Stratification and Allocation for Estimating a Complex Statistic Using Auxiliary Data

Serge Godbout[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6,serge.godbout@statcan.gc.ca

Abstract
In the literature, methods for creating strata, identifying take-all units and allocating samples generally seek to minimize the variance under a fixed cost or to minimize the costs under a fixed variance in estimating a total. For a complex statistic that corresponds to a smooth function of total, such as a ratio or a regression coefficient, a sample design based on the total of one of the variables involved in the statistic is not necessarily optimal. We propose using a Taylor linearization to build an auxiliary variable to identify take-all units, generate strata and allocate the sample. Then, we use the usual stratification and allocation methods to minimize the variance of the estimator or the size of the sample. A simulation study is used to measure the method's efficiency.

Key Words: Stratification, Sample allocation, Taylor linearization, Complex statistic

1. Introduction

In a survey, when we have an auxiliary variable for the whole population that correlates with our variable of interest, we can stratify the population in order to increase the precision of our estimator by creating homogeneous groups. Different allocation methods can be used to divide the sample efficiently among the strata (Cochran, 1977; Särndal et al, 1992; Lohr, 1999). Also, if the target population is skewed, the identification of take-all (TA) units also helps achieve a much greater precision for a given sample size (Lavallée and Hidiroglou, 1988).

However, the methods described are generally applied to the estimation of a total $t_y = \sum_U y_k$. In the case of a complex statistic or estimator θ corresponding to a smooth function of totals $t_{y_i} = \sum_U y_{ik}$, a common solution is to build a sample design aimed at optimizing one or more t_{y_i} , particularly if these t_{y_i} are also among the survey variables of interest (on the issue of allocation in the case of several variables of interest, see Cochran, 1977 and Särndal *et al*, 1992). However, this sample design may not be optimal for θ .

We propose a method based on the Taylor linearization of the complex statistic. The statistic is first linearized and the usual stratification and allocation methods are then applied. We will measure the method's efficiency through a simulation study. We will present an example of the application on Statistics Canada's *Survey of Employment, Payrolls and Hours* (Grondin *et al*., 2005).

2. Proposed Methodology

To calculate the linearized variable of a complex statistic corresponding to a smooth function of totals, we propose to use the Demnati-Rao (2004) method.

2.1 Demnati-Rao linearization method
Taylor linearization is a method that is regularly used to estimate variance (for an overview, see Woodruff, 1971; Binder, 1996; Demnati and Rao, 2004). Among the different methods of linearization, the one proposed by Demnati and Rao is of particular interest: in addition to having the properties sought to estimate variance (approximate

unbiasedness for the model variance of the estimator under a hypothetical model and validity under a conditional repeated sampling framework), this is a simple application method involving the derivation of a linearized variable z_k for all of the *k* units belonging to *s* and estimated based on the sample produced. To show this linearization method, we take a function $f(\mathbf{b})$ where $\mathbf{b}$ is a column vector *N* of weights. This means a size *N* population *U* and its parameters of interest $\theta = f(\mathbf{d}(U))$, where $\mathbf{d}(U)$ is the dimension *N* unit vector. The expansion estimator becomes $\hat{\theta} = f(\mathbf{d}(s))$ where $\mathbf{d}(s)$ is the dimension *N* vector containing the survey weights (for the sampled units) and 0s for the others. The linearized variable $\tilde{z}_k$ of parameter θ is given by $\tilde{z}_k = \partial f / \partial b_k \big|_{\mathbf{b}=\mathbf{d}(U)}$. In a case where only data from sample *s* are available, we instead use z_k for the linearized variable of estimator $\hat{\theta}$, which is obtained through $z_k = \partial f / \partial b_k \big|_{\mathbf{b}=\mathbf{d}(s)}$. In the issue at hand, because we have the values for our auxiliary variable for the entire *U* universe, we will use $\tilde{z}_k$.

2.2 Example

We illustrate the problem of stratification and allocation through an example aimed at estimating a ratio. The parameter of interest is $R_{y/x} = t_y / t_x$, where $t_x = \sum_U d_k(U) x_k$ and $t_y = \sum_U d_k(U) y_k$ and the expansion estimator is given by $\hat{R}_{y/x} = \hat{t}_y / \hat{t}_x$, where $\hat{t}_x = \sum_U d_k(s) x_k$ and $\hat{t}_y = \sum_U d_k(s) y_k$. Under Demnati-Rao linearization, the linearized variable $\tilde{z}_{Rk}$ for our parameter of interest is given by:

$$\tilde{z}_{Rk} = \left. \frac{\partial R_{y/x}}{\partial b_k} \right|_{\mathbf{b}=\mathbf{d}(U)} = \frac{1}{t_x}\left(y_k - R_{y/x} x_k\right) \qquad (1)$$

This linearized variable will serve as an auxiliary variable to estimate $R_{y/x}$. By dividing the population into homogeneous groups with $\tilde{z}_{Rk}$ and identifying extreme values as being TA units by using common stratification and allocation methods, we can define a sample design that will increase the precision of the $\hat{R}_{y/x} = \hat{t}_y / \hat{t}_x$ estimator.

3. Simulation Study

To measure the efficiency of using the linearized variable to stratify a population and allocate a sample, we performed a Monte Carlo (MC) simulation study.

3.1 Description of the hypotheses

We started by generating a population of 1,000 units based on hypotheses built on observations of the real data from a survey (presented in point 4).

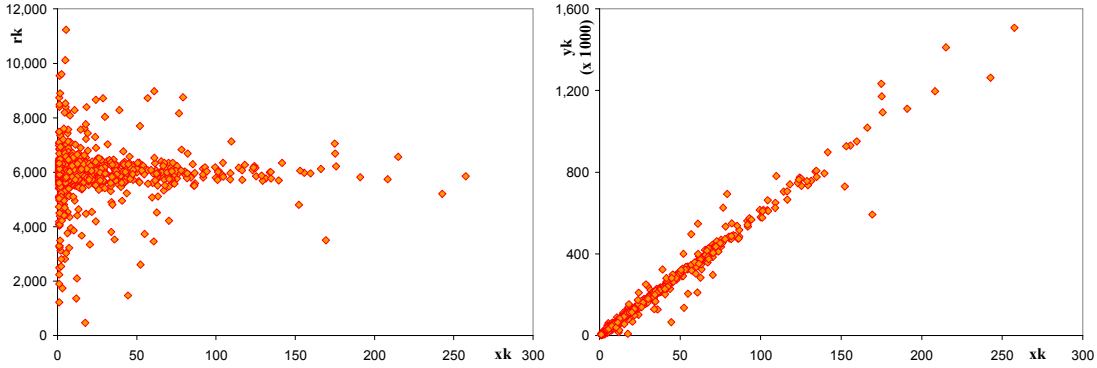
$$x_k \rightarrow \Gamma(0.4\,,\,50) + 1$$

$$r_k \rightarrow \begin{cases} N(6000\,,\,250^2) & 80\%\ of\ units \\ N(6000\,,\,250^2) + N(0\,,\,2000^2) & 20\%\ of\ units \end{cases}$$

$$y_k = x_k r_k$$

We chose the Γ function for x_k to generate a skewed distribution with a few very large units, while the addition of a unit was used to avoid values too close to 0. In addition, we decided to generate values of r_k based on a normal distribution centred at 6,000, while 20% of the units received normal additional noise at zero expectation to create a pool of outlier units. Figure 1 presents the population thus generated. Once the population was created, we independently selected 1,000 size 50 simple random samples ($i = 1,...,1000$) for each of the different stratification and allocation scenarios presented below.

Figure 1: Simulated Population (1,000 units)



Once all the samples were selected, for each $\hat{\theta}$ estimator, we calculated the 1,000 MC $\hat{\theta}_i$ estimates. We used the Monte Carlo coefficient of variation (CV^{MC}) as a measure of goodness, calculated as follows:

$$CV^{MC}(\hat{\theta}) = \frac{1}{\theta}\left(\sum_{i=1}^{1000}\frac{(\hat{\theta}_i - \Sigma\hat{\theta}_i/1000)^2}{1000}\right)^{1/2}$$

Given that the estimators studied are all unbiased, we have a property whereby the CV^{MC} measures the square root of the relative mean square error.

3.2 Parameters of interest, estimators and linearized variables

We looked at four parameters and their estimators, that is to say two simple statistics (totals t_x and t_y of variables x_k and y_k) and two complex statistics ($R_{y/x}$ ratio of the x_k and y_k variables and their coefficient of correlation $RSQ_{x,y}$).

Table 2: Parameters of Interest and Estimators for the Simulation Study

Names	Parameter	Estimator
Total X	$t_x = \sum_U d_k(U)x_k$	$\hat{t}_x = \sum_U d_k(s)x_k$
Total Y	$t_y = \sum_U d_k(U)y_k$	$\hat{t}_y = \sum_U d_k(s)y_k$
Y/X Ratio	$R_{y/x} = t_y / t_x$	$\hat{R}_{y/x} = \hat{t}_y / \hat{t}_x$
Coefficient of correlation between X and Y	$RSQ_{x,y} = \frac{\left(Nt_{xy} - t_x t_y\right)^2}{\left(Nt_{x^2} - t_x^2\right)\left(Nt_{y^2} - t_y^2\right)}$	$\hat{RSQ}_{x,y} = \frac{\left(\hat{N}\hat{t}_{xy} - \hat{t}_x\hat{t}_y\right)^2}{\left(\hat{N}\hat{t}_{x^2} - \hat{t}_x^2\right)\left(\hat{N}\hat{t}_{y^2} - \hat{t}_y^2\right)}$

where $t_{xy} = \sum_U d_k(U)x_k y_k$, $\hat{t}_{xy} = \sum_U d_k(s)x_k y_k$, $t_{y^2} = \sum_U d_k(U)y_k^2$, $\hat{t}_{y^2} = \sum_U d_k(s)y_k^2$

$N = \sum_U d_k(U)$ and $\hat{N} = \sum_U d_k(s)$

For simple estimators of totals, the linearized variables $\tilde{z}_{t_x k}$ and $\tilde{z}_{t_y k}$ are simply x_k and y_k , respectively. For the ratio estimator, the linearized variable $\tilde{z}_{Rk}$ is given by (1). In the case of the estimator for the coefficient of correlation, we get:

$$\tilde{z}_{RSQk} = \left.\frac{\partial RSQ_{x,y}}{\partial b_k}\right|_{\mathbf{b}=\mathbf{d}(U)}$$
$$= \frac{2\left(Nt_{xy} - t_x t_y\right)\left(t_{xy} - Nx_k y_k - x_k t_y y_k t_x\right)}{\left(Nt_{x^2} - t_x^2\right)\left(Nt_{y^2} - t_y^2\right)}$$
$$- \frac{RSQ\left[\left(Nt_{x^2} - t_x^2\right)\left(t_{y^2} - Ny_k^2 - 2t_y y_k\right) + \left(Nt_{y^2} - t_y^2\right)\left(t_{x^2} - Nx_k^2 - 2t_x x_k\right)\right]}{\left(Nt_{x^2} - t_x^2\right)\left(Nt_{y^2} - t_y^2\right)}$$

Table 3 provides the correlations among the four linearized variables as observed in the population created for the simulation. It is evident that x_k and y_k are strongly correlated.

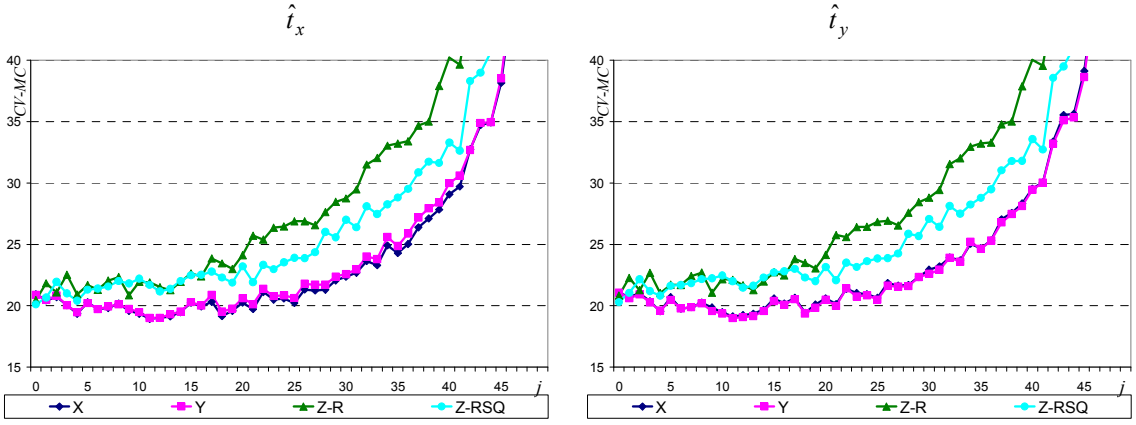
Table 3: Correlations Among the Four Linearized Variables in the Simulated Population

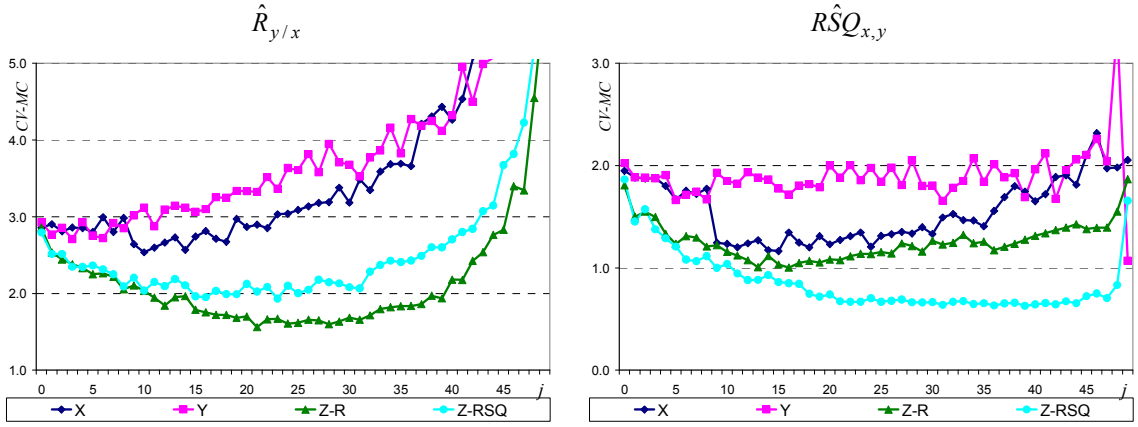
	x_k	y_k	$\tilde{z}_{Rk}$	$\tilde{z}_{RSQk}$
x_k	1.000	0.990	-0.031	0.020
y_k	0.990	1.000	0.108	0.076
$\tilde{z}_{Rk}$	-0.031	0.108	1.000	0.400
$\tilde{z}_{RSQk}$	0.020	0.076	0.400	1.000

3.3 First study: Identifying the take-all units

The first study aims at measuring the efficiency of the linearized variables in identifying the TA units based on the estimator. We started by grouping the 1,000 population units according to a linearized variable. We identified the j ($j = 0,...,49$) largest values as being the j TA units, and the other $1{,}000 - j$ units were assigned to a take-some (TS) stratum. The size $n = 50$ sample was allocated based on $n_{TA} = j$ and $n_{TS} = 50 - j$. For every linearized variable x_k , y_k , $|\tilde{z}_{Rk}|$ and $|\tilde{z}_{RSQk}|$ (the absolute values are to account for their symmetrical distributions), we produced 50 stratification scenarios ($j = 0,...,49$). The charts in Figure 4 show the CV_j^{MC} of the different estimators based on j for every linearized variable. They show that the linearized variables x_k , y_k , $\tilde{z}_{Rk}$ and $\tilde{z}_{RSQk}$ respectively generate the best CV^{MC} for the $\hat{t}_x$, $\hat{t}_y$, $\hat{R}_{y/x}$ and $\hat{RSQ}_{x,y}$ estimators. However, because the x_k and y_k variables are strongly correlated, we note that they generally produce similar results when used as auxiliary variables.

Figure 4: Monte Carlo Coefficient of Variation for Estimators of $\hat{t}_x$, $\hat{t}_y$, $\hat{R}_{y/x}$ and $\hat{RSQ}_{x,y}$ by the Number of Take-all Units for Different Linearized Variables





3.4 Second study: Stratifying the population and allocating the sample

The second study aims at measuring the efficiency of the linearized variables in stratifying the population and allocating the sample based on the estimator. We started by grouping the 1,000 population units according to a linearized variable to create up to three homogeneous strata. For the stratification, we analyzed the four linearized variables x_k, y_k, $\tilde{z}_{Rk}$ and $\tilde{z}_{RSQk}$. We then allocated the sample of $n = 50$ units according to the Neyman method (Särndal et al, 1992) based on the S_h^2 variance of a linearized variable. We also used (independently of the stratification) the four linearized variables x_k, y_k, $\tilde{z}_{Rk}$ and $\tilde{z}_{RSQk}$ for the allocation, with the following variances:

$$S_{xh}^2 = \frac{1}{N-1}\sum\nolimits_{k\in h}\left(x_k - \bar{x}_k\right)^2$$

$$S_{yh}^2 = \frac{1}{N-1}\sum\nolimits_{k\in h}\left(y_k - \bar{y}_k\right)^2$$

$$S_{Rh}^2 = \frac{t_{xh}}{N-1}\sum\nolimits_{k\in h}\left(\tilde{z}_{Rk} - \bar{\tilde{z}}_{Rk}\right)^2$$

$$S_{RSQh}^2 = \frac{t_{xh}}{N-1}\sum\nolimits_{k\in h}\left(\tilde{z}_{RSQk} - \bar{\tilde{z}}_{RSQk}\right)^2$$

Given that $t_x = \sum_h t_{xh}$ and $t_y = \sum_h t_{yh}$, the Neyman allocation under x_k or y_k yields:

$$n_h = \frac{50 N_h S_h}{N_{TS1}S_{TS1} + N_{TS2}S_{TS2} + N_{TS3}S_{TS3}}$$

Given that $R = t_x^{-1}\sum_h t_{xh}R_{y/x,h}$, the Neyman allocation under $\tilde{z}_{Rk}$ yields:

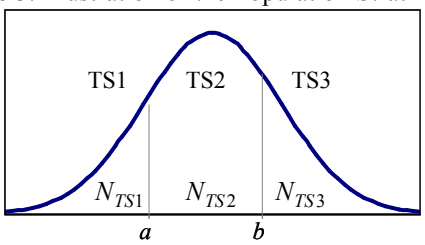
$$n_h = \frac{50 N_h t_{xh} S_{\tilde{z}_R h}}{N_{TS1}t_{xTS1}S_{\tilde{z}_R TS1} + N_{TS2}t_{xTS2}S_{\tilde{z}_R TS2} + N_{TS3}t_{xTS3}S_{\tilde{z}_R TS3}}$$

However, given that $RSQ_{x,y}$ is not a linear combination of correlations at the stratum level, the Neyman allocation under $\tilde{z}_{RSQk}$ cannot be accurately derived. We built an ad hoc allocation similar to the Neyman allocation under $\tilde{z}_{Rk}$:

$$n_h = \frac{50 N_h t_{xh} S_{\tilde{z}_{RSQ}h}}{N_{TS1} t_{xTS1} S_{\tilde{z}_{RSQ}TS1} + N_{TS2} t_{xTS2} S_{\tilde{z}_{RSQ}TS2} + N_{TS3} t_{xTS3} S_{\tilde{z}_{RSQ}TS3}}$$

For each of the 16 stratification and allocation combinations (four each), we introduced two mobile boundaries, *a* and *b*, to divide the population into three strata (including up to two empty strata), as illustrated in figure 5.

Figure 5: Illustration of the Population Stratification



Mobile boundaries *a* and *b* ($a < b$) were placed so that the size of the strata N_{TS1}, N_{TS2} and N_{TS3} would be multiples of 10 (including 0), thereby generating 4,950 groups of strata (numbered $j = 1,...,4950$). For each of these 16 combinations, we obtained 4,950 CV_j^{MC} and kept scenario *j*, with the smallest CV^{MC} to empirically identify the optimal pair of stratification boundaries. Table 6 shows for each of the 16 combinations of stratification and allocation variables the minimum value of the CV^{MC}.

Table 6: Monte Carlo Coefficients of Variation According to the Population Stratification and Sample Allocation Scenarios

Stratification Variable	Allocation Variable	Minimum CV Value Estimated Using the Monte Carlo Method			
		$\hat{t}_x$	$\hat{t}_y$	$\hat{R}_{y/x}$	$R\hat{S}Q_{x,y}$
x_k	x_k	5.87	5.99	1.72	1.07
	y_k	5.99	6.15	1.70	1.07
	$\tilde{z}_{Rk}$	6.19	6.30	1.69	0.84
	$\tilde{z}_{RSQk}$	13.42	13.60	2.06	0.99
y_k	x_k	6.11	5.79	1.89	1.17
	y_k	6.04	5.83	1.91	1.15
	$\tilde{z}_{Rk}$	6.41	6.26	1.83	1.03
	$\tilde{z}_{RSQk}$	16.70	16.84	2.47	1.70
$\tilde{z}_{Rk}$	x_k	14.72	14.81	1.58	1.41
	y_k	14.53	14.67	1.60	1.43
	$\tilde{z}_{Rk}$	20.34	20.22	1.08	0.89
	$\tilde{z}_{RSQk}$	21.76	21.77	1.37	0.86
$\tilde{z}_{RSQk}$	x_k	10.39	10.48	1.65	0.96
	y_k	10.15	10.26	1.68	0.98
	$\tilde{z}_{Rk}$	11.40	11.54	1.26	0.56
	$\tilde{z}_{RSQk}$	17.33	17.56	1.46	0.34

The results show that the linearized variable of an estimator helps stratify the population and allocate the sample in a way that reduces this estimator's CV^{MC}.

4. Application to a survey

Statistics Canada's *Survey of Employment, Payrolls and Hours* (SEPH) is a monthly survey that uses two sources of data: a census of administrative data and an establishment survey. The purpose of the SEPH is to produce estimates of levels and trends for employment, earnings, hours and other related variables, by province and industry. One of the SEPH's key variables of interest is average weekly earnings. For the domain of interest *d*, if x_k and y_k represent the number of employees and their total weekly earnings, respectively, for establishment *k*, the ratio of average weekly earnings is defined as:

$$R_d = \sum_{U_d} y_k \Big/ \sum_{U_d} x_k$$

This variable of interest is strongly correlated with an auxiliary variable, monthly earnings, which is available from the payroll deductions administrative file prepared by the Canada Revenue Agency. If y'_k represents the total monthly earnings of employees in establishment *k*, the ratio of average monthly earnings is defined as follows:

$$R'_d = \sum_{U_d} y'_k \Big/ \sum_{U_d} x_k$$

For the redesign of the SEPH, which is to be implemented by 2009, we used a combination of auxiliary variables to make the sample design as efficient as possible, while at the same time remaining robust for the survey's different variables of interest.

Table 7: Identifying Take-all Units, Stratifying the Population and Allocating the Sample in the Redesigned SEPH

Steps	Strategies
Identifying TA units	. Units representing at least 25% of the relative share of the total of x_k or y'_k in their domain of interest . Units with an extreme value of $\tilde{z}_{R'k}$ in their stratum and with a significant impact on the size of the sample needed to meet a target CV for R'_d at the level of their domain of interest
Stratifying the population	. Initial division of the population by domain of interest (industrial and geographic groupings) . Addition of subdivisions by industry and/or size x_k with a significant impact on the size of sample required to meet a target CV for R'_d at the level of their domain of interest
Allocating the sample	. Neyman allocation algorithm based on the variance of $\tilde{z}_{R'k}$. Minimum size per stratum of 12 units up to a maximum of 40% of the population

We decided to use the x_k variable, associated with the size variable for the survey units, to identify the major TA units and for the stratification because it is very stable and is somewhat correlated with most of the survey's variables of interest. We felt that the use of a derived variable, $\tilde{z}_{R'k}$, in the stratification was not appropriate because this variable is less stable, not well correlated with the other variables of interest, and rather difficult to present to all survey users. However, in order to make the strata more homogeneous for R'_d, we removed the extreme values of $\tilde{z}_{R'k}$ if the net impact was to reduce the total sample size (stratum to TS and units to TA) for a fixed CV. The sample was allocated according to the Neyman algorithm in order to minimize the size of the sample to produce a target CV for R'_d. Finally, we set a minimum size of units per stratum to meet operational constraints. One of the impacts of this minimum size is to ensure good CVs for the other variables of interest in the survey.

5. Conclusion

In the case of a complex statistic or estimator, we proposed linearizing the estimator using the Demnati-Rao method to build an auxiliary variable to identify the take-all units, stratify the population and allocate the sample. We used a simulation to show that this is a truly efficient method. Moreover, it is consistent with Taylor's method of estimating variance through linearization. However, this method requires a derivable (or smooth) statistic (or estimator). Also, the current stratification and allocation algorithms may require certain adjustments when they are applied to a linearized variable because it generally has a distribution that is symmetrical with extreme values at the two extremities. Finally, we believe that the use of a derived variable in some steps of the sample design can raise concerns on the part of data users.

Naturally, the simulated population was created to reduce the correlation among the linearized variables, highlighting the method's benefits. In conclusion, if the complex statistic θ corresponding to a smooth function of totals $t_{y_i} = \sum_U y_{ik}$ correlates with one of the y_{ik}, building an optimal sample design for t_{y_i} could suit $\hat{\theta}$. But if not, using the linearized variable $\tilde{z}_k$ as an auxiliary variable in the sample design would help increase the precision of the estimator $\hat{\theta}$.

Acknowledgements

The author is very grateful to Yves Morin and Yanick Beaucage, as well as the revisers.

References

Cochran W. G. (1977), *Sampling techniques*. 3rd edition, New York: Wiley

Binder D. A. (1996), "Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach", *Survey Methodology*. June 1996, Volume 22, no 1, pp. 17-22, Ottawa, Canada: Statistics Canada

Demnati A. and Rao J. N. K. (2004), "Linearization Variance Estimators for Survey Data", *Survey Methodology*. June 2004, Volume 30, no 1, pp. 17-27, Ottawa, Canada: Statistics Canada

Grondin C., Lavallée P. and Godbout S. (2005). *Current Methodology of the Survey of Employment, Payrolls and Hours*. Internal document, Ottawa, Canada: Statistics Canada

Lavallée P. and Hidiroglou M. (1988), "On the Stratification of Skewed Populations". *Survey Methodology*. June 1988. Volume 14, no 1, Ottawa, Canada: Statistics Canada

Lohr S. (1999), *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press

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

Woodruff R. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate", *Journal of the American Statistical Association*. June 1971, Volume 66, Number 334, Theory and Methods Section