CHAPTER 18

Variance Estimation: Complex Survey Designs

18.1 INTRODUCTION

A sampling design other than simple random sampling is known as a complex sampling design. Most real-life surveys are complex surveys, and for such surveys we often need to estimate nonlinear parametric functions such as the population ratio of the total of two characteristics, population coefficient of variation, population regression coefficient, and population correlation coefficient, among others. Variance estimation is essential for estimating the precision of the survey estimates, calculation of confidence intervals, determination of optimum sample sizes, and testing of hypotheses, among others. In particular, finding the optimum sample size is the key factor in the determination of the cost of a survey and the subsequent precision of estimates. In most situations unbiased estimators of the nonlinear parametric functions are not available. We get neither the exact expressions of the variance/mean square of the nonlinear estimators nor the exact unbiased estimators of the variance/mean square errors. In this section we will discuss a few popular methods of approximate variance estimation of nonlinear estimators that can be used for complex survey designs. The methods include (i) linearization method (LR), (ii) random group method (RG), (iii) jackknife method (JK), (iv) balanced repeated replication (BRR), (v) bootstrap method (BT), and (vi) generalized variance functions (GVF). By and large, the proposed estimators are approximately unbiased and consistent. So, in our present discussion we will use the terms variance and mean square error in the same sense.

18.2 LINEARIZATION METHOD

Let $\theta_1,\ldots,\theta_k$ be k parametric functions of a finite population parameter $\mathbf{y}=(y_1,\ldots,y_N)$ and $\widehat{\theta}_1(s)=\widehat{\theta}_1,\ldots,\widehat{\theta}_k(s)=\widehat{\theta}_k$ be the estimators of $\theta_1,\ldots,\theta_k$ based on a sample s of size n selected with probability p(s) according to some suitable sampling design. In most situations, $\widehat{\theta}_j$'s are either unbiased or

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00018-2

consistent estimators of θ_j's. Suppose we want to estimate a parametric function $f(\boldsymbol{\theta}) = f(\theta_1,\ldots,\theta_k)$. In this situation we generally choose $T(s) = T = f\left(\widehat{\theta}_1,\ldots,\widehat{\theta}_k\right)$ as an estimator of $f(\boldsymbol{\theta})$. In case T is a differentiable function of $\widehat{\theta}_j$'s, for $j = 1,\ldots,k$, we can expand $T = f\left(\widehat{\theta}_1,\ldots,\widehat{\theta}_k\right)$ around $\theta_1,\ldots,\theta_k$ by Taylor's theorem as follows:

$$T = f\left(\widehat{\theta}_1,\ldots,\widehat{\theta}_k\right)\Big|_{\hat{\theta}_j=E\left(\hat{\theta}_j\right)} + \sum_{j=1}^{k}\left(\widehat{\theta}_j - \theta_j\right)\frac{\partial f\left(\widehat{\theta}_1,\ldots,\widehat{\theta}_k\right)}{\partial\widehat{\theta}_j}\Bigg|_{\hat{\theta}_j=E\left(\hat{\theta}_j\right)} + R_2 \tag{18.2.1}$$

where R_2 is the error or remainder term.

Letting $\widehat{\theta}_j$ be an unbiased or consistent estimator for θ_j for $j = 1,\ldots,k$, and assuming R_2 is small, at least for large n, we approximate

$$T \cong f(\theta_1,\ldots,\theta_k) + \sum_{j=1}^{k}\omega_j\left(\widehat{\theta}_j - \widehat{\theta}_j\right) \tag{18.2.2}$$

where $\omega_j = \dfrac{\partial f\left(\hat{\theta}_1,\ldots,\hat{\theta}_k\right)}{\partial\hat{\theta}_j}\Bigg|_{\hat{\theta}_j=\theta_j}$.

From expression (18.2.2), we get the following:

Theorem 18.2.1

(i) $E(T) \cong f(\theta_1,\ldots,\theta_k)$

(ii) Variance of $T = V(T) \cong \sum\limits_{j=1}^{k}\omega_j^2 V\left(\widehat{\theta}_j\right) + \sum\limits_{i\neq}^{k}\sum\limits_{j=1}^{k}\omega_j\omega_k Cov\left(\widehat{\theta}_j,\widehat{\theta}_k\right)$

$$= \boldsymbol{\omega}'\boldsymbol{\Lambda}\boldsymbol{\omega}$$

(iii) An approximate unbiased estimator of $V(T)$ is

$$\widehat{V}(T) \cong \sum_{j=1}^{k}\widehat{\omega}_j^2\widehat{V}\left(\widehat{\theta}_j\right) + \sum_{i\neq}^{k}\sum_{j=1}^{k}\widehat{\omega}_j\widehat{\omega}_k C\widehat{o}v\left(\widehat{\theta}_j,\widehat{\theta}_k\right)$$
$$= \widehat{\boldsymbol{\omega}}'\widehat{\boldsymbol{\Lambda}}\widehat{\omega}$$

where $\boldsymbol{\omega}' = (\omega_1,\ldots,\omega_k)$, $\boldsymbol{\Lambda} =$ variance covariance matrix of $\widehat{\theta}_j$'s, $\widehat{\boldsymbol{\omega}}' = (\widehat{\omega}_1,\ldots,\widehat{\omega}_k) =$ unbiased or approximately unbiased estimator of $\boldsymbol{\omega}'$, $\widehat{V}\left(\widehat{\theta}_j\right)$, $C\widehat{o}v\left(\widehat{\theta}_j, \widehat{\theta}_k\right)$ are unbiased estimators of $V\left(\widehat{\theta}_j\right)$ and $Cov\left(\widehat{\theta}_j, \widehat{\theta}_k\right)$, respectively, and $\widehat{\boldsymbol{\Lambda}}$ is an unbiased estimator of $\boldsymbol{\Lambda}$.

In case $\widehat{\theta}_j = \sum\limits_{i\in s} b_{si}\xi_i(j)$ is a linear homogeneous unbiased estimator of $\theta_j = \sum\limits_{i\in U} \xi_i(j)$ for $j = 1,\ldots,k$ and b_{si} are constants free from $\xi_i(j)$'s satisfy $\sum\limits_{s\supset i} b_{si} = 1$ and $b_{si} = 0$ for $i \notin s$, we can write following Woodruff (1971)

$$\begin{aligned} V(T) &\cong V\left(\sum_{j=1}^{k}\omega_j\widehat{\theta}_j\right) \\ &= V\left[\sum_{i\in s}\left(\sum_{j=1}^{k}\omega_j b_{si}\right)\xi_i(j)\right] \\ &= V\left(\sum_{i\in s} b_{si}h_i\right) \text{ with } h_i = \sum_{j=1}^{k}\omega_j\xi_i(j) \\ &= \sum_{i\in U}\beta_{ii}h_i^2 + \sum_{i\neq}\sum_{j\in U}\beta_{ij}h_ih_j \end{aligned} \tag{18.2.3}$$

where $\beta_{ii} = \sum\limits_{s\supset i} b_{si}^2 p(s) - 1$ and $\beta_{ij} = \sum\limits_{s\supset i,j} b_{si}b_{sj}p(s) - 1$.

An approximate unbiased estimator of $V(T)$ is

$$\widehat{V}_{LR}(T) = \sum_{i\in s}\frac{\beta_{ii}}{\pi_i}\widehat{h}_i^2 + \sum_{i\neq}\sum_{j\in s}\frac{\beta_{ij}}{\pi_{ij}}\widehat{h}_i\widehat{h}_j \tag{18.2.4}$$

where $\widehat{h}_i$ is a suitable unbiased or consistent estimator of h_i.

The performance of $\widehat{V}_{LR}(T)$ obviously depends on the validity of the expansion of the Taylor series (18.2.2). The inclusion of second- and higher-order terms of the expansion of the Taylor series will certainly increase the performance of the variance estimator but at the same time yield a more complex variance formula. Wolter (1985) pointed out that the LR method provides efficient variance estimators in complex surveys when the sample size is fairly large. However, the method may provide unreliable estimators if the Taylor series is not convergent or the population is highly skewed. The LR technique fails for the parameter that cannot be expressed as a simple function of the population total, e.g., population median.

18.2.1 Ratio Estimator

The ratio estimator for the population ratio $R = Y/X$ is given by

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}}$$

where $\widehat{Y} = \sum\limits_{i\in s} b_{si}y_i$ and $\widehat{X} = \sum\limits_{i\in s} b_{si}x_i$ are unbiased estimators of $Y = \sum\limits_{i=1}^{N} y_i$ and $X = \sum\limits_{i=1}^{N} x_i$, respectively, based on a sample s with probability $p(s)$. The coefficients b_{si}'s satisfy $\sum\limits_{s\supset i} b_{si}p(s) = 1$.

Taking $y_i = \xi_i(1)$, $x_i = \xi_i(2)$, $\theta_1 = Y = \sum\limits_{i\in U}\xi_i(1)$, $\theta_2 = X = \sum\limits_{i\in U}\xi_i(2)$, $\widehat{\theta}_1 = \widehat{Y}$, $\widehat{\theta}_2 = \widehat{X}$, and $f(\theta_1,\theta_2) = \dfrac{\theta_1}{\theta_2} = R$, we can write

$$\widehat{R} = T = f\left(\widehat{\theta}_1,\widehat{\theta}_2\right) = \frac{\widehat{\theta}_1}{\widehat{\theta}_2} = f(\theta_1,\theta_2) + \omega_1\left(\widehat{\theta}_1 - \theta_1\right) + \omega_2\left(\widehat{\theta}_2 - \theta_2\right) \tag{18.2.5}$$

where

$$\omega_1 = \left.\frac{\partial T}{\partial\hat{\theta}_1}\right|_{\hat{\theta}_1=\theta_1,\hat{\theta}_2=\theta_2} = \frac{1}{\theta_2} \text{ and } \omega_2 = \left.\frac{\partial T}{\partial\hat{\theta}_2}\right|_{\hat{\theta}_1=\theta_1,\hat{\theta}_2=\hat{\theta}_2} = -\left.\frac{\hat{\theta}_1}{\hat{\theta}_2^2}\right|_{\hat{\theta}_1=\theta_1,\hat{\theta}_2=\theta_2} = -\frac{\theta_1}{\theta_2^2}.$$

Writing $h_i = \sum\limits_{j=1}^{2}\omega_j\xi_i(j) = \dfrac{1}{X}(y_i - Rx_i)$

we get $\widehat{R} \cong R + \dfrac{1}{X}\sum\limits_{i\in s} b_{si}(y_i - Rx_i)$

and

$$V\left(\widehat{R}\right) \cong \frac{1}{X^2}\left[\sum_{i\in U}\beta_{ii}(y_i - Rx_i)^2 + \sum_{i\neq}\sum_{j\in U}\beta_{ij}(y_i - Rx_i)(y_j - Rx_j)\right] \tag{18.2.6}$$

For a fixed sample size n design with $b_{si} = 1/\pi_i$, the expression (18.2.6) reduces to

$$V\left(\widehat{R}\right) \cong \frac{1}{2}\frac{1}{X^2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{d_i}{\pi_i} - \frac{d_j}{\pi_j}\right)^2 \tag{18.2.7}$$

where $d_i = y_i - Rx_i$.

An approximate unbiased estimator of $V\left(\widehat{R}\right)$ is

$$\widehat{V}_{LR}\left(\widehat{R}\right) \cong \frac{1}{2}\frac{1}{X^2}\sum_{i\neq}\sum_{j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{\widehat{d}_i}{\pi_i} - \frac{\widehat{d}_j}{\pi_j}\right)^2 \tag{18.2.8}$$

with $\widehat{d}_i = y_i - \widehat{R}x_i$.

For a simple random sampling without replacement (SRSWOR) sampling design with $b_{si} = N/n$. Hence the expressions (18.2.7) and (18.2.8) become, respectively,

$$V\left(\widehat{R}\right) \cong (1-f)\frac{1}{n\overline{X}^2}S_d^2 \text{ and } \widehat{V}_{LR}\left(\widehat{R}\right) \cong (1-f)\frac{1}{n\overline{X}^2}s_{\hat{d}}^2$$

where $f = n/N$, $(N-1)S_d^2 = \sum\limits_{i\in U}\left(d_i - \overline{D}\right)^2$, $\overline{D} = \sum\limits_{i\in U} d_i/N$,
$(n-1)s_{\hat{d}}^2 = \sum\limits_{i\in s}\left(\widehat{d}_i - \overline{\widehat{d}}\right)^2$, and $\overline{\widehat{d}} = \sum\limits_{i\in s}\widehat{d}_i\Big/n$.

18.2.2 Coefficient of Variation

The population coefficient of variation of a finite population is defined as

$$\theta = \frac{\sqrt{\sum\limits_{i\in U}\left(y_i - \overline{Y}\right)^2/N}}{\overline{Y}} = \frac{\sqrt{N\sum\limits_{i\in U}y_i^2 - \left(\sum\limits_{i\in U}y_i\right)^2}}{\sum\limits_{i\in U}y_i} = \frac{\sqrt{\theta_0\theta_2 - \theta_1^2}}{\theta_1}$$

$$= f(\theta_0, \theta_1, \theta_2)$$

where $\theta_j = \sum\limits_{i\in U} y_i^j$ for $j = 0,\ 1,\ 2$.

Let us choose $\widehat{\theta} = f\left(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2\right)$ with $\widehat{\theta}_j = \sum\limits_{i\in s} y_i^j\Big/\pi_i$. The variance of $\widehat{\theta}$ becomes

$$V\left(\widehat{\theta}\right) = \boldsymbol{\omega}'\boldsymbol{\Lambda}\boldsymbol{\omega}$$

where $\boldsymbol{\omega}' = (\omega_0, \omega_1, \omega_2)$, $\omega_0 = \left.\frac{\partial\hat{\theta}}{\partial\hat{\theta}_0}\right|_{\hat{\theta}_0=\theta_0,\hat{\theta}_1=\theta_1,\hat{\theta}_2=\theta_2} = \frac{1}{2\theta}\frac{\theta_2}{\theta_1^2}$,
$\omega_1 = \left.\frac{\partial\hat{\theta}}{\partial\hat{\theta}_1}\right|_{\hat{\theta}_0=\theta_0,\hat{\theta}_1=\theta_1,\hat{\theta}_2=\theta_2} = -\frac{\theta_0}{\theta}\frac{\theta_2}{\theta_1^3}$, $\omega_2 = \left.\frac{\partial\hat{\theta}}{\partial\hat{\theta}_2}\right|_{\hat{\theta}_0=\theta_0,\hat{\theta}_1=\theta_1,\hat{\theta}_2=\theta_2} = \frac{1}{2\theta}\frac{\theta_0}{\theta_1^2}$, and
$\boldsymbol{\Lambda} = (V_{jk})$ with

$$V_{jk} = Cov\left(\widehat{\theta}_j, \widehat{\theta}_k\right) = \sum_{r\in U}\left(\frac{1}{\pi_r} - 1\right)y_r^{j+k} + \sum_{r\neq}\sum_{t\in U}\left(\frac{\pi_{rt}}{\pi_r\pi_t} - 1\right)y_r^j y_t^k.$$

An approximate unbiased estimator of $V\left(\widehat{\theta}\right)$ is

$$\widehat{V}_{LR}\left(\widehat{\theta}\right) = \widehat{\boldsymbol{\omega}}'\widehat{\boldsymbol{\Lambda}}\widehat{\omega} \qquad (18.2.9)$$

where $\widehat{\boldsymbol{\omega}}' = (\widehat{\omega}_0, \widehat{\omega}_1, \widehat{\omega}_2)$, $\widehat{\omega}_0 = \frac{1}{2\hat{\theta}} \frac{\hat{\theta}_2}{\hat{\theta}_1^2}$, $\widehat{\omega}_1 = -\frac{\hat{\theta}_0}{\hat{\theta}} \frac{\hat{\theta}_2}{\hat{\theta}_1^3}$, $\widehat{\omega}_2 = \frac{1}{2\hat{\theta}} \frac{\hat{\theta}_0}{\hat{\theta}_1^2}$,

$\widehat{\Lambda} = \left(\widehat{V}_{jk}\right)$, and $\widehat{V}_{jk} = \sum_{r \in s} \left(\frac{1}{\pi_r} - 1\right) \frac{y_r^{j+k}}{\pi_r} + \sum_{r \neq} \sum_{t \in s} \left(\frac{\pi_{rt}}{\pi_r \pi_t} - 1\right) \frac{y_r^j y_t^k}{\pi_{rt}}$.

18.3 RANDOM GROUP METHOD

The pioneering work of the RG method was due to Mahalanobis (1946) who proposed the interpenetrating network of subsampling procedure (IPNS). The main objective of IPNS was to maintain the quality of the data collected by different investigators by comparing the responses obtained from those units that are common in two or more samples. In this method k independent samples of the same size are selected from the population using the same sampling design. For $k = 2$, the method is called half sampling. Let $\widehat{\theta}_{(i)}$ be an unbiased estimator of the parameter θ obtained from the ith sample $i = 1,\ldots,k$ and $\widehat{\overline{\theta}} = \sum_{i=1}^{k} \widehat{\theta}_{(i)} \Big/ k$ be a pooled estimator of θ, then we can easily verify the following theorem.

Theorem 18.3.1

(i) $E\left(\widehat{\overline{\theta}}\right) = \theta$

(ii) $V\left(\widehat{\overline{\theta}}\right) = \sigma^2 \Big/ k$

(iii) $\widehat{V}\left(\widehat{\overline{\theta}}\right) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\widehat{\theta}_{(i)} - \widehat{\overline{\theta}}\right)^2$

where $\widehat{V}\left(\widehat{\overline{\theta}}\right)$ is an unbiased estimator for $V\left(\widehat{\overline{\theta}}\right)$ and $\sigma^2 = V\left(\widehat{\theta}_{(i)}\right)$ for $i = 1,\ldots,k$.

This theorem may be generalized by selecting different samples independently by using different sampling procedures so that the $E\left(\widehat{\theta}_{(i)}\right)$'s are equal to θ, while the variances $V\left(\widehat{\theta}_{(i)}\right) = \sigma_i^2$ may vary. In this situation we also get the following theorem, which is similar to the previous one.

Theorem 18.3.2

(i) $E\left(\widehat{\overline{\theta}}\right) = \theta$

(ii) $V\left(\widehat{\overline{\theta}}\right) = \sum_{i=1}^{k} \sigma_i^2 / k^2$

(iii) $\widehat{V}\left(\widehat{\overline{\theta}}\right) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\widehat{\theta}_{(i)} - \widehat{\overline{\theta}}\right)^2$

Thus if the samples are selected independently by any sampling procedure, an unbiased estimator of the variance of the estimator of the parameter θ may be obtained easily by applying the aforementioned theorems. But in practice, the sample is selected once and unbiased estimators of the parameters of interest (such as population ratio, regression coefficient, coefficient of variance, etc.) cannot be obtained easily. In this situation we divide the original sample s of size n at random into k disjoint groups. The ith RG s_i is a random subsample of size $m = n/k$ (assuming m is an integer) selected from the original sample s, by SRSWOR method. Let $\widehat{\theta}$ be an estimator of θ based on the original sample s, $\widehat{\theta}_{(i)}$ estimator of θ based on the sample s_i, and $\widehat{\widehat{\theta}} = \sum\limits_{i=1}^{k} \widehat{\theta}_{(i)} \Big/ k$ be a combined estimator of θ. Then the variance of $\widehat{\theta}$ can be estimated by any of the following formulae.

$$\widehat{V}_{RG1}\left(\widehat{\theta}\right) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\widehat{\theta}_{(i)} - \widehat{\widehat{\theta}}\right)^2 \tag{18.3.1}$$

$$\widehat{V}_{RG2}\left(\widehat{\theta}\right) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\widehat{\theta}_{(i)} - \widehat{\theta}\right)^2 \tag{18.3.2}$$

None of the aforementioned estimators are unbiased for $V\left(\widehat{\theta}\right)$. The estimator $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$ is conservative in the sense that it has an upward bias. If $\widehat{\theta}$ and $\widehat{\theta}_{(i)}$ are linear unbiased estimators of θ, then $\widehat{\widehat{\theta}} = \widehat{\theta}$. But, if $\widehat{\theta}$ and $\widehat{\theta}_{(i)}$ are nonlinear estimators of θ such as the population ratio $R = Y/X$, then $\widehat{\theta} = \widehat{Y}\Big/\widehat{X}$ is not, in general, equal to $\widehat{\widehat{\theta}} = \frac{1}{k}\sum\limits_{i=1}^{K}\widehat{Y}_i/\widehat{X}_i$, where $\widehat{Y}, \widehat{X}$ and $\widehat{Y}_i$, $\widehat{X}_i$ are unbiased estimators for totals of Y and X, respectively, based on the samples s and s_i. As per the magnitude of $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ and $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$, we note that $\widehat{V}_{RG1}\left(\widehat{\theta}\right) \leq \widehat{V}_{RG2}\left(\widehat{\theta}\right)$ because $\widehat{V}_{RG2}\left(\widehat{\theta}\right) = \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(\widehat{\theta}_{(i)} - \widehat{\theta}\right)^2 = \widehat{V}_{RG1}\left(\widehat{\theta}\right) + \frac{\left(\hat{\bar{\theta}} - \hat{\theta}\right)^2}{k-1}$. However, for large sample size $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ and $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$ are approximately equal.

18.3.1 Simple Random Sampling With Replacement

Let $\widehat{\theta} = \overline{y}(s)$ be the sample mean based on a sample s of size n ($=mk$) selected by the simple random sampling with replacement (SRSWR) method. The sample s is divided into k groups at random and $\widehat{\theta}_{(i)}=\overline{y}(s_i)$ is the sample mean of the ith group s_i, $i=1,\ldots,k$. Here, $\widehat{\theta}$, $\widehat{\theta}_{(i)}$, and the combined estimator $\widehat{\overline{\theta}}=\sum_{i=1}^{k}\widehat{\theta}_{(i)}\Big/k$ are all unbiased estimators for the population mean $\overline{Y}=\sum_{i\in U}y_i/N$. Furthermore, $\widehat{\theta}=\overline{y}(s)=\widehat{\overline{\theta}}=\sum_{i=1}^{k}\overline{y}(s_i)/k$, $\widehat{V}_{RG1}\left(\widehat{\theta}\right)=\widehat{V}_{RG2}\left(\widehat{\theta}\right)=\dfrac{1}{k(k-1)}\sum_{i=1}^{k}\{\overline{y}(s_i)-\overline{y}(s)\}^2$, and

$$E\left\{\widehat{V}_{RG1}\left(\widehat{\theta}\right)\right\}=\frac{1}{k(k-1)}E\left(\sum_{i=1}^{k}\widehat{\theta}_{(i)}^{2}-k\widehat{\theta}^{2}\right)=\frac{V\left(\widehat{\theta}_{RG1}\right)-V\left(\widehat{\theta}\right)}{(k-1)}$$

$$=\frac{\sigma^2}{n}=V\left(\overline{y}_s\right)$$

where $\sigma^2=\sum_{i\in U}\left(y_i-\overline{Y}\right)^2\Big/N$. Hence $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ (i.e., $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$) is unbiased for $\dfrac{\sigma^2}{n}=V\left(\widehat{\theta}\right)$. It is important to note that $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ is different from the traditional unbiased estimator $\widehat{V}\left(\overline{y}_s\right)=s_y^2\Big/n$, where $s_y^2=\sum_{i\in s}\{y_i-\overline{y}(s)\}^2\Big/(n-1)$.

18.3.2 Simple Random Sampling Without Replacement

Let a sample s of size n be selected by SRSWOR method and be divided at random into k groups of m units each ($n=mk$). The estimator for $\overline{Y}$ based on the sample s and s_i are the sample means $\widehat{\theta}=\overline{y}(s)$ and $\widehat{\theta}_{(i)}=\overline{y}(s_i)$, respectively. In this case the combined estimator of θ is $\widehat{\theta}=\overline{y}(s)=\widehat{\overline{\theta}}=\sum_{i=1}^{k}\widehat{\theta}_{(i)}\Big/k=\sum_{i=1}^{k}\overline{y}(s_i)/k$. Each of the estimated variances $\widehat{V}_{RG1}\left(\widehat{\overline{\theta}}\right)$ and $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$ of $\widehat{\theta}$ is equal to $\dfrac{1}{k(k-1)}\sum_{i=1}^{k}\{\overline{y}(s_i)-\overline{y}(s)\}^2$ and

$$E\left\{\widehat{V}_{RG1}\left(\widehat{\theta}\right)\right\}=E\left\{\widehat{V}_{RG2}\left(\widehat{\theta}\right)\right\}=\frac{V\left(\widehat{\theta}_{(i)}\right)-V\left(\widehat{\theta}\right)}{(k-1)}$$

$$=\frac{1}{k-1}\left[\left(\frac{1}{m}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{N}\right)\right]S_y^2$$

$$=\frac{S_y^2}{n}$$

The estimator $\widehat{V}_{RG1}\left(\widehat{\theta}\right)\left(=\widehat{V}_{RG2}\left(\widehat{\theta}\right)\right)$ is therefore biased for $V\left(\widehat{\theta}\right)$ and the amount of bias is $E\left\{\widehat{V}_{RG1}\left(\widehat{\theta}\right)-V\left(\widehat{\theta}\right)\right\}=S_y^2\Big/N$. Thus the estimator $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ (i.e., $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$) overestimates the variance of $V\left(\widehat{\theta}\right)$. The amount of bias of $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ is obviously negligible if N is large.

18.3.3 Varying Probability Sampling

Let a sample s of size n be selected by the varying probability sampling scheme with positive inclusion probability for the ith unit π_i for $i=1,\ldots,N$. The Horvitz–Thompson estimator for the population total Y based on the sample s is $\widehat{\theta}=\sum_{i\in s}y_i/\pi_i$. The sample s is divided at random into k groups each of size m. The estimator for Y based on the ith group is $\widehat{\theta}_{(i)}=\frac{n}{m}\sum_{i\in s_i}y_i/\pi_i$ for $i=1,\ldots,k$ and the combined estimator is

$$\widehat{\widehat{\theta}}=\sum_{i=1}^{k}\widehat{\theta}_{(i)}\Big/k=\sum_{i=1}^{k}\sum_{i\in s_i}y_i/\pi_i=\sum_{i\in s}y_i/\pi_i=\widehat{\theta}.$$

Here $\widehat{V}_{RG1}\left(\widehat{\theta}\right)=\widehat{V}_{RG2}\left(\widehat{\theta}\right)=\dfrac{1}{k(k-1)}\sum_{i=1}^{k}\left(\dfrac{n}{m}\sum_{i\in s_i}y_i/\pi_i-\sum_{i\in s}y_i/\pi_i\right)^2$ and

$$\begin{aligned}
E\left[\widehat{V}_{RG1}\left(\widehat{\theta}\right)\right]&=\frac{1}{(k-1)}\left[V\left(\frac{n}{m}\sum_{i\in s_i}y_i/\pi_i\right)-V\left(\sum_{i\in s}y_i/\pi_i\right)\right]\\
&=\frac{1}{(k-1)}\left[E\left\{V\left(\frac{n}{m}\sum_{i\in s_i}y_i/\pi_i|s\right)\right\}\right.\\
&\qquad\left.+V\left\{E\left(\frac{n}{m}\sum_{i\in s_i}y_i/\pi_i|s\right)\right\}-V\left(\sum_{i\in s}y_i/\pi_i\right)\right]\\
&=\frac{1}{(k-1)}E\left[V\left(\frac{n}{m}\sum_{i\in s_i}y_i/\pi_i|s\right)\right]\\
&=\frac{1}{(k-1)}E\left[n^2\left(\frac{1}{m}-\frac{1}{n}\right)\frac{1}{n-1}\left\{\sum_{i\in s}y_i^2/\pi_i^2-\frac{1}{n}\left(\sum_{i\in s}y_i/\pi_i\right)^2\right\}\right]\\
&=\frac{n}{n-1}\left[\sum_{i\in U}y_i^2/\pi_i-\frac{Y^2}{n}-\frac{1}{n}V\left(\sum_{i\in s}y_i/\pi_i\right)\right]
\end{aligned}$$

The amount of bias of the estimator $\widehat{V}_{RG1}\left(\widehat{\theta}\right)$ for estimation of $V\left(\widehat{\theta}\right)$ is given by

$$\begin{aligned} B\left[\widehat{V}_{RG1}\left(\widehat{\overline{\theta}}\right)\right] &= E\left[\widehat{V}_{RG1}\left(\widehat{\overline{\theta}}\right)\right] - V\left(\widehat{\theta}\right) \\ &= \frac{n}{n-1}\left[\sum_{i\in U} y_i^2/\pi_i - \frac{Y^2}{n} - V\left(\sum_{i\in s} y_i/\pi_i\right)\right] \end{aligned} \tag{18.3.3}$$

For an inclusion probability proportional to size (IPPS or πps) sampling design $\pi_i = np_i$, the expression of bias (18.3.3) reduces to

$$B\left[\widehat{V}_{RG1}\left(\widehat{\overline{\theta}}\right)\right] = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2 - V\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)\right] \tag{18.3.4}$$

The expression $\frac{1}{n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2$ is the variance of the Hansen–Hurwitz estimator based on a sample of size n selected by probability proportional to size with replacement sampling (PPSWR) sampling scheme, which is expected to be higher than the variance of the Horvitz–Thompson estimator based on a πps design of the same sample size n. Hence for a πps sampling design $\widehat{V}_{RG1}\left(\widehat{\overline{\theta}}\right)$, i.e., $\widehat{V}_{RG2}\left(\widehat{\theta}\right)$ overestimates the variance $V\left(\widehat{\theta}\right)$ in general.

18.3.4 Multistage Sampling

In the BAIS II survey (Botswana HIV/AIDS Impact Survey II) each district was divided into a number of enumeration areas. At first, a sample of s of n enumeration areas of a district is selected from a total of N enumeration areas by the IPPS (or πps) sampling scheme using Goodman and Kish (1950) sampling taking number of households x_i as measure of size for the ith enumeration area. Here, $\pi_i = np_i$, $\left(p_i = x_i/X, X = \sum_{i\in U} x_i\right)$ is the inclusion probability of the ith enumeration area. If the ith enumeration area is included in the sample, a subsample s_i of m_i households is selected from the M_i households by a systematic sampling procedure. Let y_{ij} be the value of the variate of interest for the jth household of the ith enumeration area, then the population total is $Y = \sum_{i\in U} Y_i$, where $Y_i = \sum_{j=1}^{M_i} y_{ij}$. An unbiased estimator for the population total Y is

$$\widehat{\theta} = \sum_{i\in s} \frac{\widehat{Y}_i}{\pi_i} \tag{18.3.5}$$

where $\widehat{Y}_i = M_i \overline{y}(s_i)$ and $\overline{y}(s_i) = \sum_{j\in s_i} y_{ij}/m_i$. The variance of $\widehat{Y}$ is

$$\begin{aligned} V\left(\widehat{\theta}\right) &= V\left[\left(E \sum_{i\in s} \frac{\widehat{Y}_i}{\pi_i}\middle| s\right)\right] + E\left[V\left(\sum_{i\in s} \frac{\widehat{Y}_i}{\pi_i}\middle| s\right)\right] \\ &= V\left(\sum_{i\in s} \frac{Y_i}{\pi_i}\right) + E\left(\sum_{i\in s} \frac{\sigma_i^2}{\pi_i^2}\right) \end{aligned} \tag{18.3.6}$$

$\left(\text{where } \sigma_i^2 \text{ is the variance of } \widehat{Y}_i \text{ given } s\right).$

$$= \frac{1}{2} \sum_{i\neq} \sum_{j\in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i\in U} \frac{\sigma_i^2}{\pi_i}$$

The variance $V\left(\widehat{\theta}\right)$ cannot be estimated unbiasedly because an unbiased estimator of the variance σ_i^2, based on a single systematic sample, is not available. However, if we divide the sample into n groups, taking only one enumeration unit in a group and $\widehat{\theta}_{(i)} = \frac{\hat{Y}_i}{p_i}$ be an estimator of the total Y based on the ith enumeration area, then the combined estimator for θ is

$$\widehat{\widehat{\theta}} = \frac{1}{n} \sum_{i\in s} \frac{\widehat{Y}_i}{p_i} = \widehat{\theta} \tag{18.3.7}$$

The RG variance estimator of $V\left(\widehat{\theta}\right)$ is given by

$$\widehat{V}_{RG1}\left(\widehat{\theta}\right) = \widehat{V}_{RG2}\left(\widehat{\theta}\right) = \frac{1}{n(n-1)} \sum_{i\in s} \left(\frac{\widehat{Y}_i}{p_i} - \widehat{\theta}\right)^2 \tag{18.3.8}$$

$$
\begin{aligned}
\text{The bias of } \widehat{V}_{RG1}\left(\widehat{\theta}\right) &= E\left[\widehat{V}_{RG1}\left(\widehat{\theta}\right)\right] - V\left(\widehat{\theta}\right)\\
&= \frac{1}{n(n-1)} E\left[\sum_{i\in s}\left(\frac{\widehat{Y}_i}{p_i}\right)^2 - n\widehat{\theta}^2\right] - V\left(\widehat{\theta}\right)\\
&= \frac{1}{n(n-1)}\left[E\left(\sum_{i\in s}\frac{Y_i^2+\sigma_i^2}{p_i^2}\right) - n\left(V\left(\widehat{\theta}\right)+Y^2\right)\right] - V\left(\widehat{\theta}\right)\\
&= \frac{1}{n(n-1)}\left[\sum_{i\in U}\frac{Y_i^2+\sigma_i^2}{p_i^2}\pi_i - n\left(V\left(\widehat{\theta}\right)+Y^2\right)\right] - V\left(\widehat{\theta}\right)\\
&= \frac{1}{(n-1)}\left[\left(\sum_{i\in U}\frac{Y_i^2}{p_i} - Y^2\right) + \sum_{i\in U}\frac{\sigma_i^2}{p_i} - V\left(\widehat{\theta}\right)\right]\\
&\qquad (\text{since } \pi_i = np_i)\\
&= \frac{n}{(n-1)}\left[\frac{1}{n}\left(\sum_{i\in U}\frac{Y_i^2}{p_i} - Y^2\right) - V\left(\sum_{i\in s}\frac{Y_i}{\pi_i}\right)\right]
\end{aligned}
\tag{18.3.9}
$$

Because $V\left(\sum_{i\in s}\frac{Y_i}{\pi_i}\right)$, the variance of the Horvitz–Thompson estimator based on a πps sampling scheme, is expected to be smaller than $\frac{1}{n}\left(\sum_{i\in U}\frac{Y_i^2}{p_i} - Y^2\right)$, the variance of the Hansen–Hurwitz estimator based on the PPSWR sampling scheme of the same size, the RG estimator $\widehat{V}_{RG1}\left(\widehat{\theta}\right)\left(\widehat{V}_{RG2}\left(\widehat{\theta}\right)\right)$ is expected to overestimate the variance.

18.3.5 Numerical Example

A sample *s* of size 15 is selected from a population of 50 households by the SRSWOR method. The medical expenditures (*y*) and family sizes (*x*) of households are given in the following Table 18.3.1. Our objective is to estimate the average medical consumption per household given that the average household size of the population is 3.5.

Table 18.3.1 Medical Expenditures and family size of the sampled households

Households	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
y	1500	6000	4500	4000	8000	6800	9750	8800
x	1	2	2	2	4	2	3	4
Households	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
y	7620	7500	4500	5000	6500	7500	4500	
x	3	3	4	3	3	4	2	

Suppose that the sample s is divided at random into three groups each of size 5 as follows; $s_1 = (1, 2, 3, 4, 5)$, $s_2 = (6, 7, 8, 9, 10)$, and $s_3 = (11, 12, 13, 14, 15)$.

(i) **Estimates without using auxiliary information**

The estimator of the population mean medical expenditure per household $= \theta = \overline{Y}$ is $\widehat{\theta} = \overline{y}(s) = \sum\limits_{i \in s} y_i/15 = 6164.667$ and an unbiased estimate of the variance $V\left(\widehat{\theta}\right)$ is $\widehat{V}\left(\widehat{\theta}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 = 2, 18, 367.244$ (taking $n = 15$, $N = 50$, and $s_y^2 = \frac{1}{n-1} \sum\limits_{i \in s} \{y_i - \overline{y}(s)\}^2 = 46, 79, 298.095$). To use RG method, the three group estimates are computed as $\widehat{\theta}_{(1)} = \overline{y}(s_1) = 4800$, $\widehat{\theta}_{(2)} = \overline{y}(s_2) = 8094$, and $\widehat{\theta}_{(3)} = \overline{y}(s_3) = 5600$. The RG estimator for θ is $\widehat{\widehat{\theta}} = \frac{4800 + 8094 + 5600}{3} = 6164.667 = \widehat{\theta}$. The estimator of $V\left(\widehat{\theta}\right)$ by RG method is $\widehat{V}_{RG1}\left(\widehat{\theta}\right) = \widehat{V}_{RG2}\left(\widehat{\theta}\right) = \frac{1}{3 \times 2} \sum\limits_{i=1}^{3} \left(\widehat{\theta}_{(i)} - \widehat{\widehat{\theta}}\right)^2$ $= 9, 83, 918.111$, which is much larger than $\widehat{V}\left(\widehat{\theta}\right) = 2, 18, 367.2$. This is because the estimated amount of bias of $\widehat{V}_{RG1}\left(\widehat{\theta}\right) = S_y^2 \big/ N = 93, 585.961$ is large.

(ii) **Estimates using auxiliary information**

The ratio estimator for the average expenditure $\theta = \overline{Y}$ is $\widehat{\theta} = \frac{\overline{y}(s)}{\overline{x}(s)} \overline{X} = \frac{6164.67}{2.8} \times 3.5 = 7705.833$. The conventional estimate of variance of $\widehat{\theta}$ based on LR method is $\widehat{V}\left(\widehat{\theta}\right) = (1/n - 1/N) s_{\hat{d}}^2 =$ $(1/15 - 1/60) \times 31, 55, 850 = 1, 47, 273$. Ratio estimates based on the three groups are $\widehat{\theta}_{(1)} = 4800 \times 3.5 \big/ 2.2 = 7636.36$, $\widehat{\theta}_{(2)} = 8094 \times 3.5 \big/ 3$ $= 9443.000$, and $\widehat{\theta}_{(3)} = 5600 \times 3.5 \big/ 3.2 = 6125.000$. The RG estimate of θ is $\widehat{\widehat{\theta}} = 7734.788$, which is different from $\widehat{\theta} = 7705.83$. The two estimates of the variance of $\widehat{\theta}$ are $\widehat{V}_{RG1}\left(\widehat{\theta}\right) = \frac{1}{3 \times 2} \sum\limits_{i=1}^{3} \left(\widehat{\theta}_{(i)} - \widehat{\widehat{\theta}}\right)^2 = 9, 19, 848.433$ and $\widehat{V}_{RG2}\left(\widehat{\theta}\right) = \frac{1}{3 \times 2} \sum\limits_{i=1}^{3} \left(\widehat{\theta}_{(i)} - \widehat{\theta}\right)^2 = 9, 20, 268.016$.

18.4 JACKKNIFE METHOD

The most widely used method of estimation of variance in a complex survey design is the Jackknife (JK) method. The JK method was introduced by

Quenouille (1949) for reduction of bias of an estimator of a serial correlation coefficient. Quenouille (1956) extended this technique for bias reduction, in general, under infinite population setup. Tuckey (1958) used the JK technique for estimation of variance under an infinite population setup while Durbin (1959) proposed the JK method for estimation of variance in finite population sampling. Good details of the JK method are given by Gray and Schucany (1972), Miller (1974), Efron (1982), and Wolter (1985), among others.

18.4.1 Jackknife Method for an Infinite Population

Let a random sample s of size n be selected from an infinite population with distribution function $\mathscr{F}$ and let $\widehat{\theta}(s) = \widehat{\theta}$ be an estimator for the population parameter θ based on the full sample s. The sample s is partitioned into k disjoined groups, $s_1,\ldots,s_k$ each of size $m(=n/k$, assuming an integer). Let $\widehat{\theta}_{(-j)}$ be an estimator of θ, which is the same functional form as $\widehat{\theta}$, but based on the reduced sample $s - s_j$ of size $n - m$, obtained by deleting the sample s_j from s. Let us define the pseudovalue computed from $s - s_j$ as

$$\widehat{\theta}^{(j)} = k\widehat{\theta} - (k-1)\widehat{\theta}_{(-j)} \tag{18.4.1}$$

The Jackknife (JK) estimator for θ is defined as the average of the pseudovalues $\widehat{\theta}^{(j)}$, and is denoted by

$$\begin{aligned}\widehat{\theta}_J &= \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}^{(j)} \\ &= k\widehat{\theta} - (k-1)\ \widehat{\theta}_{\bullet}\end{aligned} \tag{18.4.2}$$

where

$$\widehat{\theta}_{\bullet} = \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}_{(-j)} \tag{18.4.3}$$

The JK variance estimators of the variances of both the estimators of $\widehat{\theta}$ and $\widehat{\theta}_J$ are given by

$$\begin{aligned}\widehat{V}_J(1) &= \frac{1}{k(k-1)}\sum_{j=1}^{k}\left(\widehat{\theta}^{(j)} - \widehat{\theta}_J\right)^2 \\ &= \frac{(k-1)}{k}\sum_{j=1}^{k}\left(\widehat{\theta}_{(-j)} - \widehat{\theta}_{\bullet}\right)^2\end{aligned} \tag{18.4.4}$$

An alternative JK variance estimator is given by

$$\widehat{V}_J(2) = \frac{1}{k(k-1)} \sum_{j=1}^{k} \left(\widehat{\theta}^{(j)} - \widehat{\theta}\right)^2 \tag{18.4.5}$$

The estimator $\widehat{V}_J(2)$ is conservative in the sense $\widehat{V}_J(2) \geq \widehat{V}_J(1)$.

Theorem 18.4.1

If $B\left(\widehat{\theta}\right)$, the bias of $\widehat{\theta}$ is of order $1/n$ and can be expressed as

$$B\left(\widehat{\theta}\right) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \cdots$$

then the bias of $\widehat{\theta}_J$ is of order $1/n^2$ and is given by

$$B\left(\widehat{\theta}_J\right) = \frac{c_2(\theta)}{n^2} + \frac{c_3(\theta)}{n^3} + \cdots$$

where the constants $b_1(\theta)$, $b_2(\theta)$, $b_3(\theta)$,... and $c_2(\theta)$, $c_3(\theta)$,... are free from n but may depend on θ.

Proof

Given $E\left(\widehat{\theta}\right) = \theta + B\left(\widehat{\theta}\right) = \theta + \dfrac{b_1(\theta)}{n} + \dfrac{b_2(\theta)}{n^2} + \cdots$

Hence $E\left(\widehat{\theta}_{(-j)}\right) = \theta + \dfrac{b_1(\theta)}{m(k-1)} + \dfrac{b_2(\theta)}{\{m(k-1)\}^2} + \cdots$

and

$$\begin{aligned}
E\left(\widehat{\theta}^{(j)}\right) &= kE\left(\widehat{\theta}\right) - (k-1)E\left(\widehat{\theta}_{(-j)}\right) \\
&= k\left(\theta + \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \cdots\right) \\
&\quad - (k-1)\left(\theta + \frac{b_1(\theta)}{m(k-1)} + \frac{b_2(\theta)}{\{m(k-1)\}^2} + \cdots\right).
\end{aligned}$$

The bias of $\widehat{\theta}_J$ is

$$\begin{aligned}
B\left(\widehat{\theta}_J\right) &= E\left(\widehat{\theta}_J\right) - \theta \\
&= \frac{1}{k}\sum_{j=1}^{k} E\left(\widehat{\theta}^{(j)}\right) - \theta \\
&= \frac{c_2(\theta)}{n^2} + \cdots
\end{aligned}$$

where $c_2(\theta) = -kb_2(\theta)/(k-1)$.

Remark 18.4.1

As for the optimum number of the group k that maximizes efficiency of the JK estimator, no general rule is available. However, studies by Rao and Webster (1966), Chakrabarty and Rao (1968), and Rao and Rao (1971) reveal that for ratio estimation, the bias and variance of the JK estimator decreases with an increase in k when the sample size n is small or moderate. More detailed discussions are given by Wolter (1985).

Example 18.4.1.1

Let X_1, X_2,…,X_n be a random sample from a population with mean μ and variance σ^2. Consider a biased estimator for $\theta = \sigma^2$ as

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2.$$

The bias of $\widehat{\theta}$ is $B\left(\widehat{\theta}\right) = -\sigma^2/n$. Let $m = 1$ (i.e., $k = n$), then

$$\begin{aligned}
\widehat{\theta}_{(-j)} &= \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n}\left(X_i - \frac{n\overline{X} - X_j}{n-1}\right)^2\\
&= \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n}\left(X_i - \overline{X} - \frac{\overline{X} - X_j}{n-1}\right)^2\\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i - \overline{X} - \frac{\overline{X} - X_j}{n-1}\right)^2 - \left(\frac{n}{n-1}\right)^2\left(X_j - \overline{X}\right)^2\right]\\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \frac{n}{(n-1)^2}\left(X_j - \overline{X}\right)^2 - \left(\frac{n}{n-1}\right)^2\left(X_j - \overline{X}\right)^2\right]\\
&= \frac{1}{n-1}\left[n\widehat{\theta} - \frac{n}{n-1}\left(X_j - \overline{X}\right)^2\right],
\end{aligned}$$

$$\begin{aligned}
\widehat{\theta}_{\bullet} &= \frac{1}{n}\sum_{j=1}^{n}\widehat{\theta}_{(-j)} = \frac{1}{n-1}\left[n\widehat{\theta} - \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2\right]\\
&= \frac{n(n-2)}{(n-1)^2}\widehat{\theta}
\end{aligned}$$

and

$$\begin{aligned}\widehat{\theta}_J &= n\widehat{\theta} - (n-1)\widehat{\theta}_{\bullet}\\ &= \frac{n}{n-1}\widehat{\theta}\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\end{aligned}$$

Clearly, $\widehat{\theta}_J$ is an unbiased estimator of θ.

Now,

$$\begin{aligned}\sum_{j=1}^{n}\left(\widehat{\theta}_{(-j)} - \widehat{\theta}_{\bullet}\right)^2 &= \frac{n^2}{(n-1)^4}\sum_{j=1}^{n}\left(\left(X_j - \overline{X}\right)^2 - \sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2/n\right)^2\\ &= \frac{n^2}{(n-1)^4}\left[\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^4 - \left\{\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2\right\}^2 \Big/ n\right]\\ &= \frac{n^3}{(n-1)^4}\left(m_4 - m_2^2\right)\end{aligned}$$

where $m_4 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4$ and $m_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

The JK variance estimator of $\widehat{\theta}_J$ is

$$\begin{aligned}\widehat{V}_J(1) &= \frac{n-1}{n}\sum_{j=1}^{n}\left(\widehat{\theta}_{(-j)} - \widehat{\theta}_{\bullet}\right)^2\\ &= \frac{n^2}{(n-1)^3}\left(m_4 - m_2^2\right)\end{aligned}$$

Example 18.4.1.2

Let a random sample $X_1,\ldots,X_n$ of size n be selected from a uniform distribution with density $f(x,\theta) = 1/\theta;\ 0 < x < \theta$. The nth order statistic $X_{(n)}$ is the maximum likelihood estimator of θ. The estimator $\widehat{\theta} = X_{(n)}$ is a biased estimator of θ. Noting $E\{X_{(n)}\} = n\theta/(n+1)$, we find the bias of $\widehat{\theta}\left(= X_{(n)}\right)$ as $B\left(\widehat{\theta}\right) = E\left\{X_{(n)}\right\} - \theta = -\theta/(n+1)$, which is of order n^{-1}. Now,

$$\widehat{\theta}_{(-j)} = \begin{cases} X_{(n)} & \text{for } j = 1,\ldots,n-1\\ X_{(n-1)} & \text{for } j = n\end{cases}$$

where $X_{(n\ -\ 1)}$ is the $(n-1)$th order statistic.

Furthermore, noting $\widehat{\theta}_{\bullet}=\frac{1}{n}\sum_{j=1}^{n}\widehat{\theta}_{(-j)}=\frac{(n-1)X_{(n)}+X_{(n-1)}}{n}$, we find the JK estimator of θ is

$$\begin{aligned}\widehat{\theta}_J&=n\widehat{\theta}-(n-1)\widehat{\theta}_{\bullet}\\&=X_{(n)}+\frac{n-1}{n}\left(X_{(n)}-X_{(n-1)}\right)\end{aligned}$$

Noting $E\{X_{(n-1)}\}=(n-1)\theta/(n+1)$, the bias of θ_J is obtained as

$$B\left(\widehat{\theta}_J\right)=E\left(\widehat{\theta}_J\right)-\theta=-\frac{\theta}{n(n+1)}\ \text{ is of order }\ n^{-2}.$$

The JK variance estimator of $\widehat{\theta}_J$ is

$$\begin{aligned}\widehat{V}_J(1)&=\frac{n-1}{n}\sum_{j=1}^{n}\left(\widehat{\theta}_{(-j)}-\widehat{\theta}_{\bullet}\right)^2\\&=\frac{n-1}{n}\left[(n-1)\left\{X_{(n)}-\frac{(n-1)X_{(n)}+X_{(n-1)}}{n}\right\}^2\right.\\&\qquad\left.+\left\{X_{(n-1)}-\frac{(n-1)X_{(n)}+X_{(n-1)}}{n}\right\}^2\right]\\&=\frac{(n-1)^2}{n^2}\left(X_{(n)}-X_{(n-1)}\right)^2\\&\cong\left(X_{(n)}-X_{(n-1)}\right)^2\ \text{ for large }\ n.\end{aligned}$$

18.4.1.1 Higher-Order Jackknife Estimator

Let $\widehat{\theta}_{J(-j)}$ be the estimator of θ based on the sample size $n-1$ obtained from $\widehat{\theta}_J$ by deleting *j*th observation. The second-order JK estimator of θ was proposed by Quenouille (1956) as

$$\begin{aligned}\widehat{\theta}_J^{(2)}&=\frac{n^2\widehat{\theta}_J-\frac{(n-1)^2}{n}\sum_{j=1}^{n}\widehat{\theta}_{J(-j)}}{n^2-(n-1)^2}\\&=\frac{1}{(2n-1)}\left[n^3\widehat{\theta}-(n-1)\{2n(n-1)+1\}\widehat{\theta}_{\bullet}\right.\\&\qquad\left.+(n-1)^2(n-2)\left\{\frac{1}{n(n-1)}\sum_{i\neq}^{n}\sum_{j=1}^{n}\widehat{\theta}_{-i,j}\right\}\right]\end{aligned}\qquad(18.4.6)$$

where $\widehat{\theta}_{-i,j}$ is the same functional as $\widehat{\theta}$ obtained from the sample size $n-2$ after deleting the ith and $j(\neq i)$th observations from the original sample.

The second-order JK estimator $\widehat{\theta}_J^{(2)}$ eliminates the order $1/n^2$ term from the bias. Similarly, the higher-order JK estimators are defined to eliminate third-, fourth-, and higher-order bias terms.

18.4.1.2 Generalized Jackknife Estimator

Schucany et al. (1971) generalized the biased reduction procedure further. Consider two biased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of θ with

$$E\left(\widehat{\theta}_1\right) = \theta + f_1(n)a(\theta),$$
$$E\left(\widehat{\theta}_2\right) = \theta + f_2(n)a(\theta)$$

and

$$\begin{vmatrix} 1 & 1 \\ f_1(n) & f_2(n) \end{vmatrix} \neq 0.$$

The generalized JK estimator (defined by Wolter, 1985) as

$$GJ\left(\widehat{\theta}_1, \widehat{\theta}_2\right) = \frac{\begin{vmatrix} \widehat{\theta}_1 & \widehat{\theta}_2 \\ f_1(n) & f_2(n) \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ f_1(n) & f_2(n) \end{vmatrix}} \qquad (18.4.7)$$

is strictly unbiased for θ.

If we take $\widehat{\theta}_1 = \widehat{\theta}$, $\widehat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n}\widehat{\theta}_{(-j)} = \widehat{\theta}_{\bullet}$, $f_1(n) = 1/n$, and $f_2(n) = 1/(n-1)$, the $GJ\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$ reduces to $\widehat{\theta}_J$.

More general, consider $p+1$ estimators of θ based on a sample size n such that

$$B\left(\widehat{\theta}_j\right) = E\left(\widehat{\theta}_j\right) - \theta = f_{1j}(n)b_1(\theta) + \cdots + f_{kj}(n)b_k(\theta) + \cdots$$

for $j = 1, 2, \ldots, p+1$

$$\text{with} \quad \begin{vmatrix} 1 & \cdot & \cdot & 1 \\ f_{11}(n) & \cdot & \cdot & f_{1p+1}(n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ f_{p1}(n) & \cdot & \cdot & f_{pp+1}(n) \end{vmatrix} \neq 0. \qquad (18.4.8)$$

Schucany et al. (1971) showed that the generalized JK estimator

$$GJ\left(\widehat{\theta}_1, \ldots, \widehat{\theta}_{p+1}\right) = \frac{\begin{vmatrix} \widehat{\theta}_1 & \cdot & \cdot & \widehat{\theta}_{p+1} \\ f_{11}(n) & \cdot & \cdot & f_{1p+1}(n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ f_{p1}(n) & \cdot & \cdot & f_{pp+1}(n) \end{vmatrix}}{\begin{vmatrix} 1 & \cdot & \cdot & 1 \\ f_{11}(n) & \cdot & \cdot & f_{1p+1}(n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ f_{p1}(n) & \cdot & \cdot & f_{pp+1}(n) \end{vmatrix}} \tag{18.4.9}$$

eliminates the first p terms of the bias of Eq. (18.4.8).

18.4.2 Jackknife Method for a Finite Population

Here a sample s of size n is selected from a finite population using some sampling design. The sample s is divided at random into k groups each of size m assuming $n/m = k$ is an integer. The formulae $\widehat{\theta}_{(-j)}$, $\widehat{\theta}^{(j)}$, $\widehat{\theta}_J$, $\widehat{\theta}_\bullet$, and the JK variance estimators $\widehat{V}_J(1)$ and $\widehat{V}_J(2)$ used for infinite population in Section 18.4.1 remain the same for the finite population.

18.4.2.1 Probability Proportional to Size With Replacement Sampling

Let a sample s of size n be selected from a finite population U of N units by PPSWR method using $p_i\left(>0, \sum_{i\in U} p_i = 1\right)$ as the normed size measure for the ith unit. The Hansen—Hurwitz estimator for the population total Y is

$$\widehat{\theta} = \widehat{Y}_{hh} = \frac{1}{n}\sum\nolimits_{i\in s}\frac{y_i}{p_i} \tag{18.4.10}$$

where $\sum_{i\in s}$ denotes the sum over units in s with repetition.

The expressions for the variance of $\widehat{\theta}$ and its unbiased estimators are, respectively, given by $V\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i\in U} p_i\left(\frac{y_i}{p_i} - Y\right)^2$ and $\widehat{V}\left(\widehat{\theta}\right) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i} - \widehat{Y}_{hh}\right)^2$.

Let the sample s be divided at random into k groups each of size m. The estimator for the total Y based on the sample after deleting the jth group (i.e., based on $s - s_j$) is given by $\widehat{\theta}_{(-j)} = \frac{1}{n-m}\sum_{i\in s-s_j}\frac{y_i}{p_i}$. The pseudovalue computed from $s - s_j$ is

$$\widehat{\theta}^{(j)} = k\widehat{\theta} - \frac{k-1}{n-m}\sum_{i\in s-s_j}\frac{y_i}{p_i}$$

Hence the JK estimator for Y is given by

$$\begin{aligned}\widehat{\theta}_J &= \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}^{(j)} = k\widehat{\theta} - \frac{k-1}{k(n-m)}\sum_{j=1}^{k}\sum_{i\in s-s_j}\frac{y_i}{p_i} \\ &= k\widehat{\theta} - \frac{k-1}{k(n-m)}\sum_{j=1}^{k}\left(\sum_{i\in s}\frac{y_i}{p_i} - \sum_{i\in s_j}\frac{y_i}{p_i}\right) = \widehat{\theta} = \widehat{\theta}_{\bullet}.\end{aligned} \tag{18.4.11}$$

The JK variance estimator is given by

$$\begin{aligned}\widehat{V}_J(1) = \widehat{V}_J(2) &= \frac{k-1}{k}\sum_{j=1}^{k}\left(\widehat{\theta}_{(-j)} - \widehat{\theta}_{\bullet}\right)^2 \\ &= \frac{k-1}{k}\sum_{j=1}^{k}\left\{\frac{1}{n-m}\sum_{i\in s-s_j}\left(\frac{y_i}{p_i} - \widehat{Y}_{hh}\right)\right\}^2\end{aligned} \tag{18.4.12}$$

Clearly, the JK variance estimator $\widehat{V}_J(1)\left(\text{i.e., } \widehat{V}_J(2)\right)$ is not equal to $\widehat{V}\left(\widehat{\theta}\right)$ in general. However, for $m = 1$, i.e., $k = n$, we get

$$\begin{aligned}\widehat{V}_J(1) = \widehat{V}_J(2) &= \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i} - \widehat{Y}_{hh}\right)^2 \\ &= \widehat{V}\left(\widehat{\theta}\right)\end{aligned} \tag{18.4.13}$$

18.4.2.1.1 Bias of Jackknife Variance Estimator

$$\begin{aligned}E\left\{\widehat{V}_J(1)\right\} &= \frac{k-1}{k}\left\{\sum_{j=1}^{k}E\left(\widehat{\theta}_{(-j)}\right)^2 - k\left(\widehat{Y}_{hh}\right)^2\right\} \\ &= \frac{k-1}{k}\left\{k\left(Y^2 + \frac{\sigma_0^2}{n-m}\right) - k\left(Y^2 + \frac{\sigma_0^2}{n}\right)\right\} \\ &\qquad \left(\text{where } \sigma_0^2 = \sum_{i\in U}p_i\left(\frac{y_i}{p_i} - Y\right)^2\right) \\ &= \frac{1}{n}\sum_{i\in U}p_i\left(\frac{y_i}{p_i} - Y\right)^2 = V\left(\widehat{\theta}\right)\end{aligned}$$

Thus for PPSWR sampling the JK estimator is unbiased for $V\left(\widehat{\theta}\right)$.

18.4.2.2 Simple Random Sampling With Replacement

PPSWR reduces to simple random sampling (SRSWR) if $p_i = 1/N$ for $\forall\ i \in U$. Hence the JK estimator $\widehat{\theta}_J$ for Y is equal to the full sample estimator $N\,\overline{y}(s)$.

18.4.2.3 Inclusion Probability Proportional to Size or πps Sampling Design

Suppose a sample s of size n is selected from a population with $\pi_i = np_i$ as the inclusion probability for the ith unit with $p_i\left(>0, \sum_{i\in U} p_i = 1\right)$ as the normed size measure for the ith unit. The Horvitz–Thompson estimator for the total Y based on the full sample s is given by

$$\widehat{\theta} = \widehat{Y}_{ht} = \sum_{i\in s}\frac{y_i}{\pi_i} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} \tag{18.4.14}$$

The variance of $\widehat{\theta}$ and its unbiased estimator are, respectively, given by

$$V\left(\widehat{\theta}\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{18.4.15}$$

and

$$\widehat{V}\left(\widehat{\theta}\right) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{18.4.16}$$

The sample s is partitioned at random in to k groups each of size m $(n = mk)$. The Horvitz–Thompson estimator for the total Y based on the sample $s - s_j$ is given by

$$\widehat{\theta}_{(-j)} = \frac{1}{n-m}\sum_{i\in s-s_j}\frac{y_i}{p_i}$$

The ith pseudovalue is

$$\begin{aligned}\widehat{\theta}^{(j)} &= k\sum_{i\in s}\frac{y_i}{np_i} - (k-1)\sum_{i\in s-s_j}\frac{y_i}{(n-m)p_i}\\ &= k\sum_{i\in s}\frac{y_i}{np_i} - k\left(\sum_{i\in s}\frac{y_i}{np_i} - \sum_{i\in s_j}\frac{y_i}{np_i}\right)\\ &= k\sum_{i\in s_j}\frac{y_i}{np_i}\end{aligned}$$

The JK estimator of θ is the same as the original estimator $\widehat{Y}_{hte}$ because

$$\begin{aligned}\widehat{\theta}_J &= \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}^{(j)} \\ &= \sum_{i\in s}\frac{y_i}{np_i} \\ &= \widehat{\theta} = \widehat{Y}_{ht} = \widehat{\theta}\boldsymbol{.}\end{aligned} \tag{18.4.17}$$

The JK variance estimator is given by

$$\widehat{V}_J(1) = \widehat{V}_J(2) = \frac{k-1}{k}\sum_{j=1}^{k}\left(\widehat{\theta}_{(-j)} - \widehat{\theta}\boldsymbol{.}\right)^2 \tag{18.4.18}$$

The estimators $\widehat{V}_J(1)\left(\text{ i.e., } \widehat{V}_J(2)\right)$ are not equal to $\widehat{V}\left(\widehat{\theta}\right)$ in general.

18.4.2.3.1 Bias of Jackknife Variance Estimator

The bias of the JK estimator is given by

$$B\left[\widehat{V}_J(1)\right] = E\left\{\widehat{V}_J(1)\right\} - V\left(\widehat{Y}_{ht}\right) \tag{18.4.19}$$

Now,
$$\begin{aligned}E\left[\widehat{V}_J(1)\right] &= \frac{k-1}{k}E\left[\sum_{j=1}^{k}\left(\widehat{\theta}_{(-j)}\right)^2 - k\left(\widehat{Y}_{ht}\right)^2\right] \\ &= \frac{k-1}{k}\left[E\left\{\sum_{j=1}^{k}V\left(\widehat{\theta}_{(-j)}\middle|s\right)\right\}\right] \\ &= (k-1)\left(\frac{1}{n-m}-\frac{1}{n}\right)\frac{1}{n-1}E\left(\sum_{i\in s}\frac{y_i^2}{p_i^2} - n\widehat{Y}_{ht}^2\right) \\ &= \frac{1}{n-1}\left[\sum_{i\in U}p_i\left(\frac{y_i}{p_i}-Y\right)^2 - V\left(\widehat{Y}_{ht}\right)\right]\end{aligned} \tag{18.4.20}$$

Substituting Eq. (18.4.20) in Eq. (18.4.19) we get

$$B\left[\widehat{V}_J(1)\right] = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i\in U}p_i\left(\frac{y_i}{p_i}-Y\right)^2 - V\left(\widehat{Y}_{ht}\right)\right] \tag{18.4.21}$$

From expression (18.4.21), we note that the JK estimator overestimates or underestimates the variance if the variance of the Hansen–Hurwitz estimator based on PPSWR sampling is more or less than the variance of Horvitz–Thompson estimator based on a πps sampling scheme of the same sample size. In practice Eq. (18.4.21) is expected to be positive because the Horvitz–Thompson estimator is expected to be more efficient than the Hansen–Hurwitz estimator for estimating total Y.

18.4.2.4 Simple Random Sampling Without Replacement

On substituting $\pi_i = n/N$ in Eq. (18.4.17), we find that the JK estimators for SRSWOR sampling scheme are equal to full sample estimate $N\overline{y}(s)$ and that the JK variance estimator overestimates the variance as variance of the sample mean based on SRSWR is larger than that of the sample mean based on SRSWOR.

18.4.2.5 Regression Estimator

Let a sample s of size n be selected by the SRSWOR method. Let y_i and x_i be the value of the study (y) and auxiliary variable (x) for the ith unit and $\overline{X} = \sum\limits_{i\in U} x_i/N$ be the population mean of x, which is assumed to be known. The regression estimator for the population mean $\theta = \overline{Y}$ based on the full sample s is given by

$$\widehat{\theta} = \widehat{Y}_{reg}(s) = \overline{y}(s) - \widehat{\beta}(s)\left(\overline{x}(s) - \overline{X}\right) \tag{18.4.22}$$

where $\overline{x}(s)$, $\overline{y}(s)$, and $\widehat{\beta}(s) = \dfrac{\sum\limits_{i\in s} y_i x_i - n\overline{x}(s)\overline{y}(s)}{\sum\limits_{i\in s} x_i^2 - n\{\overline{x}(s)\}^2}$ are, respectively, the sample mean of x, sample mean of y, and the sample regression coefficient of y on x.

The regression estimator after deleting the group s_j is given by

$$\widehat{\theta}_{(-j)} = \overline{y}_{(-j)} - \widehat{\beta}_{(-j)}\left(\overline{x}_{(-j)} - \overline{X}\right)$$

where $\overline{x}_{(-j)}$, $\overline{y}_{(-j)}$, and $\widehat{\beta}_{(-j)} = \dfrac{\sum\limits_{i\in s-s_j}\left(y_i - \overline{y}_{(-j)}\right)\left(x_i - \overline{x}_{(-j)}\right)}{\sum\limits_{i\in s-s_j}\left(x_i - \overline{x}_{(-j)}\right)^2}$ are, respectively, the sample mean of x, sample mean of y, and sample regression coefficient obtained from $s - s_j$. Noting

$$\widehat{\theta}^{(j)} = k\left\{\overline{y}(s) - \widehat{\beta}(s)\left(\overline{x}(s) - \overline{X}\right)\right\} - (k-1)\left(\overline{y}_{(-j)} - \widehat{\beta}_{(-j)}\left(\overline{x}_{(-j)} - \overline{X}\right)\right) \tag{18.4.23}$$

we find the JK estimator of θ as

$$\widehat{\theta}_J = \frac{1}{k}\sum_{j=1}^{k}\widehat{\theta}^{(j)} \tag{18.4.24}$$

The estimator $\widehat{\theta}_J$ is quite different from $\widehat{\theta}$. The JK variance estimators

$$\begin{aligned}\widehat{V}_J(1) &= \frac{1}{k(k-1)}\sum_{j=1}^{k}\left(\widehat{\theta}^{(j)} - \widehat{\theta}_J\right)^2 \text{ and}\\ \widehat{V}_J(2) &= \frac{1}{k(k-1)}\sum_{j=1}^{k}\left(\widehat{\theta}^{(j)} - \widehat{\theta}\right)^2\end{aligned} \tag{18.4.25}$$

are also different.

18.4.2.6 Numerical Example

The following data give the number of patients treated and the cost of treatment per day at 15 clinics selected at random from 50 clinics by the SRSWOR method (Table 18.4.1).

Let Y and X be the total cost per day and the total number of patients in 50 clinics, respectively, and we want to estimate $R = \dfrac{Y}{X} =$ cost per patient based on the selected sample s of 15 (=n) clinics. Here estimated cost per patient θ, based on the full sample, is $\widehat{\theta} = \widehat{R} = \dfrac{\overline{y}(s)}{\overline{x}(s)} = 90.361,$ where $\overline{y}(s) = 5000$ and $\overline{x}(s) = 55.333$ are the sample means of x and y, respectively. The estimated approximate variance of $\widehat{R}$ based on the full sample of size 15 is $\widehat{V}\left(\widehat{\theta}\right) = \widehat{V}\left(\widehat{R}\right) = \frac{1}{\overline{x}_s^2}\left(\frac{1}{15} - \frac{1}{50}\right)\frac{1}{14}\sum_{i\in s}\widehat{d}_i^2 = 24.390,$ where $\widehat{d}_i = y_i - \widehat{R}x_i$. Let us divide the sample s into 15 (=k) groups each group consisting of just one (=m) unit. Writing $\widehat{\theta}_{(-j)} = \widehat{R}_{(-j)} = \dfrac{n\overline{y}(s) - y_j}{nx(s) - x_j}$ and $\widehat{\theta}^{(j)} = k\widehat{\theta} - (k-1)\widehat{\theta}_{(-j)}$, we arrive at the JK estimator for θ as $\widehat{\theta}_J = \dfrac{1}{k}\sum_{j=1}^{k}\widehat{\theta}^{(j)} = 89.491.$

Two alternative JK variance estimators are, respectively, given by $\widehat{V}_J(1) = \dfrac{1}{k(k-1)}\sum_{j=1}^{k}\left(\widehat{\theta}^{(j)} - \theta_J\right)^2 = 40.526$ and $\dfrac{1}{k(k-1)}\sum_{j=1}^{k}\left(\widehat{\theta}^{(j)} - \widehat{\theta}\right)^2 = 40.579.$

Here we note that the full sample estimate and JK estimate of the population ratio R are very close to each other and hence the two JK variance estimators are almost equal but the JK variance estimates are much larger than the full sample estimate. Numerical computations are shown in Table 18.4.2.

Table 18.4.1 Number of patients and cost of treatment of selected clinics

Clinic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of patients (x)	100	40	50	70	50	60	45	150	60	15	20	40	30	45	55
Cost in rands (y)	8,000	4,000	5,000	5,000	6,000	6,000	4,000	10,000	5,000	2,000	3,000	5,000	3,000	4,000	5,000

Table 18.4.2 Compution of Jackknife variance estimation

j	x	y	$n\bar{y}(s) - y_j$	$n\bar{x}(s) - x_j$	$\widehat{R}_{(-j)}$	$\widehat{\theta}^{(j)}$	$\widehat{\theta}^{(j)} - \widehat{\theta}$	$\left(\widehat{\theta}^{(j)} - \widehat{\theta}\right)^2$	$\widehat{d}_j = y_j - \widehat{R}x_j$	$\widehat{d}_j^2$
1	100	8,000	67,000	730	91.781	70.488	−19.873	394.936	−1,036.145	1,073,596.461
2	40	4,000	71,000	790	89.873	97.20000	6.839	46.772	385.542	148,642.634
3	50	5,000	70,000	780	89.744	99.006	8.645	74.736	481.928	232,254.597
4	70	5,000	70,000	760	92.105	65.952	−24.409	595.799	−1,325.301	1,756,422.741
5	50	6,000	69,000	780	88.462	116.954	26.593	707.188	1481.928	2,196,110.597
6	60	6,000	69,000	770	89.61	100.882	10.521	110.691	578.313	334,445.926
7	45	4,000	71,000	785	90.446	89.178	−1.183	1.399	−66.265	4,391.050
8	150	10,000	65,000	680	95.588	17.1900	−73.171	5,353.995	−3,554.217	12,632,458.480
9	60	5,000	70,000	770	90.909	82.696	−7.665	58.752	−421.687	177,819.926
10	15	2,000	73,000	818	89.571	101.428	11.067	122.478	644.578	4,18,480.798
11	20	3,000	72,000	810	88.889	110.976	20.618	424.978	1,192.771	1,422,702.658
12	40	5,000	70,000	790	88.608	114.91	24.549	602.653	1,385.542	1,919,726.634
13	30	3,000	72,000	800	90.000	95.422	5.061	25.614	289.187	83,611.771
14	45	4,000	71,000	785	90.446	89.178	−1.183	1.399	−66.265	4,391.050
15	55	5,000	70,000	775	90.323	90.900	0.539	0.291	30.120	907.214
Total	830	75,000				1,342.360		8,523.681		22,402,962.540

18.5 BALANCED REPEATED REPLICATION METHOD

Let a population of N units be stratified into L strata and N_h be the size of the $h(=1,\ldots,L)$th stratum. Let y_{hi} be the value of the variable of interest y for u_{hi}, the ith unit of the hth stratum, $W_h = N_h/N$, $\overline{Y}_h = \sum_{i=1}^{N_h} y_{hi}/N_h$, and $\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$. Suppose from each of the strata a sample $\Im_h$ of size n_h units is selected by some suitable sampling scheme. Let $\widehat{\theta}_{st}$ be an estimator of the population parameter of interest θ based on the full sample $s = \bigcup_{h=1}^{L} \Im_h$ of size $n = \sum_h n_h$. In this section we shall consider the various methods of estimation of variance of $\widehat{\theta}_{st}$ using the BRR method. The BRR method was introduced by McCarthy (1969).

18.5.1 Stratified Sampling With $n_h = 2$

Suppose a sample $\Im_h = (u_{h1}, u_{h2})$ of size $n_h = 2$ is selected from the hth stratum by the SRSWOR method, $h = 1,\ldots,L$. Let y_{h1} and y_{h2} be the value of the study variable y for the units u_{h1} and u_{h2}, respectively. In this case, an unbiased estimator for the population mean $\overline{Y}$ is $\widehat{\overline{Y}}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h$, where $\overline{y}_h = (y_{h1} + y_{h2})/2$. If the finite population correction term $f_h = n_h/N = 2/N$ is neglected for each of the strata, then the variance of $\widehat{\overline{Y}}_{st}$ reduces to

$$V\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{2}\sum_{h=1}^{L} W_h^2\, S_h^2 \tag{18.5.1}$$

where $S_h^2 = \sum_{j=1}^{N_h}\left(y_{hj} - \overline{Y}_h\right)^2 \Big/ (N_h - 1)$.

An unbiased estimator of $V\left(\widehat{\overline{Y}}_{st}\right)$ is given by

$$\begin{aligned}\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) &= \frac{1}{2}\sum_{h=1}^{L} W_h^2 s_h^2 \\ &= \frac{1}{4}\sum_{h=1}^{L} W_h^2 d_h^2\end{aligned} \tag{18.5.2}$$

where $s_h^2 = \sum_{i=1}^{n_h}\left(y_{hi} - \overline{y}_h\right)^2/(n_h - 1) = (y_{h1} - y_{h2})^2/2$, $n_h = 2$, and $d_h = y_{h1} - y_{h2}$.

Let us divide the selected sample $\Im = (\Im_1, \ldots, \Im_L)$ into two half-samples $\alpha_1 = (u_{11},\ldots,u_{h1},\ldots,u_{L1})$ and $\alpha_2 = (u_{12},\ldots,u_{h2},\ldots,u_{L2})$. From each of the half-samples α_1 and α_2, we can construct estimators for the mean $\overline{Y}$ as

$$t_{\alpha_1} = \sum_{h=1}^{L} W_h y_{h1} \text{ and } t_{\alpha_2} = \sum_{h=1}^{L} W_h y_{h2} \tag{18.5.3}$$

From the estimators t_{α_1} and t_{α_2}, a combined estimator of $\overline{Y}$ is obtained as

$$\bar{t} = \frac{t_{\alpha_1} + t_{\alpha_2}}{2} = \widehat{\overline{Y}}_{st} \tag{18.5.4}$$

Assuming that the half-samples are independent, we find an unbiased estimator of $V\left(\widehat{\overline{Y}}_{st}\right)$ using the RG method as

$$\begin{aligned} \widehat{V}_{RG}(\bar{t}) &= \frac{1}{2}\sum_{j=1}^{2}\left(t_{\alpha_j} - \bar{t}\right)^2 \\ &= \frac{1}{4}(t_{\alpha_1} - t_{\alpha_2})^2 \end{aligned} \tag{18.5.5}$$

Clearly, $\widehat{V}_{RG}(\bar{t})$ may be different from $V\left(\widehat{\overline{Y}}_{st}\right)$ given in Eq. (18.5.2).

In general, we can construct 2^L half-samples by taking one unit from each of the stratum. Let us select one half-sample α (say) by choosing one unit at random from each of the stratum and construct an unbiased estimator of $\overline{Y}$ as follows:

$$t_\alpha = \sum_{h=1}^{L} W_h\{\phi_{h\alpha} y_{h1} + (1 - \phi_{h\alpha}) y_{h2}\} \tag{18.5.6}$$

where $\phi_{h\alpha} = 1$ if the unit u_{h1} belongs to the half-sample α and $\phi_{h\alpha} = 0$ otherwise. Let $\widehat{V}(t_\alpha) = \left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2$. Then, we have the following theorem:

Theorem 18.5.1

(i) $E(t_\alpha) = \overline{Y}$ and (ii) $E\left[\widehat{V}(t_\alpha)\right] = V\left(\widehat{\overline{Y}}_{st}\right)$

Proof

(i) $$\begin{aligned} E(t_\alpha) &= E[E(t_\alpha|\Im)] \\ &= E\left[\frac{1}{2^L}\sum_{\alpha=1}^{2^L} t(\alpha)\right] \\ &= E\left[\frac{1}{2^L}\sum_{h=1}^{L} W_h\left\{y_{h1}\sum_{\alpha=1}^{2^L}\phi_{h\alpha} + y_{h2}\sum_{\alpha=1}^{2^L}(1 - \phi_{h\alpha})\right\}\right] \end{aligned}$$

Now noting $\sum_{\alpha=1}^{2^L} \phi_{h\alpha}$ = number of times the unit u_{h1} is repeated in all the 2^L half-samples $= 2^{L-1}$, we find

$$\begin{aligned} E(t_\alpha) &= E\left\{\frac{1}{2}\sum_{h=1}^{L} W_h(y_{h1}+y_{h2})\right\} \\ &= E\left(\widehat{\overline{Y}}_{st}\right) \\ &= \overline{Y} \end{aligned}$$

(ii)
$$\begin{aligned} E\left[\widehat{V}(t_\alpha)\right] &= E\left[E\left(\widehat{V}(t_\alpha)|\mathfrak{J}\right)\right] \\ &= E\left[\frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2\right] \end{aligned} \tag{18.5.7}$$

Now,

$$\begin{aligned} \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2 &= \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left[\sum_{h=1}^{L} W_h\{\phi_{h\alpha}y_{h1} + (1-\phi_{h\alpha})y_{h2}\} - \frac{1}{2}(y_{h1}+y_{h2})\right]^2 \\ &= \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left[\sum_{h=1}^{L} W_h\left\{y_{h1}\left(\phi_{h\alpha}-\frac{1}{2}\right) + \left(\frac{1}{2}-\phi_{h\alpha}\right)y_{h2}\right\}\right]^2 \\ &= \frac{1}{2^L}\frac{1}{4}\sum_{\alpha=1}^{2^L}\left(\sum_{h=1}^{L} W_h d_h \psi_{h\alpha}\right)^2 \end{aligned} \tag{18.5.8}$$

where

$$\psi_{h\alpha} = 2\phi_{h\alpha} - 1 \tag{18.5.9}$$

From expression (18.5.9) we note that $\psi_{h\alpha} = 1$ if the unit u_{h1} is selected in the half-sample α and $\psi_{h\alpha} = -1$ otherwise. Furthermore,

$$\begin{aligned} \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(\sum_{h=1}^{L} W_h d_h \psi_{h\alpha}\right)^2 &= \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(\sum_{h=1}^{L} W_h^2 d_h^2 + \sum_{h\neq}^{L}\sum_{h'=1}^{L} W_h W_{h'} d_h d_{h'} \psi_{h\alpha}\psi_{h'\alpha}\right) \\ &= \sum_{h=1}^{L} W_h^2 d_h^2 \\ &= \widehat{V}\left(\widehat{Y}_{st}\right) \end{aligned} \tag{18.5.10}$$

because $\sum_{\alpha=1}^{2^L} \psi_{h\alpha}\psi_{h'\alpha} = 0$ for $h \neq h'$.

Finally, from Eqs. (18.5.7), (18.5.8), and (18.5.10), we find

$$\begin{aligned} E\left[\widehat{V}(t_\alpha)\right] &= E\left(\frac{1}{4}\sum_{h=1}^{L} W_h^2 d_h^2\right) \\ &= E\left[\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)\right] \\ &= V\left(\widehat{\overline{Y}}_{st}\right) \end{aligned}$$

Hence $\widehat{V}(t_\alpha) = \frac{1}{2^L}\sum_{\alpha=1}^{2^L}\left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2$ is an unbiased estimator of $V\left(\widehat{\overline{Y}}_{st}\right)$. The estimator $\widehat{V}$ cannot be used in practice because one cannot compute all possible 2^L half-samples as 2^L is a huge number, e.g., $L = 6$ produces $2^L = 64$ half-samples. Furthermore, if we choose any random subset of $k(\leq 2^L)$ half-samples and construct an estimator of $V\left(\widehat{\overline{Y}}_{st}\right)$ as

$$\widehat{V} = \frac{1}{k}\sum_{\alpha=1}^{k}\left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2$$

then $\widehat{V}$ will be less efficient than $\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)$ because

$$\begin{aligned} V\left(\widehat{V}(t)\right) &= V\left[E\left(\widehat{V}\middle|\Im\right)\right] + E\left[\widehat{V}\left(\widehat{V}\middle|\Im\right)\right] \\ &= V\left[\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)\right] + E\left[V\left(\widehat{V}\middle|\Im\right)\right] \\ &\geq V\left[\widehat{V}\left(\widehat{\overline{Y}}_{st}\right)\right] \end{aligned} \tag{18.5.11}$$

To overcome this difficulty, we choose a specific subset $\boldsymbol{\Psi}$ of k half-samples from 2^L half subsamples for which

$$\widehat{V}_{BRR} = \frac{1}{k}\sum_{\alpha\in\boldsymbol{\Psi}}\left(t_\alpha - \widehat{\overline{Y}}_{st}\right)^2 = \widehat{V}\left(\widehat{\overline{Y}}_{st}\right) \tag{18.5.12}$$

Now noting from Eq. (18.5.8)

$$\begin{aligned} \widehat{V}_{BRR} &= \frac{1}{4k}\sum_{\alpha\in\boldsymbol{\Psi}}\left(\sum_{h=1}^{L} W_h d_h \psi_{h\alpha}\right)^2 \\ &= \frac{1}{4}\left[\sum_{h=1}^{L} W_h^2 d_h^2 + \frac{1}{k}\sum_{h\neq}^{L}\sum_{h'=1}^{L} W_h W_{h'} d_h d_{h'} \sum_{\alpha\in\boldsymbol{\Psi}}\psi_{h\alpha}\psi_{h'\alpha}\right] \end{aligned}$$

we find that the condition (18.5.12) holds if we choose a subset Ψ of k half-samples that satisfies

$$\sum_{\alpha\in\Psi}\psi_{h\alpha}\psi_{h'\alpha}=0 \tag{18.5.13}$$

A set of half-samples satisfying property (18.5.13) is called balanced half-samples. The method of variance estimation based on a balanced set Ψ of half-samples is known as the Balanced Repeated Replication (BRR) method. The balanced half-samples can be constructed by using Hadamard matrices. A Hadamard matrix is a square matrix of order multiple of 4 with elements $+1$ and -1. The columns of the Hadamard matrices are orthogonal to each other. Details are given by Plackett and Burman (1946). An example of 8×8 Hadamard matrix is given as follows.

Hadamard matrix of order 8

$$\begin{matrix}
+1 & +1 & +1 & +1 & +1 & +1 & +1 & +1\\
+1 & -1 & +1 & -1 & +1 & -1 & -1 & +1\\
+1 & -1 & -1 & +1 & +1 & -1 & +1 & -1\\
+1 & +1 & -1 & -1 & +1 & +1 & -1 & -1\\
+1 & +1 & +1 & +1 & -1 & -1 & -1 & -1\\
+1 & -1 & +1 & -1 & -1 & +1 & +1 & -1\\
+1 & -1 & -1 & +1 & -1 & +1 & -1 & +1\\
+1 & +1 & -1 & -1 & -1 & -1 & +1 & +1
\end{matrix} \tag{18.5.14}$$

We take the rows of a Hadamard matrix as half-samples and columns as strata. The entry $+1$ of the stratum h indicates that the unit u_{h1} is included in the half-sample and -1 indicates the inclusion of u_{h2} in the half-sample. Thus an 8×8 Hadamard matrix can be used for a maximum number of strata $L=8$. If we want to consider any smaller number of strata, say 5, we take just any of the five columns of the Hadamard matrix. Thus, the rows of the Hadamard matrix (18.5.14) form the half-samples

$$\begin{aligned}
&\alpha_1=(u_{11},u_{21},u_{31},u_{41},u_{51},u_{61},u_{71},u_{81}),\ \alpha_2=(u_{11},u_{22},u_{31},u_{42},u_{51},u_{62},u_{72},u_{81}),\\
&\alpha_3=(u_{11},u_{22},u_{32},u_{41},u_{51},u_{62},u_{71},u_{82}),\ \alpha_4=(u_{11},u_{21},u_{32},u_{42},u_{51},u_{61},u_{72},u_{82}),\\
&\alpha_5=(u_{11},u_{21},u_{31},u_{41},u_{52},u_{62},u_{72},u_{82}),\ \alpha_6=(u_{11},u_{22},u_{31},u_{42},u_{52},u_{61},u_{71},u_{82}),\\
&\alpha_7=(u_{11},u_{22},u_{32},u_{41},u_{52},u_{61},u_{72},u_{81}),\ \alpha_8=(u_{11},u_{21},u_{32},u_{42},u_{52},u_{62},u_{71},u_{81}).
\end{aligned}$$

For estimation of variance we choose a set of balanced half-samples consisting of k half-samples from the totality of 2^L half-samples so that

$\widehat{V}_{BRR}$ becomes exactly unbiased for $V\left(\widehat{\overline{Y}}_{st}\right)$. The number k should be as small as possible. A minimal set of balanced samples can be readily obtained from a Hadamard matrix of order $k \times k$ by choosing any L columns excluding the column of all $+1$'s, where $L+1 \leq k \leq L+4$ (Rao and Shao, 1999).

An estimator for $\overline{Y}$ based on k half-samples is given by

$$\begin{aligned} t_{BRR} &= \frac{1}{k}\sum_{\alpha \in \Psi} t(\alpha) \\ &= \frac{1}{k}\sum_{h=1}^{L} W_h\left\{ y_{h1}\sum_{\alpha \in \Psi}\left(\frac{1}{2}+\frac{\psi_{h\alpha}}{2}\right) + y_{h2}\sum_{\alpha \in \Psi}\left(\frac{1}{2}-\frac{\psi_{h\alpha}}{2}\right)\right\} \\ &= \widehat{\overline{Y}}_{st} + \frac{1}{2k}\sum_{h=1}^{L} W_h(y_{h1}-y_{h2})\sum_{\alpha \in \Psi}\psi_{h\alpha} \end{aligned}$$

Thus $t_{BRR} = \widehat{\overline{Y}}_{st}$ if

$$\sum_{\alpha \in \Psi}\psi_{h\alpha} = 0 \qquad (18.5.15)$$

A set of balanced half-samples satisfying the properties of Eqs. (18.5.13) and (18.5.15) is known as full orthogonal balanced half-samples. For full orthogonal balanced half-samples, k should be a multiple of 4 and $k > L$. For example if $L = 8$, one needs minimum 12 replicates to achieve full orthogonal balance. Thus for reduction of computation labor, we should choose k as a smallest multiple of 4 but greater than L. If $k = L$, then only the balance is achieved but the full orthogonal balance is not. In case k is less than L, neither the balance nor the full orthogonal balance is achieved. Because Hadamard matrices are not unique, full orthogonal balance may be achieved with alternative sets of half-samples.

18.5.2 Methods of Variance Estimation

Let $\widehat{\theta}_{st}$ be an estimator (not necessarily linear) of a population parameter θ based on a stratified sample of 2 units per stratum using any sampling design. Here we choose a set of k balanced half-samples. Let $\widehat{\theta}_{\alpha}$ be an estimator of θ based on the αth half-sample, $\alpha = 1,\ldots,k$ and $\overline{\theta} = \sum_{\alpha=1}^{k}\widehat{\theta}_{\alpha}\Big/k$. The estimators $\widehat{\theta}_{\alpha}$ should necessarily be the same functional form of the estimator

$\widehat{\theta}_{st}$. Then the variance or mean square estimator of $\widehat{\theta}_{st}$ can be estimated using any of the following formula.

$$\widehat{V}_{BRR}(1) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}_\alpha - \widehat{\theta}_{st}\right)^2 \tag{18.5.16}$$

$$\widehat{V}_{BRR}(2) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}_\alpha - \overline{\theta}\right)^2 \tag{18.5.17}$$

Let $\widehat{\theta}^*_\alpha$ be an estimator of θ based on the complement of the αth half-sample. We can then also get the following alternative estimators

$$\widehat{V}_{BRR}(3) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}^*_\alpha - \widehat{\theta}_{st}\right)^2 \tag{18.5.18}$$

$$\widehat{V}_{BRR}(4) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}^*_\alpha - \overline{\theta}^*\right)^2 \tag{18.5.19}$$

$$\widehat{V}_{BRR}(5) = \frac{1}{4k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}_\alpha - \widehat{\theta}^*_\alpha\right)^2 \tag{18.5.20}$$

where $\overline{\theta}^* = \sum_{\alpha=1}^{k}\widehat{\theta}^*_\alpha \Big/ k$.

The estimators of the variance of a linear estimator described previously based on a set of balanced half-samples do not provide a more improved estimator than the conventional variance estimator of the stratified sampling. However, the BRR method can be gainfully used for nonlinear estimators where elegant and unbiased variance estimators are not available. For very large L it is difficult to find balanced replicates and in this situation one may use partially balanced replicates by dividing L into a number of groups. Details are given by Wolter (1985).

18.5.3 Applications

18.5.3.1 Population Ratio

Consider a stratified sampling where two units are selected at random from each of the stratum. The conventional combined stratified ratio estimator of the population ratio $R = Y/X$ is $\widehat{R}_{com} = \widehat{Y}_{st}\big/\widehat{X}_{st}$, where $\widehat{Y}_{st} = \sum_{h=1}^{L} N_h\overline{y}_h$, $\widehat{X}_{st} = \sum_{h=1}^{L} N_h\overline{x}_h$. An estimator R based on αth half-sample is $\widehat{R}_{\alpha(com)} = \widehat{Y}_\alpha\big/\widehat{X}_\alpha$, where $\widehat{Y}_\alpha = \sum_{h=1}^{L} N_h\ \{\phi_{h\alpha}y_{h1} + (1-\phi_{h\alpha})y_{h2}\}$ and

$\widehat{X}_{\alpha} = \sum_{h=1}^{L} N_h\{\phi_{h\alpha}x_{h1} + (1-\phi_{h\alpha})x_{h2}\}$. The estimator for R based on a set of k balanced half-samples is given by

$$\widehat{R}_{BRR} = \frac{1}{k}\sum_{\alpha=1}^{k} \widehat{Y}_{\alpha}/\widehat{X}_{\alpha}$$

An estimator for the variance of $\widehat{R}_{com}$ based on k balanced half-samples is

$$\widehat{V}_{BRR} = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{R}_{\alpha(com)} - \widehat{R}_{com}\right)^2$$

Clearly, $\widehat{V}_{BRR}$ is quite different from the conventional variance estimator

$$\widehat{V}_{st}\left(\widehat{R}_{com}\right) = \frac{1}{4\widehat{X}_{st}^{2}}\sum_{h=1}^{L} N_h^2\left\{(y_{h1} - y_{h2}) - \widehat{R}_{com}(x_{h1} - x_{h2})\right\}^2.$$

18.5.3.2 Inclusion Probability Proportional to Size Sampling Scheme

Suppose from each of the stratum two units are selected with IPPS sampling design. Let the inclusion probability for the ith unit of the hth stratum be $\pi_{hi} = 2p_{hi}$. Here the conventional estimator for the total Y and its variance are, respectively, given by

$$\widehat{Y}_{st} = \frac{1}{2}\sum_{h=1}^{L}\left(\frac{y_{h1}}{p_{h1}} + \frac{y_{h2}}{p_{h2}}\right)$$

and

$$V\left(\widehat{Y}_{st}\right) = \frac{1}{4}\sum_{h=1}^{L}\sum_{i\neq}^{N_h}\sum_{j=1}^{N_h}\frac{(p_{hi} - p_{hj})^2}{\pi_{hij}}\left(\frac{y_{hi}}{p_{hi}} - \frac{y_{hj}}{p_{hj}}\right)^2$$

The estimator based on the αth half-sample is

$$t_{\alpha} = \sum_{h=1}^{L}\left\{\phi_{h\alpha}\frac{y_{h1}}{p_{h1}} + (1-\phi_{h\alpha})\frac{y_{h2}}{p_{h2}}\right\} \quad \text{for } \alpha = 1, \ldots, k$$

The variance estimator of $\widehat{Y}_{st}$ based on BRR is

$$\widehat{V}_{BRR} = \frac{1}{k}\sum_{\alpha=1}^{k}\left(t_{\alpha} - \widehat{Y}_{st}\right)^2$$

18.5.4 Numerical Example

The following table relates to the number of patients (y) treated per day and the number of doctors (x) at the clinics in a certain city. The clinics are classified into five different zones (strata). From each of the zones two clinics are selected by the SRSWR method (Table 18.5.1).

Table 18.5.1 Number of patients treated and number of doctors

Zone	**1**		**2**		**3**		**4**		**5**	
Number of clinics	15		25		20		30		10	
Sampled clinics	1	2	1	2	1	2	1	2	1	2
y	30	20	30	25	40	15	40	50	20	20
x	2	1	3	2	2	1	2	3	2	1

18.5.4.1 Population Mean

Here the parameter of interest is $\theta = \sum W_h \overline{Y}_h = \overline{Y} =$ average number of patients treated per clinic per day. An unbiased estimator of θ is $\widehat{\theta} = \overline{y}_{st} = \sum W_h \overline{y}_h = 0.15 \times 25 + 2.5 \times 27.5 + 0.2 \times 27.5 + 0.3 \times 45 + 0.1 \times 20 = 31.625$. The estimated variance of $\overline{y}_{st}$ is

$$\widehat{V}\left(\overline{y}_{st}\right) = \sum W_h^2 (y_{h1} - y_{h2})^2 / 4 = 9.453$$

Let us now consider the BRR method with $L = 5$ and $k = 8$. For this method we delete the first and the last two columns of the Hadamard matrix (18.5.14) to achieve orthogonal balance. The Hadamard matrix (after deletion of columns) and the units belonging to the replicates are given in Tables 18.5.2 and 18.5.3.

Table 18.5.2 BRR method of selection of sample

						Units belong to the replicates				
	Stratum *h*					**Stratum *h***				
Replicates (α)	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	+1	+1	+1	+1	+1	u_{11}	u_{21}	u_{31}	u_{41}	u_{51}
2	−1	+1	−1	+1	−1	u_{12}	u_{21}	u_{32}	u_{41}	u_{52}
3	−1	−1	+1	+1	−1	u_{12}	u_{22}	u_{31}	u_{41}	u_{52}
4	+1	−1	−1	+1	+1	u_{11}	u_{22}	u_{32}	u_{41}	u_{51}
5	+1	+1	+1	−1	−1	u_{11}	u_{21}	u_{31}	u_{42}	u_{52}
6	−1	+1	−1	−1	+1	u_{12}	u_{21}	u_{32}	u_{42}	u_{51}
7	−1	−1	+1	−1	+1	u_{12}	u_{22}	u_{31}	u_{42}	u_{51}
8	+1	−1	−1	−1	−1	u_{11}	u_{22}	u_{32}	u_{42}	u_{52}

Table 18.5.3 BRR esimates

(α)	$\overline{y}_{st}(\alpha)$	$\overline{x}_{st}(\alpha)$	$\widehat{R}_{sep}(\alpha)$	$\widehat{R}_{com}(\alpha)$
1	34.00	2.25	15.750	15.1111
2	27.50	1.80	16.500	15.2777
3	31.25	1.75	18.125	17.8571
4	27.75	1.80	15.375	15.4166
5	37.00	2.45	15.750	15.1020
6	30.50	2.20	14.500	13.8636
7	34.25	2.15	16.125	15.9302
8	30.75	2.00	15.375	15.3750
Mean	31.625	2.05	15.9375	15.4917

$\overline{y}_{BRR} = \sum_{\alpha=1}^{8} \overline{y}_{st}(\alpha)/8 = 31.625 = \overline{y}_{st}$. An estimated variance of $\overline{y}_{st}$ by the BRR method is $\widehat{V}\left(\overline{y}_{BRR}\right) = \sum_{\alpha=1}^{k} \left\{\overline{y}_{st}(\alpha) - \overline{y}_{st}\right\}^2/k = 9.453 \; = \widehat{V}\left(\overline{y}_{st}\right)$. Because the replicates are not only balanced but also orthogonally balanced, we find for linear estimates $\overline{y}_{BRR} = \overline{y}_{st}$ and $\widehat{V}\left(\overline{y}_{BRR}\right) = \widehat{V}\left(\overline{y}_{st}\right)$.

18.5.4.2 Population Ratio

Here the parameter of interest is $\theta =$ average number of patients treated by a doctor per day $= R = Y/X$. The combined and separate ratio estimators of R are, respectively, given by $\widehat{R}_{com} = \overline{y}_{st}/\overline{x}_{st} = 31.625/2.05 = 15.426$ and $\widehat{R}_{sep} = \sum W_h \widehat{R}_h = 15.65$, where $\widehat{R}_h = \overline{y}_h/\overline{x}_h$.

The estimated approximate variances or mean square errors of $\widehat{R}_{com}$ and $\widehat{R}_{sep}$ are, respectively,

$$\widehat{V}\left(\widehat{R}_{com}\right) = \frac{1}{4(\overline{x}_{st})^2} \sum W_h^2 \left[\left(y_{h1} - \widehat{R}_{com} x_{h1}\right) - \left(y_{h2} - \widehat{R}_{com} x_{h2}\right)\right]^2 = 0.9609$$

$$\widehat{V}\left(\widehat{R}_{sep}\right) = \frac{1}{4(\overline{x}_{st})^2} \sum W_h^2 \left[\left(y_{h1} - \widehat{R}_{sep} x_{h1}\right) - \left(y_{h2} - \widehat{R}_{sep} x_{h2}\right)\right]^2 = 0.9890$$

The combined ratio estimates based on the balanced sample is given by $\widehat{R}_{BRR(com)} = \sum_{\alpha=1}^{8} \widehat{R}_{com}(\alpha)/8 = 15.4917,$ which is not equal to $\widehat{R}_{com} = 15.426$. The variance of $\widehat{R}_{com}$ can be estimated using any of the following formulae

$$\widehat{V}_{BRR(com)}(1) = \frac{1}{8} \sum_{\alpha=1}^{8} \left(\widehat{R}_{com}(\alpha) - \widehat{R}_{com}\right)^2 = 1.1041$$

$$\widehat{V}_{BRR(com)}(2)=\frac{1}{8}\sum_{\alpha=1}^{8}\left(\widehat{R}_{com}(\alpha)-\overline{\widehat{R}}_{com}\right)^2=1.099$$

The separate ratio estimates based on the balanced sample is given by $\widehat{R}_{BRR}(sep)=\sum_{\alpha=1}^{8}\widehat{R}_{sep}(\alpha)/8=15.9375$, which is not equal to $\widehat{R}_{sep}=15.65$. The variance of $\widehat{R}_{sep}$ can be estimated by using any of the following formulae

$$\widehat{V}_{BRR(sep)}(1)=\frac{1}{8}\sum_{\alpha=1}^{8}\left(\widehat{R}_{sep}(\alpha)-\widehat{R}_{sep}\right)^2=1.071$$

$$\widehat{V}_{BRR(sep)}(2)=\frac{1}{8}\sum_{\alpha=1}^{8}\left(\widehat{R}_{sep}(\alpha)-\overline{\widehat{R}}_{sep}\right)^2=0.998$$

Thus we see for the nonlinear statistics that the BRR estimates are quite different from the conventional estimates based on the full sample.

18.5.4.3 Correlation Coefficient

The sample correlation coefficient between x and y is given by

$$r=\frac{\sum_h N_h(y_{h1}x_{h1}+y_{h2}x_{h2})-2N\overline{y}_{st}\overline{x}_{st}}{\sqrt{\left(\sum_h N_h\left(y_{h1}^2+y_{h2}^2\right)-2N\overline{y}_{st}^2\right)}\sqrt{\left(\sum_h N_h\left(x_{h1}^2+x_{h2}^2\right)-2N\overline{x}_{st}^2\right)}}$$
$$=0.717$$

The correction coefficient r_α and r_α^* based on the αth half-sample and its complementary sample are given in Table 18.5.4.

Table 18.5.4 BRR method of variance estimation

α	r_α	r_α^*	$\frac{(r_\alpha-r_\alpha^*)^2}{4}$
1	−0.3481	0.9543	0.42406
2	0.6356	0.8405	0.01049
3	0.7276	0.8012	0.00135
4	0.6866	0.7003	0.00004
5	0.5726	0.6960	0.00380
6	0.8141	0.5547	0.01682
7	0.8480	0.5824	0.01763
8	0.9604	0.2414	0.12924
Mean	0.6121	0.6713	0.0754

The estimates of variance of r obtained by using different formulae are as follows: $\widehat{V}_{BRR}(1)=\frac{1}{k}\sum_{\alpha=1}^{k}(r_\alpha-r)^2=0.1562$, $\widehat{V}_{BRR}(2)=\frac{1}{k}\sum_{\alpha=1}^{k}(r_\alpha-\bar{r})^2=0.1451$, $\widehat{V}_{BRR}(3)=\frac{1}{k}\sum_{\alpha=1}^{k}\left(r_\alpha^*-r\right)^2=0.0438$, $\widehat{V}_{BRR}(4)=\frac{1}{k}\sum_{\alpha=1}^{k}\left(r_\alpha^*-\bar{r}^*\right)^2=0.0416$, $\widehat{V}_{BRR}(5)=\frac{1}{4k}\sum_{\alpha=1}^{k}\left(r_\alpha-r_\alpha^*\right)^2=0.0754$, $\widehat{V}_{BRR}(6)=\left(\widehat{V}_{BRR}(1)+\widehat{V}_{BRR}(3)\right)/2=0.1000$, and $\widehat{V}_{BRR}(7)=\left(\widehat{V}_{BRR}(2)+\widehat{V}_{BRR}(4)\right)/2=0.0933$, where $\bar{r}=\frac{1}{k}\sum_{\alpha=1}^{k}r_\alpha=0.6121$, $\bar{r}^*=\frac{1}{k}\sum_{\alpha=1}^{k}r_\alpha^*=0.6713$, and $k=8$.

18.5.5 Stratum Size $n_h \geq 2$

18.5.5.1 Grouped Balanced Half-Sample Method

Let $\overline{Y}_h$ and $\overline{y}_h$ be the population mean and sample mean of the hth stratum, respectively. The parameter of interest is $\theta=g\left(\overline{\mathbf{Y}}\right)$, a function of $\overline{\mathbf{Y}}=\left(\overline{Y}_1,\ldots,\overline{Y}_L\right)$ the vector of the strata means. Let a full sample estimate of θ is $\widehat{\theta}=g(\overline{\mathbf{y}})$, where $\overline{\mathbf{y}}=\left(\overline{y}_1,\ldots,\overline{y}_L\right)$. In the grouped balanced half-sample (GBHS) method, the selected sample $n_h(>2)$ units from the stratum h are divided at random into two groups of sizes $m_{h1}=[n_h/2]$ and $m_{h2}=n_h-[n_h/2]$, respectively. A set of k balanced half-samples of groups are selected. Let $\overline{y}_{h1}$ and $\overline{y}_{h2}$ denote the sample means of the first and second groups, respectively. The estimator of the population mean θ based on the αth half-sample is denoted by

$$\widehat{\theta}_\alpha=g(\overline{\mathbf{y}}_\alpha^{'})$$

where $\overline{\mathbf{y}}_\alpha^{'}=(\overline{\mathbf{y}}_{1\alpha}^{'},\ldots,\overline{\mathbf{y}}_{h\alpha}^{'},\ldots,\overline{\mathbf{y}}_{L\alpha}^{'})$, $\overline{\mathbf{y}}_{h\alpha}^{'}=\phi_{h\alpha}\overline{y}_{h1}+(1-\phi_{h\alpha})\overline{y}_{h2}$, $\phi_{h\alpha}=1$ if the group 1 of the stratum h is selected in the half-sample and zero otherwise.

A GBHS variance estimator of $\widehat{\theta}$ is given by

$$\widehat{V}_{GB}\left(\widehat{\theta}\right)=\frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}_\alpha-\widehat{\theta}\right)^2 \tag{18.5.21}$$

The variance estimator $\widehat{V}_{GB}\left(\widehat{\theta}\right)$ in general is quite different from the usual unbiased variance estimator $\widehat{V}_{st}=\sum_{h=1}^{L}W_h^2s_h^2/n_h$, where s_h^2 is the hth stratum sample variance. However, it should be noted that for the linear function $\theta = \sum_{h=1}^{L} W_h \overline{Y}_h = \overline{Y}$, $\widehat{\theta} = \sum_{h=1}^{L} W_h \overline{y}_h = \widehat{\overline{Y}}_{st}$ with $\overline{y}_h$ as the hth stratum sample mean, the GBHS variance estimator (18.5.21) reduces to

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{4}\sum_{h=1}^{L} W_h^2 \left(\overline{y}_{h1} - \overline{y}_{h2}\right)^2 = \widehat{V}_{st}$$

Rao and Shao (1996) have shown that the GBHS method leads to asymptotically incorrect inferences as strata sample sizes $n_h \to \infty$ with L fixed. To overcome this difficulty, they proposed an alternative method known as repeatedly grouped balanced half-sample (RGBHS) method where the random grouping is repeated T times independently and then taking average of the GBHS variance estimators $\widehat{V}^{t}_{GB}\left(\widehat{\theta}\right)$, obtained from the tth grouping. The resulting RGBHS variance estimator

$$\widehat{V}_{RGB}\left(\widehat{\theta}\right) = \frac{1}{T}\sum_{t=1}^{T} \widehat{V}^{t}_{GB}\left(\widehat{\theta}\right) \tag{18.5.22}$$

possesses asymptotic validity. Rao and Shao (1996) proposed modification of $\widehat{V}_{GB}\left(\widehat{\theta}\right)$ by replacing W_h by $W_h^{'} = W_h\sqrt{\lambda_h}$ with $\lambda_h = 1 - n_h/N_h$ in the calculation of $\widehat{\theta}_{\alpha}$ and $\widehat{\theta}$.

18.5.5.2 Subdivision of Strata

This method is suitable with a small number of strata and relatively large sample sizes within strata. Let us assume that the stratum size $n_h = 2m_h$ with m_h is an integer for all h. The hth stratum is subdivided into m_h artificial strata each of size 2 so that the total number of strata is $H = \sum_h m_h$. A balanced set of k half-samples is obtained by using a Hadamard matrix $k \times k$ $\left(\sum_h m_h \le k \le \sum_h m_h + 3\right)$. The proposed variance estimator of $\widehat{\theta}$ is

$$\widehat{V}_{SG}\left(\widehat{\theta}\right) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}_{\alpha^*} - \widehat{\theta}\right)^2 \tag{18.5.23}$$

where $\widehat{\theta}_{\alpha^*} = g(\overline{\mathbf{y}}^{'}_{\alpha^*})$, $\overline{\mathbf{y}}^{'}_{\alpha^*} = (\overline{\mathbf{y}}^{'}_{1\alpha^*}, \ldots, \overline{\mathbf{y}}^{'}_{h\alpha^*}, \ldots, \overline{\mathbf{y}}^{'}_{H\alpha^*})$, $\overline{\mathbf{y}}^{'}_{h\alpha^*} = \frac{1}{m_h}\sum_{i=1}^{m_h}\{\phi_{hi\alpha} y_{hi1} + (1 - \phi_{hi\alpha}) y_{hi2}\}$, $\phi_{hi\alpha} = 1$ if the first unit of the artificial strata with y-value y_{hi1} belongs to the half-sample α and $\phi_{hi\alpha} = 0$ if the second unit of the artificial strata with y-value y_{hi2} belongs to the half-sample α.

For the linear estimator $\widehat{\theta} = \sum_h W_h \bar{y}_h = \widehat{\bar{Y}}_{st}$, the variance estimator (18.5.23) reduces to

$$\widehat{V}_{SG}\left(\widehat{\theta}\right) = \sum_{h=1}^{L} \frac{W_h^2}{m_h^2} \sum_{i=1}^{m_h} \frac{\left(y_{hi1} - y_{hi2}\right)^2}{4}$$

The variance estimator $\widehat{V}_{SG}\left(\widehat{\theta}\right)$ does not agree with the usual unbiased variance estimator $\widehat{V}_{st} = \sum_{h=1}^{L} W_h^2 s_h^2 / n_h$.

18.5.6 Stratified Multistage Sampling

Consider a stratified multistage sampling design where the population comprises of large number of L strata. The hth strata consists of N_h clusters. The ith cluster of the hth stratum comprises with M_{hi} ultimate units. A sample of n_h clusters is selected from the hth stratum with probability proportional to sizes without replacement (PPSWOR) method and each of the selected clusters are subsampled independently using the PPSWOR method again. Let y_{hik} be the value of the study variable y associated with (h,i,k), the kth ultimate unit of the ith cluster of the hth stratum. Let w_{hik} be the survey weights attached to (h,i,k) if it is included in the selected sample s. Consider the class of unbiased estimators of the population total $Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{ijk}$ of the form

$$\widehat{Y} = \sum_{(h,i,k) \in s} w_{hik} y_{hik} \tag{18.5.24}$$

where w_{hik} is the weight associated with (h,i,k).

Suppose we are interested in estimating parametric function of the form $\theta = g(\mathbf{A})$ where $\mathbf{A}$ is a vector of population totals. For example, $\mathbf{A} = (A_1, A_2, A_3, A_4, A_5)$ with $A_1 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} x_{hik}$, $A_2 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik}$, $A_3 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} x_{hik}^2$, $A_4 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik}^2$, and $A_5 = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} x_{hik} y_{hik}$, we can express θ as a population ratio, variance, correlation coefficient, and coefficient of variation of two characters x and y, among others. Let $\widehat{\mathbf{A}} = \left(\widehat{A}_1, \ldots, \widehat{A}_5\right)$ be a consistent or unbiased estimator of $\mathbf{A}$, where $\widehat{A}_1 = \sum_{ijk \in s} w_{ijk} x_{ijk}$, $\widehat{A}_2 = \sum_{ijk \in s} w_{ijk} y_{ijk}$, $\widehat{A}_3 = \sum_{ijk \in s} w_{ijk} x_{ijk}^2$, $\widehat{A}_4 = \sum_{ijk \in s} w_{ijk} y_{ijk}^2$, $\widehat{A}_5 = \sum_{ijk \in s} w_{ijk} x_{ijk} y_{ijk}$, and w_{ijk} are suitably chosen weight. Consider the estimator $\widehat{\theta}^* = g\left(\widehat{\mathbf{A}}\right)$ of θ based on the full sample s as $\widehat{\theta}^* = g^*\left(\widehat{\mathbf{A}}\right)$. For variance estimation, we select two units ($n_h = 2$) from each of the stratum $h(=1,\ldots, L)$ and construct k balanced half-samples using Hadamard matrix as described in Section 18.5.1. The estimator $\widehat{\theta}^*_\alpha$ is obtained from the αth half-sample by using the same formula $\widehat{\theta}^*$ with weight w_{hik} changed to w^α_{hik}, which is equal to $2w_{hik}$ or 0 according to whether or not (h,i) cluster is selected in the αth half-sample. The variance estimator of $\widehat{\theta}^*$ is given by

$$\widehat{V}\left(\widehat{\theta}^*\right) = \frac{1}{k}\sum_{\alpha=1}^{k}\left(\widehat{\theta}^*_\alpha - \widehat{\theta}^*\right)^2 \qquad (18.5.25)$$

For $n_h > 2$, variance estimator can be obtained by constructing k balanced half-samples of clusters following Sections 18.5.5.1 or 18.5.5.2 and adjusting weight appropriately.

18.5.7 Fay's Method

Consider stratified sampling where $n_h = 2$ units are selected from each of the strata by the SRSWOR method. In this case the conventional estimator for the population mean $\overline{Y}$ is $\overline{Y}_{st} = \sum_{h=1}^{L} W_h \dfrac{(y_{h1} + y_{h2})}{2}$. The weight associated with each of the units of the hth stratum is $W'_h = W_h/2$. The estimator for the population mean based on the half-sample α defined in Eq. (18.5.6) is

$$t_\alpha = \sum_{h=1}^{L} 2W'_h\{\phi_{h\alpha}y_{h1} + (1-\phi_{h\alpha})y_{h2}\}$$

Here, a weight $2W'_h$ is attached to each of the units selected in the half-sample α and a zero weight is attached to the units not selected in the half-sample α. In Fay's (1989) adjustment of weights, the selected unit in the half-sample is given a less weight $W'_h(1+\in)$ with $0 < \in \leq 1$ if it is selected in the half-sample α while a positive weight $W'_h(1-\in)$ is assigned to the units not selected in the half-sample. So, Fay's estimator based on the half-sample α is

$$t^F_\alpha = \sum_{h=1}^{L}\frac{W_h}{2}[\{1+\in(2\phi_{h\alpha}-1)\}y_{h1} + \{1-\in(2\phi_{h\alpha}-1)\}y_{h2}] \quad (18.5.26)$$

where $\phi_{h\alpha} = 1$ if the unit u_{h1} is selected in the half-sample α and $\phi_{h\alpha} = 0$ if u_{h2} is selected in the half-sample.

The Fay's adjusted variance estimator of $\widehat{\overline{Y}}_{st}$ based on a set k balanced half-samples is given by

$$\widehat{V}^F_{BRR} = \frac{1}{k\in^2}\sum_{\alpha=1}^{k}\left(t^F_\alpha - \widehat{\overline{Y}}_{st}\right)^2 \qquad (18.5.27)$$

Theorem 18.5.2

For a set of k balanced half-samples

$$\widehat{V}^{F}_{BRR} = \frac{1}{4}\sum_{h=1}^{L} W_h^2 (y_{h1} - y_{h2})^2 = \widehat{V}\left(\overline{Y}_{st}\right)$$

Proof

$$\widehat{V}^{F}_{BRR} = \frac{1}{k\in^2}\sum_{\alpha=1}^{k}\left[\sum_{h=1}^{L}\frac{W_h}{2}\{(1+\in(2\phi_{h\alpha}-1))y_{h1} + (1-\in(2\phi_{h\alpha}-1))y_{h2} - (y_{h1}+y_{h2})\}\right]^2$$

$$= \frac{1}{4k}\sum_{\alpha=1}^{k}\left[\sum_{h=1}^{L} W_h\psi_{h\alpha}(y_{h1}-y_{h2})\right]^2 \quad \text{where } \psi_{h\alpha} = 2\phi_{h\alpha} - 1$$

$$= \frac{1}{4k}\sum_{\alpha=1}^{k}\left[\sum_{h=1}^{L} W_h^2\psi_{h\alpha}^2(y_{h1}-y_{h2})^2 + \sum_{h\neq}^{L}\sum_{h'}^{L} W_h\psi_{h\alpha}W_{h'}\psi_{h'\alpha}(y_{h1}-y_{h2})(y_{h'1}-y_{h'2})\right]$$

$$= \frac{1}{4k}\left[\sum_{h=1}^{L} W_h^2(y_{h1}-y_{h2})^2\sum_{\alpha=1}^{k}\psi_{h\alpha}^2 + \sum_{h\neq}^{L}\sum_{h'}^{L} W_hW_{h'}(y_{h1}-y_{h2}) (y_{h'1}-y_{h'2})\sum_{\alpha=1}^{k}\psi_{h\alpha}\psi_{h'\alpha}\right]$$

Now, noting $\psi_{h\alpha} = 1$ if u_{h1} belongs to the half-sample α and $\psi_{h\alpha} = -1$ if u_{h1} does not belong to the half-sample α, and $\sum_{\alpha=1}^{k}\psi_{h\alpha}\psi_{h'\alpha} = 0$ (vide Eq. 18.5.13) for the set of k balanced half-samples, we find $\widehat{V}^{F}_{BRR} = \widehat{V}\left(\overline{Y}_{st}\right)$.

For the multistage sampling design described in Section 18.5.6, the Fay-adjusted variance estimator based on a set of k balanced half-samples is given by

$$\widehat{V}^{F}\left(\widehat{\theta}^{*}\right) = \frac{1}{k\in^2}\sum_{\alpha=1}^{k}\left(\widehat{\theta}^{*F}_{\alpha} - \widehat{\theta}^{*}\right)^2$$

where $\widehat{\theta}^{*F}_{\alpha}$ is computed using the formula for $\widehat{\theta}^{*}$ but with w_{hik} changed to $w^F_{hik} = w_{hik}(1+\in)$ if the (hik) is included in the half-sample α otherwise $w^F_{hik} = w_{hik}(1-\in)$. When $\widehat{\theta}^{*}$ is linear, $\widehat{V}^{F}\left(\widehat{\theta}^{*}\right)$ reduces to the standard variance estimator (Judkins, 1990).

18.6 BOOTSTRAP METHOD

18.6.1 Bootstrap for Infinite Population

The Bootstrap (BT) method was introduced by Efron (1979). This method can be used for the estimation of variance of an estimator and determination of confidence interval of the parameter of interest. Let $X_1, \ldots, X_n$ be a random sample from a population with distribution function F_θ indexed by a parameter θ. Let $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ be an estimator of the parameter θ. The empirical distribution of F_n of F is obtained by assigning a mass $1/n$ to each of the observations $X_1, \ldots, X_n$. From the empirical distribution F_n, a sample $X_1^*, \ldots, X_n^*$ of size n is selected with replacement. Let $\widehat{\theta}^* = \widehat{\theta}^*\left(X_1^*, \ldots, X_n^*\right)$ be an estimator of θ based on the BT sample $X_1^*, \ldots, X_n^*$. The selection of the BT sample is then repeated independently a large number of times B (at least 1000 times). Let $\widehat{\theta}^{b}_{*}$ be the value of $\widehat{\theta}^*$ based on the BT sample $b(=1, \ldots, B)$. The variance of this distribution of $\widehat{\theta}^*$ is considered as an estimator of the variance of $\widehat{\theta}$ and it is given by

$$\widehat{V}_B\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \overline{\theta}^*\right)^2 \tag{18.6.1}$$

where $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*$.

In the Eq. (18.6.1), $\overline{\theta}^*$ may be replaced by $\widehat{\theta}$.

18.6.1.1 Bootstrap Confidence Interval

To determine the confidence interval of θ, we may use (i) percentile method and (ii) BT t-method.

18.6.1.1.1 Percentile Method

In the percentile method, we arrange the values of $\widehat{\theta}_b^*$ in the ascending order of magnitude. Let $\widehat{\theta}_{b,U}^*$ and $\widehat{\theta}_{b,L}^*$ be the upper and lower $(\alpha/2)100$ percent points of the distribution of $\widehat{\theta}_b^*$, respectively. Then 100 $(1 - \alpha)\%$ BT confidence interval of θ is given by $\left(\widehat{\theta}_{b,L}^*, \widehat{\theta}_{b,U}^*\right)$.

18.6.1.1.2 Bootstrap *t*-Method

In BT t-method we compute $t_b = \left(\widehat{\theta}_b^* - \widehat{\theta}\right)\Big/\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)}$ for $b = 1, \ldots, B$ and assume distribution of t_b approximately the same as $t = \left(\widehat{\theta} - \theta\right)\Big/\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)}$. The $100(1 - \alpha)\%$ confidence interval of θ is given by $\left(\widehat{\theta} - t_U^*\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)}, \widehat{\theta} - t_L^*\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)}\right)$, where t_L^* and t_U^* are lower and upper $100\alpha/2$ percent points of the BT histogram generated by the values of t_b. Although the BT t-method involves tedious computation, it is better in terms of coverage probability.

18.6.2 Bootstrap for Finite Population

Application of BT technique in finite sampling, especially in complex surveys, is not straightforward as demonstrated by Rao and Wu (1988), Gross (1980), Bickel and Freedman (1984), and Sitter (1992a,b), among others. However, few applications are presented here.

18.6.2.1 Bootstrap for Simple Random Sampling With Replacement

Suppose a sample s of n units is selected from a finite population of N units by the SRSWR method and let $\mathbf{y}(n) = (y_1,\ldots, y_i,\ldots,y_n)$ be the observed values of the character y. Our objective is to estimate a population parameter θ, such as population mean, median, coefficient of variation, etc. Suppose that $\widehat{\theta} = \widehat{\theta}(\mathbf{y}(n))$ is an estimator of θ. We select a with-replacement random sample of size n from s assuming all the elements in s are distinct. Let us denote the selected sample by s^* and let the observed y-values be denoted by $\mathbf{y}^*(n) = \left(y_1^*,\ldots,y_i^*,\ldots,y_n^*\right)$. Let $\widehat{\theta}_{s^*} = \widehat{\theta}\left(\mathbf{y}^*(n)\right)$ be an estimator of θ based on the sample s^*. The selection of sample s^* is then repeated independently a large number of B times. The BT variance of $\widehat{\theta}$ may be calculated by using any of the following formulae

$$\widehat{V}_B(1) = \sum_{b=1}^{B}\left(\widehat{\theta}_{s^*}^{b} - \widehat{\overline{\theta}}\right)^2\Big/B \text{ or } \widehat{V}_B(2) = \sum_{b=1}^{B}\left(\widehat{\theta}_{s^*}^{b} - \widehat{\theta}\right)^2\Big/B \quad (18.6.2)$$

where $\widehat{\theta}_{s^*}^{b}$ is the value of $\widehat{\theta}_{s^*}$ based on bth BT sample $b = 1,\ldots,B$ and $\widehat{\overline{\theta}} = \sum_{b=1}^{B}\widehat{\theta}_{s^*}^{b}\Big/B.$

The BT confidence interval of θ may be computed using any of the formulae

$$\left(\widehat{\theta}_L^*, \widehat{\theta}_U^*\right) \text{ and } \left(\widehat{\theta} - t_U^*\sqrt{\widehat{V}_B(\bullet)}, \widehat{\theta} - t_L^*\sqrt{\widehat{V}_B(\bullet)}\right) \quad (18.6.3)$$

where $\widehat{\theta}_L^*$ and $\widehat{\theta}_U^*$ are the lower and upper $(\alpha/2)100$ percent points of the values of $\widehat{\theta}_{s_*}^{b}$; t_L^* and t_U^* denote lower and upper $(\alpha/2)100$ points of BT t distribution; $\widehat{V}_B(\bullet)$ may be taken as $\widehat{V}_B(1)$ or $\widehat{V}_B(2)$.

Example 18.6.1

Let the daily wages (in US$) of random sample s of five factory workers be as follows:

$$20, 30, 50, 80, \text{ and } 40$$

Suppose we want to estimate the population coefficient of variation (θ) and confidence intervals of θ by BT method, we proceed as follows:

First, we compute the coefficient of variation based on the full sample s as $\widehat{\theta} = 0.5232$. Then we select $B = 60$ BT samples each of size 5 from the sample s by the SRSWR method and compute the sample coefficient of variation $\widehat{\theta}_{s^*}^{b}$ as follows:

b	**bootstrap sample *s****					$\widehat{\theta}^{b}_{s^*}$	t_b	***b***	**bootstrap sample *s****					$\widehat{\theta}^{b}_{s^*}$	t_b	***b***	**bootstrap sample *s****					$\widehat{\theta}^{b}_{s^*}$	t_b
1	30	80	30	80	40	0.498	−0.195	21	30	20	20	50	80	0.637	0.879	41	40	50	80	20	40	0.476	−0.365
2	30	40	50	50	20	0.343	−1.393	22	50	20	40	30	80	0.523	0.000	42	40	50	50	40	20	0.306	−1.679
3	40	80	50	50	30	0.374	−1.153	23	80	20	40	50	50	0.452	−0.550	43	40	20	80	30	30	0.586	0.485
4	20	40	50	80	30	0.523	−0.002	24	50	30	40	50	80	0.374	−1.150	44	20	30	80	30	20	0.697	1.343
5	40	20	80	80	30	0.566	0.331	25	50	20	30	50	40	0.343	−1.390	45	30	50	80	30	30	0.498	−0.195
6	40	40	50	40	50	0.124	−3.085	26	80	30	20	80	20	0.681	1.220	46	20	30	40	30	20	0.299	−1.733
7	50	20	80	50	30	0.500	−0.179	27	80	40	40	50	40	0.346	−1.370	47	20	80	40	20	50	0.593	0.539
8	50	40	80	80	50	0.312	−1.632	28	40	40	20	80	20	0.612	0.686	48	20	50	40	50	50	0.310	−1.648
9	80	30	80	40	50	0.411	−0.867	29	20	50	30	80	30	0.568	0.346	49	30	40	20	50	80	0.523	−0.002
10	40	30	40	40	50	0.177	−2.675	30	20	80	20	80	50	0.600	0.594	50	50	80	20	30	40	0.523	−0.002
11	80	50	80	20	80	0.433	−0.697	31	30	40	30	40	80	0.471	−0.400	51	30	40	20	50	30	0.335	−1.454
12	80	80	40	20	80	0.471	−0.403	32	30	80	20	80	30	0.614	0.702	52	30	40	20	30	80	0.586	0.485
13	80	30	20	20	40	0.655	1.019	33	50	80	30	40	50	0.374	−1.150	53	30	20	20	80	40	0.655	1.019
14	40	50	50	20	20	0.421	−0.790	34	80	40	30	20	80	0.566	0.331	54	50	20	30	80	40	0.523	−0.002
15	20	50	30	20	30	0.408	−0.890	35	80	20	30	30	50	0.568	0.346	55	80	30	30	40	20	0.586	0.485
16	30	80	50	30	50	0.427	−0.743	36	50	40	50	20	40	0.306	−1.680	56	50	20	50	80	20	0.570	0.362
17	30	40	80	20	30	0.586	0.485	37	80	30	30	80	40	0.498	−0.200	57	50	30	30	40	30	0.248	−2.127
18	20	40	80	30	80	0.566	0.331	38	80	40	50	40	20	0.476	−0.370	58	50	30	30	40	20	0.335	−1.454
19	30	20	80	30	50	0.568	0.346	39	80	20	40	20	40	0.612	0.686	59	80	30	50	40	30	0.451	−0.558
20	30	40	40	30	50	0.220	−2.343	40	40	30	50	80	20	0.523	0.000	60	40	50	50	20	20	0.421	−0.790

BT variance of the estimator of the sample coefficient of variation $\widehat{\theta}$(=sample standard deviation/sample mean) is

$$\widehat{V}_B(1) = \sum_{b=1}^{B} \left(\widehat{\theta}_{s^*}^{b} - \widehat{\overline{\theta}}\right)^2 \Big/ B = 0.0167$$

where $\widehat{\overline{\theta}} = \frac{1}{60} \sum\limits_{b=1}^{60} \widehat{\theta}_{s^*}^{b} = 0.4708$.

To determine the confidence interval for θ by percentile method, we arrange $\widehat{\theta}_{s^*}^{b}$ in the ascending order of magnitude as follows:

0.124	0.177	0.220	0.248	0.299	0.306	0.306	0.310	0.312	0.335	0.335	0.343
0.343	0.346	0.374	0.374	0.374	0.408	0.411	0.421	0.421	0.427	0.433	0.451
0.452	0.471	0.471	0.476	0.476	0.498	0.498	0.498	0.500	0.523	0.523	0.523
0.523	0.523	0.523	0.566	0.566	0.566	0.568	0.568	0.568	0.570	0.586	0.586
0.586	0.586	0.593	0.600	0.612	0.612	0.614	0.637	0.655	0.655	0.681	0.697

Here, lower and upper 5% points of the $\widehat{\theta}_{s^*}^{b}$ values are 0.220 and 0.655, respectively. Hence 90% confidence interval for θ obtained by the percentile method is (0.220, 0.655).

For BT t-method, we arrange t_b values in the ascending order as follows:

−3.085	−2.675	−2.343	−2.127	−1.733	−1.679	−1.679	−1.648	−1.632	−1.454	−1.454	−1.393
−1.393	−1.369	−1.153	−1.153	−1.153	−0.890	−0.867	−0.790	−0.79	−0.743	−0.697	−0.558
−0.550	−0.403	−0.403	−0.365	−0.365	−0.195	−0.195	−0.195	−0.179	−0.002	−0.002	−0.002
−0.002	−0.002	−0.002	0.331	0.331	0.331	0.346	0.346	0.346	0.362	0.485	0.485
0.485	0.485	0.539	0.594	0.686	0.686	0.702	0.879	1.019	1.019	1.22	1.343

The lower and upper 5% percentile points of the BT t-distribution are −2.343 and 1.019, respectively. Hence 90% confidence interval for θ is

$$\left(\widehat{\theta} - (1.019)\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)} - (-0.243)\sqrt{\widehat{V}_B\left(\widehat{\theta}\right)}\right) = (0.391, 0492).$$

18.6.2.2 Rescaling Bootstrap

Rao and Wu (1988) proposed various rescaling BT procedures for estimation of variance for stratified, multistage, and the Rao—Hartley—Cochran (1962) method of sampling designs. They all reproduce to usual variance estimators in the linear case. Consider the case where a sample s of size n is selected by the SRSWR sampling design. Then the variance of the sample mean $\overline{y}_s = \sum_{i\in s} y_i/n$ is unbiasedly estimated by $\widehat{V}\left(\overline{y}_s\right) = s_y^2\Big/n$, where $s_y^2 = \dfrac{\sum_{i\in s}\left(y_i - \overline{y}_s\right)^2}{n-1}$ and $\sum_{i\in s}$ denote the sum over units in s with repetition. In BT method we select B independent samples each of size n from s by SRSWR method treating all the units in s as distinct. Then the BT estimator of the variance of $\overline{y}_s$ is given by

$$\widehat{V}_B(1) = \sum_{b=1}^{B}\left(\overline{y}_b^* - \overline{y}^*\right)^2\Big/B \tag{18.6.4}$$

where $\overline{y}_b^*$ is the sample mean of the bth BT sample and $\overline{y}^* = \sum\limits_{b=1}^{B}\overline{y}_b^*\Big/B$. Let $E*$ denote the conditional expectation with respect to BT sampling given s, then $E_*\left(\widehat{V}_B(1)\right) = \dfrac{(n-1)}{n}\dfrac{s_y^2}{n}$, which is not equal to $\widehat{V}\left(\overline{y}_s\right) = \dfrac{s_y^2}{n}$, particularly if n is small. Rao and Wu (1988) proposed a rescaling BT method where the BT variance estimator for the variance of $\overline{y}_s$ is exactly equal to $\widehat{V}\left(\overline{y}_s\right)$. The proposed rescaling method is as follows:

Step 1: Draw BT sample (with SRSWR method) $s^* = \left(y_1^*, \ldots, y_i^*, \ldots, y_n^*\right)$ and calculate

$$\widetilde{y}_i = \overline{y}_s + \{n/(n-1)\}^{1/2}\left(y_i^* - \overline{y}_s\right) \tag{18.6.5}$$

and

$$\widetilde{\theta}^* = \widetilde{y}_{s^*} = \sum\nolimits_{i\in s^*}\widetilde{y}_i/n \tag{18.6.6}$$

Step 2: Repeat step 1 independently B times and compute the corresponding estimates $\widetilde{y}_{s^*}^1, \ldots, \widetilde{y}_{s^*}^B$.

Step 3: The BT variance for $\overline{y}_s$ based on the rescaling technique is given by

$$\widehat{V}_{BRS}(1) = \sum_{b=1}^{B}\left(\widetilde{y}_{s^*}^b - \widetilde{y}^*\right)^2\Big/B \text{ or } \widehat{V}_{BRS}(2) = \sum_{b=1}^{B}\left(\widetilde{y}_{s^*}^b - \overline{y}_s\right)^2\Big/B \tag{18.6.7}$$

where $\widetilde{y}^* = \sum\limits_{b=1}^{B}\widetilde{y}_{s^*}^b\Big/B$.

We note that $E_*\left(\widehat{V}_{BRS}(1)\right) = E_*\left(\widehat{V}_{BRS}(2)\right) = \widehat{V}\left(\overline{y}_s\right) = s_y^2\Big/n$.

18.6.2.3 Bootstrap Without Replacement Method

Bootstrap without replacement (BWO) was proposed by Gross (1980) for variance estimation for the SRSWOR sampling design. Suppose a sample $s = (y_1,\ldots, y_n)$ of size n is selected from a finite population of size N by SRSWOR method and $N/n = k$ is an integer. Let the estimator of the parameter of interest θ, based on the sample s be $\widehat{\theta} = \widehat{\theta}(s)$. The variance of $\widehat{\theta}$ is computed by the BWO procedure is as follows:

Step 1: Generate a pseudopopulation by replication of the sample s, k times. The pseudopopulation is denoted by

$$U^* = \left(\overbrace{y_1,\ldots,y_n}^{1},\ldots,\overbrace{y_1,\ldots,y_n}^{k} \right).$$

Step 2: Draw an SRSWOR sample $s^* = \left(y_1^*,\ldots,y_n^*\right)$ of size n from the population U^* of size $N^* = nk$ assuming all elements in U^* are distinct. Let $\widehat{\theta}^*$ be the value of $\widehat{\theta}(s^*)$ based on the sample s^*.

Step 3: Repeat step 2 a large number times B and compute $\widehat{\theta}_1^*,\ldots, \widehat{\theta}_B^*$. The BT variance estimator of $\widehat{\theta}$ is given by

$$\widehat{V}_{BWO}\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \overline{\theta}^*\right)^2 \tag{18.6.8}$$

where $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*$ ($\overline{\theta}^*$ may be replaced by $\widehat{\theta}$).

Sitter (1992a) observed that the $\widehat{V}_{BWO}\left(\widehat{\theta}\right)$ is quite different from $\widehat{V}\left(\overline{y}_s\right) = (1-f)s_y^2\big/n$, where $\widehat{\theta} = \sum_{i\in s} y_i/n$ and $f = n/N$.

18.6.2.4 Mirror-Match Bootstrap

Sitter (1992b) proposed the mirror-match BT method, which is applicable to the stratified random sampling, two-stage cluster sampling, and Rao—Hartley—Cochran (1962) sampling designs. All the variance estimates based on this method reduce to the standard ones in the linear case. For simplicity, let us consider the SRSWOR sampling scheme where a sample s of size n is selected from a finite population and let $d = (y_1,\ldots,y_n)$ be the observed values of the study variable y. Let $\widehat{\theta} = \theta(d)$ be an estimator for the parameter of interest θ. The mirror-match BT method is described as follows:

Step 1: Select a sample s' of size $n'(<n)$ from s by SRSWOR method. Let corresponding observations be $d = \left(\widetilde{y}_1,\ldots,\widetilde{y}_{n'}\right)$.

Step 2: Repeat step 1, $k = n(1-f^*)/\{n'(1-f)\}$ (assuming integer) times independently to generate y-values $\widetilde{\mathbf{y}}^* = \left(\widetilde{y}_1, \ldots, \widetilde{y}_{n^*}\right)$ where $f^* = n'/n$ and $n^* = n(1-f^*)/(1-f)$. Let $\widehat{\theta}^m = \widehat{\theta}^m\left(\widetilde{\mathbf{y}}^*\right)$ be an estimator of θ based on $\widetilde{\mathbf{y}}^*$.

Step 3: Repeat steps 1 and 2, a large number of times B, and compute $\widehat{\theta}_1^m, \ldots, \widehat{\theta}_B^m$, where $\widehat{\theta}_j^m$ is the value of $\widehat{\theta}^m$ based on the jth iteration. The mirror-match estimator for the variance of $\widehat{\theta}$ is given by

$$\widehat{V}_{BM}\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^m - \overline{\theta}^m\right)^2 \text{ with } \overline{\theta}^m = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^m \qquad (18.6.9)$$

($\overline{\theta}^m$ may be replaced by $\widehat{\theta}$).

In case $\widehat{\theta} = \overline{y}_s =$ the sample mean, $E_*\left[\widehat{V}_{BM}\left(\widehat{\theta}\right)\right] = \widehat{V}\left(\overline{y}_s\right) = (1-f)s_y^2\big/n.$

18.6.2.5 Bootstrap for Varying Probability Sampling Without Replacement

Särndal et al. (1992) proposed the following method of BT for varying probability sampling schemes. Suppose a sample s of size n is selected from a finite population U of size N using a varying probability sampling scheme. Let π_i be the inclusion probability of the ith unit and $\widehat{\theta} = \widehat{Y}_{ht} = \sum_{i\in s} y_i/\pi_i$ be the Horvitz–Thompson estimator of the total $\theta = \sum_{i\in U} y_i$. We generate a population U^* by repeating units from the selected sample s. If a unit u_j is included in the sample s, then the unit u_j will appear $f_j = 1/\pi_j$ times in the population U^* assuming f_j is an integer for $j \in s$ so that the total number of units in U^* is $\widehat{N} = \sum_{j\in s} 1/\pi_j$. Now, from the population U^*, a sample s^* of size n is selected by PPSWR method using normed size measure $p_j = \pi_j/n$ for the unit $u_j \in s$. Clearly, $\sum_{k\in U^*} p_k = 1$. From the sample s^*, an estimator for θ is obtained as $\widehat{\theta}^* = \sum_{j\in s^*}\frac{y_j}{np_j}$. In the proposed BT method, we repeat the selection of sample s^* a large number of B times. Then the BT estimator of the variance of $\widehat{\theta}$ is

$$\widehat{V}_{BVP}\left(\widehat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \overline{\theta}^*\right)^2 \qquad (18.6.10)$$

where $\widehat{\theta}_b^*$ is the value of $\widehat{\theta}^*$ based on the bth BT sample and $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^*$.

18.7 GENERALIZED VARIANCE FUNCTIONS

In many large-scale surveys such as US Current Population Surveys (CPS), National Health Improvement Surveys (USA), Medical Expenditure Panel Surveys (USA), Botswana Aids Impact Surveys (BAIS), hundreds or even thousands of estimates of population characteristics are calculated. In such a situation direct computation of standard errors of each of the estimates may not be possible even with modern computers because of cost and time. Even if the cost is affordable, timely presentation of hundreds or thousands of estimates with their standard errors is quite impractical because of limitations of tabular presentations. However, such estimates are of interest to users. The GVFs provide variances for groups of statistics rather than individual variances of each of the statistic.

18.7.1 Generalized Variance Function Model

A GVF is a simple mathematical relationship between the variance or relative variance of a statistic and its expectation. Let $\widehat{\theta}$ be an unbiased estimator of a parameter θ of interest and $V\left(\widehat{\theta}\right)$ be the variance of $\widehat{\theta}$. Then the relative variance is defined by

$$\Delta_\theta^2 = \frac{V\left(\widehat{\theta}\right)}{\theta^2} \tag{18.7.1}$$

The GVF may be modeled as follows:

$$\Delta_\theta^2 = \alpha + \frac{\beta}{\theta} \tag{18.7.2}$$

$$\Delta_\theta^2 = \alpha + \frac{\beta}{\theta} + \frac{\gamma}{\theta^2} \tag{18.7.3}$$

$$\Delta_\theta^2 = \frac{1}{\alpha + \beta\theta + \gamma\theta^2} \tag{18.7.4}$$

$$log\left(\Delta_\theta^2\right) = \alpha - \beta \log \theta \tag{18.7.5}$$

The model parameters α, β, and γ are unknown and these are estimated from the survey data. The simplest and most popular model is (18.7.2) where relative variance Δ_θ^2 is a decreasing function of θ when β is positive. The US Census Bureau used this model for CPS survey.

18.7.2 Justification of Generalized Variance Function Model

There is little theoretical justification of the aforementioned models available in the literature. However, some justifications have been provided by Wolter (1985) and Kish (1965). These are described as follows.

Let a sample s of n clusters be selected from a population of N clusters each of size M by the SRSWOR method. Then an unbiased estimator of the population total $Y = \theta$ is $\widehat{\theta} = \frac{N}{n}\sum_{i\in s} Y_i$, where $Y_i = \sum_{j=1}^{M} y_{ij}$ and y_{ij} is the value of the characteristic y for the jth unit of the ith cluster. The variance of $\widehat{\theta}$ is obtained from Eq. (12.2.5) as

$$V\left(\widehat{\theta}\right) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{NM-1}{N-1}S_y^2\{1+(M-1)\rho_c\}$$

where ρ is the intracluster correlation coefficient and S_y^2 is the population variance of y.

For Large M and N, $V\left(\widehat{\theta}\right)$ reduces to

$$V\left(\widehat{\theta}\right) = \frac{N^2M}{n}S_y^2\{1+(M-1)\rho\}$$

The relative variance is given by

$$\Delta_\theta^2 = \frac{N^2M}{nY^2}S_y^2\{1+(M-1)\rho\}$$

Let y_{ij} be binary that takes value 1 if jth unit of the ith cluster possesses a certain attribute and y_{ij} is 0 otherwise, and $\pi = Y/(NM)$ is the proportion of persons possessing the attribute, then

$$\Delta_\theta^2 = \frac{1}{nM}\,\frac{1-\pi}{\pi}\{1+(M-1)\rho\}$$
$$= \alpha + \frac{\beta}{\theta}$$

where $\alpha = -\frac{1+(M-1)\rho}{nM}$, $\beta = \frac{N\{1+(M-1)\rho\}}{n}$, and $\theta = Y = NM\pi$.

Let $\theta(=\pi)$ be the population proportion and $\widehat{\pi}$ be the sample proportion based on an SRSWR sample of size n. Then the design effect (*Deff*) of fixed sample size n design (p) with respect to the estimator $\widehat{\theta}$ is defined by Kish (1965) as

$$Deff = \frac{V_p\left(\widehat{\theta}\right)}{V(\widehat{\pi})} = \frac{V_p\left(\widehat{\theta}\right)}{\pi(1-\pi)/n}$$

In this case the relative variance is

$$\begin{aligned}\Delta_\pi^2 &= \frac{V_p\left(\widehat{\theta}\right)}{\pi^2}\\ &= \frac{\mathit{Deff}\,(1-\pi)}{n\pi}\end{aligned}$$

If the design effect for a class of statistics happens to be independent of the parameter π, then the relative variance Δ_π^2 can be written as

$$\Delta_\pi^2 = \alpha + \frac{\beta}{\pi}$$

with $\alpha = -\mathit{Deff}/n$ and $\beta = \mathit{Deff}/n$.

18.7.3 Generalized Variance Function Method for Variance Estimation

First, we need to group the survey estimates so that they follow a common model. The group may comprise estimates related to the same demographic or economic characteristics, similar geographical areas, the same race ethnicity, the same intraclass correlation ρ, or similar *Deffs*. The items to be included in a group may also be based on the past experience.

After forming a group, some estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ belonging to a group are chosen and their variances $V\left(\widehat{\theta}_1\right), \ldots, V\left(\widehat{\theta}_k\right)$ are estimated using a direct method such as LR, RG, BRR, JK, or BT method. Then estimates of GVFs $\left(\Delta_\theta^2\right)$ are computed using the formula $\widehat{\Delta}_{\theta_j}^2 = \widehat{V}\left(\widehat{\theta}_j\right)\Big/\widehat{\theta}_j^2$.

To determine appropriate model, scatter plots $\widehat{\Delta}_{\theta_j}^2$ against $\widehat{\theta}_j$ may be useful. One may remove statistics from the group that do not seem to follow the same model.

After selection of the appropriate model, the model parameters may be determined by the ordinary least square (OLS) procedure. However, the OLS procedure may not be totally satisfactory because some of the estimates with low value may have high value of the estimated relative variance. This difficulty can be overcome by using the weighted least square method with weights inversely proportional to the variance of the estimates.

Finally, the estimated GVF of $\widehat{\theta}_j$ based on the model is obtained from the estimated model, e.g., for the model (18.7.2) GVF of $\widehat{\theta}_j$ is

$$\widehat{\Delta}_{\hat{\theta}_j}^2 = \widehat{\alpha} + \frac{\widehat{\beta}}{\widehat{\theta}_j} \qquad (18.7.6)$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the estimates of the model parameters α and β.

The variance of $\widehat{\theta}_j$ based on the GVF method is given by

$$\widehat{V}_{gvf}\left(\widehat{\theta}_j\right) = \widehat{\Delta}^2_{\hat{\theta}_j} \times \widehat{\theta}^2_j \tag{18.7.7}$$

18.7.4 Applicability Generalized Variance Function Model

There is a limited theoretical justification of the GVF models discussed earlier. Furthermore, grouping of estimates following the same model is not an easy task. However, the GVF method can be used when sufficient information is not available for direct computation of variance. GVF saves time and cost for production of reports. For further information readers are referred to Wolter (1985).

18.8 COMPARISON BETWEEN THE VARIANCE ESTIMATORS

For most complex survey designs, unbiased variance estimators of statistics that are expressible as linear functions of the observations can be derived. For nonlinear statistics and functional this is generally not the case. Various methods of obtaining variance estimation have been proposed in the literature. The merits of the proposed variance estimators can be judged with respect to the criteria of bias, mean square error, setting the confidence interval of the estimators, and applicability. It is very difficult to make theoretical comparisons among estimators. Little has been done to compare relative performances of the proposed estimators. Limited empirical studies reveal that none of the proposed variance estimators is the best in all situations.

The biases of LR, RG, JK, BRR, and BT are equal to a first-order of approximation, but regarding the mean square errors of the variance estimators, no definite conclusion can be made. Kreweski and Rao (1981) proved that the pivotal quantity $z = \dfrac{\hat{\theta} - \theta}{\sqrt{\hat{V}\left(\hat{\theta}\right)}}$ asymptotically follows standardized normal distribution for all the proposed methods (LR, RG, JK, BRR and BT) where $\widehat{\theta}$ and $\widehat{V}\left(\widehat{\theta}\right)$, respectively, denote estimator of θ and $V\left(\widehat{\theta}\right)$. So, all the variance estimators can be used for determination of confidence interval as well as testing hypothesis regarding the parameter θ when the sample size is large, but little is known about the distribution z when the sample size is small.

The RG method is easy to apply and works for any sampling design. The LR method needs theoretical derivation of variance and hence no standard programming can be used. This can make it cumbersome to implement. JK method is easy to apply and routine programming is available. BRR method is useful only for stratified sampling and can be complex for large stratum sample size. BT method is easy to apply with replacement sampling scheme but not readily applicable for varying probability without replacement sampling scheme. Although the GFS method does not possess adequate theoretical justification, it is useful where the variances of a large number of estimators are required to be computed or when sufficient information is not available for direct computation. Further details are given by Rao and Shao (1996, 1999), Rao and Wu (1988), Shao et al. (1998), Sitter (1992a,b), Saigo et al. (2001), and Wolter (1985).

18.9 EXERCISES

18.9.1 What is meant by a complex sampling design? Describe different methods of variance estimation in complex survey designs.

18.9.2 Explain the linearization method of variance estimation. Describe merits and demerits of the linearization method.

18.9.3 Let a sample s of size n be selected by a varying probability sampling scheme with inclusion probability π_i attached to the ith unit. Determine the expressions of bias and approximate mean square error of $t = \dfrac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} 1/\pi_i}$ as an estimator of the population mean $\overline{Y}$.

18.9.4 The population correlation coefficient between x and y is defined by

$$\rho = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}\sqrt{\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2}}$$

Using LR method, derive approximate expressions of bias and mean square error of r when a sample of size n selected by the SRSWR method.

18.9.5 Express the population regression coefficient

$$\beta = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}}$$ as a function of $\theta_0 = N$,

$\theta_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \theta_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \theta_3 = \frac{1}{N}\sum_{i=1}^{N} y_i,$ and

$\theta_4 = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$. Estimate the bias and mean square error of the estimate of β by LR technique based on a sample selected by varying probability sampling scheme with inclusion probabilities π_i and π_{ij} for the ith and ith and jth $(i \neq j)$th units.

18.9.6 The following table gives the number of cows and daily production of milk of 10 farms selected by SRSWOR method from 125 farms in a certain agricultural block.

Farm	1	2	3	4	5	6	7	8	9	10
No of cows	45	25	50	30	45	10	30	45	35	40
Production of milk in (L)	180	40	100	70	80	30	45	75	50	80

Estimate the average daily production of milk per cow. Obtain the standard error of the estimator used by using (i) linearization and (ii) jackknife method.

18.9.7 The following data give the monthly income distribution of 1000 families selected at random from 10,000 families in Durban. Estimate the income inequality using (i) sample coefficient of variation and (ii) sample variance as measure of inequality. Estimate the standard errors of your estimates by using (i) jackknife and (ii) random group method.

Income (in $)	0 −1,000	1,001 −2,500	2,501 −5,000	5,001 −8,000	8,001 −18,000	18,001 −30,000	30,001 −50,000	50,001 −80,000
No. of families	100	250	300	180	100	75	20	5

18.9.8 The following table gives the IQ of 10 students selected at random from a school. Estimate the median IQ of the students and its standard error by bootstrap and jackknife method. Obtain 90% confidence interval of the population median.

Students	1	2	3	4	5	6	7	8	9	10
IQ	55	75	80	80	95	80	90	75	65	60

18.9.9 Samples each of size 2 are selected by SRSWOR from each of the six strata and information of the expenditure on food (x) and total income (y) are gathered.

Strata 1		Strata 2		Strata 3		Strata 4		Strata 5		Strata 6	
Size: 20		Size: 30		Size: 10		Size: 25		Size: 18		Size: 20	
x	y	x	y	x	y	x	y	x	y	x	y
1,000	2,500	1,250	3,000	3,000	5,000	3,500	6,000	4,000	8,000	4,000	9,500
1,500	2,000	2,500	4,000	2,500	5,500	3,000	7,500	5,000	8,500	6,000	10,000

(i) Estimate the proportion of income expended on food.
(ii) Estimate the standard error of the proportion by (a) BRR, (b) linearization, and (c) random group method.

18.9.10 The following data give the distribution of daily wages of 1000 factory workers.

Daily wages (in $)	0—50	51—100	101—200	201—250
No. of workers	18	200	300	250
Daily wages (in $)	251—300	301—500	501—600	601—800
No. of workers	100	30	25	20

Estimate the median income and coefficient of variation of the income. Estimate the variances of the median income and coefficient of variance by (i) jackknife, (ii) random group, and (iii) bootstrap methods.