SECOND ORDER ACCURATE VARIANCE ESTIMATION IN POSTSTRATIFIED TWO-STAGE SAMPLING

A Dissertation

by

KYONG RYUN KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2006

Major Subject: Statistics

SECOND ORDER ACCURATE VARIANCE ESTIMATION IN POSTSTRATIFIED TWO-STAGE SAMPLING

A Dissertation

by

KYONG RYUN KIM

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Suojin Wang
Committee Members,	Bani Mallick
	Johan Lim
	Lihong Wang
Head of Department,	Simon J. Sheather

May 2006

Major Subject: Statistics

ABSTRACT

Second Order Accurate Variance Estimation in Poststratified Two-stage Sampling.

(May 2006)

Kyong Ryun Kim, B.S., Sungkyukwan University, Korea;

B.S., Michigan State University;

M.S., Michigan State University

Chair of Advisory Committee: Dr. Suojin Wang

We proposed new variance estimators for the poststratified estimator of the population total in two-stage sampling. The linearization or Taylor series variance estimator and the jackknife linearization variance estimator are popular for the poststratified estimator. The jackknife linearization variance estimator utilizes the ratio, $\hat{R}_c$, which balances the weights for the poststrata while the linearization or Taylor series estimator does not. The jackknife linearization variance estimator is equivalent to Rao's (1985) adjusted variance estimator. Our proposed estimator makes use of the ratio, $\hat{R}_c$, in a different shape which is naturally derived from the process of expanding to the second-order Taylor series linearization, while the standard linearization variance estimator is only expanded to the first-order. We investigated the properties and performance of the linearization variance estimator, the jackknife linearization estimator, the proposed variance estimator and its modified version analytically and through simulation study. The simulation study was carried out on both artificially generated data and real data. The result showed that the second order accurate variance estimator and its modified version could be very good candidates for the variance estimation of poststratified estimator of population total.

To my wife, my son and parents

ACKNOWLEDGEMENTS

I would like to express special thanks to my wife, Sungeun. Without her guidance and support this dissertation would not have existed. Her kindness and encouragement made the entire journey of this research very smooth and delightful. I am also grateful for Dr. Suojin Wang, Bani Mallick, Johan Lim, and Lihong Wang as committee members. Their critical reading and helpful comments have greatly improved the dissertation.

I thank my parents and my brother, Kyunghoon. I also thank my parents-in-law, sister-in-law, Hyeun, and brother-in-law, Ikro. I also thank my friends in College Station, Sukho Lee and Hyejin Shin, and Deukwoo Kwon, in Washington D.C., who helped and encouraged me in many ways.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Poststratified estimation of population total is one of popular methods in point estimation used in survey sampling. Poststratification can improve accuracy of estimate by using demographic information of population level that are already known. However, poststratum identifiers of indivisual units are not usually available at the design stage. Therefore, the number of sampling units from each poststratum is random. That implies that can be made both unconditionally and conditionally (Yung and Rao 1996). As for the variance estimation, previous research has adopted two principal approaches, linearization methods and resampling methods. A linearization method involves the analytic calculations of linearizing procedure for a new variable. An advantage of the linearization method is that it is applicable to general sampling design, but it requires the derivation of a separate standard error formula and can be tedious, especially for nonlinear statistics. For example, when estimating ratio or regression coefficients, linearization method is very common (see Rao 1988). Resampling procedures, such as the jackknife, balanced repeated replicaton (BRR) and bootstrap, reuse the procedure for computing the point estimator repeatedly, using computing power to reduce the theoretical work. The jackknife variance estimator is one of the most frequently used method in practice. By linearizing the jackknife variance estimator, jackknife linearization variance estimator can be obtained which is identical to Rao's (1985) variance estimator when estimating the variance of the

The format and style follow that of *Journal of the American Statistical Association*.

poststratified estimator. Valliant (1993) studied the standard linearization variance estimator, the balanced repeated replication, and the jackknife linearization variance estimator to determine if they estimate the conditional variance of the poststratified estimator of a finite population total under a super population model. Yung and Rao (1996) studied the standard linearization variance estimator, the jackknife and the jackknife linearization variance estimators for both the poststratified estimator and the regression estimator. Their simulation study suggested that the three variance estimators perform similarly, while an incorrect jackknife procedure which does not recalculate the regression weights each time when a cluster is deleted performs poorly. The jackknife linearization variance estimator has the adjustment factor that plays a role of balancing the weights for poststrata while the standard linearization variance estimator doesn't have such a feature. We propose a second-order accurate variance estimator by extending the linearization step to the next order. We will study the standard linearization variance estimator, the jackknife linearization variance estimator, proposed variance estimator and the adjusted proposed linearization variance estimator for the poststratified point estimator.

CHAPTER II

SAMPLING DESIGN

2.1 Survey Sampling

A survey concerns a finite set of elements called a finite population. The goal of a survey is to provide information about the finite population in question or about subpopulations of special interest, for example, “men” and “women” as two subpopulations of “all persons”. Such populations are called domains of study or just domains. A value of one or more variables of study is associated with each population element. The goal of a survey is to get information about unknown population characteristics or parameters. Parameters are functions of the study variable values. They are unknown, quantitative measures of interest to the investigator, such as total revenue, mean revenue, total yield, number of unemployed, for the entire population or for the specified domains.

In most surveys, access to and observation of the individual population elements are established through a sampling frame that associates the elements of the population with sampling units in the frame. From the population, a sample of elements is selected in the frame. A sample is a probability sample to be realized by a chance mechanism. The sample elements are observed. That is, for each element in the sample, the variables of study are measured and the values recorded. The recorded variable values are used to calculate estimates of the finite population parameters of interest. Estimates of the precision of the estimates are also calculated.

In the sample survey, observation is limited to a subset of the population. A special type of survey where the whole population is observed is called a census or a complete enumeration.

2.2 Probability Sampling

Probability sampling is an approach to sample selection that satisfies certain conditions, which, for the case of selecting elements directly from the population is described:

(1) we can define the set of samples, $S = \{s_1, s_2, ..., s_M\}$, that are possible to obtain with the sampling procedure;

(2) each possible sample s is associated with a known selection probability $p(s)$;

(3) every element in the population has a nonzero probability of selection through the procedure;

(4) one sample is selected by a random mechanism under which each possible s receives exactly the probability $p(s)$.

A sample under these conditions is called a probability sample. The function $p(\cdot)$ defines a probability distribution on $S = \{s_1, s_2, ..., s_M\}$. It is called a sampling design, or just design. The probability referred (3) is called the inclusion probability of the element. Under a probability sampling design, every population element has a strictly positive inclusion probability. This is a strong requirement, but one that plays an important role in the probability sampling approach. Sampling is often carried out in two or more stages. Clusters of elements are selected in an initial stage. This may be followed by one or more subsampling stages. The elements themselves are sampled at the ultimate stage. To have a probability sampling design, those conditions must apply to each stage. The procedure as a whole must give every population element a strictly positive inclusion probability.

The frame or the sampling frame is any material or device used to obtain observational access to the finite population of interest. It must be possible with the aid of the frame to identify and select a sample in a way that respects a given probability sampling design.

2.3 Inclusion Probability

An interesting feature of a finite population of N labeled elements is that the elements can be given different probabilities of inclusion in the sample. The sampling statistician often takes advantage of the identifiability of the elements by deliberately attaching different inclusion probabilities to the various elements. This is one way to obtain more accurate estimates. Suppose that a certain sampling design has been fixed. That is, $p(s)$, the probability of selecting s, has a given mathematical form. The inclusion of a given element k in a sample s is a random event indicated by the random variable I_k, defined as

$$I_k = \begin{cases} 1 & \text{if } k \in S \\ 0 & \text{oterwise .} \end{cases}$$

Note that $I_k = I_k(s)$ is a function of the random variable S. We call I_k the sample membership indicator of element k.

The probability that element k will be included in a sample, denote π_k, is obtained from the given design $p(\cdot)$ as

$$\pi_k = Pr(k \in s) = Pr(I_k = 1) = \sum_{k \in s} p(s).$$

Here, k denotes that the sum is over those samples s that contain the given k. The probability that both of the elements k and l will be included is denoted π_{kl} and is obtained from the given $p(\cdot)$ as follows:

$$\pi_{kl} = Pr(k, l \in s) = Pr(I_k I_l = 1) = \sum_{k,l \in s} p(s).$$

If the study variable y is approximately proportional to a positive and known auxiliary variable x, there are some advantages in selecting the elements with probability proportional to x. By choosing π_k proportional to the known value x_k will lead to approximately constant ratio y_k/π_k. As a result, the variance of the estimator will be small (Sarndal et al., 1992).

2.4 Horvitz-Thompson Estimator

Let's consider the estimator of the population total T,

$$\hat{T}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

This estimator can be expressed with indicator functions I_k:

$$\hat{T}_\pi = \sum_{k \in U} I_k \frac{y_k}{\pi_k}.$$

Because $E(I_k) = \pi_k$ and $\pi_k \geq 0$ for all $k \in U$, it follows that $\hat{T}_\pi$ is an unbiased estimator of $T = \sum_{k \in U} y_k$. The quantity y_k/π_k can be called the "π-expanded y-value for the k-th element" (Sarndal et al., 1992). The given estimator will be referred as the π estimator of the population total. The π expansion has the effect of increasing the importance of the elements in the sample. Because the sample contains fewer elements than population, an expansion is required to reach the level of the whole population. The k-th element, when present in the sample, will, as it were, represent $1/\pi_k$ population elements. The above formula embody extremely important principle, namely, the use of π-expanded sample values to obtain an unbiased estimator of a population total when sampling is done with arbitrary positive inclusion probabilities.

Horvitz and Thompson (1952) used the principle of π expansion to estimate the total $t = \sum_U y_k$, and is often called the Horvitz-Thompson estimator.

A probability sample s is drawn from U, the set of all possible samples from population, by any sampling design which induces the inclusion probabilities $\pi_k =$

$P(k \in s)$. Let $1/\pi_k$ be the sampling weight associated with the k-th unit. Then an unbiased estimator of population total, Y, introduced by Horvitz-Thompson (1952), can be given. This estimator does not depend on the number of times a unit may be selected. Each distinct unit of the sample is utilized only once. Let the probability that both k and l are included in the sample be denoted by π_{kl}. The variance estimate is

$$var(\hat{T}_{HT}) = \sum_{k=1}^{N}(\frac{1-\pi_k}{\pi_k})y_k^2 + \sum_{k=1}^{N}\sum_{l \neq k}(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l})y_k y_l.$$

In most experiments, it is no necessary to actually compute the probability of selecting the entire sample. For each sample unit k, the probability of selecting that particular unit, π_k, is only needed to be calculated. For simple random sampling without replacement, each of n units has the same selection probability, $\pi_k = \frac{n}{N}$.

CHAPTER III

POINT ESTIMATION IN FINITE POPULATION

3.1 Calibration Estimation

It is often desirable to make use of several data sources when producing statistical estimates. First, a more accurate estimate may be achievable from a combination of sources than from any single source. Second, presence of common variables in different data sources may lead to incoherence if estimates from the different sources are produced separately.

Calibration estimation (Deville and Sarndal, 1992) provides a valuable class of techniques for combining data sources. The basic idea is to use estimates from one set of sources, which may be treated as sufficiently accurate to act as 'benchmarks'. Estimates based on data from a further sample source are then adjusted so as to agree with these benchmarks. The process of adjustment is called 'calibration'. The constraints that the estimates of the benchmarks based on this source should agree with the benchmarks are called 'calibration constraints'.

Simple examples of calibration estimation are provided by ratio estimation and poststratification. In the classical case it is assumed that population values are available for an auxiliary variable and that these data are combined with sample data on some survey variable to estimate the mean or total of this variable.

In ratio estimation it is assumed that the population total or mean of continuous auxiliary variable is known. In poststratification it is assumed that population proportions falling into the categories of a discrete auxiliary variable are known.

3.2 Auxiliary Variable

Generally speaking, an auxiliary variable is any variable about which information is available prior to sampling. Ordinarily, we assume that a priori information for an auxiliary variable is complete. The value of the variable, say x, is known for each of N population elements so that the values $x_1, ..., x_N$ are at our disposal prior to sampling. An auxiliary variable assists in the estimation of the study variable. The goal is to obtain an estimator with increased accuracy.

Some sampling frames are equipped from the outset with one or more auxiliary variables, or with information that can be transformed into auxiliary variables through simple numerical manipulations. This is, the frame provides not only the identification characteristics of the units, but attached to each unit is also the values of one or more auxiliary variables. For example, a register of farms may contain information about the area of each farm. A list of district may contain information about the number of people living in each district at the time of the latest population census.

Auxiliary variable values can be transferred to the sampling frame from administrative or other registers by matching these registers to the sampling frame. There are practical problems associated with matching. For instance, the frame and register may date from different periods in time, elements may be coded differently or erroneously, and so on. In these cases, an element in the frame cannot always be unambiguously identified in the register. These are sometimes difficult problems.

We already noted that auxiliary variables can be used at the design stage of a survey to create a sampling design that increases the precision of the π estimator. One approach is the probability proportional-to-size sampling, that is to make the inclusion probabilities $\pi_1, ..., \pi_N$ of the design proportional to known, positive values $x_1, ..., x_N$ of an auxiliary variable. The π estimator will then have a small variance if

x is more or less proportional to y, the study variable.

Another approach is to use auxiliary information to construct strata such that the π estimator for a stratified simple random sampling design,

$$\hat{T}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

obtains a small variance. However, the stratification that is efficient for one study variable may be inefficient for another.

The auxiliary information can also be used at the estimation stage. The auxiliary variable will enter explicitly into the estimator formula, not only through π_k. That is, for a given sampling design, we construct estimators that utilize information from auxiliary variables and bring considerable variance reduction compared to the π estimator.

The basic assumption behind the use of auxiliary variables is that they covary with the study variable and thus carry information about the study variable. Such covariation is used advantageously in the regression estimator.

3.3 Generalized Regression Estimator

Consider a finite population $U = \{1, ..., k, ..., N\}$, from which a probability sample $s(s \subseteq U)$ is drawn with a given sampling design, $p(\cdot)$. That is, $p(s)$ is the probability that s is selected. The inclusion probabilities $\pi_k = Pr(k \in s)$ and $\pi_{kl} = Pr(k, l \in s)$ are assumed to be strictly positive. Let y_k be the values of the variable of interest, y, for the kth population element, with which also associated an auxiliary vector value, $\mathbf{x}_k = (x_{k1}, ..., x_{kj}, ...x_{kJ})'$. The population total of $\mathbf{x}$, $\mathbf{X} = \sum_U \mathbf{x}_k$, is assumed to be accurately known. The incorporation of auxiliary information can be reflected in the creation of new weights, denoted by w_k, $k \in s$. The new estimator is

$$\hat{T}_w = \sum_{k \in s} w_k y_k,$$

where the weights w_k are chosen to minimize $\sum_{k\in s} d_k(w_k, a_k)$,which measures the distance between the w_k and the design weights $a_k = 1/\pi_k$, subject to the following calibration constraints,

$$\sum_{k\in s} w_k \mathbf{x}_k = \mathbf{X}.$$

The approach of calibration involves determining these new weights $\{w_k : k \in s\}$ by making them as close as possible to the original sampling weights $\{a_k : k \in s\}$ according to a specified distance function. Constraints placed on the new weights are such that, when applied to each of the auxiliary variables, the known population total $\mathbf{X}$ is reproduced.

Suppose $\mathbf{x}' = (x_{1k}, x_{2k}, \cdots, x_{pk})$ is a vector of length p containing the values of auxiliary variables for the k-th indivisual, and the auxiliary information available from an external source is summarized by the known vector total $\sum_{k\in U} \mathbf{x}_k = X$.

The choice of the function d_k will lead to different estimators. The choice $d(w_k, a_k) = (w_k - a_k)^2/2a_k$ leads to the generalized regression (GREG) estimator.

By use of lagrange Multiplier with the above constraint, we have the following

$$\phi(\lambda) = \sum_{k\in s} \frac{(w_k - a_k)^2}{2a_k} - \lambda'(\sum_{k\in s} w_k \mathbf{x}_k - \mathbf{X}).$$

Differenciating $\phi(\lambda)$ with respect to w_k, equating the result to zero

$$\frac{\partial}{\partial w_k} = 0$$

gives

$$\sum_{k\in s} \Big(\frac{w_k - a_k}{a_k} - \lambda' \mathbf{x}_k\Big) = 0,$$

$$\begin{aligned} w_k &= a_k + a_k \lambda' \mathbf{x}_k \\ &= a_k + a_k \mathbf{x}_k' \lambda. \end{aligned}$$

Plug w_k into the equation of the constraint and solve for λ as follows:

$$\begin{aligned}
\sum_{k\in s}\mathbf{x}_k(a_k + a_k\mathbf{x}_k'\lambda) - \mathbf{X} &= 0,\\
(\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')\lambda + \sum_{k\in s}\mathbf{x}_k a_k - \mathbf{X} &= 0,\\
\lambda = (\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}(X - \sum_{k\in s}\mathbf{x}_k a_k).
\end{aligned}$$

Then we have

$$\begin{aligned}
w_k &= a_k + a_k\lambda'\mathbf{x}_k\\
&= a_k(1+\lambda'\mathbf{x}_k)\\
&= a_k\big\{1 + (X - \sum_{k\in s}\mathbf{x}_k a_k)'(\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}\mathbf{x}_k'\big\}\\
&= a_k\big\{1 + (X - \hat{X}_a)'(\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}\mathbf{x}_k'\big\}.
\end{aligned}$$

New calibration estimator for the population total, T, which is adjusted by the auxiliary information vector $\mathbf{x}$ is following,

$$\begin{aligned}
\hat{T}_w &= \sum_{k\in s}w_k y_k\\
&= \sum_{k\in s}a_k\big\{1 + (X - \hat{X}_a)'(\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}\mathbf{x}_k'\big\}y_k\\
&= \sum_{k\in s}a_k y_k + (X - \hat{X}_a)'(\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}\sum_{k\in s}a_k\mathbf{x}_k y_k\\
&= \hat{T}_a + (X - \hat{X}_a)'\hat{\beta},
\end{aligned}$$

where $\hat{\beta} = (\sum_{k\in s}a_k\mathbf{x}_k\mathbf{x}_k')^{-1}\sum_{k\in s}a_k\mathbf{x}_k y_k$.

3.4 Raking

Raking has been widely used for many years for benchmarking sample distributions to external distributions. When benchmarking to population distributions from external sources, sometimes only the marginal distributions of the auxiliary variables are available. Raking operates only on the marginal distributions of the auxiliary variables. Raking is an iterative proportional fitting procedure (IPF):

(1) Sample row totals are forced to conform to the population row totals; then the sample adjusted column totals are forced to conform to population column totals.

(2) Then the row totals are adjusted to conform and so on until convergence is reached.

Consider a two dimensional table with observed cell counts, n_{ij}, unknown population cell counts, N_{ij} and estimates of the population cell counts, $\hat{N}_{ij}$. Marginal sums $\sum_j N_{ij} = N_{i\cdot}$(i-th row total) and $\sum_i N_{ij} = N_{\cdot j}$(j-th column total) are known. As pointed out in Little and Rubin (1987), raking applies to the individual cell counts, n_{ij}, to iteratively calculate estimates that satisfy marginal constraints $\hat{N}_{i\cdot} = \sum_j \hat{N}_{ij} = N_{i\cdot}$ and $\hat{N}_{\cdot j} = \sum_i \hat{N}_{ij} = N_{\cdot j}$.

IPF is used to adjust the cells to marginal totals. At the first step of the procedure, estimators are calculated $\hat{N}^(1)_{ij} = \frac{n_{ij}N_{i\cdot}}{n_{i\cdot}}$. This matches the row marginals exactly, but the column marginals are unlikely to agree with the known values. Then next iteration adjusts the individual cells to the column marginals by $\hat{N}^(2)_{ij} = \frac{\hat{N}^1_{ij}N.j}{\hat{N}^1_{\cdot j}}$. Then the row marginals are adjusted by $\hat{N}^(3)_{ij} = \frac{\hat{N}^(2)_{ij}N_{i\cdot}}{\hat{N}^(2)_{i\cdot}}$. Iteration between rows and columns continues until convergence is achieved, where convergence is defined as $\mid \hat{N}_{i\cdot} - N_{i\cdot} \mid \leq \epsilon$ and $\mid \hat{N}_{\cdot j} - N_{\cdot j} \mid \leq \epsilon$ for some small value ϵ. Both iterative proportional fitting and raking are attributed to Deming and Stephan (1940).

The next few tables show an hypothetical example (Micheal A. Greene, Linda E. Smith, Mark S. Levenson, Singne Hiser and Jean C. Mah) for a 2×2 problem with an additional unknown row and unknown column. The example adjusts columns first instead of rows, but the principles are the same.

Before raking, the unknown marginal are distributed to the known marginal in proportional to the value of the known marginal and shown in Table 1. In the first

Table 1: 2×2 problem with additional unknown row and column

	female	male	unknown	pop.marginal
old	65	30	5	100
Young	25	50	25	100
unknown	10	2000	70	2080
pop.marginal	100	2080	100	2280

column of female ($100 + 100 \times 100/2180 = 104.6$) and in the second column of male ($2080 + 100 \times 2080/2180 = 2175.4$). The table without the values of unknown is shown in Table 2. This is now ready for raking.

Table 2: Distribution of unknowns to the known marginals

	female	male	sample marginal	pop.marginal
old	65	30	95	1140
Young	25	50	75	1140
sample marginal	90	80	170	
pop.marginal	104.6	2175.4		2280

Population totals of 104.6 and 2175.4 for the columns are shown in Table 2 and are different from the computed marginals by 14.6 and 2095.4, respectively. The first iteration involves multiplying the entries in the first row by ratio of population to computed marginals as follows

$$\begin{aligned}\hat{N}_{11}^{1} &= \frac{n_{11}N_{1\cdot}}{n_{1\cdot}} = \frac{65 \times 1140}{95} = 780,\\ \hat{N}_{12}^{1} &= \frac{n_{12}N_{1\cdot}}{n_{1\cdot}} = \frac{30 \times 1140}{95} = 360.\end{aligned}$$

Also, in the second row

$$\begin{aligned}\hat{N}_{21}^{1} &= \frac{n_{21}N_{1\cdot}}{n_{2\cdot}} = \frac{25 \times 1140}{75} = 380,\\ \hat{N}_{22}^{1} &= \frac{n_{22}N_{1\cdot}}{n_{2\cdot}} = \frac{50 \times 1140}{75} = 760.\end{aligned}$$

Then we have following results in Table 3 after first adjustment procedure.

Table 3: First step in raking with rows

	female	male	pop.marginal
old	780	360	1140
Young	380	760	1140
adjusted marginal	1160	1120	2080
pop.marginal	104.6	2175.4	2280

While the row marginal have been adjusted to the population totals, the column marginal are now off. The appropriate multipliers for the column marginal are 104.6/1160 and 2175.4/1120, respectively. This results in Table 4.

Table 4: First step in raking with columns

	female	male	adjusted marginal	pop.marginal
old	70.33	699.24	769.57	1140
Young	34.27	1476.16	1510.43	1140
pop.marginal	104.6	2175.4	2280	

Table 5: Second step in raking with rows & columns

	female	male	adjusted marginal	pop.marginal
old	83.79	1048.1	1131.8	1140
Young	20.81	1127.3	1148.1	1140
pop.marginal	104.6	2175.4	2280	

The application of column multipliers perfectly aligns the columns at the expense of the rows. The next iteration multiplier carries in the first row by 1140/769.57 and the second row by 1140/1510.43. In Table 5, it can be verified that one more adjustment to the columns, using multipliers 104.6/130.05 and 2175.4/2149.9 bring the population and calculated marginals within 8.2 in both dimensions. More iterative adjustments will lead the difference to converge into very small value, 'ϵ', which is a given stopping rule before the procedure starts.

CHAPTER IV

STRATIFICATION AND CLUSTERING

4.1 Stratified Sampling

Many populations are, in effect, collections of populations, and a target variable in a survey may follow a different model in each of subpopulations. In a survey of households to estimate average income, for example, the income levels may vary widely among different demographic groups and regions of a country. The sample data from one subgroup may be of limit use in making estimates for other subgroups. In these populations, estimates mat be required both the full population and for some or all of the subpopulations. In either case, it is desirable to take each subpopulations as a stratum and so require a sample in each subpopulation.

The cost of conducting a survey may differ substancially among the strata. An optimum allocation of sample to the strata will consider both the cost and the variability of the target variable in each stratum. Practical problems related to response and measurement may differ considerably among subpopulations. Stratification allows some flexibility in the choice of data-collection procedures that are used for different subpopulations. Telephone data collection may be adequate for some groups while personal interviews may be needed for others, for instance. For operational convenience, the survey organization may also be divided into geographic district with a field office supervising work in each district.

In stratified sampling, the population is divided into nonoverlapping subpopulations called strata. A probability sample is selected in each stratum. The selections in different strata are independent. Stratified sampling is powerful and flexible method that is widely used in practice. In a survey, practical aspects related to response, measurements, and auxiliary information may differ considerably from one subpopulation to another. Nonresponse rates and measurements problems may be more pronounced in some subpopulations than in others. The extent of the auxiliary information may differ greatly. These factors suggest that the choice of sampling design and estimator perhaps should be made difficulty in different subpopulations to increase the efficiency of the estimation. One may thus want to treat each subpopulation as a separate stratum. For administrative reasons, the survey organization may have divided its total territory into several geographic district with a field office in each district. Here, it is natural to let each district be a stratum. An additional reason in favor of stratified sampling is that most of the potential gain in efficiency of probability proportional-to-size sampling can be captured through stratified selection with simple random sampling within well-constructed strata. Stratified sampling in several respects simpler than and consequently preferred to proportional-size sampling. Let us first introduce some notation and definitions. By a $stratification$ of a finite population $U = \{1, ..., k, ..., N\}$ we mean a partitioning of U into H subpopulations, called strata and denoted $U_1, ..., U_h, ..., U_H$, where $U_h = \{k : k \text{ belongs to stratum } h\}$. By stratified sampling we mean that a probability sample s is selected from U_h according to a design $p_h(\cdot)$ $(h = 1, ..., H)$ and that the selection in one stratum is independent of selections in all other strata.

The resulting total sample, denoted by s as usual, will thus be composed as

$$s = s_1 \cup s_2 \cup \cdots \cup s_H$$

and, because of the independence feature,

$$p(s) = p_1(s_1)p_2(s_2)\cdots p_H(s_H).$$

The number of elements in stratum h, called the size of stratum h, is denoted

N_h, which is assumed to be known. Since the strata form a partition of U we have

$$N = \sum_{h=1}^{H} N_h.$$

Furthermore, the population total can be decomposed as

$$T = \sum_{U} y_k = \sum_{h=1}^{H} T_h = \sum_{h=1}^{H} N_h \bar{y}_h,$$

where T_h is the stratum total, and $\bar{y}_h$ the stratum mean. Finally, let $W_h = N_h/N$ denote the relative size of the stratum U_h. Then the population mean has the decomposition

$$\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h.$$

4.2 Cluster Sampling

Many naturally occuring populations exhibit clustering in which units that are near to each other (geographically or in some other respect) have similar characteristics. Households in the same neighborhood tend to have similar incomes, educational levels of the heads of household, and amounts of expenditures on food and clothing. Business establlishments in the same industry and geographic area will pay similar wages to a guven occupation because of competition.

In cluster population, the methods of data collection may also differ from the methods used in other populations. In household survey, for example, a complete list of households to use for sampling is usually not available, especially if the population is large. In the united states, for instance, there were nearly 100 million households in 1995. The households of interest may be geographically dispersed; field work can be more economically done when sample units are clustered together to limit travel costs. A practical and widely used technique is to select the sample in stages, using at each stage, sampling units for which a complete list is available. In the household example, geographic areas may be selected at first stage. At the second stage, each first stage sample unit may be further subdivided and a sample of the subdivisions selected. A list of the households in each sample subdivision is then compiled and data collected from each. In a business population, establishments may be selected at the first stage, a list of occupations compiled in each sample establishment, and a sample of occupations then drawn from each list. Although occupations are the units ultimately sampled, a complete list of occupations for each establishment inthe universe is unlikely to be available, whereas a list of establishments often is. Selecting occupations in two stages can also allow better control over survey costs. Travel and extracting data from personnel records may be referred to the number of sample occupations. Two-stage sampling can also allow fine-tuning of survey budget.

A probability sample of clusters is selected, and every population element in the selected cluster is surveyed. In the single-stage cluster sampling, the finite population $U = \{1, ..., k, ..., N\}$ is partitioned into N_I clusters, and denoted $U_1, ..., U_{N_I}$. The set of clusters is symbolically represented as

$$U_I = \{1, ..., i, ..., N_I\}.$$

The number of population elements in the ith cluster U_i is denoted N_i. The partitioning of U is expressed by the equations

$$U = \bigcup_{i \in \mathrm{U_i}} \mathrm{U_i} \quad \text{and} \quad \mathrm{N} = \sum_{\mathrm{i} \in \mathrm{U_i}} \mathrm{N_i}.$$

Cluster sampling is now defined in the following way:

(1) A probability sample s_I of clusters is drawn from U_I according to the design $p_I(\cdot)$. The size of s_I is denoted by n_I, for a fixed size design, or by n_{s_I} for a variable size design.

(2) Every population element in the selected clusters is observed. Here, $p_I(\cdot)$ may be any of conventional designs, that is, simple random sampling without replacement, systematic sampling, stratified sampling and so on.

The strategy of simple random cluster sampling is likely to be inefficient in many situations, especially if the clusters are heterogeneous or of unequal sizes. However, from a cost efficiency point of view, the strategy may have advantages, since it is often much cheaper to survey clusters of elements than to survey the geographically scattered sample that may arise from a simple random selection of elements.

However, the efficiency of cluster sampling can be improved when auxiliary information is available. The choice of strategy then depends on the information available. A simple case is when an approximate measure of size u_i is available for each cluster $i = 1, ..., N_I$. If u_i is roughly proportional to t_i which is the ith cluster total, we can reduce the variance of the π- estimator (or Horvits-Thompson estimator which will be discussed later) by using probability proportional-to-size cluster sampling with inclusion probabilities $\pi_{Ii} \propto u_i$. An alternative is to use stratified cluster sampling with strata of clusters formed so that the variation of u_i is small in each stratum.

4.3 Two-stage Sampling

Cluster sampling is also called single-stage cluster sampling. By contrast, in two-stage cluster sampling, the sample of elements is obtained as result of two stages of sampling.

(1) The population elements are first grouped into disjoint subpopulations, called primary sampling units (PSUs). A probability sample of PSUs is drawn (first-stage sampling).

(2) For each PSU in the first-stage sampling, the type of sampling unit to be used in the second-stage sampling is decided upon. These second-stage sampling units may

be elements or clusters of elements. A probability sample of second-stage sampling units from each PSU in the first stage sample. When the second-stage sample units are clusters, every element in the selected second-stage sampling units is surveyed.

It is noted that variance of simple random sampling is smaller than simple random cluster sampling (Sarndal et al., 1992). This is explained by the tendency for elements in the same cluster to resemble each other, which implies that the homogeneous measure is positive, and by the variation in the cluster sizes. The variance of the π estimator under simple random cluster sampling can always be reduced by selecting more clusters. However, the increased cost of taking a bigger sample may be unacceptable under the variable budget.

To control the cost and at the same time increase the number of selected clusters, we may subsample within the selected clusters, instead of surveying all elements in the selected clusters. Then we must estimate the cluster total T_{hi} from the subsamples. If the variation within the clusters is small, the estimates $\hat{T}_{hi}$ have a smaller variance, even for rather modest subsample sizes. It often pays to use two-stage sampling instead of cluster sampling.

Notation and estimation in two-stage sampling are slightly more complex than in cluster sampling. There are two sources of variation. The first-stage sampling variation arises from the selection of primary sampling units. The second-stage sampling variation arises from the subsampling of secondary sampling units within selected PSUs.

Multistage sampling consists of three or more stages of sampling. There is a hierarchy of sampling units: primary sampling units, secondary sampling units within the PSUs, tertiary sampling units within secondary sampling units and so on. The sampling units in the last-stage sampling are called ultimate sampling units and those in the text to the last stage sampling are called penultimate sampling units.

A variety of sampling designs are available for surveys in which direct element sampling is impossible or impractical. These range from cluster sampling to highly complex multistage sampling designs using unequal probability sampling at the various stages of selection. In cluster sampling, the finite population is grouped into subpopulations called clusters. Stratification and clustering both divide the population into mutually exclusive groups. Whether those groups are strata or clusters depends on how the sample is selected. If at least one sample unit is drawn from each group, they are strata. Otherwise, the groups are clusters.

CHAPTER V

POSTSTRATIFIED POINT ESTIMATOR

5.1 Population Total

Before we start to talk about point estimators, we need to define some definitions which will be used throughout later chapters. The following indexes and notations will be used in the remainder of the dissertation.

$h =$ index for strata $(h = 1, \ldots, L)$,

$i =$ index for cluster $(i = 1, \ldots, N_h)$,

$k =$ index for units $(k = 1, \ldots, M_{hi})$,

$c =$ index for poststrata $(c = 1, \ldots, C)$,

$y_{hikc} =$ a variable of interest,

$N = \sum_h N_h$, number of clusters in population,

$M = \sum_h M_h$, population size.

We consider a clustered finite population with L strata. Let N_h be the number of primary sampling units(PSU), or clusters in h-th stratum and w_{hik} be the survey weight associated with y_{hik}, k-th unit within i-th cluster in h-th stratum. (An unbiased estimator of the cluster total, T_{hi} $(i = 1, ..., n_h)$ by subsampling in a sampled cluster is assumed.) A stratum h contains N_h clusters. Cluster hi contains M_{hi} units with $M_h = \sum_{i=1}^{N_h} M_{hi}$ and $M = \sum_{h=1}^{L} M_h$. In the same manner, the number of sampled clusters in h-th stratum and sampled units in hi-th cluster are n_h and m_{hi} respectively. Assume that units in different clusters and strata are iid and srs(simple random sampling) without repllacement is also assumed. But units in different post-strata may not be *iid*.

However, at variance estimation stage, *srs* with replacement can be assumed only for computational convenience but all the assumptions maybe considered to be still valid if first-stage sampling fraction is small enough. The finite population total is

$$\begin{aligned} T &= \sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}} y_{hik} \\ &= \sum_{h=1}^{L} T_h, \end{aligned}$$

where $T_h = \sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}} y_{hik}$ is the h-th stratum total.

An customary estimator of population total T is expressed as

$$\begin{aligned} \hat{T} &= \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}y_{hik} \\ &= \sum_{h=1}^{L} \bar{r}_h \Big(= \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} r_{hi} \Big) \\ &= \sum_{h=1}^{L} \hat{T}_h, \end{aligned}$$

where $\hat{T}_h = \sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}} w_{hik}y_{hik}$ is estimate of the stratum total and $r_{hi} = n_h \sum_k w_{hik}y_{hik}$ (one of the stratum total estimate among n_h ones based on only i-th sampled cluster in h stratum). Note that r_{hi} are *iid* with mean $\hat{T}_h$, h stratum total, and same variance in each stratum under with replacement sampling scheme.

5.2 Poststratified Estimator

Auxiliary variables used in the regression estimator can be both quantitative variables and qualitative variables. Actually, the poststratified estimator is a special case of the regression estimator when the auxiliary variables are the indicator variables for the poststrata. Define the population total as

$$T = \sum_{hik \in s} y_{hik}.$$

We assume that the population is divided into C poststrata with size, M_c. Then the number of units in c-th post-stratum is

$$M_c = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} \delta_{hikc},$$

where (assume design weight $w_{hik} = w_{hi}$ for all k) δ_{hikc} is the indicate function which identifies if each y_{hik} is in that poststratum or not. That is, it is defined as

$$\delta_{hikc} = \begin{cases} 1 & \text{if } y_{hik} \in \text{ c-th poststratum} \\ 0 & \text{if not .} \end{cases}$$

In figure 1, the structure of poststratification is graphically expressed.

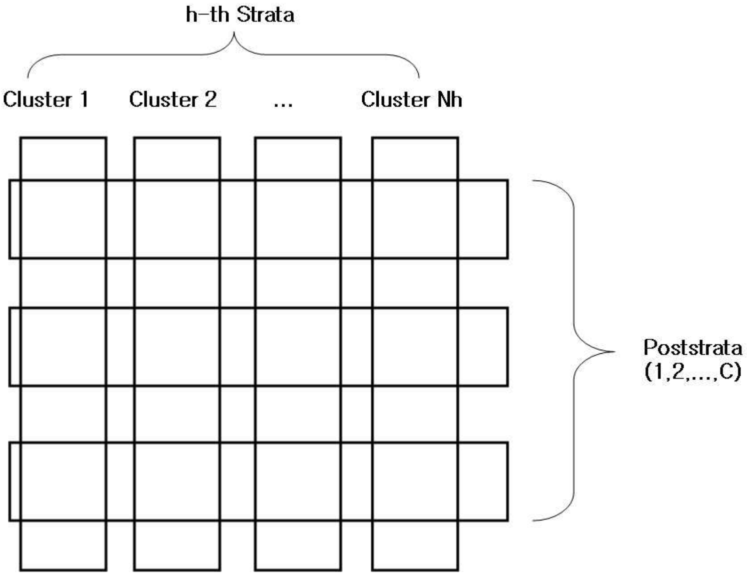


Figure 1: Poststratification

Population total can be described as sum of poststrata total

$$\begin{aligned}
\hat{T} &= \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}\hat{T}_{hi} \\
&= \sum_{h}\sum_{i} w_{hi}M_{hi}\bar{y}_{hi} \\
&= \sum_{h}\sum_{i} w_{hi}M_{hi}\frac{1}{m_{hi}}\sum_{c}\sum_{k=1}^{m_{hi}} y_{hik}\delta_{hikc} \\
&= \sum_{h}\sum_{i} \frac{w_{hi}M_{hi}m_{hic}}{m_{hi}}\sum_{c}\bar{y}_{hic} \\
&= \sum_{h}\sum_{i}\sum_{c} S_{hic}\bar{y}_{hic} \\
&= \sum_{c}\hat{T}_{c},
\end{aligned}$$

where $\hat{T}_c = \sum_h \sum_i S_{hic}\bar{y}_{hic}$. And if replace y_{hik} by δ_{hik} which is indicator variable, then we obtain estimator of M_c.

$$\hat{M}_c = \sum_{h}\sum_{i} S_{hic}\Big(= \sum_{h}\sum_{i} S_{hic}\delta_{hikc}\Big),$$

where $S_{hic} = w_{hi}M_{hi}m_{hic}/m_{hi}$ and m_{hic} is the number of units in poststratum c among m_{hi} units in (hi)-cluster M_c is assumed to be known and seems to be used for better estimation. Then the poststratified estimator of the total is suggested as follows

$$\hat{T}_{pst} = \sum_{c}\hat{R}_c\hat{T}_c,$$

where $\hat{R}_c = M_c/\hat{M}_c$.

Adjustment factor $\hat{R}_c$ plays a very important role as balancing the weights of each poststrata estimate cause when too many elements are selected from a poststratum, $\hat{R}_c$ gets smaller then gives a smaller weight to poststratum estimate, $\hat{T}_c$ and too small size sample from the poststratum adjusts $\hat{R}_c$ to be bigger for more weight. So poststratified estimator is calculated based on the combination of both sample and population level information. GREG estimator can be expressed as following

$$\begin{aligned}\hat{T}_w &= \sum_{k\in s} a_k y_k + (X - \hat{X}_a)'(\sum_{k\in s} a_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{k\in s} a_k \mathbf{x}_k y_k \\ &= \hat{T}_a + (X - \hat{X}_a)'\hat{\beta},\end{aligned}$$

where $\hat{X}_a = \sum_{k\in s} \mathbf{x}_k a_k$, $X = \sum_{k\in U} \mathbf{x}_k$.

If we assume the auxiliary vector, $\mathbf{x}_k$, to be $(\overbrace{0, ..., 0}^{c-1}, 1, 0, ..., 0)'$ when y_k is in c-th poststratum, $\mathbf{x}_k$ is the indicator variable and $X = \sum_{k\in U} \mathbf{x}_k$ is the vector of known population total of poststrata $= (M_1, M_2, ..., M_c)'$. Let the weight a_k to be $1/\pi_k$, then $\hat{X}_a$ becomes $(\hat{M}_1, \hat{M}_2, ..., \hat{M}_c)'$, where $\hat{M}_c = \sum_{k\in s_c} \frac{1}{\pi_k}\mathbf{x}_k$ whic is the Horvitz-Thompson estimate of poststrata size, $(M_1, M_2, ..., M_c)'$ and $\hat{\beta} = (\hat{T}_1/\hat{M}_1, \hat{T}_2/\hat{M}_2, ..., \hat{T}_c/\hat{M}_c)'$. Then we are ready to show that poststratified estimator is the special case of the GREG estimator (Yung and Rao 1996). The following justifies the poststratified total estimator is the special case of the generalized regression estimator:

$$\begin{aligned}\hat{T}_w &= \hat{T}_\pi + (X - \hat{X}_\pi)'\hat{\beta} \\ &= \hat{T}_\pi + \left(\begin{pmatrix} M_1 \\ \vdots \\ M_C \end{pmatrix} - \begin{pmatrix} \hat{M}_1 \\ \vdots \\ \hat{M}_c \end{pmatrix}\right)' \begin{pmatrix} \hat{T}_1/\hat{M}_1 \\ \vdots \\ \hat{T}_c/\hat{M}_c \end{pmatrix} \\ &= \hat{T}_\pi + \frac{M_1 - \hat{M}_1}{\hat{M}_1}\hat{T}_1 + \cdots + \frac{M_c - \hat{M}_c}{\hat{M}_c}\hat{T}_c \\ &= \hat{T}_\pi - \sum_c \hat{T}_c + \frac{M_1}{\hat{M}_1}\hat{T}_1 \cdot\cdot + \frac{M_c}{\hat{M}_c}\hat{T}_c \\ &= \sum_c \frac{M_c}{\hat{M}_c}\hat{T}_c \\ &= \hat{T}_{pst}.\end{aligned}$$

CHAPTER VI

VARIANCE ESTIMATION

6.1 Variance Estimator in Estimating Population Total

Then variance of $\hat{T}$ is

$$\begin{aligned} var(\hat{T}) &= var(\sum_{h=1}^{L} \bar{r}_h) \\ &= \sum_{h=1}^{L} \frac{1}{n_h} var(r_{hi}) \\ &\approx \sum_{h=1}^{L} \frac{1}{n_h} s^2_{r_{hi}} \\ &= \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2. \end{aligned}$$

The variance estimator can be expressed in another way as follows:

$$\begin{aligned} v\hat{a}r(\hat{T}) &= \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (r_{hi} - \bar{r}_h)^2 \\ &= \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (\sum_{k=1}^{m_{hi}} n_h w_{hik} y_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} n_h w_{hik} y_{hik})^2 \\ &= \sum_{h=1}^{L} \frac{n_h}{n_{h-1}} \sum_{i=1}^{n_h} (\sum_{k=1}^{m_{hi}} w_{hik} y_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} w_{hik} y_{hik})^2 \\ &= \sum_{h=1}^{L} \frac{n_h}{n_{h-1}} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2, \end{aligned}$$

where $z_{hi} = \sum_k w_{hik} y_{hik}$.

6.2 The Jackknife Method

An subsample replication technique, called the jackknife, has been suggested as a broadly useful method of variance estimation. The jackknife derives estimates of the parameter of interest from each of several subsamples of the parent sample, and then estimates the variance of the parent sample estimate from the variability between the subsamples estimates.

The jackknife is less dependent on model assumptions and does not require the formula which is usually needed by the traditional way. However, it needs repeatedly calculating the statistic n times, which was practically impossible in the old days. The latest computing technique has made it possible for us to use the jackknife method. The jackknife has become a popular and useful tool in statistical way. Many agencies have computer software to implement the computation of the jackknife method.

The jackknife method was originally introduced to estimate the bias of an estimator by Quenouille(1949). It can be calculated by deleting one datum value each time from n sampled values and reproducing the estimator using $n-1$ remaining sample data. Let T_n be the estimator of unknown parameter θ based on n sample data such as $T_n = T_n(x_1, x_2, ..., x_{n-1}, x_n)$. And the bias of T_n is $bias(T_n) = E(T_n) - \theta$. Here we need to define one more variable $T_{n-1,i} = T_{n-1}(x_1, x_2, .., x_{i-1}, x_{i+1}, .., x_n)$ which is based on $n-1$ observations. Now we have Quenouille's jackknife bias estimator as

$$b_j = (n-1)(\bar{T_n} - T_n),$$

where $\bar{T_n} = \frac{1}{n}\sum_{i=1}^{n} T_{n-1,i}$. Also the bias reduced jackknife estimator of θ is

$$T_{jack} = T_n - b_j = nT_n - (n-1)\bar{T_n}.$$

The jackknife estimators b_j and T_{jack} can be justified as

$$bias(T_{jack}) = bias(T_n) - E(b_j) = -\frac{b}{n(n-1)} + \mathrm{O}(\frac{1}{n^2}).$$

The bias of T_{jack} is of order $\frac{1}{n^2}$. The jackknife method produces a bias reduced estimator by removing the first order term in bias(T_n). Furthermore, it can lead to the following (Shao, 1995)

$$\begin{aligned}
T_{jack} &= nT_n - (n-1)\bar{T}_n \\
&= nT_n - (n-1)\frac{1}{n}\sum_{i=1}^{n} T_{n-1,i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\{nT_n - (n-1)T_{n-1,i}\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_{n-1,i}.
\end{aligned}$$

Tukey(1958) established that the jackknife also can be used in variance estimation and, for finite population, the jackknife technique was first introduced by Durbin (1959). Tukey suggested two conjectures to justify the jackknife variance estimation:

- $\tilde{T}_{n,i}$,$i = 1, ..., n$ are *iid*

- var$(\tilde{T}_{n,i}) \approx$ var$(\sqrt{n}T_n)$

If these two conjectures are satisfied, the $var(T_n) \approx \frac{1}{n}var(\tilde{T}_{n,i})$. Then the jackknife variance estimator is

$$\begin{aligned}
v_{jack} &= \frac{1}{n}v\hat{a}r(\tilde{T}_{n,i}) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{T}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\tilde{T}_{n,j})^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\{nT_n - (n-1)T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}(nT_n - (n-1)T_{n-1,j})\}^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\{-(n-1)T_{n-1,i} + \frac{1}{n}\sum_{j=1}^{n}(n-1)T_{n-1,j}\}^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}(T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}T_{n-1,j})^2.
\end{aligned}$$

According to formula of $var(\hat{T})$, variance of the total estmator, $\hat{T}$, equals to sum of variances of h strata, $\sum_{h=1}^{L} var(\bar{r}_h)$, which can be produced based on the assumption that samplings between strata are independent. So jackknife method can be applied to estimate each variance of strata then jackknife variance estimate is produced by summing them up. For stratum h, jackknife variance estimator is

$$v_{jack}(\hat{T}_h) = \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{T}_{(hi)} - \hat{T}_h)^2.$$

Then we have the jackknife variance estimator of population total estimate

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

where $\hat{T}_{(hi)} = T_{n-1,i}$ which is calculated based on $n-1$ remaining sample data after deleting i-th cluster in h-th stratum. Furthermore, we can show that the jackknife estimator above is equivalent to the customary variance estimator by the following justification:

$$\begin{aligned}
v_{jack}(\hat{T}_h) &= \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{T}_{(hi)} - \hat{T}_h)^2 \\
&= \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \{ \frac{1}{n_h - 1} (z_{h,1} + z_{h,2} + \cdots + z_{h,n_h-1} + z_{h,n_h} - n_h z_{h,i}) \}^2 \\
&= \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (n_h z_{hi} - \sum_{i=1}^{n_h} z_{hi})^2 \\
&= \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (n_h \sum_{k} w_{hik} y_{hik} - \sum_{i=1}^{n_h} \sum_{k} w_{hik} y_{hik})^2 \\
&= \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\sum_{k} w_{hik} y_{hik} - \frac{1}{n_h} \sum_{i=1}^{n_h} \sum_{k} w_{hik} y_{hik})^2 \\
&= \frac{n_h}{n_{h-1}} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2,
\end{aligned}$$

where $z_{hi} = \sum_k w_{hik} y_{hik}$.

We note that in the linear case such as the population total, the customary variance estimator is equal to the jackknife estimator.

6.3 Linearization Variance Estimator

1st order Taylor series expansion for $\hat{R}_c\hat{T}_c$ at (M_c, T_c) is

$$\begin{aligned}\hat{R}_c\hat{T}_c = M_c\frac{\hat{T}_c}{\hat{M}_c} &\approx M_c\Big\{\frac{T_c}{M_c}+\frac{1}{M_c}(\hat{T}_c-T_c)-\frac{T_c}{M_c^2}(\hat{M}_c-M_c)\Big\}\\ &= T_c+\hat{T}_c-\frac{\hat{M}_c}{M_c}T_c.\end{aligned}$$

Then we have the following

$$\begin{aligned}\hat{T}_{pst}-T &= \sum_c(\hat{R}_c\hat{T}_c-T_c)\\ &= \sum_h\sum_i\frac{1}{m_{hi}}\sum_k\sum_c w_{hi}M_{hi}\delta_{hikc}(y_{hik}-\frac{\hat{T}_c}{\hat{M}_c})\\ &= \sum_h\sum_i g_{hi}\Big(=\sum_h\frac{1}{n_h}\sum_i n_hg_{hi}\Big)\\ &= \sum_h\bar{g}_h^*,\end{aligned}$$

where

$$\begin{aligned}g_{hi} &= \frac{1}{m_{hi}}\sum_k\sum_c w_{hi}M_{hi}\delta_{hikc}(y_{hik}-\frac{\hat{T}_c}{\hat{M}_c})\\ &= \sum_c S_{hic}(\bar{y}_{hic}-\hat{\mu}_c)\\ &= \sum_c S_{hic}g_{hic}\end{aligned}$$

and $d_{hi}^*(=n_hd_{hi})$ are iid, $i=1,...,n_h$. Then we can obtain the variance estimator of $\hat{T}_{pst}$,

$$\begin{aligned}v_L(\hat{T}_{pst}) &= \sum_h\frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}(g_{hi}^*-\bar{g}_h^*)^2\\ &= \sum_h\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(g_{hi}-\bar{g}_h)^2,\end{aligned}$$

provided that $v_L(\hat{T}_{pst}-T)\approx v_L(\hat{T}_{pst})$. However, $v_L(\hat{T}_{pst})$ actually estimates $v(\hat{T})$, but not $v(\hat{T}_{pst})$ (Valliant, 1993). Rao (1985) suggested another estimator which is adjusted by $\hat{R}_c = M_c/\hat{M}_c$

$$v_L^*(\hat{T}_{pst}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \Big\{ \sum_c \hat{R}_c (g_{hic} - \bar{g}_{hc}) \Big\}^2,$$

where $g_{hic} = \frac{1}{m_{hi}} \sum_{k \in s_{hi}} w_{hi} M_{hi} \delta_{hikc} (y_{hik} - \frac{\hat{T}_c}{\hat{M}_c}) = S_{hic}(\bar{y}_{hic} - \hat{\mu}_c)$.

6.4 Jackknife Linearization Variance Estimator

The following is the jackknife variance estimator for the poststratified total estimator:

$$\begin{aligned} v_J(\hat{T}_{pst}) &= \sum_h \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \Big(\hat{T}_{pst(hi)} - \hat{T}_{pst} \Big)^2 \\ &= \sum_h \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \Big(\sum_c \hat{R}_{c(hi)} \hat{T}_{c(hi)} - \sum_c \hat{R}_c \hat{T}_c \Big)^2 \\ &= \sum_h \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \Big\{ \sum_c (\hat{R}_{c(hi)} \hat{T}_{c(hi)} - \hat{R}_c \hat{T}_c) \Big\}^2, \end{aligned}$$

where

$$\hat{T}_{pst(hi)} = \sum_c \hat{R}_{c(hi)} \hat{T}_{c(hi)} = \sum_c \frac{M_c}{\hat{M}_{c(hi)}} \hat{T}_{c(hi)}.$$

Note that $\hat{M}_{c(hi)}$ and $\hat{T}_{c(hi)}$ are estimated after deleting (hi)-cluster and (adjusted) linearization variance estimator is

$$v_L^*(\hat{T}_{pst}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \Big\{ \sum_c \hat{R}_c (g_{hic} - \bar{g}_{hc}) \Big\}^2.$$

According to Valliant (1993), the standard Taylor expansion of $\hat{R}_{c(hi)} \hat{T}_{c(hi)}$ at ($\hat{M}_c$,$\hat{T}_c$) is

$$\begin{aligned} \hat{R}_{c(hi)} \hat{T}_{c(hi)} &= \frac{M_c}{\hat{M}_{c(hi)}} \hat{T}_{c(hi)} \\ &\approx \frac{M_c}{\hat{M}_c} \hat{T}_c + \frac{M_c}{\hat{M}_c} (\hat{T}_{c(hi)} - \hat{T}_c) - \frac{M_c}{\hat{M}_c^{\,2}} \hat{T}_c (\hat{M}_{c(hi)} - \hat{M}_c) \\ &= \hat{R}_c \hat{T}_c + \hat{R}_c (\hat{T}_{c(hi)} - \hat{T}_c) - \hat{R}_c \frac{\hat{T}_c}{\hat{M}_c} (\hat{M}_{c(hi)} - \hat{M}_c) \\ &= \hat{R}_c \hat{T}_c + \hat{R}_c (\hat{T}_{c(hi)} - \hat{T}_c) - \hat{R}_c \hat{\mu}_c (\hat{M}_{c(hi)} - \hat{M}_c). \end{aligned}$$

Also we can rewrite $\hat{T}_c$ as

$$\begin{aligned}
\hat{T}_c &= \sum_{h=1}^{L}\sum_{i=1}^{n_h} S_{hic}\bar{y}_{hic} \\
&= \sum_h\sum_i \tilde{y}_{hic} = \sum_h n_h \frac{1}{n_h}\sum_i \tilde{y}_{hic} \\
&= \sum_h n_h \bar{\tilde{y}}_{hc}.
\end{aligned}$$

Then the estimate of T_c without one missing cluster is computed as

$$\begin{aligned}
\hat{T}_{c(hi)} &= n_h\bar{\tilde{y}}_{hc(hi)} + \sum_{h\neq h'} n_{h'}\bar{\tilde{y}}_{h'c} \\
&= n_h\Big(\frac{n_h\bar{\tilde{y}}_{hc} - \tilde{y}_{hic}}{n_h - 1}\Big) + \sum_{h\neq h'} n_{h'}\bar{\tilde{y}}_{h'c} \\
&= \frac{n_h}{n_h - 1}(n_h\bar{\tilde{y}}_{hc} - \bar{\tilde{y}}_{hc} + \bar{\tilde{y}}_{hc} - \tilde{y}_{hic}) + \sum_{h\neq h'} n_{h'}\bar{\tilde{y}}_{h'c} \\
&= \frac{n_h}{n_h - 1}(\bar{\tilde{y}}_{hc} - \tilde{y}_{hic}) + n_h\bar{\tilde{y}}_{hc} + \sum_{h\neq h'} n_{h'}\bar{\tilde{y}}_{h'c} \\
&= \frac{n_h}{n_h - 1}(\bar{\tilde{y}}_{hc} - \tilde{y}_{hic}) + \hat{T}_c.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\hat{T}_{c(hi)} - \hat{T}_c &= \frac{n_h}{n_h - 1}(\bar{\tilde{y}}_{hc} - \tilde{y}_{hic}) \\
&= -\frac{n_h}{n_h - 1}(\tilde{y}_{hic} - \frac{1}{n_h}\sum_i \tilde{y}_{hic}) \\
&= -\frac{n_h}{n_h - 1}(S_{hic}\bar{y}_{hic} - \frac{1}{n_h}\sum_i S_{hic}\bar{y}_{hic}).
\end{aligned}$$

By just replacing y_{hik} by δ_{hikc}, we obtain

$$\hat{M}_{c(hi)} - \hat{M}_c = -\frac{n_h}{n_h - 1}(S_{hic} - \frac{1}{n_h}\sum_i S_{hic}).$$

Then plug these expressions into the standard Taylor expansion, we have

$$\begin{aligned}
\hat{R}_{c(hi)}\hat{T}_{c(hi)} &- \hat{R}_c\hat{T}_c = \hat{R}_c(\hat{T}_{c(hi)} - \hat{T}_c) - \hat{R}_c\hat{\mu}_c(\hat{M}_{c(hi)} - \hat{M}_c) \\
&= -\hat{R}_c\frac{n_h}{n_h-1}(S_{hic}\bar{y}_{hic} - \frac{1}{n_h}\sum_i S_{hic}\bar{y}_{hic}) \\
&\quad +\hat{R}_c\hat{\mu}_c\frac{n_h}{n_h-1}(S_{hic} - \frac{1}{n_h}\sum_i S_{hic}) \\
&= -\hat{R}_c\frac{n_h}{n_h-1}\Big(S_{hic}\bar{y}_{hic} - \frac{1}{n_h}\sum_i S_{hic}\bar{y}_{hic} \\
&\quad -S_{hic}\hat{\mu}_c + \frac{1}{n_h}\sum_i S_{hic}\hat{\mu}_c\Big) \\
&= -\hat{R}_c\frac{n_h}{n_h-1}\Big\{S_{hic}(\bar{y}_{hic} - \hat{\mu}_c) - \frac{1}{n_h}\sum_i S_{hic}(\bar{y}_{hic} - \hat{\mu}_c)\Big\}.
\end{aligned}$$

Also, pluging this equation into the formula of the jackknife variance estimator, we finally obtain that

$$\begin{aligned}
v_{JL}(\hat{T}_{pst}) &= \sum_h \frac{n_h-1}{n_h}\sum_i\Big\{\sum_c(\hat{R}_{c(hi)}\hat{T}_{c(hi)} - \hat{R}_c\hat{T}_c)\Big\}^2 \\
&= \sum_h \frac{n_h-1}{n_h}\sum_i\Big[\sum_c -\hat{R}_c\frac{n_h}{n_h-1}\Big\{S_{hic}(\bar{y}_{hic} - \hat{\mu}_c) \\
&\quad -\frac{1}{n_h}\sum_i S_{hic}(\bar{y}_{hic} - \hat{\mu}_c)\Big\}\Big]^2 \\
&= \sum_h \frac{n_h}{n_h-1}\sum_i\Big\{\sum_c \hat{R}_c(g_{hic} - \frac{1}{n_h}\sum_{i\in s_h} g_{hic})\Big\}^2 \\
&= \sum_{h=1}^{L}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\Big\{\sum_c \hat{R}_c(g_{hic} - \bar{g}_{hc})\Big\}^2 \\
&= v_L^*(\hat{T}_{pst}).
\end{aligned}$$

The essential steps of the derivation above can be found in Yung and Rao (1996).

6.5 New Proposed Linearization Variance Estimator

Furthermore, we considered 2nd order Taylor series expansion for $\hat{R}_c\hat{T}_c$ at $(M_c,\, T_c)$ which is evaluated as

$$
\begin{aligned}
\hat{R}_c\hat{T}_c = M_c\frac{\hat{T}_c}{\hat{M}_c} \;\approx\;& M_c\Big\{\frac{T_c}{M_c} + \frac{1}{M_c}(\hat{T}_c - T_c) - \frac{T_c}{{M_c}^2}(\hat{M}_c - M_c) \\
& + \frac{T_c}{{M_c}^3}(\hat{M}_c - M_c)^2 - \frac{1}{{M_c}^2}(\hat{T}_c - T_c)(\hat{M}_c - M_c)\Big\} \\
=\;& T_c + \hat{T}_c - \frac{\hat{M}_c}{M_c}T_c \\
& + \frac{T_c}{{M_c}^2}(\hat{M}_c - M_c)^2 - \frac{1}{M_c}(\hat{T}_c - T_c)(\hat{M}_c - M_c) \\
=\;& T_c + \hat{T}_c(2 - \frac{\hat{M}_c}{M_c}) + T_c\Big\{\Big(\frac{\hat{M}_c}{M_c}\Big)^2 - 2\frac{\hat{M}_c}{M_c}\Big\} \\
=\;& T_c + (2 - \frac{\hat{M}_c}{M_c})(\hat{T}_c - \frac{\hat{M}_c}{M_c}T_c).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\hat{T}_{pst} - T \;=\;& \sum_c (\hat{R}_c\hat{T}_c - T_c) \\
=\;& \sum_c (2 - \frac{\hat{M}_c}{M_c})(\hat{T}_c - \frac{\hat{M}_c}{M_c}T_c) \\
=\;& \sum_h\sum_i\sum_c\sum_k \frac{w_{hi}M_{hi}}{m_{hi}}(y_{hik}\delta_{hikc} - \frac{\hat{T}_c}{\hat{M}_c})(2 - \frac{1}{\hat{R}_c}) \\
=\;& \sum_h\sum_i \frac{1}{m_{hi}}\sum_c (2 - \frac{1}{\hat{R}_c})S_{hic}(\bar{y}_{hic} - \hat{\mu}_c) \\
=\;& \sum_h\sum_i g_{hi} = \sum_h \frac{1}{n_h}\sum_i n_h g_{hi} \\
=\;& \sum_h \bar{g}^*_h,
\end{aligned}
$$

where $g_{hi} = \frac{1}{m_{hi}}\sum_c (2 - \frac{1}{\hat{R}_c})S_{hic}(\bar{y}_{hic} - \hat{\mu}_c)$ and $\tilde{g}^*_{hi}(= n_h\tilde{g}_{hi})$ are iid, $i = 1, ..., n_h$.

Therefore, variance of $\hat{T}_{pst}$ is

$$\begin{aligned} v_L^{**}(\hat{T}_{pst}) &= \sum_h \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (g_{hi}^* - \bar{g}_h^*)^2 \\ &= \sum_h \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (\tilde{g}_{hi} - \bar{g}_h)^2 \\ &= \sum_h \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \Big\{ \sum_c (2 - \frac{1}{\hat{R}_c})(g_{hic} - \bar{g}_{hc}) \Big\}^2 . \end{aligned}$$

Comparing this to the standard linearization estimator, the second-order linearization variance estimator has the function of the adjustment factor, $2 - 1/R_c$. This function works like the ratio, R_c, but slightly different. If the value of R_c is around 1, both have the values close to 1. But for extremely unbalanced case such that the values are far from 1, $2 - 1/R_c$ gives smaller weights for each poststratum. So $2 - 1/R_c$ also has the functionality of balancing weights for poststrata which can reduce the bias from the unbalanced sampling. Rao's adjusted variance estimator can be obtained by adding the ratio adjustment factor, R_c to the standard linearization variance estimator. We note that $M_c/\hat{M}_c$ converges in probability to 1. So, there is no harm in switching v_L to v_L^* since v_L^* is asymptotically equivalent to v_L. If the Taylor expansion is expanded to the second-order, we have the new linearization variance estimator with the function, $2 - 1/R_c$. This function came from the process of the second-order Taylor approximation. So the second-order estimator has the function which balances weights for the poststrata while the standard linearization variance estimator, v_L, does not have. We know that the new variance estimator is equivalent to v_L^* and v_L. Because the function, $2 - 1/R_c$, also converges in probability to 1. Therefore, its adjusted version also can be suggested as

$$v_{adj,L}^{**}(\hat{T}_{pst}) = \sum_h \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} \Big\{ \sum_c \hat{R}_c (2 - \frac{1}{\hat{R}_c})(g_{hic} - \bar{g}_{hc}) \Big\}^2 ,$$

where $g_{hic} = \frac{1}{m_{hi}} \sum_{k \in s_{hi}} w_{hi} M_{hi} \delta_{hikc}(y_{hik} - \frac{T_c}{M_c}) = S_{hic}(\bar{y}_{hic} - \hat{\mu}_c)$. $v_L(\hat{T}_{pst})$, $v_L^{**}(\hat{T}_{pst})$

and $v^{**}_{adj,L}(\hat{T}_{pst})$ are all asymptotically equivalent, because $\hat{R}_c \xrightarrow{p} 1$.

We also can consider that poststratification is made across on the clusters, not the units within cluster.

$$\begin{aligned} M_c &= \sum_h \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} \delta_{hic} \\ &= \sum_h \sum_{i=1}^{N_h} \delta_{hic} m_{hi}. \end{aligned}$$

Then, population total estimator, $\hat{T}$ can be expressed in different way as follows,

$$\begin{aligned} \hat{T} &= \sum_h \sum_{i=1}^{n_h} w_{hi} \hat{T}_{hi} \\ &= \sum_h \sum_i w_{hi} M_{hi} \bar{y}_{hi} \\ &= \sum_c \sum_h \sum_i S_{hi} \delta_{hic} \bar{y}_{hi} \\ &= \sum_c \hat{T}_c. \end{aligned}$$

Now we know

$$\hat{T}_c = \sum_h \sum_i S_{hi} \delta_{hic} \bar{y}_{hi}.$$

By replacing $\bar{y}_{hi}$ by δ_{hic}, then

$$\hat{M}_c = \sum_h \sum_i S_{hi} \Big(= \sum_h \sum_i S_{hi} \delta_{hic} \Big),$$

which is identical to $\hat{M}_c$ when $m_{hic} = m_{hi}$ $(S_{hi} = w_{hi} M_{hi}$ $)$. Therefore,

$$\begin{aligned} v_L(\hat{T}_{pst}) &= \sum_h \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (g^*_{hi} - \bar{g}^*_{h\cdot\cdot})^2 \\ &= \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (g_{hi} - \bar{g}_{h\cdot})^2. \end{aligned}$$

We also apply second order linearization to jackknife variance estimator. By linearizing, second-order jackknife linearization variance estimator can be obtained.

Second-order Taylor expansion of $\hat{R}_{c(hi)}\hat{T}_{c(hi)}$ at ($\hat{M}_c$,$\hat{T}_c$) is

$$\begin{aligned}
\hat{R}_{c(hi)}\hat{T}_{c(hi)} &= \frac{M_c}{\hat{M}_{c(hi)}}\hat{T}_{c(hi)} \\
&\approx \frac{M_c}{\hat{M}_c}\hat{T}_c + \frac{M_c}{\hat{M}_c}(\hat{T}_{c(hi)} - \hat{T}_c) - \frac{M_c}{\hat{M_c}^2}\hat{T}_c(\hat{M}_{c(hi)} - \hat{M}_c) \\
&\quad + \frac{\hat{T}_c}{\hat{M}_c^3}(\hat{M}_{c(hi)} - \hat{M}_c)^2 - \frac{1}{\hat{M}_c^2}(\hat{T}_{c(hi)} - \hat{T}_c)(\hat{M}_{c(hi)} + \hat{M}_c) \\
&= \hat{R}_c\hat{T}_c + \hat{R}_c\Big\{\hat{T}_{c(hi)} - \hat{T}_c - \frac{\hat{T}_c}{\hat{M}_c}\hat{M}_{c(hi)} + \hat{T}_c \\
&\quad + \frac{\hat{T}_c}{\hat{M}_c^2}(\hat{M}_{c(hi)} - \hat{M}_c)^2 - \frac{1}{\hat{M}_c}(\hat{T}_{c(hi)} - \hat{T}_c)(\hat{M}_{c(hi)} + \hat{M}_c)\Big\} \\
&= \hat{R}_c\hat{T}_c + \hat{R}_c\Big\{\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c + \frac{\hat{M}_{c(hi)}^2}{\hat{M}_c^2}\hat{T}_c + \hat{T}_c - 2\frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c \\
&\quad - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_{c(hi)} + \hat{T}_{c(hi)} + \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c - \hat{T}_c\Big\} \\
&= \hat{R}_c\hat{T}_c + \hat{R}_c\Big\{2\Big(\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c\Big) - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\Big(\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}}\hat{T}_c\Big)\Big\} \\
&= \hat{R}_c\hat{T}_c + \hat{R}_c\Big\{\Big(2 - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\Big)\Big(\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c\Big)\Big\}.
\end{aligned}$$

Then we have

$$\hat{R}_{c(hi)}\hat{T}_{c(hi)} - \hat{R}_c\hat{T}_c = \hat{R}_c\Big(2 - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\Big)\Big(\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c\Big).$$

Therefore,

$$\begin{aligned}
v_{JL}^{*}(\hat{T}_{pst}) &= \sum_h \frac{n_h - 1}{n_h}\sum_{i=1}^{n_h}\Big\{\sum_c(\hat{R}_{c(hi)}\hat{T}_{c(hi)} - \hat{R}_c\hat{T}_c)\Big\}^2 \\
&\approx \sum_h \frac{n_h - 1}{n_h}\sum_{i=1}^{n_h}\Big\{\sum_c \hat{R}_c\Big(2 - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\Big)\Big(\hat{T}_{c(hi)} - \frac{\hat{M}_{c(hi)}}{\hat{M}_c}\hat{T}_c\Big)\Big\}^2.
\end{aligned}$$

The second-order jackknife linearization variance estimator is very similar to the adjusted version of the second order linearization variance estimator that we proposed before. It has $\hat{M}_{c(li)}/\hat{M}_c$ and $\hat{T}_{c(li)}$ instead of $\hat{M}_c/M_c$ and $\hat{T}_{c(li)}$. However, v^*_{JL} needs to compute $\hat{M}_{c(li)}$ and $\hat{T}_{c(li)}$ which require extensive calculations as the standard jackknife estimator does. So it may not be preferred to the jackknife variance estimator with respect to time and cost.

CHAPTER VII

SIMULATION STUDY I

To observe and compare the performances of variance estimators which include the standard linearization estimtor v_L, the jackknife linearization estimator v_{JL}, the second order linearization estimator v_L^{**} and its adjusted version $v_{adj,L}^{**}$, we generate a population with $50,000$ with four poststrata. The values of y_{hik} are generated from four different normal distributions (poststrata) with given means (μ_1, μ_2, μ_3, μ_4)=(40, 60, 80, 100) and standard deviations (σ_1, σ_2, σ_3, σ_4)=(8.94, 10.95, 12.65, 14.14). Poststrata sizes are randomly determined and assigned as (9,561, 18,800, 6,163, 15,476) respectively. All 50,000 units are randomly apportioned into 10 strata and 800 clusters with equal probabilities. Consequently, each stratum has 80 clusters and cluster size varies from 40 to 89. After the clustered population is obtained, iterative drawings of sample should be carried under designed sampling plan.

First, we consider one-stage sampling scheme. One-stage is a special case of two-stage sampling design. Because if all the units are selected within sampling clusters under two-stage sampling design, this becomes a single stage. The largest cluster size of the generated population is 89. If we select 89 units within all the sampling clusters, which covers all the units in, that is equivalent to one-stage cluster sampling. Hence, calculations of variance estimators for both one-stage and two-stage are carried in the same manner. In one-stage sampling, we selected 1,000 independent samples. At each sample, n_h clusters were selected from each hth-stratum with probability proportional to cluster size. We repeated this with four different numbers of sampling clusters $n_h = 4, 6, 8, 10$ respectively, for $i = 1, ..., 10$ per stratum with selecting all the units in the clusters.

The sampling method used under two-stage cluster sampling plan is that *pps* at first stage and *srs* at second stage. This yields equal selection probabilities for all units. The selection probability of j-th unit in i-th cluster is

$$\pi_{ij} = \frac{nM_i}{M}\frac{m_{hi}}{M_i} = \frac{nm_{hi}}{M}.$$

Eight clusters per stratum are selected with *pps* so total eighty clusters out of eight hundreds population clusters are sampled at the first stage. For each sampling cluster, $m_{hi} = 6, 10, 14, 18$ units within cluster are drawned respectively. If number of units in a cluster is smaller m_{hi}, all the units in that cluster are taken. So total sample size for each time of sampling is not fixed but similar. Empirical mean sqaure error or say 'v_E' is calculated for each variance estimator based on 1,000 samples defined by

$$v_E = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{T}_{pst,i} - T)^2,$$

where $\hat{T}_{pst,i}$ is the estimated total for the i-th generated sample $(i = 1, 2, ..., 1000)$. Mean sqaure error and relative bias are used to measure the precision and performance for each variance estimator based on the sample size.

$$\begin{aligned} \text{MSE} &= \frac{1}{1000}\sum_{i=1}^{1000}(\hat{v}_i - v_E)^2, \\ \text{Relative bias} &= \frac{1}{1000}\frac{\sum_i v_i}{v_E} - 1, \end{aligned}$$

where $\hat{v}_i$ is the variance estimate for the i-th generated sample $(i = 1, 2, ..., 1000)$.

The second order linearization estimator, v_L^{**}, performs as well as v_L and v_{JL} for both one and two-stage. We also used a real finite population which is called third grade population. It consists of 2,427 students who participated in the Third International Mathematics and Science Study (Caslyn, Gonzales and Frase, 1999). The methods used in conducting the original study are given in TIMSS International Study Center (1996). The population consists of only students from the United States and it has four regions which are strata. Clusters are schools while units within clusters are the students. We limit the variable of interest be the total math score of the population and let the poststrata to be the ethics which has eight categories in this study. $n_1 = 11, n_2 = 16, n_3 = 10, n_4 = 23$ clusters are selected from stratum with proportional allocation and $m = 4, 8, 12, 16$ units are sampled within each cluster, respectively. 1,000 simulations for each different number of sampling units shows similar result to simulated population. v_L^{**} still estimates the variance of the poststratified estimator well.

In Tables 6 and 7 and from Figures 2 to 9, poststratified estimator shows much better performance than standard estimator. Performances of the variance estimators are shown in Tables 8 and 9 and from Figures 10 to 14. We also recorded the results of the simulations on the third grade data in Tables 10 and 11.

Table 6: Point estimators at one-stage on the simulated population

		$\hat{T}$	$\hat{T}_{pst}$
Relative bias	$m = 4$	-0.01	0.00
	$m = 6$	0.00	0.00
	$m = 8$	-0.01	0.00
	$m = 10$	-0.01	0.00
$MSE(\div 10^7)$	$n = 4$	51.7	8.09
	$n = 6$	33.3	5.19
	$n = 8$	25.8	3.87
	$n = 10$	18.2	2.80

Table 7: Point estimators at two-stage on the simulated population

		$\hat{T}$	$\hat{T}_{pst}$
Relative bias	$m = 6$	-0.01	0.00
	$m = 10$	-0.01	-0.01
	$m = 14$	-0.01	0.00
	$m = 18$	0.00	0.00
$MSE(\div 10^8)$	$m = 6$	30.34	4.00
	$m = 10$	17.96	2.71
	$m = 14$	12.55	1.73
	$m = 18$	8.74	1.39

Table 8: Variance estimators at one-stage on the simulated population

		V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
Relative bias	$m = 4$	0.002	0.002	0.001	0.003
	$m = 6$	0.021	0.021	0.020	0.022
	$m = 8$	0.022	0.022	0.021	0.022
	$m = 10$	0.076	0.076	0.075	0.076
$MSE(\div 10^{13})$	$m = 4$	16.17	16.09	16.04	16.18
	$m = 6$	4.300	4.284	4.238	4.335
	$m = 8$	1.890	1.894	1.876	1.922
	$m = 10$	2.841	2.838	2.814	2.860

Table 9: Variance estimators at two-stage on the simulated population

		V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
Relative bias	$m = 6$	0.014	0.017	0.011	0.026
	$m = 10$	-0.035	-0.028	-0.037	-0.028
	$m = 14$	0.025	0.023	0.026	0.029
	$m = 18$	0.007	0.005	0.007	0.011
$MSE(\div 10^{14})$	$m = 6$	27.57	28.68	26.97	34.20
	$m = 10$	11.54	11.30	11.94	10.87
	$m = 14$	4.999	5.092	4.894	5.499
	$m = 18$	2.399	2.415	2.387	2.510

Table 10: Point estimators at two-stage on the third grade population

		$\hat{T}$	$\hat{T}_{pst}$
Relative bias	$m = 4$	-0.01	-0.01
	$m = 12$	-0.01	-0.01
$MSE(\div 10^8)$	$n = 4$	2.45	2.00
	$n = 12$	1.75	1.26

Table 11: Variance estimators at two-stage on the third grade population, $n_1 = 11, n_2 = 16, n_3 = 10, n_4 = 23$

		V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
Relative bias	$m = 4$	0.055	0.056	0.037	0.075
	$m = 12$	0.105	0.101	0.089	0.109
$MSE(\div 10^{15})$	$m = 4$	2.039	2.207	1.807	3.378
	$m = 12$	1.420	1.349	1.169	1.618

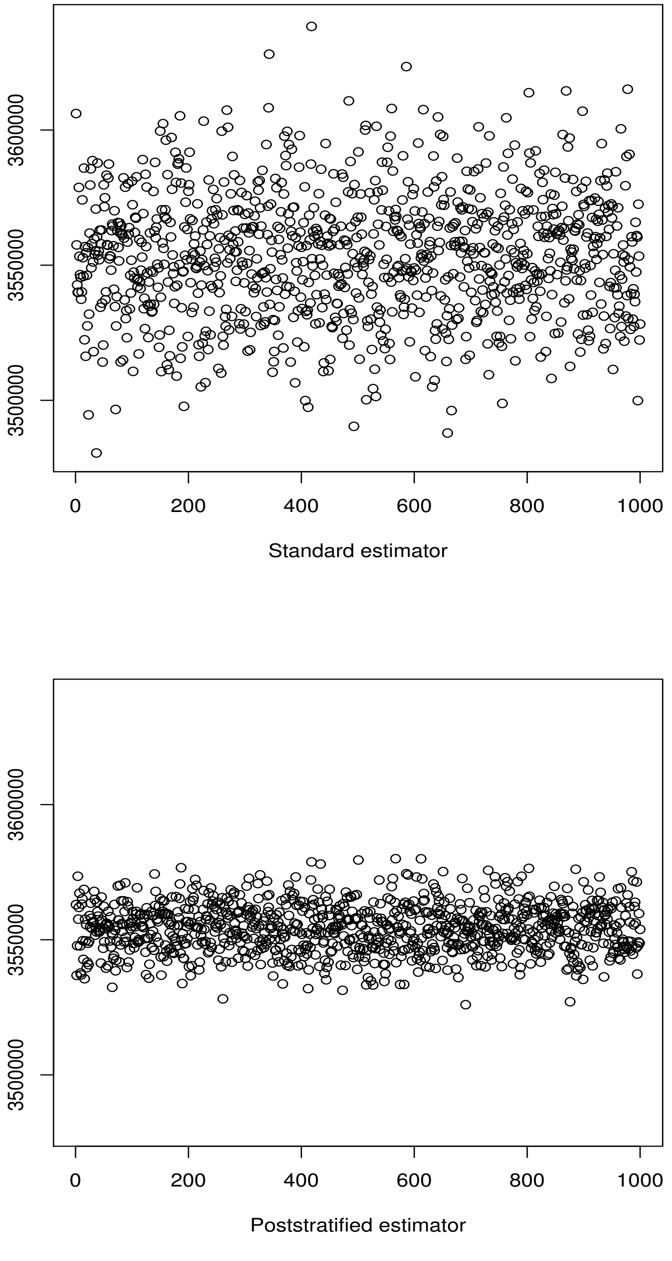


Figure 2: Point estimates of population total on 1,000 samples from the simulated population under one-stage cluster sampling, $n = 4$

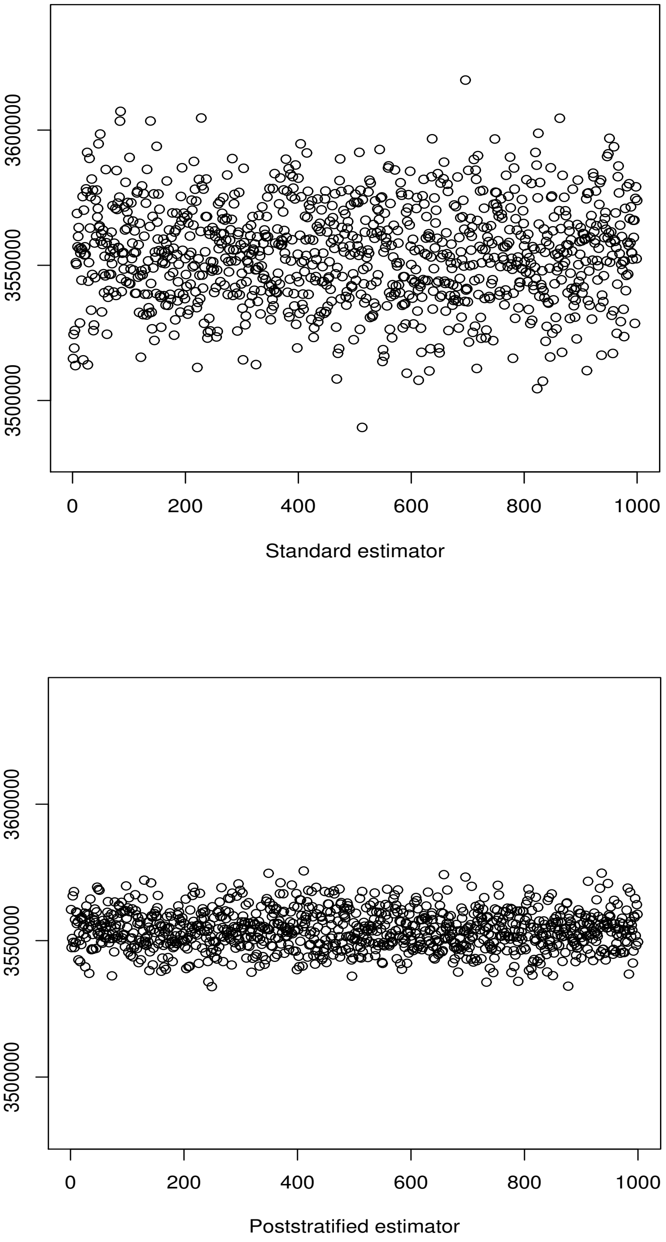


Figure 3: Point estimates of population total on 1,000 samples from the simulated population under one-stage cluster sampling, $n = 6$

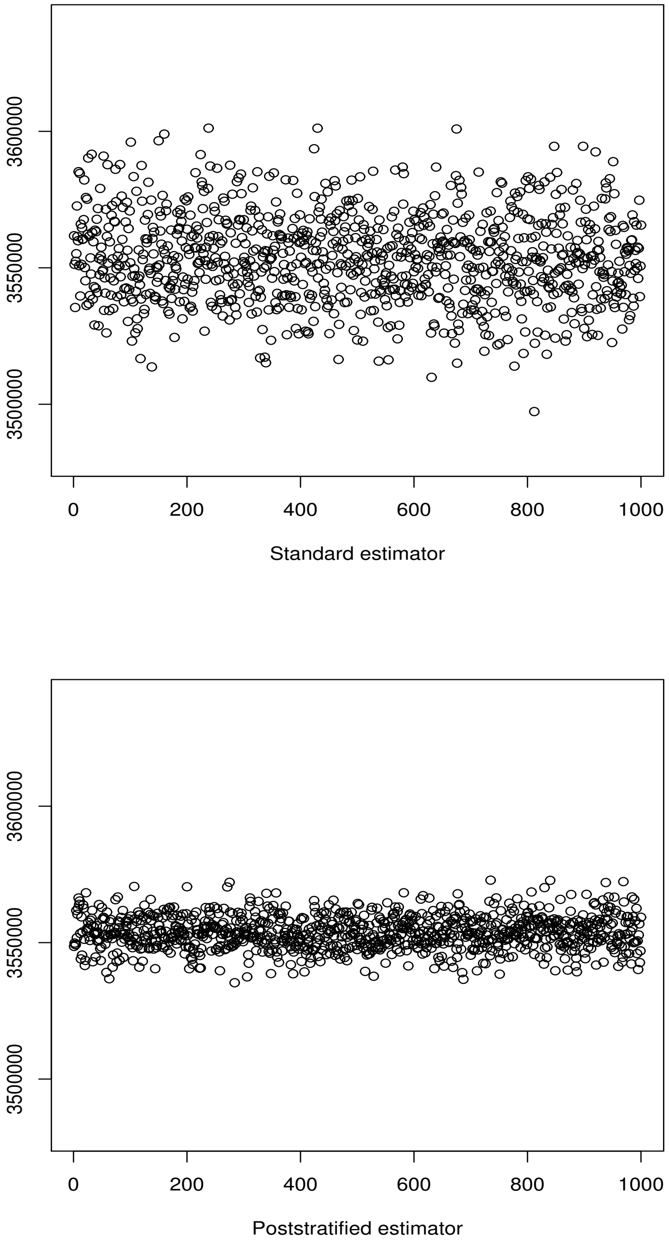
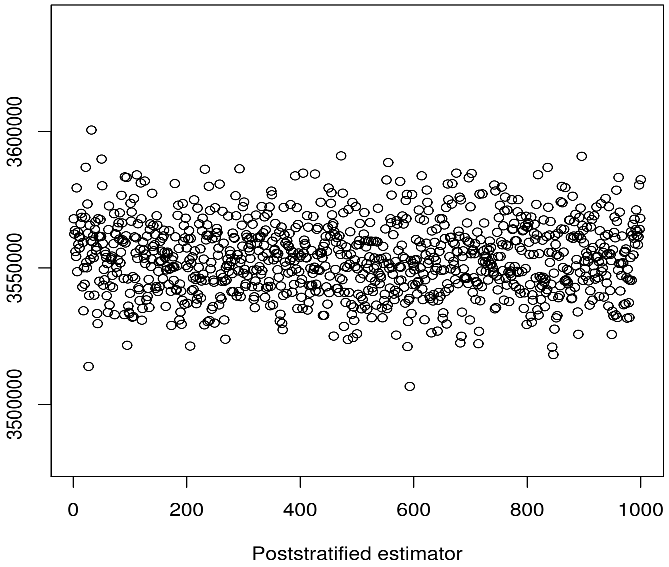


Figure 4: Point estimates of population total on 1,000 samples from the simulated population under one-stage cluster sampling, $n = 8$



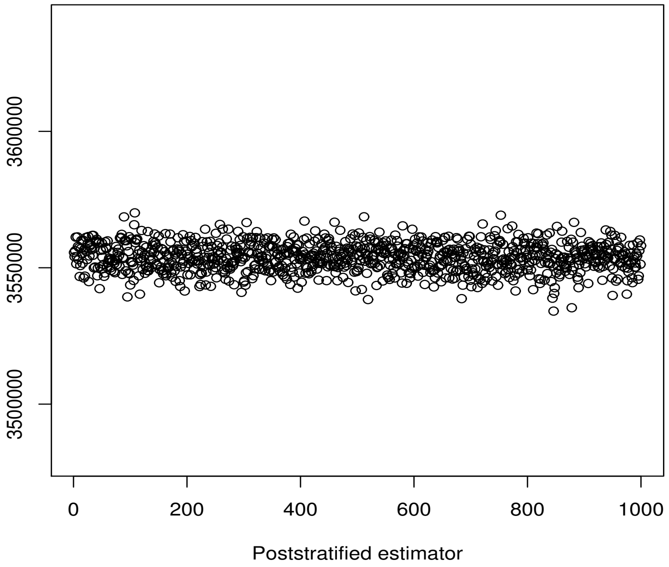


Figure 5: Point estimates of population total on 1,000 samples from the simulated population under one-stage cluster sampling, $n = 10$

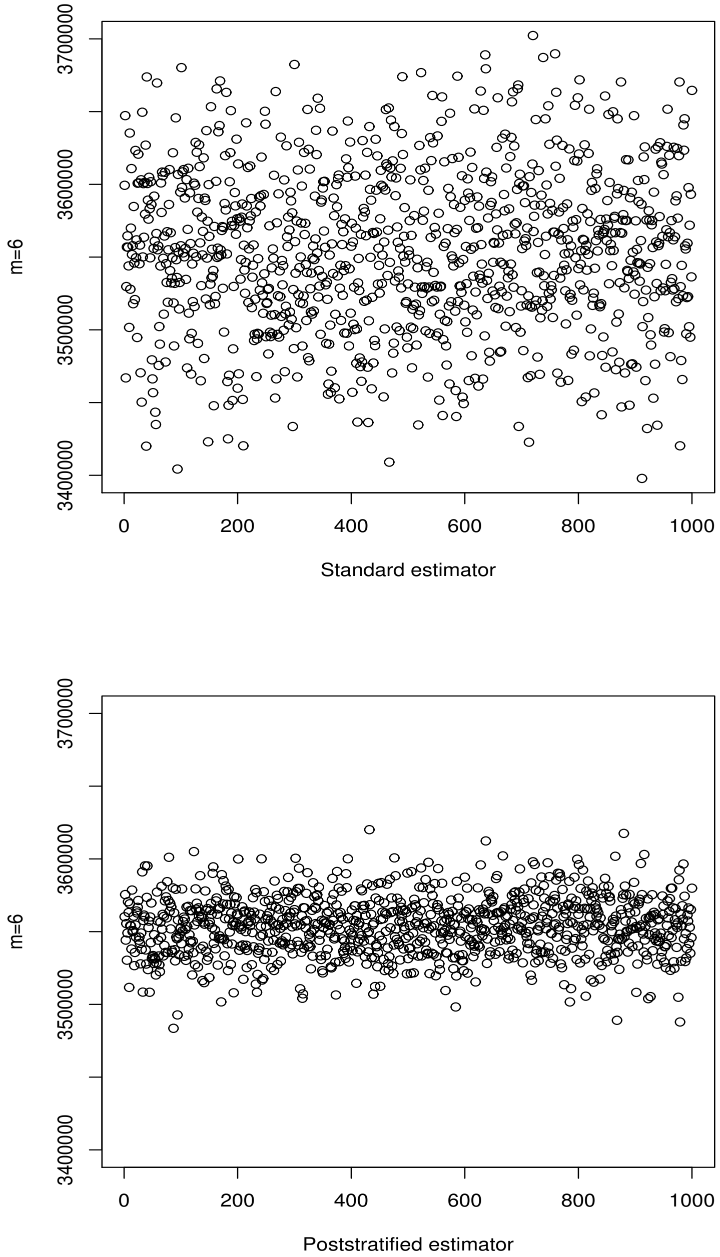


Figure 6: Point estimates of population total on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 6$ units in each cluster

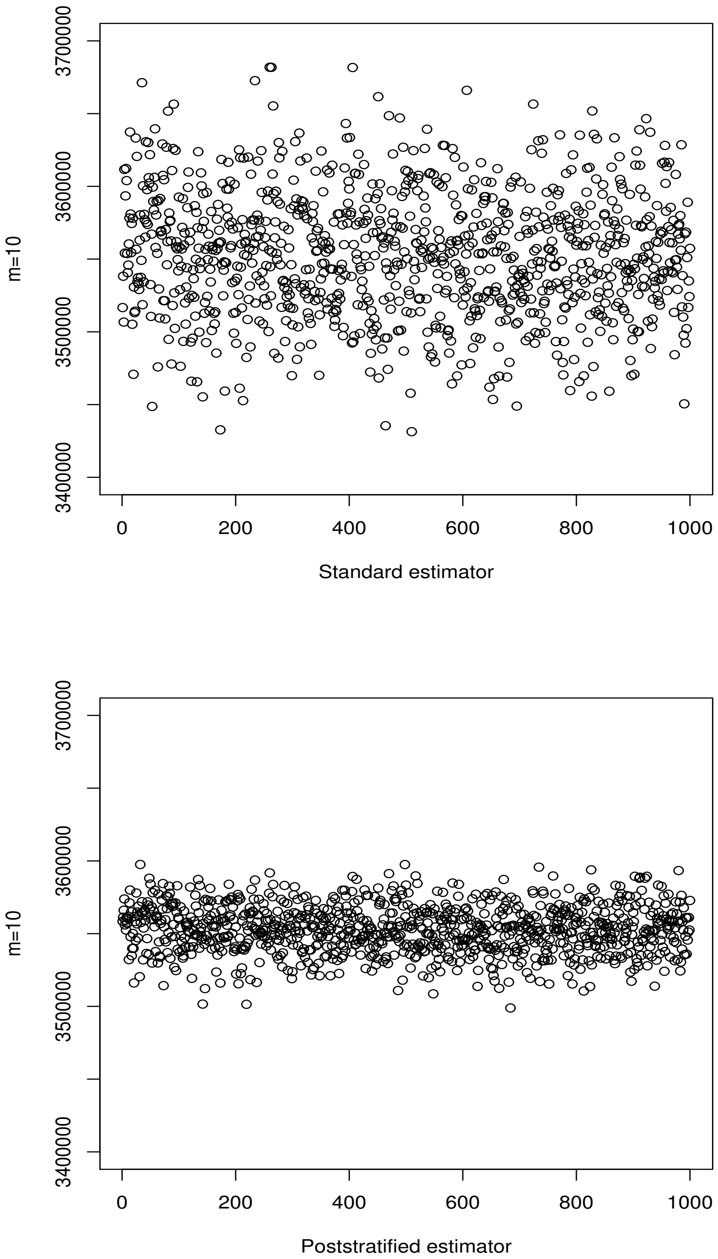


Figure 7: Point estimates of population total on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 10$ units in each cluster

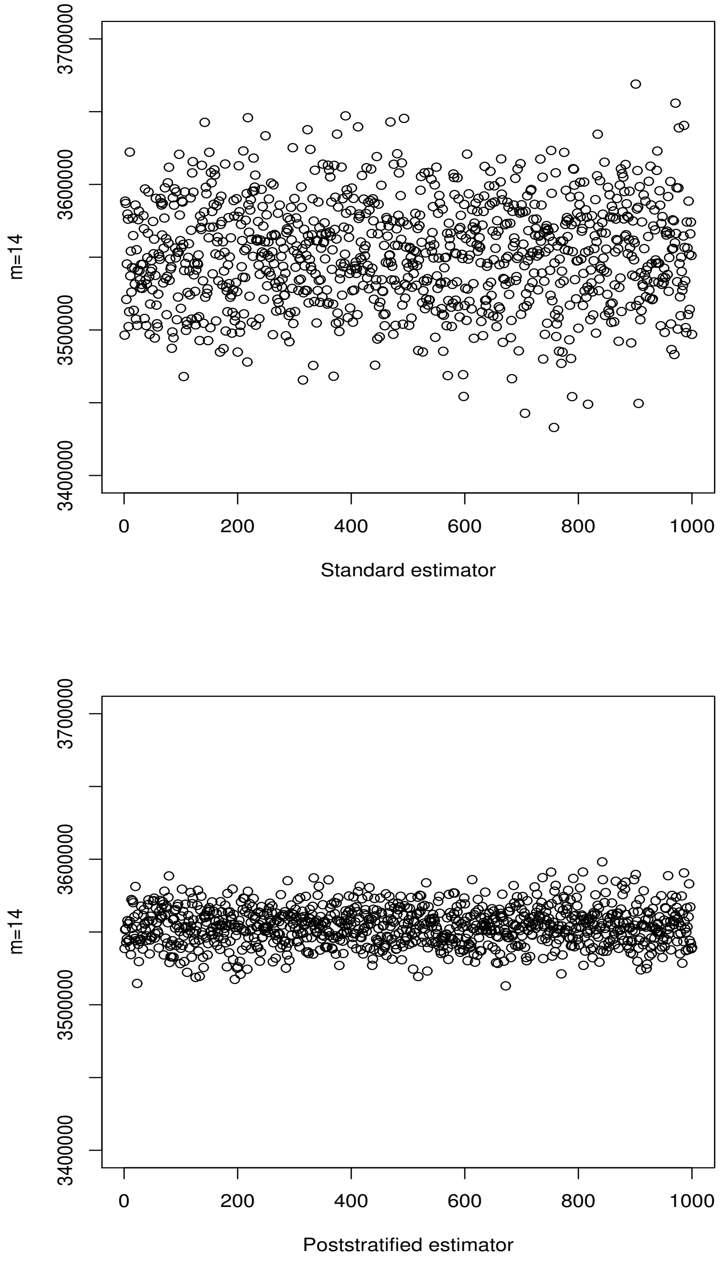


Figure 8: Point estimates of population total on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 14$ units in each cluster

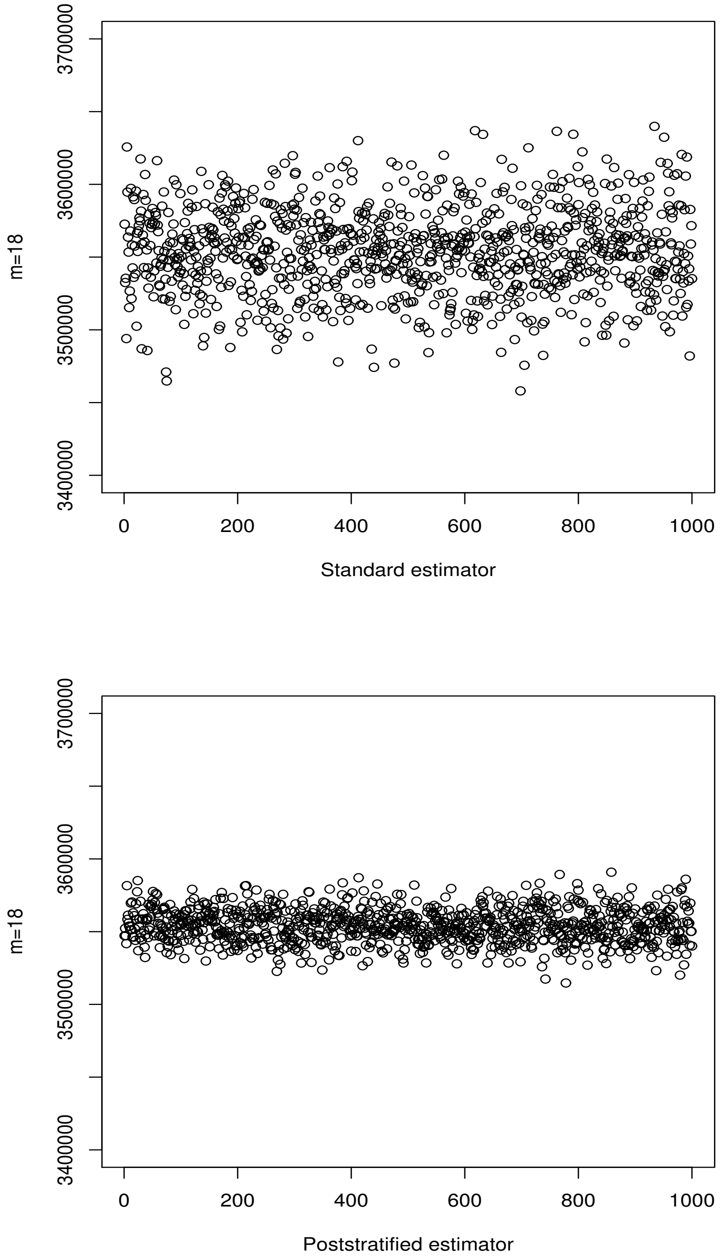
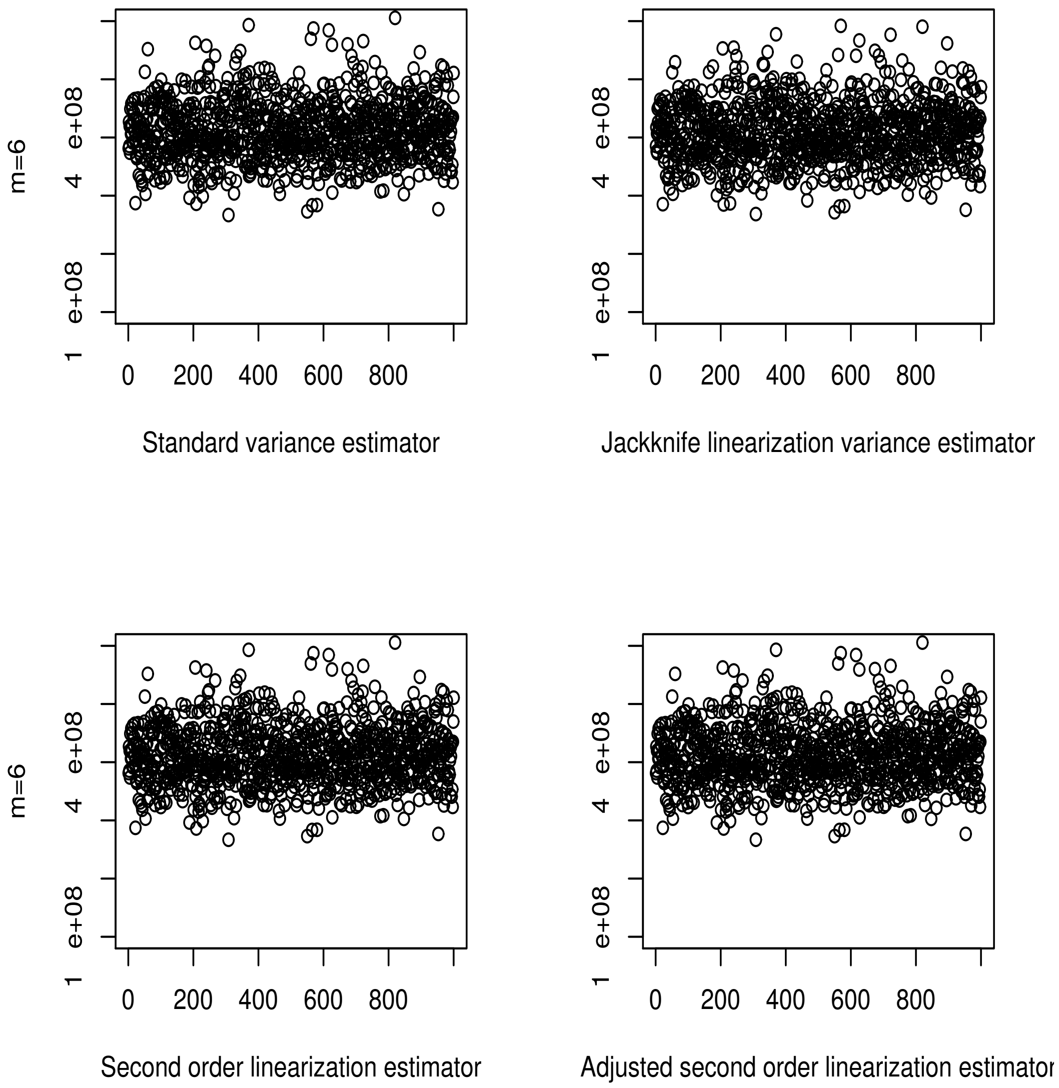


Figure 9: Point estimates of population total on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 18$ units in each cluster



Standard variance estimator

Jackknife linearization variance estimator

Second order linearization estimator

Adjusted second order linearization estimator

Figure 10: Variance estimates based on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 6$ units in each cluster

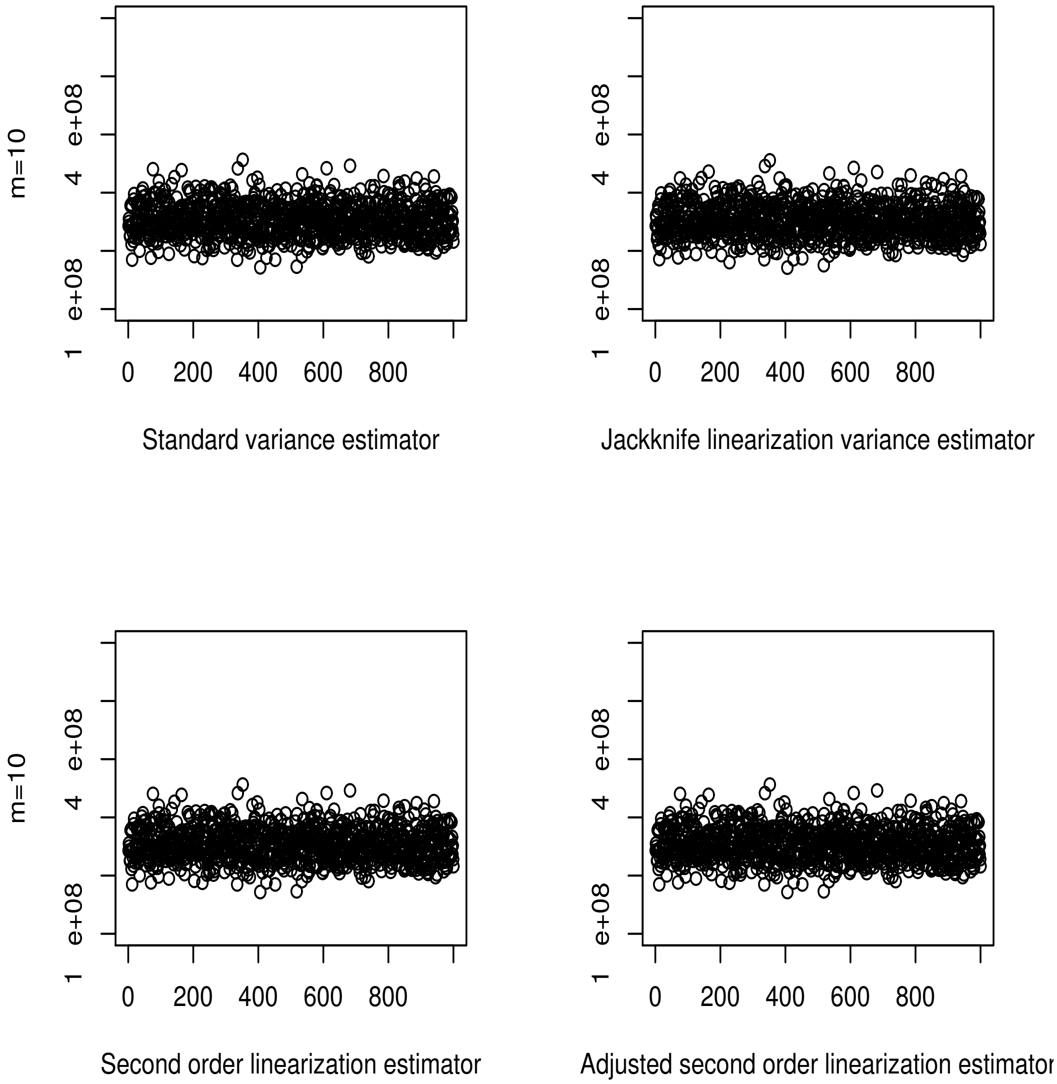
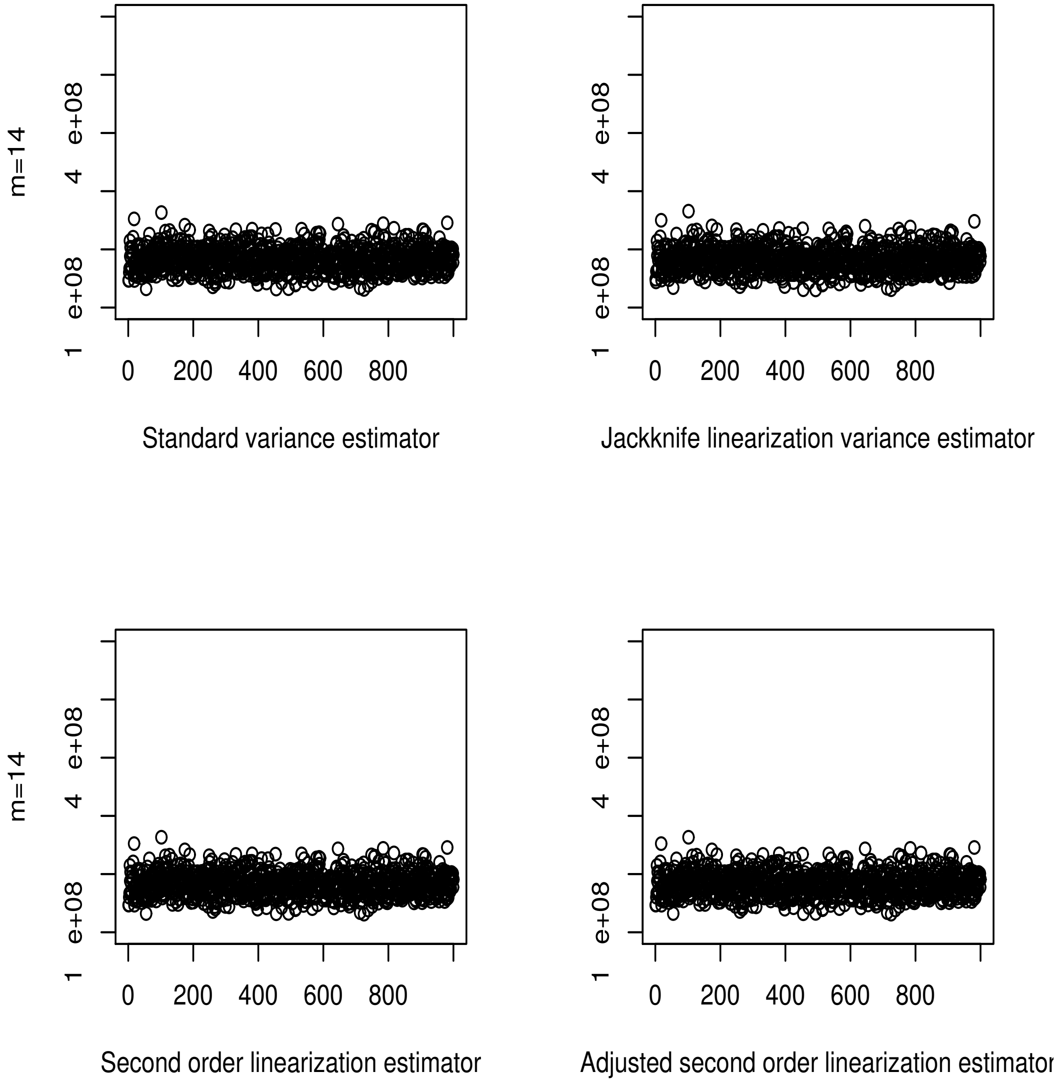


Figure 11: Variance estimates based on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 10$ units in each cluster



Standard variance estimator

Jackknife linearization variance estimator

Second order linearization estimator

Adjusted second order linearization estimator

Figure 12: Variance estimates based on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 14$ units in each cluster

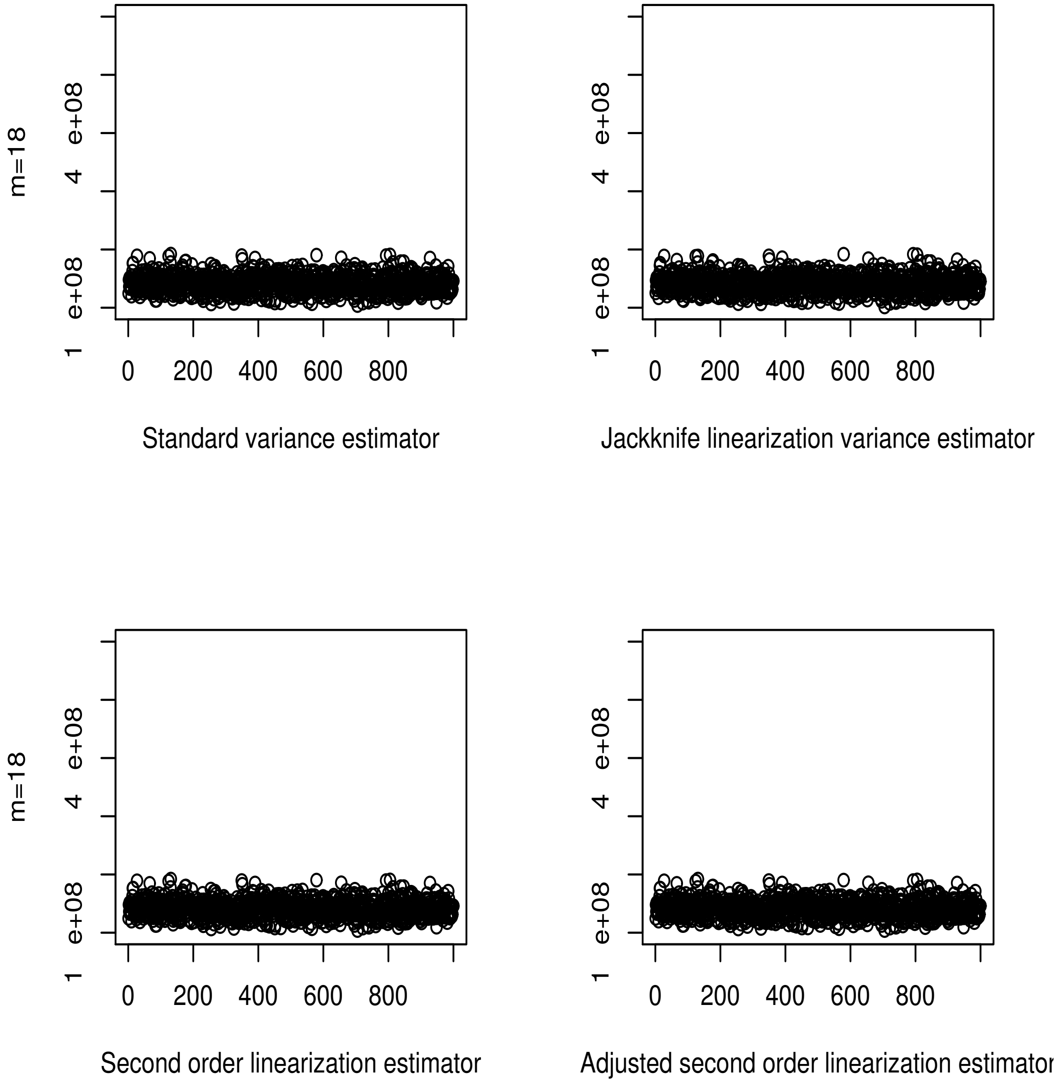


Figure 13: Variance estimates based on 1,000 samples from the simulated population under two-stage cluster sampling, selecting eight clusters per stratum and $m = 18$ units in each cluster

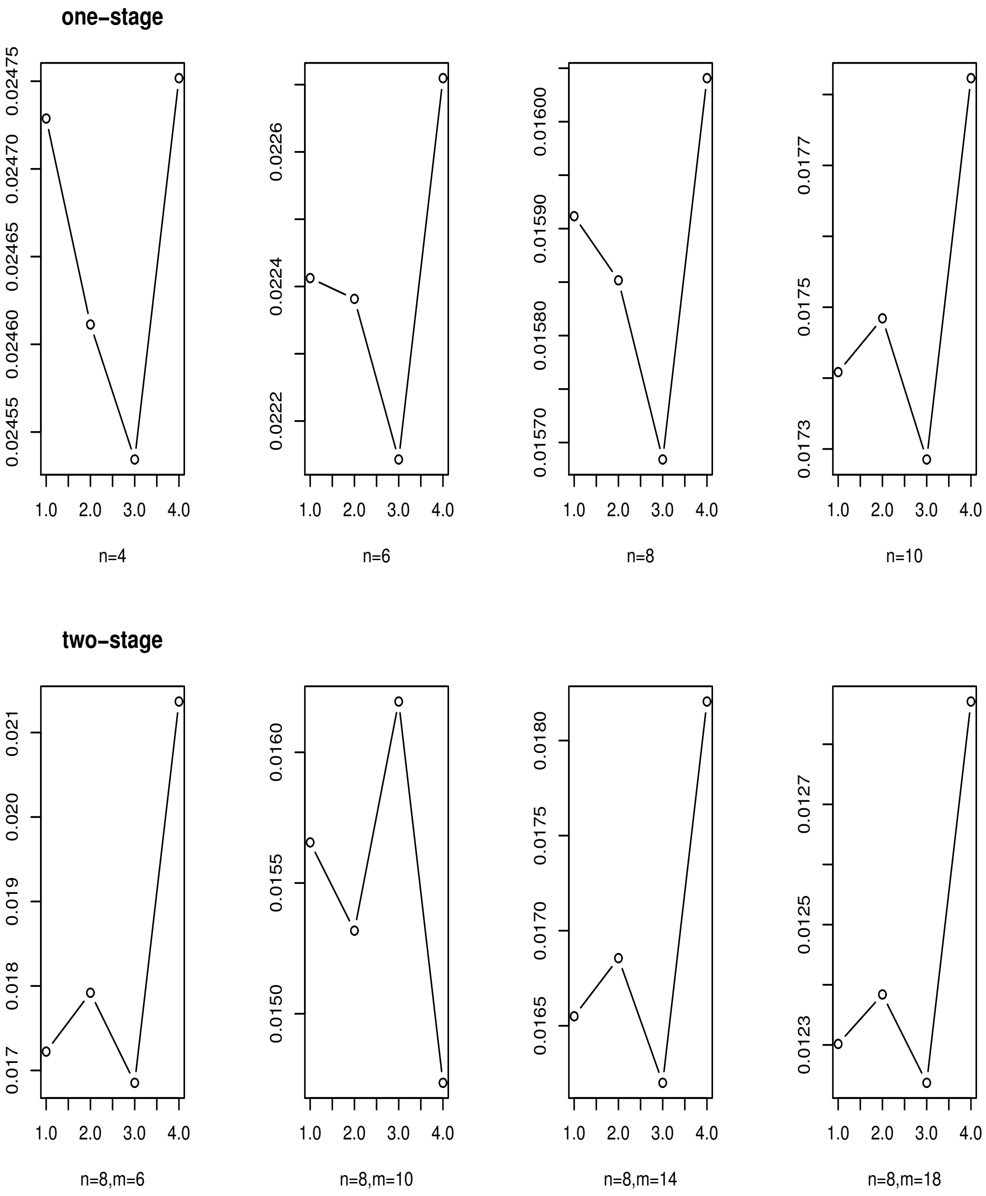


Figure 14: Relative mse, $\sum_{i=1}^{1000}(\frac{v_i - v_E}{v_E})^2/1,000$, of variance estimators in one-stage and two-stage where n = number of sampling cluster per stratum, m = number of units within each sampling cluster: in order of standard variance estimator(1), jackknife linearization estimator(2), second order linearization estimator(3) and its adjuster version(4)

CHAPTER VIII

SIMULATION STUDY II

8.1 Homogeneous vs Nonhomogeneous

We generate two populations of size, $40,483$ and $59,891$, both of which have eight poststrata. The values of y_{lik} are generated from each of different poststrata. We assumed each poststrata follows normal distribution with given mean and standard deviation. And the sizes of poststrata are randomly assigned. For the first generated population, all 40,483 units are apportioned into 20 strata and 800 clusters. Consequently, each stratum has 80 clusters and cluster size varies from 40 to 89. After the clustered population is obtained, iterative drawings of sample should be carried under designed sampling plan. The second population is created exactly same way except it has 10 strata. It has also 800 clusters. However, the distribution of cluster means in the first population is quite homogeneous but pretty heterogeneous in the second population. So we can compare the performance of the variance estimators if they are applied to different situations.

We considered 2 different configurations of sampling for each of two generated populations denoted by case A and B for the first population and case C and D for the second one. A is sampling 10 clusters with *pps* per each stratum and subsampling 10 units within each sampled cluster. B is $(7, 10)$ which is 7 clusters per stratum and 10 units in each cluster. For the second population, $C = (10, 15)$ and $D = (15, 10)$. If number of units in a cluster is smaller m_{li}, all the units in that cluster are taken. So total sample size for each time of sampling is not fixed but similar.

The result for the variance estimators in the heterogeneous case which is population 2 are recorded in Tables 14 and 15. With respect to MSE, v_L^{**} performs very well here. It has 16.9% and 12.4% smaller MSE than v_{JL}. It also has the shortest interval among the 4 competitors. However, 4 estimators perform similarly in the categories of the relative bias and the coverage. Every estimators shows almost 95% coverage and good relative bias between 1.7% and 3.2%.

In the population 1, the homogeneous one, v_L^{**} has the smallest MSE but the differences are smaller than in the population 2. Its MSE is just 2.7% and 1.2%.

It is difficult to drive any solid conclusion such that one of the estimators is superior to others based on the results. Because, the difference is not significant. But it looks clearly that the two new estimators, v_L^{**} and $v_{adj,L}^{**}$ can be considered as good candidates for the variance estimators of the poststratified estimator.

The results of the simulations for homogeneous case and nonhomogeneous case are shown in from Tables 12 to 15 and from Figures 15 to 22.

Table 12: Case A in homogeneous population

	V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
$MSE(\div 10^{13})$	2.93	3.03	2.87	3.53
Relative bias	0.006	0.007	0.004	0.011
Coverage	0.947	0.947	0.946	0.949
Lengths	38,502	38,534	38,425	38,676

Table 13: Case B in homogeneous population

	V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
$MSE(\div 10^{13})$	9.23	9.38	9.15	10.59
Relative bias	0.003	0.004	0.001	0.009
Coverage	0.948	0.948	0.947	0.950
Lengths	37,664	37,699	37,595	37,839

Table 14: Case C in nonhomogeneous population

	V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
$MSE(\div 10^{14})$	1.108	1.193	1.020	1.617
Relative bias	0.020	0.017	0.023	0.032
Coverage	0.956	0.954	0.953	0.956
Lengths	44,595	44,719	44,444	45,116

Table 15: Case D in nonhomogeneous population

	V_L	V_{JL}	V_L^{**}	$V_{adj,L}^{**}$
$MSE(\div 10^{13})$	6.399	6.792	6.044	9.008
Relative bias	0.009	0.011	0.006	0.019
Coverage	0.944	0.944	0.942	0.946
Lengths	44,506	44,604	44,364	45,941

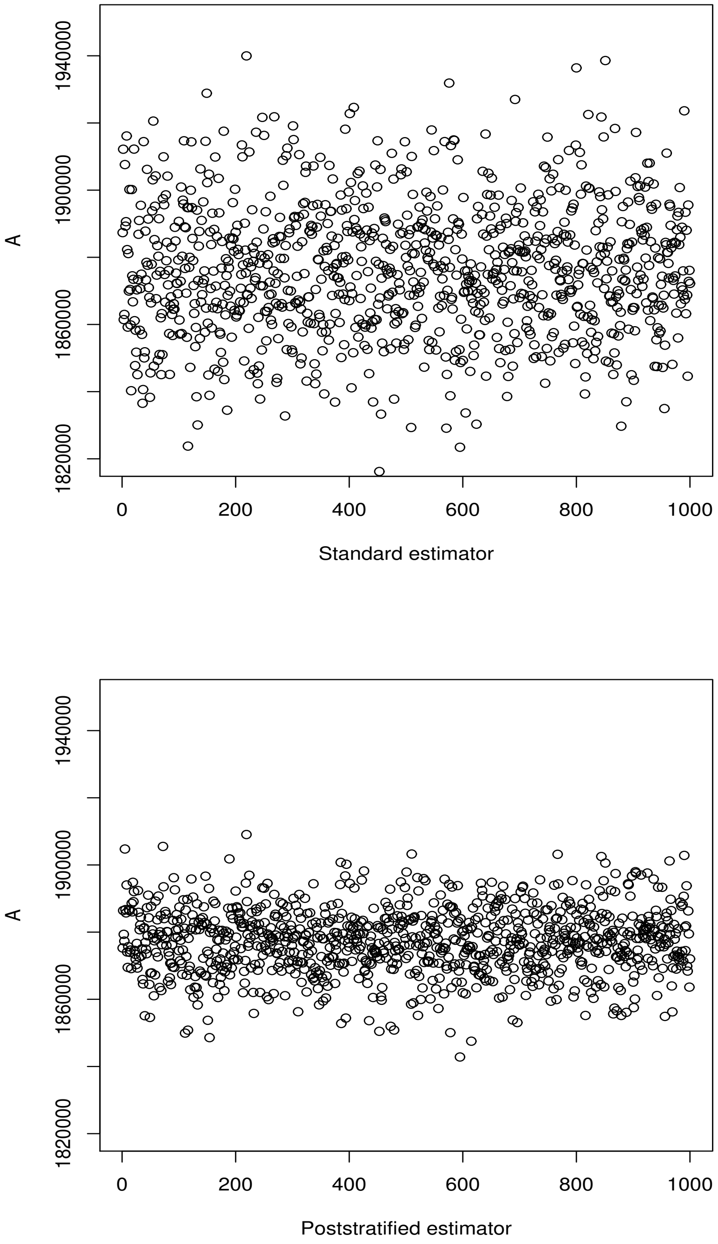


Figure 15: Point estimates of population total on 1,000 samples from the homogeneous simulated population when A configuration

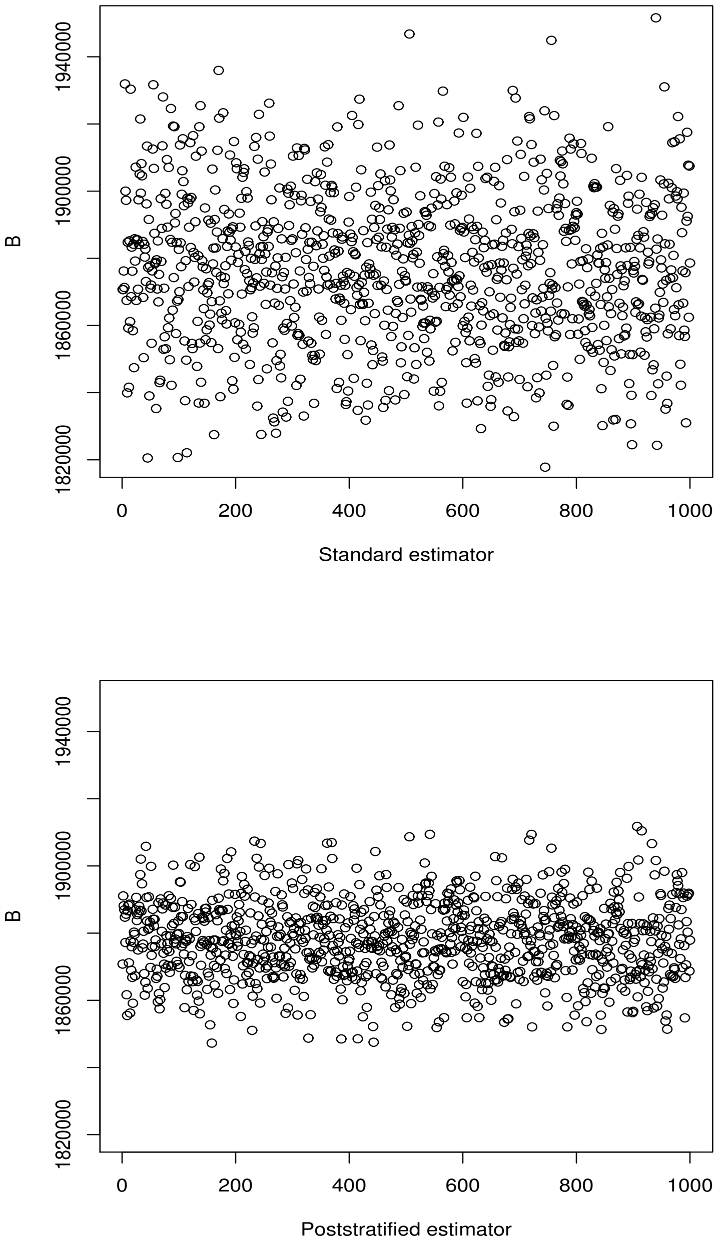


Figure 16: Point estimates of population total on 1,000 samples from the homogeneous simulated population when B configuration

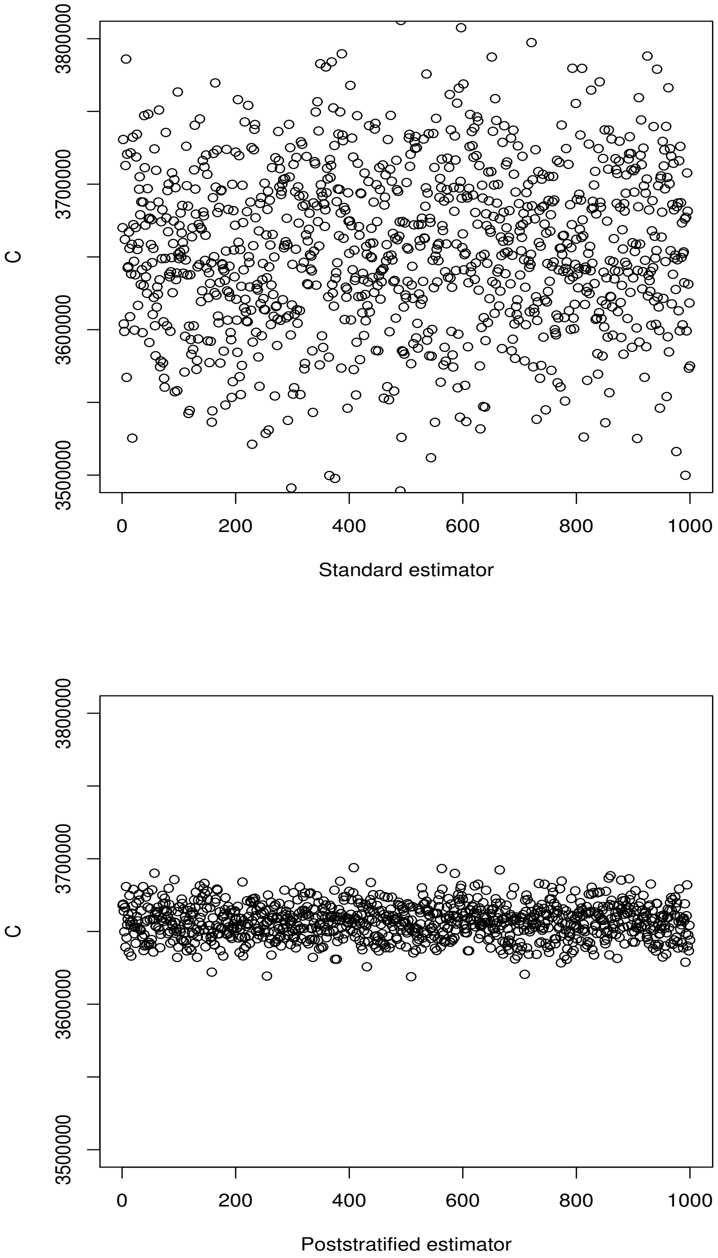


Figure 17: Point estimates of population total on 1,000 samples from the heterogeneous simulated population when C configuration

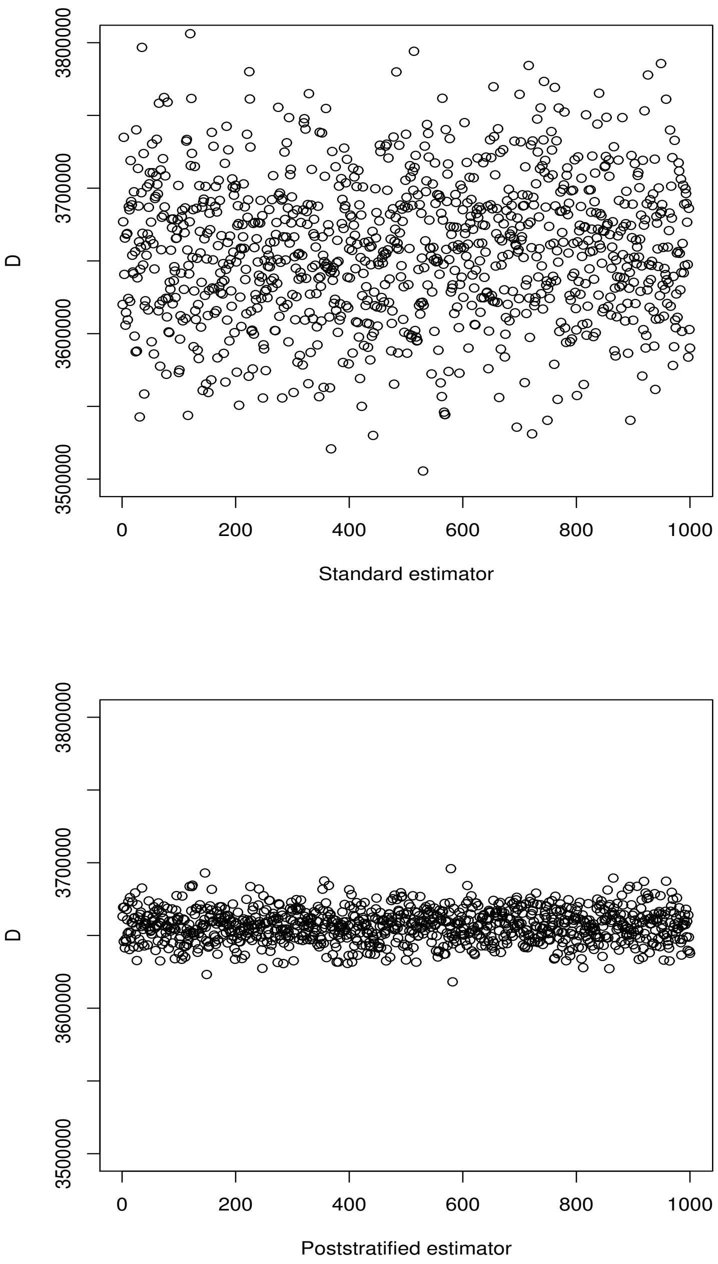


Figure 18: Point estimates of population total on 1,000 samples from the heterogeneous simulated population when D configuration

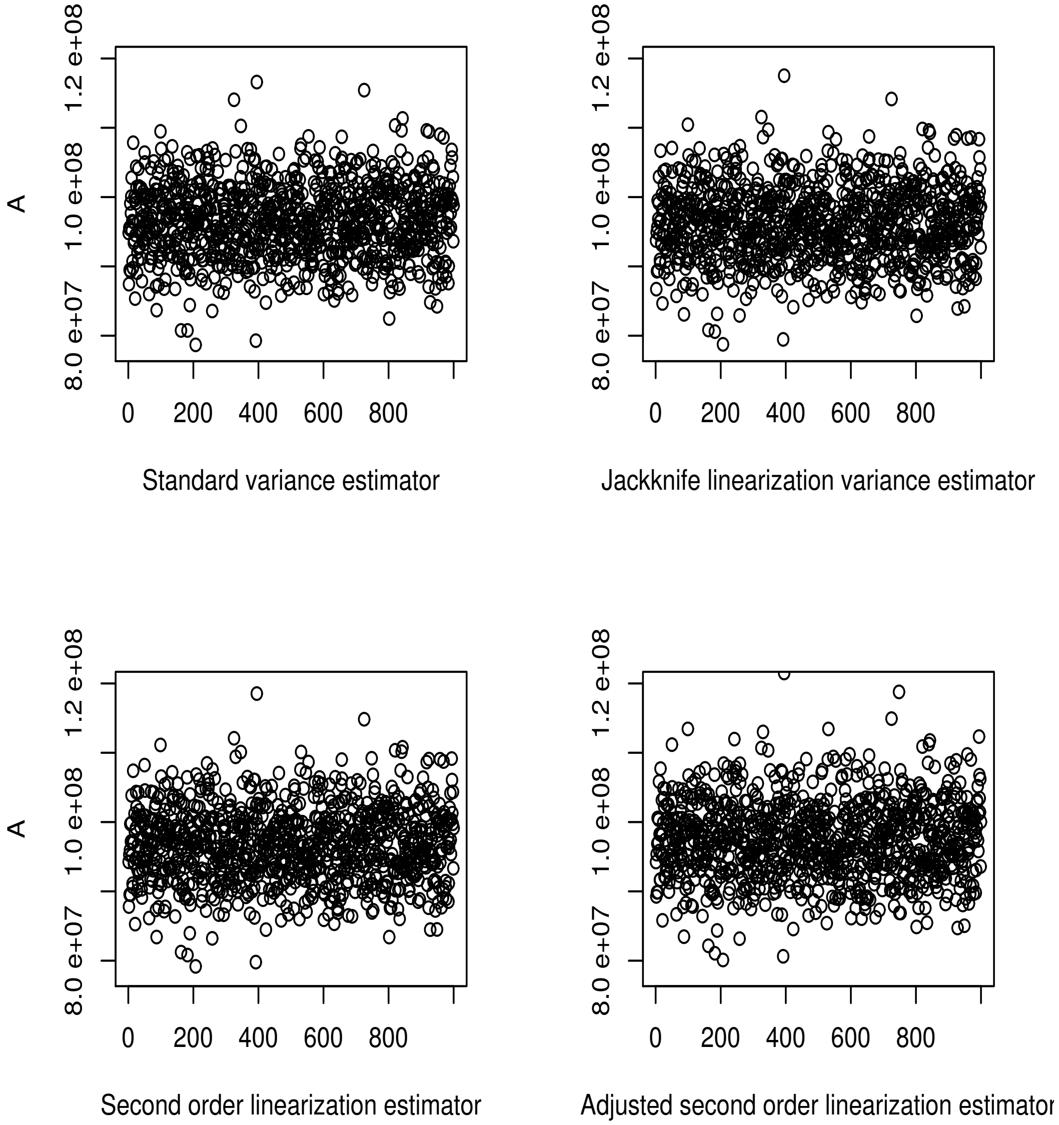


Figure 19: Variance estimates of population total on 1,000 samples from the homogeneous simulated population when A configuration

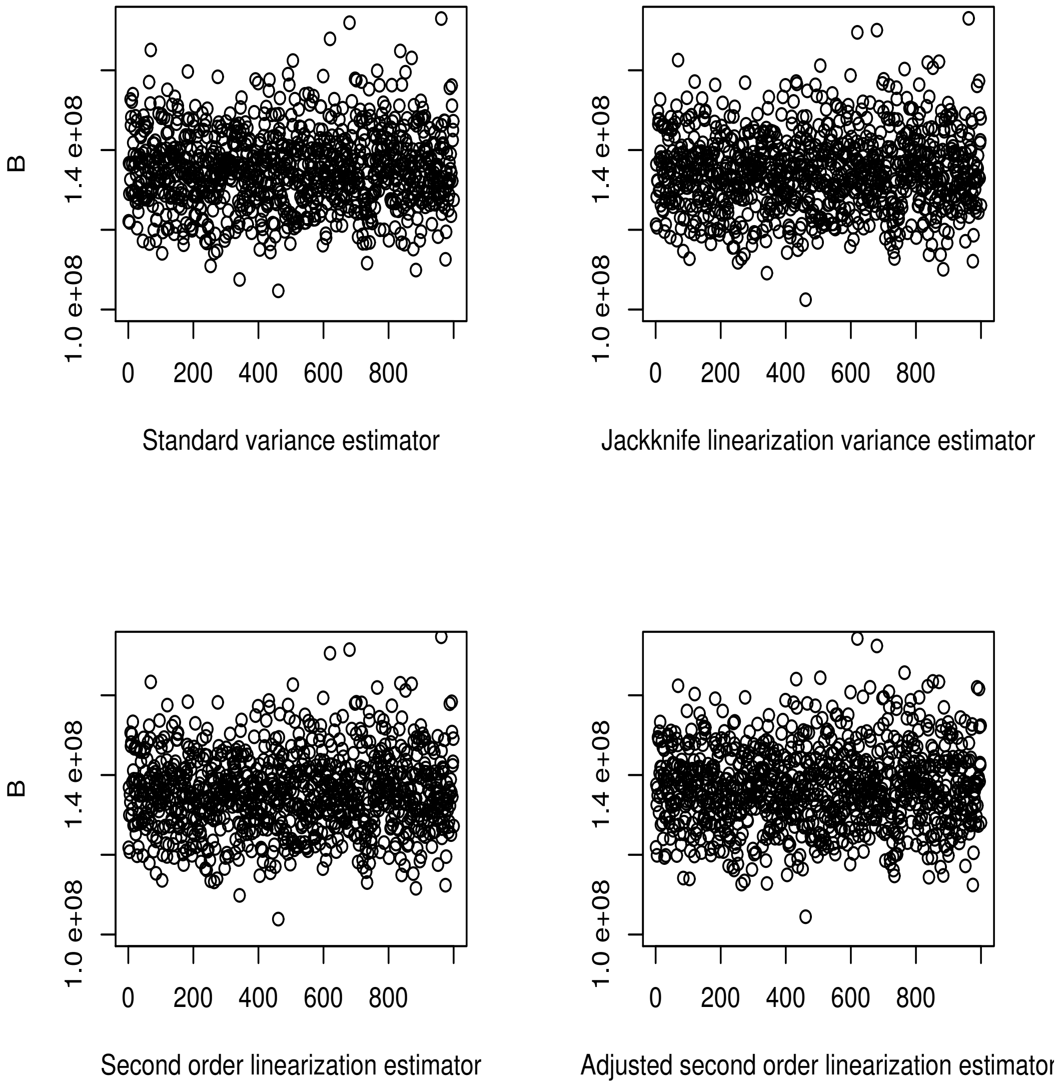


Figure 20: Variance estimates of population total on 1,000 samples from the homogeneous simulated population when B configuration

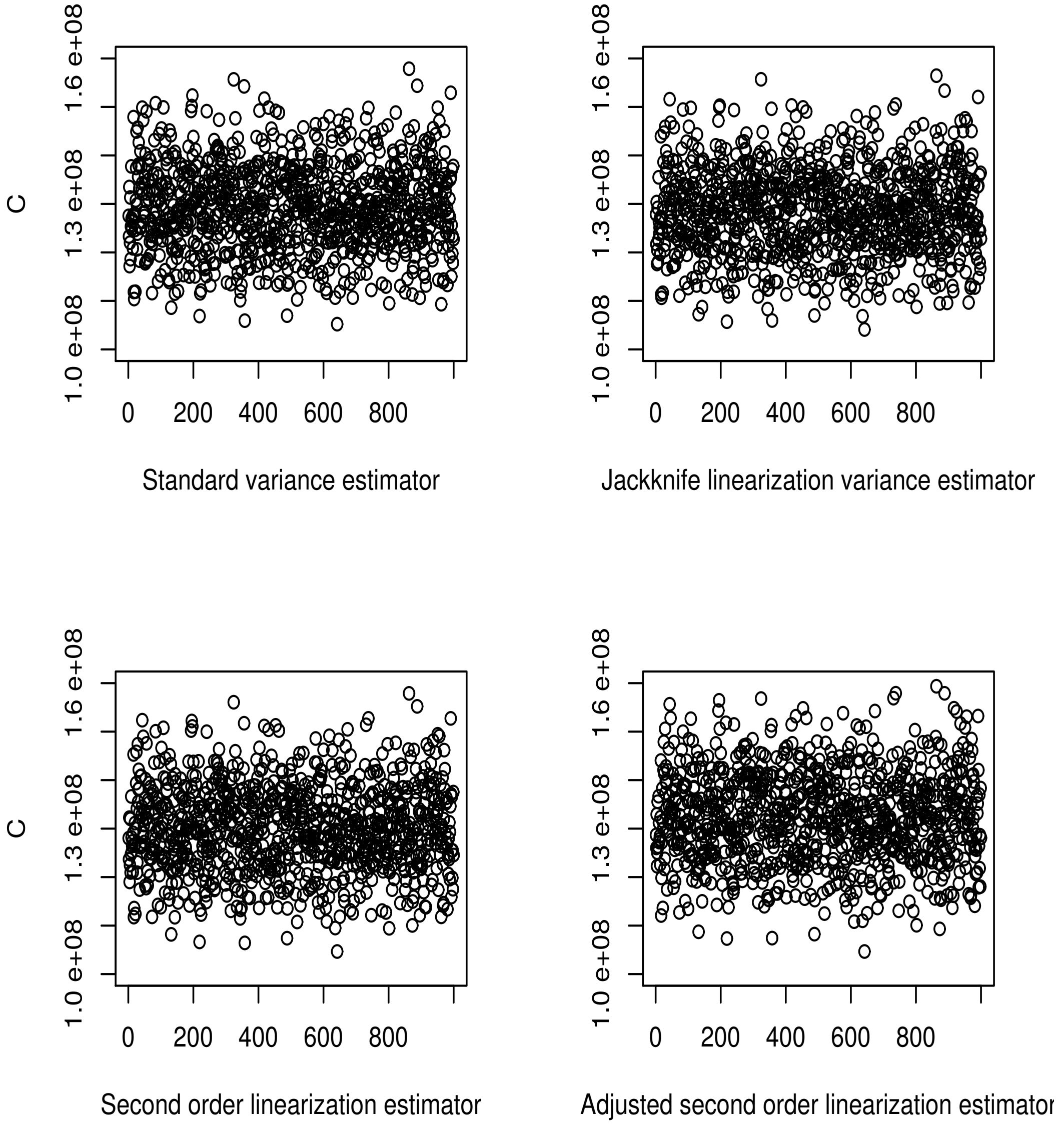


Figure 21: Variance estimates of population total on 1,000 samples from the heterogeneous simulated population when C configuration

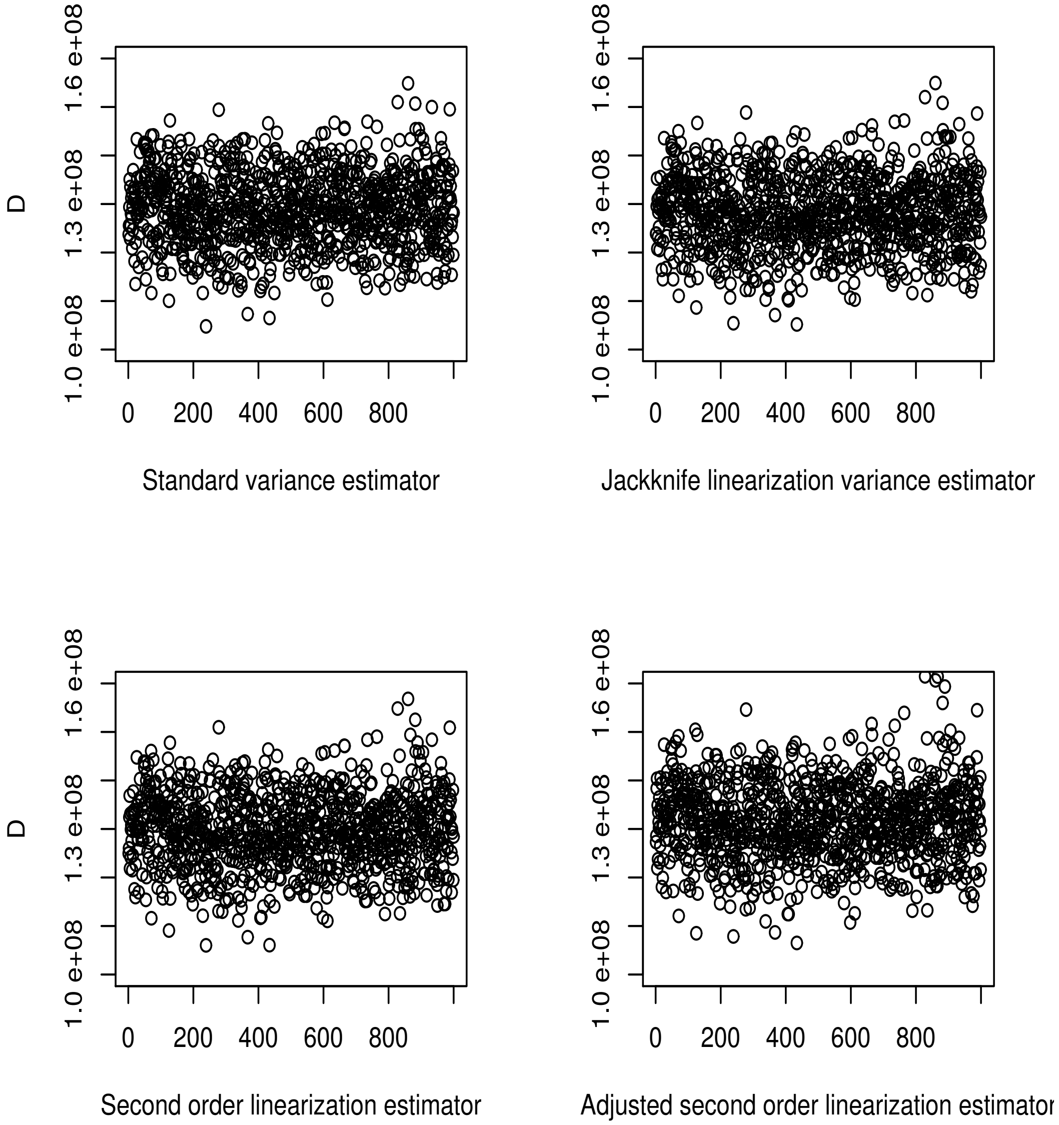


Figure 22: Variance estimates of population total on 1,000 samples from the heterogeneous simulated population when D configuration

CHAPTER IX

CONCLUDING REMARKS

Yung and Rao (1996) studied the jackknife variance estimator, the jackknife linearization estimator and the standard linearization estimator through a simulation study based on the finite population of 10,841 persons included in the september 1988 Current Population Survey (CPS). Their results showed that these methods estimate the mean squared error of the poststratified estimator well unconditionally. v_{JL} is more preferred to v_J for its simplicity and it also performs as well as v_J. However, v_{JL} and v_L has the disadvantage that both need the separate formula derivations. The second-order linearization variance estimator which we proposed in this paper, is as simple to be computed as the standard linearization estimator. Its performance is as good as v_L and v_{JL} based on the simulation result in both generated populations and real population. As for the estimating variance of the poststratified population total estimator, the second-order linearization estimator is preferred because it is computationally as simple as the standard linearization estimtor v_L, and the jackknife linearization estimator v_{JL}. Also theoretically, it might have more accuracy from the extension to the second order of the Taylor linearization.

The use of the second order Taylor linearization created the new adjustment factor for the poststrata weights and it is the function of R_c which also plays role of balancing weights for the poststrata in the standard estimator. This can also be applied to the linearization estimator of another type of generalized regression estimator which might be associated with different distance function other than the poststratified estimator for the population total that we considered here.

REFERENCES

Caslyn, C., Gonzales, P., and Frase, M. (1992). *Highlights from TIMSS*. Washington, D. C.: National Center for Education Statistics.

Deming, W. E. and Stephan, F. F. (1940). "On a least squares adjustment of a simpled frequency table when the expected marginal totals are known." *Annals of Mathematical Statistics*, 11, 427–444.

Deville, J. C. and Sarndal, C. E. (1992). "Calibration estimation in survey sampling." *Journal of The American Statistical Association*, 87, 376–382.

Durbin, J. (1959). "A note on the application of Quenouille's method of bias reduction to the estimation of ratios." *Biometrika*, 46, 477–480.

Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of The American Statistical Association*, 47, 663–685.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York, Wiley.

TIMSS International Study Center (1996). *Third International Mathematics and Science Study: Technical Report, Volume 1 Design and Development*. Chestnut Hill, MA: Boston College.

Quenouille, M. (1949). "Approximation tests of correlation in time series." *Journal of The Royal Statistical Society B*, 11, 18–84.

Rao, J. N. K. (1985). “Conditional inference in survey sampling.” *Survey Methodology*, 11, 15–31.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, Springer.

Shao, J. and Tu, D. (1995). *The Jacknife and Bootstrap*. New York, Springer.

Tukey, J. (1958). “Bias and confidence in not quite large samples.” *Annals of Mathematical Statistics*, 29, 614.

Valliant, R. (1990). “Comparisons of variance estimators in stratified random and sytematic sampling.” *Journal of Official Statistics*, 6, 115–131.

——— (1993). “Poststratification and conditional variance estimation.” *Journal of The American Statistical Association*, 88, 89–96.

Yung, W. and Rao, J. N. K. (1996). “Jackknife linearization variance estimators under stratified multi-stage sampling.” *Survey Methodology*, 22, 23–31.

VITA

Kyongryun Kim was born in Taegu, Korea. He received a Bachelor of Science degree in civil engineering from Sungkyunkwan University in Seoul, Korea in 1995 and received a Bachelor of Science in mathematics from Michigan State University in 1998. He received a Master of Science degree in statistics from Michigan State University, in 2000. He continued his studies under the direction of Dr. Suojin Wang, and received a Doctor of Philosophy degree from Texas A&M University in May 2006. His permanent address is Woobang apartment 113-802, Yongheung dong, Buk-ku, Pohang, Kyungsangbuk-do, Korea.