1. Introduction

Welcome to the series of e-learning modules on Non-Sampling Errors. In this module we are going cover the types of errors in a survey, various sources of non-sampling errors and the different types of non sampling errors

By the end of this session, you will be able to explain:

- Types of errors in a survey
- Various sources of non-sampling errors
- Types of non-sampling errors

In any survey two types of errors are likely to occur
A, Sampling Errors and b, Non-sampling errors.

Sampling Errors are the errors which are introduced due to the errors in the selection of the sample, or the differences between population's parameters and estimates which are derived from a random sample.

Non-Sampling Errors are the errors, other than those attributable to sampling, that arise during the course of almost all survey, such as respondents' different interpretation of questions, mistakes in processing results, or errors in the sampling frame.

Both Sampling and Non-sampling errors need to be controlled and reduced to a level at which their presence does not defeat or obliterate the usefulness of the final sample results.
There are however, other sources of variation in surveys caused by non-sampling errors which are particularly harmful when they are non-random and cause biased estimates from household surveys.

All survey data are subject to error from various sources.
The broad fundamental distinction of errors is between:

1. Errors in the measurement process, and
2. Errors in the estimation of population values from measurement of a sample drawn from the population.

Sampling errors arise solely as a result of drawing a probability sample, rather than conducting a complete enumeration.
Non-sampling errors, on the other hand, are mainly associated to data collection and processing procedures.
The former is as a result of selecting a sample instead of canvassing the whole population, while the latter is mainly due to adopting wrong procedures in the system of data collection and, or processing.
Total survey error, therefore is equal to sampling error plus non sampling error.

Non sampling errors, therefore, arise mainly due to:

- Misleading definitions and concepts
- Inadequate frames
- Unsatisfactory questionnaires

- Defective methods of data collection, tabulation, coding,
- Incomplete coverage of sample units etc.

Unlike in the case of sampling error, this error may increase with an increase in sample size. If not properly controlled, non sampling errors can be more damaging than sampling errors for large-scale household surveys.

2. Sources of Non-Sampling Errors

Here is a look at the sources of non-sampling errors.
A difference in the estimated values and the actual values of the parameters occur due to many errors which are termed as Non-Sampling Errors.
Various sources of such errors are:
One,
Observational or Response errors
If the observations are taken repeatedly on the same unit, the observed values generally differ, or otherwise if the same respondent is asked the same question repeatedly his response may differ.

Two,
Lack of preciseness in definition also adds to non-sampling errors:
Eg: Judging the loss of crop due to disease like wilt or rust, will be subject to error due to definitions of what we call severely diseased, moderately diseased and low intensity of disease.
More often this measure of intensity will vary from person to person depending on the maturity, qualifications and training the person has.

Three,
Errors are also introduced in editing and tabulation of data:
We reiterate that non sampling errors arise due to various causes right from initial stage when the survey is being planned and designed to the final stage when data are processed and analyzed.

A household survey programme is a set of rules, which specify various operations.
The rules, for instance, describe the population under coverage, specifies concepts, the definitions to be used, methods of data collection and measurements to be made.
However, even though the various survey operations are strictly carried out and expected to yield the true value 'y_i' of the characteristics under study, this is rarely achieved in practice.

In general, non-sampling errors may arise from one or more of the following factors:

- Data specification being inadequate or inconsistent with respect to the objectives of the survey.
- Duplication or omission of units due to imprecise definition of the boundaries of area units, incomplete or wrong identification of particulars of units or faulty methods of enumeration.
- Inappropriate methods of interview, observation or measurement using ambiguous questionnaires, definitions or instructions.
- Lack of trained and experienced field enumerators including lack of good quality field supervision.
- Inadequate scrutiny of the basic data.
- Errors in data processing operations such as coding, keying, verification, tabulation etc.
- Errors during presentation and publication of tabulated results.

This list is by no means exhaustive.

3. Components of Non-Sampling Errors

Brieumer and Lyberg (2003) identified five components of non-sampling errors.

1. Specification Error
2. Coverage/ Frame Error
3. Non-Response Error
4. Measurement Error
5. Processing Error

We may add that estimation error is another error, which should be considered.
We shall have a look at the details of these errors in slides to come.

Specification error occurs when the concept implied by the question is different from the underlying construct that should be measured.
A simple question such as '*how many children does a person have*' can be subject to different interpretations in different cultures.
In households with extended family members, biological children may not be distinguished from children of brothers or sisters living in the same household.

In a disability survey, a general question asking people whether or not they have a disability can be subject to different interpretations depending on the severity of the impairment or the respondent's perception of disability.
People with minor disabilities may perceive themselves to have no disability. Unless the right screening and filter questions are included in the questionnaire, the answers may not fully bring out the total number of people with disabilities.

Specification Errors are the errors which arise due to faulty planning. They arise due to three causes:

1. Wrong data specification which may be inconsistent with the objectives of the survey or census
2. Sampling units may be inadequately defined or they may be wrongly identified. Sometimes units may be duplicated or may be overlapping. The methods of enumeration may be faulty.
3. The questionnaire/ schedules may be faulty. The methods of interview or observation and measurement may be inappropriate.

In most area surveys, primary sampling units comprises of clusters of geographic units generally called Enumeration Areas (EAs).
It is not uncommon that the demarcation of these EAs is not properly carried out during census mapping. Thus households may be omitted or duplicated in the second stage frame.

Frame imperfections can bias the estimates in the following ways:

- If units are not represented in the frame, but should have been part of the frame, it results in zero probability of selection for those units omitted from the frame.

- On the other hand if some units are duplicated, this results in over coverage with such units having larger probabilities of selection.

Under-coverage is the most common in large-scale surveys in most African countries.
In multi-stage household surveys, which are common in the Southern African Development Community region, sampling involves a number of stages, such as:

- Selection of area units in one or more stages;
- Listing and selection of households; and
- Listing and selection of persons within selected households.

Coverage error can arise in any of these stages.

Non-coverage denotes failure to include some sample units of a defined survey population in the sampling frame.
Because such units have zero probability of selection, they are effectively excluded from the survey results.
Here, it is important to note that we are not referring to deliberate and explicit exclusion of sections of a larger population from the survey population.
Survey objectives and practical difficulties determine such deliberate exclusions.

For example:
Attitudinal surveys on marriage may exclude:

- Persons under the minimum legal age for marriage.
- Residents of institutions are often excluded because of practical survey difficulties.
- Areas in a country infested with landmines may be excluded from a household survey to safeguard the safety of field workers.

When computing non-coverage rates, members of the group deliberately and explicitly excluded should not be counted either in the survey population or under non-coverage.
In this regard, defining the survey population should be a part of the clearly stated essential survey conditions

Non-coverage is often associated with problems of incomplete frames.
Examples are: Omissions in preparing the frames as well as missed units, implying omissions due to faulty execution of survey procedures.
Thus non-coverage refers to the negative errors resulting from failure to include elements that would, under normal circumstances, belong in the sample.
Positive errors of over-coverage also occur due to inclusion in the sample of elements that do not belong there.

- The term *Gross Coverage Error* refers to the sum of the absolute values of non-coverage and over-coverage error rates.
- Most household surveys in developing countries suffer mainly from under-coverage errors.
- Most survey research practitioners agree that in most social surveys, under-coverage is a much more a common problem. Corrections and weighting for non-coverage are much more difficult than for non-responses.

The non-coverage errors may be caused by the use of faulty frames of sampling units. If the frames are not updated or old frames are used as a device to save time or money, it may lead to serious bias.

For example, in a household survey, if an old list of housing units is not updated from the time of its original preparation, say 10 years prior to the current survey, newly added housing units in the selected enumeration area will not be part of the second stage frame of housing units.

4. Non-Coverage Errors and Non-Response Errors

Non-coverage errors differ from Non-response errors.

Non-response errors result from failure to obtain observations on some sample units, due to refusals, failure to locate addresses or find respondents at home and losses of questionnaires.

The extent of non-response can be measured from the sample results.

By contrast, the extent of non-coverage can only be estimated by some kind of check external to the survey operations, sample selections and implementation errors.

Non-response refers to the failure to measure some of the sample units, and thus the failure to obtain observations on some units selected for the sample.

It is instructive to think of the sample population as split into two strata, one consisting of all sample units for which measurements can be obtained and the second for which no measurements could be obtained.

In most cases, non-response is not evenly spread across the sample units but is heavily concentrated among sub-groups. As a result of differential non-response, the distribution of the achieved sample across the subgroups will deviate from that of the selected sample.

This deviation is likely to give rise to non-response bias if the survey variables are also related to the sub-groups.

The non-response rate can be accurately measured if accounts of all eligible elements that fall into the sample are kept.

Response rate for a survey is defined as:

The ratio of the number of questionnaires completed for sample units to the total number of sample units.

Reporting of non-response is good practice in surveys. Non response can be due to:

- Respondents not being-at-home,
- Refusing to participate in the survey,
- Respondents being incapacitated to answer questions
- Lost schedules/ questionnaires.

All categories of non-response refer to eligible respondents and exclude ineligibles.

For example: If a survey is on fertility, frame in the selected EAs will comprise of only women in the reproductive age groups , and exclude females who are not in this group.

There are two types of non-responses: Unit non-response and Item non-response.

Unit non-response implies that no information is obtained from certain sample units.

This may be because respondents refuse to participate in the survey when contacted or they cannot be contacted.

Item non-response refers to a situation where for some units the information collected is incomplete.

Item non-response is therefore, evidenced by gaps in the data records for responding sample

units.
Reasons may be due to refusals, omissions by enumerators and incapacity.

The magnitude of unit non-response, among other reasons, is indicative of the general receptivity, complexity, organization and management of the survey.
The extent of item non-response is indicative of the complexity, clarity and acceptability of particular items sought in a questionnaire and the quality of the interviewer's work in handling those items.

In summary, the types of non respondents include:

- Not-at-homes: Prospective respondents who may not be at home when enumerators visit their households.
- Refusals: Respondents who refuse to give information for whatever reasons.
- Not identifiable respondents.

5. Measurement Errors

Measurement errors arise from the fact that, what is observed or measured departs from the actual values of sample units.
These errors centre on the sustentative content of the survey such as:

- Definition of survey objectives,
- Transformation into usable questions, and
- The obtaining, recording, coding and processing of responses.
  These errors concern the accuracy of measurement at the level of individual units.

For example:
In the initial stage, wrong or misleading definitions and concepts on frame construction and questionnaire design leads to incomplete coverage and varied interpretations by different enumerators leading to inaccuracies in the collected data.

Inadequate instructions to field staff are another source of error.
For some surveys, instructions are vague and unclear, leaving enumerators to use their own judgment in carrying out fieldwork.
At times, sample units in the population lack precise definition, thereby resulting in defective and unsatisfactory frames.
The enumerators themselves can be a source of error.

Depending on the type and nature of the enquiry or information collected, these errors may be assigned to respondents or enumerators or both.
At times, there may be interaction between the two, which may contribute to inflation of such errors.
Likewise, the measurement device or technique may be defective and may cause observational errors.

Reasons for such errors are:

- Inadequate supervision of enumerators
- Inadequately trained and inexperienced field staff
- Problems involved in data collection and other type of errors on the part of respondents
- Failure to understand the question

Errors in data collection are the errors that creep in during fieldwork.
These can be any of the following:

- The investigators may lack training and experience. They may also lack the will and enthusiasm to collect the data properly. Sometimes they are likely to replace one sampling unit by another that is easily approachable.
- *There may be practical difficulties in the collection of the correct data.*

For example: In the collection of data on farm yields, there is likely to be genuine errors of measurement.
In the case of mailed questionnaires, errors may be introduced due to wrong understanding of the questions.

- *There may be accidental loss of information or recording of incorrect data*

- *Field inspection and supervision may be inadequate*

At times, respondents may introduce errors because of the following reasons:

1. Careless and incorrect answers from respondent
2. Respondents answering questions when they do not know the correct answer
3. Deliberate inclination to give wrong answers
4. Memory lapses if there is a long reference period

Processing Errors.
These are the errors in classification and tabulation.
These errors are introduced while processing the data before they are subjected to mathematical analysis. We may classify them as follows:

- Errors due to insufficient scrutiny of the collected data
- Errors in coding , punching, verification and tabulation
- Errors may also be introduced while programming, presenting the data, in printing and other operation

Errors of Estimation are errors that arise in the process of extrapolation of results from the observed sample units to the entire target population.
This group of errors centers on the process of sample design, implementation and estimation.
Biases of the estimating procedure may either be deliberate, due to the use of biased estimation procedures, or it may be due to inadvertent use of wrong formula.

To conclude with, Non-sampling errors can be defined as errors arising during the course of survey activities rather than resulting from the sampling procedure due to mistakes and inaccuracies in the collection and processing of data which can never be eliminated.
However, they can be kept in such limits that they may be ignored in the final analysis.

It is important to note that, the non-sampling errors cannot be avoided in complete enumeration as well as in sample surveys.
These can be minimized through superior management of the survey or investigation, employing fitting personnel and by using modern computational aids.

Here’s a summary of our learning in this session:

- Introduction to types of errors
- Types of non-sampling errors
- Sources of non-sampling errors