Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 213–228

An Implementation Strategy for Efficient Convergence of the Lavallée and Hidiroglou Stratification Algorithm

Patricia Gunning[1], *Jane M. Horgan*[1], *and Gary Keogh*[1]

The iterative procedure of Lavallée and Hidiroglou (1988) for stratifying skewed populations into a certainty stratum, where all the units are examined, and a number of noncertainty strata, which are sampled, has been found to have convergence problems. In this article we present a strategy for implementing the algorithm, which improves its convergence and in many cases results in smaller sample sizes than those obtained with the traditional implementation.

Key words: Boundaries; geometric progression; Neyman allocation; noncertainty stratum; power allocation; skewness.

1. Introduction

Lavallée and Hidiroglou (1988) have presented an iterative procedure for stratifying skewed populations into a certainty stratum, where all the units are examined, and a number of noncertainty strata, which are sampled. The stratum boundaries are derived in terms of a known auxiliary variable assumed to be closely related to the information being collected by the survey. The algorithm starts with a set of initial boundaries and replaces them iteratively, using a procedure suggested by Sethi (1963), until boundaries are obtained that minimise the sample size for a given level of precision.

Many researchers have encountered numerical difficulties when using this Lavallée and Hidiroglou algorithm. Detlefsen and Veum (1991) showed that the final boundaries depend on where the initial boundaries are set, often so that the minimum sample size attained is a local but not necessarily a global minimum. Slanta and Krenzke (1994) reported slow convergence, also not converging to the true minimum sample size. Rivest (2002) had similar problems, with failure to reach the global minimum sample size.

In this article we propose a strategy for improving the convergence of the Lavallée and Hidiroglou algorithm. The key lies in the choice of initial boundaries, which we simply take in geometric progression. We illustrate using real populations that the convergence and efficiency of the algorithm improves with the new strategy. In Section 2 we give a brief overview of stratification, and in Section 3 we outline the implementation of the Lavallée and Hidiroglou algorithm. In Section 4 we describe our new strategy, and in

[1] School of Computing, Dublin City University, Dublin 9, Ireland. Email: pgunning@computing.dcu.ie; jhorgan@computing.dcu.ie; gkeogh@computing.dcu.ie

Acknowledgments: This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology. The constructive suggestions of the referees greatly improved the article.

Section 5 we compare its performance with that of the traditional implementation of the algorithm. A summary of the conclusions is given in Section 6.

2. Stratification

A stratified sample design partitions a population of size N into L mutually exclusive strata, containing N_h $(h = 1, 2, \ldots, L)$ units. The population mean is

$$\bar{X} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi} \tag{1}$$

where X_{hi} is the value of the i^{th} unit in the h^{th} stratum and $N = \sum_{h=1}^{L} N_h$.

From each stratum a simple random sample of size $n_h \leq N_h$ is drawn without replacement. The total sample size is the sum $n = \sum_{h=1}^{L} n_h$ of the units selected from each stratum.

The mean of the sample selected from stratum h is

$$\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi} \tag{2}$$

where x_{hi} is the value of the i^{th} unit selected from the h^{th} stratum. The overall stratified sample mean is obtained by

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h \tag{3}$$

where W_h is the weight of stratum h, given by

$$W_h = \frac{N_h}{N} \tag{4}$$

It is easy to show (Cochran 1977) that (3) is an unbiased estimator of the population mean $\bar{X}$, with variance given by

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{x}_h) \tag{5}$$

Simple random sampling of n_h units from N_h yields

$$V(\bar{x}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \tag{6}$$

where S_h^2 is the variance of the h^{th} stratum:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 \tag{7}$$

and $\bar{X}_h$ is the mean of all the units in the h^{th} stratum:

$$\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} X_{hi} \tag{8}$$

Putting (6) into (5) we have

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \tag{9}$$

For any given number of strata, L, stratification designs may differ with respect to

1. where the stratum boundaries are put;
2. how the sample is allocated among the strata.

Both the boundaries and the sample allocation are chosen either to minimise (9), the variance of the sample mean, for a fixed sample size, or to minimise the sample size for a fixed precision of the mean.

3. The Lavallée and Hidiroglou Algorithm

Designed specifically for stratifying skewed populations, the Lavallée and Hidiroglou algorithm is an iterative procedure which arranges the data in ascending order, and obtains L strata defined by the cut-off points $k_0 < k_1 < k_2 <, \ldots, < k_{L-1} < k_L$ where $k_0 = min(X)$ and $k_L = max(X)$. The boundary point k_{L-1} creates the top stratum for which all of the N_L units are examined. The remaining $n - N_L$ units are allocated to the $L - 1$ strata. It is obviously assumed that $n > N_L$.

The objective is to choose the boundaries to minimise the sample size n for a given level of precision usually stated by requiring the coefficient of variation, $cv(\bar{x}_{st})$, of the sample stratified mean $\bar{x}_{st}$ to be a specified level between 1% and 10%, where

$$cv(\bar{x}_{st}) = \frac{\sqrt{V(\bar{x}_{st})}}{\bar{X}} \tag{10}$$

With a top certainty stratum, the variance in (9) may be written as

$$V(\bar{x}_{st}) = \sum_{h=1}^{L-1} W_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \tag{11}$$

If we denote $a_h = n_h/(n - N_L)$, the proportion of the $n - N_L$ sampling units allocated to the stratum h, $1 \le h \le L - 1$, then a_h satisfies $\sum_{h=1}^{L-1} a_h = 1$, $n_h = a_h(n - N_L)$ and (11) can be written as

$$V(\bar{x}_{st}) = \sum_{h=1}^{L-1} W_h^2\left(1 - \frac{(n - N_L)a_h}{N_h}\right)\frac{S_h^2}{(n - N_L)a_h} \tag{12}$$

Solving (12) for the sample size n yields:

$$n = N_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_h^2 / a_h}{V(\bar{x}_{st}) + \sum_{h=1}^{L-1} W_h S_h^2 / N} \tag{13}$$

Writing $V(\bar{x}_{st}) = \bar{X}^2 cv(\bar{x}_{st})^2$, (13) becomes:

$$n = N_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_h^2 / a_h}{\bar{X}^2 cv(\bar{x}_{st})^2 + \sum_{h=1}^{L-1} W_h S_h^2 / N} \tag{14}$$

Lavallée and Hidiroglou (1988) suggested that, in (14), n be treated as a function of the stratum boundaries $k_1, k_2, \ldots, k_{L-1}$. Then the optimum boundary points, i.e., those that minimise n for a given $cv(\bar{x}_{st})$, are obtained by getting the partial derivatives of (14) with respect to k_h and setting them equal to zero:

$$\frac{\partial n}{\partial k_1} = \frac{\partial n}{\partial k_2} = \cdots = \frac{\partial n}{\partial k_{L-1}} = 0 \tag{15}$$

From (15) Lavallée and Hidiroglou (1988) obtained a series of what look like quadratic equations for the boundary points k_h:

$$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0, \quad 1 \leq h \leq L-1 \tag{16}$$

The solution of these equations, however, is no easy matter, since the coefficients α_h, β_h and γ_h are functions of W_h, S_h, $\bar{X}_h$ and the allocation method a_h, which in turn depend not only on k_h but also on k_{h-1} and k_{h+1}. The situation is much more complicated than when it comes to solving a quadratic equation. Lavallée and Hidiroglou (1988) provided the following iterative procedure for solving these equations.

1. Sort the population values in ascending order.
2. Start with arbitrary initial boundaries $k_1' < k_2' <, \ldots, < k_{L-1}'$.
3. Based on these boundaries, calculate the stratum weights W_h', the stratum means $\bar{X}_h'$and the stratum variances $(S_h')^2$ given in (4), (8), and (7), respectively, for each stratum $h = 1, 2, \ldots, L-1$.
4. The sample size n is calculated using (14).
5. All the N_L units in the top stratum are selected into the sample, and the remaining $n - N_L$ sample units are selected from the $L-1$ lower strata using an appropriate allocation method.
6. Replace the initial set of boundaries by taking the larger root of (16):

$$k_h'' = \frac{-\beta_h' + \sqrt{(\beta_h')^2 - 4\alpha_h'\gamma_h'}}{2\alpha_h'}$$

7. Repeat Steps 3, 4, 5, and 6 with the new set of boundaries, continuing until two consecutive sets are either identical or differ by negligible quantities.

In practice the stratum boundaries are derived in terms of a known auxiliary variable X. Since this is assumed to be closely related to the unknown variable being estimated, it is not unreasonable to deduce that the minimum n obtained for a fixed $V(\bar{x}_{st})$, will approximate the required minimum sample size.

There is SAS program available for the implementation of this algorithm at www.mat.ulaval.ca/pages/lpr. Steps 2 and 5 in the algorithm are discretionary. Regarding Step 2, the SAS program allows user specified starting points for the stratum boundaries, and the default situation is that the initial boundaries are taken in such a way that each stratum has the same number of population units. With respect to Step 5, the sample allocation method, Neyman (1934) showed that for a fixed sample size the variance $V(\bar{x}_{st})$ is minimised provided the n_h are allocated among the noncertainty strata so that the proportion a_h of sampled units satisfy

$$a_h = \frac{N_h S_h}{\sum_{j=1}^{L-1} N_j S_j}, \quad 1 \le h \le L-1 \tag{17}$$

Neyman allocation is commonly used when the cost of sampling each unit is constant and when the stratum variances are likely to differ substantially, which is usual in skewed populations. Since the Lavallée and Hidiroglou algorithm is designed specifically for skewed populations, one might expect Neyman allocation to be used.

Users of the algorithm have had problems with this approach. Slanta and Krenzke (1994) encountered numerical difficulties with Neyman allocation. Specifically they found convergence was slow, and that sometimes the algorithm did not converge to the true minimum sample size n. Rivest (2002) reported similar difficulties, and observed that the algorithm became more stable if Neyman allocation was replaced by power allocation. Power allocation determines the allocation n_h so that the proportion $a_h = n_h/(n - N_L)$ of sampled units satisfies

$$a_h = \frac{(N_h \bar{X}_h)^p}{\sum_{j=1}^{L-1} \left(N_j \bar{X}_j\right)^p}, \quad 1 \le h \le L-1 \tag{18}$$

where p is some value in [0,1].

Power allocation is the method of allocation most commonly invoked by users of the Lavallée and Hidiroglou algorithm; indeed it was also the preferred option of Lavallée and Hidiroglou themselves, who noted that power allocations have the

> "peculiarity that under relatively simple assumptions and for a suitable choice of p, the coefficients of variation for the non-certainty strata tend to be equalized without a significant increase in the overall coefficient of variation."

Lavallée and Hidiroglou (1988) tested their algorithm by taking $p = 0.25,\ 0.50$ and 1 in (18) and showed that the variation in the value of p has only a minor effect on the resulting sample size for any given level of accuracy.

In what follows we propose a strategy for the Lavallée and Hidiroglou algorithm which uses Neyman allocation and which overcomes the convergence problems experienced by users.

4. Geometric Starts

Lavallée and Hidiroglou (1988) observe that

> "for skewed populations stratum coefficients of variation tend to be equalised with optimal design."

When we (Gunning and Horgan 2004) investigated the consequence of this assumption, we made a curious discovery: setting the coefficients of variation in each stratum $cv_h = S_h/\bar{X}_h$ equal, i.e.,

$$\frac{S_1}{\bar{X}_1} = \frac{S_2}{\bar{X}_2} = \ldots = \frac{S_L}{\bar{X}_L} \tag{19}$$

produces boundaries that are in geometric progression. We briefly outline the argument.

Following Dalenius and Hodges (1959), we assume that X is approximately uniformly distributed in each stratum. Uniform density of X in stratum h implies

$$\bar{X}_h \approx \frac{k_h + k_{h-1}}{2} \tag{20}$$

and

$$S_h \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1}) \tag{21}$$

The coefficient of variation of stratum h is therefore

$$cv_h = \frac{S_h}{\bar{X}_h} \approx \frac{2(k_h - k_{h-1})}{\sqrt{12}(k_h + k_{h-1})} \tag{22}$$

With approximately equal cv_h it follows that

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}} \tag{23}$$

which reduces to

$$k_h^2 = k_{h+1}k_{h-1} \tag{24}$$

and means that the stratum boundaries are the terms of a geometric progression.

Gunning, Horgan, and Keogh (2006) generalised this result and showed that when the data follow a Pareto distribution (Evans, Hastings and Peacock 2000), commonly used to model skewed distributions, geometric breaks give exactly equal coefficients of variation in the different strata.

Lavallée and Hidiroglou (1988) were not alone in their assertion that optimum strata have equal coefficients of variation. Cochran (1961) also observed that for skewed populations:

> "with near-optimum boundaries the coefficients of variation are often found to be the same in each stratum."

Our proposed new strategy is to start with geometric breaks and use Neyman allocation.

5. The Comparisons

In this section, we implement our new strategy on real skewed data, and compare its performance with some of the traditional implementations of the Lavallée and Hidiroglou algorithm.

5.1. The Data

The data used in the comparisons are four specific positively skewed populations, detailed in Gunning and Horgan (2004):

1. An accounting population of debtors in an Irish firm (Population 1);
2. The number (in thousands) of inhabitants of U.S. cities (Population 2);
3. The number of students in four-year U.S. colleges (Population 3); and
4. The resources (in millions) of dollars of a large commercial bank in the U.S. (Population 4).

These four populations are illustrated and summarised in Table 1 in decreasing order of skewness.

The populations given in Table 1 are divided into 4, 5, and 6 strata using two different sets of starting points:

1. Geometric starting points where the initial stratum breaks are chosen so that they are in geometric progression;
2. Default starting points where the initial stratum breaks are chosen so that each stratum has the same number of units.

These breaks are presented in Table 2, along with the coefficients of variation of the strata (cv_h).

Examining Table 2, we note that the actual initial boundaries obtained with the two methods differ considerably.

The default initial bounds of the Lavallée and Hidiroglou algorithm method put 25% of the populations in each of the four strata when $L = 4$, 20% in each of the five strata when $L = 5$ and approximately 17% in each of the six strata when $L = 6$. Since the populations are not uniform, it is unlikely that the optimum stratum breaks will be such that the strata contain an equal number of units, consequently the default breaks are unlikely to be near the optimum.

On the other hand with positively skewed populations containing a large number of small units and a small number of large units, the optimum stratum breaks are more

Table 1. Summary statistics for four real populations

Population	N	Range	Skewness	Mean	Variance
1	3,369	40–28,000	6.44	838.64	3,511,827
2	1,038	10–200	2.88	32.57	924
3	677	200–10,000	2.46	1,563.00	3,236,602
4	357	70–1,000	2.08	225.62	36,274

Table 2. Stratum coefficients of variation with initial starting boundaries

Population	Starting Method	$L = 4$				$L = 5$					$L = 6$					
		1	2	3	4	1	2	3	4	5	1	2	3	4	5	6
1	Geometric															
	k_h	205	1,057	5,443		147	549	2,037	7,552		119	355	1058	5153	9397	
	% of N	42%	41%	14%	3%	31%	37%	22%	8%	2%	25%	31%	27%	11%	5%	1%
	cv_h	0.45	0.44	0.48	0.50	0.37	0.38	0.40	0.37	0.41	0.32	0.32	0.30	0.31	0.26	0.35
	Default															
	k_h	117	290	700		673	198	410	888		81	151	290	500	1088	
	% of N	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	cv_h	0.32	0.27	0.25	1.18	0.28	0.20	0.22	0.22	1.07	0.24	0.18	0.19	0.17	0.22	0.99
2	Geometric															
	k_h	20	43	93		17	32	59	108		16	27	44	73	120	
	% of N	44%	38%	13%	5%	35%	40%	13%	8%	4%	26%	41%	15%	9%	6%	3%
	cv_h	0.22	0.20	0.22	0.22	0.18	0.14	0.15	0.16	0.15	0.15	0.14	0.14	0.13	0.14	0.12
	Default															
	k_h	16	23	33		15	20	26	40		14	18	23	27	46	
	% of N	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	cv_h	0.16	0.11	0.10	0.56	0.13	0.08	0.07	0.16	0.50	0.12	0.08	0.08	0.05	0,16	0.46
3	Geometric															
	k_h	526	1386	3,653		433	941	2,043	4.434		381	727	1387	2646	5046	
	% of N	20%	51%	19%	10%	14%	38%	29%	11%	8%	11%	26%	34%	14%	8%	7%
	cv_h	0.27	0.26	0.26	0.27	0.22	0.21	0.24	0.21	0.21	0.18	0.16	0.16	0.18	0.20	0.19
	Default															
	k_h	566	911	1,673		520	763	1,080	1,973		469	673	911	1175	2508	
	% of N	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	cv_h	0.29	0.14	0.17	0.56	0.27	0.12	0.10	0.20	.49	0.25	0.10	0.09	0.07	0.21	0.42
4	Geometric															
	k_h	134	261	504		118	200	339	576		109	169	262	406	630	
	% of N	44%	30%	18%	8%	32%	32%	18%	11%	7%	25%	34%	15%	11%	10%	5%
	cv_h	0.18	0.19	0.19	0.20	0.14	0.14	0.17	0.12	0.16	0.12	0.11	0.10	0.11	0.12	0.11
	Default															
	k_h	106	1,447	262		100	131	176	318		96	122	144	208	354	
	% of N	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	cv_h	0.13	0.08	0.17	0.41	0.11	0.08	0.08	0.18	0.36	0.10	0.07	0.05	0.12	0.17	0.32

likely to be such that a large proportion of the population is in the lowest stratum, and the highest stratum contains fewer but very large units. We see from Table 2 that geometric breaks do exactly that: in all cases the lowest stratum contains substantially more and the highest stratum contains substantially fewer units with geometric breaks than with the default. With geometric divisions, there is never more than 10% of the population in the top stratum and always larger proportions than the default in the lower strata. Geometric breaks tend to yield initial bounds which are higher than the default break points.

But the most important point to note from Table 2 is that geometric starting points give near equal stratum coefficients of variation cv_h in the initial strata, which is not at all the case with the default method. This is illustrated further in Figure 1.

Recalling the observation of Lavallée and Hidiroglou (1988) that stratum coefficients of variation tend to be equalised with optimum design, we might reasonably expect the geometric breaks, which give near-equal coefficients of variation, to get us near to the optimum at the first stage, avoiding the danger of the algorithm converging to a local optimum and hence allowing the use of the Neyman allocation without encountering the instability problems experienced by users of the algorithm (e.g., Rivest 2002; Slanta and Krenzke 1996). We investigate this in what follows.

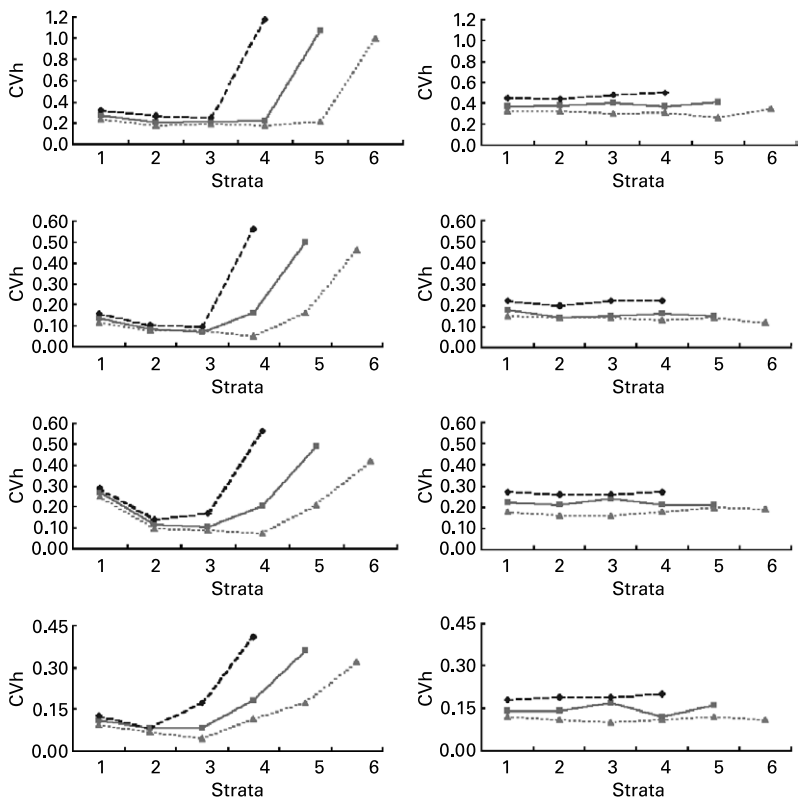

Fig. 1. Initial strata coefficients of variation with geometric and default starting points

5.2. Final Results

We implement the Lavallée and Hidiroglou algorithm using geometric starting points with Neyman allocation (*geomney*) and compare its performance with the traditional implementation which takes initial values so that the strata have an equal number of units, using two methods of allocation: Neyman (*defney*) and power (*defpower1*). When using power allocation (18), we took $p = .5$, $p = .7$, $p = .9$ and $p = 1$ and, similar to Lavallée and Hidiroglou (1988), we found that, for any given level of accuracy, the value of p has only a minor effect on the resulting sample sizes. We present the results for $p = 1$ (*defpower1*), i.e.,

$$\frac{n_h}{N - N_L} = \frac{N_h \bar{X}_h}{\sum_{j=1}^{L-1} N_j \bar{X}_j}, \quad 1 \le h \le L - 1 \tag{25}$$

This is referred to as X-proportional allocation and should be near to Neyman allocation if the Lavallée and Hidiroglou algorithm succeeds in obtaining breaks so that the stratum coefficients of variation $S_h/\bar{X}_h$ are equal.

5.2.1. The Final Break Points

Table 3 gives the final boundaries obtained with each strategy for 4, 5, and 6 strata, and with accuracy levels of $cv(\bar{x}_{st}) = .05, .025,$ and $.01$.

We see that, with $cv(\bar{x}_{st}) = .05$, the final boundaries with *geomney* and *defney* are identical in most cases. The discrepancies between *defpower1* and the other two are not large, and in most cases the top certainty strata are identical. When $cv(\bar{x}_{st}) = .025$, the differences between the final boundaries with each strategy are somewhat larger, and when $cv(\bar{x}_{st}) = .01$, substantial differences in the final break points occur.

5.2.2. Sample Sizes

While the differences in break points are interesting to observe, of greater importance is the sample size required to obtain a given level of precision with each strategy. The final sample sizes necessary for $cv(\bar{x}_{st}) = .05, .025$ and $.01$ are presented in Table 4, along with the number of iterations necessary to arrive at the final result.

We see from Table 4 that the sample sizes needed to obtain a given level of precision with *geomney* are less than or equal to those of *defney* and *defpower1* for most levels of precision, with the largest decreases in sample sizes occurring for $cv(\bar{x}_{st}) = .01$. For example, in Population 2 with four strata and $cv(\bar{x}_{st}) = .01$, *defgeom* returned a sample size of $n = 213$, compared to $n = 247$ for *defney* in Population 2, 34 units of difference. Also with six strata and $cv(\bar{x}_{st}) = .01$, sample sizes of $n = 146, 126,$ and 74 are required with *geomney* in Populations 2, 3, and 4, respectively, compared to $n = 163, 141,$ and 81, respectively, with *defney*; $n = 168, 145,$ and 81, respectively, with *defpower1*.

There was just one case in which the *defney* and *defpower1* yielded sample sizes substantially less than the *geomney*: in Population 4 with four strata and $cv(\bar{x}_{st}) = .01$, the *geomney* returned a sample size $n = 124$ compared to $n = 113$ with *defney*, and $n = 114$ with *defpower1*. Notably, this is the least skewed of the populations. Gunning and Horgan

Table 3. Final boundaries

Population	Strategy	$cv(\bar{x}_{st}) = .05$	$cv(\bar{x}_{st}) = .025$	$cv(\bar{x}_{st}) = .01$
Four strata				
1	Geomney	498, 2216, 10133	387, 1476, 5382	333, 1029, 2569
	Defney	498, 2216, 10133	367, 1476, 5382	284, 845, 2238
	Defpower1	463, 2154, 10227	345, 1422, 5386	245,793, 2221
2	Geomney	21, 53, 195	20, 41, 112	20, 33, 63
	Defney	21, 53, 195	19, 39, 110	15, 23, 45
	Defpower1	23, 59, 195	20, 43, 110	17, 30, 56
3	Geomney	1366, 3757, 9466	744, 1574, 4171	731, 1328, 2350
	Defney	1366, 3757, 9446	744, 1574, 4171	722, 1297, 2300
	Defpower1	1279, 3674, 9446	723, 3633, 4159	654, 1266, 2368
4	Geomney	174, 387, 968	150, 277, 566	141, 245, 359
	Defney	174, 387, 968	150, 277, 566	116, 171, 279
	Defpower1	174, 385, 927	148, 282, 566	112, 172, 278
Five strata				
1	Geomney	367, 1248, 3757, 13226	339, 1090, 2970, 7513	249, 670, 1565, 3288
	Defney	367, 1238, 3752, 13226	239, 1092, 2972, 7514	230, 572, 1262, 2977
	Defpower1	319, 1181, 3761, 13293	282, 995, 2831, 7685	187,495, 1192, 2938
2	Geomney	19, 34, 73, 195	19, 31, 58, 132	19, 31, 55, 91
	Defney	19, 34, 73, 195	19, 22, 42, 116	14, 21, 33, 66
	Defpower1	20, 42, 95, 195	17, 31, 58, 127	14, 21, 33, 59

Table 3. Continued

Population	Strategy	$cv(\bar{x}_{st}) = .05$	$cv(\bar{x}_{st}) = .025$	$cv(\bar{x}_{st}) = .01$
3	Geomney	742, 1534, 3807, 9466	735, 1432, 3049, 6485	579, 925, 1440, 2673
	Defney	742, 1534, 3807, 9466	740, 1505, 3566, 7204	511, 857, 1370, 2456
	Defpower1	715, 1591, 3932, 9446	709,1551, 3525, 7500	443, 798, 1362, 2463
4	Geomney	118, 195, 405, 968	118, 189, 353, 651	117, 185, 348, 503
	Defney	117, 195, 405, 968	116, 173, 289, 599	99, 129, 178,297
	Defpower1	149, 283, 557, 968	117, 195, 351, 651	98, 131, 185, 293
Six strata				
1	Geomney	269, 741, 1767, 4378, 14915	267, 732, 1688, 3700, 8894	199, 484, 1044, 2125, 3936
	Defney	240, 639, 1619, 4295, 14829	267, 732, 1687, 3700, 8893	189, 438, 849, 1722, 3551
	Defpower1	229, 678, 1753, 4679, 14961	221, 618, 1479, 3593, 8944	148, 374, 812, 1695, 3586
2	Geomney	19, 31, 57, 110, 195	16, 25, 40, 69, 144	16, 25, 40, 67, 99
	Defney	14, 31, 34, 73, 195	13, 20, 31, 58, 139	13, 17, 22, 34, 68
	Defpower1	17, 31, 57, 109, 195	14, 21,33, 60, 129	13, 17, 23, 35, 62
3	Geomney	523, 909, 1665, 4133, 9446	512, 869, 1580, 3643, 7789	511, 857, 1363, 2240, 3496
	Defney	523, 909, 1665, 4133, 9446	512, 869, 1580, 3643, 7789	428, 683, 969, 1480, 2839
	Defpower1	663, 1270, 2321, 4624, 9446	472, 865, 1645, 3623, 7692	415, 682, 991, 1541, 2699
4	Geomney	116, 172, 289, 567, 968	116, 170, 257, 387, 680	116, 170, 256, 380, 516
	Defney	116, 172, 289, 567, 968	93, 120, 172, 289,607	93, 120, 179, 256, 387
	Defpower1	117, 194, 342, 600, 968	112, 169, 257, 395, 667	92, 121, 171, 256, 384

Table 4. Final sample size and number of iterations

Population	Strategy	n	$cv(\bar{x}_{st}) = .05$ Iterations	n	$cv(\bar{x}_{st}) = .025$ Iterations	n	$cv(\bar{x}_{st}) = .01$ Iterations
Four strata							
1	Geomney	92	25	212	16	497	12
	Defney	92	30	212	25	498	29
	Defpower1	94	32	216	22	509	23
2	Geomney	36	10	88	4	213	11
	Defney	38	14	90	12	247	7
	Defpower1	37	14	89	13	219	18
3	Geomney	37	25	98	11	188	13
	Defney	37	27	98	14	187	25
	Defpower1	38	24	99	15	195	33
4	Geomney	24	18	55	9	124	8
	Defney	24	26	55	24	113	10
	Defpower1	25	20	54	21	114	8
Five strata							
1	Geomney	57	24	146	29	384	12
	Defney	57	34	146	48	384	37
	Defpower1	59	29	154	55	403	34
2	Geomney	20	8	62	5	171	6
	Defney	20	20	77	12	179	16
	Defpower1	19	24	64	19	182	9
3	Geomney	23	18	70	36	159	18
	Defney	23	23	70	53	160	14
	Defpower1	23	21	72	48	163	8
4	Geomney	17	9	41	5	103	5
	Defney	18	19	43	16	105	7
	Defpower1	15	31	41	26	104	4
Six strata							
1	Geomney	43	29	109	43	318	16
	Defney	43	44	109	65	316	52
	Defpower1	44	36	112	46	327	48
2	Geomney	11	23	53	4	146	6
	Defney	16	18	55	18	163	11
	Defpower1	12	32	55	16	168	11
3	Geomney	20	19	58	11	126	16
	Defney	20	27	58	35	143	16
	Defpower1	16	41	59	29	145	19
4	Geomney	10	12	32	6	74	6
	Defney	10	33	39	9	81	10
	Defpower1	11	36	31	30	81	9

(2004) showed that geometric break points work best when L is large and when the populations are highly skewed.

The boxplots in Figure 2 detail the differences in sample sizes between *geomney* and *defney* (*geomney-defney*) and the differences between *geomney* and *defpower1* (*geomney-defpower1*) for each of the 4, 5, and 6 strata, and for each of the accuracy levels of $cv(\bar{x}_{st}) = .01$ (cv01), $cv(\bar{x}_{st}) = .025$ (cv025), $cv(\bar{x}_{st}) = .05$ (cv05).

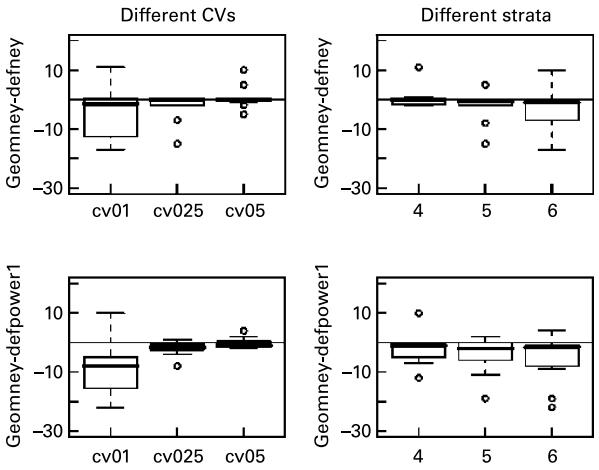

Fig. 2. Differences in sample sizes

We observe from Figure 2 that the greatest improvements in the sample sizes for *geomney* over the other two strategies occur when the number of strata is six and when $cv(\bar{x}_{st}) = .01$ and when the number of strata is six.

5.2.3. Iterations

Our final comparison concerns the convergence rate of the different strategies. The SAS program for the implementation of this algorithm, www.mat.ulaval.ca/pages/lpr, allows up to 30 iterations before coming to a halt. We changed this to 100. Looking at the number of iterations needed to converge to the optimum sample size in Table 4, we see that more than 30 iterations were necessary many times with *defney* and *defpower1*, particularly in the larger number of strata, while the *geomney* strategy used less than 30 in all cases.

We see from Table 4 that the number of iterations necessary to converge is smaller in most cases with *geomney*. The overall average number of iterations used with *geomney* is calculated to be 14.7, while it is 24.5 with *defney* and 25.1 with *defpower1*, with increases of 67% for *defney* and 71% in *defpower1* over *geomney*. Duncan's multiple comparison test indicated that these differences are highly significant ($p \leq .001$), while the mean difference between the two default strategies *defney* and *defpower1* is not significant.

Of course an increase in the number of iterations necessary for convergence may not be all that important, and indeed it may even go unnoticed by the user, since the Lavallée and Hidiroglou algorithm is implemented by the computer. More important is whether or not the Lavallée and Hidiroglou algorithm converges to the true minimum sample size. Our results indicate that the geometric strategy appears to converge more quickly to a lower sample size than the two implementations of the traditional strategy that we examined.

6. Summary and Discussion

Users of the Lavallée and Hidiroglou (1988) iterative stratification algorithm have experienced convergence problems which we show can be overcome by taking initial

starting points in geometric progression and proceeding iteratively using Neyman allocation at each stage to allocate the sample units among the noncertainty strata. It is appropriate that the initial geometric set of boundaries yields approximately equal stratum coefficients of variation, as it was Lavallée and Hidiroglou themselves who found equal coefficients of variation in the strata to be desirable and "often asked for by the users of survey data." What is more important, though, is that these geometric break points tend to already be close to the global optimum, and hence keep us away from the instability and convergence problems experienced in the past by users of the algorithm.

Using real populations, and dividing them into 4, 5, and 6 strata, we compare the geometric approach with the default strategy of Lavallée and Hidiroglou, which starts with initial breaks so that strata contain equal number of units. We use two methods to allocate the sample units among the strata with the default strategy: Neyman allocation and power allocation with $p = 1$.

The geometric approach converged more quickly to the optimum bounds, and in many cases resulted in smaller sample sizes than those obtained with the default strategies. The greatest improvements in sample size were observed for six strata with coefficient of variation $cv(\bar{x}_{st}) = .01$. The number of iterations needed to converge is smaller in most cases with the geometric strategy than with the default strategies.

Geometric starting points are extremely easy to implement, and with optimum allocation overcome the instability and convergence problems experienced by users of the Lavallée and Hidiroglou algorithm.

7. References

Cochran, W.G. (1961). Comparison of Methods for Determining Stratum Boundaries. Bulletin of the International Statistical Institute, 32, 345–358.

Cochran, W.G. (1977). Sampling Techniques. New York: Wiley.

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. Journal of the American Statistical Association, 54, 88–101.

Detlefsen, R., Veum, L. (1991). Design Issues for the Retail Trade Survey in the U.S. Bureau of the Census. Proceedings of the American Statistical Association, Survey Research Methods Section, 214–219.

Evans, M., Hastings, N.A.J., and Peacock, J.B. (2000). Statistical Distributions (3rd Ed.). New York: Wiley.

Gunning, P. and Horgan, J.M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. Survey Methodology, 30, 1–18.

Gunning, P., Horgan, J.M., and Keogh, G. (2006). Efficient Pareto Stratification. Mathematical Proceedings of the Royal Irish Academy, 2, 131–138.

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

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, Series A, 97, 558–606.

Rivest, L. (2002). A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys. Survey Methodology, 28, 191–198.

Sethi, V.K. (1963). A Note on the Optimum Stratification of Populations for Estimating the Population Means. Australian Journal of Statistics, 5, 20–33.

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.

Received May 2006
Revised December 2007