Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal. (2011), Vol. 4, Issue 1, 71 – 82
e-ISSN 2070-5948, DOI 10.1285/i20705948v4n1p71
© 2008 Università del Salento – http://siba-ese.unile.it/index.php/ejasa/index

GEOMETRIC PROGRAMMING APPROACH TO OPTIMUM ALLOCATION IN MULTIVARIATE TWO- STAGE SAMPLING DESIGN

Showkat Maqbool [*], Abdul H. Mir, Shakeel A. Mir

Division of Agricultural Statistics, SK University of Agricultural Sciences & Technology of Kashmir, India.

Received 22 April 2010; Accepted 24 November 2010
Available online 26 April 2011

***Abstract**: Geometric programming provides a powerful tool for solving non-linear problems where non-linear relations can be well presented by an exponential or power function. In real life situations applications of geometric programming are sound in engineering design, sampling design etc. In this paper, the problem of allocation in first stage and second stage units in multivariate two stage sampling is considered. The problem is formulated as a convex programming problem with linear objective function. A solution procedure is developed to solve the resulting mathematical programming problem by using geometric programming technique. The computational details of the procedure are illustrated through a numerical example.*

***Keywords**: Two Stage sampling, Non-linear programming, convex programming, geometric programming.*

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. Analysis of two-stage designs are well documented when a single variable is measured and the methods to obtain the optimum allocations of sampling units to each stage are readily available The problem of optimum allocation in two-stage sampling with a single character is described in standard texts on sampling (see Cochran

[*] E-mail: smaqbool2007@yahoo.co.in

[1]). However when more than one characteristic 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 characteristic. Moreover, in the absence of strong positive correlation between the characteristics under study the individual optimum allocations may differ a lot and there may be no obvious compromise. In certain situations some criterion is needed to work out an acceptable sampling design which is optimum in some sense for all the characteristics. Geometric programming (GP), a systematic method for solving the class of mathematical programming problems that tend to appear mainly in engineering design, was first developed by Duffin and Zener in the early 1960s, and further extended by Duffin *et al.* [4]. Davis and Rudolph [3] use geometric programming to optimal allocation of integrated samples in quality control. Shiang [6] and Shaojian et.al [7] used G.P for engineering design problems. The paper is presented as follows: First an allocation problem is formulated in a two-stage sampling design in section 2 and geometric programming approach is used to solve it section 3. A numerical illustration is then presented in section 4 and the final comments and conclusion is given in section 5.

2. Formulation of the Problem

Let us assume that the population consists of NM elements grouped into N first-stage units of M second-stage units each. Let n and m be the corresponding sample sizes selected with equal probability and without replacement at each stage. Let y_{hrj} be the value of the population at r^{th} secondary stage unit in the h^{th} primary stage unit for j^{th} character, $\left(h=1,2,...,N,\quad r=1,2,...,M,\quad j=1,...,p\right)$.

We define for j^{th} character:

- $\bar{y}_{hj}=\sum_{r=1}^{m}\frac{y_{hrj}}{m}$ = Sample mean per sub unit at the h^{th} primary stage unit.
- $\bar{\bar{y}}_{j}=\sum_{h=1}^{n}\frac{\bar{y}_{hj}}{n}$ = Overall sample mean per sub unit (element).
- $\bar{Y}_{hj}=\sum_{r=1}^{M}\frac{y_{hrj}}{M}$ = Mean per element at the h^{th} first stage unit.
- $\bar{\bar{Y}}_{j}=\sum_{h=1}^{N}\frac{\bar{Y}_{hj}}{N}$ = Mean per element in the population.
- $S_{bj}^{2}=\sum_{h=1}^{N}\frac{\left(\bar{Y}_{hj}-\bar{\bar{Y}}_{j}\right)^{2}}{N-1}$ = True variance between first stage unit means.

- $S_{wj}^2 = \sum_{h=1}^{N} \sum_{r=1}^{M} \frac{\left(y_{hrj} - \overline{Y}_{hj}\right)^2}{N\left(M-1\right)}$ = True variance within first stage units.

In case of equal first-stage units an unbiased estimate of $\overline{\overline{Y}}_j$ is $\overline{\overline{y}}_j$ with its sampling variance as,

$$V\left(\overline{\overline{y}}_j\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{bj}^2 + \left(\frac{1}{nm} - \frac{1}{NM}\right)S_{wj}^2, \quad j = 1,\ldots,p \qquad (1)$$

(*see proof in Appendix*)

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

$$C = C_1 n + C_2 mn \qquad (2)$$

Where:

- $C_1 =$ The cost of the survey in approaching a single primary stage unit.
- $C_{2j} =$ The cost of enumerating the j^{th} character per element.
- $C_2 = \sum_{j=1}^{p} C_{2j}$ = The cost of enumerating all the p characters per SSu.

Suppose that it is required to find the values of *n* and *m* so that the cost *C* is minimized, subject to the upper limits on the variances. If *N* and *M* are large, then from (1), the limits on the variances may be expressed as:

$$\frac{S_{bj}^2}{n} + \frac{S_{wj}^2}{nm} \le v_j, \quad j = 1,\ldots,p \qquad . \qquad (3)$$

Where v_j is the upper limits on the variances of various characters. Here S_{bj}^2 is the variance among primary stage units means and S_{wj}^2 is the variance among subunits within primary units for *j*th characteristic respectively.

The problem therefore reduces to find *n* and *m* which:

$$\textit{Minimize} \quad C = C_1 n + C_2 nm \qquad (4)$$

$$\textit{Subject to} \quad \frac{S_{bj}^2}{n} + \frac{S_{wj}^2}{nm} \le v_j, \quad j = 1,\ldots,p \qquad (5)$$

$$n \ge 1, \qquad m \ge 1 \qquad (6)$$

(In each primary stage unit at least one secondary stage unit has to be enumerated as negative values of PSUs and SSUs are of no practical use).

3. Geometric programming approach

Geometric programming (GP) is a technique for minimizing a function called a "posynomials" subject to several constraints. A posynomial is a polynomial in several variables with positive coefficients in all terms and the power to which the variables are raised can be any real numbers. Both the cost function and the variance constraint functions are posynomials. G.P transforms the primal problem of minimizing a "posynomial" subject to "posynomial" constraints to a dual problem of maximizing a function of the weights on each constraint. Usually there are fewer constraints than strata, so the transformation simplifies the procedure. The problem (4) - (6) as such takes the following mathematical form:

Find the vector $\underline{x}=\left(x_1,x_2\right)$ ($x_1=n$ and $x_2=nm$)

which minimizes $C(\underline{x}) = \sum_{i=1}^{2} C_i x_i = C_1 n + C_2 nm$ (7)

subject to $g(\underline{x}) = \sum_{i=1}^{2} \frac{a_{iq}}{x_q} \le v_q, \quad q=1,...,p$ (8)

and $x_i \ge 0, \quad i=1,2$ (9)

We have substituted in the above equations:

$x_1=n\ , x_2=nm\ , S_{bq}^2=a_{1q}\ , S_{wq}^2=a_{2q} \quad for\ q=1,...,p$

It may be noted that the objective function (7) is linear and the constraints (8) are nonlinear and the standard GP (Primal) problem stated with two subscripts is reduced to:

Minimize $f_0(x)$

Subject to $fq(x) \le 1, q=1,...p$ (10)

$x_j > 0, j=1,...n$

Where posynomial q is:

$$f_q(x) = \sum_{i=q} d_i \left[\prod_{j=1}^{n} x_j^{p_{ij}} \right], d_i > 0,\ x_j > 0, q=0,1,...,p, \quad (11)$$

where k denotes the number of posynomial terms in the function, n is the number of variables and the exponents p_{ij} are real constants. For our allocation problem, the objective function C($\underline{x}$) given in (7) and (8) has $k=2,\ n=2\ , p_{11}=p_{22}=1\ , p_{12}=p_{21}=0,\quad d_i=C_i\ , i=1,2$, and the q^{th} constraint has $k=2,\ n=2\ , p_{11}=p_{22}=-1\ , p_{12}=p_{21}=0$ and $d_i=a_{iq}\ , i=1,2$. (see Maqbool & Pirzada[5]). The dual of GP problem stated in (10) is given by:

$$Maximize\left[\prod_{q=0}^{p}\prod_{i\in[q]}\left(\frac{d_i}{w_i}\right)^{w_i}\right]\prod_{q=1}^{p}\left(\sum_{i\in[q]}w_i\right)^{\sum_{i\in[q]}w_i} \tag{12}$$

$$\text{subject to } \sum_{i\in[0]}w_i=1 \tag{13}$$

$$\sum_{q=0}^{p}\sum_{i\in[q]}P_{ij}w_i=0 \tag{14}$$

$$\begin{aligned}&w_i\geq 0,\ q=0,.....,p\\ &\quad i=1,.....,k_p\end{aligned} \tag{15}$$

Following Woolsey and Swanson [9] and Duffin et al [4], the allocation problem (7) & (8) will be solved in four steps as follows:

Step 1: The Optimum value of the objective function is always of the form

$$\begin{aligned}C_0(x^*)=&\left(\frac{Coeff.\ \ of\ \ \ first\ \ term}{w_1}\right)^{w_1}\times\left(\frac{Coeff.\ \ of\ \ \ Second\ \ \ term}{w_2}\right)^{w_2}\\ &\times......\times\left(\frac{Coeff.\ \ of\ \ \ last\ \ term}{w_K}\right)^{w_K}\\ &\left(\sum\ w's\ \ in\ \ the\ \ first\ \ constra\text{int}\,s\right)^{\sum\ w's\ \ in\ \ the\ \ first\ \ constra\text{int}\,s}\\ &\times\left(\sum\ w's\ \ in\ \ the\ \ last\ \ constra\text{int}\,s\right)^{\sum\ w's\ \ in\ \ the\ last\ \ constra\text{int}\,s}\end{aligned} \tag{16}$$

For our problem the objective function is:

$$\text{Cost} = \left(\frac{C_1}{w_1}\right)^{w_1}\left(\frac{C_2}{w_2}\right)^{w_2}\left(k_1\right)^{w_3}\left(k_2\right)^{w_4} \tag{17}$$

Where $k_1=\dfrac{a_1}{v_1}\ ,k_2=\dfrac{a_2}{v_2}$

Step 2: The equations generated for geometric program for the weights are

$$\sum w's \ in \ the \ objective \ \ function \ = 1 \tag{18}$$

and for each primal variable x_j given *n* variables and *k* terms

$$\sum_{i=1}^{m} (w_i \ \ for \ \ each \ \ term) \times (\exp onent \ \ on \ \ x_j \ \ in \ that \ \ term) = 0 \tag{19}$$

In our case:

$$w_1 + w_2 = 1 \text{ (Normalization condition, see (13))} \tag{20}$$

$$(1)\, w_1 + (0)\, w_2 + (-1)\ w_3 + (0) w_4 = 0 \tag{21}$$

$$(0)\, w_1 + (1)\, w_2 + (0)\ w_3 + (-1) w_4 = 0 \tag{22}$$

Equations (21) & (22) are Orthogonality conditions, see (14). Collectively, these conditions are referred to as dual constraints. For more details see Duffin *et al.* [4]. Now combining (20), (21) and (22), we get:

$$w_1 + w_2 = 1, \quad w_1 - w_3 = 0\,, \quad w_2 - w_4 = 0$$

which is a set of three linear equations in four unknowns. The above set of equations may be solved in terms of one w, say w_1.

$$w_2 = 1 - w_1\,, \qquad w_3 = w_1\,, \qquad w_4 = w_2 = 1 - w_1$$

Step 3: The contribution of terms in the constraints to optimal solution is always proportional to their weights. In this case:

$$\frac{k_1}{x_1} = \frac{w_3}{w_3 + w_4} = w_1 \tag{23}$$

$$\frac{k_2}{x_2} = \frac{w_4}{w_3 + w_4} = 1 - w_1 \tag{24}$$

From the above equations (23) & (24), we get:

$$\frac{k_2}{x_2} = 1 - \frac{k_1}{x_1}$$

which implies: $x_2^* = \dfrac{k_2 x_1}{x_1 - k_1}$ (25)

Step 4: The primal variables may be found by:

$$C_0(x^*) = \frac{\textit{first term in objective function}}{w_1} = \frac{\textit{second term in objective function}}{w_2}$$
$$= ...... = \frac{\textit{last term in objective function}}{w_K}$$

In this case:

$$\frac{C_1 x_1}{w_1} = \frac{C_2 x_2^{\ *}}{1-w_1}, \text{ here } \left[w_1 = \frac{k_1}{x_1} \right] \quad \textit{and} \left[1 - w_1 = \frac{k_2}{x_2} \right] \tag{26}$$

Since w_1 and $x_2^{\ *}$ are already known from (23) and (25), the above equation can be solved for $x_1^{\ *}$ in terms of the constants C and k, then:

$$\frac{C_1 x_1^{\ *}}{\frac{k_1}{x_1}} = C_2 \frac{\frac{k_2 x_1}{x_1^* - k_1}}{1 - \frac{k_1}{x_1^*}} \tag{27}$$

The above equation implies that:

$$x_1^* = k_1 + \sqrt{\frac{C_2 k_1 k_2}{C_1}} \tag{28}$$

From equations (25) and (28), we can easily calculate the optimum values for are n and m.

4. Numerical illustration

We consider Chakravarthy [2] for numerical illustration, where dispersion matrix for 2 characters in a sample of 20 PSUs and 8 SSUs in a situation when each PSU was drawn with equal probability at each stage is given below; the cost of enumerating a PSU is estimated as and that of SSU as:

Table 1. Dispersion Matrix.

Dispersion due to	Degree of Freedom	S.P. Matrix	Covariance Matrix
Between PSUs	19	$\begin{bmatrix} 0.5592 & 0.2993 \\ --- & 1.1026 \end{bmatrix}$	$\begin{bmatrix} 0.0294 & 0.0157 \\ --- & 0.0580 \end{bmatrix}$
Within PSUs ***Between SSUs***	140	$\begin{bmatrix} 1.0872 & 0.3568 \\ --- & 3.4041 \end{bmatrix}$	$\begin{bmatrix} 0.0078 & 0.0025 \\ --- & 0.0243 \end{bmatrix}$
Total	159	$\begin{bmatrix} 1.6454 & 0.6561 \\ --- & 4.5067 \end{bmatrix}$	

In this case the values of N and M are not known, they may be assumed to be infinite. Also the data may be taken as derived from a pilot survey and a similar survey is to be planned for which we require the best values of n and m which minimizes the total cost. Therefore from the above dispersion matrix, we have:

Sw_1^2 =0.0078, Sw_2^2 =0.0243

Sb_1^2 =0.0037, Sb_2^2 =0.0073

These are sample estimates and are subject to the sampling fluctuations. Now our problem is to minimize:

$$Minimize = C_1 n + C_2 nm \qquad (29)$$

$$Subject \;\; to \; \frac{Sb_1^2}{x_1} + \frac{Sw_1^2}{x_2} \le v_1 \qquad (30)$$

$$\frac{Sb_2^2}{x_1} + \frac{Sw_2^2}{x_2} \le v_2 \qquad (31)$$

$$x_1 \ge 1 , \quad x_2 \ge 1$$

The upper bound of v is calculated using the lower 5 percent point of χ^2 distribution with 19 d.f. =10.117, we have:

$$v_1 = 0.000345, \; v_2 = 0.000681 \qquad (32)$$

The upper confidence bounds of Sw_j^2 at 95 percent confidence level are:

$$\left.\begin{array}{l}\text{Upper bound of } Sw_1^2 = 0.009589 \\ \text{Upper bound of } Sw_2^2 {=} 0.030025\end{array}\right] \qquad (33)$$

The upper confidence bounds of Sb_j^2 at 95 percent confidence level are:

$$\left.\begin{array}{l}\text{Upper bound of } Sb_1^2 = 0.006129 \\ \text{Upper bound of } Sb_2^2 {=} 0.011180\end{array}\right] \qquad (34)$$

Using the values (32), (33) and (34), our problem becomes:

Minimize $\quad C = 8.7x_1 + 2.5x_2 \qquad (35)$

Subject to $\quad \dfrac{0.006129}{x_1} + \dfrac{0.009589}{x_2} \le 0.000345 \qquad (36)$

$$\frac{0.011180}{x_1} + \frac{0.030025}{x_2} \le 0.000681 \qquad (37)$$

$$x_1, x_2 \ge 0$$

The normalized constraints are:

$$\frac{0.006129}{0.000345x_1} + \frac{0.009589}{0.000345x_2} \le 1 \qquad (38)$$

$$\frac{0.011180}{0.000681x_1} + \frac{0.030025}{0.000681x_2} \le 1 \qquad (39)$$

Which gives:

$$\frac{17.76}{X_1} + \frac{27.79}{X_2} \le 1 \qquad (40)$$

$$\frac{16.41}{X_1} + \frac{44.08}{X_2} \le 1 \qquad (41)$$

Let us take the constraint (41) as active (if both constraints were active, then one would not be able to find an optimal dual solution nor an optimal solution to the original solution, see Shiang[8] and Maqbool & pirzada [5]. Then $K_1 = 16.41$ and $K_2 = 44.08$. Substituting the values of K_1, K_2, C_1 and C_2 in equation (28) and (25), we get:

$$x_1^* = 16.41 + \sqrt{\frac{2.5 \times 16.41 \times 44.08}{8.7}}$$

$$x_1^*(= n) = 30.82 \cong 31$$

and compute $m = \frac{93.27}{31} = 3.009 \cong 3$ and rounding yields:

$$x_2^*(= nm) = 31 * 3 = 93,$$

using the values of x_1^* and x_2^*, we get the total cost as:

$$C = 8.7 \times 31 + 2.5 \times 93 = 502.2.$$

Therefore the optimum values are $n = 31$ and $nm = 93$, i.e $m = 3$. This shows that the solution is feasible. Thus, we require a sample of 31 primary stage units and 3 secondary stage units in each primary stage unit giving us a total of $nm = 93$ elementary units for the sample. The above results can easily be verified through GP optimization algorithms available on internet (see GPGLP [10] & XGP [11].

4. Comments and Conclusion

Optimum allocation in two stage sampling is easy when dealing with one variable. However a simple technique has not been available when one is interested in estimating more than one variable. This paper is an attempt to utilize geometric programming approach to the solution of optimum allocation problems in multivariate two-stage sampling. The solution described here is much simpler than complex analytical techniques described in statistical literature. Geometric programming has already shown its power in practice in the past. In real world applications, the parameters in the geometric program may not be known precisely due to insufficient information. The numerical result illustrates the feasibility and effectiveness of the present approach. With the availability of GP optimization software, the wider applications of the proposed approach can be utilized in double sampling design having multiple characters and in case of response error (interviewer variability), various agricultural surveys where two stage sampling designs are frequently employed for different research studies.

Appendix: Proof of equation (1)

$$V\left(\bar{\bar{y}}_j\right)=\left(\frac{1}{n}-\frac{1}{N}\right)S_{bj}^2+\left(\frac{1}{nm}-\frac{1}{NM}\right)S_{wj}^2,$$

With simple random sampling at both the stages:

$$E\left(\bar{\bar{y}}_j\right)=E_1\left[E_2\left(\bar{\bar{y}}\right)\right]=E_1\left[\frac{1}{n}\sum\bar{Y}_i\right]=\left(\frac{1}{N}\sum\bar{Y}_i\right)=\bar{\bar{Y}}$$

$V\left(\bar{\bar{y}}_j\right)=V_1\left[E_2\left(\bar{\bar{y}}\right)\right]+E_1\left[V_2\left(\bar{\bar{y}}\right)\right]$, because:

$$\left\lfloor V\left(\hat{\theta}\right)\right\rfloor=V_1\left\lfloor E_2\left(\hat{\theta}\right)\right\rfloor+E_1\left\lfloor V_2\left(\hat{\theta}\right)\right\rfloor \quad .$$

Since $E_2\left(\bar{\bar{y}}\right)=\sum\frac{\bar{y}_i}{n}$, the first term on the right is the variance of the mean per subunit for a one stage simple random sample of n units, hence by using the basic theorems of SRS (see Cochran [1]):

$$V_1\left[E_2\left(\bar{\bar{y}}\right)\right]=\left(\frac{N-n}{N}\right)Sb_j^2 \tag{42}$$

Furthermore, with $\left(\bar{\bar{y}}\right)=\sum^{n}\frac{\bar{y}_i}{n}$ and simple random sampling used at the second stage:

$$V_2\left(\bar{\bar{y}}\right)=\left(\frac{M-m}{Mn^2}\right)\sum Sw_r^2/m \tag{43}$$

Where $Sw_r^2=\sum\left(y_{rj}-\bar{y}_r\right)^2/(M-1)$ is the variance among subunits for the h th primary unit. When we average over the first stage samples:

$\sum_{h=1}^{n}\frac{Sw_r^2}{n}$ averages to $\sum_{h=1}^{N}\frac{Sw_r^2}{N}=Sw_j^2$,

Hence:

$$E_1\left[V_2\left(\bar{\bar{y}}\right)\right]=\left(\frac{M-m}{M}\right)Sw_j^2\,/\,mn \tag{44}$$

Adding (42) and (44) gives:

$$V\left(\bar{\bar{y}}_j\right)=\left(\frac{1}{n}-\frac{1}{N}\right)S_{bj}^2+\left(\frac{1}{nm}-\frac{1}{NM}\right)S_{wj}^2,$$

If we ignore the terms independent of n and m, we get the variance:

$$V\left(\bar{\bar{y}}\right)=\frac{S_{bj}^2}{n}+\frac{S_{wj}^2}{nm}$$

References

[1]. Cochran, W.G. (1977). *Sampling techniques*. New York: John Wiley & Sons.

[2]. Chakravarthy, I. M. (1955). On the problem of planning a multistage survey for multiple correlated characters. *Shankhya*, 14,211-216.

[3]. Davis, M., Rudolf, E.S. (1987). Geometric programming for optimal allocation of integrated samples in quality control. *Comm.Stat.Theo.Meth.*,16 (11),3235-3254.

[4]. Duffin, R.J., Peterson, E.L., Zener, C. (1967). *Geometric programming: Theory & applications*. New York: John Wiley & Sons.

[5]. Maqool, S., Pirzada, S. (2007). Optimum allocation in multivariate two-stage sampling: An analytical solution. *Jour.of analysis & Computation*, Vol 3, No,1. 87-92.

[6]. Shaojian, Qu, Kecun, Z., Fusheng, W. (2008). A global optimization using linear relaxation for generalized geometric programming. *Europ. Jour.of Oper. Res.*,190, 345-356.

[7]. Shiang-Tai Liu (2008). Posynomial geometric programming with interval exponents and coefficients. *Europ. Jour.of Oper. Res.*,186,17-27.

[8]. Sukhatme, P.V., Sukhatme, B.V. Sukhatme, S. (1954). *Sampling theory of surveys with applications*. Iowa State University Press.,USA.

[9]. Woolsey, R.E., Swanson, H.S. (1975). *Operations Research for immediate applications: A quick and dirty manual*. New York: Harper & Row.

[10]. GPGLP: ftp://ftp.pitt.edu/dept/ie/GP/.

[11]. XGP: ftp://col.biz.uiowa.edu/dist/XU/doc/software.html.