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

\[
(1.1) \quad \text{minimize} \sum_{j=1}^{k} \frac{|V_{\text{comp}}(\bar{y}_j) - V_{\text{d}}(\bar{y}_j)|}{V_{\text{d}}(\bar{y}_j)},
\]

where \(V_{\text{comp}}(\bar{y}_j)\) is the variance of the sample mean of the \(j\)th characteristic under optimum compromise integer strata sample sizes \((n_h)\), and \(V_{\text{d}}(\bar{y}_j)\) is the desired variance of the sample mean of the \(j\)th characteristic under optimum individual strata sample sizes \((n_{j,h})\) in the \(h\)th strata. The desired individual variance \(V_{\text{d}}(\bar{y}_j)\) can be either predetermined or calculated. If one has no idea how to predetermine \(V_{\text{d}}(\bar{y}_j)\), the minimum value of the individual variances \(V_{\text{min}}(\bar{y}_j)\) can be used instead of the desired variance \(V_{\text{d}}(\bar{y}_j)\).

In the first step, the desired individual optimal variances should be predetermined or calculated as \(V_{\text{min}}(\bar{y}_j)\) 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 \(j\)th characteristic in the \(h\)th strata is to use the equation

\[
(2.1) \quad n_{j,h} = \frac{CW_h S_{j,h}}{\sqrt{c_h} \sum_{h=1}^{L} W_h S_{j,h} \sqrt{c_h}},
\]
as indicated by Cochran [2], where \(c_h\) is the cost of a sample taken from the \(h\)th strata, \(W_h\) is the weight of the size of the \(h\)th strata, \((N_h/\sum_{h} N_h)\) and \(S_{j,h}\) is the standard deviation of the \(h\)th strata for the \(j\)th characteristic. The solution of equation (2.1) depends on the total sampling cost function \(f = \sum_{h=1}^{L} c_h n_{j,h}\) 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 \(j\)th individual variance, \(V(\bar{y}_j)\), subject to \(f \leq C\), \(n_{j,h} \leq N_h\), where \(n_{j,h}\) 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 \(j\)th characteristic is formulated by the following model:

\[
(2.2) \quad \text{minimize} \quad V_{\text{min}}(\bar{y}) = \sum_{h=1}^{L} \frac{W_h^2 S_{j,h}^2}{n_{j,h}}
\]
subject to \(f \leq C\)

\[
0 \leq n_{j,h} \leq N_h, \quad h = 1, 2, \ldots, L, \quad n_{j,h} \text{ are integers, } h = 1, 2, \ldots, L.
\]

Individual optimum integer values \(n_{j,h}\) 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.

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

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.

<table>
<thead>
<tr>
<th>$h/j$</th>
<th>Cochran (2.1)</th>
<th>NIP (2.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_h$</td>
<td>$S_{1h}$</td>
</tr>
<tr>
<td>1</td>
<td>336</td>
<td>341</td>
</tr>
<tr>
<td>2</td>
<td>209</td>
<td>240</td>
</tr>
<tr>
<td>3</td>
<td>208</td>
<td>176</td>
</tr>
<tr>
<td>4</td>
<td>135</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>187</td>
<td>198</td>
</tr>
</tbody>
</table>

Total cost ($)$: $\text{Cochran} = 5,000; \text{NIP} = 5,000$

$V_{\text{min}}(\hat{y}_j) = \begin{cases} 0.01210 & \text{Cochran} \\ 0.07595 & \text{NIP} \end{cases}$

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 $V_{\text{min}}(\hat{y}_j)$ are considered as the individual desired variances $V_d(\hat{y}_j)$.

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

\[
V_{\text{comp}}(\hat{y}_j) = d_j^- - d_j^+ = V_d(\hat{y}_j), \quad j = 1, 2, \ldots, k,
\]

where

\[
\begin{align*}
V_{\text{comp}}(\hat{y}_j) < V_d(\hat{y}_j) & \implies d_j^- > 0, \quad d_j^+ = 0, \\
V_{\text{comp}}(\hat{y}_j) > V_d(\hat{y}_j) & \implies d_j^- = 0, \quad d_j^+ > 0, \\
V_{\text{comp}}(\hat{y}_j) = V_d(\hat{y}_j) & \implies d_j^- = 0, \quad d_j^+ = 0.
\end{align*}
\]

For the \( j \)th 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

\[
(3.1) \quad \frac{d_j^+}{V_d(\hat{y}_j)} \quad \text{or} \quad \frac{d_j^-}{V_d(\hat{y}_j)}, \quad j = 1, 2, \ldots, k,
\]

respectively. As seen in Equation (3.1), the \( j \)th unit free standardized deviation is equal to the \( j \)th proportional increase in variance resulting from the absolute deviation between \( V_{\text{comp}}(\hat{y}_j) \) and \( V_d(\hat{y}_j) \) 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:

\[
(3.2) \quad \text{minimize } \sum_{j=1}^{k} w_j \frac{|V_{\text{comp}}(\hat{y}_j) - V_d(\hat{y}_j)|}{V_d(\hat{y}_j)} \equiv \text{minimize } \sum_{j=1}^{k} w_j \frac{V_{\text{comp}}(\hat{y}_j) - V_d(\hat{y}_j)}{V_d(\hat{y}_j)} \\
\text{subject to } V_{\text{comp}}(\hat{y}_j) + d_j^- - d_j^+ = V_d(\hat{y})_j, \quad j = 1, 2, \ldots, k, \\
f_c \leq C, \\
1 \leq n_h^j \leq N_h, \quad h = 1, 2, \ldots, L, \\
n_h^j \text{ are integers, } h = 1, 2, \ldots, L,
\]

where \( f_c = \sum_{h=1}^{L} n_h^j \) can be of any form. The problem (3.2) may accept many constraints, and \( w_j \) can be added as the weight of the \( j \)th 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

\[
n_1^c = 329, \quad n_2^c = 246, \quad n_3^c = 195, \quad n_4^c = 123, \quad n_5^c = 188,
\]

and the compromise variances of characteristics as

\[
V_{\text{comp}}(\hat{y}_1) = 0.012197215, \quad V_{\text{comp}}(\hat{y}_2) = 0.076172628, \quad V_{\text{comp}}(\hat{y}_3) = 72.93787706.
\]

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.**

<table>
<thead>
<tr>
<th>Methods and compromise integer strata sample sizes</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_4$</th>
<th>$n_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Minimizing the trace</td>
<td>314</td>
<td>283</td>
<td>200</td>
<td>108</td>
<td>182</td>
</tr>
<tr>
<td>(ii) Cochran’s Average</td>
<td>330</td>
<td>244</td>
<td>195</td>
<td>123</td>
<td>189</td>
</tr>
<tr>
<td>(iii) Chatterjee’s Method</td>
<td>330</td>
<td>245</td>
<td>195</td>
<td>123</td>
<td>189</td>
</tr>
<tr>
<td>(iv) Integer DP</td>
<td>331</td>
<td>246</td>
<td>195</td>
<td>123</td>
<td>187</td>
</tr>
<tr>
<td>(v) Proposed integer GP</td>
<td>329</td>
<td>246</td>
<td>195</td>
<td>123</td>
<td>188</td>
</tr>
</tbody>
</table>

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

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

**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).**

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

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