Chapter 2
The Design Effects and Misspecification Effects

Abstract It is known that the classical statistical models are based on the assumptions that the observations are obtained from samples drawn by simple random sampling with replacement (*srswr*) or equivalently the observations are independently and identically distributed (IID). As such the conventional formulae for standard statistical packages which implement these procedures are also based on IID assumptions. In practice, in large-scale surveys samples are generally selected using a complex sampling design, such as a stratified multistage sampling design and this implies a situation different from an IID setup. Again, in large-scale sample surveys the finite population is often considered as a sample from a superpopulation. Survey data are commonly used for analytic inference about model parameters such as mean, regression coefficients, cell probabilities, etc. The sampling design may entail the situation that the sample observations are no longer subject to the same superpopulation model as the complete finite population. Thus, even if the IID assumption may hold for the complete population, the same generally breaks down for sample observations. The inadequacy of IID assumption is well known in the sample survey literature. It has been known for a long time, for example, that the homogeneity which the population clusters generally exhibit tend to increase the variance of the sample estimator over that of the estimator under *srswr* assumption, and further estimates of this variance wrongly based on IID assumptions are generally biased downwards. In view of all these observations it is required to examine the effects of a true complex design on the variance of an estimator with reference to a *srswr* design or an IID model setup. Section 2.2 examines these effects, *design effect*, and *misspecification effect* of a complex design for estimation of a single parameter θ. The effect of a complex design on the confidence interval of θ is considered in the next section. Section 2.4 extends the concepts in Sect. 2.2 to multiparameter case and thus defines multivariate design effect. Since estimation of variance of estimator of θ, $\hat{\theta}$ (covariance matrix when θ is a vector of parameters) is of major interest in this chapter we consider different methods of estimation of variance of estimators, particularly nonlinear estimators in the subsequent section. The estimation procedures are very general; they do not depend on any distributional assumption and are therefore nonparametric in nature. Section 2.5.1 considers in detail a simple method of estimation of variance of a linear statistic. In Sects. 2.5.2–2.5.7 we consider Taylor series linearization procedure, random group (RG) method, balanced repeated replication (BRR), jackknife

P. Mukhopadhyay, *Complex Surveys*, DOI 10.1007/978-981-10-0871-9_2

(JK) procedure, JK repeated replication, and bootstrap (BS) techniques of variance estimation. Lastly, we consider the effect of a complex survey design on a classical test statistic for testing a hypothesis regarding a covariance matrix.

Keywords IID · Design effect · Misspecification effect · Design factor · Effective sample size · Multivariate design effect · Generalized design effect · Variance estimation · Linearization method · Random group · Balanced repeated replication · Jackknife (JK) procedure · JK repeated replication · Bootstrap · Wald statistic

2.1 Introduction

In analysis of data collected through sample surveys standard statistical techniques are generally routinely employed. However, the probabilistic assumptions underlying these techniques do not always reflect the complexity usually exhibited by the survey population. For example, in the classical setup, the log-linear models are usually based upon distributional assumptions, like Poisson, multinomial, or product-multinomial. The observations are also assumed to be independently and identically distributed (IID). On the other hand, survey populations are often complex with different cell probabilities in different subgroups of the population and this implies a situation different from the IID setup. A cross-tabulation of the unemployment data, for example, by age-group and level of education would not support the IID assumption of sample observations but would exhibit a situation far more complex in distributional terms. However, the conventional formulae for standard errors and test procedures, as implemented in standard statistical packages such as SPSS X or SAS are based on assumptions of IID observations or equivalently, that samples are selected by simple random sampling with replacement, and these assumptions are almost never valid for complex survey data.

Longitudinal surveys where sample subjects are observed over two or more time points typically lead to dependent observations over time. Moreover, longitudinal surveys often have complex survey designs that involve clustering which results in cross-sectional dependence among samples.

The inadequacy of IID assumption is well known in the sample survey literature. It has been known for a long time, for example, that the homogeneity which the population clusters generally exhibit tends to increase the variance of the sample estimator over that of the estimator under *srswr* assumption, and further estimates of this variance wrongly based on IID assumptions are generally biased downwards (Example 2.2.1). Hence consequences of wrong use of IID assumptions for cluster data are: estimated standard errors of the estimators would be too small and confidence intervals too narrow. For analytic purposes test statistic would be based on downwardly biased estimates of variance and the results would, therefore, appear to be more significant than was really the case. Hence such tests are therefore conservative in nature.

Again, in large-scale sample surveys the finite population is usually considered as a sample from a superpopulation. Survey data are commonly used for analytic inference about model parameters such as mean, regression coefficients, cell probabilities, etc. The sampling design may entail the situation that the sample observations are no longer subject to the same superpopulation model as the complete finite population. To illustrate the problem suppose that with each unit i of a finite population $\mathcal{P}$ is a vector $(Y_i, Z_i)'$ of measurements. Assume that $(Y_i, Z_i)'$ are independent draws from a bivariate normal distribution with mean $\boldsymbol{\mu}' = (\mu_Y, \mu_Z)$ and variance–covariance matrix $\boldsymbol{\Sigma}$. The values (y_i, z_i) are observed for a sample of n units selected by a probability sampling scheme. It is desirable to estimate mean μ_Y and variance σ_Y^2 of the marginal distribution of Y. We consider the following two cases.

(A) The sample is selected by *srswr* and only the values $\{(y_i, z_i), i \in s\}$ are known. This is the case of IID observations. Here, the maximum likelihood estimators (MLE's) of the parameters are

$$\hat{\mu}_Y = \bar{y}_s = \sum_{i \in s} y_i/n; \quad \hat{\sigma}_Y^2 = \sum_{i \in s}(y_i - \bar{y}_s)^2/n. \tag{2.1.1}$$

Clearly, $\mathcal{E}(\hat{\mu}_Y) = \mu_Y$ and $\mathcal{E}[n\hat{\sigma}_Y^2/(n-1)] = \sigma_Y^2$ where $\mathcal{E}(.)$ defines expectation with respect to the bivariate normal model. Thus, standard survey estimators are identical with the classical estimators in this case.

(B) The sample is selected with probability proportional to Z_i with replacement such that at each draw $i = 1, \ldots, n, P_i =$ Prob. $(i \in s) = Z_i / \sum_{i=1}^N Z_i$. The data known to the statistician are $\{(y_i, z_i), i \in s; z_j, j \notin s\}$. Suppose that the correlation coefficient $\rho_{Y,Z} > 0$. This implies that Prob.$(Y_i > \mu_Y | i \in s) > 1/2$ since the sampling scheme tends to select units with larger values of Z and hence large values of Y. Clearly, the distribution of the sample Y values, in this case, is different from the distribution in the population and the estimators defined in (2.1.1) are no longer MLE.

Recently, researchers in the social science and health sciences are increasingly showing interest in using data from complex surveys to conduct same sorts of analyses that they traditionally conduct with more straightforward data. Medical researchers are also increasingly aware of the advantages of well-designated subsamples when measuring novel, expensive variables on an existing cohort. Until recent times they would be analyzing the data using softwares based on the assumption that the data are IID.

In the very recent years, however, there have been some changes in the situation. All major statistical packages, like, STATA, SUDAAN, now include at least some survey analysis components and some of the mathematical techniques of survey analysis have been incorporated in widely used statistical methods for missing data and causal inference. The excellent book by Lumley (2010) provides a practical guide to analyzing complex surveys using R.

The above discussions strongly indicate that the standard procedures are required to be modified to be suitable for analysis of data obtained through sample surveys.

In Sects. 2.2–2.4 we consider the effects of survey designs on standard errors of estimators, confidence intervals of the parameters, tests of significance as well as the multivariate generalizations of these design effects.

Since the estimation of variance of an estimator under complex survey designs is one of the main subjects of interest in this chapter and in subsequent discussions we make a brief review of different nonparametric methods of estimation of variance in Sect. 2.5. Section 2.5.1 considers in detail a simple method of estimation of variance of a linear statistic. In Sects. 2.5.2–2.5.6 we consider Taylor series linearization procedure, random group method, balanced repeated replication, jackknife, and bootstrap techniques of variance estimation. All these procedures (except the bootstrap resampling) have been considered in detail in Wolter (1985). In this treatise we do not consider estimation of superpopulation-based variance of estimators. Interest readers may refer to Mukhopadhyay (1996) for a review in this area.

2.2 Effect of a Complex Design on the Variance of an Estimator

Let $\hat{\theta}$ be an estimator of a finite population parameter θ induced by a complex survey design of sample size n with $Var_{true}(\hat{\theta})$ as the actual design variance of $\hat{\theta}$. Let $Var_{SRS}(\hat{\theta})$ be the variance of $\hat{\theta}$ calculated under a hypothetical simple random sampling with replacement (*srswr*) (also, stated here as SRS) design of the same sample size (number of draws) n. The effect of the complex design on the variance of $\hat{\theta}$ (relative to the *srswr* design) is given by the design effect (*deff*) developed by Kish (1965),

$$\text{deff }(\hat{\theta})_{Kish} = \frac{Var_{true}(\hat{\theta})}{Var_{SRS}(\hat{\theta})}. \tag{2.2.1}$$

Clearly, if deff $(\hat{\theta})_{Kish} < 1$, the true complex design is a better design than a corresponding *srswr* design with respect to $\hat{\theta}$, the estimator of θ under the true design. Note that Kish's deff (2.2.1) is completely a design-based measure.

At the analysis stage one is, however, more interested in the effect of the design on the estimator of the variance. Let $v_0 = \hat{V}ar_{SRS}(\hat{\theta}) = \hat{V}ar_{IID}(\hat{\theta})$ be an estimator of $Var_{SRS}(\hat{\theta})$ which is derived under the SRS assumption or under the equivalent IID assumption, that is $E(v_0|SRS) = E(v_0|IID) = Var_{SRS}(\hat{\theta})$. Clearly, v_0 may be a design-based estimator or a model-based estimator. The effect of the true design on the estimator pair $(\hat{\theta}, v_0)$ is given by the bias of v_0,

$$E_{true}(v_0) - Var_{true}(\hat{\theta}), \tag{2.2.2}$$

where expectation in (2.2.2) is with respect to the actual complex design. However, for the sake of comparability with (2.2.1) we define the *misspecification effect* (*meff*) of $(\hat{\theta}, v_0)$ as

$$\text{meff}\ (\hat{\theta}, v_0) = \frac{Var_{true}(\hat{\theta})}{E_{true}(v_0)}. \tag{2.2.3}$$

This measure is given by Skinner (1989).

Note 2.2.1 Kish's design effect (2.2.1) is a design-based measure, while Skinner's misspecification effect may be defined either as a design-based measure or a as a model-based measure. When taken as a model-based measure, the quantities E_{true} and Var_{true} in (2.2.3) should be based on the true model distribution. Thus the measure (2.2.3) can also be used to study the effect of the assumed model on the variance of the estimator relative to the IID assumption. Clearly, under model-based approach meff$(\hat{\theta}, v_0)$ depends only on the model relationship between the units in the actual sample selected and not on how the sample was selected. □

It has been found that in large-scale sample surveys using stratified multistage sampling design with moderate sample sizes, $E_{true}(v_0) \approx Var_{SRS}(\hat{\theta})$. Hence, for such designs, the values of measures (2.2.1) and (2.2.3) are often very close. Also,

$$\hat{\text{deff}}\ (\hat{\theta}) \approx\ \hat{\text{meff}}\ (\hat{\theta}, v_0) = \frac{v}{v_0} \tag{2.2.4}$$

where v is a consistent estimator of $Var_{true}(\hat{\theta})$ under the true sampling design. Thus, even though the values of (2.2.1) and (2.2.3) may be unequal, the estimated values of *deff*$_{Kish}$ and *meff* are often equal.

We shall henceforth, unless stated otherwise, assume that all the effects on the variance are due to sampling designs only. The misspecification effect may now be called a *design effect*. Following Skinner (1989) we shall now define the design effect (deff) of $(\hat{\theta}, v_0)$ as

$$\text{deff}\ (\hat{\theta}, v_0) =\ \text{meff}\ (\hat{\theta}, v_0) = \frac{Var_{true}(\hat{\theta})}{E_{true}(v_0)}. \tag{2.2.5}$$

Note that measures in (2.2.5) may be based on both models and designs.

We can generalize (2.2.1) to define the *general design effect* (deff) of an estimator $\hat{\theta}$ as

$$\text{deff}\ (\hat{\theta}) = \frac{Var_{true}(\hat{\theta})}{Var_*(\hat{\theta})} \tag{2.2.6}$$

where $Var_*(\hat{\theta})$ is the variance of $\hat{\theta}$ under some benchmark design representing IID situations.

Example 2.2.1 In this example we shall clarify the distinction between design-based deff of Kish and model-based misspecification effect of Skinner.

Consider an infinite population of clusters of size 2 (elementary units) with mean θ (per elementary unit), variance σ^2 (per elementary unit), and intracluster correlation

(correlation between two units in a cluster) τ. Suppose that a sample of one cluster is selected for estimating θ and observations y_1, y_2 on the units in the cluster are noted. An estimator of θ under this cluster sampling design is $\hat{\theta} = (y_1 + y_2)/2$. The true variance of $\hat{\theta}$ is

$$Var_{true}(\hat{\theta}) = V\left[\frac{y_1 + y_2}{2}\right] = \frac{\sigma^2}{2}(1 + \tau).$$

Under the hypothetical assumption that the two elementary units have been drawn by *srswr* (or under the model assumption that y_1, y_2 are IID with mean and variance as above) from the population of elementary units in the hypothetical population,

$$Var_{SRS}(\hat{\theta}) = Var_{SRS}\left(\frac{y_1 + y_2}{2}\right) = \frac{\sigma^2}{2}.$$

Again, an estimator of $Var_{SRS}(\hat{\theta}) = Var_{IID}(\hat{\theta})$, also based on a srswr design, is

$$\begin{aligned} v_0 &= \tfrac{1}{2}[(y_1 - \hat{\theta})^2 + (y_2 - \hat{\theta})^2] \\ &= \frac{(y_1 - y_2)^2}{4}. \end{aligned}$$

Also,

$$E_{true}(v_0) = E_{true}\left[\frac{(y_1 - y_2)^2}{4}\right] = \sigma^2\frac{1 - \tau}{2}.$$

Therefore, by (2.2.1), Kish's design effect is

$$\text{deff }(\hat{\theta})_{Kish} = \frac{Var_{true}(\hat{\theta})}{Var_{SRS}(\hat{\theta})} = 1 + \tau.$$

Also, at the analysis stage, by (2.2.3),

$$\text{meff }(\hat{\theta}, v_0) = \frac{V_{true}(\hat{\theta})}{E_{true}(v_0)} = \frac{1 + \tau}{1 - \tau}.$$

Thus, if $\tau = 0.8$, deff $(\hat{\theta})_{Kish} = 1.8$, meff $_{Skinner}(\hat{\theta}, v_0) = 9$. This means that the true design variance is 80 % higher than the SRS-based variance under the design-based approach; but its true-model-based variance is 800 % higher than the average value of the IID-based variance estimator v_0. □

Example 2.2.2 Consider the problem of estimating the population mean θ by a simple random sample without replacement (*srswor*)-sample of size n. Here $\hat{\theta} = \bar{y}_s = \sum_{i \in s} y_i/n$, the sample mean with

$$Var_{true}(\hat{\theta}) = (N - n)\sigma^2/\{n(N - 1)\}, \; Var_{SRS}(\hat{\theta}) = \sigma^2/n,$$

$$v_0 = \hat{V}ar_{SRS}(\hat{\theta}) = \text{ estimator of } V_{SRS}(\bar{y}) \text{ under srswr } = s^2/n,$$

$$E_{true}(v_0) = N\sigma^2/\{(N-1)n\},$$

where $\sigma^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2/N$, $s^2 = \sum_{i\in s}(y_i - \bar{y})^2/(n-1)$. Hence

$$\text{deff } (\hat{\theta}) = \frac{Var_{true}(\bar{y})}{Var_{SRS}(\bar{y})} = \frac{N-n}{N-1},$$

$$\text{deff } (\bar{y}, v_0) = \frac{Var_{true}(\bar{y})}{E_{true}(v_0)} = \frac{N-n}{N},$$

which is the finite population correction factor.

Example 2.2.3 Suppose we want to estimate $\theta = \bar{Y} = \sum_h W_h\bar{Y}_h$ by stratified random sampling of size n with proportional allocation, where $W_h = N_h/N$, etc. Here $\hat{\theta} = \bar{y}_{st} = \bar{y}$, the sample mean and the true variance,

$$Var_{true}(\bar{y}) = \frac{N-n}{nN}\sum_h W_h S_h^2,$$

where S_h^2 is the population variance of the hth stratum. If we assume that the sample has been drawn by *srswor*, an unbiased estimator of $Var_{SRS}(\bar{y})$, also calculated under *srswor*, is

$$v_{SRS} = \frac{N-n}{nN}s^2 = v_0.$$

Note that *srswor* is the benchmark design here. Its expectation under the true design is

$$E_{true}(v_0) \approx \frac{N-n}{nN}S^2 = \frac{N-n}{nN}\sum_h W_h\{S_h^2 + (\bar{Y}_h - \bar{Y})^2\}$$

where $S^2 = (N-1)^{-1}\sum_h\sum_i(y_{hi} - \bar{Y})^2$ is the finite population variance. Hence the design effect is

$$\text{deff } (\bar{y}, v_0) = \frac{\sum_h W_h S_h^2}{\sum_h W_h[S_h^2 + (\bar{Y}_h - \bar{Y})^2]}.$$

The deff is always less than or equal to one and can be further reduced by the use of an appropriate allocation rule.

Example 2.2.4 Consider the cluster sampling design in which n clusters are selected by *srswor* from a population of N clusters each of size M. An unbiased estimator of population mean (per element) $\theta = \sum_{c=1}^{N}\sum_{l=1}^{M} y_{cl}/(MN)$ is

$$\bar{y} = \sum_{c=1}^{n}\sum_{l=1}^{M} y_{cl}/(nM). \tag{2.2.7}$$

Hence,

$$V_{true}(\bar{y}) = \frac{N-n}{n(N-1)}\sigma_b^2 \text{ where } \sigma_b^2 = \frac{1}{N}\sum_{c=1}^{N}(\bar{y}_c - \theta)^2,\ \bar{y}_c = \frac{1}{N}\sum_{l=1}^{M} y_{cl}.$$

Now,

$$\sigma_b^2 = \frac{\sigma^2}{M}\{1 + (M-1)\tau\},\ \sigma^2 = \frac{1}{MN}\sum_{c=1}^{N}\sum_{l=1}^{M}(y_{cl} - \theta)^2$$

where τ is the intraclass correlation among units belonging to the same cluster (vide, Mukhopadhyay 2009). Hence,

$$\begin{aligned} Var_{true}(\bar{y}) &= \tfrac{N-n}{n(N-1)}\tfrac{\sigma^2}{M}\{1 + (M-1)\tau\} \\ &= \left(\tfrac{1}{n} - \tfrac{1}{N}\right)\tfrac{N\sigma^2}{(N-1)M}\{1 + (M-1)\tau\}. \end{aligned}$$

Also,

$$\begin{aligned} \tau &= \tfrac{E(y_{cl}-\theta)(y_{cm}-\theta)}{E(y_{cl}-\theta)^2} \\ &= \tfrac{1}{(M-1)MN\sigma^2}\sum_{c=1}^{n}\sum\sum_{l\neq m=1}^{M}(y_{cl} - \theta)(y_{cm} - \theta). \end{aligned} \tag{2.2.8}$$

In *srswor* of *nM* elements from the population of *MN* elements,

$$V_{wor}(\bar{y}) = \frac{N-n}{nNM}\sigma^2.$$

An estimator of $V_{wor}(\bar{y})$ based on without replacement sampling is

$$v_{wor}(\bar{y}) = \left(1 - \frac{n}{N}\right)\sum_{c=1}^{n}\sum_{l=1}^{M}(y_{cl} - \bar{y})^2/[nM(MN-1)], \tag{2.2.9}$$

(assuming $MN \approx MN - 1$). Again,

$$E_{true}[v_{wor}(\bar{y})] = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{(nM-1)}\left[1 - \frac{(N-n)\{1 + (M-1)\tau)\}}{Mm(N-1)}\right]. \tag{2.2.10}$$

Hence,

$$\begin{aligned}\text{deff }(\bar{y}, v_{wor}) &= \frac{Var_{true}(\bar{y})}{E_{true}[v_{wor}(\bar{y})]}\\ &= \frac{N(nM-1)\{1+(M-1)\tau\}}{nM(N-1)\{1-[(N-n)/(nN)(N-1)][1+(M-1)\tau]\}}.\end{aligned} \tag{2.2.11}$$

If n is large, this gives approximately,

$$\text{deff }(\bar{y}, v_{wor}) = 1 + (M-1)\tau.$$

The above derivation of deff is based on the randomization due to sampling design. We now consider the corresponding result in a model-based setup. Consider the one-way random effect superpopulation model

$$y_{cl} = \theta + \alpha_c + \epsilon_{cl}, c = 1, \ldots, N; \quad l = 1, \ldots, M, \tag{2.2.12}$$

where θ is a constant overall effect, α_c is a random effect due to cluster, and ϵ_{cl} is a random error effect. We assume that α_c, ϵ_{cl} are mutually independent random variables with zero means and

$$V(\alpha_c) = \tau\sigma_0^2, \quad V(\epsilon_{cl}) = (1-\tau)\sigma_0^2.$$

The quantity τ can be interpreted as the intraclass correlation coefficient among the units belonging to the same cluster.

Here, $\hat{\theta} = \bar{y}$, the mean of the nM sampled elements. Under model (2.2.12),

$$\begin{aligned}Var_{true}(\bar{y}) &= V\left(\frac{1}{nM}\sum_{c=1}^{n}\sum_{l=1}^{M} y_{cl}\right)\\ &= \frac{1}{n^2M^2}[nM\sigma_0^2 + \sum_{c=1}^{n}\sum_{l\neq l'=1}^{M} \text{ Cov }(y_{cl}, y_{cl'})]\\ &= \frac{1}{n^2M^2}[Mn\sigma_0^2 + nM(M-1)\tau\sigma_0^2]\\ &= \frac{\sigma_0^2[1+(M-1)\tau]}{nM}.\end{aligned} \tag{2.2.13}$$

On the contrary, the IID model is

$$y_{cl} = \theta + e_{cl} \tag{2.2.14}$$

where e_{cl} are independently distributed random variables with mean 0 and variance σ_0^2. Hence

$$V_{IID}(\bar{y}) = \frac{\sigma_0^2}{nM}.$$

An unbiased estimator of $V_{IID}(\bar{y})$ under the IID assumption is

$$v_{IID}(\bar{y}) = \frac{1}{(nM-1)nM}\sum_{c=1}^{n}\sum_{l=1}^{M}(y_{cl}-\bar{y})^2. \tag{2.2.15}$$

$$\begin{aligned}
E_{true}[v_{IID}(\bar{y})] &= \frac{1}{(nM-1)nM}\sum_{c=1}^{n}\sum_{l=1}^{M}E\{y_{cl}^2+\bar{y}^2-2\bar{y}(y_{cl})\}\\
&= \frac{1}{(nM-1)nM}\sum_{c=1}^{n}\sum_{l=1}^{M}\left\{\theta^2+\sigma_0^2+\frac{\sigma_0^2}{nM}[1+(M-1)\tau]\right.\\
&\quad\left.+\theta^2-2\left(\frac{\sigma^2}{nM}+\frac{(M-1)\tau\sigma^2}{nM}+\theta^2\right)\right\}\\
&= \frac{\sigma_0^2}{nM(nM-1)}\{nM-1-(M-1)\tau\}\\
&\approx \frac{\sigma_0^2}{nM}
\end{aligned} \tag{2.2.16}$$

if n is large. Hence,

$$\text{deff }(\bar{y}, v_{IID}) \approx 1+(M-1)\tau$$

as in the case of design-based approach.

Example 2.2.5 Consider the linear regression model

$$E(Y|\mathbf{X}=\mathbf{x}) = \alpha + \mathbf{x}'\beta \tag{2.2.17}$$

where Y is the main variable of interest, $\mathbf{X} = (X_1, \ldots, X_k)'$, a set of k auxiliary variables. The ordinary least square (OLS) estimator of β which is best linear unbiased estimator (BLUE) when $V(Y|X)$ is a constant (model A) is

$$\hat{\beta}_{OLS} = \hat{\beta} = \mathbf{V}_{xx}^{-1}\mathbf{V}_{xy} \tag{2.2.18}$$

where

$$\mathbf{V}_{xx} = n^{-1}\sum_{i\in s}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})',\ \mathbf{V}_{xy} = n^{-1}\sum_{i\in s}(\mathbf{x}_i-\bar{\mathbf{x}})y_i,\ \bar{\mathbf{x}} = n^{-1}\sum_{i\in s}\mathbf{x}_i,$$

n being the size of the sample s, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$, y_i being observations on unit $i \in s$.

It is known that

$$v_{OLS}(\hat{\beta}) = \{n(n-k)\}^{-1}\left(\sum_{i\in s}e_i^2\right)\mathbf{V}_{xx}^{-1} \tag{2.2.19}$$

where $e_i = y_i - \bar{y} - (\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\beta}, \bar{y} = \sum_{i \in s} y_i/n$, has expectation

$$E\{v_{OLS}(\hat{\beta})|A\} = V(\hat{\beta}|A). \tag{2.2.20}$$

Now, if heteroscedasticity is present, i.e., if $V(Y_i|\mathbf{X} = \mathbf{x}_i) = \sigma^2(\mathbf{x}_i)$ (model B), then $v_{OLS}(\hat{\beta})$ may be inconsistent for $V(\hat{\beta}|B)$ even under simple random sampling and

$$E[v_{OLS}(\hat{\beta})|B] \approx \{n(n-k)\}^{-1} \left\{ \sum_{i \in s} \sigma^2(\mathbf{x}_i) \right\} \mathbf{V}_{xx}^{-1}. \tag{2.2.21}$$

Hence, in the multivariate case,

$$meff(\hat{\beta}, v_{OLS}(\hat{\beta})) = (E\{v_{OLS}(\hat{\beta})|B\})^{-1} V(\hat{\beta}|B). \tag{2.2.22}$$

For $k = 1$,

$$meff(\hat{\beta}, v_{OLS}(\hat{\beta})) = \frac{V(\hat{\beta}|B)}{E(v_{OLS}(\hat{\beta})|B)} \approx 1 + \rho C_\sigma C_x \tag{2.2.23}$$

where C_σ is the coefficient of variation (cv) of $\sigma^2(x_i)$, C_x is the cv of $(x_i - \bar{x})^2$, and ρ is the mutual correlation between x and y. This misspecification effect is due to the inconsistency of $v_{OLS}(\hat{\beta})$ under heteroscedastic model B. This inconsistency occurs even under simple random sampling and hence it is not proper to call Eq. (2.2.23) a design effect.

Now, under simple random sampling with replacement, a linearization estimator which is unbiased for $V(\hat{\beta}|B)$ in large samples is

$$v_B(\hat{\beta}) = n^{-2}\mathbf{V}_{xx}^{-1} \sum_{i \in s} (\mathbf{x}_i - \bar{\mathbf{x}}) e_i^2 (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{V}_{xx}^{-1}. \tag{2.2.24}$$

Therefore, in large samples,

$$meff(\hat{\beta}, v_B(\hat{\beta})) = \{E\{v_B(\hat{\beta})|B)\}^{-1} V(\hat{\beta}|B) = \{V(\hat{\beta}|B)\}^{-1} V(\hat{\beta}|B) = \mathbf{I}_k. \tag{2.2.25}$$

Hence, there is no inconsistency under simple random sampling in this case.

2.3 Effect of a Complex Design on Confidence Interval for θ

Let $\tilde{\theta}$ be an unbiased estimator of θ under the hypothetical SRS design (IID model assumption) and v_0 an estimate of $Var_{SRS}(\tilde{\theta})$. Then

$$t_0 = \frac{\tilde{\theta} - \theta}{\sqrt{v_0}} \tag{2.3.1}$$

is approximately distributed as a $N(0, 1)$ variable and 95 % confidence interval for θ under the IID assumption is

$$C_0 = \{\theta : |\tilde{\theta} - \theta| \leq 1.96\sqrt{v_0}\}. \tag{2.3.2}$$

Our aim is to study the properties of C_0 under the effect of true complex design.

Under the true design $\tilde{\theta}$ may be assumed to be normal with mean θ and variance $Var_{true}(\tilde{\theta})$. Again, in large samples, $v_0 \approx E_{true}(v_0)$ so that, from (2.3.1),

$$t_0 \approx \frac{\tilde{\theta} - \theta}{\sqrt{E_{true}(v_0)}} = \frac{\tilde{\theta} - \theta}{\sqrt{Var_{true}(\tilde{\theta})}} \sqrt{\frac{Var_{true}(\tilde{\theta})}{E_{true}(v_0)}}.$$

Hence the distribution of t_0 under the true design would be approximately

$$t_0 \sim_{true} N\left(0, \frac{Var_{true}(\tilde{\theta})}{E_{true}(v_0)} = \text{ deff } (\tilde{\theta}, v_0)\right). \tag{2.3.3}$$

Therefore, under the complex design, true 95 %-confidence interval for θ is

$$\tilde{\theta} \pm 1.96\sqrt{\text{ deff } (\tilde{\theta}, v_0) \cdot E_{true}(v_0)}. \tag{2.3.4}$$

Hence, the actual coverage probability of a confidence interval obtained from the IID assumption would be different from its nominal value depending on the deff $(\hat{\theta}, v_0)$. If an estimated $\widehat{\text{deff}}$ $(\hat{\theta}, v_0)$ is available then an adjusted confidence interval for θ with approximately 95 % coverage is

$$\tilde{\theta} \pm 1.96\sqrt{v_0 \cdot \widehat{\text{deff}}}\,. \tag{2.3.5}$$

Thus the deff$(\hat{\theta}, v_0)$ measures the inflation or deflation of IID-based pivotal statistic due to the use of true design. Table 2.1 adopted from Skinner et al. (1989) shows some such values.

We note that if deff $= 1$, C_0 has the same coverage probability as its nominal value. If deff $> (<) 1$, C_0 has coverage less (more) than its nominal value and hence its significance level is more (less) than the nominal significance level.

Suppose we want to test the null hypothesis $H_0 : \theta = \theta_0$ using data collected through a sampling design whose design effect is 1.5 and we shall use tests with nominal level 95 %. If we assume *srswr* or IID assumption, ignoring the true complex design we will use the confidence interval C_0 whose true coverage probability is 89 %, much below the nominal 95 % value. Therefore, in many cases H_0 will be

Table 2.1 Coverage of IID-based confidence intervals C_0

Design effect	Nominal level 95 %	Nominal level 99 %
0.9	96	99.3
1.0	95	99
1.5	89	96
2.0	83	93
2.5	78	90
3.0	74	86

rejected though we should have accepted the same in those cases. Test based on IID assumption is therefore conservative.

In practice, it is generally considered more desirable to have a conservative test (actual coverage probability less than the nominal coverage probability), than to use a liberal test. Therefore when using data from a complex survey, one should be careful of the large design effect. Even a design effect of 1.5 can make the actual significance level more than double its nominal value.

We now consider two definitions.

Definition 2.3.1 The *design factor* (deft) of a survey design is defined as

$$\text{deft} = \sqrt{\text{deff}}. \tag{2.3.6}$$

This is the appropriate inflation factor for standard errors and confidence intervals.

Definition 2.3.2 The *IID-effective sample size* or simply, *effective sample size* is defined as

$$n_e = \frac{n}{\text{deff}}, \tag{2.3.7}$$

and has the property that the SRS formula given by (2.3.2) becomes correct for the true design if n is replaced by n_e. (This definition is not to be confused with the Definition 1.2.3 which is concerned with the with-replacement sampling.)

Say

$$v_0 = \frac{A}{n}.$$

Then, if we replace n by n_e, v_0 becomes

$$v_0' = \frac{A}{n_e} = \left(\frac{A}{n}\right)(\text{deff}).$$

Therefore, if we use v_0' in place of v_0 in (2.3.1), and use the modified statistic

$$t_0' = (\hat{\theta} - \theta)/v_0'^{1/2}$$

the adjusted confidence interval $(\tilde{\theta} - 1.96\sqrt{v_0'}, \tilde{\theta} + 1.96\sqrt{v_0'})$ obtained from (2.3.2) has approximately the correct coverage probability.

2.4 Multivariate Design Effects

Suppose now that $\boldsymbol{\theta}$ is a $p \times 1$ vector, $\hat{\boldsymbol{\theta}}$ an estimator of $\boldsymbol{\theta}$ under the true design, and $\mathbf{V}_0$ a $p \times p$ matrix of estimators of covariance matrix of $\hat{\boldsymbol{\theta}}$ derived under the IID assumption or equivalently under the simple random sampling with replacement (SRS) assumption. The estimator $\mathbf{V}_0$ is also derived under the IID assumption. We may define the *multivariate design effects matrix* (in Skinner's sense) of the estimator-pair $\hat{\boldsymbol{\theta}}$ and $\mathbf{V}_0$ as

$$\text{deff}\ (\hat{\boldsymbol{\theta}}, \mathbf{V}_0) = (E_{true}(\mathbf{V}_0))^{-1} Cov_{true}(\hat{\boldsymbol{\theta}}). \tag{2.4.1}$$

The eigenvalues of this matrix $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$ are called the *generalized design effects* (in Skinner's sense) and has the property that δ_1, δ_p denote the bounds for the univariate design effects of any linear combination $\mathbf{c}'\hat{\boldsymbol{\theta}}$ of elements of $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned} \delta_1 &= \max\ \text{deff}_{\mathbf{c}}(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c}), \\ \delta_p &= \min\ \text{deff}_{\mathbf{c}}(\mathbf{c}'\hat{\boldsymbol{\theta}}, \mathbf{c}'\mathbf{V}_0\mathbf{c}). \end{aligned} \tag{2.4.2}$$

In the special case when the deff $(\hat{\boldsymbol{\theta}}, \mathbf{V}_0)$ is a $p \times p$ identity matrix, $\delta_1 = \cdots = \delta_p = 1$ so that the univariate design effects of all linear combinations of elements of $\hat{\boldsymbol{\theta}}$ are unity.

Note 2.4.1 The calculation of the design effect involves variance estimation and hence requires second-order inclusion probabilities. It also depends on how auxiliary information is used, and needs to be estimated one at a time for different scenarios. Wu et al. (2010) present bootstrap procedures (discussed in Sect. 2.5.7) for constructing pseudo empirical likelihood ratio confidence intervals for finite population parameters. The proposed method bypasses the need for design effects and is valid under general single-stage unequal probability sampling designs with small sampling fractions. Different scenarios in using auxiliary information are handled by simply including the same type of benchmark constraints with the bootstrap procedures.

Since estimation of variance of $\hat{\theta}$ (covariance matrix of $\hat{\theta}$, when θ is a vector parameter) is of major interest in this context we shall in the next section consider different methods of estimation of variance of estimators, particularly for nonlinear estimators. The estimation procedures are very general, they do not depend on any distributional assumption and are, therefore, nonparametric in nature.

2.5 Nonparametric Methods of Variance Estimation

Modern complex surveys often involve estimation of nonlinear functions, like population ratio, difference of ratios, regression coefficient, correlation coefficient, etc. Therefore, the usual formulae for unbiased estimation of sampling variance of simple (linear) estimators of, say, totals and means are inadequate for such surveys. There are two approaches to the estimation of variance of a nonlinear estimator. One is linearization, in which the nonlinear estimator is approximated by a linear one for the purpose of variance estimation. The second is replication in which several estimators of the population parameter are derived from different comparable parts of the original sample. The variability of these estimators is then used to estimate the variance of the parameter estimator.

We review these results in this chapter. Section 2.5.1 considers in detail a simple method of estimation of variance of a linear statistic. In Sects. 2.5.2–2.5.7 we consider Taylor series linearization procedure, random group (RG) method, balanced repeated replication (BRR), jackknife (JK) procedure, JK repeated replication, and bootstrap (BS) techniques of variance estimation. All these procedures (except the bootstrap resampling) have been considered in detail in Wolter (1985). Sarndal et al. (1992) have also considered in detail the problem of variance estimation in their wonderful book. We do not consider estimation of superpopulation-based variance, the topic being outside the scope of this book. Interested readers may refer to Mukhopadhyay (1996) for a review in this area. We review these results in this section.

2.5.1 A Simple Method of Estimation of Variance of a Linear Statistic

In a stratified three-stage sampling consider a linear statistic of the form

$$\hat{\theta} = \sum_{h=1}^{H} \sum_{a=1}^{n_h} \sum_{b=1}^{m_{ha}} \sum_{c=1}^{k_{hab}} u_{habc} \tag{2.5.1}$$

where u_{habc} is the value associated with the cth unit (ultimate-stage unit) belonging to the bth sampled ssu (second-stage unit) in the ath sampled fsu (first-stage unit) belonging to the hth stratum. For example, $\hat{\theta}$ may be the estimator of a population mean of a variable 'y', when

$$u_{habc} = \frac{y_{habc}}{N\pi_{habc}} \tag{2.5.2}$$

where N is the number of ultimate units and π_{habc} is the inclusion probability of the unit $(habc)$.

A simple unbiased estimator of the design variance of $\hat{\theta}$ can be obtained under the following assumptions:

(1) Samples are selected independently from one stratum to the other.
(2) The n_h sampled psu's within stratum h are selected with replacement (wr). (At each of the n_h draws there is a finite probability p_{ha} of selecting the ath psu, $\sum_{a=1}^{N_h} p_{ha} = 1$, where N_h is the total number of psu's in the hth stratum.)
(3) $n_h \geq 2$.

We may rewrite Eq. (2.5.1) as

$$\hat{\theta} = \sum_{h=1}^{H} \sum_{a=1}^{n_h} u_{ha} \tag{2.5.3}$$

where

$$u_{ha} = \sum_{b=1}^{m_{ha}} \sum_{c=1}^{k_{hab}} u_{habc}. \tag{2.5.4}$$

By assumption (2), the variables $u_{h1}, \ldots, u_{hn_h}$ are identically and independently distributed (IID) random variables within stratum h and therefore, by virtue of assumption (1),

$$\text{Var}\ (\hat{\theta}) = \sum_{h=1}^{H} n_h\ \text{Var}\ (u_{ha}). \tag{2.5.5}$$

Therefore, by (1) and (3), an unbiased estimator of Var $(\hat{\theta})$ is

$$v(\hat{\theta}) = \sum_{h=1}^{H} n_h \frac{1}{n_h - 1} \sum_{a=1}^{n_h} (u_{ha} - \bar{u}_h)^2, \tag{2.5.6}$$

where $\bar{u}_h = \sum_{a=1}^{n_h} u_{ha}/n_h$.

The estimator $v(\hat{\theta})$ can be readily computed from the aggregate quantities u_{ha} formed from the ultimate units. If the psu's are selected with replacement, one need not care about in how many subsequent stages sampling is carried out and/or if the sampling at the ultimate stage is by systematic sampling or any other procedure.

For the special case where $n_h = 2 \forall\ h$,

$$v(\hat{\theta}) = \sum_{h=1}^{H} (u_{h1} - u_{h2})^2. \tag{2.5.7}$$

Even in surveys with $n_h > 2$, the ultimate sampled units can often be grouped in two groups (on the basis of some criteria), the assumptions (1) and (2) made and the formula (2.5.7) applied. The groups are often called *Keyfitz groups* after Keyfitz (1957). The grouping of clusters, however, lead to some loss of efficiency.

The assumption (1) is often valid. In case $n_h = 1$ for some strata, such strata are often collapsed to form the new strata for which $n_h \geq 2$. Defining $v(\hat{\theta})$ with respect to the new strata then gives a conservative variance estimator.

Assumption (2) is almost always violated since the n_h psu's are generally selected by some without replacement procedure. In this case an unbiased variance estimator of $\hat{\theta}$ often involves complex formula with components for each stage of sampling. Some simplified procedures for the case $n_h = 2$ have been proposed by Durbin (1967) and Rao and Lanke (1984). One approximation is based on the assumption that the n_h psu's within stratum h are selected by srswor $(h = 1, \ldots, H)$. In this case an estimator of $\text{Var}(\hat{\theta})$ is

$$v_{wor}(\hat{\theta}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{a=1}^{n_h} (u_{ha} - \bar{u}_h)^2. \tag{2.5.8}$$

obtained by inserting a finite population correction factor in (2.5.6). Often the sampling fraction n_h/N_h is small and the difference between $v(\hat{\theta})$ and $v_{wor}(\hat{\theta})$ is negligible. In any case, $v(\hat{\theta})$ is often a conservative estimator.

In analytic surveys, where the parameter of interest is often a superpopulation model parameter, the finite population correction n_h/N_h is inappropriate and $v(\hat{\theta})$ is to be used.

We shall now show that under a measurement error model, the estimator $v(\hat{\theta})$ is a better estimator of the total variance rather than $v_{wor}(\hat{\theta})$. Consider the model

$$u_{ha} = U_{ha} + \epsilon_{ha} \tag{2.5.9}$$

where U_{ha} are the true values and ϵ_{ha} are random variables distributed independently with mean 0 and variance σ_h^2. The errors ϵ_{ha} arise, for example, from interviewers' errors and other non-sampling errors. Now, by (2.5.3),

$$\mathcal{V}_p(\hat{\theta}) = \mathcal{V}_p\left(\sum_{h=1}^{H}\sum_{a=1}^{n_h} U_{ha}\right) + \mathcal{V}_p\left(\sum_{h=1}^{H}\sum_{a=1}^{n_h} \epsilon_{ha}\right) \tag{2.5.10}$$

where $\mathcal{V}_p$ means variance due to joint randomization of sampling design and measurement error distribution. Now, if the psu's are selected by *srswosr*,

$$\begin{aligned}\mathcal{V}_p\left(\sum_{h=1}^{H}\sum_{a=1}^{n_h} U_{ha}\right) &= \sum_{h=1}^{H} n_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) S_h^2 \\ &= \sum_{h=1}^{H} n_h \left(1 - \frac{n_h}{N_h}\right) S_h^2 \end{aligned} \tag{2.5.11}$$

where

$$S_h^2 = (N_h - 1)^{-1} \sum_{a=1}^{N_h} (U_{ha} - \bar{U}_h)^2, \ \bar{U}_h = \sum_{a=1}^{Nh} U_{ha}/N_h.$$

Also

$$\begin{aligned} \mathcal{V}_p\left(\sum_{h=1}^{H} \sum_{a=1}^{n_h} \epsilon_{ha}\right) &= \mathcal{V}\left(\sum_{h=1}^{H} \sum_{a=1}^{n_h} \epsilon_{ha}\right) \\ &= \sum_{h=1}^{H} n_h \sigma_h^2, \end{aligned} \tag{2.5.12}$$

where $\mathcal{V}(.)$ denotes variance wrt error distribution and σ_h^2 denotes the fixed variance of ϵ_{ha}. Hence,

$$\mathcal{V}_p(\hat{\theta}) = \sum_{h=1}^{H} n_h \left[\left(1 - \frac{n_h}{N_h}\right) S_h^2 + \sigma_h^2\right]. \tag{2.5.13}$$

Here, from (2.5.6)

$$Ev(\hat{\theta}) = \sum_{h=1}^{H} n_h \left(S_h^2 + \sigma_h^2\right). \tag{2.5.14}$$

From (2.5.8)

$$Ev_{wor}(\hat{\theta}) = \sum_{h=1}^{H} n_h \left(1 - \frac{n_h}{N_h}\right) \left(S_h^2 + \sigma_h^2\right). \tag{2.5.15}$$

Therefore,

$$E[v_{wor}(\hat{\theta})] \le \mathcal{V}_p(\hat{\theta}) \le E(v(\hat{\theta})). \tag{2.5.16}$$

The estimator $v(\hat{\theta})$ is preferred, since it is a conservative estimator.

In the multivariate case, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$, we can write $\hat{\boldsymbol{\theta}}$ as

$$\hat{\boldsymbol{\theta}} = \sum_h \sum_a \sum_b \sum_c \mathbf{u}_{habc} \tag{2.5.17}$$

where $\mathbf{u}_{habc}$ is a vector of values associated with the unit '*habc*'. Corresponding to $v(\hat{\theta})$ in (2.5.6) in the univariate case, we have the covariance matrix estimator

$$v(\hat{\boldsymbol{\theta}}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{a=1}^{n_h} (\mathbf{u}_{ha} - \bar{\mathbf{u}}_h)(\mathbf{u}_{ha} - \bar{\mathbf{u}}_h)' \tag{2.5.18}$$

where

$$\mathbf{u}_{ha} = \sum_b \sum_c \mathbf{u}_{habc}, \quad \bar{\mathbf{u}}_h = \sum_{a=1}^{n_h} \mathbf{u}_{ha}/n_h.$$

Clearly, assumptions (1) and (2) above are of vital importance and the procedure can be applied to any sampling design based on sampling at any arbitrary number of stages. The above results are derived following Wolter (1985) and Skinner et al (1989).

2.5.2 Linearization Method for Variance Estimation of a Nonlinear Estimator

We now consider the problem of estimation of variance of a nonlinear estimator, like ratio estimator, regression estimator. In the estimation of variance of a nonlinear estimator we adopt the method based on Taylor series expansion. The method is also known as *linearization method.*

Let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ where Y_j is a population total (or mean) of the jth variable and let $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_p)'$ where $\hat{Y}_j$ is a linear estimator of Y_j. We consider a finite population parameter $\theta = f(\mathbf{Y})$ with a consistent estimator $f(\hat{\mathbf{Y}})$. A simple example is a population subgroup ratio, $\theta = Y_1/Y_2$ with $\hat{\theta} = \hat{Y}_1/\hat{Y}_2$, Y_1, Y_2 are population totals for groups 1 and 2.

Suppose that continuous second-order derivatives exist for the function $f(\mathbf{Y})$. Now,

$$f(\hat{\mathbf{Y}}) = f(\mathbf{Y}) + \sum_{j=1}^{p} (\hat{Y}_j - Y_j) \frac{\partial f}{\partial Y_j} + \sum \sum_{j,k=1}^{p} (\hat{Y}_j - Y_j)(\hat{Y}_k - Y_k) \frac{\partial f}{\partial Y_j} \frac{\partial f}{\partial Y_k} + \ldots . \tag{2.5.19}$$

Thus using only the linear terms of the Taylor series expansion, we have an approximate expression

$$\hat{\theta} - \theta = \sum_{j=1}^{p} (\hat{Y}_j - Y_j) \frac{\partial f}{\partial Y_j}. \tag{2.5.20}$$

Using the linearized equation (2.5.20), an approximate expression for variance of $\hat{\theta}$ is

$$E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) \approx \sum_{j=1}^{p} \left(\frac{\partial f}{\partial Y_j} \right)^2 V(\hat{Y}_j) + \sum \sum_{j \neq k = 1}^{p} \left(\frac{\partial f}{\partial Y_j} \right) \left(\frac{\partial f}{\partial Y_k} \right) \text{Cov}\,(\hat{Y}_j, \hat{Y}_k). \tag{2.5.21}$$

We have thus reduced the variance of a nonlinear estimator to the function of the variance and covariance of p linear estimators $\hat{Y}_j$. A variance estimator $v(\hat{\theta})$ is obtained from (2.5.21) by substituting the variance and covariance estimators $v(\hat{Y}_j, \hat{Y}_k)$ for the corresponding parameters $V(\hat{Y}_j, \hat{Y}_k)$. The resulting variance estimator is a first-order Taylor series approximation. The justification for ignoring the remaining higher order terms has to be sought from practical experience derived from various complex surveys in which sample sizes are sufficiently large. Krewski and Rao (1981) have shown that the linearization estimators are consistent.

Basic principles of the linearization method for the variance estimation of a nonlinear estimator under complex sampling designs are due to Keyfitz (1957) and other. A criticism against the method is about the convergence of the Taylor series used to develop (2.5.20). For ratio estimator Koop (1972) gave a simple example where the convergence condition is violated. Again, for complex estimators, the analytic partial differentiation needed to derive the linear substitute has been found to be intractable. Woodruff and Causey (1976) describes a solution to this problem that uses a numerical procedure to obtain the necessary partial derivative. Binder (1983) provides a general approach to the analytic derivation of variance estimators for linear Taylor series approximations for a wide class of estimators. Empirical evidences have shown, however, that the linearization variance estimators are generally of adequate accuracy, particularly, when the sample size is large. The approximation may be unreliable in the case of highly skewed population.

Example 2.5.1 Ratio Estimator: Let

$$\mathbf{Y} = (Y_1, Y_2)', \theta = f(\mathbf{Y}) = \frac{Y_1}{Y_2}, \hat{\theta} = f(\hat{\mathbf{Y}}) = \frac{\hat{Y}_1}{\hat{Y}_2},$$

$$\frac{\partial f(\mathbf{Y})}{\partial Y_1} = \frac{1}{Y_2}, \frac{\partial f(\mathbf{Y})}{\partial Y_2} = -\frac{Y_1}{Y_2^2}, \frac{\partial f(\mathbf{Y})}{\partial Y_1}\frac{\partial f(\mathbf{Y})}{\partial Y_2} = -\frac{Y_1}{Y_2^3}.$$

Hence,

$$\begin{aligned} V(\hat{\theta}) &= \frac{V(\hat{Y}_1)}{Y_2^2} + \frac{Y_1^2 V(\hat{Y}_2)}{Y_2^4} - \frac{2Y_1}{Y_2^3} \text{Cov}\,(\hat{Y}_1, \hat{Y}_2) \\ &= \frac{Y_1^2}{Y_2^2}\left[\frac{V(\hat{Y}_1)}{Y_1^2} + \frac{V(\hat{Y}_2)}{Y_2^2} - \frac{2\,\text{Cov}\,(\hat{Y}_1, \hat{Y}_2)}{Y_1 Y_2}\right]. \end{aligned} \quad (2.5.22)$$

Example 2.5.2 Combined and Separate Ratio Estimator in Stratified Two-Stage Sampling: The population consists of H strata, the hth stratum containing N_h clusters (which consist of M_h elements, $\sum_{h=1}^{H} M_h = M$). A first-stage sample of $n_h(\geq 2)$ clusters is drawn from the hth stratum and a second-stage sample of m_h elements is drawn from the n_h sampled clusters, $m = \sum_h m_h$. The quantity m_h is a random variable.

We assume that the sampling design is self-weighing, i.e., inclusion probability of each of M elements in the population is a constant over the strata and adjustments for nonresponse is not necessary. Let

m_{ha} = number of elements in the ath cluster in the hth stratum in the sample;
$y_{ha} = \sum_{b=1}^{m_{ha}} y_{hab}$ = sum of the response variable y over the m_{ha} elements in the ath cluster belonging to the hth stratum in the sample ($a = 1, \ldots, n_h$; $h = 1, \ldots, H$).

Let Y_{ha}, M_{ha} denote the respective population totals. A combined ratio estimator of population ratio (mean per element)

$$r = \frac{\sum_{h=1}^{H} \sum_{a=1}^{N_h} Y_{ha}}{\sum_{h=1}^{H} \sum_{a=1}^{N_h} M_{ha}} = \frac{T}{M} \tag{2.5.23}$$

is

$$\hat{r}_{com} = \frac{\sum_{h=1}^{H} \sum_{a=1}^{n_h} y_{ha}}{\sum_{h=1}^{H} \sum_{a=1}^{n_h} m_{ha}} = \frac{\sum_{h=1}^{H} y_h}{\sum_{h=1}^{H} m_h} = \frac{y}{m} \tag{2.5.24}$$

where $y = \sum_{h=1}^{H} y_h$, $y_h = \sum_{a=1}^{n_h} y_{ha}$ is the sample sum of the response variable for the hth stratum and $m = \sum_{h=1}^{H} m_h$ and m_h the number of sample elements in the hth stratum. For a binary 0–1 variable, $r = P = M_1/M$, the population proportion where M_1 is the count of elements each having value 1. In estimator (2.5.24) not only the quantities y_{ha} vary, but also the quantities m_{ha} and m in the denominator. Hence, $\hat{r}_{com} = \hat{r}$ is a nonlinear estimator.

A separate ratio estimator is a weighted sum of stratum sample ratios, $\hat{r}_h = y_h/m_h$ which themselves are ratio estimators of the population stratum ratios, $r_h = \sum_{a=1}^{N_h} Y_{ha}/\sum_{a=1}^{N_h} M_{ha}$. Thus,

$$\hat{r}_{sep} = \sum_{h=1}^{H} W_h r_h \tag{2.5.25}$$

with $W_h = M_h/M$. A linearized variance for the combined ratio estimator $\hat{r}_{com} = \hat{r} = y/x$ in (2.5.24) is, according to (2.5.22),

$$V(\hat{r}) = r^2 \left[\frac{V(y)}{y^2} + \frac{V(m)}{m^2} - \frac{2\,\mathrm{Cov}\,(y, m)}{ym} \right]. \tag{2.5.26}$$

Hence, an estimator of $V(\hat{r})$ can be written as

$$v_{des} = \hat{r}^2[y^{-2}\hat{V}(y) + m^{-2}\hat{V}(m) - 2m^{-1}y^{-1}\hat{Cov}(y, m)] \tag{2.5.27}$$

as the design-based variance estimator of $\hat{r}$ is based on the linearization method. The estimators $\hat{V}$'s depend on the sampling design.

The variance estimator (2.5.27) is a large-sample approximation in that good performance can be expected if not only the number of sampled elements is large, but also the number of sampled clusters is so. In case of a small number of sampled clusters the variance estimator can be unstable.

The variance estimator v_{des} is consistent if $\hat{V}(y)$, $\hat{V}(m)$, $\hat{V}(m, y)$ are consistent estimators. The cluster sample sizes should not vary too much for the reliable performance of the approximate variance estimator (2.5.27). The method can be safely used if the coefficient of variation of m_{ha} is less than 0.2. If the cluster sample sizes m_{ha} are all equal, $\hat{V}(m) = 0$, $\hat{V}(y, m) = 0$, and $\hat{V}(\hat{r}) = \hat{V}(y)/m^2$. For a 0–1 binary response variable and for sampling under IID conditions, $\hat{r} = p = m_1/m$, sample proportion, where m_1 is the number of elements in the sample each having value 1 and m is the number of elements in the sample. Assuming m is a fixed quantity, the variance estimator (2.5.27) reduces to the binomial variance estimator $v_{des}(p) = v_{bin}(\hat{p}) = p(1-p)/n$.

Assuming that $n_h(\geq 2)$ clusters are selected by srswr from each stratum we obtain relatively simple variance and covariance estimators in (2.5.27). We have

$$\hat{V}(y) = \sum_h n_h s_{yh}^2, \quad \hat{V}(m) = \sum_h n_h s_{mh}^2,$$

$$\hat{V}(y, m) = \sum_h n_h s_{y,mh}^2$$

where

$$\begin{aligned} s_{yh}^2 &= (n_h - 1)^{-1} \textstyle\sum_{a=1}^{n_h} (y_{ha} - \bar{y}_h)^2, \\ s_{y,mh}^2 &= (n_h - 1)^{-1} \textstyle\sum_{a=1}^{n_h} (y_{ha} - \bar{y}_h)(m_{ha} - \bar{m}_h)^2, \end{aligned} \tag{2.5.28}$$

$\bar{y}_h = \sum_a y_{ha}/m_h$ and s_{mh}^2, $\bar{m}_h$ have similar meanings. Note that by assuming srswr of clusters we only estimate the between-cluster components of variances and do not account for the within-cluster variances. As such the variance estimator (2.5.27) obtained using (2.5.28) will be an underestimate of the true variance. This bias is negligible if the first-stage sampling fraction n_h/N_h in each stratum is small. This happens if N_h is large in each stratum.

2.5.3 Random Group Method

The random group (RG) method was first developed at the US Bureau of Census. Here, an original sample and other $k(\geq 2)$ samples, also called random groups, are drawn from the population, usually using the same sampling design. The task of these last k random samples or random groups is to provide an estimate for the variance of an estimator of population parameter of interest based on the original sample. We shall distinguish two cases:

(a) *Samples or Random Groups are mutually independent*: Let $\hat{\theta}, \hat{\theta}_1, \ldots, \hat{\theta}_k$ be the estimators obtained from the original sample and k random groups respectively, all the estimators using the same estimating procedure. Here $\hat{\theta}_1, \ldots, \hat{\theta}_k$ are mutually independent. We want to estimate Var$(\hat{\theta})$, variance of the estimator $\hat{\theta}$ based on the original sample.

The RG estimate of θ is $\bar{\hat{\theta}} = \sum_i \hat{\theta}_i/k$. If $\hat{\theta}$ is linear, $\bar{\hat{\theta}} = \hat{\theta}$. Now an estimate of Var $(\bar{\hat{\theta}})$ is

$$v(\bar{\hat{\theta}}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\theta}_i - \bar{\hat{\theta}})^2. \tag{2.5.29}$$

Note that for the above formula to hold it is neither required to assume that all $\hat{\theta}_i$'s have the same variance nor to assume that $\hat{\theta}_i$ are independent. It is sufficient to assume that all $\hat{\theta}_i$'s have finite variances and that they are pairwise uncorrelated.

Now, by Cauchy–Schwarz inequality

$$0 \le \left[\sqrt{Var(\bar{\hat{\theta}})} - \sqrt{Var(\hat{\theta})}\right]^2 \le Var[\bar{\hat{\theta}} - \hat{\theta}] \tag{2.5.30}$$

and Var$(\bar{\hat{\theta}} - \hat{\theta})$ is generally small relative to both Var$(\bar{\hat{\theta}})$ and Var$(\hat{\theta})$. Thus, the two variances are usually of similar magnitude.

To estimate Var $(\hat{\theta})$ one may use either $v_1(\hat{\theta}) = v(\bar{\hat{\theta}})$ or

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta})^2. \tag{2.5.31}$$

Note that $v_2(\hat{\theta})$ does not depend on $\bar{\hat{\theta}}$. When the estimator of θ is linear $v_1(\hat{\theta})$ and $v_2(\hat{\theta})$ are identical. For nonlinear estimators we have,

$$\sum_{i=1}^{k} (\hat{\theta}_i - \hat{\theta})^2 = \sum_{i=1}^{k} (\hat{\theta}_i - \bar{\hat{\theta}})^2 + k(\bar{\hat{\theta}} - \hat{\theta})^2. \tag{2.5.32}$$

Thus,

$$v_1(\hat{\theta}) \le v_2(\hat{\theta}). \tag{2.5.33}$$

If a conservative estimator of Var $(\hat{\theta})$ is desired, $v_2(\hat{\theta})$ is, therefore, preferable to $v_1(\hat{\theta})$. However, as noted above, Var$(\bar{\hat{\theta}} - \hat{\theta}) = E(\bar{\hat{\theta}} - \hat{\theta})^2$ will be unimportant in many complex surveys and there should be little difference between v_1 and v_2. It has been shown that the bias of v_1 as an estimator of Var$(\hat{\theta})$ is less than or equal to the bias of v_2.

Inferences about parameter θ are usually based on normal theory or Student's t distribution. The results are stated in the following theorem.

Theorem 2.5.1 *Let $\hat{\theta}_1, \ldots, \hat{\theta}_k$ be independently and identically distributed (iid) $N(\theta, \sigma^2)$ variables. Then*

(i) $\frac{\sqrt{k}(\bar{\hat{\theta}}-\theta)}{\sigma}$ *is distributed as a N(0,1) variable. (Obvious modification will follow if $\hat{\theta}_i$'s have different but known variances.)*

(ii) $\frac{\sqrt{k}(\bar{\hat{\theta}}-\theta)}{\sqrt{v_1(\hat{\theta})}}$ *is distributed as a $t_{(k-1)}$ variable.*

If Var $(\bar{\hat{\theta}}) = \sigma^2/k$ is known, or k is large, $100(1-\alpha)$ % confidence interval for θ is

$$\bar{\hat{\theta}} \pm \tau_{\alpha/2}\sigma/\sqrt{k}$$

where $\tau_{\alpha/2}$ is the upper $100(\alpha/2)$ percentage point of the $N(0, 1)$ distribution. When Var $\hat{\theta}_i$ is not known or k is not large $100(1-\alpha)$ % confidence interval for θ is

$$\bar{\hat{\theta}} \pm t_{k-1;\alpha/2}\sqrt{v(\bar{\hat{\theta}})}$$

where $t_{k-1;\alpha/2}$ is the upper $100(\alpha/2)$ percentage point of the $t_{(k-1)}$ distribution.

(b) *Random groups are not independent*: In practical sample surveys, samples are often selected as a whole using some form of without replacement sampling instead of in the form of a series of independent random groups. Random groups are now formed by randomly dividing the parent sample into k groups. The random group estimators $\hat{\theta}_i$'s are no longer uncorrelated because sampling is performed without replacement. Theorem 2.5.1 is no longer valid. Here also $\bar{\hat{\theta}}$, $v_1(\hat{\theta})$, $v_2(\hat{\theta})$ as defined above are used for the respective purposes. However, because the random group estimators are not independent, $v_1(\hat{\theta}) = v(\bar{\hat{\theta}})$ is not an unbiased estimator of Var $(\bar{\hat{\theta}})$. The following theorem describes some properties of $v(\bar{\hat{\theta}})$.

Theorem 2.5.2 *If $E(\hat{\theta}_i) = \mu_i (i = 1, \ldots, k)$,*

$$E\{v(\bar{\hat{\theta}})\} = \mathit{Var}\ (\bar{\hat{\theta}}) + \frac{1}{k(k-1)}\left[\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2 - 2\sum\sum_{i<j=1}^{k}\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j)\right]. \tag{2.5.34}$$

Proof It is obvious that

$$E(\bar{\hat{\theta}}) = \bar{\mu} = \sum_{i=1}^{k}\mu_i/k.$$

Again,

$$\begin{aligned} v(\bar{\hat{\theta}}) &= \tfrac{1}{k(k-1)}\left[\sum_{i=1}^{k}\hat{\theta}_i^2 - k\bar{\hat{\theta}}^2\right] \\ &= \bar{\hat{\theta}}^2 - \tfrac{2}{k(k-1)}\sum\sum_{i<j=1}^{k}\hat{\theta}_i\hat{\theta}_j. \end{aligned}$$

Now,

$$E[\bar{\hat{\theta}}^2] = Var(\bar{\hat{\theta}}) + \bar{\mu}^2$$

and

$$E[\hat{\theta}_i\hat{\theta}_j] = \text{ Cov } (\hat{\theta}_i, \hat{\theta}_j) + \mu_i\mu_j.$$

Therefore, the result follows. □

Theorem 2.5.2 gives the bias of $v(\bar{\hat{\theta}})$ as an estimator of Var $(\bar{\hat{\theta}})$. For large populations and small sampling fractions, the term $2\sum\sum_{i<j}$ Cov $(\hat{\theta}_i, \hat{\theta}_j)$ will tend to be a relatively small negative quantity. The quantity

$$\frac{1}{k(k-1)}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2$$

will also be relatively small if $\mu_i \approx \bar{\mu}(i = 1, \ldots, k)$. Thus the bias of $v(\bar{\hat{\theta}})$ will be unimportant in many large-scale sample surveys and will tend to be slightly positive. Work by Frankel (1971) suggests that the bias of $v(\bar{\hat{\theta}})$ is often small and decreases as the size of the groups increase (or equivalently as the number of groups decreases).

The RG procedure was initiated by Mahalanobis (1946) and Deming (1956). Mahalanobis called the various samples as *Interpenetrating samples*, Deming proposed the term *replicated samples*. They selected k independent samples using the same sampling design and used the estimator of the type (2.5.29) to estimate the variance of the overall estimator. In RG method, the major difference is that the replicates are not necessarily formed independently.

It has been found that if $\hat{\theta}_1, \ldots, \hat{\theta}_k$ are independently and identically distributed random variables, then coefficient of variation (cv) of the RG estimator $v(\bar{\hat{\theta}})$, which measures its stability is

$$\begin{aligned} cv[v(\bar{\hat{\theta}})] &= [Var\{v(\bar{\hat{\theta}})\}]^{1/2}/Var(\bar{\hat{\theta}}) \\ &= \left\{\tfrac{\beta_4(\hat{\theta}_1)-(k-3)/(k-1)}{k}\right\}^{1/2}. \end{aligned}$$

The cv is thus an increasing function of kurtosis $\beta_4(\hat{\theta}_1)$ of the distribution of $\hat{\theta}_1$ and a decreasing function of k for a wide range of complex surveys (when N, n are large and $n/N \approx 0$, the result holds even for nonindependent RG's). As a result, the larger the number (k) of groups, the higher the precision, though computational cost will increase at the same time. The optimum value of k is a trade-off between cost and precision. The RG method is suitable for surveys using a large number of primary-stage units (psu's) where many psu's are selected per stratum.

2.5.4 Balanced Repeated Replications

The method of balanced half-sample repeated replications (BRR) has proved very useful for surveys in which two primary-stage units (psu's) are selected per stratum. Following Plackett and Burman (1946), McCarthy (1966, 1969a, b) introduced the concept of BRR, also known as balanced half-samples, balanced fractional samples, and pseudoreplication.

Suppose that two units are selected by *srswr* from each of H strata for estimating $\bar{Y} = \sum_h W_h \bar{Y}_h$ where $W_h = N_h/N, N_h(n_h)$ is the stratum population (sample) size, $\bar{Y}_h = \sum_{i=1}^{N_h} Y_{hi}/N_h$, Y_{hi}, being the value of 'y' on the ith unit in stratum h.

By selecting one unit from the sampled units in each stratum at random we can form 2^H sets of two half-samples (HS's) each such that each set forms a complete replicate.

In a set α denote the two HS's as $S_\alpha, S_{\alpha'}$ with the corresponding estimates $\bar{y}_{st,\alpha} = \sum_h W_h y_{h1,\alpha}$ and $\bar{y}_{st,\alpha'} = \sum_h W_h y_{h2,\alpha}$. (A more complicated notation is given below). The customary estimator is

$$\bar{y}_{st(\alpha)} = \frac{\bar{y}_{st,\alpha} + \bar{y}_{st,\alpha'}}{2}.$$

The αth replicate estimate of $V(\bar{y}_{st})$ is

$$\begin{aligned} v_\alpha(\bar{y}_{st}) &= \tfrac{1}{2}[(\bar{y}_{st,\alpha} - \bar{y}_{st(\alpha)})^2 + (\bar{y}_{st,\alpha'} - \bar{y}_{st(\alpha)})^2] \\ &= \tfrac{1}{4}(\bar{y}_{st,\alpha} - \bar{y}_{st,\alpha'})^2. \end{aligned} \tag{2.5.35}$$

The estimator v_α is unbiased for $V(\bar{y}_{st})$ (Exercise 2.2).

Now,

$$\bar{y}_{st,\alpha} = \sum_h W_h y_{h1} = \sum_h W_h\{Y_{h1}\delta_{h1\alpha} + Y_{h2}\delta_{h2\alpha}\} \tag{2.5.36}$$

wherein we denote the values of y on the two units selected from the hth stratum as Y_{h1}, Y_{h2}, respectively, in some well-defined manner. The term Y_{h1} becomes y_{h1} if the corresponding unit goes to S_α. Also,

$$\begin{aligned} \delta_{h1\alpha} &= 1(0) \text{ if the unit } (h, 1) \in S_\alpha \text{ (otherwise)} \\ \delta_{h2\alpha} &= 1 - \delta_{h1\alpha}. \end{aligned} \tag{2.5.37}$$

Now,

$$\bar{y}_{st,\alpha} - \bar{y}_{st} = \frac{1}{2}\sum_h W_h \delta_h^\alpha d_h \tag{2.5.38}$$

where

$$\begin{aligned}\delta_h^\alpha &= 2\delta_{h1\alpha} - 1\\ d_h &= y_{h1} - y_{h2}.\end{aligned} \tag{2.5.39}$$

Hence,

$$\begin{aligned} v_\alpha &= \tfrac{1}{4}\left(\sum_h W_h \delta_h^\alpha d_h\right)^2\\ &= \tfrac{1}{4}\left[\sum_h W_h^2 d_h^2 + 2\sum\sum_{h<h'} W_h W_{h'} \delta_h^\alpha \delta_{h'}^\alpha d_h d_{h'}\right]. \end{aligned} \tag{2.5.40}$$

It follows that

$$\frac{1}{2^H}\sum_{\alpha=1}^{2^H} \bar{y}_{st,\alpha} = \bar{y}_{st}, \quad \frac{1}{2^H}\sum_{\alpha=1}^{2^H} v_\alpha = v(\bar{y}_{st}). \tag{2.5.41}$$

When H is large, computation of $v(\bar{y}_{st})$ as the average of v_α over 2^H HS's becomes formidable. However, if we choose a set η of K HS's such that

$$\sum_{\alpha\in\eta} \delta_h^\alpha \delta_{h'}^\alpha = 0, h < h' = 1, \ldots, H, \tag{2.5.42}$$

then

$$\bar{v}_{(K)} = \sum_{\alpha\in\eta} v_\alpha / K = v(\bar{y}_{st}). \tag{2.5.43}$$

Plackett and Burman (1946) developed a method for constructing $m \times m$ orthogonal matrices with entries $+1, -1$ where m is a multiple of 4. These can be used directly to obtain values of δ_h^α satisfying (2.5.42). The orthogonal matrix of size K where K is a multiple of 4, between H and $H + 3$, can be used dropping the last $K - H$ columns. The entries in the matrix can be substituted as δ_h^α, each column standing for a stratum. McCarthy referred to the set η as balanced. If, further, the condition

$$\sum_{\alpha\in\eta} \delta_h^\alpha = 0, h = 1, \ldots, H \tag{2.5.44}$$

is satisfied, then $\bar{y}_{st,\alpha}/K = \bar{y}_{st}$. The set of replicates satisfying (2.5.42) and (2.5.44) is set to be in full orthogonal balance.

For designs with *wor* sampling of psu's, v_α is positively biased. In this case, a separate adjustment is necessary to account for this bias, though the bias is generally negligible.

In the nonlinear case, in which the BRR is most useful, let $\hat{\theta}$, $\hat{\theta}_\alpha$, $\hat{\theta}_{\alpha'}$ be the estimates of θ based on the whole sample, S_α, and $S_{\alpha'}$, respectively. Let $\bar{\hat{\theta}}_\alpha = (\hat{\theta}_\alpha + \hat{\theta}_{\alpha'})/2$. We note that even for a balanced set of HS's, $\sum_{\alpha\in\eta}\hat{\theta}_\alpha/K = \hat{\hat{\theta}}_\alpha \neq \hat{\theta}$ in general. Empirical studies by Kish and Frankel (1970), among others, however, show that $\bar{\hat{\theta}}_\alpha$ is very close to $\hat{\theta}$ in general. Writing

$$\begin{aligned}\bar{v}_{(K)}(\hat{\theta}) &= \sum_{\alpha\in\eta}(\hat{\theta}_\alpha - \hat{\theta})^2/K\\ \bar{v}'_{(K)}(\hat{\theta}) &= \sum_{\alpha\in\eta}(\hat{\theta}_{\alpha'} - \hat{\theta})^2/K,\end{aligned} \tag{2.5.45}$$

we have the following alternative variance estimators:

$$\begin{array}{ll}\text{(i)} & \bar{v}_{(K)}(\hat{\theta})\\ \text{(ii)} & \bar{v}'_{(K)}(\hat{\theta})\\ \text{(iii)} & [\bar{v}_{(K)}(\hat{\theta}) + \bar{v}'_{(K)}(\hat{\theta})]/2 = \bar{\bar{v}}_{(K)}(\hat{\theta})\\ \text{(iv)} & \sum_{\alpha\in\eta}(\hat{\theta}_\alpha - \hat{\theta}_{\alpha'})^2/(4K) = \bar{v}^{+}_{(K)}(\hat{\theta}).\end{array} \tag{2.5.46}$$

The estimators (i), (ii), (iii) are sometimes regarded as estimators of mse$(\hat{\theta})$, while (iv) is regarded as estimator of Var $(\hat{\theta})$.

Since (iii) is the average of (i) and (ii), it is at least as precise as the others and equally biased. However, (iii) is comparatively costlier than (i) (and (ii)) and perhaps, significantly so, when many estimators are produced.

Another set of variance estimators can be attained by replacing $\hat{\theta}$ by $\hat{\hat{\theta}} = \sum_\alpha \hat{\theta}_\alpha/K$ or $\sum_\alpha \hat{\theta}_{\alpha'}/K$ in (i), (ii), and (iii) of (2.5.46). Such estimators are unbiased for linear $\hat{\theta}$ only if the number of HS's is $T > H$. If H is a multiple of 4, $T(= H + 4)$ HS's must be used to maintain the unbiasedness (Lemeshow and Epp 1977). The estimators using $\hat{\hat{\theta}}$ are generally not preferred to those using $\hat{\theta}$, since they give smaller and less conservative estimates of mse, as they do not include the components for bias of $\hat{\theta}$. Empirical works of McCarthy (1969a, b), Kish and Frankel (1970), Levy (1971), Frankel (1971) and others show that BRR provides satisfactory estimates of the true variance.

All the above-mentioned BRR estimators become identical in the linear case.

Two modifications of BRR has been proposed, that require fewer replicates but the corresponding estimates are less precise and equally biased as the full BRR estimate. In one modification strata are combined into groups, not necessarily of the same size. For each replicate all strata into a group g are assigned the same value δ_h^α. The constraints (2.5.42) are imposed for pairs h, h' of strata which are not in the same group g. Thus if G groups are formed, the number of replicates required is the multiple of 4, which lies in the range G to $G + 3$. The Plackett and Burman matrices of size K may then be used to derive the values of δ_h^α.

The second procedure for reducing the number of replicates, discussed by McCarthy (1966) and developed by Lee (1972, 1973) is the method of partially balanced repeated replications (PBRR). Here the strata are divided into groups and full balancing are applied to the strata within each group. If T replicates are required for H strata, $G = H/T$ groups are formed with T strata in each. A $T \times T$ orthogonal matrix is then used to ensure a full balance within each group. Lee (1972, 1973), Rust (1984) suggested methods of implementing PBIB that would minimize the loss in precision over fully balanced BRR.

Rust (1984, 1986) shows that the method of PBRR and combined strata are equivalent. However, the combined strata method has a greater flexibility in the sense that the number of strata per group varies.

For general designs in which strata sample sizes vary, BRR can be implemented by dividing the psu's in each stratum into two groups of equal sizes (assuming $n_h = 2m_h$, m_h an integer), and then using these groups as units (Kish and Frankel 1970). In this case the BRR variance estimator is somewhat less precise than the customary variance estimator. Valliant (1987) considers the large-sample prediction properties of the BRR separate ratio and regression estimator under a superpopulation model when n_h is large and compares these with jackknife and linearization procedure.

Example 2.5.3 Let us consider BRR for estimation of population ratio $R = Y/X$. A ratio estimator of R based on the set S_α is

$$\hat{r}_\alpha = \frac{\sum_h y_{h1}}{\sum_h x_{h1}} = \frac{\sum_h (Y_{h1}\delta_{h1\alpha} + Y_{h2}\delta_{h2\alpha})}{\sum_h (X_{h1\alpha}\delta_{h1\alpha} + X_{h2\alpha}\delta_{h2\alpha})}, \quad \alpha = 1, \dots, 2^H.$$

Consider variance estimator for the mean of α-HS estimators

$$\bar{\hat{r}}_\alpha = \sum_{\alpha=1}^{2^H} \hat{r}_\alpha / 2^H.$$

The parent estimator of population ratio R is

$$\hat{r} = \frac{\sum_h (y_{h1} + y_{h2})}{\sum_h (x_{h1} + x_{h2})}.$$

Estimator of $V(\hat{r})$ is

$$\begin{aligned}
(i)\ \ \bar{v}(\hat{r}) &= \sum_{\alpha=1}^{2^H} (\hat{r}_\alpha - \hat{r})^2/2^H, \\
(ii)\ \ \bar{v}'(\hat{r}) &= \sum_{\alpha'=1}^{2^H} (\hat{r}_{\alpha'} - \hat{r})^2/2^H, \\
(iii)\ \ \bar{\bar{v}}(\hat{r}) &= [\bar{v}(\hat{r}) + \bar{v}'(\hat{r})]/2, \\
(iv)\ \ \bar{v}^{+}(\hat{r}) &= \sum_{\alpha=1}^{2^H} (\hat{r}_\alpha - \hat{r}_{\alpha'})^2/[4(2^H)].
\end{aligned} \qquad (2.5.47)$$

Three other estimators are obtained by replacing in (i), (ii), and (iii) of (2.5.47) $\hat{r}$ by $\hat{\hat{r}}(=\sum_{\alpha}\hat{r}_{\alpha}/2^H$ or $\sum_{\alpha'}\hat{r}_{\alpha'}/2^H)$. Since these estimators are nonlinear, they are not identical. For example,

$$\bar{\bar{v}}(\hat{r})=\bar{v}^{+}(\hat{r})+\sum_{\alpha=1}^{2^H}(\bar{\hat{r}}_{\alpha}-\hat{r})^2/(2^H)$$

where $\bar{\hat{r}}_{\alpha}=(\hat{r}_{\alpha}+\hat{r}_{\alpha'})/2$ and hence

$$\bar{\bar{v}}(\hat{r})\geq\bar{v}^{+}(\hat{r}).$$

One problem that occasionally arises in BRR is that one or more replicate estimates will remain undefined due to division by zero. This happens particularly often when ratio estimator has been used with very small cell sizes. Fay suggested a solution to this problem: Instead of increasing the weight of one HS by 100 % and decreasing the weight of the other HS to zero, he recommended perturbing the weights by +/− 50 %. Judkins (1990) evaluated Fay's techniques through simulation and also discusses further modification to the techniques that are used for variance estimation when only one psu is selected per stratum.

2.5.5 The Jackknife Procedures

Quenouille (1949, 1956) originally introduced jackknife (JK) as a method of reducing the bias of an estimator. Tukey (1958) suggested the use of this technique for variance estimation. Durbin (1953) first considered its use in finite population. Extensive discussion of JK method is given in Miller (1964, 1974), Gray and Schucany (1972) and in a monograph by Efron (1982).

Let θ be the parameter to be estimated. An estimator $\hat{\theta}$ is obtained from the full sample. Assuming $n=mk(m,k$ integers), we partition the sample into k groups of m original observations each. Let $\hat{\theta}_{(\alpha)}$ be the estimator of θ computed from the whole sample except the αth group. Define pseudo-values $\hat{\theta}_{\alpha}$ as

$$\hat{\theta}_{\alpha}=k\hat{\theta}-(k-1)\hat{\theta}_{(\alpha)}. \qquad (2.5.48)$$

Quenouille's estimator is

$$\bar{\hat{\theta}}=\frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_{\alpha}. \qquad (2.5.49)$$

Tukey suggested that $\hat{\theta}_\alpha$ are approximately independently and identically distributed. The JK estimator of variance is

$$\begin{aligned} v_1(\hat{\theta}) &= \frac{1}{k(k-1)} \sum_{\alpha=1}^{k} (\hat{\theta}_\alpha - \bar{\hat{\theta}})^2 \\ &= \frac{k(k-1)}{k} \sum_{\alpha=1}^{k} (\hat{\theta}_{(\alpha)} - \hat{\theta}_{(.)})^2 \end{aligned} \tag{2.5.50}$$

where $\hat{\theta}_{(.)} = \sum_{\alpha=1}^{k} \hat{\theta}_{(\alpha)}/k$. In practice, $v_1(\bar{\hat{\theta}})$ is used not only to estimate the variance of $\bar{\hat{\theta}}$, but also of $\hat{\theta}$. Alternatively, one may use

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} (\hat{\theta}_\alpha - \hat{\theta})^2 \tag{2.5.51}$$

which is always at least as large as $v_1(\hat{\theta})$.

The number of groups k is determined from the point of view of computational cost and the precision of the resulting estimator. The precision is maximized when each dropout group is of size one and each unit is dropped only once. Rao (1965), Rao and Webster (1966), Chakraborty and Rao (1968), Rao and Rao (1971) in their studies on ratio estimator based on superpopulation models showed that both bias and variance of $\bar{\hat{\theta}}$ are maximized for the choice $k = n$.

Brillinger (1966) showed that both v_1 and v_2 give plausible estimates of the asymptotic variance. Shao and Wu (1989), Shao (1989) considered the efficiency and consistency of JK variance estimators. For nonlinear statistic $\hat{\theta}$ that can be expressed as functions of estimated means of p variables, such as ratio, regression, correlation coefficient, Krewski and Rao (1981) established the asymptotic consistency of variance estimators from JK, the linearization, and BRR methods. In the case of two samples psu's per stratum, Rao and Wu (1985) showed that the linearization and JK variance estimators are asymptotically efficient. In case of item nonresponse in sample surveys, Rao and Shao (1999) considered jackknife variance estimation for stratified multistage surveys which is obtained by first adjusting the hot deck imputed values for each pseudoreplicate and then applying the standard jackknife formulae. Rao and Tausi (2004) considered variance estimation for the *generalized regression estimator* (GREG) of a total based on p auxiliary variables under stratified multistage sampling. Customary resampling procedures, like jackknife, balanced repeated replication, and bootstrap (Sect. 2.5.7) for estimating the variance of a GREG estimator requires the inversion of a $p \times p$ matrix for each subsample. This may result in illconditioned matrices for some subsamples. The authors applied the estimating function resampling methods to obtain variance estimators using jackknife resampling.

2.5.6 The Jackknife Repeated Replication (JRR)

This is a combination of JK and BRR techniques. We assume as in the case of BRR, that two clusters are selected with replacement from each of H strata. We construct the pseudo-samples following the method suggested by Frankel (1971).

For the first pseudo-sample, we exclude the cluster (1, 1) (i.e., the cluster 1 in stratum 1), weigh the second cluster (1, 2) by 2 and leave the sampled clusters in the remaining $H-1$ strata unchanged. By repeating the procedure for all the strata we get a total of H pseudo-samples. These are

$$\begin{aligned}\text{First Pseudo-sample } &: \{(2y_{12});\ (y_{21}, y_{22}),\ (y_{31}, y_{32}), \ldots (y_{H1}, y_{H2})\},\\ \text{Second Pseudo-Sample } &: \{(y_{11}, y_{12}),\ (2y_{22}),\ (y_{31}, y_{32}), \ldots, (y_{H1}, y_{H2})\},\\ &\ldots\\ H\text{th Pseudo-Sample } &: \{(y_{11}, y_{12}),\ (y_{21}, y_{22}),\ (y_{31}, y_{32}), \ldots, (2y_{H2})\}.\end{aligned}$$

Changing the order of excluded clusters we get another set of H pseudo-samples.

The JRR variance estimators are derived using these two sets of pseudo-samples. We illustrate this by the example of finding the variance estimator of combined ratio estimator $\hat{r}$.

For this, we first construct ratio estimator for each pseudo-sample. The estimator of population ratio r based on the hth pseudo-sample in the first set is

$$\hat{r}_h = \frac{2y_{h2} + \sum_{h'(\neq h)=1}^{H} \sum_{\alpha=1}^{2} y_{h'\alpha}}{2x_{h2} + \sum_{h'(\neq h)=1}^{H} \sum_{\alpha=1}^{2} x_{h'\alpha}}, \quad h = 1, \ldots, H. \tag{2.5.52}$$

Similarly, the estimator of the population ratio based on the hth pseudo-sample of the second set is

$$\hat{r}_h^c = \frac{2y_{h1} + \sum_{h'(\neq h)=1}^{H} \sum_{\alpha=1}^{2} y_{h'\alpha}}{2x_{h1} + \sum_{h'(\neq h)=1}^{H} \sum_{\alpha=1}^{2} x_{h'\alpha}}, \quad h = 1, \ldots, H. \tag{2.5.53}$$

Using the pseudo-sample estimators $\hat{r}^h$, $\hat{r}_h^c$, $h = 1, \ldots, H$ we get different JRR estimators of variance of $\hat{r}$. These are

$$v_{1,jrr}(\hat{r}) = \frac{1}{H} \sum_{h=1}^{H} (\hat{r}_h - \hat{r})^2, \tag{2.5.54}$$

$$v_{2,jrr}(\hat{r}) = \frac{1}{H} \sum_{h=1}^{H} (\hat{r}_h^c - \hat{r})^2, \tag{2.5.55}$$

$$v_{3,jrr}(\hat{r}) = \frac{1}{2}(v_{1,jrr} + v_{2,jrr}). \tag{2.5.56}$$

Another set of variance estimators can be obtained by using the estimator $\hat{r}$ first corrected for its bias using pseudo-values as in the JK procedure. The pseudo-value of $\hat{r}_h$ is, following (2.5.52),

$$\hat{r}_h^p = 2\hat{r} - \hat{r}_h, \quad h = 1, \dots, H. \tag{2.5.57}$$

A bias-corrected estimator of r is, therefore,

$$\bar{\hat{r}}^p = \frac{1}{H}\sum_{h=1}^{H} \hat{r}_h^p. \tag{2.5.58}$$

Similarly, the pseudo-value of $\hat{r}_h^c$ is

$$\hat{r}_h^{pc} = 2\hat{r} - \hat{r}_h^c, \quad h = 1, \dots, H. \tag{2.5.59}$$

A bias-corrected estimator based on $\hat{r}_h^c (h = 1, \dots, H)$ is, therefore,

$$\bar{\hat{r}}^{pc} = \frac{1}{H}\sum_{h=1}^{H} \hat{r}_h^{pc}. \tag{2.5.60}$$

Following (2.5.54)–(2.5.56) we, therefore, get the following variance estimators of $\hat{r}$:

$$v_{4,jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^p - \bar{\hat{r}}^p)^2 / \{H(H-1)\}, \tag{2.5.61}$$

$$v_{5,jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^{pc} - \bar{\hat{r}}^{pc})^2 / \{H(H-1)\}, \tag{2.5.62}$$

$$v_{6,jrr}(\hat{r}) = (v_{4,jrr} + v_{5,jrr})/2. \tag{2.5.63}$$

Finally, from all the $2H$ pseudo-samples we obtain

$$v_{7,jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r}_h^c)^2/4. \tag{2.5.64}$$

For a nonlinear estimator, the bias-corrected JRR estimators and the parent estimator coincide. In practice all the JRR variance estimators should give closely related results.

The method can be extended to a more general case where more than two clusters are selected from each stratum without replacement (see Wolter 1985, Sect. 4.6).

2.5.7 The Bootstrap

The bootstrap (BS) method is the most recent technique of variance estimation for complex sample surveys. The technique uses a highly computer-intensive resampling procedure to mimic the theoretical distribution from which the sample is derived. The method does not need any prior assumption about the distribution of observations or the estimators. It provides estimates of bias and standards errors and other distributional properties of the estimators, however complex it may be.

The naive BS technique was suggested by Efron (1979) who indicated that the method may be better than its competitors. The BS method for finite population sampling was introduced and discussed by Gross (1980), Bickel and Freedman (1984), Chao and Lo (1985), McCarthy and Snowdon (1985), Booth et al. (1991), among others. Rao and Wu (1988) showed the application of the BS in design-based survey sampling under different sampling designs including stratified cluster sampling with replacement, stratified simple random sampling without replacement, unequal probability random sampling without replacement, and two-stage cluster sampling with equal probabilities and without replacement. Rao (2006) showed that BS method provides an alternative option to the analysis of complex surveys for taking account of the design effects and weight adjustments. We consider here the elements of variance estimation by BS.

Suppose we have p variables $y_1, \dots, y_p$ with $Y_{hij}(y_{hij})$, $\bar{Y}_j$, $\bar{Y}_{hj}$, $\bar{y}_{hj}$ as the value of y_j on the ith unit in the hth stratum in the population (sample), population mean, stratum population mean, stratum sample mean of y_j, respectively ($h = 1, \dots, H; j = 1, \dots, p$; $\bar{Y}_j = \sum_h W_h \bar{Y}_{hj}$, $\bar{Y}_{hj} = \sum_{i=1}^{N_h} Y_{hij}/N_h$, $\bar{y}_{hj} = \sum_{i \in s_h} y_{hij}/n_h$, N_h, n_h being, respectively, the size of sample s_h and population in stratum h). Suppose we want to estimate $\theta = g(\bar{Y}_1, \dots, \bar{Y}_p) = g(\bar{\mathbf{Y}})$, a nonlinear function of $\bar{\mathbf{Y}} = (\bar{Y}_1, \dots, \bar{Y}_p)'$. This includes population ratio, regression, correlation coefficient, etc. A natural estimator of θ, whenever $n_h(\geq 2)$ psu's are selected with replacement from each stratum is $g(\hat{\bar{\mathbf{Y}}}) = g(\bar{\mathbf{y}})$ where $\bar{\mathbf{y}} = (\bar{y}_1, \dots, \bar{y}_p)'$, $\bar{y}_j = \sum_h W_h \bar{y}_{hj}$. We denote by $\bar{\mathbf{y}}_h = (\bar{y}_{h1}, \dots, \bar{y}_{hp})'$.

In this case, the random vector $\mathbf{y}_{hi} = (y_{hi1}, \dots, y_{hip})'$, $i = 1, \dots, n_h$ are iid with $E(\mathbf{y}_{hi}) = \bar{\mathbf{Y}}_h = (\bar{Y}_{h1}, \dots, \bar{Y}_{hp})'$. The vectors $\mathbf{y}_{hi}, \mathbf{y}_{h'k} (h \neq h')$ are independently but not necessarily identically distributed. The BS sampling procedure is as follows:

(a) Draw a random sample wr $\{\mathbf{y}^*_{hi}, i = 1, \dots, n_h\}$ of size n_h from the given sample $\{\mathbf{y}_{hi}, i = 1, \dots, n_h\}$ independently from each stratum. Calculate $\bar{\mathbf{y}}^*_h = \sum_i \mathbf{y}^*_{hi}/n_h$, $\bar{\mathbf{y}}^* = \sum_h W_h \bar{\mathbf{y}}^*_h$ and $\hat{\theta}^* = g(\bar{\mathbf{y}}^*)$.

(b) Repeat step (a) a large number of times, say B times and calculate the corresponding estimates $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ of θ.

(c) Calculate the Monte Carlo estimate of $V(\hat{\theta})$,

$$v_b(a) = \sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2/(B-1) \tag{2.5.65}$$

where $\hat{\theta}^{*\cdot} = \sum_{b=1}^{B} \hat{\theta}^{*b}/B$.

The estimator $v_b(a)$ is a fair approximation to the BS variance estimator of $\hat{\theta}$,

$$v_b = var_*(\hat{\theta}^*) = E_*(\hat{\theta}^* - E_*(\hat{\theta}^*))^2 \tag{2.5.66}$$

where E_* denotes expectation with respect to BS sampling. The BS estimator $E_*(\hat{\theta}^*)$ of θ is approximated by $\hat{\theta}^{*\cdot}$.

In the linear case with $p = 1$, $\theta = \bar{Y}$, $\hat{\theta}^* = \sum_h W_h \bar{y}_h^* = \bar{y}^*$ and v_b reduces to

$$var_*(\bar{y}^*) = \sum_h W_h^2 \sigma_h^2/n_h \tag{2.5.67}$$

where $\sigma_h^2 = \sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2/n_h$. Comparing (2.5.67) with the customary estimator $v(\bar{y}_{st}) = \sum_h W_h^2 s_h^2/n_h$ where $s_h^2 = \sum_i (y_{hi} - \bar{y}_h)^2/(n_h - 1)$, it follows that $var_*(\bar{y}^*)/v(\bar{y}_{st})$ does not converge in probability to 1 when n_h is bounded. Hence, $var_*(\bar{y}^*)$ is not a consistent estimator of $V(\bar{y}_{st})$ unless n_h and $f_h = n_h/N_h$ are constants for all h. Moreover, v_b in (2.5.66) is not a consistent estimator of the variance of a general nonlinear estimator.

Recognizing this problem Efron (1982) suggested to draw BS sample of size $n_h - 1$ with *srswr* sampling design instead of n_h independently from each stratum. The rest of the procedure is the same as before.

To get rid of this difficulty, Rao and Wu (1988) suggested the following resampling procedure.

(i) Draw a random sample $\{y_{hi}^*, i = 1, \ldots, m_h\}$ with replacement ($m_h \geq 1$) from the original sample $\{y_{hi}, i = 1, \ldots, n_h\}$. Calculate

$$\begin{aligned}
\tilde{y}_{hi} &= \bar{y}_h + \tfrac{\sqrt{m_h}}{\sqrt{n_h - 1}}(y_{hi}^* - \bar{y}_h) \\
\tilde{y}_h &= \sum_i \tilde{y}_{hi}/m_h = \bar{y}_h + \tfrac{\sqrt{m_h}}{\sqrt{n_h - 1}}(\bar{y}_h^* - \bar{y}_h) \\
\tilde{y} &= \sum_h W_h \tilde{y}_h, \quad \tilde{\theta} = g(\tilde{y}).
\end{aligned} \tag{2.5.68}$$

(ii) Repeat the step (i) B times independently and calculate the corresponding estimates $\tilde{\theta}^1, \ldots, \tilde{\theta}^B$. The BS estimator $E_*(\tilde{\theta})$ of θ is approximated by $\tilde{\theta}^{\cdot} = \sum_b \tilde{\theta}^b/B$.

(iii) The BS variance estimator of $\hat{\theta}$,

$$\tilde{v}_b = E_*(\tilde{\theta} - E_*(\tilde{\theta}))^2$$

is approximated by the Monte Carlo estimator

$$\tilde{v}_b(a) = \sum_{b=1}^{B} (\tilde{\theta}^b - \tilde{\theta}^{\cdot})^2/(B-1).$$

In the linear case of $\theta = \bar{Y}$ with $p = 1$, $\tilde{v}_b$ reduces to the customary unbiased variance estimator $v(\bar{y}_{st})$ for any choice of m_h, because,

$$\tilde{v}_b = E_*(\tilde{y} - \bar{y})^2 = \sum_h W_h^2 \cdot \tfrac{m_h}{n_h - 1} E_*(\bar{y}_h^* - \bar{y}_h)^2$$

$$= \sum_h W_h^2 \cdot \tfrac{m_h}{n_h - 1} \cdot \tfrac{(n_h-1)s_h^2}{m_h n_h} = \sum_h W_h^2 \tfrac{s_h^2}{n_h} = v(\bar{y}_{st}).$$

Thus, Rao and Wu (1988) applied the previously stated algorithm of a naive BS procedure with a general sample size m_h not necessarily equal to n_h, but rescaled the resampled values appropriately so that the resulting variance estimator is the same as the usual unbiased variance estimator in the linear case.

In the nonlinear case it has been shown that under certain conditions

$$\tilde{v}_b = v_L + 0(n^{-2})$$

where v_L is the customary linearization variance estimator,

$$v_L = \sum \sum_{j,k=1}^{p} g_j(\bar{\mathbf{y}}) g_k(\bar{\mathbf{y}}) \sum_{h=1}^{H} \frac{W_h^2}{n_h} s_{hjk},$$

where for $\mathbf{t} = (t_1, \ldots, t_p)' g_j(\mathbf{t}) = \frac{\partial g(\mathbf{t})}{\partial t_j}$, $s_{hjk} = \sum_{i \in s_h} (y_{hij} - \bar{y}_{hj})(y_{hik} - \bar{y}_{hk})$. Since v_L is a consistent estimator of the variance of $\hat{\theta}$, $\tilde{v}_b$ is consistent for Var $(\hat{\theta})$.

It has also been found that the estimate of bias of $\hat{\theta}$, $B(\hat{\theta}) = E(\hat{\theta}) - \theta$, based on the suggested BS procedure, which is $\tilde{B}(\hat{\theta}) = E_*(\tilde{\theta}) - \hat{\theta} = \tilde{\theta}^{\cdot} - \hat{\theta}$, is consistent, while the same based on the naive BS procedure is not consistent.

The choice $m_h = n_h - 1$ gives $\tilde{y}_{hi} = y_{hi}^*(i = 1, \ldots, m_h)$ and the method reduces to the naive BS. For $n_h = 2$ and $m_h = 1$, the method reduces to the well-known random half-sample replication. For $n_h \geq 5$, $m_h \approx n_h - 3$ Rao and Wu made some empirical studies on the choice of m_h.

The method can be easily extended to simple random sampling within stratum by changing $\tilde{y}_{hi}$ in (2.5.68) to

$$\tilde{y}_{hi} = \bar{y}_h + \frac{\sqrt{m_h}}{\sqrt{n_h - 1}} \cdot \sqrt{1 - f_h}(y^*_{hi} - \bar{y}_h) \tag{2.5.69}$$

where $f_h = n_h/N_h, h = 1, \ldots, H$. Here, even by choosing $m_h = n_h - 1$, we do not get $\tilde{y}_{hi} = y^*_{hi}$. Hence, the naive BS using y^*_{hi} will still give a wrong scale.

The method has been extended to any unbiased sampling strategy including Rao–Hartley–Cochran sampling procedure.

Apart from these with replacement procedures, a without replacement BS technique was proposed by Gross (1980) in the case of a single stratum. His method assumes that $N = Rn$ for some integer R and creates a pseudopopulation of size N by replicating the data R times. However, the method does not yield the usual unbiased estimate of variance in the linear case. The difficulty was corrected by Bickel and Freedman (1984) who proposed a randomization between two pseudopopulations and also allowed an extension of the method for $H > 1$. Sitter (1992) developed a BS procedure which retains the desirable properties of both with replacement BS and without replacement BS techniques but extends to more complex without replacement sampling designs.

Hall (1989) considered three efficient bootstrap algorithms: these are balanced bootstrap and the linear approximation method proposed by Davison et al. (1986) and a centering method proposed by Efron (1982) in the context of bias estimation. He compares the asymptotic performance of these methods and show that they are asymptotically equivalent. Hall prove that the variances and mean square errors of all these three algorithms are asymptotic to the same constant multiple of $(Bn^2)^{-1}$, where B denotes the number of bootstrap resamples and n is the size of the original sample. The convergence rate $(Bn^2)^{-1}$ represents a significant improvement on that for the more usual, unbalanced bootstrap algorithm, which has mean square error of only $(Bn)^{-1}$. These results apply to smooth functions of means.

Ahmad (1997) suggested a new bootstrap variance estimation technique, *rescaling bootstrap without replacement* technique and also proposed an optimum choice of bootstrap sample size for his proposed procedure. Canty and Davison (1999) considered labor force surveys to demonstrate the advantages of resampling methods in estimation of variance. Labor force surveys are conducted to estimate, among others, quantities such as unemployment rate and the number of people at work. Interest focusses typically both in estimates at a given time and in changes between two successive time points. Calibration of the sample to ensure agreement with the known population margins results in random weights being assigned to each response, but the usual method of variance estimation do not account for this. The authors describe how resampling methods, such as jackknife, jackknife linearisation, balanced repeated replication and bootstrap can be used to do so. Robert et al. (2004) suggested a design-based bootstrapping method for the estimation of variance of an estimator in longitudinal surveys.

Multiple imputation is a method of estimating the variance of estimators that are constructed with some imputed data. Kim et al. (2006) give an expression for the bias of the multiple imputation variance estimator for data that are collected with a complex survey design. A bias-adjusted variance estimator is also suggested.

2.6 Effect of Survey Design on Inference About Covariance Matrix

In this section we shall look into the effect of sampling design on a classical test statistic for testing a hypothesis regarding a covariance matrix.

Let $\mathbf{V} = ((v_{ij}))$ be a consistent estimator of a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ under the IID assumption. For example, for a self-weighing design, $\mathbf{V}$ may be the usual sample covariance matrix. Let

$$\tilde{\boldsymbol{\omega}} = Vech(\mathbf{V}) = (v_{11}, v_{21}, v_{22}, v_{31}, v_{32}, v_{33}, \ldots, v_{pp})' \tag{2.6.1}$$

be the $u \times 1$ vector of distinct elements of $\mathbf{V}$ where $u = p(p+1)/2$. Suppose we may write $\tilde{\boldsymbol{\omega}}$ as

$$\tilde{\boldsymbol{\omega}} = \frac{1}{n}\sum_{i\in s}\boldsymbol{\omega}_i \tag{2.6.2}$$

where $\boldsymbol{\omega}_i$ is a vector of sample square and cross-product terms, each term centered around the corresponding sample mean, and also possibly weighted, say, for unequal selection probabilities and n is the sample size. Thus, for equal selection probability sampling,

$$\boldsymbol{\omega}_k(k \in s) = ((y_{1k} - \bar{y}_1)^2, \ldots, (y_{pk} - \bar{y}_p)^2, (y_{1k} - \bar{y}_1)(y_{2k} - \bar{y}_2),$$
$$\ldots, (y_{p-1,k} - \bar{y}_{p-1})(y_{p,k} - \bar{y}_p))'$$

where y_{jk} denotes the value of y_j on the unit k in the sample. Then $\tilde{\boldsymbol{\omega}}$ is consistent for $Vech\boldsymbol{\Sigma} = \mu$ (say).

Under IID assumptions, $\tilde{\boldsymbol{\omega}}$ will generally be asymptotically normally distributed with mean μ and the linearization asymptotic covariance matrix estimator of $\tilde{\boldsymbol{\omega}}$, $\text{Var}(\tilde{\boldsymbol{\omega}})$ may be expressed as the $u \times u$ matrix

$$\mathbf{V}^* = \frac{1}{n(n-1)}\sum_{i\in s}(\boldsymbol{\omega}_i - \tilde{\boldsymbol{\omega}})(\boldsymbol{\omega}_i - \tilde{\boldsymbol{\omega}})'. \tag{2.6.3}$$

Consider a linear hypothesis about $\boldsymbol{\Sigma}$ which may be expressed as, say, $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ where $\mathbf{A}$ is a given $q \times u$ matrix of rank q. An IID procedure for testing H_0 is the Wald statistic

$$X_W^2 = (\mathbf{A}\tilde{\boldsymbol{\omega}})'[\mathbf{A}\mathbf{V}^*\mathbf{A}']^{-1}(\mathbf{A}\tilde{\boldsymbol{\omega}}), \tag{2.6.4}$$

which follows the central chi-square distribution, $\chi^2_{(q)}$ in large sample. (Wald statistic has been introduced in Chap. 4).

Under a complex design this procedure may be modified by replacing $\mathbf{V}^*$ by the linearization estimator $\mathbf{V}_L$ of $Var(\tilde{\boldsymbol{\omega}})$ which accommodates the sampling design. This gives a modified Wald statistic X^2_{W0} (Pervaiz 1986). This approach which also assumes near normality of $\tilde{\boldsymbol{\omega}}$ is constrained by the fact that the d.f. ν of $V_L(\tilde{\boldsymbol{\omega}})$ may be low compared to u, particularly if p is moderate to large, in which case $\mathbf{V}_L(\tilde{\boldsymbol{\omega}})$ may become very unstable and even singular. As a result X^2_W may deviate considerably from $\chi^2_{(q)}$. It is therefore suggested to correct X^2_W for its first moment, that is, to refer

$$X^2_{W1} = \frac{(q)X^2_W}{tr[(\mathbf{AV}^*\mathbf{A}')^{-1}\mathbf{AV}_L(\tilde{\boldsymbol{\omega}})\mathbf{A}']} \tag{2.6.5}$$

as $\chi^2_{(q)}$ (Layard 1972).

Note 2.6.1 Graubard and Korn (1993) considered the problem of testing the null hypotheses $H_0 : \theta = \mathbf{0}$, where the p-dimensional parameter $\theta = \mathbf{g}(\boldsymbol{\lambda})$ and $\boldsymbol{\lambda}$ is a r-dimensional vector of means. The authors used replicated estimates of the variances that take into account the complex survey design. The Wald statistic can be used to test H_0, but inference for θ may have very poor power. They used an alternative procedure based on classical quadratic test statistic. Another reference in the area is due to Korn and Graubard (1990).

2.7 Exercises and Complements

2.1 Suppose that y_{abc} is the value of the cth sampled unit belonging to the bth second-stage unit sampled from the ath sampled first-stage unit in a three-stage sample ($a = 1, \ldots, n$; $b = 1, \ldots, m$; $c = 1, \ldots, k$.) Consider the superpopulation model

$$y_{abc} = \theta + \alpha_a + \beta_b + \epsilon_{abc}$$

where α_a, β_b, ϵ_{abc} are independent random variables with mean zero and

$$\text{Cov}\,(y_{abc}, y_{a'b'c'}) = \begin{cases} \sigma_0^2 & \text{if } (a, b, c) = (a', b', c') \\ \tau_2\sigma_0^2 & \text{if } (a, b) = (a' \cdot b'), c \neq c' \\ \tau_1\sigma_0^2 & \text{if } a = a', b \neq b', c \neq c' \\ 0 & \text{if } a \neq a', b \neq b', c \neq c' \end{cases}$$

Here τ_2 is the inter-second-stage-unit correlation, τ_1 is the inter-first-stage-unit correlation. Then show that

$$\begin{aligned} Var_{true}(\bar{y}) &= Var_{true}\left[\sum_{a=1}^{n}\sum_{b=1}^{m}\sum_{c=1}^{k} y_{abc}/nmk\right] \\ &= \tfrac{\sigma_0^2}{nmk}[1 + (m-1)k\tau_1 + (k-1)\tau_2]. \end{aligned}$$

For the IID model

$$y_{abc} = \theta + e_{abc}$$

with e_{abc} independently distributed with zero mean and variance σ_0^2, show that (using the notations of Example 2.2.4)

$$E_{true}[v_{IID}(\bar{y})] \approx \frac{\sigma_0^2}{nmk},$$

if n is large. Hence, deduce that

$$\text{deff}\ (\bar{y}, v_{IID}) \approx 1 + (m-1)k\tau_1 + (k-1)\tau_2.$$

(Skinner 1989)

2.2: Show that the estimator v_α given in (2.5.35) is unbiased for $V(\bar{y}_{st})$.

Springer

http://www.springer.com/978-981-10-0870-2

Complex Surveys
Analysis of Categorical Data
Mukhopadhyay, P.
2016, XV, 248 p., Hardcover
ISBN: 978-981-10-0870-2