Proceedings of the VII International Symposium on Optimization and Statistics & III National Conference on Statistical Inference, Sampling Techniques and Related Areas, Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, December 21 – 23, 2012

ON CONSTRUCTING OPTIMUM STRATA AND DETERMINING OPTIMUM ALLOCATION

M. G. M. Khan and Sushita Sharma

ABSTRACT

The problem of constructing optimum stratum boundaries (OSB) and the problem of determining sample allocation to different strata are well known in the sampling literature. To increase the efficiency in the estimates of population parameters these problems must be addressed by the sampler while using stratified sampling. There were several methods available to determine the OSB when the frequency distribution of the study (or its related) variable is known. Whereas, the problem of determining optimum allocation was addressed in the literature mostly as a separate problem assuming that the strata are already formed and the stratum variances are known. However, many of these attempts have been made with an unrealistic assumption that the frequency distribution and the stratum variances of the target variable are known prior to conducting the survey. Moreover, as both the problems are not addressed simultaneously, the OSB and the sample allocation so obtained may not be feasible or may be far from optimum.

In this paper, the problems of finding the OSB and the optimum allocation are discussed simultaneously when the population mean of the study variable y is of interest and its frequency distribution $f(y)$ or the frequency distribution $f(x)$ of its auxiliary variable x is available. The problem is formulated as a Nonlinear Programming Problem (NLPP) that seeks minimization of the variance of the estimated population parameter of the target variable, which is subjected to a fixed total sample size. The formulated NLPP is then solved by executing a program coded in a user's friendly software, LINGO. Two numerical examples, when the study variable or its auxiliary variable has respectively a uniform and a right-triangular distribution in the population, are presented to demonstrate the practical application of the proposed method and its computational details. The proposed technique can easily be applied to other frequency distributions.

1. INTRODUCTION

Stratified random sampling is the most commonly used sampling technique in sample surveys. It is used when the units vary considerably (heterogeneous) so that the population parameters (mean or total) are estimated with greater precision.

To gain the precision in the estimates, the use of stratified sampling in sample surveys needs the solution of the following three basic problems:

(1) The determination of the number of strata.

(2) The determination of the optimum strata boundaries.

(3) The determination of the optimum sample sizes in various strata.

Assuming that the number of strata in problem (1) is predetermined, the present paper deals with the problems (2) and (3), that is, the determination of optimum strata boundaries and the determination of the optimum sample sizes to be drawn from various strata in a survey.

It is known that stratified random sampling will be efficient if the strata are so constructed within which the units are internally homogeneous as much as possible with respect to the characteristics under study. In other words, the maximum precision can be achieved if the strata are formed in such a way that the stratum variances $\left(S_{hy}^2\right)$ are as small as possible for a given type of sample allocation. Then the problem of constructing such strata is known as *optimum stratum boundaries* (OSB), which can be achieved effectively when the frequency distribution of $\left(y\right)$ is known. This problem was first discussed by Dalenius in [1950], when a single characteristic is under study and the study variable (y) itself is used as stratification variable. He presented a set of minimal equations for finding the OSB. Unfortunately these equations could not usually be solved because of their implicit nature. Hence attempts have been made by several authors to obtain the strata boundaries including Dalenius and Gurney [1951], Mahalanobis [1952], Hansen, et al. [1953], Aoyama [1954], Dalenius and Hodges [1959], Ekman [1959], Sethi [1963], Unnithan [1978], Lavallée and Hidiroglou [1988], Hidiroglou and Srinath [1993], Sweet and Sigman [1995], Hedlin [2000], Rivest [2002], Lednicki and Wieczorkowski [2003], Gunning and Horgan [2004], Kozak [2004], Keskintürk and Er [2007], etc. They used the frequency distribution of the main study variable and proposed different techniques 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 of OSB under certain assumptions.

When the frequency distribution of the auxiliary variable x is known, many authors such as Dalenius [1957], Taga [1967], Singh and Sukhatme [1969, 1972, 1973], Singh and Prakash [1975], Singh [1971, 1975], Mehta et al. [1996], Rizvi et al. [2002], and Gupta et al. [2005] have suggested different approximation method of determining OSB.

Another kind of stratification method that has been proposed in the literature is due to Bühler and Deutler [1975] and Khan, et al. [2002, 2005, 2008, 2009]. They formulated the problem of determining OSB as an optimization problem and developed a computational technique to solve the problem by using dynamic programming. Considering the problem as an equivalent problem of determining optimum strata width, Khan, et al. [2002, 2005, 2008, 2009] formulated the problem as a mathematical programming problem (MPP) for the populations with different frequency distributions and developed solution procedures using dynamic programming techniques. 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.

The main problem for many of the techniques discussed above is that the authors ignored the sample allocation problem while constructing the OSB, assuming that the proportional, Neyman or optimum allocation will be used. Such OSB may not always be feasible and thus optimal, especially when the populations are small and/or skewed. The final stratification should be considered a combination of stratum boundaries as well as stratum sample sizes.

In this present paper, both the problems of determining OSB and sample allocation are considered simultaneously. The problems are formulated as a nonlinear programming problem that minimizes the variance of the estimate which is subjected to a fixed total sample size.

2. FORMULATION OF THE PROBLEM AS AN NLPP

2.1 Problem of Determining Optimum Sample Sizes

Let the population be stratified into L strata based on a study variable y and the estimate of its mean $\overline{Y}$ is of interest. If n_h be a simple random sample drawn independently from different strata and $\overline{y}_h$ denotes an unbiased sample estimate of $\overline{Y}_h$ for the characteristic in stratum h, then an unbiased estimate of the population mean $\overline{Y}$ is given by

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h, \tag{1}$$

where, $W_h = \frac{N_h}{N}$, N_h = number in stratum h, and $N = \sum_{h=1}^{L} N_h$.

The variance of the estimate $\overline{y}_{st}$ is given by

$$\text{var}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N}. \tag{2}$$

Where, S_{hy}^2 denotes the stratum variance of the characteristic y in the $h-$th stratum.

If the total sample size n for a stratified survey is predetermined, a reasonable criterion for obtaining the optimum allocation n_h is to minimize the variance, $\text{var}(\overline{y}_{st})$, of the stratified sample means of the characteristic given in (2).

Further, for practical application of any allocation the integer values of the sample sizes are required. They could be obtained by simply rounding off non-integer sample sizes to their nearest integral values. However, in many situations the rounded-off sample allocation may be infeasible or sub-optimal.

Thus, the problem of finding the allocation $(n_1,...,n_L)$ for a fixed sample size n may be given as the following all integers NLPP:

$$\text{Minimize } \operatorname{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N}$$

$$\text{subject to } \quad \sum_{h=1}^{L} n_h = n;$$

$$1 \le n_h \le N_h \text{ and } n_h \text{ are integers}, (h = 1,2,...,L). \tag{3}$$

The restrictions $n_h \le N_h$ are imposed to avoid oversampling and the restrictions $1 \le n_h$ are imposed so that all the L strata with at least one unit are formed.

2.2 Problem of Determining OSB

Let the population be stratified into L strata based on the study variable y and $f(y)$ denotes frequency function of y. If y_0 and y_L be the smallest and largest values of y, then a problem of determining the strata boundaries is to cut up the range,

$$y_L - y_0 = d \text{ (say)}, \tag{4}$$

at intermediate points $y_1 \le y_2 \le,...,\le y_{L-1}$ such that the variance of the stratified sample mean, $\operatorname{var}(\bar{y}_{st})$, given in (2) is minimum.

Thus, the problem of determining the OSB can be stated as:

$$\text{Minimize } \operatorname{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N}$$

$$\text{subject to } y_0 \le y_1 \le y_2 \le,...,\le y_{L-1} \le y_L. \tag{5}$$

It can be noted that, if the target population is small, one cannot expect the sampling fraction n_h/N_h is negligible and hence the finite population correction factor in the objective function of (5) cannot be ignored. Also, when frequency function of the study variable $f(y)$ is known, the values of W_h and S_{hy}^2 can be expressed as a function of boundary points (y_{h-1}, y_h) of $h-$th stratum by

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy \tag{6}$$

$$S_{hy}^2 = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y^2 f(y)dy - \mu_{hy}^2 \tag{7}$$

Where, $$\mu_{hy}=\frac{1}{W_h}\int\limits_{y_{h-1}}^{y_h} yf(y)dy \tag{8}$$

Let $d_h = y_h - y_{h-1} \geq 0$ denotes the width of the $h-$th ($h=1,2,...,L$) stratum.

With the above definition, (4) is expressed as

$$\sum_{h=1}^{L} d_h = \sum_{h=1}^{L}\left(y_h - y_{h-1}\right) = y_L - y_0 = d$$

Thus, the $k-$th stratification point y_k ; $k=1,2,...,L-1$ can be expressed as:

$$\begin{aligned} y_k &= y_0 + d_1 + d_2 + ...+ d_k \\ &= y_{k-1} + d_k \end{aligned}$$

Then, the problem of determining OSB in (5) can also be considered as the problem of determining optimum strata widths and may be expressed as the following NLPP:

$$\left.\begin{array}{ll} \text{Minimize} & \text{var}(\overline{y}_{st}) = \sum\limits_{h=1}^{L}\frac{W_h^2 S_{hy}^2}{n_h} - \sum\limits_{h=1}^{L}\frac{W_h S_{hy}^2}{N} \\ \text{subject to} & \sum\limits_{h=1}^{L} d_h = d, \\ \text{and} & d_h \geq 0;\ h=1,2,...,L. \end{array}\right\} \tag{9}$$

2.3 Problem of Determining OSB and Sample Sizes:

Thus, merging (3) and (9), the problem of determining OSB and the optimum allocation of sample size is formulated as an NLPP as follows:

$$\left.\begin{array}{ll} \text{Minimize} & \text{var}(\overline{y}_{st}) = \sum\limits_{h=1}^{L}\frac{W_h^2 S_{hy}^2}{n_h} - \sum\limits_{h=1}^{L}\frac{W_h S_{hy}^2}{N} \\ \text{subject to} & \sum\limits_{h=1}^{L} d_h = d, \\ & \quad \sum\limits_{h=1}^{L} n_h = n, \\ \text{and} & d_h \geq 0;\ 1 \leq n_h \leq N_h;\ n_h \text{ integers, } h=1,2,...,L. \end{array}\right\} \tag{10}$$

2.4 Problem of Determining OSB and Sample Sizes using Auxiliary Variable

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. In Section 2.3, the problem of determining OSB and sample allocation is formulated based on the study variable (y) itself and its frequency distribution $f\left(y\right)$ is assumed to be known. However, this may not be feasible in practice since in many situations the study variable is unknown prior to conducting the survey, which leads to use of the distribution of closely related variable (x), called auxiliary variable. 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. By and large, auxiliary data are readily available or can be made available easily with minimum cost and effort.

Moreover, 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.

Consider the regression model:

$$y = \lambda\left(x\right) + \varepsilon\,, \tag{11}$$

where $\lambda\left(x\right)$ is a linear or nonlinear function of x and ε is an error term such that $E\left[\varepsilon|x\right]=0$ and $\mathrm{var}\left[\varepsilon|x\right]=\phi\left(x\right)$ for all x.

Under the model (11), the stratum mean μ_{hy} and the stratum variance S_{hy}^2 can be expressed as (see Singh and Sukhatme, 1969):

$$\mu_{hy} = \mu_{h\lambda} \tag{12}$$

and

$$S_{hy}^2 = S_{h\lambda}^2 + \mu_{h\phi} \tag{13}$$

where $\mu_{h\lambda}$ and $\mu_{h\phi}$ are the expected values of $\lambda\left(x\right)$ and $\phi\left(x\right)$ respectively, and $S_{h\lambda}^2$ denotes the variance of $\lambda\left(x\right)$ in the *h*-th stratum.

If λ and ε are uncorrelated, from the model (11), S_{hy}^2 can also be expressed as (see Dalenius and Gurney [1951]):

$$S_{hy}^2 = S_{h\lambda}^2 + S_{h\varepsilon}^2\,, \tag{14}$$

where $S_{h\varepsilon}^2$ is the variance of ε in the *h*th stratum. It is, therefore, minimizing (2) is equivalent to minimizing

$$\text{var}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2\left(S_{h\lambda}^2 + \mu_{h\phi}\right)}{n_h} - \sum_{h=1}^{L} \frac{W_h\left(S_{h\lambda}^2 + \mu_{h\phi}\right)}{N}. \tag{15}$$

Let $f(x);\ a \le x \le b$ be the frequency function of the auxiliary variable *x*, which is used for determining OSB by cutting its range $d = b - a$ at $(L-1)$ intermediate points $a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b$ such that (15) is minimum.

Thus, from (10), the OSB and the optimum allocation can be formulated as the following NLPP:

$$\left.\begin{array}{ll} \text{Minimize} & \text{var}(\overline{y}_{st}) = \displaystyle\sum_{h=1}^{L} \frac{W_h^2\left(S_{h\lambda}^2 + \mu_{h\phi}\right)}{n_h} - \sum_{h=1}^{L} \frac{W_h\left(S_{h\lambda}^2 + \mu_{h\phi}\right)}{N} \\ \text{subject to} & \displaystyle\sum_{h=1}^{L} d_h = d, \\ & \displaystyle\sum_{h=1}^{L} n_h = n \\ \text{and} & d_h \ge 0;\ 1 \le n_h \le N_h;\ n_h \text{ integers, } h = 1, 2, ..., L. \end{array}\right\} \tag{16}$$

For a known $f(x)$, the values of $S_{h\lambda}^2$ and $\mu_{h\phi}$ can be expressed as a function of boundary points (x_{h-1}, x_h) of $h-$th stratum by

$$S_{h\lambda}^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda(x)^2 f(x)dx - \mu_{h\lambda}^2 \tag{17}$$

and

$$\mu_{h\phi} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \phi(x) f(x)dx \tag{18}$$

Where,

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx \tag{19}$$

and

$$\mu_{h\lambda} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda(x) f(x) dx \tag{20}$$

3. THE SOLUTION PROCEDURE USING LINGO

When the number of strata (L) and the total sample size (n) are predetermined, the NLPP (10) and (16) may be solved by executing a program developed in the LINGO software package for a known $f(y)$ or $f(x)$.

4. NUMERICAL ILLUSTRATIONS

In order to demonstrate the computational details of the proposed technique, two sets of populations that follow respectively uniform and right-triangular distributions are considered.

4.1 Population 1: Uniform Distribution

The uniform distribution is a family of continuous probability distributions. It is frequently a probability model of many events of items that has equal probability of occurrence over a given range. The distribution is defined by the two parameters, a and b, which are its minimum and maximum values.

The general formula for the probability density function of the uniform distribution is

$$f(x) = \begin{cases} \dfrac{1}{b-a}; & a \le x \le b \\ 0; & \text{otherwise} \end{cases} \tag{21}$$

Some continuous variables in the engineering, industry, management, and biological sciences have uniform probability distributions. For example, in a survey of telecom industry, 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, the delivery time of equipment in an interval, or selecting a location to observe the work habit of workers in a certain assembly line, etc. are uniformly distributed (see Wackerly et al. [2008]).

Estimating the distribution of the population:

The data set of Population 1 of size $N=1000$ provides the information of the study variable y that has the minimum value y_0 = 0.000391 and the maximum value $y_L = 0.998604$.

Thus, the dataset gives the range of the distribution as $d = y_L - y_0 = 0.998213$. To determine the distribution, we construct a relative frequency histogram of y.

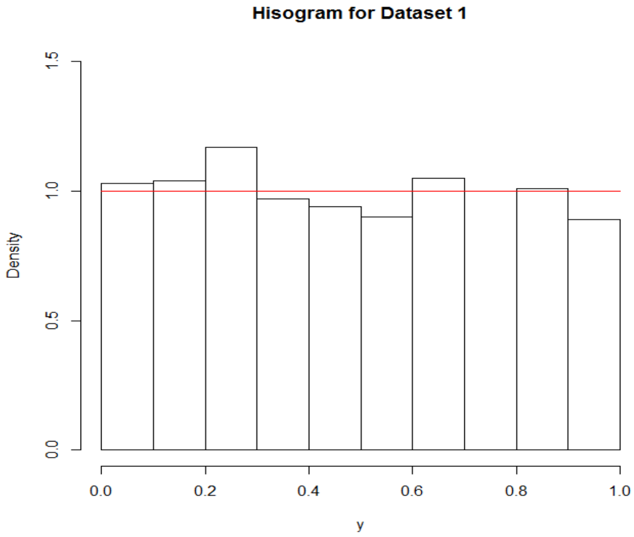


Figure 1: Frequency Histogram for *y*.

The Q-Q plot of *y* in Figure 2 and the Kolmogrov-Smirnov test (D=0.0291, p-value = 0.3653) also confirm that *y* has a uniform distribution.

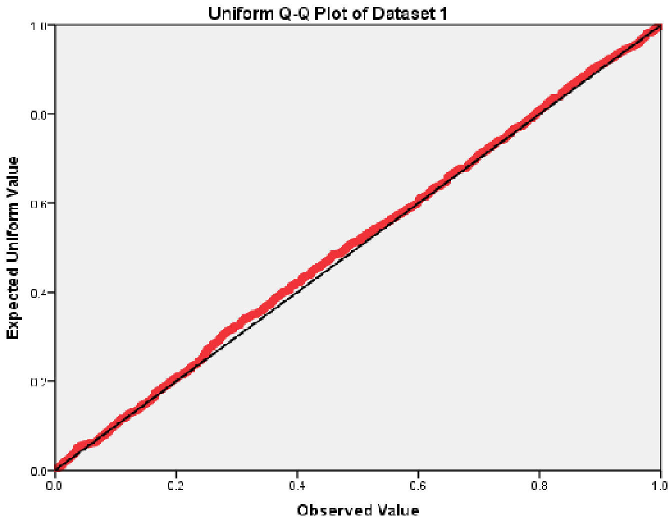


Figure 2: Q-Q Plot of *y*.

For testing whether a particular distribution, if needed, one can use some method of goodness of fit. Also there are some software (such as, EasyFit at mathwave.com: http://www.mathwave.com/easyfit-distribution-fitting.html), which allow to automatically or manually fit a large number of distributions including uniform distribution to the data.

Formulation of the Problem Determining OSB and Sample sizes as an NLPP:

As the study variable *y* has uniform distribution with density function given in (21), using (6), (8) and (7), we obtain:

$$W_h = \frac{d_h}{b-a}, \quad \text{where } d_h = y_h - y_{h-1} \qquad (22)$$

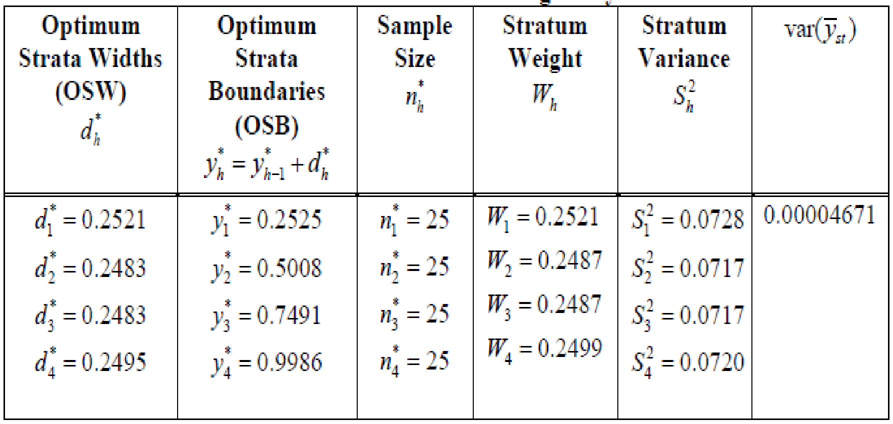
and

$$S_{yh}^2 = \frac{d_h^2}{12} \qquad (23)$$

Substituting (22) and (23), the NLPP (10) is solve for constructing a predetermined number of strata, say $L=$ 4, with a fixed total sample size $n=100$ for Population 1 using LINGO. For this population, $a = y_0$ = 0.000391, $b = y_L$ = 0.998604 and $d = y_L - y_0$ = 0.998213.

Then, the OSB $\left(y_h^*\right)$ and the optimum sample sizes $\left(n_h^*\right)$ using the proposed method as discussed earlier are obtained as shown in Table 1. The stratum width $\left(d_h^*\right)$, stratum weight $\left(W_h\right)$, stratum variance $\left(S_h^2\right)$ and the variance of the estimate, $\text{var}(\overline{y}_{st})$, are also presented in the table.

Table 1: Results for Uniform Distribution using Study Variable for n = 100



Optimum Strata Widths (OSW) d_h^*	Optimum Strata Boundaries (OSB) $y_h^* = y_{h-1}^* + d_h^*$	Sample Size n_h^*	Stratum Weight W_h	Stratum Variance S_h^2	$\text{var}(\overline{y}_{st})$
$d_1^* = 0.2521$	$y_1^* = 0.2525$	$n_1^* = 25$	$W_1 = 0.2521$	$S_1^2 = 0.0728$	0.00004671
$d_2^* = 0.2483$	$y_2^* = 0.5008$	$n_2^* = 25$	$W_2 = 0.2487$	$S_2^2 = 0.0717$	
$d_3^* = 0.2483$	$y_3^* = 0.7491$	$n_3^* = 25$	$W_3 = 0.2487$	$S_3^2 = 0.0717$	
$d_4^* = 0.2495$	$y_4^* = 0.9986$	$n_4^* = 25$	$W_4 = 0.2499$	$S_4^2 = 0.0720$	

Suppose that the information on the study variable is not available but its auxiliary variable *x*. For a sample data, it has been seen that *y* has a linear regression model with *x*, that is:

$$\lambda(x) = \alpha + \beta x \tag{24}$$

For this population, the auxiliary variable follows uniform distribution of the form (21) with

$a = x_0 = 0.002877$, $b = x_L = 1.999727$ and $d = x_L - x_0 = 1.998685$.

The estimated regression coefficients are $\hat{\alpha} = -0.026$ and $\hat{\beta} = 0.505$.

Using (17), (21) and (24) we obtain:

$$S^2_{h\lambda} = \frac{\beta_h^h d_h^2}{12}$$

Assumed that the regression model is common across the strata, that is,

$$\beta_h = \hat{\beta} = 0.505.$$

The expected stratum variance of the error is obtained as:

$$\mu_{h\phi} = \text{ MSE} = 0.000101,$$

Where MSE is the mean square of residuals for the regression model. Then, solving the MPP (16), the OSB of the auxiliary variable $\left(x_h^*\right)$ and hence the OSB of study variable $\left(y_h^*\right)$ along with the optimum sample sizes $\left(n_h^*\right)$ and variance of the estimate of study variable are obtained as shown in Table 2.

Table 2: Results for Uniform Distribution using Auxiliary Variable for *n* = 100

Optimum Strata Widths (OSW) d_h^*	OSB for x $x_h^* = x_{h-1}^* + d_h^*$	OSB for y $y_h^* = \hat{\alpha}_h + \hat{\beta}_h x_h^*$	Sample Size n_h^*	$\text{var}(\overline{y}_{st})$
$d_1^* = 0.4752$	$x_1^* = 0.4781$	$y_1^* = 0.2154$	$n_1^* = 22$	0.00004868
$d_2^* = 0.5104$	$x_2^* = 0.9885$	$y_2^* = 0.4732$	$n_2^* = 27$	
$d_3^* = 0.5103$	$x_3^* = 1.4988$	$y_3^* = 0.7309$	$n_3^* = 26$	
$d_4^* = 0.5009$	$x_4^* = 1.9997$	$y_4^* = 0.9838$	$n_4^* = 25$	

4.2 Population 2: Right-Triangular Distribution

The right-triangular distribution is a family of continuous probability distribution, which models many observable phenomena that shows the number of successes when the most likely success falls at the maximum and the least likely success falls at the minimum values. For example; less income earned by a larger portion of families in a society, whereas a very few families earns larger income.

The distribution is defined by two parameters a and b, which are its minimum and maximum values where respectively the most likely and the least likely number of items fall.

The general formula for the probability density function of a right-triangular distribution is given by

$$f(x)=\begin{cases}\dfrac{2(b-x)}{(b-a)^2}; & a\le x\le b\\ 0; & \text{otherwise.}\end{cases} \tag{25}$$

Estimating the distribution of the population:

The data set of Population 2 of size $N=800$ provides the information of the study variable y that follows a right-triangular distribution with the minimum value y_0 = 0.003716 and the maximum value y_L = 1.943307.

Thus, the dataset gives the range of the distribution as

$$d = y_L - y_0 = 1.939591\,.$$

Formulation of the Problem Determining OSB and Sample sizes as an MPP:

As the study variable y has right-triangular distribution with density function given in (25), using (6), (8) and (7), we obtain:

$$W_h=\frac{d_h\left(2a_h-d_h\right)}{\left(b-a\right)^2},$$

where

$$a_h=b-y_{h-1}=b-\left(a+\sum_{l=1}^{h-1}d_l\right) \tag{26}$$

and

$$S_{yh}^2=\frac{d_h^2\left(d_h^2-6a_hd_h+6a_h^2\right)}{18\left(2a_h-d_h\right)^2} \tag{27}$$

Substituting (26) and (27), the MPP (10) is solve for constructing $L=$ 4 with a fixed total sample size $n=150$ using LINGO.

Then, the OSB $\left(y_h^*\right)$ and the optimum sample sizes $\left(n_h^*\right)$ using the proposed method as discussed earlier are obtained as shown in Table 3. The stratum width

$\left(d_h^*\right)$, stratum weight $\left(W_h\right)$, stratum variance $\left(S_h^2\right)$ and the variance of the estimate, $\text{var}(\overline{y}_{st})$, are also presented in the table

Table 3: Results for Right-Triangular Distribution using Study Variable for n = 150

Optimum Strata Widths (OSW) d_h^*	Optimum Strata Boundaries (OSB) $y_h^* = y_{h-1}^* + d_h^*$	Sample Size n_h^*	Stratum Weight W_h	Stratum Variance S_h^2	$\text{var}(\overline{y}_{st})$
$d_1^* = 0.2807$	$y_1^* = 0.2844$	$n_1^* = 36$	$W_1 = 0.2685$	$S_1^2 = 0.1535$	0.000171
$d_2^* = 0.3237$	$y_2^* = 0.6081$	$n_2^* = 36$	$W_2 = 0.2576$	$S_2^2 = 0.1614$	
$d_3^* = 0.4028$	$y_3^* = 1.0109$	$n_3^* = 36$	$W_3 = 0.2428$	$S_3^2 = 0.1742$	
$d_4^* = 0.9324$	$y_4^* = 1.9433$	$n_4^* = 42$	$W_4 = 0.2311$	$S_4^2 = 0.2122$	

When the information on the study variable is not available, an auxiliary variable x is considered, which is found to be linearly regressed with y for a sample data, that is:

$$\lambda(x) = \alpha + \beta x$$

The estimated regression coefficients are found as: $\hat{\alpha}$ = -0.023 and $\hat{\beta}$ = 0.675.

Using (17), (25) and (24) we obtain:

$$S_{yh}^2 = \beta_h^2 \frac{d_h^2\left(d_h^2 - 6a_h d_h + 6a_h^2\right)}{18\left(2a_h - d_h\right)^2}$$

Assumed that the regression model is common across the strata, that is,

$$\beta_h = \hat{\beta} = 0.675.$$

The expected stratum variance of the error is obtained as:

$$\mu_{h\phi} = \text{ MSE} = 0.000157,$$

Where MSE is the mean square of residuals for the regression model. Then, solving the MPP (16), the OSB of the auxiliary variable $\left(x_h^*\right)$ and hence the OSB of study variable $\left(y_h^*\right)$ along with the optimum sample sizes $\left(n_h^*\right)$ and variance of the estimate of study variable are obtained as shown in Table 4.

Table 4: Results for Right-Triangular Distribution using Auxiliary Variable for $n = 150$

Optimum Strata Widths (OSW) d_h^*	OSB for x $x_h^* = x_{h-1}^* + d_h^*$	OSB for y $y_h^* = \hat{\alpha}_h + \hat{\beta}_h x_h^*$	Sample Size n_h^*	$\text{var}(\overline{y}_{st})$
$d_1^* = 0.4278$	$x_1^* = 0.4323$	$y_1^* = 0.2688$	$n_1^* = 36$	0.00002773
$d_2^* = 0.4935$	$x_2^* = 0.9258$	$y_2^* = 0.6019$	$n_2^* = 36$	
$d_3^* = 0.6139$	$x_3^* = 1.5397$	$y_3^* = 1.0163$	$n_3^* = 36$	
$d_4^* = 1.4208$	$x_4^* = 2.9605$	$y_4^* = 1.9753$	$n_4^* = 42$	

5. CONCLUSION

The problem of determining OSB and the allocation of sample size are discussed by many authors mostly either separately or they determined the OSB under a particular allocation. The OSB so obtained may be infeasible or sub-optimum, especially for small and skewed populations.

In this research, an attempt is made to solve both the problems simultaneously. The problems are formulated as an NLPP, which is solved by developing a program coded in a user friendly software LINGO.

RFERENCES

Aoyama, H. (1954): A Study of Stratified Random Sampling. *Ann. Inst. Stat. Math.*, **6**, 1-36.

Bühler, W. and Deutler, T. (1975): Optimum Stratification and Grouping by Dynamic Programming. *Metrika*, **22**, 161-175.

Dalenius, T. (1950): The problem of Optimum Stratification-II. *Skand. Aktuartidskr*, 33, 203-213.

Dalenius, T. (1957): *Sampling in Sweden*. Almqvist & Wiksell, Stockholm.

Dalenius, T. and Gurney, M. (1951): *The problem of optimum stratification-II, Skand. Akt.*, **34**, 133-148.

Dalenius, T. and Hodges, J. L. (1959): Minimum variance stratification. *J. Amer. Statist. Assoc.* **54**, 88-101.

Ekman, G. (1959): Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Ann. Inst. Stat. Math.*, **30**, 1131-1134.

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.

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.

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.

Hedlin, D. (2000): A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, **16**, 15-29

Hidiroglou, M.A. and Srinath, K.P. (1993): Problems associated with designing subannual business surveys. *Journal of Business & Economics Statistics*, **11**, 397- 405.

Keskintürk, T., Er, Ş. (2007): A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. *Computational Statistics & Data Analysis*, **52**(1), 53-67.

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.

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.

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.

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.

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

Lavallée, P. and Hidiroglou, M. (1988): On the Stratification of Skewed Populations. *Survey Methodology*, **14**, 33-43.

Lednicki, B., Wieczorkowski, R. (2003): Optimal Stratification and Sample Allocation between Subpopulations and Strata. *Statistics in Transition*, **6**, 287-306.

Mahalanobis, P. C. (1952): Some Aspects of the Design of Sample Surveys. *Sankhya*, **12**, 1-7.

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.

Neyman, J. (1934): On the Two Different Aspects of the Representatives Methods: the Method Stratified Sampling and the Method of Purposive Selection. J. *Roy. Stat. Soc.* **97**, 558-606.

Rivest, L.P. (2002): A Generalization of Lavallée and Hidiroglou algorithm for stratification in business survey. *Survey Methodology*, **28**, 191-198.

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.

Serfling, R. J. (1968): Approximately Optimum Stratification. *Journal of American Statistical Association*, **63**, 1298-1309.

Sethi, V. K. (1963): A Note on Optimum Stratification of Population for Estimating the Population Mean. *Aust. J. Statist.,* 5, 20-33.

Singh, R. and Parkash, D. (1975). Optimum Stratification for Equal Allocation. *Annals of the Institute of Statistical Mathematics*, **27**, 273-280.

Singh, R. (1971). Approximately Optimum Stratification on the Auxiliary Variable. *J. Amer. Stat. Assc.,* **66**, 829-833.

Singh, R. (1975): An Alternate Method of Stratification on the Auxiliary Variable. *Sankhya.* C, **37**, 100-108.

Singh, R. and Sukhatme, B. V. (1969): Optimum Stratification for Equal Allocation. *Ann. Inst. Stat. Math.*, **27**, 273-280.

Singh, R. and Sukhatme, B. V. (1972): Optimum Stratification in Sampling with Varying Probabilities. *Ann. Inst. Stat. Math.*, **24**, 485-494.

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.

Sweet, E.M., and Sigman, R.S. (1995): Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data, *Proceedings of the Survey Research Methods Section,* ASA, 491-496.

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.

Unnithan, V.K.G. (1978): The Minimum Variance Boundary Points of Stratification. *Sankhya*, **40**(C), 60-72.

Wackerly, D.W., Mendenhall, W. and Scheaffer, R. (2008): *Mathematical Statistics with Applications* (8th Eddition), Thomson Learning, Inc., USA.

M. G. M. Khan and Sushita Sharma

School of Computing, Information and
Mathematical Sciences
University of the South Pacific
Suva, Fiji.
Email: khan_mg@usp.ac.fj