İstanbul Üniversitesi İşletme Fakültesi Dergisi
Istanbul University Journal of the School of Business Administration
Cilt/Vol:39, Sayı/No:2, 2010, 195-204
ISSN: 1303-1732 – www.ifdergisi.org © 2010

Choosing the stratification boundaries: The elusive optima

Jane M. Horgan[1]
School of Computing
Dublin City University, Dublin, Ireland

Abstract

Since Dalenius (1950) provided a set of equations for the determination of the stratum boundaries, there has been a proliferation of attempts to obtain the optimum stratum boundaries, those that minimise the variance of the Horvitz-Thompson estimator of mean or total. In this paper, we track the progress of such methods, and ask where we are now and where to go from here.

Keywords: *Approximation methods, numerical optimisation, skewness.*

Zümre sınırlarının seçilmesi: Bulunması zor optimumlar

Özet

Dalenius'un (1950) zümre sınırlarının belirlenmesi için bir denklem kümesi sağlamasının ardından, ortalamanın ya da toplamın Horvitz-Thompson tahminleyeninin varyansını minimize eden optimum zümre sınırlarını elde etmek için yapılan çalışmalar giderek yaygınlaşmıştır. Bu makalede, gerçekleştirilen yöntemlerin gelişimi takip edilerek şu anda hangi aşamada olduğumuz ve bulunduğumuz noktadan nereye gidebileceğimiz sorularına cevap aranmaktadır.

Anahtar Sözcükler: *Yaklaşım yöntemleri, nümerik optimizasyon, çarpıklık.*

1. Introduction

Survey populations are often highly positively skewed where a small number of high-valued units account for a large share of the total value, and a large number of low-valued units account for a small share of the total. Such populations arise in business enterprises and agriculture, as well as in surveys of personal income and other financial applications. They are also natural in establishment survey populations that often have distributions that are skewed to the right. In populations such as these, stratification can lead to a substantial improvement in the precision of the sample estimators.

A stratified sample design partitions a population *U* into *L* mutually exclusive groups called strata:

$$U = \bigcup_{i=1}^{L} U_i \quad U_i \cap U_j = \phi \ , \ i \neq j$$

[1] *jhorgan@computing.dcu.ie (J.M. Horgan)*

The population mean is

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

where X_{hi} is the i^{th} unit in the h^{th} stratum which contains N_h, $h = 1, 2, \ldots L$ units, and $N = \sum_{h=1}^{L} N_h$.

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

The mean of the sample selected from stratum h is

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

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

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

where $W = N_h/N$ is the weight of stratum h. It is easy to show [1] that (3) is an unbiased estimator of the population mean $\overline{X}$, with variance

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \tag{4}$$

The objective of stratification is to choose the boundaries to minimise (4).

Sixty years ago, Dalenius [2] showed that when the stratum boundaries k_h satisfy

$$\frac{S_h^2 + (k_h - \overline{X}_h)^2}{S_h} = \frac{S_{h+1}^2 + (k_h - \overline{X}_{h+1})^2}{S_{h+1}} \tag{5}$$

(4) is minimized.

However, these equations are ill adapted to practical computations because $\overline{X}_h$ and S_h depend on k_h. To this day, they remain intractable, and researchers have had to concentrate on obtaining approximations to (5) or by applying iterative, computational algorithms to arrive at a solution that minimise the variance given in (4).

In the next Section 2, we look at the cumulative root frequency approximation suggested by Dalenius and Hodges [2], arguably the most commonly used method of stratum construction. After that in Section 3, we examine some of the more recent iterative algorithms that have been developed for stratifying skewed populations, and go on in Section 4 to outline our own contributions, geometric stratification and its applications. Finally in Section 5, we attempt to determine where we are now and where to go from here in our continued efforts to find that elusive set of optimum stratification bounds.

2. Cumulative Root Frequency

Dalenius and Hodges [2] were the first to develop an approximation to (5), the cumulative root frequency method of stratum construction, which for decades has been the main method of obtaining stratification boundaries in finite populations. The approximation is obtained by first dividing the frequencies of the variable into a fairly large number of classes M, counting the number f_j of units within the interval j, $j=1,2,\ldots,M$. Then one calculates $\sqrt{f_j}$, and forms strata by joining the adjacent intervals into L groups (strata) in which the $\sum\sqrt{f_j}$ are to be equal or near equal.

The main problem with this method is the arbitrariness in deciding the value of M. Cochran [3] cautions that it is advisable to have a substantial number of classes in the original frequency, otherwise the true optimum stratification may be missed and the calculation of the within-stratum boundaries becomes affected by grouping errors. Hedlin [4] notes that the final stratum boundaries depend on the initial choice of the number of classes M, and there is no theory which gives the best number of classes.

3. Iterative Procedures

Since survey populations are finite, optimal strata bounds could be obtained by considering all the possible divisions of the population associated with the number of strata, by calculating the variance in (4) of all the solutions, and selecting the one with the lowest variance. However, the number of possible solutions increases rapidly with L and N, and even with the availability of today's powerful computing facilities, an exhaustive enumerating process would take too long; instead a good feasible solution is obtained by applying optimisation iterative algorithms.

3.1. Lavallée and Hidiroglou

The best known iterative procedure is that of Lavallée and Hidiroglou [5] who suggested that, when a population is skewed, the stratification should consist of a top stratum, where all the units are selected into the sample ($n_L = N_L$), and a number of take-some strata which are sampled ($n_h < N_h$ for $h < L$). The variance in (4) then becomes

$$V(\bar{x}_{st}) = \sum_{h=1}^{L-1} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \tag{6}$$

which can be written in terms of the sample size n as

$$n = N_L + \frac{N \sum_{h=1}^{L-1} W_h^2 S_h^2 / a_n}{Nc^2 \overline{X}^2 + \sum_{h=1}^{L-1} W_h S_h^2} \tag{7}$$

where c is the coefficient of variation, i.e. $c^2 = V(\bar{x}_{st})/\bar{X}^2_{st}$. The $a_h = n_h/n$ are the normalised sample sizes with power allocation,

$$a_h = \frac{(W_h \overline{X}_h)^p}{\sum_{i=1}^{L-1} (W_h \overline{X}_h)^p} \tag{8}$$

so that

$$n_h = \frac{n(W_h \overline{X}_h)^p}{\sum_{i=1}^{L-1} (W_i \overline{X}_i)^p} \tag{9}$$

where p is in (0,1). In order to find the optimal boundaries to minimise n, the partial derivatives of (7) are taken with respect to each of the k_h, and equated to zero, and the resultant equations are solved iteratively. Lavallée and Hidiroglou originally suggested that the initial values are set by taking the breaks with an equal number of elements in each group, and these boundaries are replaced iteratively, using a procedure suggested by Sethi [6], until the minimum n is obtained.

While widely used in the US and Canada, the Lavallée and Hidiroglou algorithm is not without serious implementation problems and shortcomings which may lead to non-optimal results. Detlefsen and Veum [7], who used the algorithm to analyse the US Census Bureau Monthly Retail Trade Survey, discovered that the minimum sample size attained may a local but not necessarily a global minimum, and that sometimes the algorithm does not converge at all. They also found that convergence is slow when the number of strata is large. Slanta and Krenzke [8] encountered numerical difficulties when using the Lavallée and Hidiroglou algorithm, as well as failure to reach the global minimal size, and non-convergence of the algorithm when the number of strata was large.

3.2. Recent Developments

The last decade has witnessed a proliferation of numerical optimisation-based iterative methods of stratification:

- Kozak [9] presented a random search algorithm as a method of obtaining optimal boundaries so that n given in (7) is minimised under the constraints

  $$N_h \geq 2 \qquad h = 1, \ldots, L$$

  and

  $$2 \leq n_h \leq N_h, \qquad h = 1, \ldots, L\text{-}1.$$

  Here, the n_h are determined with Neyman allocation [10]:

  $$n_h = \frac{nW_hS_h}{\sum_{i=1}^{L} W_iS_i} \tag{10}$$

  At each iteration, a set of stratum boundaries is chosen at random from all the possible alternatives. The algorithm continues to change the boundaries as long as they continue to reduce n; otherwise the boundaries are not changed, and a new iteration is started. If the sample size does not decrease for some specified m consecutive iterations, the algorithm stops.

Kozak's algorithm returns random results. A nonrandom version of the original as implemented by Baillargeon and Rivest [11], in which at each iteration of the algorithm, all the possible boundary modifications are tried and the modification giving the largest decrease in n is kept. Obviously this approach is slower then the original.

Kozak [9] tested the algorithm using data from the Polish Agricultural Census, and concluded that the efficiency of the random search methods was similar to that of the Lavallée and Hidiroglou algorithm. Baillargeon and Rivest [12] found that Kozak's algorithm produced better results than Lavallée and Hidiroglou.

A weakness of Kozak's algorithm is that it may return a sample size for the top stratum that exceeds its size ($n_L > N_L$). It such cases one adds a take-all stratum, and reruns the algorithm. Baillargeon and Rivest [11] found that for small populations, Kozak's algorithm often yields a local rather than global minimum. They argue, however that when the population is small, a complete enumeration of all sets of boundaries is feasible.

- Kestinturk and Er [13] suggested that the genetic algorithm could be used to find the optimum stratification bounds. They implemented the algorithm on a range of real and simulated populations, including some Turkish manufacturing firms. Comparisons with the cumulative root approximation lead them to the conclusion that the best results are obtained when both strata boundaries and sample sizes are determined using the genetic algorithm.
- Khan et al. [14] formulated the problem of finding the optimum stratification bounds as a mathematical programming problem and developed a solution procedure using dynamic programming. A numerical example using a hospital population data is presented to illustrate the computational details. Comparisons with the cumulative root frequency approximation revealed that the proposed method is more efficient.
- Brito et al. [15] suggested that an iterative local search (ILS) metaheuristic algorithm would obtain a good feasible solution. It is a search-based method that is intended to work for variables with any distribution. They implemented their algorithm on sixteen skewed populations; some real and some simulated, and showed that it produced better solutions than the random search algorithm of Kozak [9] in most cases.
- Brito et al. [16] suggested an algorithm based on using minimal path in a graph, and claimed that it guarantees optimum stratification boundaries. They tested the algorithm using real data from the Brazilian Central Statistics Office, and provided the CPU time for the algorithm's implementation; in some cases this was nearly three minutes.

All of these authors claim that their algorithms achieve optimal stratification, either by minimising the variance for a given sample size n or by minimising n for a given variance. All use finite populations, real or simulated to show that their algorithm outperforms one or more of the algorithms already available. All the algorithms are computer-intensive. It is clear however that while some algorithms improve certain aspects of stratification, none perform uniformly better than the others.

4. Coefficients of Variation

In this section we outline our own contribution to the problem of stratification of skewed populations. Geometric stratification [17] is based on an observation of Lavallée and Hidiroglou [5]:

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

Some years previously Dalenius and Hodges [2] hinted at the same conjecture:

"for many populations, and for reasonable locations of the stratum boundaries, the relative variance does not vary much from stratum to stratum''

When we investigated the consequence of this assumption, we made a curious discovery: setting the coefficients of variation in each stratum, i.e.

$$\frac{S_1}{\overline{X}_1} = \frac{S_2}{\overline{X}_2} = \ldots = \frac{S_L}{\overline{X}_L} \tag{11}$$

produces boundaries that are in geometric progression [18].

4.1. Geometric Stratification

We briefly outline the argument which leads to geometric stratification:

Following Dalenius and Hodges [2], we assume that X is approximately uniformly distributed in each stratum. Uniform density of X in stratum h implies

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2} \tag{12}$$

$$S_h = \frac{1}{\sqrt{12}}(k_h - k_{h-1}) \tag{13}$$

The coefficient of variation of stratum h is therefore,

$$cv_h = \frac{S_h}{\overline{X}_h} \approx \frac{2(k_h - k_{h-1})}{\sqrt{12}(k_h + k_{h-1})} \tag{14}$$

With approximately equal cv_h it follows that

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}} \tag{15}$$

which reduces to

$$k_h^2 = k_{h+1}k_{h-1} \tag{16}$$

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

$$k_h = ar^h \quad h = 0, 1, \ldots. L \tag{17}$$

Thus $a = k_0$, the minimum value of the variable, and $k_L = a.r^L$, the maximum value of the variable so that

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

An example given in [17] illustrates its simplicity:

A population ranging from 5-50,000 is to be divided into 4 strata.

$$L = 4, \quad k_0 = 5, k_4 = 50{,}000$$

Thus

$$r = (50{,}000/5)^{1/4} = 10$$

and so $k_h = 5 \,.\, 10^h$ which means the breaks are

5, 50, 500, 5,000, 50,000

Geometric stratification does not involve iteration. It overcomes the pain of optimisers, and is obtained in one run through of the data file.

Initial tests by Gunning and Horgan [17] on three of the skewed populations in Cochran [13] and an Irish population of debtors [19] showed that it compared favourably with the cumulative root frequency approximation and the Lavallée and Hidiroglou algorithm for obtaining optimum boundaries.

However, Gunning and Horgan [17] cautions:

"*the algorithm will of course not work for normal distributions. Also since the boundaries increase geometrically, it will not work with variables that have very low starting points: this will lead to too many small strata".*

4.2. The Pareto Distribution

The Pareto distribution (see [20]) with density function

$f(x) = \lambda\beta^{\lambda} x^{-\lambda-1} \qquad x \geq \beta,$

is a skewed distribution with long tails to the right, and is commonly used to model skewed data.

We showed [21] that, for Pareto distributions, geometric breaks give exactly equal coefficients of variation in the different strata. Specifically we proved that, in any finite interval (β, γ) in the range of a Pareto distribution, if the break points $\beta = k_0 < k_1 < k_2 \ldots < k_L = \gamma$ are taken in geometric progression, the successive coefficients of variation are equal.

Although, these breaks failed to satisfy Dalenius's optimum conditions for minimum variance in (5), tests illustrated that geometric breaks for Pareto-type data yielded efficient results.

4.3. Geometric Starting Points

Most iterative procedures depend critically on their starting points for convergence; the final result is affected by how the initial values are chosen. Lavallée and Hidiroglou originally used starting points with equal numbers of units in each strata, but encountered convergence problems. In [22, 23] it is illustrated that the use of geometric breaks as starting points in the Lavallée and Hidiroglou algorithm improved its efficiency, and decreased the number of iterations necessary to converge. Kozak [9] recognised that better results would be obtained by using some classical approximation method such as the cumulative root frequency method of Dalenius and Hodges [2] as initial starting points. Subsequently Kozak and Verma [24] suggested that the geometric algorithm may be seen as efficient starting points for the optimization approach.

Geometric breaks have now become a standard method for setting initial values in an iterative process for stratifying skewed populations [11].

5. Where do we go from here?

Geometric stratification is an antidote to the computationally-intensive numerical algorithms that have become available over the last decade. It is deterministic and does not involve iteration. It is unbelievably simple, using just two values of the population, the minimum and the maximum, to get the boundaries. Not surprisingly then the efficiency depends critically on these, and if these are too small or too large, things go wrong [24]. If there is a large outlier, the strata will not be optimum, because large values will drag the boundaries up. If the starting point is too small, there will be too many small strata.

Modifications of the geometric algorithm are necessary to address:

- Outliers; it is obvious that the geometric method will be far from optimum when the *X* variable has large outliers; in this case a take-all stratum will need to be considered. Kestinturk and Er [13] noticed that when the population contained a large outlier, the sample size allocated with Neyman allocation to the top stratum may exceed the total number of units in the stratum ($n\_L > N\_L$). In such cases, they added a take-all stratum, and applied the geometric breaks to the remaining set of data. This idea needs to be developed further.
- Small starting points; Baillargeon and Rivest [11] found that the geometric method yielded a poor design when very small values of *X* were present in the data set. Clearly low-valued starting points will result in too many small strata; in this case a take-none strata should be considered.
- Kurtosis; the geometric method uses the minimum and maximum to obtain the bounds. It also assumes the population is skewed. In between the minimum and the maximum, there are many possibilities. The kurtosis coefficient might be examined to establish which types of skewed populations are appropriate for geometric stratification, and which are not.

No matter what improvements we make, however, there will still be a need for a stratification algorithm that is optimum irrespective of the situation (e.g. of population size, range or kurtosis), and that provides non-random results.

As we have illustrated above, there has been an inundation of iterative methods attempting to achieve the elusive optimum set of stratification boundaries, all claiming to have reached the optimum, and most claiming to have improved on methods previously available. The authors use data usually from a local source. For example Kozak [9] uses data from the Polish Agricultural Census. Kestinturk and Er [13] use Turkish data. Brito et.al. [15, 16] implement their algorithm on Brazilian data.

While it is understandable that the algorithm should be applied to real data from the source country, simulation will not prove anything conclusively. However, if all algorithms were implemented on the same data set, valid comparisons of efficiency could be obtained. To this end, I suggest that the data bank of populations provided in the stratification package of Baillargeon and Rivest [11] be used as a base set. These include:

- Three of the skewed populations of Cochran [3], i.e. inhabitants of US cities in 1940; students in four-year US colleges in 1952-1953; resources of a large US commercial;

- A population of debtors in an Irish firm details in Horgan [19];
- The monthly retail trade survey and the household budget survey of Statistics Canada;
- The population of municipalities in Sweden from Sarndal et.al. [25].

The diversity of these populations provide a useful set on which new algorithms (and indeed those already in existence) may be tested and compared.

References

[1] W.G. Cochran, *Sampling Techniques*, New York: Wiley, 1977.

[2] T. Dalenius, and J.L. Hodges, Minimum Variance Stratification. *Journal of the American Statistical Association.* 88-101 (1959).

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

[4] D. Hedlin, A Procedure of Stratification by an Extended Ekman Rule. *Journal of Official Statistics*. 16, 15-29 (2000).

[5] P. Lavallée, and M. Hidiroglou, On the Stratification of Skewed Populations. *Survey Methodology*. 14, 33-43 (1988).

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

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

[8] J. Slanta, and T. Krenzke, Applying Lavallée and Hidiroglou Method to obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Survey Methodology.* 22, 65-75 (1996).

[9] M. Kozak, Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistics in Transition*. 6, 5, 797-806 (2004).

[10] J. Neyman, 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 (1934).

[11] S. Baillargeon, and L.P. Rivest, *Univariate Stratification of Survey Populations, R Package*, available on the CRAN website at *http://www.r-project.org/*. 2010.

[12] S. Baillargeon, and L.P. Rivest, A General Algorithm or Univariate Stratification. *International Statistical Review*. 77, 3, 331-344 (2009).

[13] T. Keskinturk, and S. Er, A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. *Computational Statistics and Data Analysis*. 52, 1, 58-67 (2007).

[14] M.G.M. Khan, N. Nand, and N. Ahmad, Determining the Optimum Stratum Boundary Points Using Dynamic Programming. *Survey Methodology.* 34, 205-214 (2008).

[15] J. Brito, et al., An ILS Approach applied to Optimum Stratification Problem, 2009.

[16] J. Brito, et al., An Exact Algorithm for the Stratification with Proportional Allocation. *Optim Lett*. 4, 185-195 (2010).

[17] P. Gunning, and J.M. Horgan, A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology.* 2, 30, 159-166 (2004).

[18] J.M. Horgan, Stratification of Skewed Populations: A Review. *The International Statistics Review*. 74, 67-76 (2006).

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

[20] M. Evans, N.A.J. Hastings, and J.B. Peacock, *Statistical Distributions*. 3rd Edition, New York: Wiley, 2000.

[21] P. Gunning, J.M Horgan, and G. Keogh, Efficient Pareto Stratification. *Mathematical Proceedings of the Royal Irish Academy.* 106, 2, 131-138 (2006).

[22] P. Gunning, and J.M. Horgan, Improving the \LH Algorithm for Stratification of Skewed Populations. *Journal of Statistical Computation and Simulation.* 4, 77, 277-291 (2007).

[23] P.Gunning, J.M, Horgan, and G. Keogh, An Implementation Strategy for Efficient Convergence of \LH Stratification. *Journal of Official Statistics*. 213-228 (2008).

[24] M. Kozak, and M. Verma, Geometric versus Optimization to Stratification: A Comparison of Efficiency. *Survey Methodology.* 32, 2, 157-183 (2006).

[25] C. Sarndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling.* Springer, 1992.