Statistics & Decisions 12, 195 - 201 (1994)

ON MINIMAX ALLOCATION OF STRATIFIED RANDOM SAMPLING WHEN ONLY THE ORDER OF STRATUM VARIANCES IS KNOWN[1]

Manoranjan Pal and Pulakesh Maiti

Received: Revised version: December 21, 1993

Abstract

This paper proposes the *minimax* criteria for obtaining the sample sizes to different strata when only the ranks of the stratum variances, apart from the stratum sizes, are known, and obtains a very simple and elegant solution to this problem.

1 Introduction

In many practical situations in sample survey it may not be possible to know the exact values of the stratum variances or it may even be very difficult to get good estimates of the stratum variances, whereas the order of the stratum variances may easily be found out from other sources. To be specific, suppose we have strata with known sizes $N_1, N_2, \ldots, N_s$ and unknown variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_s^2$. The problem is to minimize $V(\bar{y}_{st})$ with respect to $n_1, n_2, \ldots, n_s$, the respective sample sizes, where

$$(1.1) \qquad \bar{y}_{st} = \sum_1^s W_h \bar{y}_h$$

with $W_h = N_h/N$, $N = \sum N_h$ and $\bar{y}_h = \frac{\sum_i y_{hi}}{n_h}$ for all h, y_{hi} denoting the value of the i-th unit of the sample from the h-th stratum. $\bar{y}_{st}$ is unbiased for the population mean under simple random sampling scheme. Expression for variance of $\bar{y}_{st}$ is (e.g., Cochran (1974))

$$V = \sum \frac{W_h^2 \sigma_h^2}{n_h}.$$

[1]The problem of finding an optimal allocation of sample sizes to different strata under a given ordering/spacing of stratum variances was initially raised by Professor S. P. Mukhopadhyay of University of Calcutta before the audience in the "Seminar on Problems of Large Scale Sample Survey in India: 26 – 27 December, 1990" - conducted by Computer Science Unit of Indian Statistical Institute.

AMS (1980) subject classification: 62 C 20

Keywords and phrases: Stratified random sampling, minimax

Since the stratum variances are not known, minimization of V is not possible. However, if it is possible to know the order of the stratum variances, say,

$$\sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_s^2, \tag{1.2}$$

then one can hope to minimize V with respect to $n_1, n_2, \ldots, n_s$ as well as $\sigma_1^2, \sigma_2^2, \ldots, \sigma_s^2$ subject to the conditions $\sum n_h = n$, $n_h > 0$ for all h together with the condition (1.2). It is necessary to introduce further restriction such as $\sum \sigma_h^2 = K$ to avoid trivial solutions like $\sigma_h^2 = 0$ for all h for the case where the condition (1.2), say, is imposed. The value of K, as we shall see later, does not affect the optimum n_h values for the problem considered in this paper. Thus the value of K need not be known apriori. In Pal and Maiti (1991), the solution for the same problem has been obtained.

A possible criticism of the above approach stems from the fact that we have no control over the σ_h^2 values. The minimization problem described above will give some optimum values of σ_h^2's. But, there is no guarantee that the optimum values will be equal to the actual values. In fact, it is perfectly possible that the optimum values become far different from the actual values. A more reasonable approach to tackle this problem would be to get

$$\underset{\underset{\sim}{n}}{\text{Min}}\ \underset{\underset{\sim}{\sigma^2}}{\text{Max}}\ V, \tag{1.3}$$

where $\underset{\sim}{n}$ and $\underset{\sim}{\sigma^2}$ are the vectors of n_h and σ_h^2 values respectively, subject to the same conditions as described earlier. This procedure, thus, tackles the adverse situations so far as $\underset{\sim}{\sigma^2}$ is concerned. One may also find

$$\underset{\underset{\sim}{\sigma^2}}{\text{Max}}\ \underset{\underset{\sim}{n}}{\text{Min}}\ V \tag{1.4}$$

to see what will be the maximum possible value over $\underset{\sim}{\sigma^2}$ of the minimum variance, since $\underset{\sim}{\sigma^2}$ is not known.

In this paper we present the minimax solution for the problem where the condition (1.2) is imposed. It so happens that the minimax and the maximin solutions give the same optimum values for $\underset{\sim}{n}$ (and $\underset{\sim}{\sigma^2}$) and hence, also for the values of the objective function.

To summarize the above points, our object is to find

$$\underset{\underset{\sim}{n}}{\text{Min}}\ \underset{\underset{\sim}{\sigma^2}}{\text{Max}} \sum_{h=1}^{s} \frac{W_h^2 \sigma_h^2}{n_h} \tag{1.5}$$

subject to

$$0 \leq \sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_s^2 \quad \text{and} \quad \sum_{h=1}^{s} \sigma_h^2 = K, \tag{1.6}$$

$$n_h > 0 \quad \text{for} \quad h = 1, \ldots, s \quad \text{and} \quad \sum_{h=1}^{s} n_h = n. \tag{1.7}$$

Here $W_h = N_h/N,\ N = \sum_{i=1}^{s} N_i,\ N_1, \ldots, N_s,\ n$ and K are given positive constants. Even though the n_h's should be integers, the optimization problem becomes too difficult under this constraint. So we solve the problem allowing n_h's to be nonnegative and approximate the optimal n_h's by integers hoping that it will give a near optimal solution.

2 Solution of the Optimization Problem

We start with a result which can be used to find the maximum in (1.5) for given n satisfying (1.7).

Lemma 2.1 *Let $a_1, a_2, \ldots, a_s$ be positive constants. Then,* $\text{Max}_{\underset{\sim}{\sigma}^2} \sum_{h=1}^{s} a_h \sigma_h^2$ *subject to (1.6) is* $K \max_{1 \le h \le s} b_h$*, where* $b_h = \frac{1}{s-h+1} \sum_{j=h}^{s} a_j$.

Proof: Define

$$Z_h = (s-h+1)(\sigma_h^2 - \sigma_{h-1}^2) \quad \text{for} \quad h = 1, 2, \ldots, s,$$

where we take $\sigma_0^2 = 0$. Then it is easy to check that $\sum_{h=1}^{s} a_h \sigma_h^2 = \sum_{h=1}^{s} b_h Z_h$ and (1.6) is equivalent to

$$Z_h \ge 0 \quad \text{for} \quad h = 1, 2, \ldots, s \quad \text{and} \quad \sum_{h=1}^{s} Z_h = K. \tag{2.1}$$

Let $b_j = \max_{1 \le h \le s} b_h$. It is clear that $\max \sum b_h Z_h$ subject to (2.1) is $K b_j$, which is attained when $Z_j = K$ and $Z_h = 0$ for all $h \ne j$. □

Theorem 2.2 *For given $\underset{\sim}{n}$ satisfying (1.7),* $\text{Max}_{\underset{\sim}{\sigma}^2} \sum_{h=1}^{s} W_h^2 \sigma_h^2 / n_h$ *subject to (1.6) is*

$$\frac{K}{N^2} \times \max_{1 \le h \le s} f_h, \tag{2.2}$$

where

$$f_h = \frac{1}{s-h+1} \sum_{j=h}^{s} \frac{N_j^2}{n_j}. \tag{2.3}$$

This Theorem follows from Lemma 2.1 on taking $a_h = W_h^2/n_h = N_h^2/(N^2 n_h)$.

Thus, to find the minimax solution of (1.5) subject to (1.6) and (1.7), we have to solve the following problem:

$$\underset{\underset{\sim}{n}}{\text{Min}} \max_{1 \le h \le s} f_h(\underset{\sim}{n}) \tag{2.4}$$

subject to (1.7), where $f_h(\underset{\sim}{n})$ is given by (2.3).

Theorem 2.3 *The minimum in (2.4) is attained at some $\underset{\sim}{n}$ satisfying (1.7).*

Proof: If we take $n_h \alpha N_h^2$ (i.e., $n_h = nN_h^2/\sum_j N_j^2$), then $\max f_h(\underset{\sim}{n}) = \sum N_j^2/n = \Delta$ (say). Hence for the minimization in (2.4), all $\underset{\sim}{n}$'s for which $\max f_h(\underset{\sim}{n})$ is larger than Δ can be ignored. Now for any fixed h, there exists $\delta_h > 0$ such that if $n_h < \delta_h$ then $\max_i f_h(\underset{\sim}{n}) > \Delta$ whatever be the other n_j's. Thus the problem (2.4) subject to (1.7) is equivalent to (2.4) subject to

$$n_h \geq \delta_h \quad \text{for} \quad h = 1, 2, \ldots, s \quad \text{and} \quad \sum_{h=1}^{s} n_h = n. \tag{2.5}$$

Since the set of $\underset{\sim}{n}$'s satisfying (2.5) is a compact set in $\mathbb{R}^s$ and $\max_{1 \leq h \leq s} f_h(\underset{\sim}{n})$ is a continuous function on it, it follows that the minimum is attained at some $\underset{\sim}{n}$ satisfying (2.5) and so satisfying (1.7).

Choose and fix an optimal solution $\underset{\sim}{n}^0$ of (2.4) subject to (1.7). We introduce some notations. Let

$$\{m_1, m_2, \ldots, m_k\} = \{i : f_i(\underset{\sim}{n}^0) = \max_h f_h(\underset{\sim}{n}^0)\},$$

where

$$1 \leq m_1 < m_2 < \ldots < m_k \leq s.$$

We shall write $m_{k+1} = s + 1$ for convenience. Also let

$$A_r = \{m_r, m_r + 1, \ldots, m_{r+1} - 1\}$$

and $t_r = |A_r| = m_{r+1} - m_r$ for $r = 1, 2, \ldots, k$. □

Lemma 2.4 $m_1 = 1$.

Proof: Suppose not. Consider $\underset{\sim}{n}^*$ defined by

$$n_i^* = \begin{cases} n_1^0 - k\varepsilon & \text{if } i = 1 \\ n_i^0 + \varepsilon & \text{if } i \in \{m_1, \ldots, m_k\} \\ n_i^0 & \text{otherwise} \end{cases}.$$

Then for sufficiently small $\varepsilon > 0$, it is easy to see that $\underset{\sim}{n}^*$ satisfies (1.7), $f_i(\underset{\sim}{n}^*) < f_i(\underset{\sim}{n}^0)$ for $i = 2, 3, \ldots, s$ and $f_1(\underset{\sim}{n}^*) < f_{m_1}(\underset{\sim}{n}^0)$, a contradiction to the optimality of $\underset{\sim}{n}^0$. □

Lemma 2.5 *Let* $1 \leq i < j \leq s$. *Then*

$$\frac{N_i}{n_i^0} \geq \frac{N_j}{n_j^0}. \tag{2.6}$$

Moreover, if $i, j \in A_r$ *for some* r, *then equality holds in (2.6).*

Proof: Suppose (2.6) does not hold. Let $\underset{\sim}{n}^*$ be defined as

$$n_h^* = \begin{cases} n_i^0 - \varepsilon & \text{if } h = i \\ n_j^0 + \varepsilon & \text{if } h = j \\ n_h^0 & \text{otherwise} \end{cases} .$$

Then for sufficiently small $\varepsilon > 0$, $\underset{\sim}{n}^*$ satisfies (1.7) and

$$(s-h+1)(f_h(\underset{\sim}{n}^*) - f_h(\underset{\sim}{n}^0)) = \begin{cases} \frac{\varepsilon N_i^2}{(n_i^0-\varepsilon)n_i^0} - \frac{\varepsilon N_j^2}{(n_j^0+\varepsilon)n_j^0} & \text{if } h \le i \\ -\frac{\varepsilon N_j^2}{(N_j+\varepsilon)n_j^0} & \text{if } i < h \le j \\ 0 & \text{if } h > j \end{cases} .$$

Thus $f_h(\underset{\sim}{n}^*) \le f_h(\underset{\sim}{n}^0)$ for $h = 1, 2, \ldots, s$, strict inequality holding at least for $h = 1, 2, \ldots, i$. It follows that $\underset{\sim}{n}^*$ is also optimal for (2.4). So by Lemma 2.4, $\max_h f_h(\underset{\sim}{n}^*) = f_1(\underset{\sim}{n}^*)$. Since $f_1(\underset{\sim}{n}^*) < f_1(\underset{\sim}{n}^0) = \max_h f_h(\underset{\sim}{n}^0)$, we have a contradiction to the optimality of $\underset{\sim}{n}^0$. This proves the first statement. If $i, j \in A_r$ and strict inequality holds in (2.6), we arrive at a contradiction in a similar way by taking ε to be negative with sufficiently small absolute value. Here it should be noted that when $i < h \le j$, $f_h(\underset{\sim}{n}^*) \le f_h(\underset{\sim}{n}^0)$ is not true but $f_h(\underset{\sim}{n}^*) < f_1(\underset{\sim}{n}^0)$ holds since $|\varepsilon|$ is sufficiently small. This proves the lemma. □

We now introduce some more notations. We define

$$X_h = \frac{N_h^2}{n_h} \text{ for } h = 1, 2, \ldots, s.$$

Also let

$$\bar{N}(i,j) = \frac{1}{j-i+1} \sum_{h=i}^{j} N_h$$

for any i, j with $1 \le i \le j \le s$ and let

$$\bar{N}_{(r)} = \bar{N}(m_r, m_{r+1} - 1)$$

for $r = 1, 2, \ldots, k$. The quantities $\bar{X}(i,j)$ and $\bar{X}_{(r)}$ are defined analogously.

Lemma 2.6

(2.7) $$\bar{N}_{(1)} \le \bar{N}_{(2)} \le \ldots \le \bar{N}_{(k)}$$

and

(2.8) $$\bar{N}(m_r, i) \ge \bar{N}_{(r)} \quad \text{if } i \in A_r.$$

Proof: Since $f_h(\underset{\sim}{n}^0) = \bar{X}(h,s)$, we have $\bar{X}(m_1, s) = \bar{X}(m_2, s) = \ldots = \bar{X}(m_k, s)$. Hence

(2.9) $$\bar{X}_{(1)} = \bar{X}_{(2)} = \ldots = \bar{X}_{(k)} = \bar{X}(m_r, m_u - 1)$$

for any r and u such that $1 \leq r < u \leq k+1$. Let the common value of N_i/n_i^0 for all $i \in A_r$ be denoted by c_r. Then

$$\bar{X}_{(r)} = \frac{1}{t_r} \sum_{i \in A_r} \frac{N_i^2}{n_i^0} = \frac{c_r}{t_r} \sum_{i \in A_r} N_i = c_r \bar{N}_{(r)}.$$

Since $c_1 \geq c_2 \geq \ldots \geq c_k$, (2.7) follows from (2.9). To prove (2.8), let $i \in A_r$. Then $\bar{X}(i+1,s) \leq \bar{X}(m_r,s)$ gives $\bar{X}(m_r,i) \geq \bar{X}(m_r,s) = \bar{X}_{(r)}$. Since $X_h \alpha N_h$ within A_r, (2.8) follows. □

Theorem 2.7 *(i) $m_1 = 1$ and for any r with $1 \leq r \leq k$, $m_{r+1}-1$ is the smallest i in the range $m_r \leq i \leq s$ at which*

$$\min_{m_r \leq i \leq s} \bar{N}(m_r,i) \tag{2.10}$$

is attained.

(ii) If $i \in A_r$, then

$$n_i^0 = \frac{\bar{N}_{(r)} N_i}{\sum_{u=1}^k t_u \bar{N}_{(u)}^2} \times n. \tag{2.11}$$

Proof: We first show that the minimum in (2.10) is attained at $i = m_{r+1}-1$. By (2.7) and (2.8), we have $\bar{N}(m_u,i) \geq \bar{N}_{(u)} \geq \bar{N}_{(r)}$ whenever $i \in A_u$ with $u \geq r$. So it easily follows that $\bar{N}(m_r,i) \geq \bar{N}_{(r)} = \bar{N}(m_r,m_{r+1}-1)$ for $i = m_r, m_r+1, \ldots, s$. We next show that $m_{r+1}-1$ is the *smallest* i at which the minimum in (2.10) is attained. Suppose not. Let the minimum be attained also at j with $m_r \leq j < m_{r+1}-1$. Then $\bar{N}(m_r,j) = \bar{N}_{(r)}$, so $\bar{N}(j+1,m_{r+1}-1) = \bar{N}_{(r)}$ and $\bar{X}(j+1,m_{r+1}-1) = \bar{X}_{(r)}$. Hence $\bar{X}(j+1,s) = \bar{X}(m_r,s)$, a contradiction, since $j+1$ does not belong to $\{m_1, m_2, \ldots, m_k\}$. This proves (i).

To prove (ii), we first show that

$$\sum_{i \in A_r} n_i^0 = \frac{t_r \bar{N}_{(r)}^2}{\sum_{u=1}^k t_u \bar{N}_{(u)}^2} \times n. \tag{2.12}$$

For this, we have

$$\sum_{i \in A_r} n_i^0 = \sum_{i \in A_r} \frac{N_i}{c_r} = \frac{t_r \bar{N}_{(r)}}{c_r} = \frac{t_r \bar{N}_{(r)}^2}{\bar{X}_{(r)}}.$$

Since $\bar{X}_{(r)}$ is independent of r, (2.12) follows. Now since $n_i^0 \alpha N_i$ within A_r, we have

$$n_i^0 = \frac{N_i}{t_r \bar{N}_{(r)}} \sum_{j \in A_r} n_j^0 = \frac{\bar{N}_{(r)} N_i}{\sum_{u=1}^k t_u \bar{N}_{(u)}^2} \times n.$$

From Theorem 2.7 it follows that the optimal $\underset{\sim}{n}^0$ is unique and can be determined by the following **Procedure**: First find $m_1 = 1, m_2, \ldots, m_k$ using (i). Then find n_i^0 using (ii).

The optimum sample sizes have to be approximated roughly to the nearest integers (n_h^* values, say), though the optimum sample sizes that would have been obtained if we restrict the n_h values to natural numbers may not be same as the n_h^* values. It should be pointed out that a more desirable criterion for optimum allocation would be to impose the set of restrictions $2 \leq n_h(\leq N_h)$ and n_h is an integer for all h. This problem has not yet been solved. □

Acknowledgements. We gratefully acknowledge Professors Bikas K. Sinha and A. R. Rao of Indian Statistical Institute who went through an earlier draft of this paper and gave many constructive suggestios for improvement and reduction of this paper. Besides, Professor Sinha pointed out the superiority of taking minimax criterion over the earlier (Pal and Maiti (1991)) criterion and Professor Rao pointed out the necessity of proving the existence of the solution. In fact, he supplied the proof of Theorem 2.3

References

[1] Cochran, G. W. (1974): Sampling Techniques. Second Wiley Eastern Reprint. John Wiley & Sons, Inc.

[2] Pal, M. and P. Maiti (1991): Optimum Stratified Random Sampling Under a Given Ordering of Stratum Varianced. Tech. Report No. ERU/6/91

Manoranjan Pal and Pulakesh Maiti
Economic Research Unit
Indian Statistical Institute
203 Barrackpore Trunk Road
Calcutta 700 035, India