Survey Methodology, June 1996
Vol. 22, No. 1, pp. 65–75
Statistics Canada

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

JOHN G. SLANTA and THOMAS R. KRENZKE[1]

ABSTRACT

The Lavallée-Hidiroglou (L-H) method of finding stratification boundaries has been used in the Census Bureau's Annual Capital Expenditures Survey (ACES) to stratify part of its universe in the pilot study and the subsequent preliminary survey. This iterative method minimizes the sample size while fixing the desired reliability level by constructing appropriate boundary points. However, we encountered two problems in our application. One problem was that different starting boundaries resulted in different ending boundaries. The other problem was that the convergence to locally-optimal boundaries was slow, *i.e.*, the number of iterations was large and convergence was not guaranteed. This paper addresses our difficulties with the L-H method and shows how they were resolved so that this procedure would work well for the ACES. In particular, we describe how contour plots were constructed and used to help illustrate how insignificant these problems were once the L-H method was applied. This paper describes revisions made to the L-H method; revisions that made it a practical method of finding stratification boundaries for ACES.

KEY WORDS: Convergence; Contour plots; Economic surveys.

1. INTRODUCTION

The primary objectives of the sample design of the Census Bureau's Annual Capital Expenditures Survey (ACES) are to meet desired reliability levels using operationally-feasible methodology and to stay within budget limitations. To achieve these goals, we implemented a stratified simple random sample design using a modified version of Lavallée and Hidiroglou's (L-H) (1988) approach of finding stratum bounds. This stratification 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.

We considered several papers that documented other methods for finding size stratum boundaries. Hess, Sethi, and Balakrishnan (1966) compared several stratifying techniques. The popular Dalenius and Hodges method (Cochran 1977, p. 129) was considered easy to implement in our case but was initially ruled out because it was not designed with certainty strata in mind. Sethi's method (1963) 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 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 to our application. Designed specifically for skewed populations, which is often the case for economic surveys, it creates a boundary that defines the take-all stratum, and the optimal boundary point(s) for the take-some strata. It sometimes will create additional take-all strata if through Neyman Allocation, the stratum sample size is greater than or equal to the stratum size.

The L-H method goes through an iterative algorithm beginning with computing or arbitrarily setting the initial stratum boundaries. Then, stratum statistics are computed such as, the stratum size, mean, and the variance. These parameters are entered into boundary formulas that were derived from minimizing the sample size subject to a desired cv. If the new boundaries do not converge then the stratum statistics are calculated for the newly defined size strata. The cycle continues until the boundaries converge.

Schneeberger (1979) discussed the problem of finding optimal stratification boundaries. Schneeberger shows in the paper that when expressing this problem as a non-linear program, when solved by a gradient method, the solution may be relative or global minima, maxima, or saddle points of the variance of the sample mean. Detlefsen and Veum (1991) document this as a shortcoming of the L-H method when testing its application for the Census Bureau's Monthly Retail Trade Survey. In the L-H method, they found that many times the resulting boundaries differed substantially from where the initial boundaries were set,

[1] John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

so the minimum sample size attained was a local minimum. Geometrically, the sample size as a function of two strata boundaries, 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. Therefore, one objective is to have initial boundaries that are in the neighborhood of the global minimum. Using starting boundaries resulting from a technique such as the Dalenius and Hodges method may help satisfy this desire.

Detlefsen and Veum (1991) also noted instances of slow or non-convergence. However, they also noted 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 order to defend ourselves against infinite loops in the computer program or a large number of iterations, we decided on doing two things. First, we implemented a sample design in which the L-H method would create sets of only three size strata. Second, we decided to implement stopping rules so that when the convergence rate appeared to slow down, the program stopped processing.

In this work, we give background information on the ACES and briefly describe the way the L-H method was applied. We show how contour plots and three-dimensional plots gave us justification for using the L-H method to get the final boundaries. We show how the contour plots address the convergence problem by showing how constraints can be setup to be met after each iteration. This would protect us against slow or non-convergence under the assumption that the marginal gain achieved is not worth the extra effort.

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 Classification (SIC) codes for each company were 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 THE ACES

The universe of companies was classified into two major strata. Stratum I was an arbitrarily defined take-all stratum that consisted of large companies with more than 500 employees and over $100 million in assets. Stratum I companies were not classified into one ACES industry. For the estimated industry level payroll totals used in the calculation of the industry-level sample sizes, stratum I companies could contribute to more than one ACES industry depending on the number of different ACES industries the companies have payroll in, identified in the SSEL.

Stratum II contained companies that had five or more employees and had less than 500 employees. Stratum II companies were classified into one industry, even if engaged in more than one activity. Each company had frame information available for each of the ACES industries the company had activity in. However, the company's payroll contributed only to estimated total payroll for the industry that the company was classified in. Subsequently, within stratum II, for each ACES industry category, three size strata were created based on total company annual payroll using the L-H method.

A concern with the sample design is the result of companies being misclassified due to the measure of size being used. We classified each stratum II company into its highest payroll industry; however, companies self-report their capital expenditures into ACES industries on the ACES questionnaire. Companies may report in multiple

industries. If too many companies self-report into industries other than where they were classified, then control on the reliability of the estimates is lost.

A similar concern is that the variation in payroll is not the same as the variation in 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.

The application of the L-H method to the ACES 1992 preliminary survey sample design involved splitting stratum II into one take-all size stratum and two take-some size strata 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. Within stratum II for each ACES industry and assuming Neyman Allocation (Detlefsen and Veum 1991), the sample size equation that is minimized is,

$$n = n_{TA} + \frac{N\left(\sum_{j=1}^{2} W_j S_j\right)^2}{\frac{cv^2 Y^2}{N} + \sum_{j=1}^{2} W_j S_j^2}, \qquad (1)$$

where, n_{TA} is the number of companies in the take-all size stratum within stratum II defined by the L-H method, N is the number of stratum II companies in the ACES industry of interest, $W_j = N_j/N$ is the stratum proportion, N_j is the number of stratum II companies for size stratum j, cv is the desired coefficient of variation for the ACES industry of interest, Y is the total payroll for stratum I and II for the ACES industry of interest defined by,

$$Y = \sum_{k=1}^{N_I} y_k + \sum_{j=1}^{3} \sum_{i=1}^{N_j} y_{ji},$$

N_I is the number companies in stratum I, and S_j is the standard deviation of payroll from the SSEL for size stratum j in stratum II defined by,

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}},$$

where, y_{ji} is the payroll value of company i of size stratum j for the ACES industry of interest, and $\bar{Y}_j$ is the mean of payroll for size stratum j.

The reliability level for each industry was an expected 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. CONVERGENCE INTO NEIGHBORHOODS

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. Our goal was to be able to implement stopping rules so that when an area around a local minimum is reached, we can stop processing. This prompted our use of contour plots in analyzing the effect the boundaries have on the resulting sample size. It also allowed us to get a graphical view of the neighborhoods around the local minima. We will use two distributions to illustrate the benefits of reviewing contour plots.

4.1 Non-Skewed Distribution

The first example is a non-skewed distribution from Schneeberger's paper. This distribution is symmetric at $x = 1$ as shown in Figure 1.

$$f(x) = \begin{cases} 0 & x \le 0 \\ 2x & 0 < x \le 0.5 \\ 2(1-x) & 0.5 < x \le 1 \\ 2(x-1) & 1 < x \le 1.5 \\ 2(2-x) & 1.5 < x \le 2 \\ 0 & 2 < x \end{cases} \quad \text{for}$$

Schneeberger's objective was to find boundaries for three take-some strata using a gradient method. Using the objective function of $z = (\sum W_h \sigma_h)^2$, the results attained are listed in Table 1.

Table 1
Optimum Boundaries for Non-Skewed Distribution

	b_1	b_2	Optimum Point
(2a)	.50241	1.03985	Minimum
(2b)	.70910	1.29090	Saddle Point
(2c)	.96015	1.49759	Minimum

Source: Schneeberger (1979).

Table 2

L-H Boundaries for Three Take-Some Strata for Non-Skewed Distribution

N	Starting Method	1st Iteration			Iteration Within 5% of Sample Size				Final Iteration			
		b_1	b_2	n	b_1	b_2	n	iter.#	b_1	b_2	n	iter.#
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	2	.70	1.31	9.77	4
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	2	.70	1.30	10.55	5
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	2	.71	1.29	10.99	6
1000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	2	.71	1.29	11.37	7
5000	$N_1 = N_2 = N_3$	.59	1.41	14.12	.66	1.34	11.84	2	.71	1.29	11.45	9
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	1	.77	1.37	9.63	4
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	7	.93	1.47	9.65	13
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	7	.95	1.49	9.96	17
1000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	8	.96	1.50	10.27	23
5000	Dalenius-Hodges	.70	1.40	11.95	.86	1.42	10.74	8	.96	1.50	10.34	28
50	Off Line	.50	1.30	10.87	.57	1.20	9.43	3	.55	1.11	9.11	6
100	Off Line	.50	1.30	11.95	.57	1.18	10.04	3	.53	1.07	9.65	8
200	Off Line	.50	1.30	12.64	.56	1.14	10.28	4	.51	1.05	9.96	12
1000	Off Line	.50	1.30	13.24	.56	1.14	10.59	4	.50	1.04	10.27	18
5000	Off Line	.50	1.30	13.37	.56	1.14	10.67	4	.50	1.04	10.34	24

We generated five datasets of different sizes (*e.g.*, $N = 50$, 100, 200, 1000, and 5000) using the formula, $F(x) = (j - 1/2)/N$. For this example, we adapted the L-H method to construct three take-some strata and no take-all stratum in order to compare our results with the results in the Schneeberger paper. With our application of estimating totals, when minimizing the sample size subject to a c.v. = 0.05, the L-H method ran for each of the five population sizes using three different starting techniques. The results are given in Table 2.

There are three main points from the information in Table 2. First, the algorithms convergence depends on the population size. The underlying theory of the L-H method is based on continuous distributions. Our examples and any survey application has discrete data from finite populations. It is also apparent that as N gets larger, the resulting boundaries get closer to where the minimum is under an infinite population size. Figure 2 shows the roughness of the sample size surface when N is small (*i.e.*, $N = 50$). The resulting surface illustrates the saddle in three dimensions in Figure 2. In this graph, the axes are the lower and upper boundaries and the surface is the resulting sample sizes. This graph shows the saddle-point, the two local minima, and it also gives a picture of the magnitude of the sample size reductions as a result of shifting the boundaries. In contrast, Figure 3 shows the smoothness of the surface when N is large (*i.e.*, $N = 5000$). From this, it seems that the roughness of the sample size surface and consequently the population size has an effect on where the boundaries converge.

The second point of this example reemphasizes that the ending boundaries are dependent on the starting boundaries. For this example, Schneeberger describes that with a starting point symmetric to $x = 1$, where $b_1 = 1 - \lambda$ and $b_2 = 1 + \lambda (0 < \lambda < 1)$ which defines 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, which occurred when we started with the condition $N_1 = N_2 = N_3$, the L-H method also converged to the saddle point (see Table 1). With starting boundaries from the Dalenius-Hodges method, which are not on the line in the case where $b_2 > 2 - b_1$, the L-H method converged to a minimum (2c). The Dalenius-Hodges method works well in this example because of the three take-some strata. With starting boundaries which are not on the line in the case where $b_2 < 2 - b_1$ (specifically, $b_1 = .5$ and $b_2 = 1.3$), the L-H method converges to a different minimum (2a). This problem is not unique to the L-H method, as Schneeberger points out that the gradient method's resulting boundaries are also dependent of the starting boundaries.

The third point of this example is that there seems to be relatively large reductions in sample size in the first few iterations and then there are several iterations where there are small reductions in sample size. Results are shown in Table 2 from the iteration in which the algorithm produced a sample size within 5% of the final sample size. This implies that the L-H algorithm quickly goes to a neighborhood around an optimal boundary. While close to an optimal sample size, there seems to be a wide range of boundary points resulting in a small range of sample sizes. The point is that stopping rules can save computing time while not relinquishing any real reduction in sample size, since sample size is in integer values.

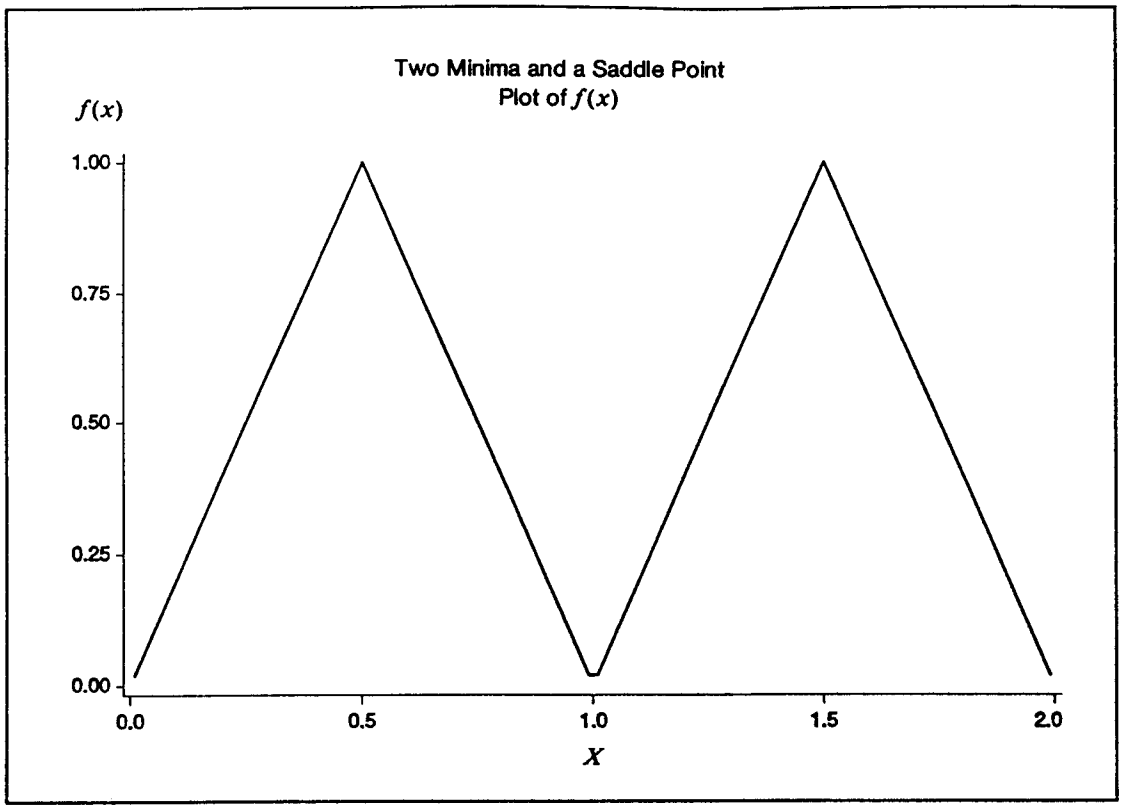

Figure 1. Graph of non-skewed distribution.

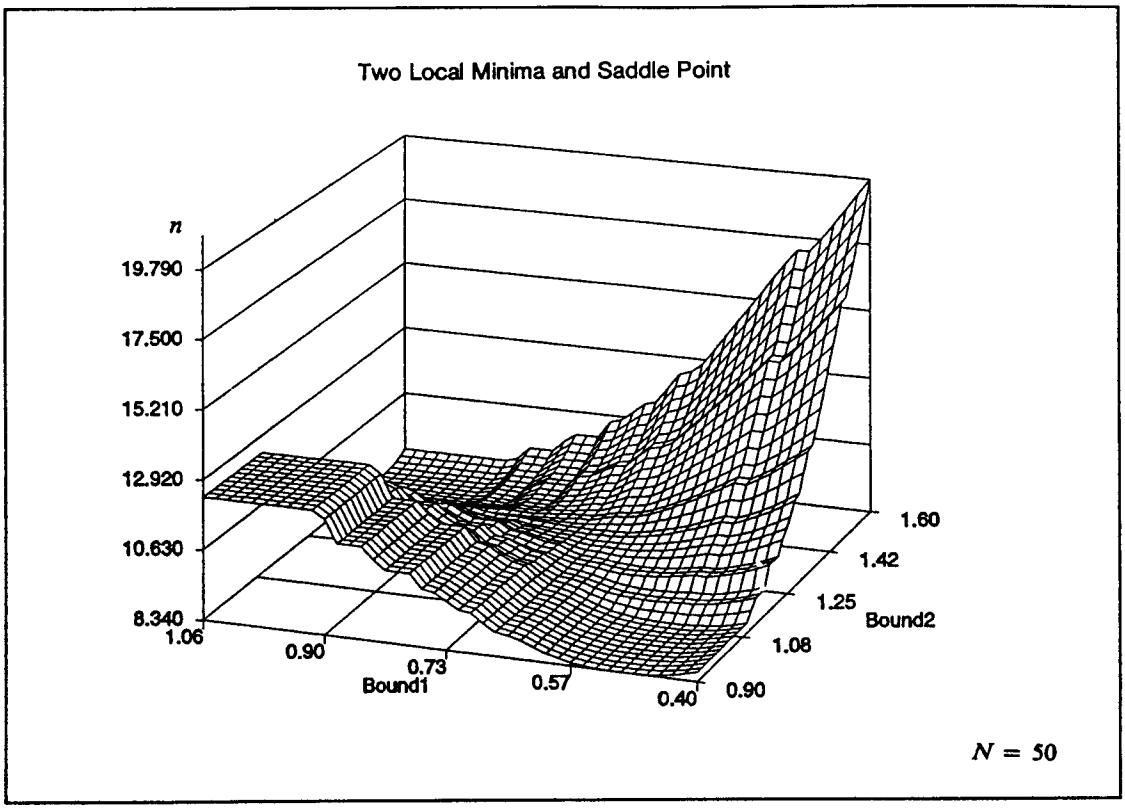

Figure 2. Sample size surface for non-skewed distribution ($N = 50$).

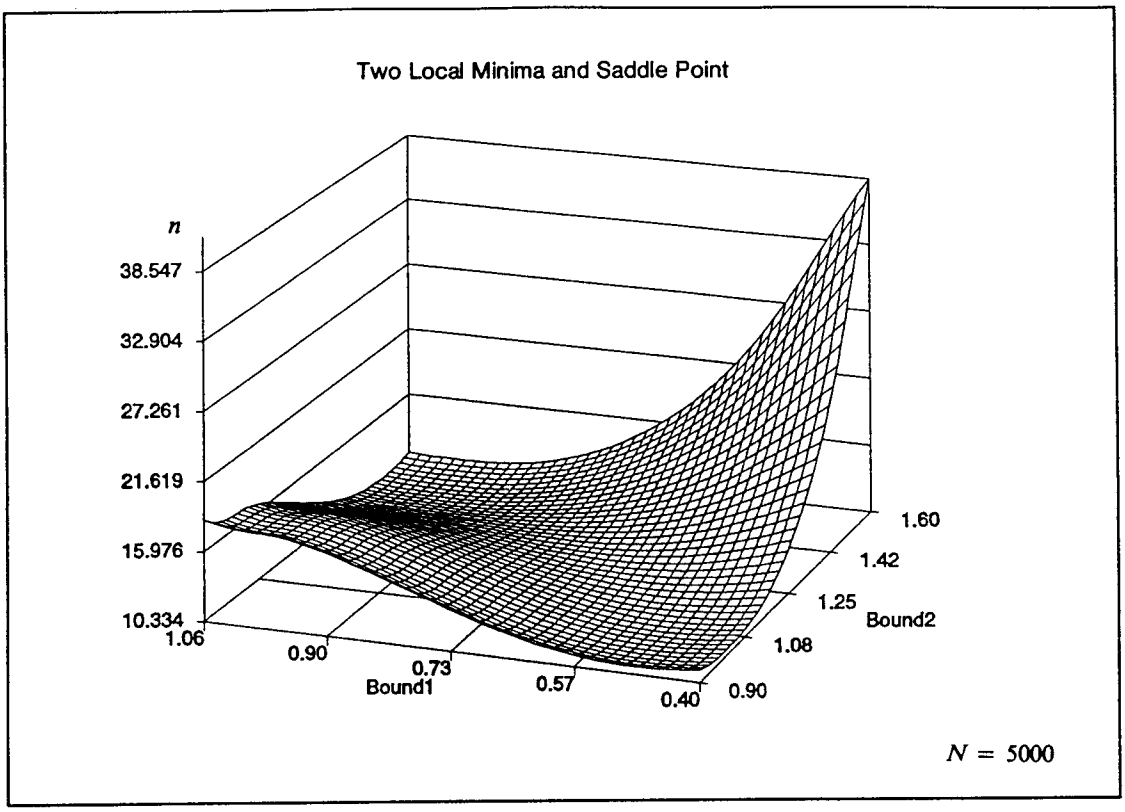

Figure 3. Sample size surface for non-skewed distribution ($N = 5000$).

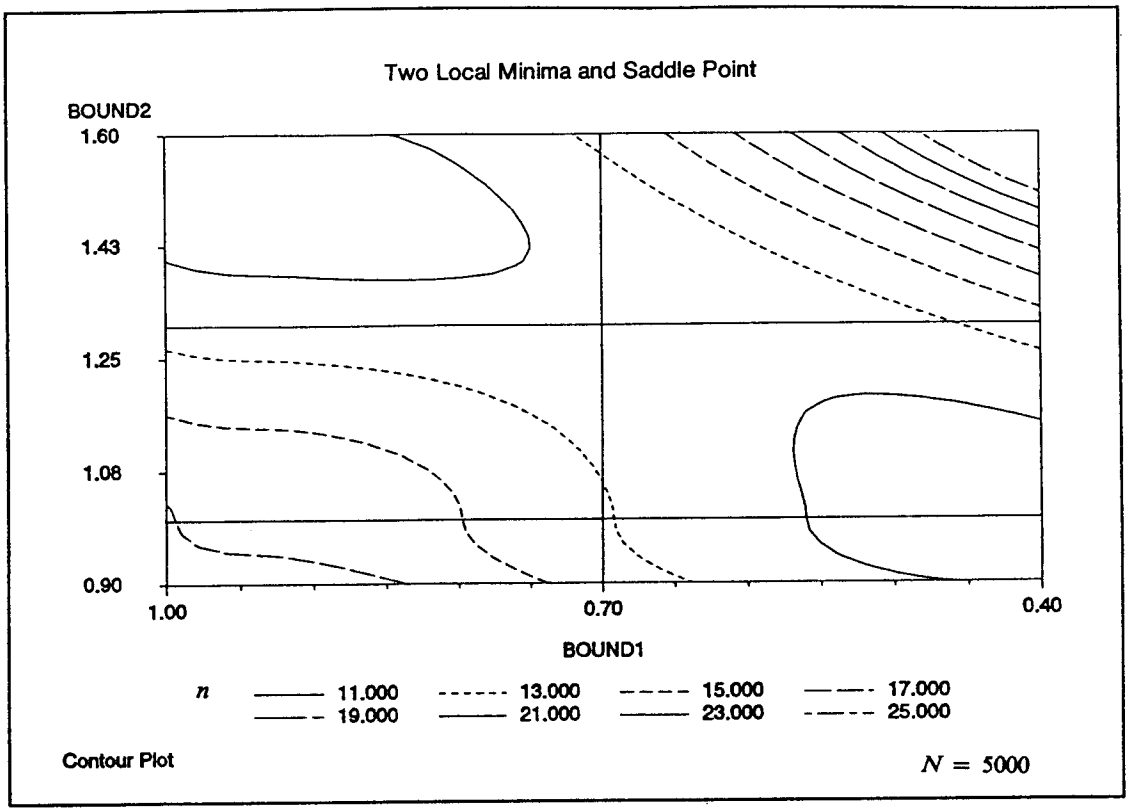

Figure 4. Contour plot for non-skewed distribution ($N = 5000$).

A contour plot of the surface shown in Figure 3 is given in Figure 4. Again, the axes are the lower and upper boundaries and the surface is defined by the resulting sample size. The lines in the plot represent a sample size value. The space between the lines gives an area that contains a range of sample size values. For example, a solid line represents a sample size of 11 and a series of short dash marks represents a sample size of 13. The area in between the solid line and the line of short dash marks contains sample sizes in the range of 11 to 13. This contour plot shows a marginal improvement in the sample size by illustrating that when an area around 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. It becomes apparent that after the first few iterations, the improvement of the sample size from iteration to iteration reduces quickly. For instance, in Table 2, where $N = 5000$ and where the Dalenius-Hodges method was used for the starting boundaries, the first eight iterations accounted for 74% of the total reduction in the sample size from iteration 1 to the 28th and final iteration.

4.2 A Skewed Distribution

Economic data are usually highly skewed and therefore it is more appealing to have a take-all stratum. The next example comes from the Pareto distribution, which is a very typical distribution of economic universes, where there are a large number of small companies and a small number of large companies.

The Pareto distribution function is defined as $F(x) = 1 - 1/(1 + x)^b$, $0 \leq x < \infty$. From this we again generated five datasets of different sizes using the formula $F(x) = (j - 1/2)/N$. We let the values of b change as the population size changed. This was done so as to keep the upper tail of the finite discrete distribution roughly the same proportion to the entire population for each population size. To do so, the parameter b was chosen in such a way that about 90% of the total sum could be accounted for in the top 20% of all possible sampling units. Since the datasets contain a finite number of discrete values there was no problem deriving variances of different strata when values of b were less than 2.

Table 3 gives the L-H results for different population sizes and starting points. The first group uses starting values which yield equal stratum populations ($N_1 = N_2 = N_3$). The second group uses the Dalenius-Hodges method to obtain all initial boundaries. The third group obtains starting boundaries by first using a method for determining the take-all boundary as presented by Hidiroglou (1986) and uses the Dalenius-Hodges method for the other boundary. Again it can be observed that the sample size surface given strata boundaries is much more choppier for smaller population sizes (see Figure 5). For example, when $N = 50$ and b_1 is fixed, there was only one sample size when b_2 varied between 11.8 and 14.7. This is because there were no values within this range in the population. As the population size increases, the data values are closer together, and the sample surface becomes very smooth (see Figure 6).

Table 3

L-H Boundaries for Skewed Distribution (one take-all stratum, two take-some strata)

N	Starting Method	1st Iteration					Iteration Within 5% of Sample Size					Final Iteration					
		b	b_1	b_2	n_{TA}	n	b_1	b_2	n_{TA}	n	iter.#	b	b_1	b_2	n_{TA}	n	iter.#
50	$N_1 = N_2 = N_3$	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4	9
100	$N_1 = N_2 = N_3$	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1	12
200	$N_1 = N_2 = N_3$	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9	13
1000	$N_1 = N_2 = N_3$	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3	18
5000	$N_1 = N_2 = N_3$	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8	22
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4	6
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1	9
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9	10
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3	15
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8	19
50	Hidiroglou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4	7
100	Hidiroglou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1	11
200	Hidiroglou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9	10
1000	Hidiroglou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3	15
5000	Hidiroglou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8	19

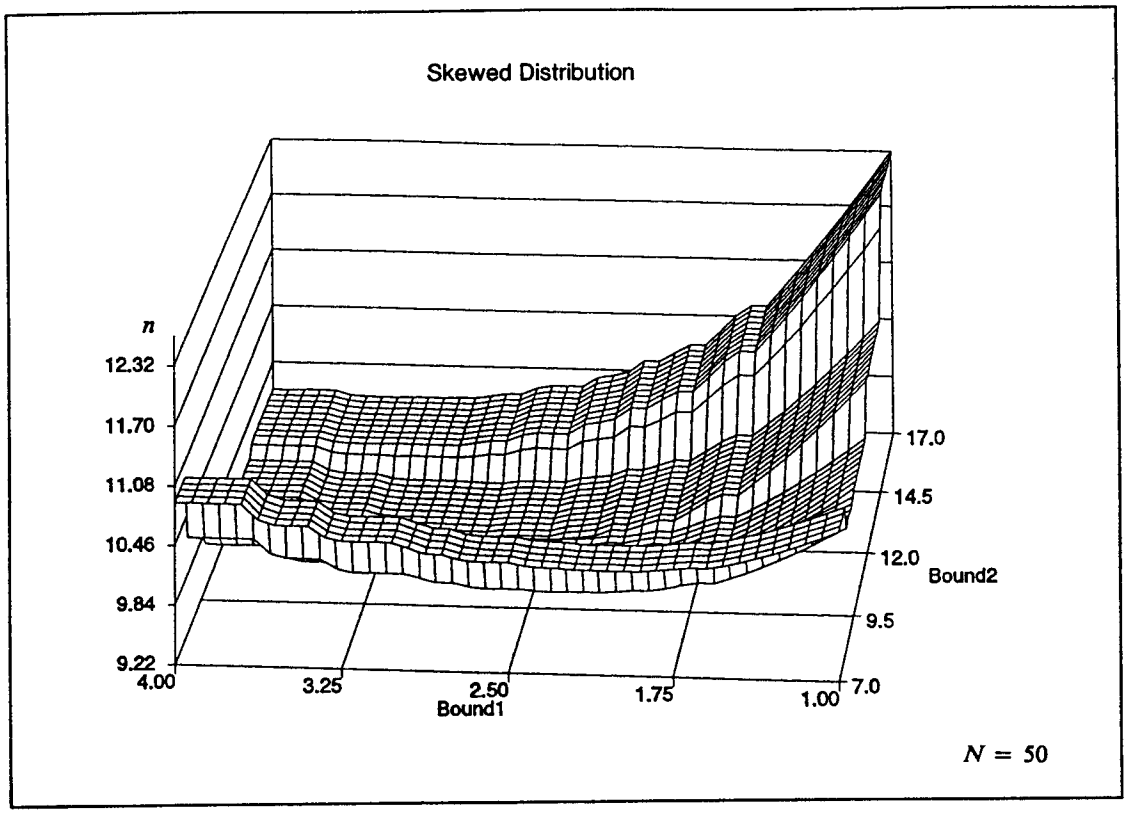

Figure 5. Sample size surface for skewed distribution ($N = 50$).

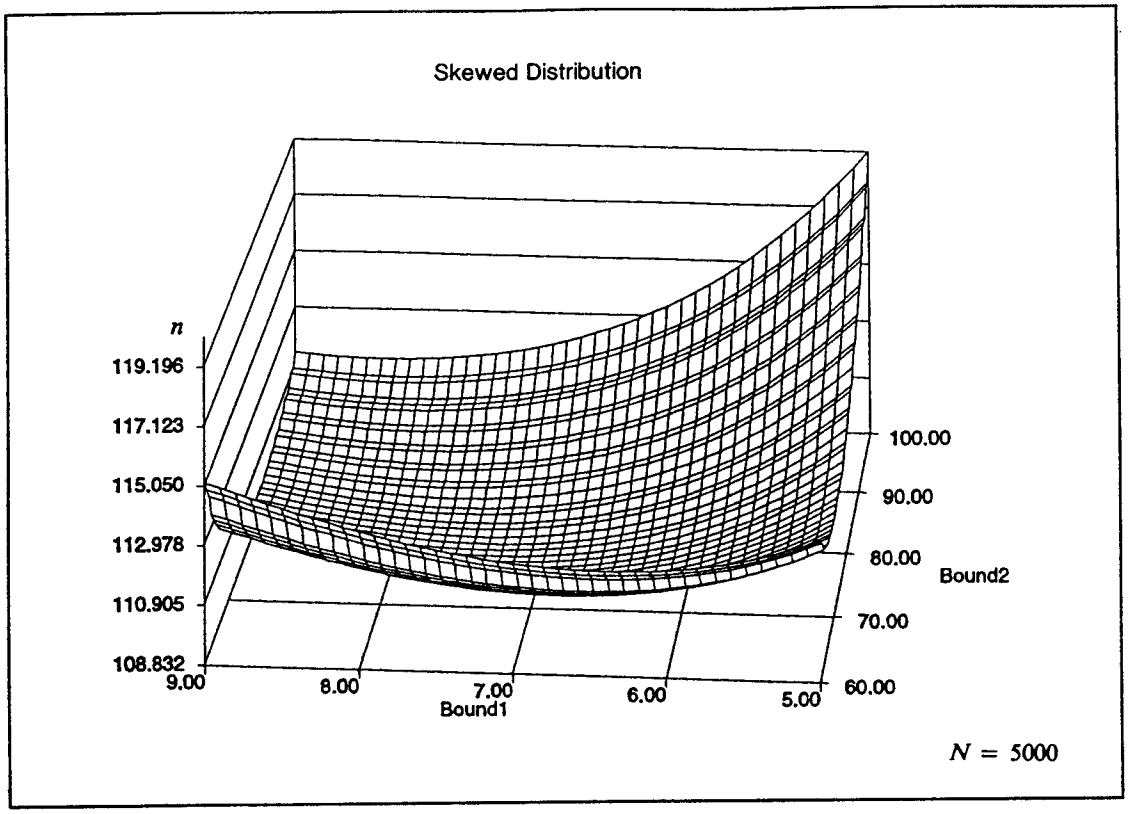

Figure 6. Sample size surface for skewed distribution ($N = 5000$).

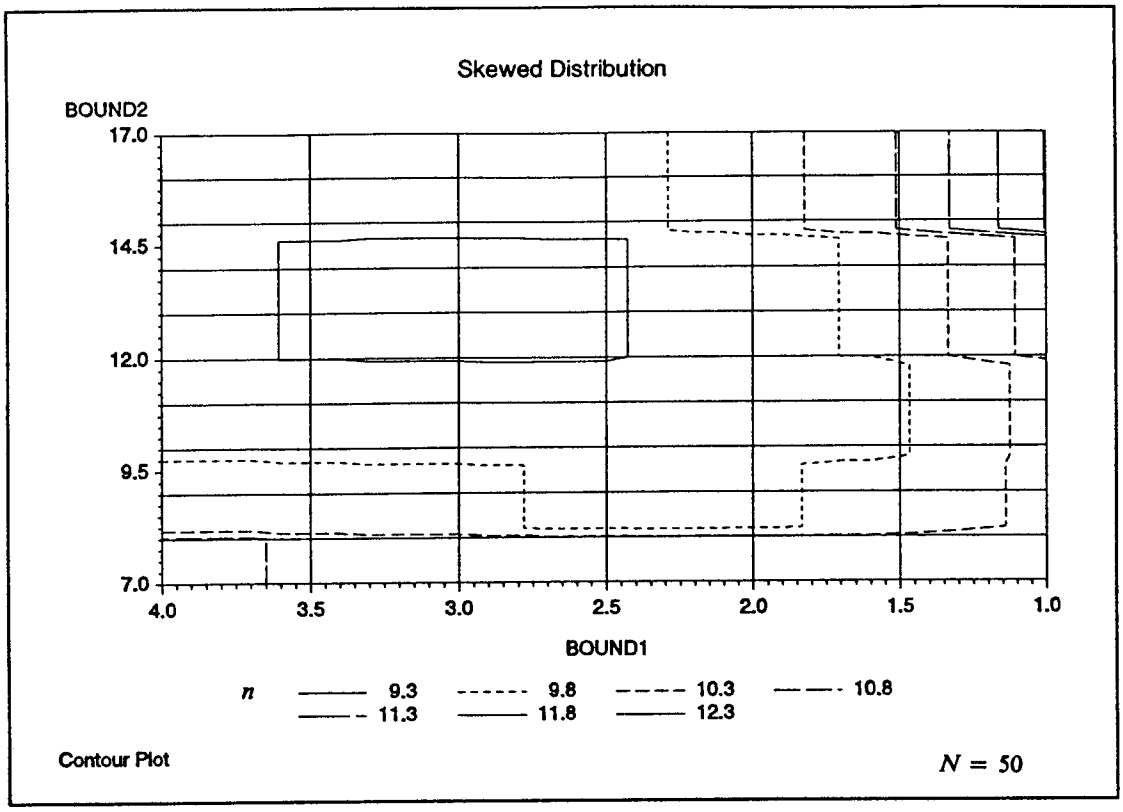

Figure 7. Contour plot of skewed distribution ($N = 50$).

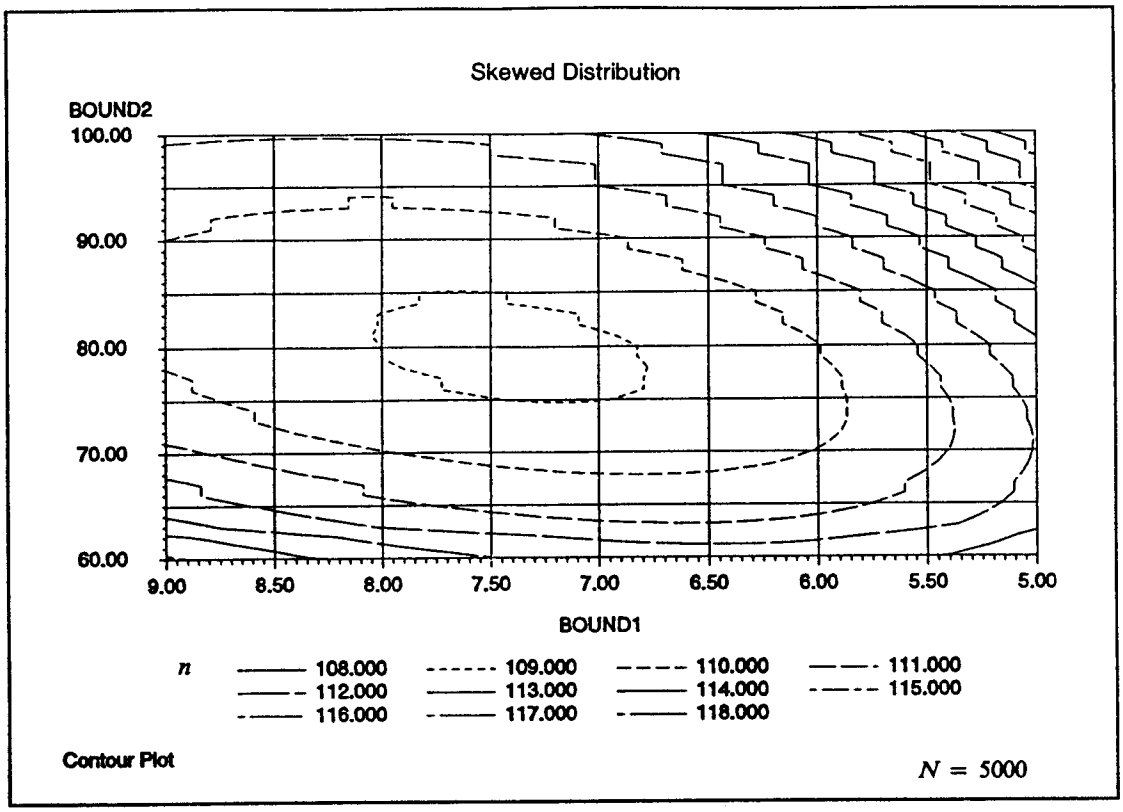

Figure 8. Contour plot of skewed distribution ($N = 5000$).

The contour plot for $N = 50$ (Figure 7) has erratic shapes defined by straight lines for contour markings. The contour plot for $N = 5000$ (Figure 8) has almost smooth concentric ellipses for contour markings. It would appear to be a desirable quality for the contour markings to be the same shape and concentric. This would imply that the global minimum is the only local minimum.

The contour plot for $N = 50$ demonstrated the case where the L-H method didn't converge to optimal boundaries. Since, for this example, we let the L-H program run until it converged the question may arise as to why the L-H method didn't converge to the optimal boundaries. The easiest way to explain this is by viewing Figure 5. We can see that when the population size is small then the sample size surface is not as smooth as in Figure 6. We see several major ridges in Figure 5 that are caused by wide gaps in the skewed discrete data ($x_{43} = 9.71$, $x_{44} = 11.81$, $x_{45} = 14.79$, $x_{46} = 19.29$). This means that for a given b_1, any value of b_2 between 11.81 and 14.79 would yield the same sample size. When we ran the L-H program for different starting boundaries other than the three listed in Table 3 we came up with the final boundaries as in Table 3 along with other boundaries and their corresponding sample sizes. It appears that the L-H method converges to a low region on one of the major ridges, provided that the region is in the neighborhood of the optimal boundaries. The minimum sample size is 9.22 and the L-H method in Table 3 yielded a sample size of 9.36. The smallest whole integer sample size for each result that meets or exceeds the constraint is 10. Here again we see that the L-H method performs exceptionally well even with discrete distributions that have small population sizes as we see that the boundaries converge within the neighborhood containing the optimal solution.

Another observation to be pointed out is that there is a broad range of values that the boundaries can take on while keeping the integer value of the sample size the same. As the size of the neighborhood expands, the range of boundary values extends as well. It should also be pointed out that even though the range of b_1 values for a given neighborhood is smaller than the range of values for b_2, there are far more sampling units in the range of b_1 than b_2 because of the skewed distribution.

5. SUMMARY

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 method 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 following stopping rules are recommended. Stop processing when:

1) the difference between the new upper boundary and the previous iteration's upper boundary is less than one. The whole number, one, is used in our case since payroll values are only available to us in whole number values and any shifting of boundaries of a value less than one does not affect any companies;
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 a small arbitrary value. We recommend a number less than one since sample sizes are usually rounded up and any fractional improvement on the sample size is negligible. One should be careful when choosing this value since it is possible that the sample size reduction rate may increase from iteration to iteration because the slope of the surface changes;
4) the program goes into the 30th iteration. Of course, this is an arbitrary value and may depend on the number of times (industries) one has to apply the L-H method.

Another note is that small population sizes may cause convergence of the boundaries to a point suboptimal, as shown in the examples. Graphs of the sample size surface show a rough surface for small populations and a smooth surface for large populations. It is this rough surface due to the discrete nature of the small population that contribute, in part, to where the L-H method converges.

Another point in conclusion, in our application, the Dalenius-Hodges method assumes that all resulting strata will be sampled. The L-H method is written to construct an analytical take-all substratum. Therefore, the top stratum developed by the Dalenius-Hodges method, 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 Dalenius-Hodges method 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.

ACKNOWLEDGEMENTS

The authors are grateful to Michel Hidiroglou for useful comments and discussion. We also thank Carol (Veum) Caldwell, Easley Hoy, the referees from *Survey Methodology*, and the Research and Methodology Branch managers of the Manufacturing and Construction Division for helpful comments during review.

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 Section on Survey Research Methods, American Statistical Association*, 214-219.

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

HESS, I., SETHI, V.K., and BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

LAVALLÉE, P., and HIDIROGLOU, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhȳa*, Series C, 41, 92-96.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.