# Optimum Allocation in Two-stage and Stratified Two-stage Sampling for Multivariate Surveys

M.G.M. Khan, Munish A. Chand, and Nesar Ahmad
School of Computing, Information and Mathematical Sciences
Faculty of Science and Technology
The University of the South Pacific
Suva, Fiji

## Abstract

When more than one characteristics are under study it is not possible for one reason or the other to use the individual optimum allocation of first-stage and second-stage sampling units to each stage and to various strata while using two-stage and stratified two-stage sampling designs. In such situations some criterion is needed to work out an acceptable allocation which is optimum for all characteristics in some sense. In this paper the problems of the optimum allocation in multivariate two-stage and multivariate stratified two-stage sampling are formulated as Nonlinear Programming Problems (NLPP). The NLPPs are then solved using Lagrange multiplier technique and explicit formulas are obtained for the optimum allocation of the first-stage and second-stage sampling units.

**Keywords:** Multivariate two-stage sampling, Multivariate stratified two-stage sampling, First-stage sampling units, Second-stage sampling units, Optimum allocation, Nonlinear programming problem.

## 1. Introduction

In many surveys the use of two-stage sampling designs often specifies two stages of selection: clusters or primary sampling units (PSUs) at the first stage, and subsamples from PSUs at second stage as a secondary sampling units (SSUs). For the large-scale surveys, stratification may precede selection of the sample at any stage. Analyses of two-stage designs are well documented when a single variable is measures and the methods to obtain the optimum allocations of sampling units to each stage are readily available (Cochran (1977), Chapter 10; Arnold (1986); Sadooghi-Alvandi (1986); Valliant and Gentle (1997); Clark and Steel (2000); Dever, *et al.* (2001)). However, when more than one characteristics are under study the procedures for determining optimum allocations are not well defined. The traditional approach is to estimate optimal sample size for each characteristic individually and then choose the final sampling design from among the individual solutions. In practice it is not possible to use this approach of individual optimum allocations because an allocation, which is optimum for one characteristic, may not be optimum for other characteristics. Moreover, in the absence of a strong positive correlation between the characteristics under study the individual optimum allocations may differ a lot and there may be no obvious compromise. In such situations some criterion is needed to work out an acceptable sampling design which is optimum, in some sense, for all characteristics. Waters and Chester (1987) proposed a graphical approach to identify the possible optimum solution for multivariate case.

In this paper a method of optimum allocation for multivariate two-stage sampling designs and multivariate stratified two-stage sampling designs is developed. The problems of determining the optimum allocations are formulated as Nonlinear Programming Problems (NLPP), in which each NLPP has a convex objective function and a single linear cost constraint. Several techniques are available for solving these NLPPs, better known as Convex Programming Problems (CPP). We used Lagrange multiplier technique to solve the formulated NLPPs and explicit formulae for the optimum allocation of PSUs and the optimum size of SSUs or the subsamples to various strata are obtained. The Kuhn-Tucker (1951) necessary conditions, which are also sufficient, for this problem, are verified at the optimum solutions.

## 2. The Formulation of the Problem in Two-stage Sampling

In a multivariate two-stage sampling, where $p$ characteristics are under study, $n$ units as PSU and $m$ subunits as SSU within each of $n$ selected PSU are drawn randomly from $N$ units in first stage and $M$ units in the second stage, respectively. Let $y_{ijk}$ , $\overline{y}_{ik} = \sum_{j=1}^{m} \frac{y_{ijk}}{m}$ and $\overline{\overline{y}}_{k} = \sum_{i=1}^{n} \frac{\overline{y}_{ik}}{n}$ denote the value obtained from $j$ th subunit in the $i$ th primary unit, the sample mean per subunit in the $i$ th primary unit, and the overall sample mean per subunit for $k$ th characteristic, respectively. It could be shown that $\bar{\bar{y}}_k$ is an unbiased estimate of the over all population mean $\bar{\bar{Y}}_k$ of $k$ th characteristic with variance

$$V(\bar{\bar{y}}_k) = \left(\frac{N-n}{N}\right)\frac{S_{1k}^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_{2k}^2}{mn}, \qquad (2.1)$$

where $S_{1k}^2$ is the variance among primary unit means and $S_{2k}^2$ is the variance among subunits within primary units for $k$ th characteristic, respectively.

The total cost function of a two-stage sampling procedure may be given as:

$$C = c_1 n + c_2 nm, \qquad (2.2)$$

where $C$ denotes the total cost of the survey, $c_1$ denotes the cost of approaching to a PSU for measurement and $c_2 = \sum_{k=1}^{p} c_{2k}$ denotes the cost of measurement all the $p$ characteristics per SSU. Also $c_{2k}$ are the per unit costs of measuring the $k$ th characteristic of a SSU.

The optimum choice of $n$ and $m$ for an individual characteristic can thus be determined by minimizing the variance in (2.1) for the given cost in (2.2), or by minimizing the cost for fixed variance.

In multivariate stratified sample surveys usually a compromise criterion is needed to work out an acceptable choice of the number of PSU's and SSU's which is optimum, in some sense, for all characteristics. However, if the total cost for the survey is predetermined, using the compromise criterion suggested by Khan, Khan and Ahsan (2003), an optimal choice may be one that minimizes the weighted sum of the sampling variances of the estimates of various characteristics within the available budget. It is, therefore, in a two-stage sampling, if the population means of $p$ characteristics are of interest, it may be a reasonable criterion for determining the optimal choice of $n$ and $m$ is to minimize a weighted sum of the variances of the two-stage sample means of all the $p$ characteristics, that is,

$$\sum_{k=1}^{p} a_k V(\bar{\bar{y}}_k), \qquad (2.3)$$

where $a_k$ is the weights assigned to the $k$ th characteristic in proportion to its importance as compared to other characteristics and $V(\bar{\bar{y}}_k)$ as given in (2.1). Ignoring the term independent of $n$ and $m$ minimizing (2.3) will be equivalent to minimize

$$\frac{A_1^2}{n} + \frac{A_2^2}{nm} - \frac{A_2^2}{nM}, \qquad (2.4)$$

where $A_1^2 = \sum_{k=1}^{p} a_k S_{1k}^2$ and $A_2^2 = \sum_{k=1}^{p} a_k S_{2k}^2$. (2.5)

For a fixed budget $C_0$ given by (2.2) the problem of finding the optimum values of $n$ and $m$ may be stated as the following NLPP – I:

$$\left.\begin{array}{ll} \text{Minimize} & Z = \dfrac{A_1^2}{n} + \dfrac{A_2^2}{nm} - \dfrac{A_2^2}{nM} \\ \text{subject to} & c_1 n + c_2 nm \le C_0 \\ \text{and} & n, m \ge 0 \end{array}\right\} \qquad (2.6)$$

The restrictions $n \ge 0$ and $m \ge 0$ are obvious because negative values of the number of PSU's and SSU's are of no practical use.

## 3. The Formulation of the Problem in Stratified Two-stage Sampling

The most common design in surveys is stratified two-stage sampling. The population of PSUs is divided into strata, within each stratum a simple random sample without replacement of PSUs is selected and each of the PSUs is further sub-sampled. Let the population of $N$ PSUs be divided into $L$ strata, each with $N_h$ PSUs such that $N = \sum_{h=1}^{L} N_h$. Also let $M_{hi}$ be the number of SSUs in the $i$ th PSU and $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$ be the total number of SSUs in the $h$ th stratum. In a multivariate stratified two-stage sampling, where $p$ characteristics are under study, let $y_{hijk}$ denotes the value of $k$ th characteristic on the

$j$ th SSU of $i$ th PSU of $h$ th stratum. A random sample of $n_h$ PSUs and $m_{hi}$ SSUs from $i$ th PSU are selected in $h$ th stratum. Let

$$\bar{\bar{y}}_{k,st} = \sum_{h=1}^{L} W_h \bar{\bar{y}}_{k,hs} ,$$

denotes the overall sample mean per SSU for $k$ th characteristic in $h$ th stratum, where $\bar{\bar{y}}_{k,hs} = \frac{1}{n_h}\sum_{i=1}^{n_i} \frac{M_{hi}}{\bar{M}_h} \bar{y}_{k,hi}$ , $\bar{y}_{k,hi} = \frac{1}{m_{hi}}\sum_{i=1}^{m_{hi}} y_{hijk}$ , $\bar{M}_h = \sum_{i=1}^{N_h} M_{hi} \big/ N_h$ , and $W_h = M_{h0} \big/ \sum_{h=1}^{L} M_{h0}$ . It could be shown that $\bar{\bar{y}}_{k,st}$ is an unbiased estimate of the over all population mean $\bar{\bar{Y}}_k$ of $k$ th characteristic with variance

$$V\left(\bar{\bar{y}}_{k,st}\right) = \sum_{h=1}^{L} W_h^2 \left[ \left( \frac{1}{n_h} - \frac{1}{N_h} \right) S_{k,hb}^2 + \frac{1}{n_h N_h} \sum_{i=1}^{N_h} \left( \frac{M_{hi}}{\bar{M}_h} \right)^2 \left( \frac{1}{m_{hi}} - \frac{1}{M_{hi}} \right) S_{k,hiy}^2 \right], \tag{3.1}$$

where $S_{k,hb}^2$ is the variance among primary unit means and $S_{k,hiy}^2$ is the variance among subunits within primary units for $k$ th characteristic, respectively.

Assume that the total cost of the survey consists of two components depending upon the numbers of PSUs and the number of SSUs in the sample. Let $c_{1h}$ and $c_{2h} = \sum_{k=1}^{p} c_{2hk}$ denote the cost per PSU and the cost of measurement all the $p$ characteristics per SSU in $h$ th stratum, respectively. Where $c_{2hk}$ are the per unit costs of measuring the $k$ th characteristic of a SSU. Thus the total cost of the survey may be expressed as a function of first and second-stage sample sizes, $n_h$ and $m_{hi}$ , as:

$$c_0 + \sum_{h=1}^{L} \left( c_{1h} n_h + c_{2h} \sum_{i=1}^{n_h} m_{hi} \right),$$

where $c_0$ is the overhead cost of the survey. The second component in $(\ )$ varies from sample to sample. It is, therefore, the expected cost function could be considered as:

$$c_0 + \sum_{h=1}^{L} \left( c_{1h} n_h + c_{2h} \cdot \frac{n_h}{N_h} \sum_{i=1}^{N_h} m_{hi} \right). \tag{3.2}$$

If the total amount available for a multivariate stratified two-stage survey is predetermined, a compromise allocation of $n_h$ and $m_{hi}$ may be one discussed in section 2 that minimizes the weighted sum of the sampling variances of the estimates of various characteristics, that is.

$$\sum_{k=1}^{p} a_k V\left(\bar{\bar{y}}_{k,st}\right), \tag{3.3}$$

where $a_k$ is the weights assigned to the $k$ th characteristic in proportion to its importance as compared to other characteristics and $V\left(\bar{\bar{y}}_{k,st}\right)$ as given in (3.1). For the purpose of minimization, the term independent of $n_h$ and $m_{hi}$ in (3.3) is ignored. Also letting

$$\begin{aligned} A_h &= \sum_{k=1}^{p} a_k \left( S_{k,hb}^2 - \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{\bar{M}_h^2} \cdot S_{k,hiy}^2 \right) \\ \text{and} \quad B_{hiy}^2 &= \sum_{k=1}^{p} a_k S_{k,hiy}^2 \end{aligned} \tag{3.4}$$

the problem of finding the compromise allocation of $n_h$ and $m_{hi}$ for a fixed cost $C_0$ may be given as the following NLPP – II:

$$\left. \begin{aligned} &\text{Min } Z = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left( A_h + \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{M_{hi}^2}{\bar{M}_h^2} \cdot \frac{B_{hiy}^2}{m_{hi}} \right) \\ &\text{s.t.} \quad \sum_{h=1}^{L} \left( c_{1h} n_h + c_{2h} \cdot \frac{n_h}{N_h} \sum_{i=1}^{N_h} m_{hi} \right) \le C_0 \\ &\text{and} \quad n_h, m_{hi} \ge 0 \\ &\qquad (i = 1,2,...,N_h;\ h = 1,2,...,L) \end{aligned} \right\}, \tag{3.5}$$

where $C_0 = C - c_0$ .

## 4. The Solution

### 4.1 NLPP – I

The objective function $Z$ of the NLPP – I given in (2.6) will be minimum when the values of $n$ and $m$ are as large as permitted by the cost constraint. This suggests that at the optimum point the cost constraint will be active, that is, it is satisfied as an equation. Then, ignoring the restrictions $n \geq 0$ and $m \geq 0$, we can use Lagrange multipliers technique to determine the optimum values of $n^*$ and $m^*$. If these values $n^*$ and $m^*$, satisfy the ignored restrictions, the NLPP (2.6) is solved completely.

The Lagrangian function $\varphi$ is defined as

$$\varphi(n,m,\lambda) = \frac{A_1^2}{n} + \frac{A_2^2}{nm} - \frac{A_2^2}{nM} + \lambda\left(c_1 n + c_2 nm - C_0\right) \ , \tag{4.1}$$

where $\lambda$ is a Lagrange multiplier.

The necessary conditions for the solution of the problem are

$$\frac{\delta\varphi}{\delta n} = -\frac{A_1^2}{n^2} - \frac{A_2^2}{n^2 m} + \frac{A_2^2}{n^2 M} + \lambda\left(c_1 + c_2 m\right) = 0 , \tag{4.2}$$

$$\frac{\delta\varphi}{\delta m} = -\frac{A_2^2}{nm^2} + \lambda c_2 n = 0 , \tag{4.3}$$

and

$$\frac{\delta\varphi}{\delta\lambda} = c_1 n + c_2 nm - C_0 = 0 . \tag{4.4}$$

(4.2) and (4.3) give

$$m^* = \sqrt{\frac{c_1 A_2^2}{c_2\left(A_1^2 - \frac{A_2^2}{M}\right)}} \text{ , provided } A_1^2 > \frac{A_2^2}{M} \tag{4.5}$$

(4.4) and (4.5) give

$$n^* = \frac{C_0}{c_1 + c_2 m^*} \tag{4.6}$$

It can be verified that the objective function $Z$ in (2.6) is convex for $A_1^2 > \frac{A_2^2}{M}$ or $\sum_{k=1}^{p} a_k S_{1k}^2 > \sum_{k=1}^{p} a_k S_{2k}^2 \Big/ M$ and the constraint is linear. Therefore, the (K-T) necessary conditions for the NLPP (2.6) are sufficient also. These conditions are

$$\nabla_{(n,m)}\varphi = \begin{pmatrix} -\frac{A_1^2}{n^2} - \frac{A_2^2}{n^2 m} + \frac{A_2^2}{n^2 M} + \lambda\left(c_1 + c_2 m\right) \\ -\frac{A_2^2}{nm^2} + \lambda c_2 n \end{pmatrix} \geq 0 ,$$

$$n\left(-\frac{A_1^2}{n^2} - \frac{A_2^2}{n^2 m} + \frac{A_2^2}{n^2 M} + \lambda\left(c_1 + c_2 m\right)\right) + m\left(-\frac{A_2^2}{nm^2} + \lambda c_2 n\right) = 0,$$

$$\nabla_\lambda \varphi = c_1 n + c_2 nm - C_0 \leq 0 ,$$

$$\lambda\left(c_1 n + c_2 nm - C_0\right) = 0 ,$$

and $\qquad n, m \text{ and } \lambda \geq 0$.

For the case $n, m$ and $\lambda > 0$ the above expressions give the same set of equations as (4.2), (4.3) and (4.4), which implies that the K-T conditions hold at the point $\left(n^*, m^*\right)$ given by (4.5) and (4.6). Hence, $\left(n^*, m^*\right)$ is optimum for NLPP (2.6).

If $A_1^2 \leq \frac{A_2^2}{M}$, one may use a single-stage sampling design instead of two-stage sampling by considering $m^* = M$.

### 4.2 NLPP – II

The objective function $Z$ of the NLPP – II given in (3.5) will be minimum when the values of $n_h$ and $m_{hi}$ are as large as permitted by the cost constraint. Therefore, this problem also suggests that at the optimum point the cost constraint will be active and one can use Lagrange multipliers technique to determine the optimum values of $n_h^*$ and $m_{hi}^*$ considering the cost constraint as an equation and ignoring the non-negativity restrictions on the variables.

The Lagrangian function $\varphi$ is defined as

$$\varphi\left(n_h, m_{hi}, \lambda\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(A_h + \frac{1}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}^2}{\bar{M}_h^2}\cdot\frac{B_{hiy}^2}{m_{hi}}\right) + \lambda\left(\sum_{h=1}^{L}\left(c_{1h}n_h + c_{2h}\cdot\frac{n_h}{N_h}\sum_{i=1}^{N_h} m_{hi}\right) - C_0\right) \tag{4.7}$$

where $\lambda$ is a Lagrange multiplier.

The necessary conditions for the solution of the problem are

$$\frac{\delta\varphi}{\delta n_h} = -\frac{W_h^2}{n_h^2}\left(A_h + \frac{1}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}^2}{\bar{M}_h^2}\cdot\frac{B_{hiy}^2}{m_{hi}}\right) + \lambda\left(c_{1h} + c_{2h}\cdot\frac{1}{N_h}\sum_{i=1}^{N_h} m_{hi}\right) = 0 , \tag{4.8}$$

$$\frac{\delta\varphi}{\delta m_{hi}} = -\frac{W_h^2}{n_h}\frac{1}{N_h}\frac{M_{hi}^2}{\bar{M}_h^2}\frac{B_{hiy}^2}{m_{hi}^2} + \lambda c_{2h}\frac{n_h}{N_h} = 0, \tag{4.9}$$

and

$$\frac{\delta\varphi}{\delta\lambda} = \sum_{h=1}^{L}\left(c_{1h}n_h + c_{2h}\frac{n_h}{N_h}\sum_{i=1}^{N_h} m_{hi}\right) - C_0 = 0 ., \tag{4.10}$$

Multiplying by $\dfrac{m_{hi}}{n_h}$ and summing over $i$ $(i = 1, 2, ..., N_h)$, (4.9) reduces to

$$-\frac{W_h^2}{n_h N_h}\sum_{i=1}^{N_h}\frac{M_{hi}^2}{\bar{M}_h^2}\frac{B_{hiy}^2}{m_{hi}} + \lambda\frac{c_{2h}}{N_h}\sum_{i=1}^{N_h} m_{hi} = 0 . \tag{4.11}$$

(4.8) and (4.11) give

$$n_h = \frac{1}{\sqrt{\lambda}}\frac{W_h\sqrt{A_h}}{\sqrt{c_{1h}}}, \text{ provided } A_h > 0 \tag{4.12}$$

Substituting the values of $n_h$ from (4.12) in (4.9), the optimum values of $m_{hi}$ are obtained as:

$$m_{hi}^* = \frac{M_{hi}B_{hiy}}{\bar{M}_h}\cdot\sqrt{\frac{c_{1h}}{A_h c_{2h}}} \tag{4.13}$$

for $i = 1, 2, ..., N_h, h = 1, 2, ..., L$.

Substituting the values of $n_h$ and $m_{hi}$ from (4.12) and (4.13) respectively, (4.10) gives

$$\frac{1}{\sqrt{\lambda}} = \frac{C_0}{\sum_{h=1}^{L}\left(W_h\sqrt{A_h c_{1h}} + \frac{W_h\sqrt{c_{2h}}}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{\bar{M}_h}B_{hiy}\right)}. \tag{4.14}$$

From (4.12) and (4.14) the optimum values of $n_h$ are obtained as:

$$n_h^* = \frac{C_0 W_h\sqrt{A_h}\Big/\sqrt{c_{1h}}}{\sum_{h=1}^{L}\left(W_h\sqrt{A_h c_{1h}} + \frac{W_h\sqrt{c_{2h}}}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{\bar{M}_h}B_{hiy}\right)}. \tag{4.15}$$

As the objective function of (3.5) is convex for $A_h = \sum_{k=1}^{p} a_k\left(S_{k,hb}^2 - \frac{1}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{\bar{M}_h^2}\cdot S_{k,hiy}^2\right) > 0$ and the constraint is linear, the (K-T) necessary conditions of the NLPP (3.5) are sufficient also. It can be easily verified that the K-T conditions hold at the point $\left(n_h^*, m_{hi}^*\right)$ given by (4.13) and (4.15). Hence, $\left(n_h^*, m_{hi}^*\right)$ is optimum for NLPP (3.5).

## References

Arnold, B. F. (1986). "Procedures to Determine Optimum Two-stage Sampling Plans by Attributes," *Metrika*, 33, 93-109.

Clark, Robert G. and Steel, D. G. (2000). "Optimum Allocation of Sample to Strata and Stages with Simple Additional Constraints," *Journal of the*

*Royal Statistical Society, Series D: The Statistician*, 49, 197-207.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons, Inc.

Dever, Jill A., Liu, Jun, Iannacchione, Vincent G. and Kendrick, Douglas E. (2001). "An Optimal Allocation Method for Two-stage Sampling Designs with Stratification at the Second Stage," *ASA Proceedings of the Joint Statistical Meetings*, *American Statistical Association (Alexandria, VA)*.

Khan, M. G. M., Khan, E. A. and Ahsan, M. J. (2003). "An Optimal Multivariate Stratified Sampling Design using Dynamic Programming," *Australian & New Zealand J. Statist*. 45 (1), 107 – 113.

Kuhn, H. W. and Tucker, A. W. (1951). "Nonlinear Programming," *Proceedings of the Second Berkley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkley, 481 – 492.

Sadooghi-Alvandi, Mohammad (1986). "The Choice of Subsample Size in Two-stage Sampling," *Journal of the American Statistical Association*, 81, 555-558.

Valliant, Richard and Gentle, James E. (1997). "An Application of Mathematical Programming to Sample Allocation," *Computational Statistics & Data Analysis*, 25, 337-360.

Waters, James R. and Chester, Alexander J. (1987). "Optimal Allocation in Multivariate, Two-stage Sampling Designs," *The American Statistician*, 41, 46-50