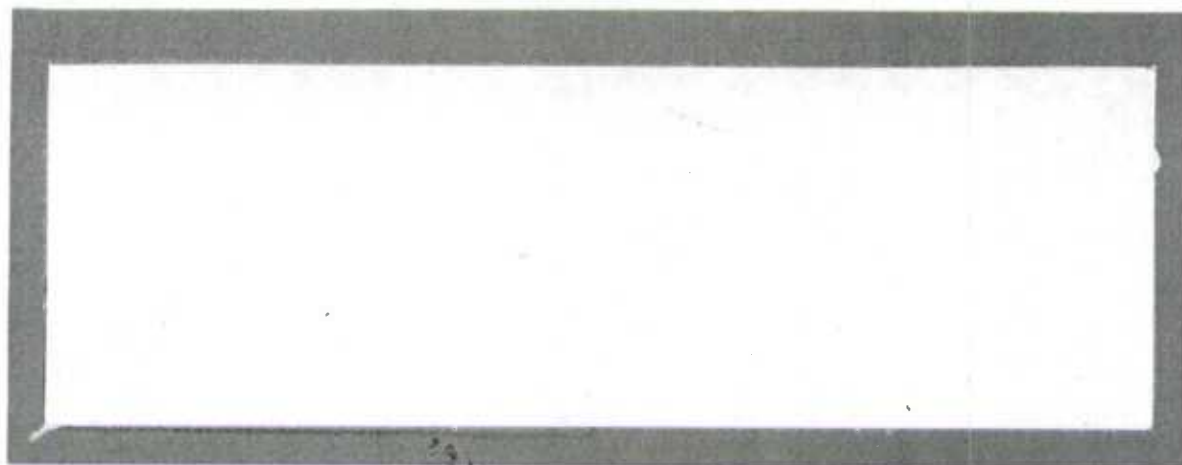


11-619E  
no.95-02  
c.1



Statistics  
Canada

Statistique  
Canada



Methodology Branch

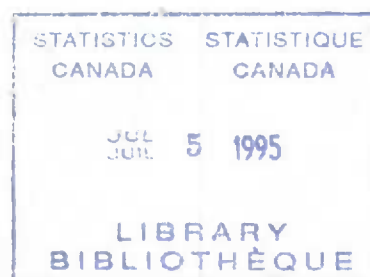
Household Survey  
Methods Division

Direction de la méthodologie

Division des méthodes  
d'enquêtes-ménages

Canada

**WORKING PAPER  
METHODOLOGY BRANCH**



**VARIANCE ESTIMATION IN  
STRATIFIED MULTISTAGE SAMPLING**

HSMD-95-002-E

**NOT FOR LOAN  
NE S'EMPRUNTE PAS**

Wesley Yung and J.N.K. Rao

Methods Development and Analysis Section  
Household Survey Methods Division, Statistics Canada  
and  
Carleton University

April 1995

---

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

# VARIANCE ESTIMATION IN STRATIFIED MULTISTAGE SAMPLING

Wesley Yung and J.N.K. Rao

Methods Development and Analysis Section

Household Survey Methods Division

Methodology Branch, Statistics Canada

and

Carleton University

## ABSTRACT

A robust Taylor linearization variance estimator under stratified multistage sampling and generalized regression estimation is proposed. The robust variance estimator is obtained by linearizing the jackknife variance estimator. Properties of the proposed variance estimator, the standard linearized variance estimator, and the jackknife variance estimator are studied through a simulation study. All of the variance estimators performed well both unconditionally and conditionally given a measure of how far away the estimated totals of auxiliary variables are from the known population totals.

# ESTIMATION DE LA VARIANCE DANS L'ÉCHANTILLONNAGE MULTIPLE STRATIFIÉ

Wesley Yung et J.N.K. Rao

Section du développement des méthodes et de l'analyse

Division des méthodes d'enquêtes-ménages

Direction de la méthodologie, Statistique Canada

et

Université Carleton

## RÉSUMÉ

On propose ici un estimateur de la variance de linéarisation de Taylor robuste pour l'échantillonnage multiple stratifié et l'estimation de régression généralisée. L'estimateur de la variance robuste est obtenu par la linéarisation de l'estimateur de la variance jackknife. Dans une étude de simulation, on examine les propriétés de l'estimateur de la variance proposé, de l'estimateur de la variance linéarisé ordinaire et de l'estimateur de la variance jackknife. Tous ces estimateurs de la variance ont donné de bons résultats, avec ou sans conditions, avec une mesure de la distance qui sépare les totaux estimés pour les variables auxiliaires des totaux de la population connue.

# VARIANCE ESTIMATION IN STRATIFIED MULTISTAGE SAMPLING

Wesley Yung and J.N.K. Rao

## 1. Introduction

Large-scale sample surveys often use stratified multistage designs with large numbers of strata,  $L$ , and relatively few primary sampling units (clusters),  $n_h (\geq 2)$ , sampled within each stratum. We assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals,  $Y_{hi}, i = 1, \dots, n_h; h = 1, \dots, L$ .

From the specification of the survey design, basic weights  $w_{hik} (> 0)$ , attached to the  $(hik)^{th}$  element (ultimate unit), are obtained. Often these basic weights  $w_{hik}$  are subjected to post-stratification adjustment to ensure consistency with known totals of post-stratification variables. In the case of a single post-stratifier, the weights are ratio-adjusted to the known population counts (e.g., age-sex counts). To handle two or more post-stratifiers with known marginal population counts, the weights  $w_{hik}$  can be calibrated through generalized regression (see section 4), as in the Canadian Labour Force Survey (LFS).

The LFS uses the jackknife method for estimating the variance of the generalized regression estimator. This method is computer-intensive but it is known to possess good conditional properties, given the estimated post-stratification counts. For example, in the context of simple random sampling, Royall and Cumberland (1981) have shown that the jackknife variance estimator is approximately equal to a robust Taylor linearization variance estimator.

The main purpose of this paper is to present a robust Taylor linearization variance estimator under stratified multistage sampling and generalized regression estimation. This is obtained by linearizing the jackknife variance estimator. In the case of a ratio-adjusted post-stratified estimator, this variance estimator is identical to Rao's (1985) variance estimator. The proposed linearization variance estimator is computationally simple and can

be implemented using software packages that employ the linearization method, such as PC CARP, SUDAAN or Statistic Canada's Generalized Estimation System (GES).

Section 2 will introduce the jackknife variance estimator for the basic expansion estimator of the total,  $Y$ . Section 3 will present the poststratified estimator along with its jackknife and linearized jackknife variance estimators. These results will be extended to the case where there are two or more poststratification variables using the generalized regression estimator in section 4. Section 5 deals with variance estimation for a ratio of two totals, both of which have been estimated using a generalized regression estimator. Finally, results of a simulation study will be presented in section 6.

## 2. Basic Estimator

Using the basic weights  $w_{hik}$ , an unbiased estimator of the population total  $Y$  is of the form

$$\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}, \quad (2.1)$$

where  $s$  denotes the sample of elements and  $y_{hik}$  is the value of the characteristic of interest associated with the sample element  $(hik) \in s$ . For simplicity, we assume complete response in this paper.

It is common practice to sample clusters without replacement. However, at the stage of variance estimation, the calculations are greatly simplified by treating the sample as if the clusters are sampled with replacement. This approximation generally leads to overestimation of the variance of  $\hat{Y}$ , but the relative bias is likely to be small if the first-stage sampling fractions are small.

An estimator of the variance of  $\hat{Y}$  is simply given by

$$v(\hat{Y}) = \sum_{h=1}^L \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2 = v(r_{hi}), \quad (2.2)$$

where  $r_{hi} = \sum_k (n_h w_{hik}) y_{hik}$ , and  $\bar{r}_h = \frac{1}{n_h} \sum_i r_{hi}$ . The operator notation  $v(r_{hi})$  denotes that  $v(\hat{Y})$  depends only on the  $r_{hi}$ 's.

To introduce the jackknife method, we need the estimator  $\hat{Y}_{(gj)}$  for each  $(gj)$  obtained from the sample after omitting the data from the the  $j^{\text{th}}$  sampled cluster in the  $g^{\text{th}}$  stratum ( $j = 1, \dots, n_g; g = 1, \dots, L$ ). It is simply obtained from (2.1) by letting  $w_{gjk} = 0$ , changing  $w_{gik} (i \neq j)$  to  $n_g w_{gik} / (n_g - 1)$  and retaining the original weights  $w_{hik}$  for  $h \neq g$ , i.e.,

$$w_{hik(gj)} = \begin{cases} 0 & \text{if } (hi) = (gj) \\ \frac{n_h}{(n_h-1)} w_{hik} & \text{if } h = g \text{ and } i \neq j \\ w_{hik} & \text{if } h \neq g \end{cases}.$$

These jackknife weights,  $w_{hik(gj)}$ , are calculated for each cluster  $(gj)$  and

$$\hat{Y}_{(gj)} = \sum_{(hik) \in s} w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then given by

$$v_J(\hat{Y}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} [\hat{Y}_{(gj)} - \hat{Y}]^2. \quad (2.3)$$

The variance estimator (2.3) is applicable to general statistics, say  $\hat{\theta} = g(\hat{Y})$ , by simply replacing  $\hat{Y}_{(gj)}$  and  $\hat{Y}$  with  $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$  and  $\hat{\theta}$  respectively. In the linear case,  $\hat{\theta} = \hat{Y}$ , the jackknife variance estimator is identical to the customary variance estimator (2.2).

### 3. Post-Stratified Estimator

Suppose the population is partitioned into  $C$  poststrata with known population counts  ${}_cM, c = 1, \dots, C$ . We will use the prescript  $c$  to denote post-strata. An estimator of  ${}_cM$  is given by

$${}_c\hat{M} = \sum_{(hik) \in {}_c s} w_{hik}, \quad (3.1)$$

where  ${}_c s$  is the sample of elements belonging to the  $c$ -th post-stratum. Similarly, an estimator of the post-stratum total  ${}_c Y$  is

$${}_c\hat{Y} = \sum_{(hik) \in {}_c s} w_{hik} y_{hik}. \quad (3.2)$$

Using the estimators  ${}_c\hat{Y}$  and  ${}_c\hat{M}$ , we obtain a post-stratified estimator of the total  $Y$  as

$$\hat{Y}_{ps} = \sum_c \frac{{}_c\hat{M}}{{}_c\hat{M}} {}_c\hat{Y}. \quad (3.3)$$

We can rewrite (3.3) as

$$\hat{Y}_{ps} = \sum_c \sum_{(hik) \in {}_c s} {}_c w_{hik} y_{hik} \quad (3.4)$$

where  ${}_c w_{hik} = w_{hik}({}_c\hat{M}/{}_c\hat{M})$  is the ratio-adjusted weight for  $(hik) \in {}_c s$ . If  $y_{hik}$  is the indicator variable for a post-stratum, say  $c$ , then  $\hat{Y}_{ps} = {}_c\hat{M}$ , thus ensuring consistency with known totals,  ${}_c\hat{M}$ .

The customary Taylor linearization variance estimator is given by (2.2) with  $r_{hi}$  changed to

$$\tilde{r}_{hi} = \sum_c \sum_{k \in {}_c s} (n_h w_{hik}) {}_c e_{hik},$$

where  ${}_c e_{hik} = y_{hik} - {}_c\hat{Y}/{}_c\hat{M}$  for the  $k^{th}$  element in the  $(hi)^{th}$  cluster belonging to  ${}_c s$ , i.e.,

$$v_L(\hat{Y}_{ps}) = v(\tilde{r}_{hi}) \quad (3.5)$$

Rao (1985) proposed a robust linearization variance estimator using the ratio-adjusted weights  ${}_c w_{hik}$ :

$$v_R(\hat{Y}_{ps}) = v(r_{hi}^*) \quad (3.6)$$

where

$$r_{hi}^* = \sum_c \sum_{k \in {}_c s} (n_h {}_c w_{hik}) {}_c e_{hik}.$$

Turning to the jackknife method, we need to recalculate the post-stratification weights  ${}_c w_{hik}$  each time a cluster  $(gj)$  is deleted. This is done by using the jackknife weights  $w_{hik(gj)}$  in (3.1) to get  ${}_c\hat{M}_{(gj)}$  and then using  ${}_c w_{hik(gj)} = ({}_c\hat{M}/{}_c\hat{M}_{(gj)})w_{hik(gj)}$  to get

$$\hat{Y}_{ps(gj)} = \sum_c \sum_{(hik) \in {}_c s} {}_c w_{hik(gj)} y_{hik}.$$

The jackknife variance estimator is then obtained as

$$v_J(\hat{Y}_{ps}) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} [\hat{Y}_{ps(gj)} - \hat{Y}_{ps}]^2. \quad (3.7)$$

By linearizing the jackknife variance estimator (3.7), we obtain a robust variance estimator which is identical to Rao's robust variance estimator (3.6). In the important special case of  $n_h = 2$  clusters per stratum, (3.6) and (3.7) are in fact asymptotically equal to higher order terms, as the number of strata  $L$  increases. Proofs of these results will be given in W. Yung's Ph.D thesis.

#### 4. Generalized Regression Estimator

To handle several post-stratifiers with known marginal population counts, we can use a generalized regression estimator of  $Y$  by using indicator auxiliary variables to denote the categories of the post-stratifiers (Huang and Fuller, 1978; Deville and Särndal, 1992).

Let  $\mathbf{x}_{hik}$  be a vector of auxiliary variables with known population totals  $\mathbf{X}$ . The generalized regression estimator of  $Y$  is then given by

$$\hat{Y}_r = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}, \quad (4.1)$$

where

$$\hat{\mathbf{X}} = \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik},$$

and  $\hat{\mathbf{B}}$  is the vector of estimated regression coefficients

$$\begin{aligned} \hat{\mathbf{B}} &= \left[ \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T \right]^{-1} \left[ \sum_{(hik) \in s} w_{hik} \mathbf{x}_{hik} y_{hik} \right] \\ &= \hat{\mathbf{A}}^{-1} \hat{\mathbf{U}}, \end{aligned}$$

where

$$\hat{\mathbf{A}}^{-1} = \sum_{(hik) \in s} w_{hik} \mathbf{V}_{hik},$$

and

$$\hat{\mathbf{U}} = \sum_{(hik) \in s} w_{hik} \mathbf{u}_{hik}$$

with  $\hat{\mathbf{V}}_{hik} = \mathbf{x}_{hik} \mathbf{x}_{hik}^T$ , and  $\mathbf{u}_{hik} = \mathbf{x}_{hik} y_{hik}$ .

It is readily verified that,

$$\hat{\mathbf{X}}_r = \mathbf{X},$$

thus ensuring consistency with known totals  $\mathbf{X}$ .

The post-stratified estimator,  $\hat{Y}_{ps}$ , is a special case of (4.1) by letting  $\mathbf{x}_{hik}$  denote the vector of indicator variables for the post-strata. In this case,  $\hat{\mathbf{X}} = ({}_1\hat{M}, \dots, {}_C\hat{M})^T$ ,  $\mathbf{X} = ({}_1M, \dots, {}_CM)^T$ , and  $\hat{\mathbf{B}} = ({}_1\hat{R}, \dots, {}_C\hat{R})^T$  with  ${}_c\hat{R} = {}_c\hat{Y}/{}_c\hat{M}$ . Thus,

$$\hat{Y}_r = \hat{Y} + \sum_c {}_c\hat{R}({}_cM - {}_c\hat{M}) = \hat{Y}_{ps}.$$

The generalized regression estimator may be rewritten as

$$\hat{Y}_r = \sum_{(hik) \in s} w_{hik}^* y_{hik}, \quad (4.2)$$

where

$$w_{hik}^* = w_{hik} a_{hik} \quad (4.3)$$

is the calibrated weight with

$$a_{hik} = 1 + \mathbf{x}_{hik}^T \hat{\mathbf{A}}^{-1} (\mathbf{X} - \hat{\mathbf{X}}).$$

In the special case of  $\hat{Y}_{ps}$ , we have  $a_{hik} = {}_cM/{}_c\hat{M}$  for  $(hik) \in {}_cs$ .

Turning to variance estimation, the customary Taylor linearization variance estimator is again given by (2.2) with  $r_{hi}$  changed to

$$\tilde{r}_{hi} = \sum_k (n_h w_{hik}) e_{hik},$$

where

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}, \quad (4.4)$$

i.e.

$$v_L(\hat{Y}_r) = v(\tilde{r}_{hi}). \quad (4.5)$$

It may be noted that the  $e_{hik}$  may be rewritten as

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{A}}^{-1} \hat{\mathbf{U}}.$$

This alternative form of  $e_{hik}$  may be computationally more convenient.

For the jackknife method we need to recalculate the calibration weights  $w_{hik}^*$  each time a cluster  $(gj)$  is deleted. These weights are given by

$$w_{hik(gj)}^* = w_{hik(gj)} a_{hik(gj)}, \quad (4.6)$$

where

$$a_{hik(gj)} = 1 + \mathbf{x}_{hik}^T \hat{\mathbf{A}}_{(gj)}^{-1} (\mathbf{X} - \hat{\mathbf{X}}_{(gj)}),$$

$$\hat{\mathbf{A}}_{(gj)} = \sum_{(hik)\epsilon s} w_{hik(gj)} \mathbf{V}_{hik},$$

and

$$\hat{\mathbf{X}}_{(gj)} = \sum_{(hik)\epsilon s} w_{hik(gj)} \mathbf{x}_{hik}.$$

Denote the resulting generalized regression estimator as

$$\begin{aligned} \hat{Y}_{r(gj)} &= \sum_{(hik)\epsilon s} w_{hik(gj)}^* y_{hik} \\ &= \hat{Y}_{(gj)} + (\mathbf{X} - \hat{\mathbf{X}}_{(gj)})^T \hat{\mathbf{B}}_{(gj)} \end{aligned}$$

where  $\hat{\mathbf{B}}_{(gj)}$  is the vector of estimated regression coefficients when the  $(gj)^{th}$  cluster is deleted.

The jackknife variance estimator of  $\hat{Y}_r$  is then given by

$$v_J(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} [\hat{Y}_{r(gj)} - \hat{Y}_r]^2. \quad (4.7)$$

Using a Taylor series expansion, we linearize the function

$$f(\hat{Y}_{(gj)}, \hat{\mathbf{X}}_{(gj)}, \hat{\mathbf{B}}_{(gj)}) = \hat{Y}_{(gj)} + (\mathbf{X} - \hat{\mathbf{X}}_{(gj)})^T \hat{\mathbf{B}}_{(gj)}$$

about the full sample estimates  $(\hat{Y}, \hat{\mathbf{X}}, \hat{\mathbf{B}})$  to obtain the linearized jackknife variance estimator. In linearizing (4.7) we use the following relationships:

$$\hat{Y}_{(gj)} = \hat{Y} + \frac{1}{n_g - 1} \left( \sum_{(gik)\epsilon s} w_{gik} y_{gik} - n_g \sum_{(gjk)\epsilon s} w_{gjk} y_{gjk} \right), \quad (4.8)$$

$$\hat{\mathbf{X}}_{(gj)} = \hat{\mathbf{X}} + \frac{1}{n_g - 1} \left( \sum_{(gik)\epsilon s} w_{gik} \mathbf{x}_{gik} - n_g \sum_{(gjk)\epsilon s} w_{gjk} \mathbf{x}_{gjk} \right), \quad (4.9)$$

and

$$\hat{\mathbf{B}}_{(gj)} \approx \hat{\mathbf{B}} + \hat{\mathbf{A}}^{-1}(\mathbf{c}_{gj} - \mathbf{D}_{gj}\hat{\mathbf{B}}), \quad (4.10)$$

where

$$\mathbf{c}_{gj} = \frac{1}{n_g - 1}(\bar{\mathbf{u}}_g - \mathbf{u}_{gj}),$$

with  $\mathbf{u}_{gj} = \sum_k (n_g w_{gjk}) \mathbf{u}_{gjk}$  and

$$\mathbf{D}_{gj} = \frac{1}{n_g - 1}(\mathbf{V}_g - \mathbf{V}_{gj}),$$

with  $\mathbf{V}_{gj} = \sum_k (n_g w_{gjk}) \mathbf{V}_{gjk}$ . Relationship (4.10) can be obtained using the matrix equality

$$(I + PQ)^{-1} = I - P(I + QP)^{-1}Q.$$

This leads to a robust linearization variance estimator

$$v_{JL}(\hat{Y}_r) = v(r_{hi}^*) \quad (4.11)$$

with

$$r_{hi}^* = \sum_k (n_h w_{hik}^*) e_{hik}$$

where  $w_{hik}^*$  is defined in (4.3) and  $e_{hik}$  is defined in (4.4). Details of the proof will be given in W. Yung's Ph.D thesis. It is interesting to note that (4.11) is similar to the model-assisted variance estimator proposed by Särndal et al. (1989) in the context of a superpopulation model appropriate for unistage sampling.

The computation of the jackknife variance estimator involves the inversion of the matrix  $\hat{\mathbf{A}}_{(gj)}$  for each  $(gj)$ . This can be avoided by retaining the inverse for the full sample,  $\hat{\mathbf{A}}^{-1}$ , and then using modified weights

$$\tilde{w}_{hik(gj)} = w_{hik(gj)} \tilde{a}_{hik(gj)} \quad (4.12)$$

with

$$\tilde{a}_{hik(gj)} = 1 + (w_{hik}/w_{hik(gj)}) \mathbf{x}_{hik}' \hat{\mathbf{A}}^{-1}(\mathbf{X} - \hat{\mathbf{X}}_{(gj)}).$$

The resulting estimator of  $Y$ , when the  $(gj)$ -th cluster is deleted, is given by

$$\tilde{Y}_{r(gj)} = \sum_{(hik) \in s} \tilde{w}_{hik(gj)} y_{hik}$$

and the corresponding jackknife variance estimator is

$$v_{J1}(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} [\tilde{Y}_{r(gj)} - \hat{Y}_r]^2. \quad (4.13)$$

It is readily seen that (4.13) is exactly equal to the customary variance estimator (4.5). However, an advantage of the jackknife method is that it is readily applicable to nonlinear statistics unlike the Taylor linearization method. An example of this is given in the next section.

## 5. Estimation of a Ratio

Often a ratio of two estimated totals is required. For example, in a family expenditure survey, one may be interested in the proportion of income spent on clothing. Let

$$\hat{Y}_r = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}_1$$

be a generalized regression estimator of the total amount spent on clothing,  $Y$ . Similarly, let

$$\hat{Z}_r = \hat{Z} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}_2$$

be a generalized regression estimator of the total income,  $Z$ . The proportion of interest is  $\theta = Y/Z$ , and can be estimated by

$$\hat{\theta} = \hat{Y}_r / \hat{Z}_r.$$

The jackknife variance estimator is simply given by

$$v_J(\hat{\theta}) = \sum_g \frac{n_g - 1}{n_g} \sum_j [\hat{\theta}_{(gj)} - \hat{\theta}]^2 \quad (5.1)$$

where

$$\hat{\theta}_{(gj)} = \hat{Y}_{r(gj)} / \hat{Z}_{r(gj)}.$$

Linearizing (5.1), we obtain a robust linearization variance estimator

$$v_{JL}(\hat{\theta}) = v(r_{hi}^{**})$$

where

$$r_{hi}^{**} = \frac{1}{\hat{Z}_r} \sum_k (n_h w_{hik}^*) e_{hik}^*$$

with

$$e_{hik}^* = e_{hik} - \frac{\hat{Y}_r}{\hat{Z}_r} \tilde{e}_{hik},$$

$$e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_1$$

and

$$\tilde{e}_{hik} = z_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_2.$$

## 6. Simulation Study

A simulation study was conducted to investigate the finite sample properties (both conditional and unconditional) of the jackknife and the linearized jackknife variance estimators for the poststratified estimator,  $\hat{Y}_{ps}$ . An artificial population was created with 100 design strata, 20 clusters per strata and 2 poststrata cutting across the design strata. This was achieved using a method similar to the one given in Casady and Valliant (1993) and is described below.

We would like a common mean within a poststratum, but different means across poststrata (i.e.,  $E(y_{hik}) = {}_c\mu$  for  $(hik) \in {}_cU$  where  ${}_cU$  denotes the population of units in the  $c^{th}$  poststratum). The size of the  $(hi)^{th}$  cluster, say  $N_{hi}$ , was generated as a realization of a Poisson random variable with mean 20. Once the  $N_{hi}$  was generated, the numbers of units in the 2 poststrata (ie.,  ${}_1N_{hi}, {}_2N_{hi}$ ) were determined using a multinomial distribution with parameters  $N_{hi}$  and  $\mathbf{p} = (0.30, 0.70)$ .

The value of the variable of interest for the  $(hik)^{th}$  element was generated as

$$y_{hik} = {}_c\mu + \epsilon_{hi1} + \epsilon_{hik2} + {}_c\epsilon_{hik3}N_{hi} \quad (hik) \in {}_cU$$

where the  $\epsilon$ 's are independent, standardized chi-square variates with 6 degrees of freedom, and  ${}_c\mu = 50c$ ,  $c = 1, 2$ . The  $\epsilon$ 's induce correlation between the  $y_{hik}$ 's in the same cluster and the correlation depends on whether or not the units are in the same poststrata. This process was repeated 20 times in each of the 100 design strata, giving a total of 2,000 clusters.

A two-stage sample with two clusters from each stratum was selected as follows: clusters were selected by simple random sampling with replacement, independently in each stratum, while within each sampled cluster, 15 units were selected by simple random sampling without replacement (SRSWOR) or if a sampled clusters contained fewer than 15 units, all units in the cluster were selected. Using this procedure, a total of 10,000 independent samples were generated from the finite population for the simulation study.

From each sample we calculated  $\hat{Y}$ ,  $\hat{Y}_{ps}$ ,  $v_L(\hat{Y}_{ps})$ ,  $v_J(\hat{Y}_{ps})$ , and  $v_{JL}(\hat{Y}_{ps})$ . We also included a shortcut jackknife variance estimator  $v_J^*(\hat{Y}_{ps})$ , obtained by using the full sample ratio adjustment,  ${}_cM/{}_c\hat{M}$ , instead of  ${}_cM/{}_c\hat{M}_{(gj)}$  when the  $(gj)^{th}$  cluster is deleted, i.e.

$$\hat{Y}_{ps(gj)} = \sum_c \sum_{(hik) \in {}_c\mathcal{S}} \frac{{}_cM}{{}_c\hat{M}} w_{hik(gj)} y_{hik}.$$

The linearized version of  $v_J^*(\hat{Y}_{ps})$  is the same as (3.6) with  $r_{hi}^*$  replaced by

$$r_{hi}^* = \sum_c \sum_{k \in {}_c\mathcal{S}} n_{hc} w_{hik} y_{hik}.$$

From this, it is clear that  $v_J^*(\hat{Y}_{ps})$  should overestimate the true variance of  $\hat{Y}_{ps}$ .

The empirical mean squared error, EMSE, was calculated as

$$EMSE = \frac{1}{10000} \sum_{t=1}^{10000} (\hat{Y}_{ps,t} - Y)^2$$

where  $Y$  is the known total. Table 1 reports unconditional results showing the Monte Carlo expectation of the variance estimators and the relative biases over the 10,000 samples. As

predicted by the theory, the jackknife and the linearized jackknife variance estimators are approximately equal and are good estimators of EMSE. The usual linearization variance estimator also appears to be a good estimator of EMSE. The shortcut jackknife variance estimator seriously overestimates the EMSE, reinforcing the fact that reweighting for each deleted replicate must be done correctly.

**Table 1**  
Unconditional Relative Biases of Variance Estimators

$E(v_L)$	$E(v_{JL})$	$E(v_J)$	$E(v_J^*)$	EMSE
16 948 (-0.7 %)	16 943 (-0.4 %)	16 943 (-0.4 %)	56 864 (236 %)	16 936

For studying conditional performances, a measure of how far away  $\hat{\mathbf{X}} = ({}_1\hat{M}, {}_2\hat{M})^T$  was from  $\mathbf{X} = ({}_1M, {}_2M)^T$  was calculated for each sample. This measure of bias is given by

$$B = \sum_{c=1}^2 ({}_cM / {}_c\hat{M})^2.$$

Based on this bias measure, the 10,000 samples were first sorted in ascending order and divided into ten groups of 1,000 samples each. EMSE's and Monte Carlo expectations of the variance estimators were then calculated within each of the 10 groups and are given in Table 2.

**Table 2**  
Conditional Properties of Variance Estimators

Group	$E(v_L)$	$E(v_{JL})$	$E(v_J)$	$E(v_J^*)$	EMSE
1	17 670	17 238	17 238	56 036	17 698
2	17 374	17 131	17 131	56 539	16 821
3	17 290	17 128	17 128	56 823	17 266
4	17 124	17 039	17 039	56 986	16 307
5	16 973	16 942	16 942	56 700	17 398
6	16 839	16 876	16 876	57 039	16 952
7	16 785	16 883	16 884	56 916	16 080
8	16 586	16 748	16 748	57 215	17 479
9	16 524	16 762	16 762	57 200	16 637
10	16 267	16 668	16 668	57 172	16 641

As we can see from Table 2,  $E(v_L)$ ,  $E(v_J)$  and  $E(v_{JL})$  are again approximately equal and appear to track the conditional EMSE.  $E(v_J)$  and  $E(v_{JL})$  are almost identical in each of the 10 groups, supporting the claim that they are equivalent to higher orders when 2 clusters are selected per stratum. The linearization variance estimator,  $v_L(\hat{Y}_{ps})$ , also performed well in tracking the conditional EMSE. On the other hand, the shortcut jackknife variance estimator,  $v_J^*(\hat{Y}_{ps})$ , seriously overestimates the conditional EMSE in each group, as in the unconditional case.

## 7. Work in Progress

The simulation study also included an optimal regression estimator, but those results are not reported here. The optimal estimator is approximately unbiased, conditionally given  $\hat{X}$ , unlike the generalized regression estimator,  $\hat{Y}_r$ , and is more efficient than  $\hat{Y}_r$ . Results from the simulation study support these claims.

Rao and Shao(1992) obtained consistent jackknife variance estimators under imputation for missing data, using the idea of adjusted imputed values. Rao (1993) obtained linearized versions of these jackknife variance estimators. However, they have considered only the basic estimator  $\hat{Y}$ . We are at present extending their work to post-stratified estimators and generalized regression estimators.



## References

- Casady, R. J. & Valliant, R. (1993). Conditional Properties of Post-Stratified Estimators Under Normal Theory, *Survey Methodology*, 19, 183-192.
- Deville, J. & Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382.
- Huang, E.T. & Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data, *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.
- Rao, J.N.K. (1985). Conditional inference in survey sampling, *Survey Methodology*, 11, 15-31.
- Rao, J.N.K. (1993) Linearization jackknife variance estimators under imputation for missing data, Paper presented at the American Statistical Association Annual Meetings, 1993.
- Rao, J.N.K. & Shao, J. (1992). Jackknife variance estimation with survey data under hotdeck imputation, *Biometrika*, 79, 811-22
- Royall, R.M. and Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimator of its variance, *Journal of the American Statistical Association*, 76, 66-88.
- Särndal, C.E., Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total, *Biometrika*, 76, 527-37.