Optimal Stratification of Univariate Populations via stratifyR Package

Karuna G. Reddy[1], M.G.M. Khan[2]
Office of Deputy Vice-Chancellor, Research, International & Innovation,
The University of South Pacific, Suva, Fiji[1]
School of Computing, Information and Mathematical Sciences,
The University of South Pacific, Suva, Fiji[2]

Abstract
Stratification reduces the variance of sample estimates for population parameters by creating homogeneous strata. Often, surveyors stratify the population using the most convenient variables such as age, sex, region, etc. Such convenient methods often do not produce internally homogeneous strata, hence, the precision of the estimates of the variables of interest could be further improved. This paper introduces an R-package called 'stratifyR' whereby it proposes a method for optimal stratification of survey populations for a univariate study variable that follows a particular distribution estimated from a data set that is available to the surveyor. The stratification problem is formulated as a mathematical programming problem and solved by using a dynamic programming technique. Methods for several distributions such as uniform, weibull, gamma, normal, lognormal, exponential, right-triangular, cauchy and pareto are presented. The package is able to construct optimal stratification boundaries (OSB) and calculate optimal sample sizes (OSS) under Neyman allocation. Several examples, using simulated data, are presented to illustrate the stratified designs that can be constructed with the proposed methodology. Results reveal that the proposed method computes OSB that are precise and comparable to the established methods. All the calculations presented in this paper were carried out using the stratifyR package that will be made available on CRAN.

Key Words: Optimal stratification, Mathematical programming problem, Dynamic programming technique, Stratified random sampling, Optimum sample sizes, Univariate populations.

1. Introduction

Determination of optimum stratum boundaries (OSB) and optimum sample sizes (OSS) to be selected from each stratum are two inherent optimization problems in optimal stratification. Once the OSB have been determined, OSS can easily be computed using a particular sample allocation method. When stratification is based on a single study variable (y), its distribution can be utilized as the best characteristic to determine the OSB, i.e., by cutting the range of the distribution at suitable points. The basic consideration involved in determining OSB is that the strata should be as internally homogenous as possible. Thus, in order to achieve maximum precision, the stratum variances should be as small as possible Cochran (1977).

This univariate problem of determining the OSB was first discussed by Dalenius (1950) and then notable contributions to these problems were made by Dalenius and Gurney (1951), Mahalanobis (1952), Hansen, Hurwitz and Madow (1953), Aoyama (1954), Ekman (1959), Dalenius and Hodges (1957, 1959), Cochran (1961), Sethi (1963). They used the frequency distribution of the study variable to determine the OSB under various allocations of the sample sizes. Most of these authors achieved calculus equations which were not suitable for practical computations. They were only able to obtain approximate solutions under certain assumptions.

In statistical literature, several techniques to determine the OSB are available to the surveyors. An early and popular method is the Cumulative Root Frequency Method (Cum $\sqrt{f}$) of Dalenius and Hodges (1959), which approximates the implicit equations derived in Dalenius (1950). To date, this is one of the most widely-used techniques available. Lavallée and Hidiroglou's (1988) algorithm (L-H) is also a popular iterative procedure which gives the OSB and OSS that minimize the total sample size required to achieve a target level of

precision.

Gunning and Horgan (2004) proposed an alternative method to approximate stratification based on a geometric progression. They demonstrated an extremely simple way of stratifying skewed populations. Horgan (2006) compared this approach with Dalenius and Hodges (1959), Ekman (1959) and Lavallée and Hidiroglou (1988) and confirmed that geometric progression method is more efficient. However, Kozak and Verma (2006) studied the usefulness of Gunning and Horgan's geometric progression method and obtained a different result that the geometric progression approach is less efficient than Lavallée and Hidiroglou's algorithm (see Kozak et al (2007)).

Kozak (2004) presented a random search algorithm as a method of obtaining OSB and determined the OSS with Neyman allocation. Each iteration produced a set of random OSB, so a non-random version of the original was implemented by Baillargeon and Rivest (2011). Kozak (2004) tested the algorithm and concluded that the efficiency of the random search methods was similar to that of the L-H algorithm (Lavallée and Hidiroglou (1988)). Baillargeon and Rivest (2009) found that Kozak's algorithm produced better results than Lavallée and Hidiroglou, however, for small populations, Kozak's algorithm often yields a local minimum rather than global.

Another method of stratification method that has been proposed in the literature is due to Khan et al (2002, 2005, 2008, 2009, 2015) and Nand & Khan (2008). When the distributions or frequency functions of the study variables are known, they formulated the problems of determining the OSB as optimization problems using many different distributions. They considered the problem of finding OSB as an equivalent problem of determining Optimum Strata Width (OSW), which is formulated as a Mathematical Programming Problem (MPP) and solved by the Dynamic Programming (DP) technique, which was first proposed by Bühler and Deutler (1975). The advantage of this method is that it gives the optimum solution of the objective function and it does not require an initial solution, if the frequency distribution of the study variable is known and the number of strata (L) is fixed in advance. Collectively, the following distributions have been addressed in all their papers thus far: uniform, cauchy, exponential, pareto, right-triangular, weibull, gamma, normal and log-normal.

It may be a somewhat unrealistic assumption that stratification can be based on the frequency distribution of study variable because often the study variable is not available to the surveyor until the survey is done. Thus, to determine the OSB in such cases, the distribution of the study variable and it's approximate parameters could be ascertained from a previous or a recent survey. In this paper, all of the above distributions are implemented in a stratifyR package for the open-source R software. The package is able to determine OSB and OSS either directly from the data or when essential information such as the type and range of the distribution are provided. Simulated data are used to demonstrate numerical illustrations of the package.

2. Formulation of the Univariate Stratification Problem

Let the target population of the variable under study be stratified into L strata where the estimation of the mean of this study variable (y) is of interest. If a simple random sample of size n_h is to be drawn from h^{th} stratum with sample mean $\bar{y}_h$, then the stratified sample mean, $\bar{y}_{st}$, is given by

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

where W_h is the proportion of the population contained in the h^{th} stratum.

When the finite population correction factors are ignored, under the Neyman (1934) allocation,

$$n_h = n \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h}. \tag{2}$$

The variance of $\bar{y}_{st}$ is given by

$$Var(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{n}, \tag{3}$$

where σ_h^2 is the stratum variance for the study variable in the h^{th} stratum and n is the preassigned total sample size.

For a single study variable, if the data is available, the nature of distribution is easily estimated and the range over which OSB are constructed is known. In many practical situations, constructing OSB based on the study variable may not be feasible since the variable of interest is unavailable prior to conducting the survey. Thus, the estimates of the distribution and its range could be obtained from recent or past surveys and left to the judgement of the surveyor.

The problem of finding the OSB is formulated as Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter under Neyman allocation. The MPP is then solved for OSB by developing a solution procedure using a dynamic programming technique.

Let $f(y);\ a \leq y \leq b$ be the frequency function of the study variable on which OSB are to be constructed. If the population mean of this study variable is estimated under Neyman allocation given in (Neyman (1934)), then the problem of determining OSB is to cut up the range, $d = b - a$, at $(L-1)$ at intermediate points $a = y_0 \leq y_1 \leq y_2 \leq, ..., \leq y_{L-1} \leq y_L = b$ such that (3) is minimum. The lower and upper bounds of the study variable are denoted by a and b respectively.

For a fixed sample size n, minimizing the expression of the right hand side of (3) is equivalent to minimizing

$$\sum_{h=1}^{L} W_h \sigma_h \tag{4}$$

If $f(y)$ for the study variable is known and if this function is integrable, W_h, σ_h^2 and μ_h can be obtained as a function of the boundary points y_h and y_{h-1} by using the following expressions:

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy\,; \tag{5}$$

$$\sigma_h^2 = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 f(y)dy - \mu_h^2 \tag{6}$$

$$\text{where}\quad \mu_h = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y f(y)dy \tag{7}$$

and (y_{h-1}, y_h) are the boundaries of h^{th} stratum.

Thus, the objective function in (4) could be expressed as a function of boundary points y_h and y_{h-1} only. Further defining $l_h = y_h - y_{h-1};\ h = 1, 2, ..., L$ where $l_h \geq 0$ denotes the range or width of the h^{th} stratum and the range of the distribution, $d = b - a$, is expressed as a function of stratum width as:

$$\sum_{h=1}^{L} l_h = \sum_{h=1}^{L}(y_h - y_{h-1}) = b - a = y_L - y_0 = d \tag{8}$$

The h^{th} stratification point $y_h;\ h = 1, 2, ..., L$ is then expressed as $y_h = y_{h-1} + l_h$ and from (8), the problem can be treated as an equivalent problem of determining optimum strata widths (OSW), $l_1, l_2, ..., l_L$. Due to the special nature of functions, the problem may be

treated as a function of l_h alone and can be expressed as:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^{L} \phi_h(l_h), \\ \text{subject to} \quad & \sum_{h=1}^{L} l_h = d, \\ \text{and} \quad & l_h \geq 0; \;\; h = 1, 2, ..., L. \end{aligned} \tag{9}$$

3. Dynamic Programming Solution Procedure

The MPP (9) is a multistage decision problem in which the objective function and the constraint are separable functions of l_h, which allows us to use a dynamic programming technique (Khan et al., 2008). Dynamic programming determines the optimum solution of the MPP by decomposing it into stages, each stage comprising of a single variable subproblem. A dynamic programming model is basically a recursive equation based on Bellman's principle of optimality (Bellman, 1957). This recursive equation links the different stages of the problem in a manner which guarantees that each stage's optimal feasible solution is also optimal and feasible for the entire problem (Chapter 10, Taha, 2007).

Consider the following subproblem of (9) for first $k(< L)$ strata:

$$\begin{aligned} \text{Minimize} \quad & \sum_{h=1}^{k} \phi_h(l_h), \\ \text{subject to} \quad & \sum_{h=1}^{k} l_h = d_k, \\ \text{and} \quad & l_h \geq 0; \;\; h = 1, 2, ..., k. \end{aligned} \tag{10}$$

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 $k = L$.

The transformation functions are given by

$$\begin{aligned} d_k &= l_1 + l_2 + ... + l_k, \\ d_{k-1} &= l_1 + l_2 + ... + l_{k-1} \;=\; d_k - l_k, \\ d_{k-2} &= l_1 + l_2 + ... + l_{k-2} \;=\; d_{k-1} - l_{k-1}, \\ &\vdots \qquad\qquad \vdots \\ d_2 &= l_1 + l_2 \;=\; d_3 - l_3, \\ d_1 &= l_1 \;=\; d_2 - l_2. \end{aligned}$$

Let $\Phi_k(d_k)$ denote the minimum value of the objective function of (10), that is, for $h = 1, 2, ..., k$ and $1 \leq k \leq L$,

$$\Phi_k(d_k) = \min \left[\sum_{h=1}^{k} \phi_h(l_h) \middle| \sum_{h=1}^{k} l_h = d_k, \text{and} \, l_h \;\geq 0 \right].$$

With the above definition of $\Phi_k(d_k)$, the MPP (10) is equivalent to finding $\Phi_L(d)$ recursively by finding $\Phi_k(d_k)$ for $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$. Hence, for $l_h \geq 0; \; h = 1, 2, ..., k$,

$$\Phi_k(d_k) = \min \left[\phi_k(l_k) + \sum_{h=1}^{k-1} \phi_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k \right].$$

For a fixed value of l_k; $0 \leq l_k \leq d_k$, and $l_h \;\geq 0; \; h = 1, 2, ...(k-1)$ and $1 \leq k \leq L$,

$$\Phi_k(d_k) = \phi_k(l_k) + \min \left[\sum_{h=1}^{k-1} \phi_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k \right].$$

Using the Bellman's principle of optimality, a forward recursive equation of the dynamic programming technique for $k \geq 2$ and minimizing on $0 \leq l_k \leq d_k$ could be written as:

$$\Phi_k(d_k) \quad = \quad \min\left[\phi_k(l_k) + \Phi_{k-1}(d_k - l_k)\right]. \tag{11}$$

For the first stage ($k = 1$),

$$\Phi_1(d_1) = \phi_1(d_1) \implies l_1^* = d_1. \tag{12}$$

where $l_1^* = d_1$ is the optimum width of the first stratum. The relations (11) and (12) 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.

4. Results and Discussion

This section demonstrates the application of the stratifyR package that implements the prosed method. Together with the simulated data sets, the number of strata (h), fixed sample size (n) and population size (N) were used as the input arguments to the **strata.dp()** function in the package. When executed, the package outputs the OSB and OSS, amongst other quantities such as stratum weight (W_h), stratum variance (S_h^2), etc.

4.1 Distributions and their Datasets

The package deals with a total of nine distributions commonly found in surveys, namely, uniform, weibull, gamma, normal, lognormal, exponential, right-triangular, cauchy and pareto. Table 1 below presents the frequency functions (pdf) for all of the above distributions found in the package. The associated parameters, to be estimated via Maximum Likelihood Estimation (MLE) method (using 'fitdistrplus' package in R), are also provided in the table.

4.2 The stratifyR Package

Under the proposed method, in order to construct optimum stratum boundaries and optimum sample sizes for a given population, its best-fit frequency distribution needs to be estimated. The problem of OSB is then formulated as a mathematical programming problem, where the objective function is minimised on the range of the data set subject to the constraints given in Equation (9). Both the estimation of the distribution and the MPP formulation (for the indicated distributions) are implemented in the proposed stratifyR package. An example of the command used and its output from the package is given below in Table 2. The problem uses UScities data (Cochran (1961)) to construct a 3-strata solution with a fixed sample size of $n = 300$ from a population of $N = 1000$:

```
> strata.dp(data = UScities, h=3, n=300, N=1000)
```

Output:

```
Optimum Strata Boundaries for h = 3
Data Range: [10, 198] with d = 188
Best-fit Frequency Distribution: lnorm
Parameter estimates are:
meanlog    sdlog
3.231552   0.642034
```

To demonstrate the useablity of the package, several data sets were simulated (using random generators in R) for uniform, exponential, normal, and weibull distributions. Only four distributions are presented to basically show the results for these four. Similar distributions and subsequent results can be presented for the remaining distributions. Table 3

Table 1: Distributions, Frequency Functions and their Datasets

Distribution	*Frequency Function (pdf)*	*Parameters*
Uniform	$f(x) = \begin{cases} \frac{1}{b-a}; & x \in [a,b] \\ 0; & otherwise \end{cases}$	min(a), max(b)
Right-triangular	$f(x) = \begin{cases} \frac{2(b-x)}{(b-a)^2}; & x \in [a,b] \\ 0; & otherwise \end{cases}$	location(a), scale(b)
Exponential	$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}; & \lambda \in (0,+\infty) \\ 0; & otherwise \end{cases}$	rate(λ)
Cauchy	$f(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}; \quad x, x_0, \gamma \in (\infty, +\infty)$	location(x_0), scale(γ)
Pareto	$f(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}; \quad x \in [\beta, \infty); \quad \alpha, \beta \in (0, +\infty)$	shape(α), scale(β)
Gamma	$f(x) = \frac{x^{r-1}}{\theta^r \Gamma(r)} e^{-\frac{x}{\theta}}; \quad x, r, \theta \in (0, +\infty)$	shape(r), scale(θ)
Weibull	$f(x) = \frac{r}{\theta}\left(\frac{x}{\theta}\right)^{r-1} e^{-\left(\frac{x}{\theta}\right)^r}; \quad x, r, \theta \in (0, +\infty)$	shape(r), scale(θ)
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad x, \mu \in R, \sigma > 0$	location(μ), scale(σ)
Log-normal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2}; \quad x, \mu \in R, \sigma > 0$	location(μ), scale(σ)

Table 2: An Example Output from stratifyR Package

Strata(L)	*OSW*	*OSB*	W_h	S_h^2	$W_h S_h$	n_h	N_h
1	18.2308	28.2308	0.493361	25.5601	2.49428	168	335
2	27.1656	55.3964	0.321352	55.5140	2.39432	161	322
3	142.6036	198.0000	0.110648	533.8481	2.55653	172	343
Total WhSh:	7.44513						

below provides the objective functions that are minimized for these distributions subject to the constraints stated in Equation (9).

Table 3: Objective Functions and Estimated Parameters of Simulated Datasets

Distribution	*Objective Function ($W_h S_h$)*
Uniform	$\sum_{h=1}^{L} \frac{y_h^2}{2\sqrt{(3)}(b-a)}$
Exponential	$\sum_{h=1}^{L} e^{-\frac{x_{h-1}}{\lambda}} \sqrt{\lambda^2\left(1-e^{-\frac{y_h}{\lambda}}\right) - y_h^2 e^{-\frac{y_h}{\lambda}}}$
Weibull	$\left\{ \sum_{h=1}^{L} \text{Sqrt}\left\{ \theta^2\,\Gamma\left(\frac{2}{r}+1\right)\left[e^{-\left(\frac{x_{h-1}}{\theta}\right)^r} - e^{-\left(\frac{x_{h-1}+l_h}{\theta}\right)^r}\right] \times \left[Q\left(\frac{2}{r}+1, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(\frac{2}{r}+1, \left(\frac{x_{h-1}+l_h}{\theta}\right)^r\right)\right] - \left[\theta\,\Gamma\left(\frac{1}{r}+1\right)\left[Q\left(\frac{1}{r}+1, \left(\frac{x_{h-1}}{\theta}\right)^r\right) - Q\left(\frac{1}{r}+1, \left(\frac{x_{h-1}+l_h}{\theta}\right)^r\right)\right]\right]^2 \right\}$
Normal	$\left\{ \sum_{h=1}^{L} \text{Sqrt}\left\{ \frac{\sigma^2}{2\sqrt{2\pi}}\left[erf\left(\frac{x_{h-1}+y_h-\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{x_{h-1}-\mu}{\sigma\sqrt{2}}\right)\right] \times \left[\left(\frac{x_{h-1}-\mu}{\sigma}\right) exp\left(-\left(\frac{x_{h-1}-\mu}{\sigma\sqrt{2}}\right)^2\right) - \left(\frac{x_{h-1}+y_h-\mu}{\sigma}\right) exp\left(-\left(\frac{x_{h-1}+y_h-\mu}{\sigma\sqrt{2}}\right)^2\right)\right] + \frac{\sigma^2}{4}\left[erf\left(\frac{x_{h-1}+y_h-\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{x_{h-1}-\mu}{\sigma\sqrt{2}}\right)\right]^2 - \frac{\sigma^2}{2\pi}\left[exp\left(-\left(\frac{x_{h-1}-\mu}{\sigma\sqrt{2}}\right)^2\right) - exp\left(-\left(\frac{x_{h-1}+y_h-\mu}{\sigma\sqrt{2}}\right)^2\right)\right]^2 \right\}$

Table 4 provides the outputs (which presents only essential items such as OSB and OSS) provided by the package for a fixed sample size of $n = 300$ from a population of $N = 1000$. The input parameters to the **strata.dp()** function of the stratifyR package are the data, number of strata, n and N. The OSB and OSS are provided for the four distributions given in Table 3.

Table 4: OSB and OSS for Uniform, Exponential, Weibull and Normal Distributions Using Proposed Method in stratifyR Package

	Uniform		Exponential		Weibull		Normal	
L	OSB	OSS	OSB	OSS	OSB	OSS	OSB	OSS
2	2.97423	150	1.18353	142	1.40886	146	4.96836	149
		150		158		154		151
3	2.43835	101	0.716667	96	1.04878	101	4.42763	103
	3.51621	98	1.893383	97	1.82909	94	5.51739	92
		101		107		106		105
4	2.14827	76	0.516607	73	0.860306	77	4.11097	78
	2.97433	74	1.229601	73	1.424337	71	4.97401	71
	3.81278	74	2.399734	73	2.089903	72	5.84278	71
		76		81		80		81
5	1.96159	61	0.404407	59	0.740139	62	3.89201	63
	2.65156	59	0.918335	59	1.198009	58	4.64429	57
	3.29814	59	1.628727	59	1.672368	58	5.30841	57
	4.00485	60	2.791513	59	2.276567	58	6.07075	57
		61		65		65		65
6	1.83	51	0.332389	49	0.655387	52	3.7276	53
	2.43035	50	0.734577	49	1.047475	48	4.41077	48
	2.97363	50	1.246974	49	1.430015	48	4.97763	48
	3.52205	50	1.954457	49	1.856256	48	5.54731	48
	4.14085	50	3.10936	56	2.42033	49	6.24443	48
		50		54		54		55

To compare the OSB and OSS obtained in Table 4 via the stratifyR package, 'stratification' package (Baillargeon & Rivest (2011)) is used to construct OSB using Cum $\sqrt{f}$ method and calculate OSS under Neyman allocation given by Equation (2). Table 5 provides these results and the computed values reveal that the OSB and OSS obtained via the proposed method implemented in the stratifyR package is very similar to the Cum $\sqrt{f}$ method, that is, they are quite comparable to each other. It is able to successfully create OSB and OSS for the simulated populations and can surely be used on any real population.

Table 5: OSB and OSS for Uniform, Exponential, Weibull and Normal Distributions Using the Cum $\sqrt{f}$ Method

	Uniform		Exponential		Weibull		Normal	
L	OSB	OSS	OSB	OSS	OSB	OSS	OSB	OSS
2	2.01	150	1.23	146	1.45	150	4.94	145
		150		154		150		155
3	2.36	102	0.77	104	1.03	90	4.4	103
	3.64	95	1.99	99	1.87	106	5.61	103
		103		97		104		94
4	1.96	71	0.46	60	0.87	77	4	64
	3	79	1.23	84	1.45	72	4.94	84
	3.96	69	2.3	67	2.2	78	5.88	74
		81		89		73		78
5	1.8	61	0.46	77	0.7	48	3.87	58
	2.6	58	0.92	48	1.2	77	4.67	66
	3.4	57	1.68	64	1.7	58	5.34	57
	4.2	63	2.76	57	2.37	57	6.14	62
		61		54		60		57
6	1.64	44	0.31	47	0.62	39	3.73	55
	2.36	58	0.77	67	1.03	55	4.4	48
	3	47	1.23	41	1.45	57	4.94	39
	3.64	47	1.84	47	1.87	42	5.61	58
	4.28	48	2.91	52	2.45	49	6.28	43
		56		46		58		57

5. Conclusion

The stratifyR package successfully implements the proposed method to construct OSB and OSS by minimizing the formulated MPP for the best-fit distribution of a given data set. Results for the four simulated data sets with different distributions illustrate that the stratified designs can be constructed with the proposed methodology. The package is able to handle a total of nine different distributions and all of their performances, in terms of precision and comparability, are on par with the established Cum $\sqrt{f}$ method. Since the study variables are not easily available in practice, the package also has the advantage of being able to create OSB and OSS based on the distribution (ascertained from past surveys) over a particular range.

REFERENCES

Aoyama, H. (1954), "A Study of Stratified Random Sampling," *Annals of the Institute of Statistical Mathematics*, **6**, 1-36.

Baillargeon, S., and Rivest, L. P. (2009). "A general algorithm for univariate stratification", *International Statistical Review*, **77(3)**, 331-344.

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

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

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

Cochran, W.G. (1977), "Sampling Techniques", 3rd edition, *John Wiley & Sons Inc.*, New York.

Cochran, W.G. (1961), "Comparison of methods for determining stratum boundaries", *Bull. Int. Stat. Inst.*, **38**, 345-58.

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

Dalenius, T., and Gurney, M. (1951), "The Problem of Optimum Stratification", *Almqvist & Wiksell*, Stockholm.

Dalenius, T. and Hodges, J.L. (1959), "Minimum Variance Stratification", *Journal of American Statistical Association*, **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.

Horgan, J.M. (2006), "Stratification of Skewed Populations: A Review", *International Statistical Review.* **74(1)**, 67-76.

Hansen, M.H. and Hurwitz, W.N. (1953), "Theory of Sampling from Finite Population", *Ann. Math. Stat.* **14**, 333-362.

Hidiroglou, M.A. and Srinath, K.P. (1993), "Problems Associated with Designing Subannual Bussiness Surveys", *Journal of Bussiness and Economic Statistics*, **11**, 397-405.

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.

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.

Khan, M.G.M., Ahmad, N. and Khan, S. (2009), "Determining the Optimum Stratum Boundaries using Mathematical Programming", *Journal of Mathematical Modelling and Algorithms*, **8(4)**, 409-423, DOI:10.1007/s10852-009-9115-3.

Khan, M.G.M., Reddy, K.G. and Rao, D.K. (2015). "Designing stratified sampling in economic and business surveys", *Journal of Applied Statistics*, **42(10)**, 2080-2099.

Kozak, M. (2004), "Optimal Stratification Using Random Search Method in Agricultural Surveys", *Statistics in Transition*, **6(5)**, 797-806.

Kozak, M. and Verma, M.R. (2006), "Geometric versus Optimisation Approach to Stratification: A Comparison of Efficiency", *Survey Methodology*, **32(2)**, 157-163.

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

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

Nand, N. and Khan, M.G.M. (2009). "Optimum Stratification for Cauchy and Power Type Study Variables", *Journal of Applied Statistical Science*, **16(4)**, 453-462.

Neyman, J. (1934), "On the Two Different Aspects of the Representative 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.

Sethi, V. K. (1963). A note on optimum stratification of populations for estimating the population means. The Australian Journal of Statistics, 5, 20-33.