2. COLLECTION, EDITING AND PRESENTATION OF DATA

2.1 Statistical data : Primary data and Secondary data

Statistical data are classified as primary and secondary, depending on the source of the data.

(i) Primary data : The statistical data which are gathered directly from the field of investigation for the desired purpose are called primary data. It is obvious that these data are original in nature.

A doctor, interested in the weights of his patients, records their weights using a machine. These data are primary to the doctor.

This type of data can be used with greater confidence, because the enquirer himself decides upon coverage of the data, the definitions used, the method of collection of the data, etc. and, as such, he has a measure of control on the reliability of the data.

But collection of such data requires more money, man-power and time.

(ii) Secondary data : The statistical data which have already been collected by some agency and are compiled from that source by the enquirer for his use are called secondary data. In other words, data collected by one when used by another, or collected for one purpose when used for a different one are termed as secondary data. It is evident that secondary data are not original.

The census data collected from census reports by a research scholar for his study are secondary data to the scholar.

Collection of secondary data is cheaper as it involves less man-power and time.

But secondary data usually contain errors due to transcription, rounding, etc. and, hence, are hardly reliable. In making use of such data, the enquirer has to be particularly careful about the coverage of the data, the definitions on which they are based, method of collection of data and their degree of reliability.

From the above discussion, it is clear that the same data which are primary for one may be secondary for someone else.

2.2 Collection of data

The collection of data is the primary task at the outset of any statistical activity. In this context, most often we come across the terms *schedule*, *questionnaire*, etc. Firstly, these commonly used terms are explained here.

A. Questionnaire : The term questionnaire means a list of certain systematically arranged questions pertaining to the subject of enquiry. It is necessary that a questionnaire is designed with due care so that necessary data may be easily collected. A draft questionnaire is usually formed in the first stage of a survey and tried over a group of people to detect any kind of fault in preparing the questions. The questionnaire is then modified, if necessary, with the help of trial data.

A good questionnaire must possess the following important characteristics :

(i) Questions should be relevant to the subject and put in simple language.

(ii) The number of questions must be limited; otherwise, the respondents may be reluctant to fill in the questionnaire.

(iii) Questions should not be dubious in meaning.

(iv) Multiple choice type questions should mostly be included to facilitate the informants in answering the queries.

(v) Questions which may hurt the sentiment or ego of the respondents should be avoided.

(vi) A few questions that enable one to cross examine the consistency of replies to some particular items (like date of birth and age in years) should be included.

B. Schedule : In the schedule one finds a list of items on which information will be collected; the exact forms of the questions to be put to the informants are not given and the task of questioning and eliciting the desired information is left to the investigator.

C. Pilot survey : For preparing the questionnaire or the schedule, occasionally, it becomes necessary to collect some broad information about the field of enquiry. To meet the purpose, a small-scale survey is conducted prior to the main survey, and this is called a pilot survey.

We now discuss the different methods adopted for the collection of primary data.

(1) Interview method : In this method, the investigators gather the desired data directly from the field of enquiry. Each investigator, carrying with him some questionnaires/schedules, meets the informants of the area which is allotted to him and tactfully collects the relevant information from them by way of interrogation. The investigator has the responsibility to convey the purpose of the enquiry to the respondents and also to record their answers to the different questions.

It is noteworthy that the investigators are imparted suitable training before performing the survey.

Advantages

(a) It can be used even if the informants are illiterate.

(b) In this method, trained investigators find the scope of cross-examining the respondents and, as its effect, it is possible to gather accurate and reliable data.

(c) The chance of non-response is fairly diminished due to the fact that the investigators personally collect the information.

Disadvantages

(a). The informants may feel hesitant in furnishing correct answers to certain typical questions (related to income, age, education, etc.) before the interviewers.

(b) It is a time consuming procedure.

(c) Some responses may be affected by interviewer-bias.

(d) This method is highly expensive.

(2) Mail questionnaire method : The questionnaire is the main tool used in this approach. A questionnaire, along with a letter informing the objective of the investigation and return postage is sent to each of the informants by post. The individuals are requested to mail the appropriately completed questionnaires within a stipulated period. Specific instructions (if any) are also sent to the informants during this enquiry. In this procedure, the respondents are generally assured that their information would remain secret.

Advantages

(a) It is not a costly method.

(b) It does not require long period for the desired purpose.

(c) The available information are free from interviewer-bias.

(d) In this technique, informants freely furnish genuine answers, especially to confidential questions.

Disadvantages

(a) There is large amount of non-response due to unwillingness or for some kind of fear of a section of informants.

(b) This method is applicable only for people who have enough education to comprehend the significance of the enquiry.

(c) It is possible to get back some questionnaires that are not fully filled in.

(3) Direct personal observation method : In this method, the enumerators visit the field of investigation and gather the necessary information by observation (i.e. by seeing, counting or measuring). Here the enumerators are not to depend on others for the information and, as such, the data are likely to be much reliable. Of course, the authenticity of data mainly depends on the honesty, sincerity and capability of keen observation of the enumerators.

Advantages

(a) It is possible to collect genuine information.

(b) The scope of bias due to respondents is completely eliminated.

Disadvantages

(a) It is an expensive method.

(b) The enumerators must be efficient and loyal to their task, otherwise this method may fail to yield correct information.

(c) This procedure is not appropriate for a large area.

(4) Indirect oral investigation method : The requisite data are collected from some indirect source. Persons possessing correct and elaborate knowledge about the problem at hand are selected and they are interrogated for gathering the desired data.

This method is mostly used by the commissions of enquiry and committees appointed by government for collecting primary data.

It should be mentioned that the accuracy of data mainly depends on the impartial attitude of the source of information and also on the honesty of the investigators.

Remarks

Apart from the pre-explained methods, in some situations, locally appointed agents and correspondents collect the necessary information and transmit the same to the appropriate agency. This method is usually applied in those cases where the information are required regularly. In particular, media sectors follow this technique.

2.3 Scrutiny of data

After collection of the data we should examine them thoroughly to see if they are correct. This scrutiny is very important, because however sophisticated statistical techniques may be used in analysing the data, the results will be misleading if the data were erroneous. When the data are of the secondary type, scrutiny assumes more importance.

In some cases, faulty data are readily detected, without any effort. Such are the cases when we find that weight of a person is stated to be 890 kg, or the birth day of a person is recorded as November 31.

Again, there may be figures which are not impossible, but very likely wrong. If a person claims that his consumption of rice per day is 1 kg, or if a student's age is stated to be 12 when his father's age is 28, or when the number of sons and daughters of a couple is given to be 15, we become suspicious and proceed for checking.

In certain situations, we may come across data which look all right at first sight but which, on through investigation, may be found to be inconsistent. When we are given two or more related series of figures, the data may be scrutinised by comparing the series for internal consistency. For example, when the area, population and density of population are given for different cities in a state, one may compute density = population/area, and compare the computed figures with the given figures for densities, and in this way the compatibility or otherwise of the three series may be judged. Similar is the case when we are given, for each family in a village, the total income, the total expenditure and the total savings in a year so that one can use their relation, savings = income – expenditure, for checking.

When figures are given in a two-way table, for example, with several rows and columns, the accuracy of the data may be checked by comparing the given row totals and column totals with computed totals of rows and columns. If computed totals of ith row and jth column do not tally with the corresponding given figures, one will suspect that the figure in the (i, j) cell is faulty.

One common problem faced by every statistician handling data is that, there may be one or more observations which are markedly different from other observations in the data. This may be due to observational error or recording error. But after scrutiny the enquirer may have the impression that the discrepant observations come from a distribution different from the distribution yielding the other observations. There are objective methods, based on the given data, for deciding whether such observations, referred to as *outliers*, can be taken as members of the same distribution with others or not.

It is to be noted that no hard and fast rules may be laid down for the scrutiny of data. One must use one's common sense, intelligence, judgement and whatever knowledge one may have about the field of investigation to assess the reliability of the data.

2.4 Frequency data and non-frequency data

It is quite natural that one cannot easily understand the significance of the collected data at a glance. To bring into focus the salient features of the data, it is necessary to present them in a neat systematic form. Different modes of presentation of data are available depending on whether they are frequency data or non-frequency data.

Suppose we are given the values of one or more variables-like yield of paddy, population, literacy rates, etc. — for different periods of time for a region. Here generally we study the nature of changes of the variables over time. Such data are called *time-series data*. Again, we may have values of one or more variables for different individuals in a group for the same period of time — like tea production in different tea-producing countries or budget estimates of receipts of a country from different heads in a year, etc. and we study the change in the values of the variables from country to country or from item to item. Data of this kind are called *spatial-series data*. In both the cases, study is conducted keeping in view the identity of individual values. These two kinds of data are referred to as *non-frequency data*.

Now consider another situation where we are given values of one or more variables for different individuals in a group – like sizes of different families in a region, marks obtained by different students in a test, etc. – but we are interested in the characteristics of the group formed by the individuals and not in the individuals themselves. For instance, we may be interested in the average size of the families or the percentage of students getting more than 75% of total marks. Here identity of the individuals is ignored. This kind of data are referred to as *frequency data*.

In the following sections of this chapter we shall discuss the different modes of presentation of non-frequency data and in the next chapter we shall deal with presentation of frequency data.

Following are the common techniques of presentation of non-frequency data.

(a) Textual presentation
(b) Tabulation
(c) Diagrammatic representation

2.5 Textual presentation of data

A very common method of presenting statistical data is to use paragraphs of text. This is the method employed in official reports, where the activities, plans or programmes are described in words, inserting relevant figures in between them. There is no hard and fast rule for this mode of presentation of data. But the writer must take adequate care to see that there is clarity and a logical sequence in the presentation. The text should be brief and precise.

Illustration

"... 2,652 men and 1,726 women participated in an opinion poll about a certain govt. measure. 1,460 persons, of whom 1,096 were male, voted against the measure. In all 2,225 persons voted for the measure, while 356 women were indifferent ..."

Advantages of the method **:**

(i) This mode of presentation of data has an appeal to people with a literary bent of mind.

(ii) The writer can draw attention of the reader to certain points which he considers to be of special importance.

Disadvantages **:**

(i) A person has to read the entire text before he can get a clear idea about the nature of the data. Thus, it will take considerable time, specially when the text is lengthy.

(ii) For presenting a large mass of data, this method is not appropriate.

(iii) This method will not work well when it becomes necessary to make a large number of comparisons.

(iv) In most cases, the text is monotonous and boring.

(v) Textual presentation will not readily provide data for statistical analysis.

2.6 Tabulation of data

Tabular presentation or tabulation of data is an useful mode of exhibiting the data in a compact form. The systematic presentation of data in the structure of a table comprising some rows and columns is called *tabulation.*

It should be mentioned that the construction of two or more tables may be necessary for presenting the entire data in simple and concise form.

A table has several parts which are described below.

1. Table number : An appropriate number should be assigned to a table for its identification and easy reference in future.

2. Title : It is a brief statement of the contents of the table, placed at the head of the table.

3. Stub : This is the extreme left part of the table, giving description of the matter presented in rows.

4. Caption : This is the upper part of the table giving description of different columns. The units of measurement (if any) and column numbers also belong to this component.

The title, stub and caption, taken together, form the *box head* of the table.

5. Body : It is the main component of the table which contains the numerical information.

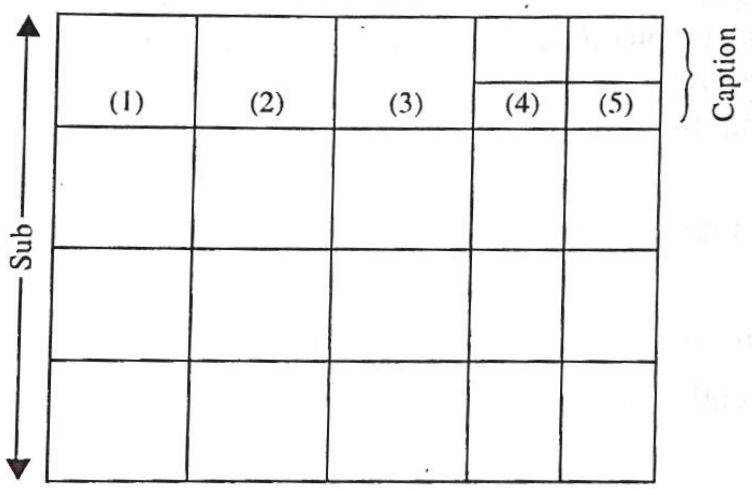
6. Footnote : This component shows the explanations (if any) of specific items.

7. Source : It refers to the origin of the available information.

Sketch of a table showing its different parts :

Table No. ....

Title ....



Source : ....

Footnote (if any) : ....

The construction of a good table is definitely an art. Although there is no rigid rule related to tabulation, a table should generally possess the following salient features.

(a) A table should be well balanced in length and breadth. In other words, it should neither be too long and narrow nor too short and broad.

(b) The data included in a table must be arranged in systematic and logical sequence.

(c) To facilitate comparison, the figures to be compared should be placed as close to each other as possible.

(d) The units of measurement (if any) for the items must be clearly specified in respective rows and columns.

(e) Generally, totals of both rows and columns should be shown in a table.

(f) If the number of rows and columns is large then each of them should be allotted a number for reference.

(g) A table must possess a simple, brief and self explanatory title.

(') A table should not involve too many abbreviations.

The merits and demerits of tabulation are as follows :

Merits

(i) It enables one to detect errors and omissions (if any) in the data.

(ii) The numerical data can be accurately presented in a table.

(iii) This technique facilitates statistical analysis.

(iv) Tabulation readily helps in comparison.

(v) It clarifies the characteristics of the data.

Demerits

(i) This mode of presentation is not comprehended by a layman.

(ii) It fails to create a lasting impression.

2.7 Diagrammatic representation of data

Graphs, charts, maps, pictures, etc. are attractive and effective means for presentation of statistical data. Diagrams are readily capable of revealing some features of the exhibited data. It should be noted that the selection of the appropriate diagram depends mainly on the nature of the given data.

Following are the important merits and demerits of the diagrammatic mode of presenting data.

Merits

(i) It is simple to understand even by laymen.

(ii) It is very essential for conveying statistical information to the general public in a short time.

(iii) In this approach, one may acquire some idea regarding the significance of the presented data at a glance.

(iv) This mode is capable of creating lasting impression.

(v) Two or more series of data can easily be compared.

Demerits

(i) Diagrams fail to represent details; they only show the general nature of the given data.

(ii) Usually, a diagram represents the figures in approximate form and, as such, in most of the cases, precision of data has to be sacrificed.

(iii) Construction of a diagram requires sufficient time.

(iv) Only limited information can be presented in a diagram.

Our subsequent discussion will relate to some of the commonly used diagrams.

1. *Line diagram* : This diagram is meant for representing chronological data. In fact, it exhibits the relationship of the variable under study with time. The successive values of the variable (e.g. sales of coffee of a company) may be specified for individual points of time (called point data) or for different periods of time (referred to as period data).