Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey

John Slanta, Thomas Krenzke, U.S. Bureau of the Census
Thomas Krenzke, U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Convergence, Contour Plots

1. Introduction

The primary objectives of the sample design of the Census Bureau's Annual Capital Expenditures Survey (ACES) are to meet the desired reliability levels using operationally-feasible methodology and to stay within budget limitations. To achieve these goals, we implemented a stratified simple random sample using a modified version of Lavallée and Hidiroglou's (1988) iterative approach of finding stratum bounds. This stratification-allocation method for skewed populations obtains optimal boundary points by minimizing the total sample size given a desired coefficient of variation (c.v.). Survey managers associated with a single-purpose survey having access to a single stratifier can benefit from its operational ease and cost reductions.

Detlefsen and Veum (1991) document two shortcomings with the Lavallée and Hidiroglou (L-H) method when tested for the Census Bureau's Monthly Retail Trade Survey. They found that the resulting boundaries depend on where the initial boundaries are set, so that the minimum sample size attained is a local minimum. Geometrically, the sample size as a function of two boundaries ($n = f(b_1,b_2)$) appears like a landscape with one or more bowl-shaped valleys. The L-H method begins in a region and descends until it reaches the lowest point. If more than one minimum exists, it will not continue to search for the global minimum. Schneeberger (1979) discussed the problem of finding optimal stratification boundaries. Schneeberger expressed this problem as a non-linear program that when solved by a gradient method, whose solution may be relative and global minima, maxima, or saddle points of the variance of the sample mean.

Detlefsen and Veum also had problems with slow or non-convergence. However, it was noted by Detlefsen and Veum that convergence occurred faster when the number of strata was reduced and when starting boundaries were the same as the previous survey's sample selection boundaries.

In this work, we describe the L-H method and the way it was applied to the Annual Capital Expenditures Survey (ACES). We show how contour plots and three-dimensional plots gave us justification for using the L-H method to get the final boundaries. We also address the convergence problem by setting up constraints to be met after each iteration that protect against slow or no convergence (under the assumption that the marginal gain achieved is not worth the extra effort) and by designing the sample to have a small number of strata.

2. ACES Background

The 1992 ACES was designed by the Census Bureau to be a large-scale operational test of the sampling, processing, programming, data entry, editing, and estimation procedures which extended beyond a 1991 pilot study, to prepare for the 1993 full-scale survey. Capital expenditure estimates for domestic activities were published at conglomerated industry levels from the 1992 survey. In addition, the 1991 and 1992 preliminary surveys provided valuable capital expenditure data that will be used in future sample design enhancements.

The sampling unit for the ACES was the company which may be comprised of several establishments. The sampled population included all active companies with five or more employees from all major industry sectors except Government. These sectors include mining, construction, manufacturing, transportation, wholesale and retail trade, finance, services, and a portion of the agriculture sector that includes agricultural services, forestry, fishing, hunting, and trapping. Only companies with domestic activity were included in the sampling frame. The Research and Methodology Staff of the Census Bureau's Industry Division constructed the sampling frame, selected the sample, and generated estimates.

The ACES sampling frame was constructed from the Census Bureau's Standard Statistical Establishment List (SSEL) in November 1992 using final 1991 data for single unit (SU) establishments and 1990 data for establishments associated with multiunit (MU) firms. Major exclusions from the frame were Public administration, U.S. Postal Service, international establishments, establishments in Puerto Rico, Guam, Virgin Islands, and the Mariana Islands. EI Submasters which are SU records on the SSEL that are associated with MU establishments, establishments associated with agricultural production, and private households were also excluded from the frame.

The establishment-based file was consolidated into a company-based file. In addition, the 4-digit Standard Industrial Classifications (SIC) for each company was recoded into ACES categories. The 80 ACES categories consisted of either 3-digit SICs or combinations of 3-digit SICs. The ACES sampling frame included approximately two million companies.

3. The L-H Method Applied to ACES

We used a stratified simple random sampling design with two major strata. Stratum I was a take-all stratum that initially consisted of large companies with more than 500 employees and over $100 million in assets. Stratum II contained the rest of the sampling frame. Each company had frame information available for each of the ACES industries the company had activity in. Each stratum II company was classified into the ACES industry with the largest payroll. There were 80 ACES codes for sampling purposes. Subsequently, for each ACES industry category, stratum II was divided into three substrata, based on company size.

We considered several papers that documented methods for finding stratum bounds. Hess, Sethi, and Balakrishnan (1966) compared several stratifying techniques. The popular cum. √f rule (Cochran (1977)) was considered easy to implement but was initially ruled out because it does not figure in certainty strata. Sethi's method of using standard distributions was not used because we thought it would be cumbersome to identify the distribution and sub-optimal to use standard distributions for each of the 80 sampling ACES industries. Eckman's rule (1959) of equalizing the product of stratum weights and stratum range seemed to require rather ominous calculations. The L-H method was the most appealing in our application. Designed specifically for skewed populations which is often the case for economic surveys, it creates a certainty cutoff for the top stratum and boundary point(s) for the noncertainty portions of the sampling frame. Starting off with a standard statistical procedure (cum √f rule), we sought improvement when possible using the L-H method. Stratum II of the ACES sample design was the focal point for this methodology.

The application of the L-H method to the ACES 1992 preliminary survey sample design involved splitting stratum II into one certainty substratum and two noncertainty substrata for each ACES industry. The boundaries were derived for each industry by taking the partial derivative of the sample size with respect to a boundary while fixing the other boundary. However, in practice, we allowed both boundaries to move simultaneously. This results in an iterative process of minimizing the sample size for each industry subject to c.v. constraints. The total sample size is equal to the number of companies in stratum I and the stratum II sample sizes as a result of controlling separately within each industry. The total sample size equation as applied to ACES is

$$n = N_I + \sum_{i=1}^{80} n_i \qquad (1)$$

where

N_I = number of companies in stratum I, and

n_i = stratum II sample size for the ith industry using the following equation:

$$n_i = N_i W_{i,3} + \frac{N_i \left(\sum_{j=1}^{2} W_{i,j} \sigma_{i,j} \right)^2}{\frac{cv_i^2 \hat{Y}_i^2}{N_i} + \sum_{j=1}^{2} W_{i,j} \sigma_{i,j}^2}$$

where

N_i = number of stratum II companies in industry i,

$W_{i,j} = N_{i,j}/N_i$ (stratum proportion),

where $N_{i,j}$ = number of stratum II companies for substratum j in industry i,

$\sigma_{i,j}$ = standard deviation of payroll from the SSEL for substratum j in industry i of stratum II,

cv_i = desired coefficient of variation for industry i,

$\hat{Y}_i$ = estimated total payroll for industry j.

This equation uses optimum allocation and assumes a fixed cost for each substratum in stratum II (i.e. Neyman's allocation).

For the estimated total payroll, $\hat{Y}$, stratum I companies could contribute to more than one industry depending on the number of different activities in which the companies were involved. However, stratum II companies contributed differently. Stratum II companies were classified into one industry, even if engaged in more than one activity, and the company's payroll contributed to the estimate only from that industry.

The reliability level for each industry was an expected coefficient of variation (c.v.) of 5% on payroll. It was not known, however, what standard errors would result for capital expenditures, as no capital expenditures data exist for the frame records. Companies responding in ACES industries different from the ones they contributed to in the sample design also caused the c.v.'s to fluctuate. The total number of companies selected for the ACES 1992 preliminary survey was 11,194, consisting of 1,500 stratum I companies and 9,694 stratum II companies.

4. L-H Algorithm for the ACES

1. The process begins by generating starting boundaries from some other source. These could have been arbitrarily selected or calculated by another method. In ACES the cum. $\sqrt{f}$ rule is used. Starting boundaries are important because ending L-H boundaries may differ for different starting boundaries.

2. Population means, variances, standard deviations, estimated payroll total, counts, and proportions are generated for each substratum.

3. The sample size, n_i, is calculated using the sample size equation above.

4. The sample size, n_i, is allocated to the three substrata. Neyman's Allocation is used for ACES. The allocation below allows for the analytical certainties in substratum 3 to be subtracted with no variation from substratum 3 contributing to the formula.

$$n_{i,j} = (n_i - N_{i,3}) \frac{W_{i,j}\sigma_{i,j}}{\sum_{j=1}^{2} W_{i,j}\sigma_{i,j}}$$

for i = 1, 2, ..., 80 and for j = 1, 2.

5. Two new boundaries are calculated by entering the statistics obtained in step 2 into the L-H method formulae. These formulae are the result of minimizing the sample size subject to c.v. constraints (See Lavallée and Hidiroglou (1988)).

6. Return to step 2 using the newly defined substrata if the boundaries are not close to converging. One of the problems with the L-H Method is that it sometimes takes a large number of iterations before the boundaries converge; sometimes they never converge. Generally after just a few iterations, a large proportion of the improvement in the sample size has already occurred. Contour plots also show the marginal improvement in the sample size by illustrating that when the bottom of the surface is reached, moving on is unnecessary. At this point, most of the improvement on the sample size from iteration to iteration is less than a value of one. Therefore, the computer program will stop processing when one of the following occurs:

> 1) the difference between the new upper boundary and the previous iteration's upper boundary is less than one,
>
> 2) the difference between the new lower boundary and the previous iteration's lower boundary is less than one,
>
> 3) the difference between the new sample size and the previous iteration's sample size is less than 0.1, or
>
> 4) the program goes into the 30th iteration.

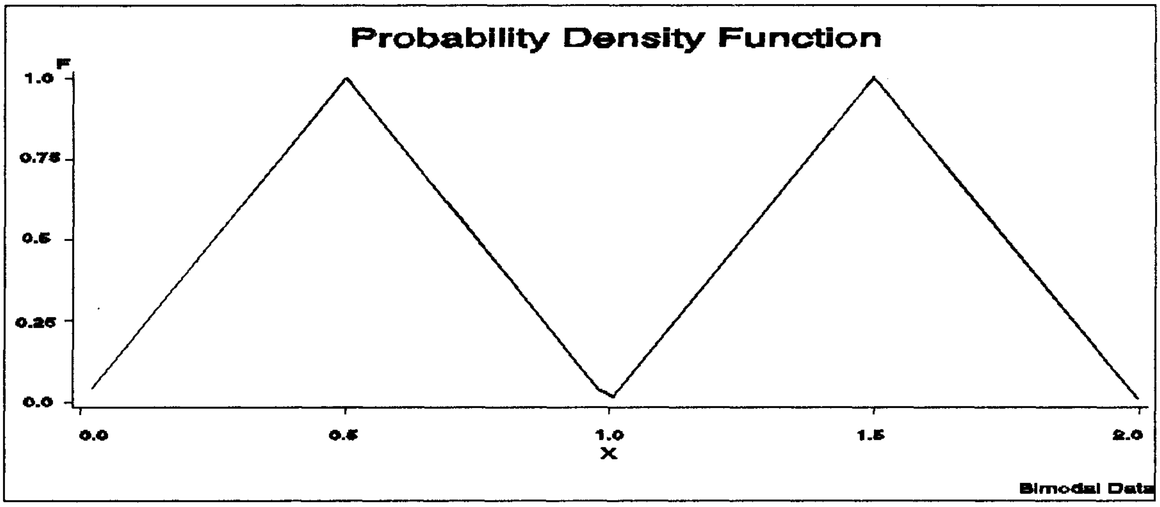
5. Illustration

The following is a distribution from Schneeberger's paper.

$$f(x) = \begin{array}{lll} 0 & & x <= 0 \\ 2x & & 0 <= x <= 0.5 \\ 2(1-x) & \text{if} & 0.5 <= x <= 1 \\ 2(x-1) & & 1 <= x <= 1.5 \\ 2(2-x) & & 1.5 <= x <= 2 \\ 0 & & 2 <= x \end{array}$$

The density function is pictured in Figure 1. With Schneeberger's application of estimating means, the objective function for the nonlinear program is $z = (\Sigma W_h \sigma_h)^2$ using L = 3 noncertainty strata. The results are: boundary points $b_1 = 0.50241$ and $b_2 = 1.03985$ yield a minimum, boundary points $b_1 = 0.7091$ and $b_2 = 1.2909$ yield a saddle-point, and boundary points $b_1 = 0.96015$ and $b_2 = 1.49759$ also yield a minimum.

Figure 1



Four thousand observations were generated with this distribution. With our application of estimating totals, when minimizing the sample size, subject to a fixed coefficient of variation (c.v. = 0.05), the resulting surface from the L-H method was created in three dimensions in Figure 2. The graph shows the saddle-point and the two minima. Figure 3 shows the same surface in the form of a contour plot.

As described in Schneeberger's paper, on the line $b_2 = 2 - b_1$, the gradient method moves the gradient along the line $b_2 = 2 - b_1$ into the saddle-point. When we set the starting boundaries on this line, specifically $b_1 = 0.6$ and $b_2 = 1.4$, the L-H method converged to a local minimum (see Table 1). The L-H method seemed to head toward the saddle-point around iteration 4 and then eventually descend toward the minimum where $b_1 = 0.967$ and $b_2 = 1.51$ at the 71st iteration. With different starting boundaries that are not on the line, specically $b_1 = 0.6$ and $b_2 = 1.2$, we got different ending results. The L-H method converged quickly (16 iterations) to a different local minimum where $b_1 = 0.514$ and $b_2 = 1.045$. This problem is not unique to the L-H method, as Schneeberger points out that the gradient method's resulting boundaries are dependent of the starting boundaries. The absolute difference in sample size is 0.107. In application, there is not very much difference between final sample sizes with either starting boundary.

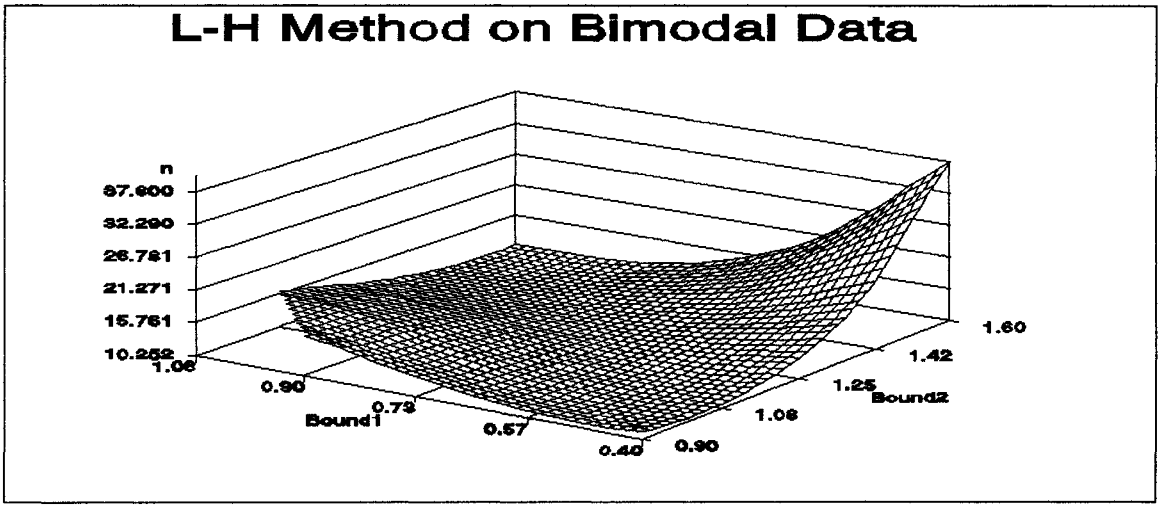
One can see from the plots how fast the L-H method converges, that is, in the first few iterations a slight shift in the boundaries result in a large reduction in sample size. We found that it took a short time to get from the starting point to flat area. The plots also show that there is a large area in which there is little gain in sample size reduction. A large proportion of the iterations usually occurs in this flat area, therefore we stop processing.

6. Sources of Error in the ACES Sample Design

In the above example, wee also used initial boundaries generated by the cum. $\sqrt{f}$ rule. The cum. $\sqrt{f}$ rule works well in the example because all three strata are noncertainty strata. However, economic data are usually highly skewed and therefore it is more appealing to have a certainty stratum. In our application, the cum. $\sqrt{f}$ rule assumes that all resulting strata will be sampled. The L-H method is written to construct an analytical certainty substratum. Therefore, the top stratum developed by the cum. $\sqrt{f}$ rule, when creating the initial boundaries for ACES industries, will be top-heavy since it will not be sampled. Improvements in the sample size were noticed from the cum. $\sqrt{f}$ rule to the first iteration of the L-H method in this situation. The error that occurs is that the starting boundaries may lead to a local minimum that is not the best solution.

Another concern is the result of companies self-reporting their capital expenditures into ACES industries on the ACES questionnaire. We classified each company into its highest payroll industry for sampling purposes, however, companies may report in multiple industries on the questionnaire and some do not report in the industry they are sampled in. If too many companies self-report into industries other than where

Figure 2



they were classified, then control on the reliability of the estimates is lost.

A similar concern is that the distribution for payroll is not the same as the distribution for expenditures. Since sample size is directly related to the variance, sample sizes may be different than what is really required. Therefore, since the correlation between payroll and expenditures is not high, the chances that reliability constraints will be met will diminish.

7. Programming

The programming of this algorithm was done using SAS. The program ran efficiently through the use of temporary arrays and also because of the four stopping rules listed above. For the 1992 preliminary survey, the L-H program was submitted for each of the 80 ACES industry groupings.

8. Conclusions and Future Research

In closing we would like to reiterate the following points of interest in our application to the ACES. The graphs presented here have shown that a wide range of boundary values result in a small range of sample sizes when in a neighborhood around an optimal value (the bowl shape bottom of the graphs). Any extraordinary improvement on the sample size, i.e. a small marginal gain, might not be worth the extra effort to obtain. This marginal gain may or may not even improve the sample size since the sample size is really an integer and the marginal gain might only be a small fraction. The L-H program proved very effective in obtaining boundary values in a desired neighborhood around an optimal value, and did it relatively fast.

By measuring the rate of convergence using the sample size instead of boundary values we were better able to determine when a desired neighborhood around an optimal value was reached. This is because boundary values vary greatly in such a neighborhood while sample size (which is of main interest) varies slightly. When the improvement in sample size from iteration to iteration was marginal or nonexistent we immediately terminated the program under the assumption that we reached the desired neighborhood.

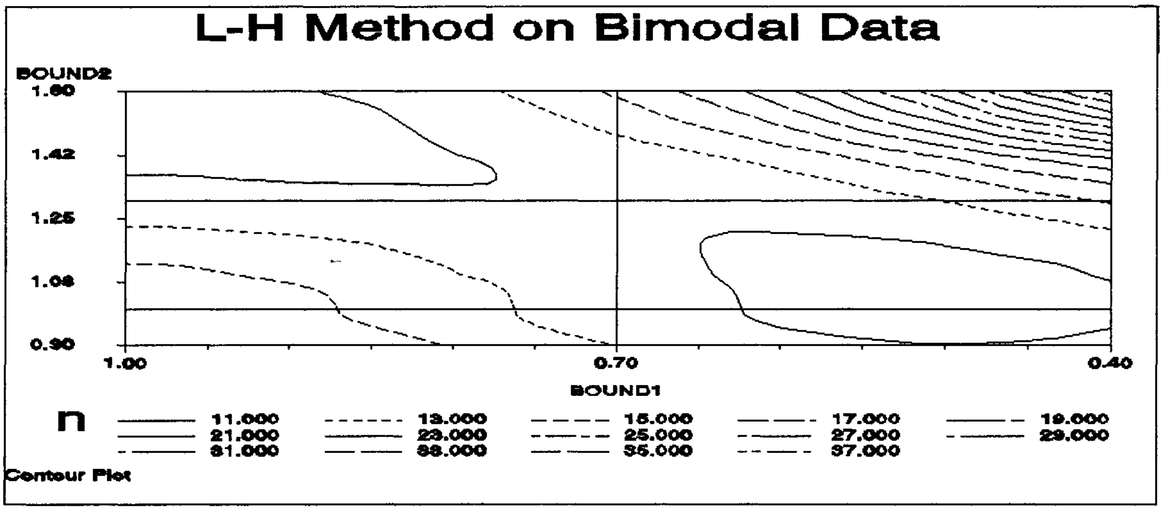
The L-H method was an ingenious way of obtaining strata boundaries based on an auxiliary variable. It's the authors' recommendation that this method not only be continued, but incorporated with other methods so that c.v. constraints based on expenditure data be met. As mentioned before there were two major reasons why c.v. constraints were not met.

1) C.V. constraints were based on payroll, not expenditures.

2) Companies were allowed to report in multiple industries, but were sampled in a single industry.

We can solve Problem 1) using expenditure data from the prior-year survey. Once the stratum boundaries have been established using the L-H program on current payroll data, we can obtain sample estimates of variance for expenditure data for each ACES industry within each stratum. We can address Problem 2) with a program that implements Chromy's

Figure 3



Algorithm (Zayatz and Sigman (1993)) for determining sample sizes subject to multiple c.v. constraints. With this algorithm, we can sample companies in multiple industries while meeting all industry c.v. constraints. To combine these procedures, we would perform the following three steps:

1) Use the L-H program to obtain stratum boundaries and population sizes within each stratum, based on current payroll data.

2) Use expenditure data from the prior year ACES to compute sample estimates of variance for each ACES industry within each stratum defined in Step 1).

3) Use the population sizes and estimates of variance in Chromy's Algorithm to obtain sample sizes for each stratum.

9. References

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed., New York: John Wiley and Sons.

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

ECKMAN, G. (1959). "An Approximation Useful in Univariate Stratification." *The Annals of Mathematical Statistics*, Vol. 30, pp. 219-229.

HESS, I., SETHI, V. K., and BALAKRISHNAN, T. R. (1966). "Stratification: A Practical Investigation," *Journal of the American Statistical Association*, March, pp. 74-90.

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

SCHNEEBERGER, H. (1979). "Saddle-Points of the Variance of the Sample Mean in Stratified Sampling." *Sankhya: The Indian Journal of Statistics*, Vol. 41, Series C, Pt. 1, pp. 92-96.

ZAYATZ, L. and SIGMAN, R. (1993). "Feasibility Study of the Use of Chromy's Algorithm in Poisson Sample Selection for the Annual Survey of Manufacturers." *Economic Statistical Methods Division (ESMD) Report Series*, Census Bureau ESMD-9305, August 1993.