A Generalization of Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys

by

Louis-Paul Rivest
Department de mathématiques et de statistique
Université Laval, Ste-Foy
Québec
Canada G1K 7P4

Abstract

This paper suggests stratification algorithms that account a discrepancy between the stratification variable and the study variable when planning a stratified survey design. Two models are proposed for the change between these two variables. One is a log-linear regression model; the other postulates that the study variable and the stratification variable coincide for most units, and that large discrepancies occur for some units. Then, the Lavallée and Hidiroglou (1988) stratification algorithm is modified to incorporate these models in the determination of the optimal sample sizes and of the optimal stratum boundaries for a stratified sampling design. An example illustrates the performance of the new stratification algorithm. A discussion of the numerical implementation of this algorithm is also presented.

Keywords: Neyman allocation, Power allocation, Stratified random sampling

1 Introduction

The construction of stratified sampling designs has a long history in the statistical sciences. Chapters 5 and 5A in Cochran (1977) review several techniques for splitting a population into strata. The construction of strata is a topic of current interest in the statistical literature. Recent contributions include Hedlin (2000) who revisits Ekman (1959) rule for stratification, and Dorfman and Valiant (2000) who compare model-based stratification with balanced sampling. Model based stratification, is discussed in Godfrey, Roshwalb, and Wright (1984) and in chapter 12 of Särndal, Swensson, and Wretman (1992).

In business surveys, populations have skewed distributions; a small number of units accounts for a large share of the total of the study variable. It is therefore appropriate to include all large units in the sample (Dalenius, 1952; Glasser, 1962). 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; small businesses get large sampling weights. The Lavallée and Hidiroglou (1988) stratification algorithm is often used to determine the stratum boundaries and the stratum sample sizes in this context (see for instance Slanta and Krenzke, 1994, 1996). This algorithm uses a stratification variable, known for all the units of the population. It gives the stratum boundaries and the stratum sample sizes that minimize the total sample size required to achieve a target level of precision. It uses an iterative procedure, due to Sethi (1963), to determine the optimal stratum boundaries. The Lavallée and Hidiroglou algorithm does not account for a difference between the stratification and the survey variables. As time goes by, this difference increases and the sampling design provided by the Lavallée and Hidiroglou algorithm may fail to meet the precision criterion.

Stratification in situations where the survey variable and the stratification variable differ is considered in Dalenius and Gurney (1951), see also Cochran (1977, chapter 5A). Many authors have studied approximate formulae for determining stratum boundaries, and for evaluating the gain in precision resulting from stratification on an auxiliary variable. Some relevant contributions are Serfling (1968), Singh and Sukatme (1969), Singh (1971), Singh and Parkash (1975), Anderson, Kish and Cornell (1976), Oslo (1976), Wang and Aggarwal (1984) and Yavada and Singh (1984). Hidiroglou and Srinath (1993) and Hidiroglou (1994) suggest techniques to update stratum boundaries using a new stratification variable. However these papers do

not explicitly provide stratification algorithms accounting for the discrepancy between the stratification variable and the survey variable. This paper fills this gap by constructing generalizations of the Lavallée and Hidiroglou (1988) algorithm that express the difference between these two variables in terms of a statistical model.

A brief review of stratified sampling and of sample allocation methods is first given. Models for the difference between stratification and survey variables are then proposed. The implementation of Sethi's algorithm, when the stratification and the survey variable differ, is then presented. Numerical illustrations are provided.

2 A Review of Stratified Random Sampling

Some of the standard notation of stratified random sampling that will be used in this paper is

L= the number of strata;

$W_h = N_h/N$ is for $h = 1, \ldots, L$ the relative weight of stratum h, 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 means of Y within stratum h;

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

In this paper the strata are constructed using X, a stratification variable. Stratum h consists of all units with an X-value in the interval $(b_{h-1}, b_h]$, where $-\infty = b_0 < b_1 < \ldots < b_{L-1} < b_L = \infty$ are the stratum boundaries.

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

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2 \tag{2.1}$$

In business surveys, all the big firms are sampled; we choose stratum L as the take-all stratum so that $n_L = N_L$. For $h < L$, n_h, the sample size in take-some stratum h, can be written as $(n - N_L)a_h$ where n is the total sample size and a_h depends on the allocation rule. The two allocation rules that are considered in this paper are

- The power allocation rule

$$a_h = \frac{(W_h \bar{Y}_h)^p}{\sum_{k=1}^{L-1} (W_k \bar{Y}_k)^p} \tag{2.2}$$

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

- The Neyman allocation rule

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}}. \tag{2.3}$$

Solving (2.1) for n leads to

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2 / a_h}{\mathrm{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2 / N} \tag{2.4}$$

The optimal stratum 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 (CV). The range $c = 1\%$ to 10% is often used for business surveys.

3 Some Models for the Discrepancy Between the Stratification and the Survey Variable

In this section $\{x_i, i = 1, \ldots, N\}$ denotes the known stratification variable for the N units in the population. Many stratification algorithms, including Lavallée and Hidiroglou, suppose that $\{x_i, i = 1, \ldots, N\}$ also represents the values of the study variable. This section suggests statistical models to account for a difference between these two variables.

For the sequel, it is convenient to look at X and Y as continuous random variables and to let $f(x),\ x \in R$ denote the density of X. The data $\{x_i, i = 1, \ldots, N\}$ can be viewed as N independent realizations of the random variable X. Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ 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, for $h = 1, \ldots, L-1$. Three models for the difference between X and Y are next given along with their conditional means and variances for Y.

3.1 A Log-linear Model

The first model considers that $\log(Y) = \alpha + \beta_{log}\log(X) + \epsilon$, where ϵ is a normal random variable with mean 0 and variance σ^2_{log}, which is independent from X, and α and β_{log} are parameters to be determined. When $\alpha = 0$, $\beta_{log} = 1$ and $\sigma^2_{log} = 0$, one has $X = Y$; the survey and the stratification variables are the same. In general, $Y = e^{\alpha}X^{\beta_{log}}e^{\epsilon}$. The conditional moments of Y can be evaluated using the basic properties of the lognormal distribution (see Johnson and Kotz, 1970), that is

$$E(e^{\epsilon}) = e^{\sigma^2_{log}/2} \quad \text{and} \quad \text{Var}(e^{\epsilon}) = e^{\sigma^2_{log}}(e^{\sigma^2_{log}} - 1).$$

One has

$$E(Y|b_h \geq X > b_{h-1}) = \exp(\alpha + \sigma^2_{log}/2)E(X^{\beta_{log}}|b_h \geq X > b_{h-1})$$

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

$$\begin{aligned} & \text{Var}(E(Y|X)|b_h \geq X > b_{h-1}) + E(\text{Var}(Y|X)|b_h \geq X > b_{h-1}) \\ = \;& \exp(2\alpha + \sigma^2_{log})\{\text{Var}(X^{\beta_{log}}|b_h \geq X > b_{h-1}) + (e^{\sigma^2_{log}} - 1)E(X^{2\beta_{log}}|b_h \geq X > b_{h-1})\} \\ = \;& \exp(2\alpha + \sigma^2_{log})\{e^{\sigma^2_{log}}E(X^{2\beta_{log}}|b_h \geq X > b_{h-1}) - E(X^{\beta_{log}}|b_h \geq X > b_{h-1})^2\} \end{aligned}$$

The parameter values β_{log} and σ_{log} can sometimes be calculated from historical data. Simple ad hoc values are $\beta_{log} = 1$ and $\sigma^2_{log} = (1 - \rho^2)\text{Var}(\log(X))$. Here ρ is the assumed correlation between $\log(X)$ and $\log(Y)$. It can be set equal to predetermined values such as 0.95 or 0.99.

3.2 A Linear Model

In the survey sampling literature, the discrepancy between Y and X is often modeled with a heteroscedastic linear model,

$$Y = \beta_{lin}X + \epsilon, \tag{3.5}$$

where the conditional distribution of ϵ, given X, has mean 0 and variance $\sigma^2_{lin}X^{\gamma}$, for some non negative parameter γ. Straightforward calculations lead to $E(Y|b_h \geq X > b_{h-1}) = \beta_{lin}E(X|b_h \geq X > b_{h-1})$ while $\text{Var}(Y|b_h \geq X > b_{h-1}) = \beta^2_{lin}\{\text{Var}(X|b_h \geq X > b_{h-1}) + (\sigma_{lin}/\beta_{lin})^2E(X^{\gamma}|b_h \geq X > b_{h-1})\}$

For an arbitrary $\gamma \geq 0$, the conditional variance of Y depends on three conditional moments of X. The generalization of Sethi's algorithm presented in Section 5 does not work in this situation. Note however that when $\gamma = 2$, the conditional mean and variance of Y are proportional to those for the log-linear model with

$$\beta_{log} = 1 \quad \text{and} \quad \sigma^2_{log} = \log(1 + (\sigma_{lin}/\beta_{lin})^2); \tag{3.6}$$

the proportionality factors are $\exp(\alpha + \sigma^2_{log}/2)/\beta_{lin}$ and $\exp(2\alpha + \sigma^2_{log})/\beta^2_{lin}$ for the conditional expectations and the conditional variances respectively. Thus the two models for the discrepancy between the stratification and the survey variable, either the log-linear model of section 3.1 or the linear model (3.5) with parameter $\gamma = 2$, lead, in Section 5, to the same stratified design provided that (3.6) holds. In the later sections, the log-linear model is used to represent the change between X and Y. It should give good results when the true relationship between Y and X is modeled by (3.5) with $\gamma \approx 2$. When model (3.5) is assumed to hold with a smaller value of γ, the algorithm of section 5 can still be implemented when γ is set to either 0 or 1. This is however not pursued in this paper.

3.3 A Random Replacement Model

This model assumes that the stratification variable is equal to the survey variable, i. e. $X = Y$, for most units. There is however a small probability ϵ that a unit changed drastically; its Y value then has $f(x)$ as density and is distributed independently of its X value. This is the approach used in Rivest (1999) to model the occurrence of stratum jumpers for which X is not representative of Y. More formally, this can be written as,

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

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

$$E(Y|b_h \geq X > b_{h-1}) = (1 - \epsilon)E(X|b_h \geq X > b_{h-1}) + \epsilon E(X),$$

while its conditional variance is equal to

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

4 An Example

Before addressing the technical details underlying the construction of the algorithms, it is convenient to look at an example. Consider the $MU284$ population of Särndal, Swensson, and Wretman (1992), presenting data on 284 Swedish municipalities.

To build a a stratified design for estimating the average of $RMT85$, the revenues from the 1985 municipal taxation, $REV84$, the real estate value according to 1984 assessment, is used as a stratification variable. One takes $L = 5$ and set the target CV at 5%. Two stratified designs obtained with the Lavallée and Hidiroglou algorithm are given in Table 1, for the power allocation with $p = 0.7$ and the Neyman allocation. Both have $n = 19$. When applied on survey variable $RMT85$, these two designs give estimators of total revenue with coefficients of variation of 8.3% and 7.3% respectively. Failing to account for a change between the survey and the stratification variables yields estimators that are more variable than expected.

To model the discrepancy between $REV84$ and $RMT85$, we use the log-linear model of Section 3.1. There are outliers in the linear regression of $\log(RMT85)$ on $\log(REV84)$; they make the least squares estimates of β_{log} and σ_{log} unrepresentative of the relationship between the two variables. Robust estimates obtained with the Splus function lmRobMM are used instead. They are given by $\hat{\beta}_{log} = 1.1$ and $\hat{\sigma}_{log} = 0.2116$. Table 2 gives the stratified designs obtained with the generalized Lavallée and Hidiroglou algorithm for two allocation rules. They both give estimators of the total of $RMT85$ having a CV of 5.7%. This CV is still larger that 5%. Since there are outliers in the log-linear regression, the assumption of normal errors made in Section 3.1 is not met. This might explain the failure to reach the target CV exactly. The increase in sample size for $n = 19$ to $n = 28$ is noteworthy! For both allocation methods the design obtained using the log-linear model has smaller take-all strata than Lavallée and Hidiroglou.

An alternative to the generalized Lavallée and Hidiroglou algorithm for the construction of stratified designs is to us their original algorithm with a smaller target CV. This increases the sample size thereby reducing the variance of the estimator of the total of the survey variable. When constructing a design for $RMT85$ using $REV84$ as a stratification variable, the standard Lavallée and Hidiroglou algorithm with power allocation rule ($p = 0.7$) and a target CV of 3.6%, yields a stratified design with $n = 28$. This design has the same sample size as those presented in Table 2. The CV of the estimator of the total $RMT85$ is 5.7%, the same as the

Power allocation with $p = 0.7$							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1251	874	56250	86	1	0.01	19
stratum 2	2352	1696	100898	82	2	0.02	19
stratum 3	4603	3114	351547	65	3	0.05	19
stratum 4	10606	6442	2027436	41	3	0.07	19
stratum 5	59878	19631	275502518	10	10	1	19
Neyman allocation							
	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	426851844	5	5	1	19

Table 1: Stratified designs obtained with the Lavallée and Hidiroglou algorithm for the MU284 population using REV84 as stratification variable and a target CV of 5%.

Log-linear model stratification algorithm with power allocation with $p = 0.7$							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1558	1023	97245	121	4	0.03	28
stratum 2	3031	2219	168204	81	5	0.06	28
stratum 3	5706	4022	464471	44	6	0.14	28
stratum 4	11107	7602	2659061	32	7	0.22	28
stratum 5	59878	25536	391314713	6	6	1	28
Log-linear model stratification algorithm with Neyman allocation							
	b_h	mean	variance	N_h	n_h	f_h	n
stratum 1	1582	1023	97245	121	4	0.03	28
stratum 2	3040	2219	168204	81	5	0.06	28
stratum 3	5608	4022	464471	44	5	0.11	28
stratum 4	11476	7709	2952313	33	9	0.27	28
stratum 5	59878	28418	426851844	5	5	1	28

Table 2: Stratified designs obtained with the generalized Lavallée and Hidiroglou algorithm for the MU284 population using REV84 as stratification variable, a log-linear with $\beta_{log} = 1.1$ and $\sigma_{log} = 0.2116$ for the discrepancy between $REV84$ and $RMT85$, and a target CV of 5%.

CVs obtained with the designs of Table 2. The main difference between these designs is the size of the take-all stratum. The design constructed with the Lavallée and Hidiroglou algorithm has a take-all stratum of size $N_5 = 13$ as compared to $N_5 = 5$ and $N_5 = 6$ for the designs of Table 2. Allowing the stratification and the survey variables to differ appears to reduce the relative importance of the take-all stratum in the sampling design. Further investigations are needed to ascertain this hypothesis.

The stratification algorithm for the random replacement model of Section 3.3 (with Neyman allocation) was also applied to $REV84$. Assuming changes in 2% of the units ($\epsilon = 0.02$), the generalized Lavallée and Hidiroglou algorithm yields a stratified design with $n = 37$ sample units; the resulting estimator of total $RMT85$ has a CV of 5.5%. An interesting property of this stratified design is that the smallest sampling fraction is $\min_h f_h = 9.3\%$; it is much larger than $\min_h f_h$ for the designs of Tables 1 and 2. Despite the presence of outliers, the random replacement model does not describe the changes between $REV84$ and $RMT85$ as well as the log-linear model. This explains why a larger sample size, 37 instead of 28, is needed to get an estimator with a variance comparable to that obtained with the stratification based on a log-linear model.

5 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 $\{x_i;\ i = 1, \ldots, N\}$ of variable X for all the units in the population. A model, such as those given in Section 3, 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 uses the log-linear model of Section 3.1 to account for the differences between Y and X. Modifications to handle the random replacement model are easily carried out (see Rivest, 1999).

5.1 A Generalization of Sethi's (1963) Stratification Method

It is convenient to consider an infinite population analogue to equation (2.4) for n. Since the random variable X has a density $f(x)$, the first two conditional moments

of Y given that $b_{h-1} < X \leq b_h$ can be written in terms of

$$W_h = \int_{b_{h-1}}^{b_h} f(x)dx, \quad \phi_h = \int_{b_{h-1}}^{b_h} x^\beta f(x)dx, \text{ and } \quad \psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x)dx,$$

where β is the slope of the log-linear model given in Section 3.1 (in this section β and σ represent parameters of the log-linear model of Section 3.1, since there is no risk of confusion the subscript *log* is not used anymore) . For stratification purposes, it is useful to rewrite (2.4) in terms of the conditional means and variances for Y,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \text{Var}(Y|b_h \geq X > b_{h-1})/a_{h,X}}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} W_h \text{Var}(Y|b_h \geq X > b_{h-1})/N}, \tag{5.7}$$

where $a_{h,X}$ denotes 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 $h = 1, \ldots, L-1$. Given a model for the relationship between Y and X, $\text{Var}(Y|b_h \geq X > b_{h-1})$ 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 derivatives of n with respect to b_h can be evaluated, for $h < L-1$, using the chain rule,

$$\frac{\partial n}{\partial b_h} = \frac{\partial n}{\partial W_h}\frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \phi_h}\frac{\partial \phi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h}\frac{\partial \psi_h}{\partial b_h} + \frac{\partial n}{\partial W_{h+1}}\frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \phi_{h+1}}\frac{\partial \phi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}}\frac{\partial \psi_{h+1}}{\partial b_h}$$

Observe that

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

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

$$\frac{\partial n}{\partial b_h} = f(b_h)\left\{\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right) + \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right) b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right) b_h^{2\beta}\right\}.$$

Similarly,

$$\frac{\partial n}{\partial b_{L-1}} = f(b_{L-1}) \left\{ -N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \phi_{L-1}} b_{L-1}^{\beta} + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^{2\beta} \right\}.$$

The Sethi's (1963) algorithm is used to solve $\partial n / \partial b_h = 0$. It considers that the partial derivatives are proportional to quadratic functions in b_h^{β}. The updated value for b_h^{β} is given by the largest root of the corresponding quadratic function. When $h < L-1$, this gives

$$b_h^{\beta\, new} = -\left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right) \Big/ \left\{ 2 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \right\}$$
$$+ \frac{\left\{ \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}} \right)^2 - 4 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}} \right) \right\}^{1/2}}{\left\{ 2 \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}} \right) \right\}},$$

while for $h = L-1$ we have

$$b_{L-1}^{\beta\, new} = \frac{-\frac{\partial n}{\partial \phi_{L-1}} + \left\{ \left(\frac{\partial n}{\partial \phi_{L-1}} \right)^2 - 4 \frac{\partial n}{\partial \psi_{L-1}} \left(\frac{\partial n}{\partial W_{L-1}} - N \right) \right\}^{1/2}}{\left(2 \frac{\partial n}{\partial \psi_{L-1}} \right)}.$$

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^{β} within stratum h. These moments are evaluated using the N x-values in the population. For instance,

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

Applications of this general method are provided next.

When using Sethi's algorithm, one typically has $L \geq 3$. Note however that it also works when $L = 2$. In this case, the algorithm is searching for the boundary between a take-all and a take-some stratum. Successive evaluations of $b_{L-1}^{\beta\, new}$ presented above yield an optimal boundary. When one assumes that the stratification and the study variable coincide, i. e. $X = Y$, this boundary is nearly identical to that obtained with the algorithm presented in Hidiroglou (1986).

5.2 An Algorithm for Power Allocation

For the log-linear model of Section 3.1, the conditional expectation is $E(Y|b_h \geq X > b_{h-1}) = C\phi_h/W_h$ while the conditional variance is

$$\mathrm{Var}(Y|b_h \geq X > b_{h-1}) = C^2\{e^{\sigma^2}\psi_h/W_h - (\phi_h/W_h)^2\},$$

where $C = \exp(\alpha + \sigma^2/2)$. Under the power allocation rule, $a_{h,X} = \phi_h^p / \sum_{h=1}^{L-1} \phi_k^p$, and formula (5.7) for n becomes

$$n = NW_L + \frac{\sum_{h=1}^{L-1} \phi_h^p \sum_{h=1}^{L-1}(e^{\sigma^2}W_h\psi_h - \phi_h^2)/\phi_h^p}{(\sum x_i^\beta/N)^2c^2 + \sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h - \phi_h^2/W_h)/N}.$$

The partial derivatives needed to implement the stratification algorithm are easily calculated; for $h \leq L-1$,

$$\begin{aligned}
\frac{\partial n}{\partial W_h} &= \frac{Ae^{\sigma^2}\psi_h/\phi_h^p}{F} - \frac{AB(\phi_h/W_h)^2/N}{F^2} \\
\frac{\partial n}{\partial \phi_h} &= \frac{A\{-pe^{\sigma^2}(W_h\psi_h - \phi_h^2)/\phi_h^{p+1} - 2/\phi_h^{p-1}\} + p\phi_h^{p-1}B}{F} + 2\frac{AB\phi_h/(NW_h)}{F^2} \\
\frac{\partial n}{\partial \psi_h} &= e^{\sigma^2}\frac{AW_h/\phi_h^p}{F} - e^{\sigma^2}\frac{AB/N}{F^2},
\end{aligned}$$

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

5.3 An algorithm for Neyman allocation

Under Neyman allocation, allocation rule (2.3) written in terms of W_h, ϕ_h, and ψ_h is

$$a_{h,X} = \frac{\{e^{\sigma^2}\psi_h W_h - \phi_h^2\}^{1/2}}{\sum_{h=1}^{L-1}\{e^{\sigma^2}\psi_h W_h - \phi_h^2\}^{1/2}}$$

and the formula for n is

$$n = NW_L + \frac{\{\sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}\}^2}{(\sum x_i^\beta/N)^2c^2 + \sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h - \phi_h^2/W_h)/N}.$$

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

$$\begin{aligned}
\frac{\partial n}{\partial W_h} &= \frac{Ae^{\sigma^2}\psi_h/(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}}{F} - \frac{A^2(\phi_h/W_h)^2/N}{F^2} \\
\frac{\partial n}{\partial \phi_h} &= \frac{-2A\phi_h/\{e^{\sigma^2}W_h\psi_h - \phi_h^2\}^{1/2}}{F} + \frac{2A^2\phi_h/(W_h N)}{F^2} \\
\frac{\partial n}{\partial \psi_h} &= e^{\sigma^2}\frac{AW_h/\{e^{\sigma^2}W_h\psi_h - \phi_h^2\}^{1/2}}{F} - e^{\sigma^2}\frac{A^2/N}{F^2},
\end{aligned}$$

where $A = \sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}$, and $F = (\sum x_i^{\beta}/N)^2 c^2 + \sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h - \phi_h^2/W_h)/N$.

6 Numerical Considerations

Slanta and Krenzke (1994, 1996) encountered numerical difficulties when using the Lavallée and Hidiroglou algorithm with Neyman 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 saddle; 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 alluded to in Slanta and Krenzke (1994). The algorithms constructed under power allocation were generally more stable than those using Neyman 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. as σ^2 increases, 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 were of paramount importance. For instance, in Table 2, the design accounting for changes between Y and X obtained under Neyman allocation depends heavily on the starting values. The one presented in Table 2 uses the boundaries presented in Table 2 for the power allocation as starting values. Starting the algorithm with the boundaries obtained in Table 1 for the Lavallée Hidiroglou algorithm with Neyman allocation yields a different sampling design having $n = 29$.

A good numerical strategy is to run the stratification algorithm for several intermediate designs to get to a final sampling design, with the stratum boundaries

obtained at one step used as starting values for the algorithm at the next step. The log-linear algorithm is always run in two steps; first run the Lavallée and Hidiroglou algorithm, setting $\sigma = 0$, and use these boundaries as starting value for the algorithm with a non null σ. Also use as starting value for Neyman allocation the corresponding boundaries found under power allocation with a p value around 0.7.

7 Conclusion

This paper has proposed generalizations of the Lavallée and Hidiroglou stratification algorithm that account for a difference between the stratification and the survey variables. Two statistical models have been introduced for this purpose. The new class of algorithms uses the Chain Rule to derive partial derivatives and Sethi's (1963) technique to find the optimal stratum boundaries.

The log-linear model stratification algorithm proposed in this paper was used successfully in several surveys designed at the Statistical Consulting Unit of Université Laval. For estimating total maple syrup production in a year, the number of sap producing maples for a producer was a convenient size variable. Historical data was used to estimate the parameters of the log-linear model linking sap producing maples and production volume. Another example is the estimation of the total maintenance deficit of hospital buildings in Quebec. The value of each building was the known stratification variable. The maintenance deficit was estimated to be in the range $(20\%, 40\%)$ by experts. Solving $4\sigma_{log} = \log(40\%) - \log(20\%)$ gives $\sigma_{log} = \log(2)/4 = 0.17$ as a possible parameter value for the log-linear model of Section 3.1. In these two examples accounting for changes between the stratification and the survey variables increased the sample size n by a fair percentage and yielded survey estimators whose estimated CVs were close to the target CVs.

Two SAS IML functions implementing the algorithm presented in this paper, for power and Neyman allocation, are available on the author's website at http://www.mat.ulaval.ca/pages/lpr/. They allow user specified starting values for the stratum boundaries; they can be used to implement the numerical strategies presented in Section 8.

Acknowledgments

I am grateful to Nathalie Vandal and to Gaétan Daigle for constructing SAS IML programs for the stratification algorithms used in the paper. The constructive comments of the associate editor and of the referee are gratefully acknowledged.

References

Anderson, D. W., Kish, L., and Cornell, R. G. (1976) Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model, *Journal of the American Statistical Association*, **71**, 887-892

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

Dalenius, T. (1952) The Problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift* **35**, 61-70

Dalenius, T. and Gurney, M. (1951) The Problem of optimum stratification II. *Skandinavisk Aktuarietidskrift* **34**, 133-148

Dorfman A. H. and Valliant, R. (2000) Stratification by size revisited. *Journal of Official Statistics*, **16**, 139-154

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

Glasser, G. J. (1962) On the complete coverage of large units in a statistical study. *Review of the International Statisticasl Institute*, **30**, 28-32

Godfrey, J. , Roshwalb, A. , and Wright, R. L. (1984) Model-based stratification in inventory cost estimation , *Journal of Business and Economic Statistics*, **2**, 1-9

Hedlin, D. (2000) A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, **16**, 15-29

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

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*, American Statistical Association (Alexandria, VA) 153-162

Hidiroglou, M. A. , and Srinath, K. P. (1993) Problems associated with designing subannual business surveys, *Journal of Business and Economic Statistics*, **11**, 397-405

Johnson, N. L. and Kotz, S. (1970) *Continuous Univariate Distribution-1.* John Wiley: New York

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

Oslo, I.T. (1976) A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*, **23**, 15-25

Rivest, L.-P. (1999) Stratum jumpers: Can we avoid them?, *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA), 64–72,

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

Serfling, R. J. (1968) Approximate optimal stratification. *Journal of the American Statistical Association*, **63**, 1298-1309

Sethi, V. K. (1963) A note on the optimum stratification of populations for estimating the population means, *Australian Journal of Statistics* **5**, 20-33

Singh, R. J. (1971) Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*, **66**, 829-834

Singh, R. J and Parkash, D. (1975) Optimal stratification for equal allocation of the auxiliary variable. *Annals of the Institute of Statistical Mathematics*, **27**, 273-280

Singh, R. and Sukatme, B. V. (1975) Optimum stratification. *Annals of the Institute of Statistical Mathematics*, **21**, 515-528

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

Slanta, J. and Krenzke, T. (1996) Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Survey Methodology*, **22**, 65-75

Wang, M. C. and Aggarwal, V. (1984) Stratification under a particular Pareto distribution. *Communications in Statistics, Part A – Theory and Methods*, **13**, 711-735

Yavada, S. and Singh, R. (1984) Optimum stratification for allocation proportional to strata totals for simple random sampling scheme. *Communications in Statistics, Part A – Theory and Methods*, **13**, 2793-2806