Statistics Canada Statistique Canada

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes-ménages

Canada

WORKING PAPER

METHODOLOGY BRANCH

A COMPARISON OF TWO VARIANCE ESTIMATION TECHNIQUES: THE JACKKNIFE METHOD AND THE LINEARIZED JACKKNIFE METHOD.

HSMD-95-005-E

Roanna M. Beebakhee

Labour Force Survey Methods Section

Household Survey Methods Division

Statistics Canada

June 1995

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

A COMPARISON OF TWO VARIANCE ESTIMATION TECHNIQUES: THE JACKKNIFE METHOD AND THE LINEARIZED JACKKNIFE METHOD.

Roanna M. Beebakhee

Labour Force Survey Methods Section
Household Survey Methods Division
Methodology Branch, Statistics Canada

ABSTRACT

The jackknife method of variance estimation is used extensively for surveys with complex designs. Although it has desirable statistical properties, it is computer intensive resulting in relatively high costs. In addition, when it is used for estimating the variance of regression estimators under a stratified multi-stage cluster design, it requires the calculation of matrix inverses which has the potential of generating computational degeneracies. An alternative method, known as the linearized jackknife method has been proposed and its theoretical properties have been investigated by Yung and Rao (1995). An empirical study has been undertaken which applies both variance methods to a number of household surveys conducted by Statistics Canada. Further, the linearized jackknife variance method and the customary Taylor linearization variance method are compared to the jackknife method.

UNE COMPARAISON DE DEUX TECHNIQUES D'ESTIMATION DE VARIANCE: LA MÉTHODE DU JACKKNIFE ET LA MÉTHODE DU JACKKNIFE LINÉARISÉ.

Roanna M. Beebakhee

Section des méthodes de l'enquête sur la population active
Division des méthodes d'enquêtes-ménages
Direction de la méthodologie, Statistique Canada

RÉSUMÉ

On utilise souvent l'estimation de la variance pa la méthode du jackknife pour les enquêtes avec un plan d'échantillonnage complexe. Même si cette méthode possède des propriétés statistiques intéressantes, elle demande de grandes ressources informatiques qui impliquent des coût relativement élevés. De plus, elle requiert l'inversion de matrices lorsqu'il s'agit d'estimer la variance d'estimateurs de régression qui proviennent d'un plan d'échantillonnage stratifié par grappes à plusieurs degrés. Il y a donc la possibilité d'obstacles de calcul majeurs. On proposera une méthode alternative connue sous le nom de méthode du jackknife linéarisé. Yung et Rao (1995) ont étudié ses propriétés théoriques et une étude empirique a été entreprise qui applique les deux méthodes d'estimation de variance à diverses enquêtes-ménages de Statistique Canada. De plus, on comparera la méthode du jackknife linéarisé et la linéarisation de Taylor traditionnelle à la méthode du jackknife.

1. INTRODUCTION

Variance estimates are produced in sample surveys in order to assess the precision of sample estimates and to inform users of data quality. The variance estimator is generally a function of both the sample design and the estimator (Wolter, 1985). Complex sample designs are used for some of the household surveys conducted by Statistics Canada, such as the Labour Force Survey (LFS), the Food Expenditure Survey (FES) and the Family Expenditure Survey (FAMEX). The estimator or sample statistic generated by these surveys is the regression estimator, which is not a linear function, but rather a non-linear function of the sample observations. An estimate of variance can be derived explicitly for linear functions; however, this is not true for non-linear functions and approximation methods must be used (Kovar et al., 1985). One such method is the jackknife variance method, which is currently used at Statistics Canada for the above mentioned household surveys.

The jackknife variance estimation method has been shown to be useful for a wide class of problems as discussed by Wolter (1985) and Rao (1985). However, a disadvantage of the jackknife method is its considerable computational cost. In order to compute the jackknife variance estimate of a regression estimator, several calculations of matrix inverses are required. These calculations are intensive, consuming a substantial portion of survey budgets and affecting the timeliness of publication releases. These factors may also discourage variance estimation for adhoc client requests. Further problems arise if a matrix is singular and the inverse can not be calculated. In order to overcome some of these problems, Yung and Rao (1995) have proposed a linearized jackknife variance estimator. This variance estimator is simply a linear approximation of the jackknife variance estimator of a regression estimator. Consequently, the two variance methods are asymptotically equivalent. Further, while investigating the finite sample properties, Yung and Rao have shown that the jackknife and the linearized jackknife variance estimators produce almost identical results.

The suitability of using the linearized jackknife method to produce variance estimates for the household surveys will be assessed through an empirical study. The goal of this study is twofold. First, the linearized jackknife variance estimates are compared to the jackknife variance estimates and secondly, the computer costs are compared. For the comparisons, variance estimates from the 1993 LFS, the 1992 FES and the 1992 FAMEX surveys are used. Section 2 profiles the two variance methods used in the comparison along with a description of the sample design and the sample statistic. Section 3 provides a brief description of the three household surveys used in the comparison. Section 4 describes the procedure taken to conduct the study. Section 5 reports the findings from the estimate comparison; section 6 reports the findings from the cost comparison. Section 7 compares the linearized jackknife variance estimate to the customary Taylor linearized variance estimate using the jackknife estimate as a base for comparison. Finally, section 8 states the conclusions of the study.

2. DESCRIPTION OF THE VARIANCE ESTIMATORS

The two variance estimators examined in the study, the jackknife and the linearized jackknife, are described in the following sub-sections. First, descriptions are provided of the sample design and the regression estimator, the sample statistic for which the variance estimators are defined.

2.1 Sample Design

Sampling units for the LFS, FES and FAMEX are selected from the same sampling frame and all three surveys generally follow the same sample design, a multi-stage stratified cluster design. A province is divided into L design strata, and n_h clusters are sampled from stratum h, where $n_h \geq 2$. For the purposes of variance estimation, the clusters are referred to as replicates. The set of sampled replicates in stratum h is denoted by s_h. For each i^{th} replicate, the ultimate sampling unit is selected, distinguished by the subscript hik. Thus the set of sampled units from the hi^{th} replicate is denoted by s_{hi}. The set of all sampled units for the province is denoted by s. Associated with each sampling unit is a variable of interest, y_{hik}, and a basic weight, w_{hik}, that is based on the sample design. In the household surveys, this basic weight is adjusted, for example for nonresponse and cluster growth. For more details on the design of the LFS, see Singh et al. (1990), for FES, see Statistics Canada (1992a), and for FAMEX, see Statistics Canada (1992b).

2.2 The Regression Estimator

For the household surveys, interest lies in estimating a population total, Y, such as total employment or expenditure for a province. The regression estimator uses auxiliary information to produce efficient estimates of Y, as described by Cochran (1977). The form of the regression estimator, $\hat{Y}_R$, is:

$$\hat{Y}_R = \hat{Y} + (\mathrm{P} - \hat{X})^T \hat{\beta}$$

where $\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}$, an unbiased estimate of Y,
P is the vector of known population totals for the auxiliary variables,
$\hat{X} = \sum_{(hik) \in s} w_{hik} x_{hik}$ is an unbiased estimate of P,
x_{hik} is the vector of auxiliary variables for hik^{th} unit,
$\hat{\beta} = (X'WX)^{-1} X'Wy$, is the vector of regression coefficients,
X is the matrix of auxiliary variables for the sample,
W is the diagonal matrix of basic weights for the sample,
and y is the vector of observed characteristic values for the sample.

In the LFS, FES and FAMEX, the form of the auxiliary matrix, X, is such that it includes an exhaustive and mutually exclusive set of indicator variables (Lemaître and Dufour, 1987), so that the regression estimator simplifies to:

$$\begin{aligned}\hat{Y}_R &= \hat{\beta}'P \\ &= y'WX(X'WX)^{-1}P = \sum_{(hik)\in s} y_{hik}w_{hik}x_{hik}(X'WX)^{-1}P \\ &= \sum_{(hik)\in s} y_{hik}w^{*}_{hik}\end{aligned} \qquad (2.2.1)$$

where $w^{*}_{hik} = w_{hik}x_{hik}(X'WX)^{-1}P$ and is called the final weight for the hik^{th} unit. For the household surveys, the auxiliary variables are modified to ensure equal final weights at the household level, as outlined by Lemaître and Dufour; however, this stage is suppressed for ease of understanding.

There is also an interest in estimating the ratio of two population totals, Y/Z, such as the unemployment rate or average expenditure per household. In this case, the ratio of two estimated totals, $\hat{Y}_R/\hat{Z}_R$ is used where both numerator and denominator are regression estimators of the form in (2.2.1). Here, $\hat{\beta}_Y$ and $\hat{\beta}_Z$, are the vectors of regression coefficients calculated separately for $\hat{Y}_R$ and $\hat{Z}_R$, respectively.

2.3 The Jackknife Variance Estimator

In order to implement the jackknife variance estimator, n_h non-overlapping replicates have been defined within each stratum. A replicate is removed and the remaining $n_h - 1$ replicates are adjusted to compensate for the removed replicate. An estimate using (2.2.1) is then calculated from this subset of the sample. The jackknife variance estimator has the following form:

$$Var_J(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{(hi)} - \hat{\theta})^2 \qquad (2.3.1)$$

where $\hat{\theta}_{(hi)}$ is the subsample estimate calculated after dropping the hi^{th} replicate. This holds for both totals and ratios of totals, where $\hat{\theta} = \hat{Y}_R \; or \; \hat{Y}_R/\hat{Z}_R$. This formula is applicable when the replicates are selected with replacement. This assumption of independently selected replicates in a stratum is not true for the household surveys, hence, a slight overestimation of the variance results (Singh et al., 1990).

The number of subsample estimates calculated is equal to the number of replicates in the sample. For the estimation of variance in a province, if there are n replicates, where $n = \sum_{h=1}^{L} n_h$, then $n + 1$ calculations of the regression estimator are required for each characteristic to be measured. For instance, there are approximately 6400 replicates for the entire LFS sample, approximately 2800 for FES and approximately 2500 for FAMEX.

The regression estimator is computed in two stages, as

$$\hat{Y}_R = \underset{(1)}{(y'WX)} \times \underset{(2)}{(X'WX)^{-1}P}.$$

The two parts (1) and (2) are "jackknifed" separately, i.e., removal of a replicate is performed on each part separately and then multiplied together to obtain the subsample estimate. Part (2) does not depend on the characteristic of interest and thus is calculated once for the full sample and once for each subsample. However, the computation of part (2) requires a matrix inverse, which is costly. For example, over 2800 inverse calculations would be performed for FES. Part (1) is not that computationally expensive since no matrix inverses are involved. However, it must be repeated for the full sample and all subsamples for each new characteristic. It is these numerous calculations for the two parts of the regression estimator that has made the jackknife variance estimator an expensive variance estimator to calculate.

2.4 The Linearized Jackknife Variance Estimator

The linearized jackknife variance estimator is based on the expression (2.3.1), the jackknife variance estimator of a regression estimator. The linearization is obtained by approximating the difference between the subsample estimate and the full sample estimate using a Taylor series expansion. See Yung and Rao (1995) for more details. This is not the same as the customary Taylor's linearized variance which linearizes the expression of the regression estimator, (2.2.1), first, and then derives the variance function around this approximate expression. The form of the linearized jackknife variance estimator for the regression estimator is:

$$\begin{aligned} Var_{LJ}(\hat{Y}_R) &= \sum_{h=1}^{L}\sum_{i=1}^{n_h} \frac{1}{n_h(n_h-1)}\left(n_h \sum_{k\in s_{hi}} w^{\bullet}_{hik}\, e_{hik} - \sum_{ik\in s_h} w^{\bullet}_{hik}\, e_{hik}\right)^2 \\ &= \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(r^{\bullet}_{hi} - \bar{r}^{\bullet}_{h}\right)^2 \end{aligned} \tag{2.4.1}$$

where $r^{\bullet}_{hi} = n_h \sum_{k\in s_{hi}} w^{\bullet}_{hik} e_{hik}$ and $\bar{r}^{\bullet}_{h} = \sum_{i=1}^{n_h} r^{\bullet}_{hi}/n_h$ with $w^{\bullet}_{hik}$ denoting the final weight for the k^{th} unit in the

hi^{th} replicate and $e_{hik} = y_{hik} - x_{hik}{}^{T}\hat{\beta}$.

This variance estimation method has fewer computations since the calculation of all subsample estimates is eliminated and the computation of $\hat{\beta}$ is performed only once, for the full sample.

In order to calculate the variance estimator for ratios, a different linearization is required. The formula for the jackknife variance estimator is more flexible since equation (2.3.1) still holds. However, the linearized jackknife variance estimator is not that flexible and an approximation has to be derived separately for the ratio of two totals. Fortunately, a closed form expression results and equation (2.4.1) is still appropriate except that the error term, e_{hik}, is now changed to:

$$\tilde{e}_{hik} = \frac{1}{\hat{Z}_R}\left[\left(y_{hik} - x_{hik}{}^T\hat{\beta}_Y\right) - \frac{\hat{Y}_R}{\hat{Z}_R}\left(z_{hik} - x_{hik}{}^T\hat{\beta}_Z\right)\right].$$

Sometimes only one replicate is available in a stratum. In this situation, no measure of variability is obtainable from either method. However, for the LFS, FES and FAMEX, the jackknife method is modified to produce variance estimates in these strata through the creation of a second replicate which has zero contribution to the sample estimate. Consequently, this modification is adopted in the linearized jackknife method. It should be noted that this modification is not theoretically justified, but the departure from the jackknife theory is maintained in the linearized jackknife method for comparison purposes.

3. DESCRIPTION OF HOUSEHOLD SURVEYS

Both variance estimation programs are similar for the three household surveys since they share common design elements such as the sampling frame and method of estimation. However, there are some differences particular to each survey that required individual customization.

3.1 Labour Force Survey

The LFS is a monthly survey that reports information about current labour market activities of the working age population in Canada. The survey is a repeated panel design, using 6 rotation groups which are used in the variance calculations as separate replicates within the strata. The LFS uses thirty age-sex breakdowns as auxiliary information as well as demographic totals in sub-provincial regions. Estimation and variance calculations are performed monthly as well as annually. Data are reported at the individual level and both the survey and reference period under study refer to 1993.

3.2 Family Expenditure Survey & Food Expenditure Survey

FAMEX provides information about the expenditure in private households and FES provides detailed data on food expenditure. These two surveys are used to monitor and update weights in the Consumer Price Index. As well, data from both surveys are used for the analysis of spending patterns, the formation of marketing decisions and the study of expenditure for low income groups. These surveys use the same sampling frame as the LFS, however, a different, smaller sample is selected. The auxiliary variables used in these two surveys are the number of individuals age 14 and under, the number of individuals age 15 and older, the number of one person households and the number of multi-person households. They are both annual surveys so estimation and variance calculations are only performed once a year and data are reported at the household level. The 1992 FES survey was collected every month in the reference year to obtain seasonal data and FAMEX was collected from January to March 1993, following the reference year.

4. STUDY METHODOLOGY

The procedure used to conduct the comparison is described in the following sub-sections. The choice of computer software used for the variance estimation is discussed along with the choice of the different measures used for the estimate and cost comparison.

4.1 Computer Software

The availability of commercial software that implements both variance estimation methods is limited. Since the LFS, FES and FAMEX use the jackknife method for their variance estimation, this production software, which was developed in-house, was used to obtain the jackknife estimates. The linearized jackknife method required the development of its own software. In order to maintain a fair comparison, the same computer platform and programming languages as the jackknife program were used. In addition, since the two methods share common mathematical concepts, i.e., matrix operations, the logic of the programs is similar.

4.2 Comparison Criteria

To compare the closeness of the linearized jackknife variance estimator for a particular characteristic relative to the jackknife variance estimator, the relative difference is used. If $\hat{\theta}$ is the estimate for the variable of interest, whether $\hat{Y}_R$ or $\hat{Y}_R/\hat{Z}_R$, then $Var_J(\hat{\theta})$ represents the jackknife variance estimate and $Var_{LJ}(\hat{\theta})$ represents the linearized jackknife variance estimate. Therefore, the relative difference between the variance estimates with respect to $\hat{\theta}$ is:

$$RD_{[J,LJ]}(\hat{\theta}) = \frac{Var_J(\hat{\theta}) - Var_{LJ}(\hat{\theta})}{Var_J(\hat{\theta})} \times 100 .$$

Similarly, the relative difference between the standard errors with respect to $\hat{\theta}$ is:

$$rd_{[J,LJ]}(\hat{\theta}) = \frac{SE_J(\hat{\theta}) - SE_{LJ}(\hat{\theta})}{SE_J(\hat{\theta})} \times 100 .$$

This expression also holds for the comparison of the coefficient of variations (CVs). A positive relative difference, $RD_{[J,LJ]}(\hat{\theta})$ or $rd_{[J,LJ]}(\hat{\theta})$, indicates that the estimate for the jackknife variance is larger than the estimate for the linearized jackknife; a negative relative different indicates that the linearized jackknife has a larger variance estimate.

In order to compare the computing costs of the two methods, factors such as the service units which are the amount of computer memory required to run the program and the execution charge are examined to provide a measure of cost. The processing time, or CPU time and total execution time provide measures of timeliness. The CPU time is the time used by the computer to interpret and process the information from the program. The execution time is the cumulative time taken on the computer to execute the program, which includes the processing time as well as the waiting time in execution queues and for system interventions.

5. ESTIMATE COMPARISON

The comparison of the two variance estimation methods is carried out and the resulting observations are described for each of the three household surveys.

5.1 LFS

For the LFS, variance estimates under the two methods were obtained for five labour force characteristics, i.e., total individuals in the labour force, employed, unemployed, not in labour force and the unemployment rate at the provincial and national levels. The comparison of variance estimates at the national level is shown in Table 5.1.1 for January 1993 and September 1993. The variance estimates are reported in millions, which are rounded to two decimal places as per the official variance publication.

TABLE 5.1.1: Comparison of Variance Estimates for LFS Characteristics at National Level

LFS CHARACTERISTIC	JANUARY 1993			SEPTEMBER 1993		
	$Var_J(\hat{\theta})$	$Var_{LJ}(\hat{\theta})$	$RD_{[J,LJ]}(\hat{\theta})$	$Var_J(\hat{\theta})$	$Var_{LJ}(\hat{\theta})$	$RD_{[J,LJ]}(\hat{\theta})$
In Labour Force	1,598.96	1,564.43	2.2%	1,583.68	1,544.10	2.5%
Employed	1,988.54	1,951.77	1.8%	2,094.12	2,047.01	2.2%
Unemployed	625.11	614.19	1.7%	689.16	673.82	2.2%
Not in Labour Force	1,598.96	1,564.43	2.2%	1,583.66	1,544.10	2.5%
Unemployment Rate	0.03	0.03	0.0%	0.04	0.04	0.0%

It is clear from Table 5.1.1 that the relative difference between the variance estimates for the unemployment rate in January and September is insignificant, as both variance estimates are equal at the level reported. The relative differences between the jackknife and the linearized jackknife variance estimates of variance range from 1.7% to 2.5%; for the characteristics other than unemployment rate, the variance estimates for the linearized jackknife are slightly smaller than the corresponding jackknife variance estimates.

Table 5.1.2 shows the standard error comparisons for the same LFS characteristics. The relative differences are again low. Hence, either standard error estimate may be used in statistical inference or in the computation of confidence intervals and the results would be close.

TABLE 5.1.2: Comparison of Standard Error Estimates for LFS Characteristics at National Level

LFS CHARACTERISTIC	JANUARY 1993			SEPTEMBER 1993		
	$SE_J(\hat{\theta})$	$SE_{LJ}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$	$SE_J(\hat{\theta})$	$SE_{LJ}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$
In Labour Force	39.99	39.55	1.1%	39.80	39.30	1.3%
Employed	44.59	44.18	0.9%	45.76	45.24	1.1%
Unemployed	25.00	24.78	0.9%	26.25	25.96	1.1%
Not in Labour Force	39.99	39.55	1.1%	39.80	39.30	1.3%
Unemployment Rate	0.17	0.17	0.0%	0.20	0.20	0.0%

The comparison of variance estimates for the characteristic employed at the provincial level is presented in Table 5.1.3 for January and September 1993.

TABLE 5.1.3: Comparison of Variance Estimates for Employed at Provincial Level

PROVINCE	JANUARY 1993			SEPTEMBER 1993		
	$Var_J(\hat{\theta})$	$Var_{LJ}(\hat{\theta})$	$RD_{[J,LJ]}(\hat{\theta})$	$Var_J(\hat{\theta})$	$Var_{LJ}(\hat{\theta})$	$RD_{[J,LJ]}(\hat{\theta})$
Newfoundland	15.45	14.95	3.2%	13.19	12.86	2.5%
Prince Edward Island	1.00	0.95	5.0%	1.26	1.12	11.1%
Nova Scotia	19.46	18.90	2.9%	18.28	17.79	2.7%
New Brunswick	11.97	11.81	1.3%	14.73	14.57	1.1%
Quebec	693.55	681.59	1.7%	708.51	695.38	1.9%
Ontario	846.56	833.77	1.5%	882.69	863.14	2.2%
Manitoba	30.65	29.66	3.2%	29.89	28.82	3.6%
Saskatchewan	28.32	27.84	1.7%	17.95	17.71	1.3%
Alberta	97.30	95.73	1.6%	85.30	83.50	2.1%
British Columbia	244.28	236.57	3.2%	322.33	312.12	3.2%

Here, the relative difference between the two methods varies over time, i.e, the relative difference for Newfoundland changed from 3.2% to 2.5% from January to September 1993. Also, the relative differences are positive so the variance estimates for the linearized jackknife are smaller than the jackknife variance estimates. Again, the relative differences are small. An exception is noted for Prince Edward

Island. There are 25 LFS design strata in Prince Edward Island, which is low in comparison to the other provinces where the number of strata ranges from 46 in Newfoundland to 277 in Ontario. The variance estimates for the linearized jackknife method are closer to the variance estimates for the jackknife method when the number of design strata is large (Yung and Rao, 1995). Hence, the low strata count in Prince Edward Island may affect estimation for the linearized jackknife method which affects the comparison of the variance estimates. For example, the relative difference of the variance estimates for employed is over 11% in September 1993 for Prince Edward Island.

Table 5.1.4 shows the CV comparison for employed, again at the provincial level.

TABLE 5.1.4: Comparison of Coefficient of Variation for Employed at Provincial Level

PROVINCE	JANUARY 1993					SEPTEMBER 1993				
	$CV_J(\hat{\theta})$		$CV_{LJ}(\hat{\theta})$		$rd_{[J,LJ]}(\hat{\theta})$	$CV_J(\hat{\theta})$		$CV_{LJ}(\hat{\theta})$		$rd_{[J,LJ]}(\hat{\theta}$
Newfoundland	2.31	C	2.27	C	1.7%	1.87	C	1.84	C	1.6%
Prince Edward Island	2.03	C	1.98	C	2.5%	1.99	C	1.88	C	5.5%
Nova Scotia	1.29	C	1.27	C	1.6%	1.16	C	1.15	C	0.9%
New Brunswick	1.27	C	1.26	C	0.8%	1.30	C	1.29	C	0.8%
Quebec	0.93	B	0.92	B	1.1%	0.88	B	0.87	B	1.1%
Ontario	0.62	B	0.62	B	0.0%	0.62	B	0.61	B	1.6%
Manitoba	1.16	C	1.14	C	1.7%	1.09	C	1.07	C	1.8%
Saskatchewan	1.25	C	1.24	C	0.8%	0.95	B	0.94	B	1.1%
Alberta	0.82	B	0.81	B	1.2%	0.73	B	0.72	B	1.4%
British Columbia	1.05	C	1.03	C	1.9%	1.14	C	1.12	C	1.8%

The relative differences are low. Since the LFS publishes letter values to represent a range of CV values, as seen in the table, the reported coefficient of variations experience little change when using one method or the other.

5.2 FES

The variance estimates for different FES estimates such as total household count, total expenditure for 7 items and average expenditure per household, were calculated at the provincial, national and regional levels. The average expenditure per household is a more meaningful estimate since it is used to correct for overall under-reporting of food and non-alcoholic beverages on the food diaries (Statistics Canada, 1992a). The comparison of standard errors between the two methods for average expenditure per household is shown in Table 5.2.1 at the national level.

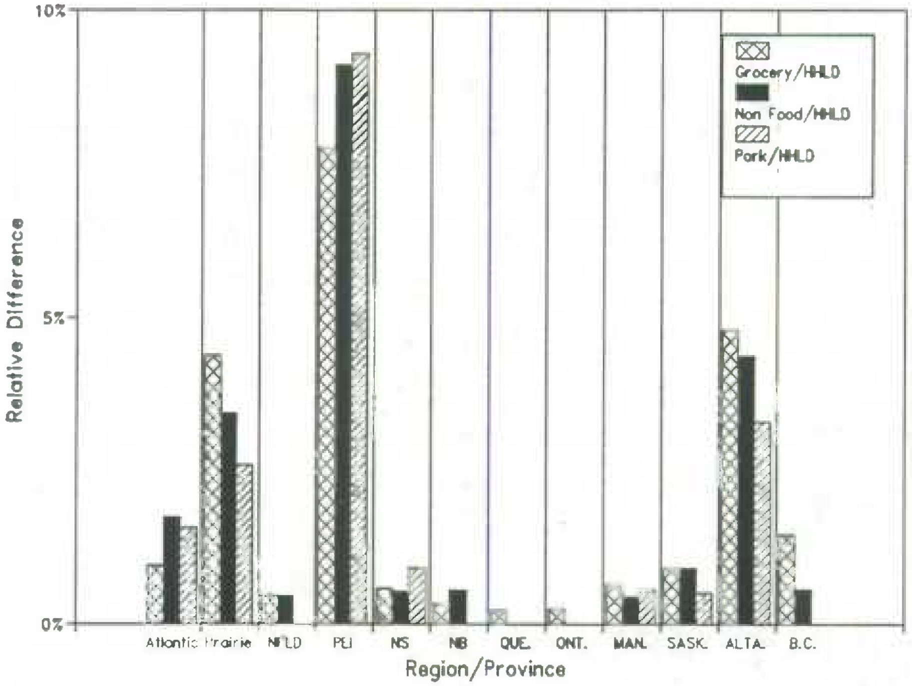
TABLE 5.2.1: Comparison of Standard Errors for FES Expenditures at National Level

AVERAGE EXPENDITURE	$SE_J(\hat{\theta})$	$SE_{LJ}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$
Grocery/Household	2.37	2.35	0.8%
Non Food/Household	0.65	0.64	1.5%
Bulk Food/Household	0.91	0.91	0.0%
Prepared Food/Household	0.31	0.31	0.0%
Beef/Household	1.64	1.63	0.6%
Pork/Household	0.57	0.56	1.8%
Other Meat/Household	0.78	0.77	1.3%

The range of the relative differences is low, varying from 0.8% to 1.8%. For average bulk food expenditure and average prepared food expenditure per household, the estimates for the linearized jackknife are equal to the jackknife estimates when rounded to two decimal places.

The relative differences between the two methods for the standard error at the provincial and regional level are plotted in Figure 5.2.2 for some FES expenditures.

FIGURE 5.2.2: Relative Difference of Standard Error Estimates for FES Expenditures



Again, many of the relative differences are low: under 5% and positive or equal to zero. Prince Edward Island is again the exception.

5.3 FAMEX

The FAMEX example reflects a large production run since it duplicated the FAMEX user's guide standard error table. Ninety-six expenditures were examined for nine different family composition categories at the provincial and national levels, producing over 19,000 variance estimates for both methods. The comparison of standard errors for the average expenditure per household for the nine family composition categories at the national level is shown in Table 5.3.1.

TABLE 5.3.1: **Comparison of Standard Errors for FAMEX Average Expenditure at National Level by Family Composition**

FAMILY COMPOSITION	$SE_J(\hat{\theta})$	$SE_{LJ}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$
One Person Household	3.29	3.29	0.0%
Married-Couple Only	2.50	2.49	0.4%
Married Couple with Children	8.90	8.71	2.1%
Married Couple with Relatives	11.87	11.66	1.8%
Married Couple with Non-relatives	20.09	19.31	3.9%
Lone Parent Family	9.75	9.77	-0.2%
Relatives only	0.40	0.40	0.0%
Other Households	1.83	1.71	6.6%
ALL	3.50	3.43	2.0%

This table shows that the linearized jackknife variance estimate can be larger than the jackknife variance estimate, as seen by the negative relative difference in the Lone Parent Family category. The relative differences are still low, ranging from -0.2% to 6.6%.

Table 5.3.2 presents standard error comparisons for the same average expenditure per married couple households at the provincial level.

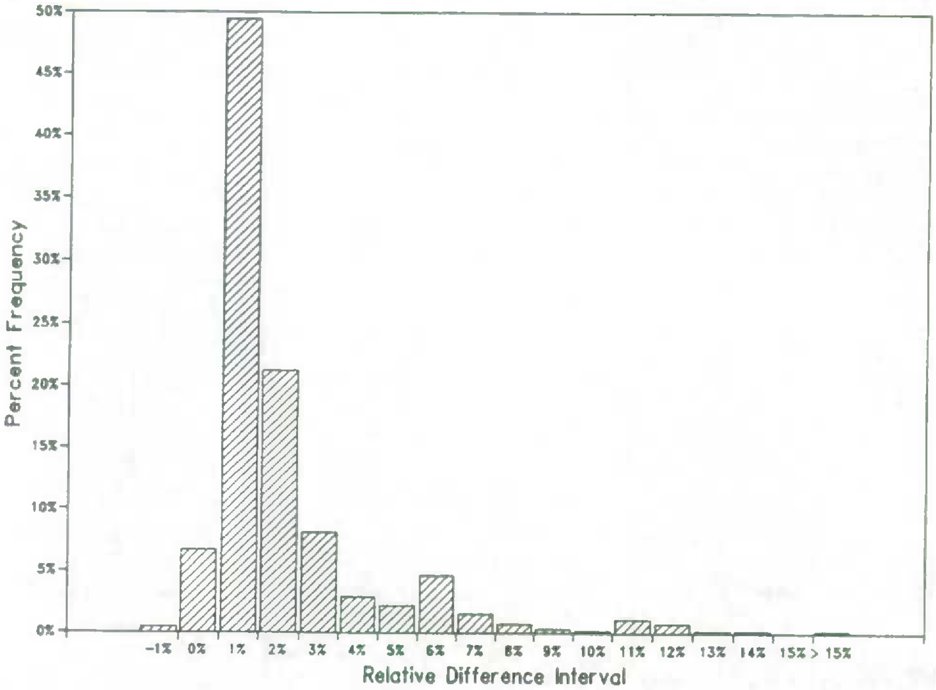
TABLE 5.3.2: Comparison of Standard Errors for FAMEX Average Expenditures at Provincial Level for Married Couple Households

PROVINCE	$SE_J(\hat{\theta})$	$SE_{LJ}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$
Newfoundland	2.73	2.66	2.6%
Prince Edward Island	2.30	2.28	0.9%
Nova Scotia	0	0	n/a
New Brunswick	19.55	18.97	3.0%
Quebec	8.15	8.14	0.1%
Ontario	0.61	0.61	0.0%
Manitoba	5.10	5.08	0.4%
Saskatchewan	0	0	n/a
Alberta	12.53	12.54	-0.1%
British Columbia	0.24	0.24	0.0%

Again the results are similar to the other surveys. In this example, the standard errors produced from the linearized jackknife method for Prince Edward Island are well-behaved, even though there is a small number of design strata for FAMEX.

Figure 5.3.3 plots the relative frequency of the relative differences for all the FAMEX average expenditure variance estimates in Ontario. Of the possible 864 average expenditures, only 741 had non-zero estimates permitting the calculation of a variance estimate by each method. Each interval in the histogram covers all values less than and equal to the interval value. It appears that many of the relative differences, approximately 77%, can be found to be within 0% to 2%.

FIGURE 5.3.3: **Percent Frequency of Relative Differences for FAMEX estimates in Ontario**



There are four points of interest from this analysis. The first point is that many of the relative differences are positive, which means that the linearized jackknife method produces variance estimates lower than the jackknife method. Second, if a discrepancy exists, the relative difference between the jackknife and the linearized jackknife is low, approximately under 5%. Third, the relative difference between the two methods is not uniform over time for a characteristic or an estimation area, so it is not predictable. And finally, the differences can be significant if the number of strata used in the computations is small, as in the case of Prince Edward Island.

6. COMPUTING COSTS

The computer comparison between the two variance methods is based on cost and timeliness. Computer costs for the LFS are shown in Table 6.1. The average costs for January 1993 and September 1993 are presented.

TABLE 6.1: Operational Cost for LFS

METHOD	COST		TIME	
	SERVICE UNITS	EXECUTION COST	CPU TIME	EXECUTION TIME
JACKKNIFE METHOD	8,624 Kb	$9.48	00:06:10	00:29:19
LINEARIZED JACKKNIFE	4,269 Kb	$4.69	00:03:03	00:23:14
RELATIVE DIFFERENCE	50%	51%	51%	21%

The linearized jackknife method achieves roughly a 50% savings in the computer resources used and its total execution charge. The processing time is also halved, from approximately 6 minutes to 3 minutes. The execution time is decreased by only 1/5, since both methods require a tape disk mount, a system intervention which is operator dependent.

The cost comparison for FES is presented in Table 6.2. Again the linearized jackknife has a lower execution cost, amount of service units and CPU time used, approximately 40% less than the jackknife method. Additionally, three different comparisons were conducted utilizing the different run classes offered by the mainframe computer. Depending whether the job is run for immediate output or delayed output, the execution cost and time are affected. Class P delivers immediate execution and is the most expensive, class N is normal daily usage and class U is an overnight execution which is the least expensive. Even though the execution cost is affected by the class type, the relative difference between the two variance methods was somewhat constant at 40%. However, the relative difference for the execution time differed. It ranged from a 76% savings in class P to a 90% savings in class U.

TABLE 6.2: Operational Cost for FES

METHOD	COST		TIME	
	SERVICE UNITS	EXECUTION COST	CPU TIME	EXECUTION TIME
	RUN CLASS: U - UNCOMMITTED			
JACKKNIFE METHOD	1,686 Kb	$1.85	00:01:08	00:25:07
LINEARIZED JACKKNIFE	975 Kb	$1.07	00:00:41	00:02:37
RELATIVE DIFFERENCE	42%	42%	40%	90%
	RUN CLASS: N - COMMITTED NON-PRIME			
JACKKNIFE METHOD	1,684 Kb	$3.36	00:01:08	00:19:38
LINEARIZED JACKKNIFE	990 Kb	$1.96	00:00:42	00:04:02
RELATIVE DIFFERENCE	41%	41%	38%	79%
	RUN CLASS: P - COMMITTED PRIME			
JACKKNIFE METHOD	1,717 Kb	$6.86	00:01:10	00:12:50
LINEARIZED JACKKNIFE	987 Kb	$3.94	00:00:41	00:03:07
RELATIVE DIFFERENCE	43%	43%	41%	76%

The cost comparison for FAMEX is shown in Table 6.3.

TABLE 6.3: Operational Cost for FAMEX

METHOD	COST		TIME	
	SERVICE UNITS	EXECUTION COST	CPU TIME	EXECUTION TIME
JACKKNIFE METHOD	65,091 Kb	$71.54	00:45:12	06:12:58
LINEARIZED JACKKNIFE	7,444 Kb	$8.18	00:05:22	00:13:47
RELATIVE DIFFERENCE	89%	89%	88%	96%

The estimation of variance for FAMEX is more intensive because several characteristics were examined at different family composition levels. The larger run dramatizes the difference between the jackknife and linearized jackknife programs. There is a 89% savings in terms of service units used, CPU time and execution charge. The cost of the job for the jackknife was approximately $70 compared to $8.18 for the linearized jackknife. The execution time decreased by 96% for the linearized jackknife because the jackknife program is divided into 7 smaller programs where each one requires two tape disk mounts. The linearized program is combined into one program requiring only one tape disk mount, which reduces the amount of system interventions required and as already noted, the number of matrix operations has substantially decreased.

7. EXAMINATION OF THE STANDARD TAYLOR LINEARIZATION METHOD

Similar to the linearized jackknife method, the customary Taylor linearization variance method requires separate variance formulas for different sample statistics. The Taylor expression for a regression estimator as given by Yung and Rao (1995) is

$$Var_L(\hat{Y}_R) = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{1}{n_h(n_h-1)} \left(n_h \sum_{k \in s_{hi}} w_{hik} e_{hik} - \sum_{ik \in s_h} w_{hik} e_{hik} \right)^2$$

where $e_{hik} = y_{hik} - x_{hik}{}^T \hat{\beta}$. The Taylor expression for the ratio of two totals is

$$Var_L(\hat{Y}_R / \hat{Z}_R) = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{1}{n_h(n_h-1)} \left(n_h \sum_{k \in s_{hi}} w_{hik} \tilde{e}_{hik} - \sum_{ik \in s_h} w_{hik} \tilde{e}_{hik} \right)^2$$

where $\tilde{e}_{hik} = \frac{1}{\hat{Z}_R}\left[\left(y_{hik} - x_{hik}{}^T\hat{\beta}_Y\right) - \frac{\hat{Y}_R}{\hat{Z}_R}\left(z_{hik} - x_{hik}{}^T\hat{\beta}_Z\right)\right]$. In this method, the linearization is applied to the sample statistic, not to the jackknife variance estimator. The Taylor method has a similar form to the linearized jackknife estimator (2.4.1), except that the basic weight, w_{hik} is used instead of the final weight, w^*_{hik}. An empirical examination of the variance and CV estimates for the three variance methods was undertaken with LFS data for January 1993. Table 7.1 shows the three variance estimates for LFS characteristics at the national level.

TABLE 7.1: Comparison of Variance Estimates for LFS Characteristics at National Level

LFS CHARACTERISTIC	JANUARY 1993				
	$VAR_L(\hat{\theta})$	$VAR_J(\hat{\theta})$	$VAR_{LJ}(\hat{\theta})$	$RD_{[J,L]}(\hat{\theta})$	$RD_{[J,LJ]}(\hat{\theta})$
In Labour Force	1,317.91	1,598.96	1,564.43	17.6%	2.2%
Employed	1,639.78	1,988.54	1,951.77	17.5%	1.8%
Unemployed	508.75	625.11	614.19	18.6%	1.7%
Not in Labour Force	1,317.91	1,598.96	1,564.43	17.6%	2.2%
Unemployment Rate	0.03	0.03	0.03	0.0%	0.0%

The relative difference between the Taylor linearized variance and the jackknife variance estimate is considerable higher than the relative difference between the linearized jackknife and jackknife variance estimates. Table 7.2 shows the CV estimates for employed at the provincial level.

TABLE 7.2: Comparison of Coefficient of Variation for Employed at Provincial Level

PROVINCE	JANUARY 1993				
	$CV_L(\hat{\theta})$	$CV_J(\hat{\theta})$	$CV_{LJ}(\hat{\theta})$	$rd_{[J,L]}(\hat{\theta})$	$rd_{[J,LJ]}(\hat{\theta})$
Newfoundland	2.23	2.31	2.27	3.5%	1.7%
Prince Edward Island	1.86	2.03	1.98	8.4%	2.5%
Nova Scotia	1.22	1.29	1.27	5.4%	1.6%
New Brunswick	1.20	1.27	1.26	5.5%	0.8%
Quebec	0.85	0.93	0.92	8.6%	1.1%
Ontario	0.56	0.62	0.62	9.7%	0.0%
Manitoba	1.05	1.16	1.14	9.5%	1.7%
Saskatchewan	1.17	1.25	1.24	6.4%	0.8%
Alberta	0.75	0.82	0.81	8.5%	1.2%
British Columbia	0.94	1.05	1.03	10.5%	1.9%

Again, the Taylor linearization shows a larger difference from the jackknife than the linearized jackknife, although, both linearization methods produce smaller estimates than the jackknife method. The three variance estimates are approximations for the true variance and are shown to be asymptotically equivalent (Kovar, 1985; Yung and Rao, 1995). The empirical data shows that the linearized jackknife variance method produces estimates closer to the jackknife method than the Taylor linearized method. There is no need to analyze the cost differences, since the computational cost for the Taylor linearized method is equivalent to the linearized jackknife method.

8. CONCLUSIONS

The linearized jackknife variance method has been compared to the jackknife variance method. Both variance methods require specialized software and both methods are approximations to the true variance of the regression estimator in a stratified multi-stage design. The jackknife method is known to over-estimate the true variance of the regression estimator. Since the linearized jackknife is an approximation to the jackknife, it tends to produce smaller estimates. The linearized jackknife is computationally simpler while the formula for the jackknife method is more flexible given the form of the estimator.

For some sample statistics no differences between the two variance estimates were found, given the reported level of precision. The linearized jackknife produced smaller estimates but these differences when they existed were small and lower than 5%. The linearized jackknife variance estimates for Prince Edward Island were somewhat problematic due to smaller number of design strata used in the surveys.

The operational costs of running the two methods were compared. Measures for cost and timeliness were examined. The linearized jackknife consistently consumed less time and money for all study surveys. The amount of time and money is affected by the length of the job, where larger runs show the greater difference between the two methods, i.e. a greater savings by the linearized jackknife method.

The customary Taylor linearization variance estimate does not approximate the jackknife variance estimate as well as the linearized jackknife variance estimate for totals, even though the computer costs for both linearization methods are comparable. However, for the ratio of two totals, all three methods produce approximately the same variance estimates, at the level reported.

ACKNOWLEDGEMENTS

The author would like to thank Professor J.N.K Rao of Carleton University and W. Yung of Statistics Canada for their assistance and suggestions in the implementation of the variance estimation techniques. The author would also like to thank U. Nevraumont and I. Clogg of Statistics Canada for their assistance with the survey data, J. Armstrong, M. Kovacevic, S. Kumar, and H. Lee of Statistics Canada for their assistance with the preparation of this report.

REFERENCES

1) Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley and Sons.

2) Kovar, J. (1985). *Variance Estimation of Nonlinear Statistics in Stratified Samples*. Statistics Canada: Working Paper No. BSMD - 85 - 052E.

3) Kovar, J., Ghangurde, P., Germain, M.-F., Lee, H. and Gray, G. (1985). *Variance Estimation in Sample Surveys*. Statistics Canada: Working Paper No. BSMD - 85 - 049E.

4) Lemaître G. and Dufour J. (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, Vol. 13, No. 2, pp. 199-207.

5) Rao, J.N.K. (1985). Variance Estimation in Sample Surveys. Technical Report Series of the Laboratory for Research in Statistics and Probability. Carleton University: No. 62.

6) Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian labour force survey*. Statistics Canada: Catalogue Number 71-526.

7) Statistics Canada. (1989). *Variance Estimation for Surveys using the LFS Frame Weighting System*. Ottawa: User Documentation.

8) Statistics Canada. (1992a). *Family Food Expenditure in Canada*. Ottawa: Catalogue Number 62-554.

9) Statistics Canada. (1992b). *Family Expenditure in Canada*. Ottawa: Catalogue Number 62-555.

10) Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

11) Yung, W. and Rao, J.N.K. (1995). *Variance Estimation in Stratified Multistage Sampling*. Statistics Canada: Working Paper No. HSMD - 95 - 002.