Optimum Stratification Using Multiple Auxiliary Variables with 3P Weibull Distributions

Karuna G. Reddy*
The University of the South Pacific, Suva, Fiji - karunaredz@gmail.com

M.G.M. Khan and Dinesh K. Rao
The University of the South Pacific, Suva, Fiji - khan_mg/dinesh.rao@usp.ac.fj

Abstract

The determination of Optimum Stratum Boundaries (OSB) based on the survey variable is not feasible in practice since the variable of interest is unavailable prior to conducting the survey. This paper proposes a method of constructing OSB for a study variable based on multiple auxiliary variables that are readily available and regressible with the study variable. The auxiliary variables used for this problem are estimated to follow skewed 3P Weibull distributions. It is formulated into a Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter. The formulated MPP is then solved for the OSB using a dynamic programming (DP) technique. A numerical example with a real data set, aiming to estimate the Haemoglobin content in women in a national Iron Deficiency Anaemia survey, is presented to illustrate the procedure developed in this paper. The results obtained by the proposed technique are compared with other univariate methods and the results reveal that the proposed approach yields a substantial gain in the precision of the estimates.

Keywords: Optimum stratification; 3P Weibull distribution; Mathematical programming problem; Dynamic programming technique

1. Introduction

In stratified random sampling, the sampling-frame is divided into non-overlapping groups or strata in such a way that the strata constructed are internally homogeneous with respect to the main variable that maximizes the precision of its estimate. The surveyors often stratify the population conveniently by using geographical or administrative regions or other natural criteria such as age, gender, etc. which are not always a reasonable criterion as the strata so obtained may not be internally as homogeneous as possible with respect to a variable of interest. When stratification is made based on a single study variable for which the distribution is known, the OSB can be determined by cutting the range of its distribution at suitable points. This problem was first discussed by Dalenius (1950) who presented a set of minimal equations which are usually difficult to solve because of their implicit nature.

Several approximation methods of determining OSB using the distribution of auxiliary variable, x, have been suggested by many authors such as Dalenius and Hodges (1959), Sethi (1963), Taga (1967), Serfling (1968), Singh and Sukhatme (1969), Cochran (1977) and Rizvi, Gupta, Bhargava (2002). Lavallee and Hidiroglou (1988) proposed an algorithm to construct stratum boundaries for a power allocated stratified sample. Sweet and Sigman (1995) and Rivest (2002) reviewed Lavallee and Hidiroglous algorithm and proposed a modified algorithm that incorporates the different relationships between the stratification and study variables.

Kozak (2004) presented a modified random search algorithm while Gunning and Horgan (2004) proposed an alternative method to approximate stratification based on a geometric progression. Kozak and Verma (2006) studied Gunning and Horgans geometric progression method and found out that the approach is less efficient than Lavallee and Hidiroglous algorithm. Another kind of stratification method is due to Bühler and Deutler (1975); Khan et. al. (2008, 2009) where they formulated the problems of determining OSB as optimization problems, which are solved by developing computational techniques using DP technique.

This paper presents a procedure for determining the OSB for the main variable in a positively skewed population using multiple auxiliary variables $(x_1, x_2, ..., x_p)$ that are estimated to follow 3P Weibull distributions.

The formulated MPP is a multistage decision problem that is solved by using a DP technique. A numerical example is presented to illustrate the application of the proposed method.

2. The General Formulation of the Problem of OSB as an MPP

Let the population be stratified into L strata based on p auxiliary variables, $x_i = x_1, x_2, ..., x_p$, and the estimation of the mean of 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 $\bar{y}_h; (h = 1, 2, ..., L)$, then the variance of the stratified sample mean, $\bar{y}_{st}$, is given by

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

where $W_h = N_h/N$ and σ_{hy}^2 is the stratum variance of y in h^{th} stratum; $h = 1, 2, ..., L$ and n is the preassigned total sample size.

Consider that the study variable has the regression model of the form:

$$y = \lambda(x_1, x_2, ..., x_p) + \epsilon, \tag{2}$$

where $\lambda(x_1, x_2, ..., x_p)$ is a linear (or nonlinear) function of $x_i; (i = 1, 2, ..., p)$ and ϵ is an error term such that $E(\epsilon|x_1, x_2, ..., x_p) = 0$ and $V(\epsilon|x_1, x_2, ..., x_p) = \sigma_{h\epsilon}^2 > 0$ for all x_i.

If λ and ϵ are uncorrelated, from model (2), σ_{hy}^2 can be expressed as (see Dalenius and Gurney (1951)) $\sigma_{hy}^2 = \sigma_{h\lambda(x_1,x_2,...,x_p)}^2 + \sigma_{h\epsilon}^2$, where $\sigma_{h\epsilon}^2$ is the variance of ϵ in the h^{th} stratum.

Let $f(x_i)$ be the frequency function of the auxiliary variables, $x_i; (i = 1, 2, ..., p)$, that are used for the stratification of the main variable. If the population mean of the study variable y is estimated over its range (a, b), the problem of determining the OSB of y is to cut up the range, (a, b) at $(L - 1)$ intermediate points $a = y_0 \leq y_1 \leq y_2 \leq, ..., \leq y_{L-1} \leq y_L = b$ such that (1) is minimum.

For a fixed sample size n, minimizing the expression of the right hand side of (1) is equivalent to minimizing $\sum_{h=1}^{L} W_h \sigma_{hy}$. Thus, we minimize

$$\sum_{h=1}^{L} W_h \sigma_{hy} = \sum_{h=1}^{L} W_h \sqrt{\sigma_{h\lambda(x_i)}^2 + \sigma_{h\epsilon}^2} = \sum_{h=1}^{L} \sqrt{W_h^2 \sigma_{h\lambda(x_i)}^2 + W_h^2 \sigma_{h\epsilon}^2} \tag{3}$$

If $f(x_i)$ are known and integrable frequency functions of the auxiliary variables, then for the given $\lambda(x_i)$ and $\sigma_{h\epsilon}^2$, the first term in (3) can be expressed as the functions of the boundary points (y_{h-1}, y_h) by finding the stratum weight, mean and variance respectively as follows:

$$W_{hx_i} = \int_{y_{h-1}}^{y_h} f(x_i)\, dx_i \tag{4}$$

$$\mu_{hx_i} = \frac{1}{W_{hx_i}} \int_{y_{h-1}}^{y_h} x_i f(x_i)\, dx_i \tag{5}$$

$$\sigma_{hx_i}^2 = \frac{1}{W_{hx_i}} \int_{y_{h-1}}^{y_h} x_i^2 f(x_i)\, dx_i - \mu_{hx_i}^2 \tag{6}$$

The second term in (3) are also obtained as a function of boundary points using the frequency function of the regression error. Thus, the objective function (3) could be expressed as a function of boundary points (y_{h-1}, y_h) only, that is, $\phi_h(y_h, y_{h-1}) = \sqrt{W_h^2 \sigma_{h\lambda(x_i)}^2 + W_h^2 \sigma_{h\epsilon}^2}$.

The h^{th} stratification point $y_h;\ h = 1, 2, ..., L$ is then expressed as $y_h = y_0 + \sum_{i=1}^{h} l_i = y_{h-1} + l_h$. With the range as a constraint, the problem of determining OSW, $l_1, l_2, ..., l_L$, can be derived. If y_0 is known, the first

term, $\phi_1(l_1, y_0)$, is a function of l_1 alone. Once l_1 is known, the second term $\phi_2(l_2, y_1)$ will become a function of l_2 alone and so on. Hence, the problem may be treated as a function of l_h alone and is expressed as:

$$\begin{aligned} Minimize \quad & \sum_{h=1}^{L} \phi_h(l_h), \\ subject\ to \quad & \sum_{h=1}^{L} l_h = d, \\ and \quad & l_h \geq 0; \ \ h = 1, 2, ..., L. \end{aligned} \tag{7}$$

3. The Solution Procedure using Dynamic Programming Technique

Dynamic programming determines the optimum solution by decomposing the multistage decision problem (7) into stages, each stage comprising of a single variable subproblem. A DP model is basically a recursive equation based on Bellman's principle of optimality (see Bellman (1957)).

Consider a subproblem of (7) for first $k(< L)$ strata 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

$$\begin{aligned} d_k &= l_1 + l_2 + ... + l_k, \ \text{ and} \\ d_{k-1} &= l_1 + l_2 + ... + l_{k-1} = d_k - l_k \end{aligned}$$

Let $\Phi_k(d_k)$ denote the minimum value of the subproblem, that is,

$$\Phi_k(d_k) = Minimize \left[\sum_{h=1}^{k} \phi_h(l_h) \middle| \sum_{h=1}^{k} l_h = d_k, \, and \, l_h \ \geq 0; \, h = 1, 2, ..., k \, and \, 1 \leq k \leq L \right].$$

With the above definition of $\Phi_k(d_k)$, the MPP (7) is equivalent to finding $\Phi_L(d)$ recursively by finding $\Phi_k(d_k)$ for $k = 1, 2, ..., L$ and $0 \leq d_k \leq d$. We can write:

$$\Phi_k(d_k) = Minimize \left[\phi_k(l_k) + \sum_{h=1}^{k-1} \phi_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k, \, and \, l_h \ \geq \ 0; \ h = 1, 2, ..., k \right].$$

For a fixed value of l_k; $0 \leq l_k \leq d_k$,

$$\Phi_k(d_k) = \phi_k(l_k) + Minimize \left[\sum_{h=1}^{k-1} \phi_h(l_h) \middle| \sum_{h=1}^{k-1} l_h = d_k - l_k, \, and \, l_h \ \geq 0; \ h = 1, 2, ...k-1 \, and \ 1 \leq k \leq L \right].$$

Thus, we write a forward recursive equation of the dynamic programming technique as:

$$\Phi_k(d_k) = \underset{0 \leq l_k \leq d_k}{Minimize} \left[\phi_k(l_k) + \Phi_{k-1}(d_k - l_k) \right], \ \ k \ \geq \ 2. \tag{8}$$

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

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

where $l_1^* = d_1$ is the optimum width of the first stratum. The relations (8) and (9) are solved in a forward manner first starting $k = 1, 2, ..., L$ to determine the optimum subproblem objective and then solved in a backward manner second to determine the OSB.

4. Formulation of the MPP with Weibull Distribution

The 3P probability Weibull density function for i^{th} auxiliary variable with a state space $x_i \geq 0$ is given by

$$f(x_i; r_i, \theta_i, \gamma_i) = \frac{r_i}{\theta_i} \left(\frac{x_i - \gamma_i}{\theta_i} \right)^{r_i - 1} e^{-\left(\frac{x_i - \gamma_i}{\theta_i} \right)^{r_i}}, \ \ x_i \geq 0 \tag{10}$$

where $r_i > 0$ is the shape parameter, $\theta_i > 0$ is the scale parameter and γ_i is the location parameter.

All auxiliary variables are standardized by subtracting its mean and dividing by its standard deviation. Considering that b_i is the upper bound and a_i is the lower bound for i^{th} auxiliary variable, the estimated range, d, can be calculated as:

$$d = \text{Max}(b_1, b_2, ..., b_p) - \text{Min}(a_1, a_2, ..., a_p) \tag{11}$$

Thus, the solution provides the OSB of the standardized study variable and the OSB for the original study variable can be obtained by the usual unstandardizing procedure.

With all the auxiliary variables, x_i, following 3P Weibull distributions, the quantities W_{hx_i}, μ_{hx_i}, and $\sigma^2_{hx_i}$ can be obtained as a function of boundary points (y_{h-1}, y_h) and presented as follows:

$$W_{hx_i} = e^{-\left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}} - e^{-\left(\frac{y_{h-1}+l_h-\gamma_i}{\theta_i}\right)^{r_i}}, \tag{12}$$

$$\mu_{hx_i} = \frac{\theta_i}{W_{hx_i}} \left[\int_{\left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}}^{\infty} t^{\frac{1}{r_i}} e^{-t}\, dt - \int_{\left(\frac{y_h-\gamma_i}{\theta_i}\right)^{r_i}}^{\infty} t^{\frac{1}{r_i}} e^{-t}\, dt \right] \tag{13}$$

where $\Gamma(r, x)$ and $Q(r, s)$ denote the upper incomplete gamma function and the regularized incomplete gamma function, respectively.

Then, using equations (4)-(6), μ_{hx_i} can be simplified as

$$\mu_{hx_i} = \frac{\theta_i \Gamma\left(1+\frac{1}{r_i}\right)}{W_{hx_i}} \left\{ \left[Q\left(1+\frac{1}{r_i}, \left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}\right) - Q\left(1+\frac{1}{r_i}, \left(\frac{y_{h-1}+l_h-\gamma_i}{\theta_i}\right)^{r_i}\right)\right]\right\} \tag{14}$$

Similarly, the quantity $\sigma^2_{hx_i}$ is reduced to

$$\sigma^2_{hx_i} = \frac{\theta_i^2\, \Gamma\left(1+\frac{2}{r_i}\right)}{W_{hx_i}} \left[Q\left(1+\frac{2}{r_i}, \left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}\right) - Q\left(1+\frac{2}{r_i}, \left(\frac{y_{h-1}+l_h-\gamma_i}{\theta_i}\right)^{r_i}\right)\right] - \mu^2_{hx_i} \tag{15}$$

where W_{hx_i} and and $\mu^2_{hx_i}$ are given by equations (12) and (14) respectively.

W_h and $\sigma^2_{h\lambda}$ in the first term of (3) are given by (12) and (15) respectively. Thus, the formulated MPP (7) could be generalised and expressed as the following MPP that will be minimized to determine the OSB:

$$\sum_{h=1}^{L} \left\{ \begin{array}{l} \text{Sqrt}\left\{ \sum_{i=1}^{p} \beta_i^2 \left[\frac{\theta_i^2\, \Gamma\left(1+\frac{2}{r_i}\right)}{W_{hx_i}} \left[Q\left(1+\frac{2}{r_i}, \left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}\right) - Q\left(1+\frac{2}{r_i}, \left(\frac{y_{h-1}+l_h-\gamma_i}{\theta_i}\right)^{r_i}\right)\right] \right.\right. \\ \quad \left.\left. - \left[\frac{\theta_i \Gamma\left(1+\frac{1}{r_i}\right)}{W_{hx_i}} \left[Q\left(1+\frac{1}{r_i}, \left(\frac{y_{h-1}-\gamma_i}{\theta_i}\right)^{r_i}\right) - Q\left(1+\frac{1}{r_i}, \left(\frac{y_{h-1}+l_h-\gamma_i}{\theta_i}\right)^{r_i}\right)\right]\right]^2 \right] \right. \\ \quad \left. + W_h^2 \sigma^2_{h\epsilon} \right\} \end{array} \right\}, \\ \textbf{Subject to} \quad \sum_{h=1}^{L} l_h = d, \\ \textbf{and} \quad l_h \geq 0; \;\; h = 1, 2, ..., L \tag{16}$$

where d in (16) is the estimated range obtained by (11), β_i are the regression coefficients, θ_i and r_i are parameters of the 3P Weibull distributions of i^{th} auxiliary variable. The term $W_h^2 \sigma^2_{h\epsilon}$ is computed by (12)-(15) since the distribution of ϵ is known to be approximately normally distributed and is given by

$$f(\epsilon) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2}{2}\right), \; -\infty < \epsilon < +\infty \tag{17}$$

5. Results and Discussion

To illustrate the application of the method, a health data set of size $N = 724$ is obtained from 2004 Fiji National Nutrition Survey which has three characteristics for each woman: Haemoglobin, Iron and Folate levels. Suppose that for a survey where a sample is be collected using stratified random sampling and Haemoglobin (y) will be the main variable, the levels of Iron and Folate may be the reasonable choice for the auxiliary variables, x_1 and x_2.

Using the recursive equations (8) and (9), the MPP (16) with $d = 3.9324 - (-1.8866) = 5.8190$ given in (11) is solved by executing a computer program developed for the proposed DP technique. The performance of the proposed method is compared against the available univariate methods such as Cum $\sqrt{f}$ method of Dalenius and Hodges (1959); Geometric method of Gunning and Horgan (2004); and Lavallee and Hidiroglou (1988) method with Kozak (2004) algorithm.

The `stratification` package recently developed by Baillargeon and Rivest (2011) in the R statistical software is used to determine the OSB for the main study variable for the three methods. The results in the form of OSB and the variances of the estimate using all methods for $L = 2, 3, ..., 6$ are given in Table 1. The efficiencies of the proposed method over the other methods are given in Table 2.

Table 1: OSB and Variances Using All Four Methods

	Proposed Method		**Cum $\sqrt{f}$**		**Geometric**		**Kozak (L-H)**	
L	**OSB**	**Variance**	**OSB**	**Variance**	**OSB**	**Variance**	**OSB**	**Variance**
2	7.58	2.0094e-07	12.15	2.1821e-07	10.15	2.1821e-07	12.35	2.1821e-07
3	7.07		11.28		8.57		11.55	
	8.26	1.6572e-07	13.23	2.1821e-07	12.03	2.1719e-07	12.75	2.1821e-07
4	6.83		10.64		7.87		11.35	
	7.64	1.3641e-07	12.15	2.1821e-07	10.15	2.1007e-07	12.35	2.1821e-07
	8.70		13.66		13.10		13.05	
5	6.68		10.20		7.48		9.25	
	7.31	1.1462e-07	11.72	2.1821e-07	9.17	1.9639e-07	11.95	2.1811e-07
	8.04		12.80		11.24		12.75	
	9.02		13.88		13.78		13.55	
6	6.59		9.77		7.23		9.35	
	7.10		11.07		8.57		12.05	
	7.66	9.8360e-08	12.15	2.1819e-07	10.15	1.8012e-07	12.65	2.1671e-07
	8.34		13.01		12.03		13.05	
	9.28		14.09		14.26		13.55	

Table 2: Efficiency of the Proposed DP Method

	Efficiency (%) of DP Method Over		
L	**Cum $\sqrt{f}$**	**Geometric**	**L-H (Kozak)**
2	108.59	108.59	108.59
3	131.67	131.06	131.67
4	159.96	153.99	159.96
5	190.38	171.34	190.34
6	221.83	183.12	221.78

Upon examination of the results, it is noted that the OSB given by the proposed DP method appear to be quite different from the other methods. The variances of Haemoglobin in all methods appear to be declining exponentially as L increases, however, the proposed method performs substantially better than other methods. For $L = 1, 2, ..., 6$, the efficiency of the OSB determined by the proposed method increases by about 109% to 222% over Cum $\sqrt{f}$ and L-H Kozak's methods while against the Geometric method, it increases from about 109% to 183%, Thus, the proposed method yields lower variances compared to other univariate

methods.

6. Conclusion

This paper presents a method which uses multiple auxiliary variables to determine the OSB. It is found out that the method substantially increases the precision of the estimates. Skewed auxiliary variables are fitted with 3P Weibull distributions and then the problem of finding the OSB is formulated as an MPP and solved using a DP technique where the variance of the estimated population parameter is minimized. A numerical example using a real data set is presented to illustrate the application and performance of the proposed technique. The proposed methods works considerably well in determining the OSB for the main variable and results in substantial gains in the precision of the estimates over other available methods.

References

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

Bellman RE (1957) *Dynamic Programming*, Princetown University Press, New Jersey.

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

Cochran WG (1977) *Sampling Techniques*, 3rd edn. John Wiley & Sons Inc.:New York.

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

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

Dalenius T and Hodges JL (1959) *Minimum Variance Stratification*, 54:88-101. J. Ame. Stat. Assn.

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

Khan MGM, Nand N and Ahmad N (2008) *Determining the Optimum Strata Boundary Points Using Dynamic Programming*, 34(2):205-214.. Survey Methodology.

Khan MGM, Ahmad N and Khan S (2009) *Determining the Optimum Stratum Boundaries using Mathematical Programming*, 8(4):409-423, DOI:10.1007/s10852-009-9115-3. J. Math. Mod. Algo.

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

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

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

Rivest LP (2002) *A Generalization of Lavalle and Hidiroglou algorithm for Stratification in Business Survey*, 28:191-198. Survey Methodology.

Rizvi SEH, Gupta JP and Bhargava M (2002) *Optimum Stratification based on Auxiliary Variable for Compromise Allocation*, 28(1):201-215. Metron.

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

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

Singh R and Sukhatme BV (1969) *Optimum Stratification*, 21(3):515-528. Ann. Inst. Stat. Math.

Sweet EM and Sigman RS (1995) *Evaluation of Model-assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data*, 491-496. Proceedings of the Survey Research Methods Section, ASA.

Taga Y (1967) *On Optimum Stratification for the Objective Variable Based on Concomitant Variables using Prior Information*, 19:101-129. Ann. Inst. Stat. Math.