Block 2

Collection of Data

BLOCK 2 INTRODUCTION

Block 2 "Collection of Data" deals with data collection and management.

Unit 4 "Data Collection" deals with the data collection process and various data collection methods.

Unit 5 "Data Management" deals with classification and tabulation of data and graphical and tabular representation of data.

Unit 6 "Geospatial Tools" deals withremote sensing and GIS.

Objectives

After studying this block, you will be able to:

- discuss the process of data collection;
- explain methods and tools of data collection;
- describe the tabulation method of representation of raw data with its advantages;
- explain the diagrammatic method of representation of data; and
- explain the concept and various features associated with frequency polygon along with its construction.

We hope that after studying this block, you will acquire an understanding of data collection and data representation.

Wishing you success in this endeavour!

UNIT 4 DATA COLLECTION

Structure

4.1 INTRODUCTION

Data Collection is a significant step in the entire research process. Once the researcher finalizes their research design, the process of data collection starts. While collecting the data, the researcher can continue his review of the literature. Two types of data are available for research, i.e., primary and secondary data. The primary data are those collected from the field by the researcher for their research. It is freshly generated, and they present the primary data in their research in an analytical form. It is original. If the researcher publishes their research after completing the entire research, others can use those primary data as secondary data. Unlike primary data, secondary data are those that have already been collected by an individual or an agency. The secondary data is sometimes available in a tabular format to interpret quickly. The researcher can also avail raw secondary data from the agency.

There are various sources for secondary data. The data collected by the government and published in a tabular format are considered accurate secondary data. As far as India is concerned, the following secondary data sources are available in the public domain and are considered authentic. The available secondary data sources are the Census of India, National Family Health Survey (NFHS), District level Household Survey (DLHS) and National Sample Survey Organization (NSSO). A private agency like the Center for Monitoring Indian Economy (CMIE) generates data sets regularly regarding the Indian economy and is available for researchers on a payment basis. There is a question about the authenticity of secondary data collected by private agencies. The researchers quote these secondary data to substantiate their arguments in their research. Many researchers develop statistical models and mathematical calculations from secondary data sources to address their objectives. The researcher could able to decide on the dataset based on their objectives. The methods of data collection differ for primary and secondary data. The researcher chooses the primary data collection method based on their objectives. The agencies collect secondary data based on finalized questionnaires and compile the same. In this unit, we learn the data collection process and various data collection methods. The unit also explains the positive and negative aspects of the various data collection process.

4.2 OBJECTIVES

After studying this unit, you should be able to:

- discuss the process of data collection;
- explain methods of data collection; and
- examine both positive and negative aspects of various data collection methods.

4.3 DATA COLLECTION

Data collection is a process of measuring and gathering information. It is part of the research component in all study fields, such as natural and physical sciences, social sciences, business administration and management, and humanities. The methods of data collection vary depending upon the discipline. For example, the researchers from the environmental study may use secondary data published by the international and transnational agencies on climate change and the environment like Intergovernmental Panel on Climate Change (IPCC). At the same time, the environmental discipline can also undertake experiments in the field to measure air and water qualities. Secondly, if researchers are interested in gender and environment and environmental movements, they have the choice to generate primary data from the field by undertaking in-depth interviews among selected samples. The goal for all data collection is to generate authentic evidence that allows the researcher to analyze and set convincing and credible answers to the research questions and objectives formulated by the researcher that the research committee has finalized. The data collection task begins after identifying and defining the research problem. Once the researcher defines the research problem, they develop a research design. There are two types of data, as mentioned in the introduction of the unit, i.e., primary and secondary data.

Primary data collection is gathering data through surveys, interviews, or experiments. The primary data is collected while carrying out experiments in experimental research. In the case of surveys (sample surveys or census surveys), the primary data will be collected through observation or direct communication with respondents through an interview schedule or questionnaire. The difference between experimental and survey research is that "An experiment refers to an investigation in which a factor or variable under test is isolated and its effect(s) measured. In an experiment, the investigator measures the effects of an experiment they conduct intentionally.

A survey refers to securing information concerning a phenomenon under study from all or a selected number of respondents of the concerned Universe. In a survey, the investigator examines those phenomena which exist in the Universe independent of their action" (Kothari and Garg 2019).

As far as India is concerned, the available secondary data sources are the Census of India, Family Health Survey, District Level House Hold Survey, National Crime Record Bureau, Periodic Labour force Survey (PLFS) of National Sample Survey Organization (NSSO), and National Sample Survey Organization (NSSO).

The private institutes that generate secondary data are India Human Development Survey (IHDS) and the Centre for Monitoring Indian Economy (CMIE). The discipline of Environmental Studies can obtain secondary data from the IPCC portal or Ministry of Environment, Forest and Climate Change.

4.4 PRIMARY DATA COLLECTION METHODS

There are various methods available for collecting primary data. As mentioned earlier in this Unit, it needs to be collected by the researcher, and it is original and first-time generated data. In this section, we will discuss various primary data collection methods and their advantages and disadvantages. The practical methods of primary data collection are given below:

1. Observation method
2. Interview method
3. Questionnaires
4. Schedules
5. Other methods include
    a. Warranty Cards
    b. Distributor Audits
    c. Pantry Audits
    d. Consumer Panels
    e. Using Mechanical Devices
    f. Using projective techniques
    g. Depth Interviews
    h. Content Analysis.

1. Observation Method

Observation is a process of gathering data by systematically observing the behaviour of individuals, groups and events or taking notes on various characteristics in their natural settings. Observations can be overt (everyone knows they are being observed) or covert (no one knows they are being observed, and the observer is concealed). The observation method is the most commonly used, especially in studies relating to behavioural sciences.

Observation becomes a scientific tool and data collection method for the researcher when it serves a formulated research purpose. It is a systematically planned activity, and it is recorded systematically. It is subjected to checks and controls on validity and reliability. Under the observation method, the information is sought through the investigator's direct observation without asking from the respondent.

Advantages

- The subjective bias will be eliminated if the observation is done accurately.
- The observation method focuses only on the present behaviour of the individual or group. This method does not bother about past behaviour or future intentions or attitudes. It does not complicate the present behaviour with past behaviour or intentions.
- This method is independent of respondents' willingness to respond, and there is less demand from the respondents for their active cooperation in this method like an interview or the questionnaire method.

Disadvantages

- It is time-consuming.
- It is an expensive method.
- The information provided by this method is minimal.
- Sometimes unforeseen factors may interfere with the observational task. Sometimes, the fact that some people are rarely accessible to direct observation creates an obstacle for this method to collect data effectively.

Types of Observation

The definition of observation is that it is a method of observing chosen units. There are various types of observation. In the structured observation, the researcher will determine the style of recording the observed information, standardize observation conditions and the selection of pertinent data. Thus, the entire process of observation is predetermined in the structured observation. The same is called unstructured observation when the researcher does not determine characters to be observed in advance. If the researcher themselves part of the group to be observed and the researcher observes the group members as well as researcher experiences everything as a group member, the observation is called participant observation.

But when the observer observes as a detached person without any attempt from the researcher to experience what group members experiences, this type is often termed non-participant observation. The researcher peacefully sits and takes notes of the activities of selected samples without interfering with their activities.

When the observer observes in such a manner that their presence may be unknown to the people they are observing, such an observation is described as disguised observation.

2. Interview Method

The interview method of data collection involves two or more people exchanging information through a series of questions and answers. The interview method may use a structured or unstructured schedule. In other words, the interview method of collecting data involves the presentation of oral-verbal stimuli and replies in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, telephonic interviews.

Advantages

- The interview method involves the intense exchange of information between interviewee and interviewer. In this process, the researcher can obtain more information in greater depth. With the in-depth knowledge of the subject and the excellent technical skills of the researcher, the interviewer can overcome the respondents' resistance. In the initial stage, the respondent may not provide the required information. Once the respondent gets confidence in the researcher, the respondent may be ready to share information. The interview method can yield an almost perfect sample of the general population.

- There is greater flexibility under this method as the opportunity to restructure questions is always there, especially in the case of unstructured interviews. In the process of the interview, the researcher can add more questions, and the researcher can obtain more data.

- Personal information can well be obtained easily under this method. The interviewer can usually control respondents, bringing the respondent back to the subject if they deviate from the subject. This is not possible in the mailed questionnaire approach. If so desired, group discussions may also be held.

Disadvantages

- It is time-consuming and expensive, especially when a large and widely spread geographical sample is taken. Due to time constraints, the researcher may limit their sample size to take in-depth interviews.

There are possibilities of the bias of the interviewer and the respondent; The respondent may bring their personal opinions in the response. If the sample size is enormous and the researcher involves more research assistants to interview, it may bring problems in supervision.

- Certain types of respondents, such as high government officials or executives or people in high-income groups, may not be easily approachable under this method, and the data may prove inadequate to that extent. This method is relatively more time-consuming, especially when the large sample and the respondents' recall are necessary.

- Under the interview method, the organization required for selecting, training and supervising the field staff is more complex with challenging problems. Interviewing at times may also introduce systematic errors. Effective interview presupposes proper rapport with respondents that would facilitate free and frank responses. This is often a challenging requirement. Once the researcher chooses an interview as a method and decides on sample size, they have to identify the sample and they have to create rapport with the sample. It takes time to identify and create rapport. Thus, it is time-consuming, and the researchers limit their samples.

Types of Interviews

The personal interview method requires a person known as the interviewer to ask questions face-to-face with the other person. Sometimes the interviewee may also ask specific questions, and the interviewer responds to these, but usually, the interviewer initiates the interview and collects the information.

In the direct personal investigation, the interviewer has to collect the information personally from the sources concerned. They have to be present at the place of the interview at a fixed time and meet people from whom data have to be collected. This method is particularly suitable for intensive investigations.

Indirect oral examination can be conducted under which the interviewer has to cross-examine other persons who are supposed to know about the problem under investigation, and the information obtained is recorded. The government appointed most of the commissions and committees to carry on investigations using this method.

Collecting information through personal interviews is usually carried out in a structured way. As such, we call the interviews structured interviews. Such interviews involve using a set of predetermined questions and highly standardized recording techniques.

The method of collecting information through a flexible approach to questioning is called unstructured interviews. Unstructured interviews do not follow a system of predetermined questions and standardized techniques for recording information. In this method, the interviewer is allowed much greater freedom to ask supplementary questions in case of need. At times, he may omit specific questions if the situation so requires.

The interview is meant to focus on the respondent's given experience and its effects. Under it, the interviewer has the freedom to decide the manner and sequence in which the questions would be asked and also has the freedom to explore reasons and motives. Such interviews are generally used in developing hypotheses and constitute a significant type of unstructured interview.

The clinical interview concerns underlying feelings or motivations or the course of an individual's life experience. The method of eliciting information is generally left to the interviewer's discretion. In the case of a non-directive interview, the interviewer's function is to encourage the respondent to talk about the given topic with a bare minimum of direct questioning. The interviewer often acts as a catalyst or facilitator to comprehensively express the respondents' feelings and beliefs and the frame of reference within which such feelings and beliefs take on personal significance.

To successfully implement the interview method, interviewers should be carefully selected, trained and briefed. They should be honest, sincere, hardworking, and impartial and must possess the technical competence and necessary practical experience.

Interviewing is an art governed by certain scientific principles. Every effort should be made to create a friendly atmosphere of trust and confidence so that respondents may feel at ease while talking to and discussing with the interviewer. The interviewer must ask questions correctly and intelligently and record the responses wholly and accurately. The interviewer should not show surprise or disapproval of a respondent's answer. Still, they must keep the direction of the interview in their hand, discouraging irrelevant conversation and making all possible efforts to keep the respondent on track.

The telephone interview method of collecting information involves contacting respondents on the telephone itself. It is not widely used but plays an integral part in industrial surveys, particularly in developed countries like the United States. The telephone interview method is widely used in advertising industries to know consumers' product preferences.

3. Questionnaires Method

The questionnaire method of data collection is popular while making extensive enquiries. This method is used mainly by research workers, private and public organizations, and government sectors. A questionnaire consists of many questions printed 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 their replies in the space for the questionnaire. The respondents have to answer the questions on their own. Most extensively, the method of data collecting by mailing the questionnaires is done in various economic and business surveys.

Advantages

- Even if it is done in a large and widely way, the cost is meagre.
- The respondents answer in their own words, free from the interviewer's bias.
- The respondents can be reached conveniently, though they are not easily approachable.
- The respondents have enough time to well thought out answers.
- Respondents, who are not easily approachable, can also be reached conveniently.
- Large samples can be made use of, and thus the results can be more dependable.

Disadvantages

- The low percentage of return of the duly filled questionnaires; bias due to no-response is often indeterminate.
- It can be used only when respondents are educated and cooperating.
- The control over the questionnaire may be lost once it is sent.
- There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been dispatched.
- There is also the possibility of ambiguous or omitted replies to particular questions; interpretation of omissions is difficult.
- It isn't easy to know whether willing respondents are truly representative.
- This method is likely to be the slowest of all.

The questionnaire is considered the heart of a survey operation. Hence it should be constructed carefully. If it is not adequately set up, the survey will fail.

The general form of a questionnaire is concerned with whether it can either be a structured or unstructured questionnaire. Structured questionnaires are those questionnaires in which there are definite, concrete and predetermined questions. The questions are presented with the exact wording and in the same order to all respondents. This sort of standardization ensures that all respondents reply to the same questions.

Structured questionnaires may also have fixed alternative questions in which the informants' responses are limited to the stated alternatives. Thus, a highly structured questionnaire is one in which all questions and answers are specified, and comments in the respondent's own words are held to the minimum. When these characteristics are not present in a questionnaire, it can be termed an unstructured or non-structured questionnaire.

Question Sequence: The questionnaire is effective, and to ensure the quality of the replies received, a researcher should pay attention to the question sequence in preparing the questionnaire. The question sequence must be precise and smoothly moving, meaning 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. The following type of questions should generally be avoided as opening questions in a questionnaire:

1. Questions that put too great a strain on the memory or intellect of the respondent;

2. Questions of a personal character;

3. Questions related to personal wealth, etc.

Question formulation and wording are essential in the entire process. The researcher should note that each question must be stated clearly. If the questions are complicated, it may lead to misunderstanding, and it will do irreparable harm to a survey.

- Questions should also be impartial to avoid a biased picture of the actual state of affairs.

- Questions should be constructed to form a logical part of a well-thought-out tabulation plan.

- Questions should be simple and easily understood.

Essentials of a Good Questionnaire: The questionnaire should be comparatively short and simple. The size of the questionnaire should be kept to a minimum. Questions should proceed logically, moving from easy to more complex questions. Personal and intimate questions should be left to the end. Technical terms and vague expressions capable of bringing different interpretations should be avoided in a questionnaire. Questions may be dichotomous, multiple choice or open-ended.

4. Schedules Method

This method of data collection is very much like the collection of data through a questionnaire, with little difference, which lies in the fact that schedules are being filled in by the enumerators who are specially appointed for the purpose. In certain situations, schedules may be handed over to respondents, and enumerators may help them record their answers to various questions in the said schedules. Enumerators explain the aims and objects of the investigation and remove the difficulties that any respondent may feel in understanding the implications of a question or the definition or concept of complex terms.

This method requires the selection of enumerators for filling up schedules. The enumerators should be trained to perform their job well, and the nature and scope of the investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule; they should be honest, sincere, hardworking and should have patience and perseverance.

This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results. It is, however, costly and is usually adopted in investigations conducted by government agencies or by some big organizations. Population census all over the world is conducted through this method.

Other Methods

Warranty cards are usually postal-sized cards which dealers of consumer durables use to collect information regarding their products.

Distributors and manufacturers perform regular distributor or store audits through their salespeople. Distributors get the retail stores audited through salespeople and use such information to estimate market size, market share, seasonal purchasing patterns etc.

The pantry audit technique is used to estimate the consumption of the basket of goods at the consumer level. In this type of audit, the investigator collects an inventory of types, quantities and prices of commodities consumed. An extension of the pantry audit approach regularly is known as the 'consumer panel', where a set of consumers are arranged to come to an understanding to maintain detailed daily records of their consumption, and the same is made available to the investigator on demands.

Use of mechanical devices: The use of mechanical devices has been widely made to collect information through indirect means. Eye camera, Pupillometric camera, Psycho galvanometer, motion picture camera and audiometer are the principal devices so far developed and commonly used by modern big business houses, mainly in the developed world, to collect the required information. In the era of the 21st century, researchers widely use their mobile phones to record their data.

Psychologists have developed projective techniques for collecting data to use respondents' projections for inferring underlying motives, urges, or intentions in which the respondent either resists revealing them or is unable to figure them out himself.

Depth interviews are designed to discover underlying motives and desires and are often used in motivational research. Such interviews are held to explore respondents' needs, desires and feelings.

The content analysis consists of analyzing the contents of documentary materials such as books, magazines, and newspapers and the contents of all other verbal materials which can be either spoken or printed.

EXPERIMENTATION OR EXPERIMENTAL RESEARCH

Experimental research, also called experimentation, is conducted using tools and techniques by developing two or more variables. The first variable is a constant you can manipulate to see the differences caused in the second variable. Most studies under quantitative research methods are experimental. The researcher may develop a tool specifically to test their study or use existing instruments to carry out the study.

Variables and Experimental Control

Variables are a measurable characteristic that varies. It may change from group to group, person to person, or even within one person. There are six common variable types:

Dependent variables show the effect of manipulating or introducing the independent variables. For example, suppose the independent variable is the use or non-use of a new language teaching procedure. In that case, the dependent variable might be students' scores on a test, and the content taught using that procedure.

Independent variables are those that the researcher has control over. This "control" may involve manipulating existing variables or introducing new variables in the research setting. Whatever the case, the researcher expects the independent variable to affect the dependent variables.

Intervening variables refer to abstract processes that are not directly observable but that link the independent and dependent variables. In language learning and teaching, they are usually inside the subjects' heads, including various language learning processes which the researcher cannot observe. For example, if using a particular teaching technique is the independent variable and mastery of the subject is the dependent variable. The language learning processes used by the subjects are the intervening variables.

Moderator variables affect the relationship between the independent and dependent variables by modifying the effect of the intervening variable. Unlike extraneous variables, moderator variables are measured and taken into consideration. Typical moderator variables include the subjects' sex, age, culture, or language proficiency.

Control variables: It is not possible to consider every variable in a single study. Therefore, the variables not measured in a particular study must be held constant, neutralized/balanced, or eliminated, so they will not have a biasing effect on the other variables. Variables that have been controlled in this way are called control variables.

Extraneous variables are those factors in the research environment that may affect the dependent variable but are not controlled. Extraneous variables are dangerous. They may damage a study's validity, making it impossible to know whether the effects were caused by the independent and moderator variables or some extraneous factor. If they cannot be controlled, extraneous variables must be considered when interpreting results.

4.4.1 Focus Group Discussion (FGD)

A focus group discussion involves gathering people from similar backgrounds or experiences to discuss a specific topic of interest. It is a form of qualitative research where questions are asked about their perceptions, attitudes, beliefs, opinions or ideas.

In the focus group discussion, participants are free to talk with other group members; unlike other research methods, it encourages discussions with other participants. It generally involves group interviewing in which a small group of usually 8 to 12 people. It is led by a moderator (interviewer) in a loosely structured discussion of various topics of interest. Like the observation method, the researcher does not insist on their opinion, and the researcher acts as a facilitator to moderate the discussion in the group. FGD is predominantly used in Gender and Natural Resource Management, Common Property Resources (CPR), Community participation in sustainable development and the role of women in managing forest resources.

Advantages

- Free and open discussion among the respondents generates new ideas that can be very useful for decision-making.

- A focus group is not static. The moderator can bring any changes to better facilitate the discussion during the group discussion. This dynamism allows better results in information derived by a focus group.

- Expressions other than those in a verbal form, such as gestures and stimulating activities, can provide the researcher with valuable insights.

Disadvantages

- Though a moderator can control the discussion, the extent to which they can control it depends on their experience. The inexperienced moderator may face problems controlling some participants who try to dominate the group.

- Respondents may be reluctant to share some sensitive ideas and concerns publicly.

- Due to the small sample size and heterogeneity of individuals, the findings may not be adequate to make projections or a composite picture of the situation.

- An FGD can be a very artificial set-up that influences the respondents to express and act unnaturally. The findings may be far from the actual.

4.4.2 Participatory Rural Appraisal

Participatory Rural Appraisal (PRA) is an effective primary data collection method. We are aware of the efforts of various development interventions in India to bring notable changes to improve individual quality of life, improve the ecosystem and bring sustainable development. Development interventions have been designed and implemented through five-year plans till the 12th Five Year Plans. After the 12th Five-year Plan, the government changed its approach and formulated and implemented sectoral plans and programmes. The five-year plans have been formulated and implemented in a top-down strategy that was target oriented. After reviewing implementation, governments and other agencies understood the limitations of top-down strategies. The top-down strategies are part of the following approaches: the Gandhian approach to development, community development approach, sector approach, target approach, area approach, minimum needs approach and integrated approach. These approaches have limitations and provide material and resources, and people may not be actively involved in the development process. Alternative to these approaches and ensuring people's involvement in the development process and collecting required data with people involved to assess their needs, Participatory Rural Appraisal (now it is called approach) emerged. Participation is emphasized in the PRA technique. According to McCraken et al. (1998), “PRA as a semi-structured activity carried out in the field, by a multidisciplinary team and designed to acquire new information on quickly, and new hypothesis about rural life”. According to the above definition, we could conclude that PRA is a field-based appraisal carried out by a team of researchers. It is a quick way of collecting data. It is a method of collecting information. It is exploratory. The existing references clearly stated that the PRA data collection method contributes to the community's emancipation and empowerment. PRA is a technique of participatory data collection. It has borrowed various methods and techniques from several sources and improvised to suit local needs. The primary sources of PRA are (Narayanasamy 2009)

1. Rural Development tourism
2. Survey
3. Participatory Action Reflection Research
4. Agro-ecosystem analysis
5. Applied Anthropology
6. Informal Survey
7. Rapid Rural Appraisal

Robert Chambers (1997), the father of the PRA technique, mentioned the first two techniques as harmful sources and the last five as positive ones.

The main features of PRA are a) flexible and adaptable to the local situation; b) It facilitates both participants and researcher to teach each other; c) Being an outsider, the researcher has space to learn from local people while collecting the required data; d) It also provides information on specific issues from another angle beyond collecting required data; e) It challenges various biases in urban researchers on rural issues and people; d) It makes the researchers visit interior villages rather than collecting data from the roadside villagers; e)It makes the researcher stay in the village for a considerable time to involve people in collecting data; f) Researcher needs a team of people to collect data through PRA techniques; g) It uses techniques and tools easily understandable by the local people; h) It helps the researcher avoid data collection, and i) PRA technique accepts unavoidable percentage inaccuracy in the data.

There are pretty good advantages to using PRA techniques. Still, the following are limitations of PRA.

a) The research team may collect lots of data, but they may not share it with the participants. As an ethical practice in research, people may not be aware of the collected data.

b) It is challenging to make participants focus on the data collection during the entire period.

c) Even though it talks about the empowerment of local communities, it may not empower local communities.

Box 4.1 GENDER ANALYSIS

Gender analysis is a process in which the inequalities and discriminations are identified in any given set-up by using various participatory tools and research methods. Since gender is a social construct, the method used for gender analysis is primarily subjective. It is more about analyzing the relationship between men and women regarding access to resources and obtaining a better standard of life, education, empowerment, etc. Gender analysis is predominantly used for development policy-making, project planning, implementation and evaluation. Hence, it has its research method for such purposes. Further, the gender construct has a historical background and some universal nature despite different contexts and realities. Gender analysis facilitates researchers to know who does what and has access to and control over resources. Gender analysis is used widely in Natural Resources Management, and it also helps researchers to know how far common property resources (CPR) accessible to women and the marginalized.

4.5 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 analyzed by someone else. Secondary data may either be published data or unpublished data.

Usually, published data are available in:

- Various publications of the central, state and local governments.
- Various publications of foreign governments or international bodies and their subsidiary organizations, technical and trade journals.
- Books, magazines and newspapers.
- Reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.
- Reports prepared by research scholars, universities, economists, etc. in different fields
- Public records and statistics, historical documents, and other sources of published information.

Reliability of data can be tested by finding out such things about the said data:

- Who collected the data?
- What were the sources of data?
- Were they collected by using proper methods
- At what time were they collected?

- Was there any bias in the compiler?
- What level of accuracy was desired? Was it achieved?

Suitability and Adequacy of Data: The data suitable for one enquiry may not necessarily be found suitable in another. If the level of accuracy achieved in data is found inadequate, then the researcher should not use it.

CHECK YOUR PROGRESS 1

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

1. Discuss the process of data collection.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

2. What are the sources of secondary data?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

3. What are the differences between interviews and questionnaires?

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

4. Explain various methods of collecting primary data.

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

……………………………………………………………………………………

5. Write short note on PRA.

……………………………………………………………………………………….

……………………………………………………………………………………….

……………………………………………………………………………………….

……………………………………………………………………………………….

4.6 LET US SUM UP

Data collection is a process of measuring and gathering information. It is part of the research component in all study fields, such as natural and physical sciences, social sciences, business administration and management, and humanities. In this unit, we have discussed the process of data collection and different methods of data collection. Further, we have discussed the advantages and disadvantages of various data collection methods.

4.7 KEY WORDS

Data : A collection of observations or scores from a survey or an experiment.

Quantitative Data : A set of observations where any single observation is a number that represents an amount or a count.

4.8 SUGGESTED FURTHER READING/ REFERENCES

Bryman, A. (2008). *Social Research Methods*. 3rd Edition, Oxford University Press., New York.

Kish, L. (1965). *Survey Sampling*, New York: Wiley.

Kothari, C.R. and Garg. G. (2019). Research Methodology: Methods and Techniques., New Delhi: New age Publishers.

Moser, C. A. and Kalton, G. (1971). *Survey Methods in Social Investigation.* The English Language Book Society: London

Narayanasamy, N. (2009*). Participatory Rural Appraisal Principles. Methods and Application.* SAGE Publications India Pvt Ltd., New Delhi
Singleton Jr., R.A. and Straits, B.C. (2017). *Approaches to Social Research*, 6th Edition. Oxford University Press, Oxford.

4.9 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress 1

1. Please refer to section 4.3

2. Please refer to section 4.5

3. Please refer to section 4.4

4. Please refer to section 4.4

5. Please refer to section 4.4.2

UNIT 5 DATA MANAGEMENT

Structure

5.1 INTRODUCTION

It is the fact that first and foremost thing in any experiment/investigation/ project is to design it in such a way that it may produce relevant and reliable data for drawing fruitful conclusions through the statistical analysis of the concerned data and for the purpose of applying different analytical tools on the data for drawing valid conclusions, it is necessary to summarize and classify the data in such a manner that some basic properties of it become apparently visible and, therefore, data may be considered to be in easy-to-read form.

This unit describes tabulation of raw data depending upon a number of factors which might affect the data in different way. Tabulation of raw data actually reveals many of the hidden things, which are otherwise not observable. Depending upon the purpose of the experiment/investigation, data may be tabulated in a number of ways.

On the other hand, if anyhow data can be represented in the form of some attracting colourful diagrams/charts/graphs/figures using only some salient features of the study, then these generally seem to attract the attention of common people and, therefore, they may be interested to spend some time to understand the facts and figures depicted through diagrams, graphs, figures whether or not they are seriously interested in it. In this unit our aim is to acquaint you with different types of diagrams/charts/figures for the purpose of representing any set of data.

After discussing the diagrammatic representation of common types of data, we now move towards the graphical presentation of a frequency distribution. Diagrams used to represent different characteristics of a frequency distribution are popularly known as 'graphical representation of data'. These graphs, plotted for frequency distributions, are used to interpolate/extrapolate items in a series including locating various partition values.

In section 5.3, method of construction of a frequency distribution and its types are discussed and described. In section 5.4, different kinds of tabulation methods, each with necessary parts of the table, are discussed and described. Types of tables are also discussed in this section. An introduction of diagrammatic representation has been provided in Section 5.5. The section also highlights significance of diagrammatic method of representation. A mention of different types of diagrams, has also been given in this section. In Section 5.6, we have discussed exclusively about bar diagrams which quite frequently used in Statistics. The various kinds of bar diagrams are also discussed. Another diagram which is considered to be a substitute of sub-divided bar diagram is "Pie Diagram" or "Pie Chart", which is discussed in Section 5.7 with its construction. Pictographs and Cartograms have been discussed respectively in Section 5.8 and Section 5.9. Section 5.10 will be devoted to explaining some commonly used graphs associated with frequency distributions, such as, Histogram, Frequency polygon, Frequency curve, Cumulative frequency curves (or Ogives) along with their significance and their construction.

5.2 Objectives

After studying this unit, you should be able to:

- describe the tabulation method of representation of raw data with its advantages;

- mention different parts of a suitable table prepared with specific purpose and difference between univariate, bi-variate, tri-variate and multi-variate tables;
- explain the diagrammatic method of representation of data;
- explain different types of bar diagrams, and the situations where these diagrams would be appropriate to use;
- explain the pie diagram and situations where it would be preferred as well as method of its construction;
- draw pictograms and cartograms, and their uses as a diagrammatic representation of a set of data;
- describe the meaning and advantages of graphical presentation;
- explain the concept of histogram and its construction for continuous frequency distribution;
- explain the concept and various features associated with frequency polygon along with its construction; and
- explain how to draw cumulative frequency curves (ogives) and their applications in order to explore some of the positional values of the data.

5.3 FREQUENCY DISTRIBUTION

When observations, whether they are discrete or continuous, available on a single characteristic of a large number of individuals, it becomes necessary to condense the data as far as possible without loosing any information of interest. Let us consider the ages of 30 students selected at random from among those studying in a certain class.

20, 22,25, 22,21, 22,25, 24,23, 22,21, 20,21, 22,23, 25,23, 24,22, 24, 21, 20, 23, 21, 22, 21, 20, 21, 22, 25.

This presentation of the data is not considered as good since for large number of observations it is not easy to handle the data in this form. A better way to express the figures is shown in table given below:

Table 5.1

Age of students	Tally Mark	Frequency
20	||||	04
21	~~||||~~ ||	07
22	~~||||~~ |||	08
23	||||	04
24	|||	03
25	||||	04
Total		30

A bar (|) called tally mark is put against the number when it occurs. After putting this mark four times against the value, a cross tally is put on these 4 tallies for the fifth mark as shown in the above table. From the sixth mark onwards, we start afresh in the similar manner. This technique facilitates easy counting of the tally marks at the end. The presentation of the data as given in Table 5.1 is known as frequency distribution.

A frequency distribution refers to the data which are classified on the basis of some variables that can be measured such as wages, age of children, etc. A variable refers to the characteristic that varies in magnitude in a frequency distribution. It may be either discrete or continuous. A discrete variable is that which generally takes integer values. For example, the number of students, the number of books, etc. A continuous variable can take integer or fractional values within the range of possibilities, such as the height or weight of individuals. Generally speaking, continuous data are obtained through measurements while discrete data are derived by counting. A series described by a continuous variable is called continuous series. Similarly, series represented by a discrete variable is called discrete series.

According to the nature of the variable, the frequency distribution may be of two types, i.e., discrete frequency distribution and continuous frequency distribution. Let us discuss them one by one.

5.3.1 Discrete Frequency Distribution

A frequency distribution in which the information is distributed in different classes on the basis of a discrete variable is known as discrete frequency distribution. For example, frequency distribution of number of children in 20 families is discrete frequency distribution as shown in table given below:

Table 5.2

No. of children	Tally Mark	Frequency
0	|||	3
1	||||	4
2	||||| \	6
3	||||	4
4	|||	3
Total		20

5.3.2 Continuous Frequency Distribution

A distribution in which the information is distributed in different classes on the basis of a continuous variable is known as continuous frequency distribution.

There may be some variables which have integer values as well as fractional values. Frequency distribution of such variables is called continuous frequency distribution. An example of a continuous frequency distribution is given below in table given below:

Table 5.3

Heights (cm)	Tally Mark	Frequency
120 -130	|||	3
130 -140	𝍸	5
140- 150	𝍸 𝍸	10
150 -160	𝍸 𝍸 ||||	14
160 -170	𝍸 𝍸 ||	12
170-180	||||	4
180-190	||	2
Total		50

After discussing the discrete and continuous frequency distributions let us discuss the Relative and Cumulative frequency distributions which are of the similar importance as analysis point of view of data is considered.

5.3.3 Relative Frequency Distribution

A relative frequency corresponding to a class is the ratio of the frequency of that class to the total frequency. The corresponding frequency distribution is called relative frequency distribution. If we multiply each relative frequency by 100, we get the percentage frequency corresponding to that class and the corresponding frequency distribution is called "Percentage frequency distribution". Let us take an example in which both relative and percentage frequency distributions are prepared.

Example 1: A frequency distribution of marks of 50 students in a subject is as given below:

Class (Marks):	0-10	10-20	20-30	30-40	40-50
Frequency:	6	10	14	18	2

Prepare relative and percentage frequency distributions.

Solution: The relative and percentage frequency distributions can be formed as given in the following table:

Class (Marks) X	**Frequency (f)**	**Relative frequency (f/N)**	**Percentage Frequency (f/N) × 100**
0-10	6	6/50 = 0.12	0.12 × 100 = 12 %
10-20	10	10/50 = 0.20	0.20 × 100 = 20 %

20-30 30-40 40-50	14 18 2	14/50 = 0.28 18/50 = 0.36 2/50 = 0.04	$0.28 \times 100 = 28\ \%$ $0.36 \times 100 = 36\ \%$ $0.04 \times 100 = 4\ \%$
Total	$\sum f = N = 50$	1.00	100

5.3.4 Cumulative Frequency Distribution

The cumulative frequency of a class is the total of all the frequencies up to and including that class. A cumulative frequency distribution is a frequency distribution which shows the observations 'less than' or 'more than' a specific value of the variable.

The number of observations less than the upper class limit of a given class is called the less than cumulative frequency and the corresponding cumulative frequency distribution is called less than cumulative frequency distribution. Similarly, the number of observations corresponding to the value of more than the lower class limit of a given class is called more than cumulative frequency and the corresponding cumulative frequency distribution is called 'more than' cumulative frequency distribution. Following is an example, wherein 'less than' and 'more than' cumulative frequency distributions have been obtained.

Example 2: For the following frequency distribution of marks of 50 students in a subject, form both types of cumulative frequency distributions.

Class (Marks)	0-10	10-20	20-30	30-40	40-50
No. of Students	7	11	15	12	5

Solution: Cumulative frequency distributions are formed as given in the following table:

Given Frequency Distribution		**Less Than Cumulative Frequency Distribution**		**More Than Cumulative Frequency Distribution**	
Classes	**No. of Students**	**Marks Less than**	**No. of students**	**Marks More than**	**No of students**
0-10	07	10	07	0	50
10-20	11	20	18	10	43
20-30	15	30	33	20	32
30-40	12	40	45	30	17
40-50	05	50	50	40	05
Total	50				

5.4 TABULATION OF DATA

Tabulation is a method of summarizing the raw data, in which the data (mostly numerical data) are presented in the form of a Table. A table is, in fact, a rectangular body divided into different rows and columns, consisting of a title to describe the main content of the table, row and column headings to provide sufficient explanatory statements about their contents and foot notes mentioning the origin, source and other necessary details related to data. This is one of the simplest and most revealing devices for summarizing and presenting data in a meaningful arrangement. Thus, tabulation is a systematic and logical method of representation of numeric data in rows and columns to facilitate comparison and statistical analysis of a number of facts. We can also define a statistical table as the logical listing of quantitative data in columns and rows of numbers with sufficient explanatory statements.

5.4.1 Objectives of Tabulation

Whatever we discussed above about the tabulation method of summarization, indicate some of the objective for preparing it. These can be listed as follows:

(a) **To save space:** A table presents several facts contained in the data in a better way without sacrificing the quality and quantity of the data. This is because, a table does not contain sentences for describing each and every fact of the data; thus, saving maximum space.

(b) **To simplify complex data:** It reduces the raw data which are haphazardly scattered in a simplified and meaningful form so as to be easily understandable and interpretable by a common man.

(c) **To make available essential features of data:** A table presents only essential and main characteristics of the data, ignoring all the textual explanations, which are needed for further statistical analysis. Thus, a table presents the facts about the data in a clear and precise way.

(d) **To facilitate mutual comparison and statistical analysis of data:** Since, a table consists of numerical data arranged adjacent to each other in rows and columns, it is very simple to make simultaneous comparisons of differences occurred in the data row-wise or column-wise on the basis of many parameters. Also, using the tabulated values, all types of statistical analysis are possible to apply.

5.4.2 Components of a Table

A table is prepared using a number of its components. Therefore, in order to construct a table, the meaning of these components and their proper places in the table must be known to everybody. However, various components of a table may vary from case to case depending upon the given data, but a good table must contain at least the following components:

1. Table Number
2. Table Heading

3. Caption
4. Stub
5. Body of Table
6. Head Note
7. Foot Note

Let us throw some light on these components one by one.

1. Table Number

A table should be given a unique table number such that while dealing with a number of tables, each of them can be easily distinguished with each other. The table number may be shown either in the centre at the top above the title or in the left-hand side of the table at the top. When there are many columns inside the table, it is desirable to number each column also so that easy reference to it is possible.

2. Table Heading or Title of the Table

A good table should have a suitable heading. The heading is a brief description of the contents of the table. It should be placed above the table, either along with the table number or just below the table number. In other words, the heading of the table should be clear, brief, crisp and carefully-worded so as to be self-explanatory and permits one and only one interpretation.

3. Caption

Caption refers to the column heading, and explains what information column present. Under a column heading, there may be two or more sub-headings. The caption should be clearly defined and placed at the middle of the column. If the different columns are expressed in different units, the unit should be specified along with the captions.

4. Stub

The stubs are row headings. They are placed at the extreme left of the table and perform the same function for the horizontal rows in the table as the captions do for the vertical columns.

5. Body

The body of the table is the central part of table that contains the numerical information presented in table arranged in rows and columns. This is the most vital part of the table.

6. Head Note

Head note is a brief explanatory statement applying to all or a major part of the material presented in the table and is placed below the title entered and enclosed in brackets. It is used to explain certain points relating to the whole table that have not been included in the title nor in the captions or stubs. For example, the unit of measurement is frequently written as the head note such as “in thousands” or “million tons” or “in crores”, etc.

7. Foot Note

Anything in a table which the reader supposed to find difficult to understand should be explained in footnotes. Footnotes may be placed directly below the body of the table. The footnotes are generally used for the following purposes:

(a) Any special circumstances affecting the data, for example, strike, fire, etc.

(b) To give the source in case of the secondary data. If any information in table obtained from some journal, its name, data publication, page number, table number, etc., should be mentioned so that if the user wishes to check the data from the original source, he could know where to look for the information.

5.4.3 Types of Tables

As mentioned earlier, a table is used mostly to classify one or more characteristics (either quantitative or qualitative); classified according to different levels of them, along with their corresponding values recorded or measured on some objects/units. Depending upon the number of characteristics used for classification, we have mainly two types of tables, which are:

1. Simple Tables
2. Complex Tables

1. Simple Tables

Tables can be differentiated with one another in terms of characteristic(s) which has (have) been classified with all of its (their) levels. If the data related to only one characteristic is presented and classified in the tabular form, the table is known as “Simple Table”. The simple table is also known as “One-Way Table”. In a simple table, obviously, there happens to be only two columns, the first consisting of the different levels of the characteristic under classification and the second, consisting corresponding values of the characteristic over different levels.

The following is an example of a simple table, in which the characteristic for classification is ‘Marks in Statistics’ obtained by 100 students, which are classified into a number of classes of marks:

Table 5.4: Marks Obtained by 100 Students in Statistics

Marks in Statistics	Number of Students
Below 10	5
10-20	8
20-30	12
30-40	10
40-50	15
50-60	18
60-70	17
70-80	13
Above 80	02
Total	100

Depending upon the available space, if needed, this table can be converted into horizontal form as follows:

Marks	Below 10	10 -20	20 -30	30 -40	40 -50	50 -60	60 -70	70 -80	Above 80	Total
No. of Students	5	8	12	10	15	18	17	13	2	100

2. Complex Tables

Contrary to simple table, in complex table, instead of considering only one characteristic, two or more characteristics are presented. These are the types of tables which represent the classification of data into two or more categories based on two or more characteristics. Complex tables are frequently used in practice because they facilitate to incorporate full information and a proper consideration of all related facts. Such tables may further be categorized as

(1) Two-Way Table

(2) Three -Way Table

(3) Many-Fold Table

Let us now discuss these categories of complex tables.

(1) Two-Way Table

Two-way tables consider two characteristics simultaneously for the purpose of classification of data and are formed when either the stub or the caption or both are divided into two categories. In the following example, you can observe that the characteristics used for classification are two in number, one, the "Age" and the second, "Sex". The data are related with age of persons and within age groups, sex of persons. Hence, it is an example of a Two-Way Classification and the table is a Two-Way Table or bi-variate table.

Table 5.5: Age and Sex-Wise Distribution of Persons Living in a Colony

Age (in years)	Sex-Wise Number of Persons		Total
	Male	Female	
Below 15	12	6	18
15-25	20	12	32
25-35	42	27	69
35-45	25	18	43
45-55	10	8	18
55-65	8	5	13
65 and Above	5	2	7
Total	122	78	200

(2) Three-Way Table

A "Three -Way Table" is an arrangement of data in a tabular form where the classification of the data is made on the basis of three characteristics considered simultaneously. The process of such classification is, therefore, called "Three-Way Classification.

Obviously, because of considering three variables (characteristics) at a time along with their different categories for tabulation of data, such a table becomes comparatively more complicated in its structure than one way and two-way tables. You can feel that the construction of such a table is not very simple and needs a lot of experience and vision. For example, let us assume that a firm, consisting of 1000 permanent employees, was asked by the Government Audit Office to submit a detailed report of the salary it paid to its employees in last month, who worked in different shifts. The firm sent the information required in the form of a table which is reproduced below:

Table 5.6: Classification of employees of a firm according to their sex, salary per month and shift of duties

Salary ('000 Rupees)	Male Worker			Female worker			Total
	I shift	II shift	III shift	I shift	II shift	III shift	
2 – 4.999	13	22	10	3	6	2	**56**
5 – 9.999	23	38	37	7	11	4	**120**

10 – 14.999	40	45	20	16	26	9	**156**
15 – 19.999	48	50	18	38	38	17	**209**
20 – 29.999	48	48	15	56	49	14	**230**
30 – 39.999	31	20	14	18	10	6	**99**
40 – 49.999	20	40	12	8	12	3	**95**
50 and above	10	5	6	9	4	1	**35**
Total	**233**	**268**	**132**	**155**	**156**	**56**	**1000**

(3) Many-fold Table

When the classification characteristics exceeds than three, the table is said to be a "Many-Fold Table" or "Multi-variate Table". You can visualize from above example of a three-way table that as the way of classification involves more and more characteristics, the construction of the table using information of all the characteristics simultaneously, becomes more and more complicated. Therefore, it is advised normally not to use more than four characteristics in the same table. When more than four characteristics are to be presented, we should form more than one table depicting relationship between different characteristics.

CHECK YOUR PROGRESS 1

Note: i) Check your answers with those given at the end of the unit.

1) In a sample survey study about the drinking habits in two cities, it is observed that, in city X 57% are male, 22% are drinkers, and 14% are male drinkers, whereas in city Y 52% are male, 28% are drinkers and 21% are male drinkers. Tabulate the above information.

2) Present the following information in a suitable tabular form: In 2009 out of a total 2000 employees in a company 1550 were members of a trade union. The number of women employees was 250, out of which 200 did not belonging to any trade union. While in 2010 the number of union employees was 1725 out of which 1600 were men. The number of none union employees was 380 among which 155 were women.

5.5 DIAGRAMMATIC REPRESENTATION OF DATA

As mentioned in section 5.1, diagrammatic representation of data means that facts inherent in any data set are represented through various diagrams, charts, graphs and figures. Before we discuss different types of diagrams, let us first see what are the significance of the diagrams?

5.5.1 Significance of Diagrammatic Representation of Data

In addition to Tabular Methods of representation of data, if we use the Diagrammatic Method of representation as the second method for summarizing and interpreting the useful facts of the data set, obviously we should highlight its significance indicating some of its salient features as:

(i) Easy to Understand

A diagram is used to carry some informative facts inherent in any set of data, collected in a study; it has more importance as a statistical tool. In fact, a large number of observations become easy to understand if they are presented through an appropriate and suitable diagram. As the number of observations increases, generally their analysis tends to be more tedious, but through suitable chosen diagrams the presented data can be understood easily. It is also said that pictures have explanation power of worth more than 10000 words.

(ii) Looks Attractive

Diagrams look attractive to the eyes and, hence, they are sometimes known as "eye-catching devices". Numerical numbers and facts presented without the help of diagrams are generally boring to common people whereas diagrams even with some sort of data looks impressive and give pleasure to the eyes. It is a well-known fact that whenever numerical facts are depicted through some colourful attractive pictures, diagrams, charts; they very easily attract the attention of common people even when they are not very much concerned with those facts which are expressed in terms of numerical numbers. That is why, readers give more attention to the diagrams rather than the numbers, if presented in text or tabular forms.

(iii) Greater Memorizing Effect

Diagrams are long lasting in memory than numbers. Numbers may not be remembered easily but diagrams have greater memorizing effect, as the impressions created by them remains in mind for long time.

(iv) Easy Comparison of Data

Through the diagrams, one can easily compare the data related to different areas and time. It is difficult to read and compare the numbers whereas diagrams can be compared easily by viewing the presented information.

5.5.2 Different Types of Diagrams

Depending upon the nature of the data, type of variations in their values, number of characteristics used for their classification and the objectives behind using the diagrams as tools of data representation; several types of diagrams, graphs, charts and figures have been suggested. In this sub-section, we shall describe several types of diagrams which are very commonly used for representing the data in diagrammatic form.

1. Bar Diagrams

Bar diagrams are the diagrams which are mostly used in Statistics for diagrammatic representation of any data set. In such diagrams, we draw either single bars or multiple bars according as the number of characteristic(s) (variable(s)) is (are) given in the data set, each of them is drawn either vertically or horizontally; to represent the values of the variable given in the data set. Here, a 'bar' means a 'rectangle' with some length and width.

There are generally different kinds of bar diagrams, which are suggested for different kinds of situations related to nature of data given and types of classification of values of the variable(s) under consideration. Accordingly, bar diagrams may be categorized as follows:

a. Simple Bar Diagrams;

b. Multiple Bar Diagrams; and

c. Sub-divided Bar Diagrams.

2. Pie Diagram or Pie Chart

This is another method of representing the data through diagrams, graphs, charts. Pie chart uses a circle of any radius, instead of rectangles for representing a specific form of data. Generally, pie chart is used to depict the data which are given in the form of proportions or percentages. It is never used for representing the actual values of the variable(s).

3. Pictographs (Pictorial Diagrams)

Sometimes, in order to represent the available data in diagrammatic form, we neither use any bar or any circle; rather we use some kinds of symbols, pictures for each and every value of the characteristic (variable) under study. Such a diagrammatic representation of data is done using what is popularly called "Pictographs" or "Pictorial Diagrams. In pictographs we generally use some figures or symbols, as mentioned above, to represent the data. The symbol or picture for each unit of the variable is so chosen that it should have some relevance with the unit measured for obtaining values of some characteristics of it.

4. Statistical Maps (Cartograms)

Representation of numerical facts of some characteristics using maps of regions, parts of a geographical place, some territories of a country or the country itself is another method of diagrammatic representation of data. The diagrams used in such methods are also called "Cartograms". Such a representation has its own importance and significance, though not very much in practice due to some reasons. We shall see afterwards in a section, while discussing cartograms, what are the merits and demerits and limitations of cartograms.

5.6 BAR DIAGRAMS

In the previous section, we presented a brief review of bar diagrams. Under the sub-section 5.5.2, a reference of different types of bar diagrams has been made. Accordingly, we shall describe different types of bar diagrams as follows:

5.6.1 Simple Bar Diagrams

Simple bar diagram means a bar diagram having a single bar for each and every value of the variable under consideration. This indicates that simple bar diagrams are used when there is only one characteristic (variable) whose values are to be represented with the help of a bar diagram. Thus, for example, if a data set show strength of students in a school over eight consecutive years, we then have a series of eight values, representing the strength of students, which are to be represented by a simple bar diagram consisting of eight bars. Similarly, other examples where simple bar diagram may be used are productions of some kind of goods produced by a firm over last six months, profits of a manufacturing company over last four years, indoor patients admitted in a hospital during last week and so on.

In simple bar diagrams and in general, in all types of bar diagrams, bars may be of different lengths and widths; but the length of a bar should represent the value of the variable for that particular item. Thus, if two values of the variable are 20 and 65 respectively for the fourth and seventh items, the length of the seventh bar must be greater than the fourth bar in the diagram. As far as the widths of bars are concerned, there is no hard and fast rule for it. However, whatever width is taken for one bar, it should be same for each and every bar in the diagram, that is, each and every bar should be of same width. Further, generally in order to make bars distinguishable from each other, equal gaps are left between all the bars of the diagram. After drawing different bars in the diagram, respective values of the variable should be written at the top of bars so that reader may be able to know the precise value without looking at the scale.

Simple bar diagrams may be either with vertical bars or with horizontal bars, due to the reasons described in the sub-section 5.5.2. If it is a vertical simple bar diagram, obviously, the items must be shown on the X-axis whereas values of the items must be shown on the Y-axis. In order to accommodate all the items on X-axis and all the item-values on Y-axis, it is necessary to choose suitable scales separately for both the axes.

On the other hand, to draw a horizontal simple bar diagram, the items have to be shown on the Y-axis and their values on the X-axis.

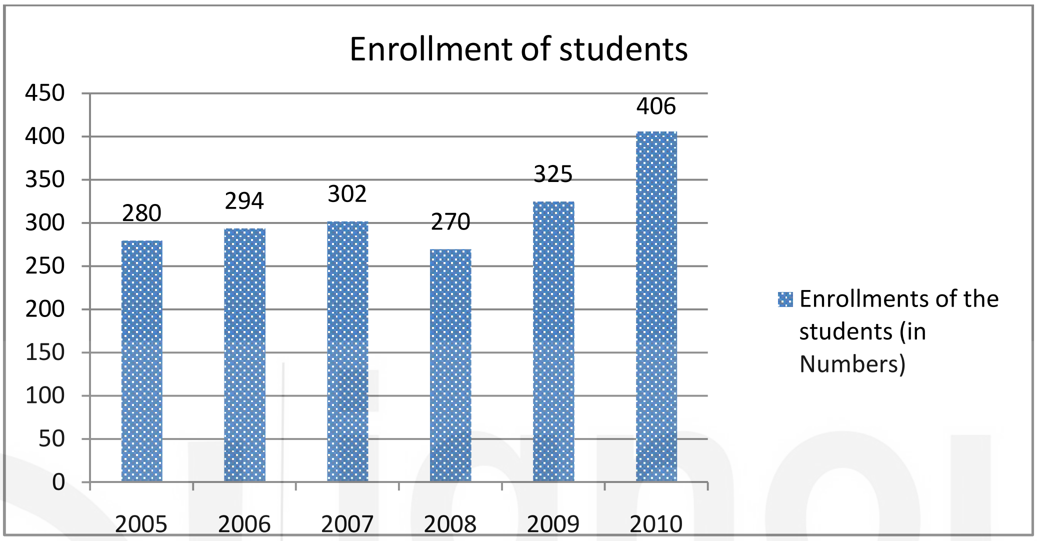
Now, we shall describe how to construct the simple bar diagrams. For this, we consider the following example for illustration:

Example 1: Represent the following data by a suitable diagram:

Years	2005	2006	2007	2008	2009	2010
Enrollment of Students	280	294	302	270	325	406

Solution: We observe that here there is only one characteristic (variable) under study which is 'enrollment of students', which vary over years, that is, it gives us values recorded in chronological order. Therefore, we must use a vertical simple bar diagram.

The number of values in the data set is six, so have to draw six bars in the diagram with their lengths equal to their corresponding values and with some width between each bar; the equal width of bars should be so chosen that each bar with that much of width along with the equal gap between each bar can be accommodated in the available space. Keeping these things in mind, the above data may be presented in the form of the following vertical simple bar diagram:



The above bar diagram consists of six vertical bars, each bar representing the value of enrollments in that particular year. Years are shown on the X-axis and number of enrollments (values of the variable) on the Y-axis. A suitable scale on both the axes has been taken such that all the bars are accommodated in the given space. The width of each bar is same and an equal gap has been left between each bar. The other components of the diagram, such as, its title, scales on both the axes, index of the diagram and footnotes can be added if the source of the data and other information are given along with the data. Clearly, scale on Y-axis can be shown only if the diagram is constructed on a graph paper.

5.6.2 Multiple Bar Diagrams

We saw that a simple bar diagram is mostly used when values of only one variable, as measured on different items, are given for their diagrammatical representation. But in practice, we sometimes have values of a number of variables, measured on the same items. These values may vary either over time or over space. As an example, say a survey was conducted to know the prices of four food items of same brand, say, Sugar, Edible oil, Ghee and Wheat flour on the first Monday of five consecutive months. Obviously then we would have five series of prices as values of four food items, measured over five consecutive months. Such a data set cannot be represented diagrammatically through a simple bar diagram at a time. Therefore, we need another type of diagram for such data sets.

Multiple bar diagrams, where we use more than one bar at a time for each and every item help us in representing these types of data sets. We know that multiple bars are constructed for either the different components of the total or for the magnitudes of several variables taken together at a time.

All the bars drawn for a particular point of time are drawn one after another without any gap between them. The next group of bars is similarly drawn for next time point or place; but there should be a gap between two consecutive such groups of bars. As a rule, similar to simple bar diagram, here also height of each bar must represent the value of the corresponding variable.

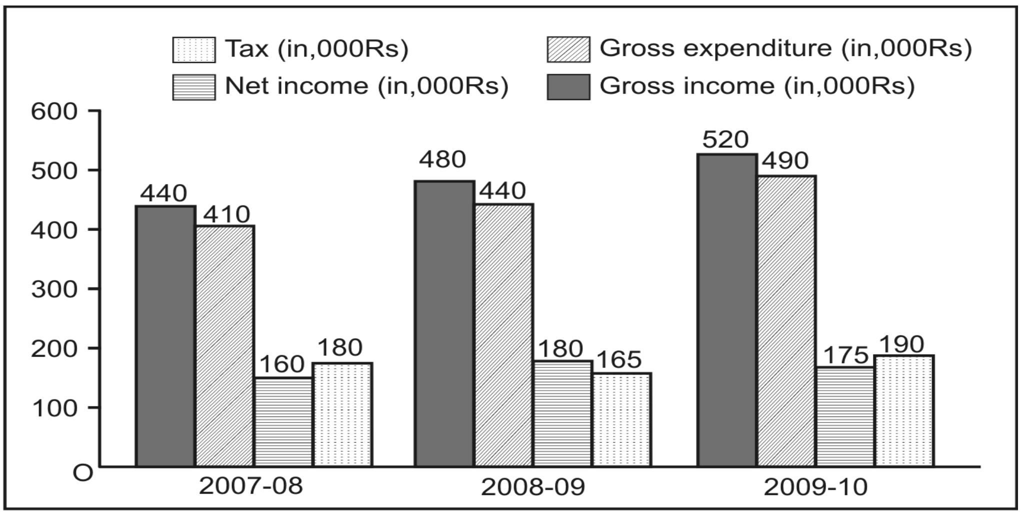
Let us now take the following example to illustrate how to construct the multiple bar diagrams:

Example 2: Represent the following data by a suitable bar diagram:

Year	2007-08	2008-09	2009-10
Gross Income (in 000 Rs.):	440	480	520
Gross Expenditure (in 000 Rs.):	410	440	490
Net Income (in 000 Rs.):	160	180	175
Tax (in 000 Rs.):	180	165	190

Solution: Since the data set consists of four characteristics (variables) at a time and values of each characteristic vary over time, we shall construct a vertical multiple bar diagram to represent diagrammatically the data set given. On Y-axis, a suitable scale is chosen so as to accommodate the entire range of values of all the four variables. On X-axis, groups of four vertical bars are drawn and there are three such groups to represent variations of the four variables over time, with equal gaps between each group. In a group there is no gap between the bars. This shows that a group of bars in multiple bar diagram is considered to be a single bar as compared to simple bar diagram. Further, to distinguish between bars drawn for four variables in a group, they are shaded in a different manner. The index given in the diagram describes which type of shading is used for a particular type of variable.

The multiple bar diagram constructed for the given data set is depicted below.



5.6.3 Sub-Divided Bar Diagrams

Sometimes values of one or more variables are not provided in absolute form, rather these are provided in the proportion or percentage form. For example, the results of a number of students belonging to a class are generally expressed in terms of percentages and not in absolute forms. This, in fact, is more so if the total marks differ either subject-wise or school-wise. Whenever values of variable(s) are given in terms of proportions or percentages, a new type of bar diagram is used to represent the data diagrammatically. It is popularly known as “Sub-Divided Bar Diagram”.

In sub-divided bar diagram, we draw a vertical bar with a good length and sufficient width. The length of the bar is kept equal to 100 as per the chosen scale on Y-axis. Thus, the length of the bar represents 100 percent. Now, either we calculate percentages of different values of the variable if the values are not provided in the form of percentages or these are already provided in the data set. Then, on the basis of percentages corresponding to different values of the variable, the length of the bar is divided into different segments.

However, it is not necessary always to convert the absolute values into percentages before we prepare a sub-divided bar diagram. We can also divide the bar using the absolute values. The only difference is that, as mentioned above, whenever we use percentage values, a scale on Y-axis is calibrated which shows the total height of the bar equal to 100%. If instead, we use the absolute values of the items for dividing the bar into segments, the scale on the Y-axis be so chosen that it covers all the values and the total height of the bar should be equal to the sum of all the values.

Let us explain the process with an example given below:

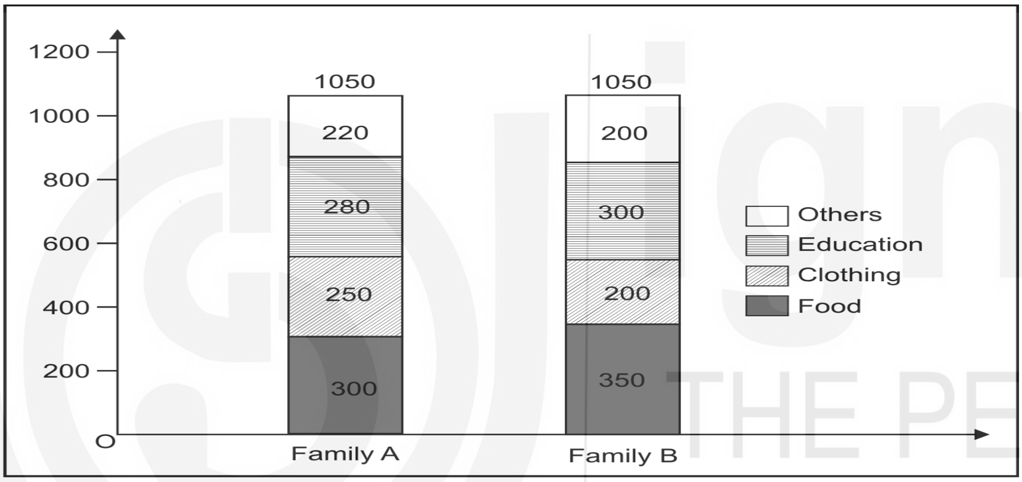
Example 3: Represent the given data, showing the expenditure (in rupees) of two families on different items:

Item	**Family A**	**Family B**
Food	300	350
Clothing	250	200
Education	280	300
Others	220	200

Solution: As required, expenditures of two families on the given items can be shown in the form of two separate sub-divided bars constructed in the same diagram. Since, the expenditures are given in their absolute values and not in the form of percentage, we may take help of the absolute values and, accordingly, can divide the bar into different segments. We simple use a scale on the Y-axis to accommodate all the values of the two families. Since the total of values is same (=1050) for both the families, the heights of both the bars must be equal to 1050 on the graph. The following table shows the conversion of values in percentages:

Items	Family A		Family B	
	Expenditure (in Rs)	Cumulative total	Expenditure (in Rs)	Cumulative total
Food	300	300	350	350
Clothing	250	550	200	550
Education	280	830	300	850
Others	220	1050	200	1050
Total	1050	----	1050	----

On the basis of the above table required sub-divided bar diagram is given below:



CHECK YOUR PROGRESS 2

Note: i) Check your answers with those given at the end of the unit.

3) The profit (in Rs crore) of a company from 1990-91 to 1999-2000 are given below:

Year	Profit (in Rs crore)	Year	Profit (in Rs crore)
1990-91	35.6	1995-96	87.2
1991-92	46.7	1996-97	113.1
1992-93	39.8	1997-98	123.6
1993-94	68.2	1998-99	119.7
1994-95	93.5	99-2000	130.8

Represent this data by a simple bar diagram.

4) Draw the multiple bar diagram for the following data.

Year	Sale (in ,000 Rs)	Gross profit (in '000 Rs)	Net profit (in, '000 Rs)
1990	100	30	10
1995	120	40	15
2000	130	45	25
2005	150	50	30
2010	200	70	30

5) Draw a sub-divided bar diagram for the following data:

Category	Cost Per Unit (1990)	Cost Per Unit (2000)
Material	20	32
Labour	25	36
Delivery	5	12
Total	50	80

5.7 PIE DIAGRAM OR PIE CHART

Pie diagram/chart may be considered as an alternative to sub-divided bar diagram. Whereas in sub-divided bar diagram, we use a bar and different components of a characteristic (variable) are represented on the basis of different segments in which a bar is divided; in pie diagram/chart we use a circle of any radius and the total area of the circle is divided into different segments of the area represent different components of the variable. Obviously, this means that depending upon the percentage of a particular component of the variable, we find the respective angle which is subtended by the respective area of the circle. Since, the total angle to be subtended by all the areas of segments will be 360^o, this calculation is equivalent to finding percentages in the bar diagram.

In order to construct a pie diagram, we do the following steps:

1) With the help of a compass, draw a perfect circle on graph paper with sufficient radius according to given space. The radius should be a horizontally drawn in order to simplify the further constructions.

2) Start with the first component of the considered variable and make an angle starting from this radius. The angle should be same as calculated for that particular component. Thus, we have a new radius where the angle ends. Draw it.

3) Next, take the second component of the variable and draw an angle equal to that obtained for this component starting from the new radius obtained as in the previous step. Find the new radius where the angle ends.

4) Repeat this process till all the components of the variable exhaust. You will see that the total area of the circle is divided into different segments.
5) Shade or colour all the segments so obtained in different ways or shades.
6) Prepare an index showing the shades/colours and corresponding components of the variable.

Let us explain the procedure with the help of the following example:

Example 4: Draw a pie diagram to represent the expenditure of Rs. 100 of a family on different budget heads as given below:

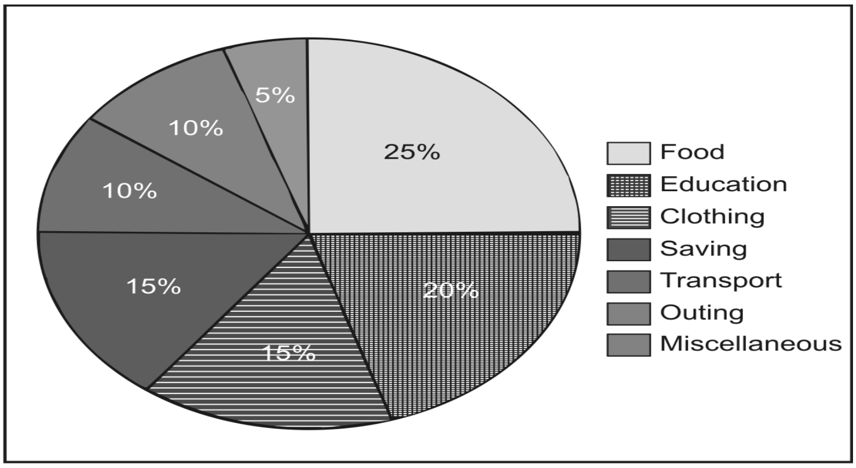
Item	Expenditure (in Rs.)
Food	25
Clothing	15
Education	20
Transport	10
Outing	10
Miscellaneous	5
Saving	15

Solution: The sum of expenditures is = 25+15+20+10+10+5+15=100 and, therefore, this total has to be equivalent to 360^o. Now, therefore, we compute the angles which different heads subtend at the center as follows:

Table 5.7: Calculation of Degrees and Percentages

Items	Expenditure (in Rs.)	Sector Angle (in degree)	Percentages
Food	25	$\frac{25}{100}\times 360 = 90$	$\frac{25}{100}\times 100 = 25$
Clothing	15	$\frac{15}{100}\times 360 = 54$	$\frac{15}{100}\times 100 = 15$
Education	20	$\frac{20}{100}\times 360 = 72$	$\frac{20}{100}\times 100 = 20$
Transportation	10	$\frac{10}{100}\times 360 = 36$	$\frac{10}{100}\times 100 = 10$
Outing	10	$\frac{10}{100}\times 360 = 36$	$\frac{10}{100}\times 100 = 10$
Miscellaneous	5	$\frac{5}{100}\times 360 = 18$	$\frac{5}{100}\times 100 = 5$
Saving	15	$\frac{15}{100}\times 360 = 54$	$\frac{15}{100}\times 100 = 15$
Total	100		100

On the basis of above calculations, pie chart which shows the expenditure of a family on different items is shown below:



CHECK YOUR PROGRESS 3

Note: i) Check your answers with those given at the end of the unit.

6) Represent the following data of utilization of 100 paise of income by XYZ company in year 2009-10.

Item/Head	Money spent (in paise)
Manufacturing Expenses	42
Salaries of employees	14
Selling and distribution Expenses	8
Interest Charges	6
Advertisement Expenses	15
Excise duty of sales	5
Taxation	10

5.8 PICTOGRAMS OR PICTORIAL DIAGRAMS

Pictograms, also known as pictorial diagrams, are very frequently used pictorial device in representing statistical data. Pictograms are drawn with the help of pictures which are generally miniatures of units on which certain measurements are taken or the number of the units it represents. It is up to the investigator to decide the units of measurement or the number of units a single miniature represents. These diagrams indicate towards the nature of the represented facts. Pictograms are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman.

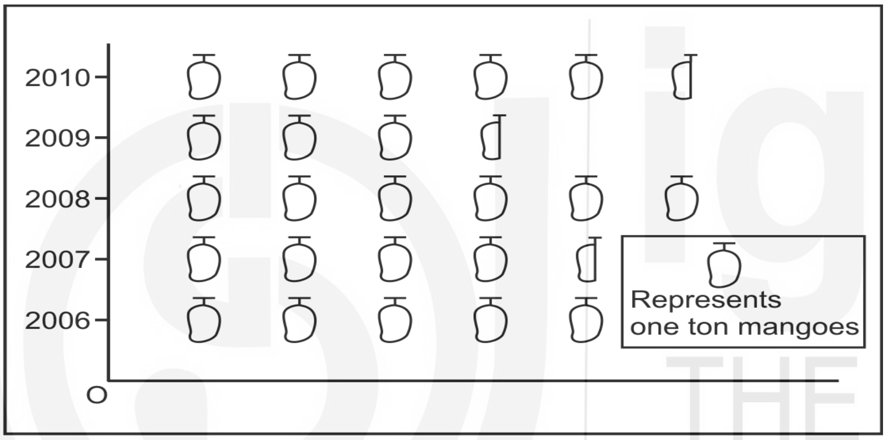
The picture which is used as symbols to represent the units or values of any variable or commodity are to be selected carefully, since to draw a complicated picture with free hand is not a job of statisticians; it requires expertise of an artist. Further, the picture symbol must be self-explanatory in nature. For example, if the increase in number of Airlines Company is to be shown over a period of time, then the appropriate symbol would be an airplane. To represent the number of houses in different wards in a city, a symbol, 'hut', may be used which is simple to draw even by anybody.

We use below an example to show how the information contained in the data set can be depicted using a pictogram:

Example 5: Draw a pictogram for production of mangoes in a particular area of Maharashtra from 2006 to 2010.

Year	2006	2007	2008	2009	2010
Production of Mangoes (in tons)	5.0	4.5	6.0	3.5	5.5

Solution: We locate the production of mangoes through the symbol (picture) of 'mango' for the different years according to different magnitude of the data (taking 1 mango = 1 tons mangoes). The pictogram for the data is shown below:



CHECK YOUR PROGRESS 4

Note: i) Check your answers with those given at the end of the unit.

7) Draw a pictogram for the data of production of tea (in hundred kg) in a particular area of Assam from year 2006 to 2010.

Year	2006	2007	2008	2009	2010
Production of Tea (in 100 kg.)	2.5	3.0	4.0	5.5	7.0

5.9 STATISTICAL MAPS OR CARTOGRAMS

Sometimes information (data) is represented diagrammatically using maps of some regions to which these data belong. Such maps are popularly known as Statistical Maps or Cartograms. This type of representation of data through statistical maps is generally used when one is interested to compare the variation or change in the values of the variable(s) under consideration in a particular country, state or region. For instance, meteorological departments of different countries use statistical maps to represent the region-wise annual rainfall during a year; state-wise population density; city-wise daily temperature, etc.

The quantities on the map can be shown in many ways, such as through shads or colours or by dots or by placing pictograms in each geographical area or by the appropriate numerical figure in each geographical area.

Let us take an example to get a look of the cartogram.

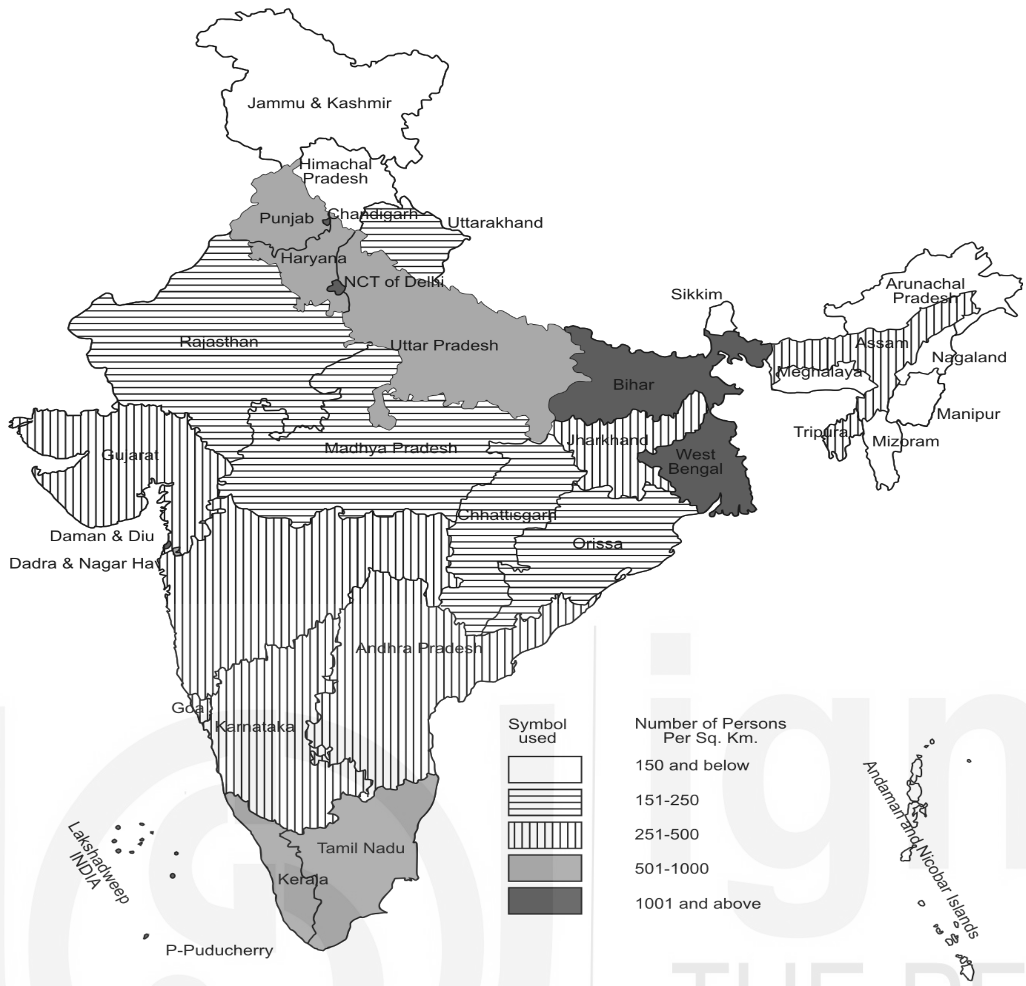
Example 6: Information on the population density per square kilometer in different states and union territories in India according to 2011 census are given below:

State/Union Territory	**Density (per sq. km.)**	**State/Union Territory**	**Density (per sq. km.)**	**State/Union Territory**	**Density (per sq. km.)**
Andhra Pradesh	308	Kerala	859	Tripura	350
Arunachal Pradesh	17	Madya Pradesh	236	Uttarakhand	189
Assam	397	Maharashtra	365	Uttar Pradesh	828
Bihar	1102	Manipur	122	West Bengal	1029
Chhattisgarh	189	Meghalaya	132	Andaman and Nicobar Islands	46
Goa	394	Mizoram	52	Chandigarh	9252
Gujarat	308	Nagaland	119	Dadar and Nagar Haveli	698
Haryana	573	Orissa	269	Daman and Diu	2169
Himachal Pradesh	123	Punjab	550	Delhi	11297
Jammu and Kashmir	124	Rajasthan	201	Lakshadweep	2013
Jharkhand	414	Sikkim	86	Pondicherry	2598
Karnataka	319	Tamil Nadu	555		

Represent the above data with the help of cartogram.

Solution: Since all the data are related to India, we need to consider the map of India with all of its states and union territories. We shade or colour each state and union territories in different styles so as to distinguish them from others. The Cartogram for the above data is given below along with the index showing the different ranges of population densities:

Density of Population, 2011



5.10 GRAPHICAL PRESENTATION

A graphical presentation is a geometric image of a set of data. Graphical presentations are very common in statistics whenever one wants to express the data, grouped into the form of a frequency distribution, in diagrammatical form. Unlike diagrams, these graphs are not only represent frequency distributions in diagrammatic form, but also very much useful in locating some partition values like median, quartiles, etc., in particular, and interpolate/extrapolate items in a series, in general. They are also used to measure absolute as well as relative changes in the data. They also help us in studying cause and effect relationship between two variables. The graph of a frequency distribution presents the huge data in an interesting and effective manner and brings to light the salient features of the data at a glance.

5.10.1 Methods of Graphical Presentation

The purpose of this section is to describe and discuss different types of graphs which are commonly used to represent frequency distributions graphically. In this context, we shall consider the following graphs:

(i) Histogram;

(ii) Frequency Polygon;

(iii) Frequency Curve;

(iv) Cumulative Frequency Curves (Ogives):

a. Less than type Ogive,

b. Greater than type Ogive.

Before we discuss above methods of graphical representation, you are advised to go through the concepts related to construction of a frequency distribution.

5.10.2 Histogram

A histogram is a graphical representation of a frequency distribution, constructed on the basis of distinct values of the variable and their corresponding frequencies in case of a discrete ungrouped frequency distribution; whereas in case of continuous grouped frequency distribution it is constructed on the basis of continuous class intervals and their corresponding frequencies. In both the cases, frequencies are shown on the Y-axis and distinct values/class intervals are plotted on X-axis.

Histogram can be defined as a diagram consisting of "a set of adjacent vertical bars drawn without allowing any gap between two consecutive bars; each bar drawn on the X-axis to represent a particular class interval of a frequency distribution, with its width equal to the length of the class and height of the bar equal to the frequency of the class (in case of equal class intervals) and equal to the frequency density of the class (in case of unequal class intervals)".

In other words, the frequencies of the class-intervals having the least width are written as they are and the frequencies of other class intervals are written as follows:

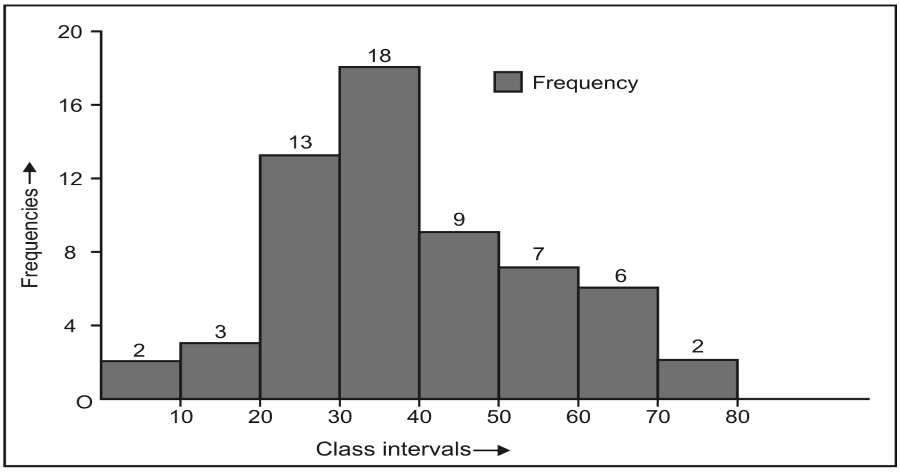
$$\frac{\text{Given frequency}}{\text{Width of its Class-interval}} \times (\text{The least width}) \qquad \ldots (5.1)$$

Let us draw a histogram to the following frequency distribution given below in Table 5.8:

Table 5.8

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	2	3	13	18	9	7	6	2

Histogram for Frequency Distribution when Class-intervals are of Equal Width is given below.



Now, let us consider the frequency distribution for unequal class intervals as given in Table 5.9:

Table 5.9

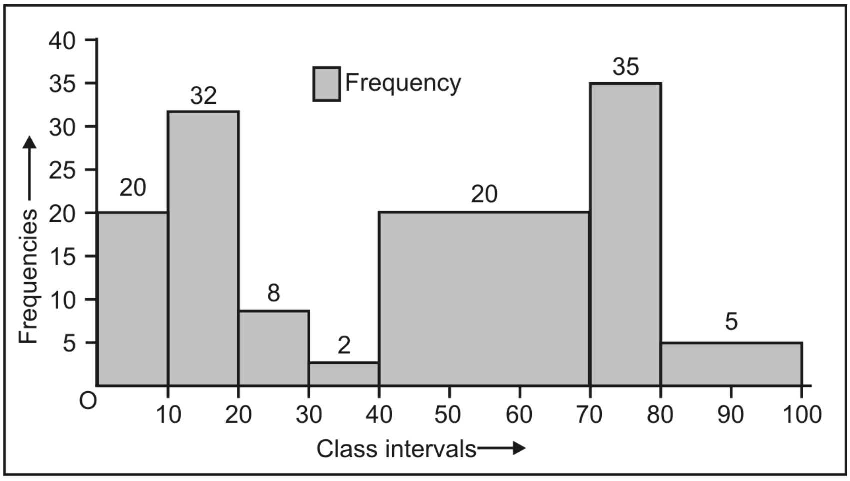
Class	0-10	10-20	20-30	30-40	40-70	70-80	80-100
Frequency	20	32	8	2	60	35	10

As it is a case of unequal class intervals, so we have to adjust the frequencies of the classes 40-70 and 80-100 by the formula suggested in equation 5.1. These calculations are shown below in the table 5.10:

Table 5.10

Class Interval (CI)	Frequency	Width of (CI)	Heights of the rectangles
0-10	20	10	20
10-20	32	10	32
20-30	8	10	8
30-40	2	10	2
40-70	60	30	$(60/30)\times 10 = 20$
70-80	35	10	35
80-100	10	20	$(10/20) \times 10 = 5$

The histogram for this frequency distribution when Class Intervals are of Unequal Width is shown in Figure given below.



Note 2: Sometimes, a histogram is also used for the frequency distribution of a discrete variable. Each value of the discrete variable is regarded as the mid-point of an interval. But generally, its use is not recommended, because in discrete case each frequency actually corresponds to a single point and not to an interval.

CHECK YOUR PROGRESS 5

Note: i) Check your answers with those given at the end of the unit.

8) Draw a histogram from the following data

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	2	3	13	18	9	7	6	2

9) Draw a histogram for the following frequency distribution

Wages (Rs)	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50
No of Workers	30	70	100	110	140	150	130	100	90	60

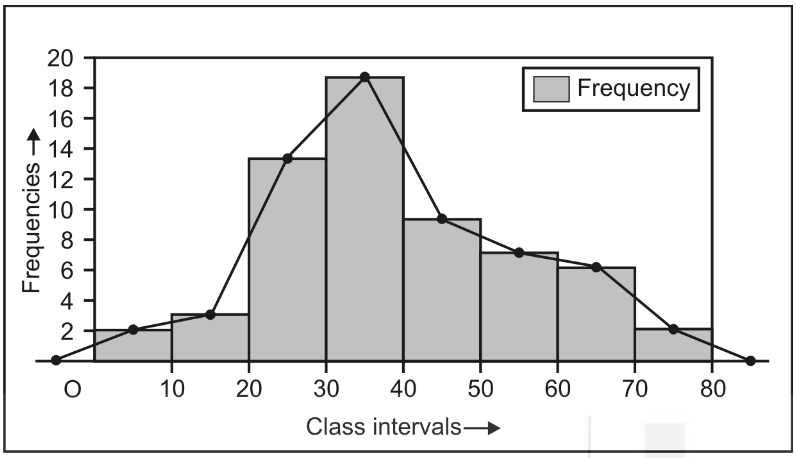
5.10.3 Frequency Polygon

Another important graphical representation of a frequency distribution is done by "Frequency Polygon". As the name implies, a frequency polygon is a polygon drawn using mid-values of each class and corresponding frequency of the class. The frequency polygon can be drawn over the histogram of the data or can be drawn independently.

In order to draw a frequency polygon, first of all mid values of all the class intervals and corresponding frequencies are plotted on X-axis and Y-axis, respectively. Secondly, we join these plotted points in sequence by line segments. The graph, thus, obtained is seen to be a jig-jag line from the first point to the last point. However, to convert this into a polygon, we join both the ends of the jig-jag line to X-axis under the rule that the area of the polygon should be equal to the area of the histogram. This is possible only when we assume two imaginary classes; one before the first class with the same width as of the first class but with frequency zero and another class just following the last with its width equal to the last class but its frequency to be zero. After this, we join the first point by a straight line with the mid-value of the first imaginary class and similarly, join the last point by a straight line with the mid-value of the last imaginary class. In this way, we see that we get a polygon which is known as "frequency polygon" of the given frequency distribution. It can be shown that the frequency polygon so obtained, would satisfy the property that

Area under the polygon = Area of the histogram.

For example, if we take the frequency distribution as given in Table 5.5 then, we have to first plot the points (5, 2), (15, 3), …, (75, 2) on graph paper along with the horizontal bars. Then we join the successive points (including the mid points of two imaginary class intervals each with zero frequency) by line segments to get a frequency polygon. The frequency polygon for frequency distribution given in Table 5.5 is shown in following Figure:

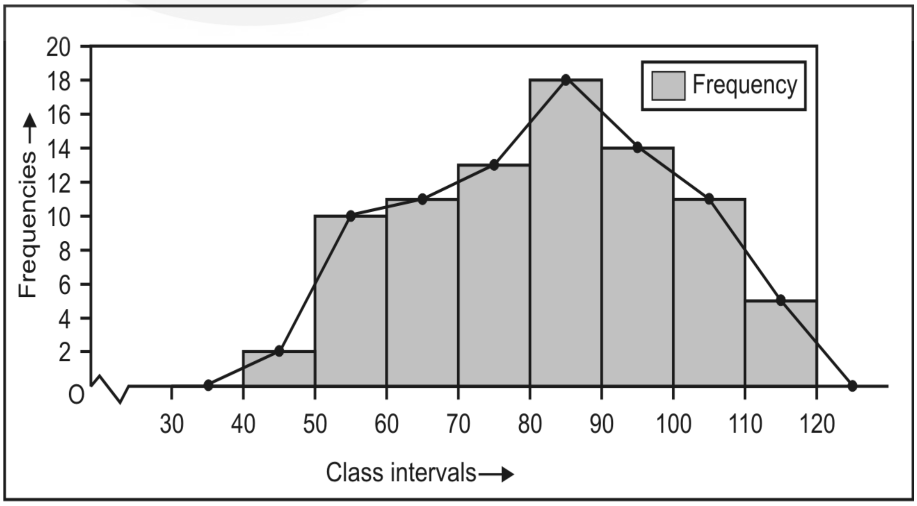


Note 3: In some cases, first class interval does not start from zero. In such situations we mark a kink on the horizontal axis, which will indicate the continuity of the scale starting from zero. Let us take an example of this type.

Example 7: Draw a frequency polygon for the following frequency distribution:

Class Interval	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120
Frequency	4	10	11	13	18	14	11	5

Solution: Frequency polygon for the given data is shown in following Figure:



5.10.4 Frequency Curve

In simple words frequency curve is a smooth curve obtained by joining the points (not necessary all points) of the frequency polygon such that

(a) like frequency curve, it also starts from the base line (horizontal axis) and ends at the base line and

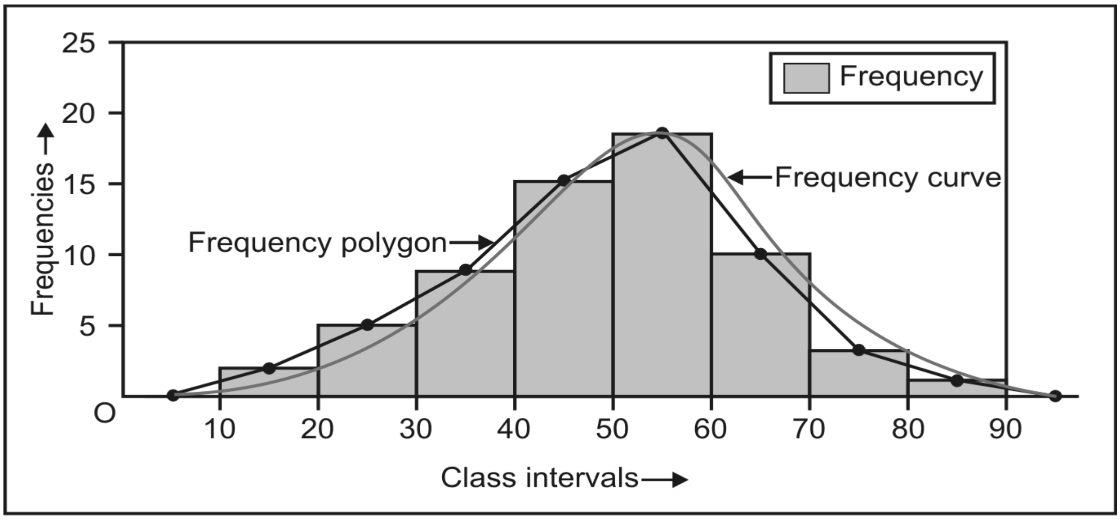
(b) area under frequency curve remains approximately equal to the area under the frequency polygon.

In other words, let us try to explain the concept theoretically. We know that grouping of values of the variable concerned in class intervals is made mainly due to the reason that infinitely large number of values are neither possible nor advisable to tackle separately. However, if class intervals are large enough, very few classes can accommodate all the values, but then we lose many of the characteristics of the variable which should not be hidden in any graphical representation. Therefore, theoretically there should be as many class intervals as possible and thus, having infinitely large number of bars in the corresponding histogram. If number of bars in the histogram is infinitely increased, the mid-values of classes would come too close to each other and this would provide the frequency polygon in the form of a smooth-hand curve. This smooth-hand curve would then be considered to be the frequency polygon, but as it would be like a smooth curve, it is popularly called "Frequency Curve" instead of frequency polygon. In fact, frequency curve is the graph of a continuous variable. So, theoretically continuity of the variable implies that whatever small class interval we take there will be some observations in that class interval. That is, in this case there will be large number of line segments and the frequency polygon tends to coincides with the smooth curve passing through these points as sample size (n) increases.

For the following example, we have drawn both frequency polygon and frequency curve to make the idea clear to you.

Class Intervals	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	2	5	8	15	18	10	3	1

Frequency polygon and frequency curve for the above data is shown below:



CHECK YOUR PROGRESS 6

Note: i) Check your answers with those given at the end of the unit.

10) Draw a frequency polygon from the following data

Class Interval	40-50	50-60	60-70	70-80	80-90	90-100	100-110	110-120
Frequency	4	10	11	13	18	14	11	5

11) Draw a frequency curve from the following frequency distribution:

Class Intervals	1- 2	2 - 3	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9
Frequency	3	9	20	25	18	12	6	3

5.10.5 Cumulative Frequency Curves (Ogives)

We shall now discuss next type of graphical presentation of a frequency distribution which utilize concepts of above three types of cumulative frequencies. These graphical presentations are called **"Cumulative Frequency Curves"** or **"Ogives"**.

(i) Less Than Type Cumulative Frequency Curve (Less Than Type Ogive):

Less than type ogive is a line diagram drawn with upper class boundary of the class interval, plotted on the X-axis and less than type cumulative frequency ($F_<$), plotted on the Y-axis by choosing suitable scales on both the axes. Then on the graph paper, locate the co-ordinates of all the pairs of upper class boundaries and corresponding cumulative frequencies on the graph paper. Finally, join these points by line segments starting from the first to last point. Such a line diagram is called "Less than type cumulative frequency curve" or "Less than type ogive".

It can be seen that less than type ogive is a non-decreasing (increasing) curve. Its shape resembles with the English capital letter 'S'.

We use the following fictitious data for explaining its construction:

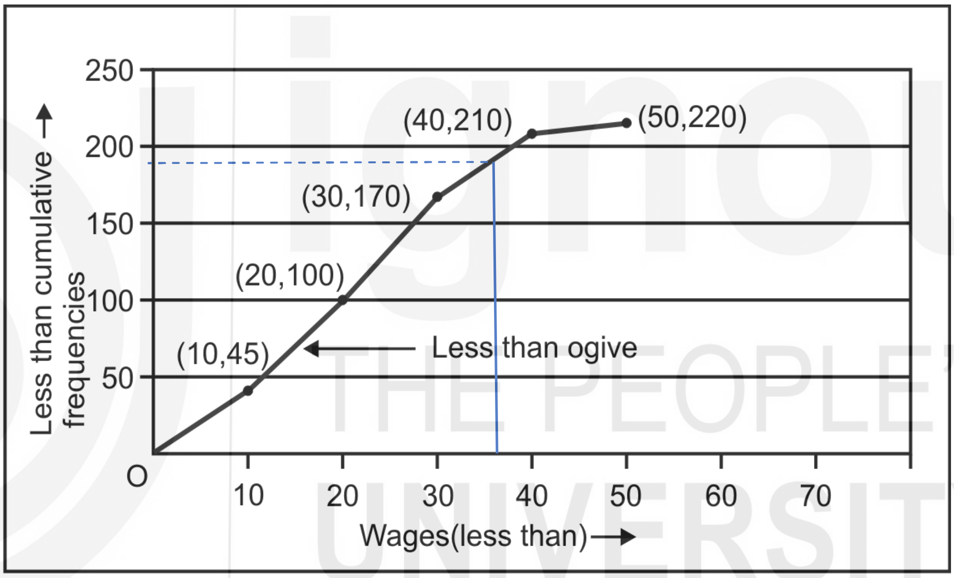
Table 5.11: Wage-Wise Distribution of Workers

Weekly Wages (in Rupees)	0-10	10-20	20-30	30-40	40-50
No. of workers	45	55	70	40	10

We observe that data are available in grouped form with continuity of class intervals. Therefore, there is no need to find class boundaries of the classes, these are already given in the table. Thus, upper class boundaries of the classes are respectively 10, 20, 30, 40 and 50. We find less than cumulative frequencies corresponding to each class in the following table along with corresponding upper class boundaries against which $F_<$ frequencies would be plotted:

Weekly Wages	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Number of Workers	45	55	70	40	10
Upper Class Boundaries	10	20	30	40	50
$F_<$	45	100	170	210	220

In the table, third row shows upper class boundaries of the classes or equivalently, the upper limits of the weekly wages of each class. For example, 45 workers are getting weekly wage less than Rs. 10; 45 +55 = 100 workers are getting weekly wage less than Rs. 20 and so on. Using the figures given in third and fourth rows of the table, we draw the less than type ogive, which is depicted below:



We see that the less than type ogive is a non-decreasing line diagram. This graph shows that how many workers are there who are getting less than a particular weekly wage. For example, in order to estimate the number of workers who are getting weekly wage less than Rs. 35, we draw a vertical line over the value 35 on X-axis which touches the curve at a certain point. From that point we draw a horizontal line which touches the Y-axis at a point (both the vertical and horizontal lines are shown in the graph in blue colour). Read the value of the point on Y-axis at which the horizontal line touches the axis; which will provide us the number of workers getting the weekly wage less than Rs. 35. From the graph we observe that it is approximately 180.

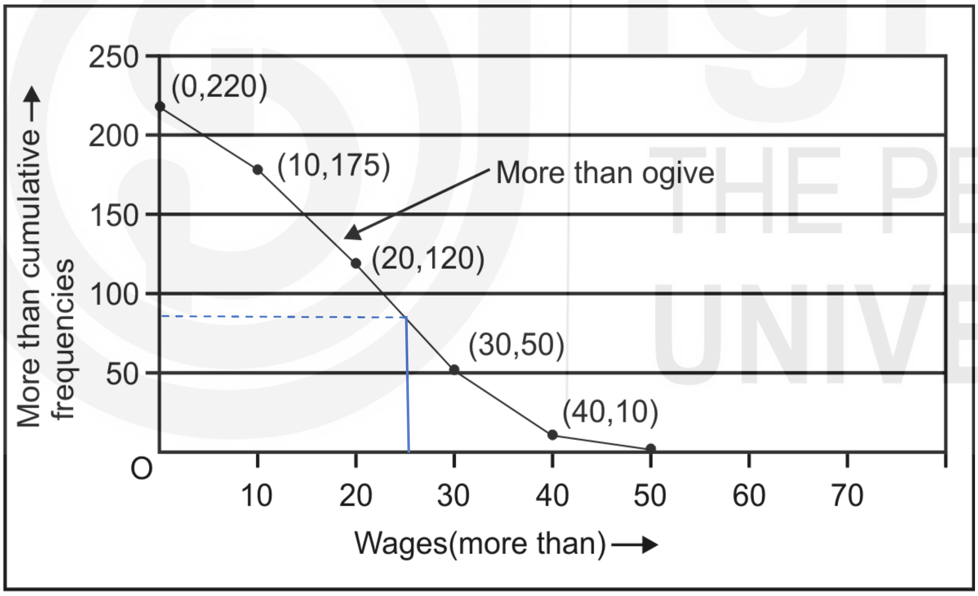
(ii) Greater Than Type Cumulative Frequency Curve (Greater Than Type Ogive):

Similar to less than type ogive, it is also a line diagram which is plotted on a graph using lower class boundaries of the class intervals and corresponding greater than type cumulative frequencies. The method of construction of this curve is same as that of less than type ogive except that on the X-axis we show lower class boundaries of the classes and on Y-axis, greater than type cumulative frequencies ($F_>$). Contrast to less than type ogive, greater than type ogive is a non- increasing (decreasing) curve.

Let us draw the greater than type ogive for the data shown in Table 4.9. We see that lower class boundaries are respectively 0, 10, 20, 30 and 40. The following table is prepared for other computing other values to draw the ogive.

Weekly Wages	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Number of Workers	45	55	70	40	10
Lower Class Boundaries	0	10	20	30	40
$F_>$	220	175	120	50	10

The desired greater than type ogive for the data given in Table 5.8 is shown below:



This seems to be a decreasing curve. This curve is helpful to find the number of workers who are getting weekly wage greater than a particular value. For example, the number of workers getting weekly wage greater than Rs. 25 is approximately 90 (see the blue lines in the figure for these values).

Remark: If both the ogives are drawn simultaneously on the same graph paper, it can be seen that the intersection point of these curves provide us approximate value of median of the given data. Also, different quartiles, deciles and percentiles can graphically be estimated using any of these ogives.

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

12) Draw less than ogive from the following frequency distribution of marks of 90 students:

Class	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79
Frequency	7	11	19	8	20	14	8	3

13) Draw the more than ogive for the following frequency distribution of the weekly wages of workers:

Class	0-10	10-20	20-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	5	15	20	30	45	35	25	15	10

5.11 LET US SUM UP

In this unit, we have discussed:

1) the meaning of tabulation and different components of a good table;
2) different types of tabular representation of data;
3) meaning of diagrammatic representation of data and the reasons for using this method of representation;
4) different methods of diagrammatic representation of data;
5) under what situations we can use simple bar diagram, multiple bar diagram, sub-divided bar diagram;
6) pie diagram/pie chart and how it is constructed;
7) pictograms & cartogram and how we can construct them;
8) various types of graphical presentation of data;
9) the concept of histogram and methods of their construction;
10) the concept of frequency polygon and frequency curves for a frequency distribution and their construction; and
11) the concepts of less than type ogive and greater than type ogive and their construction.

5.12 KEY WORDS

Data : A collection of observations or scores from a survey or an experiment.

Frequency distribution : A collection of observations produced by sorting observations into classes that show their frequency (f) of occurrence.

Histogram : A bar-type graph for quantitative data, with no gaps between adjacent bars.

5.13 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

5.14 ANSWERS TO CHECK YOUR PROGRESS

1) The following table is the representation of the data for the given information's regarding the drinkers in city X and Y.

Table 5.12: Data regarding the Drinkers in City X and City Y

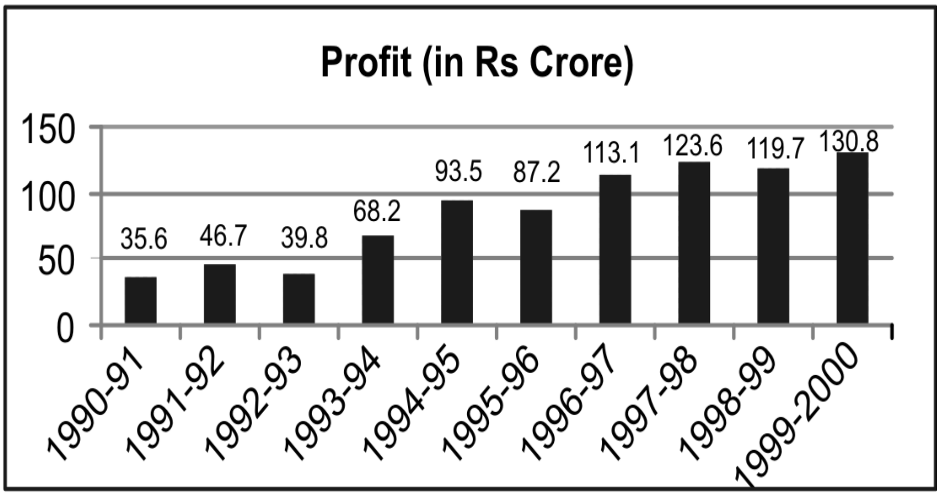
Attributes	City X		Total	City Y		Total
	Males	Females		Males	Females	
Drinkers	14	8	22	21	7	28
Non-drinkers	43	35	78	31	41	72
Total	57	43	100	52	48	100

2) The following table is showing the trade union membership:

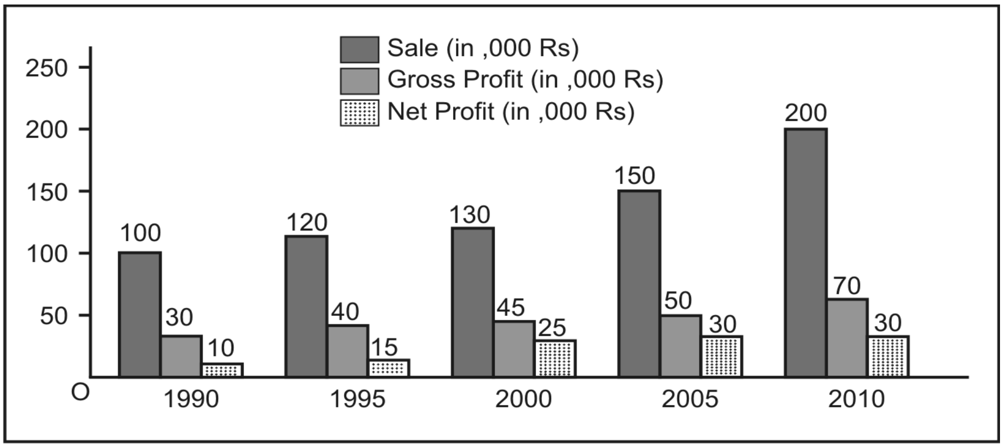
Table 5.13: Data regarding the Trade Union Membership of Workers of a Company in the Year 2009 and 2010

Category	2009			2010		Total
	Trade Union Members	None Union Members	Total	Trade Union Members	None Union Members	
Men	1500	250	1750	1600	225	1825
Women	50	200	250	125	155	280
Total	1550	450	2000	1725	380	2105

3) The simple bar diagram of the given data is shown below:



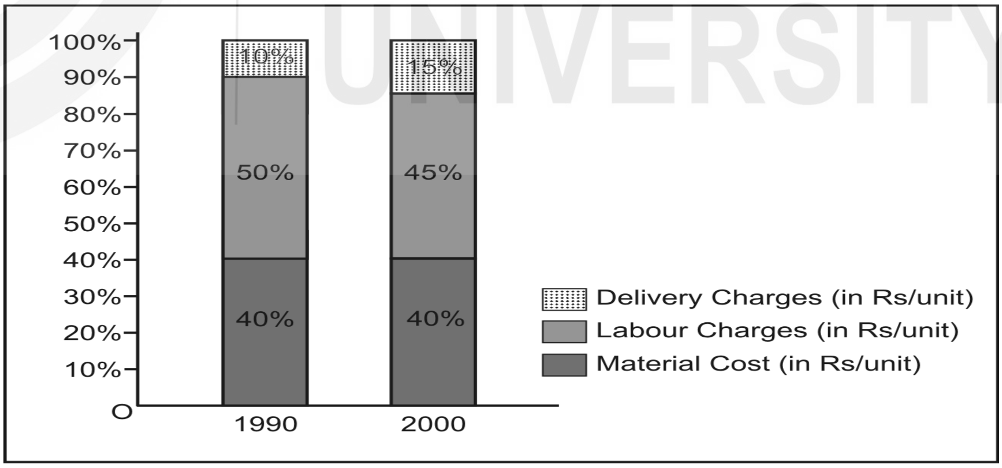
4) **Solution:** Multiple bar diagram for the given data is given below:



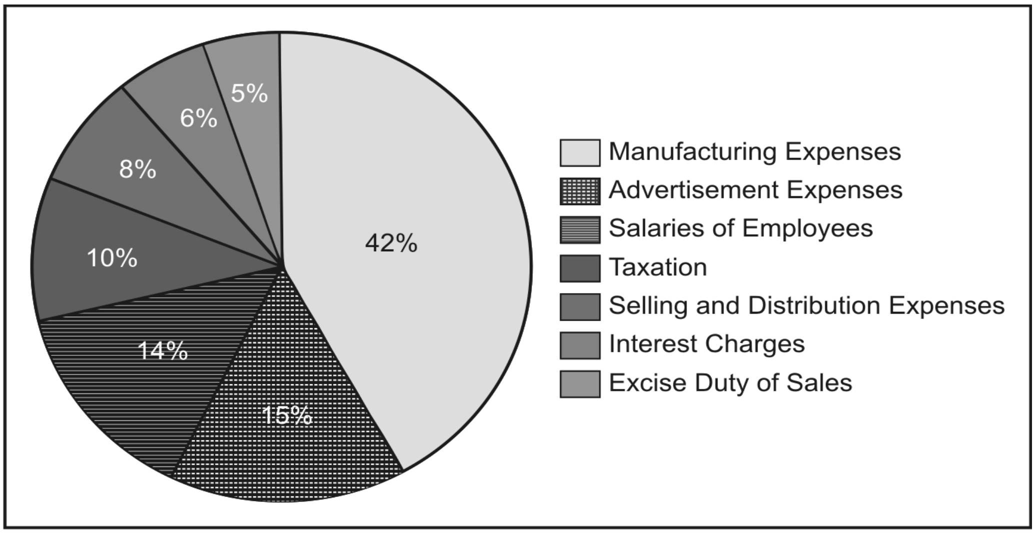
5) First of all, percentages and cumulative percentages are obtained for both the years in various category.

Category	Cost Per Unit (1990)			Cost Per Unit (2000)		
	Cost	% Cost	Cumulative % Cost	Cost	% Cost	Cumulative % Cost
Material	20	40	40	32	40	40
Labour	25	50	90	36	45	85
Delivery	5	10	100	12	15	100
Total	50	100		80	100	

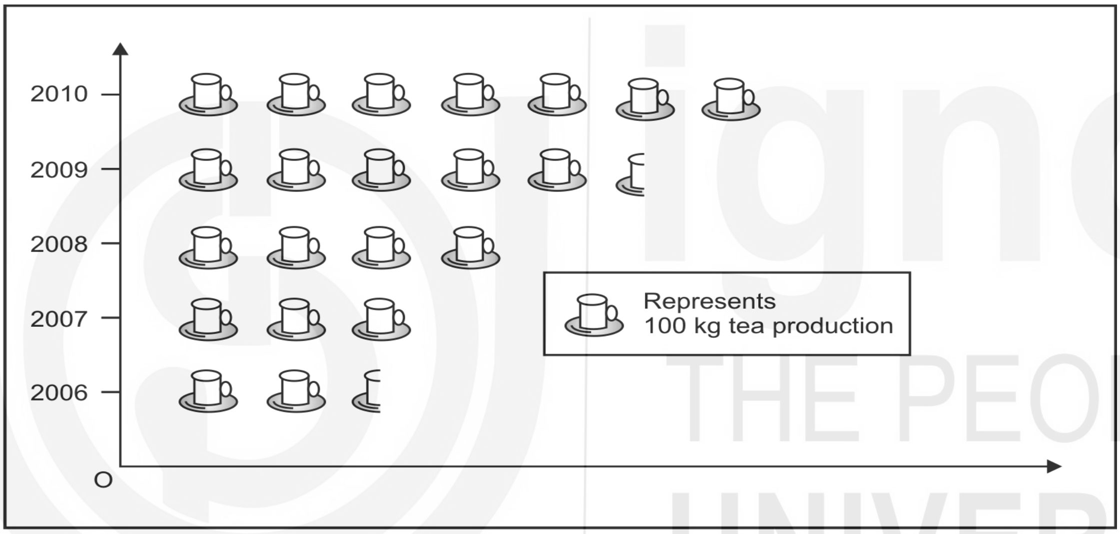
On the basis of above table required percentage bar diagram is given below:



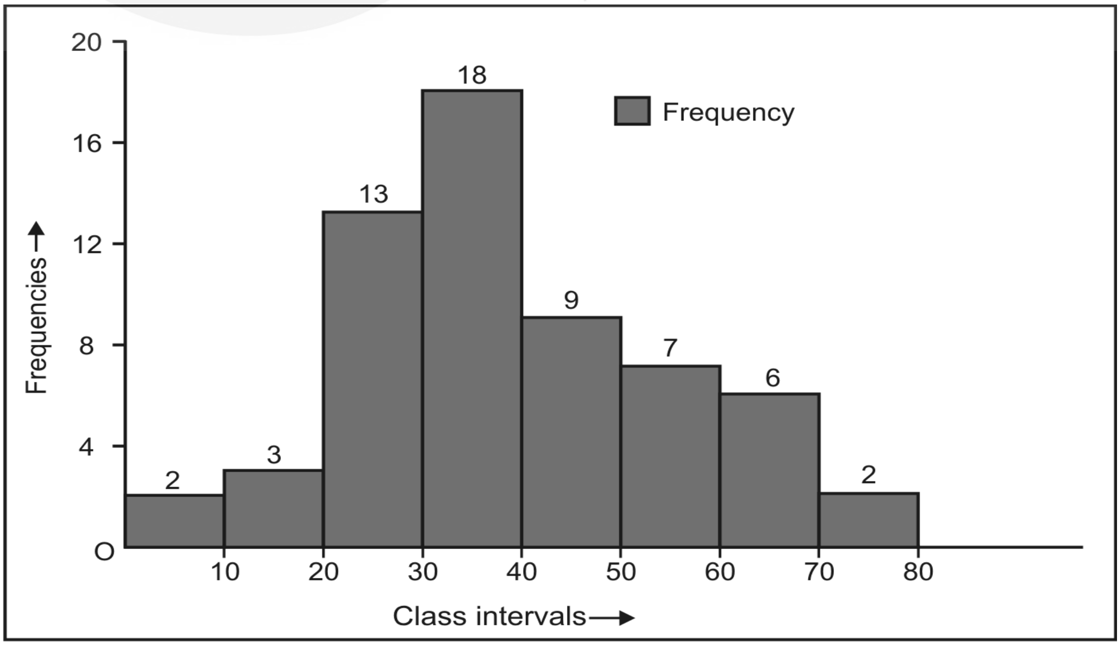
6) The suitable diagram in this case is pie diagram. On the basis of calculations of angles and percentages of each component as we did in Example 7, pie chart which shows the utilization of 100 paise of income by XYZ company in year 2009-2010 is shown below:



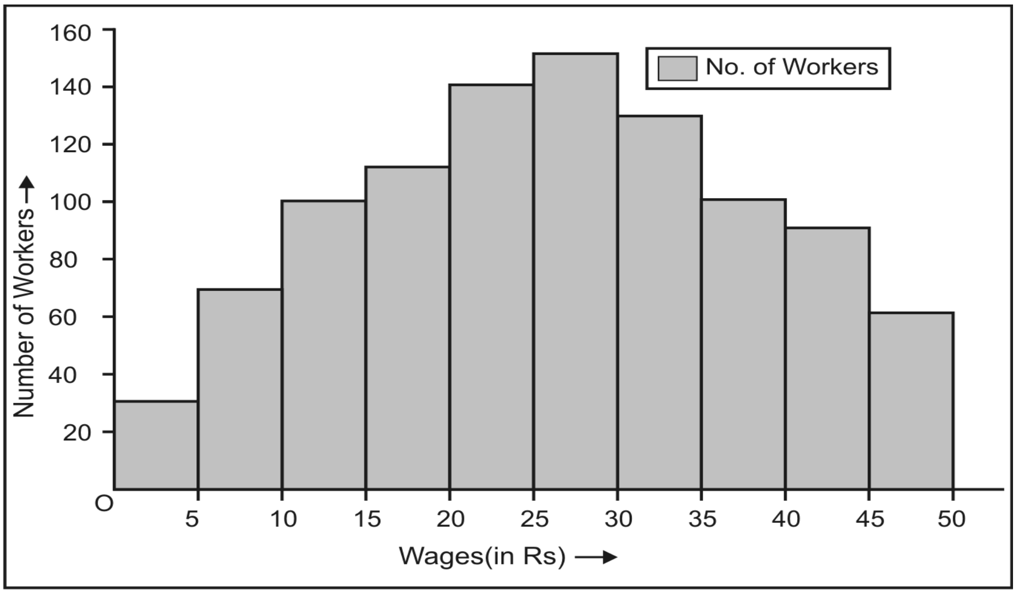
7) Pictogram for the production of tea in a particular area of Assam from year 2006 to 2010 is shown below:



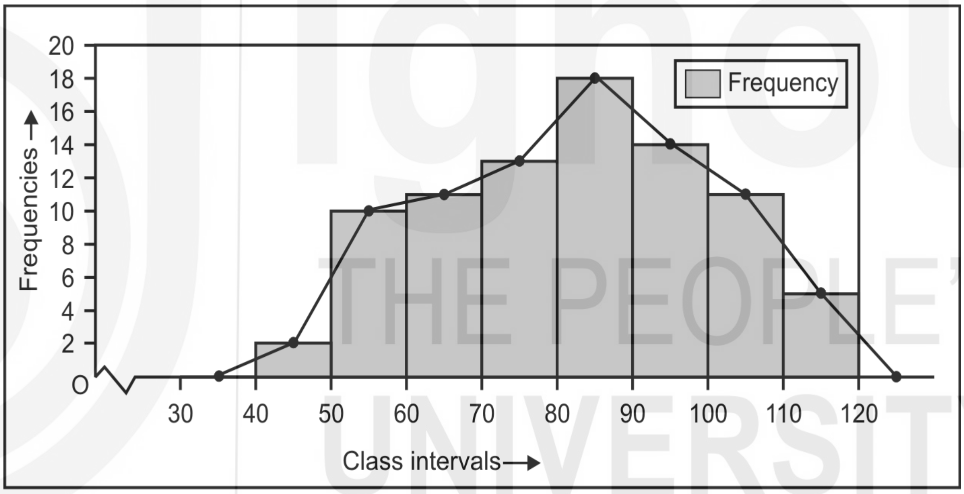
8) Histogram of the given data is given below:



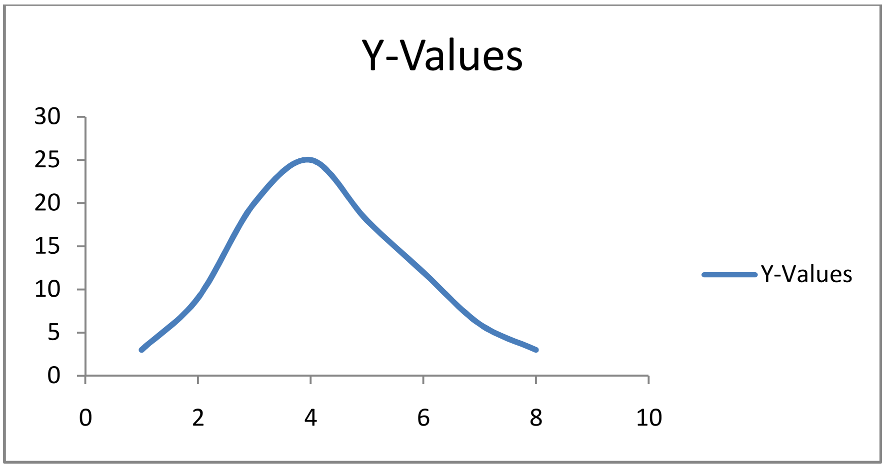
9) **Histogram** of the given data is given below**:**



10) Frequency polygon of the given data by first drawing histogram is given on the next page.

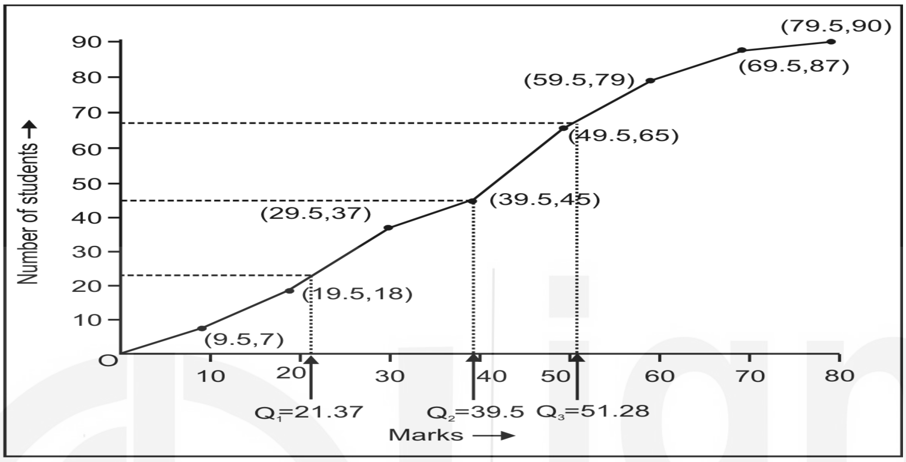


11) Frequency curve for the given data is given below.



12) Less than ogive for the given frequency distribution is given below.

To find $Q_1 : \frac{N}{4} = \frac{90}{4} = 22.5$ first of all we have to draw a dotted horizontal line starting from y-axis at 22.5 $\left(\text{as} \frac{N}{4} = 22.5 \right)$ and then from the point where it meets the less ogive, we shall draw dotted vertical line, the value corresponding to the point where it meets the horizontal axis is the value of Q_1 as shown in figure. Similarly, values of Q_2 and Q_3 can be obtained as shown in the figure below:



13) More than ogive for the given frequency distribution is given below:

UNIT 6 GEOSPATIAL TOOLS

Structure

6.1 INTRODUCTION

Geospatial data and tools have now become very significant in environmental studies in many application areas. Geospatial technologies such as "Remote Sensing", Geographic Information System (GIS) and Global Navigation Satellite System (GNSS) have very high potential in environmental applications and are being used for the creation of baseline databases, mapping and periodic monitoring of features and areas of environmental concerns, environmental impact assessment, modelling and prediction of future scenarios, etc. In this unit, you will learn the basics of the three geospatial technologies i.e. remote sensing, GIS and GNSS and their potential in environmental applications.

6.2 OBJECTIVES

After studying this unit, you should be able to:

- define remote sensing and discuss its physical basis, types, and advantages;
- explain GIS including its components and advantages;
- discuss the basics of GNSS and its functions;

- describe application potential of geospatial technologies in environmental studies; and
- list different kinds of geospatial data and their sources.

6.3 BASIC CONCEPTS

In this section, we shall learn the basics of different geospatial technologies such as remote sensing, Geographical Information System (GIS) and Global Navigation Satellite System (GNSS), which have immense potential in various environmental studies. Geospatial technologies have become important to decision-makers and planners as it enables them to acquire, process, analyse, visualise spatial information and produce outputs representing different scenarios useful for planning and management. Several national organisations, industries, and environmental agencies together with local and national governments are harnessing the advantages of geospatial technologies in their functioning.

Now, let us understand the definition of geospatial technology, its components, and geospatial data. The term *geospatial* (*geo* meaning the earth and spatial referring to space) refers to any space on the Earth. And the term *geospatial data* refers to data related to a part of the Earth. The term *geospatial technology* is relatively new and refers to the technologies, which deal with geospatial data. Remote sensing, GNSS and GIS are the three important components of geospatial technology. In this section, we would learn about the three major components and their advantages.

Let us first learn about remote sensing.

6.3.1 Remote Sensing

In simple terms, ***remote sensing*** is the process of acquiring information (i.e., *sensing*) from a distance (i.e., *remotely*) or collection of data about an object from a distance. *Remote* means without coming into physical contact with the features and *sensing* (acquisition of information) of earth surface features through some instruments placed on aerial or satellite platforms. Environmentalists use remote sensing techniques to monitor or measure phenomena found in the Earth's lithosphere, biosphere, hydrosphere, and atmosphere. When we observe an object through our eyes, we are engaged in remote sensing and our eyes act as remote sensors. However, remote sensing is usually done with the help of a mechanical device known as a remote sensor. This device has greatly improved the ability to receive and record information about an object without having any physical contact with them. Often, these sensors are positioned away from the object of interest by using helicopters, planes, and satellites. Most remote sensing devices record information about an object by measuring an object's transmission of electromagnetic energy from reflecting and radiating surfaces.

There are several formal definitions of remote sensing. According to the American Society for Photogrammetry and Remote Sensing (ASPRS), remote sensing is "*the measurement or acquisition of information of some property of an object or phenomenon, by a recording device that is not in physical or intimate contact with the object or phenomenon under study*" (Colwell, 1983).

Campbell (2002) has defined remote sensing as "*the practice of deriving information about the Earth's land and water surfaces using images acquired from an overhead perspective, using electromagnetic radiation in one or more regions of the electromagnetic spectrum, reflected or emitted from the Earth's surface*".

According to Lillesand and Kiefer (1979), remote sensing is "*the science and art of obtaining useful information about an object, area or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation*".

This definition refers to remote sensing as both an 'art' and 'science'. Remote sensing is a 'science' when information is extracted from remote sensing data using some scientific facts whereas it is an 'art' when information is derived through background or prior knowledge obtained through experience. Remote sensing analysts derive information from remote sensing data using both scientific facts and their experience.

After learning what remote sensing is let us now discuss the remote sensing process.

(i) Remote Sensing Process

The process of remote sensing involves several steps. We now know that our eyes are an example of remote sensors. We observe objects because there is a source of radiation that interacts with the objects and the radiated light is captured by our eyes (i.e., sensor). The eyes send a signal (data) to our brain (i.e., processor) that processes the data and converts it into information. These remote sensing processes are shown in Fig. 6.1 and discussed here.

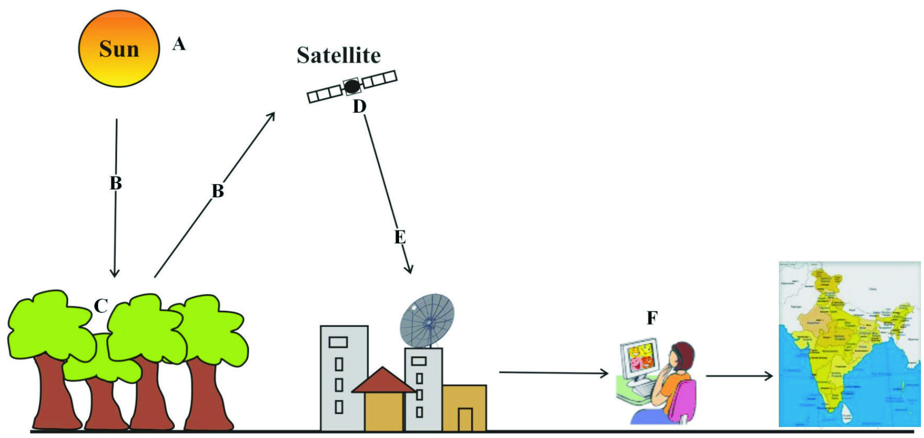


Fig. 6.1: Remote sensing process: A – Energy source, B – Energy interactions with the atmosphere, C – Interaction of EMR with Earth surface features, D – Recording of energy by the sensor, E – transmission of remotely sensed data, their reception by ground station and their processing, and F – interpretation and analysis of the data. Information derived from remote sensing data is generally presented in the form of a map or table. The softcopy of the course material has the colour version of the image.

(**Source:** http://egyankosh.ac.in/handle/123456789/39605 page 8.).

A **- Source of Energy:** The first and very important requirement for remote sensing is an energy source which provides electromagnetic energy to the Earth. It may be either from natural (e.g., solar radiation) or artificial (e.g., RADAR) sources. For remote sensing, Sun's radiations are commonly used as a source of energy.

B - **Interaction of energy with the atmosphere:** When energy travels from its source to the Earth's surface, it meets the Earth's atmosphere where it interacts with atmospheric constituents. The energy reflected from Earth's surface is received by remote sensors. In this process, the energy once again interacts with the atmosphere.

C - **Interaction of energy with Earth surface features:** Energy reaching the Earth's surface through the atmosphere interacts with the Earth's surface features. The interaction and its outcome depend on the characteristics of the features and the energy.

D **- Recording of the energy by sensor:** After interacting with Earth's surface features, the reflected and emitted energy travels to the sensor. And, the sensor records the reflected and emitted energy.

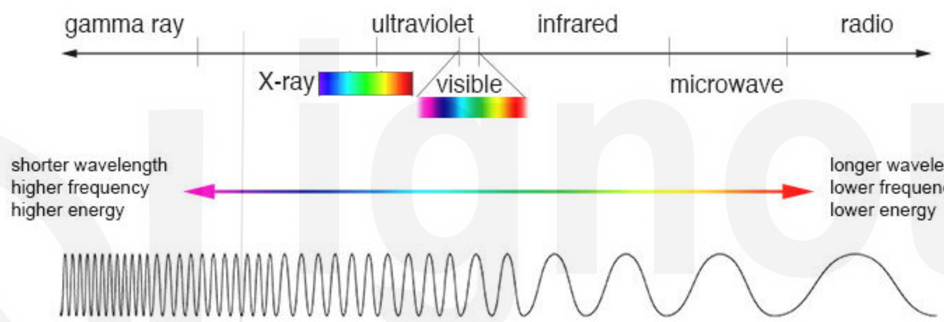
E **- Transmission, reception, and processing of the recorded signals:** The energy recorded by the sensor is transmitted in the form of signals to earth stations that receive and process the signals. The electronic signals are processed and converted into an image.

F – **Interpretation and analysis of the data:** The processed image is interpreted and analysed to extract information about the object of interest.

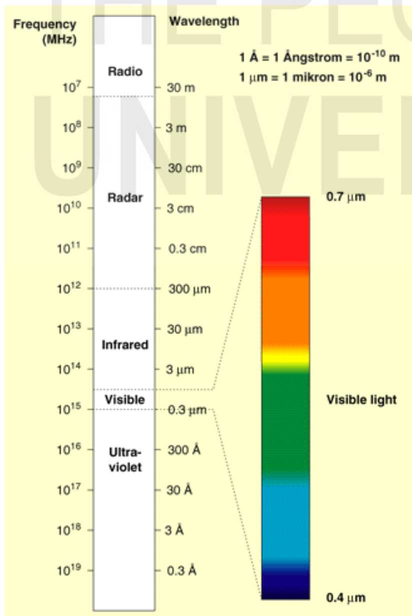
(ii) Physical Basis of Remote Sensing

We cannot see anything in dark. We can see something only in the presence of light meaning when the object is illuminated with a source of energy. The sense associated with the eyes is known as sight or vision and is stimulated by light. The eyes convert the incoming light into electrical signals and convey them to the brain, which creates pictures/images in our mind after processing. Similarly, as you see in Fig. 6.1, a remote sensing device records energy that has interacted with Earth's surface features. This energy serves as the main communication link between devices and objects. So, for remote sensing, the first requirement is a source of energy and the natural source of energy is the electromagnetic energy emanating from the Sun. As you know energy is transferred in three ways i.e., conduction, convection, and radiation. Unlike conduction and convection, radiation does not require a medium to transfer energy through. Radiant energy is defined as the energy carried by electromagnetic radiation. You know that everything that has a temperature above absolute zero radiates energy. In the context of remote sensing, our interest is in the transfer of energy by radiation because it can take place in a vacuum and the atmosphere in the region between the Sun and the Earth.

Electromagnetic radiation extends over a wide range of wavelengths. The entire range of radiant energies is called **electromagnetic spectrum** (EMS). EMR ranges from very high energy radiation such as gamma rays and X rays through ultraviolet light, visible light, infrared radiation, and microwaves to radio waves as shown in Fig. 6.2. These divisions are ordered according to their frequency, wavelength, or wave number. Our eyes are sensitive to only a small portion of the EMS called visible range. Radiations beyond the visible region are invisible to our eyes but can be detected by some remote sensing devices such as thermal sensors that can sense the radiations in the thermal infrared region. You should note that the division of the electromagnetic spectrum is for practical purposes. The gamma rays and X-rays are potentially dangerous and the ultraviolet light is powerful enough to cause sunburn. You can feel infrared radiation as heat. You know that we employ microwaves in microwave ovens and radio waves are used for communications. All types of EMR are waveforms which travel at the speed of light. The radiation can be defined in terms of either its wavelength or frequency. Shorter wavelength radiations (infrared or shorter) are generally described in terms of their wavelength, whereas longer wavelength radiations (microwave, etc.) are generally described in terms of their frequency.



(a)



(b)

Fig. 6.2: The Electromagnetic Spectrum: a) Major Divisions; b) Divisions with their Wavelengths and Frequencies. The softcopy of the course material has the colour version of the image.

(**Source:** a https://imagine.gsfc.nasa.gov/Images/science/EM_spectrum_compare_level1_lg.jpg; b- www.esa.int/images/02-spektrum.gif)

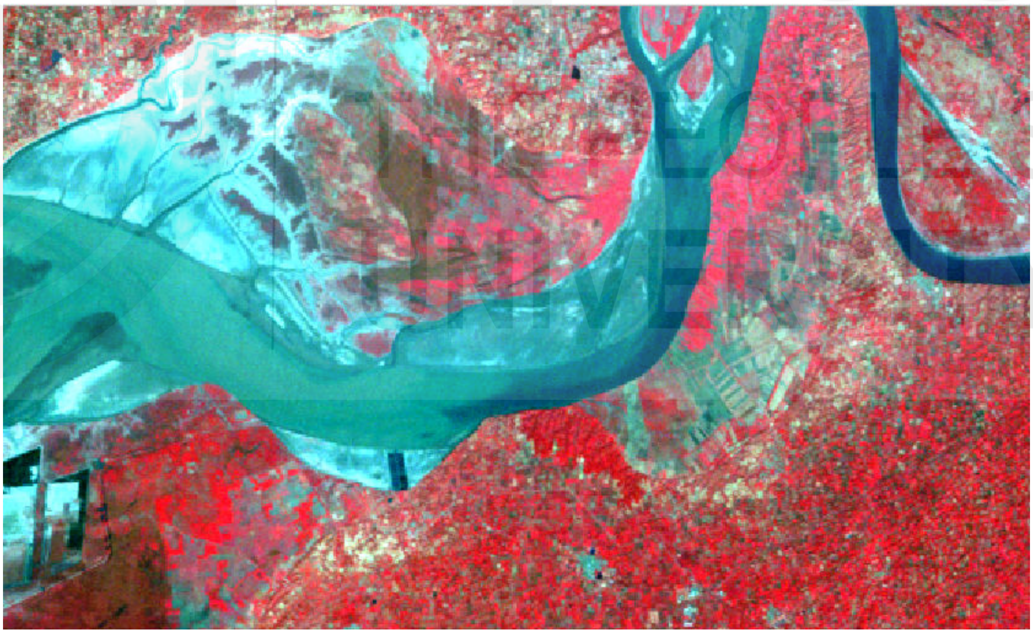
The electromagnetic spectrum is divided into different regions based on wavelength for practical purposes. Although, the electromagnetic spectrum ranges from gamma rays to radio waves, remote sensing is generally restricted to the range from ultraviolet to microwave regions as shown in Table 6.1. There are three major regions in EMR-Earth interaction that are important in remote sensing. The visible and NIR spectral band from 0.3 µm to 3 µm is known as the ***reflective region***. In this band, the Sun's radiation sensed by the sensor is reflected by the Earth's surface. The band corresponding to the atmospheric window between 8 µm and 14 µm is known as the ***thermal infrared band***. The energy available in this band for remote sensing is due to thermal emissions from the Earth's surface. Both reflection and self-emission are important in the intermediate band from 3 µm to 5.5 µm.

Table 6.1: Divisions of the electromagnetic spectrum used in remote sensing

Division	**Subdivision**	**Wavelength**
Ultraviolet (UV) (0.01 - 0.38 µm)		0.01 - 0.38 µm (10–380 nm)
Visible (VIS) (0.4 - 0.7 µm)	Blue	0.4 - 0.5 µm
	Green	0.5 - 0.6 µm
	Red	0.6 - 0.7 µm
Infrared (IR) (0.7 µm – 1 mm)	Near Infrared (NIR)	0.7 – 1.3 µm
	Middle Infrared (MIR)	1.3 - 3 µm
	Far or Thermal Infrared (TIR)	3 – 30 µm
Microwave (0.1 - 100 cm)	Ku-K-Ka	0.75-2.5 cm
	X	2.5-4 cm
	C	4-8 cm
	S	8-15 cm
	L	15-30 cm

As a result of the motion of atoms and molecules, objects at above absolute zero temperature (0ºK or -273ºC) emit electromagnetic radiation. When objects receive radiation from a natural source i.e., the Sun or artificial sources, a part of it is reflected or absorbed. The radiation absorbed is emitted at some other wavelength. Every object has a unique response to the radiation at different wavelength regions. We identify different objects on a remote sensing image (Fig. 6.3) based on their unique spectral responses. The unique spectral properties of different objects are called ***spectral signatures*** or ***spectral reflectance curve***s. For example, plant leaves appear green to us because they reflect green light the most within the visible spectrum in comparison to blue and red light. Spectral signatures are the combination of reflected, absorbed, and transmitted or emitted EMR by objects at different wavelengths, based on which we can uniquely identify an object.

(a)



(b)

Fig. 6.3: Sample colour images of a part of Gujarat in different band combinations showing various earth surface objects in (a) natural colours as visible to our eyes, created using a combination of images acquired in red, green, and blue wavelengths, and (b) false colours created using a combination of images acquired in near-infrared, red and green wavelengths. The softcopy of the course material has the colour version of the image. (**Source:** a- HereMaps.com; b-Landsat ETM^{+} image)

Understanding spectral signature i.e., the interaction of different objects (such as water, soil, and vegetation) with different wavelengths of the EMR is important to identify features in remote sensing images. We shall now briefly discuss spectral signatures of common Earth's surface materials i.e., vegetation, soil, and water.

Plant leaf appears green to our eyes because it reflects green light higher in comparison to the light of other wavelengths in the visible range. The low reflectance of vegetation in the blue and red regions is due to their absorption by chlorophyll for photosynthesis which makes it easier to distinguish it from non-living features. However, beyond the visible range, vegetation has much higher reflectance in the near-infrared (NIR) region due to the cellular structure of its leaves. The reflectance again lowers in the middle infrared (MIR) region due to absorption by moisture content in plant leaves. A typical spectral signature of green healthy vegetation is shown in Fig. 6.4. Note that the reflectance of vegetation changes with time due to the composition, maturity, and health of the vegetation. The amount of chlorophyll content determines the health of vegetation and plant leaves appear 'greenest' when chlorophyll content is at its maximum.

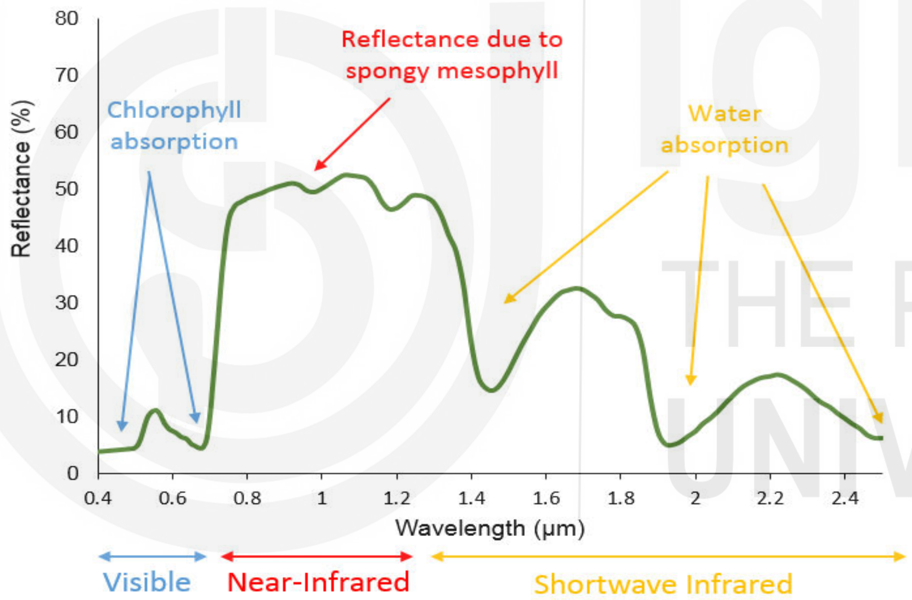


Fig. 6.4: Spectral signature of vegetation showing absorption and reflection in different spectral regions. Chlorophyll content causes higher reflectance in the NIR region and the presence of moisture content causes absorption in some wavelength regions. The softcopy of the course material has the colour version of the image.
(Source: http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson2-1/vegetation.html)

Soil is a mixture of different minerals and matter. Generally, soils appear in shades of brown or black. The appearance is due to the constituents of the soil. Soils appear brown due to relatively higher reflection of green and red energy in comparison to blue. Reflectance gradually increases with the wavelength in the visible and NIR spectral regions (Fig. 6.5) with maximum soil reflectance in NIR wavelength. The spectral signature of soil depends on moisture, iron oxide, texture, organic matter, etc. The moisture content of the soil varies from place to place and hence their appearance too. With the increase in soil moisture content, the radiation at a longer wavelength is absorbed so with the increase in the moisture content overall reflectance of the soil decrease. The reddish appearance of some soil is due to the high iron oxide content in the soil. Soil texture also affects its spectral signature. In comparison to other, sandy soil appears brighter in the image (Fig. 6.5) because radiation in the visible spectrum is almost equally reflected.

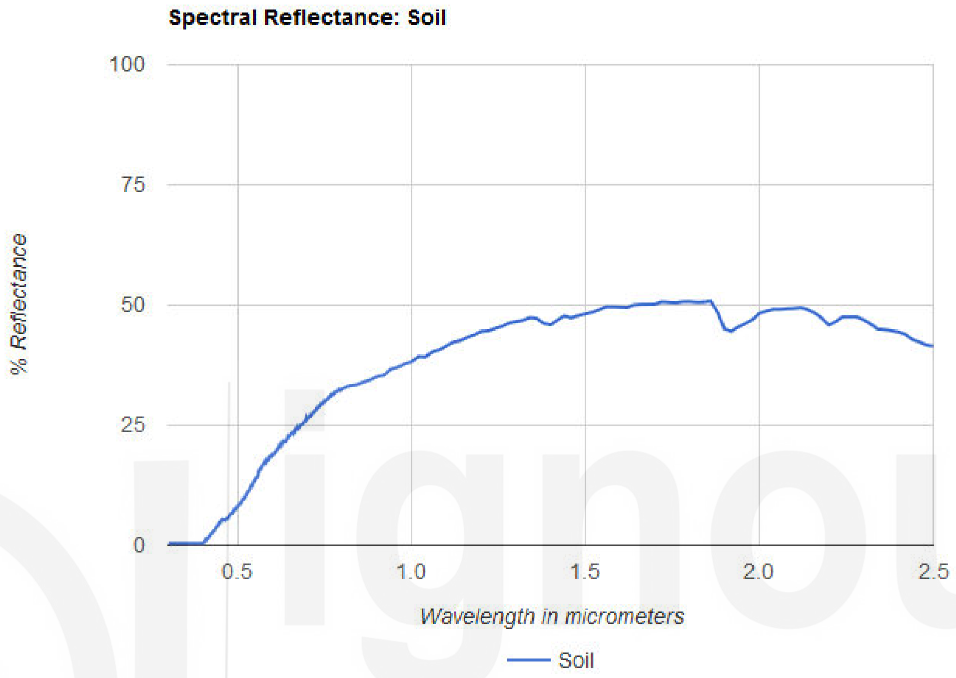


Fig. 6.5: Spectral signature of soil showing high reflectance. The softcopy of the course material has the colour version of the image.
(Source: a - http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson2-1/reflectance.html)

Clear water bodies appear dark whereas shallow and turbid water appears in shades of cyan to our eyes. In the visible region, the incident radiation is largely transmitted in a clear water body. In the NIR region, water absorbs strongly so this wavelength is used for the delineation of land and water boundary. The radiation incident upon the water is largely either absorbed or transmitted. And the reflection could be either from the surface of the water (specular reflectance) or from the bottom (bottom reflectance) or from the water column itself (volume reflectance). Of these three kinds, only the volume reflectance contains information about the water constituents and quality. The depth of the water and the water constituents are the major factors contributing to the spectral signature of water. It is the bottom materials that determine the water body's spectral property in shallow areas. Water constituents like suspended sediments, chlorophyll/phytoplankton, and dissolved organic matter also influence the spectral signature of water. The presence of suspended sediments results in relatively higher reflectivity and scattering and therefore a brighter appearance. As we have read in the spectral signature of vegetation, chlorophyll content in algae makes the water appear greener in the presence of algae due to higher absorption of the blue and red wavelengths. Water having excessive levels of chlorophyll content have a spectral signature somewhat like vegetation. And, presence of organic matter results in lower blue and higher red reflectance.

(iii) Types of Remote Sensing

Remote sensing can be categorised into various types based on different parameters. Different types of remote sensing are listed in Table 6.2.

Table 6.2: Typology of remote sensing and the basis of their classification

(Compiled from Bhatta, 2015)

Basis of classification	Typology	Description
Platform	Ground-based remote sensing	Platforms are on the ground surface including the mobile van-mounted sensors
	Air borne (/suborbital) remote sensing	Platforms are within the atmosphere. e.g., kite, balloon, pigeon, aircraft
	Space borne (Orbital or satellite) remote sensing	Platforms are in the space. e.g., Satellite. Sometimes, International Space Station and the space shuttle is used as a platform
Energy Source	Passive remote sensing	Sun's energy reflected/emitted from the earth's surface features are recorded.
	Active remote sensing	Provides their own energy source for illumination and the radiation reflected/emitted from the target is detected by the sensor
Imaging Media	Photographic imaging	Uses chemical reactions on the surface of light-sensitive photographic film to detect and record energy
	Digital Imaging	The reflected/emitted energy is imaged electronically by electric transducers e.g., charge-coupled devices
Region of the EMS used	Optical remote sensing	Performed within the optical region i.e., 0.3 to 3μm
	Photographic remote sensing	Performed within the photographic region i.e., 0.3 to 0.9 μm
	Thermal remote sensing	Performed within the thermal infrared region i.e., 3μm to 1mm
	Microwave remote sensing	Performed within the microwave region i.e., 1mm to 1m

Number and continuity of bands used	Panchromatic remote sensing	Single band image covering the entire visible spectrum or sometimes a wider region 0.3 to 0.9 µm
	Multispectral remote sensing	Multiple bands using different regions of the EMS such as optical, thermal or microwave regions in a relatively broad wavelength bands
	Hyperspectral remote sensing (narrow band imaging)	An extension of the multispectral remote sensing collected in dozens or hundreds of contiguous and narrow intervals of the spectrum generally, in the optical region

(iv) Advantages and Limitations

After understanding the concept of remote sensing, let us now look at its advantages and limitations. We observe the earth's surface features from a horizontal perspective whereas remote sensing provides us with a bird's eye view of the earth's surface. This overhead perspective gives us a very different view of the earth's surface features. Remote sensing has the following advantages:

- it provides a synoptic view.
- it can acquire data of inaccessible areas.
- it is a cost-effective means of data collection.
- it can provide data in wavelengths beyond the visible range of the human eye.
- it is an unobtrusive means of data collection which does not change the characteristics of the object or phenomenon being observed.
- it provides historical datasets that are useful to understand the characteristics of an object at each point of time in past.

Besides these advantages, there are limitations of remote sensing as listed below:

- it is sometimes found that appropriate data is not easily available, particularly in the tropical regions where cloud cover obstructs the acquisition of remote sensing data because not all sensors can 'see' through the cloud.
- remote sensing systems having their own source of energy such as radars and lasers could be intrusive and may affect the phenomenon under investigation.
- remote sensing systems can become uncalibrated with time resulting in errors in the data being collected.

Check Your Progress 1

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

1. Explain the remote sensing process.

……………………………………………………………………………….

……………………………………………………………………………….

………………………………………………………………………………

………………………………………………………………………………

2. What is the electromagnetic spectrum?

……………………………………………………………………………….

……………………………………………………………………………….

………………………………………………………………………………

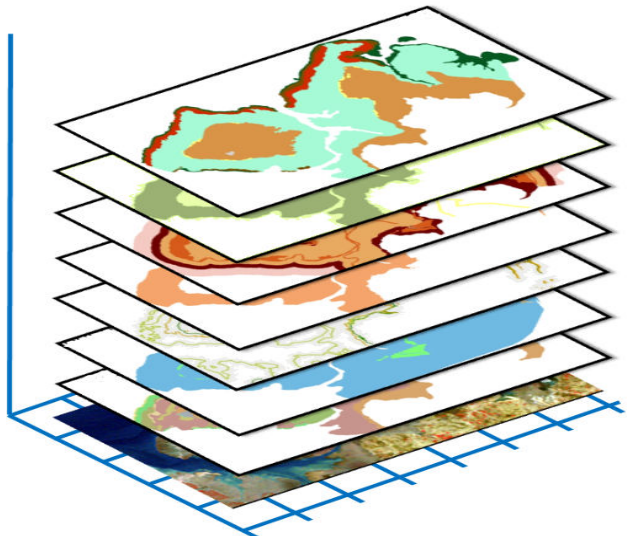
………………………………………………………………………………

In the next subsection, we shall learn about the GIS component.

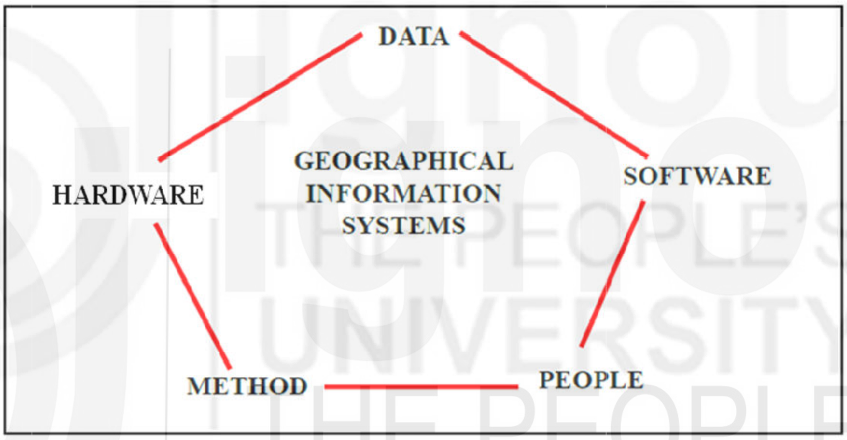
6.3.2 Geographic Information System

Geographic Information System (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. ***GIS*** has been defined based on its different aspects i.e., the tools, the organisation, and the spatial database.

The advantage of GIS is that it allows us to visualize, understand, question, and interpret data in ways that reveal relationships, patterns, and trends in the form of maps, reports, and charts. A GIS helps us answer questions and solve problems by looking at our data in a way that is quickly understood and easily shared. GIS integrates hardware, software, and data for capturing, managing, analysing, and displaying all forms of geographically referenced information (6.6a). GIS helps us locate the areas like land use and the physical infrastructure like telecommunication lines, water, sewer pipes, etc. This is achieved by the process of spatial analysis in which maps/layers of various themes are integrated and analyzed in the GIS platform. The maps have tables of data behind them so when you combine different maps you are combining the data to get the information you need. This also allows questions to be asked from the map or database side. So, it is useful for decision-makers in planning and management purposes.



(a)



(b)

Fig. 6.6: (a) Various thematic layers which are used in the GIS platform for spatial analysis; and (b) the five components of a GIS. The softcopy of the course material has the colour version of the image. (Source: http://egyankosh.ac.in/handle/123456789/39605)

The computer systems - the hardware and software, spatial data, data management and analysis procedures, and the people to operate the GIS are the five components of a GIS (Fig. 6.6b). Data forms the major component of GIS which may come from the interpretation of remote sensing data and GNSS. Computer hardware and software components are required for data entry, data storage, data processing, and analysis and for producing outputs. GIS is of no use if people and the organisation in which they work are not properly oriented towards GIS.

In the next subsection, we shall learn about the GNSS component.

6.3.3 Global Navigation Satellite System

Global Navigation Satellite System refers to a constellation of Earth-orbiting satellites that have been designed to provide autonomous geospatial positioning with global coverage in real-time. These satellites provide signals (containing positioning and timing data) to GNSS receivers on Earth, which convert the signals to determine location.

Applications of GNSS fall into the following broad categories:

- ***Location:*** determining a basic position
- ***Navigation:*** moving from one location to another
- ***Tracking:*** monitoring the movement of people, animals, and goods
- ***Mapping:*** creating a map of an area

The most widely used example of GNSS is the American Global Positioning System commonly known as NAVSTAR GPS. ***Global Positioning System*** (GPS) is a constellation of about 24 satellites which are orbiting the Earth every 12 hours at an altitude of ~20,200 km, as shown in Fig. 6.7. These satellites broadcast signals, which are used to derive precise timing, location, and velocity information. The derived information can then be clubbed with other systems, such as communication devices, computers, and software to perform a variety of functions. With equipment ranging from hand-held receivers to rack-mounted electronics, the signals of GPS can be used by anyone, anytime, anywhere in the world. GPS technology consisting of space, control and user segments enables people to precisely know where they are on the surface of the Earth. Before GPS, positions were established by the complex procedure of relative and absolute measurements of directions and distances.

Other examples of GNSS include Russia's GLONASS, Europe's Galileo, and China's BeiDou. Some other countries e.g., South Korea are planning for their own Korea Positioning System (KPS) in future. India has developed its own Indian Regional Satellite Navigation System (IRNSS) called NavIC (Navigation Indian Constellation). NavIC consists of 7 satellites with all components (i.e., space, ground, and user segments) built in India. It has a service area of ~1500 km around India. It has also developed a Satellite-based Augmentation System i.e., GAGAN to provide a navigation system for all phases of flight over the Indian airspace and in the adjoining area.

Fig. 6.7: Showing diagrammatic representation of the constellation of Global Navigation Satellites System orbiting the Earth. The softcopy of the course material has the colour version of the image.

6.4 APPLICATIONS OF GEOSPATIAL TECHNOLOGIES

In this subsection, we shall learn about the role of geospatial technologies in environmental studies.

Note that remote sensing and GNSS are the data acquisition technologies whereas GIS is the platform in which these data can be analyzed. The remote sensing data and GNSS data are used as input in GIS. With recent developments in remote sensing technologies, a large volume of geospatial data is widely available at an affordable cost. Geospatial technologies have revolutionized the way we plan and manage our natural and built environments. Geospatial tools are now used by planners and decision-makers to address a wide array of problems such as environmental mapping and monitoring, site suitability and feasibility studies, and environmental conservation. Since it gives decision-makers valuable insights into current needs and future projections, it helps make informed decisions.

Geospatial technologies are increasingly being employed to map and maintain data on land cover and land use, and hence to plan and monitor urban expansion, public transport services, finding suitable sites for different purposes including waste management, etc. To understand the role of geospatial tools in environmental studies, we need to first understand the requirements for such studies and then see how the use of geospatial technologies could be advantageous for the derivation of the information required. Broadly, it can be useful for the following:

- **Mapping** - Identification and delineation of various features.
- **Monitoring** - Periodic data analyses of the features identified.
- **Management** - Suggest measures based on the nature of changes observed.

To understand the dynamics of land, ocean and atmospheric parameters over time and space their continuous monitoring is required at different temporal and spatial scales. However, to be able to monitor these parameters, we should first be able to identify them. These earth surface (and subsurface), oceanic and atmospheric features can be identified based on their characteristic spectral signatures in different wavelength regions. Monitoring is particularly useful in functions like waste management.
We will now have an overview of the potential of geospatial technologies in different application areas:

6.4.1 Natural Resources Management

A resource is a wealth. And we know that conservation and proper management of natural resources are essential for the betterment of future generations. Because of greater areal extent, voluminous data, and uneven topography under changing climatic conditions, monitoring and managing the available resources by people is not an easy task. The following kinds of information and studies are possible using geoinformatics technology in natural resources management:

- Mapping of different kinds of soils, and their quality including degraded land.
- Identification of mineral resources and potential mineral zones.
- Mapping of water resources and potential groundwater zones.
- Mapping and monitoring of water quality.
- Identification of potential groundwater zones.

Some of the commonly observed applications of geoinformatics for land resources are land use/land cover mapping, geomorphology, land degradation, etc. Land use (LU) is the usage of land by human beings, such as houses, roads, manufacturing plants, etc. Land cover (LC) refers to a natural landscape as a surface component of a portion of the Earth's surface, such as forests, lakes, beaches, rivers, etc. While land use designates economic activity and hence is abstract, land cover designates the surface features, and is therefore directly observable. Geospatial technologies are useful in monitoring land use/land cover (LULC) and its changes over time over a variety of spatial scales.

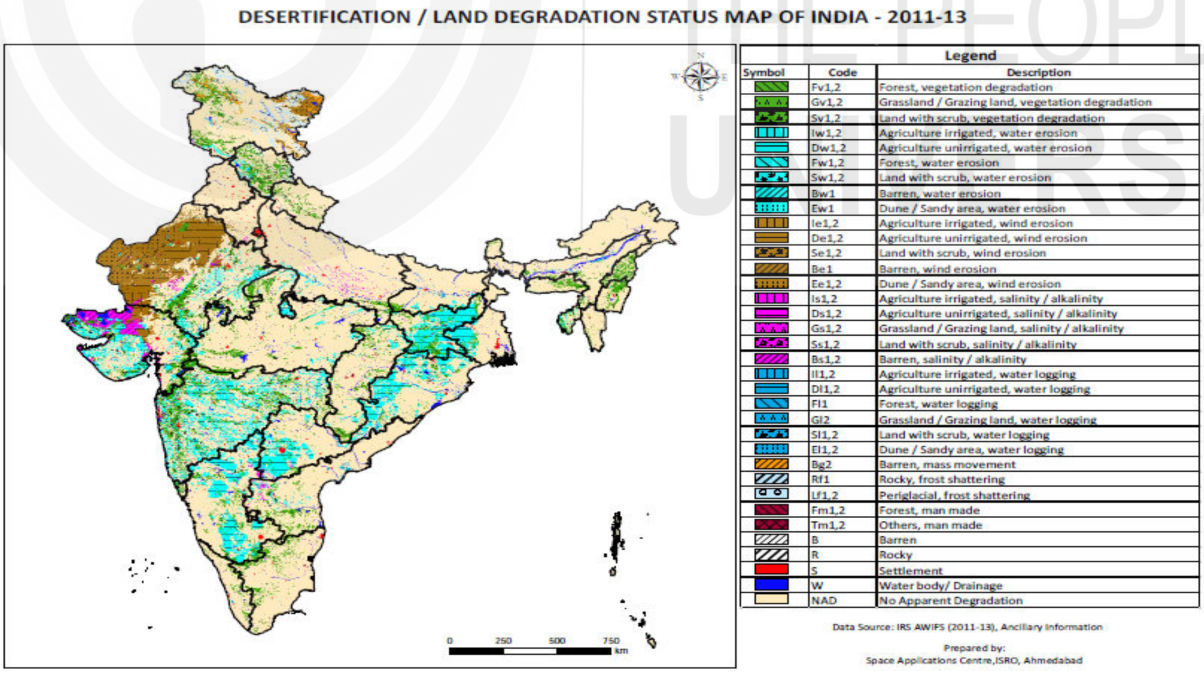


Fig. 6.8 Land degradation map of India prepared primarily using satellite remote sensing and field data.
The softcopy of the course material has the colour version of the image.
(Source: SAC Report)

6.4.2 Mapping and Monitoring Urban Environment

It is estimated that with the increasing population and their migration to urban regions driven by the factors such as employment and better living/facilities, a massive transformation would be required to accommodate an additional 10 million urban dwellers per year. It includes providing them with planned areas with adequate housing and urban services such as water supply, sewerage, drainage, solid waste management, urban infrastructure, etc. It is also important at the same time to ensure that such urbanisation is environmentally sustainable.

Satellite remote sensing and aerial photography techniques provide accurate, orderly, and reliable information repetitively for the planning and management of urban utility services. Some of the common applications are:

- Mapping and monitoring of LULC.
- Mapping and modelling urban sprawl.
- Urban heat island mapping.
- Mapping and monitoring Urban Infrastructure/Facilities.

6.4.3 Terrestrial and Aquatic Ecosystems Related Studies

Geoinformatics technologies are used for the following kinds of studies in ecosystem-related applications:

- Estimation of terrestrial and aquatic habitats, vegetation cover and historical changes.
- Mapping and monitoring of wetlands.
- Mapping and monitoring of inland and marine national parks.
- Mapping and monitoring shorelines and identification of vulnerable zones.
- Identification of different kinds of plant communities.
- Biodiversity estimation, species richness, distribution, and migration of wildlife.
- Mapping and monitoring of desertification.
- Identification and mapping of habitats and habitat suitability, etc.
- Identification and mapping of water bodies.
- Monitoring sediment discharge and water quality.
- Mapping bathymetry.
- Inland and marine pollution including oil spills.
- Mapping and monitoring forest fires and other natural or manmade disasters, and impact assessment.

- Identification of potential fishing zones.
- Identification of flood-prone areas and flood inundated areas, modelling, etc.
- Mapping and monitoring of water surface temperature, etc.
- Biodiversity characterisation.
- Preparation of ecological zone maps.
- Landscape characterisation to establish disturbance gradient.
- Prioritisation of bio-rich areas for conservation.

6.4.4 Climate Related Studies

In this subsection, we shall learn about the applications related to climate.
Data from satellite remote sensing is widely used for developing prevention, mitigation, and adaptation measures to cope with the impact of climate change. Some of the breakthroughs to which the use of satellite remote sensing data has particularly contributed are (Yang et al 2013):

- Understanding of warming characteristics of Earth's atmosphere and ocean surfaces
- Trends of acceleration in Arctic Sea ice loss, and
- Role of distant aerosol sources in the genesis of the hurricane.

Some of the climate-related phenomena and relevant remote sensing-derived climate variables are listed in Table 6.3.

Table 6.3: List of Essential Climate Variables (ECVs) that can be retrieved from remote sensing data

Atmospheric Variables	Terrestrial Variables	Oceanic Variables
Precipitation	Soil moisture	Sea level
Radiation budget (Both Surface and Upper-air)	Albedo	Sea surface temperature
Upper air temperature	Photosynthetically Active Radiation (fAPAR)	Ocean salinity
Wind speed and direction (Surface and upper-air)	Land cover	Sea state (wave height, direction, wavelength, and time)
Water vapour	Leaf Area Index (LAI)	Ocean colour
Cloud properties	Snow cover, Glaciers, and ice caps	Sea ice extent
Carbon dioxide	Biomass	
Ozone	Fire disturbance	
Aerosol properties	Lakes	

(**Source:** Adapted from Yang et al 2013 and originally compiled from GCOS, 2010)

6.5 Geospatial Data and their Sources

In this subsection, we will discuss different sources of geospatial data which you can acquire either at cost or for free. We will also briefly discuss how to find data and the kind of data available at different sources.

With the evolution of various sensors such as LANDSAT from the USA, IRS i.e., Indian Remote Sensing satellite series and Indian geostationary Satellite series i.e., INSAT from India, ENVISAT from European Space Agency, SCHYMACHY from the USA, TRMM and Terra/Aqua-MODIS from USA/NASA etc., data have been acquired and used as the primary data source for Earth's resources and atmospheric monitoring since 1972.
In India, users can acquire remote sensing data from the National Data Centre (NDC) of the National Remote Sensing Centre (NRSC), of ISRO, which is located in Hyderabad.

A variety of remote sensing data and several derived products from many satellites are available at the following portals:

- http://glcf.umiacs.umd.edu/data
- https://earthexplorer.usgs.gov/
- https://neo.sci.gsfc.nasa.gov/
- https://glovis.usgs.gov
- https://scihub.copernicus.eu/dhus/
- https://lpdaac.usgs.gov/news/?&page=13
- https://search.earthdata.nasa.gov/
- www.vito-eodata.be/
- www.class.ngdc.noaa.gov/saa/products/welcome
- https://coast.noaa.gov/digitalcoast/
- www.terrapop.org/
- https://earthdata.nasa.gov/earth-observation-data/near-real-time/rapid-response
- www.vito-eodata.be/PDF/portal/Application.html#Home
- www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html
- https://eospso.gsfc.nasa.gov/content/nasa-earth-science-data
- https://clim-engine-development.appspot.com/fewsNet.

As regards land use and land cover mapping, appropriate data is essential. This appropriateness is in the following terms:

- the details which are possible to extract from the given remote sensing data
- the time interval between the data acquisition for the same area by the same sensor
- timing of data acquisition with respect to the phenomena being observed.

At the national and state level, you do not need to use data which provide very detailed information because this would not be cost-effective. So, data should be selected keeping in view the scale at which the map is to be prepared. Another important issue is the consideration of data of appropriate timing, such as for mapping kharif crops, you need to choose data which has been acquired between June to September. You should also note that the cropping pattern and the harvesting season vary from region to region within India. Another critical point to remember is that if you are going to compare land use/land cover patterns of an area for a different date then it is best to use the data from the same sensor so that there is uniformity in the information derived.

CHECK YOUR PROGRESS 2

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

1. Write a short note on Geographical Information System.

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

2. Write a short note on Global Navigation Satellite System.

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

3. Write a short note on Geospatial Data.

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

………………………………………………………………………………….

6.6 LET US SUM UP

Let us summarise what we have learnt in this unit.

- Remote sensing is the process of acquiring information about an object or phenomenon from a distance and the Sun is the main source of energy for remote sensing.
- Plants absorb blue and red light within the visible region due to the presence of chlorophyll content, reflect very high in the NIR region due to internal cell structure and are sensitive to water content in the SWIR region.
- The spectral signature of soil depends on moisture, iron oxide, texture and organic matter contents.
- The depth of the water and the water constituents are the major factors contributing to the spectral signature of water.

6.7 KEY WORDS

Band : It refers to the specific range of the electromagnetic spectrum to which the sensor is sensitive.

Electromagnetic Energy : Energy with both electric and magnetic components. The measurement of reflected and emitted electromagnetic energy is the basis of remote sensing.

6.8 SUGGESTED FURTHER READING/ REFERENCES

Bhatta, B. (2015) Remote Sensing and GIS, Second Ed., Oxford University Press, New Delhi, 716p.

Campbell, J.B. (2002) Introduction to Remote Sensing, Third Ed., Guilford Publications, New York, 620p.

Colwell, R.N. (Ed.) (1983) Manual of Remote Sensing, Second Ed., American Society of Photogrammetry, Falls Church, Va., p.1.

http://glovis.usgs.gov

http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson2-1/reflectance.html

https://imagine.gsfc.nasa.gov/Images/science/EM_spectrum_compare_level1_lg.jpg

https://zoom.earth/#view=22.262531,72.9039,13z/layers=archive31

Lillesand, T.M. and R.W. Kiefer (1979) Remote Sensing and Image Interpretation, John Wiley & Sons, New York.

Unit-3 *Spectral Signature* in Block-1 *Introduction to Remote Sensing* of the Course *MGY-002 Introduction to Remote Sensing and Image Interpretation* [http://egyankosh.ac.in/handle/123456789/39605]

Unit-4 *Platforms and Sensors* in Block-2 *Sensors and Space Programmes* of the Course *MGY-002 Introduction to Remote Sensing and Image Interpretation* [http://egyankosh.ac.in/handle/123456789/39606]

www.esa.int/images/02-spektrum.gif

http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson2-1/reflectance.html (Spectral Reflectance)

https://crustal.usgs.gov/speclab/ (USGS Spectral Library)

https://eos.com/ndvi/ (For NDVI)

https://landsat.usgs.gov/spectral-characteristics-viewer (Spectral Characteristics Viewer)

www.asprs.org/wp-content/uploads/pers/1986journal/sep/1986_sep_1507-1511.pdf (On the definition of remote sensing)

(Websites last accessed on 12 September 2022)

6.9 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress 1

1. The process of remote sensing involves several steps. The first and very important requirement for remote sensing is an energy source which provides electromagnetic energy to the Earth. It may be either from natural (e.g., solar radiation) or artificial (e.g., RADAR) sources. For remote sensing, Sun's radiations are commonly used as a source of energy. The steps are as follows:

   - Source of Energy
   - Interaction of energy with the atmosphere
   - Interaction of energy with Earth's surface features
   - Recording of the energy by the sensor
   - Transmission, reception, and processing of the recorded signals
   - Interpretation and analysis of the data.

2. Electromagnetic radiation extends over a wide range of wavelengths. The entire range of radiant energies is called electromagnetic spectrum (EMS). EMR ranges from very high energy radiation such as gamma rays and X rays through ultraviolet light, visible light, infrared radiation and microwaves to radio waves.

Check Your Progress 2

1. Please refer to section 6.3.2
2. Please refer to section 6.3.3
3. Please refer to section 6.5