Empirical Tests of Stratum Boundary Methods in Tax Populations

David E. McGrath
Quantitative Economics and Statistics Group, Ernst & Young, LLP

Key Words: Stratification, Stratum Boundaries, Neyman Allocation, Tax, Skewed Populations

Abstract

Empirical tests were conducted to compare four methods for determining stratum boundaries for highly skewed populations. As with other economic data, tax populations generally contain highly skewed numeric variables that present special problems. The cumulative square root of the frequency method is commonly used to determine initial stratum boundaries. When paired with Neyman allocation, this method proves unsatisfactory because it often creates boundaries where one highly variable, large-dollar stratum receives a disproportionate amount of the overall sample. This has the effect of creating undesirably small sample sizes in other strata, requiring manual adjustments.

To conduct the experiment, we stratified all populations by a highly skewed dollar variable using the cumulative square root of the frequency as a control group, and the following three methods as experimental groups:

Equal dollars per stratum
Equal W_hS_h in each stratum
Lavalle/Hidiriglou method

For a fixed sample size, we selected the stratification method that produces the smallest relative precision for the simulated estimation variable.

Introduction

This paper discusses problems using the cumulative square root of the frequency (Cum$\sqrt{f}$) (Cochran, 1977), for creating strata boundaries with skewed populations, such as tax records. Three alternate methods for creating strata boundaries are compared with the Cum$\sqrt{f}$ stratification across 35 tax studies.

Background - Stratification

Prior to sample selection stratifying a population into homogeneous groups on the variable of interest for the study (hereafter called estimation variable, or Y) generally decreases sampling variability. This improvement in precision means either smaller variance or obtaining the same variance with a reduced sample size. Stratification also allows us to use high sampling rates in large-dollar strata, culminating with the identification and placement of the largest units in a certainty, or take-all stratum.

Because Y is generally unknown, researchers stratify populations on variables that they expect to be correlated with the estimation variables. These auxiliary variables may be discrete or continuous. When sampling tax populations, little auxiliary data exists that is correlated with the estimation variable. Typically, we have one continuous numeric variable (X) that is related to the estimation variable (Y).

This paper tests and compares several methods for developing strata boundaries in tax populations.

Literature

Creating optimal stratum boundaries has an active history in statistical literature. Dalenius and Hodges (1959) developed the Cum$\sqrt{f}$ rule, an optimal method for data that are normally distributed within strata.

For skewed populations, Lavallee and Hidiriglou LH (1988) developed an iterative algorithm to create optimal stratum boundaries for the stratification variable (X). Most recently, Rivest (pending publication) develops optimal stratum boundaries that account for differences between the stratification and estimation variable (Y), assuming specific relationships between the two variables.

Sampling Business Data

Administrative records are frequently sampled to estimate deductible tax amounts for businesses. Although the goal of each tax study is different, we generally sample from a population of expenses (X) and estimate an unknown tax-deductible amount (Y). The sampling population may contain employee wages, project costs, or expense items. As in typical stratification situations, we know the expense amount (X_i) for every record in the sampling population, but the deductible amount (Y_i) is known only for the sample units.

Because of our small sample sizes, often 75-200, we use few auxiliary variables for stratification. Generally, we form only the numeric strata or use one categorical variable (lines of business, departments, etc.) in combination with the numeric strata. For this study, we are *only* considering methods for creating stratum boundaries using one numeric expense variable X (in dollars). Generally, our studies have from 3 to 6 numeric, noncertainty strata. In the past, we used the Cum$\sqrt{f}$ rule to develop initial stratum boundaries, and then performed ad hoc adjustments to overcome deficiencies in the stratification (discussed later).

Because businesses are only interested in a total amount (i.e., estimates by strata are rarely required), we minimize the variance of the overall estimates by using Neyman optimum allocation to distribute the overall sample to strata.

Sampling Populations in Tax Studies

Similar to most economic data, tax populations such as business expenses or projects costs have highly skewed, non-normal distributions.

Figure 1. Median Statistics for 35 Sampling Populations

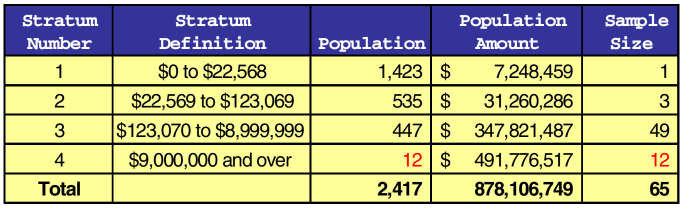
Statistic (Median across 35 Sampling Populations)	With Certainty Stratum	After Removing Certainty
Skewness	26	6
Kurtosis	2,141	84

As Figure 1 shows, the median skewness and kurtosis across 35 tax populations are 26 and 2,141 respectively. After identifying and removing the certainty stratum, the population still has positive skewness (6) and large kurtosis (84). As the statistics show, these sampling populations are far from normally distributed. The lack of normality in the sampling population causes the Cum $\sqrt{f}$ method to produce undesirable stratifications.

Stratification Problem

For a typical skewed tax population, Figure 2 shows the Cum $\sqrt{f}$ stratification paired with the Nehman allocation of sample to strata.

Figure 2. Stratification of a Tax Population using the Cum $\sqrt{f}$ rule



Stratum Number	Stratum Definition	Population	Population Amount	Sample Size
1	$0 to $22,568	1,423	$ 7,248,459	1
2	$22,569 to $123,069	535	$ 31,260,286	3
3	$123,070 to $8,999,999	447	$ 347,821,487	49
4	$9,000,000 and over	12	$ 491,776,517	12
Total		**2,417**	**878,106,749**	**65**

This population has 2,417 expense items containing about $878 million dollars. Because of the small sample size (65), we form three strata plus the certainty stratum. The above stratification is problematic because the stratum definitions generated by Cum $\sqrt{f}$ leave too much variability in the stratification variable X in Stratum 3, which is the largest non-certainty stratum. Of the 53 non-certainty sample selections, the Nehman allocation assigns 49 to stratum 3 with the Cum $\sqrt{f}$ boundaries. This stratification problem is typical when using the Cum $\sqrt{f}$ rule for tax populations, and creates small sample sizes in the first two strata.

Comparing Stratification Methods

Because the Cum $\sqrt{f}$ rule creates unsatisfactory boundaries, the initial goal of this research was to empirically test 3 competing stratification methods to compare with the Cum $\sqrt{f}$ rule. Results from the LH method are not shown because problems with the algorithm rendered the method ineffective.[1] For 35 actual sampling projects, we created strata boundaries using three methods. We then compared the methods to identify the best stratification of each population The three stratification methods compared for this paper are described below:

Cum $\sqrt{f}$ - The method equalizes the square root of the frequency (number of records) across strata Cochran (1977).

Equal Dollars per Stratum - A common method for creating stratum boundaries for tax populations is to select strata boundaries so that all non-certainty strata contain roughly equal dollars (Roberts, 1978).

Equal W_hS_h per stratum – The impetus for the Equal W_hS_h method for developing stratum boundaries comes from a comment in Cochran (1977).

'*The Dalenius-Hodges rule* (Cum $\sqrt{f}$) *is roughly equivalent to making W_hS_h constant, as conjectured earlier by Dalenius and Gurney (1951)'. If W_hS_h were constant across strata, Nehman allocation creates approximately equal sample sizes across non-certainty strata.*'

For highly skewed populations, the Cum $\sqrt{f}$ method clearly does not produce stratifications with similar W_hS_h across strata, as shown by the unequal sample sizes in Figure 2. The highly skewed data, even within strata, causes the Cum $\sqrt{f}$ rule to produce poor stratifications in tax populations. To abandon the Cum $\sqrt{f}$ rule but retain the objective in the above quote, we developed an iterative method that forces the equality of the W_hS_h across strata.

[1] The LH method appeared ideally suited for stratifying tax populations because it was designed for skewed populations. We programmed the LH algorithm in SAS and planned to compare this method with the other methods discussed in the paper. However, we encountered numerical problems with the algorithm, similar to Slanta and Krenskie (1994) and Chen (1989) that prohibited the use of the algorithm in the majority of sampling populations. The most common numerical problem that halted the program was the occurrence of negative values inside square root operators in the equations for new stratum boundaries.

Stratifying the Population (The Certainty Stratum)

The data in Figure 3 show the skewed distribution of stratification variable X, a business expense that may be tax deductible.

Figure 3. Distribution of the Sampling Population (Expenses)

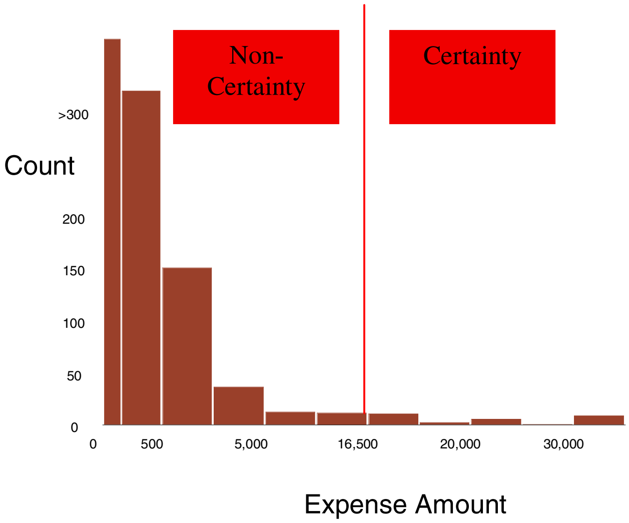


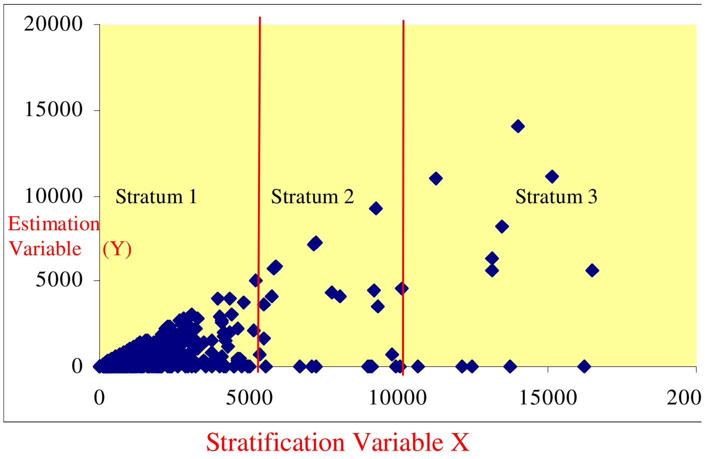
Figure 3 shows a typical distribution of business expenses in a tax population. There are hundreds of small expenses (< $500), but the majority of the dollars come from a small number of very large expenses. Because an overwhelming percent of the element variance in X comes from the large dollar records, we identify a certainty stratum and select all these records into sample. In this case, the vertical line in Figure 3 shows that all expenses larger than $16,500 were placed into the certainty stratum.

For the remainder of this paper, we ignore the certainty stratum because this paper addresses problems with methods for creating stratum boundaries in only the non-certainty strata.

Stratifying the Population (The Non-Certainty Stratum)

In Figures 4 and 5, the estimation variable Y will be the tax-deductible amount for each expense, and the estimate of interest is the overall tax-deductible amount. Figure 4 shows a scatter plot of the data from Figure 3, although now the deductible values (Y) are simulated for each expense (X). After removing the certainty stratum, the stratification variable now ranges from 0 to $16,500. For this study, we want to create three non-certainty strata, but where is the best place to make the partitions in the data? The horizontal axis contains the known expense amounts for the stratification variable (X) and the vertical axis shows simulated deductible amounts (Y). (The next section describes how we simulate the estimation variable -- the deductible amount (Y) for each expense.)

Figure 4. Joint Distribution of the Stratification and Estimation Variables.



The goal when creating stratum boundaries is to efficiently partition the population into three non-certainty strata. In other words, we need to determine where to draw the two vertical lines showing the breaks between the 3-non-certainty strata in Figure 4. These lines are arbitrarily drawn at X=5,000 and X=10,000 to show a potential stratification of the non-certainty expenses, buy where are the best partitions?

Under certain assumptions, Dalenius and Hodges (1959) have mathematically worked out equations that minimize the variability of the X values and achieve a best stratification on X. When X and Y are only loosely correlated, these stratifications are less effective. Our goal is not necessarily an optimal stratification of tax populations, but to achieve the following:

- Reasonable sample sizes in all non-certainty strata (prefer 30 or more)
- Low sampling variance for the unknown estimation variable (Y)

In Figure 5, we have added the line Y=X to Figure 4 data. Notice that the Y values are always less than or equal to the X values. Unlike typical covariates, the auxiliary variable in tax populations often constrains the estimation variable. For instance, again let the auxiliary variable be expenses and the estimation variable be the amount deductible. The amount deductible Y_i for expense X_i is distributed from 0 to X_i. It is illogical for the amount deductible to exceed the expense amount.

In tax populations, many expenses are either non-deductible (0 percent) or completely deductible (100 percent). In Figure 5, the non-deductible expenses fall on the X-axis and the completely deductible expenses fall on the line Y=X.

Figure 5. Joint Distribution of the Stratification and Estimation Variables.

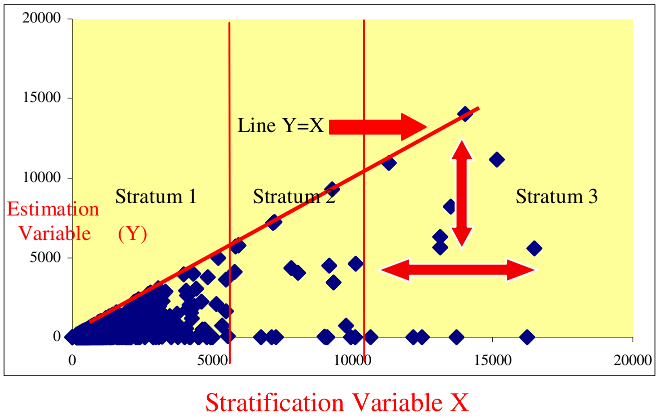


Figure 5 shows a horizontal two-sided arrow in stratum 3, indicating we could measure the effectiveness of the stratification by the variability of the X values (within each stratum). However, a better method is to measure the variability of the simulated Y values, indicated by the vertical 2-sided arrow. We know we cannot simulate the Y values accurately enough to use for stratifying the population, but we can stratify the population on X and measure the success of the stratification by the simulated variability in Y.

Experience allows accurate simulation of Y in tax populations, and therefore the variability in Y is the main criterion when evaluating competing stratifications.

When we talk about the variability in X and Y, we can measure the element variance or incorporate the finite population corrections (fpc) and measure the sampling variance. In tax populations, we create large dollar strata with very few items and sample a large proportion of items. For this reason, the fpc's play a significant role and we should always measure the sampling variance of X and Y and *not* the element variance. In this study, stratifications that produced low element variance often had high sampling variance.

The next session discusses methods for simulating the variability of the estimation variable (Y).

Simulating the Variance of the Estimation Variable

In tax populations, the estimation variable is generally more variable than the stratification variable (X). In Figure 5, X ranges from $10,000 to $16,500 in stratum 3, but the estimation variable can range from $0 to $16,500, because the deductibility for each expense ranges from 0 to 100 percent deductible. The difference between the variability of X and Y is greatest in the largest dollar strata. Given this, if samplers use the variability in X as a proxy for the variability in Y, they will underestimate the required sample size and miss the precision goals for the study.

Many methods exist for simulating the variability of the estimation variable Y, and two are discussed here.

1) Simulating the Variability in Y (Roberts Method)

For tax populations, often expenses are either non-deductible or completely deductible, in which case Y=X. For instance, in sales tax studies purchases are either zero or 100 percent deductible for tax purposes.

Given this assumption Roberts (1978) developed the following expected element variance of Y given the sampling population:

$$E\left(\sigma_Y^2\right) \approx \rho \times [\sigma_X^2 + (1-\rho)\overline{Y^2}] \qquad (1)$$

Where:

σ_X^2 is the stratum variance of the stratification variable X;

$\overline{Y}$ is the mean for the stratification variable X ;

ρ is the proportion or dollars expected to be tax deductible (supplied by subject matter experts working on the tax study).

In tax studies where Y takes on values between 0 and X , the Roberts method of variance estimation is conservative, overestimating the true variance. For these studies, Robert's formula can still be used to approximate the variance in Y during sample size calculations to provide conservative sample sizes. However, we want to use a different method for estimating the variability in Y when evaluating stratifications.

2) Simulating the Variability in Y (Truncated Normal Method)

One option is to simulate values for the estimation variable and simply compute the new variable's sampling variance. Rivest (pending publication) discusses creating optimal stratum boundaries that account for differences between the strata and estimation variables. These methods generally assume linear, or log-linear relationships between X and Y. For this study, we created the estimation variable Y by the following linear model:

$$Y_i = X_i * p * \mathbf{1.5} * N(0,1) \qquad (2)$$

Where:

ρ is the proportion or dollars expected to be tax deductible;

N(0,1) is a standard normally distributed deviate.

Formula 2 creates a random variable Y where each Yi is normally distributed around its expected value (expense value times expected qualifying rate). Multiplying the N(0,1) variable by 1.5 provides the appropriate variance to achieve the correct proportions of zero and 100 percent deductible items. Because deductible amounts in tax populations are constrained between 0 and Xi, the values of this random variable are truncated when they exceed these constraints. The joint distribution shown in Figure 5 was created by this method.

Measuring a Best Stratification

After developing 3 different stratifications (recall LH was dropped) for each project, how do we measure which stratification is best? Ideally, we would draw a sample using each stratification method and compare the sampling variance of the estimation variable. The criterion of lowest sampling variance identifies a best stratification, because all stratification methods produce unbiased mean per unit estimates.

However, we typically use variance-reducing estimators, such as regression and ratio estimators. With this added complexity, do me measure the sampling variance across a set of potential estimators? Given practical constraints, we can measure the effectiveness of the stratification only at the sample design stage. Because stratification only affects the variance of the estimates, we evaluate stratification methods solely on a variance-related statistic, the relative precision[2].

For each of the 35 tax studies, we compare the relative precision of Y across stratification methods, evaluating the stratifications using data simulated by both the Roberts and truncated normal approximations. How do we determine the best stratification using these two measures? Averaging the relative precision from the two methods is reasonable, or we could assign more weight to the method best suited for the type of tax study.

For both the Roberts and the truncated normal approximation, the relative precision computation is shown in equation 3

$$\text{Relative Precision } \hat{Y} = \frac{\sum_h t\, N_h^2\, \sigma_Y^2\, (\frac{1-f_h)}{N_h}}{\hat{Y}} \qquad (3)$$

Where:

$\hat{Y}$ *is* $X\rho$;

X is the known population expenses of the stratification variable;

ρ is the proportion or dollars expected to be tax deductible;

T is a t-statistic with appropriate degrees of freedom;

σ_Y^2 is the simulated element variance of the estimation variable Y (*This number differs for the Roberts and Normal approximations.*);

N_h^2 is the squared number of expenses in stratum h;

f_h is the finite population correction in stratum h;

For all stratification methods we design and select a sample with fixed sample size across methods. For each stratification methods, we compute the sampling variance of the two methods for estimating the sampling variability of Y (Roberts and Truncated Normal)

[2] Relative precision is defined in this paper as the margin of error of the estimate divided by the estimate.

Comparison of Stratification Methods

For each of 35 actual studies, we compared each of the three methods for creating stratum boundaries. We evaluated the quality of each stratification by the relative precision of Y using both the Roberts and Normal approximations. Figure 6 below shows the median relative precision across the 35 tax studies.

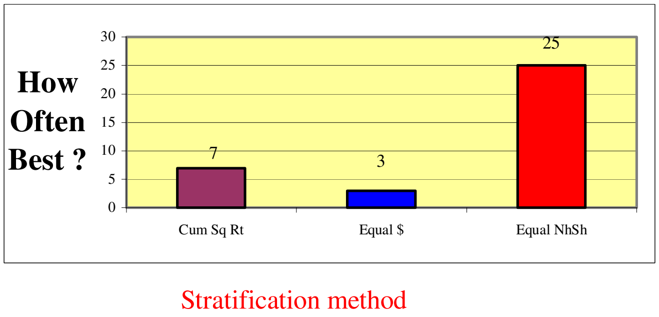
Figure 6. Median Relative Precision for Three Stratification Methods

Stratification Method	**Median Relative Precision** (Truncated Normal)	**Median Relative Precision** (Roberts Method)
Cum $\sqrt{f}$	11.5 %	13.6 %
Equal Dollars	10.9 %	12.6 %
Equal W_hS_h	10.2 %	11.8 %

As Figure 6. shows, the Equal W_hS_h stratification method had the lowest median relative precision for both the Truncated Normal and Roberts methods for simulating the variance of Y, finishing between 1 and 2 percentage points better than the Cum $\sqrt{f}$ method. The Truncated Normal and Robert's methods produced very similar variance approximations, although the estimated median relative precision was consistently about two percentage points larger for the conservative Roberts method. This trend is similar regardless of stratification method.

For the 35 studies analyzed, we also want to know how often each stratification method was best, again using the relative precision of the simulated Y as the criterion. According to the Roberts criterion, the Equal W_hS_h method was the best method in 25 of the 35 studies examined. Figure 7 results were nearly identical when the criterion for best stratification was the relative precision from the truncated normal method.

Figure 7. Number of Times Each Stratification Method was Best (criterion is the relative precision of the simulated Y variable (Roberts method)



Predicting a Best Stratification Method

If we decide on our criterion for evaluation stratification methods, it is possible to create an infinite number of different stratifications and select a best method, at least according to our criterion. Although algorithms exist to speed this process by randomizing stratifications that are similar to other 'good' stratifications, this effort rarely pays large dividends.

We want a convenient method for creating a very good, but not necessarily optimal set of stratum boundaries. We would like to predict a best stratification among a small set of competing methods, instead of attempting many methods and selecting a best method.

Therefore, we want to examine the interrelationship between characteristics of our sampling population and the performance of a set of stratification methods. In other words, can we predict a best stratification method? Because we are interested in determining a best stratification, we will attempt to examine the correlation between statistics in the sampling population and the stratification that was best, or won. To do this, we create a binomial variable (coded 0,1) for each of the stratification methods. This variable equals 1 when the stratification method was best and 0 otherwise.

Because we think the high skewness in our tax populations causes problems with the Cum $\sqrt{f}$ method, we first look at the correlation between skewness and the stratification that won. Figure 8 shows that the correlation between the Cum $\sqrt{f}$ method winning and the population skewness is -.47, indicating that as the skewness of the population increases, the Cum $\sqrt{f}$ method becomes less likely to perform as our best stratification method.

Figure 8. Median Pearson's Correlation Between Wins and Skewness in 35 Sampling Populations

	Cum $\sqrt{f}$	Equal $	Equal W_hS_h
$\rho_{\text{Skewness, Wins}}$	-.47	.65	.02

Conversely, the correlation of .65 indicates that the Equal $ method tends to be the best stratification method for the most highly skewed populations. The success of the Equal W_hS_h method depends less heavily on the population skewness.

For future study, we will develop a multinomial logistic regression model where the binomial wins variables are the dependent variables and a set of univariate statistics from the sampling population act as predictors (e.g., skewness, kurtosis, variance, mean, etc.). The prediction power of the population skewness alone, shown in Figure 8, indicates predicting a best stratification method may be successful. In other words, instead of running a set of stratification methods (for this paper 3 methods), it seems possible to run the prediction model plus one stratification method, or at least a smaller subset of stratification methods. The benefit from these predictions would be saving time and improving the variance of sample estimates.

Conclusions

The Equal W_hS_h method performed far superior to the Equal Dollar and Cum$\sqrt{f}$ stratification methods. This method was the best as measured by relative precision, regardless of the method used to approximate the variance of the estimation variable. In addition, Equal W_hS_h also produces stratifications with similar sample sizes across strata when paired with Nehman allocation. This second benefit is especially important in tax setting because the IRS imposes minimum stratum sample sizes on taxpayers making sample-based estimates. For instance, the Cum$\sqrt{f}$ stratification shown in Figure 2 is undesirable even if it has lower relative precision than the other methods because of the small sample sizes in strata 1 and 2.

Acknowledgements: - Thanks to Ryan Petska for SAS programming assistance, Mary Batcher for constructive editorial comments, Jinhee Yang, Amy Luo, and Archana Joshee for programming support.

References:

Chen, W. (1989) Stratification of a Population: Programming of Lavallee and Hidiriglou's Algorithm.

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

Dalenius, T., and Gurney, M. (1951). The problem of optimum stratification. Akandinavisk Aktuarietidskrift, 34, 133-148.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification, Akandinavisk Aktuarietidskrift, 54, 88-101.

Lavallee, P. and Hidiriglou, M.A., (1988). On the Stratification of Skewed Populations. Survey Methodology, Vol. 14, pp. 33-43.

Rivest, L.P. (pending publication) A Generalization of Lavallee and Hidiriglou Algorithm for Stratification in Business Surveys.

Roberts, D. (1978) Statistical Auditing. American Institute of Certified Public Accountants, Inc. New York

Slanta, J. and Krenzke, T. (1996) Applying the Lavallee and Hidiroglou method to obtain stratification boundaries fo rhte Census Bureau's Annual Capital Expenditures Survye. Survey Methodology, 22, 65-75.