Aligarh Journal of Statistics

Vol. 34 (2014), 85-104

OPTIMUM STRATIFICATION USING AUXILIARY VARIABLES

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

ABSTRACT

The problem of determining optimum strata boundaries (OSB), when the frequency distribution of survey (or main) information is known, is discussed by many authors and is available in sampling literature. However, many of these authors made an unrealistic assumption that the frequency distribution of study variable is known prior to conducting the survey. In this manuscript, we discuss the problems of determining the optimum stratifications, when the frequency distribution of auxiliary variable is known. If the stratification of survey variable is made using the auxiliary variable it may lead to substantial gains in precision of the estimates. Moreover, often the auxiliary information is easily available or can be made available with a minimum cost and effort. In this manuscript, the problems of constructing optimum stratification is discussed for two study variables based on the auxiliary variables that follow respectively a uniform and a right-triangular distribution. The problems of determining the OSB are formulated as Nonlinear Programming Problems $(NLPP)$, which turn out to be multistage decision problems and are solved using dynamic programming techniques.

1. 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 estimates using stratified sampling one of the basic problem is the determination of the optimum strata boundaries (OSB) 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 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 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. For example, the data on the size of land cultivated (x), which are highly correlated with the amount of crop (y), are readily available. Moreover, 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 much as possible 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 Dalenius (1950).

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 Dalenius and Gurney (1951), Mahalanobis (1952), Hansen, Hurwitz and Madow (1953), Aoyama (1954), Ekman (1959), Dalenius and Hodges (1959), Durbin (1959), Sethi (1963), Murthy (1967) 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.

Many authors such as Unnithan (1978), Lavallée and Hidiroglou (1988), Hidiroglou and Srinath (1993), Sweet and Sigman (1995) and Rivest (2002) suggested some iterative procedures to determine OSB. These algorithms require an initial approximate solution and also there is no guarantee that the algorithm will provide the global minimum. Moreover, the convergences of some of these algorithms are slow or non-existent (see Detlefsen and Veum 1991). Gunning and Horgan (2004) developed an approximate method of stratification for positively skewed populations. They showed that their algorithm is much easier and more efficient than the cum $\sqrt{f}$ method of Dalenius and Hodges (1959) and Lavallee-Hidiroglou (1988) method.

Niemiro (1999) proposed a random search method for optimum stratification but the algorithm did not guarantee that it leads to global optimum and also goes wrong in case of a large population, as it requires too many iterations. Lednicki

and Wieczorkowski (2003) presented a method of stratification based on Rivest (2002) using the simplex method of Nelder and Mead (1965) but the method was rather slow and may not provide the best solution in the case of large number of variables.

Later, Kozak (2004) presented the modified random search algorithm as a method of the optimum stratification, which was quite faster and efficient as compared to Rivest (2002), and Lednicki and Wieczorkowski (2003) but it could not guarantee that the algorithm leads to the global optimum (see Kozak 2004).

When the frequency distribution of an 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 Khan et al. (2002). 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 Khan et al. (2002) technique to deal with the problem of determining OSB for two populations using a single auxiliary variable that has Uniform and Right-Triangular frequency distributions, respectively.

2. FORMULATION OF THE PROBLEM AS AN NLPP

Let the population be stratified into L strata based on a single auxiliary variable x and the estimation of the population 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 $x_1, x_2, x_3, ..., x_{L-1}$. Then, for the $h-th$ stratum with 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

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

If x and ε are uncorrelated, from the model (1), the variance of y in $h-th$ stratum can be expressed as (see Dalenius and Gurney, 1951):

$$\sigma_{y_h}^2 = \beta^2 \sigma_{x_h}^2 + \sigma_{e_h}^2 \tag{5}$$

$$\sigma_{e_h}^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \phi(x) f(x)\, dx \tag{6}$$

where $\sigma_{e_h}^2$ is the expected variance of ε in the $h-th$ stratum.

Ignoring the finite population correction $(f.p.c.)$ and using (5) the variance of stratified mean $\bar{y}_{st}$ under Neyman allocation (Neyman, 1934) for a fixed total sample size n is given as:

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

Minimizing $V(\bar{y}_{st})$, for a fixed n is equivalent to minimize

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

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

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

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

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

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

Let $l_h = x_h - x_{h-1} \ge 0$ be the width of $h-th$ stratum and $x_L - x_0 = d$ (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 the *NLPP* (10) may be rewritten as:

$$\text{Minimize} \quad \sum_{h=1}^{L} \phi_h\left(l_h\right)$$

$$\text{subject to} \quad \sum_{h=1}^{L} l_h = d$$

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

Solving (11) the optimum strata width (OSW) l_h^* and hence the *OSB* can be obtained. In the following section, a general procedure for solving the *NLPP* (11) is discussed.

3. A GENERAL SOLUTION PROCEDURE

The *NLPP* (11) is a multistage decision problem which allows us to use the dynamic programming technique (see Khan et al., 2005, Khan et al., 2008, Khan et al., 2014). Dynamic programming determines the optimum solution of a multi-variable problem by decomposing it into stages, each stage comprising a single variable sub-problem.

Consider the following sub-problem of *NLPP* (11) for first $k\left(< L\right)$ strata:

$$\text{Minimize} \quad \sum_{h=1}^{k} \phi_h\left(l_h\right)$$

$$\text{subject to} \quad \sum_{h=1}^{k} l_h = d_k \,,$$

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

where $d_k < d$ is the total width available for division into first 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 $NLPP$ (12) and using the Bellman's principle of optimality (Bellman, 1957), 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 $k = 1$:

$$\Phi_1(d_1) = \phi_1(d_1),$$

$$\Rightarrow l_1^* = d_1 \tag{14}$$

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.

4. DETERMINATION OF OSB USING UNIFORM AUXILIARY VARIABLE

4.1 The Uniform Distribution

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 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 (see Wackerly et al. 2008).

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

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

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.

4.2 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

$$W_h = \frac{l_h}{b-a} \tag{16}$$

and

$$\sigma_{x_h}^2 = \frac{l_h^2}{12}. \tag{17}$$

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

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

$$\text{subject to} \quad \sum_{h=1}^{L} l_h = d$$

$$\text{and} \quad l_h \geq 0; h = 1,2,...,L \tag{18}$$

where $d = b-a$ is the range of the distribution, 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).

4.2.1 Estimating the Variance of the Error Term

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:

$$\phi(x) = cx^g; \quad c > 0, \quad g \geq 0, \tag{19}$$

where c and g are constants and $0 \leq g \leq 2$ (see Singh and Sukhatme (1969), Singh (1971), Rizvi et. al. (2002), Khan et. al.(2009)).

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

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

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 expressed as:

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

$$\text{subject to } \sum_{h=1}^{L} l_h = d$$

$$\text{and} \qquad l_h \geq 0; h = 1,2,...,L. \tag{21}$$

4.3 Numerical Illustration

To illustrate the computational details of the solution procedure discussed in Section 3 using a dynamic programming technique for determining the OSB with uniform distribution, we take $a=1$, $b=2$, $\beta=1.2$, $c=1$, $g=0$ and $d=1$. Then, the $NLPP$ (21) is reduced to:

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

$$\text{subject to } \sum_{h=1}^{L} l_h = 1$$

$$\text{and} \qquad l_h \geq 0; h = 1,2,...,L. \tag{22}$$

Note that the $(h-1)-th$ 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 relations (13) and (14) are reduced as:

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

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

For the stages $k \geq 2$:

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

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

TABLE 1: OSW **,** OSB **AND OPTIMUM VALUE OF THE OBJECTIVE FUNCTION FOR UNIFORM DISTRIBUTION**

No. of Strata L	OSW l_h^*	OSB x_h^*	Optimum Values of the Objective Function $\sum_{h=1}^{L} W_h \sigma_{y_h}$
2	$l_1^* = 0.500$ $l_2^* = 0.500$	$x_1^* = 1.500$	1.0148892
3	$l_1^* = 0.333$ $l_2^* = 0.333$ $l_3^* = 0.333$	$x_1^* = 1.333$ $x_2^* = 1.666$	1.0066446
4	$l_1^* = 0.250$ $l_2^* = 0.250$ $l_3^* = 0.250$ $l_4^* = 0.250$	$x_1^* = 1.250$ $x_2^* = 1.500$ $x_3^* = 1.750$	1.0037430
5	$l_1^* = 0.200$ $l_2^* = 0.200$ $l_3^* = 0.200$ $l_4^* = 0.200$ $l_5^* = 0.200$	$x_1^* = 1.200$ $x_2^* = 1.400$ $x_3^* = 1.600$ $x_4^* = 1.800$	1.0023971

6	$l_1^* = 0.167$ $l_2^* = 0.167$ $l_3^* = 0.167$ $l_4^* = 0.167$ $l_5^* = 0.167$ $l_6^* = 0.167$	$x_1^* = 1.167$ $x_2^* = 1.333$ $x_3^* = 1.500$ $x_4^* = 1.667$ $x_5^* = 1.833$	1.0016653

5. DETERMINATION OF OSB USING RIGHT-TRIANGULAR AUXILIARY VARIABLE

5.1 The 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 probability density function of a right-triangular distribution is given by

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

5.2 Formulation of the NLPP for Right Triangular distribution

Let the auxiliary variable x follow Right-Triangular Distribution within the interval $[a,b]$ given by (25). By using (2), (3), (4) and (25), the terms W_h and $\sigma_{x_h}^2$ can be expressed as

$$W_h = \frac{l_h(2a_h - l_h)}{(b-a)^2} \tag{26}$$

and

$$\sigma_{x_h}^2 = \frac{l_h^2\left(l_h^2 - 6a_h l_h + 6a_h^2\right)}{18\left(2a_h - l_h\right)^2} \tag{27}$$

where,

$a_h = b - x_{h-1}$.

Then, using (26) and (27), the *NLPP* (11) could be expressed as:

$$\text{Minimize} \quad \sum_{h=1}^{L} \frac{l_h\left(2a_h - l_h\right)}{\left(b-a\right)^2}\sqrt{\beta^2 . \frac{l_h^2\left(l_h^2 - 6a_h l_h + 6a_h^2\right)}{18\left(2a_h - l_h\right)^2} + \sigma_{e_h}^2}$$

$$\text{subject to} \quad \sum_{h=1}^{L} l_h = d$$

$$\text{and} \quad l_h \geq 0; h = 1, 2, ..., L \tag{28}$$

where $d = b - a$ is the range of the distribution, β is the regression coefficient and $\sigma_{e_h}^2$ the variance of the error function given in (6) for the error term in the regression model (1).

5.2.1 Estimating the Variance of the Error Term

Using (19), (25) and (26), we compute $\sigma_{e_h}^2$ as a function of boundary points as:

$$\sigma_{e_h}^2 = \frac{2c}{\left(2a_h - l_h\right)}\left[\frac{2a_h - g\left(2 - l_h - 2x_{h-1}\right) - l_h}{\left(g+1\right)\left(g+2\right)}\right]. \tag{29}$$

Thus, with (29), the NLPP (28) reduces to:

$$\text{Minimize} \ \sum_{h=1}^{L} \frac{l_h}{3\sqrt{2}\left(b-a\right)^2} \times$$

$$\sqrt{\beta^2\left[l_h^2\left(l_h^2 - 6a_h l_h + 6a_h^2\right)\right] + 36c\left(2a_h - l_h\right)\left[\frac{2a_h - g\left(2 - l_h - 2x_{h-1}\right) - l_h}{\left(g+1\right)\left(g+2\right)}\right]}$$

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

$$\text{and} \quad l_h \geq 0; h = 1, 2, ..., L. \tag{30}$$

5.3 Numerical Illustration

To illustrate the computational details of the solution procedure using the proposed dynamic programming technique for determining the OSB with Right-Triangular distribution, we assume that $a=1,\ b=2,\ c=1,\ g=0$ and $\beta=1.2$, which gives $a_h=2-x_{h-1}$.

Thus, the NLPP (30) reduces to:

$$\text{Minimize}\quad \sum_{h=1}^{L}\frac{l_h\sqrt{(1.2)^2\left(l_h^4-6a_hl_h^3+6a_h^2l_h^2\right)+18\left(2a_h-l_h\right)^2}}{3\sqrt{2}}$$

$$\text{subject to}\quad \sum_{h=1}^{L}l_h=1,$$

$$\text{and}\qquad l_h\geq 0;h=1,2,...,L. \tag{31}$$

Here, the $(h-1)-th$ stratification point is given by $x_{h-1}=1+d_h-l_h$.

Substituting this value of x_{h-1}, the recurrence relation (13) and (14) are reduced to:

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

$$\Phi\left(d_1\right)=\frac{d_1\sqrt{(1.2)^2\left(d_1^4-6d_1^3+6d_1^2\right)+18\left(2-d_1\right)^2}}{3\sqrt{2}}\quad \text{at}\quad l_1^*=d_1. \tag{32}$$

For the first stages $k\geq 2$:

$$\Phi_k\left(d_k\right)=\min_{0\leq l_k\leq d_k}\left[\begin{array}{l}\dfrac{l_k\sqrt{(1.2)^2\left(l_k^4-6a_kl_k^3+6a_k^2l_k^2\right)+18\left(2a_k-l_k\right)^2}}{3\sqrt{2}}\\ +\Phi_{k-1}\left(d_k-l_k\right)\end{array}\right] \tag{33}$$

where

$$a_k=1-d_h+l_h.$$

Solving the recurrence relations (32) and (33) using a $C++$ program, the $NLPP$ (31) 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 2 for $L=2,3,4,5$ and 6.

Table 2: OSW , OSB and optimum value of the objective function for right-triangular distribution

No. of Strata	OSW l_h^*	OSB x_h^*	Optimum Values of the Objective Function $\sum_{h=1}^{L} W_h \sigma_{y_h}$
2	$l_1^* = 0.618480$ $l_2^* = 0.381520$	$x_1^* = 1.618480$	1.0110791
3	$l_1^* = 0.462960$ $l_2^* = 0.285860$ $l_3^* = 0.251220$	$x_1^* = 1.462960$ $x_2^* = 1.748820$	1.0051409
4	$l_1^* = 0.375930$ $l_2^* = 0.232180$ $l_3^* = 0.204090$ $l_4^* = 0.187800$	$x_1^* = 1.37593$ $x_2^* = 1.60811$ $x_3^* = 1.81220$	1.0029562
5	$l_1^* = 0.319480$ $l_2^* = 0.197340$ $l_3^* = 0.173480$ $l_4^* = 0.159640$ $l_5^* = 0.150060$	$x_1^* = 1.319480$ $x_2^* = 1.516820$ $x_3^* = 1.690300$ $x_4^* = 1.849940$	1.0019174
6	$l_1^* = 0.279520$ $l_2^* = 0.172680$ $l_3^* = 0.151810$ $l_4^* = 0.139700$ $l_5^* = 0.131310$ $l_6^* = 0.124980$	$x_1^* = 1.279520$ $x_2^* = 1.452200$ $x_3^* = 1.604010$ $x_4^* = 1.743710$ $x_5^* = 1.875020$	1.0013436

6 A SIMULATION STUDY AND DISCUSSION

In this section, we conduct a simulation study to investigate the effectiveness of the proposed dynamic programming with the following methods that are available in stratification package in the R statistical software:

1. Dalenius and Hodges (1959) cum $\sqrt{f}$ method, which is the most frequently used and better known method.
2. Geometric method by Gunning and Horgan (2004).
3. Lavallée -Hidiroglou (1988) method with Kozak's (2004) algorithm.

In the simulation study, the uniformly distributed and right-triangular distributed populations were used.

6.1 Results for Uniformly Distributed Population

In this study, a data set (with $N=1000$) following uniform distribution with $a=0$ and $b=1$ was randomly generated by the R software. The values of minimum (x_0) and the maximum (x_L) were found to be 0.000391 and 0.998604 respectively. Thus, the dataset gives the range of the distribution as $d=x_L-x_0=0.998213$. Then, the OSBs using the proposed method as discussed earlier are obtained for the three different number of strata, that is, $L=2,3$ and 4. For comparison purposes, the OSB are also determined for cum$\sqrt{f}$, geometric and the Lavallée-Hidiroglou (Kozak's) method using the **stratification** package (see Baillargeon, and Rivest, 2011) with $CV=0.75$. These are presented in Table 3. The optimum values variance$=\sum_{h=1}^{L} W_h\sigma_{y_h}$ are also presented.

Table 3: OSB and Variance for Different Methods

	Cum $\sqrt{f}$ Method		Geometric Method		L-H (Kozak's) Method		Proposed Method	
L	OSB	Variance	OSB	Variance	OSB	Variance	OSB	Variance
2	0.77103	0.14089	0.02977	0.27636	1.52876	0.28779	0.73369	0.14089
3	0.48213 1.06008	0.09683	0.00772 0.12166	0.25198	0.65312 1.52876	0.14031	0.48388 1.03418	0.09603
4	0.38723 0.77103 1.25203	0.07291	0.00525 0.02977 0.24318	0.21899	0.36641 0.81380 1.41122	0.09563	0.36208 0.75511 1.20241	0.07291

Examination of Table 3 reveals that the OSBs obtained by the cum $\sqrt{f}$ method are by far the closest to the proposed dynamic programming method. The OBSs in the other two methods, namely geometric and Lavallée-Hidiroglou method with Kozak's algorithm differ widely from that of the dynamic programming method.

It can also be seen that the dynamic programming method yields the smallest variance for all $L = 2,3$ and 4 as compared to all the other methods. Although the values of variance for the dynamic programming method are almost same as of cum $\sqrt{f}$ method, the other two methods produce a greater variance than the dynamic programming technique. Thus, the study shows that the dynamic programming technique is more efficient than the other methods discussed in the manuscript while stratifying a population with uniform distribution.

6.2 Results for Right-Triangle Distributed Population

For the study, a data set (with $N = 800$) following a Right-Triangle distribution with $a = 0$ and $b = 2$ was generated by writing an R code given in Appendix A. The values of minimum (x_0) and the maximum $((x_L))$ were found to be 0.000484 and 1.928040 respectively. Thus, the range of the distribution is $d = x_L - x_0 = 1.927556$. Then, the OSBs using the proposed method are obtained for $L = 2,3$ and 4. The OSB are determined using cum$\sqrt{f}$ method, geometric method and the Lavallée-Hidiroglou (Kozak's) method using the stratification package (see Baillargeon, and Rivest, 2011) with $CV = 0.75$. Table 4 presents the OSB and the values of variance $= \sum_{h=1}^{L} W_h \sigma_{y_h}$ for all methods.

TABLE 4: OSB and Variance for Different Methods

	Cum $\sqrt{f}$ Method		Geometric Method		L-H (Kozak's) Method		Proposed Method	
L	OSB	Variance	OSB	Variance	OSB	Variance	OSB	Variance
2	0.77103	0.24765	0.02977	0.44071	1.52876	0.39987	0.73369	0.24740
	0.48213		0.00772		0.65312		0.48388	
3	1.06008	0.16870	0.12166	0.39182	1.52876	0.21196	1.03418	0.16780
	0.38723		0.00525		0.36641		0.36208	
	0.77103		0.02977		0.81380		0.75511	
4	1.25203	0.12262	0.24318	0.32399	1.41122	0.12784	1.20241	0.12245

From the Table above, it is noted that the OSBs obtained by the cum $\sqrt{f}$ method and the proposed dynamic programming method are very close to each other. Whereas the OBS's in the other two methods, geometric and Lavallée-Hidiroglou method with Kozak's algorithm differ widely from that of the proposed method. However, the table reveals that the proposed method yields the smallest variances of the estimate for all $L = 2,3$ and 4 as compared to all the other methods. Although the variances for the dynamic programming method are closed to the cum $\sqrt{f}$ method, the other two methods produce a greater variance than the dynamic programming technique. Thus, the study reveals that the proposed dynamic programming technique is more efficient than the other methods while stratifying a population with a Right-Triangular distribution.

7 CONCLUSION

Numerical examples using two sets of simulated data are presented to illustrate the applications and the computational details of the proposed technique. The results are presented together with the results of the cum $\sqrt{f}$ method of Dalenius and Hodges (1959), the geometric method by Gunning and Horgan (2004) and the generalized method of Lavallée and Hidiroglous (1988) for a comparative analysis. It is found that the construction of strata using auxiliary variable for the populations with Uniform and Right-Triangular distributions, leads to substantial gains in the precision of the estimates while using the proposed technique. Therefore, in many surveys, this research, especially finding the *OSB* for the study variable by using frequency functions of the auxiliary variables will be very useful in gaining the precision in the estimates of the survey.

APPENDIX A

R Code for Generating Random numbers Using Right Triangular Distribution:

The density function is

$$f(x) = 2(b-x)/(b-a)^2 \quad \text{for } a \le x \le b.$$

Here, a and b are fixed positive parameters, where $a < b$ is the minimum possible value. Let the default values of a and b are 0 and 2, respectively. Then, we define the density function as:

```
> drighttriangle=function(x, a=0, b=2) 2*(b-x)/(b-a)^2
```

Next, we integrate the density function to obtain the distribution function, which is

$$F(x) = 1-(b-x)^2/(b-a)^2 \text{ for } x \ge a.$$

Thus

> prighttriangle =function(x, a=0, b=2) (x > a)*(1-(b-x)^2/(b-a)^2)

Inverting the distribution function (i.e. $u = 1-(b-x)^2/(b-a)^2$) gives the quantile function $x = b-(b-a)(1-u)^{1/2}$.

That is,

> qrighttriangle=function(u, a=0, b=2) b-(b-a)*(1-u)^(1/2)

Finally, to simulate random Right-Triangle random variables, we use the fact that whenever the quantile function is applied to a uniform random variable, the result is a random variable with the desired distribution. Thus, the R function to generate the random numbers is

> rrighttriangle=function(n, a=0, b=2) qrighttriangle (runif(n),a,b)

REFERENCES

Aoyama, H. (1954): A study of stratified random sampling. *Ann. Inst. Statist. Math.* **6**, 1-36.

Baillargeon, S., and Rivest, L. P. (2011): The construction of stratified designs in R with package stratification. Survey Methodology, 37(1), 53-65.

Bellman, R.E. (1957): *Dynamic Programming*. Princetown University Press, New Jersey.

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. Aktuartidskr, **34**, 133-148.

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

Detlefsen, R.E., and Veum, C.S. (1991): Design issues for the retail trade sample surveys of the u.s. bureau of the census. *Proceedings of the Survey Research Methods Section*, ASA, 214-219.

Durbin, J. (1959): Review of sampling in Sweden. *J. Roy. Statist. Soc. Ser. A*, **122**, 146-148.

Ekman, G. (1959): Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Ann. Inst. Statist. 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 optimum strata boundaries for ratio and regression estimators. *Aligarh J. Statist.*, **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.

Hidiroglou, M. A. and Srinath, K. P. (1993): Problems associated with designing subannual business surveys. *J. Bus. Econom. Statist.*, **11**, 397-405.

Khan, M. G. M., Ahmad, N. and Khan, Sabiha (2009): Determining the optimum stratum boundaries using mathematical programming. *J. Math. Model. 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 Statist. Assoc. Bull.*, **52** (special Volume), 323-333.

Khan, M. G. M., Najmussehar and Ahsan, M. J. (2005): Optimum stratification for exponential study variable under Neyman allocation. *J. Indian Soc. Agricultural Statist.*, **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.

Khan, M.G.M., Rao, D., Ansari, A.H. and Ahsan, M. J. (2014): Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution. *Comm. Statist. Simulation Comput.* (To appear), DOI # 10.1080/03610918.2013.819917.

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. *J. Indian Statist. Assoc.*, **34**, 9-19.

Murthy, M. N. (1967): *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Nelder, J. A. and Mead, R. (1965): *A Simplex Method for Function Minimization*. *Computer Journal*, **7**, 308-313.

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.

Niemiro, W. (1999): Konstrukcja optymalnej stratyfikacja metoda poszukiwan losowych. (Optimal Stratification using Random Search Method). *Wiadomosci Statystyczne*, **10**, 1-9.

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.

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. *Ann. Inst. Statist. Math.*, **27**, 273-280.

Singh, R. (1971): Approximately optimum stratification on the auxiliary variable. *J. Amer. Statist. Assoc.*, **66**, 829-833.

Singh, R. (1975). An alternate method of stratification on the auxiliary variable. *Sankhya*, **37**, 100-108.

Singh, R. and Sukhatme, B. V. (1969): Optimum stratification for equal allocation. *Ann. Inst. Statist. Math.*, **27**, 273-280.

Singh, R. and Sukhatme, B. V. (1972): Optimum stratification in sampling with varying probabilities. *Ann. Inst. Statist. Math.*, **24**, 485-494.

Singh, R. and Sukhatme, B. V. (1973): Optimum stratification with ratio and regression methods of estimation. *Ann. Inst. Statist. Math.*, **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. *Ann. Inst. Statist. Math.*, 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.

Received : 10.12.2013
Revised : 28.04.2014

M. G. M. Khan, V. D. Prasad and D. K. Rao
School of Computing
Information and Mathematical Sciences
Faculty of Science Technology and Environment
The University of the South Pacific
Suva, Fiji Islands
email: khan_mg@usp.ac.fj