



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Votre référence*

Our file *Notre référence*

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

BALANCED REPEATED REPLICATION
VARIANCE ESTIMATORS FOR SURVEY DATA
UNDER IMPUTATION

by

YING CHEN

A M.Sc. Thesis

submitted to the School of Graduate Studies and Research
in partial fulfilment of the requirements for
the Master's degree in Mathematics *

University of Ottawa
Ottawa, Ontario
Canada
October, 1993

*The M.Sc. Program is a joint program with
Carleton University, administered by the Ottawa-Carleton
Institute of Mathematics and Statistics

© Ying Chen, Ottawa, Canada, 1993



National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file *Voire référence*

Our file *Notre référence*

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-95897-9

Canada



UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

Imputation is commonly used for missing data in sample surveys. Usually The imputed values are treated as true values and variance estimates are computed using standard variance formulas. But this procedure can lead to serious underestimation of the true variance. Rao and Shao(1992) proposed a new consistent jackknife variance estimator based on adjusting the imputed values. This thesis applies their idea to construct two adjusted Balanced Repeated Replication(BRR) variance estimators for stratified multistage surveys. Under a uniform response mechanism, the adjusted BRR variance estimators are shown to be consistent for a particular simple hot deck imputation and ratio hot deck imputation. Also, the relationship between jackknife variance estimators and BRR variance estimators which was established by Rao and Wu (1985) for complete data set, is shown to be still held for data set with imputed values. The performances of these variance estimates are compared through some simulation studies.

Acknowledgements

I would like to express my sincere gratitude to Dr. Jun Shao for the valuable guidance. His time and energy spent in supervising this thesis was very much appreciated. I would like to thank the University of Ottawa which provided me financial support. The thesis was also supported by a research grant from Dr. Shao. Finally, I would like to thank my parents, my daughter Ellen who was born in the middle of my two years of study and my husband, Hao for their support.

Contents

1	Introduction	1
2	The Balanced Repeated Replication Method	5
2.1	The Stratified Multistage Sampling Design	5
2.2	The BRR Variance Estimators	6
3	The Adjusted BRR Variance Estimators under Imputation	9
3.1	The BRR Variance Estimator under Simple Hot Deck Imputation	9
3.1.1	The Imputed Estimator	9
3.1.2	The Adjusted BRR Variance Estimator	11
3.1.3	Asymptotic Properties	12
3.2	The BRR Variance Estimator under Ratio Hot Deck Imputation	21
3.2.1	The Imputed Estimator	22
3.2.2	The Adjusted BRR Variance Estimator	23
3.2.3	Asymptotic Properties	24
4	Comparison of BRR and Jackknife Variance Estimators	30
4.1	The Original BRR and Jackknife Variance Estimators	30
4.2	The Adjusted BRR and Jackknife Variance Estimators	32
5	Simulation Studies	37
5.1	The Finite Population	37

5.2	The Sample and Estimators	38
5.3	The Measures of The Performance	39
Appendix A Computer program for simulation		42
Bibliography		49

Chapter 1

Introduction

One important problem in sample surveys is the estimation of variances of statistics from survey data with item nonresponse. Item nonresponse occurs when the sampled unit fails to provide information on some items in the survey. A common way of handling this kind of problem is to impute each missing datum under some model for nonresponse. Imputation is popular because it has the following advantages:

1. It creates a complete data set which allows us to use many established standard complete-data methods, so that much processing time will be saved.
2. The results obtained from different analyses are consistent with one another, which may not apply to results from an incomplete data set.
3. It permits the use of the same survey weight for all items.

There are many types of imputation, for example, mean imputation which replaces all missing response by the respondent mean, ratio imputation which uses an auxiliary variable observed for all units, nearest neighbour imputation which assigns to a nonrespondent the value of the "nearest" respondent, and hot deck imputation which draws a value randomly from respondents to replace a missing datum. Most

surveys prefer to use hot deck imputation because it preserves the distribution of the item values unlike mean and ratio imputation. However, if we simply treat the imputed values as true values and compute the variance estimator using standard formulas, the resulting estimator may seriously underestimate the true variance when the response rate is low, because the variability from not knowing the missing data is ignored.

Some research has been done in solving this problem. Rubin (1978) provided a multiple imputation method to account for the inflation in the variance due to imputation. He presented his method in the context of simple random sampling. Suppose $m(\geq 2)$ independent simple random hot deck imputations are performed to get m complete data set. Let $\bar{y}_{I1}, \dots, \bar{y}_{Im}$ be the estimators of population mean \bar{Y} from m complete data sets, and

$$\bar{y}_I = \frac{1}{m} \sum_{t=1}^m \bar{y}_{It}$$

be the final estimator of \bar{Y} . The variance of \bar{y}_I is estimated by

$$v_M = \sum_{t=1}^m \frac{s_{It}^2}{mn} + \frac{m+1}{m} \sum_{t=1}^m \frac{(\bar{y}_{It} - \bar{y}_I)^2}{m-1}, \quad (1.1)$$

where n is the sample size, s_{It}^2 is the sample variance for the t th complete data set. The second term in (1.1) makes up for the inflation in the variance due to imputation. Based on this idea, Rubin and Schenker (1986) proposed several other multiple imputation methods by modifying the imputation scheme. But multiple imputation may not provide consistent variance estimators for stratified multistage survey, even for large m (Fay, 1991). Moreover, it is hard to operate due to the

difficulty in maintaining the multiple complete data sets, especially in large scale surveys.

Burns (1990) presented a pseudo-replicate imputation for stratified multistage surveys using jackknife variance estimators. Suppose that we have L strata with n_h first-stage units sampled from stratum h , $\sum n_h = n$. Let A_r , and $A_r(-hi)$ denote the set of all respondents and the set of respondents which are not in the (h, i) th first-stage unit, respectively. Also let y_I be the estimator of the population total Y using some imputation method based on A_r , and $y_I(-hi)$ be the estimator of Y using the same imputation method based on $A_r(-hi)$, $i = 1, \dots, n_h$; $h = 1, \dots, L$. Then Burns' jackknife variance estimator of y_I is given by

$$v_B = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} [y_I(-hi) - y_I]^2.$$

Unfortunately, under simple random sampling this variance estimator can lead to serious overestimation when n is large (Rao and Shao 1992).

Rao and Shao (1992) proposed a new jackknife variance estimator which is obtained by first adjusting the imputed value for each pseudo-replicate and then applying the standard jackknife formula (details are in section 4.2). This estimator is shown to be consistent under some regularity conditions.

The main contribution of this thesis is to extend Rao and Shao's idea to construct balanced repeated replication(BRR) variance estimators. Following a brief overview of the balanced repeated replication method in chapter 2, chapter 3 presents two adjusted BRR variance estimators under simple hot deck imputation and ratio hot deck imputation. The proofs of consistency of these estimators can be found in

section 3.1.3 and section 3.2.3. Furthermore, Rao and Shao's adjusted jackknife variance estimator and the adjusted BRR variance estimator under simple hot deck imputation are compared in chapter 4. The result coincides with what Rao and Wu (1985) showed for the jackknife variance estimator and BRR variance estimator in the case of a complete data set. The performances of the naive BRR variance estimator, the adjusted BRR variance estimator and adjusted jackknife variance estimators are compared through simulations in chapter 5.

Chapter 2

The Balanced Repeated Replication Method

2.1 The Stratified Multistage Sampling Design

Throughout this thesis we will assume the following commonly used stratified multistage sampling design. The population has been stratified into L strata with N_h first-stage clusters in the h th stratum. $n_h (\geq 2)$ clusters are selected from N_h clusters with probability p_{hi} ($h = 1, 2, \dots, L; i = 1, 2, \dots, N_h$), where $\sum_{i=1}^{N_h} p_{hi} = 1$. These clusters are selected independently and with replacement. Within the i th cluster in the h th stratum n_{hi} ultimate units are selected from N_{hi} population units using some sampling methods ($i = 1, 2, \dots, n_h, h = 1, 2, \dots, L$). The population size is $N = \sum_{h=1}^L \sum_{i=1}^{N_h} N_{hi}$, and the sample size is $n_T = \sum_{h=1}^L \sum_{i=1}^{n_h} n_{hi}$. Suppose that \hat{Y}_{hi} is a linear unbiased estimator of the total Y_{hi} for a selected cluster based on the second and subsequent stages. Then a linear unbiased estimator of the stratum total Y_h is given by

$$\hat{Y}_h = \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{n_h p_{hi}}.$$

A linear unbiased estimator of population mean is given by $\bar{y} = \sum \hat{Y}_h/N$, which may be written as

$$\bar{y} = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} y_{hij}$$

or

$$\bar{y} = \sum_{(hij) \in A_n} w_{hij} y_{hij},$$

where A_n is the total sample of n units, and w_{hij} and y_{hij} denote the survey weight and the y -value attached to the (hij) -th unit respectively ($j = 1, 2, \dots, n_{hi}; i = 1, 2, \dots, n_h; h = 1, 2, \dots, L$). Letting $r_{hi} = \sum_{j=1}^{n_{hi}} w_{hij} y_{hij}$, we note that, for fixed h , the r_{hi} 's are independent identically distributed, whereas for $h \neq g$, r_{hi} and r_{gi} are independent but not necessarily identically distributed.

2.2 The BRR Variance Estimators

In order to measure the uncertainty of a given estimator or to compare the efficiencies of sampling designs, we need to derive a suitable estimator for the variance of this estimator. The balanced repeated replication method is one of the useful methods to get a variance estimator. This method was first proposed by McCarthy(1969) for the case where $n_h = 2$ clusters per stratum are selected in the first stage, and has been extended to the general case of $n_h \geq 2$ clusters per stratum.

Suppose that for the h th stratum and r th replication, a set s_{rh} of m_h integers is selected from $\{1, 2, \dots, n_h\}$. $\{s_{rh} : r = 1, 2, \dots, R, h = 1, 2, \dots, L\}$ constitutes a BRR if

- 1) for fixed h , the number of elements in $\{r : i \in s_{rh}, i' \in s_{rh}, i \neq i'\}$ is a constant.
- 2) for fixed h and h' , the number of elements in $\{r : i \in s_{rh}, i' \in s_{rh'}, i \neq i'\}$ is a constant.

Let

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad y_{hi} = n_h \sum_{j=1}^{n_{hi}} w_{hij} y_{hij},$$

$$\bar{y} = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} y_{hij} = \sum_{h=1}^L \bar{y}_h$$

and

$$\bar{y}_h^{(r)} = \frac{1}{m_h} \sum_{i \in s_{r,h}} y_{hi}, \quad r = 1, 2, \dots, R.$$

We can define the r th balanced sample mean by

$$\bar{y}^{(r)} = \sum_{h=1}^L \left[\sqrt{\frac{m_h}{n_h - m_h}} \bar{y}_h^{(r)} + \left(1 - \sqrt{\frac{m_h}{n_h - m_h}} \right) \bar{y}_h \right]. \quad (2.1)$$

$\bar{y}^{(r)}$ is also an unbiased estimator of \bar{Y} because

$$E(\bar{y}_h^{(r)} | A_n) = \bar{y}_h.$$

Therefore

$$E(\bar{y}^{(r)}) = E(\bar{y}) = \bar{Y}.$$

The BRR variance estimator of \bar{y} is then given by

$$v_{BRR} = \frac{1}{R} \sum_{r=1}^R (\bar{y}^{(r)} - \bar{y})^2.$$

This BRR variance estimator is unbiased due to the balancedness properties and can be shown to be consistent under some conditions.

To construct the BRR variance estimator, we need to find a BRR with the number of replications R as small as possible. In the case of $n_h = 2$ for all h , we can do this using a Hadamard matrix. Note that a balanced set of R half samples can be defined as a $R \times L$ matrix with the (r, h) th element $\varepsilon_{rh} = +1$ or -1 according

to whether the first or the second first-stage unit in h th stratum is in the r th half sample, and the balanced properties can be expressed as

$$\sum_{r=1}^R \varepsilon_{rh} = 0, \quad \text{for all } h$$

and

$$\sum_{r=1}^R \varepsilon_{rh} \varepsilon_{rh'} = 0, \quad \text{for all } h \neq h',$$

i.e., the sum of the elements in every column equals 0 and the columns of the matrix are orthogonal, hence, we can construct a minimal set of R balanced half sample from a $R \times R$ Hadamard matrix by choosing any L columns excluding the column of all +1's, where $L + 1 \leq R \leq L + 4$.

Unfortunately, it is not easy to obtain a BRR with a feasible R in the general case of $n_h \geq 2$. Gurney and Jewett (1975) proposed a method to get a BRR in the case of $n_h = q > 2$, where q is a prime or a power of prime, by using orthogonal arrays of strength two. Gupta and Nigam (1987) and Wu (1991) used mixed level orthogonal arrays of strength two to obtain BRR in the case of unequal n_h with $m_h = 1$ for all h . More methods can be found in Sitter (1993). In this thesis we assume that a BRR exists and R is of order n .

Chapter 3

The Adjusted BRR Variance Estimators under Imputation

3.1 The BRR Variance Estimator under Simple Hot Deck Imputation

3.1.1 The Imputed Estimator

We now consider stratified multistage surveys with random nonresponse, i.e., the units are missing randomly with the same probability. Let y_{hij}^* be the imputed values for nonrespondents using a hot deck imputation. The imputed estimator of \bar{Y} is then given by

$$\bar{y}_I = \sum_{(hij) \in A_r} w_{hij} y_{hij} + \sum_{(hij) \in A_m} w_{hij} y_{hij}^*, \quad (3.1)$$

where A_r and A_m denote the sample of respondents and the sample of nonrespondents respectively. The estimator \bar{y}_I is no longer unbiased if simple random sampling is used to select the donors from A_r , unless y_{hij}^* is chosen as $y_{gk}(w_{gk}/w_{hij})$, where $(gk) \in A_r$ is the selected donor (Platek and Gray 1983). But this y_{hij}^* may not be appealing when the characteristic of interest takes only integer values. A simple alternative (Rao and Shao 1992) is to choose donor $(gk) \in A_r$, with replacement

and with probability

$$\frac{w_{gk}}{\sum_{(hij) \in A_r} w_{hij}},$$

and take $y_{hij}^* = y_{gk}$. Under this hot deck imputation scheme, we get

$$\begin{aligned} E_*(\bar{y}_I) &= \sum_{(hij) \in A_r} w_{hij} y_{hij} + \sum_{(hij) \in A_m} w_{hij} E_*(y_{hij}^*) \\ &= \sum_{(hij) \in A_r} w_{hij} y_{hij} + \sum_{(hij) \in A_m} \left(w_{hij} \sum_{(gk) \in A_r} \frac{w_{gk} y_{gk}}{\sum_{(hij) \in A_r} w_{hij}} \right) \\ &= \left(\sum_{(hij) \in A_r} w_{hij} y_{hij} \right) \left(\sum_{(hij) \in A_n} w_{hij} \right) \left(\sum_{(hij) \in A_r} w_{hij} \right)^{-1} \\ &= \frac{SU}{T}, \text{ say,} \end{aligned} \tag{3.2}$$

where E_* denotes the expectation with respect to hot deck imputation given the sample of respondents, A_r .

Define a response indicator variable by

$$a_{hij} = \begin{cases} 1 & \text{if } (hij) \in A_r \\ 0 & \text{if } (hij) \in A_m \end{cases}.$$

Under the uniform response mechanism, $P(a_{hij} = 1) = p$ for all $(hij) \in A_n$ and a_{hij} 's are independent. Then

$$\begin{aligned} E(S) &= E \left(\sum_{(hij) \in A_n} a_{hij} w_{hij} y_{hij} \right) \\ &= p E \left(\sum_{(hij) \in A_n} w_{hij} y_{hij} \right) \\ &= p E(\bar{y}) = p \bar{Y}. \end{aligned}$$

Similarly, $E(U) = 1$ and $E(T) = p$. Therefore the imputed estimator \bar{y}_I is asymptotically unbiased for \bar{Y} .

3.1.2 The Adjusted BRR Variance Estimator

In order to construct the adjusted BRR variance estimator we need to define some terms. First, we rewrite the r th balanced sample mean $\bar{y}^{(r)}$ (2.1) by

$$\bar{y}^{(r)} = \sum_{(hij) \in A_n} (1 + d_{hi}^r) w_{hij} y_{hij}, \quad (3.3)$$

where $d_{hi}^r = \frac{c_h n_h}{m_h} \delta_{hi}^r - c_h$, $c_h = \sqrt{\frac{m_h}{n_h - m_h}}$ and

$$\delta_{hi}^r = \begin{cases} 1 & \text{if } (hi) \in s_{rh} \\ 0 & \text{if } (hi) \notin s_{rh} \end{cases}.$$

Then the r th balance sample mean with imputed values can be written as

$$\bar{y}_I^{(r)} = \sum_{(hij) \in A_r} (1 + d_{hi}^r) w_{hij} y_{hij} + \sum_{(hij) \in A_m} (1 + d_{hi}^r) w_{hij} y_{hij}^*. \quad (3.4)$$

Let

$$S^{(r)} = \sum_{(hij) \in A_r} (1 + d_{hi}^r) w_{hij} y_{hij},$$

$$U^{(r)} = \sum_{(hij) \in A_n} (1 + d_{hi}^r) w_{hij},$$

$$T^{(r)} = \sum_{(hij) \in A_r} (1 + d_{hi}^r) w_{hij}.$$

Define the adjusted r th balanced sample mean by

$$\bar{y}_{I_a}^{(r)} = \sum_{(hij) \in A_r} (1 + d_{hi}^r) w_{hij} y_{hij} + \sum_{(hij) \in A_m} (1 + d_{hi}^r) w_{hij} \left(y_{hij}^* + \frac{S^{(r)}}{T^{(r)}} - \frac{S}{T} \right).$$

That is, in the r th balanced sample we adjusted the imputed value y_{hij}^* by an amount $S^{(r)}/T^{(r)} - S/T$. The inflation in the BRR variance estimator caused by this adjustment accounts for the inflation in the variance due to missing values and imputation.

It is easy to see that

$$E_* \bar{y}_{I_a}^{(r)} = \frac{S^{(r)} U^{(r)}}{T^{(r)}}$$

by noting that $E_* y_{hij}^* = \frac{S}{T}$. Under the uniform response assumption we have

$$E(S^{(r)}) = E \left(\sum_{(hij) \in A_n} a_{hij} (1 + d_{hi}^r) w_{hij} y_{hij} \right) = p E(\bar{y}^{(r)}) = p \bar{Y}.$$

Similarly, $E(U^{(r)}) = 1$ and $E(T^{(r)}) = p$. So $\bar{y}_{Ia}^{(r)}$ is also an asymptotically unbiased estimator for \bar{Y} .

Now the adjusted BRR variance estimator under simple hot deck imputation can be defined by

$$v_{BRR}^a = \frac{1}{R} \sum_{r=1}^R (\bar{y}_{Ia}^{(r)} - \bar{y}_I)^2.$$

The asymptotic consistency of v_{BRR}^a will be established in the next section. Combining this result and the asymptotic normality of \bar{y}_I (Rao and Shao 1992) we can obtain the approximate $(1 - \alpha)$ level confidence interval for \bar{Y} .

3.1.3 Asymptotic Properties

We are ready to study the asymptotic properties of the imputed estimator \bar{y}_I and the adjusted BRR variance estimator v_{BRR}^a . While showing these properties, we assume the following regularity conditions:

Condition 1: $n^{1+\delta} \sum_{h=1}^L \sum_{i=1}^{n_h} E |r_{hi}^{(l)} - E r_{hi}^{(l)}|^{2+\delta} = O(1)$ for some $\delta > 0, l = 1, 2, 3$ as $n \rightarrow \infty$, where $n = \sum_{h=1}^L n_h$, $r_{hi}^{(1)} = \sum_{j=1}^{n_{hi}} a_{hij} w_{hij} y_{hij}$, $r_{hi}^{(2)} = \sum_{j=1}^{n_{hi}} a_{hij} w_{hij}$, $r_{hi}^{(3)} = \sum_{j=1}^{n_{hi}} w_{hij}$.

Condition 2: $n(\text{covariance matrix of } S, U, \text{ and } T) \rightarrow \text{a positive definite matrix}$ as $n \rightarrow \infty$.

Condition 3: $n \max_{h,i \in A_n} \sum_{j=1}^{n_{hi}} w_{hij} = O_p(1)$.

Condition 4: $\sum_{(hij) \in A_n} w_{hij} |y_{hij} - \bar{Y}|^{2+\delta} = O_p(1)$, as $n \rightarrow \infty$.

Condition 5: $R = O(n)$

Condition 6: $0 < \varepsilon \leq \frac{m_h}{n_h} \leq \frac{1}{2}$ for all h , where ε is a constant.

Condition 1 is a standard Liapunov-type condition on the $2+\delta$ absolute moments used to establishing a central limit theorem for independent but not necessarily identically distributed random variables. Condition 2 assumes that the limit of the covariance matrix of S, U and T exists when multiplied by the normalizing factor n . Because we focus on survey with large numbers of strata with relatively few first-stage clusters selected from each stratum, Condition 3 is required, which says no cluster contribution is of disproportionate size. Condition 3 and 4 imply a Liapunov-type condition on the $2 + \delta$ absolute moment, $E_* |w_{hij}(1 - a_{hij})(y_{hij}^* - E_* y_{hij}^*)|^{2+\delta}$. Condition 5 is needed in establishing lemma 4 and consistency of the adjusted BRR variance estimator. Condition 6 is a necessary condition for lemma 4.

The following lemmas are useful:

Lemma 1 (Central Limit Theorem) Suppose $\{X_t\}_{t=1}^T$ are independent but not necessarily identically distributed random variables with $E(X_t) = \mu_t$ and $V(X_t) = \sigma_t^2$. If $T^{-1} \sum_{t=1}^T \sigma_t^2 \rightarrow \sigma^2$ and $T^{-1} \sum_{t=1}^T E |X_t - \mu_t|^{2+\delta} = O(1)$ as $T \rightarrow \infty$ for some $\delta > 0$ then

$$\sqrt{T}(\bar{X} - \bar{\mu}) \longrightarrow_d N(0, \sigma^2),$$

where $\bar{X} = T^{-1} \sum_{t=1}^T X_t$ and $\bar{\mu} = T^{-1} \sum_{t=1}^T \mu_t$.

Lemma 2 (Law of Large Numbers) Suppose $\{X_t\}_{t=1}^T$ are independent but not necessarily identically distributed random variables with $E(X_t) = \mu_t$ and $V(X_t) = \sigma_t^2$. If $T^{-1} \sum_{t=1}^T E |X_t|^{1+\delta} = O(1)$ as $T \rightarrow \infty$ for some $\delta > 0$, then for any $\varepsilon > 0$,

$$P\{|\bar{X} - \bar{\mu}| \geq \varepsilon\} = O(T^{-r}),$$

where $r = \delta$ when $\delta \leq 1$ and $r = (1 + \delta)/2$ when $\delta \geq 1$.

Lemma 3 Suppose that X_n are random variables and a_n are positive constants such that $X_n/a_n \rightarrow_d N(0, 1)$. Let W_n be random variables, b_n be positive constants and Z_n be random elements such that

$$\frac{W_n}{b_n} \rightarrow_{d|Z_n} N(0, 1).$$

Assume also that X_n is a function only of Z_n . Then

$$c_n^{-1}(X_n + W_n) \rightarrow_d N(0, 1),$$

where $c_n^2 = a_n^2 + b_n^2$.

Lemma 4 Let \bar{X} be a vector of population mean, \bar{x} be a linear unbiased estimator of \bar{X} , $\theta = g(\bar{X})$, $\hat{\theta} = g(\bar{x})$. Assume condition 3, 5, 6 hold and condition 1 holds for $r_{hi} = \sum_{j=1}^{n_{hi}} w_{hij} x_{hij}$. Assume further that g is continuously differentiable with nonzero ∇g in a neighbourhood of \bar{X} and $n\text{var}(\bar{x}) < \infty$. Then

$$n(v_{BRR} - \sigma^2) \rightarrow_p 0,$$

where $\sigma^2 = \text{var}(\hat{\theta})$, $v_{BRR} = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2$, and $\hat{\theta}^{(r)} = g(\bar{x}^{(r)})$.

Lemmas 1 and 2 are a central limit theorem and a law of large numbers for independent but not necessarily identically distributed random variables (Hoadley,1971 and Sen 1970). Lemma 3 from Rao and Shao's unpublished technical report(1992) is used to show the asymptotic normality of \bar{y}_I . Lemma 4 (Shao 1993) is an extension of the consistency of BRR variance estimator in the case of $n_h = 2$ for all h (Krewski and Rao, 1981).

The Asymptotic Normality of \bar{y}_I

The asymptotic normality of \bar{y}_I is established by Rao and Shao (1992). First, let $z_{hij}^* = w_{hij}(1 - a_{hij})(y_{hij}^* - E_*y_{hij}^*)$, $(hij) \in A_n$. Then we can write

$$\bar{y}_I - \bar{Y} = \bar{z}^* + \left(\frac{SU}{T} - \bar{Y} \right),$$

where $\bar{z}^* = \sum_{(hij) \in A_n} z_{hij}^*$.

Under the condition 1, 2, 4, 5, using lemma 1 and the delta method, we can show that

$$a_n^{-1} \left(\frac{SU}{T} - \bar{Y} \right) \longrightarrow_d N(0, 1),$$

where a_n^2 is the asymptotic variance of SU/T , and

$$b_n^{-1} \bar{z}^* \longrightarrow_{d|y_{obs}, a} N(0, 1),$$

where $y_{obs} = y_{hij}$, $(hij) \in A_r$, $a = a_{hij}$, $(hij) \in A_n$, and b_n^2 is the asymptotic expectation of $var_*(\bar{z}^*)$, var_* denotes the variance with respect to hot deck imputation given y_{obs} and a .

Now applying lemma 3 with $X_n = (SU)/T - \bar{Y}$, $W_n = \bar{z}^*$ and $Z_n = (y_{obs}, a)$ we conclude that

$$c_n^{-1}(\bar{y}_I - \bar{Y}) \longrightarrow_d N(0, 1),$$

where $c_n^2 = a_n^2 + b_n^2$ is the asymptotic variance of \bar{y}_I , i.e.,

$$\text{var}(\bar{y}_I) = \text{var} E_*(\bar{y}_I) + E \text{var}_*(\bar{y}_I) = a_n^2 + b_n^2 = c_n^2.$$

So that \bar{y}_I is asymptotically normal.

Asymptotic Consistency of v_{BRR}^a

After some straightforward calculation, we get

$$\bar{y}_{Ia}^{(r)} - \bar{y}_I = \left(\frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T} \right) + \sum_{(hij) \in A_n} d_{hi}^r z_{hij}^*.$$

So the adjusted BRR variance estimator v_{BRR}^a can be expressed as

$$\begin{aligned} v_{BRR}^a &= \frac{1}{R} \sum_{r=1}^R \left(\bar{y}_{Ia}^{(r)} - \bar{y}_I \right)^2 \\ &= \frac{1}{R} \sum_{r=1}^R \left(\frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T} \right)^2 + \sum_{r=1}^R \left(\sum_{(hij) \in A_n} d_{hi}^r z_{hij}^* \right)^2 \\ &\quad + \frac{2}{R} \sum_{r=1}^R \left(\frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T} \right) \left(\sum_{(hij) \in A_n} d_{hi}^r z_{hij}^* \right) \\ &= A + B + C, \text{ say.} \end{aligned} \tag{3.5}$$

We claim that

1. $n\{A - \text{var}[E_*(\bar{y}_I)]\} \rightarrow_p 0$;
2. $n\{B - E[\text{var}_*(\bar{y}_I)]\} \rightarrow_p 0$;
3. $nC \rightarrow_p 0$.

The first assertion is a direct consequence of Lemma 4 with

$$x_{hi} = \left(\sum_{j=1}^{n_{hi}} a_{hij} w_{hij} y_{hij}, \sum_{j=1}^{n_{hi}} w_{hij}, \sum_{j=1}^{n_{hi}} a_{hij} w_{hij} \right),$$

$$\bar{x} = (S, U, T) \text{ and } \hat{\theta} = g(S, U, T) = \frac{SU}{T}.$$

Considering the second assertion, we first note that

$$E_*(z_{hij}^*) = E_*[(1 - a_{hij})w_{hij}(y_{hij}^* - E_*y_{hij}^*)] = 0,$$

hence,

$$\begin{aligned} E_*B &= \frac{1}{R} \sum_{r=1}^R E_* \left(\sum_{(hij) \in A_n} d_{hi}^r z_{hij}^* \right)^2 \\ &= \frac{1}{R} \sum_{r=1}^R \text{var}_* \left(\sum_{(hij) \in A_n} d_{hi}^r z_{hij}^* \right) \\ &= \sum_{(hij) \in A_n} \left[\frac{1}{R} \sum_{r=1}^R (d_{hi}^r)^2 \right] \text{var}_*(z_{hij}^*). \end{aligned}$$

Using the balanceness properties, we have

$$\begin{aligned} \frac{1}{R} \sum_{r=1}^R (d_{hi}^r)^2 &= \frac{1}{R} \sum_{r=1}^R \frac{m_h}{n_h - m_h} \left(\frac{n_h}{m_h} \delta_{hi}^r - 1 \right)^2 \\ &= \frac{1}{R} \frac{m_h}{n_h - m_h} \left[\frac{Rm_h}{n_h} \left(\frac{n_h}{m_h} - 1 \right)^2 + \frac{R(n_h - m_h)}{n_h} \right] \\ &= 1. \end{aligned} \tag{3.6}$$

Thus

$$E_*B = \sum_{(hij) \in A_n} \text{var}_*(z_{hij}^*) = \text{var}_* \left(\sum_{(hij) \in A_n} z_{hij}^* + \frac{SU}{T} \right) = \text{var}_*(\bar{y}_I),$$

and

$$E(B) = E[\text{var}_*(\bar{y}_I)].$$

We now apply Lemma 2 (Law of Large Number) to prove

$$n[B - E(B)] \longrightarrow_p 0.$$

Let $t_{hi}^* = \sum_{j=1}^{n_{hi}} z_{hij}^*$, then nB can be expressed as

$$\begin{aligned} nB &= \frac{n}{R} \sum_{r=1}^R \left(\sum_{h=1}^L \sum_{i=1}^{n_h} d_{hi}^r t_{hi}^* \right)^2 \\ &= \frac{n}{R} \sum_{r=1}^R \sum_{h=1}^L \left(\sum_{i=1}^{n_h} d_{hi}^r t_{hi}^* \right)^2 \\ &\quad + \frac{2n}{R} \sum_{r=1}^R \sum_{h < h'} \left(\sum_{i=1}^{n_h} d_{hi}^r t_{hi}^* \right) \left(\sum_{i=1}^{n_{h'}} d_{h'i}^r t_{h'i}^* \right) \end{aligned}$$

By the balancedness properties, we have

$$\sum_{h < h'} \left(\sum_{i=1}^{n_{hi}} d_{hi}^r t_{hi}^* \right) \left(\sum_{i=1}^{n_{h'i}} d_{h'i}^r t_{h'i}^* \right) = 0.$$

Hence

$$\begin{aligned} nB &= \frac{n}{R} \sum_{r=1}^R \sum_{h=1}^L \left(\sum_{i=1}^{n_h} d_{hi}^r t_{hi}^* \right)^2 \\ &= \frac{n}{R} \sum_{r=1}^R \sum_{h=1}^L \left[\sum_i (d_{hi}^r t_{hi}^*)^2 + \sum_{i \neq i'} d_{hi}^r t_{hi}^* d_{h'i}^r t_{h'i}^* \right] \\ &= n \sum_{h=1}^L \sum_{i=1}^{n_{hi}} \left(\frac{1}{R} \sum_{r=1}^R d_{hi}^{r2} \right) t_{hi}^{*2} + n \sum_{h=1}^L \sum_{i \neq i'} \left(\frac{1}{R} d_{hi}^r d_{h'i}^r \right) t_{hi}^* t_{h'i}^* \\ &= n \sum_{h=1}^L \sum_{i=1}^{n_h} t_{hi}^{*2} - n \sum_{h=1}^L \sum_{i \neq i'} \frac{2}{n_{hi} - 1} t_{hi}^* t_{h'i}^*, \end{aligned} \tag{3.7}$$

where the last equality follows from (3.6) and

$$\begin{aligned}
\frac{1}{R} \sum_{r=1}^R d_{hi}^r d_{hi'}^r &= \frac{1}{R} \sum_{r=1}^R \frac{m_h}{n_h - m_h} \left(\frac{n_h}{m_h} \delta_{hi}^r - 1 \right) \left(\frac{n_h}{m_h} \delta_{hi'}^r - 1 \right) \\
&= \frac{m_h}{R(n_h - m_h)} \left[\frac{Rm_h(m_h - 1)}{n_h(n_h - 1)} \left(\frac{n_h}{m_h} - 1 \right)^2 \right. \\
&\quad \left. - \frac{2Rm_h(n_h - m_h)}{n_h(n_h - 1)} \left(\frac{n_h}{m_h} - 1 \right) + \frac{R(n_h - m_h)(n_h - m_h - 1)}{n_h(n_h - 1)} \right] \\
&= -\frac{2}{n_h}.
\end{aligned}$$

In order to apply Lemma 2 to the two terms in (3.7) separately, we need to establish

$$E_* |t_{hi}^*|^{2+2\delta} \leq O_p(1) \left(\sum_{j=1}^{n_{hi}} w_{hij} \right)^{2+2\delta}. \quad (3.8)$$

By definition,

$$t_{hi}^* = \sum_{j=1}^{n_{hi}} z_{hij}^* = \sum_{j \in A_n} (1 - a_{hij}) w_{hij} (y_{hij}^* - \frac{S}{T}).$$

Hence,

$$\begin{aligned}
E_* |t_{hi}^*|^{2+2\delta} &= \sum_{(gk) \in A_r} \left| \sum_{j \in A_n} (1 - a_{hij}) w_{hij} (y_{gk} - \frac{S}{U}) \right|^{2+2\delta} \frac{w_{gk}}{T} \\
&\leq \frac{1}{T} \left(\sum_{j \in A_n} w_{hij} \right)^{2+2\delta} \sum_{(gk) \in A_n} w_{hij} \left| y_{gk} - \frac{S}{T} \right|^{2+2\delta}.
\end{aligned}$$

But

$$\begin{aligned}
\sum_{(gk) \in A_n} \left| y_{gk} - \frac{S}{T} \right|^{2+2\delta} &\leq 2^{1+\delta} \left(\sum_{(gk) \in A_n} w_{gk} \left| y_{gk} - \bar{Y} \right|^{2+2\delta} \right. \\
&\quad \left. + \sum_{(gk) \in A_n} w_{gk} \left| \bar{Y} - \frac{S}{T} \right|^{2+2\delta} \right).
\end{aligned}$$

Also,

$$T - p = O_p(1), \quad \frac{S}{T} - \bar{Y} = O_p(1).$$

Then (3.8) follows from these results and condition 3.

Now we turn to (3.7). Let

$$\bar{X}_1 = n \sum_{h=1}^L \sum_{i=1}^{n_h} (t_{hi}^*)^2 = \frac{1}{n} \sum_t X_{1t},$$

where $X_{1t} = (nt_{hi}^*)^2$ and

$$\bar{X}_2 = n \sum_{h=1}^L \sum_{i=1}^{n_h} \left(\frac{2}{n_h - 1} t_{hi}^* \sum_{j \neq i} t_{hj}^* \right) = \frac{1}{n} \sum_t X_{2t},$$

where $X_{2t} = \frac{2n^2}{n_h - 1} t_{hi}^* \sum_{j \neq i} t_{hj}^*$. It follows from the (3.8) and condition 3 that

$$\begin{aligned} \frac{1}{n} \sum_t E_* |X_{1t}|^{1+\delta} &\leq O_p(1) n^{1+2\delta} \sum_h \sum_i (\sum_j w_{hij})^{2+2\delta} \\ &\leq O_p(1) n^{2+2\delta} \max_j (\sum_j w_{hij})^{2+2\delta} \\ &= O_p(1). \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n} \sum_t E_* |X_{2t}|^{1+\delta} &\leq n^{1+2\delta} \sum_h \sum_i \frac{2^{1+\delta}}{n_h - 1} E_* \sum_{i \neq j} |t_{hi}^* t_{hj}^*|^{1+\delta} \\ &\leq O_p(1) n^{1+2\delta} \sum_h \sum_i \frac{1}{n_h - 1} \{ (n_h - 1) E_* |t_{hi}^*|^{2+2\delta} \\ &\quad + \sum_{j \neq i} E_* |t_{hj}^*|^{2+2\delta} \} \\ &\leq O_p(1) n^{1+2\delta} \sum_h \sum_i (\sum_{j=1}^{n_{hi}} w_{hij})^{2+2\delta} \\ &= O_p(1). \end{aligned}$$

Applying lemma 2 to X_{1t} and X_{2t} , we complete the proof of the second assertion.

The third assertion remains to be shown. Let

$$A^{(r)} = \frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T}.$$

Then

$$\begin{aligned}
nC &= \frac{2n}{R} \sum_{r=1}^R \left(A^{(r)} \sum_{(hij) \in A_n} d_{hi}^* z_{hij}^* \right) \\
&= \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} \left(\frac{2n^2}{R} \sum_{r=1}^R A^{(r)} d_{hi}^* t_{hi}^* \right) \\
&= \frac{1}{n} \sum_{t=1}^n X_t, \text{ say,} \tag{3.9}
\end{aligned}$$

where $X_t = \frac{2n^2}{R} \sum_r A^{(r)} d_{hi}^* t_{hi}^*$. Note that $E_* X_t = 0$. Hence $nC \rightarrow_p 0$ follows from Lemma 2 and

$$\frac{1}{n} \sum_{t=1}^n E_* |X_t|^2 = O_p(1). \tag{3.10}$$

To do this, first we note that $\frac{1}{R} \sum_r (A^{(r)})^2 = O_p(\frac{1}{n})$ by the first assertion. With condition 3, 5, (3.8) and the fact, $d_{hi}^r = O_p(1)$, we get

$$\begin{aligned}
\frac{1}{n} \sum_{t=1}^n E_* |X_t|^2 &= \frac{1}{n} \sum_{h,i} \left[2^2 \frac{n^4}{R} E_* |t_{hi}^*|^2 \frac{1}{R} \left| \sum_r A^{(r)} d_{hi}^r \right|^2 \right] \\
&\leq O_p(1) n \sum_{h,i} \left(\sum_{j=1}^{n_{hi}} w_{hij} \right)^2 \\
&= O_p(1).
\end{aligned}$$

Combining the three assertions, we obtain that

$$n[v_{BRR}^a - \text{var}(\bar{y}_I)] \rightarrow_p 0,$$

i.e., the adjusted BRR variance estimator, v_{BRR}^a , is an asymptotically consistent estimator for the variance of the imputed estimator, $\text{var}(\bar{y}_I)$.

3.2 The BRR Variance Estimator under Ratio Hot Deck Imputation

Hot deck imputation has various versions (Kalton, 1981; Sedransk, 1985). In this section we will consider the BRR variance estimator under ratio hot deck imputation.

Like ratio estimators, when the item y has an auxiliary variable x , using ratio hot deck imputation will increase the precision of the estimator by taking advantage of the correlation between y and x .

3.2.1 The Imputed Estimator

Suppose now y_{hij} has an auxiliary variable x_{hij} , and x_{hij} 's are observed for all units. The imputed estimator of \bar{Y} is then given by

$$\bar{y}_I = \sum_{(hij) \in A_r} w_{hij} y_{hij} + \sum_{(hij) \in A_m} w_{hij} y_{hij}^*,$$

where y_{hij}^* takes the value

$$\hat{\rho} x_{hij} + (y_{glk} - \hat{\rho} x_{glk}), \quad (hij) \in A_m, \quad (glk) \in A_r$$

with probability

$$\frac{w_{glk}}{\sum_{(hij) \in A_r} w_{hij}},$$

and

$$\hat{\rho} = \frac{\sum_{(hij) \in A_r} w_{hij} y_{hij}}{\sum_{(hij) \in A_r} w_{hij} x_{hij}}.$$

Note that the simple hot deck imputation described previously is a special case of the ratio hot deck imputation with $x_{hij} = 1$ for all (hij) . Because

$$\begin{aligned} E_* y_{hij}^* &= x_{hij} \hat{\rho} + \sum_{(glk) \in A_m} (y_{glk} - x_{glk} \hat{\rho}) \frac{w_{glk}}{\sum_{(hij) \in A_r} w_{hij}} \\ &= x_{hij} \hat{\rho}, \end{aligned}$$

we have

$$\begin{aligned}
E_* \bar{y}_I &= \sum_{A_r} w_{hij} y_{hij} + \hat{\rho} \sum_{A_r} w_{hij} x_{hij} \\
&= \frac{(\sum_{A_r} w_{hij} y_{hij})(\sum_{A_n} w_{hij} x_{hij})}{\sum_{A_r} w_{hij} x_{hij}} \\
&= \frac{\hat{S}\hat{U}}{\hat{T}}, \text{ say.}
\end{aligned}$$

Using the uniform response assumption again, we get $E\hat{S} = p\bar{Y}$, $E\hat{U} = \bar{X}$, and $E\hat{T} = p\bar{X}$. Therefore $E\bar{y}_I = \bar{Y}$. The imputed estimator \bar{y}_I is asymptotically unbiased for \bar{Y} .

3.2.2 The Adjusted BRR Variance Estimator

We need to define the following terms to construct the adjusted BRR variance estimator. Let

$$\hat{S}^{(r)} = \sum_{A_r} (d_{hi}^r + 1) w_{hij} y_{hij},$$

$$\hat{U}^{(r)} = \sum_{A_n} (d_{hi}^r + 1) w_{hij} x_{hij},$$

$$\hat{T}^{(r)} = \sum_{A_r} (d_{hi}^r + 1) w_{hij} x_{hij},$$

and adjust the r th balanced sample mean with imputed values (3.4) by

$$\bar{y}_{Ia}^{(r)} = \sum_{A_r} (d_{hi} + 1) w_{hij} y_{hij} + \sum_{A_m} (d_{hi} + 1) w_{hij} \left(y_{hij}^* + x_{hij} \frac{\hat{S}^{(r)}}{\hat{T}^{(r)}} - x_{hij} \frac{\hat{S}}{\hat{T}} \right).$$

Because $E_* y_{hij}^* = x_{hij} \hat{\rho}$ and $\hat{\rho} = \frac{\hat{S}}{\hat{T}}$, we have

$$\begin{aligned} E_* \bar{y}_{I_a}^{(r)} &= \sum_{A_r} (d_{hi}^r + 1) w_{hij} y_{hij} + \sum_{A_m} (d_{hi}^r + 1) w_{hij} x_{hij} \frac{\hat{S}^{(r)}}{\hat{T}^{(r)}} \\ &= \hat{S}^{(r)} + (\hat{U}^{(r)} - \hat{T}^{(r)}) \frac{\hat{S}^{(r)}}{\hat{T}^{(r)}} \\ &= \frac{\hat{S}^{(r)} \hat{U}^{(r)}}{\hat{T}^{(r)}}. \end{aligned}$$

It is easy to show that

$$E(\hat{S}^{(r)} | A_n) = p\bar{y}, \quad E(\hat{U}^{(r)} | A_n) = \bar{x}, \quad \text{and} \quad E(\hat{T}^{(r)} | A_n) = p\bar{x}.$$

Thus $E \bar{y}_{I_a}^{(r)} \doteq \bar{Y}$, i.e., the adjusted r th balanced sample mean is asymptotically unbiased for \bar{Y} .

The adjusted BRR variance estimator is then defined by

$$v_{BRR}^a = \frac{1}{R} \sum_{\tau=1}^R (\bar{y}_{I_a}^{(\tau)} - \bar{y}_I)^2.$$

We will establish the asymptotic normality of \bar{y}_I and asymptotic consistency of v_{BRR}^a in the next section.

3.2.3 Asymptotic Properties

Besides the conditions given in section 3.1.3, we need to assume another two conditions about x in order to study the asymptotic properties of \bar{y}_I and v_{BRR}^a .

Condition 1' $n^{1+\delta} \sum_{h=1}^L \sum_{i=1}^{n_h} E | r_{hi} - E r_{hi} |^{2+\delta} = O_p(1)$, for some $\delta > 0$, as $n \rightarrow \infty$, where $n = \sum_{h=1}^L n_h$, $r_{hi} = \sum_{j=1}^{n_{hi}} w_{hij} x_{hij}$.

Condition 4' $\sum_{A_n} w_{hij} | x_{hij} - \bar{X} |^{2+\delta} = O_p(1)$

The method of proof we used here is very similar to that used in section 3.1.3.

Asymptotic Normality of \bar{y}_I

As in the case of $x_{hij} \equiv 1$, we can write

$$\bar{y}_I - \bar{Y} = \left(\frac{\hat{S}\hat{U}}{\hat{T}} - \bar{Y} \right) + \bar{z}^*,$$

where $\bar{z}^* = \sum_{A_n} z_{hij}^*$, $z_{hij}^* = (1 - a_{hij})w_{hij}(y_{hij}^* - E_r y_{hij}^*)$. To prove

$$a_n^{-1} \left(\frac{\hat{S}\hat{U}}{\hat{T}} - \bar{Y} \right) \longrightarrow_d N(0, 1)$$

we define

$$\delta_S = \frac{\hat{S} - p\bar{Y}}{p\bar{Y}}, \quad \delta_U = \frac{\hat{U} - \bar{X}}{\bar{X}}, \quad \delta_T = \frac{\hat{T} - p\bar{X}}{p\bar{X}}.$$

Under condition 1 and 1', we can show by Lemma 1 that δ_S, δ_U and δ_T converge to normal random variables with mean zero and are all $O_p(1)$. Then

$$\begin{aligned} \frac{\hat{S}\hat{U}}{\hat{T}} - \bar{Y} &= \frac{p\bar{Y}(\delta_S + 1)\bar{X}(\delta_U + 1)}{p\bar{X}(\delta_T + 1)} - \bar{Y} \\ &= \bar{Y}(\delta_S + \delta_U + \delta_S\delta_U + 1)(1 - \delta_T + \delta_T^2 - \dots) - \bar{Y} \\ &= \bar{Y}(\delta_S + \delta_U - \delta_T) + O_p\left(\frac{1}{n}\right). \end{aligned}$$

Hence,

$$a_n^{-1} \left(\frac{\hat{S}\hat{U}}{\hat{T}} - \bar{Y} \right) \longrightarrow_d N(0, 1)$$

by Slutsky's theorem, where $a_n^2 \doteq \text{var}\left(\frac{\hat{S}\hat{U}}{\hat{T}}\right) = \text{var}(E_*\bar{y}_I)$.

Turning to $\bar{z}^* = \sum_{A_n} z_{hij}^* = \sum_{h=1}^L \sum_{i=1}^{n_h} t_{hi}^*$, we also need to establish a useful inequality

$$E_* |t_{hij}^*|^{2+\delta} \leq O_p(1) \left(\sum_j w_{hij}\right)^{2+\delta}. \quad (3.11)$$

To do this, first by definition we have

$$\begin{aligned} E_* |t_{hi}^*|^{2+\delta} &= \sum_{(gk) \in A_r} \left[\frac{w_{gk}}{\sum_{(hij) \in A_r} w_{hij}} \left| \sum_j w_{hij} (1 - a_{hij}) (y_{gk} - x_{gk} \frac{\hat{S}}{\hat{T}}) \right|^{2+\delta} \right] \\ &\leq \left(\sum_j w_{hij}\right)^{2+\delta} \cdot \sum_{(gk) \in A_n} w_{gk} \left| y_{gk} - x_{gk} \frac{\hat{S}}{\hat{T}} \right|^{2+\delta} \\ &\quad \cdot \frac{1}{\sum_{(hij) \in A_r} w_{hij}}. \end{aligned} \quad (3.12)$$

Then note that

$$\sum_{(hij) \in A_r} w_{hij} \xrightarrow{p} p, \quad (3.13)$$

and

$$\begin{aligned} \sum_{A_n} w_{gk} \left| y_{gk} - x_{gk} \frac{\hat{S}}{\hat{T}} \right|^{2+\delta} &\leq O_p(1) \left(\sum_{A_n} w_{gk} \left| y_{gk} - \bar{Y} \right|^{2+\delta} \right. \\ &\quad \left. + \sum_{A_n} w_{gk} \left| \bar{Y} - x_{gk} \frac{\hat{S}}{\hat{T}} \right|^{2+\delta} \right) \\ &\leq O_p(1) + O_p(1) \sum_{A_n} \left| \bar{Y} - x_{gk} \frac{\hat{S}}{\hat{T}} \right|^{2+\delta} \end{aligned} \quad (3.14)$$

by condition 1. Because we can prove

$$\frac{\hat{S}}{\hat{T}} = \frac{\bar{Y}}{\bar{X}} + O_p\left(\frac{1}{\sqrt{n}}\right)$$

using delta method, so

$$\begin{aligned}
\sum_{A_n} w_{gk} \left| \bar{Y} - x_{gk} \frac{\hat{S}}{\hat{T}} \right|^{2+\delta} &= \sum_{A_n} w_{gk} \left| \bar{Y} - x_{gk} \frac{\bar{Y}}{\bar{X}} + O_p\left(\frac{1}{\sqrt{n}}\right) \right|^{2+\delta} \\
&\leq \left(\frac{\bar{Y}}{\bar{X}} \right)^{2+\delta} + O_p\left(\frac{1}{n}\right) \\
&= O_p(1)
\end{aligned} \tag{3.15}$$

by condition 4'. Therefore (3.11) follows from (3.12), (3.13), (3.14) (3.15).

For given y_{obs} and a as defined before, using (3.11) and condition 3 we have

$$\frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} E_* \left| nt_{hi}^* \right|^{2+\delta} \leq O_p(1).$$

Hence

$$b_n^{-1} \bar{z}^* \longrightarrow_{d|y_{obs}, a} N(0, 1)$$

follows from Lemma 1, where b_n^2 is the asymptotic expectation of $var_*(\bar{z}^*)$, or $b_n^2 \doteq E(var_* \bar{y}_I)$. Once again, applying Lemma 4 with $X_n = \frac{\hat{S}\hat{U}}{\hat{T}} - \bar{Y}$, $W_n = \bar{z}^*$ and $Z_n = (y_{obs}, a)$, we conclude that

$$c_n^{-1} (\bar{y}_I - \bar{Y}) \longrightarrow_d N(0, 1),$$

where $c_n^2 = a_n^2 + b_n^2$ is the asymptotic variance of \bar{y}_I .

Asymptotic Consistency of v_{BRR}^a

First we can write

$$\bar{y}_{Ia}^{(\tau)} - \bar{y}_I = \left(\frac{\hat{S}^{(\tau)} \hat{U}^{(\tau)}}{\hat{T}^{(\tau)}} - \frac{\hat{S} \hat{U}}{\hat{T}} \right) + \sum_{A_n} d_{hi}^r z_{hij}^*.$$

Then the adjusted BRR variance estimator under ratio hot deck imputation can be expressed by

$$\begin{aligned}
v_{BRR}^a &= \frac{1}{R} \sum_{r=1}^R \left(\bar{y}_{Ia}^{(r)} - \bar{y}_I \right)^2 \\
&= \frac{1}{R} \sum_{r=1}^R \left(\hat{A}^{(r)} \right)^2 + \frac{1}{R} \sum_{r=1}^R \left(\sum_{A_n} d_{hi}^r z_{hij}^* \right)^2 \\
&\quad + \frac{2}{R} \sum_{r=1}^R \left(\hat{A}^{(r)} \sum_{A_n} d_{hi}^r z_{hij}^* \right) \\
&= \tilde{A} + \tilde{B} + \tilde{C}, \text{ say,}
\end{aligned}$$

where

$$\hat{A}^{(r)} = \frac{\hat{S}^{(r)} \hat{U}^{(r)}}{\hat{T}^{(r)}} - \frac{\hat{S} \hat{U}}{\hat{T}}.$$

The asymptotic consistency of v_{BRR}^a follows from

1. $n \left[\tilde{A} - \text{var}(E_* \bar{y}_I) \right] \longrightarrow_p 0$,
2. $n \left[\tilde{B} - E(\text{var}_* \bar{y}_I) \right] \longrightarrow_p 0$,
3. $n \tilde{C} \longrightarrow_p 0$.

The proof of 1-3 exactly follows the lines of the proof given in the section 3.1.3. We need only to apply lemma 2 by using (3.11) and $\hat{A}^{(r)} = O_p(\frac{1}{n})$ which can be obtained by expanding $f(\hat{S}^{(r)}, \hat{U}^{(r)}, \hat{T}^{(r)}) = \frac{\hat{A}^{(r)} \hat{U}^{(r)}}{\hat{T}^{(r)}}$ at $(\hat{S}, \hat{U}, \hat{T})$.

An advantage of the BRR method is that we can extend it easily to the case where the estimator is the function of the mean. Suppose that the parameter of interest is $\theta = f(\bar{Y})$ and the estimator of θ is $\hat{\theta}_I = f(\bar{y}_I)$, where \bar{Y} and \bar{y}_I are the mean vectors of the population and the sample with imputed values respectively.

Then the BRR variance estimator of $var(\hat{\theta}_I)$ is given by

$$v_{BRR}^a(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}_I^{(r)} - \hat{\theta}_I)^2$$

where $\hat{\theta}_I^{(r)} = f(\bar{y}_I^{(r)})$. The consistency of $v_{BRR}^a(\hat{\theta})$ follows directly from Taylor's expansion and the consistency of v_{BRR}^a .

Chapter 4

Comparison of BRR and Jackknife Variance Estimators

The linearization (or Taylor's expansion), the jackknife and the balanced repeated replication methods are three well-known methods of estimating variances. When there is no missing value, Rao and Wu (1985) established second-order asymptotic expansions of the three variance estimators which enable us to compare them. In this chapter, we will first present Rao and Wu's result, then compare Rao and Shao's adjusted jackknife variance estimator with the adjusted BRR variance estimator given in section 3.1, and finally show that a result similar to Rao and Wu's holds.

4.1 The Original BRR and Jackknife Variance Estimators

We are still consider the stratified multistage sampling design. Our parameter of interest is now a function of the population mean \bar{Y} , say $\theta = g(\bar{Y})$, and the characteristic y is a vector. The linearization method gives the following variance estimator of $\hat{\vartheta} = g(\bar{y})$:

$$v_L = \sum_{h=1}^L \frac{1}{n_h} \nabla g(\bar{y})' s_h^2 \nabla g(\bar{y}),$$

where

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

and

$$y_{hi} = \sum_{j=1}^{n_{hi}} n_h w_{hij} y_{hij}, \quad \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}.$$

The jackknife variance estimator for $\hat{\theta}$ is given by

$$v_{JACK} = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{hi} - \hat{\theta})^2,$$

where $\hat{\theta}_{hi} = g(\bar{y}^{hi})$ and \bar{y}^{hi} is the unbiased estimator of \bar{Y} from the sample after deleting the i th first-stage cluster in the h th stratum ($i = 1, 2, \dots, n_h, h = 1, 2, \dots, L$), i.e.,

$$\bar{y}^{hi} = \sum_{l \neq h} \sum_{i=1}^{n_l} \sum_{j=1}^{n_{li}} w_{hij} y_{hij} + \frac{n_h}{n_h - 1} \sum_{l \neq i} \sum_{j=1}^{n_{hl}} w_{hlj} y_{hlj}.$$

In the special case of $n_h = 2$ for all h , the jackknife variance estimator v_{JACK} reduced to

$$v_{JACK-D} = \frac{1}{4} \sum_{h=1}^L (\hat{\theta}_{h1} - \hat{\theta}_{h2})^2$$

which is given in Kish and Frankel (1974).

The BRR variance estimator is given by

$$v_{BRR} = \frac{1}{R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta})^2,$$

where $\hat{\theta}^{(r)} = g(\bar{y}^{(r)})$, $\bar{y}^{(r)}$ is the r th balanced sample mean as defined in (2.1).

If g is linear, $v_L = v_{JACK} = v_{BRR}$ and it is unbiased for $\text{var}(\hat{\theta})$. In general, v_L, v_{JACK}, v_{BRR} are asymptotically consistent for $\text{var}(\hat{\theta})$ (Krewski and Rao 1981; Bickel and Freedman 1984).

Rao and Wu (1985) derived the second-order asymptotic comparison of v_L , v_{JACK} and v_{BRR} in the case of $n_h = 2$ for all h . Their result was extended by Shao (1993) to the case of $n_h \geq 2$. This result can be written in the following theorem:

Theorem Suppose condition 1 holds with $\delta = 2$, and condition 3, 4 hold. Suppose further g'' is continuous in a neighborhood of \bar{Y} . Then

- i) $v_{JACK} = v_L + O_p(n^{-2})$;
- ii) $v_{JACK} = v_L + O_p(n^{-3})$, if $n_h = 2$ for all h ;
- iii) $v_{BRR} = v_L + O_p(n^{-\frac{3}{2}})$.

Therefore, we have

$$v_{BRR} = v_{JACK} + O_p(n^{-\frac{3}{2}}).$$

4.2 The Adjusted BRR and Jackknife Variance Estimators

We now consider the adjusted jackknife variance estimator v_{JACK}^a and adjusted BRR variance estimator v_{BRR}^a . In the presence of nonresponse, the imputed estimator of \bar{Y} is \bar{y}_I (3.1) and $E_*\bar{y}_I = \frac{SU}{T}$ (3.2). To construct an adjusted jackknife variance estimator, we need to define the following terms obtained by deleting the i th first-stage cluster in the h th stratum (Rao and Shao 1992):

$$S_{-hi} = \sum_{(gk) \in A_{rg} \neq h} w_{gk} y_{gk} + \frac{n_h}{n_h - 1} \sum_{(gk) \in A_{rg} = h, i \neq i} w_{gk} y_{gk},$$

$$T_{-hi} = \sum_{(gk) \in A_{rg} \neq h} w_{gk} + \frac{n_h}{n_h - 1} \sum_{(gk) \in A_{rg} = h, i \neq i} w_{gk},$$

$$U_{-hi} = \sum_{(gk) \in A_{ng} \neq h} w_{gk} + \frac{n_h}{n_h - 1} \sum_{(gk) \in A_{ng} = h, i \neq i} w_{gk}.$$

We assume that $T_{-hi} \neq 0$ for all h, i . Note that $T_{-hi} = 0$ implies that the respondents to item y are confined only to the (h, i) th first-stage unit which is clearly unrealistic. Define also for $(glk) \in A_m, (gl) \neq (hi)$ that

$$z_{glk}^{*(hi)} = y_{glk}^* + \frac{S_{-hi}}{T_{-hi}} - \frac{S}{T},$$

that is, whenever the (hi) th cluster is deleted, each of the imputed value y_{glk}^* is adjusted by an amount $S_{-hi}/T_{-hi} - S/T$. It follows that

$$E_*(z_{glk}^{*(hi)}) = \frac{S_{-hi}}{T_{-hi}} \quad (4.1)$$

since $E_*y_{glk}^* = S/T$. Using this adjusted imputed value $z_{glk}^{*(hi)}$ for $(glk) \in A_m$, we can construct an adjusted imputed estimator for \bar{Y} when (hi) th cluster is deleted:

$$\bar{y}_I^a(-hi) = S_{-hi} + \sum_{(glk) \in A_m, g \neq h} w_{glk} z_{glk}^{*(hi)} + \frac{n_h}{n_h - 1} \sum_{(glk) \in A_m, l \neq i} w_{glk} z_{glk}^{*(hi)} \quad (4.2)$$

$\bar{y}_I^a(-hi)$ is approximately unbiased for \bar{Y} because

$$E_*[\bar{y}_I^a(-hi)] = \frac{S_{-hi}U_{-hi}}{T_{-hi}}$$

which follows from (4.1), (4.2), and $E(S_{-hi}) = p\bar{Y}, E(U_{-hi}) = 1, E(T_{-hi}) = p$. The adjusted jackknife estimator of $var(\bar{y}_I)$ is then given by

$$v_{JACK}^a = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} [\bar{y}_I^a(-hi) - \bar{y}_I]^2.$$

After some algebra, we can write

$$\bar{y}_I^a(-hi) - \bar{y}_I = \frac{S_{-hi}U_{-hi}}{T_{-hi}} - \frac{SU}{T} + \frac{1}{n_h - 1} \sum_{i=1}^{n_h} t_{hi}^* - \frac{n_h}{n_h - 1} t_{hi}^*,$$

where $t_{hi}^* = \sum_{j=1}^{n_{hi}} w_{hij}(1 - a_{hij})(y_{hij}^* - \frac{S}{T})$. Hence v_{JACK}^a may be expressed as

$$\begin{aligned} v_{JACK}^a &= \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(\frac{S_{-hi}U_{-hi}}{T_{-hi}} - \frac{SU}{T} \right)^2 \\ &+ \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(t_{hi}^* - \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi}^* \right)^2 \\ &+ 2 \sum_{h=1}^L \sum_{i=1}^{n_h} \left(\frac{S_{-hi}U_{-hi}}{T_{-hi}} - \frac{SU}{T} \right) \left(t_{hi}^* - \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi}^* \right) \\ &= A_1 + B_1 + C_1, \text{ say.} \end{aligned}$$

Recalling (3.5), (3.6), we can express the adjusted BRR variance estimator as

$$\begin{aligned} v_{BRR}^a &= \frac{1}{R} \sum_{r=1}^R \left(\frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T} \right)^2 \\ &+ \sum_{h=1}^L \sum_{i=1}^{n_h} (t_{hi}^*)^2 - \sum_{h=1}^L \sum_{i \neq i'} \frac{2}{n_h - 1} t_{hi}^* t_{hi'}^* \\ &+ \frac{2}{R} \sum_{r=1}^R \left(\frac{S^{(r)}U^{(r)}}{T^{(r)}} - \frac{SU}{T} \right) \left(\sum_{A_n} d_{hi}^r z_{hij}^* \right) \\ &= A + B + C, \text{ say.} \end{aligned}$$

It is easy to check that $B = B_1$, and

$$A = A_1 + O_p(n^{-\frac{3}{2}})$$

by the theorem above with $\hat{\theta} = g(\bar{y}) = \frac{SU}{T}$ and

$$\bar{y} = \left(\sum_{A_n} w_{hij} a_{hij} y_{hij}, \sum_{A_n} w_{hij}, \sum_{A_n} w_{hij} a_{hij} \right).$$

We want to show that $C = C_1 + O_p(n^{-3/2})$. In section 3.1.3 we have written

$nC = \frac{1}{n} \sum_t X_t$, and proved that

$$\frac{1}{n} \sum_t E_* |X_t|^{2+\delta} \leq O_p(1).$$

We can also write

$$\begin{aligned} nC_1 &= \frac{1}{n} \sum_h \sum_i [2n^2 A_{hi} t_{hi}^* - 2n^2 A_{hi} \frac{1}{n_h} \sum_i t_{hi}^*] \\ &= \frac{1}{n} \sum_t X_t^1 - \frac{1}{n} \sum_t X_t^2, \text{ say,} \end{aligned}$$

where $X_t^1 = 2n^2 A_{hi} t_{hi}^*$, $X_t^2 = 2n^2 A_{hi} \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi}^*$ and

$$A_{hi} = \frac{S_{-hi} U_{-hi}}{T_{-hi}} - \frac{SU}{T}.$$

We can show that $A_{hi} = O_p(n^{-1})$. Using this result, condition 4 and established fact $E_* |t_{hi}^*|^{2+\delta} \leq O_p(1)(\sum_j w_{hij})^{2+\delta}$, we get

$$\begin{aligned} \frac{1}{n} \sum_t E_* |X_t^1|^{2+\delta} &= O_p(1) n^{3+2\delta} \sum_h \sum_i |A_{hi}|^{2+\delta} E_* |t_{hi}^*|^{2+\delta} \\ &\leq O_p(1) n^{1+\delta} \sum_h \sum_i (\sum_j w_{hij})^{2+\delta} \\ &= O_p(1). \end{aligned}$$

Similarly, we have

$$\frac{1}{n} \sum_t E_* |X_t^2|^{2+\delta} \leq O_p(1)$$

Therefore

$$\begin{aligned} \frac{1}{n} E_* |X_t - (X_t^1 - X_t^2)|^{2+\delta} &\leq \frac{1}{n} \sum_t E_* |X_t|^{2+\delta} + \frac{1}{n} \sum_t E_* |X_t^1|^{2+\delta} \\ &\quad + \frac{1}{n} \sum_t E_* |X_t^2|^{2+\delta} \\ &\leq O_p(1). \end{aligned}$$

Note that $nC - nC_1 = \frac{1}{n} \sum_t [X_t - (X_t^1 - X_t^2)]$ and $E_*C = E_*C_1 = 0$ since $E_*t_{hi}^* = 0$. Applying lemma 1 to the random variables $X_t - (X_t^1 - X_t^2)$, we have

$$\sqrt{n}(nC - nC_1) \longrightarrow_d N(0, \tau^2)$$

for some $\tau^2 > 0$. This implies

$$C = C_1 + O_p(n^{-\frac{3}{2}})$$

Thus we obtain

$$v_{BRR}^a = v_{JACK}^a + O_p(n^{-\frac{3}{2}})$$

which means that the relationship between the jackknife and BRR variance estimators for the complete data set is still held for the data set with imputed values.

Chapter 5

Simulation Studies

Theoretically, we have proved that the adjusted BRR and jackknife variance estimators are consistent estimators of $\text{var}(\bar{y}_I)$, the difference between them is of the order $O_p(n^{-3/2})$, and they are both better than the naive BRR and jackknife variance estimators. In this chapter we will present a simulation study to compare the performance of these variance estimators numerically.

5.1 The Finite Population

The finite population used here is similar to the population given by Hansen and Tepping (1985) in the National Assessment of Educational Progress Study. The population consists of $L = 32$ strata. There are N_h clusters in the h -th stratum with 20 ultimate units in each cluster. The stratum mean is μ_h and stratum variance is such that the correlation coefficient between any two ultimate units in one cluster is ρ ($\rho = 0.0, 0.1, 0.3, 0.5$). To do this, we first generate random variables

$$y_{hi} \stackrel{i.i.d.}{\sim} N(\mu_h, v_h^2), \quad h = 1, 2, \dots, L, \quad i = 1, 2, \dots, N_h.$$

Then generate random variables

$$\varepsilon_{hij} \stackrel{i.i.d.}{\sim} N(0, \sigma_{\varepsilon_h}^2), \quad j = 1, 2, \dots, 20$$

with $\sigma_{\varepsilon_h}^2 = \frac{1-\rho}{\rho}v_h^2$ if $\rho \neq 0$, and $\sigma_{\varepsilon_h}^2 = v_h^2$ if $\rho = 0$. Define the ultimate units by

$$y_{hij} = y_{hi} + \varepsilon_{hij} \quad \text{for } \rho \neq 0,$$

$$y_{hij} = \mu_h + \varepsilon_{hij} \quad \text{for } \rho = 0.$$

The population $\{y_{hij}\}$ is now generated. Note here the h th stratum variance σ_h^2 is $1/\rho v_h^2$ if $\rho \neq 0$, or v_h^2 if $\rho = 0$. The parameters of population are given in table 5.1.

Table 5.1 Parameters of the Population

stratum h	size N_h	μ_h	v_h	stratum h	size N_h	μ_h	v_h
1	13	100.00	10.00	2	16	95.00	9.50
3	20	90.00	9.00	4	25	98.00	9.80
5	25	93.00	9.30	6	25	98.00	9.80
7	25	96.00	9.60	8	28	94.00	9.40
9	28	92.00	9.20	10	28	96.00	9.60
11	31	94.00	9.40	12	31	92.00	9.20
13	31	90.00	9.00	14	31	96.00	9.60
15	31	94.00	9.40	16	31	92.00	9.20
17	31	90.00	9.00	18	31	88.00	8.80
18	31	86.00	8.60	20	34	84.00	8.40
21	34	82.00	8.20	22	34	80.00	8.00
23	34	90.00	9.00	24	37	85.00	8.50
25	37	80.00	8.00	26	37	90.00	9.00
27	37	85.00	8.50	28	39	80.00	8.00
29	39	75.00	7.50	30	42	75.00	7.50
31	42	75.00	7.50	32	42	75.00	7.50

5.2 The Sample and Estimators

The sample is obtained in the following way. First, we draw $n_h = 2$ clusters from stratum h with equal probability $\frac{1}{N_h}$. Whenever a cluster is selected, all the ultimate

units in this cluster are selected. Therefore the sample A_n is of the size $n = 32 \cdot 2 \cdot 20 = 1280$. Secondly, the corresponding response set A_r is generated by choosing a uniform random variable x_{hij} . If x_{hij} less than or equal to the response rate, $(hij) \in A_r$, otherwise $(hij) \in A_m = A_n - A_r$. For each $(hij) \in A_m$, $(glk) \in A_r$ is selected with probability $N_g / \sum_{(hij) \in A_r} N_h$ and missing value y_{hij} is imputed by y_{glk} . Finally, we get the stratified sample A_n^* with imputed values.

For A_n^* , \bar{y}_I is calculated according to (3.1). Using the formulas given in section 4.1, section 3.1.2, and a 31×32 BRR specification matrix which is obtained by deleting the first row of all 1 in a 32×32 hadamard matrix, we get the naive and adjusted BRR variance estimators, v_{BRR}^n and v_{BRR}^a . The adjusted jackknife variance estimator v_{JACK}^a is calculated according to the formulas given in section 4.2. Note that the naive BRR and jackknife variance estimators are equal because of $\hat{\theta} = \bar{y}$.

5.3 The Measures of The Performance

Because the true $var(\bar{y}_I)$ is unknown, we generate 10000 stratified sample A_n^* 's and use the empirical variance

$$V(\bar{y}_I) = \frac{1}{10000 - 1} \sum_{i=1}^{10000} (\bar{y}_{Ii} - \frac{1}{10000} \sum_{j=1}^{10000} \bar{y}_{Ij})^2$$

to estimate $Var(\bar{y}_I)$. The relative bias(RB) and root mean square errors(RMSE) are used to compare the performance of v_{BRR}^n , v_{BRR}^a and v_{JACK}^a through 2000 sample A_n^* 's, where

$$RB = \frac{\bar{v} - V(\bar{y}_I)}{V(\bar{y}_I)} \quad \text{and}$$

$$RMSE = \left[\frac{1}{2000 - 1} \sum_{i=1}^{2000} (v_i - \bar{Y})^2 \right]^{1/2}$$

with

$$\bar{v} = \frac{1}{2000} \sum_{i=1}^{2000} v_i, \quad v = v_{BRR}^n, v_{BRR}^a, v_{JACK}^a$$

Table 4.2 gives the simulation results which can be summarized as follows:

1. For any ρ and response rate, the RB's of v_{BRR}^n are negative which means that v_{BRR}^n does underestimate the true variance. Moreover, for fixed ρ , the underestimation grows rapidly when the response rate decreases. Approximately, the RB is $-30\% \sim -60\%$ in accordance with the response rate $80\% \sim 50\%$. For fixed response rate, the underestimation tends to increase when ρ increases. But the speed is not as high as the RB to the response rate.
2. In terms of the RMSE, the v_{BRR}^n is also getting worse when response rate decreases. For fixed ρ , the RMSE when response rate is 50% can be 3 times as much as the RMSE when response rate is 90%. For fixed response rate, the RMSE actually increases a little bit when the population variance increases. Remember that the h th stratum variance is $\frac{1}{\rho}v_h^2$ for $\rho \neq 0$, or v_h^2 for $\rho = 0$. But it is much more stable compared with the RMSE to the response rate.
3. In terms of both the RB and the RMSE, v_{BRR}^a and v_{JACK}^a have almost the same performance which ensure us that the difference between them is negligible when sample size is large.
4. The RB's of v_{BRR}^a and v_{JACK}^a have no tendency to change while the response rate or ρ changes. They are all around 8%. The RMSE of v_{BRR}^a and v_{JACK}^a are also insensitive to the response rate and have a low positive correlation with the population variance. They are all less than the RMSE of v_{BRR}^n , if response rate $\leq 80\%$. Therefore v_{BRR}^a and v_{JACK}^a are much better than v_{BRR}^n in the sense of stability and closeness to the true variance.

Table 4.2 Simulation Results

correlation ρ	response rate(%)	v_{BRR}^n		v_{BRR}^a		v_{JACK}^a	
		RB	RMSE	RB	RMSE	RB	RMSE
0	90	-0.1357	0.0251	0.0382	0.0268	0.0372	0.0265
	80	-0.2422	0.0363	0.0713	0.0355	0.0686	0.0350
	70	-0.3434	0.0540	0.0856	0.0460	0.0848	0.0452
	60	-0.4608	0.0852	0.0520	0.0539	0.0491	0.0530
	50	-0.5550	0.1264	0.0335	0.0662	0.0313	0.0653
0.1	90	-0.1422	0.5892	0.0519	0.6282	0.0508	0.6193
	80	-0.2791	0.7699	0.0912	0.7168	0.0885	0.7015
	70	-0.4130	1.0605	0.1061	0.7891	0.1051	0.7785
	60	-0.5441	1.4823	0.0834	0.8460	0.0807	0.8328
	50	-0.6384	1.8841	0.0896	0.9277	0.0868	0.9073
0.3	90	-0.1381	0.4324	0.0606	0.4712	0.0592	0.4639
	80	-0.2853	0.5485	0.0980	0.5168	0.0955	0.5087
	70	-0.4357	0.7503	0.1010	0.5267	0.1003	0.5208
	60	-0.5728	1.0103	0.0830	0.5556	0.0804	0.5473
	50	-0.6687	1.2088	0.1082	0.6068	0.1049	0.5925
0.5	90	-0.1359	0.3992	0.0643	0.4391	0.0629	0.4321
	80	-0.2872	0.5024	0.1002	0.4762	0.0978	0.4759
	70	-0.4443	0.6876	0.0970	0.4718	0.0965	0.4668
	60	-0.5825	0.9109	0.0831	0.4953	0.0806	0.4879
	50	-0.6794	1.0661	0.1169	0.5425	0.1135	0.5299

Appendix A

Computer program for simulation

The following program was used for the simulation described in chapter 5. The program was written in FORTRAN. Subroutines DRNNOA and RNUND which are used to generate normal and discrete uniform random variables are called from the IMSL FORTRAN library.

```
C ***** THE PROGRAM FOR SIMULATION *****
C
C ***** GLOBAL VARIABLES *****

C N:          NUMBER OF SIMULATION
C L(H):       H-TH STRATUM SIZE
C M(H):       H-TH STRATUM MEAN
C V(H):       H-TH STRATUM VARIANCE PARAMETER
C RHO:        CORRELATION COEFFICIENT
C B(31,32):   BRR SPECIFICATION MATRIX
C POPM:       POPULATION MEAN
C PP(32,42,20): POPULATION UNIT
C SS(32,2,20): SAMPLE UNIT
C IS(32,2,20): INDICATOR VARIABLE FOR SS(32,2,20)
C SSS(32,2,20): SAMPLE UNIT WITH IMPUTED VALUE
C RESP:       RESPONSE RATE(%)
C YBAR:       SAMPLE MEAN WITH IMPUTED VALUE
C BHM(R):     BALANCED HALF SAMPLE MEAN-YBAR
C ABHM(R):    ADJUSTED BALANCED HALF SAMPLE MEAN-YBAR
C NVBRR:      NAIVE BRR VARIANCE ESTIMATE
C AVBRR:      ADJUSTED BRR VARIANCE ESTIMATE
C AVJACK(I):  ADJUSTED JACKKINFE VARIANCE ESTIMATE .
C MEAN:       MEAN OF YBAR
C MNB:        MEAN OF NVBRR
C MAB:        MEAN OF AVBRR
```

```

C MAJ:          MEAN OF AVJACK
C VAR:          VARIANCE OF YBAR
C VNB:          MSE OF NVBRR
C VAB:          MSE OF AVBRR
C VAJ:          VARIANCE OF AVJACK
C RB1:          RELATIVE BIAS OF NAIVE BRR VAR. EST.
C RB2:          RELATIVE BIAS OF ADJUSTED BRR VAR. EST.
C RB3:          RELATIVE BIAS OF ADJUSTED JACKKINFE VAR. EST.
C
C          INTEGER N, RESP, L(32), B(31,32), IS(32,2,20)
C          DOUBLE PRECISION M(32), V(32), PP(32,42,20), POPM
C          DOUBLE PRECISION RHO, SS(32,2,20), SSS(32,2,20)
C          DOUBLE PRECISION YBAR, BHM(31), ABHM(31)
C          DOUBLE PRECISION NVBRR(2000), AVBRR(2000), AVJACK(2000)
C          DOUBLE PRECISION MEAN, MNB, MAB, MAJ, VAR, VNB
C          DOUBLE PRECISION VAB, VAJ, RB1, RB2, RB3
C
C          *** LOCAL VARIABLES ***
C
C          INTEGER RND(2), RAD(1), RBD(20), RCD(1), RDD(1), SEED
C          DOUBLE PRECISION Y(1000), YY(32)X(20000), P(32,42)
C          DOUBLE PRECISION T, S, BT(32,2), AT(31), AS(31)
C          DOUBLE PRECISION THAT(32,2), SHAT(32,2), W(31)
C          DOUBLE PRECISION YBAR2, AYBAR(32,2), WW(32,2)
C          DOUBLE PRECISION AP(32,2), BP(32,2), TT(32,2)
C
C          **** INPUT INITIAL DATA ****
C
C          READ(1,10) (L(I),I=1,32)
C          READ(1,20) (M(I),I=1,32)
C          READ(1,30) (V(I),I=1,32)
C          DO 5 K=1,31
C          READ(2,40) (B(K,J), J=1,32)
C          5 CONTINUE
C          10 FORMAT(8I3)
C          20 FORMAT(8F5.1)
C          30 FORMAT(8F4.1)
C          40 FORMAT(32I2)
C          PRINT*, 'INPUT FINISHED'
C
C          **** GENERATE POPULATION ****
C
C          PRINT*, 'PLEASE ENTER RHO'
C          READ*, RHO
C          CALL RNSET(1)
C          CALL DRNNOA(1000, Y)
C          II=0
C          DO 70 IP=1,32
C          DO 50 J=1,42

```

```

      IF (J .LE. L(IP)) THEN
      P(IP,J)=M(IP)+V(IP)*Y(J+II)
      ELSE
      P(IP,J)=0.0
      END IF
50    CONTINUE
      II=II+L(IP)
70    CONTINUE
      CALL RNSET(100)
      CALL DRNNOA(20000,X)
      KK=0
      DO 90 I=1,32
      DO 90 J=1,L(I)
      DO 80 K=1,20
80    PP(I,J,K)=P(I,J)+X(K+KK)*V(I)*SQRT((1-RHO)/RHO)
      CONTINUE
      KK=KK+20
90    CONTINUE
      POPM=0.0D0
      DO 95 I=1,32
      DO 95 J=1,L(I)
      DO 95 K=1,20
95    POPM=POPM+PP(I,J,K)
      CONTINUE
      POPM=POPM/20000
      PRINT*, 'POPULATION GENERATED'
      SEED=80
C
C    *** START SIMULATION LOOP ***
C
      N=10000
      RESP=100
      DO 2000 KN=1,5
      RESP=RESP-10
      MEAN=0.0
      YBAR2=0.0
      DO 1000 IR=1,N
C
C    *** SELECT SAMPLE ***
C
      DO 110 I=1,32
105   CALL RNSET(SEED)
      CALL RNUND(2,L(I),RND)
      CALL RNGET(SEED)
      IF(RND(1) .EQ. RND(2)) GOTO 105
      DO 110 K=1,20
      SS(I,1,K)=PP(I,RND(1),K)
      SS(I,2,K)=PP(I,RND(2),K)
110   CONTINUE
C

```

```

C   *** SIMULATE MISSING VALUES ***
C
      DO 120 I=1,32
      DO 120 J=1,2
      CALL RNSET(SEED)
      CALL RNUND(20,100,RBD)
      CALL RNGET(SEED)
      DO 120 K=1,20
        IF (RBD(K) .LE. RESP) THEN
          IS(I,J,K)=1
        ELSE
          IS(I,J,K)=0
        END IF
120    CONTINUE
C
C   *** IMPUTE MISSING VALUES ***
C
      DO 140 I=1,32
      DO 140 J=1,2
      DO 140 K=1,20
        IF (IS(I,J,K) .EQ. 1) THEN
          SSS(I,J,K)=SS(I,J,K)
        ELSE
135    CALL RNSET(SEED)
          CALL RNUND(1,32,RAD)
          CALL RNUND(1,2,RCD)
          CALL RNUND(1,20,RDD)
          CALL RNGET(SEED)
          IF (IS(RAD(1),RCD(1),RDD(1)) .EQ. 0) GOTO 135
          SSS(I,J,K)=SS(RAD(1),RCD(1),RDD(1))
        END IF
140    CONTINUE
C
C   *** CALCULATE ESTIMATORS ***
C
      IF (IR .GT. 2000) GOTO 280
      DO 155 I=1,32
      DO 155 J=1,2
      AP(I,J)=0.0
      BP(I,J)=0.0
      TT(I,J)=0.0
      WW(I,J)=0.0
      DO 155 K=1,20
      AP(I,J)=AP(I,J)+SSS(I,J,K)
      BP(I,J)=BP(I,J)+SS(I,J,K)*IS(I,J,K)
      TT(I,J)=TT(I,J)+IS(I,J,K)
      WW(I,J)=WW(I,J)+1-IS(I,J,K)
155    CONTINUE
      DO 160 I=1,32
      YY(I)=AP(I,1)-AP(I,2)
160    CONTINUE

```

```

DO 175 I=1,31
BHM(I)=0.0D0
DO 170 J=1,32
BHM(I)=BHM(I)+B(I,J)*YY(J)*L(J)
170 CONTINUE
BHM(I)=BHM(I)/40000
175 CONTINUE
NVBRR(IR)=0.0D0
DO 180 I=1,31
NVBRR(IR)=NVBRR(IR)+BHM(I)**2
C DO 180 I=1,32
C NVBRR(IR)=NVBRR(IR)+YY(I)**2
180 CONTINUE
NVBRR(IR)=NVBRR(IR)/31
C

DO 195 I=1,31
AS(I)=0.0D0
AT(I)=0.0D0
W(I)=0.0D0
DO 190 J=1,32
IF (B(I,J) .EQ. 1) THEN
AS(I)=AS(I)+L(J)*BP(J,1)
AT(I)=AT(I)+L(J)*TT(J,1)
W(I)=W(I)+L(J)*WW(J,1)
ELSE
AS(I)=AS(I)+L(J)*BP(J,2)
AT(I)=AT(I)+L(J)*TT(J,2)
W(I)=W(I)+L(J)*WW(J,2)
END IF
190 CONTINUE
AS(I)=AS(I)/20000
AT(I)=AT(I)/20000
W(I)=W(I)/20000
195 CONTINUE
T=0.0D0
S=0.0D0
DO 200 I=1,32
DO 200 J=1,2
DO 200 K=1,20
T=T+IS(I,J,K)*L(I)
S=S+IS(I,J,K)*SS(I,J,K)*L(I)
200 CONTINUE
T=T/40000

S=S/40000
DO 210 I=1,31
ABHM(I)=BHM(I)+W(I)*(AS(I)/AT(I)-S/T)
210 CONTINUE
AVBRR(IR)=0.0D0
DO 220 I=1,31
AVBRR(IR)=AVBRR(IR)+ABHM(I)**2
220 CONTINUE
AVBRR(IR)=AVBRR(IR)/31
C

```

```

DO 230 I=1,32
SHAT(I,1)=S-L(I)*(BP(I,1)-BP(I,2))/40000
SHAT(I,2)=S-L(I)*(BP(I,2)-BP(I,1))/40000
THAT(I,1)=T-L(I)*(TT(I,1)-TT(I,2))/40000
THAT(I,2)=T-L(I)*(TT(I,2)-TT(I,1))/40000
230 CONTINUE
DO 240 I=1,32
DO 240 J=1,2
BT(I,J)=0.0
DO 235 K=1,20
BT(I,J)=BT(I,J)+(1-IS(I,J,K))*(SSS(I,J,K)-S/T)
235 CONTINUE
BT(I,J)=BT(I,J)*L(I)/40000
240 CONTINUE
DO 250 I=1,32
AYBAR(I,1)=BT(I,2)-BT(I,1)+(SHAT(I,1)/THAT(I,1)-S/T)
AYBAR(I,2)=BT(I,1)-BT(I,2)+(SHAT(I,2)/THAT(I,2)-S/T)
250 CONTINUE
AVJACK(IR)=0.0D0
DO 260 I=1,32
AVJACK(IR)=AVJACK(IR)+(AYBAR(I,1)**2+AYBAR(I,2)**2)/2
260 CONTINUE
C
280 YBAR=0.0D0
DO 300 I=1,32
DO 300 J=1,2
DO 300 K=1,20
YBAR=YBAR+SSS(I,J,K)*L(I)
300 CONTINUE
YBAR=YBAR/40000
MEAN=MEAN+YBAR
YBAR2=YBAR2+YBAR**2
1000 CONTINUE
C
C **** END OF SIMULATION LOOP ****
C
C *** CALCULATE THE FINAL ESTIMATORS ***
C
MEAN=MEAN/N
MNB=0.0
MAB=0.0
MAJ=0.0
DO 1100 IA=1,2000
MNB=MNB+NVBRR(IA)
MAB=MAB+AVBRR(IA)
1100 MAJ=MAJ+AVJACK(IA)
CONTINUE
MNB=MNB/2000
MAB=MAB/2000
MAJ=MAJ/2000

```

```

VAR=(YBAR2-N*MEAN**2)/(N-1)
VNB=0.0
VAB=0.0
VAJ=0.0
DO 1200 IB=1,2000
VNB=VNB+NVBRR(IB)**2
VAB=VAB+AVBRR(IB)**2
VAJ=VAJ+AVJACK(IB)**2
1200 CONTINUE
VNB=(VNB-2000*MNB**2)/1999
VAB=(VAB-2000*MAB**2)/1999
VAJ=(VAJ-2000*MAJ**2)/1999
RB1=(MNB-VAR)/VAR
RB2=(MAB-VAR)/VAR
RB3=(MAJ-VAR)/VAR
VNB=VNB+(MNB-VAR)**2
VAB=VAB+(MAB-VAR)**2
VAJ=VAJ+(MAJ-VAR)**2
C
C **** OUTPUT FINAL RESULT ****
C
WRITE(10,*)
WRITE(10,1400) 'CORRELATION COEFFICIENT:', RHO
WRITE(10,1500) 'RESPONSE RATE:', RESP
WRITE(10,1601) 'POPULATION MEAN:', POPM
WRITE(10,1601) 'THE MEAN ESTIMATOR:', MEAN
WRITE(10,1601) 'THE TRUE VARIANCE:', VAR
WRITE(10,1600) 'NAIVE BRR VAR EST.:', MNB, RB1
WRITE(10,1600) 'ADJUSTED BRR VAR EST.:', MAB, RB2
WRITE(10,1600) 'ADJUSTED JACKKNIFE VAR EST.:', MAJ, RB3
WRITE(10,*)
WRITE(10,*) 'SIMULATION MSE OF VARIANCE ESTIMATORS'
WRITE(10,1800) 'NAIVE BRR:', VNB
WRITE(10,1800) 'ADJUSTED BRR:', VAB
WRITE(10,1800) 'ADJUSTED JACKKNIFE:', VAJ
WRITE(10,*)
1400 FORMAT(A25, F5.1)
1500 FORMAT(A15, I4)
1600 FORMAT(A30, F12.6, ' RELATIVE BIAS: ', F7.4)
1601 FORMAT(A30, F12.6)
1800 FORMAT(A20, F12.6)
C
2000 PRINT*, 'ROUND FINISHED', KN
CONTINUE
STOP
END

```

Bibliography

- [1] Bickel, P.J. and Freedman, D.A. (1984). Asymptotical normality and the bootstrap in stratified sampling. *The Annals of Statistics* , Vol.12, 2470-2482.
- [2] Burns, R.M. (1990). Multiple and replicate item imputation in a complex sample survey. *Proceedings of the Sixth Annual Research Conference*. pp 655-665. Washington DC: U.S. Bureau of the Censes.
- [3] Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Seventh Annual Research Conference* . pp 429-440. Washington, DC: U.S. Bureau of the Census.
- [4] Gupta, V.K. and Nigam, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selection per-stratum. *Biometrika*, Vol.74, 735-742.
- [5] Gurbey, M. and Jewett, R.S. (1975). Constructing orthogonal replications for variance estimation. *Journal of American Statistical Association*, Vol. 70, 819-821.
- [6] Hansen, M.H. and Tepping, B.J. (1985). Estimation of variance in NAEP. Unpublished manuscript.

- [7] Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Annals of the mathematical Statistics*, Vol. 42, 1977-1991.
- [8] Kalton, G. (1981). *Compensating for Missing Data*. ISR research report series. Ann Arbor: Survey Research Center, University of Michigan.
- [9] Kish, L. and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Russian Statistical Association*. Vol. 36, 1-37.
- [10] Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *the Annals of Statistics* Vol. 9, 1010-1019.
- [11] McCarthy, P.J. (1969). Pseudo-replication: half samples, *Reviews of the International Statistical Institute*, vol. 37, 239-264.
- [12] Platek, R. and Gray, G.B. (1983). Imputation methodology: total survey error. *Incomplete Data in Sample Surveys*, 2nd edition. W.G.Madow, I. Olkin and D.B. Rubin, pp. 142-184. New York: Academic Press.
- [13] Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, Vol. 79, 811-822.
- [14] Rao, J.N.K. and Wu, C.F.J. (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, Vol. 80, 620-630.

- [15] Rubin, D.B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp20-34
- [16] Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, Vol. 81, 366-374
- [17] Sedransk, J. (1985). The objective and practice of imputation. *Proceedings of the First Annual Research Conference*, pp. 445-452. Washington DC: Bureau of the U.S. Census
- [18] Sen, P.K. (1970). On some convergence properties of one-sample rank order statistics. *Annals of the Mathematical Statistics*, Vol. 41, 2140-2143.
- [19] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.
- [20] Shao, J. (1993). Resampling methods in sample surveys. Technical Report. University of Ottawa.
- [21] Sitter, R.R. (1993). Balanced repeated replications based on orthogonal multiarrays. *Biometrika*, Vol.80, 135-154.
- [22] Wu, C.J.F. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, Vol.78, 181-188.