CHAPTER 15

Nonsampling Errors

15.1 INTRODUCTION

Complete enumeration and sample surveys are conducted to estimate unknown population parameters. None of these procedures provide exact results. In sample surveys, we observe a part of the population and draw inferences on the basis of the observed data; hence an error is committed as the entire population is not observed. This type of error is known as the sampling error. Clearly, the sampling error is absent in a complete enumeration or census procedure. However, when we collect information from a unit in a sample survey or complete enumeration, the information regarding the value of the characteristic under study is not free from error. When we estimate population parameters from survey data, all the errors are accumulated and incorrect values of the parameters are obtained. The combination of all the kinds of errors, other than the sampling error, for which one cannot obtain the true value of the parameter by conducting a survey is known as the nonsampling error. The nonsampling error is present in both the sample survey and complete enumeration, while the sampling error is present in sample survey only. The sampling error is determined by the magnitude of mean square error, but no general method of determining the nonsampling error is available. The sampling error decreases with increase in sample size, while the nonsampling error increases with the increase in sample size. The nonsampling error arises at any stage of conducting a survey such as planning, data collection, processing the data, and tabulation as well report writing.

15.2 SOURCES OF NONSAMPLING ERRORS

The sources of nonsampling errors are numerous. So, it is very difficult to give an exhaustive list. Some of the sources are listed here. Good details are given by Murthy (1977).

1. **Inappropriate sampling frame**: In conducting a survey, by either sample survey or complete enumeration, one should completely specify the coverage of the survey and prepare a list of all the units of the target

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00015-7

population, which is technically known as the sampling frame. The list should be accurate with no duplication or omission of units. In a large-scale household survey, sometimes, some of the households on the list either do not exist or are not included in the list. Hence errors arise from the inaccuracy of the sampling frame.

2. **Response error/measurement error**: Sometimes because of error in the instrument, investigators report untrue values. In a crop cutting experiment, enumerators often report eye estimations of the crop yield, which is highly subjective and erroneous. This type of error is called measurement error. Response errors constitute wrong values obtained from the respondents. This error can be caused when the wrong method of interview is employed, for example, when the interviewers influence the respondents' answers. Respondents have a tendency to report ages as even or as multiples of five. Some items in the questionnaire may be difficult to answer because of inappropriate definition, failing memory, or even lack of knowledge of the respondents. For example, respondents are often unable to recall accurately the total amount spent on clothing in the last year. Some questions are of sensitive nature, such as income and HIV infection status, where informants may report untrue values. Response errors occur when the design of questionnaire is inappropriate. An overly long questionnaire may result in the respondent becoming tired and answering false or inaccurately to finish the survey faster. Language barrier poses a threat in the situation when the interviewer and the respondent communicate via different languages. Here, the interviewer may enter a value different from the true value. Investigators may not find the right persons for interview. For example, investigators may collect data from housewives or children for household surveys, which may be inappropriate and may cause errors.
3. **Error in data processing**: Errors can also be committed while entering data into the computer, at the scrutiny stage, as well when inappropriate formulae are used. These are all sources of nonsampling errors.
4. **Nonresponse error**: Nonresponse means failure to gather information on all or some of the items in a schedule or questionnaire. They mainly occur for two reasons: noncoverage and nonresponse. Noncoverage occurs in situations where some of the target population has no chance of being selected into the sample; this is mainly due to an incomplete sampling frame. In a questionnaire, we collect information on several items. For example, in household surveys, information on household size and

composition, age, education level, types of dwelling, income, and consumptions of different items is collected, among others. If all the items of information of a questionnaire or schedule are missing, we call it unit nonresponse, and if information on some of the items is obtained and on the rest is missing, then we call this item nonresponse. Nonresponse occurs for various reasons. Unit nonresponse may occur if the informant is not at home, is busy for other business, refuses to answer (termed hard-core), enumerators fail to make prior appointment or to locate the unit, etc. Item response may occur mainly if the respondent does not know the answer, feels the item of information is sensitive (such as status of HIV/AIDS, income), finds it difficult or does not know the answer, or the enumerator forgets to write down the answer correctly.

15.3 CONTROLLING OF NONSAMPLING ERRORS

It is difficult to have 100% control of nonsampling errors; however, we can take various steps to control them. The sampling frame should be accurate. Every unit of the population should be easily identifiable. Definitions of each of the items of investigations such as households and family should be provided unambiguously. Investigators should be provided with proper training, explaining the meaning and possible answers to each question they might ask. There should be proper supervision of enumerators to ensure collection of quality data. The questionnaire should be designed properly. The order of questioning should follow a logical sequence with easy and nonsensitive questions coming first to make the respondent comfortable. Related questions should be grouped together. The length of the questionnaire should be as short as possible. Questionnaires should have a method of consistency check whenever possible. Before conducting a survey, a pilot survey should be undertaken to test the questionnaire, time taken to administer it, cost of the proposed survey, administration of the survey, etc. The analysis of the questionnaire will help in checking the accuracy of the tables and other anticipated results. After conducting the survey, the filled questionnaires should be scrutinized properly. At least 5% of the collected data should be recollected again by different investigators to check the quality of data collected.

15.4 TREATMENT OF NONRESPONSE ERROR

In almost all large-scale surveys, nonresponse is inevitable. Nonresponse creates bias in the estimates. The amount of bias often increases with the

rate of nonresponse. In household income surveys, response from the high income group is considerably lower than the other groups and hence underrepresented in the sample. These types of data therefore produce bias in the estimation of the mean income. Researchers very often take a larger sample size than the required number, taking into account that some of them will not respond [vide Lohr, 1999]. This practice generally produces erroneous results because some of the groups of the units are then underrepresented in the sample. Several methods of handling nonresponse problems in sample surveys are available in the literature. The method of poststratification, response modeling, imputation, and randomized response techniques are popular. Good details are available in Rubin (1987).

15.4.1 Poststratification

15.4.1.1 Hansen–Hurwitz Method

Let a sample s of size n be selected by simple random sampling without replacement (SRSWOR) method. Suppose n_1 units responded and remaining $n_2(=n-n_1)$ did not respond at the first attempt. The set of responded and nonresponded units will be denoted, respectively, by s_1 and s_2. A simple random subsample s_2^* of size $m = vn_2$ (assuming integer) is selected from s_2 where v is a known fraction. Responses from all the units of s_2^* are obtained by using more intensive method, which is obviously expensive. Hansen and Hurwitz (1946) assumed that the population U under study of size N can be divided into two strata according to the nature of respondents. The first stratum H_1 consisting of N_1 (unknown) units that respond at the first attempt and the remaining set of units $N_2(=N-N_1)$ that do not respond at the first attempt comprise the stratum H_2. It is assumed that the members who belong to H_1 always respond at the first attempt while members of H_2 never respond at the first attempt, but they will always respond if a more persuasive method is employed.

Let $\overline{y}(s_1)$ and $\overline{y}(s_2^*)$ be the sample means of the variable under study y for the samples s_1 and s_2^*, respectively. Let

$$\overline{y}_w = w_1\overline{y}(s_1) + w_2\overline{y}(s_2^*) \tag{15.4.1}$$

with $w_1 = n_1/n$ and $w_2 = n_2/n$.

Let $\overline{Y}, S_y^2, S_{1y}^2$, and S_{2y}^2 denote the population mean, population variance, population variances of response, and nonresponse strata, respectively.

Substituting $n' = n,\ \gamma_1 = 1,\ \gamma_2 = v,\ n_2 = m$ in Theorem 10.3.2, we obtain the following result.

Theorem 15.4.1
(i) $\overline{y}_w$ is unbiased for $\overline{Y}$
(ii) The variance of $\overline{y}_w$ is

$$V(\overline{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \frac{W_2}{n}\left(\frac{1}{\nu} - 1\right) S_{2y}^2$$

(iii) An unbiased estimator of $V(\overline{y}_w)$ is

$$\begin{aligned}\widehat{V}(\overline{y}_w) = {} & \frac{(N-n)(n_1-1)}{N(n-1)} w_1 \frac{\widehat{S}_{1y}^2}{n_1} \\ & + \frac{(N-1)(n_2-1) - (n-1)(m-1)}{N(n-1)} w_2 \frac{\widehat{S}_{2y}^2}{m} \\ & + \frac{N-n}{N(n-1)}\left[w_1\{\overline{y}(s_1) - \overline{y}_w\}^2 + w_2\{\overline{y}(s_2^*) - \overline{y}_w\}^2\right]\end{aligned}$$

where $W_2 = N_2/N$, $\widehat{S}_{1y}^2$ and $\widehat{S}_{2y}^2$ denote the sample variance of response and nonresponse stratum, respectively.

Remark 15.4.1
The bias of $\overline{y}(s_1)$ (sample mean of the response stratum) for estimating $\overline{Y}$ is given by

$$\begin{aligned} B[\overline{y}(s_1)] &= E[\overline{y}(s_1)] - \overline{Y} \\ &= \overline{Y}_1 - \overline{Y} \\ &= W_2(\overline{Y}_1 - \overline{Y}_2) \end{aligned} \tag{15.4.2}$$

where $\overline{Y}_1$ and $\overline{Y}_2$ denote the population mean of response and nonresponse stratum, respectively. The bias in Eq. (15.4.2) is negligible if at least one of the quantities W_2 and $(\overline{Y}_1 - \overline{Y}_2)$ is negligible. From Theorem 15.4.1, we note that estimator $\overline{y}_w$ is less efficient than the sample mean based on all the n observed, and the loss of efficiency is negligible if W_2 or S_{2y}^2 is small.

15.4.1.1.1 Optimum Value of ν and *n*
Let us consider the cost function of the form

$$C = c_0 n + c_1 n_1 + c_2 m \tag{15.4.3}$$

where c_0 is the fixed per sampled unit, c_1 is the cost per unit to acquire information from the respondent stratum, and c_2 is the cost per unit of acquiring information from the nonresponse stratum. Here, the cost C is a random variable and the expected cost for a given n under SRSWOR is

$$C^* = E(C) = n(c_0 + W_1 c_1 + W_2 \nu c_2) \tag{15.4.4}$$

The optimum values of n and ν may be obtained either by minimizing the expected cost C^* keeping $V(\bar{y}_w)$ to a certain level V_0 or by minimizing $V(\bar{y}_w)$ keeping C^* to a certain level C_0^*.

For minimization consider

$$\left(V(\bar{y}_w) + \frac{S_y^2}{N}\right) C^* = \left(\frac{S_y^2 - W_2 S_{2y}^2}{n} + \frac{1}{\nu}\frac{W_2}{n} S_{2y}^2\right) \{n(c_0 + W_1 c_1) + nW_2 \nu c_2\} \tag{15.4.5}$$

Using Cauchy's inequality, we find Eq. (15.4.5) attain a minimum if

$$\frac{S_y^2 - W_2 S_{2y}^2}{(c_0 + W_1 c_1)} = \frac{S_{2y}^2}{\nu^2 c_2}$$

$$\text{i.e.,} \quad Opt\ \nu = \nu_0 = \sqrt{\frac{(c_0 + c_1 W_1) S_{2y}^2}{\left(S_y^2 - W_2 S_{2y}^2\right) c_2}} \tag{15.4.6}$$

Because ν_0 is independent of n, the optimum value of n that minimizes the cost C^* for a given variance $V(\bar{y}_w) = V_0$ is obtained from the equation

$$V_0 = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \frac{W_2}{n}\left(\frac{1}{\nu_0} - 1\right) S_{2y}^2$$

$$\text{i.e.,} \quad Opt\ n = n_0 = \frac{S_y^2 + (1/\nu_0 - 1) W_2 S_{2y}^2}{V_0 + S_y^2 \Big/ N} \tag{15.4.7}$$

Similarly, the optimum n that minimizes $V(\bar{y}_w)$ keeping cost $C = C_0^*$ can be formulated as

$$n_0 = \frac{C_0^*}{c_0 + W_1 c_1 + W_2 \nu_0 c_2}.$$

In case $\nu_0 \geq 1$, all the nonrespondents should be interviewed. If $S_{2y}^2 = S_y^2$, $\nu_0 = \sqrt{\dfrac{(c_0 + c_1 W_1)}{c_2 W_1}}$ may exceed 1 unless c_2 is much larger than c_1.

The Hansen and Hurwitz's (1946) technique was generalized further by Dalenius (1955), El-Badry (1956), Kish and Hess (1958), and Srinath (1971), among others. Särndal et al. (1992) poststratified the initial stratum by more than two, using information related to the interviewers' background and skills as well as the characteristics of respondents such as sex and dwelling conditions.

15.4.2 Use of Response Probabilities

The formation of a response sample, comprising the set of units from which all the responses are obtained, is totally unknown. Hence probabilistic models are proposed to describe the unknown response distributions. Politz and Simmons (1949) computed the response probability of a respondent as the proportion of time staying at home. The response probability may be directly related to the study variable and hence to the auxiliary variable, which is highly related to the study variable. In household income surveys, people of high income may respond with low probability and hence may be underrepresented in the sample. So, if tax return is considered as an auxiliary variable, then the response probability of an individual may be inversely proportional to the amount of the tax return. Let us define an indicator variable I_i attached to the ith unit such that

$$I_i = \begin{cases} 1 & \text{if } i\text{th unit responds} \\ 0 & \text{otherwise} \end{cases} \tag{15.4.8}$$

In response modeling, we assume

$$P\{I_i = 1\} = \theta_i \tag{15.4.9}$$

The response probability θ_i may depend on a set of known p auxiliary information $\mathbf{x}_i = (x_{1i},\ldots, x_{ji},\ldots, x_{pi})$ or the value of the study variable y_i. We classify missing observations into following categories.

15.4.2.1 Classification of Response Probabilities

Missing completely at random (MCAR): If the probability θ_i is independent of the study variable y, then the missing data are called MCAR. Here, missingness does not depend on the values of the data set, observed or unobserved. The missing data are just random subset of the data. For example, if someone lost a schedule, then it may be replaced by a schedule taking at random from the set of filled schedules.

Missing at random (MAR): If the response probability θ_i depends on the observed values of y, but not on the missing values of y, then the missing data are called MAR. Here, missingness does not depend on the unobserved values of the data set but does depend on the observed. MAR is also known as ignorable nonresponse where a model can be used to determine the response mechanism. In household surveys, if consumption on food is missing but household size is not missing, then we estimate missing consumption on food by fitting a linear regression of the consumption of food on the household size based on the observed data where both the consumption of food and household sizes are available.

Not missing at random (NMAR): If θ_i depends on the missing variable y_i, then the nonresponse is called NMAR. Here, missingness depends on the observed values of the data set and also the missing value. NMAR is known as nonignorable nonresponse. For example, respondents with high income less likely to report income or respondents infected with HIV/AIDS respond less likely to report their HIV/AIDS status.

15.4.3 Politz and Simmons Method

Politz and Simmons (1949, 1950) provided a method of finding response probabilities of the sampled households selected by simple random sampling with replacement (SRSWR) method from a population U of N households. The interviewer made a call only once to each of the selected households during the interviewing hours. The time of call may be regarded as random within the interviewing hours. If the respondent is available, then information on the variable of interest y is collected along with the information on how many days in the preceeding 5 consecutive days he/she was at home at this particular point of time. If the respondent is not available, no information from this household is collected. Here, it is assumed that if the respondent is available, then the respondent responds with probability 1. The probability of the ith sampled person being at home θ_i is estimated by $\widehat{\theta}_i = \dfrac{t_i + 1}{6}$, where t_i is the number of times the ith respondent reported to be at home in the last 5 consecutive days. The probability t_i is equal to t, given by

$$P(t_i = t) = \binom{5}{t} \theta_i^t (1 - \theta_i)^{5-t}; \quad t = 0, \ldots, 5 \tag{15.4.10}$$

Let $E_h(\cdot|i)$ denote the conditional expectation given that ith respondent was at home.

Then

$$\begin{aligned} E_h\left(\frac{1}{\widehat{\theta}_i}\middle| i\right) &= \sum_{t=0}^{5} \frac{6}{t+1} \binom{5}{t} \theta_i^t (1 - \theta_i)^{5-t} \\ &= \sum_{t=0}^{5} \binom{6}{t+1} \theta_i^t (1 - \theta_i)^{6-(t+1)} \\ &= \frac{1}{\theta_i} \sum_{x=1}^{6} \binom{6}{x} \theta_i^x (1 - \theta_i)^{6-x} \\ &= \frac{1}{\theta_i} \left(1 - q_i^6\right) \end{aligned} \tag{15.4.11}$$

(where $q_i = 1 - \theta_i$)

and

$$E_h\left(\frac{1}{\widehat{\theta}_i^2}\middle| i\right) = \sum_{t=0}^{5}\left(\frac{6}{t+1}\right)^2\binom{5}{t}\theta_i^t(1-\theta_i)^{5-t}$$

$$= \sum_{t=0}^{5}\frac{6}{t+1}\binom{6}{t+1}\theta_i^t(1-\theta_i)^{6-(t+1)} \tag{15.4.12}$$

$$= \frac{6}{\theta_i}\sum_{x=1}^{6}\frac{1}{x}\binom{6}{x}\theta_i^x(1-\theta_i)^{6-x}$$

$$= \frac{6}{\theta_i}\delta_i$$

with $\delta_i = \sum_{x=1}^{6}\frac{1}{x}\binom{6}{x}\theta_i^x(1-\theta_i)^{6-x}$.

Politz and Simmons proposed the following biased estimator of the population mean $\overline{Y}$ based on an SRSWR sample of size n:

$$\widehat{\overline{Y}}_{ps} = \frac{1}{n}\sum_{j=1}^{n}\frac{y_{(j)}}{\widehat{\theta}_{(j)}}I(j) \tag{15.4.13}$$

where $I_{(j)}$ is an indicator variable taking the value 1 if the unit selected at the jth draw was at home with probability $\theta_{(j)}$ and 0 otherwise, i.e., if the jth draw produces the ith unit, we get $y_{(j)} = y_i$ and $\widehat{\theta}_{(j)} = \widehat{\theta}_i$.

Theorem 15.4.2

(i) $E\left(\widehat{\overline{Y}}_{ps}\right) = \overline{Y} - \frac{1}{N}\sum_{i\in U} y_i q_i^6$

(ii) $V\left(\widehat{\overline{Y}}_{ps}\right) = \frac{\sigma_{ps}^2}{n}$

(iii) $\widehat{V}\left(\widehat{\overline{Y}}_{ps}\right) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{\hat{\theta}_i} - \widehat{\overline{Y}}_{ps}\right)^2$

where $q_i = 1-\theta_i$, $\sigma_{ps}^2 = \frac{6}{N}\sum_{i\in U}\delta_i y_i^2 - \left(\frac{1}{N}\sum_{i\in U} y_i\left(1-q_i^6\right)\right)^2$, and $\delta_i = \sum_{x=1}^{6}\frac{1}{x}\binom{6}{x}\theta_i^x(1-\theta_i)^{6-x}$.

Proof

Let $E_h\{\cdot|(j)\}$, E_r, and E_p denote the conditional expectation given that unit selected at the jth draw was at home, unconditional expectation with respect to the person selected at the jth draw was at home and sampling design. Then noting $E_r(I_{(j)}) = \theta_{(j)}$, we get

$$\begin{aligned} E\left(\frac{\gamma_{(j)}}{\widehat{\theta}_{(j)}} I_{(j)}\right) &= E_p\left[\gamma_{(j)}\left\{E_h \frac{1}{\widehat{\theta}_{(j)}}\middle|(j)\right\} E_r\{I_{(j)}\}\right] \\ &= E_p\left[\gamma_{(j)}\theta_{(j)} E_h\left\{\frac{1}{\widehat{\theta}_{(j)}}\right\}\right] \\ &= E_p\left(\gamma_{(j)}\left(1 - q_{(j)}^6\right)\right) \\ &= \frac{1}{N}\sum_{i\in U}\gamma_i\left(1 - q_i^6\right) \\ &= \overline{Y} - \frac{1}{N}\sum_{i\in U}\gamma_i q_i^6 \end{aligned} \tag{15.4.14}$$

$$\begin{aligned} E\left(\frac{\gamma_{(j)}}{\widehat{\theta}_{(j)}} I_{(j)}\right)^2 &= E_p E_h\left(\left(\frac{\gamma_{(j)}}{\widehat{\theta}_{(j)}}\right)^2 \middle|(j)\right) E_r\left(I_{(j)}\right) \\ &= 6E_p\left[\gamma_{(j)}^2 \delta_{(j)}\right] \\ &= \frac{6}{N}\sum_{i\in U}\gamma_i^2\delta_i \end{aligned} \tag{15.4.15}$$

and

$$\begin{aligned} V\left(\frac{\gamma_{(j)}}{\widehat{\theta}_{(j)}} I_{(j)}\right) &= \frac{6}{N}\sum_{j\in U}\gamma_i^2\delta_i - \left(\overline{Y} - \frac{1}{N}\sum_{i\in U}\gamma_i q_i^6\right)^2 \\ &= \sigma_{ps}^2 \end{aligned} \tag{15.4.16}$$

Finally, noting $\frac{\gamma_{(j)}}{\widehat{\theta}_{(j)}} I_{(j)}$'s for $j = 1,\ldots, n$ are independently and identically distributed, we verify the theorem.

Remark 15.4.2

The bias of $\widehat{\overline{Y}}_{ps}$ is $-\frac{1}{N}\sum_{i\in U}\gamma_i q_i^6$ and it is certainly negligible if the response probabilities θ_i's are high for every i.

15.4.4 Imputation

Imputation is used for item nonresponse. Here we assign one or more values to a missing item to reduce the bias and control variance due to nonresponse. If we assign a single value, it is called "single imputation." In "multiple imputations" we assign more than one value to a missing item. The data with the imputed values (single or multiple) are analyzed using a standard statistical package assuming imputed values are of true value. Single imputation analysis is very simple because we get a complete data set and assume no observation is missing. Conversely, in multiple imputations we generate more than one data set because more than one value is obtained from the missing units. From each of the data sets, a separate estimate of the parameter of interest is obtained. Finally, a single estimate and its standard error are obtained by combining all these separate estimates. The process of estimation based on multiple imputations is certainly more complex than single imputation. Statistical agencies generally find difficulty in using multiple imputations because of operational challenges in maintaining multiple data sets, especially in large-scale surveys. Several methods of imputations are available in the literature. A few of the popular methods are described as follows.

Deductive imputation: Here, the missing value is imputed through a consistency check—establishing relationship with other available items. For example, suppose the HIV status is missing but information is available that nobody in the household is suffering from HIV infection. So, the missing item of the status of HIV infection is imputed as negative. Deductive imputations can be seen as almost accurate.

Substitution: In this case, if the information of a unit is not available, then it is replaced by a unit nearest to it. In a household survey, if no information is available from the sampled household, then it is substituted by the next door neighbor. Sometimes a unit is substituted randomly from those units that are not selected in the original sample.

Cold deck imputation: The missing data are replaced using records from a recent past survey. Naturally, cold deck imputation cannot provide accurate information, as it is not based on the current survey.

Hot deck imputation: The sampled units are divided into classes using prior information. For example, in household surveys, structure and location of the house may be used to classify the household into the higher, middle, or lower income groups. In hot deck imputation a particular unit in the class is chosen whose entries are complete. Such a responding unit is

called the "donor." The information of the donor is substituted for the missing items. If the donor is selected at random, the imputation is called "random hot deck imputation." On the other hand, if the selected donor closest to the recipient (the unit with the missing item), is selected on the basis of a suitable distance function, the imputation is known as "nearest hot deck imputation." In particular, if the information on consumption of food in a middle-class household for an adult Asian male is missing, then the donor should match with the economy level (middle class) age (adult), and race (Asian) as far as possible.

Mean imputation: The mean of the study variable of all possible donors are substituted for each of the missing items.

Ratio and regression imputations: Here, we assume that the auxiliary information is available for all the sampled units (both respondent and nonrespondent units). The value of the study variable y is available for the respondent units but missing for the nonrespondents. In this situation a suitable regression of the study variable on the auxiliary variables is obtained by using the study and auxiliary variables of the respondent units. In regression imputation, the missing observations are predicted using a fitted regression. In case both the study and auxiliary variables are quantitative, a multiple regression may be fitted. Logistic regression may be used if the study variable is binary or qualitative. For a single auxiliary information, the regression imputation reduces to "ratio imputation" if the regression line passes through the origin. The regression imputation may be classified into two categories: deterministic and random imputation. In deterministic regression imputation, the imputed value of y_i is obtained by fitting a regression model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \sigma \nu_i^{1/2} \in_i \tag{15.4.17}$$

where $f(\cdot, \cdot)$ is a given function, $\mathbf{x_i} = (x_{1i}, \ldots, x_{qi})'$ is a q vector auxiliary variable whose values are available for the response sample s_r and the nonresponse sample $s_m (= s - s_r)$, $\boldsymbol{\beta}$ is a q vector regression coefficient, σ is an unknown parameter, ν_i is a known constant, and $\in_i$ is an independent identically distributed random variable with mean zero and variance unity. In deterministic imputation the missing observation y_i is imputed as

$$y_i^* = f(x_i, \widehat{\boldsymbol{\beta}}); \quad i \in s_m \tag{15.4.18}$$

where $f(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$ is the fitted regression obtained from the response sample s_r, $\widehat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. So, in deterministic imputation, the imputation value will be unaffected if the sample s_r is unchanged.

Ratio imputation is a special case of deterministic imputation that uses a single auxiliary variable x_i with $f(\mathbf{x}_i, \boldsymbol{\beta}) = \beta x_i$ and $v_i = 1$. In this case $\widehat{\beta} = \dfrac{\sum_{i \in s_r} y_i}{\sum_{i \in s_r} x_i}$ and the imputed value of y_i is

$$y_i^* = \frac{\sum_{i \in s_r} y_i}{\sum_{i \in s_r} x_i} x_i \quad \text{for } i \in s_m \tag{15.4.19}$$

The deterministic imputation reduces to the mean imputation if $f(\mathbf{x}, \boldsymbol{\beta}) = \beta$ and $x_i = 1$. In this case

$$y_i^* = \sum_{i \in s_r} y_i / r = \overline{y}_r \quad \text{for } i \in s_m \tag{15.4.20}$$

The random imputation is the modification of the deterministic imputation where a random noise $\widehat{\in}_i$ is added to the imputed y_i^* and it is given by

$$\widehat{y}_i = f(x_i, \widehat{\boldsymbol{\beta}}) + \widehat{\sigma} v_i^{1/2} \widehat{\in}_i; \quad i \in s_m \tag{15.4.21}$$

where $\widehat{\sigma}$ is an estimate of σ obtained from the model (Eq. 15.4.17) and $\widehat{\in}_i$ is a random sample either from the distribution of $\in_i$ or from the standardized residuals obtained from the fitted regression (Eq. 15.4.17).

15.4.4.1 Problems of Imputation

The main demerit of imputation is that it underestimates variance if the response rate is not high enough. To illustrate this point, the following example of Chen et al. (2000) is quite useful.

Let a sample s of size n be selected from a population of size N by SRSWOR method and let s_r be the response sample of size r and the nonresponse sample $s - s_r$. Using mean imputation the missing value of jth unit y_j is imputed as

$$\overline{y}_r = \sum_{i \in s_r} y_i / r \quad \text{for } j \in s_m \tag{15.4.22}$$

The imputed estimator of the population mean and the imputed sample variance are given, respectively, by

$$\overline{y}_I = \frac{1}{n} \left\{ \sum_{i \in s_r} y_i + (n - r) \overline{y}_r \right\} \tag{15.4.23}$$

and

$$s_{yI}^2 = \frac{1}{n-1}\sum_{i\in s}\left(y_{i0}-\overline{y}_I\right)^2 = \frac{(r-1)}{(n-1)}s_{yr}^2 \tag{15.4.24}$$

where $s_{yr}^2 = \frac{1}{r-1}\sum_{i\in s_r}\left(y_{i0}-\overline{y}_r\right)^2$ and $y_{i0} = \begin{cases} y_i & \text{for } i \in s_r \\ \overline{y}_r = \sum_{i\in s_r} y_i/r & \text{for } i \in s_m \end{cases}$.

Under uniform response

$$E\left(s_{yI}^2 \middle| r\right) = \frac{r-1}{n-1}S_y^2 \cong \frac{r}{n}S_y^2 \tag{15.4.25}$$

where S_y^2 = population variance.

If r is not large, Eq. (15.4.25) is much less than S_y^2. Thus the sample variance underestimates the population variance under mean imputation in the case that the response rate is not high. To overcome the underestimation of variance, Rao and Shao (1992), Shao and Sitter (1996), and Chen et al. (2000), among others, proposed alternative methods of imputation where the variance estimators do not underestimate the variance.

Example 15.4.1

The exam marks and CA (continuous assessment) for a sample of 7(=n) students of a class of 50(=N) students that are selected at random are given in the following table.

Student	CA (x)	Exam marks (y)
1	50	60
2	40	30
3	70	65
4	60	70
5	80	90
6	55	?
7	90	?

The exam marks of the sixth and seventh students are missing. Let the average marks of CA for the entire class is known as $\overline{X} = 70$.

Mean Imputation: Here, the sample mean of y based on the response sample is $\overline{y}_r = 63$. So the imputed values of y_6 and y_7 are $y_6^* = 63$ and

$y_7^* = 63$. Hence the estimated population mean of the exam marks is $\widehat{\overline{Y}} = \overline{y}_r = 63$. The estimated standard error is

$$\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) s_{yI}^2} = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in s} \left(y_{i0} - \overline{y}_I\right)^2}$$

$$= \sqrt{\left(\frac{1}{7} - \frac{1}{50}\right) \times 313.33} = 6.20$$

Ratio imputation: Here, $y_i^* = \dfrac{\sum_{i \in s_r} y_i}{\sum_{i \in s_r} x_i} x_i = 1.05 x_i$.

So, the imputed values $y_6^* = 1.05 \times 55 = 57.75$ and $y_7^* = 1.05 \times 90 = 94.5$.

Regression imputation: Estimated linear regression based on the response sample s_r is $y = -12 + 1.25x$. So, the imputed values are $y_6^* = -12 + 1.25 \times 55 = 56.75$ and $y_7^* = -12 + 1.25 \times 90 = 100.5$.

Random imputation: Here, we impute the values of y_6 and y_7 from the response sample s_r by SRSWR method. Let the SRSW sample be (5, 3). Then the imputed values of y_6 and y_7 are, respectively, $y_6^* = y_5 = 90$ and $y_7^* = y_3 = 65$.

Nearest hot deck imputation: Here, $x_1 = 50$ and $x_4 = 60$ are nearest to $x_6 = 55$. So, we may impute y_6 either by y_1 or by y_4, i.e., $y_6^* = y_1 = 60$ or $y_6^* = y_4 = 70$. Similarly, $y_7^* = y_5 = 90$.

Random Regression imputation:

Method 1: The residuals are $\widehat{\epsilon}_1 = 9.5$, $\widehat{\epsilon}_2 = -8$, $\widehat{\epsilon}_3 = -10.5$, $\widehat{\epsilon}_4 = 7.0$, and $\widehat{\epsilon}_5 = 2.0$. We choose two residuals at random with replacement. Suppose the selected residuals are $\widehat{\epsilon}_2$ and $\widehat{\epsilon}_1$. The imputed values of y_6 and y_7 are $y_6^* = 56.75 + \widehat{\epsilon}_2 = 48.75$ and $y_7^* = 100.5 + \widehat{\epsilon}_1 = 110.0$.

Method 2: The estimated value of σ^2 is $\widehat{\sigma}^2 = \sum_{i \in s_r} \widehat{\epsilon}_i^2 / 3 = 105.83$. We select two random samples from a uniform distribution (0, 1). Let these be 0.47 and 0.76. The normal deviates corresponding to these sample are $z_1 = -0.08$ and $z_2 = 0.71$. The imputed values $y_6^* = 56.75 + 10.288 \times (-0.08) = 55.93$ and $y_7^* = 100.5 + 10.288 \times 0.71 = 107.80$.

15.4.5 Multiple Imputation

In single imputation, the imputed value is treated as the true value, ignoring the fact that the no imputation method can provide the exact value. Single imputation does not reflect the uncertainty about the prediction of the missing values. Multiple imputation was proposed by Rubin (1987). In this method D

imputed values for each of the missing observation is generated and hence we get D complete data set. From each of the complete data set an estimate of the parameter of interest θ is obtained by using a standard technique, assuming no nonresponse is present. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values.

Let $\widehat{\theta}_d$ and $\widehat{\phi}_d$ be an estimator of θ and the estimator of the variance of $\widehat{\theta}_d$ based on the dth complete data set. The combined estimator of θ under multiple imputations is given by

$$\widehat{\theta}_D = \frac{1}{D}\sum_{d=1}^{D}\widehat{\theta}_d \tag{15.4.26}$$

The proposed variance estimator of $\widehat{\theta}_D$ is given by

$$\widehat{\Phi}_D = \overline{\phi}_D + \left(1+\frac{1}{D}\right)B_D \tag{15.4.27}$$

where $\overline{\phi}_D = \frac{1}{D}\sum_{d=1}^{D}\widehat{\phi}_d$ and $B_D = \frac{1}{D-1}\sum_{d=1}^{D}\left(\widehat{\theta}_d - \overline{\theta}_D\right)^2$. The first component $\overline{\phi}_D$ of Eq. (15.4.27) is the average within-imputation variance and the second component is the product of between-variance component B_D and correction factor $\left(1+\frac{1}{D}\right)$.

For large sample size the statistic

$$\frac{\widehat{\theta}-\theta}{\sqrt{\widehat{\Phi}_D}} \tag{15.4.28}$$

follows t distribution with ν degrees of freedom where

$$\nu = (D-1)\left(1+\frac{1}{D+1}\frac{\overline{\phi}_D}{B_D}\right)^2 \tag{15.4.29}$$

For smaller sample size an improved variance formula has been proposed by Barnard and Rubin (1999).

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability. Multiple imputation facilitates simple formula for variance estimation and interval estimation of the parameter of interest. Its only disadvantage is that it is more tedious to analyze the data than the single imputation. But modern computing techniques such as computation do not pose any problem.

15.4.6 Bayesian Imputation

Without loss of generality, let us assume that the response sample s_r comprises the first r units and let $\mathbf{Y}_r = (y_1,\ldots, y_r)'$ and $\mathbf{X}_r = (x_1,\ldots, x_r)'$.

The nonresponse sample $s_m = s - s_r$ comprises the next $n - r$ units and $\mathbf{Y}_{n-r} = (y_{r+1},\ldots,\ y_n)'$ and $\mathbf{x}_{n-r} = (x_{r+1},\ldots,\ x_n)'$. The Bayesian single or multiple imputations are obtained from the conditional distribution of $\mathbf{Y}_{n-r}$ given the sample data. Here, we will consider the methods of imputations proposed by Wang et al. (1992) and Schenker and Welsh (1988).

15.4.6.1 Wang et al. Method

In this method a linear regression between the study y and the auxiliary variable x is assumed as follows:

$$y_i = \beta_0 + \beta_1 x_i + \in_i, \quad i = 1, \ldots, n \tag{15.4.30}$$

where $\boldsymbol{\beta} = (\beta_0,\ \beta_1)'$ is the unknown model parameter, x_i's are known values of an auxiliary variable x, and $\in_i$'s are independent normal variable with mean zero and variance σ^2. Here, we assume that the response mechanism cannot be regarded as MAR and the response probability of the ith unit is

$$g(R|y_i) = 1 - e^{-\alpha y_i} \tag{15.4.31}$$

with α as a known constant.

The imputations are obtained from the conditional distribution of $\mathbf{Y}$ corresponding to the nonresponding units given the sample data. For simplicity, Wang et al. (1992) assumed that σ^2 is known and the prior distribution of $\boldsymbol{\beta}$ is noninformative. This implies that f_2, the posterior distribution of $\boldsymbol{\beta}$ given data is proportional to the likelihood of $\boldsymbol{\beta}$.

The imputations are selections, $y_{r+1},\ldots,\ y_n$ from the conditional distribution

$$\begin{aligned} &f(y_{r+1}, \ldots, y_n | y_1, \ldots, y_r; x_1, \ldots, x_n; z_1, \ldots, z_n; \sigma^2) \\ &= \int \cdots \int f_1(y_{r+1}, \ldots, y_n | \boldsymbol{\beta}, y_1, \ldots, y_r; x_1, \ldots, x_n; z_1, \ldots, z_n; \sigma^2) \\ &\quad f_2(\boldsymbol{\beta} | y_1, \ldots, y_r; x_1, \ldots, x_n; z_1, \ldots, z_n; \sigma^2) \partial \boldsymbol{\beta} \end{aligned} \tag{15.4.32}$$

where $z_1 = z_2 = \cdots = z_r = 1$ and $z_{r+1} = z_{r+2} = \cdots = z_n = 0$ are indicators for respondents and nonrespondents [vide Wang et al., 1992 and Govindrajulu, 1999]. Rubin (1987) has shown that the distribution of $f_2(\boldsymbol{\beta}|\cdot)$ given in Eq. (15.4.32) is proportional to

$$K\left[\exp\left\{-\alpha\left(\sum_{i=r+1}^{n}(\beta_0 + \beta_1 x_i)\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{r}(y_i - \beta_0 - \beta_1 x_i)^2\right)\right\}\right] \tag{15.4.33}$$

where

$$K = \prod_{i=r+1}^{n}\left[1 - \Phi\left\{-\frac{\beta_0 + \beta_1 x_i}{\sigma} + \alpha\sigma\right\}\right] \times \left[1 - \Phi\left\{-\frac{\beta_0 + \beta_1 x_i}{\sigma}\right\}\right]$$

and Φ is the cumulative distribution function of standard normal variable. For sufficiently large $\frac{\beta_0 + \beta_1 x_i}{\sigma}$, the value of K is approximately equal to 1, and in this case $f_2(\boldsymbol{\beta}|y_1,\ldots, y_r; x_1,\ldots, x_n; z_1,\ldots, z_n; \sigma^2)$ can be approximated as a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$ (vide Rubin, 1987) where

$\boldsymbol{\mu}' = -\left(\mathbf{c} - 2\widehat{\boldsymbol{\beta}}_r'\mathbf{A}\right)\mathbf{A}^{-1}/2, \mathbf{c} = 2\alpha\mathbf{1}'\mathbf{X}_r^*$, $\mathbf{1}$ is a column vector of $n-r$ elements each of which is 1,

$$(\mathbf{X}_r^*)' = \begin{pmatrix} 1 & \ldots & 1 \\ x_{r+1} & \ldots & x_n \end{pmatrix}, \quad \widehat{\boldsymbol{\beta}}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{Y}_r, \quad \mathbf{X}_r' = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_r \end{pmatrix},$$

$\boldsymbol{\Sigma} = \mathbf{A}^{-1}$, and $\mathbf{A} = (\mathbf{X}_r'\mathbf{X}_r)/\sigma^2$.

To find the imputed values of $y_{r+j}(1)$ for $j = 1,\ldots, n-r$, we follow the following steps:

Step 1: Draw a sample $\widehat{\boldsymbol{\beta}}^* = \left(\widehat{\beta}_0^*, \widehat{\beta}_1^*\right)'$ from the distribution f_2, which is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Step 2: Choose a sample $\widehat{y}_{r+j}(1)$ from the population $y_{r+j}\left|x_{r+j}, \widehat{\boldsymbol{\beta}}^*, \sigma^2\right.$, which is $N\left(\widehat{\beta}_0^* + \widehat{\beta}_1^* x_{r+j}, \sigma^2\right)$.

Step 3: Select a sample $\widehat{u}_{r+j}(1)$ from a uniform distribution $U(0, 1)$.

Step 4: Treat $\widehat{y}_{r+j}(1)$ as an imputed value of y_{r+j} if $\widehat{u}_{r+j}(1) \le e^{-\alpha\hat{y}_{r+j}(1)}$. Otherwise, i.e., if $\widehat{u}_{r+j}(1) > e^{-\alpha\hat{y}_{r+j}(1)}$, repeat Steps 1–4 until an imputed value of y_{r+j} is obtained.

Using the aforementioned method for $j = r,\ldots, n-r$, the first set of imputed values $\widehat{y}_{r+j}(1), \ldots, \widehat{y}_n(1)$ of $y_{r+j},\ldots, y_n$ are obtained.

To find the second set of imputed values $\widehat{y}_{r+j}(2), \ldots, \widehat{y}_n(2)$, go to

Step 5: Using $\widehat{y}_{r+1}(1), \ldots, \widehat{y}_n(1)$ and the observed data $y_1,\ldots, y_r$, estimate β_0, β_1, and σ^2 using the following formula

$$\widehat{\beta}_0(2) = \overline{y}_n(2) - \widehat{\beta}_1(2)\overline{x}_n, \widehat{\beta}_1(2) = \frac{\sum_{i=1}^{n} y_i^+(2)(x_i - \overline{x}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} \quad \text{and}$$

$$\widehat{\sigma}^2(2) = \frac{\sum_{i=1}^{n}\left\{y_i^+(2) - \widehat{y}_i^+(2)\right\}^2}{(n-2)}$$

where $y_i^+(2) = \begin{cases} y_i & \text{for } i = 1, \ldots, r \\ \widehat{y}_i(1) & \text{for } i = r+1, \ldots, n \end{cases}$, $\overline{y}_n(2) = \frac{\sum_{i=1}^{n} y_i^+(2)}{n}$,

$\overline{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$, and $\widehat{y}_i^+(2) = \widehat{\beta}_0(2) + \widehat{\beta}_1(2)x_i$.

Step 6: Choose a sample $\widehat{y}_{r+j}(2)$ from $N\left(\widehat{\beta}_0(2) + \widehat{\beta}_1(2)x_{r+j}, \widehat{\sigma}^2(2)\right)$.

Step 7: Select a sample $\widehat{u}_{r+j}(2)$ from a uniform distribution $U(0, 1)$.

Step 8: Treat $\widehat{y}_{r+j}(2)$ as an imputed value of y_{r+j} if $\widehat{u}_{r+j}(2) \leq e^{-\alpha\hat{y}_{r+j}(2)}$. Otherwise, i.e., if $\widehat{u}_{r+j}(2) > e^{-\alpha\hat{y}_{r+j}(2)}$, repeat Steps 5–8 until an imputed value of y_{r+j} is obtained. Using the aforementioned method the second set of imputed values $\widehat{y}_{r+j}(2), \ldots, \widehat{y}_n(2)$ of $y_{r+j}, \ldots, y_n$ are obtained.

Repeating Steps 5–8, $D-2$ times, the remaining $D-2$ sets of imputed values are obtained using $\widehat{y}_{r+1}(d), \ldots, \widehat{y}_n(d)$ and the observed data $y_1, \ldots, y_r$ to compute $\widehat{y}_{r+1}(d+1), \ldots, \widehat{y}_n(d+1)$.

15.4.6.2 Schenker and Welsh Method

Schenker and Welsh (1988) and Kim (2004) used the following multiple regression model:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \end{aligned} \tag{15.4.34}$$

where the $p+1$-dimension vector $\mathbf{x}_i' = (1, x_{1i}, \ldots, x_{pi})$, $i = 1, \ldots, n$, is known; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is unknown, $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ and σ^2 are unknown.

Let us assume that $\boldsymbol{\beta}$ and σ^2 are independent with constant prior for $(\boldsymbol{\beta}, \log\sigma^2)$, so that the marginal posterior distribution of σ^2 is $\widehat{\sigma}_r^2(r-p-1)\Big/\chi_{r-p-1}^2$ and the conditional posterior distribution of $\boldsymbol{\beta}$ given σ^2 is $N\left(\widehat{\boldsymbol{\beta}}, \sigma^2\psi_r^{-1}\right)$, where $\widehat{\sigma}_r^2 = \mathbf{Y}_r'\left[\mathbf{I} - \mathbf{X}_r\left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1}\mathbf{X}_r'\right]\mathbf{Y}_r/(r-p-1)$, $\widehat{\boldsymbol{\beta}}_r = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{Y}_r$, χ_{r-p-1}^2 is a Chi-square distribution with $r-p-1$ degrees of freedom, and $\psi_r = (\mathbf{X}_r'\mathbf{X}_r)$.

For multiple imputation of $\mathbf{Y}_{n-r}' = (y_{r+1}, \ldots, y_n)$, the following steps are followed:

Step 1: For the kth repetition, $k(=1, \ldots, D)$, draw a sample of σ^2 as

$$\sigma_{(k)}^2 = \widehat{\sigma}_r^2(r-p-1)/g_k$$

where g_k is a random sample from χ_{r-p-1}^2.

Step 2: Draw a sample $\boldsymbol{\beta}_{(k)}$ from $N\left(\widehat{\boldsymbol{\beta}}_r, (\mathbf{X}_r'\mathbf{X}_r)^{-1}\sigma_{(k)}^2\right)$.

Step 3: For each missing unit $j(=r+1, \ldots, n)$, select a random sample $\epsilon_{j(k)}$ from $N\left(0, \sigma_{(k)}^2\right)$. Then the imputed value of jth unit at kth imputation is

$$\widehat{y}_{j(k)} = \mathbf{x}_j'\boldsymbol{\beta}_{(k)} + \epsilon_{j(k)}$$

15.4.7 Subsampling Method

Let a sample s, of size n, be selected from the population $U = (1,\ldots, N)$ with probability $p(s)$ according to a sampling design p with π_i and π_{ij} as inclusion probabilities for the ith, and ith and jth $(i \neq j)$ units of U. Let $s_r(\subset s)$ be the set of respondent units of size $r(\leq n)$ for which responses y_i's are obtained and the complement $s_m = s - s_r$ (of size $n - r$) be the set of nonresponse units. The formation of the response sample s_r is totally unknown and no exact modeling of s_r is possible. Here, we therefore postulate a probabilistic model that describes the response distribution. Let θ_i be the probability that ith unit $(i = 1,\ldots, N)$ responds if it is selected in the sample s. Hence the response sample s_r is regarded as a subsample of the original sample s according to some unknown sampling design.

15.4.7.1 Arnab and Singh Method

In this method, the inclusion probability of the ith unit in the response sample s_r given s is $\pi_{i|s} = \theta_i$ (response probability of the ith unit). Similarly, the inclusion probability of the ith and jth units $(i \neq j)$ in the sample s_r is $\pi_{ij|s} = \theta_{ij} = \theta_i\theta_j$, assuming responses are independent. The total number of the respondent r, in the sample s_r, is a random variable. Hence, we consider s_r as a subsample from the original sample s selected according to the Poisson sampling design (vide Section 5.4.2.4).

In case θ_i's are known for $i \in s_r$, following Arnab and Singh (2006), we propose a Horvitz–Thompson type estimator for the total Y as follows:

$$\widehat{Y}_{ht}(nr) = \sum_{i\in s_r} \frac{y_i}{\pi_i\theta_i} = \sum_{i\in s} \frac{z_i}{\theta_i} I_{s_r i} \tag{15.4.35}$$

where $z_i = \dfrac{y_i}{\pi_i}$, $I_{s_r i} = 1$ if $i \in s_r$, and $I_{s_r i} = 0$ if $i \in s_m$.

Here, we note that

$E(I_{s_r i}|s) =$ The probability that the selected ith unit will *respond* $= \theta_i$

$E(I_{s_r i} I_{s_r j}|s) =$ The probability that the selected ith unit and jth unit $(i \neq j)$ will respond

$= \theta_i\theta_j.$

The expectations, variance, and unbiased estimators of the variance of $\widehat{Y}_{ht}(nr)$ are given in the following theorem.

Theorem 15.4.3

(i) $E\{\widehat{Y}_{ht}(nr)\} = Y$

(ii) $V\{\widehat{Y}_{ht}(nr)\} = \sum_i \dfrac{y_i^2}{\pi_i}\left(\dfrac{1}{\theta_i} - 1\right) + \dfrac{1}{2}\sum_{i\neq}\sum_j (\pi_i\pi_j - \pi_{ij})\left(\dfrac{y_i}{\pi_i} - \dfrac{y_j}{\pi_j}\right)^2$

(iii) Two unbiased estimators of $V\{\widehat{Y}_{ht}(nr)\}$ are given by

$$\widehat{V}(1) = \widehat{V}_s(1) + \sum_{i\in s_r} \frac{\gamma_i^2}{\pi_i^2\theta_i}\left(\frac{1}{\theta_i} - 1\right)$$

and

$$\widehat{V}(2) = \widehat{V}_s(2) + \sum_{i\in s_r} \frac{\gamma_i^2}{\pi_i^2\theta_i}\left(\frac{1}{\theta_i} - 1\right)$$

where

$$\widehat{V}_s(1) = \frac{1}{2}\sum_{i\neq j}\sum_{\in s_r} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}\theta_i\theta_j}\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right)^2 \text{ and}$$

$$\widehat{V}_s(2) = \sum_{i\in s_r} \frac{\gamma_i^2}{\pi_i\theta_i}\left(\frac{1}{\pi_i} - 1\right) + \sum_{i\neq j}\sum_{\in s_r}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\frac{\gamma_i}{\pi_{ij}\theta_i}\frac{\gamma_j}{\theta_j}$$

Proof

Let $E_p(E_r)$ and $V_p(V_r)$ denote expectations and variances with respect to the sampling design (response mechanism), respectively. Then we have

(i) $$E\{\widehat{Y}_{ht}(nr)\} = E_p\left\{\sum_{i\in s}\frac{z_i}{\theta_i}E_r(I_{s_r i})\right\} = E_p\left(\sum_{i\in s} z_i\right) = Y$$

(ii) $$V\{\widehat{Y}_{ht}(nr)\} = E_p\left\{\sum_{i\in s}\frac{z_i^2}{\theta_i^2}V_r(I_{s_r i})\right\} + V_p\left(\sum_{i\in s}\frac{z_i}{\theta_i}E_r(I_{s_r i})\right)$$

$$= E_p\left\{\sum_{i\in s}\frac{z_i^2}{\theta_i}(1-\theta_i)\right\} + V_p\left(\sum_{i\in s} z_i\right)$$

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

(iii) Writing

$$V_p\left(\sum_{i\in s} z_i\right) = V_p\left(\sum_{i\in s}\frac{\gamma_i}{\pi_i}\right) = \frac{1}{2}\sum_{i\neq j}\sum_{\in U}(\pi_i\pi_j - \pi_{ij})\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right)^2$$

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

we get two unbiased estimators of $V_p\left(\sum_{i\in s} z_i\right)$ as follows:

$$\widehat{V}_s(1) = \frac{1}{2}\sum_{i\neq j}\sum_{\in s_r}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}\theta_i\theta_j}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{15.4.36}$$

and

$$\widehat{V}_s(2) = \sum_{i\in s_r}\frac{y_i^2}{\pi_i\theta_i}\left(\frac{1}{\pi_i} - 1\right) + \sum_{i\neq j}\sum_{\in s_r}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\frac{y_i}{\pi_{ij}\theta_i}\frac{y_j}{\theta_j} \tag{15.4.37}$$

Further, an unbiased estimator of $\sum_{i\in U}\frac{y_i^2}{\pi_i}\left(\frac{1}{\theta_i} - 1\right)$ is $\sum_{i\in s_r}\frac{y_i^2}{\pi_i^2\theta_i}\left(\frac{1}{\theta_i} - 1\right)$ (15.4.38)

From Eqs. (15.4.36)−(15.4.38), we can verify the result.

15.4.7.1.1 Simple Random Sampling Without Replacement

For an SRSWOR sampling design $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/\{N(N-1)\}$, hence we derive.

Theorem 15.4.4

(i) $\widehat{Y}_{srs}(nr) = \frac{N}{n}\sum_{i\in s_r}\frac{y_i}{\theta_i}$ is an unbiased estimator of Y

(ii) $V\left[\widehat{Y}_{srs}(nr)\right] = \frac{N}{n}\sum_i y_i^2\left(\frac{1}{\theta_i} - 1\right) + N\frac{N-n}{n}S_y^2$

(iii) Two unbiased estimators of $V\left[\widehat{Y}_{nr}(srs)\right]$ are given by

$$\widehat{V}_{srs}(1) = \widehat{V}^*(1) + \frac{N^2}{n^2}\sum_{i\in s_r}\frac{y_i^2}{\theta_i}\left(\frac{1}{\theta_i} - 1\right) \text{ and}$$

$$\widehat{V}_{srs}(2) = \widehat{V}^*(2) + \frac{N^2}{n^2}\sum_{i\in s_r}\frac{y_i^2}{\theta_i}\left(\frac{1}{\theta_i} - 1\right)$$

where

$$\widehat{V}^*(1) = \frac{N(N-n)}{n^2(n-1)}\frac{1}{2}\sum_{i\neq j}\sum_{\in s_r}\frac{(y_i - y_j)^2}{\theta_i\theta_j} \text{ and}$$

$$\widehat{V}^*(2) = N\left(\frac{N-n}{n^2}\right)\left(\sum_{i\in s_r}\frac{y_i^2}{\theta_i} - \frac{1}{n-1}\sum_{i\neq j}\sum_{\in s_r}\frac{y_i}{\theta_i}\frac{y_j}{\theta_j}\right)$$

Corollary 15.4.1

In case all the response probabilities are equal, i.e., $\theta_i = \theta$ for all $i = 1,\ldots, N$, then $\widehat{Y}_{nr}(srs) = \dfrac{Nr}{n\theta}\overline{y}(s_r)$ becomes an unbiased estimator of Y with variance

$$V\left[\widehat{Y}_{srs}(nr)\right] = \frac{N}{n}\left(\frac{1}{\theta} - 1\right)\sum_i \gamma_i^2 + N\frac{N-n}{n}S_\gamma^2$$

where $\overline{y}(s_r) = \sum_{i\in s_r} \gamma_i/r$.

Furthermore, if the response probability θ is estimated by $\widehat{\theta} = r/n$, then $\widehat{Y}_{srs}(nr) = N\,\overline{y}(s_r)$.

15.4.7.2 Singh and Singh Method

Singh and Singh (1979) considered the response sample s_r as an SRSWOR subsample of size r from s and proposed an unbiased estimator for the population total as

$$\widehat{Y}_{ss}(nr) = \frac{n}{r}\sum_{i\in s_r}\frac{\gamma_i}{\pi_i} \tag{15.4.39}$$

The expectation, variance, and unbiased estimator of the variance of $\widehat{Y}_{ss}(nr)$ are given in the following theorem.

Theorem 15.4.5

(i) $E\left[\widehat{Y}_{ss}(nr)\right] = Y$

(ii) $V\left[\widehat{Y}_{ss}(nr)\right] = V_{ht} + \left[\sum_{i\in U}\dfrac{\gamma_i^2}{\pi_i} - \dfrac{1}{n-1}\sum_{i\neq}\sum_{j\in U}\dfrac{\pi_{ij}}{\pi_i\pi_j}\gamma_i\gamma_j\right]E_r\left(\dfrac{n-r}{r}\right)$

(iii) $\widehat{V}\left[\widehat{Y}_{ss}(nr)\right] = \dfrac{n}{r}\sum_{i\in s_r}\dfrac{\alpha_i\gamma_i^2}{\pi_i} + \dfrac{n(n-1)}{r(r-1)}\sum_{i\neq}\sum_{j\in s_r}\dfrac{\alpha_{ij}\gamma_i\gamma_j}{\pi_{ij}}$

where E_r denotes expectation with respect to the response mechanism; $\alpha_i = \dfrac{n}{r\pi_i} - 1$; $\quad \alpha_{ij} = \dfrac{n(r-1)\pi_{ij}}{r(n-1)\pi_i\pi_j} - 1$; $\quad$ and

$$V_{ht} = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j - \pi_i\pi_j)\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right)^2.$$

Proof

(i) $E\left[\widehat{Y}_{ss}(nr)\right] = E_rE_pE_p\left(\dfrac{n}{r}\sum_{i\in s_r}\dfrac{\gamma_i}{\pi_i}\middle| s\right) = E_rE_p\left(\sum_{i\in s}\dfrac{\gamma_i}{\pi_i}\right) = E_r(Y) = Y$

(ii) $V\left[\widehat{Y}_{ss}(nr)\right] = E_r\left[V_p\left\{\widehat{Y}_{ss}(nr)|r\right\}\right] + V_r\left[E_p\left\{\widehat{Y}_{ss}(nr)|r\right\}\right]$

Now, noting

$$E_r\left(V_p\left\{\widehat{Y}_{ss}(nr)|r\right\}\right) = E_r\left(E_p\left[V_p\left\{\widehat{Y}_{ss}(nr)|s_r, r\right\}\right] + V_p\left[E_p\left\{\widehat{Y}_{ss}(nr)|s_r, r\right\}\right]\right)$$

$$= E_r\left(n\frac{n-r}{r(n-1)}E_p\left\{\sum_{i\in s}\left(\frac{y_i}{\pi_i}\right)^2 - \left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)^2 n\middle| r\right\}\right) + V_p\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)$$

$$= E_r\left[n\frac{n-r}{r(n-1)}\left\{\sum_{i\in U}\frac{y_i^2}{\pi_i} - \left(\sum_{i\in U}\frac{y_i^2}{\pi_i} + \sum_{I\neq}\sum_{j\in U}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\pi_{ij}\right)\Big/ n\right\}\right] + V_{ht}$$

$$= V_{ht} + \left(\sum_{i\in U}\frac{y_i^2}{\pi_i} - \frac{1}{n-1}\sum_{I\neq}\sum_{j\in U}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\pi_{ij}\right)E_r\left(\frac{n-r}{r}\right)$$

and

$$V_r\left[E\left\{\widehat{Y}_{ss}(nr)\middle|r\right\}\right] = V_r(Y) = 0,$$

we can verify the part (ii) of the theorem.

(iii)
$$E\left[\widehat{V}\left\{\widehat{Y}_{ss}(nr)\right\}\right] = E_r\left[\sum_{i\in U}\alpha_i y_i^2 + \sum_{i\neq}\sum_{j\in s_r}\alpha_{ij}y_i y_j\right] \qquad (15.4.40)$$
$$= V\left\{\widehat{Y}_{ss}(nr)\right\}$$

Corollary 15.4.2

For an SRSWOR sampling design, $\pi_i = n/N$, $\pi_{ij} = n(n-1)/\{N(N-1)\}$, and $\widehat{Y}_{ss}(nr) = \frac{N}{r}\sum_{i\in s_r} y_i = N\,\overline{y}_r$, where $\overline{y}_r = \sum_{i\in s_r} y_i/r$. Theorem 15.4.5 therefore reduces to

Theorem 15.4.6

(i) $E\left[N\,\overline{y}_r\right] = Y$

(ii) $V\left[N\,\overline{y}_r\right] = N^2\left[E_r\left(\frac{1}{r}\right) - \frac{1}{N}\right]S_y^2$

(iii) $\widehat{V}\left[\widehat{Y}_{ss}(nr)\right] = N^2\left(\frac{1}{r} - \frac{1}{N}\right)s_{yr}^2$

where

$$s_{yr}^2 = \frac{1}{r-1}\sum_{i\in s_r}\left(y_i - \overline{y}_r\right)^2$$

15.5 MEASUREMENT ERROR

So far, we assumed that if a unit is selected in the sample, the information about the value of the variable under study based on the unit is free from error. But in reality this is far from true. The error may occur because of the use of faulty measuring instrument. For example, if a baumanometer is wrong, one may get incorrect readings of blood pressure. Correct measurement may not be possible to obtain for some of the items such as intelligent quotient (IQ), ability of teaching, and attitude toward work, among others. Faulty responses may also be obtained from the respondents, enumerators, or both because of ambiguous questionnaires, intentionally reporting untrue value from carelessness, or the sensitive nature of questions. Untrue responses can be reported by the respondents if they are influenced by enumerators.

Analysis of measurement errors is reported by various authors. Important works among them include Mahalanobis (1946), Deming (1953), Raj (1968), Sukhatme et al. (1984), and Särndal et al. (1992).

Let y_i be the true value of the ith unit of the study variable y and x_i be the response obtained by an investigator. The response or measurement error for the ith unit is $x_i - y_i$. In this study we will assume the response x_i is a random variable and we will build a model on responses. Because responses from a unit depend on the investigator, enumerator, or measuring instrument, we assume that the response x_i obtained by the ith enumerator based on a sample follows the model

$$x_i = y_i + b_i + \epsilon_{is} \tag{15.5.1}$$

where b_i is the error of measurement associated with the ith unit; ϵ_{is} is the random error component with $E_m(\epsilon_{is}) = 0$, $V_m(\epsilon_{is}) = \sigma_{is}^2$, and $C_m(\epsilon_{is}, \epsilon_{js}) = \rho_s \sigma_{is} \sigma_{js}$ for $i \neq j$; E_m, V_m, and C_m denote, respectively, expectation, variance and covariance with respect to the model. In the aforementioned model it is assumed that the distribution of ϵ_{is} depends on the sample s, i.e., environment under which the data were collected. Here, we consider more simple measurement error model

$$x_i = y_i + b_i + \epsilon_i \tag{15.5.2}$$

with $E_m(\epsilon_i) = 0$, $V_m(\epsilon_i) = \sigma_i^2$, and $C_m(\epsilon_i, \epsilon_j) = \sigma_{ij}$ for $i \neq j$.

15.5.1 Measurement Bias and Variance

Let a sample s of size n be selected using a sampling design p and let π_i and π_{ij} be the inclusion probabilities for the ith and ith and jth ($i \neq j$) units, respectively. The Horvitz–Thompson type estimator for the total Y is given by

$$\widehat{Y}_{me}(ht) = \sum_{i \in s} \frac{x_i}{\pi_i} \tag{15.5.3}$$

The expected value of $\widehat{Y}_{me}(ht)$ is

$$\begin{aligned} E\big[\widehat{Y}_{me}(ht)\big] &= E_p E_m\big[\widehat{Y}_{me}(ht)\big] \\ &= E_p \sum_{i \in s} E_m\left(\frac{x_i}{\pi_i}\right) \\ &= E_p\left(\sum_{j \in s} \frac{y_i + b_i}{\pi_i}\right) \\ &= Y + B \end{aligned} \tag{15.5.4}$$

where $B = \sum_{i=1}^{M} b_i$

The quantity

$$B = E\big(\widehat{Y}_{me}(ht)\big) - Y \tag{15.5.5}$$

is known as measurement bias.

The mean square error of $\widehat{Y}_{me}(ht)$ is given by

$$\begin{aligned} M\big(\widehat{Y}_{me}(ht)\big) &= E_p E_m\big[\widehat{Y}_{me}(ht) - Y\big]^2 \\ &= E_p E_m\big[\widehat{Y}_{me}(ht) - E_p E_m\big\{\widehat{Y}_{me}(ht)\big\} + E_p E_m\big\{\widehat{Y}_{me}(ht)\big\} - Y\big]^2 \\ &= E_p E_m\big[\widehat{Y}_{me}(ht) - E_p E_m\big\{\widehat{Y}_{me}(ht)\big\}\big]^2 + B^2 \\ &= V_{pm}\big(\widehat{Y}_{me}(ht)\big) + B^2 \end{aligned} \tag{15.5.6}$$

The quantity $V_{pm}\big(\widehat{Y}_{me}(ht)\big)$ is known as the total variance. The total variance $V_{pm}\big(\widehat{Y}_{me}(ht)\big)$ is decomposed as follows:

$$\begin{aligned} V_{pm}\big(\widehat{Y}_{me}(ht)\big) &= E_p\big[V_m\big(\widehat{Y}_{me}(ht)\big)\big] + V_p\big[E_m\big(\widehat{Y}_{me}(ht)\big)\big] \\ &= E_P\left(\sum_{i \in s} \frac{\sigma_i^2}{\pi_i^2} + \sum_{i \neq} \sum_{j \in s} \frac{\sigma_{ij}}{\pi_i \pi_j}\right) + V_p\left(\sum_{j \in s} \frac{y_i + b_i}{\pi_i}\right) \\ &= \left(\sum_{i \in U} \frac{\sigma_i^2}{\pi_i} + \sum_{i \neq} \sum_{j \in U} \frac{\sigma_{ij}}{\pi_i \pi_j} \pi_{ij}\right) \\ &\quad + \frac{1}{2} \sum_{i \neq} \sum_{j \in U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i + b_i}{\pi_i} - \frac{y_j + b_j}{\pi_j}\right)^2 \end{aligned} \tag{15.5.7}$$

The first component $\left(\sum_{i\in U}\frac{\sigma_i^2}{\pi_i}+\sum_{i\neq}\sum_{j\in U}\frac{\sigma_{ij}}{\pi_i\pi_j}\pi_{ij}\right)$ is known as measurement variance (MV) and the second component $\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{\gamma_i+b_i}{\pi_i}-\frac{\gamma_j+b_j}{\pi_j}\right)^2$ is known as sampling variance.

The summarization of the discussions yield.

Theorem 15.5.1

For the measurement model (Eq. 15.5.2)

(i) The mean square error of $\widehat{Y}_{me}(ht)$ is

$$M\left(\widehat{Y}_{me}(ht)\right)=V_{pm}\left(\widehat{Y}_{me}(ht)\right)+B^2$$

where $V_{pm}\left(\widehat{Y}_{me}(ht)\right)$ is the total variance and B is the measurement bias.

(ii) $V_{pm}\left(\widehat{Y}_{me}(ht)\right)=\text{MV}+\text{SV}$

where $\text{MV}=\left(\sum_{i\in U}\frac{\sigma_i^2}{\pi_i}+\sum_{i\neq}\sum_{j\in U}\frac{\sigma_{ij}}{\pi_i\pi_j}\pi_{ij}\right)$ and

$$\text{SV}=\text{sampling variance}=\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{\gamma_i+b_i}{\pi_i}-\frac{\gamma_j+b_j}{\pi_j}\right)^2.$$

Thus the presence of measurement errors increases the variance of the Horvitz–Thompson estimator by

$$\left(\sum_{i\in U}\frac{\sigma_i^2}{\pi_i}+\sum_{i\neq}\sum_{j\in U}\frac{\sigma_{ij}}{\pi_i\pi_j}\pi_{ij}\right)+V_{ht}(b,b)-2V_{ht}(b,\gamma)$$

where $V_{ht}(u,v)=\frac{1}{2}\sum_{i\neq}\sum_{j\in U}(\pi_i\pi_j-\pi_{ij})\left(\frac{u_i}{\pi_i}-\frac{u_j}{\pi_j}\right)\left(\frac{v_i}{\pi_i}\frac{v_j}{\pi_j}\right)$.

15.5.1.1 Simple Random Sampling Without Replacement

For SRSWOR $\pi_i=\frac{n}{N}$ and $\pi_{ij}=\frac{n(n-1)}{N(N-1)}$, and Theorem 15.5.1 reduces to

Theorem 15.5.2

For the measurement model (Eq. 15.5.2), the mean square error of $\widehat{Y}_{me}=N\,\overline{x}(s)$ is

$$M[N\,\overline{x}(s)] = N^2\left[\frac{1}{Nn}\left(\sum_{i\in U}\sigma_i^2 + \frac{n-1}{N-1}\sum_{i\neq}\sum_{j\in U}\sigma_{ij}\right)\right.$$
$$\left. + \frac{(1-f)}{n}\left(S_y^2 + S_B^2 + 2S_{by}\right)\right] + B^2$$

where $S_y^2 = \dfrac{1}{N-1}\sum_{i\in U}\left(y_i - \overline{Y}\right)^2$, $S_B^2 = \dfrac{1}{N-1}\sum_{i\in U}\left(b_i - \overline{B}\right)^2$,

$S_{yb} = \dfrac{1}{N-1}\sum_{i\in U}\left(b_i - \overline{B}\right)\left(y_i - \overline{Y}\right)$, and $\overline{B} = \sum_{i\in U} b_i/N$.

If we assume that the measurement errors are independent, i.e., $\sigma_{ij} = 0$, and measurement bias $b_i = b$ is constant then

$$M\{\overline{x}(s)\} = \left[\overline{B}^2 + \frac{\overline{\sigma}^2}{n} + \frac{(1-f)}{n}S_y^2\right] \tag{15.5.8}$$

where $\overline{\sigma}^2 = \sum_{i\in U}\sigma_i^2/N =$ the average MV.

15.5.2 Interpenetrating Subsamples

The concept of interpenetrating subsamples (IPNS) was introduced by Mahalanobis (1946). In this technique two or more subsamples are selected using the same sampling designs. The subsamples may or may not be independent. Such subsamples are called IPNS. From each of the subsamples an estimate of the population parameter is obtained. This method helps to collect improved quality of data and provides estimate of error of the population parameter irrespective of the complexity of the sampling design used. Each of the subsamples is surveyed by a batch of enumerators and processed by different group of people at tabulation and subsequent stages. If the results and tables show significant discrepancies from one subsample to another, then one have reasons to suspect something is wrong somewhere and the entire work may be checked. Thus the method reduces nonsampling errors substantially. In this section we will consider the role of investigator effects in measurement error using the following model:

$$E_m(z_{ij}) = x_{ij} = y_j + b_i,\ V_m(z_{ij}) = \sigma_{ij}^2,\ C_m(z_{ij}, z_{ik}) = \sigma_{ik} \text{ and } C_m(z_{ij}, z_{i'k})$$
$$= 0 \quad \text{for } i \neq i' \tag{15.5.9}$$

where z_{ij} be the response obtained by ith investigator from the jth unit. Let $E_m(V_m)$, $E_p(V_p)$, and $E_I(V_I)$, respectively, be the expectation (variance) with

respect to measurement model, sampling design, and investigator. The operator C_m denotes covariance with respect to the model m.

Here we will consider the following sampling schemes.

Scheme A:

In this scheme k independent samples $s_1,\ldots, s_k$ each of sizes r are selected from a population of N units by SRSWOR method and allotted at random to $k(=n/r)$ (assuming integer) different investigators selected at random from a pool of M investigators. Raj (1968) considered this sampling scheme A when M is very large.

Scheme B:

Here a sample s of size $n(=kr)$ units are selected at random from the population by SRSWOR method and the units are divided at random into k groups $s_1,\ldots, s_k$. Each group is allotted to an investigator as in scheme A.

Let $t_i = \frac{1}{r}\sum_{j\in s_i} z_{ij}$ be an estimator of the population mean $\overline{Y}$ based on the data collected by the ith investigator and

$$\bar{t} = \frac{1}{k}\sum_{i=1}^{k} t_i \tag{15.5.10}$$

be the pooled estimator of $\overline{Y}$ based on the entire sample $s = s_1 \cup \cdots \cup s_k$, then we have the following results.

Theorem 15.5.3

Under the scheme A and model (Eq. 15.5.9), we have

(i) $E(\bar{t}) = \overline{Y} + \overline{B}$

(ii) $V(\bar{t}) = \left(1 - \frac{k}{M}\right)\frac{S_B^2}{k} + \left(1 - \frac{r}{N}\right)\frac{S_y^2}{n} + \frac{A}{n}$

and

(iii) An unbiased estimator of $V(\bar{t})$ is

$$\widehat{V}(\bar{t}) = \frac{1}{k(k-1)}\sum_{i=1}^{k}(t_i - \bar{t})^2 \text{ if } M \text{ is large,}$$

where $\overline{B} = \frac{1}{M}\sum_{i=1}^{B} b_i$, $S_B^2 = \frac{1}{M-1}\sum_{i=1}^{B}\left(b_i - \overline{B}\right)^2$ and

$$A = \frac{1}{NM}\sum_{i=1}^{M}\left(\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{(N-1)}\sum_{j\neq}\sum_{l\in U}\sigma_{ij,il}\right).$$

Proof

(i) $$E(\bar{t}) = E_I\left[\frac{1}{k}\sum_{i=1}^{k} E_p\{E_m(t_i)\}\right]$$

$$= E_I\left[\frac{1}{k}\sum_{i=1}^{k} E_p\left(\frac{1}{r}\sum_{j\in s_i} x_{ij}\right)\right]$$

$$= E_I\left(\frac{1}{k}\sum_{i=1}^{k}\overline{X}_i\right)$$

$$= \frac{1}{M}\sum_{i=1}^{M}\overline{X}_i\left(\text{writing } \overline{X}_i = X_i/N, X_i = \sum_{j\in U} x_{ij}\right)$$

$$= \overline{Y} + \overline{B}$$

(ii) Let $E_{pm}(V_{pm})$ denote the expectation (variance) jointly with the sampling design p and model m. Then

$$V(\bar{t}) = V_I\left[E_{pm}(\bar{t})\right] + E_I\left[V_{pm}(\bar{t})\right]$$

$$= V_I\left(\frac{1}{k}\sum_{i=1}^{k}\overline{X}_i\right) + E_I\left[\frac{1}{k^2}\sum_{i=1}^{k}\left\{V_p\left(\frac{1}{r}\sum_{j\in s_i} x_{ij}\right)\right.\right.$$

$$\left.\left. + E_p\frac{1}{r^2}\left(\sum_{j\in s_i}\sigma_{ij}^2 + \sum_{j\neq}\sum_{l\in s_i}\sigma_{ij,il}\right)\right\}\right]$$

$$= \left(\frac{1}{k} - \frac{1}{M}\right)S_B^2 + E_I\left[\frac{1}{k^2}\sum_{i=1}^{k}\left\{\left(\frac{1}{r} - \frac{1}{N}\right)S_\gamma^2\right.\right.$$

$$\left.\left. + \frac{1}{r}\left(\frac{1}{N}\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{N(N-1)}\sum_{j\neq}\sum_{l\in U}\sigma_{ij,il}\right)\right\}\right]$$

$$= \left(\frac{1}{k} - \frac{1}{M}\right)S_B^2 + \frac{1}{k}\left[\left(\frac{1}{r} - \frac{1}{N}\right)S_\gamma^2\right.$$

$$\left. + \frac{1}{rNM}\sum_{i=1}^{M}\left\{\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{(N-1)}\sum_{j\neq}\sum_{l\in U}\sigma_{ij,il}\right\}\right]$$

$$= \left(1 - \frac{k}{M}\right)\frac{S_B^2}{k} + \left(1 - \frac{r}{N}\right)\frac{S_\gamma^2}{n} + \frac{A}{n}$$

(iii) $(k-1)E\left[\widehat{V}(\bar{t})\right] = E\left(\frac{1}{k}\sum_{i=1}^{k} t_i^2 - \bar{t}^2\right)$

$$= E_I\left[E_p\left\{\frac{1}{k}E_m\sum_{i=1}^{k}\left(\frac{1}{r}\sum_{j\in s_r} z_{ij}\right)^2\right\}\right] - \left[V(\bar{t}) + \{E(\bar{t})\}^2\right]$$

$$= E_I\left[E_p\frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{r^2}\left(\sum_{j\in s_r}\sigma_{ij}^2 + \sum_{j\neq}\sum_{k\in s_r}\sigma_{ij,ik}\right) + \left(\frac{1}{r}\sum_{j\in s_r} x_{ij}\right)^2\right\}\right]$$

$$-\left[V(\bar{t}) + \overline{X}^2\right]$$

$$\left(\text{writing } \overline{X} = X/(MN), X = \sum_{i=1}^{M} X_i\right)$$

$$= E_I\left[\frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{r}\left(\frac{1}{N}\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{N(N-1)}\sum_{j\neq}\sum_{k\in U}\sigma_{ij,ik}\right)\right.\right.$$

$$\left.\left.+\left(\frac{1}{r}-\frac{1}{N}\right)S_\gamma^2 + \overline{X}_i^2\right\}\right] - \left[V(\bar{t}) + \overline{X}^2\right]$$

$$= \frac{1}{rMN}\sum_{i=1}^{M}\left(\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{(N-1)}\sum_{j\neq}\sum_{k\in U}\sigma_{ij,ik}\right)$$

$$+\left(\frac{1}{r}-\frac{1}{N}\right)S_\gamma^2 + \frac{1}{M}\sum_{i=1}^{M}\overline{X}_i^2 - \left[V(\bar{t}) + \overline{X}^2\right]$$

$$= \frac{A}{r} + \left(1-\frac{r}{N}\right)\frac{S_\gamma^2}{r} + \frac{M-1}{M}S_B^2 - V(\bar{t})$$

$$= (k-1)\left[V(\bar{t}) + \frac{S_b^2}{M}\right]$$

Thus $E\left[\widehat{V}(\bar{t})\right] \cong V(\bar{t})$ if M is large.

Theorem 15.5.4

Under the scheme B and model (Eq. 15.5.9)

(i) $E(\bar{t}) = \overline{Y} + \overline{B}$

(ii) $V(\bar{t}) = \left(1-\frac{k}{M}\right)\frac{S_B^2}{k} + \left(1-\frac{n}{N}\right)\frac{S_\gamma^2}{n} + \frac{A}{n}$

and

(iii) An unbiased estimator of $V(\bar{t})$ is

$$\widehat{V}(\bar{t}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (t_i - \bar{t})^2$$

if M and N are large.

Proof

(i) $$E(\bar{t}) = E_I \left[\frac{1}{k} \sum_{i=1}^{k} E_p \{ E_m(t_i) \} \right]$$

$$= E_I \left[\frac{1}{k} \sum_{i=1}^{k} E_p \left(\frac{1}{r} \sum_{j \in s_i} x_{ij} \right) \right]$$

$$= E_I \left(\frac{1}{k} \sum_{i=1}^{k} \overline{X}_i \right)$$

$$= \overline{Y} + \overline{B}$$

(ii) $$V(\bar{t}) = V_I \left[E_{pm}(\bar{t}) \right] + E_I \left[V_{pm}(\bar{t}) \right]$$

$$= V_I \left(\frac{1}{k} \sum_{i=1}^{k} \overline{X}_i \right) + \frac{1}{k^2} E_I \left[V_p \left\{ \sum_{i=1}^{k} \left(\frac{1}{r} \sum_{j \in s_i} x_{ij} \right) \right\} \right.$$

$$\left. + E_p \sum_{i=1}^{k} \frac{1}{r^2} \left(\sum_{j \in s_i} \sigma_{ij}^2 + \sum_{j \neq} \sum_{l \in s_i} \sigma_{ij,il} \right) \right]$$

$$= \left(\frac{1}{k} - \frac{1}{M} \right) S_B^2 + \frac{1}{k^2} E_I \left[\left\{ \sum_{i=1}^{k} V_p \left(\frac{1}{r} \sum_{j \in s_i} x_{ij} \right) \right. \right.$$

$$\left. \left. + \sum_{i \neq}^{k} \sum_{l=1}^{k} Cov \left(\frac{1}{r} \sum_{j \in s_i} x_{ij}, \frac{1}{r} \sum_{j \in s_l} x_{ij} \right) \right\} \right.$$

$$\left. + \frac{1}{rN} \sum_{i=1}^{k} \left(\sum_{j \in s_i} \sigma_{ij}^2 + \sum_{j \neq} \sum_{l \in s_i} \sigma_{ij,il} \right) \right]$$

$$= \left(1 - \frac{k}{M} \right) \frac{S_B^2}{k} + \left(1 - \frac{n}{N} \right) \frac{S_y^2}{n} + \frac{A}{n}$$

$$\left(\text{noting } Cov\left(\frac{1}{r}\sum_{j\in s_i} x_{ij}, \frac{1}{r}\sum_{j\in s_k} x_{ik}\right) = -S_\gamma^2 \Big/ N \text{ and } n = kr\right)$$

$$\text{(iii)} \quad (k-1)E\left[\widehat{V}(\bar{t})\right] = E\left(\frac{1}{k}\sum_{i=1}^{k} t_i^2 - \bar{t}^2\right)$$

$$= E_I E_p\left[\frac{1}{k}\sum_{i=1}^{k} E_m\left(t_i^2\right)\right] - \left(V(\bar{t}) + \overline{X}^2\right)$$

$$= E_I E_p\left[\frac{1}{k}\sum_{i=1}^{k} V_m(t_i) + \left(\frac{1}{r}\sum_{j\in s_i} x_{ij}\right)^2\right] - \left(V(\bar{t}) + \overline{X}^2\right)$$

$$= E_I E_p\left[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r^2}\left(\sum_{j\in s_i}\sigma_{ij}^2 + \sum_{j\neq}\sum_{l\in s_i}\sigma_{ij,il}\right)\right.$$

$$\left. + \left(\frac{1}{r}\sum_{j\in s_i} x_{ij}\right)^2\right] - \left(V(\bar{t}) + \overline{X}^2\right)$$

$$= \frac{1}{rMN}\sum_{i=1}^{M}\left(\sum_{j\in U}\sigma_{ij}^2 + \frac{r-1}{(N-1)}\sum_{j\neq}\sum_{k\in U}\sigma_{ij,ik}\right)$$

$$+ \left(\frac{1}{r} - \frac{1}{N}\right)S_\gamma^2 + \frac{1}{M}\sum_{i=1}^{M}\overline{X}_i^2 - \left(V(\bar{t}) + \overline{X}^2\right)$$

$$= \frac{A}{r} + \left(1 - \frac{r}{N}\right)\frac{S_\gamma^2}{r} + \frac{M-1}{M}S_b^2 - V(\bar{t})$$

$$= (k-1)\left(V(\bar{t}) + \frac{S_B^2}{M} + \frac{S_\gamma^2}{N}\right)$$

Hence $E\left[\widehat{V}(\bar{t})\right] = V(\bar{t})$ if M and N are large.

Remark 15.5.1

The variance $V(\bar{t})$ based on the sampling scheme B is much lower than that based on the sampling scheme A.

Remark 15.5.2

If we suppose that there are only k investigators, i.e., $k = M$ (Särndal et al., 1992) and each of them is allotted one sample at random for collection of data, then for both the sampling schemes A and B, $V(\bar{t})$ become free of the investigator effect S_B^2 and they are, respectively, given as $\left(1 - \frac{r}{N}\right)\frac{S_\gamma^2}{n} + \frac{A}{n}$ and $\left(1 - \frac{n}{N}\right)\frac{S_\gamma^2}{n} + \frac{A}{n}$. In this situation $\widehat{V}(\bar{t})$ becomes unbiased for $V(\bar{t})$ if

S_B^2 is very small for the sampling scheme A, but for the sampling scheme B, the unbiasedness condition of $\widehat{V}(\bar{t})$ requires N should be large in addition to small value of S_B^2.

Remark 15.5.3
If $\sigma_{ij}^2 = \sigma_i^2$, $\sigma_{ij,ik} = \rho_i \sigma_i^2$, $\overline{\sigma}^2 = \sum_{i=1}^{M} \sigma_i^2 / M$, and $\overline{\rho} = \sum_{i=1}^{M} \rho_i \sigma_i^2 / \sum_{i=1}^{M} \sigma_i^2$, then $V(\bar{t})$ reduces to $\left(1 - \frac{k}{M}\right)\frac{S_B^2}{k} + \left(1 - \frac{r}{N}\right)\frac{S_y^2}{n} + \frac{(1 + (r-1)\overline{\rho})\overline{\sigma}^2}{n}$ and $\left(1 - \frac{k}{M}\right)\frac{S_b^2}{k} + (1-f)\frac{S_y^2}{n} + \frac{(1 + (r-1)\overline{\rho})\overline{\sigma}^2}{n}$ for sampling schemes A and B, respectively.

15.6 EXERCISES

15.6.1 Describe the sources of nonsampling errors and method of controlling these errors.

15.6.2 What is nonresponse? Describe a few methods of controlling nonresponses in a survey.

15.6.3 A sample of 30 households was selected from 100 households in a locality. 20 of them responded and the remaining 10 did not respond. The mean income of the responded households was \$8500. A sample of 4 households was selected from the 10 nonresponded by SRSWOR and the sample mean income was \$12,000. Estimate the mean income of the locality and its standard error. State clearly the assumption made for the estimation.

15.6.4 A sample of 10 households was interviewed in a certain region by telephone. The number of days the respondents were at home for the last 5 consecutive days at the time of the interview was recorded along with their weekly expenditure on fuel. The data collected are given below.

Households	1	2	3	4	5	6	7	8	9	10
No. of days at home	3	2	4	4	5	0	0	1	4	5
Weekly expenditure on fuel (in Rs)	325	400	125	150	215	320	420	370	120	150

Using Politz and Simmons technique, give an estimate of the average weekly consumption of fuel and an unbiased estimator of the variance of the estimator used.

15.6.5 The following table gives the production and area under wheat cultivation of 15 farms selected at random from 150 farms of a block.

Farms	Area (in acre)	Production of wheat (000 kg)
1	150	120
2	160	150
3	70	56
4	120	?
5	80	60
6	100	75
7	150	?
8	65	?
9	120	100
10	80	70
11	90	100
12	150	110
13	60	?
14	100	80
15	80	60

?, indicates missing observation.

Estimate the average production of wheat of the block by (i) mean, (ii) ratio, and (iii) regression methods of imputations. Estimate the standard errors of the estimators used.

15.6.6 A sample of 10 doctors is selected by SRSWOR method from 60 doctors in a hospital and number of patients who consulted on a certain day is given below:

Doctor	1	2	3	4	5	6	7	8	9	10
No. of patients consulted	30	25	20	25	?	10	20	?	?	30

?, indicates missing observation.

Estimate the average number of patients consulted per day by a doctor and its standard error assuming (i) nonresponse is at random,

(ii) probability of obtaining response from each of the doctors is 0.80, and (iii) response sample is an SRSWOR subsample from the selected sample.

15.6.7 A sample of s of n households is selected by some probability sampling with inclusion probability π_i for the ith unit. Selected households were interviewed once from Monday to Friday. If a respondent was found at home, he/she was asked about the expenditure on food (y) for that week along with the information regarding how many days he/she was home for the 4 preceding working days. Let k_i be the number of days the respondent was at home during the preceding 4 working days and s' be the set of respondents. Show that $t = \dfrac{\sum\limits_{i\in s'} \dfrac{5y_i}{(k_i+1)\pi_i}}{\sum\limits_{i\in s'} \dfrac{5}{(k_i+1)\pi_i}}$ is an unbiased estimator for the mean weekly consumption on food. Derive an approximate mean square error of t.

15.6.8 Let a sample s of size n be selected from a population of N units by SRSWOR method. Let the observed response y_j be obtained from the jth unit related with the true response μ_j through the model $y_i = \mu_i + a$ with a as a constant bias. Show that the sample mean $\overline{y}$, ratio estimator $\overline{y}_R = \dfrac{\overline{y}}{\overline{x}}\overline{X}$, and regression estimator $\overline{y}_{reg} = \overline{y} - b\left(\overline{x} - \overline{X}\right)$ have a constant bias while the estimated error variances of $\overline{y}$, ratio, and regression estimators are free from the measurement bias.