STRATUM JUMPERS: CAN WE AVOID THEM?

Louis-Paul Rivest, Department of Mathematics and Statistics,
Université Laval, Ste-Foy, Québec, Canada G1K 7P4
(lpr@mat.ulaval.ca)

Key Words: Neymann allocation, Power allocation, Stratified random sampling.

Abstract:

This paper suggests stratification algorithms that take into account discrepancies between the stratification variable and the study variables. Two models are proposed for the occurrence of stratum jumpers. Under these models the stratification variable and the study variable are equal for most sampling units; important discrepancies between the two occur for a limited number of units. Then, Lavallée and Hidiroglou (1988) stratification algorithm is modified to incorporate the models for the stratum jumpers in the determination of the sample sizes and of the stratum boundaries. An example illustrates the performance of the new stratification algorithms.

1 Introduction

In business surveys, the size of the enterprise is an important stratification variable. The survey populations have skewed distributions and a good sampling design has one take-all stratum for big firms, where the units are all sampled, together with take-some strata for businesses of medium and small sizes. Typically the sampling fraction goes down with the size of the unit; this gives large sampling weights to small businesses.

Stratum jumpers occur when a small business experiences a rapid growth and becomes a large firm over a short period of time. When such a unit is sampled, its large sampling weight combined to its large size unduly inflate survey estimates. Lee (1995) reviews techniques, such as winsorization and weight reduction, that have been proposed to limit the impact of stratum jumpers on survey estimates. This paper investigates techniques for dealing with stratum jumpers when designing the survey. It proposes stratification algorithms that that takes into account the possible occurrence of stratum jumpers.

Stratification in situations where the survey variable differ from the stratification variable is considered briefly in Cochran (1977, chapter 5A) and in Dalenius and Gurney (1951), see also Hidiroglou and Srinath (1993) and Hidiroglou (1994). This paper constructs generalizations of Lavallée and Hidiroglou (1988) algorithm that incorporate explicitly differences between the survey and the stratification variable. Stratification algorithms accounting for the possible presence of stratum jumpers are presented. A numerical example illustrates that some gains in precision results form using the new stratification algorithms.

2 Modeling the Occurrence of Stratum Jumpers

In this section $\{x_i, i = 1, \ldots, N\}$ denotes the known stratification variable while $\{y_i, i = 1, \ldots, N\}$ represents the unknown study variable, and N is the population size. In an ideal situation $x_i = y_i$ for each i; many stratification algorithms (see Cochran, 1977, chapter 5A) rely on this assumption. To model the presence of stratum jumpers we want x_i and y_i to be very different for at least a few data points.

For the sequel, it is convenient to look at X and Y as continuous random variables with respective density $f(x)$ and $g(y)$ for x and y in R. The data $\{x_i, i = 1, \ldots, N\}$ is looked at as N independent realizations of the random variable X. Allocations rules and stratification for Y knowing only X was considered by Dalenius and Gurney (1951), see also Hidiroglou (1994). If $-\infty = b_0 < b_1 < b_2 < \ldots < b_L = \infty$ denote stratum boundaries, the stratification process uses $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h. Two models for stratum jumpers are given next with, for each one, the conditional means and variances of Y.

This research is funded by research grants form the Natural Sciences and Engineering Research Council of Canada and by the Fonds pour la formation des chercheurs et l'aide à la recherche du Québec

2.1 A Multiplicative Model

The first model considers that $Y = XZ$ where Z is a random variable distributed independently of X with the following probability function:

$$\Pr(Z = z) = \begin{cases} 1-\epsilon & \text{if } z = 1 \\ \epsilon & \text{if } z = M \end{cases},$$

where $0 < \epsilon < 1$ is a small positive probability that unit is a stratum jumper and $M > 1$ is a multiplicative inflation factor for stratum jumpers. The conditional mean for Y under this model is easily evaluated,

$$\begin{aligned} E(Y|b_h \geq X > b_{h-1}) &= \{1 + (M-1)\epsilon\} \\ &\quad E(X|b_h \geq X > b_{h-1}) \end{aligned}$$

while $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ is equal to

$$\begin{aligned} & \mathrm{Var}(E(Y|X)|b_h \geq X > b_{h-1}) \\ & \quad + E(\mathrm{Var}(Y|X)|b_h \geq X > b_{h-1}) \\ = & \{1 + (M-1)\epsilon\}^2 \mathrm{Var}(X|b_h \geq X > b_{h-1}) \\ & \quad + (M-1)^2 \epsilon (1-\epsilon) E(X^2|b_h \geq X > b_{h-1}). \end{aligned}$$

2.2 A Random Replacement Model

The multiplicative model depends on 2 parameters. The random replacement model, on the other hand, depends only on the probability ϵ that the value of Y for unit is equal to the X-value for a randomly selected unit in the population. In other words,

$$Y = \begin{cases} X \text{ with probability } 1-\epsilon \\ X_{new} \text{ with probability } \epsilon \end{cases},$$

where X_{new} represents a random variable with density $f(x)$ distributed independently of X. The conditional mean for Y under this model is given by

$$\begin{aligned} & E(Y|b_h \geq X > b_{h-1}) \\ & \quad = (1-\epsilon)E(X|b_h \geq X > b_{h-1}) + \epsilon E(X) \end{aligned}$$

while $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ is equal to

$$\begin{aligned} & (1-\epsilon)E(X^2|b_h \geq X > b_{h-1}) + \epsilon E(X^2) \\ & \quad - \{(1-\epsilon)E(X|b_h \geq X > b_{h-1}) + \epsilon E(X)\}^2. \end{aligned}$$

3 A Review of Stratified Random Sampling

Standard notations of stratified random sampling are:

$W_h = N_h/N$ is for $h = 1, \ldots, L$ the relative weight of stratum h, thus N_h is the size of stratum h and $N = \sum N_h$ is the total population size;

n_h is for $h = 1, \ldots, L$ the sample size in stratum h and $f_h = n_h/N_h$ is the sampling fraction;

$\bar{Y}_h$ and $\bar{y}_h$ are the population and sample mean of Y within stratum h;

S_{yh} is the population standard deviation of Y within stratum h.

The survey estimator for $\bar{Y}$ can be expressed as $\bar{y}_{st} = \sum W_h \bar{y}_h$; its variance is given by:

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2 \quad (1)$$

In business surveys all the big firms are sampled; this means that stratum L is a take-all stratum and $n_L = N_L$. For $h < L$, the sample size n_h can be expressed as $(n - N_L)a_h$ where n is the total sample size and a_h depend on the allocation rule. The two allocation rules that are considered here are

- The power allocation rule for which

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^{L-1} (W_k \bar{Y}_k)^p}, \; h = 1 \ldots, L-1 \quad (2)$$

  where p is a positive number in $(0, 1]$;

- Neymann allocation rule where

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}}, \; h = 1 \ldots, L-1. \quad (3)$$

Solving (1) for n leads to

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2 / a_h}{\mathrm{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2 / N} \quad (4)$$

The optimal strata boundaries are the values of $b_1, \ldots, b_{L-1}$ that minimize n subject to a requirement on the precision of $\bar{y}_{st}$ such as $\mathrm{Var}(\bar{y}_{st}) = \bar{Y}^2 c^2$ where c is the target coefficient of variation; $c = 1\%$ or 10% are common choices.

4 A Method for Constructing Stratification Algorithms

The aim of a stratification algorithm is to determine the optimal stratum boundaries and sample sizes for sampling Y using the known values of variable X for all the units in the population, $\{x_i; i = 1, \ldots, N\}$. A model, such as those given in Section 2, characterizes the relationship between X and Y. This section extends the stratification algorithm of Lavallée and

Hidiroglou (1988) to situations where X and Y differ.

It is convenient to consider an infinite population analogue of equation (4) for n. Since random variable X has a density $f(x)$, the first three moments of Y given that $b_{h-1} < X \leq b_h$ can be written in terms of

$$W_h = \int_{b_{h-1}}^{b_h} f(x)dx, \quad \phi_h = \int_{b_{h-1}}^{b_h} xf(x)dx,$$

and

$$\psi_h = \int_{b_{h-1}}^{b_h} x^2 f(x)dx.$$

For stratification purposes, it is convenient to rewrite (4) in terms of the conditional means and variances for Y,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \mathrm{Var}(Y|b_h \geq X > b_{h-1})/a_{h,X}}{\bar{Y}c^2 + \sum_{h=1}^{L-1} W_h \mathrm{Var}(Y|b_h \geq X > b_{h-1})/N}, \tag{5}$$

where $a_{h,X}$ denote the allocation rule written in terms of the known X. For instance, under power allocation,

$$a_{h,X} = \frac{\{W_h E(Y|b_h \geq X > b_{h-1})\}^p}{\sum_{k=1}^{L-1}\{W_k E(Y|b_k \geq X > b_{k-1})\}^p},$$

for $h = 1, \ldots, L-1$. For a given model for the relationship between Y and X, $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ and $E(Y|b_h \geq X > b_{h-1})$ can be written in terms of W_h, ϕ_h, and ψ_h. Thus, the partial derivative of n with respect to b_h can, for $h < L-1$, be evaluated using the chain rule,

$$\begin{aligned}\frac{\partial}{\partial b_h} n = & \frac{\partial n}{\partial W_h}\frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \phi_h}\frac{\partial \phi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h}\frac{\partial \psi_h}{\partial b_h} \\ & + \frac{\partial n}{\partial W_{h+1}}\frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \phi_{h+1}}\frac{\partial \phi_{h+1}}{\partial b_h} \\ & + \frac{\partial n}{\partial \psi_{h+1}}\frac{\partial \psi_{h+1}}{\partial b_h}.\end{aligned}$$

Observe that

$$\begin{aligned}\frac{\partial W_h}{\partial b_h} &= -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h), \\ \frac{\partial \phi_h}{\partial b_h} &= -\frac{\partial \phi_{h+1}}{\partial b_h} = b_h f(b_h), \\ \frac{\partial \psi_h}{\partial b_h} &= -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^2 f(b_h).\end{aligned}$$

This leads to the following result, for $h < L-1$,

$$\begin{aligned}\frac{\partial}{\partial b_h} n = f(b_h) \Bigg\{ & \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right) + \\ & \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right) b_h + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right) b_h^2 \Bigg\} \\ = & \; f(b_h)(\alpha_h + \beta_h b_h + \gamma_h b_h^2).\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial}{\partial b_{L-1}} n = f(b_{L-1}) \Bigg\{ & -N + \frac{\partial n}{\partial W_{L-1}} + \\ & \frac{\partial n}{\partial \phi_{L-1}} b_{L-1} + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^2 \Bigg\} \\ = & \; f(b_{L-1})(\alpha_{L-1} + \beta_{L-1} b_{L-1} + \gamma_{L-1} b_{L-1}^2).\end{aligned}$$

To solve $\partial n/\partial b_h = 0$, thereby hopefully finding the optimal stratum boundaries, one uses Sethi's (1963) algorithm. It considers that the partial derivatives are proportional to quadratic functions in b_h. The updated value for b_h is given by the largest root of the corresponding quadratic function. For any $h < L$, this gives

$$b_h^{new} = \frac{-\beta_h + (\beta_h^2 - 4\alpha_h\gamma_h)^{1/2}}{2\alpha_h}$$

The partial derivatives of n with respect to W_h, ϕ_h, and ψ_h depend on moments of order 0, 1, and 2 of x within stratum h. When implementing a particular algorithm, they are evaluated using the N x-values in the population. For instance,

$$\phi_h = \frac{1}{N} \sum_{i: b_{h-1} < x_i \leq b_h} x_i.$$

Applications of this general method are given in the next section.

5 Construction of Stratification Algorithms

Several illustrations of the general method of Section 4 for constructing stratification algorithms are now presented.

5.1 Lavallée and Hidiroglou Algorithm

Lavallée and Hidiroglou assumed that $Y = X$. Thus one has, in the W, ϕ, ψ notation defined at the beginning of Section 4, $E(Y|b_h \geq X > b_{h-1}) = \phi_h/W_h$ while $\mathrm{Var}(Y|b_h \geq X > b_{h-1}) = \psi_h/W_h - (\phi_h/W_h)^2$. Using this notation the power allocation rule is $a_{h,X} = \phi_h^p/\sum \phi_k^p$ for $h = 1, \ldots, L-1$. Formula (5) for the optimal n then becomes

$$n = NW_L + \frac{\sum_{h=1}^{L-1}(W_h\psi_h - \phi_h^2)/\phi_h^p \sum_{h=1}^{L-1}\phi_h^p}{\bar{X}^2c^2 + \sum_{h=1}^{L-1}(\psi_h - \phi_h^2/W_h)/N}.$$

The partial derivatives needed to implement Sethi's (1963) iterative algorithm are,

$$\frac{\partial n}{\partial W_h} = \frac{A\psi_h/\phi_h^p}{F} - \frac{AB(\phi_h/W_h)^2/N}{F^2}$$
$$\frac{\partial n}{\partial \phi_h} = \frac{A\{-p(W_h\psi_h - \phi_h^2)/\phi_h^{p+1} - 2/\phi_h^{p-1}\}}{F} + \frac{p\phi_h^{p-1}B}{F} + 2\frac{AB(\phi_h/W_h)/N}{F^2}$$
$$\frac{\partial n}{\partial \psi_h} = \frac{AW_h/\phi_h^p}{F} - \frac{AB/N}{F^2},$$

where $A = \sum^{L-1}\phi_h^p$, $B = \sum^{L-1}(W_h\psi_h - \phi_h^2)/\phi_h^p$, and $F = \bar{X}^2c^2 + \sum_{h=1}^{L-1}(\psi_h - \phi_h^2/W_h)/N$.

5.2 Stratification for Stratum Jumpers: The Multiplicative Model

For the multiplicative model of Section 2.1, one has

$$W_h^2\mathrm{Var}(Y|b_h \geq X > b_{h-1}) = \{1 + (M-1)\epsilon\}^2\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\},$$

where

$$C_{\epsilon,M} = 1 + (M-1)^2\epsilon(1-\epsilon)/\{1+(M-1)\epsilon\}^2.$$

Neymann allocation gives

$$a_{h,X} = \frac{\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}}{\sum_1^{L-1}\{W_kC_{\epsilon,M}\psi_k - \phi_k^2\}^{1/2}},$$

where ϕ_h and ψ_h involve, as in Section 4, the first 2 moments of X in stratum h. For this model, (5) gives

$$n = NW_L + \frac{\left[\sum_1^{L-1}\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}\right]^2}{\bar{X}^2c^2 + \sum_{h=1}^{L-1}[C_{\epsilon,M}\psi_h - \phi_h^2/W_h]/N}.$$

The partial derivatives needed to implement Sethi's (1963) iterative algorithm are,

$$\frac{\partial n}{\partial W_h} = \frac{AC_{\epsilon,M}\psi_h/\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}}{F} - \frac{A^2(\phi_h/W_h)^2/N}{F^2}$$
$$\frac{\partial n}{\partial \phi_h} = \frac{-2A\phi_h/\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}}{F} + \frac{2A^2\phi_h/(W_hN)}{F^2}$$
$$\frac{\partial n}{\partial \psi_h} = \frac{AC_{\epsilon,M}W_h/\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}}{F} - \frac{A^2C_{\epsilon,M}/N}{F^2},$$

where $A = \sum_1^{L-1}\{W_hC_{\epsilon,M}\psi_h - \phi_h^2\}^{1/2}$, and $F = \bar{X}^2c^2 + \sum_{h=1}^{L-1}(C_{\epsilon,M}\psi_h - \phi_h^2/W_h)/N$.

Under power allocation, one has $a_{h,X} = \phi_h^p/\sum\phi_k^p$ and the partial derivatives of n with respect to W_h and ϕ_h are given by the formula of Section 5.1, with ψ_h replaced by $C_{\epsilon,M}\psi_h$, while the partial derivative with respect to ψ_h is equal to that given in Section 5.1 multiplied by $C_{\epsilon,M}$.

5.3 Stratification for Stratum Jumpers: The Random Replacement Model

The conditional variance of Y, under the random replacement model of Section 2.2, satisfies

$$\begin{aligned} & W_h^2\mathrm{Var}(Y|b_h \geq X > b_{h-1}) \\ &= (1-\epsilon)W_h\psi_h + W_h^2\epsilon E(X^2) - \{(1-\epsilon)\phi_h + \epsilon W_h\bar{X}\}^2 \\ &= t_h \end{aligned}$$

Under Neymann allocation, formula (5) gives the following for this model

$$n = NW_L + \frac{\left[\sum_1^{L-1} t_h^{1/2}\right]^2}{\bar{X}^2c^2 + \sum_{h=1}^{L-1}[t_h/W_h]/N}.$$

The partial derivatives for Sethi's (1963) algorithm are written in terms of $t_h = W_h^2\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ and $m_h = W_hE(Y|b_h \geq X > b_{h-1})$.

$$\frac{\partial n}{\partial W_h} = \frac{A\{(1-\epsilon)\psi_h + 2W_h\epsilon E(X^2) - 2\epsilon\bar{X}m_h\}/t_h^{1/2}}{F} - \frac{A^2\{\epsilon E(X^2) + (1-\epsilon)^2(\phi_h/W_h)^2 + \epsilon^2\bar{X}^2\}/N}{F^2}$$

$$\frac{\partial n}{\partial \phi_h} = \frac{-2(1-\epsilon)Am_h/t_h^{1/2}}{F} + \frac{2A^2(1-\epsilon)m_h/(NW_h)}{F^2}$$
$$\frac{\partial n}{\partial \psi_h} = \frac{A(1-\epsilon)W_h/t_h^{1/2}}{F} - \frac{A^2(1-\epsilon)/N}{F^2},$$

where $A = \sum_1^{L-1}t_h^{1/2}$, and $F = \bar{X}^2c^2 + \sum_{h=1}^{L-1}(t_h/W_h)/N$. The partial derivatives under power allocation are not given here. They can be derived in a similar way.

6 Example: Stratification of the REV84 Population with 2 Stratum Jumpers

The REV84 population contains the 1984 real estate values for 284 Swedish municipalities, see Appendix B of Särndal, Swensson and Wretman (1992). REV84 is the stratification variable for this problem, the study variable Y is equal to REV84 except for municipalities 39 and 40 whose revenues have been increased to 8644 from respectively 655 and 637. The revenues for these 2 municipalities have been increased from the 5th to the 95th percentile of REV84; these two sampling units are stratum jumpers. This section presents stratified design for Y using REV84 as a stratification variable. We use $L = 5$ strata and set the target coefficient of variation at $c = .05$.

Assuming no discrepancies between the stratification variable and the target variable, the design obtained using Neymann allocation in Lavallée and Hidiroglou algorithm has $n = 19$. It is given in Table 1. Because of the 2 stratum jumpers, the coefficient of variation calculated for Y is $c = .094$ rather that the target value of .05. Furthermore the distribution of $\bar{y}_{st}$ is presented in Figure 1. It is bimodal with a probability of about 10% of overestimating the true $\bar{Y}$ by more that 33%.

To incorporate the occurrence of stratum jumpers, three strategies are investigated, namely

1. Use Lavallée and Hidiroglou algorithm with a coefficient of variation lower than .05;
2. The stratification algorithm derived from the multiplicative model;
3. The stratification algorithm derived from the random replacement model.

The three stratified designs obtained using algorithms constructed in Section 5 are given in the Table 2.

The three strategies considered in Table 2 increase the total sample size from 19 to 28. The respective coefficients of variation for the Y stratified sample mean for the increase in sample size strategy, the multiplicative and the random replacement model are .074, .067 and .057 respectively. The two models for stratum jumpers give good results. The random replacement model is much better that the other alternatives. Note that this design has a much larger minimum sampling fraction, min $f_h = .07$, as compared to minimum values of .03 and .04 for the other two. This explains the better performance of this stratification scheme since, under the random replacement model, the maximum sampling weight is around 50% of the maximum sampling weight under the other 2 models. This example suggests that constructing a stratified design using the random replacement model to account for the possible occurrence of stratum jumpers reduces the impact of stratum jumpers on survey estimates.

7 Discussion

Slanta and Krenzke (1994) encountered some numerical difficulties when using Lavallée and Hidiroglou algorithm with Neymann allocation: convergence was slow and sometimes the algorithm did not converge to the true minimum value for n. Indeed Schneeberger (1979) and Slanta and Krenzke (1994) showed that, for a particular bimodal population, the problem has a saddlepoint; that is the partial derivatives are all null at boundaries b_h which do not give a true minimum for n.

When using the algorithms constructed in this paper, we also experienced the numerical difficulties mentioned by Slanta and Krenzke (1994). The algorithms constructed under power allocation were generally more stable than those using Neymann allocation; numerical difficulties were more frequent when the number L of strata was large. Furthermore as the distribution for Y moved away from that of X, i. e. when the constants $C_{\epsilon,M}$ and the ϵ of the multiplicative and of the random replacement model increased, non convergence of the algorithm and failure to reach the global minimum for n were more frequent. In these situations, the stratification algorithm's starting values are of paramount importance. The numerical strategy used in Section 6 is to apply a sequence of stratification algorithms to get to the sampling designs given in Table 2. The stratum boundaries obtained at one step are used as starting values for the algorithm at the next step. At step 1, the initial boundaries are such that all the strata have the same size. The designs given in Table 2 result from the following three steps:

1. Lavallée and Hidiroglou algorithm for power allocation allocation with $p = .7$;
2. Lavallée and Hidiroglou algorithm for Neymann allocation used as starting values the b_h's from step 1;
3. Sethi's algorithm for stratum jumpers with the b_h's from step 2 as starting values.

Once the numerical obstacles have been circumvented, stratification algorithms accounting for differences between the survey variable and the stratification variables give good results. Over the past years, these designs have been used successfully on several consulting projects at the Statistical Consulting Unit of Université Laval.

References

Cochran, W. G. (1977). *Sampling Techniques. Third Edition.* John Wiley: New York

Dalenius, T. and Gurney, M. (1951) The Problem of Optimum Stratification II. *Skand. Akt.* **34**, 133-148

Hidiroglou, M. (1994) Sampling and Estimation for Establishment Surveys: Stumbling Blocks and Progress. *A.S.A. 1994 Proceedings of the Section on Survey Research Methods. Volume I*, 153-162

Hidiroglou, M. and Srinath, K. P. (1993) Problems Associated with Designing Sub-Annual Business Surveys. *Journal of Business and Economic statistics*, **11**, 397-405

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

Lee, H. (1995) Outliers in Business Surveys. *in Business Survey Methods* edited by B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. Colledge, and P. S. Kott, p. 503-526

Särndal C. E., Swensson B. and Wretman J. (1992) *Model Assisted Survey Sampling.* Springer Verlag: New York

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

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

Slanta, J. and Krenzke, T. (1994) Applying the Lavallée and Hidiroglou Method to Obtain Stratification Boundaries for the Census Bureau's Annual Capital Expenditure Survey. *A.S.A. 1994 Proceedings of the Section on Survey Research Methods*, 693-698

Table 1: A stratified design for the REV84 population using Lavallée and Hidiroglou algorithm

REV84 with L=5, CV=.05 and Neymann all.							
$\epsilon = 0$	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1273	878	57260	87	2	0.02	19
stratum 2	2336	1701	99688	81	2	0.02	19
stratum 3	4619	3114	351547	65	3	0.05	19
stratum 4	11776	6921	3724610	46	7	0.15	19
stratum 5	59878	28418	10^8	5	5	1	19

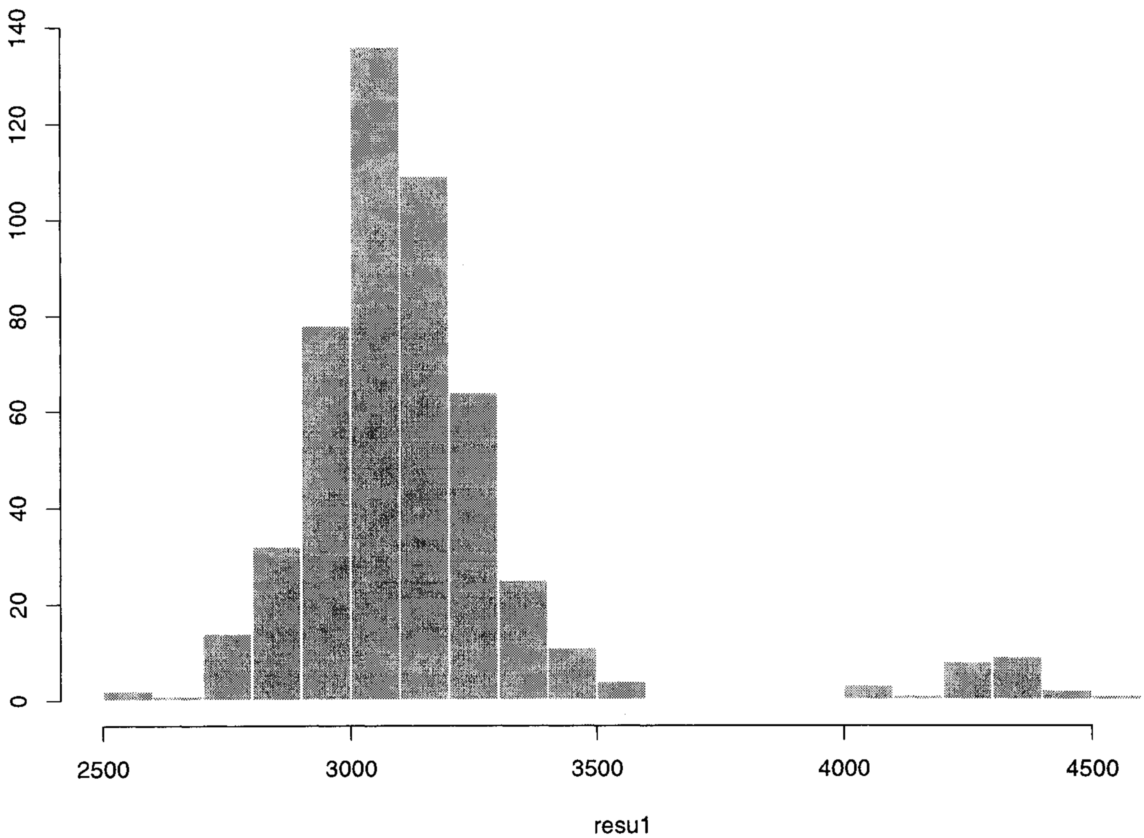

Figure 1: Distribution of the stratified mean for the population with 2 stratum jumpers under the sampling design of Table 1.

Table 2: Three Sampling Designs Accounting for the Presence of Stratum Jumpers. For each design the mean and the variance of REV84 in each stratum are given.

Lavallée and Hidiroglou algorithm with Neymann all.							
c=.037	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1273	878	57260	87	3	0.03	28
stratum 2	2335	1701	99688	81	3	0.04	28
stratum 3	4501	3114	351547	65	5	0.08	28
stratum 4	9845	6442	2027436	41	7	0.17	28
stratum 5	59878	19631	275502518	10	10	1	28
Multiplicative model with Neymann all.							
$C_{\epsilon,M} = 1.04$	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1574	1023	97245	121	5	0.04	28
stratum 2	3032	2219	168204	81	5	0.06	28
stratum 3	5597	4022	464471	44	5	0.11	28
stratum 4	11911	7709	2952313	33	8	0.24	28
stratum 5	59878	28418	426851844	5	5	1	28
Random Replacement Model with Neymann all.							
$\varepsilon = 0.011$	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1706	1070	116419	131	9	0.07	28
stratum 2	3193	2343	159431	75	5	0.07	28
stratum 3	5657	4117	410190	40	3	0.08	28
stratum 4	11797	7709	2952313	33	6	0.18	28
stratum 5	59878	28418	426851844	5	5	1	28

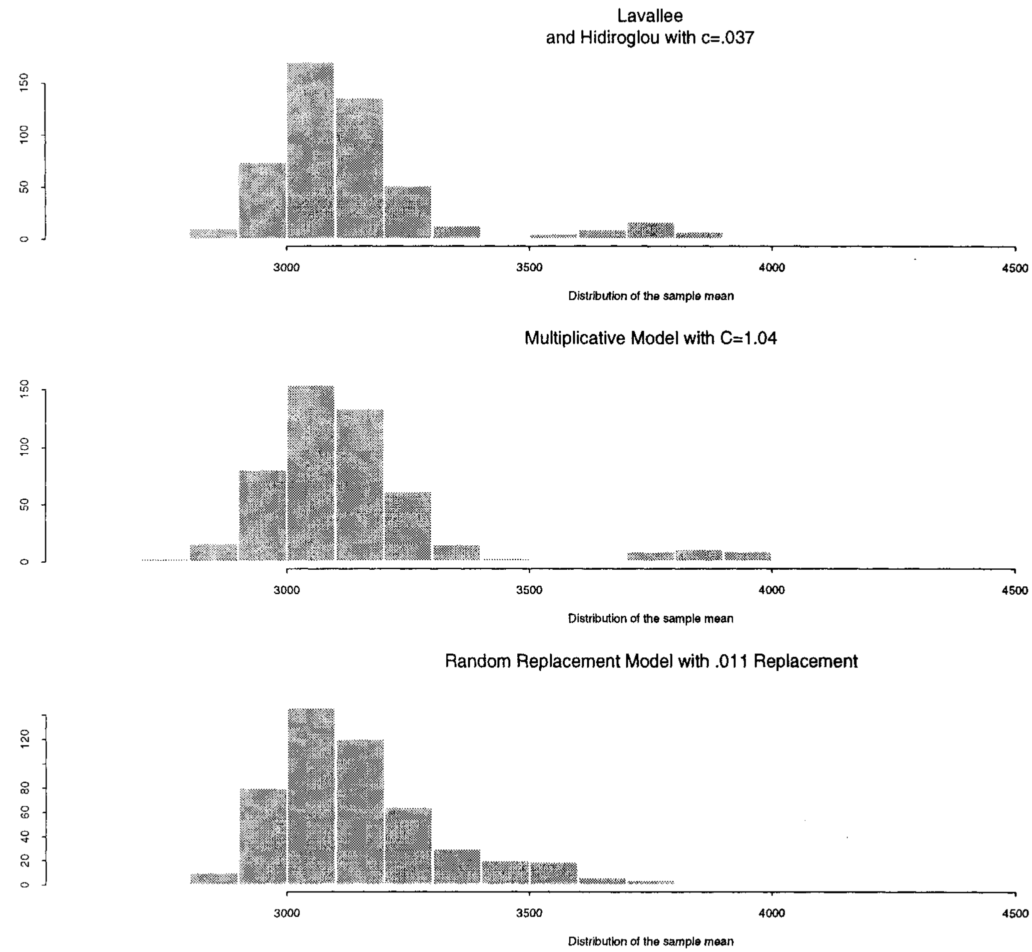

Figure 2: Distribution of the stratified means for the population with 2 stratum jumpers under the three sampling designs of Table 2.