

Example 2.14: Transforming the ratios into corresponding numbers, prepare a complete table for the following information. Give a suitable title to the table.

In the year 2000 the total strength of students of three colleges X, Y, and Z in a city were in the ratio 4 : 2 : 5. The strength of college Y was 2000. The proportion of girls and boys in all colleges was in the ratio 2 : 3. The facultywise distribution of boys and girls in the faculties of Arts, Science, and Commerce was in the ratio 1 : 2 : 2 in all the three colleges.

Solution: The data of the problem is summarized in Table 2.33.

Table 2.33 Distribution of Students According to Faculty and Colleges in the Year 2000

Colleges	Faculty									Total (1) + (2) + (3)
	Arts			Science			Commerce			
	Boys	Girls	Total (1)	Boys	Girls	Total (2)	Boys	Girls	Total (3)	
X	480	320	800	960	640	1600	960	640	1600	4000
Y	240	160	400	480	320	800	480	320	800	2000
Z	600	400	1000	1200	800	2000	1200	800	2000	5000
Total	1320	880	2200	2640	1760	4400	2640	1760	4400	11,000

Example 2.15: Represent the following information in a suitable tabular form with proper rulings and headings:

The annual report of a Public Library reveals the following information regarding the reading habits of its members.

Out of the total of 3718 books issued to the members in the month of June, 2100 were fictions. There were 467 members of the library during the period and they were classified into five classes—A, B, C, D, and E. The number of members belonging to the first four classes were respectively 15, 176, 98, and 129, and the number of fictions issued to them were 103, 1187, 647, and 58 respectively. The number of books, other than text books and fictions, issued to these four classes of members were respectively 4, 390, 217, and 341. Text books were issued only to members belonging to classes C, D, and E, and the number of text books issued to them were respectively 8, 317, and 160.

During the same period, 1246 periodicals were issued. These include 396 technical journals of which 36 were issued to member of class B, 45 to class D, and 315 to class E.

To members of classes B, C, D, and E the number of other journals issued were 419, 26, 231, and 99, respectively.

The report, however, showed an increase of 4.1 per cent in the number of books issued over last month, though there was a corresponding decrease of 6.1 per cent in the number of periodicals and journals issued to members.

Solution: The data of the problem is summarized in Table 2.35.

Table 2.35 Reading Habits of the Members of Public Library

Type of Book Issued	Class of Members					Total for the Month	
	A	B	C	D	E	June	May
Books							
Fiction	103	1187	647	58	105	2100	2018
Textbooks	—	—	8	317	160	485	466
Others	4	390	217	341	181	1133	1089
Total	107	1577	872	716	446	3718	3573
Periodicals and Journals							
Technical journals	—	36	—	45	315	396	420
Others	75	419	26	231	99	850	902
Total	75	455	26	276	414	1246	1322

Note: The figures for the month of May were calculated on the basis of percentage changes for each type of reading material given in the text.

Conceptual Questions 2B

13. What is a statistical table? Explain clearly the essentials of a good table.
14. (a) What are the different components of a table?
(b) What are the chief functions of tabulation? What precautions would you take in tabulating statistical data?
(c) What are the characteristics of a good table?
15. Explain the role of tabulation in presenting business data, and discuss briefly the different methods of presentation.
16. Explain the terms 'classification' and 'tabulation'. Point out their importance in a statistical investigation. What precautions would you take in tabulating statistical data?
17. In classification and tabulation, common sense in the chief requisite and experience the chief teacher. Comment.
18. What are the requisites of a good table? State the rules that serve as a guide in tabulating statistical data.
19. Distinguish between classification and tabulation. Mention the requisites of a good statistical table.
20. Explain how you would tabulate statistics of deaths from sexual diseases in different states of India for a period of five years.
21. Explain the purpose of tabular presentation of statistical data. Draft a form of tabulation to show the distribution of population according to (i) community by age, (ii) literacy, (iii) sex, and (iv) marital status.

Self-Practice Problems 2B

- 2.14 Draw a blank table to show the number of candidates sex-wise appearing in the pre-university, First year, Second year, and Third year examinations of a university in the faculties of Arts, Science, and Commerce in a certain year.
- 2.15 Let the national income of a country for the years 2000–01 and 2001–02 at current prices be 80,650, 90,010, and 90,530 crore of rupees respectively, and per capita income for these years be 1050, 1056, and 1067 rupees. The corresponding figures of national income and per capita income at 1999–2000 prices for the above years were 80,650, 80,820, and 80,850 crore of rupees and 1050, 1051 and 1048 respectively. Present this data in a table.
- 2.16 Present the following information in a suitable form supplying the figure not directly given. In 2004, out of a total of 4000 workers in a factory, 3300 were members of a trade union. The number of women workers employed was 500 out of which 400 did not belong to any union.
In 2003, the number of workers in the union was 3450 of which 3200 were men. The number of non-union workers was 760 of which 330 were women.
- 2.17 Of the 1125 students studying in a college during a year, 720 were SC/ST, 628 were boys, and 440 were science students; the number of SC/ST boys was 392, that of boys studying science 205, and that of SC/ST students studying science 262; finally the number of science students among the SC/ST boys was 148. Enter these frequencies in a three-way table and complete the table by obtaining the frequencies of the remaining cells.
- 2.18 A survey of 370 students from the Commerce Faculty and 130 students from the Science Faculty revealed that 180 students were studying for only C.A. Examinations, 140 for only Costing Examinations, and 80 for both C.A. and Costing Examinations. The rest had opted for part-time Management Courses. Of those studying for Costing only, 13 were girls and 90 boys belonged to the Commerce Faculty. Out of the 80 studying for both C.A. and Costing, 72 were from the Commerce Faculty amongst whom 70 were boys. Amongst those who opted for part-time Management Courses, 50 boys were from the Science Faculty and 30 boys and 10 girls from the Commerce Faculty. In all there were 110 boys in the Science Faculty.
Present this information in a tabular form. Find the number of students from the Science Faculty studying for part-time Management Courses.
- 2.19 An Aluminium Company is in possession of certain scrap materials with known chemical composition. Scrap 1 contains 65 per cent aluminium, 20 per cent iron, 2 per cent copper, 2 per cent manganese, 3 per cent magnesium and 8 per cent silicon. The aluminium content of scrap 2, scrap 3, and scrap 4 are respectively 70 per cent, 80 per cent and 75 per cent. Scrap 2 contains 15 per cent iron, 3 per cent copper, 2 per cent manganese, 4 per cent magnesium and the rest silicon. Scrap 3 contains 5 per cent iron. The iron content of scrap 4 is the same as that of scrap 3, scrap 4 contains twice as much percentage of copper as scrap 3. scrap 3 contains 1 per cent copper. Scrap 3 contains manganese which is 3 times as much copper as it contains. The percentage of magnesium and silicon in scrap 3 are respectively 3 per cent and 8 per cent. The magnesium and silicon contents of scrap 4 are respectively 2 times and 3 times its manganese contents. The company also purchases some aluminium and silicon as needed. The aluminium purchased contains 96 per cent pure aluminium, 2 per cent iron, 1 per cent copper and 1 per cent silicon respectively, whereas the purchased silicon contains 98 per cent silicon and 2 per cent iron respectively. Present the above data in a table.
- 2.20 Present the following data in a suitable tabular form with appropriate headings:

A pilot survey carried out a few years before yielded the following estimates of livestock numbers and milk production in three regions, namely, Punjab Plains (PP), Punjab Hills (PH), and Eastern U.P. (EU) (only the estimates for the rural sector are quoted here.) The total number of cows was 4396, 2098 and 15,170 thousand, respectively in three regions, namely, PP, PH and EU, and the corresponding number for buffaloes were 4,092, 765 and 5,788 thousand. The percentages of animals producing milk were 47, 40 and 37 for cows in PP, PH, and EU, respectively, the corresponding figures for buffaloes being 58, 49 and 47. The average daily milk yield per animal was 2.52, 0.51, and 0.68 kg for cows in PP, PH, and EU respectively; and for buffaloes the yield figures were 4.10, 2.35 and 1.86 kg respectively.

- 2.21** Prepare a blank tabular layout with appropriate headings for presenting the estimates of the number of unemployed persons obtained from a sample survey covering three states namely, Bihar, Orissa, and West Bengal. The estimates should be presented separately for the three states, for rural and urban areas of each State, and also separately for persons in different levels of general education (illiterate, literate below primary, primary, secondary, graduate and above). The table should show the number of unemployed persons in each region and education level as well as the percentage of such persons to the corresponding total population. It should also present relevant sub-totals and totals.
- 2.22** A state was divided into three areas: administrative district, urban district, and rural district. A survey of housing conditions was carried out and the following information was gathered:
- There were 67,71,000 buildings of which 17,61,000 were in rural district. Of the buildings in urban district 40,64,000 were inhabited and 45,000 were under construction. In the administrative district 40,000 buildings were uninhabited and 5000 were under construction of the total of 6,16,000. The total buildings in the city that are under construction are 62,000 and those uninhabited are 4,49,000. Tabulate this information.
- 2.23** Draw up a blank table to show the number of employees in a large commercial firm, classified according to (i) Sex: male and female; (ii) three age groups : below 30, 30 and above but below 45, 45 and above; and (iii) four income-groups: below Rs 400, Rs 400–750, Rs 750–1000, and above Rs 1000.
- 2.24** Transform the ratios into corresponding numbers to prepare a complete table for the following information. Give a suitable title to the table.

In the year 1997, the total strength of students of three colleges A, B, and C in a city was in the ratio

3 : 1 : 4. The strength of college B was 800. The proportion of girls and boys in all colleges was in the ratio 1 : 3. The faculty-wise distribution of girls and boys in the faculties of Arts, Science, and Commerce was in the ratio 2 : 1 : 2 in all the three colleges.

- 2.25** The 'Financial Highlights' of a public limited company in recent years were as follows:

In the year ended 31 March 1998 the turnover of the company, including other income, was Rs 157 million. The profit of the company in the same year before tax, investment allowance, reserve, and prior year's adjustment was Rs 19 million, and the profit after tax, investment allowance, reserve, and prior year's adjustment was Rs 8 million. The dividend declared by the company in the same year was 20 per cent. The turnover, including other income, for the years ended 31 March 1999, 2000, and 2001 were Rs 169, 191, and 197 million respectively. For the year ended 31 March 1999 the profit before tax, investment allowance, reserve, and prior year's adjustment was Rs 192 million and the profit after tax, and so on Rs. 7.5 million, while the dividend declared for the same year was 17 per cent. For the year ended 31 March 2000, 2001, and 2002 the profits before tax, investment allowance, reserve, and prior year's adjustment were Rs 21, 12, 13 million respectively, while the profits after tax, and so on, of the above three years were Rs 9.5, 4, and 9 million respectively. The turnover, including other income, for the year ended 31 March 2002 was Rs 243 million. The dividend declared for the year ended 31 March 2000–02 was 17 per cent, 10 per cent and 20 per cent respectively. Present the above data in a table.

- 2.26** Present the following information in suitable form:

In 1994 out of a total of 1950 workers of a factory 1400 were members of a trade union.

The number of women employed was 400 of which 275 did not belong to a trade union. In 1999, the number of union workers increased to 1780 of which 1490 were men. On the other hand, the number of non-union workers fell to 408 of which 280 were men.

In the year 2004, there were 2000 employees who belonged to a trade union and 250 did not belong to a trade union. Of all the employees in 2000, 500 were women of whom only 208 did not belong to a trade union.

- 2.27** In a trip organized by a college, there were 50 persons, each of whom paid Rs 2500 on an average. There were 40 students each of whom paid Rs 2700. Members of the teaching staff were charged at a higher rate. The number of servants was 5 and they were not charged any amount. The number of ladies was 20 per cent of the total and one was a lady teacher. Tabulate the above information.

Hints and Answers

2.14 Distribution of candidates appearing in various university examinations

Faculty	Boys					Girls				
	Pre-Univ.	First year	Second year	Third year	Total	Pre-Univ.	First year	Second year	Third year	Total
Arts										
Science										
Commerce										
Total										

2.15 National income and per capita income of the country

For the year 1999-2000 to 2001-2002

Year	National Income		Per Capita Income	
	At Current Prices (Rs in crore)	At 1999-2000 Prices (Rs in crore)	At Current Prices	At 1999-2000 Prices
1999-2000	80,650	80,650	1050	1050
2000-2001	90,010	80,820	1056	1051
2001-2002	90,530	80,850	1067	1048

2.16 Members of union by sex

Year	2003			2004		
	Males	Females	Total	Males	Females	Total
Member	3300	250	3450	3200	100	3300
Non-member	430	330	760	300	400	700
Total	3630	580	4210	3500	500	4000

2.17 Distribution of College Students by Caste and Faculty

Faculty	Boys			Girls		
	SC/ST	Non-SC/ST	Total	SC/ST	Non-SC/ST	Total
Science	148	57	205	114	121	235
Arts	244	179	423	214	48	262
Total	392	236	628	328	169	497

2.18 Distribution of students according to Faculty and Professional Courses

Faculty Courses	Commerce			Science			Total		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
Part-time Management	30	10	40	50	10	60	80	20	100
CA only	150	8	158	16	6	22	166	14	180
Costing only	90	10	100	37	3	40	127	13	140
CA and Costing	70	2	72	7	1	8	77	3	80
Total	340	30	370	110	20	130	450	50	500

2.19 The chemical composition of Scraps and Purchased Minerals

Materials	Chemical Composition (in percentage)					
	Aluminium	Iron	Copper	Manganese	Magnesium	Silicon
Scrap 1	65	20	2	2	3	8
Scrap 2	70	15	3	2	4	6
Scrap 3	80	5	1	3	3	8
Scrap 4	75	5	2	3	6	9
Aluminium	96	2	1	—	—	1
Silicon	—	2	—	—	—	98

2.25 Financial highlights of the Public Ltd.. Co.

Year Ended 31 March	Turnover Including Other Income (in Million of Rs)	Profit Before Tax, Investment Allowance Reserve and Prior Year Adjustment (in Million of Rs)	Profit After Tax Investment Allowance Reserve and Prior Year Adjustment (in Million of Rs)	Per cent
1998	157	19	8	20
1999	169	18	7.5	17
2000	191	21	9.5	17
2001	197	12	4	10
2002	243	13	9	20

2.26 Trade-union membership

Category	1994			1999			2004		
	Member	Non-member	Total	Member	Non-member	Total	Member	Non-member	Total
Men	1,275	275	1550	1490	280	1770	1708	42	1750
Women	125	275	400	290	128	418	292	208	500
Total	1400	550	1950	1780	408	2188	2000	250	2250

2.27 Type of Participants	Sex			Contribution Per Member (Rs)	Total Contribution (Rs)
	Males	Females	Total		
Students	31	9	40	2700	1,08,000
Teaching staff	4	1	5	3400	17,000
Servant	5	—	5	—	—
Total	40	10	50	—	1,25,000

Notes : 1. Total contribution = Average contribution × Number of persons in the group
= 2500 × 50 = Rs 1,25,000

2. Per head contribution of teaching staff = $\frac{\text{Total contribution} - \text{Contribution of students}}{\text{Number of teaching staff}}$
= $\frac{1,25,000 - (40 \times 2700)}{5} = 3400$

2.5 GRAPHICAL PRESENTATION OF DATA

It has already been discussed that one of the important functions of statistics is to present complex and unorganized (raw) data in such a manner that they would easily be understandable. According to King, 'One of the chief aims of statistical science is to render the meaning of masses of figures clear and comprehensible at a glance.' This is often best accomplished by presenting the data in a pictorial (or graphical) form.

The graphical (diagrammatical) presentation of data has many advantages. The following persons rightly observed that

- With but few exceptions memory depends upon the faculty of our brains possess of forming visual images and it is this power of forming visual images which lies at the root of the utility of diagrammatic presentation. —R. L. A. Holmes
- Cold figures are uninspiring to most people. Diagrams help us to see the pattern and shape of any complex situation. Just as a map gives us a bird's eye-view of the wide stretch of a country, so diagrams help us to visualise the whole meaning of the numerical complex at a single glance. —M. J. Moroney

Figures are not always interesting, and as their size and number increases they become uninteresting and confusing to such an extent that nobody would like to study them. The work of a statistician is to understand the data himself, and to put them in such a way that their importance may be known to every one. According to Calvin F. Schmid, 'Charts and graphs represent an extremely useful and flexible medium for explaining, interpreting and analysing numerical facts largely by means of points, lines, areas and other geometric forms and symbols. They make possible the presentation of quantitative data in a simple, clear, and effective manner and facilitate comparison of values, trends and relationships.'

2.5.1 Functions of a Graph

Graphic presentation of frequency distributions facilitate easy understanding of data presentation and interpretation. The shape of the graph offers easy answers to several questions. The same information can also be obtained from tabular presentation of a frequency distribution, but the same is not as effective in highlighting the essential characteristics as explicitly as is possible in the case of graphic presentation.

The shape of the graph gives an exact idea of the variations of the distribution trends. Graphic presentation, therefore, serves as an easy technique for quick and effective comparison between two or more frequency distributions. When the graph of one frequency distribution is superimposed on the other, the points of contrast regarding the type of distribution and the pattern of variation become quite obvious. All these advantages necessitate a clear understanding of the various forms of graphic representation of a frequency distribution.

2.5.2 Advantages and Limitations of Diagrams (Graphs)

According to P. Maslov, 'Diagrams are drawn for two purposes (i) to permit the investigator to graph the essence of the phenomenon he is observing, and (ii) to permit others to see the results at a glance, i.e. for the purpose of popularisation.'

Advantages Few of the advantages and usefulness of diagrams are as follows:

- (i) *Diagrams give an attractive and elegant presentation:* Diagrams have greater attraction and effective impression. People, in general, avoid figures, but are always impressed by diagrams. Since people see pictures carefully, their effect on the mind is more stable. Thus, diagrams give delight to the eye and add to the spark of interest.
- (ii) *Diagrams leave good visual impact:* Diagrams have the merit of rendering any idea readily. The impression created by a diagram is likely to last longer in the mind of people than the effect created by figures. Thus diagrams have greater memorizing value than figures.
- (iii) *Diagrams facilitate comparison:* With the help of diagrams, comparisons of groups and series of figures can be made easily. While comparing absolute figures the significance is not clear but when these are presented by diagrams, the comparison is easy. The technique of diagrammatic representation should not be used when comparison is either not possible or is not necessary.
- (iv) *Diagrams save time:* Diagrams present the set of data in such a way that their significance is known without loss of much time. Moreover, diagrams save time and effort which are otherwise needed in drawing inferences from a set of figures.

- (v) *Diagrams simplify complexity and depict the characteristics of the data:* Diagrams, besides being attractive and interesting, also highlight the characteristics of the data. Large data can easily be represented by diagrams and thus, without straining one's mind, the basic features of the data can be understood and inferences can be drawn in a very short time.

Limitations We often find tabular and graphical presentations of data in annual reports, newspapers, magazines, bulletins, and so on. But, inspite their usefulness, diagrams can also be misused. A few limitations of these as a tool for statistical analysis are as under:

- (i) They provide only an approximate picture of the data.
- (ii) They cannot be used as alternative to tabulation of data.
- (iii) They can be used only for comparative study.
- (iv) They are capable of representing only homogeneous and comparable data.

2.5.3 General Rules for Drawing Diagrams

To draw useful inferences from graphical presentation of data, it is important to understand how they are prepared and how they should be interpreted. When we say that 'one picture is worth a thousand words', it neither proves (nor disproves) a particular fact, nor is it suitable for further analysis of data. However, if diagrams are properly drawn, they highlight the different characteristics of data. The following general guidelines are taken into consideration while preparing diagrams:

Title: Each diagram should have a suitable title. It may be given either at the top of the diagram or below it. The title must convey the main theme which the diagram intends to portray.

Size: The size and portion of each component of a diagram should be such that all the relevant characteristics of the data are properly displayed and can be easily understood.

Proportion of length and breadth: An appropriate proportion between the length and breadth of the diagram should be maintained. As such there are no fixed rules about the ratio of length to width. However, a ratio of $\sqrt{2} : 1$ or 1.414 (long side) : 1 (short side) suggested by Lutz in his book *Graphic Presentation* may be adopted as a general rule.

Proper scale: There are again no fixed rules for selection of scale. The diagram should neither be too small nor too large. The scale for the diagram should be decided after taking into consideration the magnitude of data and the size of the paper on which it is to be drawn. The scale showing the values as far as possible, should be in even numbers or in multiples of 5, 10, 20, and so on. The scale should specify the size of the unit and the nature of data it represents, for example, 'millions of tonnes', in Rs thousand, and the like. The scale adopted should be indicated on both vertical and horizontal axes if different scales are used. Otherwise can be indicated at some suitable place on the graph paper.

Footnotes and source note: To clarify or elucidate any points which need further explanation but cannot be shown in the graph, footnotes are given at the bottom of the diagrams.

Index: A brief index explaining the different types of lines, shades, designs, or colours used in the construction of the diagram should be given to understand its contents.

Simplicity: Diagrams should be prepared in such a way that they can be understood easily. To keep it simple, too much information should not be loaded in a single diagram as it may create confusion. Thus if the data are large, then it is advisable to prepare more than one diagram, each depicting some identified characteristic of the same data.

2.6 TYPES OF DIAGRAMS

There are a variety of diagrams used to represent statistical data. Different types of diagrams, used to describe sets of data, are divided into the following categories:

- **Dimensional diagrams**
 - (i) One dimensional diagrams such as histograms, frequency polygons, and pie chart.
 - (ii) Two-dimensional diagrams such as rectangles, squares, or circles.
 - (iii) Three dimensional diagrams such as cylinders and cubes.
- **Pictograms or Ideographs**
- **Cartographs or Statistical maps**

2.6.1 One-Dimensional Diagrams

These diagrams are most useful, simple, and popular in the diagrammatic presentation of frequency distributions. These diagrams provides a useful and quick understanding of the *shape* of the distribution and its characteristics. According to Calvin F. Schmid, 'The simple bar chart with many variations is particularly appropriate for comparing the magnitude (or size) of coordinate items or of parts of a total. The basis of comparison in the bar is linear or one-dimensional.'

These diagrams are called one-dimensional diagrams because only the length (height) of the bar (not the width) is taken into consideration. Of course, width or thickness of the bar has no effect on the diagram, even then the thickness should not be too much otherwise the diagram would appear like a two-dimensional diagram.

Tips for Constructing a Diagram The following tips must be kept in mind while constructing one-dimensional diagrams:

- (i) The width of all the bars drawn should be same.
- (ii) The gap between one bar and another must be uniform.
- (iii) There should be a common base to all the bars.
- (iv) It is desirable to write the value of the variable represented by the bar at the top end so that the reader can understand the value without looking at the scale.
- (v) The frequency, relative frequency, or per cent frequency of each class interval is shown by drawing a rectangle whose base is the class interval on the horizontal axis and whose height is the corresponding frequency, relative frequency, or per cent frequency.
- (vi) The value of variables (or class boundaries in case of grouped data) under study are scaled along the horizontal axis, and the number of observations (frequencies, relative frequencies or percentage frequencies) are scaled along the vertical axis.

The one-dimensional diagrams (charts) used for graphical presentation of data sets are as follows:

- Histogram
- Frequency polygon
- Frequency curve
- Cumulative frequency distribution (Ogive)
- Pie diagram

Bar graph: A graphical device for depicting data that have been summarized in a frequency distribution, relative frequency distribution, or per cent frequency distribution.

Histograms (Bar Diagrams) These diagrams are used to graph both ungrouped and grouped data. In the case of an ungrouped data, values of the variable (the characteristic to be measured) are scaled along the horizontal axis and the number of observations (or frequencies) along the vertical axis of the graph. The plotted points are then connected by straight lines to enhance the shape of the distribution. The height of such boxes (rectangles) measures the number of observations in each of the classes.

Listed below are the various types of histograms:

- (i) Simple bar charts
- (ii) Grouped (or multiple) charts
- (iii) Deviation bar charts
- (iv) Subdivided bar charts
- (v) Paired bar charts
- (vi) Sliding bar charts
- (vii) Relative frequency bar charts
- (viii) Percentage bar charts

For plotting a histogram of a grouped frequency distribution, the end points of class intervals are specified on the horizontal axis and the number of observations (or frequencies) are specified on the vertical axis of the graph. Often class midpoints are posted on the horizontal axis rather than the end points of class intervals. In either case, the width of each bar indicates the class interval while the height indicates the frequency of observations in that class. Figure 2.2 is a histogram for the frequency distribution given in Table 2.12 of Example 2.1.

Remarks: Bar diagrams are not suitable to represent long period time series.

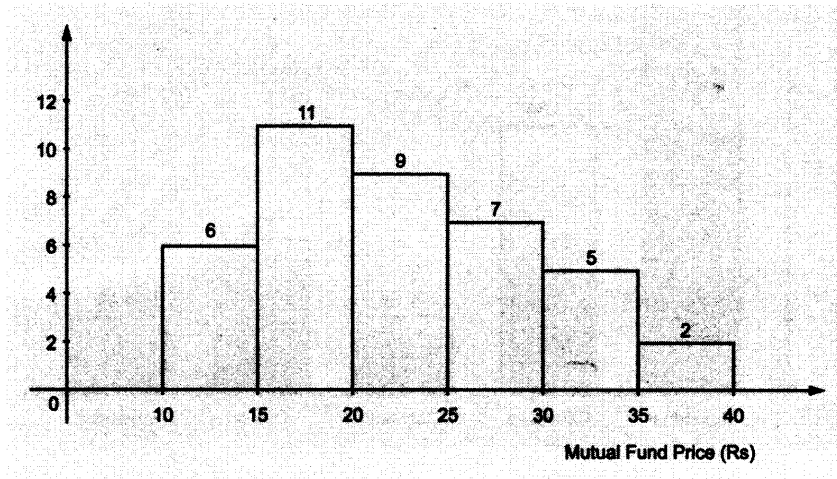


Figure 2.2
Histogram for Mutual Funds

Simple Bar Charts The graphic techniques described earlier are used for group frequency distributions. The graphic techniques presented in this section can also be used for displaying values of categorical variables. Such data is first tallied into summary tables and then graphically displayed as either *bar charts* or *pie charts*.

Bar charts are used to represent only one characteristic of data and there will be as many bars as number of observations. For example, the data obtained on the production of oil seeds in a particular year can be represented by such bars. Each bar would represent the yield of a particular oil seed in that year. Since the bars are of the same width and only the length varies, the relationship among them can be easily established.

Sometimes only lines are drawn for comparison of given variable values. Such lines are not thick and their number is sufficiently large. The different measurements to be shown should not have too much difference, so that the lines may not show too much dissimilarity in their heights.

Such charts are used to economize space, specially when observations are large. The lines may be either vertical or horizontal depending upon the type of variable—numerical or categorical.

Example 2.16: The data on the production of oil seeds in a particular year is presented in Table 2.35.

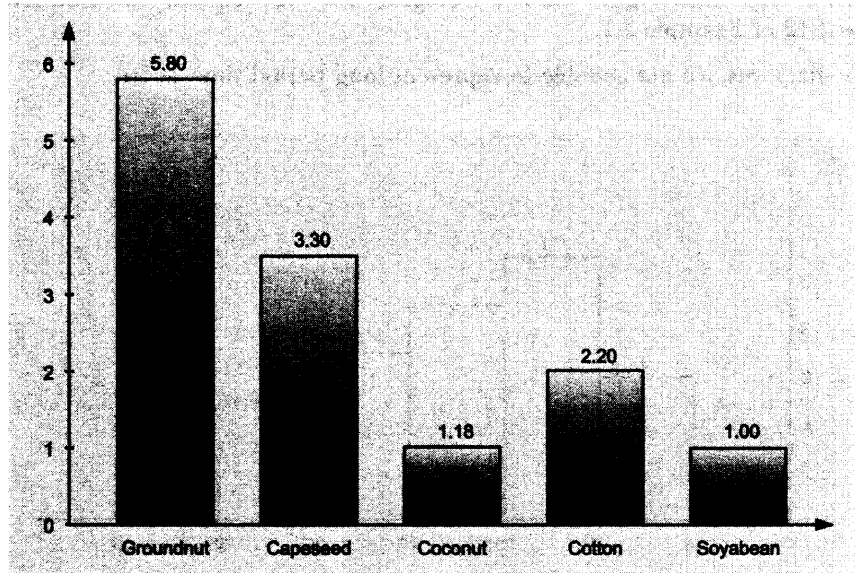
Table 2.35

Oil Seed	Yield (Million tonnes)	Percentage Production (Million tonnes)
Ground nut	5.80	43.03
Rapeseed	3.30	24.48
Coconut	1.18	8.75
Cotton	2.20	16.32
Soyabean	1.00	7.42
	<u>13.48</u>	<u>100.00</u>

Represent this data by a suitable bar chart.

Solution: The information provided in Table 2.35 is expressed graphically as the frequency bar chart as shown in Fig. 2.3. In this figure, each type of seed is depicted by a bar, the length of which represents the frequency (or percentage) of observations falling into that category.

Figure 2.3
Bar Chart Pertaining to Production of Oil Seeds



Remark: The bars should be constructed vertically (as shown in Fig. 2.3) when categorized observations are the outcome of a numerical variable. But if observations are the outcome of a categorical variable, then the bars should be constructed horizontally.

Example 2.17: An advertising company kept an account of response letters received each day over a period of 50 days. The observations were:

0	2	1	1	1	2	0	0	1	0	1	0	0	1	0	1	1	0
2	0	0	2	0	1	0	1	0	1	0	3	1	0	1	0	1	0
2	5	1	2	0	0	0	0	5	0	1	1	2	0				

Construct a frequency table and draw a line chart (or diagram) to present the data.

Solution: The observations are tallied into the summary table as shown in Table 2.35.

Figure 2.4 depicts a frequency bar chart for the number of letters received during a period of 50 days presented in Table 2.36.

Table 2.36 Frequency Distribution of Letters Received

Number of Letters Received	Days	Number of Days (Frequency)
0		23
1		17
2		7
3		2
4		0
5		1
		<hr/> 50

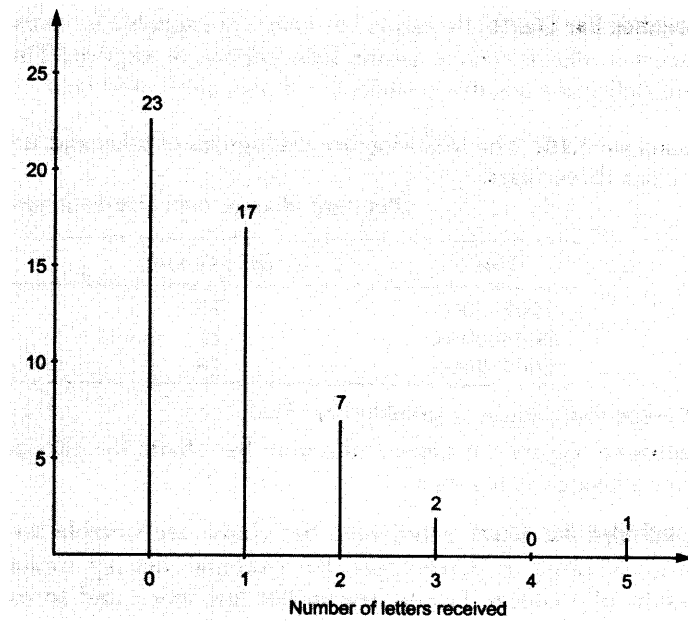


Figure 2.4
Number of Letters Received

Multiple Bar Charts A multiple bar chart is also known as grouped (or compound) bar chart. Such charts are useful for direct comparison between two or more sets of data. The technique of drawing such a chart is same as that of a single bar chart with a difference that each set of data is represented in different shades or colours on the same scale. An index explaining shades or colours must be given.

Example 2.18: The data on fund flow (Rs in crore) of an International Airport Authority during financial years 2001–02 to 2003–04 are given below:

	2001–02	2002–03	2003–04
Non-traffic revenue	40.00	50.75	70.25
Traffic revenue	70.25	80.75	110.00
Profit before tax	40.15	50.50	80.25

Represent this data by a suitable bar chart.

Solution: The multiple bar chart of the given data is shown in Fig. 2.5.

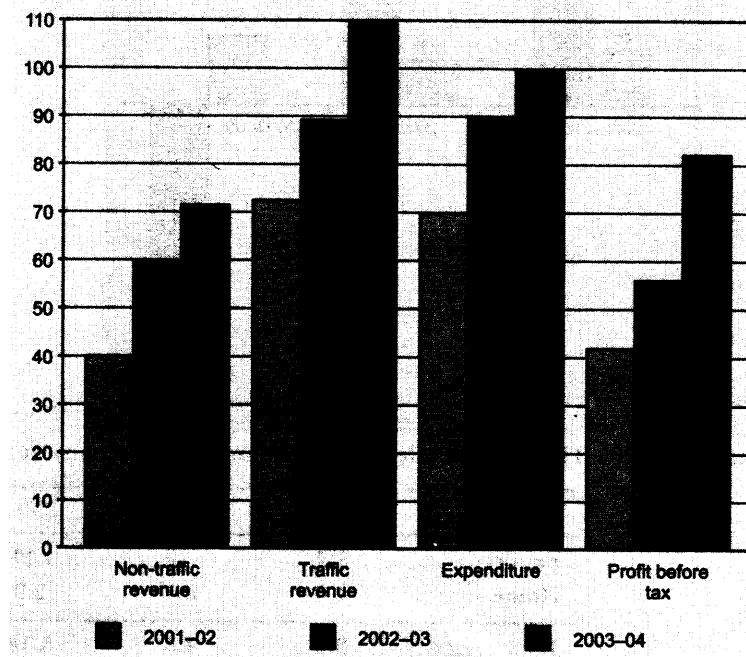


Figure 2.5
Multiple Bar Chart Pertaining to
Performance of an International
Airport Authority

Deviation Bar Charts Deviation bar charts are suitable for presentation of net quantities in excess or deficit such as profit, loss, import, or exports. The excess (or positive) values and deficit (or negative) values are shown above and below the base line.

Example 2.19: The following are the figures of sales and net profits of a company over the last three years.

(Per cent change over previous year)

Year	Sales Growth	Net Profit
2002-2003	15	30
2003-2004	12	53
2004-2005	18	-72

Present this data by a suitable bar chart.

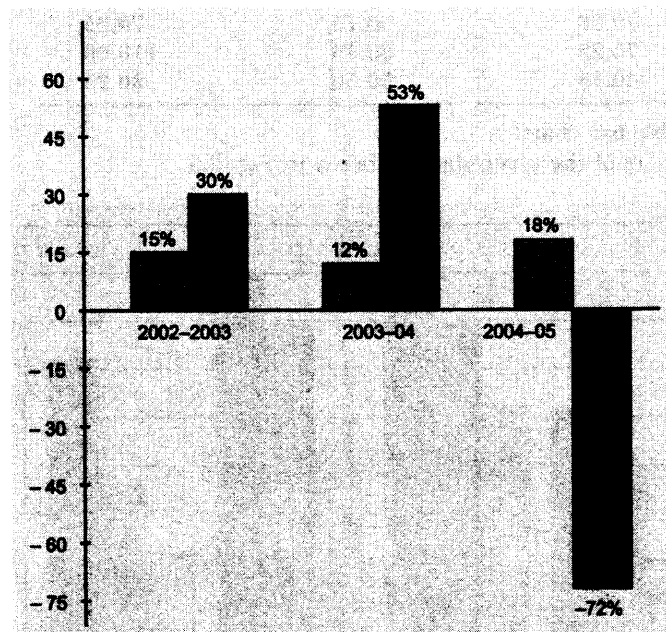
Solution: Figure 2.6 depicts deviation bar charts for sales and per cent change in sales over previous year's data.

Subdivided Bar Chart Subdivided bar charts are suitable for expressing information in terms of ratios or percentages. For example, net per capita availability of food grains, results of a college faculty-wise in last few years, and so on. While constructing these charts the various components in each bar should be in the same order to avoid confusion. Different shades must be used to represent various ratio values but the shade of each component should remain the same in all the other bars. An index of the shades should be given with the diagram.

A common arrangement while making these charts is that of presenting each bar in order of magnitude from the largest component at the base of the bar to the smallest at the end.

Since the different components of the bars do not start on the same scale, the individual bars are to be studied properly for their mutual comparisons.

Figure 2.6
Deviation Bar Chart Pertaining to
Sales and Profits



Example 2.20: The data on sales (Rs in million) of a company are given below:

	2002	2003	2004
Export	1.4	1.8	2.29
Home	1.6	2.7	2.9
Total	3.0	4.5	5.18

Solution: Figure 2.7 depicts a subdivided bar chart for the given data.

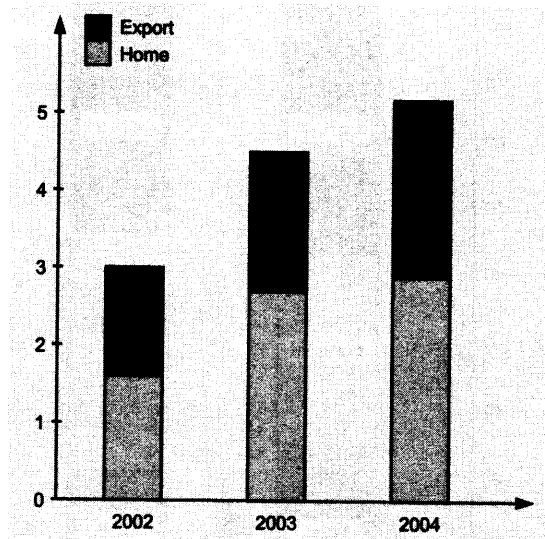


Figure 2.7
Subdivided Bar Chart Pertaining to Sales

Percentage Bar Charts When the relative proportions of components of a bar are more important than their absolute values, then each bar can be constructed with same size to represent 100%. The component values are then expressed in terms of percentage of the total to obtain the necessary length for each of these in the full length of the bars. The other rules regarding the shades, index, and thickness are the same as mentioned earlier.

Example 2.21: The following table shows the data on cost, profit, or loss per unit of a good produced by a company during the year 2003–04.

Particulars	2003			2004		
	Amount (Rs)	Percentage	Cumulative Percentage	Amount (Rs)	Percentage	Cumulative Percentage
Cost per unit						
(a) Labour	25	41.67	41.67	34	40.00	40.00
(b) Material	20	33.33	75.00	30	35.30	75.30
(c) Miscellaneous	15	25.00	100.00	21	24.70	100.00
Total cost	60	100		85	100	
Sales proceeds per unit	80	110		80	88	
Profit (+) or loss (-) per item	+ 20	+ 10		- 5	- 12	

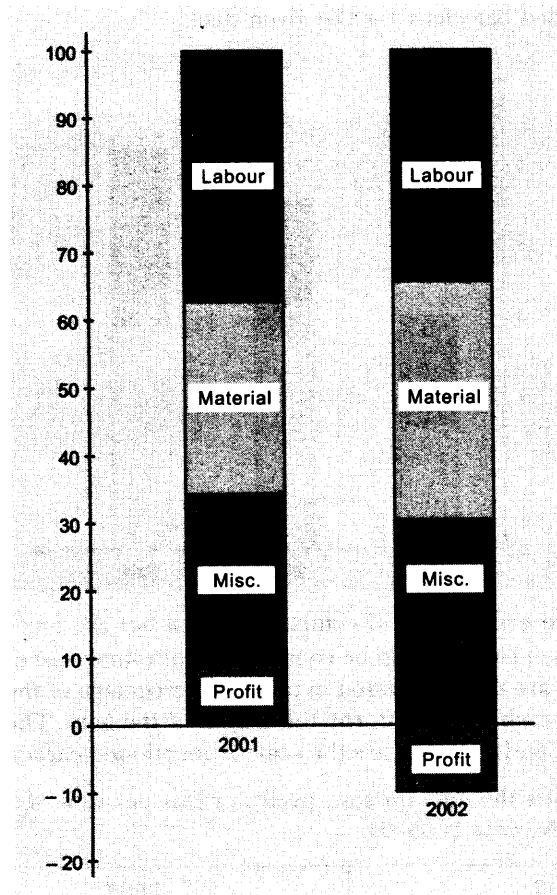
Represent diagrammatically the data given above on percentage basis.

Solution: The cost, sales, and profit/loss data expressed in terms of percentages have been represented in the bar chart as shown in Fig. 2.8.

Frequency Polygon As shown in Fig. 2.9, the frequency polygon is formed by marking the midpoint at the top of horizontal bars and then joining these dots by a series of straight lines. The frequency polygons are formed as a closed figure with the horizontal axis, therefore a series of straight lines are drawn from the midpoint of the top base of the first and the last rectangles to the mid point falling on the horizontal axis of the next outlying interval with zero frequency. The frequency polygon is sometimes jagged in appearance.

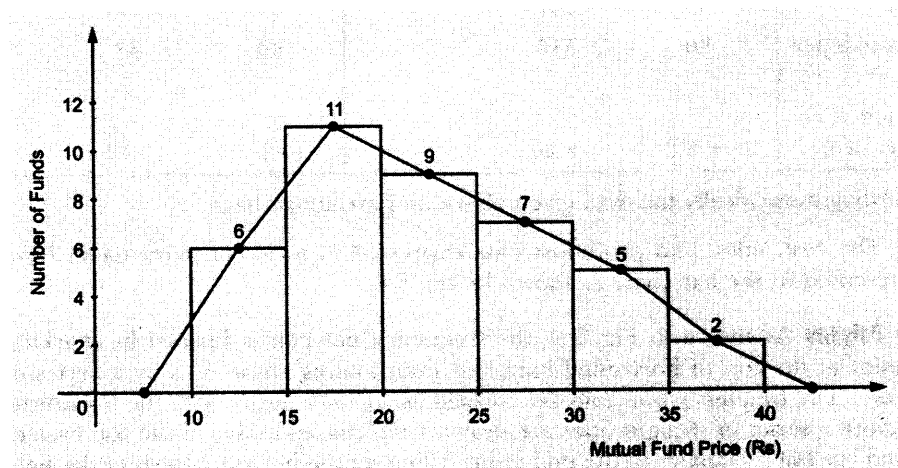
A frequency polygon can also be converted back into a histogram by drawing vertical lines from the bounds of the classes shown on the horizontal axis, and then connecting them with horizontal lines at the heights of the polygon at each mid-point.

Figure 2.8
Percentage Bar Chart Pertaining to
Cost, Sales, and Profit/Loss



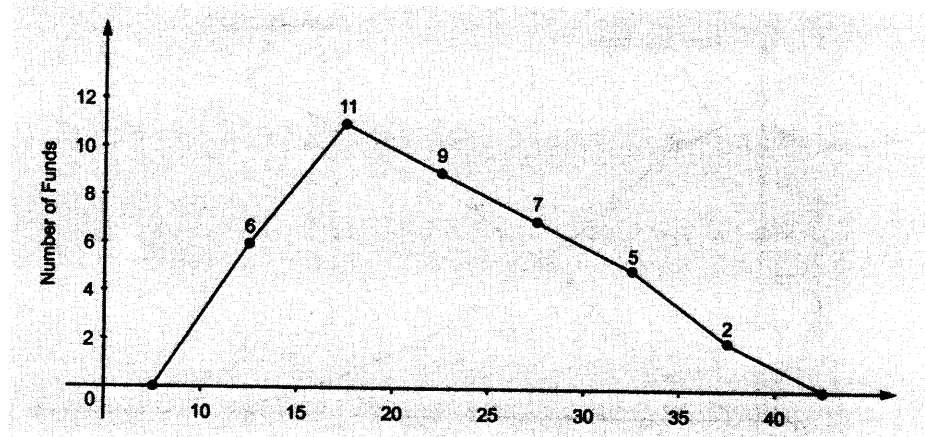
Drawing a frequency polygon does not necessarily require constructing a histogram first. A frequency polygon can be obtained directly on plotting points above each class midpoint at heights equal to the corresponding class frequency. The points so drawn are then joined by a series of straight lines and the polygon is closed as explained earlier. In this case, horizontal x -axis measures the successive class mid points and not the lower class limits. Figure 2.9 shows the frequency polygon for the frequency distribution presented by histogram in Fig. 2.2.

Figure 2.9
Frequency Polygon for Mutual Fund



Frequency Curve It is described as a smooth frequency polygon as shown in Fig. 2.10. A frequency curve is described in terms of its (i) symmetry (skewness) and its (ii) degree of peakedness (kurtosis). The concepts of skewness and kurtosis describing a frequency distribution will be discussed in Chapter 5.

Figure 2.10
Frequency Curve



Two frequency distributions can also be compared by superimposing two or more frequency curves provided the width of their class intervals and the total number of frequencies are equal for the given distributions. Even if the distributions to be compared differ in terms of total frequencies, they still can be compared by drawing per cent frequency curves where the vertical axis measures the per cent class frequencies and not the absolute frequencies.

Cumulative Frequency Distribution (Ogive) It enables us to see how many observations lie above or below certain values rather than merely recording the number of observations within intervals. Cumulative frequency distribution is another method of data presentation that helps in data analysis and interpretation. Table 2.37 shows the cumulative number of observations below and above the upper boundary of each class in the distribution.

A cumulative frequency curve popularly known as *Ogive* is another form of graphic presentation of a cumulative frequency distribution. The ogive for the cumulative frequency distribution given in Table 2.37 is presented in Fig. 2.11.

Once cumulative frequencies are obtained, the remaining procedure for drawing curve, called ogive is as usual. The only difference being that the y-axis now has to be so scaled that it accommodates the total frequencies. The x-axis is labelled with the upper class limits in the case of less than ogive, and the lower class limits in case of more than ogive.

Table 2.37 Calculation of Cumulative Frequencies

Mutual Fund Price (Rs)	Upper Class Boundary	Number of Funds (f)	Cumulative Frequency	
			Less than	More than
10-15	15	6	6	40
15-20	20	11	$6 + 11 = 17$	$40 - 6 = 34$
20-25	25	9	$17 + 9 = 26$	$34 - 11 = 23$
25-30	30	7	$26 + 7 = 33$	$23 - 9 = 14$
30-35	35	5	$33 + 5 = 38$	$14 - 7 = 7$
35-40	40	2	$38 + 2 = 40$	$7 - 5 = 2$

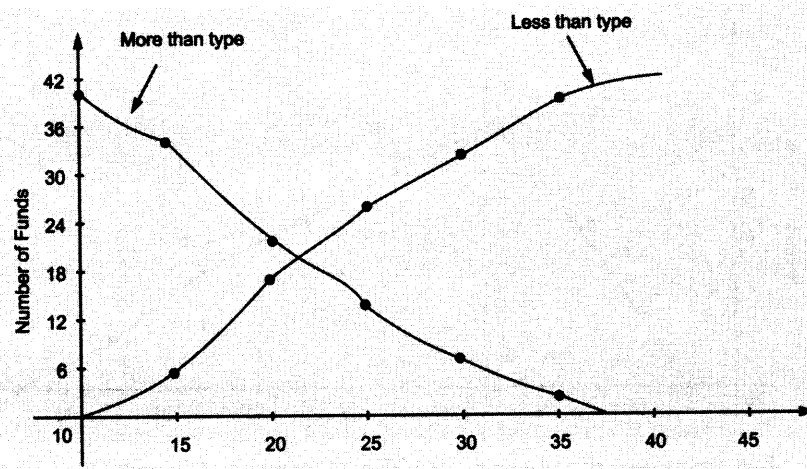


Figure 2.11
Ogive for Mutual Fund Prices

To draw a cumulative 'less than ogive', points are plotted against each successive upper class limit and the corresponding less than cumulative frequency value. These points are then joined by the series of straight lines and the resultant curve is closed at the bottom by extending it so as to meet the horizontal axis at the real lower limit of the first class interval.

To draw a cumulative 'more than ogive', points are plotted against each successive lower class limit and the corresponding more than cumulative frequency. These points are joined by the series of straight lines and the curve is closed at the bottom by extending it to meet the horizontal axis at the upper limit of the last class interval. Both the types of ogives so drawn are shown in Fig. 2.11.

It may be mentioned that a line drawn parallel to the vertical axis through the point of intersection of the two types of ogives will meet the x-axis at its middle point, and the value corresponding to this point will be the median of the distribution. Similarly, the perpendicular drawn from the point of intersection of the two curves on the vertical axis will divide the total frequencies into two equal parts.

Two ogives, whether *less than* or *more than* type, can be readily compared by drawing them on the same graph paper. The presence of unequal class intervals poses no problem in their comparison, as it does in the case of comparison of two frequency polygons. If the total frequencies are not the same in the two distributions, they can be first converted into per cent frequency distributions and then ogives drawn on a single graph paper to facilitate comparison.

Pie Diagram These diagrams are normally used to show the total number of observations of different types in the data set on a percentage basis rather than on an absolute basis through a circle. Usually the largest percentage portion of data in a pie diagram is shown first at 12 o'clock position on the circle, whereas the other observations (in per cent) are shown in clockwise succession in descending order of magnitude. The steps to draw a pie diagram are summarized below:

- (i) Convert the various observations (in per cent) in the data set into corresponding degrees in the circle by multiplying each by 3.6 ($360 \div 100$).
- (ii) Draw a circle of appropriate size with a compass.
- (iii) Draw points on the circle according to the size of each portion of the data with the help of a protractor and join each of these points to the center of the circle.

The pie chart has two distinct advantages: (i) it is aesthetically pleasing and (ii) it shows that the total for all categories or slices of the pie adds to 100%.

Example 2.22: The data shows market share (in per cent) by revenue of the following companies in a particular year:

Batata-BPL	30	Escorts-First Pacific	5
Hutchison-Essar	26	Reliance	3
Bharti-Sing Tel	19	RPG	2
Modi Dista Com	12	Srinivas	2
		Shyam	1

Draw a pie diagram for the above data.

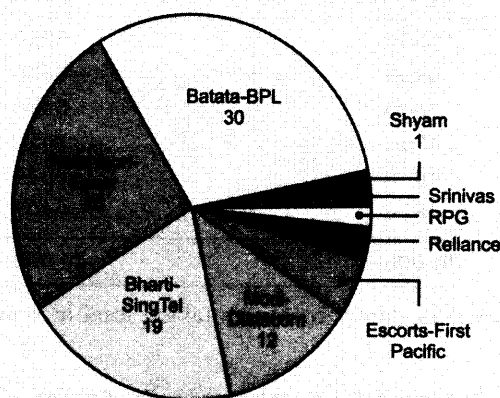
Solution: Converting percentage figures into angle outlay by multiplying each of them by 3.6 as shown in Table 2.38.

Table 2.38

<i>Company</i>	<i>Market Share (Per cent)</i>	<i>Angle Outlay (Degree)</i>
Batata-BPL	30	108.0
Hutchison-Essar	26	93.6
Bharti-Sing Tel	19	68.4
Modi Dista Com	12	43.2
Escorts First Pacific	5	18.0
Reliance	3	10.8
RPG	2	7.2
Srinivas	2	7.2
Shyam	1	3.6
Total	100	360.0

Using the data given in Table 2.38 construct the pie chart displayed in Fig. 2.12 by dividing the circle into 9 parts according to degrees of angle at the centre.

Figure 2.12
Percentage Pie Chart



Example 2.23: The following data relate to area in millions of square kilometer of oceans of the world.

<i>Ocean</i>	<i>Area (Million sq km)</i>
Pacific	70.8
Atlantic	41.2
Indian	28.5
Antarctic	7.6
Arctic	4.8

Solution: Converting given areas into angle outlay as shown in Table 2.39.

Table 2.39

<i>Ocean</i>	<i>Area (Million sq km)</i>	<i>Angle Outlay (Degrees)</i>
Pacific	70.8	$\frac{70.8}{152.9} \times 360 = 166.70$
Atlantic	41.2	$\frac{41.2}{152.9} \times 360 = 97.00$
Indian	28.5	67.10
Antarctic	7.6	17.89
Arctic	4.8	11.31
Total	152.9	360.00

Pie diagram is shown in Fig. 2.13.

2.6.2 Two-Dimensional Diagrams

In one-dimensional diagrams or charts only the length of the bar is taken into consideration. But in two-dimensional diagrams both its height and width are taken into account for presenting the data. These diagrams, also known as *surface diagrams* or *area diagrams*, are:

- Rectangles
- Squares, and
- Circles.

Rectangles Since area of a rectangle is equal to the product of its length and width, therefore while making such type of diagrams both length and width are considered.

Rectangles are suitable for use in cases where two or more quantities are to be compared and each quantity is sub-divided into several components.

Example 2.24: The following data represent the income of two families A and B. Construct a rectangular diagram.

Item of Expenditure	Family A (Monthly Income Rs 30,000)	Family B (Monthly Income Rs 40,000)
Food	5550	7280
Clothing	5100	6880
House rent	4800	6480
Fuel and light	4740	6320
Education	4950	6640
Miscellaneous	4860	6400
Total	30,000	40,000

Solution: Converting individual values into percentages taking total income as equal to 100 as shown in Table 2.40.

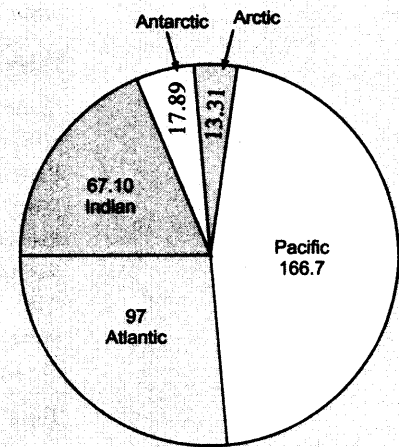
Table 2.40 Percentage Summary Table Pertaining to Expenses Incurred by Two Families

Item of Expenditure	Family A (Monthly Income Rs 30,000)			Family B (Monthly Income Rs 40,000)		
	Actual Expenses	Percentage of Expenses	Cumulative Percentage	Actual Expenses	Percentage of Expenses	Cumulative Percentage
Food	5550	18.50	18.50	7280	18.20	18.20
Clothing	5100	17.00	35.50	6880	17.20	35.40
House rent	4800	16.00	51.50	6480	16.20	51.60
Fuel and light	4740	15.80	67.30	6320	15.80	67.40
Education	4950	16.50	83.80	6640	16.60	84.00
Miscellaneous	4860	16.20	100.00	6400	16.00	100.00
Total		30,000	100.00		40,000	100.00

The height of the rectangles shown in Fig. 2.14 is equal to 100. The difference in the total income is represented by the difference on the base line which is in the ratio of 3 : 4.

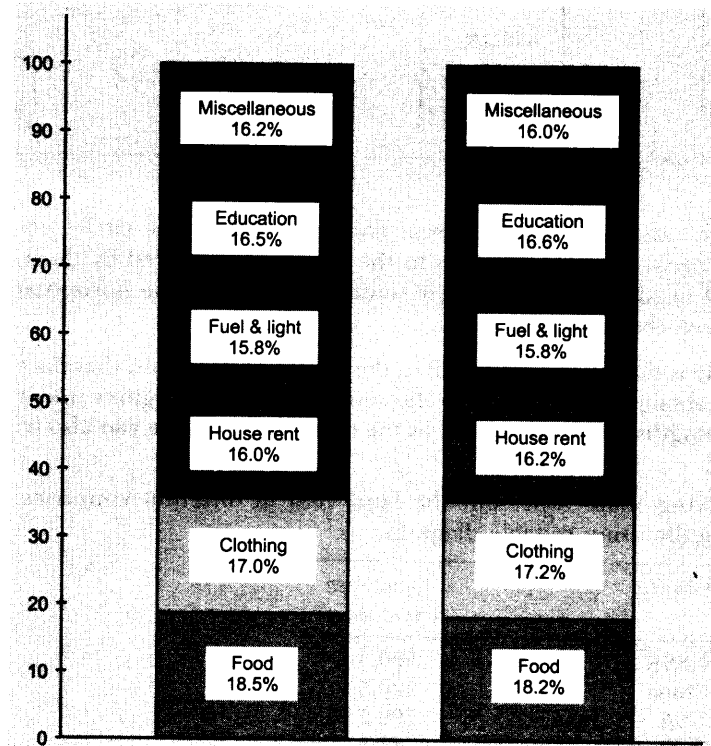
Squares Squares give a better comparison than rectangular bars when the difference of totals to be compared is large. For example, if in Example 2.24 the total expenses of families A and B are Rs 2000 and 20,000 respectively, then the width of the rectangles would be in the ratio 1 : 10. If such a ratio is taken, the diagram would look very unwieldy. Thus to overcome this difficulty squares are constructed to make use of their areas to represent given data for comparison.

Figure 2.13
Per cent Pie Diagram



To construct a square diagram, first take the square-root of the values of various figures to be represented and then these values are divided either by the lowest figure or by some other common figure to obtain proportions of the sides of the squares. The squares constructed on these proportionate lengths must have either the base or the centre on a straight line. The scale is attached with the diagram to show the variable value represented by one square unit area of the squares.

Figure 2.14
Percentage of Expenditure by Two Families



Example 2.25: The following data represent the production (in million tonnes) of coal by different countries in a particular year.

Country	Production
USA	130.1
USSR	44.0
UK	16.4
India	3.3

Represent the data graphically by constructing a suitable diagram.

Solution: The given data can be represented graphically by square diagrams. For constructing the sides of the squares, the necessary calculations are shown in Table 2.41.

Table 2.41 Side of a Square Pertaining to Production of Coal

Country	Production (Million tonnes)	Square Root of Production Amount	Side of a Square (One square inch)
USA	130.1	11.406	1.267
USSR	44.0	6.633	0.737
UK	16.4	4.049	0.449
India	3.3	1.816	0.201

The squares representing the amount of coal production by various countries are shown in Fig. 2.15.

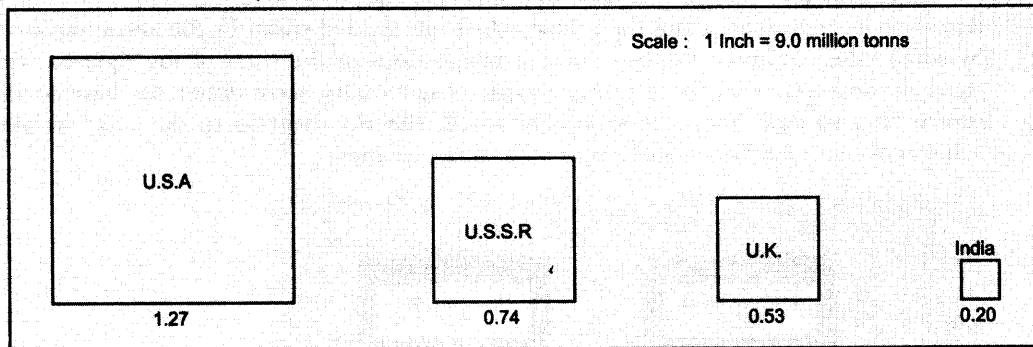


Figure 2.15
Coal Production in
Different Countries

Circles Circles are alternative to squares to represent data graphically. The circles are also drawn such that their areas are in proportion to the figures represented by them. The circles are constructed in such a way that their centres lie on the same horizontal line and the distance between the circles are equal.

Since the area of a circle is directly proportional to the square of its radius, therefore the radii of the circles are obtained in proportion to the square root of the figures under representation. Thus, the lengths which were used as the sides of the square can also be used as the radii of circles.

Example 2.26: The following data represent the land area in different countries. Represent this data graphically using suitable diagram.

Country	Land Area (crore acres)
USSR	590.4
China	320.5
USA	190.5
India	81.3

Solution: The data can be represented graphically using circles. The calculations for constructing radii of circles are shown in Table 2.42.

Table 2.42 Radii of Circles Pertaining to Land Area of Countries

Country	Land Area (crore acres)	Square Root of Land Area	Radius of Circles (Inches)
USSR	590.4	24.3	0.81
China	320.5	17.9	0.60
USA	190.5	13.8	0.46
India	81.3	9.0	0.30

The various circles representing the land area of respective countries are shown in Fig. 2.16.

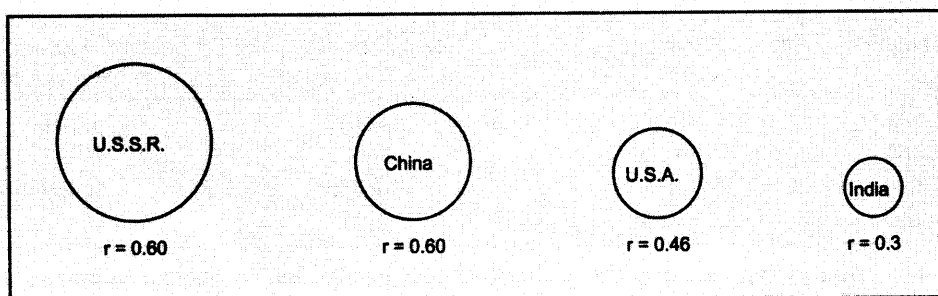


Figure 2.16
Land Area of Different Countries

2.6.3 Three-Dimensional Diagrams

Cylinders, spheres, cubes, and so on are known as three-dimensional diagrams because three dimensions—length, breadth, and depth, are taken into consideration for representing figures. These diagrams are used when only one point is to be compared and the ratio between the highest and the lowest measurements is more than 100 : 1. For constructing these diagrams, the cube root of various measurement is calculated and the side of the each cube is taken in proportion to the cube roots.

Amongst the three-dimensional diagrams, cubes are the easiest and should be used only in those cases where the figures cannot be adequately presented through bar, square, or circle diagrams.

2.6.4 Pictograms or Ideographs

A pictogram is another form of pictorial bar chart. Such charts are useful in presenting data to people who cannot understand charts. Small symbols or simplified pictures are used to represent the size of the data. To construct pictograms or ideographs the following suggestions are made:

- (i) The symbols must be simple and clear.
- (ii) The quantity represented by a symbol should be given.
- (iii) Larger quantities are shown by increasing the number of symbols, and not by increasing the size of the symbols. A part of a symbol can be used to represent a quantity smaller than the whole symbol.

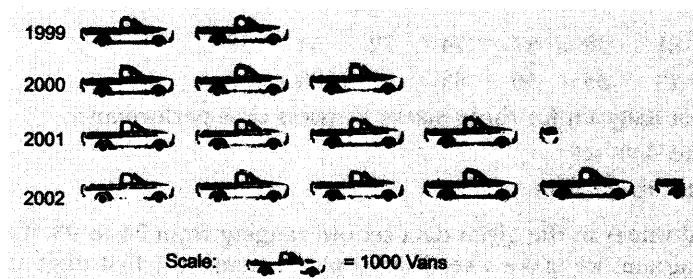
Example 2.27: Make a pictographic presentation of the output of vans during the year by a van manufacturing company.

Year	:	1999	2000	2001	2002
Output	:	2004	2996	4219	5324

Solution: Dividing the van output figures by 1000, we get 2.004, 2.996, 4.219, and 5.324 respectively.

Representing these figures by pictures of vans as shown in Fig. 2.17.

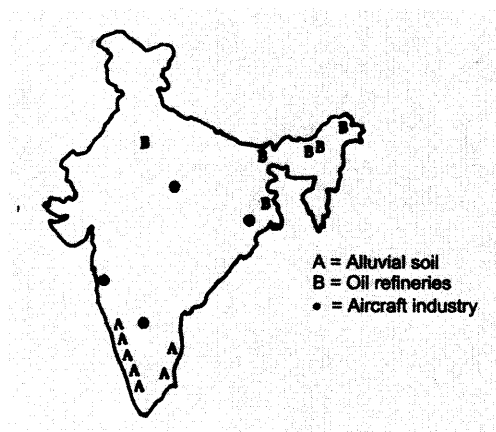
Figure 2.17
Output of Vans



2.6.5 Cartograms or Statistical Maps

Cartograms are used to represent graphical distribution of data on maps. The various figures in different regions on maps are shown either by (i) shades or colours, (ii) dots or bars, (iii) diagrams or pictures, or (iv) by putting numerical figures in each geographical area.

The following maps show the location of a particular type of soil, refineries, and aircraft industry in the country.



2.7 EXPLORATORY DATA ANALYSIS

This technique helps us to quickly describe and summarize a data set using simple arithmetic and diagrams. Such presentation of data values provides ways to determine relationships and trends, identify outliers and influential observations. In this section one of the useful techniques of exploratory data analysis, *stem-and-leaf displays (or diagrams)* technique is presented. This technique provides the *rank order* of the values in the data set and the shape of the distribution.

2.7.1. Stem-and-Leaf Displays

The stem-and-leaf display (or diagram) is another very simple but powerful technique to display quantitative data in a condensed form. The advantage of stem-and-leaf display over a frequency distribution is that the identity of each observation remains intact. This diagram provides us the rank order of the numerical values from lowest to highest in the data set and reveal the center, spread, shape and outliers (extremes) of a distribution. It is a graphical display of the numerical values in the data set and separates these values into *leading digits (or stem)* and *trailing digits (or leaves)*. The steps required to construct a stem-and-leaf diagram are as follows:

1. Divide each numerical value between the ones and the tens place. The number to the left is the stem and the number to the right is the leaf. The stem contains all but the last of the displayed digits of a numerical value. As with histogram, it is reasonable to have between 6 to 15 stems (each stem defines an interval of values). The stem should define equally spaced intervals. Stems are located along the vertical axis.

Sometimes numerical values in the data set are truncated or rounded off. For example, the number 15.69 is truncated to 15.6 but it is rounded off to 15.7.

2. List the stems in a column with a vertical line to their right.
3. For each numerical value attach a leaf to the appropriate stem in the same row (horizontal axis). A leaf is the last of the displayed digits of a number. It is standard, but not mandatory, to put the leaves in increasing order at each stem value.
4. Provide a key to stem and leaf coding so that actual numerical value can be re-created, if necessary.

Remark If all the numerical values are three-digit integers, then to form a stem-and-leaf diagram, two approaches are followed:

- (i) Use the hundreds column as the stems and the tens column as the leaves and ignore the units column.
- (ii) Use the hundreds column as the stems and the tens column as the leaves after rounding of the units column.

Example 2.28 Consider the following marks obtained by 20 students in a business statistics test:

64 89 63 61 78 87 74 72 54 88
62 81 78 73 63 56 83 86 83 93

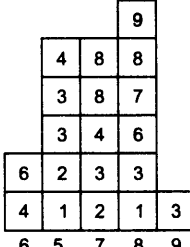
- (a) Construct a stem-and-leaf diagram for these marks to assess class performance
- (b) Describe the shape of this data set
- (c) Are there any outliers in this data set.

Solution (a) The numerical values in the given data set are ranging from 54 to 93. To construct a stem-and-leaf diagram, we make a vertical list of the stems (the first digit of each numerical value) as shown below:

Stem	Leaf
5	46
6	43123
7	84283
8	9781863
9	3

Rearrange all of the leaves in each row in rank order.

Stem	Leaf
5	46
6	12334
7	23488
8	1367889
9	3



Each row in the diagram is a stem and numerical value on that stem is a leaf. For example, if we take the row 6/12334, it means there are five numerical values in the data set that begins with 6, i.e. 61, 62, 63, 63 and 64.

If the page is turned 90 degree clockwise and draw rectangles around the digits in each stem, we get a diagram similar to a histogram.

(b) Shape of the diagram is not symmetrical.

(c) There is no outlier (an observation far from the center of the distribution).

Example 2.29 The following data represent the annual family expenses (in thousand of rupees) on food items in a city.

13.8	14.1	14.7	15.2	12.8	15.6	14.9	16.7	19.2
14.9	14.9	14.9	15.2	15.9	15.2	14.8	14.8	19.1
14.6	18.0	14.9	14.2	14.1	15.3	15.5	18.0	17.2
17.2	14.1	14.5	18.0	14.4	14.2	14.6	14.2	14.8

Construct the stem-end-leaf diagram.

Solution: Since the annual costs (in Rs 000) in the data set all have two-digit integer numbers, the tens and units columns would be the leading digits and the remaining column (the tenth column) would be trailing digits as shown below:

Stem	Leaf	Stem	Leaf
12	8	12	8
13	8	13	8
14	17999988621142628	14	11122246678889999
15	2629235	15	2223569
16	7	16	7
17	2	17	2
18	000	18	000
19	21	19	12

Rearrange all the leaves in each row in the rank order as shown above.

Conceptual Questions 2C

- What are the different types of charts known to you? What are their uses?
- Point out the role of diagrammatic presentation of data. Explain briefly the different types of bar diagrams known to you.
- Charts are more effective in attracting attention than other methods of presenting data. Do you agree? Give reasons for your answer. [MBA, HP Univ., 1998]
- Discuss the utility and limitations (if any) of diagrammatic presentation of statistical data.
- Diagrams are meant for a rapid view of the relation of different data and their comparisons. Discuss
- Write short notes on pictographic and cartographic representations of statistical data.
- What are the advantages of using a graph to describe a frequency distribution?
- When constructing a graph of a grouped frequency distribution, is it necessary that the resulting distribution be symmetric? Explain.
- Explain what is meant by a frequency polygon, a histogram, and a frequency curve.
- Define the terms relative frequency and cumulative frequency. How are these related to a frequency distribution?
- The distribution of heights of all students in the commerce department of the university has two peaks or is bimodal. The distribution of the IQ's of the same students, however, has only one peak. How is this possible since the same students are considered in both cases? Explain.

Self-Practice Problems 2C

2.28 The following data represent the gross income, expenditure (in Rs lakh), and net profit (in Rs lakh)

during the years 1999 to 2002.

	1999–2000	2000–2001	2001–2002
Gross income	570	592	632
Gross expenditure	510	560	610
Net income	60	32	22

Construct a diagram or chart you prefer to use here.

- 2.29** Which of the charts would you prefer to represent the following data pertaining to the monthly income of two families and the expenditure incurred by them.

Expenditure on	Family A (Income Rs 17,000)	Family B (Income Rs 10,000)
Food	4000	5400
Clothing	2800	3600
House rent	3000	3500
Education	2300	2800
Miscellaneous	3000	5000
Saving or deficits	+1900	-300

- 2.30** The following data represent the outlays (Rs crore) by heads of development.

Heads of Development	Centre	States
Agriculture	4765	7039
Irrigation and Flood control	6635	11,395
Energy	9995	8293
Industry and Minerals	12,770	2985
Transport and Communication	12,200	5120
Social services	8216	1420
Total	54,581	36,252

Represent the data by a suitable diagram and write a report on the data bringing out the silent features.

- 2.31** Make a diagrammatic representation of the following textile production and imports.

	Value (in crore)	Length (in hundred yards)
Mill production	116.4	426.9
Handloom production	106.8	192.8
Imports	319.7	64.7

What conclusions do you draw from the diagram?

- 2.32** Make a diagrammatic representation of the following data:

Country	Production of Sugar in a Certain Year in Quintals (10,00,000)
Cuba	32
Australia	30
India	20
Japan	5
Java	1
Egypt	1

- 2.33** The following data represent the estimated gross area under different cereal crops during a particular year.

Crop	Gross Area ('000 hectares)	Crop	Gross Areas ('000 hectares)
Paddy	34,321	Ragi	2656
Wheat	18,287	Maize	6749
Jowar	22,381	Barley	4422
Bajra	15,859	Small millets	6258

Draw a suitable chart to represent the data.

- 2.34** The following data indicate the rupee sales (in 1000's) of three products according to region.

Product Group	Sales (in Rs 1000)			Total Sales (Rs 1000)
	North	South	East	
A	70	75	90	135
B	90	60	100	250
C	50	60	40	150
	210	195	230	533

- Using vertical bars, construct a bar chart depicting total sales regionwise.
- Construct a component chart to illustrate the product breakdown of sales region-wise by horizontal bars.
- Construct a pie chart illustrating total sales.

- 2.35** The following data represent the income and dividend for the year 2000.

Year	Income Per Share (in Rs)	Dividend Per Share (in Rs)
1995	5.89	3.20
1996	6.49	3.60
1997	7.30	3.85
1998	7.75	3.95
1999	8.36	3.25
2000	9.00	4.45

- Construct a line graph that indicates the income per share for the period 1995–2000.
- Construct a component bar chart that depicts dividends per share and retained earning per share for the period 1995–2000.
- Construct a percentage pie chart depicting the percentage of income paid as dividend. Also construct a similar percentage pie chart for the period 1998–2000. Observe any difference between the two pie charts.

- 2.36** The following time series data taken from the annual report of a company represents per-share net income, dividend, and retained earning during the period 1996–2000.

Source	1996	1997	1998	1999	2000
Net income (in Rs)	67.40	67.54	66.44	67.78	14.62
Dividends (in Rs)	8.08	8.28	8.40	8.50	8.75
Retained earnings (in Rs)	66.82	66.81	65.64	66.88(-)	10.56

- (i) Construct a bar chart for pre-share income for the company during 1996–2000.
- (ii) Construct a component bar chart depicting the allocation of annual earnings for the company during 1996–2000.
- (iii) Construct a line graph for the pre-share net income for the period 1996–2000.

2.37 The following data indicate the number of foreign tourists arrived in India during the period 1998–2001.

Country	Number of Tourists Arrived		
	1998–9	1999–2000	2000–1
USA	2110	2340	2245
UK	5393	6245	6384
Middle East and Gulf	1114	1045	1097
Australia	2432	1849	2249
Western and Eastern Europe	5492	5890	5990

- (a) Construct a line graph for the arrival of tourists during 1998–2001.
- (b) Construct a histogram for the data.
- (c) Construct a percentage pie chart for the different countries.

2.38 Find a business or economic related data set of interest to you. The data set should be made up of at least 100 quantitative observations.

- (a) Show the data in the form of a standard frequency distribution.
- (b) Using the information obtained from part (i) briefly describe the appearance of your data.

2.39. The first row of a stem-end-leaf diagram appears as follows: 26/14489. Assume whole number values

- (a) What is the possible range of values in this row?

(b) How many data values are in this row?

(c) List the actual values in this row of data.

2.40. Given the following stem-end-leaf display representing the amount of CNG purchased in litres (with leaves in tenths litre) for a sample of 25 vehicles in Delhi.

9	714
10	82230
11	561776735
12	394282
13	20

(a) Rearrange the leaves and form the revised stem-end-leaf display.

(b) Place the data into an ordered array.

2.41. The following stem-end-leaf display shows the number of units produced per day of in item an a factory.

3	8
4	–
5	6
6	0133559
7	0236778
8	59
9	00156
10	36

(a) How many days were studied

(b) What are the smallest value and the largest value?

(c) List the actual values in second and fourth row.

(d) How many values are 80 or more.

(e) What is the middle value.

2.42. A survey of the number of customer used PCO/STD both located at a college gate to make telephone calls last week revealed the following information

52	43	30	38	30	42	12	46
39	37	34	46	32	18	41	5

(a) Develop a stem-and-leaf display

(b) How many calls did a typical customer made?

(c) What were the largest and the smallest number of calls made?

Hints and Answers

2.39. (a) 260 to 269 (b) 5
(c) 261, 264, 264, 268, 269

2.40. (a)

9	147
10	02238
11	135566777
12	223489
13	02

(b) 91 94 97 100 102 102 103 108 111
113 115 115 116 116 117 117 122 122
123 124 128 129 130 132

2.41. (a) 25 (b) 38, 106
(c) No values, 60, 61, 63, 65, 65, 69
(d) 9 (e) 76 (f) 16

2.42. (a)

0	5
1	28
2	–
3	0024789
4	12366
5	2

(b) 16 customers were studied

(c) Number of customers visited ranged from 5 to 52.

Formulae Used

1. Class interval for a class in a frequency distribution

$$h = \text{Upper limit} - \text{Lower limit}$$

2. Midpoint of a class in a frequency distribution

$$m = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$

3. Approximate interval size to be used in constructing a frequency distribution

$$h = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of class intervals}}$$

4. Approximate number of class intervals for constructing a frequency distribution: $2^k \geq N$, where k and N represent the number of classes and total number of observations, respectively.

Chapter Concepts Quiz

True or False

1. The midpoint of a class interval is h . (T/F)
2. Frequency distribution of continuous data can be presented graphically as histograms or frequency polygons. (T/F)
3. A frequency polygon is appropriate for graphing continuously distributed variables. (T/F)
4. The percentile rank of a score is equal to the frequency of the scores falling up to and including the score. (T/F)
5. Simple bar diagram is used only for one-dimensional comparisons. (T/F)
6. Pie diagram is a circle divided into sections with areas equal to the corresponding component. (T/F)
7. A pie diagram is inappropriate for representing nominally scaled data. (T/F)
8. The height of a bar represents the frequency rather than the value of a variable. (T/F)
9. The wider the class interval, the more specific information is lost about the actual data. (T/F)
10. We cannot construct frequency distribution tables for nominally or ordinally scaled data. (T/F)
11. The frequency distribution represents data in a compressed form. (T/F)
12. The classes in any frequency distribution are all-inclusive and mutually exclusive. (T/F)
13. A frequency polygon can always be used to construct a histogram. (T/F)
14. A histogram shows each separate class in the distribution more clearly than a frequency polygon. (T/F)
15. The data array does not allow us to locate the highest and lowest values in the data set. (T/F)

Multiple Choice

16. Which of following methods is an accurate method of classifying data
 - (a) qualitative methods
 - (b) quantitative methods
 - (c) both (a) and (b)
 - (d) A method in accordance with the information available.
17. Which of following is not an example of compressed data
 - (a) data array
 - (b) frequency distribution
 - (c) histogram
 - (d) ogive
18. For constructing a frequency distribution, the first step is to
 - (a) arrange data into an array
 - (b) decide the type and number of classes
 - (c) decide the number of class intervals
 - (d) all of these
19. The upper limit of class intervals is considered for calculating the
 - (a) 'less than' cumulative frequency
 - (b) 'more than' cumulative frequency
 - (c) relative frequency
 - (d) none of these
20. The lower limit of class intervals is considered for calculating the
 - (a) 'less than' cumulative frequency
 - (b) 'more than' cumulative frequency
 - (c) relative frequency
 - (d) none of these
21. In inclusive class intervals of a frequency distribution
 - (a) upper limit of each class interval is included
 - (b) lower limit of each class interval is included
 - (c) both (a) and (b)
 - (d) none of these
22. In exclusive class intervals of a frequency distribution
 - (a) upper limit of each class interval is excluded
 - (b) lower limit of each class interval is excluded
 - (c) both (a) and (b)
 - (d) none of these

23. The number of classes in any frequency distribution depends upon
 (a) size of the data set
 (b) size of the population
 (c) range of observations in the data set
 (d) all of these
24. As a general rule, the number of classes in a frequency distribution should be
 (a) less than five
 (b) between five and fifteen
 (c) between fifteen and twenty
 (d) more than twenty
25. Various types of graphs of frequency distributions are constructed because they
 (a) reveal data patterns
 (b) allow easy estimates of values
 (c) increase the possibility of practical applications
 (d) both (a) and (b) but not (c)
26. Frequencies in a relative frequency distribution are represented in terms of
 (a) fractions (b) percentages
 (c) both (a) and (b) (d) whole numbers
27. As the number of observations and classes increase, the shape of the frequency polygon
 (a) remains unchanged
 (b) tend to become jagged
 (c) tend to become smooth
 (d) none of these
28. A graph of a cumulative frequency distribution is called
 (a) ogive (b) frequency polygon
 (c) frequency curve (d) pie diagram
29. If a data can take on only a limited number of values, the classes of these data are called
 (a) discrete (b) continuous
 (c) inclusive (d) exclusive
30. Classes in frequency distributions are all inclusive because
 (a) no observation falls into more than one class
 (b) all observations either fit into one class or another
 (c) both (a) and (b)
 (d) none of these

Concepts Quiz Answers

1. F	2. T	3. F	4. T	5. T	6. T	7. T	8. F	9. F
10. F	11. T	12. T	13. T	14. T	15. F	16. (d)	17. (a)	18. (d)
19. (a)	20. (b)	21. (c)	22. (a)	23. (d)	24. (b)	25. (d)	26. (c)	27. (c)
28. (a)	29. (b)	30. (b)						

Review Self-Practice Problems

- 2.39 If the price of a two-bed room flat in Gurgaon varies from Rs 9,00,000 to Rs 12,00,000, then
 (a) Indicate the class boundaries of 10 classes into which these values can be grouped
 (b) What class interval width did you choose?
 (c) What are the 10 class midpoints?
- 2.40 Students admitted to the MBA programme at the FMS were asked to indicate their preferred major area of specialization. The following data were obtained
- | Area of Specialization | Number of Students |
|------------------------|--------------------|
| Marketing | 50 |
| Finance | 22 |
| HRD | 08 |
| Operations | 10 |
- (a) Construct a relative and percentage frequency distribution.
 (b) Construct a bar chart and pie chart.
- 2.41 The raw data displayed here are the scores (out of 100 marks) of a market survey regarding the acceptability of a new product launch by a company for a random sample of 50 respondents
- | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|
| 40 | 45 | 41 | 45 | 45 | 30 | 39 | 8 | 48 |
| 25 | 26 | 9 | 23 | 24 | 26 | 29 | 8 | 40 |
| 41 | 42 | 39 | 35 | 18 | 25 | 35 | 40 | 42 |
| 43 | 44 | 36 | 27 | 32 | 28 | 27 | 25 | 26 |
| 38 | 37 | 36 | 35 | 32 | 28 | 40 | 41 | 43 |
| 44 | 45 | 40 | 39 | 41 | | | | |
- (a) Form a frequency distribution having 9 class intervals
 (b) Form a percentage distribution from the frequency distribution in part (a)
 (c) Form a histogram of the frequency distribution in part (a)
- 2.42 State whether each of the following variables is qualitative or quantitative and indicate the measurement scale that is appropriate for each:
 (a) Age (b) Gender
 (c) Class rank (d) Annual sales
 (e) Method of payment (f) Earnings per share
- 2.43 The following data represent the sales of car tyres of various brands by a retail showroom of tyres during the year 2001–02.

Brand of Tyre	Tyres Sold
Dunlop	136
Modi	221
Firestone	138
Ceat	84
Goodyear	101
JK	120

- (a) Construct a bar chart and pie chart.
 (b) Which of these charts do you prefer to use? Why?

2.44 The following data represent the expenditure incurred on following heads by a company during the year 2002

Expenditure Head	Amount (Rs in lakh)
Raw materials	1,689
Taxes	582
Manufacturing expenses	543
Employees salary	470
Depreciation	94
Dividend	75
Misc. expenses	286
Retained income	51

- (a) Construct a bar chart and pie chart.
 (b) Which of these charts do you prefer to use? Why?

2.45 Draw an ogive by less than method and determine the number of companies earning profits between Rs 45 crore and Rs 75 crore:

Profit (Rs in crore)	Number of Companies
10-20	8
20-30	12
30-40	20
40-50	24
50-60	15
60-70	10
70-80	7
80-90	3
90-100	1

[Delhi Univ., MBA, 1999]

2.46 The following data represent the hottest career options in marketing:

Career Option	Percentage
Product Manager	23
Market Research Executive	10
Direct Marketing Manager	20
Manager-Events and Productions	10
VP Marketing	16
Other Marketing Careers	21

Develop the appropriate display(s) and thoroughly analyse the data.

2.47 Software engineers at a software development company want to adopt a flexitime system beginning at 7.00, 7.30, 8.00, and 9.00 a.m. The following data represent a sample of the starting times by the engineers.

7.00 8.00 7.30 8.30 8.30 9.00 8.30 7.00 7.30 8.30
 7.30 7.30 8.00 8.00 9.00 8.00 8.30 8.30 7.00 8.30

Develop the appropriate display(s) of data and summarize your conclusions about preferences in the flexitime system.

2.48 The frequency distribution of GMAT scores from a sample of 50 applicants to an MBA course revealed that none of the applicants scored below 450 and that the table was formed by choosing class intervals $450 < 500$, $500 < 550$, and so on with the last class grouping $700 < 750$. If two applicants scored between $450 < 500$ and 16 applicants scored between $500 < 550$, then construct a percentage ogive and calculate the percentage of applicants scoring below 500, between 500 and 550, and below 750.

2.49 The data represent the closing prices of 40 common stocks.

29 34 43 8 37 8 7 30 35
 19 9 16 38 53 16 1 48 18
 9 9 10 37 18 8 28 24 21
 18 33 31 32 29 79 11 38 11
 52 14 9 33

- (a) Construct frequency and relative frequency distributions for the data.
 (b) Construct cumulative frequency and cumulative relative frequency distributions of the data.

2.50 An NGO working in environmental protection took water samples from twelve different places along the route of the Yamuna river from Delhi to Agra. These samples were tested in the laboratory and rated as to the amount of solid pollution suspended in each sample. The results of the testing are given below:

35.5 50.0 65.7 51.5 47.2 30.7 37.1
 49.0 57.0 43.4 35.8 46.4

- (a) Develop an appropriate display(s) of the data.
 (b) If the pollution rating of 42 (ppm) indicates excessive pollution, then how many samples would be rated as having excessive pollution?

2.51 The noise level of aircraft departing from an airport was rounded to the nearest integer value and grouped in a frequency distribution having marks at 90 and 120 decibels.

The noise level below 90 decibels is not considered serious, while any level above 130 decibels is very serious and almost deafening. If the residents of the airport area are raising this issue and bringing it to the notice of the government, then is this distribution adequate for their concern?

2.52 The distribution of disability adjusted life year (DALY) loss by certain causes in 1990 (in percentage) is given below:

Cause	India	China	World
• Communicable diseases	50.00	25.30	45.80
• Non-communicable diseases	40.40	58.00	42.80
• Injuries	9.10	16.70	12.00

Depict this data by pie chart and bar chart.

[Delhi Univ., MBA, 1999]

- 2.53** A government hospital has the following data representing weight in kg at birth of 200 premature babies:

Weight	Number of Babies
0.5–0.7	14
0.8–1.0	16
1.1–1.3	25
1.4–1.6	26
1.7–1.9	28
2.0–2.2	36
2.3–2.5	37
2.6–2.8	18

- Develop an appropriate display(s).
- Calculate the approximate middle value in the data set.
- If a baby below 2 kg is kept in the ICU as a precaution, then what percentage of premature babies need extra care in the ICU?

- 2.54** The medical superintendent of a hospital is concerned about the amount of waiting time for a patient before being treated in the OPD. The following data of waiting time (in minutes) were collected during a typical day:

Waiting Time	Number of Patients
30–40	125
40–50	195
50–60	305
60–70	185
70–80	120
80–90	70

- Use the data to construct 'more than' and 'less than' frequency distributions and ogive.
- Use the ogive to estimate how long 75 per cent of the patients should expect to wait.

- 2.55** A highway maintenance agency has ordered a study of the amount of time vehicles must wait at a toll gate of a recently constructed highway which is severely clogged and accident-prone in the morning. The following data were collected on the number of minutes that 950 vehicles waited in line a typical day.

Waiting Time	Number of Vehicles
1.00–1.39	30
1.40–1.79	42
1.80–2.19	75
2.20–2.59	110
2.60–2.99	120
3.00–3.39	130
3.40–3.79	108
3.80–4.19	94
4.20–4.59	85
4.60–4.99	78

Construct an ogive and determine what percentage of the vehicles had to wait more than three minutes in line?

Case Studies

Case 2.1: Housing Complex

The welfare committee of a large housing complex wants to understand the possibility of appointing private security guards at the entrance gate of the complex for 24-hour duty. There are 810 flats in the housing complex, and the owners were asked to vote for or against the proposal. The following data were collected:

Should the guards be appointed	
Yes	194
No	121
Not sure	73
No response	422

Questions for Discussion

- Convert the data to percentages and construct (i) a bar chart and (ii) a pie chart. Which of these charts do you prefer to use? Why?
- Eliminating the 'no response' group, convert the remaining 388 responses to percentages and again construct bar and pie charts.

Suppose you have been designated as poll officer, based on your analysis of the data what would you like to suggest to the president of the welfare committee?

Case 2.2: Portfolio Management

A portfolio manager keeps a close watch on price-earnings ratios (defined as current market price divided by earnings for the most recent four quarters) of 200 common stocks. He reasons that, when the majority of stocks in a representative sample have low price-earnings (P-E) ratios by historical standards, it is time to become an aggressive buyer. Low P-Es may mean that investors in general are unrealistically pessimistic. Moreover stocks with low P-Es can benefit in a two-fold way when earnings increase: (a) higher earnings multiplied by a constant P-E ratio means a higher market price and (b) rising earnings are usually accompanied by rising P-E ratios.

Price-Earning Ratios for 200 Common Stocks

11.1	12.6	26.7	5.2	8.3	5.5	6.8	7.6
7.3	18.1	14.6	10.9	7.2	9.5	9.2	11.8
12.0	16.9	10.1	14.6	5.2	7.5	11.1	19.9
14.9	7.4	6.0	39.9	29.3	35.1	6.8	39.0
6.1	6.2	26.8	33.7	9.6	16.6	10.9	11.2
22.6	46.0	7.3	29.7	10.3	6.4	9.6	7.6
10.3	5.0	14.4	11.6	8.3	7.9	17.8	7.5
7.8	7.3	8.0	20.2	5.6	8.3	7.7	10.7
8.6	14.5	6.0	5.4	12.6	14.8	9.2	14.1
15.7	10.4	7.0	11.0	6.3	8.4	7.6	16.9
7.9	8.3	13.1	9.8	8.2	18.0	26.6	7.8
4.1	10.6	15.3	7.2	35.5	6.1	10.2	6.1
7.8	8.1	30.0	15.0	6.1	15.4	10.1	9.6
6.8	4.4	6.8	9.1	16.3	5.4	5.9	6.5
7.9	44.9	13.8	12.3	10.9	9.3	11.9	10.0
7.6	17.9	7.1	8.4	35.5	7.4	7.7	8.3
15.8	8.3	23.1	8.4	12.4	7.8	8.2	9.8
13.7	15.8	4.7	7.9	26.4	6.2	11.4	13.2
8.6	11.7	8.6	13.7	9.3	16.6	8.7	39.7
14.0	9.1	7.1	10.9	23.4	13.3	10.9	24.0
11.9	8.7	15.6	27.7	10.4	16.9	6.9	5.5
22.8	8.5	22.2	5.8	14.7	8.0	7.5	10.5
4.4	7.1	63.8	12.5	13.3	10.5	5.5	16.0
53.1	7.4	24.1	15.3	29.1	11.0	9.9	36.3
9.6	6.6	5.1	7.8	8.4	38.3	20.4	9.1

Questions for Discussion

- Organize the values of the variable into an array.
- Construct a frequency distribution table.
- Present the resulting frequency distribution as a histogram or frequency polygon and comment on the pattern.
- Construct a cumulative frequency distribution and ogive.

We can easily represent things as we wish them to be . . .

—Aesop

If at first you do not succeed, you are just about average.

—Bill Cosby

Measures of Central Tendency

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- understand the role of descriptive statistics in summarization, description and interpretation of the data.
- understand the importance of summary measures to describe characteristics of a data set.
- use several numerical methods belonging to measures of central tendency to describe the characteristics of a data set.

3.1 INTRODUCTION

In Chapter 2, we discussed how raw data can be organized in terms of tables, charts, and frequency distributions in order to be easily understood and analysed. Although frequency distributions and corresponding graphical representations make raw data more meaningful, yet they fail to identify three major properties that describe a set of quantitative data. These three major properties are:

1. The numerical value of an observation (also called *central value*) around which most numerical values of other observations in the data set show a tendency to cluster or group, called the *central tendency*.
2. The extent to which numerical values are dispersed around the central value, called *variation*.
3. The extent of departure of numerical values from symmetrical (normal) distribution around the central value, called *skewness*.

These three properties—*central tendency*, *variation*, and *shape* of the frequency distribution—may be used to extract and summarize major features of the data set by the application of certain statistical methods called *descriptive measures* or *summary measures*. There are three types of summary measures:

1. Measures of central tendency
2. Measures of dispersion or variation
3. Measure of symmetry—skewness

Population parameter: A numerical value used as a summary measure using data of the population.

Sample statistic: A numerical value used as a summary measure using data of the sample for estimation or hypothesis testing.

These measures can also be used for comparing two or more populations in terms of the properties mentioned in the previous page to draw useful inferences.

The term 'central tendency' was coined because observations (numerical values) in most data sets show a distinct tendency to group or cluster around a value of an observation located somewhere in the middle of all observations. It is necessary to identify or calculate this typical *central value* (also called *average*) to describe or project the characteristic of the entire data set. This descriptive value is the measure of the *central tendency* or *location* and methods of computing this central value are called *measures of central tendency*.

If the descriptive summary measures are computed using data of samples drawn from a population, then these are called **sample statistic** or simply *statistic* but if these measures are computed using data of the population, they are called **population parameters** or simply *parameters*. The population parameter is represented by the Greek letter μ (read : mu) and sample statistic is represented by the Roman letter \bar{x} (read : x bar).

3.2 OBJECTIVES OF AVERAGING

A few of the objectives to calculate a typical central value or average in order to describe the entire data set are given below:

1. It is useful to extract and summarize the characteristics of the entire data set in a precise form. For example, it is difficult to understand individual families' need for water during summers. Therefore knowledge of the average quantity of water needed for the entire population will help the government in planning for water resources.
2. Since an 'average' represents the entire data set, it facilitates comparison between two or more data sets. Such comparison can be made either at a point of time or over a period of time. For example, average sales figures of any month can be compared with the preceding months, or even with the sales figures of competitive firms for the same months.
3. It offers a base for computing various other measures such as dispersion, skewness, kurtosis that help in many other phases of statistical analysis.

3.3 REQUISITES OF A MEASURE OF CENTRAL TENDENCY

The following are the few requirements to be satisfied by an average or a measure of central tendency:

1. **It should be rigidly defined** The definition of an average should be clear and rigid so that there must be uniformity in its interpretation by different decision-makers or investigators. There should not be any chance for applying discretion; rather it should be defined by an algebraic formula.
2. **It should be based on all the observations** To ensure that it should represent the entire data set, its value should be calculated by taking into consideration the entire data set.
3. **It should be easy to understand and calculate** The value of an average should be computed by using a simple method without reducing its accuracy and other advantages.
4. **It should have sampling stability** The value of an average calculated from various independent random samples of the same size from a given population should not vary much from another. The least amount of difference (if any) in the values is considered to be the sampling error.
5. **It should be capable of further algebraic treatment** The nature of the average should be such that it could be used for statistical analysis of the data set. For example, it should be possible to determine the average production in a particular year by the use of average production in each month of that year.

6. **It should not be unduly affected by extreme observations** The value of an average should not unduly be affected by very small or very large observations in the given data. Otherwise the average value may not truly represent characteristics of the entire set of data.

3.4 MEASURES OF CENTRAL TENDENCY

The various measures of central tendency or averages commonly used can be broadly classified in the following categories:

1. **Mathematical Averages**
 - (a) Arithmetic Mean commonly called the mean or average
 - Simple
 - Weighted
 - (b) Geometric Mean
 - (c) Harmonic Mean
2. **Averages of Position**
 - (a) Median
 - (b) Quartiles
 - (c) Deciles
 - (d) Percentiles
 - (e) Mode

Notations

- m_i = mid-point for the i th class in the data set
 f_i = number of observations (or frequency) in the i th class; ($i = 1, 2, \dots, N$)
 N = total number of observations in the population
 n = number of observations in the sample (sample size)
 l = lower limit of any class-interval
 h = width (or size) of the class-interval
 cf = cumulative frequency
 Σ = summation (read: sigma) of all values of observations

3.5 MATHEMATICAL AVERAGES

Various methods of calculating mathematical averages of a data set are classified in accordance of the nature of data available, that is, ungrouped (unclassified or raw) or grouped (classified) data.

3.5.1 Arithmetic Mean of Ungrouped (or Raw) Data

There are two methods for calculating **arithmetic mean (A.M.)** for ungrouped or unclassified data:

- (i) Direct method, and
- (ii) Indirect or Short-cut method.

Direct Method

In this method A.M. is calculated by adding the values of all observations and dividing the total by the number of observations. Thus if x_1, x_2, \dots, x_N represent the values of N observations, then A.M. for a population of N observations is:

$$\text{Population mean, } \mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3-1a)$$

Mean: The sum of all the data values divided by their number.

However, for a sample containing n observations x_1, x_2, \dots, x_n , the sample A.M. can be written as:

$$\text{Sample mean, } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.1b)$$

The denominator of above two formulae is different because in statistical analysis the uppercase letter N is used to indicate the number of observations in the population, while the lower case letter n is used to indicate the number of observations in the sample.

Example 3.1: In a survey of 5 cement companies, the profit (in Rs lakh) earned during a year was 15, 20, 10, 35, and 32. Find the arithmetic mean of the profit earned.

Solution: Applying the formula (3-1b), we have

$$\bar{x} = \frac{1}{n} \sum_{i=1}^5 x_i = \frac{1}{5} [15 + 20 + 10 + 35 + 32] = 22.4$$

Thus the arithmetic mean of the profit earned by these companies during a year was Rs 22.4 lakh.

Alternative Formula

In general, when observations x_i ($i = 1, 2, \dots, n$) are grouped as a frequency distribution, then A.M. formula (3-1b) should be modified as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i x_i \quad (3-2)$$

where f_i represents the frequency (number of observations) with which variable x_i occurs

in the given data set, i.e. $n = \sum_{i=1}^n f_i$.

Example 3.2: If A, B, C, and D are four chemicals costing Rs 15, Rs 12, Rs 8 and Rs 5 per 100 g, respectively, and are contained in a given compound in the ratio of 1, 2, 3, and 4 parts, respectively, then what should be the price of the resultant compound.

Solution: Using the formula (3-2), the sample arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^4 f_i x_i = \frac{1 \times 15 + 2 \times 12 + 3 \times 8 + 4 \times 5}{1 + 2 + 3 + 4} = \text{Rs } 8.30$$

Thus the average price of the resultant compound should be Rs 8.30 per 100 g.

Example 3.3: The number of new orders received by a company over the last 25 working days were recorded as follows: 3, 0, 1, 4, 4, 4, 2, 5, 3, 6, 4, 5, 1, 4, 2, 3, 0, 2, 0, 5, 4, 2, 3, 3, 1. Calculate the arithmetic mean for the number of orders received over all similar working days.

Solution: Applying the formula (3-1b), the arithmetic mean is:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{25} x_i = \frac{1}{25} [3 + 0 + 1 + 4 + 4 + 4 + 2 + 5 + 3 + 6 + 4 \\ &\quad + 5 + 1 + 4 + 2 + 3 + 0 + 2 + 0 + 5 + 4 + 2 + 3 + 3 + 1] \\ &= \frac{1}{25} (71) = 2.84 \cong 3 \text{ orders (approx.)} \end{aligned}$$

Alternative approach: Use of formula (3-2)

Table 3.1 Calculations of Mean (\bar{x}) Value

Number of Orders (x_i)	Frequency (f_i)	$f_i x_i$
0	13	10
1	13	13
2	14	18
3	15	15
4	16	24
5	13	15
6	1	6
	25	71

$$\text{Arithmetic mean, } \bar{x} = \frac{1}{n} \sum f_i x_i = \frac{71}{25} = 2.8 \cong 3 \text{ orders (approx.)}$$

Example 3.4: From the following information on the number of defective components in 1000 boxes;

Number of defective components :	0	1	2	3	4	5	6
Number of boxes :	25	306	402	200	51	10	6

Calculate the arithmetic mean of defective components for the whole of the production line.

Solution: The calculations of mean defective components for the whole production line are shown in Table 3.2

Table 3.2 Calculations of \bar{x} for Ungrouped Data

Number of Defective Components (x_i)	Number of Boxes (f_i)	$f_i x_i$
0	25	0
1	306	306
2	402	804
3	200	600
4	51	204
5	10	50
6	6	36
	<u>1000</u>	<u>2000</u>

Applying the formula (3-2), the arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=0}^6 f_i x_i = \frac{1}{1000} (2000) = 2 \text{ defective components.}$$

Short-Cut Method (Ungrouped Data)

In this method an arbitrary *assumed mean* is used as a basis for calculating deviations from individual values in the data set. Let A be the arbitrary assumed A.M. and let

$$d_i = x_i - A \quad \text{or} \quad x_i = A + d_i$$

Substituting the value of x_i in formula (3-1b), we have

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n (A + d_i) = A + \frac{1}{n} \sum_{i=1}^n d_i \quad (3-3)$$

If frequencies of the numerical values are also taken into consideration, then the formula (3-3) becomes:

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^n f_i d_i \quad (3-4)$$

where $n = \sum_{i=1}^n f_i$ = total number of observations in the sample.

Example 3.5: The daily earnings (in rupees) of employees working on a daily basis in a firm are:

Daily earnings (Rs) :	100	120	140	160	180	200	220
Number of employees :	3	6	10	15	24	42	75

Calculate the average daily earning for all employees.

Solution: The calculations of average daily earning for employees are shown in Table 3.3.

Table 3.3 Calculations of \bar{x} for Ungrouped Data

Daily Earnings (in Rs) (x_i)	Number of Employees (f_i)	$d_i = x_i - A$ $= x_i - 160$	$f_i d_i$
100	3	-60	-180
120	6	-40	-240
140	10	-20	-200
160 ← A	15	0	0
180	24	20	480
200	42	40	1680
220	75	60	4500
	175		6040

Here A = 160 is taken as assumed mean. The required A.M. using the formula (3-4) is given by

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^7 f_i d_i = 160 + \frac{6040}{175} = \text{Rs } 194.51$$

Example 3.6: The human resource manager at a city hospital began a study of the overtime hours of the registered nurses. Fifteen nurses were selected at random, and following overtime hours during a month were recorded:

13 13 12 15 7 15 5 12 6 7 12 10 9 13 12
5 9 6 10 5 6 9 6 9 12

Compute the arithmetic mean of overtime hours during the month.

Solution: Calculations of arithmetic mean of overtime hours are shown in Table 3.4

Table 3.4 Calculations of \bar{x} for Ungrouped Data

Overtime Hours (x_i)	Number of Number (f_i)	$d_i = x_i - A$ $= x_i - 10$	$f_i d_i$
5	3	-5	-15
6	4	-4	-16
7	2	-3	-6
9	4	-1	-4
10 ← A	2	0	0
12	5	2	10
13	3	3	9
15	2	5	10
	25		-12

Here A=10 is taken as assumed mean. The required arithmetic mean of overtime using the formula (3-4) is as follows:

$$\bar{x} = A + \frac{1}{n} \sum_{i=1}^{25} f_i d_i = 10 - \frac{12}{25} = 9.52 \text{ hours}$$

3.5.2 Arithmetic Mean of Grouped (or Classified) Data

Arithmetic mean for grouped data can also be calculated by applying any of the following methods:

- (i) Direct method, and
- (ii) Indirect or Step-deviation method

For calculating arithmetic mean for a grouped data set, the following assumptions are made:

- (i) The class intervals must be closed
- (ii) The width of each class interval should be equal
- (iii) The values of the observations in each class interval must be uniformly distributed between its lower and upper limits.
- (iv) The mid-value of each class interval must represent the average of all values in that class, that is, it is assumed that all values of observations are evenly distributed between the lower and upper class limits.

Direct Method

The formula used in this method is same as formula (3-2) except that x_i is replaced with the mid-point value m_i of class intervals. The new formula becomes:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n f_i m_i \quad (3-5)$$

where m_i = mid-value of i th class interval.

f_i = frequency of i th class interval.

$n = \sum f_i$, sum of all frequencies

Mean value: A measure of central location (tendency) for a data set such that the observations in the data set tend to cluster around it.

Example 3.7: A company is planning to improve plant safety. For this, accident data for the last 50 weeks was compiled. These data are grouped into the frequency distribution as shown below. Calculate the A.M. of the number of accidents per week.

Number of accidents :	0-4	5-9	10-14	15-19	20-24
Number of weeks :	5	22	13	8	2

Solution: The calculations of A.M. are shown in Table 3.5 using formula (3-5).

Table 3.5 Arithmetic Mean of Accidents

Number of Accidents	Mid-value (m_i)	Number of Weeks (f_i)	$f_i m_i$
0-4	2	5	10
5-9	7	22	154
10-14	12	13	156
15-19	17	8	136
20-24	22	2	44
		50	500

The A.M. of the number of accidents per week is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^5 f_i m_i = \frac{500}{50} = 10 \text{ accidents per week.}$$

Step-deviation Method

The formula (3-5) for calculating A.M. can be improved as formula (3-6). This improved formula is also known as the *step-deviation method*:

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h \quad (3-6)$$

where A = assumed value for the A.M.

$n = \sum f_i$, sum of all frequencies

h = width of the class intervals

m_i = mid-value of i th class-interval

$d_i = \frac{m_i - A}{h}$, deviation from the assumed mean

The formula (3-6) is very useful in those cases where mid-values (m_i) and/or frequencies (f_i) are in three or more digits. The calculation of d_i from m_i involves reducing each m_i by an amount A (called assumed arithmetic mean) and then dividing the reduced values by h (width of class-intervals). This procedure is usually referred to as *change of location and scale or coding*.

Example 3.8: Calculate the arithmetic mean of accidents per week by the short cut method using the data of Example 3.7.

Solution: The calculations of the average number of accidents are shown in the Table 3.6.

Table 3.6 Arithmetic Mean of Accidents

Number of Accidents	Mid-value (m_i)	$d_i = (m_i - A)/h = (m_i - 12)/5$	Number of Weeks (f_i)	$f_i d_i$
0-14	2	-2	5	-10
5-19	7	-1	22	-22
10-14	12 ← A	0	13	0
15-19	17	1	8	8
20-24	22	2	2	4
			50	-20

$$\begin{aligned} \text{The arithmetic mean } \bar{x} &= A + \left\{ \frac{1}{n} \sum f_i d_i \right\} h \\ &= 12 + \left\{ \frac{1}{50} (-20) \right\} 5 = 10 \text{ accidents per week} \end{aligned}$$

Example 3.9: The following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate the average overtime work done per employee.

Overtime hours :	10-15	15-20	20-25	25-30	30-35	35-40
Number of employees :	11	20	35	20	8	6

Solution: The calculations of the average overtime work done per employee with assumed mean, $A = 22.5$ and $h = 5$ are given in Table 3.7.

Table 3.7 Calculations of Average Overtime

Overtime (hrs) x_i	Number of Employees, f_i	Mid-value (m_i)	$d_i = (m_i - 22.5)/5$	$f_i d_i$
10-15	11	12.5	-2	-22
15-20	20	17.5	-1	-20
20-25	35	22.5 ← A	0	0
25-30	20	27.5	1	20
30-35	8	32.5	2	16
35-40	6	37.5	3	18
	100			12

$$\text{The required A.M. is, } \bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 22.5 + \frac{12}{100} \times 5 = 23.1 \text{ hrs}$$

Example 3.10: The following is the age distribution of 1000 persons working in an organization

Age Group	Number of Persons	Age Group	Number of Persons
20-25	30	45-50	105
25-30	160	50-55	70
30-35	210	55-60	60
35-40	180	60-65	40
40-45	145		

Due to continuous losses, it is desired to bring down the manpower strength to 30 per cent of the present number according to the following scheme:

- Retrench the first 15 per cent from the lower age group.
- Absorb the next 45 per cent in other branches.
- Make 10 per cent from the highest age group retire permanently, if necessary.

Calculate the age limits of the persons retained and those to be transferred to other departments. Also find the average age of those retained. [Delhi Univ., MBA; 2003]

Solution: (a) The first 15 per cent persons to be retrenched from the lower age groups are $(15/100) \times 1000 = 150$. But the lowest age group 20–25 has only 30 persons and therefore the remaining, $150 - 30 = 120$ will be taken from next higher age group, that is, 25–30, which has 160 persons.

(b) The next 45 per cent, that is, $(45/100) \times 1000 = 450$ persons who are to be absorbed in other branches, belong to the following age groups:

Age Groups	Number of Persons
25–30	$(160 - 120) = 40$
30–35	210
35–40	180
40–45	$(450 - 40 - 210 - 180) = 20$

(c) Those who are likely to be retired are 10 per cent, that is, $(10/100) \times 1000 = 100$ persons and belong to the following highest age groups:

Age Group	Number of Persons
55–60	$(100 - 40) = 60$
60–65	40

Hence, the calculations of the average age of those retained and/or to be transferred to other departments are shown in Table 3.8:

Table 3.8 Calculations of Average Age

Age Group (x_i)	Mid value, (m_i)	Number of Persons (f_i)	$d_i = (x_i - 47.5)/5$	$f_i d_i$
40–45	42.5	$145 - 20 = 125$	-1	-125
45–50	47.5 ← A	105	0	0
50–55	52.5	70	1	70
		300		-55

The required average age is, $\bar{x} = A + \frac{\sum d_i f_i}{n} \times h = 47.5 - \frac{55}{300} \times 5 = 46.58 = 47$ years (approx.).

3.5.3 Some Special Types of Problems and Their Solutions

Case 1: Frequencies are Given in Cumulative Form, that is, either 'More Than Type' or 'Less Than Type'

As we know that the 'more than type' cumulative frequencies are calculated by adding frequencies from bottom to top, so that the first class interval has the highest cumulative frequency and it goes on decreasing in subsequent classes. But in case of 'less than cumulative frequencies', the cumulation is done downward so that the first class interval has the lowest cumulative frequency and it goes on increasing in the subsequent classes.

In both of these cases, data are first converted into inclusive class intervals or exclusive class intervals. Then the calculations for \bar{x} are done in the usual manner as discussed earlier.

Example 3.11: Following is the cumulative frequency distribution of the preferred length of kitchen slabs obtained from the preference study on housewives:

Length (in metres) more than	:	1.0	1.5	2.0	2.5	3.0	3.5
Preference of housewives	:	50	48	42	40	10	5

A manufacturer has to take a decision on what length of slabs to manufacture. What length would you recommend and why?

Solution: The given data are converted into exclusive class intervals as shown in Table 3.9. The frequency of each class has been found out by deducting the given cumulative frequency from the cumulative frequency of the previous class:

Table 3.9 Conversion into Exclusive Class Intervals

Length (in metres)	Preference of Housewives more than	Class Interval	Frequency
1.0	50	1.0–1.5	(50 – 48) = 2
1.5	48	1.5–2.0	(48 – 42) = 6
2.0	42	2.0–2.5	(42 – 40) = 2
2.5	40	2.5–3.0	(40 – 10) = 30
3.0	10	3.0–3.5	(10 – 5) = 5
3.5	5		

The calculations for mean length of slab are shown in Table 3.10.

Table 3.10 Calculations of Mean Length of Slab

Class Interval	Mid-value (m_i)	Preference of Housewives (f_i)	$d_i = \frac{m_i - 2.25}{0.5}$	$f_i d_i$
1.0–1.5	1.25	2	-2	-4
1.5–2.0	1.75	6	-1	-6
2.0–2.5	2.25 ← A	2	0	0
2.5–3.0	2.75	30	1	30
3.0–3.5	3.25	5	2	10
		45		30

The mean length of the slab is $\bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 2.25 + \frac{30}{45} \times 0.5 = 2.58$

metres

Example 3.12: In an examination of 675 candidates, the examiner supplied the following information:

Marks Obtained (Percentage)	Number of Candidates	Marks Obtained (Percentage)	Number of Candidates
Less than 10	7	Less than 50	381
Less than 20	39	Less than 60	545
Less than 30	95	Less than 70	631
Less than 40	201	Less than 80	675

Calculate the mean percentage of marks obtained.

Solution: Arranging the given data into inclusive class intervals as shown in Table 3.11:

Table 3.11 Calculations of Mean Percentage of Marks

Marks Obtained (Percentage)	Cumulative Frequency	Class-intervals	Frequency
Less than 10	7	0-10	7
Less than 20	39	10-20	(39 - 7) = 32
Less than 30	95	20-30	(95 - 39) = 56
Less than 40	201	30-40	(201 - 95) = 106
Less than 50	381	40-50	(381 - 201) = 180
Less than 60	545	50-60	(545 - 381) = 164
Less than 70	631	60-70	(631 - 545) = 86
Less than 80	675	70-80	(675 - 631) = 44

The calculations for mean percentage of marks obtained by the candidates are shown in Table 3.12.

Table 3.12 Calculations of Mean Percentage of Marks

Class Intervals	Mid-value (m_i)	Number of Candidates (f_i)	$d_i = \frac{m_i - 35}{10}$	$f_i d_i$
0-10	5	7	-3	-21
10-20	15	32	-2	-64
20-30	25	56	-1	-56
30-40	35 ← A	106	0	0
40-50	45	180	1	180
50-60	55	164	2	328
60-70	65	86	3	258
70-80	75	44	4	176
		675		801

The mean percentage of marks obtained is:

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 35 + \frac{801}{675} \times 10 = 46.86 \text{ marks}$$

Case 2: Frequencies are not Given but have to be Calculated From the Given Data

Example 3.13: 168 handloom factories have the following distribution of average number of workers in various income groups:

Income groups	:	800-1000	1000-1200	1200-1400	1400-1600	1600-1800
Number of firms	:	40	32	26	28	42
Average number of workers	:	8	12	8	8	4

Find the mean salary paid to the workers.

Solution: Since the total number of workers (i.e. frequencies) working in different income groups are not given, therefore these have to be determined as shown in Table 3.13:

Table 3.13

Income Group (x_i) (1)	Mid-values (m_i) (2)	$d_i = \frac{m_i - A}{h}$ $= \frac{m_i - 1300}{200}$	Number of Firms (3)	Average Number of Workers (4)	Frequencies (f_i) (5) = (3) × (4)	$m_i f_i$
800-1000	900	-2	40	8	320	-640
1000-1200	1100	-1	32	12	384	-384
1200-1400	1300 ← A	0	26	8	208	0
1400-1600	1500	1	28	8	224	224
1600-1800	1700	2	42	4	168	336
			168	40	1304	-464

The required A.M. is given by

$$\bar{x} = A + \frac{\sum m_i f_i}{n} \times h = 1300 - \frac{464}{1304} \times 200 = 1228.84$$

Example 3.14: Find the missing frequencies in the following frequency distribution. The A.M. of the given data is 11.09.

Class	Frequency	Class	Frequency
9.3 - 9.7	2	11.3-11.7	14
9.8 - 10.2	5	11.8-12.2	6
10.3 - 10.7	f_3	12.3-12.7	3
10.8 - 11.2	f_4	12.8-13.2	1
			60

Solution: The calculations for A.M. are shown in Table 3.14.

Table 3.14

Class	Frequency (f_i)	Mid-value (m_i)	$d_i = \frac{m_i - 11.0}{0.5}$	$f_i d_i$
9.3 - 9.7	2	9.5	-3	-6
9.8 - 10.2	5	10.0	-2	-10
10.3 - 10.7	f_3	10.5	-1	$-f_3$
10.8 - 11.2	f_4	11.0 ← A	0	0
11.3 - 11.7	14	11.5	1	14
11.8 - 12.2	6	12.0	2	12
12.3 - 12.7	3	12.5	3	9
12.8 - 13.2	1	13.0	4	4
	60			23 - f_3

where the assumed mean is, $A = 11$. Applying the formula

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h$$

we get
$$11.09 = 11.0 + \frac{23 - f_3}{60} \times 0.5$$

or
$$0.09 = \frac{23 - f_3}{120} \quad \text{or} \quad f_3 = 23 - 0.09 \times 120 = 12.2$$

Since the total of the frequencies is 60 and $f_3 = 12.2$, therefore

$$f_4 = 60 - (2 + 5 + 12.2 + 14 + 6 + 3 + 1) = 16.8$$

Case 3: Complete Data are Not Given

Example 3.15: The pass result of 50 students who took a class test is given below:

Marks	:	40	50	60	70	80	90
Number of students	:	8	10	9	6	4	3

If the mean marks for all the students was 51.6 find out the mean marks of the students who failed.

Solution: The marks obtained by 40 students who passed are given in Table 3.15

Table 3.15

Marks	Frequency (f_i)	$f_i x_i$
40	8	320
50	10	500
60	9	540
70	6	420
80	4	320
90	3	270
	40	2370

Total marks of all the students = $50 \times 51.6 = 2580$

Total marks of 40 students who passed = $\sum f_i x_i = 2370$

Thus marks of the remaining 10 students = $2580 - 2370 = 210$

Hence, the average marks of 10 students who failed are $210/10 = 21$ marks

Case 4: Incorrect Values have been used for the Calculation of Arithmetic Mean

Example 3.16: (a) The average dividend declared by a group of 10 chemical companies was 18 per cent. Later on, it was discovered that one correct figure, 12, was misread as 22. Find the correct average dividend.

(b) The mean of 200 observations was 50. Later on, it was found that two observations were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

Solution: (a) Given $n = 10$ and $\bar{x} = 18$ per cent. We know that

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \sum x = n\bar{x} = 10 \times 18 = 180$$

Since one numerical value 12 was misread as 22, therefore after subtracting the incorrect value and then adding the correct value in the total $n\bar{x}$, we have $180 - 22 + 12 = 170$. Hence, correct mean is $\bar{x} = \sum x/n = 170/10 = 17$ per cent.

(b) Given that $n = 200$, $\bar{x} = 50$. We know that

$$\bar{x} = \frac{\sum x}{n} \quad \text{or} \quad \sum x = n\bar{x} = 200 \times 50 = 10,000$$

Since two observations were misread, therefore the correct total $\sum x = n\bar{x}$ can be obtained as:

$$\sum x = 10,000 - (92 + 8) + (192 + 88) = 10,180$$

$$\text{Hence, correct mean is : } \bar{x} = \frac{\sum x}{n} = \frac{10,180}{200} = 50.9$$

Case 5: Frequency Distributions have Open End Class Intervals

Example 3.17: The annual salaries (in rupees thousands) of employees in an organization are given below: The total salary of 10 employees in the class over Rs 40,000 is Rs 9,00,000. Compute the mean salary. Every employee belonging to the top 25 per cent of earners has to pay 5 per cent of his salary to the worker relief fund. Estimate the contribution to this fund.

Salary (Rs '000)	Number of Employees
below 10	4
10-20	6
20-30	10
30-40	20
40 and above	10

Solution: Since class intervals are uniform, therefore we can take some width for open-end class intervals also. Calculations of mean are shown in Table 3.16

Table 3.16

Salary (Rs '000)	Mid-value (m_i)	Number of Employees (f_i)	$d_i = \frac{m_i - 25}{10}$	$f_i d_i$
0-10	5	4	-2	-8
10-20	15	6	-1	-6
20-30	25 ← A	10	0	0
30-40	35	20	1	20
40 and above	45 (given)	10	2	20
		50		26

where mid-value 25 is considered as the assumed mean. Applying the formula, we get

$$\bar{x} = A + \frac{\sum f_i d_i}{n} \times h = 25 + \frac{26}{50} \times 10 = \text{Rs } 30.2$$

The number of employees belonging to the top 25 per cent of the earners are $0.25 \times 50 = 13$ employees and the distribution of these top earners would be as follows:

Salary (Rs '000)	Number of Employees
40 and above	10
30-40	3

This calculation implies that 3 employees have been selected from the salary range 30 - 40. Under the assumption that frequencies are equally distributed between lower and upper limits of a class interval, the calculations would be as follows:

Since 20 employees have salary in the range 30 - 40 = 10 or Rs 10,000, therefore 3 employees will have income in the range $(10/20) \times 3 = 1.5$ or Rs 1,500. But we are interested in the top 3 earners in the range 30 - 40, their salaries will range from $(40 - 1.5)$ to 40, i.e., 38.5 to 40. Thus, the distribution of salaries of the top 25 persons is as follows:

Salary (Rs '000)	Mid-value (m_i)	Number of Employees (f_i)	Total Salary $m_i f_i$
40 and above	—	10	9,00,000 (given)
30-40	35	3	1,05,000
		13	10,05,000

This shows that the total income of the top 25 per cent of earners is Rs 10,05,000. Hence 5 per cent contribution to the fund is $0.05 \times 10,05,000 = \text{Rs } 50,250$.

Remark: If the width of class intervals is not same, then in accordance with the magnitude of change in the width, fix the width of last class interval.

3.5.4 Advantages and Disadvantages of Arithmetic Mean

Advantages

- The calculation of arithmetic mean is simple and it is unique, that is, every data set has one and only one mean.

- (ii) The calculation of arithmetic mean is based on all values given in the data set.
- (iii) The arithmetic mean is reliable single value that reflects all values in the data set.
- (iv) The arithmetic mean is least affected by fluctuations in the sample size. In other words, its value, determined from various samples drawn from a population, vary by the least possible amount.
- (v) It can be readily put to algebraic treatment. Some of the algebraic properties of arithmetic mean are as follows:

(a) *The algebraic sum of deviations of all the observations x_i ($i = 1, 2, \dots, n$) from the A.M. is always zero, that is,*

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = \sum_{i=1}^n x_i - n\left(\frac{1}{n}\right) \sum_{i=1}^n x_i = 0$$

Here the difference $x_i - \bar{x}$ ($i = 1, 2, \dots, n$) is usually referred to as *deviation from the arithmetic mean*. This result is also true for a grouped data.

Due to this property, the mean is characterized as a *point of balance*, i.e. sum of the positive deviations from mean is equal to the sum of the negative deviations from mean.

(b) *The sum of the squares of the deviations of all the observations from the A.M. is less than the sum of the squares of all the observations from any other quantity.*

Let x_i ($i = 1, 2, \dots, n$) be the given observations and \bar{x} be their arithmetic mean. Then this property implies that

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$$

where 'a' is any constant quantity.

This property of A.M. is also known as the *least square property* and shall be quite helpful in defining the concept of standard deviation.

(c) *It is possible to calculate the combined (or pooled) arithmetic mean of two or more than two sets of data of the same nature.*

Let \bar{x}_1 and \bar{x}_2 be arithmetic means of two sets of data of the same nature of size n_1 and n_2 respectively. Then their *combined A.M.* can be calculated as:

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \quad (3-7)$$

The result (3-7) can also be generalized in the same way for more than two sets of data of different sizes having different arithmetic means.

(d) While compiling the data for calculating arithmetic mean, it is possible that we may wrongly read and/or write certain number of observations. In such a case the correct value of A.M. can be calculated first by subtracting the sum of observations wrongly recorded from Σx_i (total of all observations) and then adding the sum of the correct observations to it. The result is then divided by the total number of observations.

Disadvantages

- (i) The value of A.M. cannot be calculated accurately for unequal and open-ended class intervals either at the beginning or end of the given frequency distribution.
- (ii) The A.M. is reliable and reflects all the values in the data set. However, it is very much affected by the extreme observations (or outliers) which are not representative of the rest of the data set. Outliers at the high end will increase the mean, while outliers at the lower end will decrease it. For example, if monthly income of four persons is 50, 70, 80, and 1000, then their A.M. will be 300, which does not represent the data.
- (iii) The calculations of A.M. sometime become difficult because every data element is used in the calculation (unless the short cut method for grouped data is used to calculate the mean). Moreover the value so obtained may not be among the observations included in the data.

- (iv) The mean cannot be calculated for qualitative characteristics such as intelligence, honesty, beauty, or loyalty.
- (v) The mean cannot be calculated for a data set that has open-ended classes at either the high or low end of the scale.

Example 3.18: The mean salary paid to 1500 employees of an organization was found to be Rs 12,500. Later on, after disbursement of salary, it was discovered that the salary of two employees was wrongly entered as Rs 15,760 and 9590. Their correct salaries were Rs 17,760 and 8590. Calculate correct mean.

Solution: Let x_i ($i = 1, 2, \dots, 1500$) be the salary of i th employee. Then we are given that

$$\bar{x} = \frac{1}{1500} \sum_{i=1}^{1500} x_i = 12,500$$

or
$$\sum_{i=1}^{1500} x_i = 12,500 \times 1500 = \text{Rs } 1,87,50,000$$

This gives the total salary disbursed to all 1500 employees. Now after adding the correct salary figures of two employees and subtracting the wrong salary figures posted against two employees, we have

$$\begin{aligned} \sum x_i &= 1,87,50,000 + (\text{Sum of correct salaries figures}) \\ &\quad - (\text{Sum of wrong salaries figures}) \\ &= 1,87,50,000 + (17,760 + 8590) - (15,760 + 9590) \\ &= 1,87,50,000 + 26,350 - 25,350 = 1,88,01,700 \end{aligned}$$

Thus the correct mean salary is given by

$$\bar{x} = 1,88,01,700 \div 1500 = \text{Rs } 12,534.46$$

Example 3.19: There are two units of an automobile company in two different cities employing 760 and 800 persons, respectively. The arithmetic means of monthly salaries paid to persons in these two units are Rs 18,750 and Rs 16,950 respectively. Find the combined arithmetic mean of salaries of the employees in both the units.

Solution: Let n_1 and n_2 be the number of persons working in unit 1 and 2 respectively, and \bar{x}_1 and \bar{x}_2 be the arithmetic mean of salaries paid to these persons respectively. We are given that:

$$\text{Unit 1: } n_1 = 760 ; \bar{x}_1 = \text{Rs } 18,750$$

$$\text{Unit 2: } n_2 = 800 ; \bar{x}_2 = \text{Rs } 16,950$$

Thus the combined mean of salaries paid by the company is:

$$\begin{aligned} \bar{x}_{12} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{760 \times 18750 + 800 \times 16950}{760 + 800} \\ &= \text{Rs } 17,826.92 \text{ per month} \end{aligned}$$

Example 3.20: The mean yearly salary paid to all employees in a company was Rs 24,00,000. The mean yearly salaries paid to male and female employees were Rs 25,00,000 and Rs 19,00,000, respectively. Find out the percentage of male to female employees in the company.

Solution: Let n_1 and n_2 be the number of employees as male and female, respectively. We are given that

Characteristics	Groups		Combined Group (Total Employees)
	Male	Female	
Number of employees	$n_1 = ?$	$n_2 = ?$	$n = n_1 + n_2$
Mean salary (Rs)	$\bar{x}_1 = 25,00,000$	$\bar{x}_2 = 19,00,000$	$\bar{x}_{12} = 24,00,000$

Applying the formula for mean of combined group:

$$\bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

or

$$(n_1 + n_2) \bar{x}_{12} = n_1 \bar{x}_1 + n_2 \bar{x}_2$$

$$(n_1 + n_2) 24,00,000 = 25,00,000 n_1 + 19,00,000 n_2$$

$$1000 n_1 = 5000 n_2$$

$$\frac{n_1}{n_2} = \frac{5000}{1000} = \frac{5}{1}$$

$$n_1 : n_2 = 5 : 1$$

Hence, male employees in the company are $\{5 \div (5 + 1)\} \times 100 = 83.33$ per cent and female employees are $\{1 \div (5 + 1)\} \times 100 = 16.67$ per cent.

3.5.5 Weighted Arithmetic Mean

The arithmetic mean, as discussed earlier, gives equal important (or weight) to each observation in the data set. However, there are situations in which value of individual observations in the data set is not of equal importance. If values occur with different frequencies, then computing A.M. of values (as opposed to the A.M. of observations) may not be truly representative of the data set characteristic and thus may be misleading. Under these circumstances, we may attach to each observation value a 'weight' w_1, w_2, \dots, w_N as an indicator of their importance perhaps because of size or importance and compute a weighted mean or average denoted by \bar{x}_w as:

$$\mu_w \text{ or } \bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

This is similar to the method for dealing with frequency data when the value is multiplied by the frequency, within each class, totalled and divided by the total number of values.

Remark: The **weighted arithmetic mean** should be used

- (i) when the importance of all the numerical values in the given data set is not equal.
- (ii) when the frequencies of various classes are widely varying
- (iii) where there is a change either in the proportion of numerical values or in the proportion of their frequencies.
- (iv) when ratios, percentages, or rates are being averaged.

Weighted mean: The mean for a data set obtained by assigning each observation a weight that reflects its importance within the data set.

Example 3.21: An examination was held to decide the award of a scholarship. The weights of various subjects were different. The marks obtained by 3 candidates (out of 100 in each subject) are given below:

Subject	Weight	Students		
		A	B	C
Mathematics	4	60	57	62
Physics	3	62	61	67
Chemistry	2	55	53	60
English	1	67	77	49

Calculate the weighted A.M. to award the scholarship.

Solution: The calculations of the weighted mean is shown in Table 3.17

Table 3.17 Calculations of Weighted Mean

Subject	Weight (w_i)	Students					
		Student A		Student B		Student C	
		Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$	Marks (x_i)	$x_i w_i$
Mathematics	4	60	240	57	228	62	248
Physics	3	62	186	61	183	67	201
Chemistry	2	55	110	53	106	60	120
English	1	67	67	77	77	49	49
	10	244	603	248	594	238	618

Applying the formula for weighted mean, we get

$$\bar{x}_{wA} = \frac{603}{10} = 60.3 ; \quad \bar{x}_A = \frac{244}{4} = 61$$

$$\bar{x}_{wB} = \frac{594}{10} = 59.4 ; \quad \bar{x}_B = \frac{248}{4} = 62$$

$$\bar{x}_{wC} = \frac{618}{10} = 61.8 ; \quad \bar{x}_C = \bar{x}_C = 59.5$$

From the above calculations, it may be noted that student B should get the scholarship as per simple A.M. values, but according to weighted A.M., student C should get the scholarship because all the subjects of examination are not of equal importance.

Example 3.22: The owner of a general store was interested in knowing the mean contribution (sales price minus variable cost) of his stock of 5 items. The data is given below:

Product	Contribution per Unit	Quantity Sold
1	6	160
2	11	60
3	8	260
4	4	460
5	14	110

Solution: If the owner ignores the values of the individual products and gives equal importance to each product, then the mean contribution per unit sold will be

$$\bar{x} = (1 \div 5) \{6 + 11 + 8 + 4 + 14\} = \text{Rs } 8.6$$

This value, Rs 8.60 may not necessarily be the mean contribution per unit of different quantities of the products sold. In this case the owner has to take into consideration the number of units of each product sold as different weights. Computing weighted A.M. by multiplying units sold (w) of a product by its contribution (x). That is,

$$\bar{x}_w = \frac{6(160) + 11(60) + 8(260) + 4(460) + 14(110)}{160 + 60 + 260 + 460 + 110} = \frac{7,080}{1,050} = \text{Rs } 6.74$$

This value, Rs 6.74, is different from the earlier value, Rs 8.60. The owner must use the value Rs 6.74 for decision-making purpose.

Example 3.23: A management consulting firm, has four types of professionals on its staff: managing consultants, senior associates, field staff, and office staff. Average rates charged to consulting clients for the work of each of these professional categories are Rs 3150/hour, Rs 1680/hour, Rs 1260/hour, and 630/hour. Office records indicate the following number of hours billed last year in each category: 8000, 14,000, 24,000, and 35,000. If the firm is trying to come up with an average billing rate for estimating client charges for next year, what would you suggest they do and what do you think is an appropriate rate?

Solution: The data given in the problem are as follows:

Staff	Consulting Charges (Rs per hour) x_i	Hours Billed w_i
Managing consultants	3150	8000
Senior associates	1680	14,000
Field staff	1260	24,000
Office staff	630	35,000

Applying the formula for weighted mean, we get,

$$\begin{aligned}\bar{x}_w &= \frac{\sum x_i w_i}{\sum w_i} = \frac{3150(8000) + 1680(14,000) + 1260(24,000) + 630(35,000)}{8000 + 14,000 + 24,000 + 35,000} \\ &= \frac{2,52,00,000 + 2,35,20,000 + 3,02,40,000 + 2,20,50,000}{81,000} \\ &= \text{Rs } 1247.037 \text{ per hour}\end{aligned}$$

However, the firm should cite this as an average rate for clients who use the four professional categories for approximately 10 per cent, 17 per cent, 30 per cent and 43 per cent of the total hours billed.

Conceptual Questions 3A

1. Explain the term *average*. What are the merits of a good average? Explain with examples.
2. What are the measures of central tendency? Why are they called measures of central tendency?
3. What are the different measures of central tendency? Mention the advantages and disadvantages of arithmetic mean.
4. What are the different measures of central tendency? Discuss the essentials of an ideal average.
5. Give a brief description of the various measures of central tendency. Why is arithmetic mean so popular?
6. What information about a body of data is provided by an average? How are averages useful as a descriptive measure?
8. It is said that the weighted mean is commonly referred to as a 'weighted average'. How is the use of this phrase inconsistent with the definition of an average?
9. How is an average considered as a representative measure or a measure of central tendency? How is the ability of an average to measure central tendency related to other characteristics of data?
10. Prove that the algebraic sum of the deviations of a given set of observations from their arithmetic mean is zero.
[MBA, UP Tech. Univ., 2000]
11. Is it necessarily true that being above average indicates that someone is superior? Explain.
[Delhi Univ., MBA, 2000]
12. What is statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is the most commonly used amongst them.
13. Distinguish between simple and weighted average and state the circumstances under which the latter should be employed.

Self-Practice Problems 3A

- 3.1 An investor buys Rs 12,000 worth of shares of a company each month. During the first 5 months he bought the shares at a price of Rs 100, Rs 120, Rs 150, Rs 200, and Rs 240 per share. After 5 months what is the average price paid for the shares by him?
- 3.2 A company wants to pay bonus to members of the staff. The bonus is to be paid as under:

Monthly Salary (in Rs)	Bonus
3000 – 4000	1000
4000 – 5000	1200
5000 – 6000	1400
6000 – 7000	1600
7000 – 8000	1800
8000 – 9000	2200
9000 – 10,000	2200
10,000 – 11,000	2400

Actual amount of salary drawn by the employees is given below:

3250	3780	4200	4550	6200	6600
6800	7250	3630	8320	9420	9520
8000	10,020	10,280	11,000	6100	6250
7630	3820	5400	4630	5780	7230
6900					

How much would the company need to pay by way of bonus? What shall be the average bonus paid per member of the staff?

- 3.3** Calculate the simple and weighted arithmetic mean price per tonne of coal purchased by a company for the half year. Account for difference between the two:

Month	Price/tonne	Tonnes Purchased	Month	Price/tonne	Tonnes Purchased
January	4205	25	April	5200	52
February	5125	30	May	4425	10
March	5000	40	June	5400	45

- 3.4** Salary paid by a company to its employees is as follows:

Designation	Monthly Salary (in Rs)	Number of Persons
Senior Manager	35,000	1
Manager	30,000	20
Executives	25,000	70
Jr Executives	20,000	10
Supervisors	15,000	150

Calculate the simple and weighted arithmetic mean of salary paid.

- 3.5** The capital structure of a company is as follows:

	Book Value (Rs)	After Tax Cost (%)
Equity	2,15,00,000	19
Preference share	2,07,00,000	11
Debt	2,11,00,000	9

Calculate the weighted average cost of capital.

- 3.6** The mean monthly salary paid to all employees in a company is Rs 16,000. The mean monthly salaries paid to technical and non-technical employees are Rs 18,000 and Rs 12,000 respectively. Determine the percentage of technical and non-technical employees in the company.
- 3.7** The mean marks in statistics of 100 students in a class was 72 per cent. The mean marks of boys was 75 per cent, while their number was 70 per cent. Find out the mean marks of girls in the class.
- 3.8** Mr. Gupta, a readymade garment store owner, advertises: 'If our average prices are not equal or lower than everyone else's, you get it free.' One customer came into the store one day and threw on the counter bills of six items he had bought from a competitor for an average price less than Gupta's. The items cost (in Rs): 201.29 202.97 203.49 205.00 207.50 210.95

The prices for the same six items at Mr. Gupta's stores

are (in Rs): 201.35, 202.89, 203.19, 204.98, 207.59 and 211.50. Mr. Gupta told the customer, My advertisement refers to a weighted average price of these items. Our average is lower because our sales of these items have been: 207, 209, 212, 208, 206, and 203 (in units).

Is Mr. Gupta getting himself into or out of trouble with his contention about weighted average?

- 3.9** The arithmetic mean height of 50 students of a college is 5'8". The height of 30 of these is given in the frequency distribution below. Find the arithmetic mean height of the remaining 20 students.

Height in inches	: 5'4"	5'6"	5'8"	5'10"	6'0"
Frequency	: 4	12	4	8	2

- 3.10** The following table gives salary per month of 450 employees in a factory:

Salary	No. of Employees
Less than 5000	80
5000–10,000	120
10,000–15,000	100
15,000–20,000	60
20,000–25,000	50
25,000–30,000	40

The total income of 6 persons in the group Rs 25,000–30,000 is Rs 1,65,000. Due to a rise in prices, the factory owner decided to give adhoc increase of 25 per cent of the average pay to the 25 per cent of the lowest paid employees, 10 per cent of the average pay to the 10 per cent highest paid employees and 15 per cent to the remaining employees.

Find out the additional amount required for the adhoc increase and after the increase, find out the average pay of an employee in the factory.

- 3.11** A professor of management has decided to use weighted average to find the internal assessment grades of his students on the basis of following parameters: Quizzes—30 per cent, Term Paper—25 per cent, Mid-term test—30 per cent and Class attendance—15 per cent. From the data below, compute the final average in the internal assessment

Student	Quizzes	Term Paper	Mid-Term	Attendance
1	55	59	64	20
2	48	54	58	22
3	64	58	63	19
4	52	49	58	23
5	65	60	62	18

- 3.12** An appliances manufacturing company is forecasting regional sales for next year. The Delhi branch, with current yearly sales of Rs 387.6 million, is expected to achieve a sales growth of 7.25 per cent; the Kolkata branch, with current sales of Rs 158.6 million, is expected to grow by 8.20 per cent; and the Mumbai branch, with sales of Rs 115 million, is expected to increase sales by 7.15 per cent. What is the average rate of growth forecasted for next year?