World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:8, No:3, 2014

Optimum Stratification of a Skewed Population

D.K. Rao, M.G.M. Khan, K.G. Reddy

Open Science Index, Mathematical and Computational Sciences Vol:8, No:3, 2014 publications.waset.org/9997637.pdf

Abstract—The focus of this paper is to develop a technique of solving a combined problem of determining Optimum Strata Boundaries (OSB) and Optimum Sample Size (OSS) of each stratum, when the population under study is skewed and the study variable has a Pareto frequency distribution. The problem of determining the OSB is formulated as a Mathematical Programming Problem (MPP) which is then solved by dynamic programming technique. A numerical example is presented to illustrate the computational details of the proposed method. The proposed technique is useful to obtain OSB and OSS for a Pareto type skewed population, which minimizes the variance of the estimate of population mean.

Keywords—Stratified sampling, Optimum strata boundaries, Optimum sample size, Pareto distribution, Mathematical programming problem, Dynamic programming technique.

I. Introduction

STRATIFIED sampling is used in sample surveys to achieve maximum precision in the estimates and it needs the solution of two basic problems that are the determination of the optimum strata boundaries (OSB) and optimum sample sizes (OSS) within each stratum, assumming that the number of strata and the total sample size are predetermined. The basic principle involved in the formation of strata is that they should be internally as homogenous as possible that is the stratum variances should be as minimum as possible, given a sample allocation. When the study variable itself is the stratification variable and its distribution is known, the OSB could be obtained by cutting the range of the distribution at suitable points. Several techniques have been proposed by many authors including [3], [4], [5], [6], [7], [13], [14], [15], [16], [17] and [18] for choosing the best strata boundaries.

There is another stratification method that exists in literature that is formulating the problem of determining OSB as an optimization problem and solving the problem using dynamic programming. The technique is useful when the frequency function of the study variable is known or can be estimated from past study. A brief review of this method can also be found in [9], [10], [11] and [12].

In this paper, the problem of finding OSB is redefined into the problem of determining Optimum Strata Width (OSW). The problem is formulated as a Mathematical Programming Problem (MPP), which minimizes the variance of the estimated population mean under Neyman allocation, subject to the constraint that the sum of the OSW be equal to the range of the distribution. The distribution of the stratification variable is considered to be continuous with Pareto distribution as in practice many populations are approximately Pareto distributed. The formulated MPP turns out to be multistage decision problem and therefore a technique using dynamic programming approach is developed to determine the OSB and OSW for each stratum. Section II provides the general formulation of the problem of finding OSW as an MPP and the solution procedure to solve the MPP is discussed in Section III. Section IV devotes the formulation of MPP and the determination of OSB for Pareto study variable.

II. Formulation of the Problem of OSW as an MPP

Let a population be divided into L non-overlapping strata and $f(x)$ denotes the probability density function of the study variable $x \in [x_0, x_L]$, where x_0 and x_L are the smallest and largest values of x. Then the problem of constructing L strata is to cut up the range of the distribution, $x_L - x_0 = d$ at intermediate points $x_1 \leq x_2 \leq, ..., \leq x_{L-1}$ such that the variance of the stratified sample mean that is $V(\overline{x}_{st}) = \sum_{h=1}^{L} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 \sigma_h^2$ is minimum.

However, ignoring the finite population correction (f.p.c.) and using neyman allocation that is $n_h = n.\frac{W_h \sigma_h}{\sum W_h \sigma_h}$, the minimization of $V(\overline{x}_{st})$ is equivalent to minimizing

$$\sum_{h=1}^{L} W_h \sigma_h. \tag{1}$$

When $f(x)$ is known and integrable, the values of W_h and σ_h in (1) can be obtained by

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

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)dx - \mu_h^2, \tag{3}$$

$$\text{where} \quad \mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x)dx \tag{4}$$

is the mean and (x_{h-1}, x_h) are the boundaries of h^{th} stratum.

Since W_h and σ_h^2 are a function of the boundary points x_{h-1} and x_h, let $\phi_h(x_{h-1}, x_h) = W_h \sigma_h$. Thus, the optimization problem to find x_1, x_2, ..., x_{L-1} can be written as:

$$\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{L} \phi_h(x_{h-1}, x_h), \\ &\text{subject to} \quad x_0 \leq x_1 \leq x_2 \leq, ..., \leq x_{L-1} \leq x_L. \end{aligned} \tag{5}$$

Using $l_h = x_h - x_{h-1} \geq 0$ to be the range or width of the h^{th} stratum, the objective function in (5) can be written as a function of x_{h-1} and l_h that is $\phi_h(l_h, x_{h-1})$. Initially, x_0 is known, therefore, the first term, that is, $\phi_1(l_1, x_0)$ is a function of l_1 alone. Once l_1 is known, the next stratification point x_1 = $x_0 + l_1$ will be known and the second term in the objective function $\phi_2(l_2, x_1)$ will become a function of l_2 alone. Thus, stating the objective function as a function of l_h alone, we may rewrite the MPP (5) as:

The authors are with the University of the South Pacific, Suva, Fiji.
(Corresponding author: email: dinesh.i.rao@usp.ac.fj, phone: +679 323 2603; fax: +679 323 1527).

$$\begin{array}{ll} \text{Minimize} & \sum_{h=1}^{L} \phi_h(l_h), \\ \text{subject to} & \sum_{h=1}^{L} l_h = d, \\ \text{and} & l_h \geq 0; \;\; h = 1, 2, ..., L. \end{array} \tag{6}$$

III. The Solution Procedure Using Dynamic Programming Technique

The MPP (6) is a multistage decision problem in which the objective function and the constraints are separable functions of l_h, which allow us to use a dynamic programming technique. A solution procedure using such a dynamic programming technique discussed in [11], which is summarized below:

Consider a subproblem of (6) of first $k(< L)$ strata, that is:

$$\begin{array}{ll} \text{Minimize} & \sum_{h=1}^{k} \phi_h(l_h), \\ \text{subject to} & \sum_{h=1}^{k} l_h = d_k, \\ \text{and} & l_h \geq 0; \;\; h = 1, 2, ..., k, \end{array} \tag{7}$$

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 $k = L$.

Using [2], we get the recursive relation of dynamic programming technique as:

$$\Phi_k(d_k) = \begin{array}{c} \min \\ 0 \leq l_k \leq d_k \end{array} \left[\phi_k(l_k) + \Phi_{k-1}(d_k - l_k)\right], \;\; k \;\geq\; 2. \tag{8}$$

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

$$\Phi_1(d_1) = \phi_1(d_1) \implies l_1^* = d_1, \tag{9}$$

where $l_1^* = d_1$ is the optimum width of the first stratum. The relations (8) and (9) are solved recursively for each $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$, 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. The details of the solution procedure can be seen in [11].

IV. The OSB for Skewed Population with Pareto Study Variable

When the study variable has a Pareto distribution, the formulation of the problem of determining OSW is expressed as an MPP and the MPP is solved using the dynamic programming technique via a numerical example.

A. The Pareto Distribution

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a skewed, heavy-tailed distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena. Outside the field of economics it is at times referred to as the Bradford distribution.

Vilfredo originally used this distribution to describe the allocation of wealth among individuals since it seemed to show that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society. This idea is sometimes expressed more simply as the Pareto principle or the 80-20 rule which says that 20% of the population controls 80% of the wealth.

If the study variable X in a survey, which is used to stratify the population, has a Pareto distribution then its probability density function is given by

$$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}; \qquad x \in [\beta, \infty), \tag{10}$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter.

B. Formulation of MPP for Pareto Study Variable

Using the definitions (2), (3), (4) and (10), the terms W_h and σ_h^2 can be expressed as

$$W_h = \frac{\beta^\alpha \left[(l_h + x_{h-1})^\alpha - x_{h-1}^\alpha\right]}{\left[x_{h-1}^\alpha (l_h + x_{h-1})^\alpha\right]}, \tag{11}$$

and

$$\begin{aligned} \sigma_h^2 \;\; = \;\; & \alpha\beta^{2\alpha} \left[2x_{h-1}^2 + 2x_{h-1}l_h - x_{h-1}^{2-\alpha}\left(l_h + x_{h-1}\right)^\alpha \right. \\ & \left. - x_{h-1}^\alpha \left(l_h + x_{h-1}\right)^{2-\alpha} + (1-\alpha)^2 l_h^2\right] / \\ & \left[(1-\alpha)^2 (2-\alpha)\, x_{h-1}^\alpha \left(l_h + x_{h-1}\right)^\alpha W_h^2\right]. \end{aligned} \tag{12}$$

Using (11) and (12), the MPP (6) may be expressed as:

$$\begin{array}{ll} \text{Minimize} & \sum_{h=1}^{L} \left\{Sqrt\left\{\alpha\beta^{2\alpha}\left[2x_{h-1}^2 + 2x_{h-1}l_h \right.\right.\right. \\ & -x_{h-1}^{2-\alpha}\left(l_h + x_{h-1}\right)^\alpha - x_{h-1}^\alpha\left(l_h + x_{h-1}\right)^{2-\alpha} \\ & \left. + (1-\alpha)^2 l_h^2\right] / \left[(1-\alpha)^2 (2-\alpha)\, x_{h-1}^\alpha \right. \\ & \left.\left.\left. (l_h + x_{h-1})^\alpha\right]\right\}\right\}, \\ \text{subject to} & \sum_{h=1}^{L} l_h = d, \\ \text{and} & l_h \geq 0; \;\; h = 1, 2, ..., L. \end{array} \tag{13}$$

C. Numerical Illustration

In this section the computational details of the solution procedure developed in Section III for the MPP (13) is presented.

Assume that x follows the Pareto distribution in the interval [1.000527, 28.147120], that is, $x_0 = 1.000527$, $x_L = 28.147120$. Also assume that $\alpha = 1.472$. This implies that

$\beta = x_0 = 1.000527$ and $d = x_L - x_0 = 27.146593$. Then the MPP (13) is expressed as:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{h=1}^{L} \Big\{ Sqrt \Big\{ 1.474285 \Big[2x_{h-1}^2 + 2x_{h-1} l_h \\
& -x_{h-1}^{0.528} \left(l_h + x_{h-1} \right)^{1.472} - x_{h-1}^{1.472} \\
& \left(l_h + x_{h-1} \right)^{0.528} + 0.222784 l_h^2 \Big] / \\
& \Big[0.117630 x_{h-1}^{1.472} \left(l_h + x_{h-1} \right)^{1.472} \Big] \Big\} \Big\}, \\
\text{subject to} \quad & \sum_{h=1}^{L} l_h = 27.146593, \\
\text{and} \quad & l_h \geq 0; \quad h = 1, 2, ..., L. \qquad (14)
\end{aligned}$$

Also

$$\begin{aligned}
x_{k-1} &= x_0 + l_1 + l_2 + ... + l_{k-1} \\
&= 1.000527 + l_1 + l_2 + ... + l_{k-1} \\
&= d_{k-1} + 1.000527 \\
&= d_k - l_k + 1.000527.
\end{aligned}$$

Substituting this value of x_{k-1} in (14) and using (8) and (9), the recurrence relations for solving MPP (14) are obtained as:
For first stage $(k = 1)$:

$$\begin{aligned}
\Phi_1(d_1) \quad = \quad & Sqrt \left\{ 1.474285 \left[2.002109 + 2.001054 d_1 \right. \right. \\
& -1.000278 \left(d_1 + 1.000527 \right)^{1.472} - 1.000776 \\
& \left. \left(d_1 + 1.000527 \right)^{0.528} + 0.222784 d_1^2 \right] / \\
& \left. \left[0.117721 \left(d_1 + 1.000527 \right)^{1.472} \right] \right\} \qquad (15)
\end{aligned}$$

at $l_1 = d_1$,
and for the stages $k \geq 2$:

$$\begin{aligned}
\Phi_k(d_k) \quad = \quad & \min_{0 \leq l_k \leq d_k} \left\{ Sqrt \left\{ 1.474285 \left[2 \left(d_k - l_k \right. \right. \right. \right. \\
& \left. +1.000527 \right)^2 + 2 \left(d_k - l_k + 1.000527 \right) l_k \\
& - \left(d_k - l_k + 1.000527 \right)^{0.528} \left(d_k + 1.000527 \right)^{1.472} \\
& - \left(d_k - l_k + 1.000527 \right)^{1.472} \left(d_k + 1.000527 \right)^{0.528} \\
& \left. +0.222784 l_k^2 \right] / \left[0.11763 \left(d_k - l_k \right. \right. \\
& \left. \left. \left. \left. +1.000527 \right)^{1.472} \left(d_k + 1.000527 \right)^{1.472} \right] \right\} \right\}. \qquad (16)
\end{aligned}$$

Solving the recursive equations (15) and (16) by executing a computer program developed for the solution procedure described in Section III, the OSWs are obtained. The results of optimum strata widths l_h^* and hence the optimum strata boundaries $x_h^* = x_{h-1}^* + l_h^*$ along with the values of the objective function $\sum_{h=1}^{L} \phi_h(l_h)$ for $L = 2, 3, 4, 5$ and 6 are presented in Table I. The table also presents the sample sizes (n_h; $h = 1, 2, ..., L$) for a fixed total sample size $n = 100$.

TABLE I
OSW, OSB, OSS AND THE OPTIMUM VALUE OF OBJECTIVE FUNCTION

L	(l_h^*)	$(x_h^* = x_{h-1}^* + l_h^*)$	n_h	$\sum_{h=1}^{L} \phi_h(l_h)$
2	y_1^* = 2.98130 y_2^* = 24.16530	x_1^*=3.98183	46 54	1.185625
3	y_1^* = 1.36677 y_2^* = 4.53950 y_3^* = 21.24033	x_1^*=2.36730 x_2^*= 6.90680	30 32 38	0.771251
4	y_1^* = 0.87031 y_2^* = 2.01873 y_3^* = 5.48669 y_4^* = 18.77087	x_1^*=1.87084 x_2^*= 3.88957 x_3^*= 9.37626	23 21 22 34	0.573397
5	y_1^* = 0.63554 y_2^* = 1.21528 y_3^* =2.54477 y_4^* = 6.00821 y_5^* = 16.74280	x_1^*=1.63607 x_2^*=2.85135 x_3^*= 5.39612 x_4^*= 11.40433	17 17 17 22 27	0.456846
6	y_1^* =0.49973 y_2^* = 0.84691 y_3^* = 1.52219 y_4^* = 2.94432 y_5^* = 6.25791 y_6^* = 15.07554	x_1^*= 1.50026 x_2^*= 2.34717 x_3^*= 3.86936 x_4^*= 6.81368 x_5^*= 13.07159	14 14 13 18 12 29	0.379856

V. SUMMARY

This paper deals with the problem of determining optimum strata boundaries (OSB) and the sample allocation to strata for a skewed population with pareto distribution. The problem is formulated as an MPP, which is solved using a dynamic programming technique. A numerical example on determining OSB is presented to show the computational details and the applications of proposed technique.

The basic advantage of the proposed method over the classical stratification techniques available in literature is that it can determine OSB efficiently, when the density function of the population is known or approximately known from previous studies. Many other iterative methods are also available for determining strata boundaries but these iterative methods require approximate initial solutions. Also there is no guarantee that an iterative method will converge and give the global minimum variance in the absence of a suitably chosen initial solution [1], [8] and [11]. Whereas, the proposed method does not require any initial approximate solution.

More importantly, the proposed technique has a wide scope of application as compared to other methods. In practice, the complete dataset of the study variable is unknown, which diminishes the uses of many stratification techniques. In such a situation, only the proposed technique can be used as it requires only the values of parameters of the population which can easily be available from the past studies. Thus, we may conclude that the proposed method is relatively efficient and may be useful for determining the OSB for any skewed population.

REFERENCES

[1] Amini, A.A., Weymouth, T.E., and Jain, R.C. (1990). Using Dynamic Programming for Solving Variational Problems in Vision.*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9), 855-867.

[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., and Gurney, M. (1951). The problem of optimum stratification. *Skand. Aktuartidskr*, 34, 133-148.

[5] Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

[6] Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Annals of the Institute of Statistical Mathematics*, 30, 1131-1134.

[7] Gunning, P and Horgan J.M. (2004) A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30(2), 159-166.

[8] Hillier, F.S., and Lieberman, G.J. (2010). *Introduction to Operations Research*. McGraw-Hill, New York.

[9] Khan, E.A., Khan, M.G.M., and Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. *Culcutta Statistical Association Bulletin*, 52 (special), 205-208.

[10] 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.

[11] 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.

[12] 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*. DOI: 10.1080/03610918.2013.819917 (To appear).

[13] Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

[14] Lavalle, P. and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

[15] Lednicki, B. and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

[16] Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. *Sankhya*, 12, 1-7.

[17] Rivest, L.P. (2002). A generalization of Lavalle and Hidiroglou algorithm for stratification in business survey. *Survey Methodology*, 28, 191-198.

[18] Sethi, V.K. (1963). A note on optimum stratification of population for estimating the population mean. *Australian Journal of Statistics*, 5, 20-33.