Chapter 1
Introduction

Statistics is the science of data.

Data are the numerical values containing some information.

Statistical tools can be used on a data set to draw statistical inferences. These statistical inferences are in turn used for various purposes. For example, the government uses such data for policy formulation for the welfare of the people, marketing companies use the data from consumer surveys to improve the company and to provide better services to the customer, etc. Such data is obtained through sample surveys. Sample surveys are conducted throughout the world by governmental as well as non-governmental agencies. For example, "National Sample Survey Organization (NSSO)" conducts surveys in India, "Statistics Canada" conducts surveys in Canada, agencies of United Nations like "World Health Organization (WHO), "Food and Agricultural Organization (FAO)" etc. conduct surveys in different countries.

Sampling theory provides the tools and techniques for data collection, keeping in mind the objectives to be fulfilled and the nature of the population.

There are two ways of obtaining the information

1. **Sample surveys**
2. **Complete enumeration or census**

Sample surveys collect information on a fraction of the total population, whereas census collects information on the whole population. Some surveys, e.g., economic surveys, agricultural surveys etc. are conducted regularly. Some surveys are need-based and are conducted when some need arises, e.g., consumer satisfaction surveys at a newly opened shopping mall to see the satisfaction level with the amenities provided in the mall .

Sampling unit:
An element or a group of elements on which the observations can be taken is called a sampling unit. The objective of the survey helps in determining the definition of the sampling unit.

For example, if the objective is to determine the total income of all the persons in the household, then the sampling unit is a household. If the objective is to determine the income of any particular person in the household, then the sampling unit is the income of the particular person in the household. So the definition of sampling unit depends and varies as per the objective of the survey. Similarly, in another example, if the objective is to study the blood sugar level, then the sampling unit is the value of the blood sugar level of a person. On the other hand, if the objective is to study the health conditions, then the sampling unit is the person on whom the readings on the blood sugar level, blood pressure and other factors will be obtained. These values will together classify the person as healthy or unhealthy.

Population:

Collection of all the sampling units in a given region at a particular point of time or a particular period is called the population. For example, if the medical facilities in a hospital are to be surveyed through the patients, then the total number of patients registered in the hospital during the time period of the survey will be the population. Similarly, if the production of wheat in a district is to be studied, then all the fields cultivating wheat in that district will constitute the population. The total number of sampling units in the population is the population size, generally denoted by N. The population size can be finite or infinite (N is large).

Census:

The complete count of the population is called a census. The observations on all the sampling units in the population are collected in the census. For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

Sample:

One or more sampling units are selected from the population according to some specified procedure.
A sample consists only of a portion of the population units. Such a collection of units is called the sample.

In the context of sample surveys, a collection of units like households, people, cities, countries etc. is called a finite population.

A census is a 100% sample, and it is a complete count of the population.

Representative sample:
When all the salient features of the population are present in the sample, then it is called a representative sample.

It goes without saying that every sample is considered as a representative sample.

For example, if a population has 30% males and 70% females, then we also expect the sample to have nearly 30% males and 70% females.

In another example, if we take out a handful of wheat from 100 Kg. bag of wheat, we expect the same quality of wheat in hand as inside the bag. Similarly, it is expected that a drop of blood will give the same information as all the blood in the body.

Sampling frame:
The list of all the units of the population to be surveyed constitutes the sampling frame. All the sampling units in the sampling frame have identification particulars. For example, all the students in a particular university listed along with their roll numbers constitute the sampling frame. Similarly, the list of households with the name of the head of family or house address constitutes the sampling frame. In another example, the residents of a city area may be listed in more than one frame - as per automobile registration as well as the listing in the telephone directory.

Ways to ensure representativeness:
There are two possible ways to ensure that the selected sample is representative.

1. Random sample or probability sample:
The selection of units in the sample from a population is governed by the laws of chance or probability. The probability of selection of a unit can be equal as well as unequal.

2. Non-random sample or purposive sample:
The selection of units in the sample from the population is not governed by the probability laws.

For example, the units are selected on the basis of the personal judgment of the surveyor. The persons volunteering to take some medical test or to drink a new type of coffee also constitute the sample on non-random laws.

Another type of sampling is Quota Sampling. The survey, in this case, is continued until a predetermined number of units with the characteristic under study are picked up.

For example, in order to conduct an experiment for rare type of disease, the survey is continued till the required number of patients with the disease are collected.

Advantages of sampling over complete enumeration:

1. **Reduced cost and enlarged scope.**

   Sampling involves the collection of data on a smaller number of units in comparison to the complete enumeration, so the cost involved in the collection of information is reduced. Further, additional information can be obtained at little cost in comparison to conducting another separate survey. For example, when an interviewer is collecting information on health conditions, then he/she can also ask some questions on health practices. This will provide additional information on health practices, and the cost involved will be much less than conducting an entirely new survey on health practices.

2. **Organizaton of work:**

   It is easier to manage the organization of a collection of a smaller number of units than all the units in a census. For example, in order to draw a representative sample from a state, it is easier to manage to draw small samples from every city than drawing the sample from the whole state at a time. This ultimately results in more accuracy in the statistical inferences because the better organization provides better data and in turn, improved statistical inferences are obtained.

3. **Greater accuracy:**

   The persons involved in the collection of data are trained personals. They can collect the data more accurately if they have to collect a smaller number of units than a large number of units.

4. **Urgent information required:**

   The data from a sample can be quickly summarized.

   For example, the forecasting of the crop production can be done quickly on the basis of a sample of data than collecting first all the observation.

5. **Feasibility:**

   Conducting the experiment on a smaller number of units, particularly when the units are destroyed, is more feasible. For example, in determining the life of bulbs, it is more feasible to fuse a minimum number of bulbs. Similarly, in any medical experiment, it is more feasible to use less number of animals.

Type of surveys:

There are various types of surveys which are conducted on the basis of the objectives to be fulfilled.

1. Demographic surveys:

These surveys are conducted to collect demographic data, e.g., household surveys, family size, number of males in families, etc. Such surveys are useful in the policy formulation for any city, state or country for the welfare of the people.

2. Educational surveys:

These surveys are conducted to collect the educational data, e.g., how many children go to school, how many persons are graduate, etc. Such surveys are conducted to examine educational programs in schools and colleges. Generally, schools are selected first, and then the students from each school constitute the sample.

3. Economic surveys:

These surveys are conducted to collect the economic data, e.g., data related to export and import of goods, industrial production, consumer expenditure etc. Such data is helpful in constructing the indices indicating the growth in a particular sector of the economy or even the overall economic growth of the country.

4. Employment surveys:

These surveys are conducted to collect the employment-related data, e.g., employment rate, labour conditions, wages, etc. in a city, state or country. Such data helps in constructing various indices to know the employment conditions among the people.

5. Health and nutrition surveys:

These surveys are conducted to collect the data related to health and nutrition issues, e.g., number of visits to doctors, food given to children, nutritional value etc. Such surveys are conducted in cities, states as well as countries by the national and international organizations like UNICEF, WHO etc.

6. Agricultural surveys:

These surveys are conducted to collect the agriculture-related data to estimate, e.g., the acreage and production of crops, livestock numbers, use of fertilizers, use of pesticides and other related topics. The government bases its planning related to the food issues for the people based on such surveys.

7. Marketing surveys:

These surveys are conducted to collect data related to marketing. They are conducted by major companies, manufacturers or those who provide services to consumer etc. Such data is used for knowing the satisfaction and opinion of consumers as well as in developing the sales, purchase and promotional activities etc.

8. Election surveys:

These surveys are conducted to study the outcome of an election or a poll. For example, such polls are conducted in democratic countries to have the opinions of people about any candidate who is contesting the election.

9. Public polls and surveys:

These surveys are conducted to collect public opinion on any particular issue. Such surveys are generally conducted by the news media and the agencies which conduct polls and surveys on the current topics of interest to the public.

10. Campus surveys:

These surveys are conducted on the students of any educational institution to study educational programs, living facilities, dining facilities, sports activities, etc.

Principal steps in a sample survey:

The broad steps to conduct any sample surveys are as follows:

1. Objective of the survey:

The objective of the survey has to be clearly defined and well understood by the person planning to conduct it. It is expected from the statistician to be well versed with the issues to be addressed in consultation with the person who wants to get the survey conducted. In complex surveys, sometimes the objective is forgotten, and data is collected on those issues which are far away from the objectives.

2. Population to be sampled:

Based on the objectives of the survey, decide the population from which the information can be obtained. For example, the population of farmers is to be sampled for an agricultural survey, whereas the population of patients has to be sampled for determining the medical facilities in a hospital.

3. Data to be collected:

It is important to decide which data is relevant for fulfilling the objectives of the survey without omitting any essential data. Sometimes, too many questions are asked, and some of their outcomes are never utilized. This lowers the quality of the responses and in turn, results in lower efficiency in the statistical inferences.

4. Degree of precision required:

The results of any sample survey are always subjected to some uncertainty. Such uncertainty can be reduced by taking larger samples or using superior instruments. This involves more cost and more time. So it is very important to decide about the required degree of precision in the data. This needs to be conveyed to the surveyor also.

5. Method of measurement:

The choice of measuring instrument and the method to measure the data from the population needs to be specified clearly. For example, the data has to be collected through interview, questionnaire, personal visit, a combination of any of these approaches, etc. The forms in which the data is to be recorded so that the data can be transferred to mechanical equipment for easily creating the data summary etc. are also needed to be prepared accordingly.

6. The frame:

The sampling frame has to be clearly specified. The population is divided into sampling units such that the units cover the whole population, and every sampling unit is tagged with identification. The list of all sampling units is called the frame. The frame must cover the whole population, and the units must not overlap each other in the sense that every element in the population must belong to one and only one unit. For example, the sampling unit can be an individual member in the family or the whole family.

7. Selection of sample:

The size of the sample needs to be specified for the given sampling plan. This helps in determining and comparing the relative cost and time of different sampling plans. The method and plan adopted for drawing a representative sample should also be detailed.

8. The Pre-test:

It is advised to try the questionnaire and field methods on a small scale. This may reveal some troubles and problems beforehand which the surveyor may face in the field in large scale surveys.

9. Organization of the fieldwork:

How to conduct the survey, how to handle business administrative issues, providing proper training to surveyors, procedures, plans for handling the non-response and missing observations etc. are some of the issues which need to be addressed for organizing the survey work in the fields. The procedure for early checking of the quality of return should be prescribed. It should be clarified how to handle the situation when the respondent is not available.

10. Summary and analysis of data:

It is to be noted that based on the objectives of the data, the suitable statistical tool is decided which can answer the relevant questions. In order to use the statistical tool, a valid data set is required, and this dictates the choice of responses to be obtained for the questions in the questionnaire, e.g., the data has to be qualitative, quantitative, nominal, ordinal etc. After getting the completed questionnaire back, it needs to be edited to amend the recording errors and delete the erroneous data. The tabulating procedures, methods of estimation and tolerable amount of error in the estimation need to be decided before the start of the survey. Different methods of estimation may be available to get the answer of the same query from the same data set. So the data needs to be collected, which is compatible with the chosen estimation procedure.

11. Information gained for future surveys:

The completed surveys work as a guide for improved sample surveys in future. Besides this, they also supply various types of prior information required to use different statistical tools, e.g., mean, variance, nature of variability, the cost involved etc. Any completed sample survey acts as a potential guide for the surveys to be conducted in the future. It is generally seen that things always do not go in the same way in any complex survey as planned earlier. Such precautions and alerts help in avoiding the mistakes in the execution of future surveys.

Variability control in sample surveys:

Variability control is an important issue in any statistical analysis. A general objective is to draw statistical inferences with minimum variability. There are various types of sampling schemes which are adopted in different conditions. These schemes help in controlling the variability at different stages. Such sampling schemes can be classified in the following way.

1. Before the selection of sampling units

- Stratified sampling
- Cluster sampling
- Two stage sampling
- Double sampling etc.

2. At the time of selection of sampling units

- Systematic sampling
- Varying probability sampling

3. After the selection of sampling units

- Ratio method of estimation
- Regression method of estimation

Note that the ratio and regression methods are the methods of estimation and not the methods of drawing samples.

Methods of data collection

There are different ways of data collection. Some of them are as follows:

1. Physical observations and measurements:

The surveyor contacts the respondent personally through the meeting. He observes the sampling unit and records the data. The surveyor can always use his prior experience to collect the data in a better way. For example, a young man telling his age as 60 years can easily be observed and corrected by the surveyor.

2. Personal interview:

The surveyor is supplied with a well-prepared questionnaire. The surveyor goes to the respondents and asks the same questions mentioned in the questionnaire. The data in the questionnaire is then filled up accordingly based on the responses from the respondents.

3. Mail enquiry:

The well-prepared questionnaire is sent to the respondents through postal mail, e-mail, etc. The respondents are requested to fill up the questionnaires and send it back. In case of postal mail, many times the questionnaires are accompanied by a self-addressed envelope with postage stamps to avoid any non-response due to the cost of postage.

4. Web-based enquiry:

The survey is conducted online through internet-based web pages. There are various websites which provide such facility. The questionnaires are to be in their formats, and the link is sent to the respondents through e-mail. By clicking on the link, the respondent is brought to the concerned website, and the answers are to be given online. These answers are recorded, and responses, as well as their statistics, is sent to the surveyor. The respondents should have an internet connection to support the data collection with this procedure.

5. Registration:

The respondent is required to register the data at some designated place. For example, the number of births and deaths along with the details provided by the family members are recorded at the city municipal office which are provided by the family members.

6. Transcription from records:

The sample of data is collected from the already recorded information. For example, the details of the number of persons in different families or number of births/deaths in a city can be obtained from the city municipal office directly.

The methods in (1) to (5) provide primary data which means collecting the data directly from the source. The method in (6) provides secondary data, which means getting the data from the primary sources.