Lavallee and Hidiroglou's (1988) developed an algorithm that minimizes the overall size of a random stratified sample by optimally choosing the boundary points of the strata. The boundary points are chosen for a given coefficient of variation for the estimator and a specific power allocation scheme. The current work presents a Fortran Program that perform the above optimal allocation. The computer algorithm is demonstrated with an application to real data.

1. INTRODUCTION

Lavallee and Hidiroglou (1988) developed an algorithm that minimizes the overall size of a random stratified sample by optimally choosing the boundary points of the strata. The boundary points are chosen for a given coefficient of variation for the estimator and a specific power allocation scheme. This allocation scheme enables estimation of the coefficients of variation among the strata that to be similar. A disadvantage of the Neyman allocation is that if we need to estimate each strata, the associated coefficients of variation may be quite different among the strata. However, an allocation that achieves equal coefficients of variation among the strata may require a much larger sample size. The approach developed by Lavallee and Hidiroglou offers a compromise between the Neyman allocation and the attainment of an equal coefficient of variation for each strata. It can be treated as a generalization of the Neyman allocation.

In section II we present the rule of optimum stratification developed by Lavallee and Hidiroglou (1988). In section III we discuss some possible computational problems that may arise when using this algorithm. In section IV we present the Fortran Program to perform the optimum stratification. In section V we give some simulated data examples as well as examples using real data to compare our results with the cumulative square root method.

II. THE ALGORITHM

Let us consider a finite ordered population of \( N \) units: \( y_1, y_2, \ldots, y_N \), with \( y_1 < y_{(i+1)} \) for \( i=1,2,\ldots,(N-1) \). This population is to be stratified into \( L \) strata. The sampling scheme calls for \( n_h \) units to be drawn from each corresponding stratum of size \( N_h \), \( h=1,2,\ldots,L \), without replacement. Cochran (1977, p. 91) defines the usual estimator of the population mean \( \bar{y} \). He gives the estimator of the population variance as:

\[
\text{Var}(\bar{y}) = \sum_{h=1}^{L} \frac{W_h}{N} \left( 1 - \frac{n_h}{N_h} \right) \sigma_h^2
\]

where

\[
W_h = \frac{N_h}{N}, \quad f_h = \frac{n_h}{N_h},
\]

and it is assumed that the desired level of precision for the estimated mean is specified by the coefficient of variation \( c \), and that the proportion of sampled units to be allocated to each of the \( L \) strata is \( a_h, (h=1,2,\ldots,L) \) where

\[
a_h = \frac{\left( W_h \mu_h \right)^2}{n_h} \quad \text{(2.3)}
\]

and substitute (2.3), (2.2) into (2.1) then simplified, this results in the overall sample size

\[
n = \frac{N f^2 \sum_{h=1}^{L} \left( W_h \sigma_h^2 \right) \left( W_h \mu_h \right)^2}{\mu^2} \left( \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{\mu^2} \right) \quad \text{(2.4)}
\]

where

\[
W_h = \int y f(y) dy, \quad \mu_h = \int y f(y) W_h/ \int f(y) dy
\]

\[
\sigma_h^2 = \int (y - \mu_h)^2 f(y) dy/W_h
\]

\[
h=1,2,\ldots,L.
\]

To simplify expression, we may also let

\[
F = N c^2 \mu^2, \quad \sigma_h^2, \quad A = \frac{1}{\mu^2} \left( W_h \mu_h \right)^2,
\]

\[
B = \frac{1}{\mu^2} \left( W_h \sigma_h^2 \right) \left( W_h \mu_h \right)^2
\]

Also, let

\[
K_h = \frac{E p \left( W_h \mu_h \right)^2 - 1}{\mu^2} A p \left( W_h \sigma_h^2 \right) \left( W_h \mu_h \right)^2 P - 1
\]

\[
T_h = A W_h \left( W_h \mu_h \right)^2
\]

Taking the partial derivative of \( n \), in equation (2.4), with respect to \( b(h) \) and equating it to zero, Lavallee and Hidiroglou (1988) obtained the following quadratic form

\[
\left( FT_h - FT(h+1) \right) b(h) + (FK(h)) 2 \mu FT_h - FK(h+1)
\]

\[
- 2 \mu A B + 2 \mu AB + 2 \mu (h+1) b(h) +
\]

\[
- 2 \mu A B + 2 \mu AB + 2 \mu (h+1) b(h)
\]

\[
F(\sigma_h^2 + \mu_h) T_h - F(\sigma_h^2(h+1) + \mu(h+1)) T(h+1)
\]

\[
- AB(\mu_h - \mu(h+1)) = 0
\]

(2.8)
Letting the coefficient of $b^2(h)$ be labeled as $\alpha_h$, the coefficient of $b(h)$ as $B(h)$, and the remaining terms as $\gamma_h$ then equation (2.8) can be written as

$$\alpha_h b^2(h) + B(h) b(h) + \gamma_h = 0$$  \hspace{1cm} (2.9)

Since the terms $\alpha_h$, $B(h)$, and $\gamma_h$ are themselves functions of $b(1), b(2), \ldots, b(L-1)$, hence they developed the following iterative process:

Step 1: Start with arbitrary boundaries.
Step 2: Compute the sample proportion weight $W_h$, the sample mean $\mu_h$, and the sample variance estimate $\sigma^2_h$ from equation (2.5) based on the boundaries defined in step 1.
Step 3: Replace the old set of boundaries by the new set of boundaries $b(1), b(2), \ldots, b(L-1)$, these $b(i)$ are roots of the quadratic equation, i.e.

$$b(h) = \frac{-B(h) \pm \sqrt{B(h)^2 - 4 \alpha_h \gamma_h}}{2 \alpha_h}$$  \hspace{1cm} (2.10)

Step 4: Repeat step 2 and step 3 until two consecutive sets of boundaries are either identical or differ by a negligible amount.

In an actual computation the parameter defined in equation (2.5), can be replaced by sample estimates.

### III. SOME DISCUSSIONS TO THE ALGORITHM

It should not be too surprising if the above suggested algorithm fails, as the algorithm has shown the following possible problems.

**Problem 1:** It is possible that the determinant given in equation (2.10) is negative. Usually this event happens when the given coefficient of variation is really small such as 0.01. If this is the case, the program given in the section iv may naturally terminate. This problem is avoided by letting the determinant equal zero. If this is the case, then some of the strata will have a zero count. This problem is avoided by using the natural condition of the root of equation, i.e. the roots $b(h)$ must lie between $\mu(h)$ and $\mu(h+1)$.

The suggested algorithm can help us to find the optimum boundaries such that the required sample sizes will be less than cumulative square root method. (see examples in section v) Another draw back of the method is that it may get larger required sample if the last strata has only one unit or very few units. Then it may be a good idea to use Hidiroglou's suggestion to reserve the last strata as taken all stratum.

### IV. THE PROGRAM

This Fortran Program can iterate the optimum boundaries for the given data files. We only assume that data file has sorted in increasing order and number of strata not exceed ten. usually, sorting the data file can be accomplished by some utility program.

A. Input data includes:

1. $C$: coefficients of variation,
2. $P$: power of allocation scheme,
3. $L$: number of strata,
4. $N$: population sizes,
5. In each iteration, it will produce the following output data information:
   1. number of iteration,
   2. the old bound of the strata,
   3. the number of counts in each strata,
   4. the weight in each strata,
   5. the strata mean values,
   6. the strata variances,
   7. the cv values in each strata,
   8. A,B,F values defined in equations (2.6)
   9. $K(h), T(h)$ values defined in equations (2.7),
   10. the coefficients of equation (2.9),
   11. the created new bounds.

### C. Possible error information includes:

1. the H strata has zero counts or use the zero as denominator: it needs to justify the bounds to some arbitrary bounds.
2. negative values in the square roots: it needs to relax the c value to some larger values.
3. how to define the initial bound: use the maximum data value subtract the minimum data value divided by the number of strata then evenly divided the range by subinterval of equal length. If this method fails it should modify the initial bounds in some random fashion.
4. The stopping rule:
   - two consecutive sets of boundaries are either identical or differ by a negligible amount.

```fortran
implicit real*8(a,h,o,z)
dimension bh(10),sum(10),sumsq(10),ybar(10),
+yvar(10),wh(10),rh(10),ch(10),alpha(10),
+beta(10),gama(10),count(10),dbh(10),cv(10),a(10)
+,(y(2000)
data c,p,sumtot,sumsq,error/0.1,0.5,0.0,0.0,
+0.001/
data 1,n,k,kk,num/3,440,1,10,1/
data fn1,fn2,fn3/0.0,0.0,0.0,0/
+open(10,file='udl2:sample.dat',status='old')
1000 read(10,100,end=99)(y(i),i=k,kk)
99 do 1020 i=1,n
do 1030 i=1,n
+sumtot=sumtot+y(i)
1020 continue
+ybar=sumtot/float(n)
+write(6,120)(y(i),i=k,kk)
+kk=k+10
+k=kk go to 1000
1030 continue
+sterr=sqrt(sumsq/(n-1))
+alcv=sterr/ybar
+write(6,130)(ybar,alcv)
130 continue
```

621
\[ \text{if}(\text{dbh}(i) \gt \text{ybar}(i+1)) \text{dbh}(i) = \text{ybar}(i+1) \\
\text{if}(\text{dbh}(i) \lt \text{ybar}(i)) \text{dbh}(i) = \text{ybar}(i) \]

write(6,440)\text{dbh}(i)
V. EXAMPLES AND APPLICATION

To illustrate results from previous section, we use data, computer generated, gross receipts of corporations, adjusted gross income of individuals in 1985. We compare the proposed method with the cumulative square root method in terms of the required sample sizes. The computed skewness for these populations is 1.464 (for population 1), 1.563 (for population 2), 1.4283 (for population 3).

Example 1. Using a computer to generate 50 Chi-Square distribution with one degree of freedom. We want to cut into two strata or three strata.

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>1.0</td>
<td>1</td>
<td>39</td>
<td>2</td>
<td>1.14</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>5</td>
<td>0.34</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 7

use optimum method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>1.0</td>
<td>1</td>
<td>34</td>
<td>1</td>
<td>1.06</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>4</td>
<td>0.43</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 5

use cum \( \frac{1}{2} \) method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>1</td>
<td>39</td>
<td>13</td>
<td>1.14</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>11</td>
<td>0.34</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 24

use optimum method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>1</td>
<td>32</td>
<td>4</td>
<td>0.98</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>11</td>
<td>0.48</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 15

use cum \( \frac{1}{2} \) method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>1</td>
<td>39</td>
<td>11</td>
<td>1.14</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>11</td>
<td>0.34</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 22

use optimum method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.25</td>
<td>1</td>
<td>34</td>
<td>7</td>
<td>1.06</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>9</td>
<td>0.43</td>
<td>5.204</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 16

* means: allocation is required to satisfy coefficient of variation.

Example 2. Using 400 values of gross receipts of corporations in the United States. The population is divided into 3 or 4 strata. (Notice that the original data has been divided by billion.)

use cum \( \frac{1}{2} \) method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>1.0</td>
<td>1</td>
<td>291</td>
<td>2</td>
<td>1.24</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>4</td>
<td>0.41</td>
<td>9.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>9</td>
<td>1.24</td>
<td>197.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 10

use optimum method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>1</td>
<td>291</td>
<td>3</td>
<td>1.24</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>98</td>
<td>5</td>
<td>0.68</td>
<td>24.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>4</td>
<td>0.90</td>
<td>197.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 10

use cum \( \frac{1}{2} \) method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>1.0</td>
<td>1</td>
<td>261</td>
<td>1</td>
<td>1.15</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>3</td>
<td>0.33</td>
<td>4.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>5</td>
<td>0.30</td>
<td>13.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>11</td>
<td>1.10</td>
<td>197.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 20

use optimum method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>1.0</td>
<td>1</td>
<td>260</td>
<td>1</td>
<td>1.15</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>3</td>
<td>0.51</td>
<td>8.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>4</td>
<td>0.42</td>
<td>39.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td>0.77</td>
<td>197.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 12

use cum \( \frac{1}{2} \) method

<table>
<thead>
<tr>
<th>C</th>
<th>P</th>
<th>Strata</th>
<th>( N_h )</th>
<th>( n_h )</th>
<th>C</th>
<th>( b(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.0</td>
<td>1</td>
<td>261</td>
<td>2</td>
<td>1.15</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>6</td>
<td>0.33</td>
<td>4.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>43</td>
<td>10</td>
<td>0.30</td>
<td>13.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>22</td>
<td>1.10</td>
<td>197.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* total: 40
Example 3: The population size of 990 adjusted individual gross income for 1985 was used. We want to cut the population into three strata for different combination of c, p and l.

Example 3: The population size of 990 adjusted individual gross income for 1985 was used. We want to cut the population into three strata for different combination of c, p and l.

VI. CONCLUDING REMARKS

From examples 1 to 3 we can see that using optimum boundaries requires less sample than using the cumulative square root method. The discrepancy of the sample sizes will depend on the distribution of the data, given coefficient of variation, c, the power of allocation, p, and the number of strata l. For both methods, the total required sample sizes will heavily depend on the given c value while less on p value and l value. If the computed determinant is negative or final sample sizes larger than the population size N it means that the current suggested algorithm cannot stratify the given population into the specified precision c value. It is clear that the cumulative square root method can make up this difficulty since in the procedure of stratification it does not depend on the c-value.

ACKNOWLEDGEMENT

I am grateful for Pierre Lavallee and Michel A. Hidiroglou for their useful comments.

REFERENCES

