Introduction

In chapter 1, we discussed Statistics as the study of collection, organization, analysis, interpretation and presentation of data. For studying statistics the first step is collection of data, which we will discuss in detail in this chapter.

2.1   Data Collection

In our day to day life we deal with different types of data collection situation. A teacher might collect information regarding the test score of a student, a journalist might collect information regarding the recent social issues, a politician collects information on how voters plan to vote in the upcoming election, etc. Data collection is the systematic gathering of **data** for a particular purpose from various sources.

Data is the plural of the term datum, which means any measurement, result, fact or observation which gives information. Statistical surveys are the most popular devices for obtaining the desired data.

Before dealing with statistical surveys, we have to familiarize with the following terms.

Statistical Investigation

Statistical investigation includes collection, classification, presentation, analysis and interpretation of data according to well defined procedures. The person authorized to make investigation is known as **Investigator**. In a statistical investigation the investigator formulates the problem, suggests the data collection methods, organises various steps in an appropriate way, analyses the data and interpret the result. Usually, the investigators depute some persons to collect the data from the field. These persons are known as **Enumerators.** The enumerator may not be aware of the investigation procedures completely.

His/her duty is to collect the data for the investigator. It is the duty of the investigator to train and supervise the work of the enumerator. The process of data collection by the enumerator is known as **Enumeration.**

Population and Sample

A **population** consists of all elements, individuals, items or objects whose characteristics are being studied. For a politician, while considering voters plan for the next election, all registered voters in the specified constituency determines the population. If data are collected from each and every unit of the population, the investigation is called **census**. Based on the number of objects in a population, we can classify the population as finite and infinite. A population is said to be **finite**, if the number of individuals involved in the population is finite. All students of Kerala University for the year 2013-14 constitute a finite population. A population which is not finite or extremely large is **infinite**. The population comprises of all people in the world above 18 years of age is considered as an infinite population.

If the population is infinite or is of extremely large size, it is not feasible or practicable to access the entire population for study. As a result, it is apt to take a representative part as a substitute for the entire population. This representative part of the population is known as **sample**. The method of collecting data from the sample is known as **sampling** or **sample survey.** Various sampling designs and their selections are discussed in the last chapter.

The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 BC. In addition, the roman Emperor Augustus (27 BC to 17 AD) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly. In India about 2000 years ago we had an efficient system of collecting administrative statistics, particularly,during the regime of Chandra Gupta Maurya (324 to 300 B.C.). The system of collecting data related to births and

deaths is mentioned in Kautilyas Arthshastra (around 300 B.C.). Ain-i-Akbari, written by Abul Fazl, gives us a detailed account of the administrative and statistical survey conducted during the reign of Emperor Akbar.

Statistical Survey

A survey is a process of collecting data either from the population or from sample units. The statistical survey may be either by Census method or by Sampling method. The purpose of conducting a sample survey is to collect information about population using sample.

Know your progress

1. Illustrate Data, Statistical Investigation and Statistical Survey.
2. Distinguish between Population and sample?
3. Explain Finite and Infinite Population with the help of examples

2.2   Variables

Consider a group of people in a locality. The members of the group are found to be varying in many factors like sex, age, eye colour, intelligence, height, weight, blood pressure etc. The factors which can vary from one object to another are called variables. Among these variables sex, eye colour and intelligence which cannot be numerically measured are called qualitative variables or attributes. A qualitative variable is one that can be identified by noting its presence or identified with different categories of the factor. The other variables height, weight, age and blood pressure which are numerically measured are called quantitative variables. A quantitative variable consists of numerical values. Depending on the values taken by a quantitative variable, it is further classified as discrete variable and continuous variable. If the variable takes specific values only, it is called discrete variable. The variable, number of children in a family, does not take values other than 0, 1, 2, 3,etc. That is, there is a specified

break between the successive values. This is an example of a discrete variable. A continuous variable takes any value within the defined range of values. Between any two values of a continuous variable, an indefinitely large number of values may occur. Height, weight, time etc are examples of continuous variables. Depending on the type of variables involved, data may also be classified as discrete or continuous.

Know your progress

1. Write examples for variables.

2. Write examples for discrete and continuous variables.

3. Give examples other than those presented in this section of a qualitative variable, discrete quantitative and continuous quantitative variable

Levels of Measurement -

Nominal, Ordinal and Cardinal Data

S.S.Stevens (1906 -1973) described the data into different scales of measurement as nominal,ordinal and cardinal data. This classification is based on the data under consideration. A nominal scale of measurement is used to name categories such as gender, nation, etc. For example the categorization like male, female is a nominal data. While filling Higher Secondary single window application form for admission we give different codes to represent data such as Thiruvanathapuram 01, Kasaragod 14 etc. This is also a nominal data. Here the number is merely a label, does not have a quantitative significance. The only effect of such a measurement is that we can count how many objects fall in each category like 10000 belongs to Thiruvanathapuram,

S.S.Stevens
Stevens classify the cardinal data into two other classes as interval and ratio scale. This classification is generally known as Stevens Taxonomy

9700 belongs to Kasaragod, etc. In the ordinal scale of measurement, we can put an order to the data according to the relation among the values of the variables. While considering the educational qualification of a group of people, we can categorise them as Secondary, Higher Secondary, Graduate, Postgraduate, etc. Here we can rank Secondary -1, Higher Secondary -2, Graduate -3, Postgraduate -4, etc. Here code 3 is surely higher than code 1 as a graduate is at a higher level than a secondary student based on educational qualifications. That is in ordinal scale, there is a specific order or rank for the codes given to each category. The data regarding a quantitative variable is a cardinal data. The height of students in a class, monthly salary of school teachers, marks of students in a class etc. are examples of cardinal data.

Know your progress

1. Compare Nominal, Ordinal, Cardinal data.
2. Give some examples Nominal, Ordinal and Cardinal data.

2.3   Types of Data

Primary and Secondary Data

Data means the raw facts and figures that have been collected. Data can be gathered by looking through existing sources, conducting experiments or by conducting surveys. Based on these sources of collections, statistical data may be classified as primary or secondary. Primary data are those which have to be collected by the investigator for the first time for his/her own purpose. It is fresh in nature. In the words of Wessel data originally collected in the process of investigation are known as primary data. It is collected by using appropriate survey techniques. Data obtained from existing sources which may be published or unpublished are known as secondary data. Secondary data may not be in the required form. These data are obtained by other persons and are being used now at second hand. According to M. M. Blair, "secondary data

those which are already in existence, and which have been collected for some other purpose than the answering of question in hand"

Comparison between Primary and Secondary data

Primary Data	Secondary Data
It is original in nature	It is not original in nature
It is in the form of raw materials	It is in the form of finished products
Collection involves more money and time	Less time and money are needed
Trained persons are required for data collection	The investigator should be vigilant while collecting secondary data
Primary data, after use become secondary data	Secondary data cannot be converted to primary data

2.4   Questionnaire and Schedule

Questionnaires and schedules are series of questions arranged in a logical order so as to collect information for a specified purpose. The purpose may be single or multiple.

Questionnaire

A questionnaire is usually mailed by post or by email to selected informants. The informants are allowed a specified time to fill up the questionnaire and have to return to the investigator. Here the quality of the obtained data depends on the quality of the questions and the honesty of the informants. As the informants are to fill up the data, they should be literate. This method is suitable in cases where the informants are widely scattered. One of the main disadvantages of this method is that the chance of getting incomplete information is large.

Schedule

If the group of informants are not widely scattered, or if they are not literate, the enumerator himself/herself can personally approach the informants with the set of question and collect information. These questions may not be in detailed manner as questionnaire. It may not contain explanatory foot notes or explanations of terms used. These set of questions used for data collection is termed as schedule. In some cases questionnaire itself can be used as a schedule.

Comparison between Questionnaire and Schedule

Questionnaire	Schedule
It is often sent by post	Enumerators carry the schedule personally to the informant
Answers are filled by the respondents	Answers are filled by the enumerators
Informants are to be literate	Informants need not be literate
Success depends on the quality of questions and sincerity of the informant	Success depends on the honesty and competence of the enumerator
Chance of getting incomplete information is more	Chance of getting incomplete information is less as enumerators explain the questions

Requisites of good questionnaire

While preparing a questionnaire we have to keep in mind the following points.

1. Questions should be capable of generating all required information

2. The language and wording of the questions should be convenient to the informant

3. Question should not contain technical terms and words with uncommon meaning, such questions leads to different information from different informants

4. Yes or No questions or multiple answer choice questions should be preferred.

5. Personal questions should be avoided

6. Necessary foot notes should be provided

7. Usually the number of questions should be 20 to 25

8. Questions should have a logical order

9. Questions should be self explanatory

10. Questionnaire should be attractive so as to impress the informant

11. Questions should be unambiguous

Once the questionnaire is ready, it is advisable to conduct a pre-test with the questionnaire for a small group. This is known as pilot survey. It helps the investigator to measure the worthiness and reliability of these questions.

Drafting of Questionnaire

A sample questionnaire is prepared below for studying the socio economic status of people residing in a village.

QUESTIONNAIRE TO COLLECT DATA ABOUT THE SOCIO-ECONOMIC STATUS OF PEOPLE

1. Name :

2. Address:

3. Members in the family : No. of Males ☐   No. of Females ☐

4. Age of family members : (write number of members in each category in Corresponding boxes)
   a) below 10 ☐   b)10-20 ☐   c)20-50 ☐   d) 50 and above ☐

5. Residing in :
   a) Own house ☐   b) Rental ☐

6. Type of House :
   a) Temporary ☐ b) Structured ☐

7. Toilet facility :
   a) Proper ☐ b) Improper ☐

8. Water Facility :
   a) own well ☐ b)water provided by panchayat ☐ c) other Sources☐

9. Electrified Home :
   Yes ☐ No ☐

10. Details of electronic items in house : (write number of items in each boxes)
    a) Bulbs ☐   b)Refrigerator ☐   c) Fan ☐   d) Television ☐
    e) Tubes ☐   f) Mixer grinder ☐   g) Others ☐

11. Mode of cooking :
    a) Wood ☐ b) Kerosene ☐ c) Gas ☐ d) Electricity ☐

12 Occupation :
a)Govt. service ☐ b) Non Govt. Service ☐ c)Own business ☐
d)Agriculture ☐ e) Others ☐

13 Monthly income of family (in rupees) :
a)Below 8000 ☐ b) 8000-20000 ☐
c) 20000-50000 ☐ d) above 50000 ☐

14 The amount you monthly spend for education of children
a)below 500 ☐ b)500-2000 ☐
c) 2000-4000 ☐ d) above 4000 ☐

15 Do you own a private vehicle? :
Yes ☐ No ☐

16 If Yes,give the number of vehicles in each category. :
a)Two wheeler ☐ b) Three wheeler ☐
c)Car ☐ d) Others ☐

17 The approximate monthly expenditure :
a) Below 5000 ☐ b)5000-15000 ☐
c) 15000-20000 ☐ d) above 20000 ☐

18 Would you prefer an outing occasionally with your family?
Yes ☐ No ☐

19 If Yes what will be your budget for a single trip?
a) below 2000 ☐ b) 2000-5000 ☐ c)above 5000 ☐

20 Are you able to properly maintain your standard of living with your
actual income?
Yes ☐ No ☐

21 Any other information regarding your family (Give details in one or
two sentences)

(** This data will be used for study purpose only)

2.5   Methods of Primary Data Collection 6

Direct Personal Interview

If the field of investigation is small, the investigator or enumerator can access all the informants personally and conduct spot enquiry. This method is called direct personal interview. The success of this method depends on the efficiency of the enumerator. The enumerator should be tactful to get all the required information. In this method the enumerator can collect all the supplementary information required for interpretation of data

Indirect Oral Investigation

Consider a situation in which the investigator wants to collect data about a resident in a city. In this case, the investigator may approach a third party, called witness, who is capable of giving sufficient information about the resident. This is a case of indirect oral investigation. Indirect oral investigation is applicable in cases where the informant is reluctant to give information or when the informant is not available. The disadvantage of this method is that the reliability of the information heavily depends on the quality and honesty of the witness or intermediate person.

Direct Observation

This method is widely used by mass media for collecting information or by journalists. Consider a situation in which the investigator wants to report the current situation in an area due to heavy rain and flood. He does not have a predetermined set of questions to collect data. The investigator collects data from what he observes. This may not be in a well defined manner. The investigator must be well equipped so as to collect maximum information from the place. The quality of the information depends mainly on the honesty of the investigator to report it to the maximum extend.

collection of data

Telephone interview

In some cases the informant may be reluctant to give answer in a face to face personal interview. In such cases it is better to select another method for data collection. Telephone interview is one such method. In this case the investigator collects data from the informant indirectly but personally. This is less time consuming and cheaper than direct personal interview. The disadvantage is that it will not worked in some rural areas were telephonic connection is very low.

Mailed Questionnaires and Schedules

Questionnaires and schedules are one of the most popular methods for collecting primary data. As the questionnaires are usually mailed to the respondents, it is known as mailed questionnaire method. The only difference between the questionnaire and schedule is that in questionnaires the answers are filled by the respondents themselves but in schedule, the answers are filled by the enumerators.

Focus Group Discussion

A Focus Group Discussion (FGD) is a small group discussion guided by a trained leader. It is used to collect more opinions about a specified topic in order to take better decisions in future plans. For example, Mrs. Rema wants to start a pre-school in an area. Her aim is to provide childcare as much as possible. For that she invites the parents who have of children under four years from that locality and arranges Focus Group Discussion. The parents have their own ideas about childcare and other locally adopted programmes. Their suggestions will help Mrs. Rema to start the institution in a well equipped manner. Here Mrs.Rema plays the role of the investigator. She collects information through focus group discussion for her specified purpose.

2.6 Sources of Secondary Data

Any published or unpublished data which are reliable for the current situation is a source of secondary data. While collecting secondary data the investigator must be aware of the following points.

- The geographical area of the collected data.

- The time at which the data was collected.

- The terms and definitions involved in the data.

- The person who collected the data and the purpose for which they are collected.

Some sources of secondary data are listed below

- Government publications.
- Office records in panchayats, municipalities etc.
- Survey reports of various research organizations.
- Survey reports in Journals, Newspapers and other publications.
- Websites.

Let us sum up

For studying statistics, the first step is collection of data. It is systematic gathering of informations. Statistical surveys are tools of data collection. Data means information regarding a variable. The variable may be qualitative and quantitative or it may be discrete or continuous. The two survey techniques are Census and sampling. Depending on the source of information, data can be classified as primary or secondary. The important primary data collection methods are direct personal investigation, indirect oral investigation, direct observation, telephone interview, mailed questionnaire or schedule sent through enumerators and focus group discussion. Any published or unpublished data which are collected by a third party, and now used by the investigator for his purpose is a secondary data. Better trained persons are required for converting the secondary data to the required form.

Learning outcomes

After transaction of this unit, the learner:-

- differentiates population and sample.
- recognises investigator, investigation, enumerator and enumeration.
- classifies variables and constants.
- distinguishes qualitative variables and quantitative variables.
- differentiates discrete and continuous variables.
- compares primary and secondary data.
- identifies questionnaire and schedule.
- constructs/drafts questionnaire.
- explains different methods of data collection.
- recognises the sources of secondary data.

Evaluation Items

1. Data that can be classified according to colour. They are measured on .... scale
   a) Nominal   b) Ordinal   c) Cardinal

2. The group of all subjects under study is called ………

3. The representative part of a population is called ………

4. The number of days of absence of a worker per year is ……… type of data
   a) Nominal   b) Qualitative   c) Discrete   d) Continuous

5. The pre-test with the questionnaire before conducting a survey is called
   ………

6. The blind population of India constitute:
   a) a hypothetical population   b) a sample   c) an infinite population   d) a finite population

7. Which of the following represents data
   a) a single value   b) only two values in a set   c) a group of values in a set
   d) all the above

8. Statistics deal with
   a) qualitative information   b) quantitative information   c) both (a) and (b) d) none of (a) and (b)

9. Data taken from a publication Agricultural situation in India will be considered as
   a) primary data   b) secondary data

10. Mailed questionnaire method of enquiry can be adopted if respondents
    a) live in cities   b) have a high income   c) are literate

11. A study based on complete enumeration is known as:
    a) sample survey   b) pilot survey   c) none of the above

12. Statistical data are collected for.
    a) no purpose   b) a given purpose   c) any purpose

13. A statistical population may consists of
    a) an infinite number of items   b) a finite number of items   c) either of (a) and (b) d) none of (a) and (b)

14. Compare Primary and Secondary data

15. Distinguish between a questionnaire and a schedule.

16. What kind of data you receive when you are told about
    (a) blood type b) house hold c)heights of waterfall

17. What are points to be remembered while drafting a questionnaire?

18. What are the points to be remembered while collecting secondary data?

19. Explain the various primary data collection methods.

20. What are the important sources of secondary data?

21. Find the discrete data and continuous data from the following list
    a. Number of shares sold each day in a stock market
    b. Temperature recorded every half an hour at a weather bureau
    c. Life time of television tubes reduced by a company
    d. Yearly income of employees in a company
    e. The age of an individual
    f. Number of petals a flower has.

22. Categorize the data obtained in the following situations as quantitative
    and qualitative
    a. Political preference of a group of a people
    b. Family size (Number of members of a family) of hundred families in a
    township
    c. IQ score of plus one students undergoing state syllabus in Thiruvananthapuram
    district d. Academic qualification of a group of unemployed youth in a
    city

23. Which of the following constitute finite or infinite population
    a. Population consisting of odd integers.
    b. Weight of 200 new born babies in a hospital.
    c. Height of fifteen year old children in a school.
    d. Number of head and tail when a coin is tossed.

24. Categorize the following under cardinal, nominal or ordinal
    a. Telephone Numbers
    b. Roll numbers given to students in a class
    c. Ranks given to a class after a test
    d. Respondents attitude towards a newly designed project in an institution on a five point scale such as 1=strongly opposed, 2= may be opposed, 3= not strongly favoured, 4= may be favoured and 5= very strongly favoured
    e. The quantity of water in a reservoir measured in every half an hour.
    f. The price of furniture exhibited in a shop.

25. It is proposed to conduct a survey to obtain information on the study habits of Higher Secondary students in Kannur district and also the facilities available to them. Prepare a questionnaire for this purpose

26. A survey is to be carried out amongst school children to study how they spent time after school hours. Prepare a questionnaire for that purpose

27. Indicate whether the following statements are true or false. If false correct the statements
    a. Secondary data are generally used in those cases where the primary data do not provide an adequate basis for analysis.
    b. Secondary data does not need much scrutiny and should be accepted at its face value
    c. The task of editing secondary data is a highly specialised one
    d. The questionnaire requires a pre-testing before putting into practice

28. Which type of study do you prefer in the following cases? (Census or Sampling). Give reason
    a. The effect of a medicine
    b. To study about the wage distribution of 250 employees in a company
    c. A study on the roll of media in the marketing of a face cream
    d. A study of a patients heart beat who is admitted to ICCU of a hospital
    e. A study on the number of petals of a flower of a special kind

f. The life span of an electric bulb.

29. Which of the primary data collection method do you suggest in the following situations?. Give reasons

a. You are appointed as marketing manager of a company. The company introduces a washing machine with many options. You are asked by your employer to prepare a datasheet regarding the opinion of your customers about the new equipment

b. To prepare a report for a media on Nehru Trophy Vallamkali in the current year.

c. To introduce shift sessions in an institution

30. As a reporter of a certain media, you got an opportunity to interview an IAS topper. Which type of data collection method will you use?. List out other primary data collection methods.

Answers:

1)a   2)population   3)sample   4)discrete   5)pilot survey   6)d   7)c
8)c   9)b   10)c   11)c   12)b   13)c

22)(a) qualitative   (b) quantitative   (c) quantitative   (d)qualitative

23)(a) infinite   (b) finite   (c) finite   (d) finite   24(a)nominai
(b) ordinal   (c)ordinal   (d) nominal   (e) cardinal   (f) cardinal
27 (a) false   (b) true   (c)true   (d) true