Survey Methodology, December 1987
Vol. 13, No. 2, pp. 147–161
Statistics Canada

Variance Estimation for the Canadian Labour Force Survey

G.H. CHOUDHRY and H. LEE[1]

ABSTRACT

The biases and stabilities of alternative variance estimators for the two stage random group design (Rao et al. 1962) are evaluated in a Monte Carlo study in the context of Canadian Labour Force Survey. The variance formula for raking ratio estimation procedure is derived using Taylor linearization method. The properties of the variance formula are investigated by a Monte Carlo simulation.

KEY WORDS: Keyfitz's variance estimator; Raking ratio estimator; Taylor linearization; Monte Carlo simulation.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is the largest monthly household survey conducted by Statistics Canada and is used to produce estimates of various labour force characteristics at national, provincial and sub-provincial levels. It follows a stratified multi-stage rotating sample design with six rotation panels (Platek and Singh 1976).

Following each decennial census of population, the LFS has undergone a sample redesign. As part of the 1981 post-censal redesign, an extensive program of research was undertaken in the areas of sampling, data collection, and estimation methodologies (Singh and Drew 1981). The post-stratified ratio estimation procedure used in the old design was replaced by a raking ratio estimation procedure to improve the reliability of subprovincial data. This paper presents the results related to variance estimation methodology.

The methodology for variance estimation for the old LFS was based on Woodruff's generalization (Woodruff 1971) of the Keyfitz procedure (Keyfitz 1957) using Taylor linearization applied to the post-stratified ratio estimates (Platek and Singh 1976). This method will be called the Keyfitz method as in Platek and Singh (1976).

There are three area types identified in the LFS design, i.e., self-representing (SR) areas consisting of major cities, non-self-representing (NSR) areas which are smaller urbans and rural areas, and special areas composed of military, institutions and remote areas. For the NSR and special areas it was decided to use the Keyfitz method with modification to incorporate the raking ratio estimation procedure.

However, for the two-stage random group design in SR areas, two alternative variance estimators given by Rao, Hartley, and Cochran (1962) and by Rao (1975) were evaluated and compared with Keyfitz's method using Monte Carlo simulation. The alternative variance estimators of estimates with and without ratio adjustment were compared with respect to their biases and stabilities. The impact on the Keyfitz variance estimator due to increase of the number of replicates was also examined. Details are reported in Section 2. Based on the results of the evaluation, the Keyfitz method was adopted for SR areas as well.

[1] G.H. Choudhry and H. Lee, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

The Keyfitz variance formula for raking ratio estimates used for all area types in the LFS is derived in Section 3 and evaluated by Monte Carlo study. Finally in Section 4, some concluding remarks are given.

2. VARIANCE ESTIMATION FOR THE SR DESIGN

2.1 SR Design

The LFS design in the SR areas is a two-stage random group design (Rao et al. 1962) with probability proportional to size (PPS) selection of primary sampling units (PSU's) and systematic selection of dwellings at the second stage such that the design becomes self-weighting. Suppose that there are N PSU's in a given stratum and let x_j and M_j, $j = 1, 2, \ldots, N$, respectively be the size measure and dwelling count for the j-th PSU in the stratum. Let $1/W$ be the sampling rate in the stratum, where W is an integer, and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i-th random group contains N_i PSU's, and $\Sigma_{i=1}^{n} N_i = N$.

Define

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t}, \quad j = 1, 2, \ldots, N,$$

and

$$\delta_{ij} = 1 \text{ if the } j\text{-th PSU is in the } i\text{-th group}$$

$$= 0 \text{ otherwise.}$$

Then $\pi_i = \Sigma_{j=1}^{N} \delta_{ij} p_j$ is the relative size of the i-th random group.

Now define W_{ij}, the sampling interval for systematic sampling, as follows: Let $a_{ij} = \delta_{ij} W p_j / \pi_i$ and $r_{ij} = a_{ij} - [a_{ij}]$ where $[a]$ is the greatest integer less than or equal to a. Without loss of generality, we assume that the set $\{r_{ij}, j = 1, 2, \ldots, N\}$ is in descending order. Then, W_{ij} is defined as

$$W_{ij} = [a_{ij}] + 1, \quad j = 1, 2, \ldots, R$$

$$= [a_{ij}], \quad j = R + 1, \ldots, N$$

where $R = \Sigma_{j=1}^{N} r_{ij}$. Then, by definition $\Sigma_{j=1}^{N} W_{ij} = W$ for the i-th random group, $i = 1, 2, \ldots, n$.

Since W_{ij} is the sampling interval for systematic sampling from the selected cluster in the i-th random group, it is defined as an integer for operational simplicity.

One PSU is selected with probability proportional to W_{ij}'s from each of the n random groups independently. The selected PSU j from the i-th random group is sub-sampled systematically at the rate $1/W_{ij}$. Then the overall sampling rate in each of the n random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W. Each random group is assigned a panel number from 1 to 6. The number of random

groups n is usually a multiple of six so that each panel has the same number of random groups.

Since only one PSU is selected from each random group, we denote by $1/W_i$ the subsampling rate in the selected PSU from the i-th random group and by m_i the number of selected dwellings from the random group i.

2.2 Alternative Variance Estimators

Suppose that we are interested in the total of a characteristic y for the stratum. Let y_{jk} be the y-value for the k-th dwelling in the j-th PSU where $k = 1, 2, \ldots, M_j$. Then the total $Y = \Sigma_{j=1}^{N} \Sigma_{k=1}^{M_j} y_{jk}$ can be estimated by $\hat{Y} = W \Sigma_{i=1}^{n} y_i$, where y_i is the sum of y-values for the m_i sampled dwellings from the PSU selected from the i-th group, $i = 1, 2, \ldots, n$. We consider the following variance estimators for estimating the variance of the estimated total $\hat{Y}$:

(1) Keyfitz's (1957) Variance Estimator

This estimator was used in the old design with two pseudo-replicates formed by collapsing the odd numbered panels into one replicate and the even into the other. Ignoring the finite population correction (fpc), the variance is obtained by

$$\hat{V}_1(\hat{Y}) = W^2 \left(\sum_o y_i - \sum_e y_i \right)^2 \tag{2.1}$$

where $\sum_o$ is the summation over all the odd numbered panels and $\sum_e$ is the summation over all the even numbered panels. Alternatively, the generalized Keyfitz variance estimator for $n(\geq 2)$ replicates which is given by

$$\hat{V}_2(\hat{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.2}$$

where $\bar{y} = (1/n) \Sigma_{i=1}^{n} y_i$, can be used. In this case each PSU or panel is taken as a replicate. $\hat{V}_2$ was considered because it was thought that this variance estimator might have better efficiency (stability) than $\hat{V}_1$ due to its larger number of degrees of freedom.

(2) Rao, Hartley, and Cochran's (1962) Variance Estimator

This variance formula is derived under the assumption that the number of secondaries m_i to be selected from the i-th group is fixed for $i = 1, 2, \ldots, n$, and simple random sampling (SRS) is also assumed at the second stage. The variance estimator is given by:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(\frac{M_i y_i}{m_i p_i} - \hat{Y} \right)^2 + \sum_1^n \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \tag{2.3}$$

where

$$A = \frac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}, \tag{2.4}$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2. \tag{2.5}$$

M_i is the number of dwellings in the selected PSU from the i-th group and m_i out of M_i dwellings are selected with systematic sampling but the variance estimate is obtained under the assumption of SRS. The y-value for the k-th selected dwelling from the selected PSU in the i-th group is y_{ik} and $\bar{y}_i = y_i / m_i$.

Since $\pi_i / p_i = W / W_i$ and $M_i / m_i = W_i$, (these equalities are not strict due to the use of integer values for W_i), the variance formula (2.2) can be written as:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 + W \sum_1^n \left(1 - \frac{m_i}{M_i} \right) M_i s_i^2. \tag{2.6}$$

(3) Rao's (1975) Variance Estimator

In this case it is assumed that m_i secondaries are selected with SRS but, since the design is self-weighting, the sample size m_i at the second stage is treated as a random variable. The variance estimator is given by:

$$\hat{V}_4(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 + \sum_1^n \left\{ \frac{\pi_i^2}{p_i^2} - A \left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2} \right) \right\} \frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2. \tag{2.7}$$

where A is defined by (2.4) and s_i^2 by (2.5). After some simplification (2.7) can be written as:

$$\hat{V}_4(\hat{Y}) = \hat{V}_3(\hat{Y}) + W^2 \sum_1^n m_i s_i^2 \left\{ \left(1 - \frac{W_i}{W} \right) - A \left(\frac{1}{\pi_i} - 1 \right) \right\}. \tag{2.8}$$

We note that there is an additional term, which could be positive or negative, in the variance formula when random sample size is assumed at the second stage.

2.3 Monte Carlo Study

In order to evaluate the biases of the four variance estimators and their relative stabilities, a Monte Carlo study was carried out with 19 Labour Force strata from the Census Metropolitan Area (CMA) of Halifax using data from the 1981 census. The census data for the purpose of this study was the census sample given the long questionnaire which is 20% systematic sample of dwellings within Enumeration Areas. The sampling rate $1 / W$ was taken to be 0.04 to obtain the same expected sample size as in the actual redesigned LFS. The number of random groups within each stratum was even and was determined so that the expected sample size within random groups would be as close to 4.5 as possible to correspond to the actual LFS. The 19 strata chosen for the study are shown in Table 1 with the number of PSU's, the number of selected PSU's, the number of dwellings, and the expected sample sizes along with the corresponding totals for all the strata. Within each of the 19 strata, 1,000

Table 1

Strata Used for the Monte Carlo Study

Stratum	No. of Dwellings	No. of PSU's	No. of Selected PSU's	Expected Sample Size
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
Total	11,057	697	100	442.3

samples were generated independently using a Monte Carlo technique, employing the random group design described in Subsection 2.1.

Let $\hat{Y}_{ht}$ be the estimate of the total Y_h for stratum h from the t-th Monte Carlo draw, $h=1, 2, \ldots, 19$, and $t=1, 2, \ldots, 1{,}000$. Similarly $\hat{V}_{jht}$, $j=1, 2, 3, 4$ are the four variance estimators of $\hat{Y}_{ht}$.

Now define

$$Y = \sum_{h=1}^{19} Y_h,$$

$$\hat{Y}_t = \sum_{h=1}^{19} \hat{Y}_{ht},$$

$$\hat{V}_{jt} = \sum_{h=1}^{19} \hat{V}_{jht}, \quad j = 1, 2, 3, 4,$$

where $t = 1, 2, \ldots, 1000$.

$\hat{Y}_t$ is the estimate of the total Y obtained from the t-th Monte Carlo draw and $\hat{V}_{jt}$, $j = 1, 2, 3, 4$ are the corresponding variance estimates.

The Monte Carlo expectation and variance denoted by E^* and V^* respectively are defined for T Monte Carlo draws as follows:

$$E^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t,$$

$$V^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}[\hat{\theta}_t - E^*(\hat{\theta})]^2,$$

where $\hat{\theta}$ is an estimator of the unknown parameter θ and $\hat{\theta}_t$ is the estimate obtained from the t-th draw. Using these definitions, we obtain the Monte Carlo variance of the estimator $\hat{Y}$, $V^*(\hat{Y})$, and the Monte Carlo expectations and variances of the variance estimators $\hat{V}_j$, $E^*(\hat{V}_j)$ and $V^*(\hat{V}_j)$ respectively for $j = 1, 2, 3, 4$.

Now define the bias of the variance estimator $\hat{V}_j$ by:

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

and percent bias as:

$$PB_j = 100\frac{B_j}{V^*(\hat{Y})}, \quad j = 1, 2, 3, 4.$$

Then the Mean Square Error (MSE) of $\hat{V}_j$ is given by:

$$MSE_j = V^*(\hat{V}_j) + B_j^2, \quad j = 1, 2, 3, 4.$$

We define the efficiency of $\hat{V}_j$, relative to the Keyfitz variance estimator with two replicates (i.e., $\hat{V}_1$) as:

$$\text{Rel. Eff}(\hat{V}_j \text{ vs. } \hat{V}_1) = (MSE_1/MSE_j)^{1/2}, \quad j = 2, 3, 4.$$

In this study, we consider three labour force characteristics: Employed, Unemployed, and In Labour Force. The relative biases and efficiencies of the variance estimators are reported in Tables 2A and 3A respectively for the three characteristics. We observe that, with respect to bias, the variance estimators 1 and 2 are similar and so are 3 and 4. The variance estimators 1 and 2 have very large positive biases notably for Employed and In Labour Force while 3 and 4 have relatively small biases. In efficiency comparison, the variance estimators 3 and 4 are much superior to 1 and 2 and very similar to each other. Moreover, the variance estimator 2 also performed better than 1.

The four variance estimators were also evaluated for ratio estimates by total population at the level of aggregation of all the strata. The corresponding variance estimators denoted by $\hat{V}_j^{(R)}$, $j = 1, 2, 3, 4$ were also obtained from each Monte Carlo draw by the Taylor linearization method. Then we obtained ratio adjusted version of percent biases of the four variance estimators (Table 2B) and relative efficiencies of the latter three variance estimators with respect to the first one (Table 3B).

We note that the biases of the variance estimators 1 and 2 were substantially reduced for ratio adjusted estimates especially for Employed and In Labour Force. For the variance estimators 3 and 4, the biases were also reduced for Employed and In Labour Force but there was very little change for Unemployed. Although the biases of the four variance estimators are small, the only nonsignificant bias at 5% level was that of the variance estimator 3 for In Labour Force. All the observed differences between biases were significant at 5% level except those of the variance estimators 1 and 2 for the three characteristics.

Table 2A

Percent Biases of the Variance Estimators of the Estimates of LF Characteristic Totals without Ratio Adjustment

Characteristic	Percent Bias			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Employed	23.4	24.5	−4.7	−6.3
Unemployed	6.3	6.6	3.7	1.2
In Labour Force	24.2	25.2	−5.1	−6.7

Table 2B

Percent Biases of the Variance Estimators of the Estimates of LF Characteristic Totals with Ratio Adjustment

Characteristic	Percent Bias			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	3.7	4.3	−1.1	−3.1
Unemployed	5.3	5.5	4.0	1.4
In Labour Force	4.5	5.0	−0.5	−2.5

Table 3A

Relative Efficiencies of $\hat{V}_2$, $\hat{V}_3$, and $\hat{V}_4$ with Respect to $\hat{V}_1$
(Rel. Eff. of $\hat{V}_j = [MSE(\hat{V}_1)/MSE(\hat{V}_j)]^{1/2}$, $j = 2,3,4$

Characteristic	Relative Efficiency		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
Employed	1.51	3.22	3.11
Unemployed	1.52	1.71	1.76
In Labour Force	1.49	3.24	3.12

Table 3B

Relative Efficiencies of $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, and $\hat{V}_4^{(R)}$ with Respect to $\hat{V}_1^{(R)}$
(Rel. Eff. of $\hat{V}_j^{(R)} = [MSE(\hat{V}_1^{(R)})/MSE(\hat{V}_j^{(R)})]^{1/2}$, $j = 2,3,4$

Characteristic	Relative Efficiency		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	2.13	2.59	2.52
Unemployed	1.57	1.71	1.76
In Labour Force	2.08	2.56	2.51

Table 4

Coverage Rates of 95% Confidence Intervals for the Estimates of LF Characteristic Totals with Ratio Adjustment

Characteristic	Coverage Rate			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	93.6	95.4	94.6	94.2
Unemployed	94.3	95.1	95.3	95.0
In Labour Force	93.2	95.3	94.6	94.2

We also computed the 95% confidence intervals (CI's) for the ratio-adjusted estimates from each Monte Carlo draw using the four variance estimators. The coverage rates were obtained as the proportion of CI's which include the true value of characteristic total. The results are given in Table 4 and show that the performances of all the 4 variance estimators are very good for all the characteristics. Since the variance estimators of ratio- adjusted estimates provide confidence intervals which have coverage rates very close to the nominal value, the small biases of the variance estimators are of no practical consequence. Thus, from the bias point of view, all four variance estimators for the ratio-adjusted estimates are not much different from each other. The relative efficiencies of the variance estimators 3 and 4 are now only marginally better than 2 regardless of characteristic. The relative efficiencies of the 3 alternatives in this case are over 2 for Employed and In Labour Force. For unemployed they are somewhat lower and lie between 1.5 and 1.8, which are almost the same as those for the unadjusted case. We should note here that the variance estimator 1 is computed with 19 degrees of freedom (1 per stratum). On the other hand, in the case of the 3 alternatives we have 81 degrees of freedom since each PSU is a replicate. Hence, we conclude that the stability of the Keyfitz variance estimator for the ratio-adjusted estimates is significantly improved by increasing the number of replicates and becomes comparable with the other two alternatives (see Table 3B).

2.4 Keyfitz's Variance Estimators with 2 vs. 6 Replicates for the LFS

The results of the Monte Carlo study reported in the previous sub-section have shown that the Keyfitz variance estimator compares well with the alternate methods for the variances of the ratio-adjusted estimates both from the bias and efficiency point of view when each method uses the same number of replicates. In addition, Keyfitz's method has the advantage of simplicity and estimating the variances of changes and averages under the alternative methods involves many complications. Therefore, the Keyfitz method was retained for the SR areas as well. In order to improve the efficiency of Keyfitz's method, 6 rotation panels were adopted as replicates as opposed to 2 replicates in the old design. One major concern with using the rotation panels as replicates was whether there would be any serious inflation of the variance estimate due to panel bias.

This aspect was investigated for the three LF characteristics by computing the variance estimates using the variance formula developed in Section 3 with 2 and 6 replicates from the actual LFS data for 24 months (March '85 - February '87). From the 24 estimated variances for each of the LF characteristics, the means and standard deviations (SD's) of the variances were obtained. The ratios of the means and SD's of the variances under the two alternatives (2 vs. 6 replicates) are averaged over 24 Census Metropolitan Areas (CMA's) and given in Table 5. The following observations can be made from the table:

Table 5

Comparison of SR Variance Estimates with 2 vs. 6 Replicates per Stratum Based on CMA Data of the LFS Mar '85 - Feb '87

Characteristic	Average Ratio of Means of Variances (2 vs. 6)	Average Ratio of SD's of Variances (2 vs. 6)
Employed	0.997	1.813
Unemployed	0.995	1.515
In Labour Force	1.003	1.833

Note: For each CMA, means and standard deviations of variance estimates were obtained from 24 months data for 2 and 6 replicates. Then the ratios (2 rep. vs. 6 rep.) of means of variances and of standard deviations (SD's) of variances were calculated for each CMA. The average ratios in the table are the averages over 24 CMA's.

(i) The effect on the levels of the variances due to using 6 replicates as compared to 2 is very minimal, which means that adopting rotation panels as replicates has little impact on the bias of the variance estimates.

(ii) As expected, the variances are more stable with 6 replicates than with 2 and the results are not much different from those of the Monte Carlo study (see the first column in Table 3B)

From the above observations, we conclude that the efficiency of the Keyfitz method is improved substantially without having serious impact on the bias by adopting the 6 rotation panels as replicates as opposed to using only 2 replicates.

3. VARIANCE ESTIMATION FOR RAKING RATIO ESTIMATES

3.1 Raking Ratio Estimation for the LFS

In the old LFS, post-stratified ratio estimation was used. The subweight, which is the design weight adjusted for non-response, was ratio-adjusted to external estimates of the LFS target population for 38 post-strata defined by age and sex at provincial level. The LFS target population is the population 15 years of age and over excluding armed forces, inmates of institutions, and population living on Indian reserves.

This ratio estimation enhanced the quality of provincial data substantially but subprovincial data still had somewhat poor reliability. In order to improve subprovincial data especially for Economic Regions (ER's) and Census Metropolitan Areas (CMA's), a raking ratio estimation procedure was adopted, through which simultaneous ratio adjustment at provincial and subprovincial levels is achieved.

The raking procedure is carried out in a sequence of adjustments: first, the subweight is adjusted to the subprovincial (CMA's and Non-CMA parts of ER's) population and then the provincial level adjustment by age / sex (the number of age / sex groups were reduced from 38 to 24 in the redesigned sample) is applied to the resulting weight. This procedure is repeated once more to obtain a second pair of weights. Note that for the ER's containing CMA(s), the CMA part is excluded when defining adjustment cells for the ER's so that the subprovincial adjustment cells are mutually exclusive. Let W_0 be the subweight and let (W_1, W_2) and (W_3, W_4) be the two pairs of weights resulting from the first and second iteration respectively. Labour force characteristics are estimated using W_4. Due to the order of adjustments, the marginal totals of W_4 at provincial age / sex groups are exactly the same as the external population estimates of the corresponding groups but the marginal totals of W_4 at

subprovincial level (ER and CMA) are not quite equal to the corresponding external population estimates. However, the differences are very small.

The special area frames, which are composed of military establishments, institutions, and remote areas, in general, do not respect the ER and CMA boundaries and hence, are treated differently during the raking procedure. Each special area type forms a stratum at the provincial level. The only exceptions are remote areas in the provinces of Quebec and Alberta where further stratification is carried out. Those ER's and CMA's which contribute to the special area frame will be called "contributing" ER's and CMA's. The special area records on the sample file are copied to each of the contributing ER's or CMA's with deflated subweights in proportion to the population of that particular type of special area in the contributing ER or CMA. The raking procedure is then carried out in the usual manner as described earlier.

3.2 Variance Formula for One-Iteration Raking Ratio Estimates

The variance formula for one-iteration raking ratio estimates is derived here. The basic methodology employed here is successive application of Taylor series approximation to the raking ratio estimates until we obtain a linear form of subweights. Then the replication formula is applied as in Woodruff (1971). The successive application of the Taylor series approximation was also used by Arora and Brackstone (1977a,b) and Brackstone and Rao (1979) to obtain variance formula of raking ratio estimates for simple random sampling of units or clusters. We have adopted this method for the stratified multi-stage PPS sampling design following Woodruff's approach.

Let $Y^{(0)}$, $Y^{(1)}$, $Y^{(2)}$ be the estimates of a labour force characteristic y in a province based on W_0, W_1, and W_2, respectively. The superscripts in parentheses correspond to the subscripts of W's.

Then $Y^{(2)}$ can be expressed as follows:

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \tag{3.1}$$

where $Y_a^{(1)}$ = W_1-weighted estimate of characteristic y for the age/sex group a in the province,

$P_a^{(1)}$ = W_1-weighted estimate of population for the age/sex group a in the province,

P_a = External estimate of population for the age/sex group a in the province.

Let $F_a = Y_a^{(1)}/P_a^{(1)}$. The first order Taylor approximation to F_a at $(E(Y_a^{(1)}), E(P_a^{(1)}))$ is

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\left\{Y_a^{(1)} - E(Y_a^{(1)})\right\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\left\{P_a^{(1)} - E(P_a^{(1)})\right\}$$

where E denotes expectation.

Then a Taylor approximation to the variance of $Y^{(2)}$ can be written as

$$V(Y^{(2)}) = V\left(\sum_a F_a P_a\right) \doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})}(Y_a^{(1)} - R_{Ya}^{(1)} P_a^{(1)})\right\} \tag{3.2}$$

where

$$R_{Ya}^{(1)} = \frac{E(Y_a^{(1)})}{E(P_a^{(1)})}.$$

Now the W_1-weighted estimates $Y_a^{(1)}$ and $P_a^{(1)}$ can be expressed in terms of W_0-weighted estimates as follows:

$$\begin{aligned} Y_a^{(1)} &= \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s, \\ P_a^{(1)} &= \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s, \end{aligned} \tag{3.3}$$

where s denotes a CMA or an ER or the complementary part of an ER after removing the CMA part and P_s is population of the subprovincial area s. Substituting the expressions for $Y_a^{(1)}$ and $P_a^{(1)}$ from (3.3) into (3.2) and applying the first order Taylor approximation to the ratios of W_0-weighted estimates, we obtain

$$V(Y^{(2)}) \doteq V\left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \left\{ \left(Y_{sa}^{(0)} - R_{Ysa}^{(0)} P_s^{(0)} \right) \right.\right.$$

$$\left.\left. - R_{Ya}^{(1)} \left(P_{sa}^{(0)} - R_{Psa}^{(0)} P_s^{(0)} \right) \right\} \right], \tag{3.4}$$

where

$$R_{Ysa}^{(0)} = \frac{E(Y_{sa}^{(0)})}{E(P_s^{(0)})} \text{ and } R_{Psa}^{(0)} = \frac{E(P_{sa}^{(0)})}{E(P_s^{(0)})}.$$

The expression in (3.4) can be written in terms of replicate level estimates. Define

$$\begin{aligned} Z_{Yshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (Y_{shia}^{(0)} - R_{Ysa}^{(0)} P_{shi}^{(0)}), \\ Z_{Pshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (P_{shia}^{(0)} - R_{Psa}^{(0)} P_{shi}^{(0)}), \end{aligned} \tag{3.5}$$

where h denotes a stratum belonging to s and i denotes a replicate in h.

Then (3.4) can be rewritten by rearranging the order of summations as follows:

$$\begin{aligned} V(Y^{(2)}) &\doteq V\left\{ \sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)} \right) \right\} \\ &= V\left(\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)} \right) \end{aligned} \tag{3.6}$$

where

$$D_{shi}^{(0)} = \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)} \right).$$

Apart from special area strata, $(\Sigma_{i=1}^{n_h} D_{pshi}^{(0)})$'s are independent because they are based on subweights. However, for the special area strata they are highly correlated because the same records are attributed to the contributing subprovincial areas.

We can rewrite (3.6) as

$$V(Y^{(2)}) \doteq V\left(\sum_{h\epsilon s}\sum_{h}\sum_{i=1}^{n_h} D_{shi}^{(0)}\right)$$

$$= V\left(\sum_{h}\sum_{i=1}^{n_h}\sum_{s\ni h} D_{shi}^{(0)}\right) \tag{3.7}$$

where $\Sigma_{s\ni h}$ is summation over all the subprovincial areas containing the stratum h. For a non-special stratum, the stratum appears only in one subprovincial area, and the summation $(\Sigma_{s\ni h})$ is redundant. However, a special area stratum could appear in several subprovincial areas and the summation $(\Sigma_{s\ni h})$ sums up all D-values $(D_{shi}^{(0)})$. belonging to the special area stratum.

Define

$$D_{hi}^{(0)} = \sum_{s\ni h} D_{shi}^{(0)}.$$

Then (3.7) becomes

$$V(Y^{(2)}) \doteq V\left(\sum_{h}\sum_{i=1}^{n_h} D_{hi}^{(0)}\right). \tag{3.8}$$

The variables, $\Sigma_i D_{hi}^{(0)}$, are independent since they are based on subweights. Then, ignoring the fpc, the variance can be estimated by

$$\hat{V}(Y^{(2)}) \doteq \sum_{h} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(0)} - \bar{D}_h^{(0)})^2 \tag{3.9}$$

where

$$\bar{D}_h^{(0)} = \frac{1}{n_h}\sum_{i=1}^{n_h} D_{hi}^{(0)}.$$

In this expression, however, expected values are involved and these are unknown. The variance can be approximated reasonably well by substituting expected values with their estimates and hence, from (3.9), we obtain the final form of $\hat{V}$ as follows:

$$\hat{V} \doteq \sum_{h} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(2)} - \bar{D}_h^{(2)})^2 \tag{3.10}$$

where

$$D_{hi}^{(2)} = \sum_{s \ni h} D_{shi}^{(2)},$$

$$\bar{D}_h^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a \left(Z_{Yshia}^{(2)} - R_{Ya}^{(2)} Z_{Pshia}^{(2)} \right),$$

$$Z_{Yshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{Pshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

and

$$R_{Ya}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

The formula (3.10) gives the variance for W_2-weighted estimates of LF characteristics and requires two weights W_0 and W_2.

3.3 Application of the One-Iteration Variance Formula to Two-Iteration Raking Ratio Estimates

The variance formula for the two-iteration raking ratio estimates can be obtained by successive application of the Taylor linearization as described in the previous section. However, the formula thus obtained is very complex. It was conjectured that the variance formula for one-iteration would be a reasonably good approximation for estimating the variance of the two-iteration raking ratio estimates. The rationale behind this conjecture was that there were only small perturbations in the weights after the first iteration. Now, the one-iteration variance formula uses the pair of weights (W_0, W_2). However, it was decided to use (W_0, W_4) instead of (W_0, W_2) since it was found that the use of W_4 instead of W_2 does not have any impact on the CV's of LF estimates which are based on W_4. The one-iteration variance

formula using the pair of weights (W_0, W_4) will be referred to as the one-iteration variance estimator.

To verify our conjecture, a Monte Carlo simulation study was carried out using the 1981 Census data from the province of Nova Scotia. In each Monte Carlo sample, the LFS design was simulated through all stages of sampling and a total of 1,000 Monte Carlo samples were selected independently. For each Monte Carlo sample, the following statistics were calculated for the three labour force characteristics at subprovincial and provincial levels;

1. Two-iteration raking ratio estimate, $Y^{(4)}$.
2. Variance estimate $\hat{V}(Y^{(4)})$ using the one-iteration variance estimator and the corresponding estimate of CV.
3. 95% confidence interval (i.e., $Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}$).

At the end of simulation, the average of 1,000 CV's was computed and compared with the Monte Carlo CV which is very close to the true value. The results are given in Table 6A. In all 21 cases (3 characteristics for each of 7 areas) the differences are less than 8% and in 13 cases less than 4%.

Also, the proportion of confidence intervals which cover the true characteristic value was obtained. The results are shown in Table 6B. Coverage rates for Employed and In Labour Force are very close to the nominal value in general, whereas those for Unemployed are somewhat lower but still acceptable.

It was also found that the two-iteration raking ratio estimate is nearly unbiased with a maximum of 0.35 percent bias in all 21 cases.

Table 6A

Average CV's Obtained by the
One-Iteration Variance Estimator and the Monte Carlo CV's

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
				Average CV's			
Employed	3.52	3.46	3.14	3.05	1.96	2.01	1.08
Unemployed	10.36	12.28	13.13	13.43	10.35	10.55	5.27
In Labour Force	2.98	3.17	2.85	2.73	1.77	1.83	0.91
				Monte Carlo CV's			
Employed	3.48	3.35	2.95	2.86	1.97	1.99	1.11
Unemployed	10.90	12.71	13.28	13.37	11.12	11.31	5.59
In Labour Force	2.76	3.08	2.76	2.53	1.72	1.74	0.92

Table 6B

Coverage Rates of 95% Confidence Intervals
Constructed by the One-Iteration Variance Estimator

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
Employed	94.5	92.8	94.0	94.7	94.7	94.9	92.5
Unemployed	92.1	90.7	91.4	91.8	92.7	92.7	93.1
In Labour Force	96.2	93.0	93.6	95.2	95.2	96.0	94.0

4. CONCLUSIONS

It has been shown that the Keyfitz variance estimation method for estimates without ratio adjustment (in this case it becomes just a replication method) has very large positive biases and low efficiencies while the alternatives have negligible biases and higher efficiencies for the labour force characteristics considered in this study.

However, for the ratio-adjusted estimates, all the methods considered here have negligibly small biases. It has also been shown that the efficiency of the Keyfitz method can be improved substantially and made comparable to the alternatives by increasing the number of replicates. It was demonstrated using actual LFS data that using 6 rotation panels as replicates in the Keyfitz variance estimator as opposed to 2 pseudo replicates does not introduce bias due to the phenomenon of rotation panel bias. As shown by Monte Carlo results, the one-iteration variance formula derived by the Keyfitz method using Taylor linearization gives reasonably good variance estimates for the two-iteration raking ratio estimates and has good coverage properties.

ACKNOWLEDGEMENT

The authors are grateful to the two referees, an Associate Editor and the Editor for their useful comments on the earlier version of the paper.

REFERENCES

ARORA, H.R., and BRACKSTONE, G.J. (1977a). An investigation of the properties of raking ratio estimators: I with simple random sampling. *Survey Methodology,* 3, 62-83.

ARORA, H.R., and BRACKSTONE, G.J. (1977b). An investigation of the properties of raking ratio estimators: II with cluster sampling. *Survey Methodology,* 3, 232-252.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 41, 97-114.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association,* 52, 503-510.

PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

RAO, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyā*, Series C, 37, 133-139.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society,* 24, 482-490.

SINGH, M.P., and DREW, J.D. (1981). Redesigning continuous surveys in a changing environment. *Survey Methodology,* 7, 44-73.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association,* 66, 411-414.