# UNIT—II DATA COLLECTION

# LESSON 1 SOURCES OF DATA

LESSON OUTLINE

- **Primary data-**
- **Methods of collecting primary data-**
- **Direct personal investigation**
- **Indirect oral interviews**
- **Information received through local agencies**
- **Mailed questionnaire method**
- **Schedules sent through enumerators**

**Learning Objectives**

**After reading this lesson you should be able to**

- **Understand the meaning of primary data**
- **Preliminaries of data collection**
- **Method of data collection**
- **Methods of collecting primary data**
- **Usefulness of primary data**
- **Merits and demerits of different methods of primary data collection**
- **Pre cautions while collecting primary data.**

**Introduction**

It is important for a researcher to know the sources of data which he requires for his different purposes. Data are nothing but the information. There are two sources of information or to say data- Primary data and Secondary data. Primary data mean the data collected for the first time, whereas secondary data mean the data that have already been collected and used earlier by somebody or some agency. For example, the statistics collected by the Government of India relating to the population, are primary data for the Government of India since it has been collected for the first time. Later when the same data are used by a researcher for his study of a particular problem, then the same data become the secondary data for the researcher.

Both the sources of information have their merits and demerits. The selection of a particular source depends upon–(a) Purpose and scope of enquiry ; (b) availability of time ;(c) availability of finance and;(d) accuracy required. (e) Statistical units to be used (f) Sources of information (data) and (g) Method of data collection. Let us discuss the above points in short.

**(a) Purpose and scope of enquiry:-**The purpose and scope of data collection or survey should be clearly set out at the very beginning. It requires the clear statement of the problem indicating the type of information which is needed and the use to which it is needed .If for example, the researcher is interested in knowing the nature of price change over a period of time, it would be necessary to collect data of commodity prices and it must be decided whether it would be helpful to study wholesale or retail prices and the possible uses to which such information could be put. The objective of an enquiry may be either to collect specific information relating to a problem or adequate data to test a hypothesis. Failure to set out clearly the purpose of enquiry is bound to lead to confusion and waste of resources.

After the purpose of enquiry has been clearly defined, the next step is to decide about the scope of the enquiry. Scope of the enquiry means the coverage with regard to the type of information, the subject-matter and geographical area. For instance, an enquiry may relate to India as a whole or a state or an industrial town where in a particular problem related to a particular industry can be studied.

**(b)Availability of time:-** The investigation should be carried out within a reasonable period of time; otherwise the information collected may become outdated, and have no meaning at all. For instance, if a producer wants to know the expected demand of a product newly launched by him and the result of the enquiry that the demand would be meager, takes two years to reach to him then the whole purpose of enquiry would become useless because by that time he would have already received a huge loss. Thus in this respect the information is quickly required and hence the researcher has to choose the type of enquiry accordingly.

**I Availability of resources:-** The investigation will greatly depend on the resources available like number of skilled personnel, the amount etc. If the number of skilled personnel who will carry out the enquiry is quite sufficient and the amount is not a problem then the enquiry can be conducted over a big area covering a good number of samples otherwise a small sample size will do.

**(d)The degree of accuracy desired:-** Deciding the degree required is must for the investigator, because absolute accuracy in statistical work is seldom achieved. This is so because (a) statistics are based on estimates, (b) tools of measurement are not always perfect and (c) there may be unintentional bias on the part of the investigator,, enumerator or informant. Therefore, a desire of

100% accuracy is bond to remain unfulfilled. Degree of accuracy desired primarily depends upon the object of enquiry. For example when we buy gold even a difference of 1/10$^{th}$ gram in its weight is significant whereas the same will not be the case when we buy rice or wheat. However, the researcher must aim at attaining a higher degree of accuracy otherwise the whole purpose of research would become meaningless.

**(e) Statistical Units to be used:** A well defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called a *statistical unit.* For example, in socio-economic survey the unit may be an individual person, a family, a household or a block of locality. A very important step before the collection of data begins is to define clearly the statistical units on which the data are to be collected. In number of situations the units are conventionally fixed like the physical units of measurement such as metres, kilometers, quintals, hours, days, week etc., which are well defined and do not need any elaboration or explanation. However in many statistical investigations, particularly relating to socio-economic studies, arbitrary units are used which must be clearly defined. This is must because in the absence of a clear cut and precise definition of the statistical units, serious errors in the data collection may be committed in the sense that we may collect irrelevant data on the items, which should have, in fact, been excluded and omit data on certain items which should have been included. This will ultimately lead to fallacious conclusions.

**(f) Sources of information (data):-** After decided about the unit, a researcher has to decide about the source from which the information can be obtained or collected. For any statistical inquiry, the investigator may collect the data first hand or he may use the data from other published sources such as the publications of the government/semi-government organizations or journals and magazines etc.

**(g) Method of data collection:-** There is no problem if secondary data are used for the research . However, if primary data are to be collected a decision has to be taken whether (i) census method or (ii) sample technique, is to be used for data collection .In census method we go for total enumeration i.e. all the units of a universe have to be investigated. But in sample technique, we inspect or study only a selected representative and adequate fraction of the population and after analyzing the results of the sample data we draw conclusions about the characteristics of the population. Selection of a particular technique becomes difficult because where population or census method is more scientific and 100% accuracy can be attained through this method, choosing this becomes difficult because it is time taking, it requires more labor and after all it is very expensive. Therefore, for a single researcher or for a small institution it proves to be unsuitable. On the other hand, sample method is less time taking, less laborious and less expensive but a 100% accuracy cannot be attained through this method because of sampling and non sampling errors attached to this method. Hence, a researcher has to be very cautious and careful while choosing a particular method.

**Methods of collecting Primary data**

Primary data may be obtained by applying any of the following methods-

1. Direct Personal Interviews
2. Indirect oral interviews.
3. Information from correspondents.
4. Mailed questionnaire methods.
5. Scheduled sent through enumerators.

**1. Direct personal interviews:-** A face to face contact is made with the informants(persons from whom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information. Thus, if a person wants to collect data about the working conditions of the workers of the Tata Iron and Steel Company, Jamshedpur, he would go to the factory, contact the workers and obtain the desired information. The information collected in this manner is first hand and also original in character.

There are many merits and demerits of this method which is discussed below:-

**Merits**

1. Most often respondents are happy to pass on the information required from them when contacted personally and thus response is encouraging.
2. The information collected through this method is normally more accurate because the interviewer can clear up doubts of the informants about certain questions and thus obtain correct information. In case the interviewer apprehends that the informant is not giving accurate information, he may cross-examine him and thereby try to obtain the information.
3. This method also provides the scope for getting the supplementary information from the informant because while interviewing it is possible to ask some supplementary questions which may be of great use latter.
4. It is experienced that there are some difficult questions which normally becomes difficult to ask directly but a trained and experienced researcher can sandwiched the difficult questions between other questions and get the desired information. He can twist the questions keeping in mind the informant's reaction. Precisely, a delicate situation can usually he

handled more effectively by a personal interview than by other survey techniques.

5. The interviewer can adjust the language according to the status and educational level of the person interviewed, and thereby can avoid inconvenience and misinterpretation on the part of the informant.

**Demerits:** There are some demerits or limitations of this method which are explained below:

1. This method can prove to be expensive if the number of informants is large and the area is wide spread
2. There is a greater chance of personal bias and prejudice under this method as compared to other method.
3. The interviewers have to be thoroughly trained and experienced otherwise they may not be able to obtain the desired information. Untrained or poorly trained interviewers may spoil the entire work.
4. This method is more time taking as compared to others. This is because interviews can be held only at the convenience of the informants. Thus, if information is required to be obtained from the working members of households, interviews will have to be held in the evening or on week end. Even during evening only an hour or two can be used for interviews and hence, the work may have to be continued for a long time, or a large staff may have to be employed which may involve huge expense.

**Conclusion:-**Though there are some demerits in this method of data collection still we cannot say that it is not useful. The matter of fact is that this method is suitable for intensive rather than extensive field surveys. Hence, it should be used only in those cases where intensive study of a limited field is desired.

In the present time of extreme advancement in the communication system, the investigator instead of going personally and conducting a face to face interview may also obtain information on telephone. A good number of surveys are being conducted every day by newspapers and television channels by sending the reply either by e-mail or SMS. This method has become very popular nowadays as it is less expensive and the response is extremely quick. But this method suffers from some serious defects as – (a) very few people own a phone or a television and hence a limited type of people can be approached by this method,(b) only few questions can be asked over phone or through television,(c) the respondents may give a vague and reckless answers because answers on phone or through SMS would have to be very short.

**2.Indirect Oral Interviews:-** Under this method of data collection, the investigator contacts third parties generally called 'witnesses' who are capable of supplying necessary information. This method is generally adopted when the information to be obtained is of a complex nature and informants are not inclined to respond if approached directly. For example, when the researcher is trying to obtain data on drug addiction or the habit of taking liquor, there is high probability that the addicted person will not supply the desired data and hence disturb the whole research process. In this situation taking the help of such persons or agency or the neighbour who know them well becomes necessary. Since these people know the person well and hence, they can supply the desired data. Enquiry Committees and Commissions appointed by the Government generally adopt this method to get people's views and all possible details of facts relating to the enquiry.

Though this method is very popular, its correctness depends upon a number of factors which is discussed below:-

1. The person or persons or agency whose help is solicited must be of proven integrity otherwise any bias or prejudiced on the part of them will not bring the correct information and the whole process of research will become useless.

2. The ability of the interviewers to draw out the information from witnesses by means of appropriate questions and cross-examination.

3. It might happen that because of bribery, nepotism or certain other reasons those who are collecting the information give it such a twist that correct conclusions are nor arrived at.

Therefore, for the success of this method it is necessary that the evidence of one person alone is not relied upon. Views from other persons and related agencies should also be ascertained to find the real position .Utmost care must be exercised in the selection of these persons because it is on their views that the final conclusions are reached.

**3. Information from Correspondents:-** The investigator appoints local agents or correspondents in different places to collect information under this method. These correspondents collect and transmit the information to the central office where data are processed. This method is generally adopted by news paper agencies. Correspondents who are posted at different places supply information relating to such events as accidents, riots, strikes, etc., to the head office. The correspondents are generally paid staff or sometimes they may be honorary correspondents also. This method is also adopted generally by the government departments in such cases where regular information is to be collected from a wide area. For example, in the construction of a wholesale price index numbers regular information is obtained from correspondents appointed in different areas. The biggest advantage of this method is that it is cheap and appropriate for extensive investigation. But a word of caution is that it may not always ensure accurate results because of the personal prejudice and bias of the correspondents.

As already stated earlier, this method is suitable and adopted in those cases where the information is to be obtained at regular intervals from a wide area.

1. **Mailed Questionnaire Method:-** Under this method, a list of questions pertaining to the survey which is known as 'Questionnaire' is prepared and sent to the various informants by post. Sometimes the researcher himself too contacts the respondents and gets the responses relating to the various questions in the questionnaire. The questionnaire contains questions and provides space for answers. A request is made to the informants through a covering letter to fill up the questionnaire and send it back within a specified time.

   The questionnaire studies can be classified on the basis of:

i. The degree to which the questionnaire is formalized or structured.
ii. The disguise or lack of disguise of the questionnaire , and
iii. The communication method used.

When no formal questionnaire is used, interviewers adapt their questioning to each interview as it progresses or perhaps elicit responses by indirect methods such as showing pictures on which the respondent comments. When a researcher follows a prescribed sequence of questions, it is referred to as *structured* study. On the other hand, when no prescribed sequence of questions exists, the study is *non-structured.*

When questionnaires are constructed so that the objective is clear to the respondents then these questionnaires are known as *non- disguised*; on the other hand, when the objective is not clear the questionnaire is a *disguised* one. On the basis of these two classifications, four types of studies can he distinguished:

i. Non-disguised structured,

ii. Non-disguised non-structured,

iii. Disguised structured, and

iv. Disguised non-structured.

There are certain merits and demerits or limitations of this method of data collection which are discussed below:

**Merits:**

2. Questionnaire method of data collection can be easily adopted where the field of investigation is very vast and the informants are spread over a wide geographical area.
3. This method is relatively cheap and expeditious provided the informants respond in time.
4. This method is proved to be superior when a question of a personal nature or questions requiring reaction by the family, than other methods as personal interviews or telephone method.

**Demerits:**

**1.** This method can be adopted only where the informants are literate people so hat they can understand written questions and wend the answers in writing.

**2.** It involves some uncertainty about the response. Co-operation on the part of informants may difficult to presume.

**3.** The information supplied by the informants may not be correct and in may be difficult to verify the accuracy.

However by following the following guidelines this method can be made more effective.

i. The questionnaire should be made in such a manner that it does not become an undue burden on the respondents; otherwise they may not return them back.

ii. Prepaid postage stamp should be affixed
iii. The sample should be large
iv. It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest in the enquiry.
v. It should be preferred in such enquiries where there could be a legal compulsion to supply the information so that the risk of non-response is eliminate.

5. **Schedules sent through Enumerators:-A**nother method of data collection is through sending schedules through the enumerators or interviewers. The enumerators contact the informants, get replies to the questions contained in a schedule and fill them in their own handwriting in the questionnaire form. There is difference between questionnaire and schedule. Questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in him self, whereas Schedule is the name usually applied to a set of questions which are asked and filed in a face-to face situation with another person. This method is free from most of the limitations of the mailed questionnaire method.

**Merits**

The main merits or advantages of this method are listed below:-

**i.** It can be adopted in those cases where informants are illiterate.
**ii.** There is very little scope of non-response as the enumerators go personally to obtain the information.
**iii.** The information received is more reliable as the accuracy of statements can be checked by supplementary questions wherever necessary.

This method too like others is not free from defects or limitations. The main limitations are listed below:-

**Demerits**

i. In comparison to other methods of collecting primary data, this method is quite costly as enumerators are generally paid persons.

ii. The success of the method depends largely upon the training imparted to the enumerators.

iii. Interviewing is a very skilled work and it requires experience and training, hut there is a tendency of statisticians to neglect this extremely important part of the data collecting process. Without good interviewing most of the information collected is of doubtful value.

iv. Interviewing is not only a skilled work but it also requires great degree of politeness and thus the way the enumerators conduct the interview would affect the data collected. When questions are asked by a number of different interviewers, it is possible that variations in the personalities of the interviewers will cause variation in the answers obtained. This variation will not be obvious. Hence every effort must be made to remove as much of variation as possible due to different interviewers.

**Secondary Data:-**As already stated earlier, secondary data are those data which have been already collected and analyzed by some earlier agency for its own use, and later the same data are used by a different agency. According to W.A.Neiswanger," A primary source is a publication in which the data are published by the same authority which gathered and analyzed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible."

**Sources of secondary data:-**The various sources of secondary data can be divided into two broad categories:

1. Published sources and,
2. Unpublished sources.

1. **Published Sources**: Various governmental, international and local agencies publish statistical data, and chief among them are explained below:

(a) International Publications: There are some international institutions and bodies like I.M.F, I.B.R.D, I.C.A.F.E, and the U.N.O etc. who publish regular and occasional reports on economic and statistical matters.

(b) Official publications of Central and State Governments: Several departments of the Central and State Governments regularly publish reports on a number of subjects. They gather additional information. Some of the important publications are: the Reserve Bank of India Bulletin, Census of India, Statistical Abstracts of States, Agricultural Statistics of India, Indian Trade Journal, etc.

(c) Semi- official publications: Semi-Government institutions like Municipal Corporations, District Boards, Panchayats, etc. publish reports relating to different matters of public concern.

(d) Publications of Research Institutions: Indian Statistical Institution (I.S.I), Indian Council of Agricultural Research (I.C.A.R), Indian Agricultural Statistics Research Institute (I.A.S.R.I), etc. publish the findings of their research programmes.

(e) Publications of various Commercial and Financial Institutions

(f) Reports of various Committees and Commissions appointed by the Government as the Raj Committee's Report on Agricultural Taxation, Wanchoo Committee's Report on Taxation and Black Money, etc. are also important sources of secondary data.

(g) Journals and News Papers: Journals and News Papers are very important and powerful source of secondary data. Current and important materials on statistics and socio-economic problems can be obtained from journals and newspapers like, Economic Times, Commerce, Capital, Indian Finance, Monthly Statistics of trade etc.

2. **Unpublished Sources:** Unpublished data can be obtained from many unpublished sources like records maintained by various government and private offices, the theses of the numerous research scholars in the universities or institutions etc.

**Precautions in the use of Secondary Data:** Since secondary data have already been obtained it is highly desirable that a proper scrutiny of such data is made before they are used by the investigator. In fact the user has to be extra-cautious while using secondary data. In this context Prof. Bowley rightly points out that "Secondary data should not be accepted at their face value." The reason being that data may be erroneous in many respects due to bias, inadequate size of the sample, substitution, errors of definition, arithmetical errors etc. Even if there is no error such data may not be suitable and adequate for the purpose of the enquiry. Prof. Simon Kuznet's view in this regard is also of great importance. According to him, "The degree of reliability of secondary source is to be assessed from the source, the compiler and his capacity to produce correct statistics and the users also, for the most part, tend to accept a series particularly one issued by a government agency at its face value without enquiring its reliability".

Therefore, before using the secondary data the investigators should consider the following factors:

6. **The suitability of data**: The investigator must satisfy him self that the data available are suitable for the purpose of enquiry. It can be judged

by the nature and scope of the present enquiry with the original enquiry. For example, if the object of the present enquiry is to study the trend in retail prices, and if the data provide only wholesale prices, such data are unsuitable.

(a) Adequacy of data: If the data are suitable for the purpose of investigation then we must consider whether the data are useful or adequate for the present analysis. It can be studied by the geographical area covered by the original enquiry. The time for which data are available is very important element. In the above example, if our object is to study the retail price trend of India, and if the available data cover only the retail price trend in the State of Bihar, then it would not serve the purpose.

(b) Reliability of data: The reliability of data is must. Without which there is no meaning in research. The reliability of data can be tested by finding out the agency that collected such data. If the agency has used proper methods in collection data, statistics may be relied upon.

It is not enough to have baskets of data in hand. In fact data in a raw form are nothing but a handful of raw material waiting for proper processing so that they can become useful. Once data have been obtained from primary or secondary source, the next step in a statistical investigation is to edit the data i.e. to scrutinize the same. The chief objective of editing is to detect possible errors and irregularities. The task of editing is a highly specialized one and requires great care and attention. Negligence in this respect may render useless the findings of an otherwise valuable study. Editing data collected from internal records and published sources is relatively simple but the data collected from a survey need excessive editing.

While editing primary data following considerations should be born in mind:

1. The data should be complete in every respect

2. The data should be accurate

3. The data should be consistent, and
4. The data should be homogeneous.

Data to posses the above mentioned characteristics have to undergo the same type of editing which are discussed below:

7. **Editing for completeness**: While editing the editor should see that each schedule and questionnaire is complete in all respects. Answers to each and every question have been furnished. If some questions are not answered and they are of vital importance, the informants should be contacted again either personally or through correspondence. Even after all the efforts it may happen that a few questions remain unanswered. In such questions, the editor should mark 'No answer' in the space provided for answers and if the questions are of vital importance then the schedule or questionnaire should be dropped.

1. **Editing for consistency**: At the time of editing the data for consistency, the editor should see that the answers to questions are not contradictory in nature. If the are mutually contradictory answers, he should try to obtain the correct answers either by referring back the questionnaire or by contacting, wherever possible, the informant in person. For example, if amongst others, two questions in questionnaire are (a) Are you a student? (b) Which class do you study and the reply to the first question is 'no' and to the latter 'tenth' then there is contradiction and it should be clarified.

2. **Editing for accuracy**: The reliability of conclusions depends basically on the correctness of information. If the information supplied is wrong, conclusions can never be valid. It is, therefore, necessary for the editor to see that the information is accurate in all respects. If the inaccuracy is due to arithmetical errors, it can be easily detected and corrected. But if

the cause of inaccuracy is faulty information supplied, it may be difficult to verify it e.g. information relating to income, age etc.

3. **Editing for homogeneity:** Homogeneity means the condition in which all the questions have been understood in the same sense. The editor must check all the questions for uniform interpretation. For example, as to the question of income, if some informants have given monthly income, others annual income and still others weekly income or even daily income, no comparison can be made. Therefore, it becomes an essential duty of the editor to check up that the information supplied by the various people is homogeneous and uniform.

**Choice between Primary and Secondary Data:-**As we have already seen, there are lot of difference in the methods of collecting Primary and Secondary data. In the case of primary data which is to be collected originally, the entire scheme of the plan starting with the definitions of various terms used, units to he employed, type of enquiry to be conducted, extent of accuracy aimed at etc. is to he formulated whereas the collection of secondary data is in the form of mere compilation of the existing data. A proper choice between the type of data needed for any particular statistical investigation is to be made after taking into consideration the nature, objective and scope of the enquiry; the time and the finances at the disposal of the agency; the degree of precision aimed at and the status of the agency (whether government- state or central-or private institution of an individual).

In using the secondary data it is best to obtain the data from the primary source as far as possible. By doing so, we would at least save ourselves from the errors of transcription which might have inadvertently crept in the secondary source. Moreover, the primary source will also provide us with detailed discussion about the terminology used, statistical units employed, size of the sample and the technique of sampling (if sampling method was used), methods of data collection and analysis of results and we can ascertain ourselves if these suit our purpose.

Now a days in a large number of statistical enquiries secondary data are generally used because fairly reliable published data on a large number of diverse fields are now available in publication of governments, private organizations and research institutions, agencies, periodicals and magazines etc. In fact primary data are collected only if there do not exist any secondary data suited to the investigation under study. In some of the investigations both primary as well as secondary data may be used.

**SUMMARY**

There are two types of data, Primary and secondary. Data which are collected first hand are called Primary data and data which have already been collected and used by some body or agency are called Secondary data. There are two methods of collecting data. They are- (a) Survey method or total enumeration method and (b) Sample method. When a researcher goes for investigating all the units of the subject, it is called as survey method and on the other hand when resorts to investigating only a few units of the subject and to give the result on the basis of that, it is known as sample survey method. There are different sources of collecting Primary and Secondary data. Some of the important sources of Primary data are—Direct Personal Interviews, Indirect Oral Interviews, Information from correspondents, Mailed questionnaire method, Schedules sent through enumerators. Though all these sources or methods of Primary data have their relative merits and demerits, a researcher should use a particular method with lot of care. There are basically two sources of collecting secondary data- (a) Published sources and (b) Un published sources. Published sources are like publications of different government and semi-government departments, research institutions and agencies etc. whereas unpublished sources are like records maintained by different government departments and unpublished theses of different universities etc. Editing of secondary data is necessary for different purposes as – editing for completeness, editing for consistency, editing for accuracy and editing for homogeneity.

It is always a tough task for the researcher to choose between primary and secondary data. Though primary data are more authentic and accurate, time, money and labor involved in obtaining these more often prompt the researcher to go for the secondary data. There are certain amount of doubt about its authenticity and suitability, but after the arrival of many government and semi government agencies and some private institutions in the field of data collection, most of the apprehensions in the mind of the researcher have been removed.

**SELF ASSESMENT QUESTIONS (SAQs)**

1. Explain primary and secondary data and distinguish between them.
   (Refer to the introduction part of this lesson.)
8. Explain different methods of collection primary data.
   (Explain direct personal, indirect oral interview, information received through agencies etc.)
3. Explain merits and demerits of different methods of collecting primary data.
   (Refer the methods of collecting primary data)
4. Explain the different sources of secondary data and precaution in using secondary data.
5. What is editing of secondary data? Why is it required?
6. What are the different types of editing of secondary data?

**GLOSSARY OF TERMS**

**Primary Source**: It is one that itself collects the data.