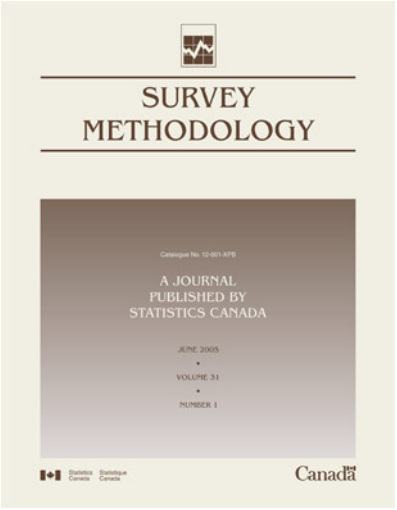
Catalogue no. 12-001-XIE

# Survey Methodology

December 2004




Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

| | |
|---|---|
| National inquiries line | 1 800 263-1136 |
| National telecommunications device for the hearing impaired | 1 800 363-7629 |
| Depository Services Program inquiries | 1 800 700-1033 |
| Fax line for Depository Services Program | 1 800 889-9734 |
| E-mail inquiries | infostats@statcan.ca |
| Website | www.statcan.ca |

## Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

# Survey Methodology

December 2004

Published by authority of the Minister responsible for Statistics Canada



April 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 30, No. 2, pp. 159-166
Statistics Canada, Catalogue No. 12-001

# A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations

**Patricia Gunning and Jane M. Horgan** [1]

Abstract

A simple and practicable algorithm for constructing stratum boundaries in such a way that the coefficients of variation are equal in each stratum is derived for positively skewed populations. The new algorithm is shown to compare favourably with the cumulative root frequency method (Dalenius and Hodges 1957) and the Lavallée and Hidiroglou (1988) approximation method for estimating the optimum stratum boundaries.

Key Words: Efficiency; Geometric progression; Neyman allocation; Stratification.

## 1. Introduction

A stratified random sampling design is a sampling plan in which a population is divided into mutually exclusive strata, and simple random samples are drawn from each stratum independently. The essential objective of stratification is to construct strata to allow for efficient estimation. In what follows $X$ represents the known stratification or auxiliary variable while $Y$ represents the unknown study variable. Suppose there are $L$ strata, containing $N_h$ elements from which a sample of size $n_h$ is to be chosen independently from each stratum $(1 \le h \le L)$. We write $N = \sum_{h=1}^{L} N_h$ and $n = \sum_{h=1}^{L} n_h$. In the case of the stratified mean estimate,

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

where $\bar{y}_h$ is the mean of the sample elements in the $h^{\text{th}}$ stratum, we need to choose the breaks in order to minimise its variance

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{yh}^2}{n_h}, \qquad (2)$$

where

$$S_{yh} = \sqrt{\sum_{i=1}^{N_h} \left(Y_{hi} - \bar{Y}_h\right)^2 / N_h},$$

is the standard deviation of $Y$ restricted to stratum and $h$, and

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi},$$

is the mean.

Dalenius (1950) derived equations for determining boundaries when stratifying variables by size, so that (2) is minimised, but these equations proved troublesome to solve because of dependencies among the components. Since then there have been numerous attempts 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 often used today. Eckman's rule (1959) of iteratively equalising the product of stratum weights and stratum ranges was found to require arduous calculations, and is less used than the method of Dalenius and Hodges method (Nicolini 2001). Lavallée and Hidiroglou (1988) derived an iterative procedure for stratifying skewed populations into a take-all stratum and a number of take-some strata such that the sample size is minimised for a given level of reliability. Other recent contributions include Hedlin (2000) who revisited Ekman's rule, Dorfman and Valliant (2000) who compared model-based stratified sampling with balanced sampling, and Rivest (2002) who constructed a generalisation of the Lavallée and Hidiroglou algorithm by providing models accounting for the discrepancy between the stratification variable and the survey variable.

In the present paper we propose an algorithm which is much simpler to implement than any of those currently available. It is based on an observation by Cochran (1961), that with near optimum boundaries the coefficients of variation are often found to be approximately the same in all strata. He concluded however that computing and setting equal the standard deviations of the strata would be too complicated to be feasible in practice. In what follows we show that, for skewed distributions, the coefficients of variation can be approximately equalised between strata

1. Patricia Gunning, School of Computing, Dublin City University, Dublin 9, Ireland; Jane M. Horgan, School of Computing, Dublin City University, Dublin 9, Ireland.

using the geometric progression. This new algorithm is derived in section 2. Section 3 compares the efficiency of the new approximation with the cumulative root frequency and the Lavallée and Hidiroglou approximations. We summarise our findings in section 4.

## 2. An Alternative Method of Stratum Construction

To stratify a population by size is to subdivide it into intervals, with endpoints $k_0 < k_1, < \ldots, < k_L$. Ideally, the division should be based on the survey variable *Y*. Such a construction is of course not possible since *Y* is unknown; if it were known we would not need to estimate it. In practice therefore we use a known auxiliary variable *X*, which is correlated with the survey variable.

In order to make the breaks $(k_0, k_1, \ldots, k_L)$ for any given $k_0$ and $k_L$, we seek to make the $CV_h = S_{xh} / \overline{X}_h$ the same for $h = 1, 2, \ldots, L$:

$$\frac{S_{x1}}{\overline{X}_1} = \frac{S_{x2}}{\overline{X}_2} = \ldots = \frac{S_{xL}}{\overline{X}_L}. \tag{3}$$

Now $S_{xh}$ is the standard deviation and $\overline{X}_h$ the mean of *X* in stratum *h*: If we make the assumption that the distribution within each stratum is approximately uniformly distributed we may write

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2}, \tag{4}$$

$$S_{xh} \approx \frac{1}{\sqrt{12}}\left(k_h - k_{h-1}\right). \tag{5}$$

As an approximation to the coefficients of variation, this gives

$$CV_h \approx \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h + k_{h-1})/2} \tag{6}$$

with equal $CV_h$ therefore we must have

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}}. \tag{7}$$

This new and exotic recurrence relation reduces however to something familiar:

$$k_h^2 = k_{h+1}\, k_{h-1}; \tag{8}$$

the stratum boundaries are the terms of a geometric progression.

$$k_h = ar^h \ (h = 0, 1, ..., L). \tag{9}$$

Thus $a = k_0$, the minimum value of the variable, and $ar^L = k_L$, the maximum value of the variable. It follows that the constant ratio can be calculated as $r = (k_L / k_0)^{1/L}$. For a numerical example take

$$L = 4;\ k_0 = 5;\ k_4 = 50{,}000: \tag{10}$$

thus $k_h = 5.10^h (h = 0,\ 1,\ 2,\ 3,\ 4)$ and the strata form the ranges

$$5 - 50;\ 50 - 500;\ 500 - 5{,}000;\ 5{,}000 - 50{,}000. \tag{11}$$

This is clearly an extremely simple method of obtaining stratum breaks.

The relationship in (8) depends on the assumption that the distributions within strata are uniform. This may be justified by the following heuristic argument. When the parent distribution is positively skewed, then the low values of the variable have a high incidence, which decreases as the variable values increase, which makes it appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of rectangular distribution in them is not unreasonable. As the value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable.

This algorithm will of course not work for normal distributions. Also since the boundaries increase geometrically, it will not work well with variables that have very low starting points: this will lead to too many small strata; the rule breaks down completely when the lower end point is zero. We expect the best results when the distribution is highly positively skewed and the upper part contains a small percentage of the total frequency.

## 3. The Performance of the Algorithm

### 3.1 Some Real Positively Skewed Populations

To test our algorithm, we implement it on four specific populations, which are skewed with positive tail:

Our first population (Population 1) is an accounting population of debtors in an Irish firm, detailed in Horgan (2003). In addition, we use three of the skewed populations that Cochran (1961) invoked to illustrate the efficiency of

the cumulative root frequency method of stratum construction. These are:

- 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).

There were five other populations in the Cochran paper, which turned out to be unsuitable for use with our algorithm. In three cases the variable was a proportion: agricultural loans, real estate loans and independent loans expressed as a percentage of the total amount of bank loans. Another, a population of farms in which the variable ranged from 1 to 18, was essentially discrete. Yet another, a population of income tax returns, was not sufficiently skewed: it owed its skewness to the top 0.05% of the population, and when this was removed, or put in a take-all stratum, the skewness disappeared.

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

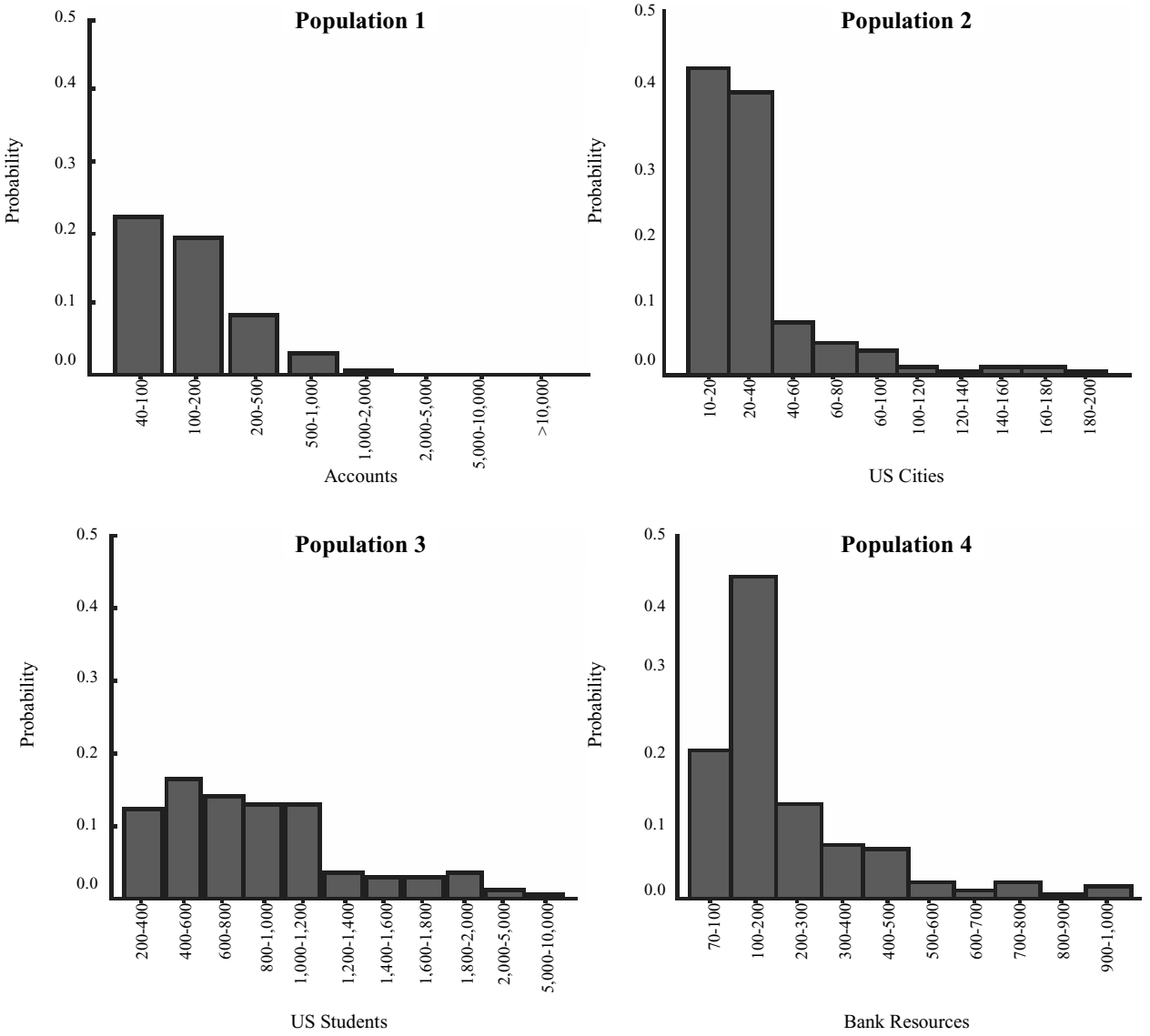



**Figure 1.** Populations

The new algorithm is implemented on these populations, and compared with the cumulative root frequency (cum $\sqrt{f}$) and the Lavallée-Hidiroglou methods of stratum construction.

### 3.2 Comparison with the Cumulative Root Frequency Method

We first compare the performance of the new algorithm with cum $\sqrt{f}$ by dividing the populations summarised in Table 1 into $L$ = 3, 4 and 5 strata, using both methods to make the breaks. The results are given in Tables 2, 3 and 4.

A cursory examination of the coefficients of variation in Tables 2, 3 and 4 suggests that, in most cases, the geometric method is more successful than cum $\sqrt{f}$ in obtaining near-equal strata $CV_h$. For example in Population 1, which has the greatest skewness, the $CV_h$ differ substantially from each other when cum $\sqrt{f}$ is used to make the breaks, while the geometric method appears to achieve near-equal $CV_h$ in all cases of 3, 4 and 5 strata: the best results are obtained with $L$ = 5. In the other three populations, the $CV_h$ are not as diverse with cum $\sqrt{f}$, but they still appear more variable than those obtained with the geometric method of stratum construction.

The $CV_h$ with the geometric 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.

A more detailed analysis of the variability of the $CV_h$ between strata is given in Table 5, where the standard deviation of the $CV_h$ is calculated for each design.

**Table 1**
Summary Statistics for 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**
The Geometric *vs* the Cum $\sqrt{f}$ : Stratum Breaks with $L$ = 3 and $n$ = 100

| Population | Stratification Method | $CV$ | | Stratum 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| 1 | Geometric | 0.0600 | $k_h$ | 354 | 3,152 | |
| | | | $N_h$ | 2,334 | 1,288 | 189 |
| | | | $n_h$ | 9 | 46 | 45 |
| | | | $CV_h$ | 0.71 | 0.68 | 0.64 |
| | Cum $\sqrt{f}$ | 0.0600 | $k_h$ | 558 | 2,236 | |
| | | | $N_h$ | 2,339 | 735 | 295 |
| | | | $n_h$ | 19 | 17 | 64 |
| | | | $CV_h$ | 0.70 | 0.42 | 0.76 |
| 2 | Geometric | 0.0270 | $k_h$ | 26 | 72 | |
| | | | $N_h$ | 701 | 243 | 94 |
| | | | $n_h$ | 36 | 29 | 35 |
| | | | $CV_h$ | 0.28 | 0.23 | 0.33 |
| | Cum $\sqrt{f}$ | 0.0269 | $k_h$ | 28 | 66 | |
| | | | $N_h$ | 729 | 208 | 101 |
| | | | $n_h$ | 40 | 22 | 38 |
| | | | $CV_h$ | 0.29 | 0.25 | 0.34 |
| 3 | Geometric | 0.0317 | $k_h$ | 726 | 2,645 | |
| | | | $N_h$ | 253 | 321 | 103 |
| | | | $n_h$ | 9 | 38 | 53 |
| | | | $CV_h$ | 0.32 | 0.37 | 0.39 |
| | Cum $\sqrt{f}$ | 0.0282 | $k_h$ | 1,179 | 3,629 | |
| | | | $N_h$ | 456 | 152 | 69 |
| | | | $n_h$ | 37 | 35 | 28 |
| | | | $CV_h$ | 0.41 | 0.31 | 0.27 |
| 4 | Geometric | 0.0184 | $k_h$ | 168 | 405 | |
| | | | $N_h$ | 211 | 93 | 53 |
| | | | $n_h$ | 27 | 27 | 46 |
| | | | $CV_h$ | 0.23 | 0.24 | 0.30 |
| | Cum $\sqrt{f}$ | 0.0198 | $k_h$ | 162 | 441 | |
| | | | $N_h$ | 207 | 107 | 43 |
| | | | $n_h$ | 25 | 39 | 36 |
| | | | $CV_h$ | 0.23 | 0.30 | 0.27 |

**Table 3**
The Geometric *vs* the Cum $\sqrt{f}$ : Stratum Breaks with $L = 4$ and $n = 100$

| Population | Stratification Method | $CV$ | | Stratum 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|
| 1 | Geometric | 0.0430 | $k_h$ | 205 | 1,057 | 5,443 | |
| | | | $N_h$ | 1,416 | 1,382 | 483 | 88 |
| | | | $n_h$ | 6 | 22 | 40 | 32 |
| | | | $CV_h$ | 0.45 | 0.44 | 0.48 | 0.50 |
| | Cum $\sqrt{f}$ | 0.0480 | $k_h$ | 558 | 1,117 | 2,795 | |
| | | | $N_h$ | 2,339 | 483 | 325 | 222 |
| | | | $n_h$ | 23 | 5 | 10 | 62 |
| | | | $CV_h$ | 0.70 | 0.19 | 0.27 | 0.69 |
| 2 | Geometric | 0.0194 | $k_h$ | 20 | 43 | 93 | 200 |
| | | | $N_h$ | 459 | 398 | 130 | 51 |
| | | | $n_h$ | 22 | 31 | 25 | 22 |
| | | | $CV_h$ | 0.22 | 0.20 | 0.22 | 0.22 |
| | Cum $\sqrt{f}$ | 0.0213 | $k_h$ | 19 | 38 | 85 | |
| | | | $N_h$ | 393 | 428 | 155 | 62 |
| | | | $n_h$ | 15 | 26 | 30 | 29 |
| | | | $CV_h$ | 0.20 | 0.17 | 0.25 | 0.26 |
| 3 | Geometric | 0.0214 | $k_h$ | 526 | 1,386 | 3,653 | |
| | | | $N_h$ | 138 | 343 | 127 | 69 |
| | | | $n_h$ | 5 | 27 | 26 | 42 |
| | | | $CV_h$ | 0.27 | 0.26 | 0.26 | 0.27 |
| | Cum $\sqrt{f}$ | 0.0230 | $k_h$ | 690 | 2,160 | 5,100 | |
| | | | $N_h$ | 235 | 319 | 75 | 48 |
| | | | $n_h$ | 13 | 43 | 21 | 23 |
| | | | $CV_h$ | 0.31 | 0.33 | 0.29 | 0.19 |
| 4 | Geometric | 0.0142 | $k_h$ | 134 | 261 | 504 | |
| | | | $N_h$ | 156 | 109 | 63 | 29 |
| | | | $n_h$ | 20 | 23 | 29 | 28 |
| | | | $CV_h$ | 0.18 | 0.19 | 0.19 | 0.20 |
| | Cum $\sqrt{f}$ | 0.0143 | $k_h$ | 162 | 255 | 488 | |
| | | | $N_h$ | 207 | 58 | 57 | 35 |
| | | | $n_h$ | 33 | 9 | 23 | 35 |
| | | | $CV_h$ | 0.23 | 0.11 | 0.18 | 0.24 |

**Table 4**
The Geometric *vs* the Cum $\sqrt{f}$ : Stratum Breaks with $L = 5$ and $n = 100$

| Population | Stratification Method | $CV$ | | Stratum 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|
| 1 | Geometric | 0.0360 | $k_h$ | 147 | 549 | 2,037 | 7,552 | |
| | | | $N_h$ | 1,054 | 1,267 | 732 | 265 | 51 |
| | | | $n_h$ | 2 | 14 | 27 | 33 | 24 |
| | | | $CV_h$ | 0.37 | 0.38 | 0.40 | 0.37 | 0.41 |
| | Cum $\sqrt{f}$ | 0.0349 | $k_h$ | 279 | 838 | 1,677 | 4,193 | |
| | | | $N_h$ | 1,644 | 1,010 | 332 | 249 | 134 |
| | | | $n_h$ | 9 | 14 | 7 | 15 | 55 |
| | | | $CV_h$ | 0.52 | 0.30 | 0.20 | 0.25 | 0.57 |
| 2 | Geometric | 0.0144 | $k_h$ | 17 | 32 | 59 | 108 | |
| | | | $N_h$ | 364 | 418 | 130 | 87 | 39 |
| | | | $n_h$ | 18 | 28 | 17 | 20 | 17 |
| | | | $CV_h$ | 0.18 | 0.14 | 0.15 | 0.16 | 0.15 |
| | Cum $\sqrt{f}$ | 0.0186 | $k_h$ | 28 | 38 | 57 | 104 | |
| | | | $N_h$ | 729 | 92 | 89 | 88 | 40 |
| | | | $n_h$ | 58 | 4 | 7 | 16 | 15 |
| | | | $CV_h$ | 0.28 | 0.08 | 0.11 | 0.16 | 0.16 |
| 3 | Geometric | 0.0184 | $k_h$ | 433 | 941 | 2,043 | 4,434 | |
| | | | $N_h$ | 100 | 255 | 1,989 | 74 | 56 |
| | | | $n_h$ | 2 | 16 | 27 | 20 | 35 |
| | | | $CV_h$ | 0.22 | 0.21 | 0.24 | 0.21 | 0.21 |
| | Cum $\sqrt{f}$ | 0.0212 | $k_h$ | 1,179 | 1,669 | 3,139 | 6,079 | |
| | | | $n_h$ | 50 | 3 | 17 | 15 | 15 |
| | | | $CV_h$ | 0.40 | 0.09 | 0.20 | 0.19 | 0.13 |
| 4 | Geometric | 0.0110 | $k_h$ | 118 | 200 | 339 | 576 | |
| | | | $N_h$ | 114 | 116 | 64 | 39 | 24 |
| | | | $n_h$ | 12 | 20 | 24 | 18 | 24 |
| | | | $CV_h$ | 0.14 | 0.14 | 0.17 | 0.12 | 0.16 |
| | Cum $\sqrt{f}$ | 0.0119 | $k_h$ | 162 | 255 | 395 | 627 | |
| | | | $N_h$ | 207 | 58 | 37 | 36 | 19 |
| | | | $n_h$ | 44 | 11 | 10 | 19 | 16 |
| | | | $CV_h$ | 0.23 | 0.11 | 0.10 | 0.13 | 0.11 |

**Table 5**
The Variability of the $CV_h$ for the Geometric and the Cum $\sqrt{f}$ Methods

| Strata | | Population | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 3 | Geometric | 0.035 | 0.050 | 0.036 | 0.038 |
| | Cum $\sqrt{f}$ | 0.181 | 0.045 | 0.072 | 0.035 |
| 4 | Geometric | 0.027 | 0.010 | 0.006 | 0.008 |
| | Cum $\sqrt{f}$ | 0.276 | 0.042 | 0.062 | 0.059 |
| 5 | Geometric | 0.018 | 0.015 | 0.013 | 0.020 |
| | Cum $\sqrt{f}$ | 0.166 | 0.076 | 0.119 | 0.054 |

We see from Table 5 that, with just two exceptions, the standard deviations of the $CV_h$ are substantially lower with the geometric method of stratum construction than with cum $\sqrt{f}$. In the two cases where the cumulative root has a lower standard deviation than the geometric, the differences between them is not great, and occur with the smallest number of strata, $L = 3$, in Populations 2 and 4. We may conclude therefore that the new algorithm is successful in breaking the strata in such a way that the $CV_h$ are near equal.

What remains is to investigate whether the geometric breaks lead to more efficient estimation than cum $\sqrt{f}$. To do this, the two methods are compared in terms of the relative efficiency or variance ratio obtained with $n = 100$ allocated optimally among the strata using *Neyman allocation* (Neyman 1934):

$$n_h = \left( \frac{N_h S_{xh}}{\sum_{i=1}^{L} N_i S_{xi}} \right) n. \qquad (12)$$

The relative efficiency is defined as

$$eff_{\text{cum, geom}} = \frac{V_{\text{cum}}(\bar{x}_{st})}{V_{\text{geom}}(\bar{x}_{st})}, \qquad (13)$$

where $V_{\text{cum}}(\bar{x}_{st})$ and $V_{\text{geom}}(\bar{x}_{st})$ are the variances of the mean respectively with the cumulative root frequency and the geometric methods, with $n = 100$ and $n_h$ allocated as in (12) for each of the stratification methods. In sample size planning the relative efficiencies may be interpreted as the proportionate increase or decrease in the sample size with cum $\sqrt{f}$ to obtain the same precision as that of the geometric method with $n = 100$.

The variance calculations are based on the auxiliary variable *X*, and since this is assumed to be highly correlated with the unknown survey variable *Y*, we can assume the relative efficiency *eff*, given in (13), will be a reasonable approximation of the relative efficiency of *Y*.

Table 6 gives the variance ratio when the number of strata $L = 3$, 4 and 5.

From Table 6 we see that, while this new method is not always more efficient than the cumulative root frequency method of stratum construction, 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 relative efficiencies are 1.69, 1.33 and 1.17 respectively indicating that samples of sizes $n = 169$, 133 and 117 are required with cum $\sqrt{f}$ to obtain the sample precision as that of the geometric method with $n = 100$.

**Table 6**
Efficiencies of the Cum $\sqrt{f}$ Relative to the Geometric Method

| Strata | Population | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 3 | 0.97 | 0.99 | 0.79 | 1.16 |
| 4 | 1.23 | 1.19 | 1.16 | 1.04 |
| 5 | 0.94 | 1.69 | 1.33 | 1.17 |

We also see from Table 6 that while there are four cases where the relative efficiency is less than 1, with one exception, all are greater than 0.9. The exception is Population 3 with $L = 3$, the smallest number of strata; the relative efficiency in this case is 0.79.

### 3.3 Comparison with the Lavallée and Hidiroglou Algorithm

With the Lavallée-Hidiroglou algorithm, the optimum boundaries $k_1, k_2 \cdots k_{L-1}$ are chosen to minimise the sample size *n* 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%. Obtaining the minimum *n* is an iterative process, and the SAS code used for implementing it was obtained from the web at *http://www.ulval.ca/pages/lpr/*.

To compare the performance of the new method with Lavallée-Hidiroglou, the *CV*s from the geometric algorithm given in Tables 2, 3 and 4 are used as input for the Lavallée-Hidiroglou algorithm, and the sample sizes required to obtain the same precision as that of the geometric method with $n = 100$ are computed. The results are given in Table 7.

The first thing to notice from Table 7 is that the sample size required with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method is greater than 100 in all but four cases. In Population 2 with 5 strata, it is necessary to increase the sample size by 36% to

$n$ = 136, to obtain the same precision as the geometric method with $n$ = 100. With three and four strata, sample sizes of $n$ = 121 and 113 are required in Population 1, and samples sizes of $n$ = 123 and $n$ = 117 are required in Population 2, to obtain the same precision as the geometric method. When the sample size falls below $n$ = 100, the drop is not as large. In Population 4, with four and five strata, $n$ = 93 and $n$ = 99 respectively, and in Population 1 with 5 strata a sample size of $n$ = 90 will suffice with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method.

The results in Table 7 might appear to indicate that the geometric method outperforms the Lavallée-Hidiroglou method in terms of the minimum sample size required for a specified precision. We observe however that the geometric 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. Often, in financial applications the top stratum is decided judgementally; for example US state taxing authorities typically decide their take-all stratum based on a total percentage of purchase amounts (Falk, Rotz and Young 2003). If after such a take-all stratum has been removed the skewness remains, the geometric method is probably the easier and more efficient way of obtaining the remaining strata.

**Table 7**
Boundaries and Sample Size Required with the Lavallée-Hidiroglou Method to Obtain the Same $CV$ as the Geometric Method when $n$ = 100

| Population | $n$ | $CV$ | | 3 Strata 1 | 2 | 3 | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 121 | 0.0600 | $k_h$ | 1,248 | 8,676 | | | |
| | | | $N_h$ | 2,867 | 464 | 38 | | |
| | | | $n_h$ | 42 | 41 | 38 | | |
| | | | $CV_h$ | 0.87 | 0.57 | 0.37 | | |
| 2 | 123 | 0.0270 | $k_h$ | 35 | 102 | | | |
| | | | $N_h$ | 795 | 202 | 41 | | |
| | | | $n_h$ | 47 | 35 | 41 | | |
| | | | $CV_h$ | 0.31 | 0.31 | 0.17 | | |
| 3 | 107 | 0.0317 | $k_h$ | 1,398 | 4,197 | | | |
| | | | $N_h$ | 481 | 135 | 61 | | |
| | | | $n_h$ | 28 | 18 | 61 | | |
| | | | $CV_h$ | 0.41 | 0.30 | 0.24 | | |
| 4 | 100 | 0.0184 | $k_h$ | 172 | 361 | | | |
| | | | $N_h$ | 212 | 85 | 60 | | |
| | | | $n_h$ | 22 | 18 | 60 | | |
| | | | $CV_h$ | 0.23 | 0.21 | 0.32 | | |
| | | | | | 4 Strata | | | |
| | | | | 1 | 2 | 3 | 4 | |
| 1 | 113 | 0.0430 | $k_h$ | 442 | 1,828 | 8,411 | | |
| | | | $N_h$ | 2,086 | 915 | 327 | 41 | |
| | | | $n_h$ | 16 | 21 | 35 | 41 | |
| | | | $CV_h$ | 0.64 | 0.41 | 0.45 | 38 | |
| 2 | 117 | 0.0194 | $k_h$ | 19 | 37 | 95 | | |
| | | | $N_h$ | 393 | 420 | 176 | 49 | |
| | | | $n_h$ | 13 | 21 | 34 | 49 | |
| | | | $CV_h$ | 0.19 | 0.16 | 0.28 | 0.21 | |
| 3 | 103 | 0.0214 | $k_h$ | 740 | 1,505 | 3,819 | | |
| | | | $N_h$ | 256 | 234 | 118 | 69 | |
| | | | $n_h$ | 9 | 10 | 15 | 69 | |
| | | | $CV_h$ | 0.32 | 0.18 | 0.25 | 0.27 | |
| 4 | 93 | 0.0142 | $k_h$ | 117 | 188 | 359 | | |
| | | | $N_h$ | 111 | 112 | 74 | 60 | |
| | | | $n_h$ | 7 | 9 | 17 | 60 | |
| | | | $CV_h$ | 0.14 | 0.12 | 0.19 | 0.32 | |
| | | | | | | 5 Strata | | |
| | | | | 1 | 2 | 3 | 4 | 5 |
| 1 | 90 | 0.0360 | $k_h$ | 342 | 1,153 | 3,431 | 10,301 | |
| | | | $N_h$ | 1,846 | 993 | 357 | 147 | 26 |
| | | | $n_h$ | 12 | 14 | 17 | 21 | 26 |
| | | | $CV_h$ | 0.58 | 0.34 | 0.31 | 0.31 | 0.32 |
| 2 | 136 | 0.0144 | $k_h$ | 14 | 21 | 35 | 80 | |
| | | | $N_h$ | 189 | 270 | 336 | 164 | 79 |
| | | | $n_h$ | 4 | 7 | 16 | 30 | 79 |
| | | | $CV_h$ | 0.12 | 0.10 | 0.12 | 0.24 | 0.30 |
| 3 | 105 | 0.0184 | $k_h$ | 512 | 869 | 1,577 | 3,675 | |
| | | | $N_h$ | 133 | 180 | 185 | 110 | 69 |
| | | | $n_h$ | 4 | 5 | 10 | 17 | 69 |
| | | | $CV_h$ | 0.27 | 0.15 | 0.16 | 0.23 | 0.27 |
| 4 | 99 | 0.0119 | $k_h$ | 99 | 130 | 189 | 339 | |
| | | | $N_h$ | 70 | 68 | 85 | 71 | 63 |
| | | | $n_h$ | 4 | 4 | 8 | 20 | 63 |
| | | | $CV_h$ | 0.10 | 0.08 | 0.10 | 0.18 | 0.33 |

## 4. Summary

This paper derives a simple algorithm for the construction of stratum boundaries in positively skewed populations, for which it is shown that the stratum breaks may be obtained using the geometric distribution. The proposed method is easier to implement than approximations previously proposed. Comparisons with the commonly used cumulative root frequency method using four positively skewed real populations divided into three, four and five strata, showed substantial gains in the precision of the estimator of the mean; the greatest gains occurring when the number of strata was five. Comparisons with the Lavallée-Hidiroglou method indicated that a greater sample size was required to obtain the same precision as the geometric method is most cases; the greatest increase in the required sample size occurred with the largest number of strata. One limitation of the new algorithm compared to the Lavallée-Hidiroglou method of stratum construction is that it does not determine a take-all top stratum.

## Acknowledgements

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

We are indebted to the referees for their helpful suggestions which have greatly improved the original paper.

## 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.

Dorfman, A.H., and Valliant, R. (2000). Stratification by size revisited. *Journal of Official Statistics*, 16, 139-154.

Eckman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219-229.

Falk, E., Rotz, W. and Young, L.L.P. (2003). Stratified sampling for sales and use tax highly skewed data-determination of the certainty stratum cut-off amount. *Proceedings of the Section on Statistical Computing*, American Statistical Association, 66-72.

Hedlin, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16, 15-29.

Horgan, J.M. (2003). A list sequential sampling scheme with applications in financial auditing. *IMA Journal of Management Mathematics*, 14, 1-18.

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 Statistics Society*, 97, 558-606.

Nicolini, G. (2001). A method to define strata boundaries. Working Paper 01-2001-marzo, Departimento di Economia Politica e Aziendale, Universita degli Studi di Milano.

Rivest, L.-P. (2002). A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.