Notes

318en06

6

COLLECTION AND CLASSIFICATION OF DATA

In the previous lesson, you have learnt about the meaning and scope of statistics and its need in Economics. In this lesson you will learn about the techniques of collecting, organizing and condensing of data. These techniques are necessary for making the statistical data meaningful.

OBJECTIVES

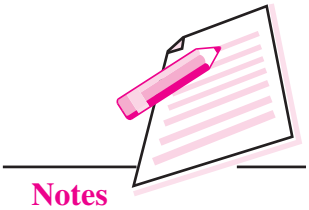
After completing this lesson, you will be able to:

- distinguish between primary and secondary data;
- list the methods of collecting primary data;
- give some examples of sources of secondary data;
- explain the concepts of an array, frequency array and frequency distribution;
- state different methods of constructing frequency distribution; and
- construct simple and cumulative frequency distributions from a given data.

6.1 COLLECTION OF DATA

(a) Primary vs. Secondary Data

Data can be collected in two different ways. One way is to collect data directly from the respondent. The person who answers the questions of the investigator is called respondent. Statistical information thus collected is called primary data and the source of such information is called primary source. This data are original because it is collected for the first time by the investigator himself. For example, if the investigator collects the information about the salaries of National Institute of



Open Schooling employees by approaching them, then it is primary data for him.

Another way is to adopt the data already collected by someone else. The investigator only adopts the data. Statistical information thus obtained is called secondary data. The source of such information is called secondary source. For example, if the investigator collects the information about the salaries of employees of National Institute of Open Schooling from the salary register maintained by its accounts branch, then it is secondary data for him.

(b) Methods for collecting primary data

There are several methods for collecting primary data. Some of which are:

1. **Direct personal interview :** In this method investigator (also called interviewer) has to be face-to-face with the person from whom he wants information. The person from whom this information is collected to called respondent.
2. **Indirect oral investigation :** Under this method data are collected through indirect sources. Under this method questions relating to the inquiry are put to different persons and their answers are recorded. This method is most suitable when the person from whom the information is sought is either unavailable or unwilling.
3. **Questionnaire method :** In this method a list of questions called questionnaire is prepared and sent to respondents either through post or given personally to them. This method is suitable where the field of inquiry is wide.

   There are some advantages of using primary data. The investigator can collect the data according to his requirement. It is reliable and sufficient for the purpose of investigation. However, it suffers from disadvantages also in that it involves a lot of cost in terms of money, time and energy. This make unsuitable when field of enquiry is very very large. Many a times with some modifications, same purpose may be served by using data collected by other persons or agencies.

(c) Sources of secondary data

As already discussed secondary data are not collected by the investigator himself but they are obtained by him from other source. Broadly, there area two sources: (a) Published data and (b) Unpublished data.

I. Published Sources

There are certain agencies which collect the data and publish them in the form of either regular journals or reports. These agencies/sources are known as published sources of data.

In India some of the published sources are:

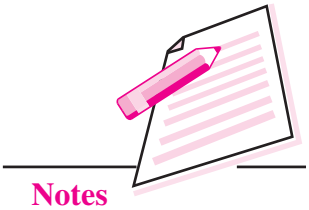
1. **Central Statistical Organisation (CSO) :** It publishes data on national income, savings, capital formation etc., in a publication called National Accounts Statistics.
2. **National Sample Survey Organisation (NSSO) :** This organization which is under the Ministry of Finance provides data on all aspects of national economy, such as agriculture, industry, employment and poverty etc.
3. **Reserve Bank of India (RBI) :** It publishes financial statistics. Its publications are Report on Currency and Finance, Reserve Bank of India Bulletin and Statistical Tables Relating to Banks in India etc.
4. **Labour Bureau :** Its publications are Indian Labour Statistics, Indian Labour Year Book and Indian Labour Journal.
5. **Population Census :** It is undertaken by the office of the Registrar General, Census of India, Ministry of Home Affairs. It provides us statistics on population, per capita income, literacy rate etc.
6. **Papers and Magazines :** Journals like 'Capital', 'Commerce', Economic and Political Weekly', and newspapers likes 'The Economic Times' etc. also publish important statistical data.

II. Unpublished Sources

Secondary data are also available from unpublished sources, because all statistical data is not always published. For example, information recorded in various government and private offices, studies made by research scholars etc. can be important sources of secondary data.

INTEXT QUESTION 6.1

1. Fill in the blanks with suitable words given in brackets against each:
   (a) ..... data are original. (Primary, Secondary)
   (b) ..... Primary data are collected by the ..... himself. (respondent, vestigator)
   (c) CSO publishes data on ..... (national income, population)
2. State whether the following statements are true or false:
   (a) Secondary data are collected by the investigator himself.
   (b) Reserve Bank of India Bulletin represents an unpublished source of data.
   (c) A person from whom an investigator tries to get information is called respondent.



6.2 ORGANISING AND CONDENSING DATA

Suppose a statistical investigator wants to analyse the marks obtained by 40 students in a class. He collects data and finds that marks obtained by 40 students in the class are:

20	25	28	27	34	31	30	32	33	40
43	43	40	43	42	43	42	45	43	47
48	46	47	48	46	49	58	54	56	50
53	51	39	38	36	38	35	35	37	

Put yourself in the position of investigator. In which aspect of this data you will be interested? Perhaps you would be interested in knowing the highest marks obtained by any student. You may also be interested to know the lowest marks obtained by a student. Another point of interest can be the marks level around which most of the students have obtained.

The above data are unorganized. To refine this data for comparison and analysis it should be arranged in an orderly sequence or into groups on the basis of some similarity. This whole process of arranging and grouping the data into some meaningful arrangement is a first step towards analysis of data. Data can be arranged in two forms: (a) Arrays and (b) Frequency distributions.

(a) Arrays

A method of presenting an individual series is a simple array of data. An orderly arrangement of raw data is called 'Array'. Arrays are of two types: (i) Simple array, and (ii) Frequency array.

(i) **Simple Array :** A simple array is an arrangement of data in ascending or descending order. Let us construct the simple arrays of the data about the marks of 40 students. The data in table 6.1 is arranged in ascending order and in table 6.2 in descending order.

Table 6.1: Ascending Array of the Marks obtained by 40 students in class

20	35	42	47
25	36	43	48
27	37	43	48
28	38	43	49
30	38	43	50

31	39	43	51
32	40	45	53
33	40	46	54
34	40	46	56
35	42	47	58

Table 6.2: Descending Array of the Marks obtained by 40 students in class

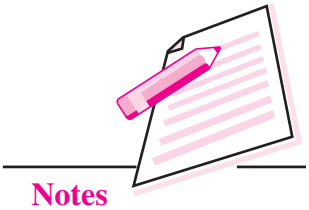
58	47	42	35
56	46	40	34
54	46	40	33
53	45	40	32
51	43	39	31
50	43	38	30
49	43	38	28
48	43	37	27
48	43	36	26
47	42	35	20

The above arrays reveal information on two points clearly. One, the highest marks obtained by any student are 58. Two, the lowest marks obtained by any student are 20.

Organising the data in the form of simple array is convenient if number of items is small. As the number of items increase the series becomes too long and unmanageable. As such there is need to condense data. Making a frequency array is one method of condensing data.

(ii) Frequency Array : Frequency array is a series formed on the basis of frequency with which each item is repeated in series. The main steps in constructing frequency array are:

1. Prepare a table with three columns-first for values of items, second for tally sheet and third for corresponding frequency. Frequency means the number of times a value appears in a series. For example in table 6.1 the marks 43 appears five times. So frequency of 43 is 5.



2. Put the items in first column in a ascending order in such a way that one item is reordered once only.
3. Prepare the tally sheet in second column marking one bar for one item. Make blocks of five tally bars to avoid mistake in counting. Note that every fifth bar is shown by crossing the previous four bars like e.g., ////.
4. Count the tally bars and record the total number in third column. This column will represent the frequencies of corresponding items.

Let us now explain construction of frequency array of the marks obtained by 40 students. In table 6.3 data about the marks is arranged in an ascending order in first column. It helps to find not only the maximum and minimum values but also makes it easy to draw bars.

Now for each mark level make one bar (/) in second column and cross the item from the data.

Table 6.3 Frequency array of marks obtained by 40 students

Marks(X)	Tally Sheet	Frequency
20	/	1
25	/	1
27	/	1
28	/	1
30	/	1
31	/	1
32	/	1
33	/	1
34	/	1
35	//	2
36	/	1
37	/	1
38	//	2
39	/	1
40	///	3
41	//	2
42	//	2
43	𝍸	5
45	/	1

46	//	2
47	//	2
48	/	1
49	/	1
50	/	1
51	/	1
53	/	1
54	/	1
56	/	1
58	/	1
		Total Frequency = 40

The main limitations of frequency array is that it does not give the idea of the characteristics of a group. For example it does not tell us that how many students have obtained marks between 40 and 45. Therefore it is not possible to compare characteristics of different groups. This limitation is removed by frequency distribution.

INTEXT QUESTIONS 6.2

Fill in the blanks with appropriate word from the brackets:

(a) A simple array is an arrangements of data in ..... (only ascending order, only descending order, either ascending or descending order).

(b) Organising data in simple array is convenient if number of items are ..... (large, small).

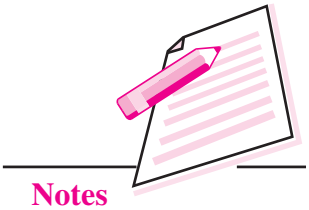
(c) Arranging the data in the form of ..... array is more convenient if number of items are large. (simple, frequency).

(d) Frequency array ..... the idea of characteristics of a group. (gives, does not give)

6.3 FREQUENCY DISTRIBUTION

Data in a frequency array is ungrouped data. To group the data we need to make a 'frequency distribution'. A frequency distribution classifies the data into groups. For example, it tells us how many students have secured marks between 40 and 45.

Before constructing frequency distribution, it is necessary to learn the following important concepts (see tables 6.4 and 6.5) :



1. **Class :** Class is a group of magnitudes having two ends called class limits. For example, 20-25, 25-30 etc. or 20-24, 25-29 etc. as the case may be, each represents a class.
2. **Class Limits :** Every class has two boundaries or limits called lower limit (L_1) and upper limit (L_2). For example in the class (20-30) $L_1 = 20$ and $L_2 = 30$.
3. **Class Interval :** The difference between two limits of a class is called class interval. It is equal to upper limit minus lower limit. It is also called class width. Class interval = $L_2 - L_1$. For $30 - 20 = 10$.
4. **Class Frequency :** Total number of items falling in a class that is having the value within L_1 and L_2 is class frequency. For example in table 6.4 class frequency in class (40-45) is 10. Similarly in class (50-55) the frequency is 4.
5. **Mid-Point/Mid-Value(M.V.) :** The mid-value of the class interval of a class also called as mid-point is obtained by dividing the sum of lower limit and upper limit of the class by 2. It is the average value of two limits of a class. It falls just in the middle of a class is

$$\text{M.V.} = \frac{L_1 + L_2}{2}$$

For example, the mid-value of class (20-30) is $\frac{20+30}{2} = 25$

Construction of Frequency Distribution

Frequency distributions can be constructed in many ways. We will explain here the construction of the following types:

(a) Exclusive series

(b) Inclusive series

(c) Open end classes

(d) Cumulative frequency

While constructing a frequency distribution same steps are to be taken which we have followed in the frequency array. The only difference is that we record classes like (20-25), (25-30), (30-35)….(55-60) etc., in first column in place of absolute items like 20, 25,..56,58 etc.

(a) Exclusive series: In this type one of the class limits (generally upper limit L_2) is excluded while making a tally sheet. Any item having the value equal to the upper limit of a class is counted in the next class. For example, in a class of (20-25) all items having the value of 20 and more but less than 25 will be counted in this class.

Notes

Item having the value of 25 will be counted in next class of (25-30) as is clear from the following example, Using the same data as given in making a frequency array and taking class interval of 5, a frequency distribution of exclusive type will be as under:

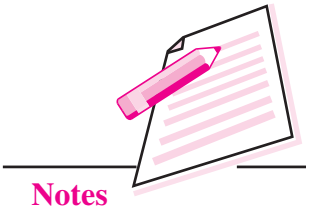
Table 6.4: Construction of Frequency Distribution – "Exclusive Type"

Class	Tally Sheet (Tallies)	Frequency (f)
20-25	/	1
25-30	///	3
30-35	𝍸	5
35-40	𝍸 //	7
40-45	𝍸 𝍸	10
45-50	𝍸 ///	8
50-55	////	4
55-60	//	2
		Total Frequency = 40

(b) Inclusive Series : In this type the lower limit of next class is increased by one over the upper limit of previous class. Both the items having value equal to lower and upper limit of a class are counted or included in the same class. That is why such a frequency distribution is called inclusive type. For example in the class (20-24) both 20 and 24 will be included in the same class. Similarly in the class (40-44) both 40 and 44 will be included. The following table has been formed on the basis of same data as taken in the exclusive type.

Table 6.5: Construction of Frequency Distribution – "Inclusive Type"

Class	Tally Sheet (Tallies)	Frequency (f)
20-24	/	1
25-29	///	3
30-34	𝍸	5
35-39	𝍸 //	7
40-44	𝍸 𝍸	10
45-49	𝍸 ///	8
50-54	𝍸	4
55-59	//	2
		Total Frequency = 40



(c) Open-end Classes : Open-end frequency distribution is one which has at least one of its ends open. You will observe that either lower limit of first class or upper limit of last class or both are not given in such series. In table 6.6 the first class and the last class i.e. below 25 and 55 and above are open-end classes.

Table 6.6: Open-end Classes Frequency Distribution

Class	Tally Sheet	Frequency (f)
Below-25	/	1
25-30	///	3
30-35	𝍸	5
35-40	𝍸 //	7
40-45	𝍸 𝍸	10
45-50	𝍸 ///	8
50-55	////	4
55 and above	//	2
		Total Frequency = 40

(d) Unequal Classes : In case of unequal classes frequency distribution, the width of different classes (i.e. L_2-L_1) need not be the same. In table 6.7, the class (30 – 40 has width 10 while the class (40-55) has width 15.

Table 6.7: Unequal Classes Frequency Distribution

Class	Tally Sheet	Frequency (f)
20-25	/	1
25-30	///	3
30-40	𝍸 𝍸 //	12
40-55	𝍸 𝍸 𝍸 𝍸 //	22
55-60	//	2
		Total Frequency = 40

(e) Cumulative Frequency : A 'Cumulative Frequency Distribution' is formed by taking successive totals of given frequencies. This can be done in two ways:

(i) From above, such as 1,4 (i.e. 1+3), 9(i.e. 4+5), 16 (i.e. 9+7), and so on.

Notes

Such a distribution is called 'Less-than' culmulative frequency distribution. It shows the total numbers of observations (frequencies) having less than a particular value of the variable (here marks). For example, there are 4 (i.e. 1+3) students who got marks less than 30; 9 (i.e. 4+5) students who got marks less than 35 and so on. Table 6.8 gives the less-than cumulative frequency distribution.

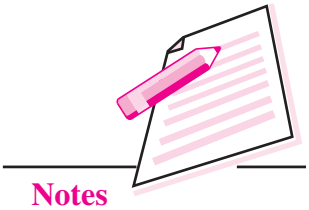
Table 6.8: 'Less-than' Cumulative Frequency Distribution

Marks	Cumulative Frequency (cf)
Less than 25	1
Less than 30	4 (1+3)
Less than 35	9 (4+5)
Less than 40	16 (9+7)
Less than 45	26 (16+10)
Less than 50	34 (26+8)
Less than 55	38 (34+4)
Less than 60	40 (38+2)

(ii) From below, such as 2,6 (i.e. 2 + 4), 14 (i.e. 6+8), 24 (i.e. 14 + 10) and so on. Such a distribution is called 'More-than' cumulative frequency distribution. It shows the total number of observations (frequencies) having more than a particular value of the variable (here marks). For example there are 6 (i.e. 2 + 4) students who got marks more than 50, 14 (i.e. 2 + 4 + 8) students who got marks more than 45 etc. See table 6.9.

Table 6.9: 'More-than' Cumulative Frequency Distribution

Marks	Cumulative Frequency (cf)
More than 20	40
More than 25	39 (40-1)
More than 30	36 (39-3)
More than35	31 (36-5)
More than 40	24 (31-7)
More than 45	14 (24-10)
More than 50	6 (14-8)
More than 55	2 (6-4)



Notes

INTEXT QUESTIONS 6.3

Fill in the blanks with appropriate word from the brackets.

(a) Frequency distribution ..... data into groups.
(classifies, does not classify)

(b) The difference between two limits of a class is called .....
(Class limit, class interval).

(c) In the exclusive type frequency distribution an item having value equal to the upper limit is counted in the ..... class. (same, next)

(d) In the inclusive type frequency distribution an item having value equal to the upper limit is counted in the ..... class.

(e) Preparing a frequency distribution by taking 'successive totals' of frequencies is called ..... frequency distribution. (open-ended, cumulative)

ACTIVITY

1. Visit children in your neighbourhood and record the age of at least 30 of them and then construct a frequency distribution of both exclusive as well as inclusive types.
2. From daily newspapers, record maximum temperature of your city for 30 days. Prepare at frequency distribution of both exclusive as well as inclusive types with a class interval of 1.5 degrees Celsius and with at least 5 classes.

WHAT YOU HAVE LEARNT

- For any statistical enquiry, data can be collected in two ways:

  (a) either by the investigator himself. This is called primary data.

  (b) or he can obtain it from other sources i.e. data already collected by others. This is called secondary data.

- In India there are several sources of getting secondary data. Some of these are: Central Statistical Organisation (CSO), National Sample Survey Organsation, (NSSO), Reserve Bank of India (RBI), etc.
- Collected data are normally in a disorderly form. Therefore, they have to be arranged in some orderly form or sequence. This is called arrangement of data.
- The various ways of arrangement of data are: a simple array, a frequency array

Notes

and frequency distribution. Arrays can be (i) simple array or (ii) frequency array.

- When simple frequencies are successively totaled, we get what is called cumulative frequency distribution.
- To get frequency distribution we have to make use of tally sheet.
- Formation of frequency distribution requires important decisions regarding number of classes, class limits and class width etc.
- A class is a group of magnitudes having two ends called class limits (L_1 and L_2), L_1 being lower limit and L_2 the upper limit.
- Total number of cases falling in a particular class is called class frequency.
- We can form the following types of frequency distributions:
  - (a) exclusive type where the upper limit of the class is excluded and put in the next class.
  - (b) inclusive type where the upper limit of the class is included in the same class.
  - (c) Open-end like (below 25) and (55 and above).
  - (d) Unequal classes where class width or class interval of different classes is different like (20 – 25), (25-30), (30-40)….
  - (e) Cumulative 'Less-than' and 'More-than' where simple frequencies are successively totaled from above and from below respectively.

Cumulative: means successive totaling. That is, something increasing in quantity by one addition after another.

Condensation: putting huge quantity of data in some useful, short or brief form without losing its utility.

Respondent: is a person who responds or answers to some questions raised. When an investigator approaches a person with a questionnaire, the person who answers these questions is called respondent.

Sequence: in ordinary language means connected line of events or ideas. In statistics it means a series formed on some such principle e.g. sequence of numbers from, say 2, with a difference of say, 5 That is 2, 7, 12, 17, 22 etc. These numbers are in ascending order of sequences. If we put them in the form (reverse sequence) …….. 22, 17, 12, 7, 2, these are called in descending order of sequence.

Tally Sheet: is a statement where occurrence of each value of a series is recorded by making one bar. (/)

Data: means statistical information on population, employment, prices, exports, imports etc. that has been collected, analysed and published by government departments, commercial and industrial associations, and other research agencies.

TERMINAL EXERCISE

1. Distinguish between primary and secondary data. Describe the methods for collecting primary data.
2. What is secondary data? Name some of its sources in India.
3. Distribution between simple array and frequency array with examples.
4. On the basis of the following data about the wages of 20 workers in a factory, prepare a frequency array; 450, 580,600, 480, 540, 620, 400, 475, 500, 480, 620, 480, 570, 600, 650, 410, 550, 600, 650, 450.
5. Explain the concept of 'frequency distribution'. How is it different from 'frequency array.?
6. On the basis of data in question 4, prepare a frequency distribution by exclusive method.
7. Distinguish between 'exclusive method' and 'inclusive method' of frequency distribution with examples.
8. Write short notes on:
   (a) Open-end frequency distribution.
   (b) Frequency distribution with unequal classes.
   (c) Cumulative frequency distribution.

ANSWERS TO INTEXT QUESTIONS

6.1

1. (a) Primary (b) Investigator (c) National income
2. (a) False (b) False (c) True

6.2

(a) either ascending or descending order (b) small
(c) frequency (d) does not give

6.3

(a) classifies (b) class interval
(c) next (d) same
(e) cumulative.

Notes

Terminal Exercise

1. Read section 6.1(a) and (b)
2. Read section 6.1 (a) and (c)
3. Read section 6.2(a)

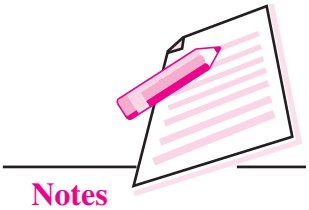
(i) Arrange the data in ascending order:

400	480	550	600
410	480	570	620
450	480	580	620
450	500	600	650
475	540	600	650

(ii) Prepare a tally sheet.

Income (₹.)	Tallies	Frequency (f)
400	/	1
410	/	1
450	//	2
475	/	1
480	///	3
500	/	1
540	/	1
550	/	1
570	/	1
580	/	1
600	///	3
620	//	2
650	//	2
		Total Frequency = 20

5. Read section 6.2 and 6.3



Notes

6. First two steps have already been explained in answer to question 4. The third step is as follows.

Income groups (Rs.)	Frequency (f)
400-450	2
450-500	6
500-550	2
550-600	3
600-650	5
650-700	2
	Total Frequency = 20

7. Read section 6.3 (a) and (b)

8. (a) Read section 6.3 (c)

   (b) Read section 6.3 (d)

   (c) Read section 6.3 (e)