Obtaining Stratum Breaks in Skewed Populations Using a Simple Method

P. Gunning[‡], **J.M. Horgan**[§] and **G. Keogh**[¶]

October 19, 2004

Abstract

A new easy-to-implement algorithm for obtaining stratum breaks is compared with the cumulative root frequency method of Dalenius and Hodges (1957) and the Lavallée and Hidiroglou (1988) approximation method and found to give favourable results. The breaks obtained using this new method are used as input values for the Lavallée and Hidiroglou method and shown to improve convergence.

KEY WORDS: Efficiency; Geometric Progression; Optimum Allocation; Stratification.

1 INTRODUCTION

In 1950 Dalenius derived equations for determining optimum boundaries when stratifying variables by size, but these equations proved troublesome to solve because of dependencies among the components. Since then numerous researchers have attempted to obtain efficient approximations to this optimum solution. The first such approximation, suggested by Dalenius and Hodges (1957, 1959), constructs the strata by taking equal intervals on the cumulative function of the square root of the frequencies; this method is still frequently used today. Lavallée and Hidiroglou (1988) derived an iterative procedure, specifically for skewed populations, for obtaining the optimum boundaries such that the sample size is minimised for a given level of reliability.

Our proposed algorithm is much simpler to implement than any of those currently available and is described in Section 2. Section 3 compares the efficiency of the new approximation with the cumulative root frequency and the Lavallée and Hidiroglou approximations. Section 4 describes how the boundaries from this new algorithm are used to improve the convergence of the Lavallée and Hidiroglou method. We summarise our findings in Section 5.

2 AN ALTERNATIVE METHOD OF STRATUM CONSTRUCTION

Gunning and Horgan (2004) showed that under the following assumptions, near-optimum boundaries can be obtained in skewed populations using a geometric progression.

- the distribution within each stratum is approximately uniformly distributed (an assumption made by Dalenius and Hodges in 1959)
- the coefficients of variation $CV_h = S_{xh}/\overline{X_h}$ are approximately the same in all strata (assumption made by Dalenius and Hodges (1959), Cochran (1961) and Lavallée and Hidiroglou (1988)) i.e.

$$\frac{S_{x1}}{\overline{X}_1} = \frac{S_{x2}}{\overline{X}_2} = \ldots = \frac{S_{xL}}{\overline{X}_L} \; . \qquad (1)$$

  where S_{xh} is the standard deviation and $\overline{X}_h$ the mean of all the elements in stratum h.

The new method divides a population into L strata for any given endpoints k_0 and k_L as follows:

Letting $a = k_0$, the minimum value of the

[‡]P. Gunning, School of Computing, Dublin City University, Dublin 9, Ireland
[§]J.M. Horgan, School of Computing, Dublin City University, Dublin 9, Ireland
[¶]G. Keogh, School of Computing, Dublin City University, Dublin 9, Ireland

variable and $ar^L = k_L$, the maximum value of the variable where r is the constant ratio $(k_L/k_0)^{1/L}$, the break for stratum h, k_h, is ar^h for $h = 1, 2, \ldots, L$.

For example, taking $L = 4$, $k_0 = 5$ and a maximum value $k_4 = 50,000$, the ratio r works out as $(50,000/5)^{1/4} = 10$. The strata form the ranges:
$5 - 50;\ 50 - 500;\ 500 - 5,000;\ 5,000 - 50,000.$
This is clearly an extremely simple method of obtaining stratum breaks.

3 THE PERFORMANCE OF THE ALGORITHM

3.1 Some Real Positively Skewed Populations

To test our algorithm, we implement it on four specific positively skewed populations, three of which (populations 2, 3 and 4) are the populations that Cochran (1961) invoked to illustrate the efficiency of the cumulative root frequency method.

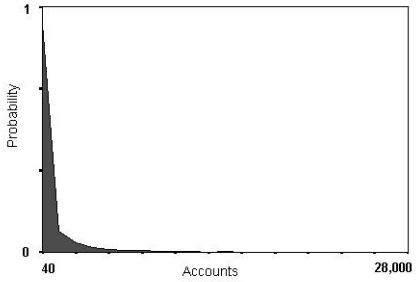
- An accounting population of debtors in an Irish firm, detailed in Horgan (2003) (Population 1).
- The population in thousands of US cities (Population 2).
- The number of students in four-year US colleges (Population 3).
- The resources in millions of dollars of a large commercial bank in the US (Population 4).

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

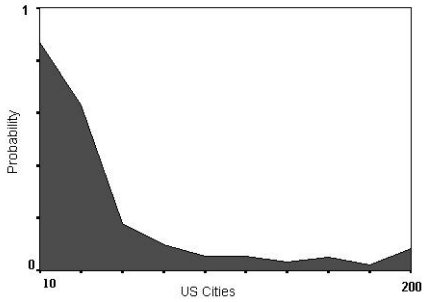
Table 1: Summary Statistics for Four Real Populations

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

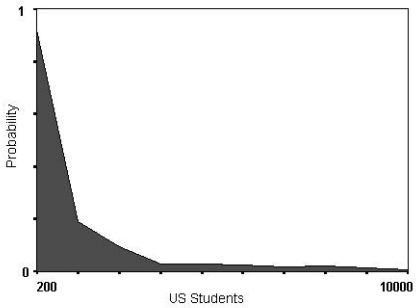
Population 1



Population 2



Population 3



Population 4

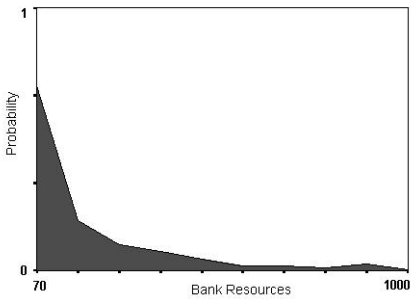


Figure 1: Four Real Populations

3.2 Comparison of New Method with the Cumulative Root Frequency Method and the Lavallée-Hidiroglou Method

The new algorithm was implemented on the populations summarised in Table 1 and compared with the cumulative root frequency (cum $\sqrt{f}$) and the Lavallée-Hidiroglou (LH) methods of stratum construction. The SAS code used for implementing the Lavallée-Hidiroglou method was obtained from the web at *http://www.ulval.ca/pages/lpr/*. Comparisons were made in terms of:

- Stratum breaks
- Equality of CVh
- Precision of estimates

3.2.1 Stratum Breaks

The three methods gave very different stratum breaks when the four populations were divided into $L = 3$, 4 and 5 strata. Table 2 shows the stratum breaks for 5 strata for Population 1 (results for all four populations are available in Gunning and Horgan, 2004).

Table 2: Comparison of Boundaries for 5 Strata using the 3 Methods on Population 1

	Stratum				
Method	1	2	3	4	5
New	147	549	2037	7552	28000
Cum $\sqrt{f}$	279	838	1677	4193	28000
LH	342	1153	3431	10301	28000

3.2.2 Equality of CVh

From Figure 2 it can be seen that for population 1 (which has the greatest skewness) the CV_h for the new method are a lot less variable than those of the cum $\sqrt{f}$ and the Lavallée-Hidiroglou methods. In the other three populations (results can be seen in Gunning and Horgan, 2004), the CV_h are not as diverse with cum $\sqrt{f}$ or the Lavallée-Hidiroglou methods but they still appear more variable than those obtained with the new method. We can conclude therefore that the new method is more successful than the cum $\sqrt{f}$ or the Lavallée-Hidiroglou method in obtaining near-equal strata CVs.

It was found that the CV_h with the new method are more homogeneous when $L = 4$ or 5 than when $L = 3$; this is to be expected since the validity of the assumption of uniformity of the distribution of elements within stratum is strengthened with increased number of strata.

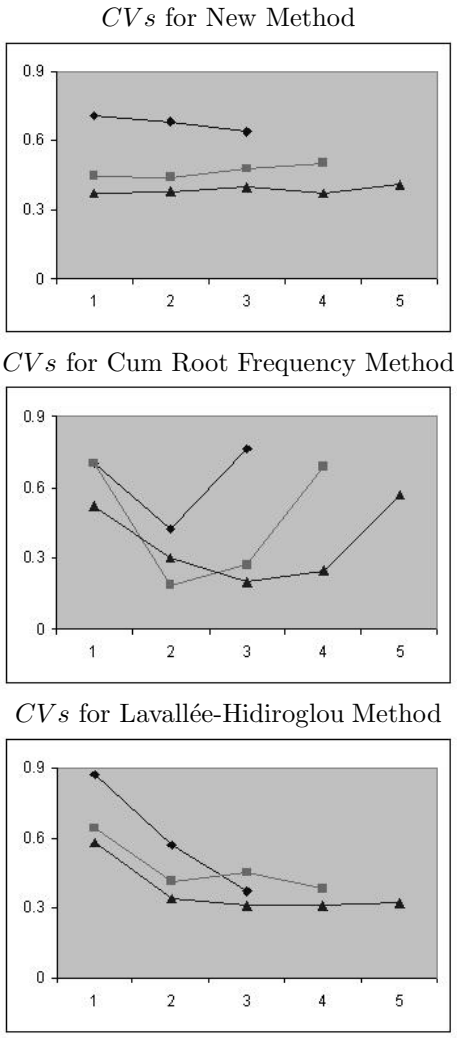


Figure 2: Variability of the between stratum CVs for the three methods for Population 1

3.2.3 Precision of Estimates

The new method was run using a sample size of $n = 100$ allocated optimally among the strata

using *Neyman allocation* (Neyman, 1934):

$$n_h = \left(\frac{N_h S_{xh}}{\sum_{i=1}^{L} N_i S_{xi}} \right) n \ , \qquad (2)$$

where n_h is the sample size for stratum h and N_h is the number of elements in the h^{th} stratum $(1 \leq h \leq L)$,

The efficiency of the new method was compared with the cumulative root frequency method in terms of relative efficiency or variance ratio which is defined as

$$eff_{cum,new} = \frac{V_{cum}(\overline{x}_{st})}{V_{new}(\overline{x}_{st})} \ , \qquad (3)$$

where $V_{cum}(\overline{x}_{st})$ and $V_{new}(\overline{x}_{st})$ are the variances of the stratified mean for the cumulative root frequency method and the new method, respectively.

Here the stratified mean is defined as:

$$\overline{x}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{x}_h, \qquad (4)$$

where $N = \sum_{h=1}^{L} N_h$ and $\overline{x}_h$ is the mean of the sample elements in the h^{th} stratum.

The variance of the stratified mean is defined as:

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_{xh}^2}{n_h}, \qquad (5)$$

where

$$S_{xh} = \sqrt{\sum_{i=1}^{N_h} \left(X_{hi} - \overline{X_h} \right)^2 / N_h},$$

is the standard deviation of X restricted to stratum h and

$$\overline{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} X_{hi}.$$

is the mean.

The Lavallée-Hidiroglou algorithm minimises the sample size n for a given level of precision. To compare the performance of the new method with the Lavallée-Hidiroglou method, the CVs from the new method are used as input for the Lavallée-Hidiroglou algorithm and the sample sizes computed.

The relative efficiencies computed for the cum $\sqrt{f}$ method may be interpreted as the proportionate increase or decrease in the sample size with the cum $\sqrt{f}$ method relative to the new method. These sample sizes along with those required by the Lavallée-Hidiroglou method to obtain the same precision as that of the new method with $n = 100$ are given in Table 3.

Table 3: Sample sizes required by the cum $\sqrt{f}$ and Lavallée-Hidiroglou methods to obtain the same precision as that of the new method with $n = 100$ for 3, 4 and 5 strata

		Population			
Strata	Method	1	2	3	4
3	Cum $\sqrt{f}$	97	99	79	116
	LH	121	123	107	100
4	Cum $\sqrt{f}$	123	119	117	103
	LH	113	117	107	93
5	Cum $\sqrt{f}$	94	168	132	117
	LH	90	136	105	99

From Table 3 it can be seen that while the new method is not always more efficient than the cumulative root frequency method, when it is, it is substantially so, and when it is not it is only marginally worse. For example, large gains in efficiency are observed when $L = 5$ in Populations 2, 3 and 4: here the samples of sizes $n = 168$, 132 and 117 are required with cum$\sqrt{f}$. While there are four cases where the sample size is less than 100, with one exception, all are greater than 90. The exception is Population 3 with $L = 3$, the smallest number of strata; the sample size in this case is 79.

The sample size required with the Lavallée-Hidiroglou algorithm is greater than 100 in all but four cases. For example, in Population 2 with 5 strata, it is necessary to increase the sample size by 36% to $n = 136$. When the sample size falls below $n = 100$, the drop is not as large. In Population 4 with 4 and 5 strata and in Population 1 with 5 strata, a sample size of $n = 93$, 99 and 90 respectively, will suffice.

These results might appear to indicate that the new method outperforms the Lavallée-Hidiroglou method in terms of the minimum

sample size required for a specified precision. We observe however that the new method unlike the Lavallée-Hidiroglou method, does not give a take-all stratum. If this is required, it is more appropriate to use the Lavallée-Hidiroglou to obtain the strata. If after such a take-all stratum has been removed the skewness remains, the new method is probably the easier and more efficient way of obtaining the remaining strata. However, often in practice the top stratum is decided judgementally.

4 Choosing Starting Boundaries for the Lavallée-Hidiroglou method

The Lavallée-Hidiroglou method starts with arbitrary initial boundaries and replaces them iteratively using a procedure suggested by Sethi (1963) until the minimum sample size is obtained for a given level of precision. The requirement on precision is usually stated by requiring the coefficient of variation to be equal to some specified level between $1\% - 10\%$.

The default initial boundaries used are equidistant points along the range. However, numerical difficulties may occur. Detlefsen and Veum (1991) found that the resulting boundaries depend on where the initial boundaries are set, so that the minimum sample size attained is a local but not necessarily a global minimum. They also found that convergence occurs faster for the lower number of strata. Rivest (2002) reported numerical difficulties, failure to reach the global minimum sample size, and non-convergence of the algorithm when the number of strata was large.

Because of the possible convergence problems of the default starting points, Slanta and Krenzke (1996) suggested that they be replaced by using the cumulative root frequency algorithm to obtain the starting points. We implemented the Lavallée and Hidiroglou method on the four populations described in Table 1 using two different sets of starting points (the default starting boundaries and the boundaries generated by the new method) for 4, 5 and 6 strata using coeffients of variation of .05, .025 and .01. The number of iterations and the sample size needed to obtain these coefficients of variation for 6 strata are given in Table 4.

Sample Sizes
The new method yields sample sizes less than or equal to that obtained with the default in most cases. The greatest improvements occur in higher number of strata. As can be seen from Table 4 in Populations 2, 3 and 4 with a $CV = .01$, $n = 146$, 126 and 74 respectively for the new method compared to $n = 163$, 143 and 81 with the default. This represents a percentage decrease in sample sizes of 10%, 12% and 9% respectively in favour of the new method starting boundaries. For Populations 1 and 4 with $CV = .025$, $n = 110$ and 32 with the new method compared to $n = 119$ and 39 with the default, a decrease of 8% and 18%.

Iterations
In most cases the set of new starting points reaches the optimum with fewer iterations than the default, and often substantially fewer. For example, in Population 3 with a $CV = .025$, the default requires 35 iterations compared to 16 for the new method to arrive at the same sample size.

It should be pointed out that an increase in the number of iterations may not be important, and indeed may even go unnoticed by the user of the Lavallée and Hidiroglou algorithm, since this work is done by the computer. More important, from the point of view of the user, is whether or not convergence occurs with a maximum of 30 iterations; if it does not, the program stops and no stratification breaks are returned. In this work, when the number of iterations exceeded 30, we adjusted the program to allow the process to continue, and counted the number of iterations necessary to arrive at a solution. It can be seen that the new method always converges in less than the 30 iterations allowed by the program, unlike the default. With the default starting points, adjustment was necessary in the following cases: Population 1 with $CV = .01$ required 49 iterations, Population 3 with $CV = .025$ required 35 iterations and Population 4 with $CV = .05$ required 33 iterations to converge.

Boundaries
It should also be pointed out that the boundaries are not always the same when different starting points are used: the discrepancies between them are greatest for the higher number of strata ($L = 6$) and the lowest coefficient of variation

($CV = .01$).

Table 4: Comparison of Starting Boundaries (Default versus New Method) for 6 Strata for the Lavallée-Hidiroglou Method

Population	CV		Default	New
1	.01	n	316	318
		Iterations	49	16
	.025	n	119	110
		Iterations	29	29
	.05	n	43	43
		Iterations	29	29
2	.01	n	163	146
		Iterations	11	6
	.025	n	55	53
		Iterations	18	4
	.05	n	16	11
		Iterations	18	23
3	.01	n	143	126
		Iterations	16	16
	.025	n	58	58
		Iterations	35	16
	.05	n	20	20
		Iterations	27	19
4	.01	n	81	74
		Iterations	10	6
	.025	n	39	32
		Iterations	9	6
	.05	n	10	10
		Iterations	33	12

5 SUMMARY

A method of stratum construction in positively skewed populations using the geometric progression is compared with the commonly used cumulative root frequency method and the Lavallée-Hidiroglou method using four positively skewed real populations divided into three, four and five strata. A greater sample size was required for the cumulative root frequency and the Lavallée-Hidiroglou methods to obtain the same precision as the new method in most cases; the greatest increase in the required sample size occurred with the largest number of strata.

The main advantage of the proposed method is that it is definitive and easier to implement. It does not have to make initial arbitrary class divisions of the population unlike the cumulative root frequency method and it does not require the creation of initial boundaries unlike the Lavallée-Hidiroglou method.

We replaced the default starting boundaries of the Lavallée-Hidiroglou method with the geometric progression starting points implementing them on the four populations for 4, 5 and 6 strata. It was found that the geometric starting points usually yielded smaller sample sizes for the same precision and quicker convergence than the default situation. The greatest improvements were recorded for the larger number of strata and higher precision levels.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology.

REFERENCES

COCHRAN, W.G. (1961). Comparison of Methods for Determining Stratum Boundaries. *Bulletin of the International Statistical Institute*, 32, 2, 345-358.

DALENIUS, T. (1950). The Problem of Optimum Stratification, *Skandinavisk Aktuarietidskrift*, 203-213.

DALENIUS, T. and HODGES, J. L. (1957). The Choice of Stratification Points. *Skandinavisk Aktuarietidskrift*, 198-203.

DALENIUS, T. and HODGES, J. L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, 88-101.

DETLEFSEN, R.E. and VEUM, C.S. (1991). Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 214-219.

GUNNING, P. and HORGAN, J.M. (2004). A New Algorithm for the Construction of

Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30, 2.

HORGAN, J.M. (2003). A List Sequential Sampling Scheme with Applications in Financial Auditing, *IMA Journal of Management Mathematics*, 14, 1-18.

LAVELLÉE, 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 Statistics Society*, 97, 558-606.

RIVEST, L.P. (2002). A Generalization of the Lavellée-Hidiroglou Algorithm for Stratiication 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 and KRENZKE (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.