Jour. Ind. Soc. Ag. Statistics
50(2), 1997 : 121–129

Determining the Optimum Cluster Size

M.G.M. Khan, Nujhat Jahan[1] and M.J. Ahsan
Aligarh Muslim University, Aligarh-202002
(Received : May, 1994)

SUMMARY

The problem of determining the optimum size of the sampling unit is formulated as a Mathematical Programming Problem (MPP). The MPP is simplified using a double transformation. The resulting MPP is then solved using Lagrange multiplier technique and an explicit formula is obtained for the optimum size of the sampling unit.

Key words: Optimum size, Sampling unit, Optimum solution.

1. Introduction

Cluster sampling is widely practiced in sample surveys. In the absence of a complete sampling frame, cluster sampling helps in reducing the cost of survey, by not requiring a complete sampling frame. In cluster sampling procedure the population is divided into groups or clusters of smaller units of the population called the sampling units. The size of these sampling units influences the efficiency of sampling. It is, therefore, advisable to use that cluster size which is optimum with respect to a carefully chosen criterion. Attempts have been made by various authors to determine the cluster size so that maximum precision is attained within the available resources. Jessen [5], [6], Mohalanobis [9], Homeyer and Black [4], Sukhatme [12], [13], Cochran [1], [2], Sukhatme and Panse [14], Hansen *et al.* [3], Murthy [10], Sheela and Unnithan [11], etc. worked on the problem of determining the optimum cluster size. Substitution, trial and error and graphical methods are commonly used to obtain the optimum cluster size. In this paper, we derive an explicit formula for the optimum cluster size by formulating the problem as a mathematical programming problem (MPP).

The original problem has a convex objective function for a specified range of measure of homogeneity and a constraint function which is indefinite, that is, neither convex nor concave, which does not allow the use of any available non-linear programming technique. Through suitable transformations the

1 *Department of Statistics, University of Dhaka, Dhaka, Bangladesh*

problem is transformed into a simple non-linear programming problem (NLPP) which has a convex objective function and a single linear constraint. Several techniques are available for solving this type of NLPP, better known as Convex Programming Problem (CPP) with linear constraints. We used Lagrange multipliers technique to solve the transformed problem and an explicit formula for the optimum cluster size is obtained. The Kuhn-Tucker (K-T) [7] necessary conditions, which are sufficient also for the transformed problem, are verified at the optimum solution. A numerical example is also presented to illustrate the computational details.

2. *Formulation*

If a simple random sample of n clusters is drawn from a population of N clusters of equal size M, the variance (f.p.c. ignored) of the sample mean per element $\bar{\bar{y}}$ which is an unbiased estimate of the population mean per element $\bar{\bar{Y}}$ is

$$V(\bar{\bar{y}}) = \frac{S^2}{nM}[1 + (M-1)\rho] \tag{2.1}$$

where ρ is the intracluster correlation coefficient *i.e.*, measure of homogeneity. Obviously $\rho \geq -\frac{1}{M-1}$, otherwise the RHS of (2.1) will become negative. Hansen *et al.* [3] expressed ρ as a function of the cluster size M as

$$\rho = aM^b$$

where a and b are constants and can be estimated for any two levels of M. Jessen [6] has shown that ρ do not vary much with small changes in M. This suggests that over a short but relevant range of M, ρ can be regarded as a constant.

Ignoring the over-head cost of planning and analysis, the total cost C of the survey may be expressed as

$$C = c_1 nM + c_2 d \tag{2.2}$$

where

c_1 = the cost of enumeration per element including the travel cost within clusters.

c_2 = the cost of travelling unit distance between clusters.

d = the total distance between n randomly selected clusters.

Mahalanobis [8] empirically showed that the expected total distance 'd' between n points located at random is proportional to $n^{1/2} - n^{-1/2}$. Jessen [5] showed that for practical purposes $n^{1/2}$ is a fairly good approximation to 'd'. Using this approximation he expressed the total cost of the survey as

$$C = c_1 nM + c_2\sqrt{n} \tag{2.3}$$

For a fixed budget C_0, the optimum values of n and M are those which minimize the variance given by (2.1) within the specified budget. This problem of determining optimum values of n and M may be expressed as the following NLPP

Minimize $$V(n, M) = \frac{S^2}{nM}\,[1 + (M-1)\rho] \tag{2.4}$$

Subject to $$c_1 nM + c_2\sqrt{n} \le C_0 \tag{2.5}$$

and $$n \ge 0,\ M \ge 0 \tag{2.6}$$

The restrictions in (2.6) are obvious because negative values of n and M are of no practical use. Also the cluster size M is regarded as a continuous variable which is a practical situation in most of the agricultural surveys.

The values of S^2, ρ, c_1 and c_2, if not known, may be estimated by conducting empirical studies on the data collected on some previous census or by conducting a pilot survey.

It is seen that the objective function V(n, M) is convex for $\rho \ge -\frac{3}{4M-3}$ and the constraint function $C(n, M) = c_1 nM + c_2\sqrt{n}$ is indefinite, that is, neither convex nor concave which implies that the feasible region defined by (2.5) and (2.6) is not convex.

The following transformations simplify the problem to a great extent, in a non-linear programming problem with convex objective function and a single linear constraint.

Letting $$x_2 = \sqrt{n} \quad \text{or} \quad x_2^2 = n \tag{2.7}$$

and separating the two terms in the objective function the problem may be rewritten as

Minimize $$V(x_2, M) = \frac{S^2(1-\rho)}{x_2^2 M} + \frac{S^2\rho}{x_2^2}$$

Subject to $c_1x_2^2M + c_2x_2 \leq C_0$

and $M,\ x_2 \geq 0$

A further transformation

$$x_1 = x_2^2M \tag{2.8}$$

converts the problem into

Minimize $$V(x_1, x_2) = \frac{S^2(1-\rho)}{x_1} + \frac{S^2\rho}{x_2^2} \tag{2.9}$$

subject to $$c_1x_1 + c_2x_2 \leq C_0 \tag{2.10}$$

and $$x_1, x_2 \geq 0 \tag{2.11}$$

It is to be noted that the additional constraint

$$\log x_1 = 2 \log x_2 + \log M \tag{2.12}$$

as a result of the transformation (2.8) is not required here because, out of the two variables x_2 and M, only x_2 appears independently in the problem while M appears only in the product term x_2^2M. Therefore, if (x_1^*, x_2^*) is an optimal solution to the NLPP (2.9) - (2.11) then the optimum value of M is given by $M^* = \frac{x_1^*}{(x_2^*)^2}$ and the additional constraint (2.12) is satisfied automatically.

The problem (2.9) - (2.11) is a non-linear programming problem and may be solved by using an appropriate non-linear programming technique.

3. *The Solution*

In NLPP (2.9) - (2.11), $V(x_1, x_2)$ will be minimum when the values of x_1 and x_2 are as large as permitted by the constraints (2.10) and (2.11) of the problem. This suggests that at the optimum point constraint (2.10) will be active, that is, it is satisfied as an equation. Ignoring restrictions (2.11), $V(x_1, x_2)$ may be minimized subject to the condition (2.10) with equality sign by Lagrange multipliers technique. If the values of x_1 and x_2 also satisfy the restrictions (2.11), they will solve the non-linear programming problem (2.9) - (2.11) provided the lagrangian function φ defined in (3.1) is convex.

The lagrangian function φ is defined as

$$\varphi(x_1, x_2, u) = \frac{S^2(1-\rho)}{x_1} + \frac{S^2\rho}{x_2^2} + u(c_1x_1 + c_2x_2 - C_0) \tag{3.1}$$

where u is the lagrange multiplier.

The necessary conditions for the solution of the problem are

$$\frac{\partial\varphi}{\partial x_1} = -\frac{S^2(1-\rho)}{x_1^2} + uc_1 = 0 \tag{3.2}$$

$$\frac{\partial\varphi}{\partial x_2} = -\frac{2S^2\rho}{x_2^3} + uc_2 = 0 \tag{3.3}$$

and

$$\frac{\partial\varphi}{\partial u} = c_1x_1 + c_2x_2 - C_0 = 0 \tag{3.4}$$

(3.2) and (3.3) give

$$u = \frac{S^2(1-\rho)}{c_1x_1^2} = \frac{2S^2\rho}{c_2x_2^3} \tag{3.5}$$

which gives

$$x_1^2 = \frac{c_2}{2c_1} \cdot \frac{1-\rho}{\rho} \cdot x_2^3 \tag{3.6}$$

By (3.4) we get

$$x_1 = \frac{C_0 - c_2x_2}{c_1} \tag{3.7}$$

Substituting the value of x_1 from (3.7) in (3.6) and simplifying we get

$$x_2^3 - \frac{2c_2}{c_1}\left(\frac{\rho}{1-\rho}\right)x_2^2 + \frac{4C_0}{c_1}\left(\frac{\rho}{1-\rho}\right)x_2 - \frac{2C_0^2}{c_1c_2}\left(\frac{\rho}{1-\rho}\right) = 0 \tag{3.8}$$

Substituting $\lambda = \frac{\rho}{1-\rho}$, we get the equation as

$$x_2^3 - \frac{2c_2\lambda}{c_1}x_2^2 + \frac{4C_0\lambda}{c_1}x_2 - \frac{2C_0^2\lambda}{c_1c_2} = 0 \tag{3.9}$$

Equation (3.9) is a cubic equation. To solve it, we substitute

$$x_2 = y + \frac{2c_2\lambda}{3c_1} \tag{3.10}$$

Then (3.9) reduces to

$$y^3 + \frac{4\lambda(3C_0c_1 - c_2^2\lambda)}{3c_1^2}\,y - \frac{2\lambda(8c_2^4\lambda^2 - 36C_0c_1c_2^2\lambda + 27C_0^2c_1^2)}{27c_1^3c_2} = 0 \tag{3.11}$$

The theory of equation gives the roots of an equation $z^3 + pz + q = 0$ as

$$z = \left\{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}\right\}^{1/3} + \left\{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}\right\}^{1/3}$$

Therefore, the roots of the equation (3.11) are

$$y = \frac{1}{3c_1}\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a + \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3} + \left(a - \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3}\right\}$$

where $a = 8c_2^4\lambda^2 - 36C_0c_1c_2^2\lambda + 27C_0^2c_1^2$ (3.12)

and $b = 3C_0c_1 - c_2^2\lambda$ (3.13)

Substituting this value of y in (3.10) we get

$$x_2^* = \frac{\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a + \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3} + \left(a - \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3}\right\} + 2c_2\lambda}{3c_1} \tag{3.14}$$

By (2.7) $n = x_2^2$, therefore

$$n = \left[\frac{\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a + \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3} + \left(a - \sqrt{a^2 + 64c_2^2\lambda b^3}\right)^{1/3}\right\} + 2c_2\lambda}{3c_1}\right]^2 \tag{3.15}$$

Substituting this solution of x_2 in (3.7) we can get the corresponding solution of x_1 as

$$x_1^* = \frac{3C_0c_1 - c_2\left[\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a+\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}+\left(a-\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}\right\}+2c_2\lambda\right]}{3c_1^2} \tag{3.16}$$

Which alongwith (2.8) and (3.14) gives after simplification the explicit expression for M as

$$M = \frac{9C_0c_1 - 3c_2\left[\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a+\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}+\left(a-\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}\right\}+2c_2\lambda\right]}{\left[\left(\frac{\lambda}{c_2}\right)^{1/3}\left\{\left(a+\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}+\left(a-\sqrt{a^2+64c_2^2\lambda b^3}\right)^{1/3}\right\}+2c_2\lambda\right]^2} \tag{3.17}$$

where λ, a and b are obtained from (3.8), (3.12) and (3.13) respectively.

However, M can be more easily obtained by treating (2.5) as an equation and solving it for M after substituting the numerical value of n obtained by (3.15).

Thus
$$M = \frac{C_0 - c_2\sqrt{n}}{c_1 n} \tag{3.18}$$

As the objective function (2.9) is convex for $\rho \geq 0$ and the constraint (2.10) is linear the (K–T) necessary conditions for the NLPP (2.9) - (2.11) are sufficient also. These conditions are

$$\nabla_{(x_1, x_2)}\varphi = \begin{pmatrix} -\frac{S^2(1-\rho)}{x_1^2} + uc_1 \\ -\frac{2S^2\rho}{x_2^3} + uc_2 \end{pmatrix} \geq 0$$

$$x_1\left[-\frac{S^2(1-\rho)}{x_1^2} + uc_1\right] + x_2\left[-\frac{2S^2\rho}{x_2^3} + uc_2\right] = 0$$

$$\nabla_u\varphi = c_1x_1 + c_2x_2 - C_0 \leq 0$$

$$u(c_1x_1 + c_2x_2 - C_0) = 0$$

and x_1, x_2 and $u \geq 0$

For the case x_1, x_2 and $u > 0$ the above expressions give rise to the same set of equations as (3.2), (3.3) and (3.4) which implies that the K–T conditions hold at the point (x_1^*, x_2^*) given by (3.16) and (3.14). Hence (x_1^*, x_2^*) is optimum for NLPP (2.9) - (2.11).

4. *A Numerical Illustration*

To illustrate the use of formulae (3.15) and (3.18) the following exercise from Cochran [2] has been worked out.

In a rural survey in which the sampling unit is a cluster of M farms, the cost of taking a sample of n units is

$$C = 4tMn + 60\sqrt{n}$$

where t is the time in hours spent getting the answers from a single farmer. If \$ 2000 is to be spent on the survey and $\rho = 0.1$, the optimum values of n and M for two different values of (a) $t = \frac{1}{2}$ hour and (b) $t = 2$ hours are worked out as follows:

(a) $t = \frac{1}{2}$ hour

We have $C_0 = 2000$, $c_1 = 4 \times \frac{1}{2} = 2$, $c_2 = 60$ and $\rho = 0.1$

The values of λ, a and b given by (3.8), (3.12) and (3.13) respectively are

$$\lambda = \frac{1}{9},\ a = 375680000 \text{ and } b = 11600$$

Using (3.15) and (3.18) we get the optimum values for n and M respectively as

$$n = 187.64 \cong 188 \text{ and } M = 3.14$$

(b) $t = 2$ hours

In this case $c_1 = 4 \times 2 = 8$

We have $\lambda = \frac{1}{9}$, $a = 6682880000$ and $b = 47600$

Using (3.15) and (3.18) we get the optimum values for n and M respectively as

$$n = 95.07 \cong 95 \quad \text{and} \quad M = 1.86$$

REFERENCES

[1] Cochran, W.G., 1948. Notes on *Sampling Survey Techniques* (Mimeographed). Institute of Statistics, Raleigh, North Carolina.

[2] Cochran, W.G., 1977. *Sampling Techniques*. 3rd ed., John Wiley and Sons Inc., New York.

[3] Hansen, M.H., Hurwitz, W.N. and Madow, W.G., 1953. *Sample Survey Methods and Theory*. 1, John Wiley and Sons, New York.

[4] Homeyer, P.G. and Black, C.A., 1946. Sampling replicated field experiments on oats for yield determinations. *Proc. Soil Sci. Soc. America*, 11, 341-344.

[5] Jessen, R.J., 1942. Statistical Investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304.

[6] Jessen, R.J., 1978. *Statistical Survey Techniques*. John Wiley and Sons Inc., New York.

[7] Kuhn, H.W. and Tucker, A.W., 1951. Nonlinear programming. *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, University of California Press, Berkeley, 481-492.

[8] Mahalanobis, P.C., 1940. A sample survey of the acreage under jute in Bengal. *Sankhya*, 4, 511-530.

[9] Mahalanobis, P.C., 1944. On large-scale sample surveys. *Phil. Trans. Roy. Soc.*, London, B231.

[10] Murthy, M.N., 1967. *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[11] Sheela, M.A. and Unnithan, V.K.G., 1992. Optimum size of plots in multivariate case. *J. Indian Soc. Agric. Statist.*, 44(3), 236-240.

[12] Sukhatme, P.V., 1947. The problem of plot size in large-scale yield surveys. *J. Amer. Statist. Assoc.*, 42, 297-310.

[13] Sukhatme, P.V., 1950. Sample surveys in agriculture. *Presidential Address to the Section of Statistics*, 37th Session, Indian Science Congress, Pomona.

[14] Sukhatme, P.V. and Panse, V.G., 1951. Crop surveys in India-II. *J. Indian Soc. Agric. Statist.*, 3, 97-168.