Some Multi Convex Programming Problems Arising in Multivariate Sampling

Manoj Kr. Sharma,
Mewar University
Chitogarh (Rajasthan India.)

M. V. Ismail, PhD.
Jamia Hamdard,
New Delhi, India.

ABSTRACT

The problems of multivariate sampling arising in the areas of stratified random sampling, two stage sampling, double sampling and response errors formulate as multiobjective convex programming problems with convex objective functions and a single linear constraint with some upper and lower bounds.

Keywords*:* Multivariate Sampling, Stratified sampling, Multiobjective convex programming, Optimization, Linear Programming.

1. Introduction

In multivariate surveys there are more than one population characteristics to be estimated and usually these characteristics are of conflicting nature. The derivation of the optimal sample numbers among various strata or various stages thus requires some special treatment. In multivariate surveys there are more than one population characteristics to be estimated and usually these characteristics are of conflicting nature. The derivation of the optimal sample numbers among various strata or various stages thus requires some special treatment.

Charles D. Day [10] suggests an alternative; a multi-objective evolutionary algorithm is used to generate multiple solutions that, taken together, describe the Pareto front for the problem. The Pareto front consists of a set of non-dominated solutions, in place of specifying a function of the objectives a priori, or choosing a set of arbitrary variance targets.

When optimize the allocation in stratification with several variables, allocation optimum for one character is not optimal for all other. This type of problems formulated as non linear multi-objective programming problem and derived an analytical solution to the problem by Kokan and Khan [2]. Charnayak and Chornous [2000] suggested new criteria and explored further the already existing criteria. M.G.M. Khan *et al.*, [9] formulated this type of problems as Nonlinear Programming Problems (NLPP) and solved by Lagrange multiplier technique. Convex programming is used to minimize the cost of survey while the sampling errors of the estimates do not exceed certain reassigned upper bounds. This approach is possible for small strata and for large strata convex programming is impractical. Hartley [1], Chatterjee, S. [3], and Bethel, J.W. [5], M.A. Rahim, [5] suggest a simple weighted Euclidean distance function is proposed as a measure of joint sampling error of all the estimates.

The problem of multiple objectives is to minimize, cost as well as the estimate the variances. This problem is solved by converting the multiple objective functions into a single scalar-valued objective in one of two methods. First method is creating a function, such as a linear combination, from the multiple objective function values and minimizing this function. A second method involves choosing one objective (usually cost minimization) and turning the rest of the objectives into constraints by setting maximum acceptable variances for each estimate of interest.

2. Multivariate Stratified Sampling

Consider a multivariate population portioned into *L* strata. Suppose that *p* characteristics are measured on each unit of the population and the strata boundaries are fixed in advance. Let n_i be the number of units drawn from ith stratum (i = 1,2,3,…….,*L*).For jth character, an unbiased estimate of the population mean $\overline{Y}_j$ (j = 1,2,3,…….p) , denoted by $\overline{y}_{ist}$, has its sampling variance

$$V(\overline{y}_{ist}) = \sum_{i=1}^{L}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)W_i^2 S_{ij}^2$$

j = 1, 2, 3…p

Where

$$W_i = \frac{N_i}{N}, S_{ij}^2 = \frac{1}{N_i - 1}\sum_{h=1}^{N_i}(y_{ijh} - \overline{Y}_{ij})^2 .$$

Substituting $a_{ij} = W_i^2 S_{ij}^2,$

it gives

$$V(\overline{y}_{ist}) = \sum_{i=1}^{L}\frac{a_{ij}}{n_i} - \sum_{i=1}^{L}\frac{a_{ij}}{N_i},$$

j = 1, 2, 3…p (1)

Let C_{ij} be the cost of enumerating the j^{th} character in the i^{th} stratum and let *C* be the upper limit on the total cost of the survey. Then assuming linear cost function one should have

$$\sum_{i=1}^{L}\sum_{j=1}^{p} C_{ij} n_i \le C,$$

or $\sum_{i=1}^{L} C_i n_i \le C,$ (2)

where $C_i = \sum_{j=1}^{p} C_{ij}$,the cost of enumeration of all the *p* characters in the i^{th} stratum.

Further one should have

$$1 \le n_i \le N_{i,} \qquad i = 1,2,3,\ldots\ldots L \quad (3)$$

It determines the optimum values of n_i , by minimizing (in some sense) all the *p* variances (1) for a fixed budget (2) i.e. it shows

$$\textit{Minimize } V_j = \sum_{i=1}^{L} \frac{a_{ij}}{N_i},$$

$$j = 1, 2, 3 \ldots p$$

$$\textit{Subject to} \quad \sum_{i=1}^{L} C_i n_i \le C \quad (4)$$

$$\textit{and} \quad 1 \le n_i \le N_{i,}$$

$$i = 1,2,3,\ldots,L$$

Since N_i*'s* are given, it is enough to minimize

$$V_j = \sum_{i=1}^{L} \frac{a_{ij}}{N_i},$$

$$j = 1, 2, 3 \ldots p$$

Using X_i *for* n_i, the problem (4) can be written as the following multiobjective non-linear programming problem:

$$\textit{Minimize } V_j = \sum_{i=1}^{L} \frac{a_{ij}}{X_i}, \qquad j = 1, 2, 3 \ldots p \quad \text{(a)} \quad (5)$$

$$\textit{Subject to} \sum_{i=1}^{L} C_i X_i \le C \quad \text{(b)}$$

$$\textit{and} \quad 1 \le X_i \le N_{i,} \quad \text{(c)}$$

$$i = 1,2,3,\ldots,L$$

The objective functions in (5) are convex Kokan and Khan (1967) [2], the single constraint is linear and the bounds are also linear. The problem (5) is, therefore a multiobjective convex programming problem.

If some tolerance limits, say v_j , are given on variances of the *p* characters then the allocation problems reduces to the single objective convex programming problem

$$\textit{Minimize } \sum_{i=1}^{L} C_i X_i$$

$$\textit{Subject to} \quad V_j = \sum_{i=1}^{L} \frac{a_{ij}}{X_i} \le v_j, \quad (6) \qquad j = 1, 2, 3 \ldots p$$

$$1 \le X_i \le N_{i,}$$

$$i = 1, 2, 3 \ldots L$$

3. Two-Stage Sampling

Let us consider a population which consists of *N* Primary Stage Units (PSU's) and the i^{th} PSU consists of M_i Secondary Stage Units (SSU's). A sample of n PSU's is to be selected and from the i^{th} selected PSU, a sample of m_i SSU's is to be selected.

Let us denote

y_{irj} = value obtained for the r^{th} SSU in the i^{th} PSU for the j^{th} character

M_i = number of SSU's in the i^{th} PSU, (i = 1, 2,…, *N*)

$M_o = \sum_{i=1}^{N} M_i$ = total number of SSU's in the population

$\overline{M} = \frac{M_o}{N}$ = average number of SSU's.

$m_o = \sum_{i=1}^{N} m_i =$ total number of SSU's in the sample j^{th} character

$\overline{Y}_{ij} = \sum_{r=1}^{M_i} \frac{y_{irj}}{M_i} =$ the i^{th} PSU population mean for

$\overline{Y}_{Nj} = \sum_{i=1}^{N} \frac{\overline{Y}_{ij}}{N}$ = the overall population mean of PSU means for j^{th} character

$\overline{Y}_j = \frac{\sum_{i=1}^{N} M_i \overline{Y}_{ij}}{M_o} = \sum_{i=1}^{N} W_i \overline{Y}_{ij}$ = population mean per SSU for j^{th} character

$$\bar{y}_{ij} = \sum_{r=1}^{m_i} \frac{y_{irj}}{m_i} = \text{sample mean per}$$

SSU for j^{th} character

$$\bar{y}_j = \frac{\sum_{i=1}^{n} M_i \bar{y}_{ij}}{n\bar{M}} = \text{sample mean}$$

per SSU in the i^{th} PSU for the j^{th} character

Define

$$S^2_{bj} = \frac{\sum_{i=1}^{N} \left(u_i \bar{Y}_{ij} - Y_j\right)^2}{N-1} = \text{population}$$

variance between PSU's mean for j^{th} character

$$S^2_{wj} = \frac{\sum_{r=1}^{M_i} \left(y_{irj} - \bar{Y}_{ij}\right)^2}{\left(M_i - 1\right)} = \text{population}$$

variance between PSU's for j^{th} character

Where

$$u_i = \frac{M_i}{\bar{M}}$$

For j^{th} character (*j = 1, 2, 3… p*), the unbiased estimate of the population mean $\bar{Y}_j$ $\bar{y}_j$ which has the sampling variance as

$$V\left(\bar{y}_j\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2_{bj} + \sum_{i=1}^{N} \frac{M_i^2}{nN\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S^2_{wj}$$

$$= \frac{1}{n} S^2_{bj} + \sum_{i=1}^{N} \frac{M_i^2}{nN\bar{M}^2} \frac{S^2_{wj}}{m_i} +$$

constant terms

$$= \frac{a_{0j}}{n} + \sum_{i=1}^{N} \frac{a_{ij}}{nm_{ij}} + \quad \text{constant}$$

terms (3.1)

Where

$$a_{0j} = S^2_{bj}, a_{ij} = \frac{M_i^2}{N\bar{M}^2} S^2_{wj}$$

Let *C* be the upper limit on total cost of the survey. Assuming the cost of the survey to be linear, it should have

$$nC_0 + \frac{nC_1}{N} \sum_{i=1}^{N} m_i \le C \quad (3.2)$$

Where C_0 the average is cost of selection per PSU and C_1 is the average cost of sampling per SSU. In practice, C_0 is likely to be larger than C_1.

Now the problem is to determine the optimum values of n and m_i so as to minimize the variances (3.1) of the various characters for a fixed budget C .Ignoring the constant terms in (3.1), and using X_0 for *n* & X_i for nm_i, this give the following multiobjective convex programming problem

$$\textit{Minimize } V_j = \sum_{i=0}^{L} \frac{a_{ij}}{X_i}, \qquad j = 1, 2, 3 \ldots p$$

$$\textit{Subject to } \sum_{i=0}^{N} C_i X_i \le C \quad (3.3)$$

and $X_0 \le N, X_i \le NM_i,$ $i = 1,2,\ldots, N$

where

$$C_i = \frac{C_1}{N}$$

for $i = 1, 2,\ldots, N$

3.1 Case of Equal Primary-Stage Units

The equal Primary-Stage Units problem can be considered as particular case of the unequal Primary-Stage Units problems where $M_i = M$ for $i = 1, 2,\ldots,N$.

Let $X_1 = n$ *and* $X_2 = nm$ then the problem in case of equal primary-stage units reduces to the following multiobjective convex programming problem in only two variables:

$$\textit{Minimize } V_j = \sum_{i=1}^{2} \frac{a_{ij}}{X_i}, \qquad j = 1, 2, 3 \ldots p$$

$$\textit{Subject to } \sum_{i=1}^{2} C_i X_i \le C \quad (3.4)$$

and $X_1 \le N, X_2 \le NM,$ $i = 1, 2,\ldots, N$

4. Double Sampling

Consider the problem of double sampling for stratification in which the population is to be stratified in to *L* strata. The first sample of size n' is selected by simple random sampling without replacement to estimate the strata weights. A second sample of *n* units with n_i

units of the i^{th} stratum is selected in which *p* characters $y_1, y_2,\ldots,y_p$ are observed. Neyman allocation uses in allocating the sample size *n* to different strata,

$W_i = \frac{N_i}{N}$ be the proportion of population units falling in the i^{th} stratum $w_i = \frac{n'_i}{n'}$ be the proportion of first sample units falling in the i^{th} stratum. W_i Being unknown is estimated by w_i .

Let $\bar{y}_{ij}$ be the sample mean of the j^{th} character in the i^{th} stratum, *i = 1, 2, L; j = 1, 2,..., p* and $\bar{Y}_{ij}$ be the population mean of the j^{th} character in the i^{th} stratum. For j^{th} character (*j = 1, 2, p*), an unbiased estimate of the population mean $\bar{Y}_j$, is $\bar{y}_j = \sum_{i=1}^{L} W_i \bar{y}_{ij}$, which, for large populations, has the sampling variance

$$V(\bar{y}_j) = \sum_{i=1}^{L}\left[W_i^2 + \frac{W_i(1-W_i)}{n'}\right]\frac{S_{ij}^2}{n_i} + \sum_{i=1}^{L}\frac{W_i(1-W_i)^2}{n'},$$

Where

$$S_{ij}^2 = \sum_{i==1}^{L}\frac{\left(y_{ijr} - \bar{Y}_{ij}\right)^2}{\left(N_i - 1\right)},\ i = 1, 2,$$

L; j = 1, 2, p.

For the proportional allocation $n_i = nW_i$, the variance $\bar{y}_j$ is approximately given by

$$V_j = \frac{1}{n}\sum_{i=1}^{L} W_i S_{ij}^2 + \frac{1}{n'}\sum_{i=1}^{L} W_i\left(\bar{Y}_{ij} - \bar{Y}\right)^2,\quad j = 1, 2, p. \qquad (4.1)$$

An approximate expression of minimum variance under Neyman allocation for j^{th} character is

$$V_j = \frac{v_{1j}}{n'} + \frac{v_{2j}}{n},$$

Where

$$v_{1j} = \sum_{i=1}^{L} W_i\left(\bar{Y}_{ij} - \bar{Y}\right)^2 \text{ and } v_{2j} = \sum_{i=1}^{L} W_i S_{ij}^2,$$

i = 1, 2

Let *C* be the upper limit on total cost of the surveys. Assuming the cost of the survey to linear, it should have

$$C_1 n' + C_2 n \le C \qquad (4.2)$$

where C_1 is the cost per unit of measuring the auxiliary variate and C_2 is the cost per unit of measuring all the study variates. C_1 Is generally smaller than C_2 .

Here it is required to find the values of n' and n so that the total cost does not exceed the given budget and at the same time, the variances for various characters are minimized.

The problem then again reduces to the following multiobjective convex programming problem in two variables:

$$\textit{Minimize } V_j = \sum_{i=1}^{2}\frac{v_{ij}}{X_i},$$

j = 1, 2,..., p.

$$\textit{Subject to } \sum_{i=1}^{2} C_i X_i \le C \qquad (4.3)$$

$$\textit{and} \qquad 1 \le X_i \le N_{i,} \qquad i = 1, 2$$

Where $n' = X_1$ and $n = X_2$.

If the upper tolerance limits v_j ,(*j = 1, 2, p*) are given on the variances of the various characters and it is required to minimize the cost of the surveys, then it gives the following single objective problem

$$\textit{Minimize } \sum_{i=1}^{2} C_i X_i$$

$$\textit{Subject to } \sum_{i=1}^{2}\frac{v_{ij}}{X_i} \le v_j, \quad (4.4)$$

j = 1, 2……, p

$1 \le X_i \le N_{i,}$

i = 1, 2

5. Response Errors

Let an individual be selected at random from the population of *N* individuals and interviewer be picked up at random out of *M* interviewers and assigned to the selected individuals. Denote by y_{abc} the response value obtained for c^{th} sample individual by b^{th} sample interviewer in the a^{th} (population) group. The expected value of y_{abc} will be $\bar{Y}$ .The sample mean is

$$\bar{y} = \sum_{a=1}^{L}\sum_{b}^{k_a}\sum_{c}^{na} y_{abc}$$

In many surveys, interviewers are available to interview only certain classes of the population and only in certain geographical areas. Therefore, conceive of our interviewers as divided into *L* groups with M_a interviewers in the a^{th} group who are available to interview a particular N_a individuals and no others.

When all the interviewers are available to interview all individuals here $L=1; M_a = M; N_a = N.$

Now *n* of *N* individuals in the population are selected at random and m_a interviewers are selected at random from the a^{th} interviewer group to interview those sample individuals who are available for interview by this interviewer group. Let $m = \sum_{a}^{L} m_a$ be the total number of interviewers selected. Hensen & Hurwitz (1951) derive the total variance of individual responses around the mean of all individual responses in the population as

$$V(\bar{y}) = \frac{\left(\sigma_y^2 - \sigma_{yI}\right)}{n} + \frac{\sigma_{yI}}{m}$$

Suppose a population of *M* interviewers is available to enumerate a population of *N* individuals on each of which *p* characters are defined. For j^{th} character (*j = 1, 2,..., p*), the total variance of the sample mean is given by $\bar{y}_j$ is given by

$$V_j = \frac{\left(\sigma_{yj}^2 - \sigma_{yjI}\right)}{n} + \frac{\sigma_{yjI}}{m} \quad (5.1)$$

where σ_{yjI} is the covariance between responses obtained from different individuals by the same interviewer for j^{th} character (this covariance being taken within interviewer groups, since independent selections of interviewers are made from each interviewer group) and σ_{yj}^2 are the variances of over all responses for all individuals to all interviewers for the j^{th} character.

With the ordinary survey which has a fixed total budget, increasing the number of interviewers will increase cost and will require a reduction of expenditure at some other point, e.g., reducing the expenditure per interviewer or per individual or reducing the number of individuals include in the sample.

Let *C* be the total budget available for field work on the survey. Assuming the cost of the survey to be linear, it should have

$$C_1 n + C_2 m \le C \quad (5.2)$$

Where C_1 is the cost of per individual in the sample and C_2 is the cost of per interviewer.

The problem is to determine the values of *n* and *m* which can be found by minimizing the variances (5.1) for a fixed cost (5.2).

The problem of finding the optimal number of interviewers who should be assigned the job and the optimal number of individuals to be selected is finally formulated as

$$\text{Minimize } V_j = \frac{v_{1j}}{n} + \frac{v_{2j}}{m},$$

$j = 1, 2, \ldots, p$

$$\text{Subject to } C_1 n + C_2 m \le C \quad (5.3)$$

$$\text{and } n \le N, m \le M$$

where

$$v_{1j} = \left(\sigma_{yj}^2 - \sigma_{yjI}\right),\ v_{2j} = \sigma_{yjI}.$$

Using X_1 for n and X_2 for *m*, the problem (5.3) reduces again to the following form of multiobjective convex programming problem:

$$\text{Minimize } V_j = \sum_{i=1}^{2} \frac{v_{ij}}{X_i}, \quad j = 1, 2, p$$

$$\text{Subject to } \sum_{i=1}^{2} C_i X_i \le C \quad (5.4)$$

$$\text{and } X_1 \le N, X_2 \le M .$$

In case of minimizing the cost of the survey while the tolerance limits are given on the variances for the various characters, the problem takes the form similar to (4.4)

6. Conclusion:

These multivariate problems can be converted to multi objective Convex programming problems and then these problems can be solved by Goal programming method.

7. References:

[1] Hartley,H.O.(1965).Multiple purpose optimum allocation in stratified sampling. Proceeding of the Social Statistics Section, American Statistical Association 258-261.

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

[3] Chatterjee, S. (1972). A Study of Optimum Allocation in Multivariate Stratified Surveys. Skandinavisk Actuarietidskrift.55, 73-80.

[4] Bethel, J.W. (1989).Sample Allocation in Multivariate Surveys. Survey Methodology, 15(1), 46-57.

[5] Rahim M.A.,Use of distance function to optimize sample allocation in multivariate surveys: a new perspective, on www.amstat.org/selections/srms/Proceedings/papers/1995-061

[6] Charnayak.O.I.& Chornous G.(2000) . Optimal allocation in Stratified sampling with a non-linear cost function. Theory of Stochastic Processes. 6(22), 3−4**,** 6−17.

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

[8] Optimum Allocation in Two-stage and Stratified Two-stage Sampling for Multivariate Surveys by 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, 2006.

[9] Khan. M. G. M., Khan, E.A. & Ahsan, M. J.(2008). Optimum allocation in multivariate stratified sampling in presence of Non-response J. Ind. Soc. Aggril. Statist. 62(1), 42-48.

[10] Charles Day D. (2010) "A Multi-Objective Evolutionary Algorithm for Multivariate Optimal Allocation", Section Survey Research Method – JSM, 3351-3358.