Optimum Stratification of a Skewed Population

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

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, assuming 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 function of the study variable is known or can be estimated in (1) can be obtained by

\[ V(\bar{x}_d) = \sum_{h=1}^{L} \left( \frac{1}{n_h} - \frac{x}{n} \right) W_h^2 \sigma_h^2 \]  

where

\[ \mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx \]  

\[ \sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x) dx - \mu_h^2 \]  

The mean and \( \sigma_h \) function of \( x \) 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:

\[ \text{Minimize } \sum_{h=1}^{L} \phi_h(x_{h-1}, x_h), \]  

subject to \( x_0 \leq x_1 \leq x_2 \leq \ldots \leq x_{L-1} \leq x_L \).
Minimize \( \sum_{h=1}^{L} \phi_h(l_h), \)
subject to \( \sum_{h=1}^{L} l_h = d, \)
and \( l_h \geq 0; \ h = 1, 2, \ldots, L. \) \( (6) \)

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:

Minimize \( \sum_{h=1}^{k} \phi_h(l_h), \)
subject to \( \sum_{h=1}^{k} l_h = d_k, \)
and \( l_h \geq 0; \ h = 1, 2, \ldots, k. \) \( (7) \)

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:

\[ \Phi_k(d_k) = \min_{0 \leq l_k \leq d_k} [\phi_h(l_h) + \Phi_{k-1}(d_k - l_k)], \quad k \geq 2. \] \( (8) \)

For the first stage, that is, for \( k = 1: \)

\[ \Phi_1(d_1) = \phi_1(d_1) \implies l_1^* = d_1, \] \( (9) \)

where \( l_1^* = d_1 \) is the optimum width of the first stratum.

The relations (8) and (9) are solved recursively for each \( k = 1, 2, \ldots, L \) and \( 0 \leq d_k \leq d, \) and \( \Phi_L(d) \) is obtained. From \( \Phi_L(d) \) the optimum width of \( L^{th} \) stratum, \( l_1^* \), is obtained. From \( \Phi_{L-1}(d - l_1^*) \) the optimum width of \( (L - 1)^{th} \) stratum, \( l_2^* \), is obtained and so on until \( l_L^* \) 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

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

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 \( \sigma_h^2 \) can be expressed as

\[ W_h = \frac{\beta^\alpha}{l_1^*(h + x_{h-1}^{-\alpha})}, \] \( (11) \)

and

\[ \sigma_h^2 = \alpha \beta^2 \left[ 2x_{h-1}^{-\alpha} + 2x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-\alpha} \right. \]

\[ \left. \right. - x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-2\alpha} + (1 - \alpha) l_h^2 \right] / \left( 1 - \alpha \right) \left( 2 - \alpha \right) x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-\alpha} W_h^2. \] \( (12) \)

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

Minimize \( \sum_{h=1}^{L} \sqrt{\{\alpha \beta^2 \left[ 2x_{h-1}^{-\alpha} + 2x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-\alpha} \right. \}
\]

\[ \left. \right. - x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-2\alpha} + (1 - \alpha) l_h^2 \right] / \left( 1 - \alpha \right) \left( 2 - \alpha \right) x_{h-1}^{-\alpha} (l_h + x_{h-1})^{-\alpha} \}
\]

subject to \( \sum_{h=1}^{L} l_h = d, \)
and \( l_h \geq 0; \ h = 1, 2, \ldots, L. \) \( (13) \)

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
\[ \beta = x_0 = 1.000527 \text{ and } d = x_L - x_0 = 27.146593. \] Then the MPP \((13)\) is expressed as:

\[
\text{Minimize} \sum_{h=1}^{L} \left\{ \sqrt{1.474285 \left[ 2x_{h-1}^2 + 2x_{h-1}l_h \right] - x_{h-1}^{0.528} (l_h + x_{h-1})^{1.472} - x_{h-1}^{1.472}} \right\} + x_{h-1} + 1.000527 - d_k - l_k + 1.000527. \]

subject to \( \sum_{h=1}^{L} l_h = 27.146593, \) and \( l_h \geq 0; \ h = 1, 2, \ldots, L. \) \((14)\)

Also

\[
x_k - 1 = x_0 + l_1 + l_2 + \ldots + l_{k-1} = 1.000527 + l_1 + l_2 + \ldots + l_{k-1} = d_{k-1} + 1.000527 = d_k - l_k + 1.000527. \]

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 \((k = 1)\):

\[
\Phi_1(d_1) = \sqrt{1.474285 \left[ 2.002109 + 2.001054d_1 - 1.000527 \left( d_1 + 1.000527 \right)^{1.472} - 1.000776 \right]} + \sqrt{1.47284d_1} / \left[ 0.117721 (d_1 + 1.000527)^{1.472} \right], \]

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

\[
\Phi_k(d_k) = \min_{0 \leq l_k \leq d_k} \left\{ \sqrt{1.474285 \left[ 2(d_k - l_k + 1.000527)^{1.472} + 2d_k - l_k + 1.000527 \right]} + \sqrt{1.47284d_k^2} / \left[ 0.11763 (d_k - l_k + 1.000527)^{1.472} \right] \right\}, \]

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 \(n_h; h = 1, 2, \ldots, L\) for a fixed total sample size \(n = 100.\]

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


