Module 3 – Data Collection and Organization

Contents

Methods of Data Collection

The task of data collection begins after a research problem has been defined and research design/
plan chalked out.

There are two types of data: primary and secondary.

The primary data are those which are collected fresh and for the first time, and thus happen to be original in character.

The secondary data are those which have already been collected by someone else and which have already been passed through the statistical process.

The researcher would have to decide which sort of data he would be using for his study and accordingly he will have to select the method of data collection.

Collection of Primary Data

We collect primary data through observation or through direct communication with respondents in one form or another or through personal interviews.

The observation method is the most commonly used method in studies relating to behavioural sciences.

Here, Observation serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability.

For instance, in a study relating to consumer behaviour, the investigator instead of asking the brand of wrist watch used by the respondent, may himself look at the watch.

Advantages:

1. Subjective bias is eliminated, if observation is done accurately.
2. Information observed relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes.
3. Independent of respondents' willingness to respond.
4. suitable in studies which deal with respondents who are not capable of giving verbal reports

Limitations:

1. Expensive method
2. Information provided is limited.
3. Unforeseen factors may interfere with the observational task
4. Some people are rarely accessible to direct observation, creating an obstacle for effective data collection.

   The following aspects should be considered during data collection by Observation:

   1. What should be observed?
   2. How the observations should be recorded?
   3. How the accuracy of observation can be ensured?

An Observation is called Structured Observation if it is characterised by

1. careful definition of the units to be observed
2. proper style of recording the observed information
3. standardised conditions of observation
4. selection of pertinent data of observation

Otherwise, the observation is called unstructured observation.

Structured observation is considered appropriate in descriptive studies.

Unstructured observation is considered appropriate in exploratory studies.

If the observer observes by making himself a member of the group he is observing so that he can experience what the members of the group experience, the observation is called as the participant observation.

When the observer observes without any attempt on his part to experience what the participants feel, the observation called as non-participant observation.

When the observer is observing in such a manner that his presence may be unknown to the people he is observing, such an observation is described as disguised observation.

The merits of the participant type of observation:

(i) The researcher is enabled to record the natural behaviour of the group.

(ii) The researcher can even gather information which could not easily be obtained if he observes in a disinterested fashion.

(iii) The researcher can even verify the truth of statements made by informants in the context of a questionnaire or a schedule.

Demerits of participant type of observation:

(i) the observer may lose the objectivity if he participates emotionally

(ii) the problem of observation-control is not solved

(iii) it may narrow-down the researcher's range of experience.

If the observation takes place according to definite pre-arranged plans, involving experimental procedure, it is termed as **controlled observation.**

In controlled observation:

1. we use mechanical (or precision) instruments as aids to accuracy and standardisation.
2. has a tendency to supply formalised data upon which generalisations can be built with some degree of assurance.

3. The interpretation is subjective.

Controlled observation takes place in various experiments that are carried out in a laboratory or under controlled conditions.

If the observation takes place in the natural setting, it may be termed as **uncontrolled observation.**

In non-controlled observation:

1. no attempt is made to use precision instruments.
2. The major aim is to get a spontaneous picture of life and persons.
3. It has a tendency to supply naturalness and completeness of behaviour, allowing sufficient time for observing.

Uncontrolled observation is resorted to in case of exploratory researches.
Primary data may be obtained by applying any of the following methods:

1. Direct Personal Interviews.
2. Telephone Interviews.
3. Information from Correspondents.
4. Mailed Questionnaire Methods.
5. Schedule Sent Through Enumerators.

Interview Method

The interview method of collecting data can be used through personal interviews and, if possible, through telephone interviews.

(a) Personal interviews: It requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons.
This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation.
He has to be on the spot and has to meet people from whom data have to be collected.
Most of the commissions and committees appointed by government to carry on investigations make use of this method.

The method of collecting information through personal interviews is usually carried out in a structured way.

Such interviews involve the use of a set of predetermined questions and of highly standardised techniques of recording.

The major advantages of personal interviews are:

(i) More information and that too in greater depth can be obtained.

(ii) Interviewer by his own skill can overcome the resistance, if any, of the respondents; the interview method can be made to yield an almost perfect sample of the general population.

(iii) There is greater flexibility under this method as the opportunity to restructure questions is always there, specially in case of unstructured interviews.

(iv) Observation method can as well be applied to recording verbal answers to various questions.

(v) Personal information can as well be obtained easily under this method.

(vi) Samples can be controlled more effectively as there arises no difficulty of the missing returns; non-response generally remains very low.

(vii) The interviewer can usually control which person(s) will answer the questions. This is not possible in mailed questionnaire approach. If so desired, group discussions may also be held.

(viii) The interviewer may catch the informant off-guard and thus may secure the most spontaneous reactions than would be the case if mailed questionnaire is used.

(ix) The language of the interview can be adopted to the ability or educational level of the person interviewed and as such misinterpretations concerning questions can be avoided.

(x) The interviewer can collect supplementary information about the respondent's personal characteristics and environment which is often of great value in interpreting results.

Drawbacks of Personal interview method are:

(i) It is a very expensive method, specially when large and widely spread geographical sample is taken.

(ii) There remains the possibility of the bias of interviewer as well as that of the respondent; there also remains the headache of supervision and control of interviewers.

(iii) Certain types of respondents such as important officials or executives or people in high income groups may not be easily approachable under this method and to that extent the data may prove inadequate.

(iv) This method is relatively more-time-consuming, specially when the sample is large and recalls upon the respondents are necessary.

(v) The presence of the interviewer on the spot may over-stimulate the respondent, sometimes even to the extent that he may give imaginary information just to make the interview interesting.

(vi) Under the interview method the organisation required for selecting, training and supervising the field-staff is more complex with formidable problems.

(vii) Interviewing at times may also introduce systematic errors.

(viii) Effective interview presupposes proper rapport with respondents that would facilitate free and frank responses.

(b) Telephone interviews: This method of collecting information consists in contacting respondents on telephone itself.

It is not a very widely used method, but plays important part in industrial surveys, particularly in developed regions.

The chief merits of such a system are:

1. It is more flexible in comparison to mailing method.
2. It is faster than other methods i.e., a quick way of obtaining information.
3. It is cheaper than personal interviewing method; here the cost per response is relatively low.
4. Recall is easy; callbacks are simple and economical.
5. There is a higher rate of response than what we have in mailing method; the non-response is generally very low.
6. Replies can be recorded without causing embarrassment to respondents.
7. Interviewer can explain requirements more easily.
8. At times, access can be gained to respondents who otherwise cannot be contacted for one reason or the other.
9. No field staff is required.

10. Representative and wider distribution of sample is possible.

The demerits of Telephone interviews are:

1. Little time is given to respondents for considered answers; interview period is not likely to exceed five minutes in most cases.
2. Surveys are restricted to respondents who have telephone facilities.
3. Extensive geographical coverage may get restricted by cost considerations.
4. It is not suitable for intensive surveys where comprehensive answers are required to various questions.
5. Possibility of the bias of the interviewer is relatively more.
6. Questions have to be short and to the point; probes are difficult to handle.

Collection of data through questionnaires

This method of data collection is adopted by private individuals, research workers, private and public organisations and even by governments.

In this method a questionnaire is sent (usually by post) to the persons concerned with a request to answer the questions and return the questionnaire.

A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms.

The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself.

The merits claimed on behalf of this method are as follows:

1. There is low cost even when the universe is large and is widely spread geographically.
2. It is free from the bias of the interviewer; answers are in respondents' own words.
3. Respondents have adequate time to give well thought out answers.
4. Respondents, who are not easily approachable, can also be reached conveniently.

5 Large samples can be made use of and thus the results can be made more dependable and reliable.

The main demerits of this system can also be listed here:

1. Low rate of return of the duly filled in questionnaires; bias due to no-response is often indeterminate.

2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.
4. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched.
5. There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

Researcher should note the following with regard to these three main aspects of a questionnaire:

1. General form: It can either be structured or unstructured questionnaire.
2. Question sequence: A proper sequence of questions reduces considerably the chances of individual questions being misunderstood. The question-sequence must be clear and smoothly-moving, meaning thereby that the relation of one question to another should be readily apparent to the respondent, with questions that are easiest to answer being put in the beginning.
3. Question formulation and wording: Question should also be impartial in order not to give a biased picture of the true state of affairs.

Collection of data through schedules

This method of data collection is very much like the collection of data through questionnaire, with little difference that schedules (proforma containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose.

These enumerators along with schedules, go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma.

Differences between questionnaires and schedules

1. The questionnaire is generally sent through mail to informants to be answered as specified in a covering letter, but otherwise without further assistance from the sender.

The schedule is generally filled out by the research worker or the enumerator, who can interpret questions when necessary.

2. To collect data through questionnaire is relatively cheap and economical since we have to spend money only in preparing the questionnaire and in mailing the same to respondents. Here no field staff required.
To collect data through schedules is relatively more expensive
since considerable amount of money has to be spent in appointing enumerators and in importing training to them.
Money is also spent in preparing schedules.
3. Non-response is usually high in case of questionnaire as many people do not respond and many return the questionnaire without answering all questions.
As against this, non-response is generally very low in case of schedules because these are filled by enumerators who are able to get answers to all questions. But there remains the danger of interviewer bias and cheating.

4. In case of questionnaire, it is not always clear as to who replies, but in case of schedule the identity of respondent is known.

5. The questionnaire method is likely to be very slow since many respondents do not return the questionnaire in time despite several reminders, but in case of schedules the information is collected well in time as they are filled in by enumerators.

6. Personal contact is generally not possible in case of the questionnaire method as questionnaires are sent to respondents by post who also in turn return the same by post.
But in case of schedules direct personal contact is established with respondents.

7. Questionnaire method can be used only when respondents are literate and cooperative, but in case of schedules the information can be gathered even when the respondents happen to be illiterate.

8. Wider and more representative distribution of sample is possible under the questionnaire method, but in respect of schedules there usually remains the difficulty in sending

enumerators over a relatively wider area.

9. Risk of collecting incomplete and wrong information is relatively more under the questionnaire method, particularly when people are unable to understand questions properly. But in case of schedules, the information collected is generally complete and accurate as enumerators can remove the difficulties, if any, faced by respondents in correctly understanding the questions.
As a result, the information collected through schedules is relatively more accurate
than that obtained through questionnaires.

10. The success of questionnaire method lies more on the quality of the questionnaire itself, but in the case of schedules much depends upon the honesty and competence of enumerators.

11. In order to attract the attention of respondents, the physical appearance of questionnaire must be quite attractive, but this may not be so in case of schedules as they are to be filled in by enumerators and not by respondents.

12. Along with schedules, observation method can also be used but such a thing is not possible while collecting data through questionnaires.

Collection of secondary data

Secondary data means data that are already available i.e., they refer to the data which have already been collected and analysed by someone else.
When the researcher utilises secondary data, then he has to look into various sources from where he can obtain them.
Secondary data may either be published data or unpublished data.
Usually published data are available in:

(a) various publications of the central, state are local governments;
(b) various publications of foreign governments or of international bodies and their subsidiary organisations;
(c) technical and trade journals;
(d) books, magazines and newspapers;

(e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.;

(f) reports prepared by research scholars, universities, economists, etc. in different fields;

(g) public records and statistics, historical documents, and other sources of published information.

The sources of unpublished data are :
Diaries
Letters
unpublished biographies
autobiographies
from scholars and research workers, trade associations, labour bureaus and other public/ private individuals and organisations.

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study.

In this connection Dr. A.L. Bowley very aptly observes that it is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticise arguments that can be based on them.

By way of caution, the researcher, before using secondary data, must see that they possess following characteristics:

1. Reliability of data: The reliability can be tested by finding out such things about the said data:

(a) Who collected the data?

(b) What were the sources of data?

(c) Were they collected by using proper methods ?

(d) At what time were they collected?

(e) Was there any bias of the compiler?

(t) What level of accuracy was desired? Was it achieved ?

2. Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry.

Hence, if the available data are found to be unsuitable, they should not be used by the researcher.

In this context, the researcher must very carefully scrutinise the definition of various terms and units of collection used at the time of collecting the data from the primary source originally.

Similarly, the object, scope and nature of the original enquiry must also be studied.

If the researcher finds differences in these, the data will remain unsuitable for the present enquiry and should not be used.

3. Adequacy of data: If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher.

The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry.

The already available data should be used by the researcher only when he finds them reliable, suitable and adequate.

But he should not blindly discard the use of such data if they are readily available from authentic sources and are also suitable and adequate for in that case it will not be economical to spend time and energy in field surveys for collecting information.

Data Organization

A systematically organized data is very important for future analysis. When you work with data every day, its organization is obvious to you, but it may be hard to understand to others who know nothing about the project.

A good logical system of data organization helps to share and exchange data. When deciding how to organize your data, think about the nature of data. It is also possible to organize chronologically, especially when you work with both real-time and historical data.

The data organization methods include:

1. **Alphabetical organization**

This is probably the first option that many people consider.

It is easy and fast to organize the data alphabetically.

However, data organization is not just for the purpose of storage of the information, but its retrieval is important as well.

Organizing information by alphabetical order works excellently only if people know some specific terms or topics that they are looking for. In such a case, accessing some particular content will be more like **'Taking a walk in the park'**. Since you know that the topic you are searching for begins with letter 'T', you will not have to waste time going through the contents beginning with the other letters.

This method of data organization will be more or less useless if the individual does not know the topic that they are looking for.

2. **Location**

Data can be organized by showing a visual depiction of some physical space.

Maps are the most common ways to organize information based on location.

Consider maps like those of some college campus or the shopping mall directories.

They give you a mental image of where a particular shop or lecture hall is located in relation to another.

Organizing data based on location helps to show the relationships between the various types of content that are relevant to each other.

3. **Time (Chronological order)**

If you want to find information in a chronological order, then organizing the information based on the time it was created is the best method to use.

This method is fantastic because it can show you how things happen over a fixed period of time.

4. **Hierarchy**

Hierarchies are beneficial when you want to show how one piece of information is related to another one in the order of importance or their ranks.

They are used in organizational charts when you want to show who should report to whom in a human resource department.

They can also be used to show scale, for instance things like biggest to smallest or lightest to heaviest.

You can organize the data in ascending or descending order although many people choose to work with descending order.

5. Category

Categories are very useful for a variety of purposes, for example describing different types of data that are being generated by an institution.

The problem with this method is that it is so broad compared to the other methods.

You can organize the data in just about any way imaginable- by color, gender, price, shape, model etc. The options are infinite.

Data Grouping

We can group quantitative data into three different kinds of groups.

1. Single value grouping
2. Using qualitative characteristics
3. Using some commonly used numerical values

Single value grouping

In single value grouping each class has one distinct value. We can organize quantitative data into a frequency distribution like the qualitative data was organized.

Grouped data are data formed by aggregating individual observations of a variable into groups, so that a frequency distribution of these groups serves as a convenient means of summarizing or analyzing the data.

The idea of grouped data can be illustrated by considering the following raw dataset:

Table 1: Time taken (in seconds) by a group of students to answer a simple math question

20	25	24	33	13

26	8	19	31	11
16	21	17	11	34
14	15	21	18	17

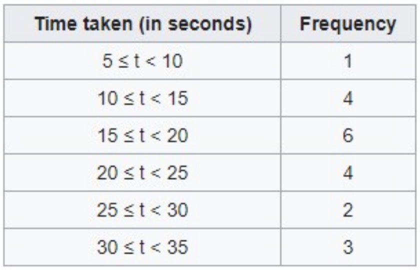
The above data can be grouped in order to construct a frequency distribution in any of several ways. One method is to use intervals as a basis.

The smallest value in the above data is 8 and the largest is 34. The interval from 8 to 34 is broken up into smaller subintervals (called class intervals).

For each class interval, the amount of data items falling in this interval is counted. This number is called the frequency of that class interval.

The results are tabulated as a frequency table as follows:

Table 2: Frequency distribution of the time taken (in seconds) by the group of students to answer a simple math question



Time taken (in seconds)	Frequency
5 ≤ t < 10	1
10 ≤ t < 15	4
15 ≤ t < 20	6
20 ≤ t < 25	4
25 ≤ t < 30	2
30 ≤ t < 35	3

2. Using qualitative characteristics

Another method of grouping the data is to use some qualitative characteristics instead of numerical intervals.

For example, suppose in the above example, there are three types of students:

1) Below normal, if the response time is 5 to 14 seconds,

2) normal if it is between 15 and 24 seconds, and

3) above normal if it is 25 seconds or more, then the grouped data looks like:

Table 3: Frequency distribution of the three types of students

	Frequency
Below normal	5
Normal	10
Above normal	5

3.Using some commonly used numerical values

Some commonly used numerical values are in fact "names" we assign to the categories.

For example, let us look at the age distribution of the students in a class. The students may be 10 years old, 11 years old or 12 years old.

These are the age groups, 10, 11, and 12.

Note that the students in age group 10 are from 10 years and 0 days, to 10 years and 364 days old, and their average age is 10.5 years old if we look at age in a continuous scale. The grouped data looks like:

Table 4: Age distribution of a class of students

Age	Frequency
10	10
11	20
12	10

Mean of grouped data

An estimate, $\bar{x}$, of the mean of the population from which the data are drawn can be calculated from the grouped data as:

$$\bar{x} = \frac{\sum f x}{\sum f}$$

In this formula, x refers to the midpoint of the class intervals, and f is the class frequency. Note that the result of this will be different from the sample mean of the ungrouped data. The mean for the grouped data in the above example, can be calculated as follows:

Class Intervals	Frequency (f)	Midpoint	f x

		(x)	
5 and above, below 10	1	7.5	7.5
10 ≤ t < 15	4	12.5	50
15 ≤ t < 20	6	17.5	105
20 ≤ t < 25	4	22.5	90
25 ≤ t < 30	2	27.5	55
30 ≤ t < 35	3	32.5	97.5
TOTAL	**20**		**405**

Thus, the mean of the grouped data is

$$\bar{x} = \frac{\sum f x}{\sum f} = \frac{405}{20} = 20.25$$

The mean for the grouped data in example 4 above can be calculated as follows:

Age Group	**Frequency (f)**	**Midpoint** **(x)**	**f x**
10	10	10.5	105
11	20	11.5	230
12	10	12.5	125
TOTAL	**40**		**460**

Thus, the mean of the grouped data is

$$\bar{x} = \frac{\sum f x}{\sum f} = \frac{460}{40} = 11.5$$

<u>Diagrammatic Representation of Data</u>

An attractive representation of statistical data is provided by charts, diagrams and pictures.

Diagrammatic representation can be used for both the educated section and uneducated section of the society. Furthermore, any hidden trend present in the given data can be noticed only in this mode of representation.

However, compared to tabulation, this is less accurate. So if there is a priority for accuracy, we have to recommend tabulation.

We are going to consider the following types of diagrammatic representation :

1. Bar diagram
2. Component Bar Charts
3. Comparative Bar Charts or Multiple Bar Charts
4. Pie chart
5. Pictogram Chart

Bar diagram

There are two types of bar diagrams namely, Horizontal Bar diagram and Vertical bar diagram.

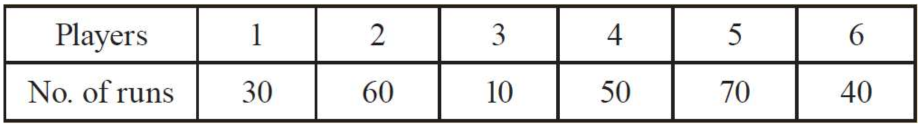
While horizontal bar diagram is used for qualitative data or data varying over space, the vertical bar diagram is associated with quantitative data or time series data.

Bars i.e. rectangles of equal width and usually of varying lengths are drawn either horizontally or vertically.

Bar diagrams for comparing different components of a variable and also the relating of the components to the whole. For this situation, we may also use Pie chart or Pie diagram or circle diagram.

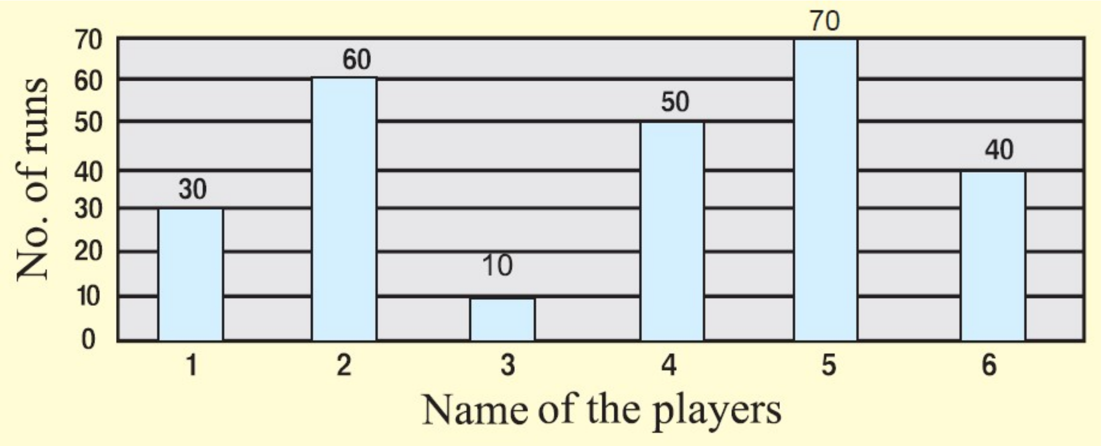
Example :

The total number of runs scored by a few players in one-day match is given.



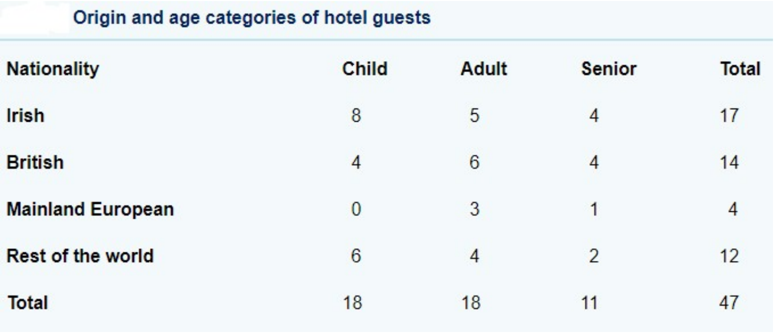
Players	1	2	3	4	5	6
No. of runs	30	60	10	50	70	40

Solution :



Component bar charts

Example: the quantity of guests from different nationalities in a hotel, broken down into the different age groups.

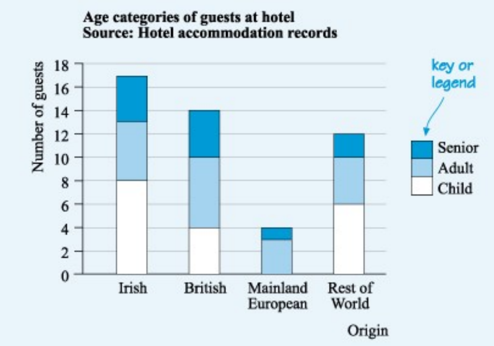


Origin and age categories of hotel guests

Nationality	Child	Adult	Senior	Total
Irish	8	5	4	17
British	4	6	4	14
Mainland European	0	3	1	4
Rest of the world	6	4	2	12
Total	18	18	11	47

To diagrammatically represent this you can split each bar into three different sections, where the length of each section represents the number of guests in that age group.

The resulting bar chart would then be a **component bar chart**, as shown in Figure. Notice that a key (or 'legend') has been added to the graph to explain the shading for the different age categories.



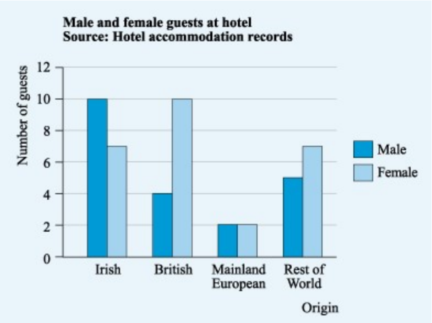
When you read a component bar chart, you need to find the height of the relevant section to determine the quantity that it represents.

For instance, the top of the section representing British adults is opposite the 10 on the vertical scale. The bottom of that section is opposite the 4. Thus, the number of British adults is the difference between the top and the bottom of the section, which is 6.

Comparative bar charts or Multiple bar charts

A comparative bar chart places bars representing sections from the same category adjacent to each other. This allows for a quick visual comparison of the data.

For example, another way of categorising the data that you have been working with so far would be to split each nationality into the number of females and males. The resulting comparative bar chart would be as shown in Figure.



It is easy to see why this is called a comparative bar chart, as it is straightforward to compare the data – for example it is easy tell that many more visitors from Britain were female.

Pie chart

In a pie chart, the various observations or components are represented by the sectors of a circle and the whole circle represents the sum of the value of all the components.

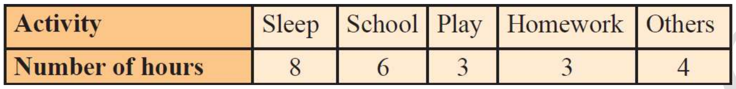
Clearly, the total angle of 360° at the center of the circle is divided according to the values of the components .

The central angle of a component is = [Value of the component / Total value] x 360°

Sometimes, the values of the components are expressed in percentages. In such cases,
The central angle of a component is = [Percentage value of the component / 100] x 360°

Pie chart - Example

The number of hours spent by a school student on various activities on a working day, is given below. Construct a pie chart using the angle measurement.



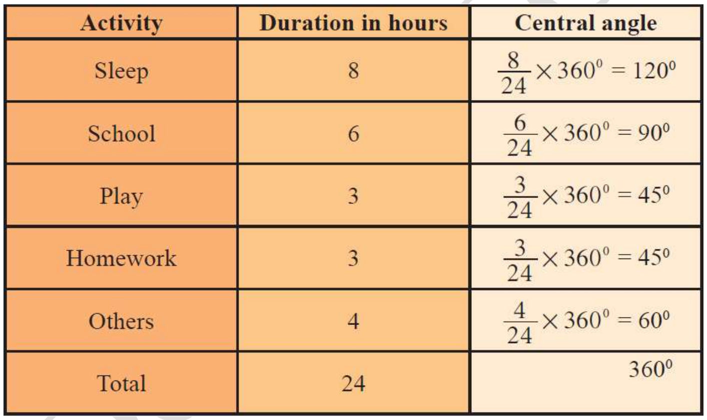
Activity	Sleep	School	Play	Homework	Others
Number of hours	8	6	3	3	4

Draw a pie chart to represent the above information.

Solution :

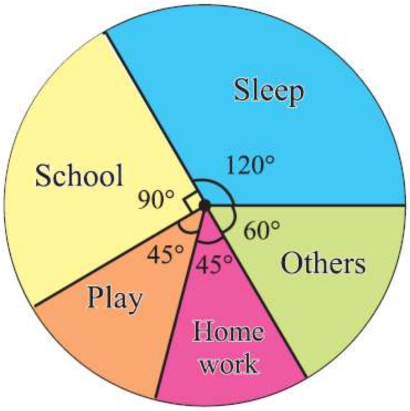
The central angle of a component is = [Value of the component / Total value] x 360°

We may calculate the central angles for various components as follows :



Activity	Duration in hours	Central angle
Sleep	8	$\frac{8}{24}\times 360^{0} = 120^{0}$
School	6	$\frac{6}{24}\times 360^{0} = 90^{0}$
Play	3	$\frac{3}{24}\times 360^{0} = 45^{0}$
Homework	3	$\frac{3}{24}\times 360^{0} = 45^{0}$
Others	4	$\frac{4}{24}\times 360^{0} = 60^{0}$
Total	24	360^{0}

From the above table, clearly, we obtain the required pie chart as shown below.

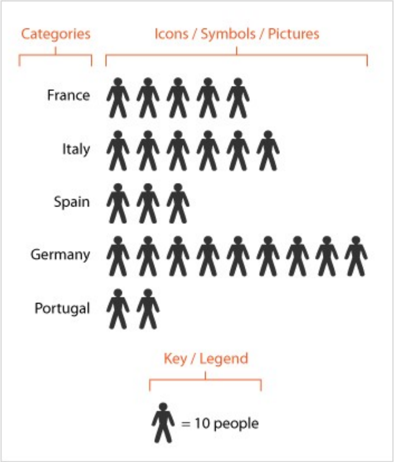


Pictogram Chart

Also known as Pictograph Chart, Pictorial Chart, Pictorial Unit Chart, Picture Graph.

Pictogram Charts use icons to give a more engaging overall view of small sets of discrete data.

Typically, the icons represent the data’s subject or category, for example, data on population would use icons of people. Each icon can represent one unit or any number of units (e.g. each icon represents 10).



Data sets are compared side-by-side in either columns or rows of icons, to compare each category to one another.

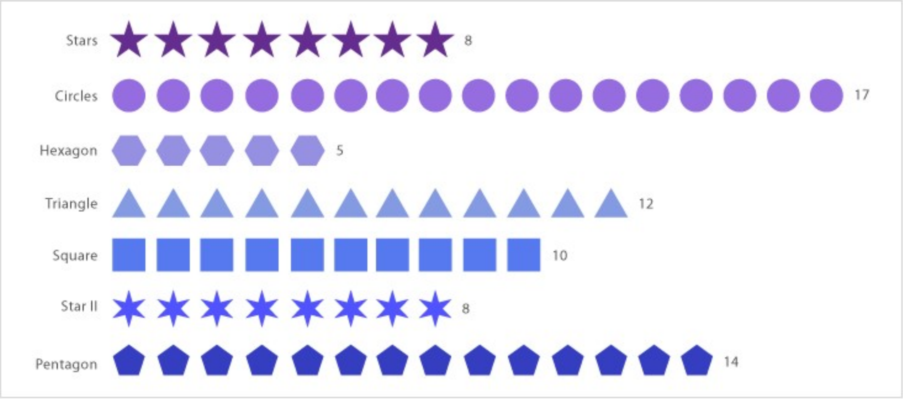
The use of icons can sometimes help overcome differences in language, culture and education. Icons can also give a more representational view of the data.

So for example, if your data is of 5 cars, you show 5 icons of cars in the chart.

Two things to avoid when using Pictogram Charts are:

Using them for large data sets, this makes values on the chart hard to count.

Displaying partial icons, as this can add confusion to what they represent.



Graphical Representation of Data

Graphical Representation is a way of analyzing numerical data.

It exhibits the relation between data, ideas, information and concepts in a diagram.

It is easy to understand and it is one of the most important learning strategies.

In Mathematics, a graph is a chart with statistical data that are represented in the form of curves or lines drawn across the coordinate point plotted on its surface.

It helps to study the relationship between two variables where it helps to measure the change of amount of variable with respect to another variable within a given interval of time.

It helps to study the series distribution and frequency distribution for a given problem.

General Rules for Graphical Representation of Data

There are certain rules to effectively present the data and information in the graphical representation. They are:

- Suitable Title: Make sure that the appropriate title is given to the graph which indicates the subject of the presentation.
- Measurement Unit: Mention the measurement unit in the graph
- Proper Scale: To represent the data in an accurate manner, choose a proper scale.
- Index: Index the appropriate colours, shades, lines, design in the graphs for better understanding

- Data Sources: Include the source of information wherever it is necessary at the bottom of the graph.
- Keep it Simple: Construct a graph in an easy way that everyone can understand.
- Neat: Choose the correct size, lettering, colours etc in such a way that the graph should be a visual aid for the presentation of information.

Merits of Using Graphs

- The graph is easily understood by everyone without any prior knowledge.
- It saves time
- It allows to relate and compare the data for different time periods
- It is used in statistics to determine the mean, median and mode for different data, as well as in interpolation and extrapolation of data.
- The graph presents data in a manner which is easier to understand.
- It allows us to present statistical data in an attractive manner as compared to tables. Users can understand the main features, trends, and fluctuations of the data at a glance.
- A graph saves time.
- It allows the viewer to compare data relating to two different time-periods or regions.
- The viewer does not require prior knowledge of mathematics or statistics to understand a graph.
- We can use a graph to locate the mode, median, and mean values of the data.
- It is useful in forecasting, interpolation, and extrapolation of data.

Limitations of a Graph

- A graph lacks complete accuracy of facts.
- It depicts only a few selected characteristics of the data.
- We cannot use a graph in support of a statement.
- A graph is not a substitute for tables.
- Usually, laymen find it difficult to understand and interpret a graph.
- Typically, a graph shows the unreasonable tendency of the data and the actual values are not clear.

Types of Graphs

Graphs are of two types:

1. Time Series graphs
2. Frequency Distribution graphs

Time Series graphs are Line graphs.

Line graph

Linear graphs are used to display the continuous data and it is useful for predicting the future events over time.

We use Multiple line chart for representing two or more related time series data expressed in the same unit and multiple – axis chart in somewhat similar situations, if the variables are expressed in different units.

When the time series exhibit a wide range of fluctuations, we may think of logarithmic or ratio chart where "Log y" and not "y" is plotted against "t".

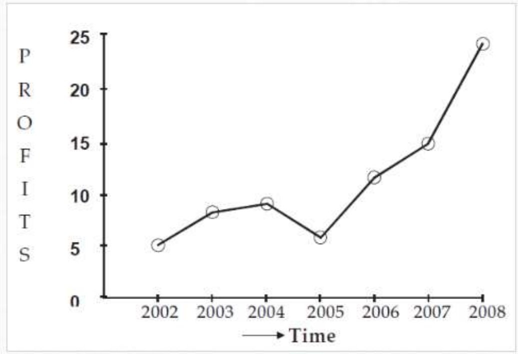
Line graph - Example

The profits in thousands of dollars of an industrial house for 2002, 2003, 2004, 2005, 2006, 2007 and 2008 are 5, 8, 9, 6, 12, 15 and 24 respectively. Represent these data using a suitable diagram.

Solution :

We can represent the profits for 7 consecutive years by drawing either a line diagram as given below.

Let us consider years on horizontal axis and profits on vertical axis.



Frequency distribution graphs are represented in four methods, namely

1. **Histograms**
2. **Frequency Polygon**

3. **Frequency Curve**
4. **Cumulative Frequency Curves**

Histogram

This graph uses bars to represent the frequency of numerical data that are organised into intervals. Since all the intervals are equal and continuous, all the bars have the same width.

A two dimensional graphical representation of a continuous frequency distribution is called a histogram.

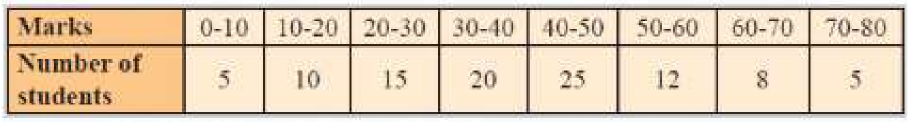
In histogram, the bars are placed continuously side by side with no gap between adjacent bars.

That is, in histogram rectangles are erected on the class intervals of the distribution. The areas of rectangle are proportional to the frequencies.

Histogram - Example

Example 1 :

Draw a histogram for the following table which represent the marks obtained by 100 students in an examination :



Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of students	5	10	15	20	25	12	8	5

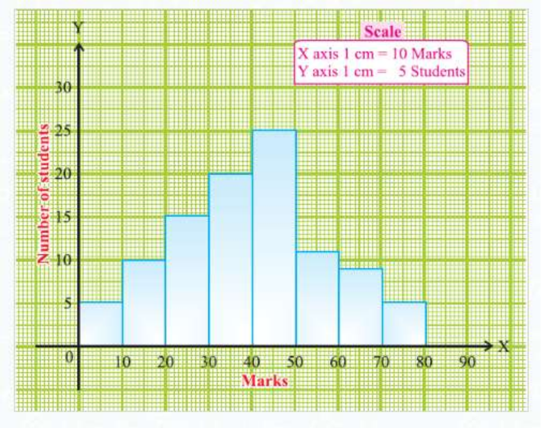
Solution :

The class intervals are all equal with length of 10 marks.

Let us denote these class intervals along the X-axis.

Denote the number of students along the Y-axis, with appropriate scale.

The histogram is given below.



Frequency polygon

Here are the steps to follow for finding the frequency distribution of a frequency polygon and it is represented in a graphical way.

- Obtain the frequency distribution and find the midpoints of each class interval.
- Represent the mid points along X-axis and frequencies along Y-axis.
- Plot the points corresponding to the frequency at each mid point.
- Join these points, using lines in order.
- To complete the polygon, join the point at each end immediately to the lower or higher class marks on the X-axis.

Draw the frequency polygon for the following data

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	4	6	8	10	12	14	7	5

Solution :

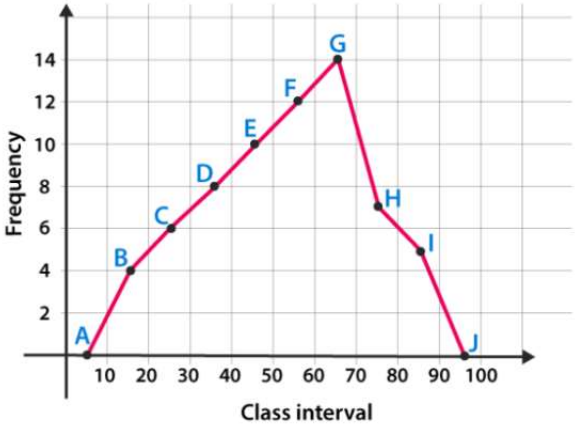
Mark the class interval along x – axis and frequency along y – axis.

Let assume that class interval 0-10 with frequency zero and 90-100 with frequency zero.

Now calculate the midpoint of the class interval.

Class Intervals	**Midpoints**	**Frequency**
0-10	5	0
10-20	15	4
20-30	25	6
30-40	35	8
40-50	45	10
50-60	55	12
60-70	65	14

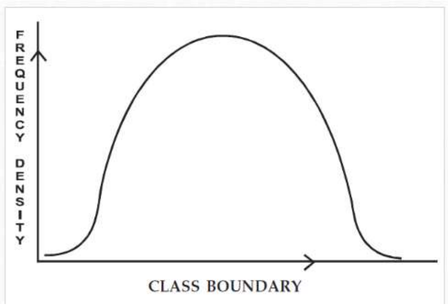
70-80	75	7
80-90	85	5
90-100	95	0



Frequency Curve

When you join the verticals of a polygon using a smooth curve, then the resulting figure is a Frequency Curve. It is a limiting form of a histogram or frequency polygon. The frequency-curve for a distribution can be obtained by drawing a smooth and free hand curve through the mid-points of the upper sides of the rectangles forming the histogram. A frequency-curve is a smooth curve for which the total area is taken to be unity.

As the number of observations increase, we need to accommodate more classes. Therefore, the width of each class reduces. In such a scenario, the variable tends to become continuous and the frequency polygon starts taking the shape of a frequency curve.



Cumulative Frequency Curve or Ogive

A cumulative frequency curve or Ogive is the graphical representation of a cumulative frequency distribution. Since a cumulative frequency is either of a ‘less than’ or a ‘more than’ type, Ogives are of two types too – ‘less than ogive’ and ‘more than ogive’.

What is Cumulative Frequency?

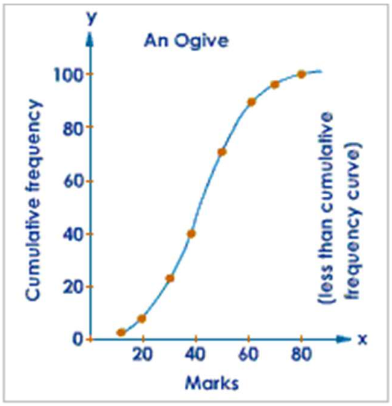
The frequency is the number of times an event occurs within a given scenario. Cumulative frequency is defined as the running total of frequencies. It is the sum of all the previous frequencies up to the current point. It is easily understandable through a Cumulative Frequency Table.

Marks	Frequency (No. of Students)	Cumulative Frequency
0 – 5	2	2
5 – 10	10	12
10 – 15	5	17
15 – 20	5	22

Cumulative Frequency is an important tool in Statistics to tabulate data in an organized manner. Whenever you wish to find out the popularity of a certain type of data, or the likelihood that a given event will fall within certain frequency distribution, a cumulative frequency table can be most useful.

Say, for example, the Census department has collected data and wants to find out all residents in the city aged below 45. In this given case, a cumulative frequency table will be helpful.

Cumulative Frequency Curve



A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results.

There are two types of Cumulative Frequency Curves (or Ogives) :

- More than type Cumulative Frequency Curve
- Less than type Cumulative Frequency Curve

More Than Type Cumulative Frequency Curve

Here we use the lower limit of the classes to plot the curve.

How to plot a More than type Ogive:

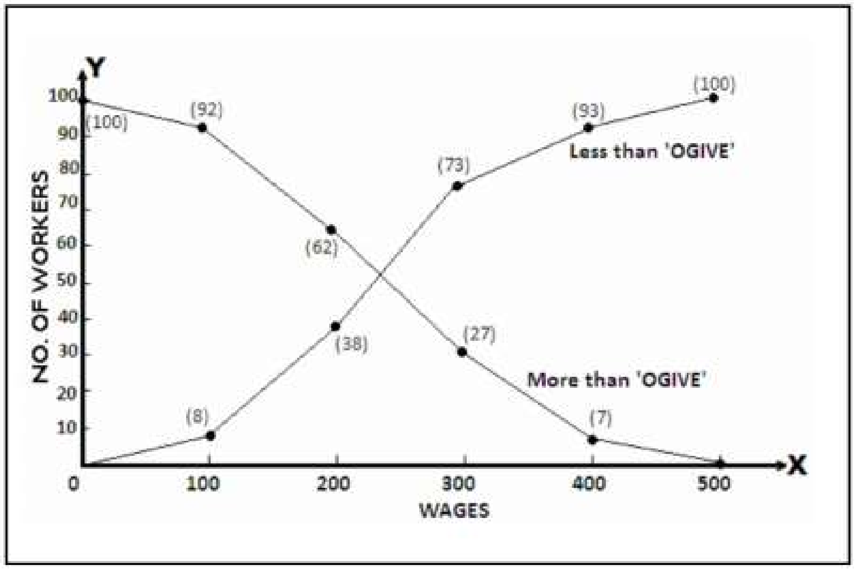
1. In the graph, put the lower limit on the x-axis
2. Mark the cumulative frequency on the y-axis.
3. Plot the points (x,y) using lower limits (x) and their corresponding Cumulative frequency (y)
4. Join the points by a smooth freehand curve. It looks like an upside down S.

Less Than Type Cumulative Frequency Curve

Here we use the upper limit of the classes to plot the curve.

How to plot a Less than type Ogive:

1. In the graph, put the upper limit on the x-axis
2. Mark the cumulative frequency on the y-axis.
3. Plot the points (x,y) using upper limits (x) and their corresponding Cumulative frequency (y)
4. Join the points by a smooth freehand curve. It looks like an elongated S.


Y
NO. OF WORKERS
100
90
80
70
60
50
40
30
20
10
0
100
200
300
400
500
X
WAGES
(100)
(92)
(62)
(27)
(7)
(8)
(38)
(73)
(93)
(100)
Less than 'OGIVE'
More than 'OGIVE'