East-West J. of Mathematics: Vol. 8, No 1 (2006) pp. 77-82

INTEGER SOLUTION TO THE ALLOCATION PROBLEM IN STRATIFIED SAMPLING

S. Pirzada[*] and **S. Maqbool**[†]

[*] *Department of Mathematics,*
University of Kashmir, Srinagar-190006, India
e-mail: sdpirzada@yahoo.co.in

[†] *Department of Statistics*
SK University of Agricultural Sciences
Srinagar, India

Abstract

In this paper, we formulate the stratified sampling problem with linear and non-linear sampling costs and use branch and bound technique to obtain the integer solution. In numerical illustration, it is seen that the situation of more than hundred percent sampling can be tackled using branch and bound technique.

1.Introduction

For practical applications of any allocations, integer values of the sample sizes are required. This could be done by simply rounding off the non integer sample sizes to the nearest integral values. When the sample sizes are large enough or the measurement cost in various strata are not too high, the rounded off sample allocation may work well. However, for small samples in some situations the rounding off allocations may become infeasible and non-optimal. This means that rounded off values may violate some of the constraints of the problem or there may exist other sets of integer sample allocations with a lesser value of the objective function. In such situations, we have to use some integer programming technique to obtain an optimum integer solution. The dynamic programming approach to obtain an integer solution is used by several authors

Key words: Stratified sampling, non linear integer programming, allocation problem, langrangian multiplier.
2000 AMS Mathematics Subject Classification:

such as, Arthanari and Dodge [1] and Khan [3]. However, the dynamic programming approach is too inefficient as is evident by the numerical examples solved in these references.

In this paper, we use the branch and bound technique for obtaining the integer solution to the allocation problems in stratified sampling with linear and non-linear sampling costs formulated as non-linear integer programming problems. The basic idea of branch and bound is to partition a given problem into a number of sub problems. This process of partitioning is usually called branching and its purpose is to establish sub-problems that are easier to solve than the original problem, because of their smaller size or amenable structure. A numerical illustration is also presented and it is seen that the optimal (non-integer) solution requires more than hundred percent sampling. This situation is then tackled by using branch and bound technique.

2. Problem formulation

The following notation will be used to define the sample allocation problem. The decision variable of interest is the sample size for each stratum. The suffix h stands for h^{th} stratum, $h = 1, 2, ..., L$, where L denotes the total number of strata into which the population has been divided.

N_h = Total number of units in the stratum h.
n_h = Number of units selected in the sample from the stratum h.
$W_h = \frac{N_h}{N}$ = Proportion of population units falling in the stratum h.
$\bar{y}_h$ = hth Stratum mean.
S_h^2 = hth Stratum variance.
C_h = Cost of surveying one unit in stratum h, $(C_h > 0, h = 1, 2, ..., L)$.
The stratified sample mean is defined as

$$\bar{y}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h,$$

and the variance of $\bar{y}_{st}$ is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}.$$

For large strata sizes the second term on the right may be ignored and we get

$$V(\bar{y}_{st}) \tilde{} \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}.$$

The problem of optimal sample allocation involves determining the sample sizes $n_1, n_2, ..., n_h$ that minimize the variance $V(\bar{y}_{st})$ subject to a given sampling budget C, or determining $n_1, n_2, ..., n_h$ that minimize the sampling cost subject to an upper bound on the variance. The simplest cost function is of

the form $\sum_{h=1}^{L} C_h n_h$. Within any stratum the cost is proportional to the size of sample, but the cost per unit C_h may vary from stratum to stratum. This cost function is appropriate when the major item of cost is that of taking the measurements on each unit. If travel costs between units in a given stratum are substantial, empirical and mathematical studies indicate that costs are better represented by the expression $\sum_{h=1}^{L} t_h \sqrt{n_h}$, where t_h is the travel cost per sample unit within the stratum h.

Below we consider the integer allocation problems for both the types of costs. Both the problems are non- linear integer programming problems and are solved by using branch and bound approach of Land and Doig [4].

For linear cost function and fixed budget the problem of sample allocation is formulated as

$$\text{Minimize } \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}, \quad (2.1)$$
$$\text{Subject to } \sum_{h=1}^{L} C_h n_h \leq C, \quad (2.2)$$
$$1 \leq n_h \leq N_h, \quad (2.3)$$
$$n_h \text{ integers.} \quad (2.4)$$

For the non linear cost function considered above the budget constraints (2.2) takes the form

$$\sum_{h=1}^{L} t_h \sqrt{n_h} \leq C. \quad (2.5)$$

3. Solution procedure

We first derive the solution of problem (2.1) to (2.2), that is, by ignoring the upper and lower bounds (2.3) and the integer requirements (2.4).

Forming the Lagrangian

$$\phi = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} + \lambda \left[\sum_{h=1}^{L} C_h n_h - C \right]$$

and differentitating with respect to n_h and λ, we get

$$\frac{\partial \phi}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda C_h = 0,$$
$$\frac{\partial \phi}{\partial \lambda} = \sum_{h=1}^{L} C_h n_h - C = 0,$$

which on simplification give the initial solution

$$n_h = \frac{C W_h S_h \sqrt{C_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{C_h}}, \; h = 1, 2, ..., L. \quad (3.1)$$

Now, the Land and Doig approach of the branch and bound technique will require the solution of subproblems in which some of the n_h are fixed. Suppose that at K^{th} node, the fixed values of n_h are for $h \in I_k$. Then at K^{th} node, we form the lagrangian

$$\phi = \sum_{h=I_k}^{L} \frac{W_h^2 S_h^2}{n_h} + \lambda \left[\sum_{h=I_k}^{L} C_h n_h - C \right].$$

Equating to zero the differentials of ϕ with respect to n_h and λ, we obtain the solution at K node as

$$n_h = \frac{(C - \sum C_i \ n_i) W_h S_h \ / \sqrt{C_h}}{\sum_{h \neq I_k}^{L} W_h S_h \sqrt{C_h}}, \ h = 1, 2, ..., L. \ (3.2)$$

Next, we derive the solution of the sub-problem of minimizing (2.1) subject to non linear constraint (2.5).

The corresponding lagrangian is given by

$$\phi = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} + \lambda \left[\sum_{h=1}^{L} t_h \sqrt{n_h} - C \right].$$

Equating to zero the differentials of ϕ with respect to n_h and λ, we get

$$\frac{\partial \phi}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda \left[t_h \times \frac{1}{2} \times n_h^{\frac{-1}{2}} \right] = 0,$$

which implies $n_h = \left(\frac{2 W_h^2 S_h^2}{t_h} \right)^{\frac{2}{3}} \left(\frac{1}{\lambda} \right)^{\frac{2}{3}}$ (3.3)

and $\frac{\partial \phi}{\partial \lambda} = \sum_{h=1}^{L} t_h \sqrt{n_h} - C = 0,$

which implies $\sum_{h=1}^{L} t_h \sqrt{n_h} = C.$ (3.4)

Now from (3.3), we have

$\sum_{h=1}^{L} t_h \sqrt{n_h} = \sum_{h=1}^{L} \left(\frac{2 W_h^2 S_h^2}{\lambda} \right)^{\frac{1}{3}} t_h^{\frac{2}{3}}.$ (3.5)

Combining (3.4) and (3.5), we get

$$C = \left(\frac{1}{\lambda} \right)^{\frac{1}{3}} \sum_{h=1}^{L} \left(2 W_h^2 S_h^2 \right)^{\frac{1}{3}} t_h^{\frac{2}{3}},$$

or $\left(\frac{1}{\lambda} \right)^{\frac{2}{3}} = \frac{C^2}{\left[\sum_{h=1}^{L} \left(2 W_h^2 S_h^2 \right)^{\frac{1}{3}} t_h^{\frac{2}{3}} \right]^2}.$

Substituting the value of $\left(\frac{1}{\lambda} \right)^{\frac{2}{3}}$ in (3.3), we obtain the initial solution for the non-linear cost case as

$$n_h = \frac{C^2 \left(\frac{W_h^2 S_h^2}{t_h} \right)^{\frac{2}{3}}}{\left[\sum_{h=1}^{L} \left(W_h^2 S_h^2 \right)^{\frac{1}{3}} (t_h)^{\frac{2}{3}} \right]^2}, \ h = 1, 2, ..., L. \ (3.6)$$

The formula for the K^{th} node solution is obtained in a parallel manner to the linear cost case as

$$n_h = \frac{\left(C - \sum_{i \in I_k} t_i \sqrt{n_i} \right)^2 \left(\frac{W_h^2 S_h^2}{t_h} \right)^{\frac{2}{3}}}{\left[\sum_{h=I_k}^{L} \left(W_h^2 S_h^2 \right)^{\frac{1}{3}} (t_h)^{\frac{2}{3}} \right]^2}, \ h = 1, 2, ..., L. \ (3.7)$$

For branching from each node of the (Land and Doig Branch and Bound) tree we will choose an at the current node which either violates the integer requirements (2.4) or which voilates the upper or lower bounds (2.3). Whenever the branching is done on the bounds, then one branch will fix the corresponding n_h on the voilated bound and the other on the next feasible integer value.

4. Numerical Illustration

The data in table below is related to the number of inhabitents of 64 large cities in the U.S., in thousands for the year 1930 [Arthanari and Dodge[1]. The cities are grouped into three strata. There are 16,20 and 28 cities respectively in the first, second and third stratum. Suppose that the total budget available for the sample survey is 80 (in hundred dollars).

Table 4.1

Stratum	$\mathbf{N_h}$	$\mathbf{S_h^2}$	$\mathbf{W_h}$	$\mathbf{C_h}$
1.	16	540.0625	0.25	3.5
2.	20	14.6737	0.3125	1.5
3.	28	7.2540	0.4375	01

It is required to find the sample numbers for the three strata so as minimize the variance of the estimate while remaining within the budgetary limits.

The allocation problem may be stated as follows.

$$\text{Minimize } Z = \tfrac{33.75}{n_1} + \tfrac{1.43}{n_2} + \tfrac{1.39}{n_3}, \text{ Subject to } 3.5n_1 + 1.5n_2 + n_3 \leq 80,$$
$$1 \leq n_1 \leq 16,$$
$$1 \leq n_2 \leq 20,$$
$$1 \leq n_3 \leq 28, n_1, n_2, n_3 \text{ integers.}$$

Using (3.1), we get the infeasible solution as

$$n_1 = 18.39,\ n_2 = 5.78,\ n_3 = 6.97,\ Z^* = 2.28.$$

The solution does not satisfy the upper bound on n_1 . This requires more than hundred percent sampling in the first stratum. We create two branches- one leading to node 2 by fixing $n_1 = 16$ and other leading to node 3 by fixing $n_1 = 15$. By using (3.2) we obtain

$$\text{node 2:}\ \ n_1 = 16,\ n_2 = 8.88,\ n_3 = 10.68,\ Z_2^* = 2.404,$$
$$\text{node 3:}\ \ n_1 = 15,\ n_2 = 10.18,\ n_3 = 12.24,\ Z_3^* = 2.50.$$

As $Z_2^* < Z_3^*$, next we branch from node 2 by fixing $n_2 = 8$ leading to node 4 and by fixing $n_2 = 9$ leading to node 5.

The first integer solution is obtained at node 4 as

$$n_1 = 16,\ n_2 = 8,\ n_3 = 12,\ Z^* = 2.404.$$

This solution also happens to be the optimal solution of our allocation problem as can be seen in the branch and bound tree.

1 5. Conclusion

The optimal non-integer solution requires more than hundred percent sampling in stratum 1. The integer solution obtained by branch and bound method (16,8,12) has an objective value as 2.404. The solution (16,6,7) obtained from node 1 by rounding to the nearest integers in stratum 2 and 3 and fixing at the upper limit 16 in stratum 1 gives the objective function value as 2.546. We thus obtain 106 percent efficiency, remaining within given budgetory limits, by usuing the branch and bound technique. The branch and bound solution provides a guaranteed optimal solution within a very reasonable amount of computing time. Presumably another heuristic can be devised to handle non identical costs and non linear cost functions, but we are not aware of any such approach in the statistical literature.

References

[1] T. S. Arthanari and Y.Dodge, "Mathematical programming in Statistics", Wiley, New York, 1981.

[2] A.T.Hamdy, "Integer programming theory, applications and computations", Academic press, Inc. London, 1975.

[3] E.A.Khan, *On use of mathematical programming techniques in some optimization problems arising in stratified sample surveys,* Ph. D. Thesis, AMU, Aligarh, India (1997).

[4] A.H.Land and A.G.Doig, *An automatic method for solving discrete programming problems,* Econometrica, 28(1960), 497-520.

[5] S.Maqbool, *Optimization techniques in sample surveys,* Ph.D. Thesis, AMU, Aligarh, India (2001).