VARIANCE ESTIMATION FOR TWO-PHASE STRATIFIED SAMPLING

David A. Binder, Colin Babyak, Marie Brodeur, Michel Hidiroglou and Wisner Jocelyn (Statistics Canada)
Business Survey Methods Division, Statistics Canada, Ottawa, ON, Canada K1A 0T6

Key Words: Double Expansion Estimator, Reweighted Expansion Estimator, Stratification, Linearization, Retail Commodity Survey.

1. Introduction

Two-phase sampling was first introduced by Neyman (1938) and was better known in the literature as Double Sampling. Cochran (1963) devoted a whole chapter of his book to the subject but at this point, the formulas were developed only for simple random sampling at the first phase and stratified random sampling at the second phase. Särndal, Swensson and Wretman (1992), derived variance estimators for the Double Expansion Estimator for two-phase sampling. Hidiroglou and Särndal (1996) have extended this work by providing a unified theory for two-phase sampling with auxiliary information. These papers do not provide, explicitly, variance estimation methods for a two-phase stratified sampling design when both phases are stratified.

In this paper, we consider the variance estimation problem for two different stratified estimators: the Double Expansion Estimator and the Reweighted Expansion Estimator suggested by Kott and Stukel (1994). The latter is used more often when non-response is treated as a second phase of sampling. They have derived a Jackknife variance for this estimator. We also consider a combined ratio version for each of the estimators. Some of these estimation methods will be applied to the Canadian Retail Trade Survey.

The paper is organized as follows. After a brief description of the Retail Commodity Survey (RCS) at the beginning of section 2, we will present the estimators considered in the rest of the section. Variance estimators for the point estimators considered in section 2 will be discussed in section 3. In section 4, we will present results of the simulation conducted to compare the estimators. We will conclude with some recommendations in section 5.

2.RCS and Two-Phase Stratified estimators

2.1 Retail Commodity Survey

Statistics Canada launched a new Retail Commodity Survey in January 1997. The survey, whose purpose is to gather estimates of retail trade by commodity, is based on a two-phase sample design. This design was chosen to reduce collection costs by taking as the first phase sample the Canadian Monthly Retail Trade Survey (MRTS), which is a stratified sample that has been in place since 1988. Information from the first phase sample is used in all the RCS design steps in order to maximize the efficiency of the design. RCS is an independently stratified sub-sample of the first phase sample. A model based on MRTS estimates of sales by Gross Business Income was used for stratification. Allocation was done using a multivariate algorithm based on the MRTS estimates of sales by commodity (Jocelyn and Brodeur (1996)).

2.2 Notation

The parameter of interest to be estimated is the total sales $Y = \sum_{i=1}^{N} y_i$ where y_i is the sales for unit i. The problem we wish to look at is the variance of the two-phase estimator of a total. We denote the population values by y_i, $i = 1,...,N$. In the first phase of sampling, we take a simple random sample independently from each of the H strata from a population U. The resulting first-phase sample is $s_1 = \bigcup_{h=1}^{H} s_{1h}$, where s_{1h} is the set of units drawn from the population in the hth stratum U_h at the first-phase. We denote by a_{ih} the indicator variable, taking the value 1 when unit i is in stratum h and 0 otherwise. We define $N_h = \sum_{i=1}^{N} a_{ih}$ and we note that $\sum_{h=1}^{H} a_{ih} = 1$ since each unit in the population belongs to exactly one first-phase stratum. We denote by z_i the indicator variable, taking the value 1 when unit i is in the first-phase sample and 0 otherwise. We see, therefore, that the sample size from the h th stratum is $n_h = \sum_{i=1}^{N} z_i a_{ih}$.

The first-phase sample s_1 is further stratified into G strata (where this stratification can be independent of the

first-phase stratification), $s_1 = \bigcup_{g=1}^{G} s_{1g}$. We take a simple random sample of units selected in the first phase, independently from each of the G strata, we denote by $a_{ig}^{(2)}$ the indicator variable, taking the value 1 when unit i is in the second-phase stratum g and 0 otherwise. A simple random sample s_{2g} of size m_g is drawn from s_{1g} , where s_{1g} consists of M_g units. We denote by $z_i^{(2)}$ the indicator variable, taking the value 1 when unit i is in the second-phase sample and 0 otherwise. Note that units selected in the second phase must have been selected in the first phase so that $z_i^{(2)} = z_i\, z_i^{(2)}$. We have $M_g = \sum_{i=1}^{N} z_i a_{ig}^{(2)}$, the number of units selected in the first phase that are in second-phase g stratum. We see, therefore, that the sample size from the g th stratum is $m_g = \sum_{i=1}^{N} z_i^{(2)} a_{ig}^{(2)}$.

2.3 Estimators

The usual *Double Expansion Estimator* is given by

$$\hat{Y}_{DE} = \sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{N_h}{n_h}\frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i$$

The *Reweighted Expansion Estimator*, uses the first-phase estimated estimate population size for group stratum g as auxiliary information. It is given by

$$\hat{Y}_{RW} = \sum_{g=1}^{G} \hat{N}_g^{(1)} \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}}$$

where

$$\hat{N}_g^{(1)} = \sum_{i=1}^{N}\sum_{h=1}^{H} z_i\, a_{ih}\, a_{ig}^{(2)} \frac{N_h}{n_h}$$

$$\hat{N}_g^{(2)} = \sum_{i=1}^{N}\sum_{h=1}^{H} z_i^{(2)}\, a_{ih}\, a_{ig}^{(2)} \frac{N_h}{n_h}\frac{M_g}{m_g}$$

and

$$\hat{Y}_g^{(2)} = \sum_{i=1}^{N}\sum_{h=1}^{H} z_i^{(2)}\, a_{ih}\, a_{ig}^{(2)} \frac{N_h}{n_h}\frac{M_g}{m_g} y_i$$

2.4 General Two-Phase Linear Estimator

After linearization, estimates which use an auxiliary variable can usually be expressed as linear combinations of x_i and y_i . We consider in this section, a general linear estimate from a two-phase sample which includes components from both the first-phase and second-phase data. Assume that an auxiliary variable x_i is available for all units belonging to the first-phase sample s_1 , and that y_i is only available from the second-phase sample s_2. We consider,

$$\hat{T} = \sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{g=1}^{G} \frac{N_h}{n_h}\frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i + \sum_{i=1}^{N}\sum_{h=1}^{H} \frac{N_h}{n_h} z_i a_{ih} x_i$$

which is an unbiased estimate of

$$T = \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} x_i .$$

Note that if $x_i = 0$ for all i , then $\hat{T}$ is the usual *Double Expansion Estimator* . A *Combined Ratio Estimator* could be derived from $\hat{T}$ and is effectively given by

$$\hat{Y}_{DERAT} = \hat{R}\, \hat{X}^{(1)}$$

where

$$\hat{R} = \left[\frac{\sum_{i=1}^{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} y_i}{\sum_{i=1}^{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} x_i} \right] = \frac{\hat{Y}_{DE}}{\hat{X}_{DE}}$$

and

$$\hat{X}^{(1)} = \sum_{i=1}^{N} \sum_{h=1}^{H} \frac{N_h}{n_h} z_i a_{ih} x_i$$

Similarly, a Reweighted Ratio Expansion Estimator could be defined as

$$\hat{Y}_{RWRAT} = \frac{\hat{Y}_{RW}}{\hat{X}_{RW}} \hat{X}^{(1)} = \hat{R}_{RW} \hat{X}^{(1)}$$

Where $\hat{Y}_{RW}$ is as defined in section 2.3 and $\hat{X}_{RW}$ is defined similarly.

3. Variance Estimators

The variance of $\hat{T}$, as well as an estimator of that variance is derived by noting that the population variance is given by

$$V(\hat{T}) = E_z V(\hat{T}|z) + V_z E(\hat{T}|z)$$

where each component represents the first-phase and second-phase contributions. The corresponding variance estimator is given by

$$\hat{V}(\hat{T}) = \hat{V}_1(\hat{T}) + \hat{V}_2(\hat{T})$$

with

$$\hat{V}_1(\hat{T}) = \sum_{g=1}^{G} M_g^2 (1 - f_g^{(2)}) \frac{s_{2g}^2}{m_g}$$

and

$$\hat{V}_2(\hat{T}) = \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h^2 (1 - f_h) M_g^2 (1 - f_g^{(2)})}{n_h^2 (n_h - 1)} \frac{s_{1(h)g}^2}{m_g} + \sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

where

(i) s_{2g}^2 is the within- g sample variance of $\sum_{h=1}^{H} \frac{N_h}{n_h} a_{ih} y_i$,

(ii) $s_{1(h)g}^2$ is the within-g sample variance of $a_{ih}(x_i + y_i)$

(iii)

$$s_h^2 = \frac{1}{(n_h - 1)} \left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} (x_i + y_i)^2 \right] - \frac{\left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} (x_i + y_i) \right]^2}{n_h (n_h - 1)}.$$

(iv) $f_g^{(2)} = m_g / M_g \quad f_h = n_h / N_h$.

A detailed derivation of the variance formula and the corresponding variance estimator is available on request. This variance can also be obtained from Särndal, Swensson, and Wetman (1992, pg. 348 equation (9.3.7)). Variance estimators are derived using linearization methods suggested by Binder (1996). The variance estimator of the Double Expansion Estimator can be obtained by substituting $x_i = 0$ in the variance equation given previously. The variance estimator of $\hat{Y}_{DERAT}$ is also obtained from the same equation. For this case, s_{2g}^2 is the within-g sample variance of $\sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\hat{X}^{(1)}}{\hat{X}} (y_i - \hat{R} x_i) a_{ih}$, $s_{1(h)g}^2$ is the within-g sample variance of $a_{ih} \left[\frac{\hat{X}^{(1)}}{\hat{X}} (y_i - \hat{R} x_i) + \hat{R} x_i \right]$,

and

$$s_h^2 = \frac{1}{(n_h-1)} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} \left(\frac{\hat{X}^{(1)}}{\hat{X}} (y_i - \hat{R}x_i) + \hat{R}x_i \right)^2 - \frac{\left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} \left(\frac{\hat{X}^{(1)}}{\hat{X}} (y_i - \hat{R}x_i) + \hat{R}x_i \right) \right]^2}{n_h(n_h-1)}.$$

Similarly, the variance estimator of the Reweighted Expansion Estimator $\hat{Y}_{RW}$ defined in section 2.3 , is obtained from the general formula of $\hat{V}(\hat{T})$.

Now, s_{2g}^2 is the within-g sample variance of $\sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\hat{N}_g^{(1)}}{\hat{N}_g^{(2)}} a_{ih} \left(y_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} \right)$, and $s_{1(h)g}^2$ is the within-g sample variance of $a_{ih} \left(\frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} + \frac{\hat{N}_g^{(1)}}{\hat{N}_g^{(2)}} \left(y_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} \right) \right)$,

$$s_h^2 = \frac{1}{(n_h-1)} \left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} v_i^2 \right] - \frac{\left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} v_i \right]^2}{n_h(n_h-1)},$$

with $v_i = \left[\frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} + \frac{\hat{N}_g^{(1)}}{\hat{N}_g^{(2)}} \left(y_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} \right) \right]$.

The variance estimator of the Reweighted Ratio Estimator $\hat{Y}_{RWRAT}$ given in section 2.4 , is also obtained from the formula of $\hat{V}(\hat{T})$. Here, s_{2g}^2 is the within-g sample variance of

$$\sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\hat{N}_g^{(1)}}{\hat{N}_g^{(2)}} \frac{\hat{X}^{(1)}}{\hat{X}_{RW}} \left[\left(y_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}} \right) - \hat{R}_{RW} \left(x_i - \frac{\hat{X}_g^{(2)}}{\hat{N}_g^{(2)}} \right) \right] a_{ih},$$

$s_{1(h)g}^2$ the within-g sample variance of $a_{ih} u_i$ with

$$u_i = \left(\frac{\hat{X}^{(1)}}{\hat{X}_{RW}} \right) \left[\hat{T}_g a_{ig}^{(2)} + \frac{\hat{N}_g^{(1)}}{\hat{N}_g^{(2)}} \left(y_i^r - \hat{R}_{RW} x_i^r \right) \right] + \hat{R}_{RW} x_i$$

with $\hat{T}_g = \sum_{g=1}^{G} \frac{\hat{Y}_g^{(2)} - \hat{R}_{RW} \hat{X}_g^{(2)}}{\hat{N}_g^{(2)}}$,

$y_i^r = y_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}}$ and $x_i^r = x_i - \frac{\hat{Y}_g^{(2)}}{\hat{N}_g^{(2)}}$

where

$$s_h^2 = \frac{1}{(n_h-1)} \left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} u_i^2 \right] - \frac{\left[\sum_{i=1}^{N} \sum_{g=1}^{G} \frac{M_g}{m_g} z_i^{(2)} a_{ih} a_{ig}^{(2)} u_i \right]^2}{n_h(n_h-1)}$$

4. Results of the simulation

We carried out a simulation to compare the properties of the variance estimators. The simulation population was made of 408 household retailers in three Canadian Provinces. Stratification and sample size are similar to those used in the real survey. The size of the first phase sample exceeded 200 units, while the size of the second phase sample was close to 140 units. We drew 10,000 samples and we computed various statistics. We looked at both unconditional and conditional results. Although we are mostly interested in commodity data for the real survey, since none were available for the simulation, we used a linear regression model to simulate sales. Therefore, the auxiliary variable used for the simulation was the Gross Business Income (GBI) and the variable of interest was simulated Sales.

Table 1 on the next page summarizes the unconditional results obtained from the simulation. The stratified full first phase estimator was included into this table for comparisons' purposes. Relative bias is close to zero for all the estimators including the ratio estimators. The efficiency of the Mean Square Error (MSE) of the estimators, which is the MSE of each estimator divided by the MSE of the full first phase estimator, is not different from one estimator to the other. The relative bias of the variance estimators and the coverage for all estimators are very similar. Generally speaking, the unconditional results for all the estimators are fairly similar. No one estimator stands out.

Table 1: UNCONDITIONAL RESULTS OF THE MONTE CARLO STUDY FOR THE ESTIMATORS AND THEIR VARIANCE ESTIMATORS

Estimator	Relative Bias(%)	MSE Efficiciency (Comp to one phase)	Relative Bias of the Variance estimators (%)	Coverage rate (95%)
Full one phase	0.0	1.000	0.3	95.0
Double Expansion	0.0	0.680	0.4	95.1
Ratio Double Exp	0.1	0.596	-1.9	95.0
Reweighted Exp	0.2	0.630	0.1	95.0
Ratio Reweighted	0.0	0.584	-13.5	95.2

CHART 1

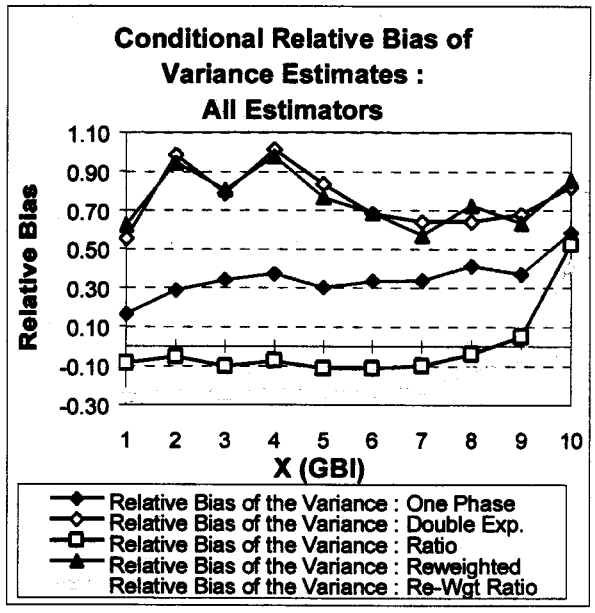
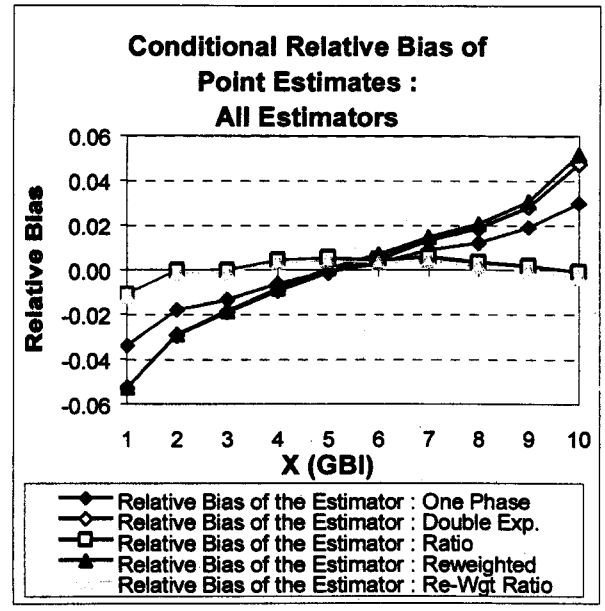


CHART 2



We now look at some conditional results. The 10,000 samples were divided in ten groups of one thousand units each. Group membership of the units was determined using the estimated totals of the auxiliary variable. This procedure is similar to the one used by Royall and Cumberland (1981). Charts 1 and 2 show respectively the conditional relative bias for the variance estimates and the point estimates. We notice that the variance estimators of the ratio estimators are closer and more stable around the zero bias line. The relative biases of the point estimates are essentially close to zero for the ratio estimators. We observe similar results for the coverages. These results suggest that the ratio estimators are doing better than the non ratio estimators. This supports and validates the use of auxiliary information in that particular survey.

5. Conclusion

The small biases observed for the variance estimators suggest that the Taylor linearization method is performing well. The conditional results clearly demonstrate the superiority of the ratio estimators over the non ratio estimators. However, they do not discriminate between the two ratio estimators considered. We recommend the use of the Double Expansion Ratio estimator because of the simplicity of the variance formula.

Acknowledgements

The authors would like to thank Eric Rancourt and Alain Théberge for their constructive suggestions. We also thank Carole Jean-Marie who typed the first draft of the paper.

6. Bibliography

Binder, D.A. (1996). Linearization Methods for Single Phase and Two Phase Samples. A Cookbook Approach. *Survey Methodology*, **22**, 17 - 22.

Cochran, W.G (1963). Sampling Techniques. John Wiley.

Hidiroglou, M.A., and Särndal C.-E. (1995). Use of Auxiliary Information for Two-phase Sampling. Proceedings of the Section on Survey Research Methods, *Annual American Statistical Association,* 873-878.

Jocelyn, W., Brodeur, M. (1996). Méthodes de répartition multivariées pour l'échantillonnage à deux phases: Application à l'enquête trimestrielle sur les marchandises. *Recueil des communications des XXVIIIe Journées de Statistiques de l'ASU, 433-436.*

Kott, S., P, Stukel, M., D (1994). Can the Jackknife be used with a two-phase sample? Submitted to *Survey Methodology.*

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association,* **33**, *101-116.*

Royall, R., M., Cumberland, W., G. (1981). An Empirical Study of the Ratio Estimator and Estimators of its Variance. *Journal of the American Statistical Association*, **76**, 66-88.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. Springer-Verlag.