CHAPTER 1

# Collection of Survey Data

## 1.1 NEED FOR STATISTICAL DATA

The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conversation and controversy require numerical data for their resolution. Data resulting from the physical, chemical, and biological experiments in the form of observations are used to test different theories and hypotheses. Various social and economic investigations are carried out through the use and analysis of relevant data. The data collected and analyzed in an objective manner and presented suitably serve as basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations, and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment, meteorology, and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country, is of importance in shaping its policies in respect of wages and price levels. Data on health, mortality, and population could be used for formulating policies for checking population growth. Similarly, information on forestry and environment is needed to plan strategies for a cleaner and healthier life. Agricultural production data are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labor, cost and quality of production, stock, and demand and supply positions for proper planning of production levels and sales campaigns.

The research institutions need data to verify the earlier findings or to draw new inferences. The data are used by public organizations to assess the state policies, and to point it out to the administration if these are not up to the expectations of the people. The international organizations collect data to present comparative positions of different countries in respect of economy, education, health, culture, etc. Besides, they also use it to frame their policies at the international level for the welfare of people.

## 1.2 TYPES OF DATA

The collection of required information depends on the nature, object, and scope of study on the one hand and availability of financial resources, time, and man power on the other. The statistical data are of two types: (1) primary data, and (2) secondary data.

**Definition 1.1** The data collected by the investigator from the original source are called *primary data.*

**Definition 1.2** If the required data had already been collected by some agencies or individuals and are now available in the published or unpublished records, these are known as *secondary data.*

Thus, the primary data when used by some other investigator/agency become secondary data. There could be large number of publications presenting secondary data. Some of the important ones are given below:

1. Official publications of the federal, state, and local governments.
2. Reports of committees and commissions.
3. Publications and reports of business organizations, trade associations, and chambers of commerce.
4. Data released by magazines, journals, and newspapers.
5. Publications of different international organizations like United Nations Organization, World Bank, International Monetary Fund, United Nations Conference on Trade and Development, International Labor Organization, Food and Agricultural Organization, etc.

Caution must be exercised in using secondary data as they may contain errors of transcription from the primary source.

## 1.3 METHODS OF COLLECTING PRIMARY DATA

There are variety of methods that may be used to collect information. The method to be followed has to be decided keeping in view the cost involved and the precision aimed at. The methods usually adopted for collecting primary data are: (1) direct personal interview, (2) questionnaires sent through mail, (3) interview by enumerators, and (4) telephone interview.

### 1.3.1 *Direct Personal Interview*

In this, the investigator contacts the respondents personally and interviews them. The interviewer asks the questions pertaining to the objective of survey and the information, so obtained, is recorded on a *schedule* ( a questionnaire form) already prepared for the purpose. Under this method, the response rate is usually good, and the information is more reliable and correct. However, more expenses and time is required to contact the respondents.

### 1.3.2 *Questionnaires Sent Through Mail*

In this method, also known as *mail inquiry*, the investigator prepares a questionnaire and sends it by mail to the respondents. The respondents are requested to complete the questionnaires and return them to the investigator by a specified date. The method is suitable where respondents are spread over a wide area. Though the method is less expensive, normally it has a poor response rate. Usually, the response rate in mail surveys has been found to be about 40 percent. The other problem with this method is that it can be adopted only where the respondents are literate and can understand the questions. They

should also be able to send back their responses in writing. The success of this method depends on the skill with which the questionnaire is drafted, and the extent to which willing co-operation of the respondents is secured.

### 1.3.3 *Interviews by Enumerators*

This method involves the appointment of enumerators by the surveying agency. Enumerators go to the respondents, ask them the questions contained in the schedule, and then fill up the responses in the schedule themselves. For example, the method is used in collecting information during population census. For success of this method, the enumerators should be given proper training for soliciting co-operation of the respondents. The enumerators should be asked to carry with them their identity cards, so that, the respondents are satisfied of their authenticity. They should also be instructed to be patient, polite, and tactful. This method can be usefully employed where the respondents to be covered are illiterate.

### 1.3.4 *Telephone Interview*

In case the respondents in the population to be covered can be approached by phone, their responses to various questions, included in the schedule, can be obtained over phone. If long distance calls are not involved and only local calls are to be made, this mode of collecting data may also prove quite economical. It is, however, desirable that interviews conducted over the phone are kept short so as to maintain the interest of the respondent.

Payne (1951) and Hyman (1954) have made detailed study of various methods of data collection and associated problems. The books by Murthy (1967) and Des Raj (1968) also contain comments of relevance.

## 1.4 FRAMING OF QUESTIONNAIRE/SCHEDULE

The *questionnaire* is a channel through which the needed information is elicited. The success of eliciting information, to a considerable extent, depends on the tactful drafting of the questionnaire (or the schedule). The way in which questions are presented affects the quality of response. It is, therefore, important to ensure that not only the right questions are asked but also that they are asked in the right way. This aspect has been dealt with in detail by Murthy (1967) and Des Raj (1968).

Persons framing the questionnaire need to have detailed knowledge of the field of inquiry. While preparing and mailing the questionnaire, following points should be kept in mind:

1. The person conducting the survey must introduce himself and state the objective of the survey. For this purpose, a short letter conveying how the respondent would be benefitted from the survey being conducted, should be enclosed. Also, the enclosing of a self-addressed stamped envelope for the respondent's convenience in returning the questionnaire, will help in improving the response rate.
2. The questions forming the questionnaire/schedule should be clear, unambiguous, and to the point. Vague questions do not bring forth clear and correct answers. As far as possible, questions should be made capable of objective answers. The

language used in the questions should be easy to understand, and the technical terms used should be properly defined.

3. Questions affecting prestige and sentiments of the people and those involving calculations should be avoided while framing the questionnaire. The order of questions should be relevant to generate a logical flow of thought in the minds of respondents. These should not skip back and forth from one topic to another. It will facilitate the answering of each question in turn. It is always advisable to start with simplest questions.
4. Precise and definite instructions for filling the questionnaire and about units of measurement should also be given.
5. The questionnaire should not be lengthy, otherwise, the respondents begin to loose interest in answering them. On the other hand, no important item should be left uncovered.
6. The outlook of questionnaire should be attractive. The printing and the paper used should be of good quality. Sufficient space should be left for answers depending on the type of questions. A pretest of the questionnaire with a group, before its actual use, helps to discover the shortcomings therein. There may be ambiguous questions, the ordering of questions may require change, and some questions may have to be asked in alternate forms. This gives an opportunity to improve the questionnaire in the light of tryout.

## 1.5 SOME TECHNICAL TERMS

In order to define few technical terms which will be used in the book quite frequently, we consider an example. Let us assume that we wish to find out the proportion of votes a particular political party A, is expected to get in an election in a particular constituency.

**Definition 1.3** An *element* is a unit for which information is sought.

In the example considered above, the element will be a registered voter of the constituency. The study variable in this case will be voter's preference for the party A. The variable will be measured as 1 if the voter prefers to vote for party A, otherwise, the measurement will be taken as zero.

**Definition 1.4** The *population* or *universe* is an aggregate of elements, about which the inference is to be made.

For the example considered above, population will be the collection of all registered voters of the constituency. It should be noted here that the same population will have different set of measurements for a different study variable.

Populations are called *finite* or *infinite*, depending on the number of units constituting it. The population of registered voters in the above example is finite. Whereas, the populations like water in a tank or a sheet of metal could be considered as infinite populations of their respective molecules. In sample surveys, we shall usually deal with

finite populations. The results for infinite populations could, however, be used for finite populations with very large number of units.

**Definition 1.5** *Sampling units* are nonoverlapping collections of elements of the population.

As pointed out earlier, a registered voter is an element in the above example. However, due to convenience or cost considerations, one could sample households in the constituency in place of registered voters, and ask for the preferences of all the registered voters in the sample households. In such a situation, household will be the sampling unit and it may be noted that the number of elements in any sampling unit could be zero, one, or more depending on the number of registered voters in any particular sampled household. If each sampling unit contains one element of the population, then both sampling unit and element are identical.

**Definition 1.6** A list of all the units in the population to be sampled is termed *frame* or *sampling frame.*

If individual voter is taken as the sampling unit then a list of all registered voters will constitute the frame. On the other hand, if the households are taken as sampling unit then the list of all households, obtained after properly arranging the list of households in different villages and towns of the constituency, could serve as frame for the selection of a sample of households.

It may be pointed out that the frame may not include all the sampling units of the population at any particular time as the lists of these units are not updated everyday. If frame is the list of registered voters, it may include some voters who have died now, and might not include the names of persons who became eligible to vote after the list of voters was last prepared.

**Definition 1.7** A subset of population selected from a frame to draw inferences about a population characteristic is called a *sample.*

In practice number of units selected in a sample is much less than the number in the population. Inferences about the entire population are drawn from the observations made on the study variable for the units selected in the sample. In the example considered above, preference for party A will be asked only from the registered voters selected in the sample. This information will then be used to determine the proportion of all votes that party A is expected to get in the election.

## 1.6 NEED FOR A SAMPLE

Collection of information on every unit in the population for the characteristics of interest is known as *complete enumeration* or *census*. The money, manpower, and time required for carrying out a census will generally be large, and there are many situations where with limited means complete enumeration is not possible. There are also instances where it is not

feasible to enumerate all units due to their perishable nature. In all such cases, the investigator has no alternative except resorting to a sample survey.

The number of units (not necessarily distinct) included in the sample is known as the *sample size* and is usually denoted by n, whereas the number of units in the population is called *population size* and is denoted by N. The ratio n/N is termed as *sampling fraction.*

There are certain *advantages of a sample survey* over complete enumeration. These are given below :

### 1.6.1 *Greater Speed*

The time taken for collecting and analyzing the data for a sample is much less than that for a complete enumeration. Often, we come across situations where the information is to be collected within a specified period. In such cases, where time available is short or the population is large, sampling is the only alternative.

### 1.6.2 *Greater Accuracy*

A census usually involves a huge and unwieldy organization and, therefore, many types of errors may creep in. Sometimes, it may not be possible to control these errors adequately. In sample surveys, the volume of work is considerably reduced. On account of this, the services of better trained and efficient staff can be obtained without much difficulty. This will help in producing more accurate results than those for complete enumeration.

### 1.6.3 *More Detailed Information*

As the number of units in a sample are much less than those in census, it is, therefore, possible to observe/interview each and every sample unit intensively. Also, the information can be obtained on more number of variables. However, in complete enumeration such an effort becomes comparatively difficult.

### 1.6.4 *Reduced Cost*

Because of lesser number of units in the sample in comparison to the population, considerable time, money, and energy are saved in observing the sample units in relation to the situation where all units in the population are to be covered.

From the above discussion, it is seen that the sample survey is more economical, provides more accurate information, and has greater scope in subject coverage as compared to a complete enumeration. It may, however, be pointed out here that *sampling errors* are present in the results of the sample surveys. This is due to the fact that only a part of the whole population is surveyed. On the other hand, *nonsampling errors* are likely to be more in case of a census study than these are in a sample survey. Merits and demerits of sample surveys have been discussed in detail by Zarkovich (1961) and Lahiri (1963).

## 1.7 SAMPLING PROCEDURES

The method which is used to select the sample from a population is known as *sampling procedure.* These procedures can be put into two categories - probability sampling and nonprobability sampling. These two types of surveys are not distinguished by the questionnaire and instructions to be followed, but by the methods of selecting the sample for obtaining the estimates of the population characteristics of interest and their precision.

### 1.7.1 *Probability Sampling*

**Definition 1.8** If the units in the sample are selected using some probability mechanism, such a procedure is called *probability sampling.*

This type of survey assigns to each unit in the population a definite probability of being selected in the sample. Alternatively, it enables us to define a set of distinct samples which the procedure is capable of selecting if applied to a specific population. The sampling procedure assigns to each possible sample a known probability of being selected. One can build suitable estimators for different population characteristics for probability samples. For any sampling procedure of this type, one is in a position to develop theory by using probability apparatus. It is also possible to obtain frequency distribution of the estimator values it generates if repeatedly applied to the same population. The measure of the sampling variation can also be obtained for such procedures, and the proportion of estimates that will fall in a specified interval around the true value can be worked out. The procedures such as these will only be considered in this book.

### 1.7.2 *Nonprobability Sampling*

**Definition 1.9** The procedure of selecting a sample without using any probability mechanism is termed as *nonprobability sampling.*

The convenience sampling and the purposive sampling belong to this category. In *convenience sampling,* the sample is restricted to a part of the population that is readily accessible. For example, a sample of coal from an open wagon may be taken from the depth of up to 50 cm from the top. In studies where the process of taking observations is inconvenient, unpleasant, or troublesome to the selected person, only the volunteers may constitute the sample.

*Purposive sampling* (also termed *Judgement sampling*) is common when special skills are required to form a representative subset of population. For instance, auditors often use judgement samples to select items for study to determine whether a complete audit of items may be necessary. Sometimes, quotas are fixed for different categories of population based on considerations relevant to the study being conducted, and selections within the categories are based on personal judgement. This type of sampling procedure is also termed *quota sampling.*

Obviously, these methods are subject to human bias. In appropriate conditions, these methods can provide useful results. They are, however, not amenable to the development of relevant theory and statistical analysis. In such methods, the sampling error can not be objectively determined. Hence, they are not comparable with the available probability sampling methods.

## 1.8 WITH AND WITHOUT REPLACEMENT SAMPLING

**Definition 1.10** In *with replacement* (WR) *sampling,* the units are drawn one by one from the population, replacing the unit selected at any particular draw before executing the next draw.

As the constitution of population remains same at each draw, some units in the with replacement sample may get selected more than once. This procedure gives rise to $N^n$ possible samples when order of selection of units in the sample is taken into account, where N and n denote the population and sample sizes respectively.

**Example 1.1**

Given below are the weights (in pounds) of 4 children at the time of birth in a hospital:

| Child : | A | B | C | D |
|---|---|---|---|---|
| Weight : | 5.5 | 8.0 | 6.5 | 7.0 |

Enumerate all possible WR samples of size 2. Also, write values of the study variable (weight) for the sample units.

**Solution**

Here, N=4 and n=2. There will, therefore, be $4^2=16$ possible samples. These are enumerated below along with the weight values for the units included in the sample.

**Table 1.1** Possible samples along with their variable values

| Sample | Children in the sample | Weight for the sampled children | Sample | Children in the sample | Weight for the sampled children |
|---|---|---|---|---|---|
| 1 | A, A | 5.5, 5.5 | 9 | C, A | 6.5, 5.5 |
| 2 | A, B | 5.5, 8.0 | 10 | C, B | 6.5, 8.0 |
| 3 | A, C | 5.5, 6.5 | 11 | C, C | 6.5, 6.5 |
| 4 | A, D | 5.5, 7.0 | 12 | C, D | 6.5, 7.0 |
| 5 | B, A | 8.0, 5.5 | 13 | D, A | 7.0, 5.5 |
| 6 | B, B | 8.0, 8.0 | 14 | D, B | 7.0, 8.0 |
| 7 | B, C | 8.0, 6.5 | 15 | D, C | 7.0, 6.5 |
| 8 | B, D | 8.0, 7.0 | 16 | D, D | 7.0, 7.0 |

■

**Definition 1.11** In *without replacement* (WOR) *sampling*, the units are selected one by one from the population, and the unit selected at any particular draw is not replaced back to the population before selecting a unit at the next draw.

Obviously, no unit is selected more than once in a WOR sample. If the order of selection of units in the sample is ignored, then there are $\binom{N}{n}$ possible samples for this selection procedure.

**Example 1.2**

Using data of example 1.1, enumerate all possible WOR samples of size 2, and also list the weight values for the respective sample units.

**Solution**

In this case, number of possible samples will be $\binom{4}{2} = 6$. These are enumerated below. Note that no samples like AA or BB appear in the list of possible samples, and also the ordered samples like AB and BA are treated as the same sample.

| Sample | Children in the sample | Weight for the sample children | Sample | Children in the sample | Weight for the sample children |
|---|---|---|---|---|---|
| 1 | A, B | 5.5, 8.0 | 4 | B, C | 8.0, 6.5 |
| 2 | A, C | 5.5, 6.5 | 5 | B, D | 8.0, 7.0 |
| 3 | A, D | 5.5, 7.0 | 6 | C, D | 6.5, 7.0 ■ |

## 1.9 PLANNING AND EXECUTION OF SAMPLE SURVEYS

Sample survey techniques are used widely as an organized and fact finding instrument. The quality of the inferences drawn about the population characteristics from the sample data is related to, how well, the sample represents the population. It requires to select a suitable sampling plan, and implement it in a way that ensures the sample to be a good representative of the population under study. It is, therefore, essential to describe briefly the steps involved in the *planning* and *execution* of a survey. Surveys vary greatly in their scope and complexity. Problems that are baffling in one survey, may be trivial or nonexistent in another. Some of the important aspects requiring attention at the planning stage are grouped under the following heads:

### 1.9.1 *Objectives*

The first task is to lay down, in concrete terms, the objectives of the survey. The investigator should ensure that these *objectives* are commensurate with available resources in terms of money, manpower, and the time limit specified for the survey.

### 1.9.2 *Population to be Studied*

The *population* to be covered by the survey should be clearly defined. An exact description should be given of the geographical region and the categories of the material to be covered by the survey. For instance, in a survey of human population, it is necessary to specify whether such categories as hotel residents, institutions, military personnel, etc., were to be included or not.

Population to be sampled should coincide with the *target population* about which inferences are to be drawn. However, sometimes impracticability and inconvenience may result in the leaving out of certain segments of the population from the scope of the survey. If so, the conclusions drawn will apply only to the *population* actually *sampled.* Any supplementary information gathered for the omitted sectors, which can throw some light on the subject matter of the survey, will be useful.

### 1.9.3 *Sampling Unit*

The population should be capable of being divided into *sampling units*, and these should be properly defined. For example, a human population can be considered to be built up of villages, localities, households, persons, etc. The division of population into sampling units should be unambiguous. Every element of the population should correspond to just one and only one sampling unit. The border line cases can be handled by framing some appropriate rules.

### 1.9.4 *The Sampling Frame*

In surveys, as already discussed, it is always desired that the sampled and the target population should coincide. It should, therefore, be ensured that all the sampling units of the population under study are included in the frame. The frame should be up to date and free from errors of omission and overlapping.

### 1.9.5 *Sample Selection*

The size of the sample and manner of selecting the sample should receive careful attention. After taking various technical, operational, and risk factors into consideration, an optimum size of the sample and sampling procedure need to be decided upon. While doing so, the aim of achieving either a given degree of precision with a minimum cost, or the maximum precision with a fixed cost, should be kept in mind. It should also be ensured that the sample is representative of the population.

### 1.9.6 *Methods of Collecting Information*

After a careful examination of the frame, the method of sample selection, available resources, and the objectives of survey, one should decide about the type of data to be used, that means, whether to collect primary data or to use secondary data. In case the primary data are to be collected, the investigator should decide whether data are to be collected by personal face-to-face interview, by mail, through enumerators, or by telephone interview. These methods have already been briefly discussed in section 1.3.

### 1.9.7 *Handling of Nonresponse*

Procedures should be devised to deal with the respondents, who do not give information by choice, or are not found at home . The reason for nonresponse should also be ascertained. This helps in assessing the effect of refusals and random nonresponse on the conclusions to be drawn.

### 1.9.8 *Pilot Survey*

Where some prior information about the nature of population under study, and the operational and cost aspects of data collection and analysis, is not available from past surveys, it is desirable to design and carry out a *pilot survey*. It will be useful for : (1) discovering shortcomings in the questionnaire/ schedule, (2) evolving suitable strategies for field and analysis work, and (3) training the staff for the purpose.

### 1.9.9 *Organization of Field Work*

Different aspects of *field work* such as recruitment and training of investigators, and inspection and supervision of field staff should be given due consideration in the light of the prevailing operational conditions. The personnel engaged in the survey must receive training, not only in the purpose of the survey and in the methods of measurement to be employed, but also in the art of eliciting acceptable responses. The investigators should be able to withstand long and arduous travel, sometimes in inhospitable conditions. The work must be adequately supervised, as it is important for the investigator to adhere to procedures and tact in answering the questions raised by respondents. Besides, it will help in resolving unusual or unforeseen problems in the field.

A quality check and editing need to be instituted to make careful review of questionnaires received. It will be valuable in amending the recording errors and deleting the data that are erroneous and superfluous.

### 1.9.10 *Analysis of Data and Preparation of Report*

The stage of analysis of collected data and drawing inferences from a sample is a vital issue, as the results of survey are the backbone of the policies to be framed. The errors creeping in the tabulation and statistical analysis of data should be kept under control.

Last, but not the least, comes the *report writing*. While writing the report, the objectives, the scope, and the subject coverage must be mentioned. It is also essential to clarify the method of data collection, estimation procedure including tabulation and analysis, and cost structure in the report. A brief description of the organizations sponsoring and conducting the survey should also be included. Relevant published papers and reports should be cited for reference. The report should conclude with a summary of findings and suggestions for possible action to be taken. For a report on an actual sample survey, the reader may refer to Des Raj (1968).

Many interesting examples showing the range of applications of the sampling methods in business have been given by Deming (1960) and Slonim (1960).

## LET US DO

1.1 Discuss the statement: "The need to collect statistical information arises in almost every conceivable sphere of human activity."

1.2 Describe briefly each of the following terms:

| | |
|---|---|
| a. Primary data | b. Secondary data |
| c. Mail inquiry | d. Questionnaire/schedule |
| e. Population | f. Census |
| g. Element | h. Sample |
| i. Sampling unit | j. Sampling frame |

1.3 Distinguish between primary and secondary data. Give examples. Which of the two data are more reliable and why ?

1.4 What precautions should one take in making use of published data for further studies ?

1.5 Discuss the methods usually employed in collection of primary data, and state briefly their respective merits and demerits.

1.6 In the situations given below, define universe and indicate which method of data collection - personal interview, mail inquiry, or telephone interview - would you prefer, keeping the cost involved and other relevant factors in view:

a. The investigator is interested in estimating the percent loss caused to wheat crop by hail storm in 500 villages. The elected head of the village (known as *sarpanch*) is the sampling unit.
b. To estimate consumer acceptance of the newly developed car model, before it is introduced in the market.
c. The investigator wishes to estimate the average time taken by a telephone company to attend to the complaints of its customers.
d. A firm wishes to estimate the gas mileage for its newly developed model of car. The addresses of the buyers of all the cars of this model are available in the head office of the firm.

1.7 Differentiate between target and sampled population. What problem arises if two populations are not same ?

1.8 What is a questionnaire/schedule? What are the various considerations one should keep in mind while framing it ?

1.9 Distinguish between a questionnaire and a schedule. What is each used for ?

1.10 A publisher wishes to determine, whether his writers prefer title of their books in bright or dull colors? How would you define population, and what sort of questionnaire would you draft to get an answer to the problem?

1.11 The authorities of a certain university wish to conduct a survey to elicit views of its faculty about introducing five days week in the university. Suggest possible sampling unit and frame for the purpose. Also, draft a suitable questionnaire for use in this survey.

1.12 Discuss relative merits of using a sample survey in relation to a census.

1.13 Distinguish between a sample inquiry and complete enumeration. Under what circumstances can the latter be recommended in preference to sample survey ?

1.14 In the following situations, indicate whether a sample study or a complete enumeration should be undertaken, and why ?

a. A firm wishes to estimate the longevity of 800 electric bulbs manufactured of a new design.
b. A tractor manufacturing company is interested in obtaining information on customer preferences in respect of parasol that protects the driver from the sun and rain.
c. A university has 600 teachers. The administration wishes to determine the acceptability of subscribing to a new group insurance scheme.

1.15 Do you agree that it is possible for sample results to be more accurate than the census results ? If so, explain.

1.16 What is the primary advantage of probability sampling over the nonprobability sampling ? Cite three situations where nonprobability sampling is to be preferred.

1.17 Describe briefly the difference between with and without replacement sampling.

1.18 Consider a population consisting of 6 villages, the areas (in hectares) of which are given below :

| Village : | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Area : | 760 | 343 | 657 | 550 | 480 | 935 |

a. Enumerate all possible WR samples of size 3. Also, write the values of the study variable for the sampled units.

b. List all the WOR samples of size 4 along with their area values.

1.19 The Department of Agriculture desires to estimate yield per hectare of wheat in a district. Describe the various steps that may be involved in the planning and execution of a sample survey for this purpose.

CHAPTER 2

# Elementary Concepts

## 2.1 INTRODUCTION

Knowledge of basic concepts is a prerequisite for an insight into the sample survey designs. Assuming some exposure to elementary probability theory on the part of the reader, we present in this chapter, a rapid review of some of these concepts. To begin with, preliminary statistical concepts including those of expectation, variance, and covariance, for random variables and linear functions of random variables will be defined. The idea of sampling distribution, being basic to the sampling theory, has been briefly explained. The concepts of measure of error, interval estimation, and sample size determination, which are related to sampling distribution, have also been discussed. The chapter concludes with a brief introduction to the sampling and nonsampling errors.

The *notations* to be used in the book will be defined, as far as possible, wherever they first appear. Usually, the uppercase letters will be used to denote the values of variables for population units, whereas lowercase letters will denote the corresponding values for sample units. The population parameters are denoted either by the uppercase letters of English alphabet or by Greek letters. The respective point estimators of these parameters will be denoted either by lowercase English letters, or by putting caps (^) on the corresponding symbols of parameters.

## 2.2 STATISTICAL PRELIMINARIES

Assuming some knowledge of the probability theory on the part of the reader, we briefly review here the concepts of expectation, variance, and covariance.

### 2.2.1 *Expectation*

**Definition 2.1** If x is a random variable which assumes values $x_1, x_2, ..., x_k$ with probabilities $p_1, p_2, ..., p_k$ respectively with $\Sigma\ p_i = 1$, i= 1,2, ..., k, then *expectation* of the variable x is defined as

$$E(x) = \sum_{i=1}^{k} p_i x_i \qquad (2.1)$$

In physical sense, expected value of the random variable x stands for the center of gravity of the probability distribution, where a probability mass $p_i$ is located at the point $x=x_i$, i=1, 2, ..., k.

**Example 2.1**
A fair die is rolled once. If x denotes the number on the upper face of the die, find expected value of x.

**Solution**
Here, the random variable x is the number on the upper face of the die. Thus, x can take any one of the values 1, 2,...,6, each with probability 1/6. Hence using (2.1),

$$E(x) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6)$$

$$= 3.5 \blacksquare$$

We now state some more useful related results.

**Result 2.1** Let a and b be two constants and x a random variable, then

$$E(ax+b) = aE(x)+b$$

**Result 2.2** If $x_1, x_2,...,x_k$ are random variables, and $a_1, a_2,..., a_k$ are constants, then

$$E(a_1x_1+a_2x_2+...+a_kx_k) = a_1E(x_1)+a_2E(x_2)+...+a_kE(x_k)$$

**Result 2.3** If $x_1,x_2,...,x_k$ are k mutually independent random variables, then

$$E(x_1.x_2...x_k) = E(x_1).E(x_2)...E(x_k)$$

**Result 2.4** Let x and y be two random variables which are not independent, then

$$E(xy) = E_1[xE_2(y)]$$

where $E_2$ is the conditional expectation of y for a given value of x, and $E_1$, the expectation over all values of x.

### 2.2.2 *Variance and Covariance*

If x is a random variable and E(x) its expected value, then variance of x is defined as

$$\left.\begin{aligned} V(x) &= E[x-E(x)]^2 \\ &= E(x^2)-[E(x)]^2 \end{aligned}\right] \qquad (2.2)$$

When the investigator is interested in linear dependence of pairs of random variables, such as income x and expenditure y, then it is indicated by a measure called *covariance* of x and y. This term is defined as

$$Cov(x,y) = E[\{x-E(x)\}\{y-E(y)\}]$$

A zero value of the covariance indicates no linear dependence between x and y.

**Example 2.2**
For the experiment given in example 2.1, determine the variance of random variable x.

**Solution**
First we work out the term $E(x^2)$. This will be

$$E(x^2) = \frac{1}{6}(1)^2 + \frac{1}{6}(2)^2 + \frac{1}{6}(3)^2 + \frac{1}{6}(4)^2 + \frac{1}{6}(5)^2 + \frac{1}{6}(6)^2$$
$$= \frac{91}{6}$$
$$= 15.167$$

On substituting in (2.2) the value of $E(x^2)$ obtained above, and of $E(x)$ from example 2.1, one gets

$$V(x) = 15.167 - (3.5)^2$$
$$= 2.917 \blacksquare$$

### 2.2.3 *Variance of Linear Functions of Random Variables*

In the theory of sample surveys, the investigator often requires the variance of a linear function of random variables to determine the amount of error in the estimator. Following result will be helpful in such a case.

**Result 2.5** Let $x_1, x_2,...,x_k$ be k random variables, then

$$V\left(\sum_{i=1}^{k} a_i x_i\right) = \sum_{i=1}^{k} a_i^2 V(x_i) + 2\sum\sum_{i<j} a_i a_j \text{Cov}(x_i, x_j)$$

where $a_1, a_2,..., a_k$ are constants. The double summation taken over all pairs $(x_i, x_j)$, with $i < j$, will involve $k(k-1)/2$ covariance terms.

For independence of random variables $x_i$ and $x_j$, $\text{Cov}(x_i, x_j)=0$. Thus, result 2.5 gets simplified in case of mutually independent random variables as

$$V\left(\sum_{i=1}^{k} a_i x_i\right) = \sum_{i=1}^{k} a_i^2 V(x_i)$$

The above expression holds even when the random variables $x_1, x_2,..., x_k$, are mutually uncorrelated. The reader should note, that the condition of no correlation is much less stringent than the condition of independence.

## 2.3 ESTIMATOR AND ITS SAMPLING DISTRIBUTION

The ultimate objective of any sample survey is to make inferences about a population of interest. Such inferences are based on information contained in a sample selected from that population. The investigator usually aims at the estimation of certain unknown features of the population. These population characteristics are called parameters.

**Definition 2.2** Any real valued function of variable values for all the population units is known as a *population parameter* or simply a *parameter.*

For any given variable, the population value of a parameter is constant. Some of the important parameters frequently required to be estimated in surveys are total, mean, proportion, and variance. For instance, if $Y_1, Y_2, ..., Y_N$ are the values of the variable y for the N units in the population, then

$$\text{Population mean} = \bar{Y} = \frac{Y_1 + Y_2 + ... + Y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{2.3}$$

$$\text{Population variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right) \tag{2.4}$$

**Definition 2.3** A real valued function of variable values for the units in the sample is called a *statistic.* If it is used to estimate a parameter, it is termed as *estimator.*

The particular value taken by the estimator for a given sample, is known as *estimate* or *point estimate.* For instance, the mean for a given sample, provides an estimate of population mean. The *sample mean* $\bar{y}$ and *sample mean square* $s^2$ for a sample of size n, are respectively given by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.5}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \tag{2.6}$$

where $y_1, y_2, ..., y_n$ are the values of the variable y for the n units in the sample. Note that the *standard deviation* is the positive square root of the variance. It will, therefore, be denoted by $\sigma$ for the population.

The probability mechanism underlying the process of sample selection usually gives rise to different samples. The estimates based on sample observations may differ from sample to sample, and also from the true value of the parameter. The estimator, therefore, is a *random variable.* It leads us to the concept of sampling distribution.

**Definition 2.4** For a given population, sampling procedure, and sample size, the array of possible values of an estimator each with its probability of occurrence, is the *sampling distribution* of that estimator.

**Example 2.3**

Four cows in a household marked A, B, C, and D respectively yield 5.00, 5.50, 6.00, and 6.50 kg of milk per day. Obtain the sampling distribution of average milk yield based on samples of n=2 cows, when the cows are selected with equal probabilities and WR. The procedure for drawing a sample has been explained in chapter 3.

**Solution**

Here N=4 and n=2. The number of possible samples in this case will be $4^2=16$. The mean of each sample, along with the cows included in the sample is given in table 2.1.

**Table 2.1** Means of all possible samples

| Sample | Cows in the sample | Sample mean $\bar{y}$ | Sample | Cows in the sample | Sample mean $\bar{y}$ |
|---|---|---|---|---|---|
| 1 | A, A | 5.00 | 9 | C, A | 5.50 |
| 2 | A, B | 5.25 | 10 | C, B | 5.75 |
| 3 | A, C | 5.50 | 11 | C, C | 6.00 |
| 4 | A, D | 5.75 | 12 | C, D | 6.25 |
| 5 | B, A | 5.25 | 13 | D, A | 5.75 |
| 6 | B, B | 5.50 | 14 | D, B | 6.00 |
| 7 | B, C | 5.75 | 15 | D, C | 6.25 |
| 8 | B, D | 6.00 | 16 | D, D | 6.50 |

From the above table, we get the following sampling distribution :

**Table 2.2** Distribution of sample mean

| Sample mean ($\bar{y}$) | Frequency (f) | Probability (p) |
|---|---|---|
| 5.00 | 1 | .0625 |
| 5.25 | 2 | .1250 |
| 5.50 | 3 | .1875 |
| 5.75 | 4 | .2500 |
| 6.00 | 3 | .1875 |
| 6.25 | 2 | .1250 |
| 6.50 | 1 | .0625 |
| Total | 16 | 1 |

The sampling distribution for any estimator can be used for finding out probabilities for various statements about the values taken by the estimator. In this particular case, one can easily verify that the probability P ($\bar{y} \geq 6$) = 6/16. Similarly, P ( $\bar{y} < 5.25$) =1/16, or P ($5.25 \leq \bar{y} \leq 5.75$) = 9/16. ■

In this case, we had considered a population of just four cows (units). However, in practice the number of units in the population will generally be quite large, and so will be the number of units to be selected in the sample. This will make the number of possible samples, and consequently the number of values taken by the sample mean (estimator), very large. In such cases, the enumeration of all possible samples, and also of the values taken by the estimator, will be quite difficult. One will, therefore, not be in a position to easily determine the exact sampling distribution of an estimator.

Also in the example considered above, all possible samples had equal chance of being selected. In practice, however, there may be situations where different samples will have different probabilities of selection. To find out exact probabilities for statements regarding the estimator values in such cases, one will be required to add the probabilities of selection of all those samples which yield estimator values satisfying the statement. This may not be an easy job to do in most of the real life situations. Exact sampling distribution of estimators in all such cases may, therefore, have to be approximated by some known continuous probability density functions to make the calculation of probabilities for various statements about the estimator value, easier.

For continuous random variables, if the study variable is normally distributed in the population, then the distribution of sample mean is exactly normal for any sample size. However, if the study variable is not normally distributed in the population, the distribution of sample mean approaches *normal distribution* as the sample size n is increased.

## 2.4 UNBIASED ESTIMATOR

In survey sampling, a good estimator is expected to have mainly two properties. One of these properties is known as unbiasedness. The other one, which we consider in the next section, is the closeness of the values taken by the estimator for different possible samples, to the actual unknown value of the parameter.

For discussion in the remaining part of this chapter, we shall denote the population parameter in general by $\theta$, whereas its estimator will be denoted by $\hat{\theta}$. In the preceding section, we have observed that the estimator $\hat{\theta}$ is a random variable, and it takes different values for different possible samples that can be selected from the population under consideration. The sample selection procedure generates a sampling distribution for $\hat{\theta}$. The unbiasedness of the estimator $\hat{\theta}$ is then defined as follows :

**Definition 2.5** The estimator $\hat{\theta}$ is said to be *unbiased* for the parameter $\theta$, if $E(\hat{\theta}) = \theta$.

If the sample selection procedure is such that all possible samples are equally likely to get selected, $E(\hat{\theta})$ becomes the simple average of the values that the estimator $\hat{\theta}$ takes

for different possible samples which can be selected from the population under study. However, if the probabilities of selection for the different possible samples are not equal, then E ($\hat{\theta}$) will be the weighted average (weights being the probabilities of selection for different possible samples) of the values of $\hat{\theta}$. Thus unbiasedness of an estimator $\hat{\theta}$ ensures, that on the average it will take value equal to the unknown population parameter $\theta$, although for most of the samples, the values taken by $\hat{\theta}$ will be either less or more than $\theta$.

**Example 2.4**

For the data in example 2.3, verify whether the sample mean based on n=2 cows is an unbiased estimator for the average milk yield in the population ? Assume that the cows in the sample are selected with equal probabilities and with replacement.

**Solution**

In table 2.1 are given the sample mean values for 16 possible samples of size n=2 that can be selected from a population of size N=4 with equal probabilities and with replacement. Here, sampling procedure ensures that all these samples have same chance of being selected. Hence, expected value of sample mean $\bar{y}$ will be the simple average of 16 possible sample mean values. Thus,

$$E(\bar{y}) = \frac{1}{16}(5.00+5.25+...+6.50)$$

$$= \frac{92.00}{16}$$

$$= 5.75$$

Using (2.3), it can be easily seen that the population mean $\bar{Y}$ (the parameter $\theta$ in this case) is also equal to 5.75. Hence, the sample mean $\bar{y}$ is unbiased for the population mean $\bar{Y}$. ■

In case where E ($\hat{\theta}$) is not equal to the value of population parameter $\theta$, the estimator $\hat{\theta}$ is said to be a biased estimator.

**Definition 2.6** If for an estimator $\hat{\theta}$, $E(\hat{\theta}) \neq \theta$,the estimator $\hat{\theta}$ is called a *biased estimator* of $\theta$. The magnitude of the bias in $\hat{\theta}$ is given by

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

The ratio

$$RB(\hat{\theta}) = \frac{B(\hat{\theta})}{\theta}$$

is called the *relative bias* of the estimator $\hat{\theta}$.

## 2.5 MEASURES OF ERROR

As discussed in section 2.4, it is not sufficient that an estimator be unbiased for it to qualify as a good estimator. In addition to the property of unbiasedness, the estimator should also have small sampling variance. As pointed out earlier, the value of the estimator $\hat{\theta}$ may differ from sample to sample and also from its parameter value $\theta$. The differences $(\hat{\theta}_s - \theta)$ and $[\hat{\theta}_s - E(\hat{\theta})]$, $\hat{\theta}_s$ being the estimator's value based on the s-th sample, denote the error of the estimate $\hat{\theta}_s$. On pooling such errors for all possible samples that can be selected from the population under consideration, one gets a single measure of error for the sampling situation at hand. Some commonly used measures of this error are presented here.

### 2.5.1 *Sampling Variance*

**Definition 2.7** The *sampling variance* is a measure of the divergence of the estimator values from its expected value. Alternatively, it is the variance of the sampling distribution of an estimator. In the light of (2.1) and (2.2), one gets it as

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$
$$= E(\hat{\theta})^2 - [E(\hat{\theta})]^2$$

The positive square root of sampling variance is termed *standard error* (SE). Thus,

$$SE(\hat{\theta}) = +\sqrt{V(\hat{\theta})}$$

In other words, SE is the standard deviation of the sampling distribution. It is also an important measure of the fluctuations in the estimator values due to specific sampling design.

**Example 2.5**
For the data in example 2.3, compute the sampling variance and standard error of sample mean for the WR equal probability samples of size 2.

**Solution**
The sampling distribution of mean $\bar{y}$ (here the estimator $\hat{\theta}$ is the sample mean $\bar{y}$ ) for n=2 has been obtained in example 2.3. Also from example 2.4, $E(\bar{y}) = 5.75$. Thus from table 2.1, we have

$$V(\bar{y}) = \frac{1}{16}[(5.00)^2 + (5.25)^2 + \ldots + (6.50)^2] - (5.75)^2$$

which from table 2.2, is equivalent to

$$V(\bar{y}) = \frac{1}{16}[(5.00)^2(1) + (5.25)^2(2) + ... + (6.50)^2(1)] - (5.75)^2$$
$$= .15625$$

It gives

$$SE(\bar{y}) = \sqrt{.15625}$$
$$= .39528. \blacksquare$$

It should be noted that the sampling variance $V(\hat{\theta})$ and $SE(\hat{\theta})$ are still of no practical use, because their values depend on the study variable values for all the population units which are usually not available in practice. Thus to get an idea about the magnitude of the error involved in $\hat{\theta}$ values, one needs to estimate $V(\hat{\theta})$ and $SE(\hat{\theta})$ from the sample data. Their estimators are respectively denoted by $v(\hat{\theta})$ and $se(\hat{\theta})$. The term $se(\hat{\theta})$, called the *estimate of standard error* of estimator $\hat{\theta}$, is the positive square root of $v(\hat{\theta})$. Thus,

$$se(\hat{\theta}) = + \sqrt{v(\hat{\theta})}$$

### 2.5.2 *Mean Square Error*

In case the estimator is biased, we use mean square error for measuring the variability of sampling distribution.

**Definition 2.8** The *mean square error* (MSE) measures the divergence of the estimator values from the true parameter value. This can be put as

$$MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2$$

The positive square root of MSE is termed as *root mean square error*. The MSE and the sampling variance are related as

$$MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$$

where $B(\hat{\theta})$ is the bias of the estimator $\hat{\theta}$. Thus, for an unbiased estimator , the $MSE(\hat{\theta})$ and the $V(\hat{\theta})$ are equivalent. The above discussion about MSE now enables us to talk about relative efficiency.

**Definition 2.9** If $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of the parameter $\theta$, the *relative efficiency* of the estimator $\hat{\theta}_2$ with respect to the estimator $\hat{\theta}_1$ , is defined as

$$RE = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)} \qquad (2.7)$$

Thus for the estimator $\hat{\theta}_2$ to be more efficient than the estimator $\hat{\theta}_1$, the RE defined above will be more than one. Since the mean square errors, MSE($\hat{\theta}_1$) and MSE($\hat{\theta}_2$), are not known in practice, their respective sample estimates denoted by mse($\hat{\theta}_1$) and mse($\hat{\theta}_2$) are used in their place.

**Example 2.6**

The estimated mean square errors of two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, are 4861.79 and 5258.62 respectively. Estimate percent relative efficiency of estimator $\hat{\theta}_2$ with respect to $\hat{\theta}_1$. Also point out, which of the two estimators is more efficient ?

**Solution**

We have mse($\hat{\theta}_1$)=4861.79 and mse($\hat{\theta}_2$)=5258.62. Through relation (2.7), the estimate of RE of estimator $\hat{\theta}_2$ in relation to $\hat{\theta}_1$ would be

$$RE = \frac{\text{mse}\,(\hat{\theta}_1)}{\text{mse}\,(\hat{\theta}_2)}$$

On making substitutions, one gets

$$RE = \frac{4861.79}{5258.62}$$

$$= .925$$

Percent RE = .925 (100)

= 92.5

As the RE obtained is less than 1, the estimator $\hat{\theta}_2$ is less efficient as compared to the estimator $\hat{\theta}_1$. ■

In addition to the variance and the mean square error, there are two more terms that are sometimes used in literature. These are *accuracy* and *precision* which refer to the deviations of $\hat{\theta}$ values from $\theta$ and $E(\hat{\theta})$ respectively.

**Remark 2.1** Another property of a good estimator, besides unbiasedness and small mean square error, is *consistency*. An estimator $\hat{\theta}$ is said to be a *consistent estimator* of parameter $\theta$, if it approaches $\theta$ with probability tending to unity as the sample size tends to infinity. This definition of consistency, thus, strictly applies to estimators based on samples drawn from infinite populations. In case of finite populations, the estimator $\hat{\theta}$ is said to be a consistent estimator of $\theta$ if it assumes value $\theta$ when the entire population is taken as the sample. An easy way to find out whether any particular estimator $\hat{\theta}$ is a consistent estimator or not, is to find the limit of MSE ($\hat{\theta}$) as the sample size tends to infinity. If this limit is zero, then the estimator is consistent.

## 2.6 CONFIDENCE INTERVALS

The point estimates yielded by different samples are rarely equal to the parameter value. But we can enhance the usefulness of estimation by calculating a specific region, known

as *confidence interval*, in which the population parameter probably lies. We may also be able to give probability of this confidence interval covering the parameter. The confidence interval (also called *interval estimate*) of a parameter $\theta$ is an interval of the form $\hat{\theta}_l \le \theta \le \hat{\theta}_u$. The values of $\hat{\theta}_l$ and $\hat{\theta}_u$, the *lower* and *upper confidence limits* of the interval, depend on the value taken by the estimator $\hat{\theta}$ for a given sample, and also on the sampling distribution of $\hat{\theta}$ (or of a suitable function $h(\hat{\theta})$ of $\hat{\theta}$ for which the sampling distribution is already known). Based on the known sampling distribution of the function $h(\hat{\theta})$, we choose $h_l(\hat{\theta})$ and $h_u(\hat{\theta})$, such that for any specified probability $(1-\alpha)$, where $0 < \alpha < 1$, $P[h(\hat{\theta}) < h_l(\hat{\theta})] = P[h(\hat{\theta}) > h_u(\hat{\theta})] = \alpha/2$ and $P[h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})] = 1-\alpha$. Knowing the form of the function $h(\hat{\theta})$, the double inequality $h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})$ can, through algebraic manipulation, be put as $\hat{\theta}_l \le \theta \le \hat{\theta}_u$. Thus, whenever the double inequality $h_l(\hat{\theta}) \le h(\hat{\theta}) \le h_u(\hat{\theta})$ holds, the inequality $\hat{\theta}_l \le \theta \le \hat{\theta}_u$ also holds. Hence, the probability $P(\hat{\theta}_l \le \theta \le \hat{\theta}_u)$ is also equal to $(1-\alpha)$. Such an interval $\hat{\theta}_l \le \theta \le \hat{\theta}_u$, computed from a particular sample, is known as $(1-\alpha)100$ percent confidence interval for $\theta$. The probability $(1-\alpha)$ is called the *confidence coefficient* for the interval.

To illustrate, let us assume that the estimator $\hat{\theta}$ is normally distributed with mean $\theta$ and known variance $V(\hat{\theta})$. The variance of the estimator $\hat{\theta}$ may be known from some previous survey, or from a pilot study. The simple function $Z=(\hat{\theta}-\theta)/\sqrt{V(\hat{\theta})}$ of $\hat{\theta}$,will then be following standard normal distribution which is extensively tabulated. An abridged version of the tables of probabilities, for this distributon, are given in appendix A. From the standard normal probability tables, the value of $Z_{\alpha/2}$ that satisfies the relation $P(Z<-Z_{\alpha/2}) = P(Z>Z_{\alpha/2}) = \alpha/2$, can be easily determined for any value of $\alpha$. If $\alpha=.05$, it is then easily seen that $Z_{\alpha/2}=Z_{.025}=1.96$, since $P(Z < -1.96) = P(Z > 1.96) = .025$. Thus, $P[-1.96 \le Z= (\hat{\theta}-\theta)/\sqrt{V(\hat{\theta})} \le 1.96]=1-.05 =.95$. This probability statement is equivalent to $P[\hat{\theta}-1.96\sqrt{V(\hat{\theta})} \le \theta \le \hat{\theta}+1.96\sqrt{V(\hat{\theta})}]=.95$. Therefore,$\hat{\theta}_l = \hat{\theta}-1.96\sqrt{V(\hat{\theta})}$ and $\hat{\theta}_u = \hat{\theta}+1.96\sqrt{V(\hat{\theta})}$, and the confidence interval $[\hat{\theta}-1.96\sqrt{V(\hat{\theta})}, \hat{\theta}+1.96\sqrt{V(\hat{\theta})}]$ has the confidence coefficient of .95. That means, it will cover the unknown population parameter $\theta$ with probability .95. Using the standard normal probability tables, such confidence intervals can be similarly obtained for any other value of $\alpha$. These confidence intervals will have confidence coefficient equal to $(1-\alpha)$ when the sampling distribution of the known function $h(\hat{\theta})$ of $\hat{\theta}$ is exactly normal, and it will be approximately $(1-\alpha)$ in other cases.

In case $V(\hat{\theta})$ is not already known, it has to be estimated from the sample data. Let $v(\hat{\theta})$ denote the estimator of $V(\hat{\theta})$. If in the function of $\hat{\theta}$ considered above, $V(\hat{\theta})$ is replaced by $v(\hat{\theta})$, the resulting statistic $(\hat{\theta}-\theta)/\sqrt{v(\hat{\theta})}$ will no longer have standard normal distribution as its sampling distribution. We shall, therefore, have to first determine the sampling distribution and then use it for obtaining confidence interval for $\theta$. Further, the sampling distribution of an estimator differs from estimator to estimator. It also depends on the method of sample selection, sample size, and the distribution of the study variable in the population. It is, therefore, not possible to specify a single sampling distribution for the construction of confidence intervals for all the situations. It is beyond the scope

of this book to work out sampling distribution for every estimator and every sampling situation. However, central limit theorem (Fisz, 1963) ensures that the sampling distribution for most of the estimators $\hat{\theta}$ can be approximated by normal distribution with mean $E(\hat{\theta})$ and variance $V(\hat{\theta})$, when the sample size is large.

The point to be emphasized here is that many estimators we shall use in the text, will not be precisely following normal distribution. However, from Tchebysheff's theorem (Fisz, 1963), at least 75% of the observations from any probability distribution will be within 2 standard deviations of their mean. For the sake of simplicity and to avoid confusion, we shall, therefore, use multiplier 2 in place of 1.96 for building up confidence intervals. Thus, the confidence interval for $\theta$ will, in general, be given by

$$\hat{\theta} \pm 2\sqrt{v(\hat{\theta})} \tag{2.8}$$

where $v(\hat{\theta})$ is the estimator of the variance $V(\hat{\theta})$. Such confidence intervals shall provide about .95 confidence coefficient for the estimators $\hat{\theta}$ following approximately normal distribution and a confidence coefficient of at least .75 for any other situation.

The numerical value of the confidence interval in (2.8), for any particular case, will be obtained by using the values of the estimator $\hat{\theta}$ and its variance estimator $v(\hat{\theta})$ computed for that situation.

**Example 2.7**

In a survey, the sample mean was computed as 796.3, and the value of the variance estimator came out to be 1016.9. Build up the confidence interval for population mean and interpret the results.

**Solution**

From the statement of the example, $\bar{y}$ =796.3 and $v(\bar{y})$=1016.9. Using relation (2.8), the confidence interval is computed as

$$\bar{y} \pm 2\sqrt{v(\bar{y})}$$

$$= 796.3 \pm 2\sqrt{1016.9}$$

$$= 732.5, 860.1$$

The lower limit 732.5 and the upper limit 860.1, obtained above, provide a reasonable assurance to the investigator that the population mean would lie in the closed interval [732.5, 860.1]. ■

**Remark 2.2** To examine the *effect of bias* in the estimator $\hat{\theta}$ on the confidence intervals for the parameter $\theta$, let us assume that the estimator $\hat{\theta}$ is normally distributed with mean $E(\hat{\theta})$ ( $\theta$) and variance $V(\hat{\theta})$. Thus, the statistic $Z=[\hat{\theta}-E(\hat{\theta})]/\sqrt{V(\hat{\theta})} = [\hat{\theta}-B(\hat{\theta})-\theta]/\sqrt{V(\hat{\theta})}$, where $B(\hat{\theta})$ is the bias of the estimator $\hat{\theta}$, will be normally distributed with mean zero and variance one. Therefore, the interval

$$[\hat{\theta}-B(\hat{\theta})-Z_{\alpha/2}\sqrt{V(\hat{\theta})},\ \hat{\theta}-B(\hat{\theta})+Z_{\alpha/2}\sqrt{V(\hat{\theta})}] \tag{2.9}$$

will cover parameter $\theta$ with probability $(1-\alpha)$. However, the confidence coefficient of the usual confidence interval $[\hat{\theta}-Z_{\alpha/2}\sqrt{V(\hat{\theta})}, \hat{\theta}+Z_{\alpha/2}\sqrt{V(\hat{\theta})}]$, when used in place of the interval in (2.9) for the parameter $\theta$, decreases with the increase in the bias $B(\hat{\theta})$ of the estimator $\hat{\theta}$ (Cochran, 1977).

**Remark 2.3** Throughout the book, we shall be using approximately .95 confidence coefficient level for the purpose of illustration. In case it is desired to have confidence intervals for some other confidence level, coefficient 2 in (2.8) is to be replaced by the corresponding standard normal variable value.

**Remark 2.4** Some of the situations require to state either a lower or an upper *confidence bound* on a population parameter rather than an interval. For instance, the investigator may wish to state that he/she is reasonably confident that the value of the parameter would not be less than a specified value. In this case, he/she intends to place a lower confidence bound on the parameter. Similarly, if he/she wishes to make a statement that the parameter value is not likely to exceed a specified value, he/she is going to set an upper confidence bound. Lower and upper confidence bounds are determined in the same way as the confidence interval is set. The difference is that the experimenter is providing only one end of the confidence interval. Thus the Z value to be used gets changed, since in this case one does not split $\alpha$ to $\alpha/2$. However, so far as the discussion in this book is concerned, we shall usually confine ourselves to confidence intervals only.

## 2.7 SAMPLE SIZE DETERMINATION

One of the important aspects in planning a sample survey is to decide about the size of the sample required for estimating the population parameter with a specified precision. The maximum difference between the estimate and the parameter value that can be tolerated on considerations of loss or gain due to policy decisions based on the sample results is termed as *permissible error, tolerable error*, or the *bound on the error of estimation*. Once the permissible error has been specified, the next objective is to determine a sample size that meets these requirements. Since the amount of error differs from sample to sample, the margin of error is specified by the probability statement

$$P[\,|\,\hat{\theta}-\theta\,|\,<B] = 1-\alpha \qquad (2.10)$$

where $(1-\alpha)$ may be taken as 95%, 99%, or some other desired level of confidence, and B is the permissible error. Sometimes, the permissible error is specified in terms of percentage of the value of parameter $\theta$. Such a specification of permissible error can, however, be easily converted into a statement of type (2.10). Theoretically, the required sample size is then determined by equating half width of the confidence interval to the permissible error B, and solving the resulting equation for the sample size n. An analogous approach, for determining the sample size, is followed in case of categorical data. Throughout this book, we shall determine the required sample size by following a two step approach proposed by Stein (1945) and Cox (1952). In the first step, we shall select a small preliminary sample of size $n_1$. Observations made on the units selected in this sample, will be used to estimate various parameters involved in the expression for the

half width of the confidence interval. After replacing the parameters by their respective estimates obtained from the preliminary sample, half width of the confidence interval will be equated to the permissible error B. The equation is then solved for n, the required sample size. If $n>n_1$, then $(n-n_1)$ additional units are selected, which along with the preliminary sample yield a pooled sample of n units. If $n<n_1$, no more units are selected and preliminary sample is taken as the final sample.

As the expression for estimator of variance ( or mean square error) to be used in the confidence interval also varies with the sampling schemes, formulas for determining sample size will change with the sampling procedure. These formulas will, therefore, be presented separately for each sampling and estimation situation.

## 2.8 SAMPLING AND NONSAMPLING ERRORS

The probability mechanism inherent in the sampling procedure usually selects different units in different samples. The estimates based on the sample observations, as already discussed, will, therefore, differ in general from sample to sample and also from the value of the parameter under consideration. The resultant discrepancy between the sample estimate and the population parameter value is the error of the estimate. Such an error is inherent and unavoidable in any and every sampling scheme, and is termed *sampling error.* This error, however, has the favorable characteristic of being controllable through the size and design of the sample. This kind of error usually decreases with increase in sample size, and shall theoretically become nonexistent in case of complete enumeration. In many situations, the decrease is inversely proportional to the square root of sample size (figure 2.1).

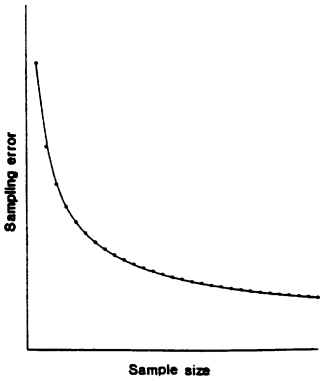


Fig 2.1 Relationship between sampling error and sample size

In any survey study, besides sampling error, there are also errors arising due to defective sampling procedures, ambiguity in definitions, faulty measurement techniques, mistakes in recording, errors in coding-decoding, tabulation and analysis, etc. These errors are known as *nonsampling errors.* For instance, data may be wrongly fed to the computer, or decimal places may be inadvertently changed. If the analysis of data requires transformation of per acre yield to per hectare basis, the multiplier used might be wrong. Sometimes, the respondent may also deliberately respond erroneously while reporting

his answer. A statistician must be aware of this source of error. Unlike the sampling errors, the nonsampling errors are likely to increase with increase in sample size. It is quite possible that nonsampling errors in a complete enumeration survey are greater than both the sampling and nonsampling errors taken together in a sample survey. One should, therefore, be careful in evaluating and checking the processing of the sample data right from its collection to its analysis to minimize the occurrence of nonsampling errors.

## LET US DO

2.1 Define expectation and variance of a random variable. Illustrate the definitions with the help of a numerical example.

2.2 From the below given distribution of random variable x, find (a) E(x), (b) E(5x), (c) E(5-x), and (d) $E[x-E(x)]^2$.

| x | Probability |
|---|---|
| 11 | .08 |
| 17 | .12 |
| 21 | .15 |
| 24 | .30 |
| 28 | .15 |
| 31 | .12 |
| 49 | .08 |

2.3 A player rolls an unbiased die. If a prime number occurs, he wins an equal number of dollars. Showing up of a nonprime number results in a loss of that number of dollars to him. Is the game favorable to the player ?

2.4 What is a parameter ? Work out mean and variance for the following given population values :
44, 56, 60, 48, 55, 50, 58, 62, 60, 40.

2.5 How does a statistic differ from a parameter ? Explain.

2.6 Explain, in what sense the statistic $s^2$, the sample mean square, is a random variable?

2.7 What do you understand by the sampling distribution of a statistic ? The knowledge of sampling distribution is important to statistical inference. Explain.

2.8 To what different uses, can the calculated values of sample mean and sample mean square be put ? Discuss.

2.9 Assuming that 20, 12, 15, 16, 18, 14, 22, 28, 24, and 26 are the observations for a sample of 10 units, calculate sample mean and sample mean square.

2.10 Prove algebraically that

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

2.11 What is a point estimator ? Discuss the desirable properties it should possess.

2.12 Five babies were born in a particular year in village Beonhin of Mathura district. The age (in years) of mothers at the time of child birth were 29, 32, 26, 28, and 36. Enumerate all possible WR equal probability samples of size 3, and show numerically that the sample mean age is an unbiased estimator of population mean age of the mothers.

2.13 How does sampling variance differ from mean square error ? State the situation where both are equivalent.

2.14 Distinguish between the population variance and the sampling variance. Illustrate it with a suitable example.

2.15 Describe briefly each of the following terms :

a. Unbiasedness
b. Consistency
c. Relative efficiency
d. Confidence coefficient

2.16 What do you understand by a confidence interval ? Why does an experimenter need it ? Also, discuss the effect of bias in the estimator used to build up the confidence interval for the parameter.

2.17 When we construct a 95% level confidence interval for the population mean, what does it mean ?

2.18 While defining approximately 95% level confidence interval, we have used multiplier 2 in place of exact standard normal value. How will it affect the confidence coefficient for different kinds of sampling distributions ?

2.19 An investigator is to construct confidence interval for population proportion P. Under what circumstances can the normal distribution be used as an approximating sampling distribution for sample proportion p to work out confidence limits ? (*Hint* : For a variable y taking only 0 and 1 values, the sample mean $\bar{y}$ reduces to sample proportion p).

2.20 Why is the determination of sample size important in designing a survey ?

2.21 An investigator has randomly selected 2000 families following WR procedure from a population of 10,000 families. For working out a sufficiently accurate confidence interval for population mean, he/she is to guess the distribution of sample mean in absence of any information regarding the distribution of study variable in the population. Is it reasonable to assume that the sampling distribution is (a) exactly normal, (b) approximately normal, or (c) not at all normal ?

2.22 Distinguish between sampling and nonsampling errors. Which of these errors are more likely to be present in a census or a sample survey ?