The Korean Journal of Applied Statistics (2009)
22(6), 1277–1287

On Convergence of Stratification Algorithms for Skewed Populations

Inho Park[1]

[1]Economic Statistics Department, The Bank of Korea

(Received August 2009; accepted September 2009)

Abstract

For stratifying skewed populations, the Lavallée-Hidiroglou(LH) algorithm is often considered to have a take-all stratum with the largest units and some take-some strata with the middle-size and small units. Related to its iterative nature have been reported some numerical difficulties such as the dependency of the ultimate stratum boundaries to a choice of initial boundaries and the slow convergence to locally-optimum boundaries. The geometric stratification has been recently proposed to provide initial boundaries that can avoid such numerical difficulties in implementing the LH algorithm. Since the geometric stratification does not pursuit the optimization but the equalization of the stratum CVs, the corresponding stratum boundaries may not be (near) optimal. This paper revisits these issues concerning convergence and near-optimality of optimal stratification algorithms using artificial numerical examples. We also discuss the formation of the strata and the sample allocation under the optimization process and some aspects related to discontinuity arisen from the finiteness of both population and sample as well.

Keywords: Lavallée-Hidiroglou(LH) algorithm, geometric stratification, random search algorithm, stratum boundaries, optimization.

1. Introduction

In many economic surveys, data are usually positively skewed so that a small number of large units account for the most share of the population total of a study variable. Thus, it is more appealing to survey planners to have a take-all stratum with the largest units and some take-some strata with the middle-size and small units. See, for example, Sigman and Monsour (1995) and Slanta and Krenzke (1996).

The LH algorithm (Lavallée and Hidiroglou, 1988) is aimed to implement the above sampling design scheme in the stratification. The algorithm iteratively updates stratum boundaries (or breaks) based on the stratum boundary formulas that are driven to minimize the total sample size for a target CV(coefficient of variation), where any sample allocation rule can be built-in through the stratum sample sizes.

[1]Statistician, Economic Statistics Department, The Bank of Korea, Namdaemun-Ro 106, Jung-Gu, Seoul 100-794, Korea. E-mail: ipark@bok.or.kr

A few numerical difficulties, however, have been addressed in relation to the LH algorithm. Slanta and Krenzke (1996), for example, encountered two problems: the dependency of the ultimate boundaries to a choice of initial boundaries and the slow convergence to locally-optimum boundaries. Gunning *et al.* (2008) argued that the boundaries obtained from the geometric progression can avoid such numerical difficulties since they are already near the optimum boundaries. The geometric stratification is to provide stratum boundaries so as to have equal stratum CVs of the design variable assuming that the stratification variable is uniformly distributed within strata. For the details, see Gunning and Hogan (2004). Hogan (2006) further argued that "none incorporated the equality of the coefficients of variation in deriving their algorithms; had they done so, they would have arrived at a definitive simple algorithm of getting stratum breaks using the geometric progression: free of implementation problems caused by having to choose initial values for iterative procedures, free of convergence problems inherent in iterative procedures, and free of the arbitrariness of grouping into initial classes." However, Kozak and Verma (2006) pointed out that the geometric stratification algorithm does not pursuit the optimization but the equalization of the stratum CVs, thus not necessarily being (near) optimal. They further addressed that a construction of a take-all stratum is not considered in the development of the geometric stratification.

In this paper, we revisit the above issues. In Section 2, we briefly overview the stratified random sampling design. In Section 3, we discuss two optimum stratification algorithms, the LH algorithm and its alterative proposed by Kozak (2004) and Kozak and Verma (2006) and also discuss the geometric stratification algorithm. In Section 4, we compare these algorithms concerning their convergence and near-optimality using artificial numerical examples. A summary is given in Section 5.

2. Stratified Random Sampling Design

Consider that a sample of size n is to be selected from the population $U = \{i : i = 1, \dots, N\}$ of size N to estimate a population mean $\bar{X} = (x_1 + \cdots + x_N)/N$ of the design variable x (or an auxiliary variable that may be closely related to a study variable). When the distribution of x is positively skewed, survey planners often adopt a stratified random sampling design in a way that a take-all stratum consists of the largest units and some take-some strata of the middle-size and small units. Suppose that the population is divided into H strata, which is assumed to be given a priori for simplifying our discussion. Then, the strata can be defined with a total of $H-1$ stratum boundaries $k_1 < \cdots < k_h < \cdots < k_{H-1}$ as

$$U_h = \{i : k_{h-1} < x_i \le k_h\} \tag{2.1}$$

of size N_h, where $k_0 = -\infty$, $k_H = \infty$ and $N = \sum_{h=1}^{H} N_h$.

The largest stratum H is taken as the take-all stratum and a simple random sample of size

$$n_h = (n - N_H)a_h \tag{2.2}$$

is selected independently from each of the $H-1$ take-some strata with the sample allocation rule a_h. For example, under Neyman allocation, the sample allocation rates for the $H-1$ take-some strata are determined as

$$a_h = \frac{N_h S_h}{\sum\limits_{h=1}^{H-1} N_h S_h} \tag{2.3}$$

and, under power allocation,

$$a_h = \frac{(N_h \bar{X}_h)^p}{\sum\limits_{h=1}^{H-1} (N_h \bar{X}_h)^p} \tag{2.4}$$

for $0 < p \leq 1$. The total sample size is $n = \sum_{h=1}^{H-1} n_h + N_H$.

A stratified mean estimator is $\bar{x}_{str} = \sum_{h=1}^{H} W_h \bar{x}_h$ and its variance is

$$V_{str}(\bar{x}_{str}) = \sum_{h=1}^{H-1} W_h^2 V_{srs}(\bar{x}_h), \tag{2.5}$$

where $W_h = N_h/N$ is the stratum weight, $V_{srs}(\bar{x}_h) = (1 - f_h) n_h^{-1} S_h^2$ is the variance of the stratum sample mean $\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi}$, $f_h = n_h/N_h$ is the stratum sampling fraction, and $\bar{X}_h = N_h^{-1} \sum_{i=1}^{N_h} x_{hi}$ and $S_h^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2$ are the stratum mean and variance of x, respectively.

Expression (2.5) indicates that the efficiency of the stratified random sampling design depends on (1) the choice of the stratum boundaries(k_h) and (2) the sample allocation(n_h), since the formers determines the stratum homogeneity breaks(S_h^2) of the finite population U and the latter the sample information shares among strata(a_h). Furthermore, by solving $V_{str}(\bar{x}_{str})$ in (2.5) for n, the coefficient of variation of the mean estimator $\bar{x}_{str}$ can be obtained at the level of c with the total sample size

$$n = N_H + \frac{\sum\limits_{h=1}^{H-1} \dfrac{W_h^2 S_h^2}{a_h}}{c^2 \bar{X}^2 + \sum\limits_{h=1}^{H-1} \dfrac{W_h S_h^2}{N}}. \tag{2.6}$$

That is, $V_{str}(\bar{x}_{str}) = c^2 \bar{X}^2$ with the sample size n in (2.6).

3. Stratification Algorithms

3.1. The LH algorithm

Lavallée and Hidiroglou (1988) suggested an iterative algorithm that searches for stratum boundaries that minimizes n in (2.6) (for a required CV at c). Note that n is a function of stratum quantities $W_h, \bar{X}_h, S_h$ and allocation rule a_h. If the allocation rule a_h is determined based on stratum quantities as in (2.3) or (2.4), then n is also a function of the $H - 1$ stratum boundaries $\mathbf{k} = (k_1, \ldots, k_h, \ldots, k_{H-1})'$:

$$n = n(\mathbf{k}), \tag{3.1}$$

since U_h and thus k_h in (2.1) determines the aforementioned stratum quantities. Therefore, the optimum boundaries can be obtained by solving the first derivatives of n with respect to k_h at zero:

$$\frac{\partial n(\mathbf{k})}{\partial k_1} = \cdots = \frac{\partial n(\mathbf{k})}{\partial k_h} = \cdots = \frac{\partial n(\mathbf{k})}{\partial k_{H-1}} = 0. \tag{3.2}$$

Equations (3.2) can be rewritten as a series of quadratic equations in k_h in the following form:

$$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0. \tag{3.3}$$

The larger roots of equations (3.3) are taken as the solutions that can be obtained iteratively because n and thus the coefficients α_h, β_h and γ_h are all functions of $\mathbf{k}$.

For a given set of initial boundaries $\mathbf{k}^{(0)} = (k_1^{(0)}, \ldots, k_h^{(0)} \ldots, k_{H-1}^{(0)})'$, the coefficients $\alpha_h^{(0)}$, $\beta_h^{(0)}$ and $\gamma_h^{(0)}$ are computed first based on $U_h(\mathbf{k}^{(0)})$ and the boundaries are then replaced by

$$k_h^{(1)} = \frac{-\beta_h^{(0)} + \sqrt{\left(\beta_h^{(0)}\right)^2 - 4\alpha_h^{(0)}\gamma_h^{(0)}}}{2\alpha_h^{(0)}}.$$

This process is continued until $\mathbf{k}^{(r)}(r = 1, 2, \ldots)$ converge. See Lavallée and Hidiroglou (1988) and Rivest (2002) for the details.

3.2. The random search algorithm

Kozak (2004) and Kozak and Verma (2006) suggested an alternative optimization algorithm to the LH algorithm, where, in each iteration, a stratum and its boundary are changed in a random fashion provided that a chosen objective function gets improved. The algorithm can be described using an alternative definition of the strata based on the indices of units in the population as follows:

$$U_h = \{i : b_{h-1} < i \leq b_h\}, \tag{3.4}$$

where $b_0 = 0$ and $b_H = N$. In other words, the strata can be defined using the indices that are assigned to units in the ascending order of the associated x values in the population U. Both k_h and b_h in (2.1) and (3.4) are related as follows:

$$\frac{b_h}{N} \leq F_N(x \leq k_h) < \frac{b_h + 1}{N}, \tag{3.5}$$

where $F_N(x \leq q)$ denotes the finite population cumulative distribution function of x at $q \in (-\infty, \infty)$. It is clear from expression (3.5) that both n and $V_{str}(\bar{x}_{str})$ are all functions of $\mathbf{b} = (b_1, \ldots, b_h, \ldots, b_{H-1})'$.

For a given set of the initial stratum boundaries $\mathbf{b}^{(0)}$, the random search algorithm updates the set $\mathbf{b}^{(r-1)}$ iteratively as follows: first, a stratum h is randomly selected among the first $H-1$ (take-some) strata. Second, the right-end boundary of the stratum h is replaced by $b_h^{(r-1)} + j$ if the objective function gets improved with the resulting strata, where j is a random integer selected at each iteration from $-p$ to p excluding zero (*i.e.*, no change) and p is a predetermined positive integer. Kozak (2004) uses the sample size as its objective function for a given level c of CV:

$$n = n(\mathbf{b}). \tag{3.6}$$

This random search algorithm is simple and flexible in a sense that its objective function can be replaced by

$$V = V_{str}(\bar{x}_{str}|\mathbf{b}) \tag{3.7}$$

for a given sample size n. See Kozak and Verma (2006).

3.3. Geometric stratification algorithm

Gunning and Horgan (2004) developed a very simple algorithm to determine stratum boundaries which require no iteration process. Their motivation is based on an observation by Cochran (1961)

Table 4.1. Summary statistics for four populations

Population	Minimum	Maximum	Skewness	Mean	Variance
1	10.081	1,695.214	2.137	187.984	179.842
2	86.902	3,272.735	0.812	991.290	438.977
3	1.818	1,173.212	3.757	88.867	109.594
4	10.243	257.549	1.474	61.714	32.258

that "with near optimum boundaries the coefficients of variation are often found to be approximately the same in all strata." That is,

$$\frac{S_1}{\bar{X}_1} = \frac{S_1}{\bar{X}_1} = \cdots = \frac{S_H}{\bar{X}_H}. \tag{3.8}$$

Assuming further that the distribution of x is uniform within strata, they arrived at the following relationship $k_h^2 = k_{h+1}k_{h-1}$ and thus the stratum boundaries are defined in terms of the geometric progression:

$$k_h = ar^h, \tag{3.9}$$

where $a = x_{\min}$ and $r = (x_{\max}/x_{\min})$ for $h = 0, 1, \ldots, H$.

4. Numerical Examples

In this section, we examine how efficient the geometric stratification is when its boundaries are used for the initial evaluation in implementing the LH algorithm, and how the optimal strata are formed and the sample is allocated into the strata.

4.1. Finite populations

For the study, we generated four populations of the same size $N = 3{,}000$. The first two populations were created from gamma distributions and the other two from lognormal distributions with each of the pair being slightly different in their skewness. Table 4.1 presents their summary statistics and Figure 4.1 displays their histograms. The four populations are all positively skewed with their skewness ranged from 0.812 to 3.757. Populations 1 and 3 are all of the exponential density type. Populations 2 and 4 are all unimodal with two tails.

4.2. Efficiency of the geometric stratification in implementing the LH algorithm

Four populations are all divided into $H = 5$ strata. To examine the efficiency of the geometric stratification in implementing the LH algorithm, we used two sets of the initial stratum boundaries following Gunning *et al.* (2008): one by equal-size stratification(ES) and the other by the geometric progression(GP). We implemented the LH algorithm by using these two sets as the initial boundaries to obtain the optimal boundaries(LH-ES, LH-GP) with Neyman allocation for a target CV of 1%. Table 4.2 displays the resulting stratum boundaries. Figure 4.2 also compares those boundaries for each of the four populations by stratification algorithms. Different starting boundaries lead to the same ultimate boundaries, indicating no dependency of the ultimate boundaries to a choice of initial boundaries. The LH algorithm gives the same ultimate boundaries regardless of a choice of the initial boundaries and of the populations. In addition, the largest boundaries of the geometric stratification are closer to the optimal boundaries, while the smaller boundaries of the equal-size stratification are closer to the optimal boundaries. Different from the observation of Gunning *et al.*

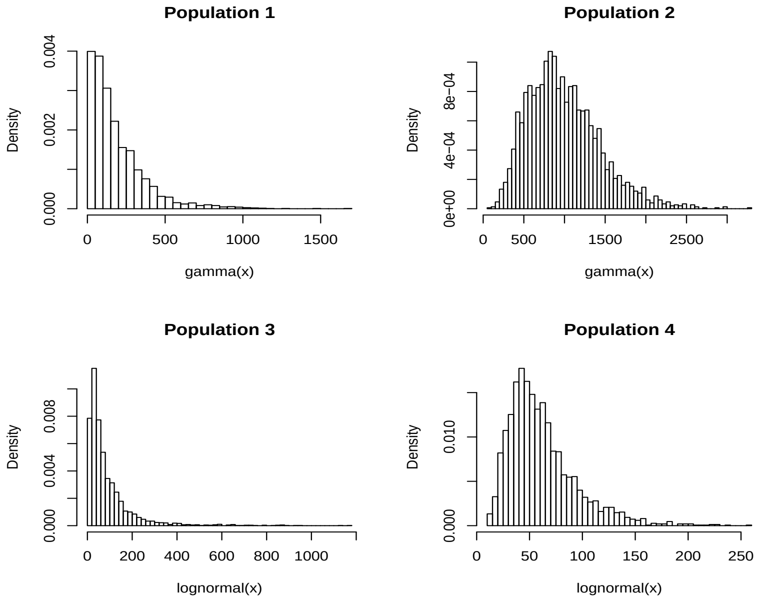

Figure 4.1. Histograms of four populations

Table 4.2. Comparisons of stratum boundaries

Population	Stratification Algorithm	Stratum boundaries			
		k_1	k_2	k_3	k_4
1	ES	50.198	102.585	171.513	293.710
	GP	28.096	78.306	218.242	608.249
	LH-ES	83.474	179.144	311.467	560.227
	LH-GP	83.474	179.144	311.467	560.227
2	ES	604.165	829.353	1,052.275	1,330.809
	GP	179.559	371.008	766.583	1,583.926
	LH-ES	680.081	1,028.411	1,466.065	2,531.188
	LH-GP	680.081	1,028.411	1,466.065	2,531.188
3	ES	23.737	41.751	70.534	126.734
	GP	6.632	24.186	88.207	321.692
	LH-ES	40.339	84.363	154.369	282.880
	LH-GP	40.339	84.363	154.369	282.880
4	ES	36.285	47.797	62.046	82.531
	GP	19.521	37.204	70.905	135.135
	LH-ES	39.661	59.358	89.183	164.322
	LH-GP	39.661	59.358	89.183	164.322

(2008), there is no clear evidence that the geometric strata(GP) are near the optimal(LH).

Recalling that the geometric stratification is to equalize stratum CVs, it would also be interesting to see how strata are formed by the other two algorithms under consideration as well. Figure 4.3 provides the CVs of the strata constructed by each algorithm for the four populations. Again, different from the observation of Gunning *et al.* (2008), near-equal CVs do not appear in the optimum stratification. Instead, the equal-size strata(ES) and the optimal strata(LH-ES and LH-GP) are more similar in their patterns in the stratum CVs.

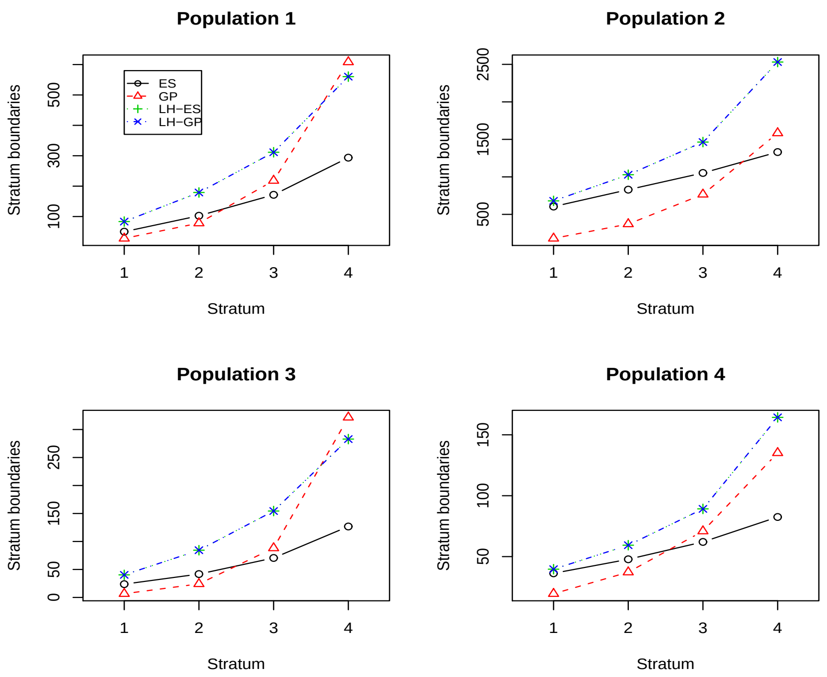

Figure 4.2. Stratum boundaries

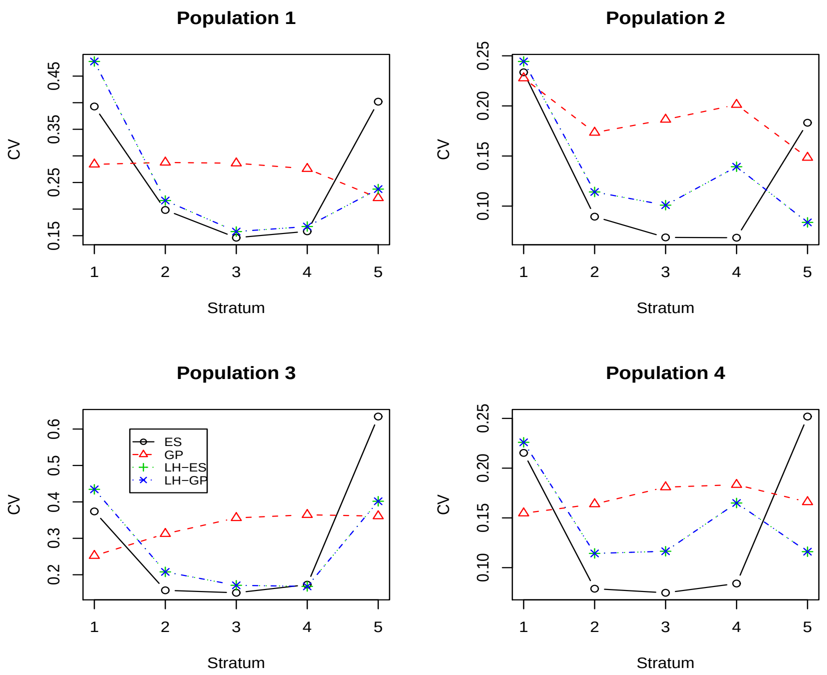

Figure 4.3. Stratum coefficients of variation

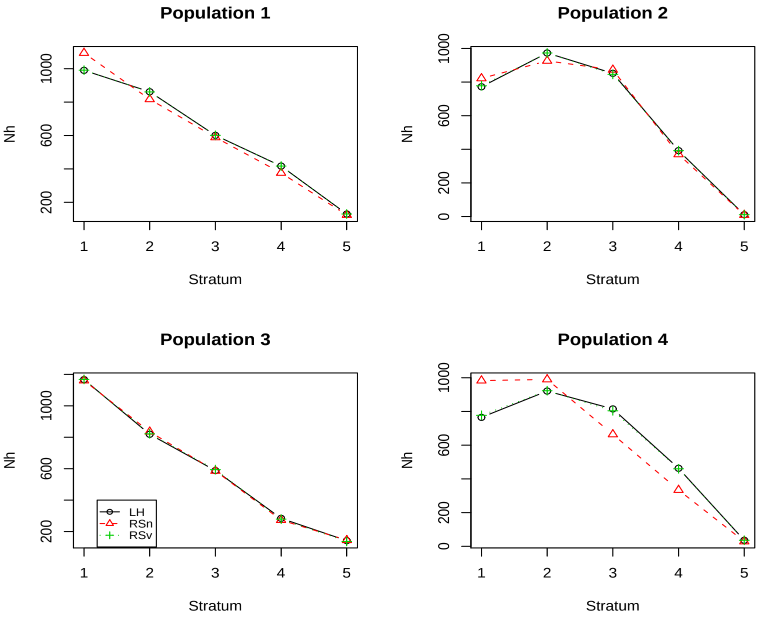

Figure 4.4. Stratum sizes

4.3. Near optimality

We compare three optimization approaches, the LH algorithm(LH) and the random search algorithm with differing in their objective functions using Neyman allocation. The first two approaches are to minimize the total same size for the required CV level, while the last approach is to minimize the variance of the stratified sample mean for the required sample size: that is,

(1) LH: to minimize $n = n(\mathbf{k})$ for a given level c of CV,

(2) RSn: to minimize $n = n(\mathbf{b})$ for a given level c of CV,

(3) RSv: to minimize $V = V_{str}(\bar{x}_{str}|\mathbf{b})$ for a given sample size n.

Due to difference in the objective function, we first obtained the sample sizes by the LH algorithm and the first type of the random search algorithm(RSn) for 1% level of CV, respectively. In order to make comparable results, we used the sample size determined from the LH algorithm to get the optimum stratum boundaries for the second type of the random search algorithm(RSv). Table 4.3 gives the results by each of the three optimum stratification approaches including sample size, variance and stratum sizes for each population. It can be observed that the first two approaches(LH and RSn) do not necessarily produce the same sample sizes for all of the four populations. They obtain the same sample sizes for Populations 1 and 3 but LH gives slightly smaller sample size for Population 2 but the larger size for Population 4 as compared to RSn. The variance may not be the same with the same sample size. Smaller sample size even produces smaller variance as seen in Population 4. These results are arisen from the different stratum sizes by approaches. LH tends to give slightly larger take-all stratum, while RSn tends to produce a bit smaller take-some strata. RSv tends to give the results similar to that of the LH algorithm. These patterns in stratum sizes are also illustrated in Figure 4.4. Such discrepancies in those quantities(sample size, variance,

Table 4.3. Stratum sample sizes, variances and stratum sizes by stratification approaches

Population	Stratification Approach	Sample size	Variance	Stratum Size				
				N_1	N_2	N_3	N_4	N_5
1	LH	384	3.5260	990	862	601	417	130
	RSn	384	3.5314	1078	836	585	372	129
	RSv	384	3.5260	991	861	601	417	130
2	LH	172	98.4109	772	973	852	392	11
	RSn	173	98.1198	782	992	814	404	8
	RSv	172	98.4098	773	973	850	393	11
3	LH	375	0.7893	1167	819	588	284	142
	RSn	375	0.7892	1172	814	590	284	140
	RSv	375	0.7887	1170	820	592	284	137
4	LH	214	0.3810	765	921	815	463	36
	RSn	212	0.3808	975	1008	663	325	29
	RSv	214	0.3809	769	916	816	463	36

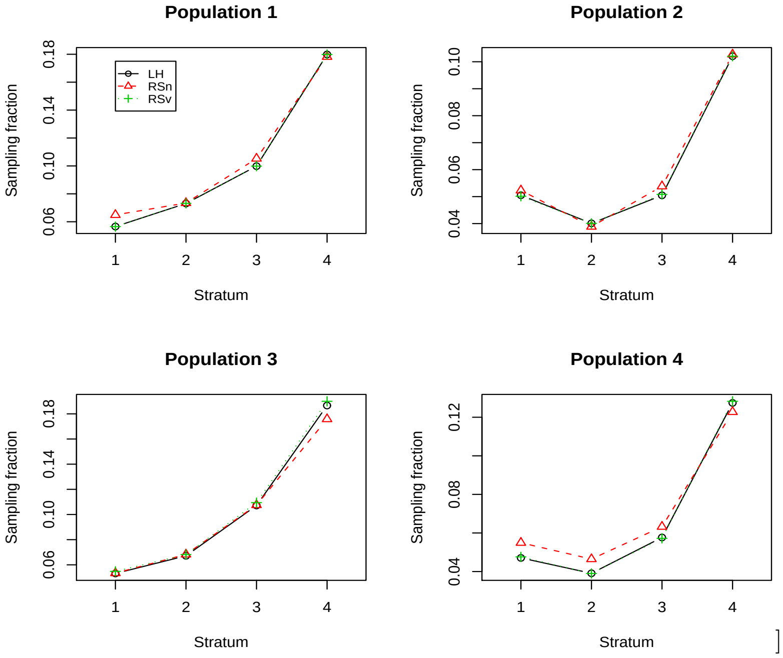

Figure 4.5. Sampling fractions by strata

stratum sizes) show that a potential improvement of the optimum stratification process may be marginal, similarly seen in Slanta and Krenzke (1996). Interestingly, the shapes of the stratum sizes are similar to those of the corresponding population distributions shown in Figure 4.1.

Another interesting observation to be pointed out is that the pattern of the relative sizes of the sampling fractions is about opposite to that of the corresponding stratum sizes. Figure 4.5 present the sampling fraction for each stratum by approaches. Therefore, the sample sizes and the allocation rates for the take-some strata tend to be roughly similar in size regardless of the population as seen in Table 4.4 and 4.5, respectively.

Table 4.4. Variance and stratum sample sizes by stratification approaches

Population	Stratification Approach	Variance	Sample Size				
			n_1	n_2	n_3	n_4	n_5
1	LH	3.5260	56	63	60	75	130
	RSn	3.5314	67	63	61	64	129
	RSv	3.5260	56	63	60	75	130
2	LH	98.4109	39	39	43	40	11
	RSn	98.1198	40	41	40	44	8
	RSv	98.4098	39	39	43	40	11
3	LH	0.7893	62	55	63	53	142
	RSn	0.7892	64	54	63	54	140
	RSv	0.7887	64	55	65	54	137
4	LH	0.3810	36	36	47	59	36
	RSn	0.3808	52	48	43	40	29
	RSv	0.3809	36	35	47	60	36

Table 4.5. Variance and stratum sample allocation rates by stratification approaches

Population	Stratification Approaches	Variance	Sample Size				
			a_1	a_2	a_3	a_4	a_5
1	LH	3.5260	0.220	0.248	0.236	0.295	-
	RSn	3.5314	0.263	0.247	0.239	0.251	-
	RSv	3.5260	0.220	0.248	0.236	0.295	-
2	LH	98.4109	0.242	0.242	0.267	0.248	-
	RSn	98.1198	0.242	0.248	0.242	0.267	-
	RSv	98.4098	0.242	0.242	0.267	0.248	-
3	LH	0.7893	0.266	0.236	0.270	0.227	-
	RSn	0.7892	0.272	0.230	0.268	0.230	-
	RSv	0.7887	0.269	0.231	0.273	0.227	-
4	LH	0.3810	0.202	0.202	0.264	0.331	-
	RSn	0.3808	0.284	0.262	0.235	0.219	-
	RSv	0.3809	0.202	0.197	0.264	0.337	-

5. Summary

Due to its simplicity in the formation of the stratum boundaries, the geometric stratification is very handy and appealing to survey planners. However, no clear evidence is shown with our numerical examples in Section 4 that the stratum boundaries by the geometric stratification algorithm are not (near) optimal, indicating that the equality of the stratum coefficients of variation may not be a better strategy in pursuing the optimum stratification. Instead, the shapes of the stratum sizes are rather similar to those of the corresponding population distributions. Furthermore, the sample allocation rates and the stratum sample sizes tend to be roughly the same in their magnitude for the take-some strata.

In addition, numerical difficulties seen in some of the literature may not be serious much in implementing the LH algorithm. However, slow convergence problems were found when comparing various optimization approaches. This observation may arise from the fact that both the sample size and the variance of the stratified sample mean are functions of stratum boundaries and sample sizes, which are both discrete in nature.

References

Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries, *Bulletin of the International Statistical Institute*, **32**, 345–358.

Gunning, P. and Horgan, J. M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations, *Survey Methodology*, **30**, 159–166.

Gunning, P., Horgan, J. M. and Keogh, G. (2008). An implementation strategy for efficient convergence of the Lavallée and Hidiroglou stratification algorithm, *Journal of Official Statistics*, **24**, 213–228.

Horgan, J. M. (2006). Stratification of skewed populations, *International Statistical Review*, **74**, 67–76.

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

Kozak, M. and Verma, M. R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency, *Survey Methodology*, **32**, 157–163.

Lavallée, P. and Hidiroglou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

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

Sigman, R. and Monsour, N. (1995). Selecting samples from list frame of business, In *Business Survey Methods*, John Wiley & Sons, New York.

Slanta, J. and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey, *Survey Methodology*, **22**, 65–75.