CHAPTER - 1

INTRODUCTION #

Survey is a form of data collection or the action of discover and establish facts regarding condition(s) of something to provide accurate information to persons responsible or interested. After analysis, interpretation and evaluation of survey data it becomes relevant source of information for us. The exactness or accuracy of the information is determined by the methodology adopted by the investigators. Methodical, organized and systematic collection of data, followed by a watchful analysis and evaluation with predefined objectives are the pillars for the exactness of data. On the basis of their broadness surveys can be divided into two general categories as:

1. Complete survey (Census or Complete Enumeration)
2. Survey on a part of entire units (Sampling)

1.1 COMPLETE ENUMERATION

One way of obtaining the required information is to collect the data for each and every unit (person, household, field, factory, shop, etc as the case may be) belonging to the population or universe, which is the accumulate of all units of a given type under consideration. This procedure of obtaining information is called complete enumeration survey or 100% inspection.

1.2 SAMPLING

First question clicks to mind regarding sampling is “Why Sampling?” Sampling involves observing some phenomenon of interest from a part and drawing an inference about the phenomenon as it applies to the whole. There are so many situations, where only summary results or figures or both are required for the domain of study as a whole or for group of units and in such situations collection of data for every unit is only a means to an end and not the end itself.

It is worth mentioning that exact planning for the future is not possible, since this would need accurate information on the resources that would be available and on the needs that would have to be satisfied in future. In general, past data are used to forecast the resources and the needs of the future and hence there is some element of uncertainty in planning. Because of this uncertainty, only broad allocations of the resources are usually attempted. Thus some margin of error may be permitted in the

“By a small sample we may judge the whole piece.” —Miguel de Cervantes (1547-1616)
data needed for planning, provided this error is not large enough to affect the broad allocations.

**The principal reasons for sampling are as follows:**

- To save money when absolute precision is not necessary. This is important when one takes into account the cost for survey.
- To save time.
- To elaborate deeper into the situation of individual cases. A sample can make it possible.
- To enumerate infinite populations. Some populations are infinite and can only be sampled,
- To improve efficiency of the survey.

### 1.3 PLANNING OF SURVEY SAMPLING

To obtain desire results from a survey, it is required to pay particular attention to the preparations for the field work. In this regard all surveys require careful and judicious preparations if they have to be successful. The amount of planning will depend on the type of survey, materials and information required.

________________________

**Source:**

- Errors in sample surveys by Tauqueer ahmad, IASRI, New Delhi.
1.4 OBJECTIVE OF THE SURVEY

The initial approach in any survey is to define the purpose or objective of the survey. We are not able to decide exactly what kind of data to collect or what to do with it once you have it without knowledge of the exact nature of the problem (objective). But this knowledge does not provide a relief to esurience of starting of the problem, problem understanding by investigator(s), problem taken is real etc.

In some other words one can define the objective as the justification or explanation of whole project. In practice it is not only a justification but also a guideline to treat the work in proper direction. The objective provides the answers of these questions:

★ What information is to be collect from the survey?
★ From where and whom?
★ At what level of precision?

1.5 SURVEY AREA

When planning to carry out a survey, it is necessary to define the geographical areas and the target population to be covered. In a household income and expenditure survey, for instance, the survey may cover the urban areas and perhaps exclude rural areas. It should be noticed out in practice that; the target population is smaller than the population forming the total population area. Usually because of many reasons the target population becomes restricted. For example, in some surveys, some military households in barracks may be excluded from the survey. In labour-force surveys,
children below a specified age may be shown as members of households surveyed, but would not be part of the labour-force.

1.6 SURVEY BUDGET

A usual or simple meaning by the word ‘Budget’ is related to economy and some mathematical calculations. The budget is not a parameter that figures in the mathematical calculation of sample size, it does figure prominently at a practical level. The survey budget indicates the financial requirements of the survey which is to be conducted. The budget is necessary to support and guide the implementation of the survey and the construction of the timetable for producing the survey results. Cost estimates must be as detailed as possible. It is therefore necessary to understand all the detailed process involved in the survey operation. The budget shows cost of personnel, equipments and all other items of expense. If there is a pre-allocated amount of funds available, the overall survey budget must be within the pre-determined framework. It is also advisable to follow the general guidelines of the financing agency in preparing the budget. The financial requests of the survey should be prepared at an early stage. In general, the budget will depend largely on the survey design, precision required and geographical coverage.

1.7 SAMPLING FRAME

A sampling frame is a list or set of total source materials from which the units are selected in the sample. More than one set of materials may be necessary. In the context of survey research, a population is a theoretically specified aggregate of study
elements or units. We can categorize a target population, a frame population and a 
survey population. The difference between the target population and the frame 
population is non-coverage or perhaps, in some instances, over-coverage. The 
difference between the target population and the survey population is non-coverage, 
over-coverage and non-response.

A perfect sample frame is one that is:

1. Complete
2. Accurate and up-to-date

1.7.1 COMPLETENESS OF SAMPLING FRAME

The ideal frame would be deemed complete with respect to the target population 
if all its members are covered by the frame. Coverage of the target population(s) is 
therefore an essential feature in judging whether the frame is suitable for a survey. If not 
suitable, then whether it can be repaired or further developed to make it more complete 
must be assessed by the survey team.

1.7.2 ACCURACY OF SAMPLING FRAME

Accuracy is an important feature as well in sampling frames, although 
inaccuracies are more likely to occur in frames other than those used for household 
surveys. A frame can be said to be accurate if each member of the target population is 
included once and only once.
With regard to sampling frames or frame populations, the ideal situation the idea is to establish one to one correspondence between the list and the elements. However, that is mainly not the case as there is hardly or never such a one-to-one correspondence in large scale sample surveys. Instead, the following problems become apparent:

- Non-Coverage, Incomplete Frame
- Cluster of Elements in one list
- Blanks or Foreign Elements and
- Duplicate Listings

1.8 HYPOTHESIS AND QUESTIONS FOR RESEARCH

Once the problem has been clearly stated, the next suitable approach is to form one or more hypotheses. The hypothesis is actually assumption or guess about the solution to the problem on the basis of our knowledge (not be a capricious guess). It ought to be based on prior experience related to the problem, or perhaps any knowledge may have of previous research done on the topic. Without such a framework in which to make an educated guess, we have no basis for making a guess at all.

1.9 SURVEY EFFORT

The next approach after determining the purpose and hypotheses is constructing the survey plan. The target of the survey plan is to ensure that the survey results will provide sufficient data to provide an answer (solution) to the problem which we are investigating. A survey plan comprises of three different parts:
None of these plans are able for decision. Decisions we make on how we will analyze which affects our data collection plan. The type of data reduction we do will affect not only the types of analyses we can do, but also the amount and types of data need to collect. Because these plans are closely interrelated, they should be developed concurrently.

1.9.1 DATA COLLECTION

The purpose of the data collection plan is to ensure that proper data are collected in the right amounts. The hypothesis and data analysis plan determine the appropriateness of the data. For example, an agency is going to plan for analyze results by age group to test a hypothesis, then agency must collect data from each age group whose opinions they want to know. The right amount applies to sample data. The use of sample data involves risk, and the amount of that risk is determined by the size of sample. The amount of risk is willing or able to accept should be stated in analysis plan.

1.9.2 DATA REFORMATTING

Purpose of the data reduction and reformatting plan is to identify up front and to decrease as much as possible the amount of data handling (reduction and reformatting) will have to do. This plan is highly dependent on the other two plans. If collection plan calls for a great deal of data, one should plan to use a computer to analyze the data.
Now a day’s ADP scanner sheets are to be used to record respondents' answers, include the sheets with the questionnaire so the respondent can fill out the scanner sheet. This will save a great deal of time that would have to spend if transferred the survey data to the scanner sheets. It also eliminates the possibility of making errors in transferring data.

1.9.3 DATA ANALYSIS

Analysis plan ensures that the information produced by the analysis will adequately address the originally stated hypotheses, objectives, or questions. It also ensures an analysis that is compatible with the data collected during the survey. In the analysis plan, one can determine which statistics to use and how much risk one can take in stating the conclusions.

1.10 QUESTIONNAIRE

The last step in preparing the survey is developing the data collection instrument. The most common means of collecting data are the interview and the self- or group-administered questionnaire. The questionnaire provides a standardized data-gathering procedure. The questionnaire plays a central role in the survey process in which information is transferred from those who have it (the respondents) to those who need it (the users). It is the instrument through which the information needs of the users are expressed in operational terms as well as the main basis of input for the data processing system for the particular survey.
It is to understand the comparison regarding the advantages and disadvantages of the questionnaire as opposed to the personal interview. This knowledge allows maximizing the strengths of the questionnaire while minimizing its weaknesses. The advantages of administering a questionnaire instead of conducting an interview are:

- **Lower costs**
- **Better samples**
- **Standardization**
- **Privacy of the respondent(s)**

The primary advantage is lower cost, in time as well as money. The questionnaire can be administered simultaneously to large groups whereas an interview requires each individual to be questioned separately. This allows the questions to reach a given number of respondents more efficiently than is possible with the interview. Many surveys are constrained by a limited budget. Since a typical questionnaire usually has a lower cost per respondent, it can reach more people within a given budget (or time) limit. This can enhance the conduct of a larger and more representative sample.

Using a well-constructed questionnaire can minimize the effects of potential human errors (for example, altering the pattern of question asking, calling at inconvenient times, and biasing by “explaining”. The use of a questionnaire also eliminates any bias introduced by the feelings of the respondents towards the interviewer and vice-versa.

The points of attention while using the questionnaire are:
Non-response (Non-returns)

Measurement error

Mis-interpretation

Validity problems

Non-response or non-returns is the case where the respondent did not answered to questionnaires or individual questions what were asked to them. For example, assume, an agency is conducting a survey to determine the attitude of a group about a new policy. Some of those opposed to it might be afraid to speak out, and they might comprise the majority of the non-returns. This would introduce non-random (or systematic) bias into the survey results. The leading question in this situation comes about minimize these errors.

Mis-interpretation occurs when the respondent does not understand either the survey instructions or the survey questions. If respondents become confused, they will either give up on the survey (becoming a non-return) or answer questions in terms of the way they understand it, but not necessarily the actual way. The questionnaire instructions and questions must be able to stand on their own and must use terms that have commonly understood meanings throughout the population under study.

An important issue related to using a questionnaire is to check the validity of the answer. For example, did the person wanted to survey give the questionnaire to a friend or complete it personally? Did the individual respond indiscriminately? Did the respondent deliberately choose answers to mislead the surveyor? Without observing the
respondent's reactions (as would be the case with an interview) while completing the questionnaire, there is no way of knowing the true answers to these questions.

There are a lot of facts available in the literature to make an effective questionnaire. Some of them are:

- Simple language
- Short questions
- Question sequence
- Limited/less number of questions
- Limit each question to one idea or concept
- Ignorance of leading questions
- Include all possible answers
- Avoid emotional or morally charged questions
- Understand the question
- Formulate questions and answers to obtain exact information and to minimize confusion
- Include a few questions that can serve as checks on the accuracy and consistency of the answers as a whole
- Organize the pattern of the questions
- Pretest the questionnaire
1.11 DATA COLLECTION PLAN

The purpose of the data collection plan is to ensure that proper data are collected in the right amounts. The different methods of collecting data are:

- Physical observation or measurement
- Personal interview
- Mail enquiry
- Telephonic enquiry
- Web-based enquiry
- Method of Registration

These all methods relate to collection of primary data from the units/respondents directly. All these methods have their respective merits and therefore sufficient thought should be given in selection of an appropriate method(s) of data collection in any survey.

1.11.1 PHYSICAL OBSERVATIONS (OR MEASUREMENT)

Data collection by physical observation consists of procedure of physically examining the units/respondents and recording data as a result of personal judgment or using a measuring instrument by the investigator. For instance, in a crop cutting experiment, while estimating the yield of a crop, the plot is demarcated, the crop in the selected plot is harvested and the produce is weighted to estimate the production per unit area. Data obtained by this method are likely to be more accurate but may often prove expensive.
1.11.2 PERSONAL INTERVIEW

The method of personal interview consists in contacting the respondents and collecting statistical data by individual questioning. In this case, the investigator can clearly explain to the respondents the objectives of the survey and the exact nature of the data requirements for giving the required information thus reducing the possibility of non-response arising from non-cooperation, indifference etc. Further, this method is most suitable for collecting data on conceptually difficult items from respondents. However, this method depends heavily on the availability of well trained interviewer.

1.11.3 MAIL ENQUIRY

In a mail enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being sent and collected back through an agency such as the postal department. This method is likely to cost much less as compared to personal interview methods. However, the response may not always be satisfactory depending upon the cooperation of the respondents, the type of questionnaire and the design of the questionnaire. In developing countries where a large proportion of the population is illiterate, the method of mailed questionnaire may not even be feasible.

1.11.4 TELEPHONIC ENQUIRY

In a telephonic enquiry, data are collected by questioning the respondents. This method provides an opportunity of two-way communication and thus can reduce the possibility of item non-response. However, this method can be used only for those
surveys in which all units of target populations have telephone, otherwise it will cause bias in the results.

1.11.5 WEB-BASED ENQUIRY

The increasing popularity and wide availability of World Wide Web technologies provide a new mode of data collection. In web-based enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being posted on the net. One important advantage of using computer technology in data collection is to minimize the loss of data owing to incomplete or incorrectly completed data sets by using Client side validation. In an era of information superhighway, this method is one of the fastest means of data collection. However, in developing countries where a large proportion of the population does not have access to Internet, the method of web-based enquiry may not serve the purpose for most of the surveys. Various Internet sites are using this method for opinion poll on certain issues.

1.11.6 METHOD OF REGISTRATION

In the registration method, respondents are required to register the required information at specified places. The vital statistics registration system followed in many countries provides an illustration of the registration method. The main difficulty with this method, as in the case of the mail enquiry, is the possibility of non-response due to indifference, reluctance, etc. on the part of informants to visit the place of registration and supply the required data.
Apart of these methods one more method is available and relates to the extraction of secondary data, collected earlier generally by one or more of the above discussed methods called transcription from records. The method of transcription from records is used when the data needed for a specific purpose are already available in registers maintained in one or more places, making it no more necessary to collect them directly from the original units at much cost and effort. The method consists in compiling the required information from the registers for the concerned units. This method is extensively used since a lot of government and business statistics are collected as a product of routine administrative operations.

1.12 STRATEGIES FOR SAMPLE SELECTION IN SURVEYS

A number of methods are available in the literature regarding sample selection from a population. These methods are divided into two parts:

* Probability Sampling
* Non Probability Sampling

1.12.1 PROBABILITY SAMPLING

Probability sampling is process of selecting sampling units when there is a known non-zero probability that every element in a population is likely to be selected in a sample. It is the primary method of selection in the social sciences and involves random selection which is a scientific process. Some of the main principles underlying probability sampling would ensure that researchers avoid introducing their own personal biases.
The different procedures for probability based sample selection are

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Cluster Sampling
5. Multistage Sampling

1.12.1.1 SIMPLE RANDOM SAMPLING

Simple random sampling is primarily a method of element sampling that is important not because it is widely used but because it is the basis upon which sampling theory is constructed. It is a basic selection process and all other procedures can be viewed as modifications of it, each of these other procedures or designs being introduced to provide more practical, economic and precise designs. In practice, however, one may hardly ever have to select a design based upon simple random sampling.

1.12.1.1.1 SIMPLE RANDOM SAMPLING WITH REPLACEMENT AND WITHOUT REPLACEMENT

When reference is made to simple random sampling, it is understood to be sampling without replacement. Simple Random Sampling with Replacement implies that an element could appear more than once in a sample. It is also referred to as unrestricted random sampling. This approach is of interest primarily for theoretical
reasons. In practice, simple random sample is usually executed without replacement. In such a case, an element can only appear once in a sample.

**Fig. 1.1** Probability sampling

1.12.1.1.2 SELECTION OF RANDOM SAMPLE

There are three basic steps which needs to follow in selecting a simple random sample. They are:

★ Assign unique identifiers 1 to N to each element in a frame population,
Select n elements out of N using a random process, for example, a table of random numbers, a computer program or a calculator with a random number generator.

The probability of any element being selected is equal to n/N, the ratio of sample size to population size.

1.12.1.2 SYSTEMATIC SAMPLING

Systematic sampling is an alternative method of element sampling. It is used especially on occasions when simple random sampling might considered to be burdensome. In order to undertake systematic sampling, there is a need to select a sample of size n from a population of N elements. This is done by first determining a sampling interval k and a random start a between 1 and k. Note that the sampling interval \( k = N/n \) and that \( a \) is a randomly chosen number that lies between 1 and k. The value \( a \) is ordinal and represents the ordinal rank of the very first selection. Having made the first selection, every \( k^{th} \) element is selected after the first selection. The ordinal rank of the selections follows an arithmetic progression. That is, \( a, a + k, a + 2k, a + 3k, a + 4k... a + (n-1)k \). The sampling fraction is \( n/N \) or \( 1/k \).

1.12.1.3 STRATIFIED SAMPLING

Stratified sampling combines conceptual simplicity with the simple random sampling resulting in potentially significant gains in the reliability of estimates. It allows one to derive separate estimates for parameters for each sub-domain within the overall population and is perhaps the best design for approximating what might be
considered to be a representative sample. The stratified sampling designs elimination of those samples that result in either extremely high or extremely low estimates of the parameters under review.

In Stratified sampling the population is divided into $L$ mutually exclusive and exhaustive strata (on the basis of homogeneity) so that the population size of the stratum $h$ is $N_h$. From each of the $L$ strata, a simple random sample of $n_h$ elements is taken, there being independent selection from each stratum. A stratified sampling design could be conceived of as $L$ separate simple random samples. The total number of possible stratified samples is $T$ where $T$ is as follows:

$$T = \binom{N_1}{n_1} \binom{N_2}{n_2} \binom{N_3}{n_3} \cdots \binom{N_L}{n_L}$$

1.12.1.4 CLUSTER SAMPLING

While earlier designs focused upon sampling techniques that required sampling frames, there are times when the compilation of a sampling frame is not feasible and perhaps not even possible. Cluster sampling is consistent with the construction of sampling frames in order to identify groups or clusters of enumeration units without explicitly listing individual units.

Cluster sampling is a sampling plan which uses a frame consisting of clusters of listing units. It is a hierarchical kind of sampling and hinges upon the idea of a cluster as any sampling unit within which one or more sampling units are contained.
Sampling is done in stages, for example, one stage, two stage and multi-stage. The clusters used at the first stage are called primary sampling units. At each stage, sample frame is involved and developed particularly for those units selected at the various stages. To select a sample of clusters, we use Simple Random Sampling, Systematic Sampling and Stratified Sampling.

**Some examples of cluster sampling**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Listing unit</th>
<th>Elementary unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>City Block</td>
<td>Household</td>
<td>Person</td>
</tr>
<tr>
<td>County</td>
<td>Hospital</td>
<td>Patient</td>
</tr>
<tr>
<td>School</td>
<td>Class Room</td>
<td>Student</td>
</tr>
<tr>
<td>Page of Text</td>
<td>Line of Text</td>
<td>Word</td>
</tr>
<tr>
<td>Week</td>
<td>Day</td>
<td>Hour</td>
</tr>
</tbody>
</table>

### 1.12.1.5 MULTISTAGE SAMPLING

In multistage sampling, the sampling process unfolds at different stages, each stage having its own set of units that are often referred to as the primary sampling units (PSUs). These might include larger spatial units such as counties, enumeration districts, school districts, etc. This is followed be second stage with secondary sampling units and may even consist of further ordinal stages with their respective units. Ultimate or final stage sampling units are the sampling units at the ultimate stage.

### 1.12.2 NON PROBABILITY SAMPLING

This family of survey sampling designs is not based upon random selection but has logic of its own. It is extremely useful in some instances, particularly when a frame does not exist or is impossible to develop given the nature of the target population. Just
as there are different types of probability sampling designs, there are also different types of non-probability sampling designs which include the following:

1. Quota Sampling
2. Snowball Sampling
3. Convenience Sampling
4. Purposive Sampling

1.12.2.1 QUOTA SAMPLING

This is a procedure to maintenance for balance between specific characteristics and most widely used method of non-probability sampling, the population is split up into non-overlapping subgroups, as for stratified sampling. Quotas of the desired number of sample cases are then calculated proportionally to the number of elements in these subgroups. These quotas are then divided up among the interviewers, who simply set out to find individuals who fit the required quota criteria. In quota sampling, we select people non-randomly according to some fixed quota. There are two types of quota sampling like: proportional and non proportional.
1.12.2.2 SNOWBALL SAMPLING

In snowball sampling, the investigator begins by identifying someone who meets the criteria for inclusion in study. Then ask them to recommend others who may know and meet the criteria. This method is useful when trying to reach populations that are inaccessible or hard to find.

1.12.2.3 PURPOSIVE SAMPLING

In this method the sampling is done with a purpose in mind. The investigator has one or more specific predefined groups he is seeking. Purposive sampling can be very useful for situations where investigator need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern. With a purposive
sample, one is likely to get the opinions of target population, but also likely to overweight subgroups in population that are more readily accessible.

1.12.2.4 CONVENIENCE SAMPLING

One of the most common methods of sampling goes under the various titles listed here. This method is based on the availability and hence Potential bias of self-selection.

1.13 ERRORS IN SURVEY SAMPLING

The errors involved in the survey are categorized into two parts as follows:

★ Sampling error

★ Non-sampling error

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.

1.13.1 SAMPLING ERROR
Fig. 1.3 Sampling Error

These are the errors originated due to the procedure of sampling. These errors may occur because of following reasons:

1. Error in selection of sample.
2. Error due to improper choice of the statistic etc. Sampling errors are directly related to the survey sampling and able to minimize by incensement of sample.

As we increase the size of sample these errors becomes minimum.

1.13.2 NON-SAMPLING ERROR

All the errors in estimation, which are not the result of sampling, are called non-sampling errors. Non sampling errors arise due to numerous factors and almost at every stage of survey from planning of the survey to report writing.

In order to study different aspects of non sampling errors, it is desirable to classify the non sampling errors according to the source or the stage of the survey or type of error. There may possible ways to classify the different approaches of non-sampling error. One approach is by the stage of the survey the where non sampling errors occur. The main stages in any survey are

1. Survey design and preparation

2. Data collection and

3. Data Processing and Analysis

Another approach to classify the non sampling errors is due to source or type of error. Categories under such classifications are
1. Coverage errors
2. Non-response errors
3. Measurement or response errors

1.13.2.1 COVERAGE ERRORS

To make inference about the desired population, selection is done by applying appropriate randomized procedure to sampling frame in which all the units of the Target Population are supposed to be represented uniquely. The coverage errors arise mainly due to the use of faulty frame of sampling units. For example, in a household survey if the old list of households prepared for the population census a few years ago is used for selection of the sample, some newly added households will not form a part of the sampling frame whereas a number of households which might have already migrated will remain in the frame.

Coverage errors may also arise due to

* Incorrect specifications or ignorance of correct procedure by field workers.
* Failure to identify actual units selected.
* Enumerating wrong units intentionally or unintentionally by the investigator etc.

1.13.2.2 NON RESPONSE ERRORS

Non-response refers to the failure to measure some of the sample units. Thus 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. There may possible so many reasons for non-response in any sample survey. Some of them are

1. **Not-at-home**
2. **Refusal**
3. **Lake of time**
4. **Lake of understanding**
5. **Self interest**
6. **Bias due to interviewer**
7. **Failure of respondent’s memory**

**1.14 TYPES OF NON-RESPONSES**

There are two types of non-responses:

1. **Unit non-response and**
2. **Item non-response**

**1.15 MISSING DATA IN SAMPLE SURVEYS**

When values of sample under investigation face to generate data or destroy due to unavoidable reason, it causes the missing data problem. As discussed earlier the missing data occurs in the survey because of non-coverage. Non-coverage represents a failure to include some units of the target population in the sampling frame. Total non-response and item non-response are the two other factors responsible for missing ness in
data. Total non-response occurs when no information is collected from a sample unit, and item non-response occurs when some not all the required information is collected from a sample unit.

Non-response cannot be completely eliminated in practice; however it can be minimized by persuasion through repeated visits or other methods.

**1.16 IMPUTATION TO DEAL WITH MISSING DATA**

As a pointed definition, imputation is a technique to replace/impute the missing observation from the survey data. In large scale surveys, it is not possible to visit again and again to the units which are non respondent and hence we need a strong mechanism to cope up from this kind of situation. Imputation provides a strong base to survey when non-response taken account into consideration. Many imputation methods are available in literature of survey sampling.

These Imputation methods can be divided into two groups:

- **By their use of auxiliary variables**
- **By the value assigned to the residuals**

**1.17 AUXILIARY VARIABLES IN SURVEY SAMPLING**

In survey research, often the investigators use supplementary or auxiliary information (say $X$) available for various units in the population to calculate an estimate of a parameter of study variable (say $Y$). Such information is generally based on previous census or large-scale surveys. Auxiliary variables are usually employed to
record the history of a computation, thereby, allow reasoning over the entire computation history.

Supplementary information can be used to improve the precision of the estimates, and may be used in a number of following ways:

1. Selecting the samples with probability proportional to \( x \),

2. Stratifying the population on the basis of \( x \), or,

1.18 MEASUREMENT ERROR IN SURVEY SAMPLING

Measurement error is related to the observation of the variables being measured in any survey, and is also called observational error. A large amount of research literature exists on measurement error. Measurement error can occur at the stage of data collection or the time of non-response, coverage, or data processing. It may arise from various sources mainly from the questionnaire, due to data collection method, the interviewer, and the respondent as suggested by Biemer et al. (1991). Measurement error can be characterized as the difference between the value of a variable provided by the respondent and the true value of that variable. The total survey error of a statistic with measurement error has (I) fixed errors (bias) and (II) variable errors (variance) over repeated trials of the survey. Measurement bias or response bias reflects a systematic pattern or direction in the difference between the respondents’ answers to a question and the correct answer; for example, respondents may tend to forget to report a certain type of income such as interest, resulting in reported income lower than the actual income.
Simple response variance reflects the random variation in the respondent’s answer to a survey question over repeated questioning (i.e., respondents may provide different answers to the same question if they are asked the question several times). Interviewers can be a major source of this type of variable error. Interviewer variance, the variable effects interviewers have on the respondents’ answers, is one form of correlated response variance, a component of total survey error that occurs because response errors might be correlated for sample units interviewed by the same interviewer. One approach to estimating measurement error is to compare the responses received from a survey respondent for specific questions against measures of the same variable from an independent source. As a simple example, if respondents were asked their age, responses could be verified against birth records. However, this true value can be elusive. Even in the simple example of verifying age, one cannot assume for certain that birth records are without errors.

Nonetheless, we seek to assess the measurement error present in the survey measures by comparing them to measures from an independent and reasonably valid source. Another approach frequently used involves comparing responses from an original interview to those obtained in a second interview conducted soon after the original interview. As discussed above the measurement error comes from four primary sources [Biemer et al. (1991)]. These are:

★ Questionnaire
★ Data Collection Method
★ Interviewer
★ **Respondent**

These four sources are the elements that comprise data collection. The questionnaire is the presentation of the request for information. The data collection mode is how the questionnaire is delivered or presented (self-administered, telephone or in person). The interviewer, in the case of telephone or in-person mode, is the deliverer of the questionnaire. The respondent is the recipient of the request for information. Each can introduce error into the measurement process. While we generally address these sources separately, they can also interact. For example, interviewers’ and respondents’ characteristics may interact to introduce errors that would not be evident from either source alone. The sections that follow describe in more detail how each of these sources of errors affects data quality and methods for assessing and reducing their effect.

6.18.1 SOURCES OF MEASUREMENT ERROR

6.18.1 QUESTIONNAIRE EFFECTS

The questionnaire is designed to communicate with the respondent in an unambiguous manner. It represents the survey designer’s request for information. A substantial literature exists on questionnaire effects; for more information, see Groves (1989) and Biemer *et al.* (1991).

The causes for these effects can be:

★ **Specification problems**

★ **Question wording**
★ Length of the questions
★ Length of the questionnaire
★ Order of questions
★ Response categories
★ Open and closed formats
★ Questionnaire format

6.2.2 DATA COLLECTION MODE EFFECTS

Various methods or modes are available for collecting data for a survey. The selection of the data collection mode is a complex decision that depends on the methodological goals of the survey as well as consideration of various factors such as funds available, the questionnaire content, the population covered, expected response rates, length of the collection period, and expected level of measurement error.

★ Face-to-face interviewing
★ Telephone interviewing
★ Self-administered mail surveys
★ Diary surveys
★ Computer assisted self-interviewing (CASI)
★ Direct observation
★ Mixed data collection mode
1.19 REVIEW OF LITERATURE

The review of literature in this content is divided into four segments as review due to survey sampling, due to imputation, due to measurement error and author’s contribution review.

1.19.1 RESEARCH CONTRIBUTION IN SURVEY SAMPLING

Abu-Dayyeh et al. (2003) has studied some estimators of finite population mean using auxiliary information. They have considered two classes of estimators to estimate the population mean for the variable of interest using two auxiliary variables. It turns out that the newly suggested estimators dominate all other well-known estimators in terms of mean squared error and bias. They also presented the content how to extend the two classes of estimators if more than two auxiliary variables are available. Adhikari (2009) has worked on improving the Hansen-Hurwitz estimator in PPSWR Sampling. Al-khasaw (2008) discussed a general class of estimators for estimating the population mean in two-stage cluster sampling with unequal first-stage units. Al-Nasser (2004) proposed a study for estimation of multiple linear functional relationships.

Berger and Skinner (2005) tackled with a jackknife variance estimator for unequal probability sampling. Breidt and Opsomer (2000) described theoretical properties of a new type of model-assisted nonparametric regression estimator for the finite population total, based on local poly- local polynomial regression estimators nominal smoothing. Local polynomial regression is a generalization of kernel regression. In another contribution, Breidt et al. (1996) has given three estimators of mean for this purpose like (i) General linear estimator (ii) Linear composite estimator
and (iii) Ratio type composite estimator. Further, the above three estimators were compared empirically with the help of real survey data.

Brewer (1999) has done an elaborate analysis on design-based or prediction-based inference. A comparison has been made between traditions sampling schemes v/s balanced sampling. They suggested a wide ground to statisticians for choice of the sampling procedure. A simple procedure has been suggested by Choudhury (2004) to estimate unbiased a survey population total and the variance of the estimator for the total based on an unequal probability sub-sample from an initially drawn sample from the population.

Demnati and Rao (2004) proposed a unified approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations for general designs. They illustrated the methods using ratio estimators and estimators defined as solutions to calibration weighted estimating equations. A discussion on consequences of over-stratification is given by Destavola and Cox (2008). The issues of over-stratification and under-stratification are studied analytically by them in their study.

A generalized estimator of a population mean under any sampling design, which utilizes the knowledge of a population mean and variance of auxiliary variable, has been studied by Dubey (2006). His proposed class of estimators been discussed in probability proportional to size sampling. Duchesne (2000) derived explicit jackknife variance
estimators of the general regression estimator (GREG) using the random group technique to removes a large part of the positive model bias using some simulations.

The use of auxiliary information in design-based estimation for domains is advised by Estevao and Särndal (1999). In an other study Estevao and Sarndal (2002) have given the ten cases of auxiliary information for calibration in two-phase sampling. They shown that there are exactly nine different subsets of the complete information, for a total of ten different cases of auxiliary information and proposed one calibration estimator in each of these ten cases. A study is made by them for the precision of the calibration estimators in the ten cases, both theoretically (by deriving the sum of the two variance components) and empirically (by repeated sampling from different types of populations).

Graubard and Korn (2002) have given inference for super population parameters using sample surveys. Hidiroglou and Sarndal (1998) presented use of auxiliary information for two-phase sampling. In one other research contribution Hidiroglou (2001) discussed similar problem in double sampling. Hu and Schennach (2008) established availability of instruments enables the identification of a large class of non-classical nonlinear errors-in-variables models with continuously distributed variables. Their main assumption for the study was that, conditional on the value of the true regresses, some “measure of location” of the distribution of the measurement error (e.g., its mean, mode, or median) is equal to zero. A convenient sieve based estimator is proposed and the behavior is investigated using monte-carlo simulations.

Singh and Singh (2001) have discussed a general class of estimators for population mean along with their efficiency, under the assumption of the knowledge of population variance. Singh and Agnihotri (2008) presented a general procedure for estimating the population mean through defining a class of estimators. It has been shown to the first degree of approximation that the MSE of the estimator based on the estimated optimum value is same as the minimum MSE of the proposed class of estimators in their study. Singh and Vishwakarma (2006) introduced two classes of estimators of population mean of the study variable using auxiliary variable in stratified random sampling. They shown that the first degree of approximation that the variances of the estimators based on estimated optimum values are same as that of optimum estimators. Singh, Mathur and Chandra (2009) have suggested a chain-type estimator of population variance by making use of a known coefficient of kurtosis of second auxiliary variable in two-phase sampling and they have shown their estimator better than usual estimators under some realistic conditions. Likewise, Singh et al. (2009) considered a family of estimators of population means of the main variable under study.
using information on two auxiliary variables under simple random sampling without replacement (SRSWOR). Singh and Singh (1999) defined a class of unbiased estimators for population mean in cluster sampling and studied its properties.


Ahmed et al. (2006) has been studied several methods of imputation and their corresponding estimators of the population mean are considered. They compared these estimators with each other and with other well known estimators using their biases and mean squared errors. They found that some of the new estimators are more efficient than the well known estimators. Brick et al. (2004) has a discussion on variance estimation with hot deck imputation using a model.

Dalabehara and Sahoo (2000) presented a new unbiased estimator for the population mean when the population mean of the main auxiliary variable is unknown but the emphasis is laid on the use of an additional auxiliary variable. Diana and Tommasi (2003) explained optimal estimation for finite population mean in two-phase sampling. A general class of estimators for finite population mean is proposed and
minimum variance bound for estimator in the class is provided. As a conclusion they have shown that proposed estimators can reach the minimum variance bound,

Kim and Fuller (2004) discussed fractional hot deck imputation. They suggested a consistent replication variance estimation procedure for estimators computed with fractional imputation. Kim and Sitter (2003) investigated variance estimation for the regression estimator for a two-phase sample. A replication variance estimator with number of replicates equal to or slightly larger than the size of the second-phase sample is developed by them with the conclusion that proposed method can be directly applied to variance estimation for the double expansion estimator and the re-weighted expansion estimator. Kim et al. (2006) constructed multiple-imputation variance estimator for data that are collected with a complex sample design. The bias-adjusted variance estimator is suggested by them. Kim et al. (2009) considered variance estimation for nearest neighbor imputation for U.S. Census long form data in his doctoral thesis.

Mathiowitz (2000) suggested the effect of length of recall on the quality of survey data. Micklewright and Schnepf (2006) performed a study on response bias where country England is taken into consideration. Montaquila and Jernigan (1997) have useful contribution on variance estimation in the presence of imputed data. Muñoz and Rueda (2009) proposed a novel imputation method based on quintiles, which can be implemented with or without the presence of auxiliary information. They extended the method to unequal sampling designs and non-uniform response mechanisms.

1.19.3. RESEARCH CONTRIBUTION OVER MEASUREMENT ERROR

Allen et al. (2003) has a work on a family of estimators of population mean using multiauxiliary information in presence of measurement errors. Al-Nasser (2003) has a vital contribution on customer satisfaction measurement models. Al-Nasser (2005) derived entropy type estimator to simple linear measurement error models. He stated that the classical maximum likelihood estimation fails to estimate the simple linear measurement error model, with or without equation error, unless additional assumptions are made about the structural parameters.

Binder and Roberts (2003) advised design-based and model-based methods for estimating model parameters. Black et al. (2000) has an elaborate discussion on bounding parameter estimates with non-classical measurement error. They considered the case of a noisily measured variable with a negative covariance between the measurement error and the true value of the variable. They identified point estimates using a method-of-moments framework and then extended their bounding results to simple multivariate models with measurement error. Moreover, they also illustrated the potential usefulness of correcting for measurement error as a complement to other approaches.


1.19.4 AUTHOR'S CONTRIBUTION

Singh and Shukla (1987) suggested a strategy for one parameter family of factor-type estimator and extended the approach for application to the case of negative correlation by Shukla et al. (1991), Singh and Shukla (1993). An elaborate discussion on super-population model approach is due to Singh et al. (1993) and chaining technique application over F-T estimator is due to Singh et al. (1994).

population. In another contribution Shukla and Trivedi (2001) studied an estimation strategy using the post-stratified sampling scheme.

Shukla and Dubey (2000) tackled the complete non-response in one of their useful contribution. Shukla and Dubey (2001) presented a general class of PSNR type sampling scheme by introducing three groups of earmarked strata based on response pattern along-with two parameters of the class. Shukla et al. (2002) presented an estimator in a post stratified set-up of sampling assuming prior knowledge of Population Proportion of Mean Matrix (PPM) and coefficients of variation of strata. The method of choice of weights for combining post-stratified sample means is proposed along with their optimum selection by them.

Shukla (2002) proposed some estimation strategies for estimating the population mean under this set up. The Factor-Type (F-T) estimator is considered throughout with the derivation of properties of suggested estimators. Shukla, Dubey, Trivedi (2003, 2004) have a useful contribution in sampling for small area estimation and non-response in mail surveys.