Survey Methodology, December 1997
Vol. 23, No. 2, pp. 81-89
Statistics Canada

Can the Jackknife Be Used With a Two-Phase Sample?

PHILLIP S. KOTT and DIANA M. STUKEL[1]

ABSTRACT

The jackknife variance estimator has been shown to have desirable properties when used with smooth estimators based on stratified multi-stage samples. This paper focuses on the use of the jackknife given a particular two-phase sampling design: a stratified with-replacement probability cluster sample is drawn, elements from sampled clusters are then restratified, and simple random subsamples are selected within each second-phase stratum. It turns out that the jackknife can behave reasonably well as an estimator for the variance for one common "expansion" estimator but not for another. Extensions to more complex estimation strategies are then discussed. A Monte Carlo study supports our principal findings.

KEY WORDS: Stratified; Reweighted expansion estimator; Double expansion estimator; Asymptotic.

1. INTRODUCTION

Krewski and Rao (1981) and Rao and Wu (1985) explore the design-based properties of the jackknife variance estimator given a stratified multi-stage sample incorporating with-replacement sampling in the first stage. Their results, although fairly general, cannot be directly applied to many multi-phase sampling designs. See also Wolter (1985; Chapter 4.5).

In this paper, we consider a simple example of two-phase sampling. A stratified with-replacement probability cluster sample is selected in a first phase of sampling. The elements in sampled clusters are then restratified, perhaps using information gathered from the first-phase sample, and a stratified simple random subsample is drawn without replacement.

One can estimate a total without auxiliary information in one of two ways. In the *double expansion estimator* – called "the π^* estimator" in Särndal, Swensson, and Wretman (1992, p. 347) – the value of each subsampled element is simply multiplied by the product of its expansion factor at each phase (*i.e.*, the inverses of its first-phase and second-phase selection probabilities) and then summed.

Although the double expansion estimator is more easily located in text books, the *reweighted expansion estimator* may be more common in practice, especially when element nonresponse is treated as a second phase of sampling, as in the weighting class estimator of Oh and Scheuren (1983, p. 150). An estimator for the population size of each second-phase stratum is computed by summing the first-phase expansion factors of all the elements in the second-phase stratum before subsampling. This value is then multiplied by the estimated second-phase stratum mean based on the subsample to yield an estimated stratum total. The second-phase estimated stratum totals are finally added together to produce the reweighted expansion estimator for the population total.

We are more concerned here with real two-phase sampling, rather than the artifice of treating nonresponse as an additional sampling phase. The National Agricultural Statistics Service (NASS) presently uses the double expansion estimator in its Quarterly Agricultural Surveys (QAS). A stratified area cluster sample is enumerated in June. Farms identified in the June survey are restratified based on their June responses and then subsampled for enumeration in September, December, and March.

NASS uses a two-phase design and the reweighted expansion estimator for its on-farm chemical use surveys. The first phase of sampling identifies farms with specific crops, and the second phase measures pesticide use on those crops.

This paper shows that although the jackknife may be used to estimate the variance of the reweighted expansion estimator under certain conditions, it is not generally effective as a variance estimator for the double expansion estimator. Section 2 introduces the reweighted expansion estimator and discusses its mean squared error. Section 3 shows that the jackknife variance estimator can be nearly unbiased for the reweighted variance estimator, while Section 4 addresses the jackknife's failings as a variance estimator for the double expansion estimator. Section 5 describes a simulation study that appears to confirm the main assertions of the previous sections. Section 6 discusses extensions of the reweighted expansion estimator, and Section 7 offers some concluding remarks. An appendix provides an outline of our assumed asymptotic framework and some proofs.

2. THE REWEIGHTED EXPANSION ESTIMATOR

2.1 The Estimator

Let $h(= 1, ..., H)$ denote the first-phase strata of a stratified with-replacement probability cluster sample, n_h the number of sampled clusters in stratum h, and F_h the set of those clusters. Let $g(= 1, ..., G)$ be the second-phase

[1] Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, VA 22030; Diana M. Stukel, Household Survey Methods Division, Statistics Canada, Ottawa, Canada K1A OT6.

strata from which a stratified simple random subsample is drawn without replacement. An element in a cluster sampled p times in the first phase is treated as p distinct elements for the subsample. Let M_g be the number of elements in g before subsampling and m_g the number of subsampled elements in g. In practice, the G second-phase strata are often not defined until after the first-phase sample has been drawn.

Let S_g be the set of elements in g before subsampling, s_g the set of subsampled elements in g, s the entire set of subsampled elements, and $m = \sum_g m_g$ the subsample size. Finally, let y_i be the value of interest for element i, and w_i the first-phase expansion factor for i (*i.e.*, the inverse of the selection probability for the cluster containing i).

The estimator for the population total, T, one would use if all the elements in the first-phase sample were enumerated can be written as

$$t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i y_i. \tag{1}$$

Let the *reweighted expansion estimator* for T be:

$$t_2 = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} (M_g/m_g) w_i y_i}{\sum_{i \in s_g} (M_g/m_g) w_i} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} \right\}. \tag{2}$$

An alternative expression for t_2 is

$$t_2 = \sum_{g=1}^{G} \sum_{i \in s_g} a_i y_i = \sum_{i \in s} a_i y_i, \tag{3}$$

where

$$a_i = \left[\sum_{k \in S_g} w_k \Big/ \sum_{k \in s_g} w_k \right] w_i \text{ for } i \in s_g$$

is the *adjusted weight* for element i. Equation (3) is what gives the reweighted expansion estimator its name.

2.2 Its Mean Squared Error (Some Theory)

Now t_2 is not, in general, an unbiased estimator of T. Nevertheless, under certain mild conditions specified in the appendix, it is a design consistent estimator for T; that is, $\text{plim}_{m \to \infty} (t_2 - T)/T = 0$ (Isaki and Fuller 1982). For the exposition in the text, it suffices to say that the m_g are assumed to be large.

Observe that

$$E[(t_2 - T)^2] = E[(\{t_1 - T\} + \{t_2 - t_1\})^2]$$

$$\approx \text{Var}_1(t_1) + E_1\{E_2[(t_2 - t_1)^2]\},$$

where the subscripts on Var and E denote the phase of sampling. Since the m_g are assumed to be large, $E_2[t_1(t_2 - t_1)] = t_1 E_2(t_2 - t_1) \approx 0$. Also, $E(t_2 - T) = E_1[E_2(t_2 - T)] \approx 0$, and the mean squared error of t_2 is effectively its (asymptotic) variance.

Since first phase of sampling was conducted with replacement, $\text{Var}_1(t_1)$ can, in principle, be estimated by

$$v_{L1} = \sum_{h=1}^{H} (n_h/[n_h - 1])$$

$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i y_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i y_i \right]^2 / n_h \right), \tag{4}$$

where U_{hj} is the set the elements in sampled cluster j of first-phase stratum h. The subscript L denotes "linearization" for historical reasons although there is nothing to linearize in this context. Note that when there is a second phase of sampling, it will generally not be possible to compute v_{L1} in practice.

Now

$$t_2 - t_1 = \sum_{g=1}^{G} \sum_{i \in S_g} w_i \left\{ \frac{\sum_{i \in s_g} w_i y_i}{\sum_{i \in s_g} w_i} - \frac{\sum_{i \in S_g} w_i y_i}{\sum_{i \in S_g} w_i} \right\}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i \frac{\sum_{i \in s_g} w_i r_i}{\sum_{i \in s_g} w_i},$$

where

$$r_i = y_i - \sum_{k \in S_g} w_k y_k \Big/ \sum_{k \in S_g} w_k \text{ for } i \in S_g.$$

It is crucial for the arguments below to realize that r_i has been defined so that $\sum_{i \in S_g} w_i r_i = 0$ for all g.

Continuing,

$$t_2 - t_1 \approx \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i, \tag{5}$$

since $\sum_{i \in S_g} w_i \approx \sum_{i \in s_g} (M_g/m_g) w_i$ (see equation (A1) of the appendix). This implies

$$E_2[(t_2 - t_1)^2] \approx \text{Var}_2 \left\{ \sum_{g=1}^{G} \sum_{i \in s_g} (M_g/m_g) w_i r_i \right\}$$

$$= \sum_{g=1}^{G} (M_g^2/[\{M_g - 1\} m_g])(1 - m_g/M_g)$$

$$* \left\{ \sum_{i \in S_g} (w_i r_i)^2 - \left(\sum_{i \in S_g} w_i r_i \right)^2 / M_g \right\}$$

$$\approx \sum_{g=1}^{G} ([M_g/m_g] - 1) \left\{ \sum_{i \in S_g} (w_i r_i)^2 \right\}. \tag{6}$$

Observe that equation (6) does *not* ignore the finite population corrections from the second phase of sampling.

3. THE JACKKNIFE VARIANCE ESTIMATOR

3.1 The Variance Estimator

We are now ready to discuss the jackknife. For $j \in F_h$, define the jackknife replicate $t_{(hj)2}$ as

$$t_{(hj)2} = \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} \right\}, \qquad (7)$$

where

$$w_{hji} = \begin{cases} w_i n_h/(n_h - 1) & \text{when } i \in U_{hj'} \text{ and } j' \neq j \\ 0 \text{ when } i \in U_{hj} & \\ w_i \text{ when } i \in U_{h'j'} & \text{and } h' \neq h. \end{cases}$$

Similarly, we define

$$t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} y_i.$$

Following Rust (1985), the *jackknife variance estimator*, $v_{Jf}(f = 1 \text{ or } 2)$, is defined here simply as

$$v_{Jf} = \sum_{h=1}^{H} (n_h - 1)/n_h \sum_{j \in F_h} (t_{(hj)f} - t_f)^2. \qquad (8)$$

This form is labeled $v_J^{(2)}$ in Krewski and Rao (1981, equation (2.4)). It is easy to show that $v_{J1} = v_{L1}$.

3.2 Why it Works (More Theory)

We will soon see that v_{J2} provides a nearly unbiased estimator for the variance of the reweighted expansion estimator in equation (2). Rao and Shao (1992) indirectly make the same claim (our equation (2) is the expectation of their estimator in Section 3.3, pp. 818-819). Their work, however, treats nonresponse as an additional phase of sample selection in which Poisson sampling (Särndal *et al.* 1992, p. 85) is used in place of stratified simple random sampling. Each first-phase sample element in the Rao and Shao (1992) setup is effectively a second-phase stratum. Consequently, the near unbiasedness of v_{J2} reduces to a special case of a result in Krewski and Rao (Rao and Shao 1992, p. 821).

What we have called the second-phase strata are reweighting classes in the Rao and Shao (1992) setup. Elements in the same class are assumed to have the same unknown probability of selection/response. *Conditional* on the realized subsample sizes within reweighting classes, Poisson sampling is equivalent to stratified simple random sampling. Rao and Shao's (1992) treatment, however, is *unconditional*.

Returning to the problem at hand, observe that

$$t_{(hj)2} - t_{(hj)1} = \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} \left\{ \frac{\sum_{i \in s_g} w_{hji} y_i}{\sum_{i \in s_g} w_{hji}} - \frac{\sum_{i \in S_g} w_{hji} y_i}{\sum_{i \in S_g} w_{hji}} \right\}$$

$$= \sum_{g=1}^{G} \left\{ \sum_{i \in S_g} w_{hji} \frac{\sum_{i \in s_g} w_{hji} r_{hji}}{\sum_{i \in s_g} w_{hji}} \right\},$$

where

$$r_{hji} = y_i - \sum_{k \in S_g} w_{hjk} y_k \Big/ \sum_{k \in S_g} w_{hjk} \quad \text{for} \quad i \in s_g.$$

Under mild conditions (see equations (A2) and (A3) in the appendix), we have the following analogue to equation (5):

$$t_{(hj)2} \approx t_{(hj)1} + \sum_{g=1}^{G} (M_g/m_g) \sum_{i \in s_g} w_{hji} r_{hji}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + [M_g/m_g] c_i r_{hji}), \qquad (9)$$

where c_i is an indicator variable equal to 1 when i is in the subsample and zero otherwise.

Continuing,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} (y_i + \{[M_g/m_g] c_i - 1\} r_{hji})$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_{hji}, \qquad (10)$$

where $z_{hji} = y_i + \{[M_g/m_g] c_i - 1\} r_{hji}$. Again, since every m_g is large, it is not unreasonable to assume $r_{hji} \approx r_i$ (see equation (A4) in the appendix). Thus,

$$t_{(hj)2} \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_{hji} z_i,$$

where $z_i = y_i + \{[M_g/m_g] c_i - 1\} r_i$. Using similar arguments, $t_2 \approx \sum_{g=1}^{G} \sum_{i \in S_g} w_i z_i$. Since t_2 is linear in the z_i,

$$v_{J2} \approx v_{L1}\left(\sum \sum w_i z_i\right) = \sum_{h=1}^{H} (n_h/[n_h - 1])$$

$$* \left(\sum_{j \in F_h} \left[\sum_{i \in U_{hj}} w_i z_i \right]^2 - \left[\sum_{j \in F_h} \sum_{i \in U_{hj}} w_i z_i \right]^2 / n_h \right). \qquad (11)$$

Let $e_i = M_g/m_g$ be the second-phase expansion factor for $i \in S_g$. Observe that c_i is a random variable with $E(c_i) = m_g/M_g$ and $E(c_i c_k) = (m_g/M_g)(m_g - 1)/(M_g - 1)$ for $i, k \in S_g$, $i \neq k$.

Now

$$E_2\left[\left(\sum_{i\in U_{hj}} w_i z_i\right)^2\right] \approx \left(\sum_{i\in U_{hj}} w_i y_i\right)^2 + \sum_{i\in U_{hj}} (e_i - 1)(w_i r_i)^2 - \sum_{g=1}^{G} \sum_{\substack{i,k\in S_g \cap U_{hj} \\ i\neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (12)$$

Similarly, letting F_h^* be the set of elements from selected clusters in the first-phase stratum h before subsampling, we have

$$E_2\left[\left(\sum_{j\in F_h} \sum_{i\in U_{hj}} w_i z_i\right)^2\right] = E_2\left[\left(\sum_{i\in F_h^*} w_i z_i\right)^2\right] \approx \left(\sum_{i\in F_h^*} w_i y_i\right)^2 + \sum_{i\in F_h^*} (e_i - 1)(w_i r_i)^2 - \sum_{g=1}^{G} \sum_{\substack{i,k\in S_g \cap F_h^* \\ i\neq k}} [(1 - m_g/M_g)/m_g] w_i r_i w_k r_k. \quad (13)$$

In the appendix, it is argued that under mild conditions that the last term in both equations (12) and (13) is negligible. As a result,

$$E_2(v_{J2}) \approx v_{J1} + \sum_{h=1}^{H} \sum_{i\in F_h^*} (e_i - 1)(w_i r_i)^2 = v_{J1} + \sum_{g=1}^{G} \sum_{i\in S_g} ([M_g/m_g] - 1)(w_i r_i)^2 \approx v_{L1} + E_2[(t_2 - t_1)^2], \quad (14)$$

which in turn implies that v_{J2} is a nearly unbiased estimator for $E[(t_2 - T)^2]$.

4. THE DOUBLE EXPANSION ESTIMATOR

An alternative to t_2, the *double expansion* estimator, has the form:

$$t_3 = \sum_{g=1}^{G} \sum_{i\in s_g} (M_g/m_g) w_i y_i. \quad (15)$$

The definition of a jackknife replicate for t_3 is unclear. One simple possibility is

$$t_{(hj)3} = \sum_{g=1}^{G} \sum_{i\in s_g} w_{hji} (M_g/m_g) y_i. \quad (16)$$

Another, perhaps more in the spirit of "replication", is

$$t^*_{(hj)3} = \sum_{g=1}^{G} \sum_{i\in s_g} w_{hji} (M_{ghj}/m_{ghj}) y_i, \quad (17)$$

where M_{ghj} is the number of elements in the first-phase sample (*i.e.*, in a cluster in the first-phase sample) that are in S_g but *not* U_{hj}. Similarly, m_{ghj} is the number of elements in the second-phase sample that are in s_g but *not* U_{hj}. Through counter-examples given in the appendix, we show that neither version of the replicate produces a jackknife variance estimator (v_{J3} from equation (8)) that is asymptotically unbiased in general.

5. A MONTE CARLO SIMULATION STUDY

5.1 Design of the Study

The results given so far in the text are asymptotic. In order to assess the accuracy of the jackknife as a variance estimator for the reweighted expansion estimator in a finite world, we undertook a Monte Carlo simulation study. At the same time, we assessed the accuracy of the two jackknife estimators suggested for the double expansion estimator in Section 4.

We used December 1990 Canadian Labour Force Survey (LFS) sample data for the province of Newfoundland 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) and Stukel and Boyer (1992). 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. The lowest level of stratification in Newfoundland yielded 45 strata, each of which contained less than 6 clusters or *primary sampling units* (PSU's), 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 PSU's. In collapsing the strata, economic regions were kept intact, as were the Census Metropolitan Areas of St. John's and Cornerbrook.

For the Monte Carlo study, R = 4,000 samples were drawn from the Newfoundland "population" (which was 9,152 individuals), according to the following two-phase design: within each first-phase stratum, two PSU's were selected at the first phase using simple random sampling (SRS) *with* replacement. This yielded a total of 36 PSU's. All households within selected first-phase PSU's (as well as individuals within those households) were selected, resulting in a single-stage take-all cluster sample. At the second phase, all selected first-phase elements (individuals, treating each person in a PSU selected twice as two separate individuals) were restratified according to five age categories (< = 14, 15-24, 25-44, 45-64, > = 65), and second-phase sample elements (*i.e.*, individuals) were drawn using SRS *without* replacement sampling within each of the five second-phase strata.

We varied the second-phase stratum 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. When the number of first-phase-sampled individuals in a second-phase stratum was less than our target m_g value, we planned to set $m_g = M_g$, but that event never occurred.

A popular rule of thumb for a "separate ratio estimator" such as the reweighted expansion estimator in equation (2) is that there should be at least 20 individuals within each second-phase stratum (see, for example, Särndal, Swensson and Wretman 1992, p. 270). By allowing m_g to be as small as 5 and 10, we are checking whether this rule is really necessary.

We considered two parameters of interest: T_y, the total number of employed, and T_y/T_z the employment rate. Here $T_y = \sum_{i\in U} y_i$, where $y_i = 1$ when individual i is employed; 0 otherwise. Similarly, $T_z = \sum_{i\in U} z_i$, where $z_i = 1$ when individual i is in the labour force (*i.e.*, either employed or unemployed); 0 otherwise. For each of the R = 4,000 samples, we calculated the reweighted expansion estimator (REE), t_2, given by equation (2), the double expansion estimator (DEE), t_3, given by equation (15), and the full first-phase expansion estimator (FFPE), t_1 given by equation (1). Although these estimators are defined for totals (applicable for total number of employed), it is a simple matter to extend them to ratios of totals (applicable for employment rate).

For each of the R = 4,000 second-phase samples, we calculated the jackknife variance corresponding to the reweighted expansion estimator and the double expansion estimator, given by equation (8) with $f = 2$ and $f = 3$ respectively. In the case of the double expansion estimator, we attempted both the replicates defined in equations (16) and (17), which we will refer to as variant 1 and 2, respectively.

For each of the R = 4,000 first-phase samples, we also calculated the jackknife variance corresponding to the full first-phase estimator for comparison purposes. This is given by equation (8) with $f = 1$.

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

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

$$\mathrm{PRB}(t^*) = \{[E_M(t^*)/T_y] - 1\} \times 100, \qquad (18)$$

where

$$E_M(t^*) = (1/4{,}000) \sum_{r=1}^{4{,}000} t_r^*$$

is the Monte Carlo expectation of the point estimator t^* taken over the 4,000 samples. Here t^* can be either t_1, t_2, or t_3, and t_r^* is the value of t^* for sample r.

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

$$\mathrm{PRB}[v_{Jf}(t^*)] = \{\{E_M[v_{Jf}(t^*)] - \mathrm{MSE}_{\mathrm{true}}\}/\mathrm{MSE}_{\mathrm{true}}\} \times 100, \qquad (19)$$

where

$$E_M[v_{Jf}(t^*)] = (1/4{,}000) \sum_{r=1}^{4{,}000} v_{Jfr}(t^*),$$

$$\mathrm{MSE}_{\mathrm{true}} = (1/4{,}000) \sum_{r=1}^{4{,}000} (t_r^* - T_y)^2,$$

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

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

$$\mathrm{CV}[v_{Jf}(t^*)] = \{\{(1/4{,}000)\sum [v_{Jfr}(t^*) - \mathrm{MSE}_{\mathrm{true}}]^2\}^{1/2}/\mathrm{MSE}_{\mathrm{true}}\} \times 100; \qquad (20)$$

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

5.2 Results of the Study

Table 1A gives the estimated percent relative biases of the three point estimates for the total number of employed using equation (18), and Table 1B gives the same for the employment rate. All biases are less than 1% in absolute value.

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	–	0.14	-0.3	-0.29	-0.56
DEE	–	0.16	-0.01	0.03	0.115
FFPE	0.04	–	–	–	–

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	-0.12	-0.13
FFPE	-0.09	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)

Not displayed are the Monte Carlo estimates of the mean squared errors (*i.e.*, the values of MSE_{true}) and the corresponding coefficients of variation from using either the reweighted or double expansion estimator. This is because the focus in this article is on mean squared error *estimation*. The mean squared errors (and coefficients of variation) from using the two estimators are comparable for each sample size (a relative difference in the coefficient of variation is roughly half of the corresponding relative difference in mean squared error). The reweighted expansion estimator is slightly more efficient when estimating the total number of employed individuals (*e.g.*, when $m_g = 5$, the double expansion estimator has 17% more mean squared error). There is less than a 1% difference in the mean squared errors from using the two approaches when estimating the employment rate. Not surprisingly, the mean squared errors for all estimators increase as the second-phase sample size decreases.

Table 2A gives the estimated percent relative biases of the jackknife variances for the total number of employed using equation (19), 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%. The jackknife for the reweighted expansion estimator works well, having small negative biases in the variances always less than -6%. The biases tend to become more negative (although not uniformly) as the second-phase sample sizes diminish.

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	–	–	–	–

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)	–	-0.36	4.91	9.09	30.46
FFPE	2.08	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)
Variant 1 uses the jackknife replicates in equation (16)
Variant 2 uses the jackknife replicates in equation (17)

In contrast, both jackknife variants for 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 behavior.

Table 2B repeats the analysis for the ratio estimate of employment rate. The results here are surprising since all variance estimators 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 10% in absolute value.

Overall, Table 2A and 2B provide strong support for using the jackknife variance estimator with a reweighted expansion estimator even when second-phase sample sizes are surprisingly small. By contrast, the jackknife can fail miserably for the double expansion estimator when estimating totals. Sometimes, however, variant 1 can also work reasonably well depending on the estimator and the data.

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 estimated (percent) coefficients of variation corresponding to the total number of employed and the employment rate, respectively. In equation (20), 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 estimated coefficients of variation corresponding to the reweighted expansion estimator range between 46.86% and 53.42%. Coefficients of variation of the magnitude exhibited here are typical for variance estimators, and have been encountered in other simulation studies relating to variances. See, for example, Kovačević and Yung (1997). To that end, note that even the estimated 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 all 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 estimators than it is for the estimators of the total. The very high coefficients of variation in the column $m_g = 5$ for both tables is not surprising, since the overall second-phase sample size (25) is actually smaller than the number of PSU's drawn in the first phase of sampling (36). In fact, a

Table 3A
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	49.3	46.86	53.42
DEE (Variant 1)	–	*	*	*	*
DEE (Variant 2)	–	*	*	*	*
FFPE	56.71	–	–	–	–

Table 3B
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	99.1
DEE (Variant 2)	–	60.94	73.2	92.71	*
FFPE	78.42	–	–	–	–

REE - Reweighted Expansion Estimator (t_2)
DEE - Double Expansion Estimator (t_3)
FFPE - Full First Phase Estimator (t_1)
Variant 1 uses the jackknife replicates in equation (16)
Variant 2 uses the jackknife replicates in equation (17)

more relevant realized sample count for the ratio estimator is the number of sampled individuals in the labour force (*i.e.*, in the denominator). This value varies from sample to sample and is often considerably less than 25.

6. EXTENDING THE REWEIGHTED EXPANSION ESTIMATOR

6.1 The Reweighted Expansion Estimator

It is not that difficult to develop a linearization variance estimator for the reweighted expansion estimator in equation (2). Suppose, however, one had a sample design with more than two phases or was interested in estimating the ratio of two totals. Linearization, although still possible, becomes increasingly cumbersome. The jackknife, on the other hand, does not.

It is a simple matter to generalize the results in Section 3 to p-phase sampling by induction. The h still refer the first-phase strata, but the g now denote the p-th-phase strata; S_g is the set of elements in the (p-1)th-phase sample from stratum g while s_g is the pth-phase subsample from g. The w_i in equation (2) are replaced with the a_i from (3) for the (p-1)th-phase estimator. Similarly, the $t_{(hj)2}$ in the jackknife are computed using a_{hji} from the (p-1)th phase in place of the w_{hji}.

It is also a simple matter (left to the reader) to replace the stratified cluster sample in the first phase of selection with a stratified multi-stage sample. The results in Section 3 follow as long as the first stage of the multi-stage sample is drawn with replacement.

Finally, it is not difficult to extend the results of Section 3 to more complicated estimators. Let $\boldsymbol{U}_2$ be a vector of estimators each in the form of t_2 from equation (2). The mean squared error of any estimator $\Theta = g(\boldsymbol{U}_2)$, where g is a smooth function, can be estimated with a jackknife in a nearly unbiased manner whenever the members of $\boldsymbol{U}_2$ can be. This follows the proofs in the literature. Rao and Wu (1985), for example, address the asymptotic framework where the n_h are all bounded, while Wolter (1985; Chapter 4.5) treats the case where the n_h grow arbitrarily large.

6.2 Regression in the Second Phase

The estimator t_2 can be generalized into the regression estimator:

$$t_{2\text{reg}} = \sum_{i\in S} w_i \boldsymbol{x}_i \left(\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i\in s} w_i e_i d_i \boldsymbol{x}_i' y_i \right), \quad (21)$$

where S denotes the original sample, $\boldsymbol{x}_i$ is a row vector, d_i is a scalar, and there exists a row vector $\boldsymbol{\gamma}$ such that $d_i \boldsymbol{\gamma} \boldsymbol{x}_i' = 1$ for all i. In practice, d_i is usually 1 for all i. A popular exception occurs when $\boldsymbol{x}_i = x_i$ and $d_i = 1/x_i$. In equation (2), $d_i = 1$ for all i, and $\boldsymbol{x}_i$ is a G-vector with a value of 1 in the g-th position and 0's elsewhere for $i \in S_g$.

Let

$$r_i = y_i - \boldsymbol{x}_i \left(\sum_{i\in S} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i\in S} w_i d_i \boldsymbol{x}_i' y_i \right).$$

The replicate $t_{2\text{reg}(hj)}$ has the same form as $t_{2\text{reg}}$ except that w_{hji} replaces w_i everywhere. Similarly, r_{hji} has the same form as r_i except that w_{hji} replaces w_i. Note that the e_i are unchanged from $t_{2\text{reg}}$ to $t_{2\text{reg}(hj)}$.

Since the sampling design hasn't changed, most of equation (6) stays as is except that now $(\sum_{i\in S_g} w_i r_i)^2$ is nonnegative rather than strictly zero. The interested reader can verify that equations (10) through (13) remain in their present form. It turns out that the jackknife has, if anything, an (approximate) upward bias in equation (14). That is to say, the jackknife is a *conservative* estimator of variance. Again, see the apppendix (equations (A6) through (A9)) for a formal statement of the asymptotic assumptions.

The bias in the jackknife disappears when $\sum_{i\in S_g} w_i r_i = 0$ for all g. Formally, this will happen when there exists G row vectors $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_G$ such that $d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' = 1$ when $i \in S_g$ and 0 otherwise (since $\sum_{i\in s} w_i r_i = \sum_{i\in s} d_i \boldsymbol{\gamma}_g \boldsymbol{x}_i' w_i r_i = \boldsymbol{\gamma}_g \sum_{i\in s} w_i d_i \boldsymbol{x}_i' r_i = \boldsymbol{\gamma}_g \{\sum_{i\in s} w_i d_i \boldsymbol{x}_i' (y_i - \boldsymbol{x}_i [\sum_{i\in s} w_i d_i \boldsymbol{x}_i' \boldsymbol{x}_i]^{-1} \sum_{i\in s} w_i d_i \boldsymbol{x}_i' y_i)\} = 0$). When all $d_i = 1$, the existence of $\boldsymbol{\gamma}_g$

means that either one member of x_i is an indicator variable equal to 1 when $i \in S_g$ and 0 otherwise, or one member of a linear transform of x_i is such an indicator variable.

7. CONCLUDING REMARKS

The main purpose of this paper was to show that a simple jackknife variance estimator can be nearly unbiased for an estimation strategy involving two-phase sampling as long as that strategy employs a reweighted expansion estimator and not a double expansion estimator. Since the theoretical results for the reweighted expansion estimator rely on asymptotic arguments, their practical application will depend on the context. Nevertheless, a Monte Carlo simulation study performed here suggests that the jackknife can be an effective estimator for the variance of a reweighted expansion estimator even with surprisingly small second-phase stratum sample sizes, that is, sizes of 5 and 10.

APPENDIX

The Design Consistency of the Reweighted Expansion Estimator

To establish the design consistency of t_2 in equation (2) it is sufficient to assume that the sample design and population values of the y_i are such that

$$\left\{\sum_{g=1}^{G} (M_g/m_g)\sum_{i\in s_g} w_i y_i/T\right\} - 1 = O_p(1/\sqrt{m}),$$

and, given *any* first-phase sample,

$$\left(\sum_{k\in S_g} w_k \Big/ \sum_{k\in s_g} w_k\right)(m_g/M_g) - 1 = O_p(1/\sqrt{m}) \qquad \text{(A1)}$$

for all g. These assumptions justify equation (5) in the text.

We assume in our analysis that G is bounded and that each m_g has the same asymptotic order as m. This is only possible when the S_g are determined *after* the first-phase sample has been drawn. Otherwise, the M_g would be random variables, and a minimum size for each m_g could not be guaranteed for all possible first-phase samples. In principle, we are assuming the existence of a mechanism for determining the S_g and the second-phase sampling fractions given any first-phase sample. By contrast, the exact values of G and the m_g can but need not be fixed before the first-phase sample is drawn.

A Comment on the Asymptotic Framework

Recall that the text showed that the jackknife contains a component that estimates the second-phase variance (*i.e.*, $E_2[(t_2 - t_1)^2]$) in an asymptotically unbiased manner given *any* first- phase sample (see equation (14)). As a result, that component also estimates the average (*i.e.*, unconditional) second-phase variance across all possible first-phase samples (*i.e.*, $E_1\{E_2[(t_2 - t_1)^2]\}$) in an asymptotically unbiased manner.

In our empirical work, we strayed from the sampling framework described above so that the results could be easily summarized. In particular, we defined the S_g beforehand, and let the M_g be random. When the first-phase sample was such that M_g was less than the desired m_g (say 50) in some second-phase stratum, we planned to choose all the individuals in S_g for the second-phase sample. As a result, there would be no contribution to the mean squared error (or bias) of t_2 from second-phase stratum g when that particular first-phase sample was selected, and so no asymptotic assumptions about m_g would be necessary. As it happened, in no simulation was M_g actually less than 50. Nevertheless, a decision rule about the second-phase sampling fractions was in place for every possible first-phase sample.

Jackknife Replicates

There are (at least) two distinct asymptotic frameworks for the first-phase sample. In the first, there is an arbitrarily large number of first-phase strata each of which is bounded in size; that is, each $1/n_h = O(1)$ while $1/H = O(1/m)$. In the second, all the first-phase strata are arbitrarily large; that is, $1/n_h = O(1/m)$. Under either framework, we assume that the number of elements in each cluster is O(1); that is to say, bounded.

Since every m_g is of the same asymptotic order as m, it is not unreasonable to assume under either regime that, given any first-phase sample,

$$\sum_{i\in S_g} w_{hji} \Big/ \sum_{i\in S_g} w_i - 1 = O_p(1/m), \qquad \text{(A2)}$$

and

$$\sum_{i\in s_g} w_{hji} \Big/ \sum_{i\in s_g} w_i - 1 = O_p(1/m), \qquad \text{(A3)}$$

which can be used to establish equation (9). Similarly, we assume that given any first-phase sample

$$\sum_{i\in S_g} w_{hji} y_i \Big/ \sum_{i\in S_g} w_i y_i - 1 = O_p(1/m), \qquad \text{(A4)}$$

which assures us that $r_{hji} - r_i = O_p(1/m)$.

Equations (12), (13), and (14)

Since the number of elements in each cluster is bounded, say by B. The third term on the right hand side of equation (12) has at most GB^2 terms, a bounded number.

Each of these terms is of order $1/m_g$ (formally, the probability that any one term is of asymptotic order greater than $1/m_g$ is zero). Consequently, the second line of equation (12) is asymptotically ignorable.

Equation (14) holds when each $1/n_h = O(1)$, because if each n_h is less than C (say), then the third term on the right hand side of equation (13) will be the sum of at most $G(BC)^2$ terms, a bounded number. Each of these terms is again of order $1/m_g$. Consequently, the second line of equation (13) is asymptotically ignorable.

Alternatively, suppose each $1/n_h$ were $O(1/m)$. We will assume that the sample design and population is such that, given any first-phase sample,

$$A_h = \sum_{i \in F_h^*} w_i(e_i c_i - 1) r_i \Big/ \sum_{i \in F_h^*} w_i y_i = O_p(1/\sqrt{m}) \qquad \text{(A5)}$$

for all h. To see why this is a reasonable assumption, observe that conditioned on the first-phase sample, the denominator of A_h is a domain total – the sum of the $w_i y_i$ among the elements in F_h^*. Consequently, it is $O(m)$ (without loss of generality we can assume that all the w_i are $O(1)$). The numerator of A_h is the difference between an expansion estimator (the sum of the $w_i e_i c_i r_i$ in F_h^*) based on a stratified simple random sample and its target (the sum of the $w_i r_i$ in F_h^*). Equation (A.5) makes the modest assumption that the sampling design and population is such that this difference is $O_p(\sqrt{m})$ for every possible first-phase sample.

Under assumption (A5), $\sum_{i \in F_h^*} w_i z_i = \sum_{F_h^*} w_i y_i (1 + A_h)$ is approximately equal to $\sum_{i \in F_h^*} w_i y_i$, which implies $E_2[(\sum_{i \in F_h^*} w_i z_i)^2]/n_h \approx (\sum_{i \in F_h^*} w_i y_i)^2/n_h$. Equation (14) follows from this near equality and from equations (11) and (12) (since n_h is large, $n_h/(n_h - 1) \approx 1$).

Counter-examples to the Jackknifes for the Double Expansion Estimator

As a counter-example to the replicate form in equation (16), consider the situation where each cluster contains a single element, $H = G = 1$, and all the y_i values are equal to 1. As a result, $t_3 = T$, which means that t_3 has no variance. Unfortunately $t_{(1j)3} = T[n_1/(n_1 - 1)](m - 1)/m$ when $j \in s$ and $Tn_1/(n_1 - 1)$ otherwise. Thus, $(t_{(1j)3} - T)/T = O_p(1/m)$. Now v_{J3}/T^2 computed from the $t_{(1j)3}$ would also be $O(1/m)$ since it is the sum of n_1 terms of order $O(1/m^2)$.

Although v_{J3}/T^2 is $O(1/m)$, v_{J3} is not close enough to zero for our purposes. To see why, observe that if the y_i were all $N(1,1)$, then the relative variance of t_3 would be $1/m$, which is also $O(1/m)$. Thus, for v_{J3} to be nearly zero, v_{J3}/T^2 would have to be smaller than $O(1/m)$. It is not, and the jackknife variance estimator is not nearly unbiased.

As a counter-example to the replicate form in equation (17), consider the situation where each cluster is again a single element and all y_i values are equal to 1, but now $H = m$, $G = 1$, the population size in each h is N_0, $n_h = 2$ for all h, and $M_1 = 2m$. As a result, $T = t_3 = mN_0$, so that t_3 has no variance. The replicate $t^*_{(hj)3}$ can take on four possible values. If $hj \in s$ and $hj' \in s (j \neq j')$, then $t^*_{(hj)3} = [(m/2)(2m - 1)/(m - 1)]N_0$. If $hj \in s$ and $hj' \notin s$, then $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/(m - 1)]N_0$. If $hj \notin s$ and $hj' \in s$, then $t^*_{(hj)3} = [(m/2)(2m - 1)/m]N_0$. If $hj \notin s$ and $hj' \notin s$, then $t^*_{(hj)3} = [(\{m - 1\}/2)(2m - 1)/m]N_0$. In all cases, $(t^*_{(hj)3} - T)/T = O_p(1/m)$, and so the jackknife variance estimator fails to be nearly unbiased.

The Two-phase Regression Estimator

To support the arguments in the text about the regression estimator in equation (21), we assume the sampling design and population values are such that the following asymptotic relationships hold. First,

$$\sum_{i \in S} w_i \boldsymbol{x_i} (\sum_{i \in s} w_i e_i d_i \boldsymbol{x_i}' \boldsymbol{x_i})^{-1} d_i \boldsymbol{x_i}' - 1 = O_p(1/\sqrt{m}), \qquad \text{(A6)}$$

which is a generalization of equation (A1). Likewise, equations (A2) and (A3) generalize to

$$\sum_{i \in S_g} w_{hji} d_i q_i \Big/ \sum_{i \in S_g} w_i d_i q_i - 1 = O_p(1/m), \qquad \text{(A7)}$$

and

$$\sum_{i \in s_g} w_{hji} e_i d_i q_i \Big/ \sum_{i \in s_g} w_i e_i d_i q_i - 1 = O_p(1/m) \qquad \text{(A8)}$$

for all q_i, where q_i is an element of the matrix $\boldsymbol{x_i}' \boldsymbol{x_i}$. Finally, the assumption in equation (A4) generalizes to

$$\sum_{i \in S_g} w_{hji} d_i p_i \Big/ \sum_{i \in S_g} w_i d_i p_i - 1 = O_p(1/m) \qquad \text{(A9)}$$

for all p_i, where p_i is an element of the matrix $\boldsymbol{x_i}' y_i$.

REFERENCES

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOVAČEVIĆ, M.S., and YUNG, W. (1997). Variance estimation for measures of income inequality and polarization – an empirical study. *Survey Methodology*, 23, 1, 41-52.

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.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. *Incomplete Data and Sample Surveys, Volume 2: Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin). New York: Academic Press, 143-184.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4, 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.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey:* 1984-1990. Catalogue No. 71-526, Statistics Canada.

STUKEL, D.M., and BOYER, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. Methodology Branch Working Paper, SSMD, 92-009E. Statistics Canada.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.