

Methodology Branch

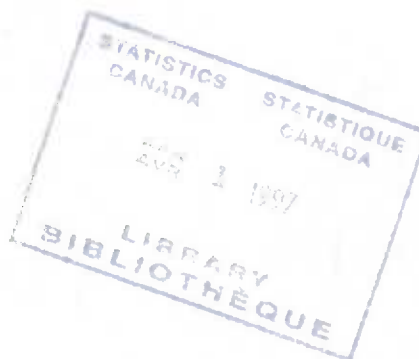
Household Survey  
Methods Division

Direction de la méthodologie

Division des méthodes  
d'enquêtes des ménages



WORKING PAPER  
METHODOLOGY BRANCH



**JACKKNIFE VARIANCE ESTIMATION UNDER TWO-PHASE SAMPLING:  
AN EMPIRICAL INVESTIGATION**

HSMD - 97 - 001E

D.M. Stukel and P.S. Kott

Household Survey Methods Division  
Statistics Canada

February 1997

---

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada



# JACKKNIFE VARIANCE ESTIMATION UNDER TWO-PHASE SAMPLING: AN EMPIRICAL INVESTIGATION

Diana M. Stukel and Phillip S. Kott<sup>1</sup>

## ABSTRACT

The jackknife variance estimator has been shown to have desirable properties when used with smooth estimators based on stratified multi-stage designs. However, its use in the context of two-phase sampling is relatively unexplored. This paper focuses on a particular two-phase design: at the first phase, a stratified one-stage take-all cluster sample is drawn using simple random sampling with replacement. The sampled first phase units are restratified, and second phase units are drawn using stratified simple random sampling without replacement. Kott (1995) has suggested that under this design, the jackknife variance will behave reasonably well for the "reweighted expansion estimator", but not for the more commonly used "double expansion estimator". In this paper, we also discuss simple poststratified versions of these estimators and their corresponding jackknife variances; poststratified estimators are appealing since it is common for many surveys to benchmark their final weights to known external control totals. We give the results of a simulation study in which we investigate the finite sample frequentist properties of all of the point estimators given above, as well as their corresponding jackknife variance estimators. The parameters of interest considered are population totals and ratios.

**KEY WORDS:** Double Expansion Estimator, Poststratification, Reweighted Expansion Estimator, Stratification

---

<sup>1</sup> Diana M. Stukel is Senior Research Methodologist, Household Survey Methods Division, Statistics Canada and Phillip S. Kott is Chief Research Statistician, National Agricultural Statistics Service, US Department of Agriculture, 3201 Old Lee Highway, Fairfax, Virginia, 22030



# ESTIMATION DE LA VARIANCE SELON LA MÉTHODE JACKKNIFE DANS UN ÉCHANTILLONNAGE À DEUX DEGRÉS : ÉTUDE EMPIRIQUE

Diana M. Stukel et Phillip S. Kott<sup>1</sup>

## RÉSUMÉ

On sait que l'estimateur de la variance jackknife présente des propriétés intéressantes lorsqu'on l'utilise avec des estimateurs continus dans un plan d'échantillonnage stratifié à plusieurs degrés. Toutefois, on connaît relativement peu de choses sur son utilisation dans un plan d'échantillonnage à deux degrés. Le présent document examine un plan d'échantillonnage à deux degrés particulier : en premier lieu, on crée un échantillon stratifié à un degré par simple tirage au sort non exhaustif de grappes complètes. Les unités prélevées lors de la première étape subissent une restratification, puis on procède à un second échantillonnage aléatoire simple stratifié, mais exhaustif celui-là. Kott (1995) croit qu'avec un tel plan d'échantillonnage, l'estimation de la variance selon la méthode jackknife donnera raisonnablement de bons résultats pour «l'estimateur avec facteur d'extension repondéré», mais pas pour «l'estimateur avec facteur d'extension double», d'usage plus courant. Le document traite aussi des simples variantes des estimateurs obtenues par stratification a posteriori et de la variance correspondante selon la méthode jackknife; les estimateurs de stratification a posteriori sont attrayants, car dans maintes enquêtes, il est fréquent d'aligner les poids finals sur des totalisations de contrôle indépendantes connues. Les auteurs présentent les résultats d'une simulation qui les a aidés à examiner les propriétés fréquentistes de l'échantillon fini composé des estimateurs ponctuels mentionnés plus haut, et des estimateurs jackknife de la variance correspondants. Les paramètres d'intérêt considérés sont des totaux et des ratios de la population.

**MOTS CLÉS :** estimateur à facteur d'extension double, stratification a posteriori, estimateur à facteur d'extension repondéré, stratification.

---

<sup>1</sup> Diana M. Stukel est chercheur principal en méthodologie à la Sous-division des méthodes d'enquêtes-ménages de Statistique Canada. Phillip S. Kott est statisticien de recherche principal au National Agricultural Statistics Service, US Department of Agriculture, 3201 Old Lee Highway, Fairfax, Virginie, 22030.



## 1. INTRODUCTION

Krewski and Rao (1981) and Rao and Wu (1985) explore the design based properties of the jackknife variance estimator under a stratified multi-stage design using with replacement sampling at the first stage. Their results, although fairly general, cannot be directly applied to multi-phase sampling designs.

In this paper, we consider a specific two-phase design: at the first phase, primary sampling units (PSUs) are drawn within each first phase stratum using Simple Random Sampling (SRS) with replacement (WR). Then, all units within the sampled PSUs are selected, resulting in a one-stage take-all cluster design. At the second phase, the entire first phase sample is restratified, and second phase units are drawn according to SRS without replacement (WOR) from each of the second phase strata. Several surveys at Statistics Canada use designs very similar to this one, such as the International Adult Literacy Survey on the social side, and the Quarterly Retail Commodity Survey on the business side.

To estimate a total in this context, it is common to use the Double Expansion Estimator (or  $\pi^*$ -Estimator, in the parlance of Särndal, Swensson and Wretman (1992)). For this estimator, each of the subsampled units is multiplied by the product of its inverse sampling rates at each phase and then summed. Although the Double Expansion Estimator is more easily located in text books, an estimator that is more commonly used in practice is the Reweighted Expansion Estimator, especially when unit nonresponse is treated as a second phase of sampling. (See Rao and Shao (1992)). Although both of these estimators behave well from the standpoint of point estimation, Kott (1995) has suggested that under the above design, the jackknife variance will behave reasonably well for the Reweighted Expansion Estimator but not for the Double Expansion Estimator. The investigation of that conjecture for estimating totals and ratios is the focus of this paper.

The organization of this paper is as follows: Section 2 introduces all of the point estimators and Section 3 gives their corresponding jackknife variance estimators. In Section 4, the results of a simulation study are given, in which the finite sample properties of the point estimators and their corresponding jackknife variance estimators are investigated. Finally, in Section 5, some concluding remarks are made.



## 2. THE POINT ESTIMATORS

In survey sampling, one is often interested in estimating parameters of a finite population, such as totals and ratios of some characteristic. In this section, for the sake of simplicity, we confine ourselves to the case of totals, since all the estimators considered can be easily extended to the case of ratios. However, we revisit the case of ratios in the Monte Carlo simulation study in Section 4. Therefore, suppose the parameter of interest is the population total,  $T = \sum_{i \in U} y_i$ , where  $y_i$  is the value of interest for unit  $i$  and  $U$  is the set of all finite population units. Suppose, further, that the two-phase design described in the introduction is assumed.

If the entire first phase sample is available, one could use a Full First Phase Estimator (FFPE) given here in terms of two-phase notation as:

$$t_1 = \sum_{g=1}^G \sum_{i \in S_g} w_i y_i \quad (1)$$

where  $g (=1, \dots, G)$  is the index for the second phase strata,  $S_g$  is the set of sampled first phase units that fall in second phase stratum  $g$ , and  $w_i$  is the first phase weight for sampled unit  $i$ .

On the other hand, if only second phase units are available, one could use the Double Expansion Estimator (DEE) or  $\pi^*$ -estimator, given by:

$$t_2 = \sum_{g=1}^G \sum_{i \in s_g} \frac{M_g}{m_g} w_i y_i \quad (2)$$

where  $s_g$  is the set of sampled second phase units in second phase stratum  $g$ ,  $M_g$  is the number of sampled first phase units in second phase stratum  $g$ , and  $m_g$  is the number of sampled second phase units in second phase stratum  $g$ . Thus,  $M_g/m_g$  is the second phase inverse sampling rate within second phase stratum  $g$ .



Kott (1995) has suggested that, in terms of jackknife variance estimation, a better choice of estimator to use would be the Reweighted Expansion Estimator (REE), given by:

$$t_3 = \sum_{g=1}^G \left( \sum_{i \in S_g} w_i \frac{\sum_{i \in S_g} \frac{M_g}{m_g} w_i y_i}{\sum_{i \in S_g} \frac{M_g}{m_g} w_i} \right) = \sum_{g=1}^G \sum_{i \in S_g} w_{ig}^* y_i \quad (3)$$

where

$$w_{ig}^* = w_i \frac{\sum_{i \in S_g} w_i}{\sum_{i \in S_g} w_i}; \quad i \in S_g. \quad (4)$$

It is the formulation on the right hand side of equation (3), in terms of  $w_{ig}^*$ , that gives the Reweighted Expansion Estimator its name. Notice that the second phase inverse inclusion probabilities,  $M_g/m_g$ , cancel out, so that this formulation is reminiscent of a "reweighting" within classes that one would use if one had unit nonresponse and were treating it as a second phase of sampling.

Since it is common for many household and business surveys to benchmark their final weights to known external control totals, it is of interest to consider a simple poststratified version of the Reweighted Expansion Estimator (SP-REE) as well as a simple poststratified version of the Double Expansion Estimator (SP-DEE). The former is given by:

$$t_3(SP) = \sum_p (N_p / \hat{N}_p) \sum_g \sum_{i \in S_{pg}} w_{ig}^* y_i \quad (5)$$

where  $\hat{N}_p = \sum_g \sum_{i \in S_{pg}} w_{ig}^*$

and where  $p$  is the index for the poststrata (different from  $g$  which is the index for second phase strata). Here,  $N_p$  represents the known external count for poststratum  $p$ . In addition,  $S_{pg}$  represents that part of the second phase sample which falls into poststratum  $p$  and second phase stratum  $g$ , and  $w_{ig}^*$  is given in equation (4). The Simple Poststratified Double Expansion Estimator is given by:



$$t_2(SP) = \sum_p (N_p / \hat{N}_p^*) \sum_g \sum_{i \in S_{pg}} w_{ig} y_i \quad (6)$$

where  $\hat{N}_p^* = \sum_g \sum_{i \in S_{pg}} w_{ig}$

and where  $w_{ig} = w_i(M_g/m_g)$ ;  $i \in S_g$ . Note that if poststrata are defined to be the same as second phase strata, then  $t_2(SP) = t_3(SP)$ . Finally, a simple poststratified version of the Full First Phase Estimator (SP-FFPE) is given by:

$$t_1(SP) = \sum_p (N_p / \hat{N}_p^{**}) \sum_{i \in S_p} w_i y_i \quad (7)$$

where  $\hat{N}_p^{**} = \sum_{i \in S_p} w_i$

and where  $S_p$  represents that part of the first phase sample which falls into poststratum  $p$ .

### 3. THE JACKKNIFE VARIANCE ESTIMATORS

Following Rust (1985), the jackknife variance estimator,  $v_{Jf}$ ; ( $f=1$  or 2 or 3), is defined here as:

$$v_{Jf} = \sum_{h=1}^H \frac{n_h - 1}{n_h} \sum_{j \in F_h} (t_f(hj) - t_f)^2, \quad f=1,2,3 \quad (8)$$

where  $h (=1, \dots, H)$  is the index for the first phase strata,  $n_h$  is the number of PSUs selected in stratum  $h$  at the first phase, and  $F_h$  the set of sampled PSUs in stratum  $h$ . Finally,  $t_f(hj)$  is called the replicate estimator, and will be defined next.

For the Reweighted Expansion Estimator ( $f=3$ ), the replicate estimator is formed by recalculating the Reweighted Expansion Estimator,  $t_3$ , after removing the  $j^{th}$  PSU from the  $h^{th}$  stratum by reweighting to reflect the removal. That is,

$$t_3(hj) = \sum_{g=1}^G \left( \sum_{i \in S_g} w_{h/j'i}(hj) \frac{\sum_{i \in S_g} w_{h/j'i}(hj) y_i}{\sum_{i \in S_g} w_{h/j'i}(hj)} \right) \quad (9)$$



where

$$w_{h'j'i}(hj) = \begin{cases} 0 & \text{if } h'=h, j'=j \\ \frac{n_{h'}}{n_{h'}-1} w_{h'j'i} & \text{if } h'=h, j' \neq j \\ w_{h'j'i} & \text{if } h' \neq h \end{cases}$$

and where  $w_{h'j'i} (\equiv w_i)$  is the first phase weight for individual  $i$  in PSU  $j'$  and first phase stratum  $h'$ .

For the Double Expansion Estimator ( $f=2$ ), in defining the replicate estimator, it is not clear whether it is better to reweight only the first phase weights,  $w_{h'j'i}(hj)$ , as in Variant 1 given below, or both the first phase weights and the second phase weights  $M_{gj}/m_{gj}$ , as in Variant 2 given below.

Variant 1 is given by:

$$t_2(hj) = \sum_{g=1}^G \sum_{i \in s_g} \frac{M_g}{m_g} w_{h'j'i}(hj) y_i \quad (10)$$

and Variant 2 is given by:

$$t_2^*(hj) = \sum_{g=1}^G \sum_{i \in s_g} \frac{M_{gj}}{m_{gj}} w_{h'j'i}(hj) y_i \quad (11)$$

where

$M_{gj} = M_g$  minus the number of selected first phase individuals falling in second phase stratum  $g$  and PSU  $j$

and

$m_{gj} = m_g$  minus the number of selected second phase individuals falling in second phase stratum  $g$  and PSU  $j$ .

As we shall see, neither produces a jackknife variance which tracks the true mean squared error (MSE) well.



The replicate estimator for the Full First Phase Estimator is straightforward and is given by:

$$t_1(hj) = \sum_{g=1}^G \sum_{i \in S_g} w_{h/j/i}(hj) y_i \quad (12)$$

It is also possible to define jackknife variance estimators for  $t_1(SP)$ ,  $t_2(SP)$ , and  $t_3(SP)$  in an analogous way. The only noteworthy difference is that the poststratification must be recalculated after each PSU removal, to ensure a proper jackknife.

## 4. A MONTE CARLO SIMULATION STUDY

### 4.1 Design of the Study

The main contention of this paper is that the jackknife based on the Reweighted Expansion Estimator should behave much better than that based on the Double Expansion Estimator. In order to see if this was the case, we undertook a Monte Carlo simulation study in which we investigated the finite sample frequentist properties of both jackknife variance estimators.

December 1990 Canadian Labour Force Survey (LFS) sample data for the province of Newfoundland was used to simulate a finite population, from which repeated samples were drawn. The LFS is the largest ongoing household sample survey conducted by Statistics Canada. Monthly data relating to the labour market is collected using a complex multi-stage sampling design with several levels of stratification. The details of the design of the survey prior to the 1991 redesign can be found in Singh, Drew, Gambino and Mayda (1990). In general, provinces are stratified into "economic regions", which are large areas of similar economic structure; Newfoundland has four such economic regions. The economic regions are further substratified into lower level substrata. Now, the lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 primary sampling units (PSUs), which was an insufficient number from which to sample for the purposes of the simulation. Thus, the 45 strata were collapsed down to 18, each containing between 6 and 18 PSUs. In collapsing the strata, economic regions were kept intact, as were the Census Metropolitan Areas (CMAs) of St. John's and Cornerbrook.

For the Monte Carlo study,  $R = 4000$  samples, were drawn from the Newfoundland "population"



(consisting of 9152 individuals), according to the following two-phase design: within each first phase stratum, two PSUs were selected at the first stage using simple random sampling (SRS) with replacement (WR), yielding a total of 36 PSUs. All households within selected first phase PSUs (as well as individuals within those households) were selected, resulting in a one-stage take-all cluster sample. At the second phase, all selected first phase units (individuals) were restratified according to five age categories ( $\leq 14$ , 15-24, 25-44, 45-64,  $\geq 65$ ), and second phase units (individuals) were drawn according to SRS without replacement (WOR) sampling within each of the five second phase strata. We varied the second phase sample size to take on values  $m_g = 5, 10, 20$ , and  $50$ , yielding overall second phase sample sizes of  $m = 25, 50, 100$ , and  $250$ . We even drew 4000 full first phase samples ( $m_g = M_g$ ), in order to calculate full first phase estimators for the sake of comparison.

We considered two parameters of interest:  $T$ , the total number of employed, and  $T/Z$ , the employment rate. Here,  $T = \sum_{i \in U} y_i = \sum_{i=1}^{9152} y_i$  where  $y_i = 1$  if individual  $i$  was employed; 0 else, and

$Z = \sum_{i \in U} z_i = \sum_{i=1}^{9152} z_i$  where  $z_i = 1$  if individual  $i$  was in the labour force (employed or unemployed); 0

else. For each of the  $R = 4000$  samples, we calculated the Reweighted Expansion Estimator (REE),  $t_3$ , given by equation (3) and the Double Expansion Estimator (DEE),  $t_2$ , given by equation (2), as well as their poststratified counterparts, given in equations (5) and (6), respectively. For the poststratified versions, we took the poststrata to be the four economic regions of Newfoundland; these economic regions are aggregates of first phase strata. Given that one can never improve on an estimator based on the full first phase sample, for the sake of comparison we also considered the full first phase sample estimator (FFPE),  $t_1$ , given in equation (1), as well as its poststratified counterpart, given in equation (7). Note that, although all of the estimators so far have defined for totals (applicable for total number of employed), it is easy to define them for ratios (applicable for employment rate) by estimating both the numerator and denominator parts separately as totals.

For each of the  $R = 4000$  second phase samples, we calculated the jackknife variance corresponding to the Reweighted Expansion Estimator and the Double Expansion Estimator, given by equation (8) with  $f = 3$  and  $f = 2$  respectively. In the case of the Double Expansion Estimator, we



attempted both variants defined in equations (10) and (11). We also attempted jackknife variances for simple poststratified versions each of the above (SP-REE, SP-DEE (variant 1) and SP-DEE (variant 2)). For each of the  $R = 4000$  *first* phase samples, we calculated the jackknife variance corresponding to the full first phase estimator, given by equation (8) with  $f = 1$ . We also attempted the jackknife variance of the simple poststratified full first phase estimator (SP-FFPE).

For all of the above estimators and their corresponding jackknife variances, a number of frequentist properties were investigated. These are given below, expressed only in terms of estimates of the total number of employed, for the sake of simplicity.

(A) The percent relative bias of the estimated number of employed with respect to the population value is estimated by:

$$\frac{E_M(t^*) - T}{T} * 100 \quad (13)$$

where

$$E_M(t^*) = \frac{1}{4000} \sum_{r=1}^{4000} t_r^*$$

is the Monte Carlo expectation of the point estimator  $t^*$  taken over the 4000 samples. Here  $t^*$  can be either  $t_1$ ,  $t_1(SP)$ ,  $t_2$ ,  $t_2(SP)$ ,  $t_3$  or  $t_3(SP)$ , and  $t_r^*$  is the value of  $t^*$  for sample  $r$ .

(B) The percent relative bias of the jackknife variance estimator with respect to the estimated true mean squared error is estimated by:

$$\frac{(E_M(v_{Jf}(t^*)) - MSE_{true})}{MSE_{true}} * 100 \quad (14)$$

where

$$E_M(v_{Jf}(t^*)) = \frac{1}{4000} \sum_{r=1}^{4000} v_{Jf_r}(t^*)$$



and

$$MSE_{true} = \frac{1}{4000} \sum_{r=1}^{4000} (t_r^* - T)^2$$

and  $v_{jf_r}(t^*)$  is the value of  $v_{jf_r}(t^*)$  for sample  $r$ .

(C) The percent coefficient of variation of the jackknife variance with respect to the estimated true MSE is estimated by:

$$\sqrt{\frac{\frac{1}{4000} \sum_{r=1}^{4000} (v_{jf_r}(t^*) - MSE_{true})^2}{MSE_{true}}} * 100, \quad (15)$$

i.e., the root mean squared error of the variance estimator divided by the estimated true MSE, expressed as a percentage.

(D) The percent coverage of a confidence interval having a nominal value or 95% is estimated by:

$$\left( \frac{1}{4000} \sum_{r=1}^{4000} I_r \right) * 100 \quad (16)$$

where  $I_r = 1$  if  $T$  is inside the interval  $t_r^* \pm z_{\alpha/2} [v_{jf_r}(t^*)]^{1/2}$  and  $I_r = 0$  otherwise, and  $z_{\alpha/2} = 1.96$  is the upper  $\alpha/2$  point of the standard normal distribution.

## 4.2 Results of the Study

Table 1A ahead gives the percent relative biases of the six point estimates for the total number of employed using equation (13), and Table 1B gives the same for the employment rate. In both tables, all biases are less than 1% in absolute value, except for the two poststratified second phase estimators, SP-REE and SP-DEE, when  $m_g = 10$  and 5 in Table 1A. Since both of these estimators can be written as separate ratio estimators within poststrata, it seems reasonable that the increased (absolute) bias is due to the small number of second phase units within each poststratum. According to Särndal, Swensson and Wretman (1992), page 270, "a rule of thumb is to keep the number of groups (poststrata) sufficiently



small so that no group sample count is less than 20". This is clearly violated in these two cases. Even so, all estimators behave reasonably in terms of point estimation, as expected.

Table 2A ahead gives the percent relative biases of the jackknife variances for the total number of employed using equation (14), and Table 2B gives the same for the employment rate. Focusing first on Table 2A, the Full First Phase Estimator's variance is almost perfectly unbiased, at 0.94%. Among the second phase estimators, the Reweighted Expansion Estimator clearly comes out the winner, having small negative biases in the variances always less than 6% in absolute value. The biases become increasingly negative as the second phase sample sizes diminish. Both variants of the Double Expansion Estimator fail miserably, with very large positive biases in the variances ranging from 46.35% to 1997.51%! The second variant is worse than the first, but both are well beyond the realm of acceptable behaviour. In the case of simple poststratification, the variance of the Full First Phase Estimator exhibits a small positive bias of 3.3%. The poststratified version of the Reweighted Expansion Estimator still behaves reasonably well, exhibiting biases in the variances between 4.88% and 12.03%. However, both variance variants of the poststratified versions of the Double Expansion Estimator behave poorly, although not as poorly as in the cases without poststratification. Here, the biases in the variances range between 22.39% and 50.03%. As before, variant 2 is worse than variant 1.

Table 2B repeats the analysis for the ratio estimate of employment rate. The results here are rather curious and difficult to explain since all variances behave reasonably well, with the exception of variant 2 of the double expansion estimator when  $m_g = 5$ . Other than this case where the bias in the variance is 30.46%, all other biases are less than 9.64% in absolute value. The biases in the variances of the reweighted expansion estimator and its poststratified counterpart tend to mirror their earlier behaviour in Table 2A. That is, the former tends to have small negative biases while the latter tends to have small positive biases. Even more curious is the fact that, in some instance, the variances of the two full first phase estimators are outperformed by some of the variances of the second phase estimators (DEE-variant 2, SP-REE and SP-DEE-variant 1 when  $m_g = 50$ , as well as DEE-variant 1 when  $m_g = 20$  and SP-DEE-variant 1 when  $m_g = 10$ ). Further analysis is required in order to explain the general behaviour of the ratio estimator. Nevertheless, Table 2A and 2B provide sufficient evidence that the jackknife variance of the reweighted expansion estimator is the correct one to use, since it behaves well in all cases, where that of the double expansion estimator fails badly at least for totals.



Although most studies focus on the *bias* of the variance estimators, it is also of secondary interest to look at the *coefficient of variation* of the variance estimators to see how stable the variance estimates themselves are. In Tables 3A and 3B, we investigate the coefficients of variation corresponding to the total number of employed and the employment rate, respectively. In equation (15), the expression under the square root in the numerator gives the MSE of the variance, whose component parts are the square of the bias of the variance and the variance of the variance. For those entries in Tables 2A and 2B where the bias of the variance has been determined to be exceedingly large (say larger than 20%), the corresponding entries in Tables 3A and 3B are not reported (indicated by a \*), since it is clear that those entries will be excessively large. In Table 3A, the coefficients of variation corresponding to the Reweighted Expansion Estimator range between 46.86% and 53.42%, while those of its poststratified counterparts range between 50.26% and 71.03%. There seems to be a tendency for the variances to become more unstable as the second phase sample sizes diminish, which is not surprising. Coefficients of variation of the magnitude exhibited here, although large, are typical for variance estimators, and have been encountered in other simulation studies relating to variances. See, for example, Kovačević, Yung and Pandher (1995). To that end, note that even the coefficients of variation corresponding to the Full First Phase Estimators are in the same range, and in fact, somewhat higher than those of the second phase estimators in certain cases.

Table 3B, which gives the coefficients of variation for the variances of the estimated employment rates, are entry by entry higher than their counterparts in Table 3A. In addition, all estimators exhibit the pattern that their corresponding coefficients of variation increase, quite substantially in fact, as the second phase sample sizes diminish. This effect is more pronounced for the ratio estimator than it is for the estimator of the total. The very high coefficients of variation in the column  $m_g = 5$  of both tables is not surprising, since the overall second phase sample size (25) is actually smaller than the overall number of PSUs drawn (36).

Lastly, Table 4A gives the percent coverages of confidence intervals having nominal value 95% for the estimator of total number of employed using equation (16), and Table 4B gives the same for the employment rate. In both tables, the coverage corresponding to the reweighted expansion estimator and its poststratified counterpart appears fairly close to the full first phase analogues, except for the case of estimating the employment rate when  $m_g = 5$  or  $m_g = 10$ . However, all of the above coverage rates



exhibit a tendency to underestimate the true coverage rather severely; the coverage for the cases REE and SP-REE ranges from 88.1% to 92.63% for Table 4A and from 74.1% to 90.05% for Table 4B. The increased coverage of both variants of the double expansion estimator and their poststratified counterparts in Table 4A (from 90.48% to 99.50%) over the reweighted and poststratified reweighted cases is not surprising; the large positive biases corresponding to these entries in Table 2A indicate that the variance is being severely overestimated, resulting in wider intervals. The results in Table 4B present an even more grim picture, with severe undercoverage across the entire table. Here, the coverage barely breaks the 90% mark, and in the most severe cases, plummets as low as 73.7%. Certainly some improvement would follow in both tables from using a t-statistic based on 18 degrees of freedom ( $t_{.025}(18) = 2.1$ ), since wider intervals would be achieved. However, the use of the t-statistic is questionable in this context, since the independence of the numerator and denominator  $\chi^2$  random variables in a t-statistic no longer holds, due to the complexity of the design. In fact, the traditional approximation to the degrees of freedom, given by number of replicants - number of strata =  $36 - 18 = 18$ , is probably not very good. Using the ideas of Satterthwaite (1941, 1946) that the degrees of freedom can be viewed as  $2/(CV(\text{variance}))^2$  in tandem with the results of Tables 3A and 3B gives a more realistic approximation to the degrees of freedom - that is, between 1 (corresponding to a coefficient of variation of 200.85%) and 9 (corresponding to a coefficient of variation of 46.86%). In conclusion, perhaps the use of confidence intervals should be avoided in this context, as illustrated by the results in Tables 4A and 4B. The results have been included here mainly as a cautionary note.

## 5. SUMMARY

The main purpose of this paper was to show that a simple jackknife variance estimator works well under a specific two-phase sampling strategy, provided the Reweighted Expansion Estimator is used in the estimation strategy and not the Double Expansion Estimator. A Monte Carlo simulation study supported these results, particularly for the case of estimating totals, even using small second phase sample sizes of magnitude 5 and 10.



**Table 1A - Percent Relative Bias of the Point Estimates for Total Number of Employed**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	.14	-0.3	-0.29	-0.56
DEE	--	0.16	-0.01	0.03	0.115
FFPE	0.04	--	--	--	--
SP-REE	--	-0.08	-0.93	-1.96	-4.44
SP-DEE	--	-0.05	-0.71	-1.67	-3.98
SP-FFPE	0.06	--	--	--	--

**Table 1B - Percent Relative Bias of the Point Estimates for Employment Rate**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	-0.09	-0.31	-0.19	-0.26
DEE	--	-0.08	-0.27	-0.119	-0.13
FFPE	-0.09	--	--	--	--
SP-REE	--	-0.28	-0.52	-0.43	-0.78
SP-DEE	--	-0.26	-0.49	-0.35	-0.67
SP-FFPE	-0.26	--	--	--	--



**Table 2A - Percent Relative Bias of Jackknife Variances for Total Number of Employed**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	-0.99	-2.51	-5.81	-5.13
DEE (Variant 1)	--	46.35	68.24	78.18	86.22
DEE (Variant 2)	--	101.59	278.44	654.99	1997.51
FFPE	0.94	--	--	--	--
SP-REE	--	4.88	6.42	12.03	9.20
SP-DEE (Variant 1)	--	28.52	32.04	35.33	22.39
SP-DEE (Variant 2)	--	33.50	40.62	50.03	46.41
SP-FFPE	3.3	--	--	--	--

**Table 2B - Percent Relative Bias of Jackknife Variances for Employment Rate**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	-3.53	-3.45	-7.09	-6.55
DEE (Variant 1)	--	-2.46	-1.53	-5.21	-7.41
DEE (Variant 2)	--	-0.357	4.91	9.09	30.46
FFPE	2.08	--	--	--	--
SP-REE	--	1.67	2.95	2.97	-4.25
SP-DEE (Variant 1)	--	1.65	2.49	0.93	-8.74
SP-DEE (Variant 2)	--	2.98	6.59	9.64	6.60
SP-FFPE	2.33	--	--	--	--



**Table 3A - Percent Coefficient of Variation of Jackknife Variances for Total Number of Employed**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	51.33	49.30	46.86	53.42
DEE (Variant 1)	--	*	*	*	*
DEE (Variant 2)	--	*	*	*	*
FFPE	56.71	--	--	--	--
SP-REE	--	50.26	58.26	63.82	71.03
SP-DEE (Variant 1)	--	*	*	*	*
SP-DEE (Variant 2)	--	*	*	*	*
SP-FFPE	58.10	--	--	--	--

**Table 3B - Percent Coefficient of Variation of Jackknife Variances for Employment Rate**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	59.28	65.66	74.26	103.06
DEE (Variant 1)	--	59.24	66.16	72.89	99.10
DEE (Variant 2)	--	60.94	73.20	92.71	*
FFPE	78.42	--	--	--	--
SP-REE	--	63.53	72.12	90.08	106.22
SP-DEE (Variant 1)	--	62.79	71.15	86.48	100.59
SP-DEE (Variant 2)	--	47.96	58.92	76.75	200.85
SP-FFPE	82.21	--	--	--	--



**Table 4A - Percent Coverage of 95% Confidence Intervals for Total Number of Employed**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	91.40	91.32	90.32	88.10
DEE (Variant 1)	--	95.60	96.38	95.95	94.43
DEE (Variant 2)	--	97.60	98.98	99.50	99.10
FFPE	89.25	--	--	--	--
SP-REE	--	92.63	91.80	91.08	89.13
SP-DEE (Variant 1)	--	95.12	94.30	93.53	90.48
SP-DEE (Variant 2)	--	95.48	94.98	94.43	91.83
SP-FFPE	91.03	--	--	--	--

**Table 4B - Percent Coverage of 95% Confidence Intervals for Employment Rate**

ESTIMATOR	$m_g = M_g$	$m_g = 50$	$m_g = 20$	$m_g = 10$	$m_g = 5$
REE	--	89.22	87.68	82.95	75.88
DEE (Variant 1)	--	89.28	87.60	83.30	75.25
DEE (Variant 2)	--	89.70	88.55	84.55	76.25
FFPE	89.30	--	--	--	--
SP-REE	--	90.05	87.78	83.20	74.10
SP-DEE (Variant 1)	--	90.15	87.75	83.15	73.70
SP-DEE (Variant 2)	--	90.22	88.10	84.33	75.35
SP-FFPE	90.28	--	--	--	--



## REFERENCES

- KOTT, P.S. (1995). Can the Jackknife be Used With a Two-Phase Sample? *Proceedings of the Survey Methods Section of the 1995 Statistical Society of Canada Meetings*, Montreal.
- KOVAČEVIĆ, M.S., YUNG, W. and PANDHER, G.S. (1995). Estimating the Sampling Variances of Measures of Income Inequality and Polarization - An Empirical Study. *Statistics Canada Branch Working Paper*, HSMD # 95-007E.
- KREWSKI, D. and RAO, J.N.K. (1981). Inferences from Stratified Samples: Properties of Linearization, Jackknife, and Balanced Repeated Replication Methods. *Annals of Statistics*, 9, 1010-1019.
- RAO, J.N.K. and SHAO, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 811-822.
- RAO, J.N.K. and WU, C.F.J. (1985). Inferences from Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, 80, 620-630.
- RUST, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1, 381-397.
- SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- SATTERTHWAITE, F.E. (1941). Synthesis of Variance. *Psychometrika*, 6, 309-316.
- SATTERTHWAITE, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2, 110-114.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G. and MAYDA, F. (1990). Methodology of the Canadian Labour Force Survey: 1984-1990. *Statistics Canada publication*, Catalogue 71-526.



Ca 005

STATISTICS CANADA LIBRARY  
BIBLIOTHEQUE STATISTIQUE CANADA



1010234875

