total relative increase in the variances, as proposed by Chatterjee [1],
iv) Minimizing the total relative increase in the variances with integer restrictions, as proposed by Khan, Ahsan, and Jahan [4], and
v) Minimizing the total proportional increase in variances, as proposed by the author (3.2).

Since every characteristic can have different units, in (i), the appropriateness of the sum of the variances should be reevaluated carefully. The compromise strata sample sizes are presented in Table 3 for each method.

Table 3. Compromise Strata Sample Sizes for the Methods Compared.

Methods and compromise integer strata sample sizes	n_1^*	n_2^*	n_3^*	n_4^*	n_5^*
(i) Minimizing the trace	314	283	200	108	182
(ii) Cochran's Average	330	244	195	123	189
(iii) Chatterjee's Method	330	245	195	123	189
(iv) Integer DP	331	246	195	123	187
(v) Proposed integer GP	329	246	195	123	188

The comparison is based on the mean sum of relative efficiencies (MSRE) of each method:

$$\text{MSRE} = \frac{1}{k}\sum_{j=1}^{k}\frac{V_{\text{comp}}(\bar{y}_j)}{V_{\min}(\bar{y}_j)} = \frac{1}{k}\text{SRE}. \tag{4.1}$$

Table 4. Sum of relative efficiencies (SRE) and mean sum of relative efficiencies (MSRE) as compared to the optimal individual variances determined by NIP (2.2).

Methods and compromised variances	$V(\bar{y}_1)$	$V(\bar{y}_2)$	$V(\bar{y}_3)$	SRE	MSRE (4.1)	Cost
Optimal Integer Individual (NIP) (2.2)	0.0121	0.0759	72.3927	3.0000	1.0000	$*$
(i) Minimizing the Trace	0.0124	0.0771	72.4506	3.0414	**1.0138**	4996
(ii) Cochran's Average	0.0122	0.0761	72.9586	3.0187	1.0062	**5002**
(iii) Chatterjee's Method	0.0122	0.0761	72.8808	3.0176	1.0059	**5006**
(iv) Integer DP	0.0122	0.0762	72.9551	3.0200	1.0067	4999
(v) Proposed integer GP (3.2)	0.0122	0.0762	72.9379	3.0197	***1.0066***	***5000***

Method (i) does not directly provide integer strata sample sizes. Solutions of (i) have to be rounded. The mean sum of the relative efficiencies of (i) is greater than for the proposed method (v). Also, the trace concept of the variance terms, which have different units, is still in question.

Methods (ii) and (iii) have lower MSRE values than the proposed method (v). However, these strata sample size solutions require to be rounded off. As seen in Table 4, these

solutions result in a cost over the budget and these methods may not provide feasible solutions.

Method (iv), proposed by Khan, Ahsan, and Jahan [4], is as efficient as the proposed method (v). However, the solutions of (iv) are determined by an algorithm for a fixed problem subject to the fixed cost function ($f_c = \sum_{h=1}^{L} c_h n_h^*$) and for a limited and specified set of constraints. Therefore, method (iv) is not flexible for different multivariate survey problems.

Method (v) is a mathematical programming model which optimizes the goal programming model subject to the constraints, which are the cost function and the integer strata sample sizes. Therefore, Method (v) always provides integer and feasible solutions for the strata sample sizes for compromise situation in multivariate stratified sampling problems. In addition this proposed goal programming method (v) has a flexible structure because it can accept different kinds of restrictions appropriate to different problems. Depending upon the problem structure, constraints may be deleted, added or changed in the new method (3.2). In addition to these advantages, the proposed goal programming solution method (v) has, for this specific example, the best MSRE value among the methods (i), (iv) and (v) that provide feasible solutions. In conclusion, as seen in this example, this goal programming method seems to provide a flexible approach as well as feasible and efficient integer compromise strata sample sizes in multivariate stratified sampling.

References

[1] Chatterjee, S. *A note on optimum allocation*, Scandinavian Actuarial Journal **50**, 40–44, 1967.

[2] Cochran, W. G. *Sampling Techniques*, 2nd Ed. (John Wiley and Sons Inc., New York, 1963).

[3] Jahan, N., Khan, M. G. M. and Ahsan, M. J. *A generalized compromise allocation*, Journal of the Indian Statistical Association **32**, 95–101, 1994.

[4] 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.

[5] Kokan, A. R. and Khan, S. *Optimum allocation in multivariate surveys: An analytical solution*, Journal of the Royal Statistical Society, Series B **29**, 115–125, 1967.

[6] Neyman, J. *On the two different aspects of the representative methods: The method of stratified sampling and the method of purposive selection*, Journal of the Royal Statistical Society **97**, 558–606, 1934.

[7] Semiz, M. and Esin, A. A. *Tabaka örnek hacimlerinin doğrusal olmayan tamsayılı programlama ile Belirlenmesi*, İstatistik Sempozyumu-2000, pp. 199–207, (Gazi University, Department of Statistics, Ankara, Turkey, 2000).

[8] Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. *Sampling Theory of Surveys with Applications*, (Indian Society of Agricultural Statistics, New Delhi, India, and Iowa State University Press, Ames, IA, 1984).

[9] Winston, W. L. *User's Guide for Lindo and Lingo: Operations Research: Applications and Algorithms, Introduction to Mathematical Programming: Applications and Algorithms*, (Duxbury Press, New York, 1997).