EVALUATION OF MODEL-ASSISTED PROCEDURES FOR STRATIFYING SKEWED POPULATIONS USING AUXILIARY DATA

Elizabeth M. Sweet and Richard S. Sigman

Elizabeth M. Sweet, U.S. Bureau of the Census, Washington DC 20233

The views expressed are of the authors and do not reflect those of the Census Bureau

KEY WORDS: Establishment Surveys

1. Introduction

In real world stratification applications, two scenarios are common. Often the survey variable of interest is not available prior to conducting the survey, leaving the statistician to stratify with an auxiliary variable. And secondly, many times data are skewed, suggesting that a certainty stratum is necessary. Singh (1971) suggested a modification to the Dalenius and Hodges (D&H) (1959) cum-$\sqrt{f}$ stratification rule when auxiliary data are used. Lavallée and Hidiroglou (L&H) (1988) proposed a method for determining stratification boundaries with a certainty stratum. This paper separately documents the effect of these two approaches. In addition, a modification to the L&H procedure when stratifying with auxiliary data is proposed and examined. As a result of this research, a generalized stratification program in SAS was developed. The code can be obtained from the authors.

Section 2 provides additional details on (a) Singh's procedure, (b) the L&H algorithm, and (c) our modification of the L&H algorithm. Topics (a) and (c) postulate a stochastic model relating the survey variable to the auxiliary variable. In Section 2 we also discuss the estimation of parameters for such a model and the use of models in allocation. Sections 3 and 4 discuss the methodology and results of a simulation study that we performed involving skewed populations. Section 5 presents our conclusions. The simulation study allowed us to address the following questions:

How does the method of forming intervals for the auxiliary data affect the cum-$\sqrt{f}$ rule procedure?

How does the method of determining stratum boundaries affect the total sample size resulting from a fixed-precision Neyman allocation?

How does the method of determining initial stratum boundaries affect the L&H algorithm and modified L&H algorithm?

What is the effect of model-assisted allocation on the expansion-estimator variance under a fixed-sample-size Neyman allocation?

What is the effect of the number of strata on total sample size and expansion-estimator variance?

2. Background

2.1. Cum-$\sqrt{f}$ rule

Cochran (1977) provides a detailed description of the cum-$\sqrt{f}$ rule of Dalenius and Hodges (1959), which constructs strata from the frequency distribution of the stratification variable. Cochran (1961) and Hess, Sethi, and Balakrishnan (1966) compare the performance of the cum-$\sqrt{f}$ rule with other methods of constructing strata. The latter authors report that "[t]here is some evidence that boundary construction by cum-$\sqrt{f}$ rule is sensitive to discontinuities or zero classes in the frequency distribution." They recommend introducing "class intervals of unequal width whenever classes with zero frequencies occur, the width of the class being such as to bridge the gap from one non-zero class to succeeding non-zero class." Hess, Sethi, and Balakrishnan find that this use of unequal intervals in the frequency distribution, along with the corresponding adjustment to the cum-$\sqrt{f}$ rule (see Cochran, 1977) , results in a lower expansion-estimator variance than when equal intervals are used.

Cochran (1977) discusses the effect of the number of cum-$\sqrt{f}$ strata on the expansion-estimator variance when the stratification variable is used for Neyman allocation. He concludes that more than six strata produce very little additional variance reduction unless the correlation between the stratification variable and the survey variable exceeds 0.95. This conclusion, however, follows from the assumption that a regression model with homoscedastic errors relates the survey variable to the stratification variable. Hess, Sethi, Balakrishnan (1966) describe a population in which the correlation between the survey variable and the stratification variable is 0.91 and Cochran's conclusion appears not to hold. They attribute this to the presence of heteroscedasticity of the errors about the regression line.

2.2. Singh's procedure

Singh's procedure assumes there exists a model $y_i = \lambda(x_i) + e_i$, $E_m(e|x)=0$, $V_m(e|x)= \phi(x)$, $i=1,\ldots,N$, where E_m and V_m denote the expectation and variance, respectively, under the assumed model. Let $V_d[\hat{M}(y)]$ denote the sample-design variance of the expansion estimator for $M(y) = \frac{1}{N}\sum_{i=1}^{N} y_i$. Singh's procedure finds the stratum boundaries that minimizes $E_m V_d[\hat{M}(y)]$ under a fixed-cost optimal allocation. It is a cum-$\sqrt[3]{p(x)}$ rule, where $p(x)=q(x)f(x)$, with $f(x)$ being the frequency of units in the particular interval and $q(x) = \{\phi^2(x)\, C'(x) + C^2(x)\, [\phi'(x)]^2 + 4\, C^2(x)\, \phi(x)\, [\lambda'(x)]^2 - 2\phi(x)\, C(x)\phi'(x)\, C'(x)\} \,/\, [\phi(x)C(x)]^{3/2}$, where $C(x)$ is the survey-data collection "cost" for a unit having auxiliary data equal to x. If $C(x)$ is constant, $q(x)$ simplifies (disregarding multiplicative constant terms) to $q(x) = \{[\phi'(x)]^2 + 4\phi(x)[\lambda'(x)]^2\} \,/\, [\phi(x)]^{3/2}$. In the simulation study that we describe below., we used $\lambda(x)=\alpha+\beta x$ and $\phi(x)=\delta^2 x^{2g}$.

2.3. L&H algorithm

Sigman and Monsour (1995) review methods for constructing strata for economic surveys. One of these methods is the algorithm by Lavallée and Hidiroglou (1988), which constructs a certainty stratum and determines stratum boundaries and stratum sample sizes for a power-allocated stratified sample of non-certainty sample units. Hidiroglou and Srinath (1993) present a more general form of the algorithm, which by assigning different values to operating parameters yields a power allocation, a Neyman allocation, or a combination of these allocations. We confine our discussion to the use of the L&H algorithm with Neyman allocation.

Assume that one wants to construct L strata from stratification data x_i, $i=1,2,\ldots,N$, by using the x_i to determine stratum boundaries b_h, $h=0,1,\ldots,L$, such that $\min\{x_i\} = b_0 < b_1 < \ldots < b_{L-1} < b_L = max\ \{x_i\}$ and b_{L-1} and b_L are the boundaries of the certainty stratum. Let c denote the fixed coefficient of variation for estimating. From equation 2.2 of Hidiroglou and Srinath (1993) it follows that the needed sample size under fixed-precision Neyman allocation is $n = NW_L + NA^2/F$ where W_h = proportion of the population belonging to stratum h

$$A = \sum_{h=1}^{L-1} W_h S_h^{(y)} \tag{1}$$

$$F = N[cM(y)]^2 + \sum_{h=1}^{L-1} W_h [S_h^{(y)}]^2 \tag{2}$$

and $S_h^{(y)}$ = the standard deviation of the y_i belonging to stratum h.

When constructing strata, however, one has only x_i, $i=1,\ldots,N$, and not the y_i, $i=1,\ldots,N$. Thus the L&H algorithm replaces M(y) and $S_h^{(y)}$ with M(x) and $S_h^{(x)}$, respectively, and then finds the b_h that minimize n by setting $\partial n/\partial b_h=0$ for $h=1,\ldots,L$-1. The solution which produces the minimizing b_h involves calculating the first and second stratum moments of the x_i and does not involve the frequency distribution of the x_i as the Singh and D&H methods do.

A number of papers describe applications of the L&H algorithm. Detlefsen and Veum (1991), who investigated the L&H algorithm for 3, 6, 9, and 12 strata, observed that the algorithm's convergence was slow (often 50 to 100 iterations) or non-existent. They also found that different starting values of stratum boundaries for the same population resulted in different ending boundaries, and many times the boundaries differed substantially. Slanta and Krenzke (1994) carefully studied this latter problem and concluded that convergence of the algorithm should be determined on the basis of the sample size instead of the boundary values. They found that the boundary values can vary greatly in the neighborhood of the minimum sample size, whereas the sample size varies slightly.

2.4. Modified L&H algorithm

We developed a model-assisted version of the L&H algorithm, which we call modified L&H by assuming there exists the following model:

$$y_i = \alpha+\beta(x_i+\sqrt{d_0+d_1x_i+d_2x_i^2}e_i),\ \mathrm{E_m}(e|x)=0,\ \mathrm{V_m}(e|x)=1, \tag{3}$$

for i=1,2,...,N. (Hidiroglou (1994) discusses a similar modification to the L&H algorithm for $\mathrm{d}_1 = \mathrm{d}_2 = 0$.) Then, $z_i = (y_i-\alpha)/\beta = x_i + \sqrt{d_0+d_1x_i+d_2x_i^2}e_i$. (4)

The modified L&H algorithm replaces M(y), $[\mathrm{S_h}^{(y)}]^2$, and c in equations (1) and (2) with $\mathrm{E_m}$M(z), $\mathrm{E_m}[\mathrm{S_h}^{(z)}]^2$, and c', respectively, where

$$(c')^2 = E_m V_d(\hat{M}(z))/[E_m M(z)]^2 \tag{5}$$

and then finds the b_h that minimize n by setting $\partial n/\partial b_h=0$ for $h=1,\ldots,L$-1. From equation (4) it follows that $E_m[S_h^{(z)}]^2 = [S_h^{(x)}]^2+M_h(d_0+d_1x+d_2x^2)$, where $\mathrm{M_h}(\,)$ denotes the mean of values in stratum h, and $E_m M(z) = M(x)$. Since $z_i = (y_i-\alpha)/\beta$, it follows that $V_d(\hat{M}(z)) = V_d(\hat{M}(y))/\beta^2 = c^2(M(y))^2/\beta^2$. Substituting this into equation (5) and simplifying yields

$$(c')^2 = c^2(\alpha+\beta M(x))^2/[\beta M(x)]^2+c^2\beta^2V_m[M(x)]/[\beta M(x)]^2$$

But $V_m M(x)=0(N^{-1})$. Hence $c' \approx c(\alpha+\beta M(x))/[\beta M(x)]$. A straight-forward but tedious derivation yields that each b_h, h = 1, 2,..., L-1, is the solution of the quadratic equation $\alpha_h^* b_h^2 + \beta_h^* b_h + \gamma_h^* = 0$.

The coefficients α_h^*, β_h^*, and γ_h^* are in terms of $M_h=\mathrm{M_h}(x)$, $S_h^2=\mathrm{E_m}[S_h^{(z)}]^2$, A, F, d_0. d_1, and d_2, and for $h=1, 2, \ldots, L$-2 are

$$\alpha_h^* = \epsilon_h(1+d_2)$$

$$\beta_h^* = 2\left[A(M_h - M_{h+1}) - F\left(\frac{M_h}{S_h} - \frac{M_{h-1}}{S_h}\right)\right] + d_1\epsilon_h$$

$$\gamma_h^* = F\left(\frac{M_h^2+S_h^2}{S_h} - \frac{M_{h+1}^2 + S_{h+1}^2}{s_{h+1}}\right) - A\left(M_h^2 - M_{h+1}^2\right) + d_0\epsilon_h$$

where $\epsilon_h = F\left(1/S_h - 1/S_{h+1}\right)$, and for $h = L$-1,

$$\alpha_{L-1}^* = \epsilon_{L-1}(1 + d_2)$$

$$\beta_{L-1}^* = -2M_{L-1}\left(F/S_{L-1} - A\right) + d_1\epsilon_{L-1}$$

$$\gamma_{L-1}^* = F\left(\frac{M_{L-1}^2 + S_{L-1}^2}{S_{L-1}}\right) - AM_{L-1}^2 - F^2/A + d_0\ \epsilon_{L-1}$$

where $\epsilon_{L-1} = F/S_{L-1} - A$. Since M_h, S_h, A, and F depend on the b_h, it is necessary to solve for the b_h in an iterative fashion.

2.5 Model-assisted allocation

Let n_h denote the sample size in stratum h. If the $S_h^{(y)}$ are known, then the Neyman allocation $n_h \alpha\ W_h S_h^{(y)}$ can be calculated. A model-assisted Neyman allocation, as discussed by Dayal (1985), is $n_h\alpha \; W_h\sqrt{E_m\left[S_h^{(y)}\right]^2}$.

2.6. Parameter estimation

If g is known then the model $y_i = \alpha + \beta x_i + \delta x_i^g e_i$, can be transformed to $v_i = \alpha u_{1i} + \beta \, u_{2i} + \delta e_i$, where $v_i = y_i/x_i^g$, $u_{1i}=x_i^{-g}$ and $u_{2i} = x_i^{1-g}$. Then α and β can be estimated via ordinary least squares and δ^2 is estimated by the mean sum of squares for error. Brewer (1963), Harvey (1976), and Knaub (1993) discuss ways to estimate g. Knaub observes that for economic data g is frequently close to 0.5.

3. Methodology

For our simulation study, we generated two sets of skewed populations, which we refer to as study populations. The model $y_i = \alpha + \beta x_i + \delta x_i^g e_i$, was used to generate 10,000 observations, where log x is distributed N(0,1), e_i~N(0,1), α=7424, β=17.7, g=.5, δ=225. Using this data set and the model in equation (3), we solved for the parameters d_0=0, d_1=161.59, and d_2=0. This process was repeated 10 times, creating 10 populations, each containing a true survey variable of interest (y) and an auxiliary or stratification variable (x). Using these parameters the average correlation between x and y was high at .9979 with a standard deviation of the correlations of .0006. The process was repeated using the parameters, α=7424, β=17.7, g=.5, δ=2200, d_0=0, d_1=15448.95, and d_2=0. Ten additional populations were generated having an average correlation of .8932 with a standard deviation of the correlations of .0330. This process created ten highly-correlated study populations and ten moderately-correlated study populations.

In this document we compared four stratification procedures: D&H, Singh, L&H and our modified L&H procedure using Neyman allocation with a fixed coefficient of variation (CV) of 0.01. We also examined the effect of model-assisted allocation with a fixed-sample size. At minimum, we compared these procedures when assigning 5 or 10 strata. Paired t-tests were used to see if differences are significant at the 0.10 level. The results are documented for the 10 high-correlated study populations and the 10 moderately-correlated study populations.

To obtain estimated parameters needed for the Singh and L&H stratification methods and model-assisted allocation, additional populations, which we refer to as estimation populations, were generated following the same distribution as the study populations. The average correlation of the first set of 10 estimation populations was .9976 with a standard deviation of .0007. The average correlation of the second set of 10 estimation populations was .8922 with a standard deviation of .0308. Using the D&H cum √f(x) rule, 5 strata were created. Using Neyman allocation to meet a CV of 0.01, a sample was drawn from each population. To find the parameter estimates, we assumed the sample data fit the model: $y_i = \alpha + \beta x_i + \delta x_i^g e_i$, where e_i~N(0,1). Although we attempted to use the plots discussed in the Knaub paper to estimate g, the plots generated were flat, and Knaub didn't present specific criteria to use when plots were flat. Assuming g=0.5, we used the sample from each estimation population to estimate an α, β, δ^2 for each corresponding study population using the methods described in Section 2.6. The parameters d_0 and d_2 were set to 0 and $d_1 = \delta^2/\beta^2$.

Mean estimated parameters and their standard deviations are found in Tables 1a and 1b.

Table 1a: Parameters and Standard Deviations for Populations with mean correlation = .9979

Parameter	Actual	Mean of Est.	Std. Dev.
α	7424	7442.11	92.90
β	17.7	17.74	0.06
g	.5	.5	0
δ	225	220.3	9.88
d_0	0	0	0
d_1	161.59	154.65	14.26
d_2	0	0	0

Table 1b: Parameter and Standard Deviations for Populations with mean correlation = .8932

Parameter	Actual	Mean of Est.	Std. Dev.
α	7424	9322.12	315.36
β	17.5	30.47	0.75
g	.5	.5	0
δ	2200	1635.06	28.51
d_0	0	0	0
d_1	15448.95	2888.37	228.27
d_2	0	0	0

As discussed in the background section, the L&H methods use an iterative procedure. Due to time and resource limitations, we limited the number of iterations to 30. For those nonconverging runs, the boundaries at the 29th iteration were used to partition the observations into strata.

4. Results

4.1 Interval Class Creation

Prior to stratification, the data had to be partitioned into interval classes. Initially we partitioned each population into 500 equal sized intervals based on x. Because the data were skewed, often the first stratum contained only one interval class. This was not a problem when 5 strata were requested, but with a large number of strata, the situation occurred where the first stratum contained over half of the observations and the remaining strata had varying numbers of observations. As a result, there was uneven allocation across the strata. To remedy this, instead of creating interval classes of equal sizes on x, we created 500 equal sized interval classes on the natural log of x. When we stratified using the x data, we adjusted our procedures by the method proposed in Cochran (1977) for unequal interval lengths.

Table 2 shows the effect of interval classes on the D&H stratification with Neyman allocation to meet the 0.01 CV requirement on the 10 populations with average correlation of .9979. As the number of strata increase beyond 5, the total sample size using interval classes created on the log of x is significantly less than the total sample size using interval classes created on x. The stratum sample sizes were more equally distributed using the intervals created using the log of x. Because of this trend, all analysis presented in this paper used the interval classes created on the log of x.

Table 2: Average Total Sample Size using strata created using 500 equal sized intervals on the log of x vs. Average Total Sample Size from strata created using 500 equal intervals on x.

Strata	Interval Class log (x)	Interval Class with x
5	811.01	754.73
10	440.91	533.05
15	350.37	485.42

Some preliminary work was done to find a "reasonable" number of intervals to use. When less than 500 intervals were used, often, because the data were skewed, the first stratum contained only one interval, even when we used log(x) to partition the data. Using 1,000 intervals produced similar results to using 500 intervals, so we decided to use 500 intervals.

4.2 Comparison of D&H, Singh, L&H and ML&H Stratification Procedures with fixed-precision Neyman Allocation

Table 3 documents the effect of the D&H and Singh stratification methods on our populations for Neyman allocation. Table 3 also contains results from Neyman allocation using the L&H procedure and our modified L&H procedure. All allocations met the 0.01 CV requirement. Estimates for the actual parameters are used for Singh, L&H and modified L&H (ML&H) methods. See Table 1 for actual parameters. Strata were created using 500 equal sized intervals on the log(x). Initial boundaries are needed for both L&H and modified L&H procedures. To determine initial boundaries, we selected the method that proved best for each of the populations when comparing the D&H and Singh methods. For the populations with average correlation of .9979, initial boundaries were found by the D&H stratification method. For the populations with average correlation of .8932, initial boundaries were found by the Singh stratification method.

Table 3: Average Total Sample Sizes for four Stratification Procedures

Population with average correlation of .9979				
Strata	D&H	Singh	L&H	ML&H
5	811.01	1093.27	629.34	629.78
10	440.91	519.75	413.60	419.73
Population with average correlationof .8932				
Strata	D&H	Singh	L&H	ML&H
5	2801.71	2615.91	2534.75	2501.75
10	2370.42	2286.22	2314.66	2290.17

For the highly-correlated populations with average correlation of .9979, the total sample size for Neyman allocation using D&H cum $\sqrt{f(x)}$ stratification method was significantly smaller than the total sample size using the Singh cum $\sqrt[3]{p(x)}$ stratification boundaries. This difference between D&H and Singh was significant when both 5 strata and 10 strata were created. However, the opposite occurred for the moderately-correlated populations. When the correlation was .8932, significantly smaller total sample size occurred when the Singh method of stratification was used compared to the D&H stratification. This, too, was significant for both 5 and 10 strata.

For the highly-correlated populations, there was no significant difference in total sample size between L&H and modified L&H when either 5 or 10 strata were requested. For the moderately-correlated populations there was no significant difference in total sample size between L&H and modified L&H when 5 strata were requested, but there was a significant difference when 10 strata were requested. When 10 strata were selected, significantly smaller total sample size occurred when the modified L&H method of stratification was used.

These L&H procedures were compared to the D&H and Singh method. Since the required CV was small, all observations in the last stratum were selected when using both the D&H and Singh methods, hence we could compare across all four methods. The D&H and Singh programs both ran much faster than either of the L&H iterative programs. Savings in total sample size would warrant using L&H or modified L&H over D&H or Singh stratification. In the highly-correlated populations the total sample size from Neyman allocation with L&H and modified L&H stratification boundaries was significantly smaller than the total sample size with both D&H and Singh stratification boundaries for both 5 and 10 strata. In the moderately-correlated populations the total sample size from Neyman allocation for L&H and modified L&H was significantly smaller than both D&H and Singh for 5 strata. But for 10 strata, the total sample size from Neyman allocation with Singh boundaries was significantly smaller than with L&H boundaries, and there was no significant difference in total sample size between Singh and modified L&H boundaries.

4.3 Effect of Initial Boundaries on L&H and ML&H

To test the effect of initial boundaries on the L&H procedures, we compared the allocation needed with Singh initial bounds and either L&H or ML&H procedure to the corresponding allocation when D&H initial bounds were used instead. We did this for the moderately-correlated populations. Table 4 provides results when the boundaries differ.

Table 4: Effect of Initial Boundaries between L&H and ML&H on Total Sample Sizes from Neyman Allocation

Initial BoundarieS: D&H

Strata	L&H	ML&H
5	2545.78	2495.76
10	2358.69	2294.03

Initial Boundaries: Singh

Strata	L&H	ML&H
5	2534.75	2501.75
10	2314.66	2290.17

Unlike when the Singh boundaries were used, when D&H initial boundaries are used, there is significantly smaller total sample size with the modified L&H (ML&H) for both 5 and 10 strata compared to the total sample size with L&H stratification. The total sample size with D&H initial bounds and modified L&H stratification is not significantly different than the total sample size using Singh initial bounds and modified L&H stratification. The total sample size with D&H initial bounds and L&H stratification is significantly larger than the total sample size using Singh initial bounds and L&H stratification. This seems to imply that for the moderately-correlated populations the modified L&H procedure can compensate for non-optimum initial bounds.

4.4 Effect of Model-Assisted Allocation on the Expansion-Estimator Variance

Table 5 provides average CVs when model-assisted allocation with estimated parameters is used with the four different stratification procedures: D&H, Singh, L&H and modified L&H. The sample size was fixed for each of the two sets of populations. For the highly-correlated populations we fixed the sample at 811. This sample size was the average total sample size from Neyman allocation for the D&H procedure with 5 strata. For the moderately-correlated population we fixed the sample at 2802. This sample size was the average total sample size from Neyman allocation for the D&H procedure with 5 strata. As in section 4.3, the L&H and modified L&H stratification procedures use initial boundaries of D&H or Singh. For the highly-correlated populations initial boundaries were found by the D&H stratification method. For the moderately-correlated populations initial boundaries were found by the Singh stratification method.

Table 5: Average CV using the Model-Assisted Allocation on D&H, Singh, L&H and ML&H Stratification Methods

Population with correlation of .9979 and sample size of 811

Strata	D&H	Singh	L&H	ML&H
5	.0100011	.0124100	.0080977	.0082000
10	.0066900	.0074957	.0065214	.0065789

Population with correlation of .8932 and sample size of 2802

Strata	D&H	Singh	L&H	ML&H
5	.0100267	.0098264	.0091849	.0084718
10	.0085798	.0084599	.0085695	.0081913

For the highly-correlated populations D&H stratification provided an estimate with significantly lower variance than Singh stratification. Modified L&H stratification provided an estimate with significantly lower variance than both D&H and Singh stratification. And L&H stratification provided an estimate with the lowest variance of all the stratification methods. The same results apply when both 5 and 10 strata were requested. But for the moderately-correlated populations, slightly different results were found. When 5 strata were requested, the Singh stratification procedure provided an estimate with significantly lower variance than the D&H stratification. The L&H procedure provided an estimate with significantly lower variance than the D&H and Singh stratification. And the modified L&H procedure provided an estimate with the lowest variance of all the stratification methods. However, when the number of strata increased to 10, the L&H procedure provided no different results than the D&H procedure. The Singh procedure provided a significantly lower variance than the D&H and L&H procedure. And the modified L&H procedure would provided an estimate with the lowest variance.

4.5 Effect of the Number of Strata on Total Sample Size

As stated in the background section, Cochran stated that beyond six strata there is little reduction in variance unless the correlation is greater than .95. In the analysis presented in this document, we found continued decrease in sample size beyond six strata, even with the population with average correlation of .8932. Cochran assumed the model $y_i = \alpha + \beta x_i + \delta x_i^g e_i$, e_i ~N(0,1), and g=0. This is the model we used to generate our study populations except that g was greater than 0. Hess, Sethi, and Balakrishnan (1966), noted that with a population with correlation of .91, heteroscedasticity was a probable reason for Cochran's conclusion not to hold. A "g" greater than 0 causes heteroscedasticity about the regression line of x and y.

To test the heteroscedasticity theory, we generated 10 new populations. Except for δ=90,000 and g=0, the same parameters were used as in our populations with average correlation of .8932. Table 6 provides the total sample size from Neyman allocation for D&H stratification for the two sets of populations. The allocation meets a 0.01 CV for 5, 10, and 15 strata using the D&H stratification method. The populations with g=.5 have an average correlation of .8932. The populations with g=0 have an average correlation of .9013 between the x and y variables.

Table 6: Average Total Sample Sizes from Neyman Allocation using D&H stratification with 500 equal sized intervals on the log (x)

Strata	Populations where g=.5	Populations were g=0
5	2801.71	3110.20
10	2370.42	3063.30
15	2273.30	3047.57

The average percent decrease in total sample size when increasing the number of strata from 5 to 10 was 18% for the population with g=0.5 and was 1.5% for the population with g=0. The average percent decrease in total sample size when increasing the number of strata from 10 to 15 was 1.5% for the population with g=0.5 and only 0.5% for the population with g=0.

5. Conclusions

- The method of forming interval classes effects the ultimate allocation. For skewed data, we found intervals based on the log of x improved our stratification.
- For populations with high correlation between the survey variable of interest and the stratification variable, the Singh and modified L&H stratification procedures do not produce more optimum stratum boundaries than the D&H and L&H procedures. When the correlation was lower, in our example the correlation was close to .9, there was definite advantages to using the Singh and modified L&H procedures over the D&H and L&H procedures. In fact, with a low correlation and a CV that guaranteed a certainty stratum, the Singh procedure appeared to produce stratum boundaries that lead to significantly lower total sample size than those boundaries produced by the L&H procedure.
- With a population with correlation around .9, we found that the modified L&H procedure could compensate for poor initial boundaries.
- With a fixed-sample-size, model assisted allocation and a correlation close to 1, for 5 strata, the CV associated with either the L&H or modified L&H was approximately 20% lower than the CV associated with a D&H stratification. When the number of strata increase to 10, the savings was only 2%. The Singh procedure produced a higher CV than the D&H method under those circumstances. When the correlation between the survey variable and the stratification variable decreased to .9, for 5 strata, the CV associated with modified L&H was approximately 18% lower than the CV associated with a D&H. When the number of strata increase to 10, the savings was approximately 5%. The CV associated with the L&H procedure for 5 strata was approximately 9% lower than the CV associated with the D&H method, but when the number of strata increased to 10, there was no significant difference between the CVs.
- For populations with correlation of .9 and heteroscedasticity between the survey variable and stratification variable, we found the total sample size continues to decrease beyond Cochran's six-strata rule.

6. Acknowledgements

The authors would like to thank Thomas R. Krenzke, Philip Steel, James Fagan and Patrick Cantwell of the Census Bureau.

7. References

Brewer, K.R.W. (1963) "Ratio Estimation and Finite Populations: Some Results Deducible from the Assumptions of an Underlying Stochastic Process," Australian Journal of Statistics. Vol 5, pp. 93-105.

Cochran, W.G. (1961) "Comparison of Methods for Determining Stratum Boundaries," Bull. Int. Stat. Inst., Vol. 38. Part 2, pp. 345-358.

Cochran, W.G. (1977) Sampling Techniques, 3rd edition, N.Y.: Wiley.

Dalenius, T., and J. L. Hodges, (1959) "Minimum Variance Stratification," JASA, Vol. 54, pp. 88-101.

Dayal, Shambhu. (1985) "Allocation of Sample Using Values of Auxiliary Characteristic." Joint Statistical Planning and Inference. Vol.11, pp, 321-328.

Detlefsen, R.E. and C.S. Veum (1991), "Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census," Proceedings of the Survey Research Methods Section, ASA, pp. 214-219.

Harvey, A.C. (1976), "Estimating Regression Models with Multiplicative Heteroscedasticity," Econometrica. 44, pp. 461-465.

Hess, I., V. K. Sethi, and T.R. Balakrishnan. "Stratification: A Practical Investigation," JASA, 61. pp. 71-90.

Hidiroglou, M.A. (1994) "Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress." Proceedings of the Survey Research Methods Section, ASA, pp. 693-698.

Hidiroglou, M.A. and K.P. Srinath (1993) "Problems Associated with Designing Subannual Business Surveys," Journal of Business and Economic Statistics, Vol. 11, pp. 397-405.

Knaub, James R. Jr. (1993) "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling." Proceedings of the Internatiomal Conference on Establishment Surveys, Invited and Contributed Papers. ASA.

Lavallée, Pierre and Michel A. Hidiroglou. (1988) "On the Stratification of Skewed Populations." Survey Methodology. Vol. 14. No. 1. pp. 33-43.

Slanta, John G. and Thomas R. Krenzke. (1994) "Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditures Survey." Proceedings of the Survey Research Methods Section, ASA, pp. 693-698.

Singh, Ravindra. (1971). "Approximately Optimum Stratification on the Auxiliary Variable," JASA, Vol. 66, pp. 829-833.

Sigman, Richard and R. Monsour. (1994) "On Selecting Samples from Frames Based on Lists of Establishments," Forthcoming in Surveys of Businesses, Farms, and Institutions. Wiley.