The Irish Accounting Review (2007), Vol. 14, No. 1, 51-64

CHOOSING THE STRATA WHEN AUDITING: AN EXAMPLE USING EXCEL

Patricia Gunning and Jane M. Horgan

Dublin City University

ABSTRACT

A new method of stratifying skewed populations is adapted to accounting populations and an Excel program provided which breaks the population into the required number of strata, summarises the book values in each of the strata, and selects the items to be sampled. An illustration is given using an Irish accounting population of debtors. The new method obtains the boundaries so that the coefficients of variation are the same in all strata, a property which indicates that the boundaries are optimum.

INTRODUCTION

One of the objectives of substantive testing is to examine the accounts of an organisation in order to collect sufficient evidence to give reasonable assurance that the financial statements are not materially misstated. For large organisations, it is often too costly and time-consuming to examine all items in the accounting population. A sample of line items is selected instead and subjected to detailed testing in order to estimate the total amount.

Accounting populations, such as debtors or creditors, are usually highly skewed with skewness coefficients often greater than two (Horgan 2003; Neter and Loebbecke, 1977). In such populations, a small proportion of large items contain a large proportion of the total monetary amount, while a large proportion of small items account for a small proportion of the total monetary amount. For these settings, the typical sample design is stratification (Liu, Batcher and Scheuren, 2005; Falk and Rotz, 2003). Usually the auditor determines a cut-off point for the very large items, marking the limit of a take-all stratum, which are to be audited on a 100% basis, and a cut-off point for very low-valued items, marking the limit of a take-none stratum, that are not audited at all. The remaining items are then sampled, the book values are ordered in increasing size, divided into mutually exclusive subgroups known as strata and separate random samples are drawn independently from each stratum. One of the advantages of stratifications is that a separate analysis can be carried out in each stratum: for example, if a population of debtors is divided into three strata, with low, medium and high book values, then the design guarantees that a number of items will be examined from each of the three levels of book values, enabling a comparative analysis of the error rates and amounts between strata.

One of the objectives of stratification is to choose the stratum boundaries that are optimum, in the sense that the variance of the stratified mean or total for a given sample size is minimised, or alternatively the sample size required for a given variance is minimised. While many algorithms have been proposed for approximating such optimum boundaries, most have implementation problems which make them difficult to program.

In this paper a new method for determining boundaries in accounting populations is proposed, which is much easier to use than currently available methods. First, an outline of the methods currently in use is given and their limitations discussed. Then the new method is described, and an Excel program is developed which enables the auditor to divide the population into strata, to allocate the sample among the strata and to ascertain which items are to be audited. After that, the optimality of the method is discussed. The paper concludes with a summary of the conclusions and plans for future research.

LITERATURE REVIEW

The most well-known procedure for obtaining stratum boundaries is the *cumulative root frequency method* of Dalenius and Hodges (1959), but this is tedious to implement and involves some arbitrariness which may be off-putting to the user. It is carried out by first sorting the population by increasing size of the book values and then dividing the sorted items into an arbitrary but fairly large number M of classes and counting the number f_i of items within each class $i = 1, 2,\ldots,M$. One then forms strata by joining the adjacent classes into L groups in which the sum of the square root of the frequencies, f_i, in each stratum are approximately equal, i.e. $\sum_i \sqrt{f_i}$ are to be equal or near equal in each stratum. Cochran (1961) cautions that it is advisable to have a substantial number M of classes to start, lest the true optimum stratification may be missed and the calculation of the stratum boundaries become affected by grouping errors. With the availability of user-friendly computing facilities, the computational tediousness of this method is not as much a problem today as when it was originally proposed, but the arbitrariness still remains: the choice of the initial number of classes, M, is a subjective decision; it is left to the discretion of the user to decide what M should be and there is no theory that gives the best number of classes (Hedlin, 2000). In addition, the final stratum boundaries depend on the initial choice of the number of classes.

In an effort to improve on *the cumulative root frequency method*, Eckman (1959) proposed an algorithm for stratification that uses iteration to equalise the product of stratum weights and stratum ranges. Cochran (1961) attempted to apply this method to real data but failed, and admitted that he had to resort to trial and error methods to arrive at an approximation. Slanta and Krenzke (1996) observed that the calculations required in the implementation of the Eckman rule are "ominous". Nicolini (2001) found that even after carrying out such arduous calculations, one may not arrive at the optimum solution.

Lavallée and Hidiroglou (1988) have derived an iterative procedure for skewed populations, such that the sample size is minimised for a given level of precision. This involves starting with an arbitrary set of initial boundaries, and then applying the algorithm iteratively to arrive at the optimum values. Detlefsen and Veum (1991) found that the final 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. Sometimes the algorithm does not converge at all. Rivest (2002) reported numerical difficulties, failure to reach the global minimum sample size and non-convergence of the algorithm when the number of strata is large.

Recent developments in boundary construction include the use of mathematical programming (Khan, Khan and Ahsan, 2002), and random search methods Kozak (2004) but all are too complicated to be of any practical use. Despite its limitation the *cumulative root frequency method* of Dalenius and Hodges (1959) is still the most frequently used in practice (Hedlin, 2000).

A NEW METHOD

The method of stratifying accounting populations which is now being proposed is much easier to implement than those currently in use and eliminates the arbitrariness. It is based on an observation by Cochran (1961), that when the optimum boundaries are achieved, the coefficients of variation, i.e. standard deviation relative to the mean, are often found to be approximately the same in all strata. In the case of accounting populations, the book values are stratified so that

$$\frac{S_1}{\overline{X}_1} = \frac{S_2}{\overline{X}_2} = \dots = \frac{S_L}{\overline{X}_L}$$

where $\overline{X}_h$ and S_h ($h = 1, \dots L$) are the mean and standard deviation of the book values respectively in stratum h, and $CV_h = S_h/\overline{X}_h$ is the stratum coefficient of variation.

Gunning and Horgan (2004) showed that, if it can be assumed that the data are approximately uniformly distributed within each stratum, near-equal coefficients of variation are achieved by constructing the stratum boundaries in geometric progression.

We briefly outline the argument: Suppose the boundaries are denoted by.

$$\text{Minimum} = k_0, k_1, k_2 \dots k_L = \text{maximum}.$$

Following Dalenius and Hodges (1959), the distribution of X is assumed to be uniformly distribution in each stratum, which implies that in stratum h, i.e. in the interval $[k_{h-1} \;\; k_h]$:

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

$$S_h = \frac{k_h - k_{h-1}}{\sqrt{12}}$$

The coefficient of variation of stratum h, is therefore

$$CV_h = \frac{S_h}{X_h} = \frac{2(k_h - k_{h-1})}{\sqrt{12}(k_h + k_{h-1})}$$

With equal stratum coefficients of variation, we must have

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

which, after some cross multiplication, reduces to

$$k_h^2 = k_{h+1} k_{h-1}.$$

This equation means that the stratum boundaries are the terms of a geometric progression:

$$k_h = ar^h \qquad (h = 1, 2,\ldots, L),$$

where a = k_0, the minimum book value, and $ar^L = k^L$ the maximum book value. It follows that the constant ratio can be calculated as

$$r = (k_L/k_0)^{1/L}.$$

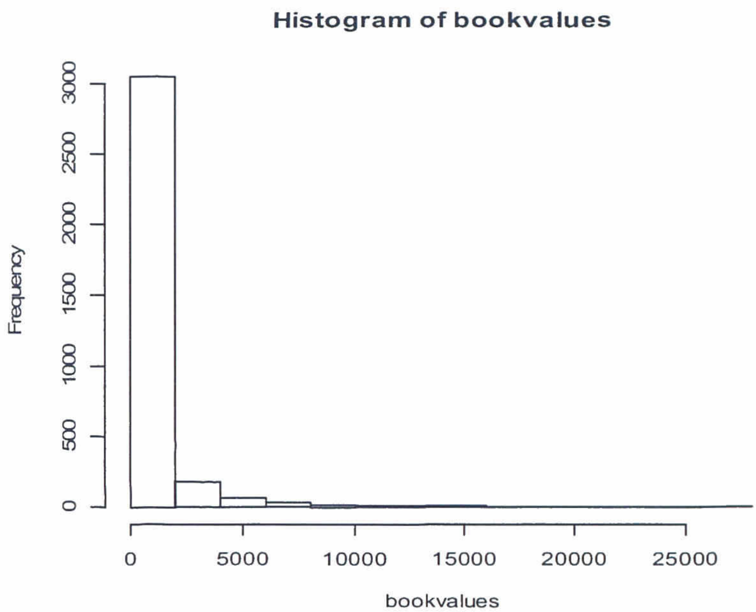
To use this geometric stratification method to divide a population of book values into L strata proceed as follows:

1. Arrange the book values in ascending order.
2. Take the minimum book value as the first term, and the maximum book value as the last term of the geometric series with $L+1$ terms: minimum = a; maximum = ar^L.
3. Calculate the common ratio: **r** = (maximum/minimum) $^{1/L}$.
4. Take the boundaries of each stratum to be the book values corresponding to the terms in the geometric progression with this common ratio:
5. minimum = a, ar, ar^2 ….. ar^L = maximum.

To illustrate how this is done on real data, consider the real accounting population of Irish debtors given in Figure 1 which consists of N = 3,369 items with book values in the range €40-€28,000, from which a sample is to be chosen for auditing. Items over €28,000 are inspected on 100% basis, and those under €40 are not inspected at all.

FIGURE 1: POPULATIONS OF IRISH DEBTORS

(N =3,369; TOTAL BOOK AMT=2,825,374, MEAN = 838; SKEWNESS = 6.44)



As can be seen from Figure 1, the population is positively skewed, with a long tail to the right, consisting of a small number of very large items. This population is discussed in detail in Horgan (2003).

To stratify into three strata, we simply take the extremes of the range of book values as the first and last term in the geometric progression, and calculate the common ratio r as follows;

$$r = (28,000/40)^{1/3} = 8.88,$$

which immediately gives the boundaries:

$a = 40$; $ar = 40(8.88) = 355$; $ar^2 = 40(8.88)^2 = 3,152$; $ar^3 = 40(8.88)^3 = 28,000$.

The population breaks are 40, 355, 3,152 and 28,000, giving the strata:

1. Stratum 1 consists of all values from 40 up to, but not including, 355:
2. Stratum 2 consists of all values from 355 up to, but not including, 3,151:
3. Stratum 3 consists of all values from 3,152 – 28,000.

With such breaks, the strata are narrow in the lower range of the book values, so that the assumption of a uniform distribution is not unreasonable. In the upper range the strata are wide but the number of line items is relatively small, with little change in the frequency over the range. Therefore, it is also not unreasonable to assume a uniform distribution in this range.

This is clearly a very simple way of forming the strata. It does not involve the choice of initial classes as does the *cumulative root frequency method*, and so overcomes its arbitrariness. It is not iterative like the Eckman and the Lavallée and Hidiroglou methods. It is objective and definitive.

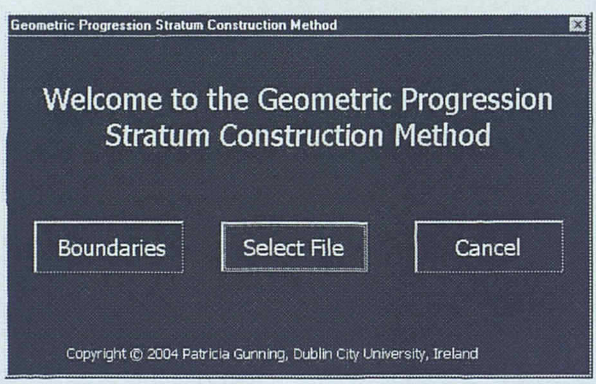
Geometric stratification is extremely easy to implement once the minimum and the maximum book values have been determined. In a geometric series with common ratio $r > 1$, the early intervals are short and later intervals progressively longer. Geometric breaks are appropriate when the distribution is highly positively skewed with the upper part containing a small percentage of the total frequency and a large percentage of the monetary amount. Again this can be said to be an apt description of accounting data. Gunning and Horgan (2004) showed that geometric stratification leads to substantial gains in the precision of the estimator in populations that have skewness coefficients greater than 2, which is the case in most populations of debtors and creditors (see Horgan 2003; Neter and Loebeckke, 1977 and Ramage, Krieger and Spero, 1979).

STRATIFICATION USING EXCEL

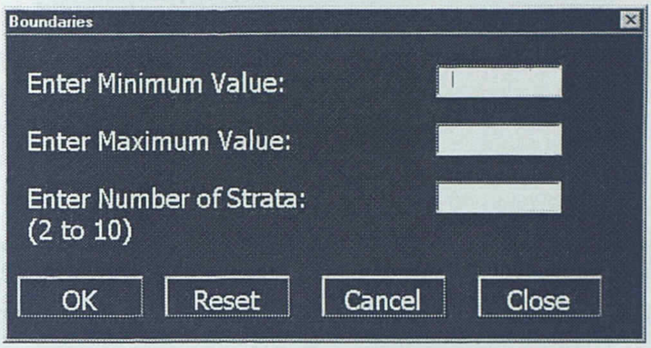
In this section, an Excel program is developed which not only obtains the stratum breaks, but, for data stored on computer file, provides summaries of the book values in each stratum, decides how to allocate the sample items among the strata, and selects the book values that should be audited.

The Excel program called GeometricProgressionMethod.xls is available at **http://www.computing.dcu.ie/~jhorgan**.

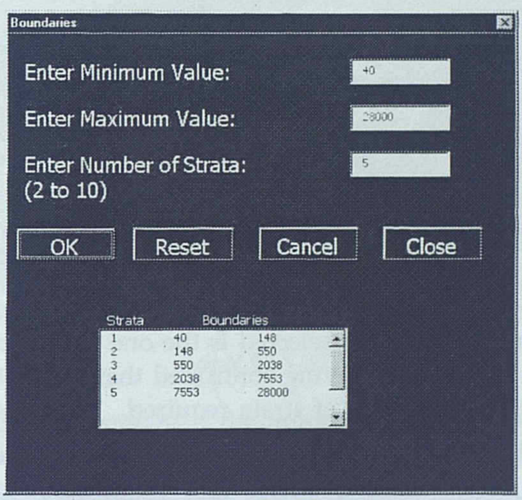
Clicking on GeometricProgressionMethod.xls gives



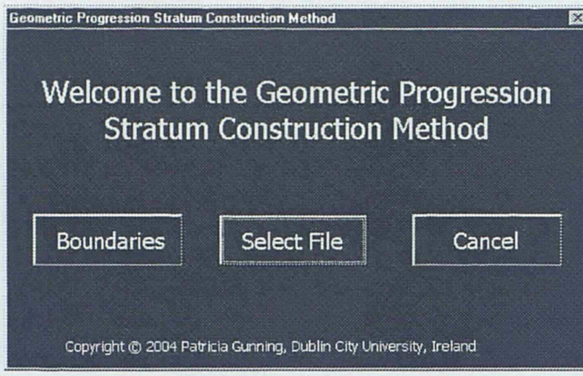
If the book values are not on computer file, click on the "Boundaries" button. The following dialog box appears.



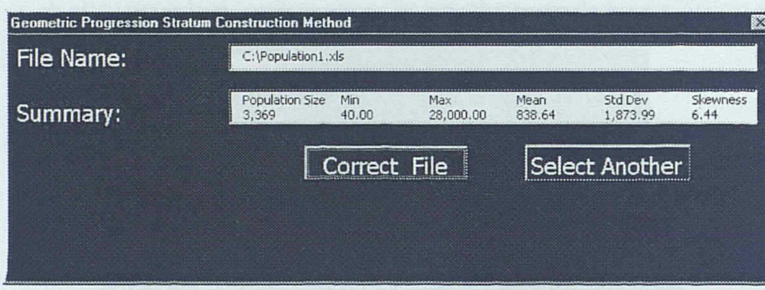
This allows you to enter the minimum and maximum value of your populations, and to decide on the number of strata. Regarding the number of strata, Cochran (1977) developed a model representing the approximate reduction in the variance over simple random sampling gained by stratification, and deduced that there is little to be gained from having more than six strata. Our program allows up to ten strata. Clicking the 'OK' button gives the stratum boundaries. If the Population in Figure 1 is to be divided into five strata, the following output is obtained.



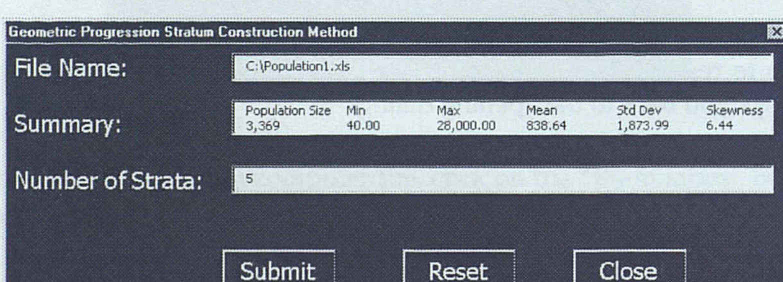
The option to reset on this dialog simply clears all the textbox entries. The cancel option brings you back to the opening dialog.



If the data file that you wish to stratify is stored on computer, then click the "Select File" button to select the file containing the data. For the file to be read successfully, each item needs to be stored on a separate row, and the maximum number of rows allowed in the file is restricted by Excel to 65,536. The next screen that appears gives a summary of the population you have selected. The population described in Figure 1 which is stored in the file called Population1.xls is used: this is available at **http://www.computing.dcu.ie/~pgunning/**.

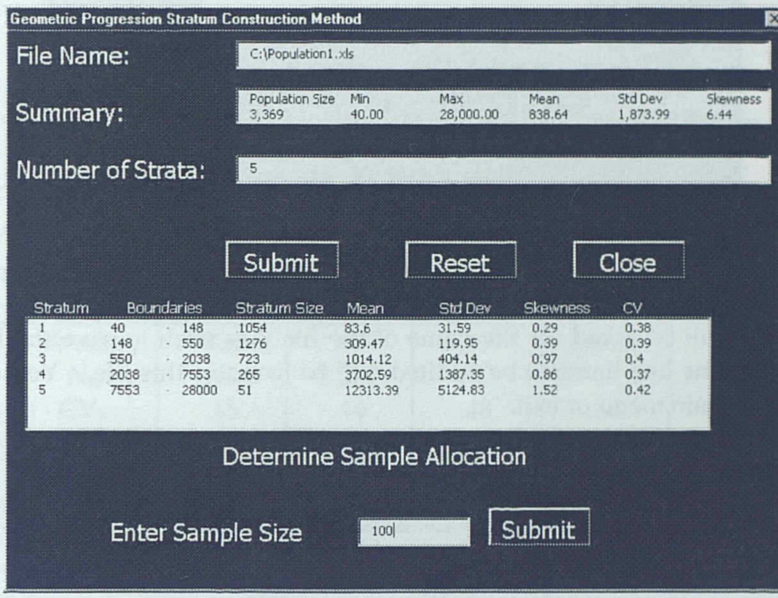


Here you can confirm that the file selected is the one you want to stratify or you may select another population. Having confirmed that this is the correct file, you will be asked to enter the number of strata required. The number 5, which is the number of strata is entered.



The "Reset" on this dialog enables you to start again and select another population.

When you press the "Submit" button, you will be presented with the boundaries for each stratum along with the number of items in each stratum. In addition the program provides summary statistics of the book values, i.e. the mean, the standard deviation, the skewness coefficient and the coefficient of variation (CV) in each stratum.

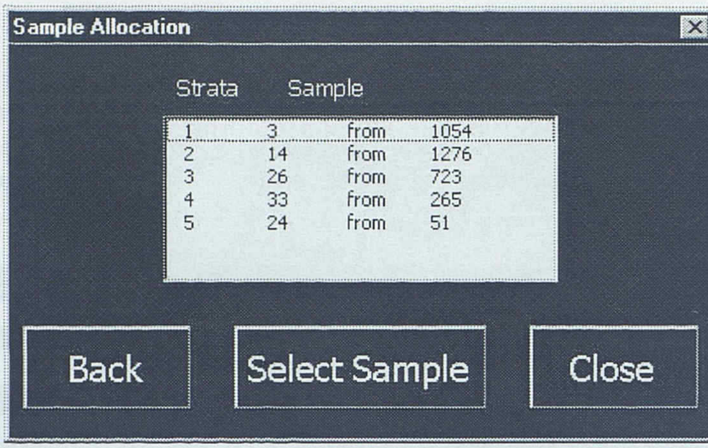


You will then have the option to determine sample allocation by entering the sample size and pressing "Submit". The sample size in each stratum is determined using Neyman allocation: for a given sample size n, the n_h are chosen so that

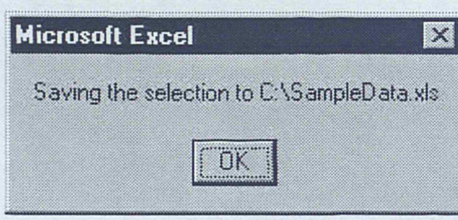
$$n_h = \left(N_h S_h \Big/ \sum_{i=1}^{L} N_i S_i \right) n.$$

Here L is the number of strata, n_h is the number of items to be chosen from each stratum, and S_h is the standard deviation of stratum h. Neyman (1934) proved that this allocation of the sample items to the strata ensures that the variance of the stratified mean or total is minimised; it is referred to as *optimum allocation.*

The following screen displays how many items from each stratum should be sampled for a total sample size of $n = 100$ .drawn from the population described in Figure 1.



At this stage you can either return to the previous screen by pressing 'Back' or get the program to select the sample ('Select Sample'). If you decide to select a sample, you will be asked for the name of the file you wish to save the line item numbers to. The line items to be audited will be listed in this file. You can then return to the main menu or exit.



OPTIMALITY OF GEOMETRIC STRATIFICATION

The question remains as to whether geometric stratification provides boundaries which are optimum in the sense that the variance of the sample mean or total is minimised. Recall that geometric stratification was derived by attempting to get near equal coefficients of variation in each stratum, and various authors have suggested that, when this is achieved, optimum stratification has been attained. Dalenius and Hodges (1959, p. 90) give the rule of thumb that 'for many populations, and for reasonable locations of the stratum boundaries, the relative variance does not vary much from stratum to stratum when L is large.' Cochran (1961, p. 171) observed that when the population is highly skewed that 'with near optimum boundaries the coefficients of variation are often found to be approximately the same in all strata.' Lavallée and Hidiroglou (1988, p. 798) noted that for skewed populations that 'with optimum allocation of the sample elements, the coefficients of variation for the take-some strata tend to be equalised without a significant increase in the overall coefficient of variation.' To illustrate how successful our algorithm is in achieving equal coefficients of variation between strata, the Excel program is used to divide the accounting population given in Figure 1 into 3, 4, 5 and 6 strata. Table 1 gives the stratum breaks, the number of elements in each stratum, N_h, the coefficient of variation in each stratum, CV_h, and the sample size to be allocated to each stratum, n_h, when the total sample size $n = 100$.

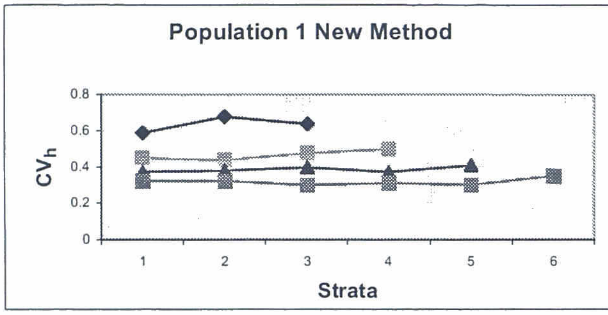
TABLE 1: STRATA CONSTRUCTION

No. of Strata		1	2	3	4	5	6
3	Breaks	354	3,152	28,000			
	N_h	2,334	1,288	189			
	CV_h	.71	.68	.68			
	n_h	9	46	45			
4	Breaks	205	1,057	5,443	28,000		
	N_h	1,416	1,382	483	88		
	CV_h	.45	.44	.48	.50		
	n_h	6	22	40	32		
5	Breaks	147	549	2,037	7,552	28,000	
	N_h	1,054	1,267	732	265	51	
	CV_h	.37	.38	.40	.37	.41	
	n_h	2	14	27	33	24	
6	Breaks	118	354	1, 057	3,152	9,396	28,000
	N_h	853	1039	906	382	157	32
	CV_h	.32	.32	.31	.31	.30	.35
	n_h	2	8	25	25	28	18

It can be seen from Table 1 that, in each case, the first stratum consists of a large number of small line items in all cases, while the last stratum consists of a relatively small number of large items. It can also be seen that the sample size is allocated so that the top stratum is sampled more intensively: the n_h form a greater proportion of the N_h than for the lower strata.

An examination of the coefficients of variation in Table 1 confirms that our method has been successful in obtaining near-equal coefficients of variation CV_h in all cases of 3, 4, 5 and 6 strata: the difference is never greater than .06. The similarity of the CV_h is illustrated further in Figure 2, which gives a line graph for each set of strata. This implies that the stratification method is near the optimum in the sense that the variance of the sample mean or total is near its minimum.

FIGURE 2: VARIABILITY OF THE BETWEEN STRATUM *CVs*



SUMMARY

The International Standard of Auditing 530 (IAASB, 2004) suggests that audit efficiency may be improved if the auditor stratifies a population by dividing it into discrete sub-populations. While many methods of stratification exist, all have implementation problems which make them difficult to implement in practice. This paper gives a new algorithm for the construction of stratum boundaries in accounting populations, which is much simpler to use than currently available procedures and overcomes the arbitrariness inherent in many of those currently available.

The new method is designed specifically for stratifying skewed populations, and in this paper it has been adapted to populations of debtors or creditors for substantial auditing. The procedure chooses the stratum breaks so that the coefficients of variation are equal in each stratum; this ensures that the variance of the mean or total is minimized when the population is positively skewed, which is the norm in accounting populations.

A real accounting population of debtors is used to illustrate the effectiveness of the new method. The population is divided into three, four, five and six strata, and it is shown that the geometric method provides strata with equal coefficients of variation.

This geometric method of obtaining stratum breaks is so simple that it can be even done by hand, no matter how large the population. The excel program developed not only constructs the stratum boundaries but also determines the number of units to be selected from each stratum and provides the auditor with a file containing the line items that should be audited.

We plan to develop the program further to allow for post audit analysis. It will be extended to allow the input of the audited values, and the calculation of the estimate of error along with the precision of the estimate. We also plan to analyse the effectiveness and efficiency of different estimators of error used with stratification.

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*, Vol. 32, No. 2, pp. 345-358.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, New York, NY: Wiley.

Dalenius, T and Hodges, J.L. (1959). Minimum Variance Stratification, *Journal of the American Statistical Association*, No. 54, No. 258, pp. 88-101.

Detlefsen, R. and Veum, L. (1991). Design issues for the Retail Trade Survey in the U.S. Bureau of the Census, *Proceedings of the Research Method Section, American Accounting Association*, pp. 214-219.

Eckman, G. (1959). An Approximation Useful in Univariate Stratification, *The Annals of Mathematical Statistics*, Vol. 30, No. 1, pp. 219-229.

Falk, E. and Rotz, W. (2003). Stratified Sampling for Sales and Use Tax Highly Slewed Data, *Proceedings of the American Statistical Association Section on Statistical Computing*, pp. 1365-1370.

Gunning, P. and Horgan, J.M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations, *Survey Methodology*, Vol. 30, No. 2. pp. 159-166.

Hedlin, D. (2000). A Procedure for Stratification by an Extended Ekman Rule, *Journal of Official Statistics*, Vol. 16, No. 1, pp. 15-29.

Horgan, J.M. (2003). A List Sequential Sampling Scheme with Applications in Financial Auditing, *IMA Journal of Management Mathematics*, Vol. 14, No. 1, pp. 1-18.

International Auditing and Assurance Standards Board (IAASB) (2004). *International Standard of Auditing (UK and Ireland) 530: Audit Sampling and Other Means of Testing*, New York, NY: International Federation of Accountants.

Khan, E.A.. Khan, K.G.M. and Ahsan, M.J. (2002). Optimum Stratification: A Mathematical Programming Approach, *Calcutta Statistical Association Bulletin*, Vol.52 (Special), pp. 323-333.

Kozak, M. (2004). Optimal Stratification using Random Search Methods in Agricultural Surveys, *Statistics in Transition*, Vol. 6, No. 5, pp. 797-806.

Lavallée, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations, *Survey Methodology*, Vol.14, No. 1, pp. 33-43.

Liu, Y., Batcher, M., and Scheuren, F. (2005). Efficient Sampling Design in Audit, *Data, Journal of Data Science*, Vol. 3, No. 3, pp. 213-222.

Neter, J and Loebbecke, J.K (1977). On the Behaviour of Major Statistical Estimators when Sampling Accounting Populations, *Journal of the American Statistical Association*, Vol. 72, No. 359, pp. 501-507.

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.

Ramage, J.G, Krieger, A.M. and Spero, L.L. (1979). An Empirical Study of Error Characteristics in Audit Populations, *Journal of Accounting Research*, Vol. 17 (Supplement), pp. 72-101.

Rivest, L. (2002). A Generalisation of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys, *Survey Methodology*, Vol. 28, No. 2, pp. 191-198.

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*, Vol. 22, No. 1, pp. 65-75.