Chapter 13
Non Sampling Errors

It is a general assumption in the sampling theory that the true value of each unit in the population can be obtained and tabulated without any errors. In practice, this assumption may be violated due to several reasons and practical constraints. This results in errors in the observations as well as in the tabulation. Such errors which are due to the factors other than sampling are called **non-sampling errors**.

The non-sampling errors are unavoidable in census and surveys. The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both – sampling errors as well as non-sampling errors. The non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample.

In general, the sampling errors decrease as the sample size increases, whereas non-sampling error increases as the sample size increases.

In some situations, the non-sampling errors may be large and deserve greater attention than the sampling error.

In any survey, it is assumed that the value of the characteristic to be measured has been defined precisely for every population unit. Such a value exists and is unique. This is called the **true value** of the characteristic for the population value. In practical applications, data collected on the selected units are called **survey values** and they differ from the true values. Such difference between the true and observed values is termed as the **observational error** or **response error.** Such an error arises mainly from the lack of precision in measurement techniques and variability in the performance of the investigators.

Sources of non-sampling errors:

Non sampling errors can occur at every stage of planning and execution of survey or census. It occurs at the planning stage, fieldwork stage as well as at tabulation and computation stage. The main sources of the nonsampling errors are

- lack of proper specification of the domain of study and scope of the investigation,
- incomplete coverage of the population or sample,
- faulty definition,
- defective methods of data collection and
- tabulation errors.

More specifically, one or more of the following reasons may give rise to nonsampling errors or indicate its presence:

- The data specification may be inadequate and inconsistent with the objectives of the survey or census.
- Due to the imprecise definition of the boundaries of area units, incomplete or wrong identification of units, faulty methods of enumeration etc., the data may be duplicated or may be omitted.
- The methods of interview and observation collection may be inaccurate or inappropriate.
- The questionnaire, definitions and instructions may be ambiguous.
- The investigators may be inexperienced or not trained properly.
- The recall errors may pose difficulty in reporting the true data.
- The scrutiny of data is not adequate.
- The coding, tabulation etc. of the data may be erroneous.
- There can be errors in presenting and printing the tabulated results, graphs etc.
- In a sample survey, the non-sampling errors arise due to defective frames and faulty selection of sampling units.

These sources are not exhaustive but surely indicate the possible source of errors.

Non-sampling errors may be broadly classified into three categories.

(a) **Specification errors**: These errors occur at planning stage due to various reasons, e.g., inadequate and inconsistent specification of data with respect to the objectives of surveys/census, omission or duplication of units due to imprecise definitions, faulty method of enumeration/interview/ambiguous schedules etc.

(b) **Ascertainment errors**: These errors occur at field stage due to various reasons e.g., lack of trained and experienced investigations, recall errors and other type of errors in data collection, lack of adequate inspection and lack of supervision of primary staff etc.

(c) **Tabulation errors**: These errors occur at tabulation stage due to various reasons, e.g., inadequate scrutiny of data, errors in processing the data, errors in publishing the tabulated results, graphs etc.

Ascertainment errors may be further sub-divided into

(i) **Coverage errors** owing to over-enumeration or under-enumeration of the population or the sample, resulting from duplication or omission of units and from the non-response.

(ii) **Content errors** relating to the wrong entries due to the errors on the part of investigators and respondents.

Same division can be made in the case of tabulation error also. There is a possibility of missing data or repetition of data at tabulation stage which gives rise to coverage errors and also of errors in coding, calculations etc. which gives rise to content errors.

Treatment of non-sampling errors:

Some conceptual background is needed for the mathematical treatment of non-sampling errors.

Total error: Difference between the sample survey estimate and the parametric true value being estimated is termed as total error.

Sampling error:

If complete accuracy can be ensured in the procedures such as determination, identification and observation of sample units and the tabulation of collected data, then the **total error** would consist only of the error due to sampling, termed as sampling error.

The measure of sampling error is mean squared error (*MSE*). The *MSE* is the difference between the estimator and the true value and has two components:

- square of sampling bias.
- sampling variance.

If the results are also subjected to non-sampling errors, then the total error would have both sampling and non-sampling error.

Total bias:

The difference between the expected value and the true value of the estimator is termed as total bias. This consists of **sampling bias** and **nonsampling bias**.

Non-sampling bias:

For the sake of simplicity, assume that the two following steps are involved in the randomization:

(i) for selecting the sample of units and

(ii) for selecting the survey personnel.

Let $\hat{\bar{Y}}_{sr}$ be the estimate of population mean $\bar{Y}$ based on s^{th} sample of units supplied by the rth sample of the survey personnel. The conditional expected value of $\hat{\bar{Y}}_{sr}$ taken over the second step of randomization for a fixed sample of units is

$$E_r(\hat{\bar{Y}}_{sr}) = \hat{\bar{Y}}_{so},$$

which may be different from $\hat{\bar{Y}}_s$ based on true values of the units in the sample.

The expected value of $\hat{\bar{Y}}_{so}$ over the first step of randomization gives

$$E_s(\hat{\bar{Y}}_{so}) = \bar{Y}^*,$$

which is the value for which an unbiased estimator can be had by the specified survey process. The value $\bar{Y}^*$ may be different from true population mean $\bar{Y}$ and the total bias is given as

$$Bias_t(\hat{\bar{Y}}_{sr}) = \bar{Y}^* - \bar{Y}.$$

The sampling bias is given by

$$Bias(\hat{\bar{Y}}) = E_s(\hat{\bar{Y}}_s) - \bar{Y}.$$

The non-sampling bias is

$$\begin{aligned} Bias_r(\hat{\bar{Y}}_{sr}) &= Bias_t(\hat{\bar{Y}}_{sr}) - Bias_s(\hat{\bar{Y}}_s) \\ &= \bar{Y}^* - E_s(\hat{\bar{Y}}_s) \\ &= E_s(\hat{\bar{Y}}_{so} - \hat{\bar{Y}}_s) \end{aligned}$$

which is the expected value of the **non-sampling deviation**.

In the case of complete enumeration, there is no sampling bias and the total bias consists only of non-sampling bias.

In the case of sample surveys, the total bias consists only of the non-sampling bias.

The non-sampling bias in a census can be estimated by surveying a sample of units in the population using better techniques of data collection and compilation than those adopted under general census condition. The surveys are called **post-enumeration surveys**, which are usually conducted just after the census for studying the quality of census data, may be used for this purpose.

In a large scale sample survey, the ascertainment bias can be estimated by resurveying a sub-sample of the original sample using better survey techniques.

Another method of checking survey data is to compare the values of the units obtained in the two surveys and to reconcile the discrepant figures by further investigation. This method of checking is termed **reconciliation (check) surveys**.

Non-sampling variance:

The MSE of $\hat{\bar{Y}}_{sr}$ based on s^{th} sample of units and supplied by r^{th} sample of the survey personnel is

$$MSE(\hat{\bar{Y}}_{sr}) = E_{sr}(\hat{\bar{Y}}_{sr} - \bar{Y})^2$$

where $\bar{Y}$ is the true value being estimated. This takes into account both the sampling and the non-sampling errors, i.e.,

$$\begin{aligned} MSE(\hat{\bar{Y}}_{sr}) &= Var(\hat{\bar{Y}}_{sr}) + \left[Bias(\hat{\bar{Y}}_{sr}) \right]^2 \\ &= E(\hat{\bar{Y}}_{sr} - \bar{Y}^*)^2 + (\bar{Y}^* - \bar{Y})^2 \end{aligned}$$

where $\bar{Y}^*$ is the expected value of the estimator taken over both steps of randomization.

Taking the variance over the two steps of randomization, we get

$$\begin{aligned} Var_{sr}(\hat{\bar{Y}}_{sr}) &= Var_s\left[E_r(\hat{\bar{Y}}_{sr}) \right] + E_s\left[Var_r(\hat{\bar{Y}}_{sr}) \right] \\ &= \underbrace{Var_s\left[\hat{\bar{Y}}_{so} \right]}_{\text{sampling variance}} + \underbrace{E_s\left[E_r(\hat{\bar{Y}}_{sr} - \hat{\bar{Y}}_{so})^2 \right]}_{\text{non-samping variance}}. \end{aligned}$$

Note that

$$\hat{\bar{Y}}_{sr} - \hat{\bar{Y}}_{so} = (\hat{\bar{Y}}_{sr} - \hat{\bar{Y}}_{so} - \hat{\bar{Y}}_{or} + \bar{Y}^*) + (\hat{\bar{Y}}_{or} - \hat{\bar{Y}}^*)$$

where $\hat{\bar{Y}}_{or} = E_s(\hat{\bar{Y}}_{sr})$.

$$E(\hat{\bar{Y}}_{sr} - \hat{\bar{Y}}_{so})^2 = \underbrace{E_{sr}(\hat{\bar{Y}}_{sr} - \hat{\bar{Y}}_{so} - \hat{\bar{Y}}_{or} + \hat{\bar{Y}}^*)^2}_{\text{Interaction between sampling and non-sampling errors}} + \underbrace{E_r(\hat{\bar{Y}}_{or} - \hat{\bar{Y}}^*)^2}_{\text{Variance between survey personnel}}$$

The *MSE* of an estimator consists of

- sampling variance,
- interaction between the sampling and the non-sampling errors,
- the variance between survey personnel and
- square of the sum of sampling and non-sampling biases.

In the complete census, the MSE is composed of only the non-sampling variance and square of the non-sampling bias.

Non-response error:

The non-response error may occur due to refusal by respondents to give information or the sampling units may be inaccessible. This error arises because the set of units getting excluded may have characteristic so different from the set of units actually surveyed as to make the results biased. This error is termed as non-response error since it arises from the exclusion of some of the anticipated units in the sample or population. One way of dealing with the problem of non-response is to make all the efforts to collect information from a sub-sample of the units not responding in the first attempt.

Measurement and control of errors:

Some suitable methods and adequate procedures for control can be adopted before initiating the main census or sample survey. Some separate programmes for estimating the different types of non-sampling errors are also required. Some such procedures are as follows:

1. Consistency checks:

Certain items in the questionnaires can be added, which may serve as a check on the quality of the collected data. To locate the doubtful observations, the data can be arranged in increasing order of some basic variable. Then they can be plotted against each sample unit. Such graph is expected to follow a certain pattern and any deviation from this pattern would help in spotting the discrepant values.

2. Sample check

An independent duplicate census or sample survey can be conducted on a comparatively smaller group by trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, then it is possible to detect the presence of non-sampling errors and to get an idea of their magnitude . Such a procedure is termed as the **method of sample check.**

3. Post-census and post-survey checks:

It is a type of sample check in which a sample (or subsample) is selected of the units covered in the census (or survey) and re-enumerate or re-survey it by using better trained and more experienced survey staff than those involved in the main investigation. This procedure is called as post-survey check or post-census. The effectiveness of such check surveys can be increased by

- re-enumerating or re-surveying immediately after the main census to avoid recall error
- taking steps to minimize the conditioning effect that the main survey may have on the work of the check-survey.

4. External record check:

Take a sample of relevant units from a different source, if available, and to check whether all the units have been enumerated in the main investigation and whether there are discrepancies between the values when matched. The list from which the check-sample is drawn for this purpose, need not be a complete one.

5. Quality control techniques:

The use of tools of statistical quality control like control chart and acceptance sampling techniques can be used in assessing the quality of data and in improving the reliability of final results in large scale surveys and census.

6. Study or recall error:

Response errors arise due to various factors like the attitude of respondents towards the survey, method of interview, skill of the investigators and recall errors. Recall error depends on the length of the reporting period and on the interval between the reporting period and data of survey. One way of studying recall error is to collect and analyze data related to more than one reporting period in a sample (or sub-sample) of units covered in the census or survey.

7. Interpenetrating sub-samples:

The use of interpenetrating sub-sample technique helps in providing an appraisal of the quality of information as the interpenetrating sub-samples can be used to secure information on non-sampling errors such as differences arising from differential interviewer bias, different methods of eliciting information etc. After the sub-samples have been surveyed by different groups of investigators and processed by different team of workers at the tabulation stage, a comparison of the final estimates based on the sub-samples provides a broad check on the quality of the survey results..