International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:8, No:3, 2014

On Optimum Stratification

M. G. M. Khan, V. D. Prasad, D. K. Rao

Abstract**—**In this manuscript, we discuss the problem of determining the optimum stratification of a study (or main) variable based on the auxiliary variable that follows a uniform distribution. If the stratification of survey variable is made using the auxiliary variable it may lead to substantial gains in precision of the estimates. This problem is formulated as a Nonlinear Programming Problem (NLPP), which turn out to multistage decision problem and is solved using dynamic programming technique.

Keywords**—**Auxiliary variable, Dynamic programming technique, Nonlinear programming problem, Optimum stratification, Uniform distribution.

I. Introduction

STRATIFIED random sampling is the most commonly used sampling technique for estimating population parameters (mean or total) with greater precision in sample surveys. To gain the precision in the estimates using stratified sampling one of the basic problem is the determination of the optimum strata boundaries and the research carried out in this paper is to deal with this problem.

Indisputably, optimum stratification could be achieved effectively by having the distribution of the main study variable known, and create strata by cutting the range of the distribution at suitable points. However, it is an unrealistic assumption that stratification can be made based on the frequency distribution of study variable (y), which is unknown prior to conducting the survey. Thus the non-availability of knowledge about the main study variable forces one to substitute for it the distribution of another known closely related variable (x), called auxiliary variable which is easily available with minimum cost and effort. Often y is highly correlated with x such that the regression of y upon x has homoscedastic errors. In situations like this, stratification can be achieved using the auxiliary variable.

If the stratification is made based on x, it may lead to substantial gains in precision in the estimate, although it will not be as efficient as the one based on y. However, if the regression of y on x fits well within all strata, the boundary points for both the variables should be nearly the same.

The construction of strata has a long history in the statistical sciences dating back to 1950. It is known that stratified random sampling will be efficient if the strata are internally homogeneous as possible as with respect to the characteristics under study in other words. In order to achieve maximum precision, the stratum variances should be as small as possible for a given type of sample allocation. One way to achieve this is to use the available prior information about the population to form the groups of similar units and take the groups as strata. This problem of determining the OSB, when both the estimation and stratification variables are the same, was first discussed by [3].

When a single variable is under study and its frequency distribution is known it can be used for determining the strata boundaries. Several authors including [1], [5]-[7], [9], [15], [17], and [20] used the frequency distribution of the main study variable for determining the strata boundaries under various allocations of the sample sizes.

Most of these authors achieved the calculus equations for the strata boundaries which are not suitable to adopt for practical computations. They obtained only the approximate solutions under certain assumptions.

When the frequency distribution of an auxiliary variable, x, is known, many authors such as [4], [8], [16], [19], [21]-[28] have suggested different approximation method of determining OSB.

Another kind of stratification method that has been proposed in the literature is due to [11]. They formulated the problem of determining OSB as an optimization problem and developed a computational technique to solve the problem by using dynamic programming. This procedure could give exact solution, if the frequency distribution of the study variable is known and the number of strata is fixed in advance.

In this paper, we extend the [11] technique to deal with the problem of determining OSB for a population using a single auxiliary variable with Uniform frequency distribution.

II. Formulation of the Problem of Determining OSB as an NLPP Based on Auxiliary Variable

Let the population be stratified into L strata based on a single auxiliary variable x and the estimation of the mean of study variable y is of interest.

Let the regression model of y on x be:

$$y = a + bx + e, \tag{1}$$

where $E(e|x) = 0$, $V(e|x) = \varphi(x)$ for all x.

Assuming that the variable x has a continuous frequency function $f(x)$, $a \le x \le b$ and the stratification points forming L strata are $x_1, x_2, x_3, ..., x_L$. Then, for the h^{th} stratum with

M. G. M. Khan, V. D. Prasad, and D. K. Rao are with the School of Computing, Information and Mathematical Sciences, Faculty of Science, Technology and Environment, The University of the South Pacific, Suva, Fiji Islands (Corresponding author: e-mail: rao_di@usp.ac.fj; phone: +679 3232603; fax: +679 3231527).

boundary points x_{h-1} and x_h, the proportion (W_h), the stratum mean (μ_{x_h}) and the stratum variance $(\sigma^2_{x_h})$ are given by

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx \tag{2}$$

$$\mu_{x_h} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} xf(x)dx \tag{3}$$

and

$$\sigma^2_{x_h} = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} x^2 f(x)dx - \mu^2_{x_h}. \tag{4}$$

If x and ε are uncorrelated, from the model (1), the variance of y in h^{th} stratum can be expressed as

$$\sigma^2_{y_h} = \beta^2\sigma^2_{x_h} + \sigma^2_{e_h}, [5] \tag{5}$$

where $\sigma^2_{e_h}$ is the expected variance of ε in the h^{th} stratum and can be obtained as

$$\sigma^2_{e_h} = \int_{x_{h-1}}^{x_h} \varphi(x)\ f(x)\ dx. \tag{6}$$

Ignoring the finite population correction (*f.p.c.*) and using (5) the variance of stratified mean $\bar{y}_{st}$ under Neyman allocation [18] for a fixed total sample size n is given as:

$$V(\bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L}\left[W_h\sqrt{\beta^2\sigma^2_{x_h} + \sigma^2_{e_h}}\right]^2. \tag{7}$$

In order to minimize $V(\bar{y}_{st})$, for fixed n it is sufficient to minimize

$$V(\bar{y}_{st}) = \sum_{h=1}^{L}\left[W_h\sqrt{\beta^2\sigma^2_{x_h} + \sigma^2_{e_h}}\right]. \tag{8}$$

Clearly, from (1) - (6), the $V(\bar{y}_{st})$ is a function of boundary points x_{h-1} and x_h, that is,

$$W_h\sqrt{\beta^2\ \sigma^2_{x_h} + \sigma^2_{e_h}} = \phi_h(x_{h-1}, x_h). \tag{9}$$

Then the optimization problem to determine $x_1, x_2, ..., x_L$ can be expressed as:

Minimize $\sum_{h=1}^{L}\phi(x_{h-1}, x_h)$,

subject to

$$x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L. \tag{10}$$

Let $l_h = x_h - x_{h-1} \ge 0$ be the width of h^{th} stratumand $x_L - x_0 = d$ (say) be the range of the distribution. Then, the objective function in (10) can be written as a function of l_h alone. Thus stating the objective function as a function of l_h we may rewrite NLPP (10) as:

Minimize $\sum_{h=1}^{L}\phi_h(l_h)$,

subject to

$$\sum_{h=1}^{L} l_h = d,$$

and

$$l_h \ge 0\ ;\ h = 1, 2, ..., L. \tag{11}$$

Solving (11) the OSW l_h^* and hence the OSB are obtained. In the following section, a general solution procedure for NLPP (11) is discussed.

III. The General Solution Procedure

The NLPP (11) is a multistage decision problem which allows us to use the dynamic programming technique [12]-[14]. Dynamic programming determines the optimum solution of a multi-variable problem by decomposing it into stages, each stage comprising a single variable sub-problem. A dynamic programming model is basically a recursive equation based on Bellman's principle of optimality [2]. This recursive equation links the different stages of the problem in a manner which guarantees that at each stage's optimal feasible solution is also optimal and feasible for the entire problem [27].

Consider the following sub-problem of (11) for first $k(<L)$ strata:

Minimize $\sum_{h=1}^{k}\phi_h(l_h)$,

subject to

$$\sum_{h=1}^{k} l_h = d_k,$$

and

$$l_h \ge 0\ ;\ h = 1, 2, ..., k, \tag{12}$$

where $d_k < d$ is the total width available for division into k strata or the state value at stage k. Note that $d_k = d$ for $k = L$.

Let $\Phi_k(d_k)$ denote the minimum value of the objective function of (12) and using the Bellman's principle of optimality, we write a forward recursive equation of the dynamic programming technique as:

$$\Phi_k(d_k) = \min_{0\le l_k\le d_k}\left[\phi_k(l_k) + \Phi_{k-1}(d_k - l_k)\right], k \ge 2. \tag{13}$$

For the first stage, that is, for $k = 1$:

$$\Phi_1(d_1) = \phi_1(d_1) \Rightarrow l_1^* = d_1, \quad (14)$$

where $l_1^* = d_1$ is the optimum width of the first stratum. The relations (13) and (14) are solved recursively for each $k = 1,2,...,L$ and $0 \le d_k \le d$, and $\Phi_L(d)$ is obtained. From $\Phi_L(d)$ the optimum width of L^{th} stratum, l_L^*, is obtained. From $\Phi_{L-1}(d - l_L^*)$ the optimum width of $(L-1)^{th}$ stratum, l_{L-1}^*, is obtained and so on until l_1^* is obtained.

IV. Determination of OSB of Uniform Auxiliary Variable

A. The Uniform Distribution

In probability theory and statistics, the uniform distribution is a family of continuous distributions and is frequently a probability model of many events of items that has equal probability of occurrence over a given range. Many continuous variables in the engineering, industry, management, and biological sciences have uniform probability distributions. For example, in a survey of telecom industry, the number of telephone calls coming into a switchboard that has a Poisson distribution is known exactly, the actual time of occurrence of one telephone call arrived at switchboard within one interval, say $(0,t)$ is distributed uniformly over this interval. Similarly, many other variables such as the delivery time of equipment in an interval, selecting a location to observe the work habit of workers in a certain assembly line, etc. are uniformly distributed [29].

The general formula for the probability density function (p.d.f) of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a}; & a \le x \le b \\ 0; & \text{otherwise} \end{cases} \quad (15)$$

where a is the location parameter and $(b-a)$ is the scale parameter. For the case, where $a = 0$ and $b = 1$, (15) is called the standard uniform distribution.

B. Formulation of NLPP for Uniform Auxiliary Variable

Let the auxiliary variable x follow Uniform Distribution with the p.d.f given in (15). By using (2), (3), (4) and (15), the terms W_h and $\sigma_{x_h}^2$ can be expressed as

$$W_h = \frac{l_h}{b-a} \quad (16)$$

and

$$\sigma_{x_h}^2 = \frac{l_h^2}{12}. \quad (17)$$

Using (16) and (17) the NLPP (11) could be expressed as:
Minimize

$$\sum_{h=1}^{L} \frac{l_h}{b-a} \sqrt{\beta^2 . \frac{l_h^2}{12} + \sigma_{e_h}^2}$$

subject to

$$\sum_{h=1}^{L} l_h = d,$$

and

$$l_h \ge 0; h = 1,2,...,L, \quad (18)$$

where β is the regression coefficient and $\sigma_{e_h}^2$ is the variance of the error function given in (6) for the error term in the regression model (1).

In the regression model given in (1), it is assumed that the variance of the error term is $V(e|x) = \phi(x)$ for all x in the range (a,b) and the expected value of the function $\phi(x)$ given by $\sigma_{e_h}^2$ is obtained by (6).Many authors have assumed that $\phi(x)$ may be of the form:

$$\phi(x) = cx^g; \quad c > 0, \quad g \ge 0, \quad (19)$$

where c and g are constants and in many populations $0 \le g \le 2$ [10], [19], [22] and [24].

Thus, from (15), (16) and (19), we may compute $\sigma_{e_h}^2$ as a function of boundary points as follows:

$$\sigma_{e_h}^2 = \frac{c.l_h^{g+1}}{l_h(b-a)(g+1)}. \quad (20)$$

Therefore, one can determine the expected value of the stratum variance of the error term using (20), if the values of the constants c and g are known.

Thus using (20), the NLPP (18) can be written as:
Minimize

$$\sum_{h=1}^{L} \frac{l_h}{b-a} \sqrt{\frac{\beta^2 l_h^2 (b-a)(g+1) + 12c}{12(b-a)(g+1)}}$$

subject to

$$\sum_{h=1}^{L} l_h = d,$$

and

$$l_h \ge 0; h = 1,2,...,L \quad (21)$$

C. Numerical Illustration of the Solution Procedure

This subsection illustrates the computational details of the solution procedure discussed in Section III using a dynamic programming technique for determining the OSB with uniform distribution.

To illustrate the computational procedure we take $a=1,\ b=2,\ \beta=1.2$, $c=1$, $g=0$ and $d=1$. Then, the NLPP (21) is reduced to:
Minimize

$$\sum_{h=1}^{L}\frac{l_h\sqrt{(1.2)^2 l_h^2+12}}{2\sqrt{3}}$$

subject to

$$\sum_{h=1}^{L} l_h = 1,$$

and

$$l_h \geq 0; h=1,2,...,L \qquad (22)$$

Note that the $(h-1)$ th stratification point is given by

$$x_{h-1} = x_0 + l_1 + l_2 + \cdots + l_{h-1} = d_h - l_h.$$

Substituting this value of x_{h-1}, the recurrence relation (13) and (14) are reduced as:
For the first stage, that is, $k=1$:

$$\Phi_1(d_1) = \frac{d_1\sqrt{(1.2)^2 d_1^2+12}}{2\sqrt{3}} \quad \text{at } l_1^* = d_1. \qquad (23)$$

For the stages $k \geq 2$:

$$\Phi_k(d_k) = \min_{0\leq l_k \leq d_k}\left[\frac{l_k\sqrt{(1.2)^2 l_k^2+12}}{2\sqrt{3}} + \Phi_{k-1}(d_k - l_k)\right] \qquad (24)$$

Solving the recurrence relations (23) and (24), for which a C++ program is coded, the NLPP (22) is solved. Executing the computer program, the optimum strata width l_h^* and hence the optimum strata boundaries $x_h^* = x_{h-1}^* + l_h^*$ are obtained. The results are presented in Table I for five different number of strata, that is, $L=2,3,4,5$ and 6.

V. Conclusion

Often, the surveyors encounter some difficulties prior to drawing the sample while sing the stratified sampling. One of such problems is how they construct the optimum strata within which the units are homogeneous as much as possible. In this paper, we address this problem and proposed a technique that can be used to estimate parameters more accurately.

While dealing with the problem constructing OSB it can be noted that the optimum stratification based on the study variable is not feasible in practice since it is unknown prior to conducting the survey. However, many a time the study variable (y) is closely related to an auxiliary variables(x) and data on x are either readily available or can easily be collected. In such situations, it is customary to consider the estimators of y that use the data on x and are more efficient than the estimators which use data on the variable y alone. Thus, the technique proposed in this paper uses auxiliary information for determining the optimum strata boundaries of the population that has uniformly distributed auxiliary variable. The problem is formulated as Nonlinear Programming Problems (NLPP)that seek minimization of the variance of the estimated population parameter under Neyman allocation. The NLPP is then solved by developing a solution procedure using a dynamic programming technique.

Numerical example is also presented to illustrate the application and also the computational details of the proposed technique.

TABLE I
OSW, OSB and Optimum Value of the Objective Function

L	l_h^*	x_h^*	$\sum_{h=1}^{L}\phi_h(l_h)$
2	$l_1^*=0.500$ $l_2^*=0.500$	$x_1^*=1.500$	1.015
3	$l_1^*=0.333$ $l_2^*=0.333$ $l_3^*=0.333$	$x_1^*=1.333$ $x_2^*=1.666$	1.007
4	$l_1^*=0.250$ $l_2^*=0.250$ $l_3^*=0.250$ $l_4^*=0.250$	$x_1^*=1.250$ $x_2^*=1.500$ $x_3^*=1.750$	1.004
5	$l_1^*=0.200$ $l_2^*=0.200$ $l_3^*=0.200$ $l_4^*=0.200$ $l_5^*=0.200$	$x_1^*=1.200$ $x_2^*=1.400$ $x_3^*=1.600$ $x_4^*=1.800$	1.002
6	$l_1^*=0.167$ $l_2^*=0.167$ $l_3^*=0.167$ $l_4^*=0.167$ $l_5^*=0.167$ $l_6^*=0.167$	$x_1^*=1.167$ $x_2^*=1.333$ $x_3^*=1.500$ $x_4^*=1.667$ $x_5^*=1.833$	1.001

References

[1] Aoyama, H. (1954). A Study of Stratified Random Sampling. *Ann. Inst. Stat. Math.*, 6, 1-36.
[2] Bellman, R.E. (1957). Dynamic Programming .Princetown University Press, New Jersey.
[3] Dalenius, T. (1950). The Problem of Optimum Stratification-II. *Skand. Aktuartidskr*, 33, 203-213.
[4] Dalenius, T. (1957). *Sampling in Sweden.* Almqvist & Wiksell, Stockholm.
[5] Dalenius, T. and Gurney, M. (1951). *The Problem of Optimum stratification-II*, Skand.Akt., 34, 133-148.

[6] Dalenius, T. and Hodges, J. L. (1959): Minimum Variance Stratification. *J. Amer. Statist. Assoc.* 54, 88-101.
[7] Durbin, J. (1959): Review of Sampling in Sweden. *J. Roy. Statist. Soc.* (A) 122, 146-148.
[8] Gupta, R. K., Singh, R. and Mahajan, P. K. (2005). Approximate Opimumum Strata Boundaries for Ratio and Regression Estimators. *Aligarh Journal of Statistics*, 25, 49-55.
[9] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sample Survey Methods and Theory*. Vol. I & II, John Wiley and Sons, Inc., New York.
[10] Khan, M. G. M., Ahmad, N. and Khan, Sabiha (2009). Determining the Optimum Stratum Boundaries using Mathematical Programming. *Journal of Mathematical Modelling and Algorithms*", Springer, Netherland, DOI 10.1007/s10852-009-9115-3, 8(4), 409-423.
[11] Khan, E. A., Khan, M. G. M. and Ahsan, M. J. (2002). *Optimum Stratification: A Mathematical Programming Approach*, Calcutta Statistical Association Bulletin, 52 (special Volume), 323-333.
[12] Khan, M. G. M., Najmussehar and Ahsan, M. J. (2005). Optimum Stratification for Exponential Study Variable under Neyman Allocation. *Journal of Indian Society of Agricultural Statistics*, 59(2), 146-150.
[13] Khan, M. G. M., Nand, N. and Ahmad, N. (2008). Determining the Optimum Strata Boundary Points Using Dynamic Programming. *Survey Methodology*, 34(2), 205-214.
[14] Khan, M.G.M.; Rao, D.; Ansari, A.H. and Ahsan, M.J. (2013). Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution. Journal of Communications in Statistics - Simulation and Computation. (To appear).
[15] Mahalanobis, P. C. (1952). Some Aspects of the Design of Sample Surveys. *Sankhya*, 12, 1-7.
[16] Mehta, S. K., Singh, R. and Kishore, L. (1996). On Optimum Stratification for Allocation Proportional to Strata Totals. *Journal of Indian Statistical Association*, 34, 9-19.
[17] Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.
[18] Neyman, J. (1934). On the Two Different Aspects of the Representatives Methods: t*he Method Stratified Sampling and the Method of Purposive Selection.* J. Roy. Stat. Soc. 97, 558-606.
[19] Rizvi, S. E. H., Gupta, J. P. and Bhargava, M. (2002). Optimum Stratification based on Auxiliary Variable for Compromise Allocation. Metron, 28(1), 201-215.
[20] Sethi, V. K. (1963). A Note on Optimum Stratification of Population for Estimating the Population Mean. *Aust. J. Statist.*, 5, 20-33.
[21] Singh, R. and Parkash, D. (1975). Optimum Stratification for Equal Allocation. *Annals of the Institute of Statistical Mathematics*, 27, 273-280.
[22] Singh, R. (1971). Approximately Optimum Stratification on the Auxiliary Variable. J. Amer. Stat. Assc., 66, 829-833.
[23] Singh, R. (1975). An Alternate Method of Stratification on the Auxiliary Variable. Sankhya. C, 37, 100-108.
[24] Singh, R. and Sukhatme, B. V. (1969). Optimum Stratification for Equal Allocation. Ann. Inst. Stat. Math., 27, 273-280.
[25] Singh, R. and Sukhatme, B. V. (1972). Optimum Stratification in Sampling with Varying Probabilities. Ann. Inst. Stat. Math., 24, 485-494.
[26] Singh, R. and Sukhatme, B. V. (1973). Optimum Stratification with Ratio and Regression Methods of Estimation. *Annals of the Institute of Statistical Mathematics*, 25, 627-633.
[27] Taha, H. A. (2007), Operations Research: An Introduction, 8th edition, Pearson Education, Inc., New Jersey.
[28] Taga, Y. (1967). On Optimum Stratification for the Objective Variable Based on Concomitant Variables using Prior Information. *Annals of the Institute of Statistical Mathematics*, 19, 101-129.
[29] Wackerly, D.W., Mendenhall, W. and Scheaffer, R. (2008). *Mathematical Statistics with Applications* (8thEddition), Thomson Learning, Inc., USA.