CHAPTER 13

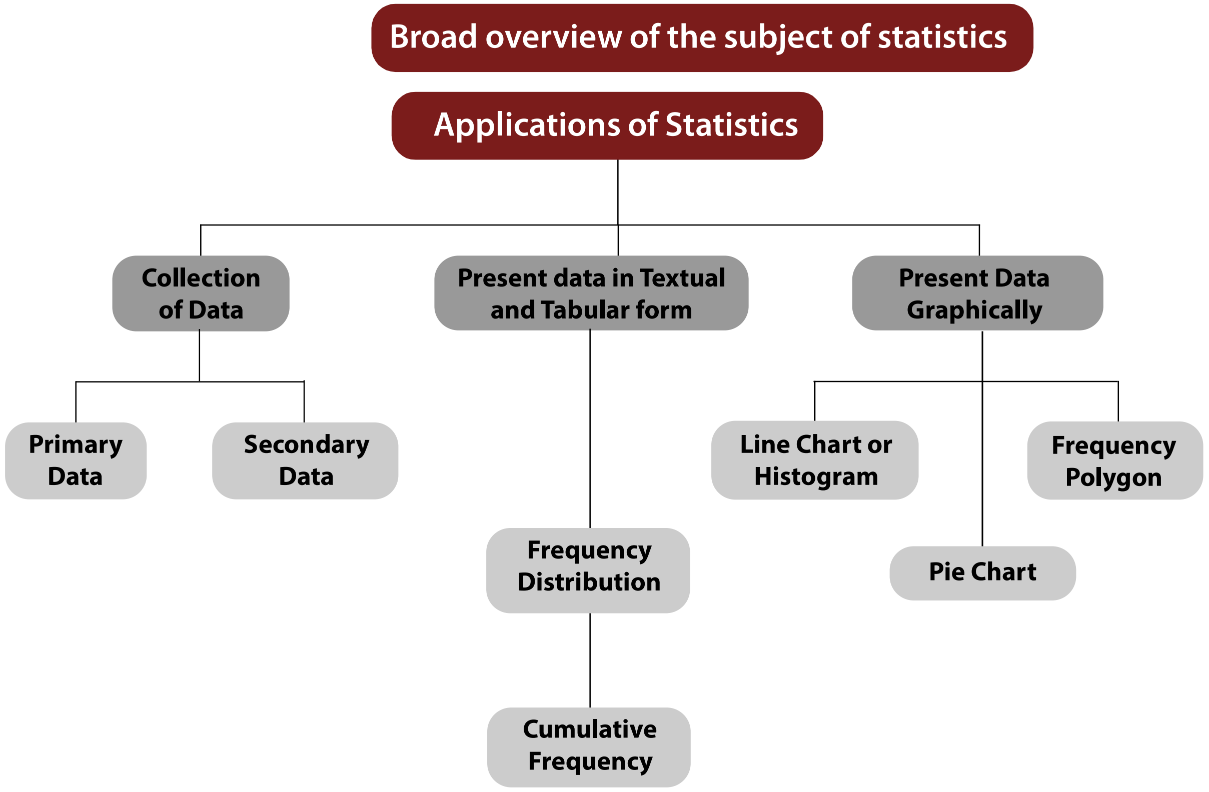
UNIT 1: STATISTICAL DESCRITPION OF DATA

LEARNING OBJECTIVES

After reading this chapter, students will be able to understand:

- Have a broad overview of the subject of statistics and application thereof.
- Know about data collection technique including the distinction of primary and secondary data.
- Know how to present data in textual and tabular format including the technique of creating frequency distribution and working out cumulative frequency.
- Know how to present data graphically using histogram, frequency polygon and pie chart.

UNIT OVERVIEW



13.1.1 INTRODUCTION OF STATISTICS

The modern development in the field of not only Management, Commerce, Economics, Social Sciences, Mathematics and so on but also in our life like public services, defence, banking, insurance sector, tourism and hospitality, police and military etc. are dependent on a particular subject known as statistics. Statistics does play a vital role in enriching a specific domain by collecting data in that field, analysing the data by applying various statistical techniques and finally making statistical inferences about the domain. In the present world, statistics has almost a universal application. Our government applies statistics to make the economic planning in an effective and a pragmatic way. The businessman plan and expand their horizons of business on the basis of the analysis of the feedback data. The political parties try to impress the general public by presenting the statistics of their performances and accomplishments. Most of the research scholars of today also apply statistics to present their research papers in an authoritative manner. Thus the list of people using statistics goes on and on and on. Due to these factors, it is necessary to study the subject of statistics in an objective manner.

History of Statistics

Going through the history of ancient period and also that of medieval period, we do find the mention of statistics in many countries. However, there remains a question mark about the origin of the word 'statistics'. One view is that statistics is originated from the Latin word 'status'. According to another school of thought, it had its origin in the Italian word 'statista'. Some scholars believe that the German word 'statistik' was later changed to statistics and another suggestion is that the French word 'statistique' was made as statistics with the passage of time.

In those days, statistics was analogous to state or, to be more precise, the data that are collected and maintained for the welfare of the people belonging to the state. We are thankful to Kautilya who had kept a record of births and deaths as well as some other precious records in his famous book 'Arthashastra' during Chandragupta's reign in the fourth century B.C. During the reign of Akbar in the sixteenth century A.D. We find statistical records on agriculture in Ain-i-Akbari written by Abu Fazl. Referring to Egypt, the first census was conducted by the Pharaoh during 300 B.C. to 2000 B.C.

Definition of Statistics

We may define statistics either in a singular sense or in a plural sense Statistics, when used as a plural noun, may be defined as data qualitative as well as quantitative, that are collected, usually with a view of having statistical analysis.

However, statistics, when used as a singular noun, may be defined, as the scientific method that is employed for collecting, analysing and presenting data, leading finally to drawing statistical inferences about some important characteristics it means it is 'science of counting' or 'science of averages'.

Application of statistics

Among various applications of statistics, let us confine our discussions to the fields of Economics, Business Management and Commerce and Industry.

Economics

Modern developments in Economics have the roots in statistics. In fact, Economics and Statistics are closely associated. Time Series Analysis, Index Numbers, Demand Analysis etc. are some overlapping areas of Economics and Statistics. In this connection, we may also mention Econometrics – a branch of Economics that interact with statistics in a very positive way. Conducting socio-economic surveys and analysing the data derived from it are made with the help of different statistical methods. Regression analysis, one of the numerous applications of statistics, plays a key role in Economics for making future projection of demand of goods, sales, prices, quantities etc. which are all ingredients of Economic planning.

Business Management

Gone are the days when the managers used to make decisions on the basis of hunches, intuition or trials and errors. Now a days, because of the never-ending complexity in the business and industry environment, most of the decision making processes rely upon different quantitative techniques which could be described as a combination of statistical methods and operations research techniques. So far as statistics is concerned, inferences about the universe from the knowledge of a part of it, known as sample, plays an important role in the development of certain criteria. Statistical decision theory is another component of statistics that focuses on the analysis of complicated business strategies with a list of alternatives – their merits as well as demerits.

Statistics in Commerce and Industry

In this age of cut-throat competition, like the modern managers, the industrialists and the businessmen are expanding their horizons of industries and businesses with the help of statistical procedures. Data on previous sales, raw materials, wages and salaries, products of identical nature of other factories etc are collected, analysed and experts are consulted in order to maximise profits. Measures of central tendency and dispersion, correlation and regression analysis, time series analysis, index numbers, sampling, statistical quality control are some of the statistical methods employed in commerce and industry.

Limitations of Statistics

Before applying statistical methods, we must be aware of the following limitations:

I Statistics deals with the aggregates. An individual, to a statistician has no significance except the fact that it is a part of the aggregate.

II Statistics is concerned with quantitative data. However, qualitative data also can be converted to quantitative data by providing a numerical description to the corresponding qualitative data.

III Future projections of sales, production, price and quantity etc. are possible under a specific set of conditions. If any of these conditions is violated, projections are likely to be inaccurate.

IV The theory of statistical inferences is built upon random sampling. If the rules for random sampling is not strictly adhered to, the conclusion drawn on the basis of these unrepresentative samples would be erroneous. In other words, the experts should be consulted before deciding the sampling scheme.

13.1.2 COLLECTION OF DATA

We may define 'data' as quantitative information about some particular characteristic(s) under consideration. Although a distinction can be made between a qualitative characteristic and a quantitative characteristic but so far as the statistical analysis of the characteristic is concerned, we need to convert qualitative information to quantitative information by providing a numeric descriptions to the given characteristic. In this connection, we may note that a quantitative characteristic is known as a variable or in other words, a variable is a measurable quantity. Again, a variable may be either discrete or continuous. When a variable assumes a finite or a countably infinite number of isolated values, it is known as a discrete variable. Examples of discrete variables may be found in the number of petals in a flower, the number of misprints a book contains, the number of road accidents in a particular locality and so on. A variable, on the other hand, is known to be continuous if it can assume any value from a given interval. Examples of continuous variables may be provided by height, weight, sale, profit and so on. Finally, a qualitative characteristic is known as an attribute. The gender of a baby, the nationality of a person, the colour of a flower etc. are examples of attributes.

We can broadly classify data as

(a) Primary;

(b) Secondary.

Collection of data plays the very important role for any statistical analysis. The data which are collected for the first time by an investigator or agency are known as primary data whereas the data are known to be secondary if the data, as being already collected, are used by a different person or agency. Thus, if Prof. Das collects the data on the height of every student in his class, then these would be primary data for him. If, however, another person, say, Professor Bhargava uses the data, as collected by Prof. Das, for finding the average height of the students belonging to that class, then the data would be secondary for Prof. Bhargava.

Collection of Primary Data

The following methods are employed for the collection of primary data:

(i) Interview method;

(ii) Mailed questionnaire method;

(iii) Observation method;

(iv) Questionnaires filled and sent by enumerators.

Interview method again could be divided into (a) Personal Interview method, (b) Indirect Interview method and (c) Telephone Interview method.

In personal interview method, the investigator meets the respondents directly and collects the required information then and there from them. In case of a natural calamity like a super cyclone or an earthquake or an epidemic like plague, we may collect the necessary data much more quickly and accurately by applying this method.

If there are some practical problems in reaching the respondents directly, as in the case of a rail accident, then we may take recourse for conducting Indirect Interview where the investigator collects the necessary information from the persons associated with the problems.

Telephone interview method is a quick and rather non-expensive way to collect the primary data where the relevant information can be gathered by the researcher himself by contacting the interviewee over the phone. The first two methods, though more accurate, are inapplicable for covering a large area whereas the telephone interview, though less consistent, has a wide coverage.

The nuculer of non-responses is maximum for this third method of data collection.

Mailed questionnaire method comprises of framing a well-drafted and soundly-sequenced questionnaire covering all the important aspects of the problem under consideration and sending them to the respondents with pre-paid stamp after providing all the necessary guidelines for filling up the questionnaire. Although a wide area can be covered using the mailed questionnaire method, the amount of non-responses is likely to be maximum in this method.

In observation nuculer, data are collected, as in the case of obtaining the data on the height and weight of a group of students, by direct observation or using instrument. Although this is likely to be the best method for data collection, it is time consuming, laborious and covers only a small area. Questionnaire form of data collection is used for larger enquiries from the persons who are surveyed. Enumerators collects information directly by interviewing the persons having information : Question are explained and hence data is collected.

Sources of Secondary Data

There are many sources of getting secondary data. Some important sources are listed below:

(a) International sources like WHO, ILO, IMF, World Bank etc.

(b) Government sources like Statistical Abstract by CSO, Indian Agricultural Statistics by the Ministry of Food and Agriculture and so on.

(c) Private and quasi-government sources like ISI, ICAR, NCERT etc.

(d) Unpublished sources of various research institutes, researchers etc.

Scrutiny of Data

Since the statistical analyses are made only on the basis of data, it is necessary to check whether the data under consideration are accurate as well as consistence. No hard and fast rules can be recommended for the scrutiny of data. One must apply his intelligence, patience and experience while scrutinising the given information.

Errors in data may creep in while writing or copying the answer on the part of the enumerator. A keen observer can easily detect that type of error. Again, there may be two or more series of figures which are in some way or other related to each other. If the data for all the series are provided, they may be checked for internal consistency. As an example, if the data for population, area and density for some places are given, then we may verify whether they are internally consistent by examining whether the relation

$$\text{Density} = \frac{\text{Population}}{\text{Area}} \quad \text{holds.}$$

A good statistician can also detect whether the returns submitted by some enumerators are exactly of the same type thereby implying the lack of seriousness on the part of the enumerators. The bias of the enumerator also may be reflected by the returns submitted by him. This type of error can be rectified by asking the enumerator(s) to collect the data for the disputed cases once again.

13.1.3 PRESENTATION OF DATA

Once the data are collected and verified for their homogeneity and consistency, we need to present them in a neat and condensed form highlighting the essential features of the data. Any statistical analysis is dependent on a proper presentation of the data under consideration.

Classification or Organisation of Data

It may be defined as the process of arranging data on the basis of the characteristic under consideration into a number of groups or classes according to the similarities of the observations. Following are the objectives of classification of data:

(a) It puts the data in a neat, precise and condensed form so that it is easily understood and interpreted.

(b) It makes comparison possible between various characteristics, if necessary, and thereby finding the association or the lack of it between them.

(c) Statistical analysis is possible only for the classified data.

(d) It eliminates unnecessary details and makes data more readily understandable.

Data may be classified as -

(i) Chronological or Temporal or Time Series Data;

(ii) Geographical or Spatial Series Data;

(iii) Qualitative or Ordinal Data;

(iv) Quantitative or Cardinal Data.

When the data are classified in respect of successive time points or intervals, they are known as time series data. The number of students appeared for CA final for the last twenty years, the production of a factory per month from 2000 to 2015 etc. are examples of time series data.

Data arranged region wise are known as geographical data. If we arrange the students appeared for CA final in the year 2015 in accordance with different states, then we come across Geographical Data.

Data classified in respect of an attribute are referred to as qualitative data. Data on nationality, gender, smoking habit of a group of individuals are examples of qualitative data. Lastly, when the data are classified in respect of a variable, say height, weight, profits, salaries etc., they are known as quantitative data.

Data may be further classified as *frequency data* and *non-frequency data*. The qualitative as well as quantitative data belong to the frequency group whereas time series data and geographical data belong to the non-frequency group.

Mode of Presentation of Data

Next, we consider the following mode of presentation of data:

(a) Textual presentation;

(b) Tabular presentation or Tabulation;

(c) Diagrammatic representation.

(a) Textual presentation

This method comprises presenting data with the help of a paragraph or a number of paragraphs. The official report of an enquiry commission is usually made by textual presentation. Following is an example of textual presentation.

'In 2009, out of a total of five thousand workers of Roy Enamel Factory, four thousand and two hundred were members of a Trade Union. The number of female workers was twenty per cent of the total workers out of which thirty per cent were members of the Trade Union.

In 2010, the number of workers belonging to the trade union was increased by twenty per cent as compared to 2009 of which four thousand and two hundred were male. The number of workers not belonging to trade union was nine hundred and fifty of which four hundred and fifty were females.'

The merit of this mode of presentation lies in its simplicity and even a layman can present data by this method. The observations with exact magnitude can be presented with the help of textual presentation. Furthermore, this type of presentation can be taken as the first step towards the other methods of presentation.

Textual presentation, however, is not preferred by a statistician simply because, it is dull, monotonous and comparison between different observations is not possible in this method. For manifold classification, this method cannot be recommended.

(b) Tabular presentation or Tabulation

Tabulation may be defined as systematic presentation of data with the help of a statistical table having a number of rows and columns and complete with reference number, title, description of rows as well as columns and foot notes, if any.

We may consider the following guidelines for tabulation :

I A statistical table should be allotted a serial number along with a self-explanatory title.

II The table under consideration should be divided into caption, Box-head, Stub and Body. Caption is the upper part of the table, describing the columns and sub-columns, if any. The Box-head is the entire upper part of the table which includes columns and sub-column numbers, unit(s) of measurement along with caption. Stub is the left part of the table providing the description of the rows. The body is the main part of the table that contains the numerical figures.

III The table should be well-balanced in length and breadth.

IV The data must be arranged in a table in such a way that comparison(s) between different figures are made possible without much labour and time. Also, the row totals, column totals, the units of measurement must be shown.

V The data should be arranged intelligently in a well-balanced sequence and the presentation of data in the table should be appealing to the eyes as far as practicable.

VI Notes describing the source of the data and bringing clarity and, if necessary, about any rows or columns known as footnotes, should be shown at the bottom part of the table.

The textual presentation of data, relating to the workers of Roy Enamel Factory is shown in the following table.

Table 13.1.1

Status of the workers of Roy Enamel factory on the basis of their trade union membership for 2009 and 2010.

Status	Member of TU			Non-member			Total		
Year	M (1)	F (2)	T (3)=(1)+ (2)	M (4)	F (5)	T (6)=(4)+ (5)	M (7)	F (8)	T (9)=(7)+ (8)
2009	3900	300	4200	300	500	800	4200	800	5000
2010	4200	840	5040	500	450	950	4700	1290	5990

Source:

Footnote: TU, M, F and T stand for trade union, male, female and total respectively.

The tabulation method is usually preferred to textual presentation as

(i) It facilitates comparison between rows and columns.

(ii) Complicated data can also be represented using tabulation.

(iii) It is a must for diagrammatic representation.

(iv) Without tabulation, statistical analysis of data is not possible.

(c) Diagrammatic representation of data

Another alternative and attractive representation of statistical data is provided by charts, diagrams and pictures. Unlike the first two methods of representation of data, diagrammatic representation can be used for both the educated section and uneducated section of the society. Furthermore, any hidden trend present in the given data can be noticed only in this mode of representation. However, compared to tabulation, this is less accurate. So, if there is a priority for accuracy, we have to recommend tabulation.

We are going to consider the following types of diagrams :

I Line diagram or Historiagram;

II Bar diagram;

III Pie chart.

I Line diagram or Historiagram

When the data vary over time, we take recourse to line diagram. In a simple line diagram, we plot each pair of values of (t, y_t), y_t representing the time series at the time point t in the $t–y_t$ plane. The plotted points are then joined successively by line segments and the resulting chart is known as line-diagram.

When the time series exhibit a wide range of fluctuations, we may think of logarithmic or ratio chart where Log y_t and not y_t is plotted against t. We use Multiple line chart for representing two or more related time series data expressed in the same unit and multiple – axis chart in somewhat similar situations if the variables are expressed in different units.

II **Bar diagram**

There are two types of bar diagrams namely, Horizontal Bar diagram and Vertical Bar diagram. While horizontal bar diagram is used for qualitative data or data varying over space, the vertical bar diagram is associated with quantitative data or time series data. Bars i.e. rectangles of equal width and usually of varying lengths are drawn either horizontally or vertically. We consider Multiple or Grouped Bar diagrams to compare related series. Component or sub-divided Bar diagrams are applied for representing data divided into a number of components. Finally, we use Divided Bar charts or Percentage Bar diagrams for comparing different components of a variable and also the relating of the components to the whole. For this situation, we may also use Pie chart or Pie diagram or circle diagram.

ILLUSTRATIONS:

Example 13.1.1: The profits in lakhs of Rupees of an industrial house for 2009, 2010, 2011, 2012, 2013, 2014, and 2015 are 5, 8, 9, 6, 12, 15 and 24 respectively. Represent these data using a suitable diagram.

SOLUTION:

We can represent the profits for 7 consecutive years by drawing either a line chart or a vertical bar chart. Fig. 13.1.1 shows a line chart and figure 13.1.2 shows the corresponding vertical bar chart.

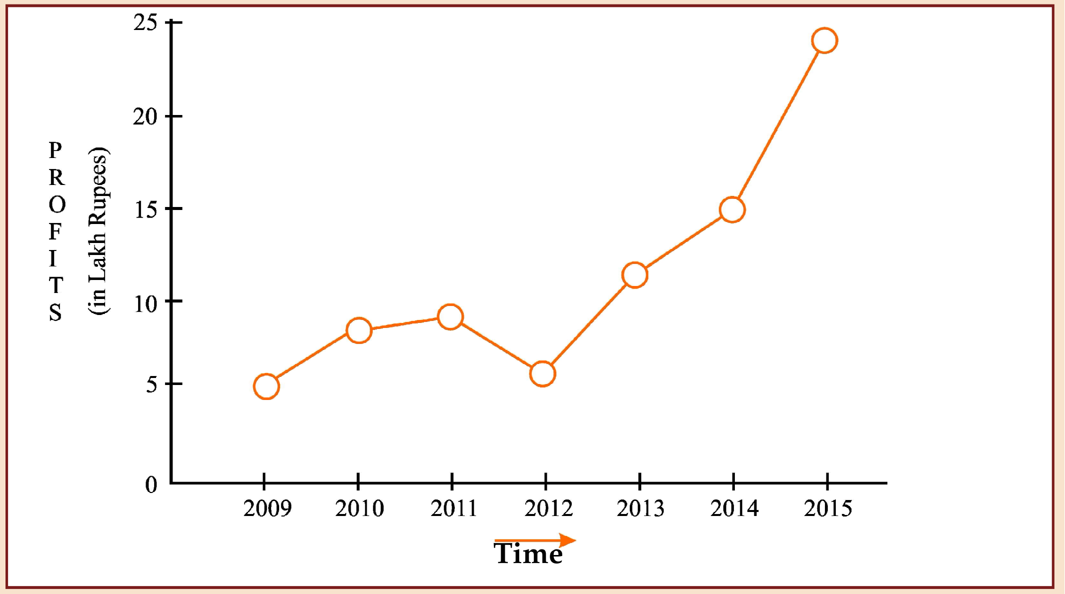


Figure 13.1.1

Showing line chart for the Profit of an Industrial House during 2009 to 2015.

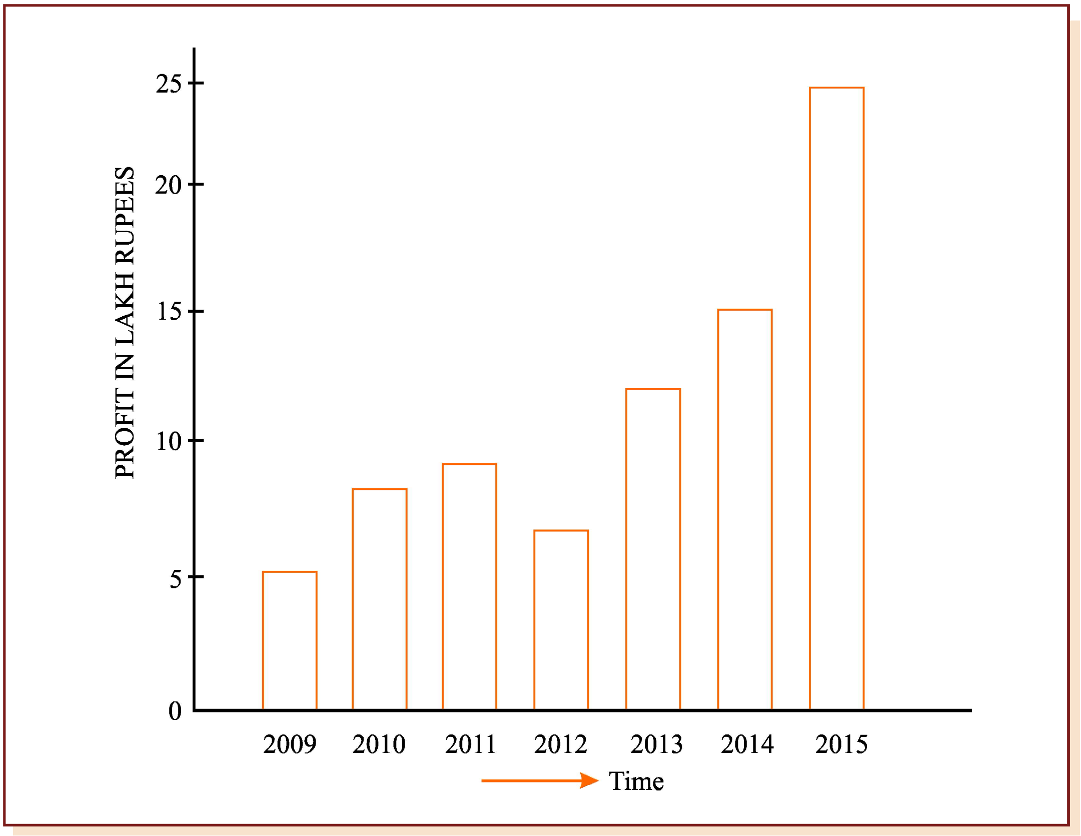


Figure 13.1.2

Showing vertical bar diagram for the Profit of an Industrial house from 2009 to 2015.

Example 13.1.2: The production of wheat and rice of a region are given below :

Year	Production in metric tones	
	Wheat	Rice
2012	12	25
2013	15	30
2014	18	32
2015	19	36

Represent this information using a suitable diagram.

Solution:

We can represent this information by drawing a multiple line chart. Alternately, a multiple bar diagram may be considered. These are depicted in figure 13.1.3 and 13.1.4 respectively.

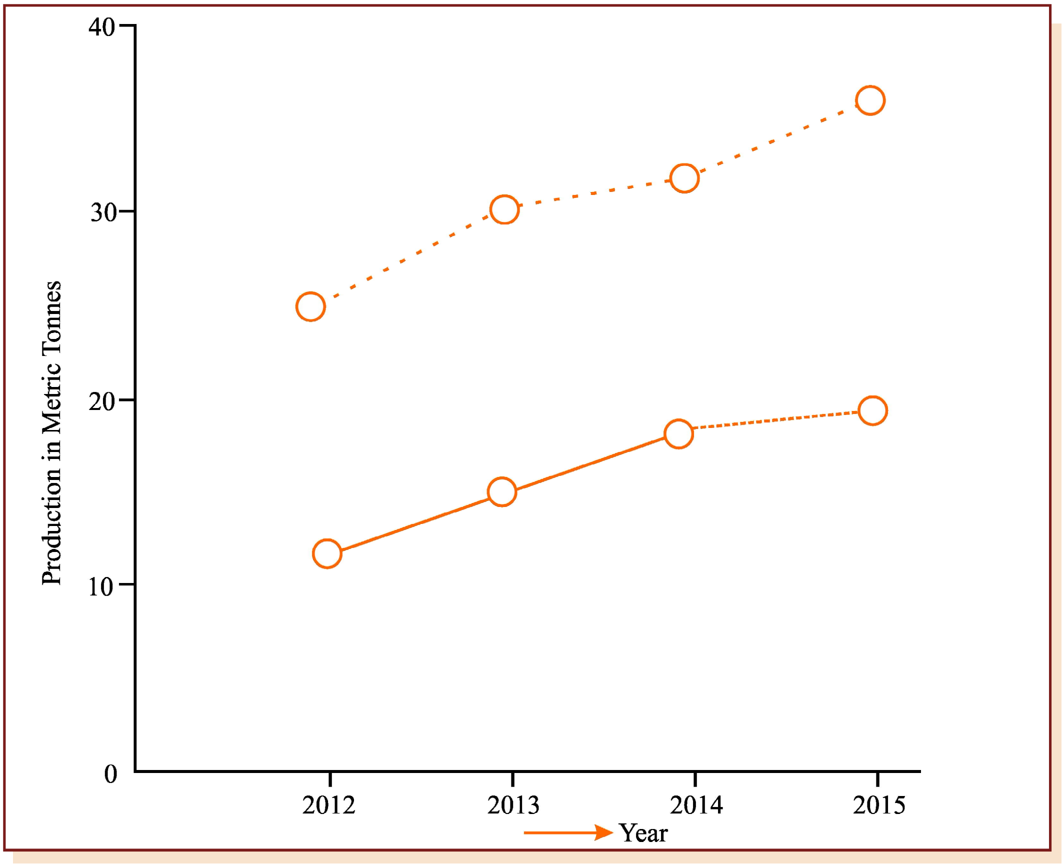


Figure 13.1.3

Multiple line chart showing production of wheat and rice of a region during 2012–2015. (Dotted line represent production of rice and continuous line that of wheat).

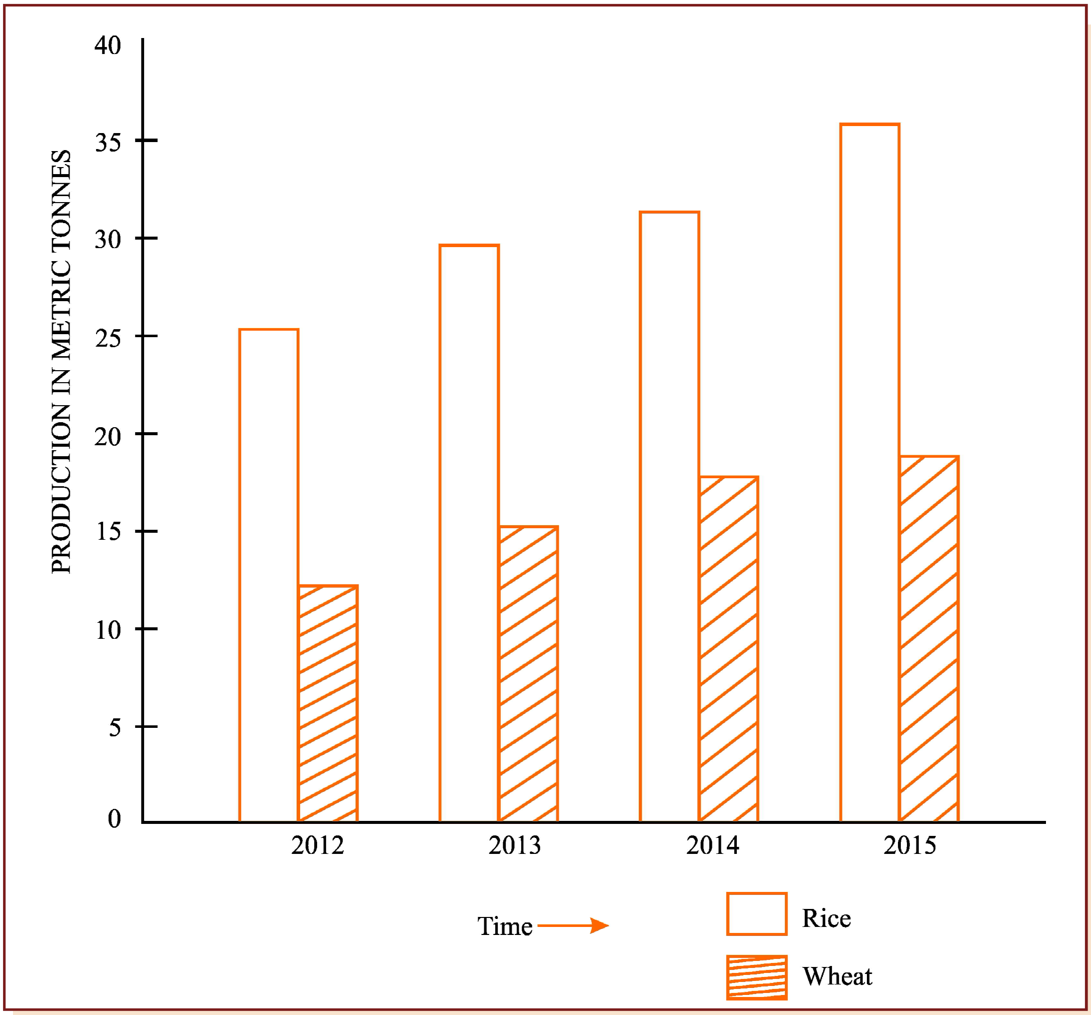


Figure 13.1.4

Multiple bar chart representing production of rice and wheat from 2012 to 2015.

Example 13.1.3: Draw an appropriate diagram with a view to represent the following data :

Source	Revenue in millions of (₹)
Customs	80
Excise	190
Income Tax	160
Corporate Tax	75
Miscellaneous	35

Solution:

Pie chart or divided bar chart would be the ideal diagram to represent this data. We consider Pie chart.

Table 13.1.2

Computation for drawing Pie chart

Source (1)	Revenue in Million rupees (2)	Central angle $= \frac{(2)}{\text{Total of (2)}} \text{x} \, 360^{o}$
Customs	80	$\frac{80}{540} \text{x} \, 360^{o} = 53^{o}$ (approx.)
Excise	190	$\frac{190}{540} \text{x} \, 360^{o} = 127^{o}$
Income Tax	160	$\frac{160}{540} \text{x} \, 360^{o} = 107^{o}$
Corporate Tax	75	$\frac{75}{540} \text{x} \, 360^{o} = 50^{o}$
Miscellaneous	35	$\frac{35}{540} \text{x} \, 360^{o} = 23^{o}$
Total	540	360^{0}

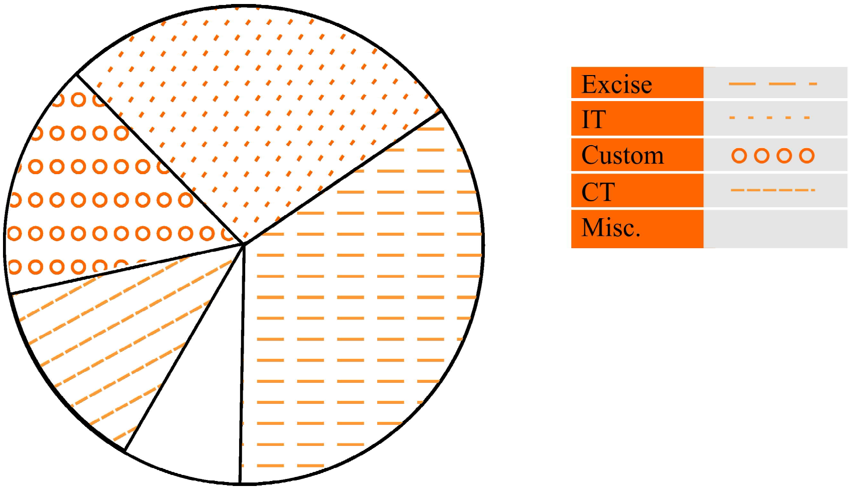


Figure 13.1.5

Pie chart showing the distribution of Revenue

13.1. 4 FREQUENCY DISTRIBUTION

As discussed in the previous section, frequency data occur when we classify statistical data in respect of either a variable or an attribute. A frequency distribution may be defined as a tabular representation of statistical data, usually in an ascending order, relating to a measurable characteristic according to individual value or a group of values of the characteristic under study.

In case, the characteristic under consideration is an attribute, say nationality, then the tabulation is made by allotting numerical figures to the different classes the attribute may belong like, in this illustration, counting the number of Indian, British, French, German and so on. The qualitative characteristic is divided into a number of categories or classes which are mutually exclusive and exhaustive and the figures against all these classes are recorded. The figure corresponding to a particular class, signifying the number of times or how frequently a particular class occurs is known as the frequency of that class. Thus, the number of Indians, as found from the given data, signifies the frequency of the Indians. So frequency distribution is a statistical table that distributes the total frequency to a number of classes.

When tabulation is done in respect of a discrete random variable, it is known as Discrete or Ungrouped or simple Frequency Distribution and in case the characteristic under consideration is a continuous variable, such a classification is termed as Grouped Frequency Distribution. In case of a grouped frequency distribution, tabulation is done not against a single value as in the case of an attribute or a discrete random variable but against a group of values. The distribution of the number of car accidents in Delhi during 12 months of the year 2005 is an example of a ungrouped frequency distribution and the distribution of heights of the students of St. Xavier's College for the year 2004 is an example of a grouped frequency distribution.

Example 13.1.4: Following are the records of babies born in a nursing home in Bangalore during a week (B denoting Boy and G for Girl) :

B G G B G G B B G G

G G B B B G B B G B

B B G B B B G G B G

Construct a frequency distribution according to gender.

Solution:

In order to construct a frequency distribution of babies in accordance with their gender, we count the number of male births and that of female births and present this information in the following table.

Table 13.1.3

Frequency distribution of babies according to Gender

Category	Number of births
Boy (B)	16
Girl (G)	14
Total	30

Frequency Distribution of a Variable

For the construction of a frequency distribution of a variable, we need to go through the following steps :

I Find the largest and smallest observations and obtain the difference between them, known as Range, in case of a continuous variable.

II Form a number of classes depending on the number of isolated values assumed by a discrete variable. In case of a continuous variable, find the number of class intervals using the relation, No. of class Interval × class length $\cong$ Range.

III Present the class or class interval in a table known as frequency distribution table.

IV Apply 'tally mark' i.e. a stroke against the occurrence of a particulars value in a class or class interval.

V Count the tally marks and present these numbers in the next column, known as frequency column, and finally check whether the total of all these class frequencies tally with the total number of observations.

Example 13.1.5: A review of the first 30 pages of a statistics book reveals the following printing mistakes:

0	1	3	3	2	5	6	0	1	0
4	1	1	0	2	3	2	5	0	4
2	3	2	2	3	3	4	6	1	4

Make a frequency distribution of printing mistakes.

Solution:

Since x, the printing mistakes, is a discrete variable, x can assume seven values 0, 1, 2, 3, 4, 5 and 6. Thus we have 7 classes, each class comprising a single value.

Table 13.1.4

Frequency Distribution of the number of printing mistakes of the first 30 pages of a book

Printing Mistake	Tally marks	Frequency (No. of Pages)
0	~~IIII~~	5
1	~~IIII~~	5
2	~~IIII~~ I	6
3	~~IIII~~ I	6
4	IIII	4
5	II	2
6	II	2
Total	–	30

Example 13.1.6: Following are the weights in kgs. of 36 BBA students of St. Xavier's College.

70	73	49	61	61	47	57	50	59
59	68	45	55	65	68	56	68	55
70	70	57	44	69	73	64	49	63
65	70	65	62	64	73	67	60	50

Construct a frequency distribution of weights, taking class length as 5.

Solution:

We have, Range = Maximum weight – Minimum weight

= 73 kgs. – 44 kgs.

= 29 kgs.

No. of class interval × class lengths ≅ Range

$\Rightarrow$ No. of class interval $\times\ 5 \cong 29$

$\Rightarrow$ No. of class interval $= \dfrac{29}{5} \cong 6.$

(We always take the next integer as the number of class intervals so as to include both the minimum and maximum values).

Table 13.1.5

Frequency Distribution of weights of 36 BBA Students

Weight in kg (Class Interval)	Tally marks	No. of Students (Frequency)
44-48	III	3
49-53	IIII	4
54-58	~~IIII~~	5
59-63	~~IIII~~ II	7
64-68	~~IIII~~ IIII	9
69-73	~~IIII~~ III	8
Total	–	36

Some important terms associated with a frequency distribution

Class Limit (CL)

Corresponding to a class interval, the class limits may be defined as the minimum value and the maximum value the class interval may contain. The minimum value is known as the lower class limit (LCL) and the maximum value is known as the upper class limit (UCL). For the frequency distribution of weights of BBA Students, the LCL and UCL of the first class interval are 44 kgs. and 48 kgs. respectively.

Class Boundary (CB)

Class boundaries may be defined as the actual class limit of a class interval. For overlapping classification or mutually exclusive classification that excludes the upper class limits like 10–20, 20–30, 30–40, ……… etc. the class boundaries coincide with the class limits. This is usually done for a continuous variable. However, for non-overlapping or mutually inclusive classification that includes both the class limits like 0–9, 10–19, 20–29,…… which is usually applicable for a discrete variable, we have

$$\text{LCB} = \text{LCL} - \frac{\text{D}}{2}$$

$$\text{and UCB} = \text{UCL} + \frac{\text{D}}{2}$$

where D is the difference between the LCL of the next class interval and the UCL of the given class interval. For the data presented in table 10.5, LCB of the first class interval

$$= 44\text{ kgs.} - \frac{(49 - 48)}{2}\text{kgs.}$$

$$= 43.50\text{ kgs.}$$

and the corresponding UCB

$$= 48\text{ kgs.} + \frac{49 - 48}{2}\text{ kgs.}$$

= 48.50 kgs.

Mid-point or Mid-value or class mark

Corresponding to a class interval, this may be defined as the total of the two class limits or class boundaries to be divided by 2. Thus, we have

$$\text{mid-point} = \frac{\text{LCL} + \text{UCL}}{2}$$

$$= \frac{\text{LCB} + \text{UCB}}{2}$$

Referring to the distribution of weight of BBA students, the mid-points for the first two class intervals are

$$\frac{44\text{ kgs.} + 48\text{ kgs.}}{2} \text{ and } \frac{49\text{ kgs.} + 53\text{ kgs.}}{2}$$

i.e. 46 kgs. and 51 kgs. respectively.

Width or size of a class interval

The width of a class interval may be defined as the difference between the UCB and the LCB of that class interval. For the distribution of weights of BBA students, C, the class length or width is 48.50 kgs. – 43.50 kgs. = 5 kgs. for the first class interval. For the other class intervals also, C remains same.

Cumulative Frequency

The cumulative frequency corresponding to a value for a discrete variable and corresponding to a class boundary for a continuous variable may be defined as the number of observations less than the value or less than or equal to the class boundary. This definition refers to the less than cumulative frequency. We can define more than cumulative frequency in a similar manner. Both types of cumulative frequencies are shown in the following table.

Table 13.1.6

Cumulative Frequency Distribution of weights of 36 BBA students

Weight in kg	Cumulative Frequency	
(CB)	Less than	More than
43.50	0	33 + 3 or 36
48.50	0 + 3 or 3	29 + 4 or 33
53.50	3 + 4 or 7	24 + 5 or 29
58.50	7 + 5 or 12	17 + 7 or 24
63.50	12 + 7 or 19	8 + 9 or 17
68.50	19 + 9 or 28	0 + 8 or 8
73.50	28 + 8 or 36	0

Frequency density of a class interval

It may be defined as the ratio of the frequency of that class interval to the corresponding class length. The frequency densities for the first two class intervals of the frequency distribution of weights of BBA students are 3/5 and 4/5 i.e. 0.60 and 0.80 respectively.

Relative frequency and percentage frequency of a class interval

Relative frequency of a class interval may be defined as the ratio of the class frequency to the total frequency. Percentage frequency of a class interval may be defined as the ratio of class frequency to the total frequency, expressed as a percentage. For the last example, the relative frequencies for the first two class intervals are 3/36 and 4/36 respectively and the percentage frequencies are 300/36 and 400/36 respectively. It is quite obvious that whereas the relative frequencies add up to unity, the percentage frequencies add up to one hundred.

13.1.5 GRAPHICAL REPRESENTATION OF A FREQUENCY DISTRIBUTION

We consider the following types of graphical representation of frequency distribution :

(i) Histogram or Area diagram;

(ii) Frequency Polygon;

(iii) Ogives or cumulative Frequency graphs.

(i) **Histogram or Area diagram**

This is a very convenient way to represent a frequency distribution. Histogram helps us to get an idea of the frequency curve of the variable under study. Some statistical measure can be obtained using a histogram. A comparison among the frequencies for different class intervals is possible in this mode of diagrammatic representation.

In order to draw a histogram, the class limits are first converted to the corresponding class boundaries and a series of adjacent rectangles, one against each class interval, with the class interval as base or breadth and the frequency or frequency density usually when the class intervals are not uniform as length or altitude, is erected. The histogram for the distribution of weight of 36 BBA students is shown below. The mode of the weights has also been determined using the histogram.

i.e. Mode = 66.50 kgs.

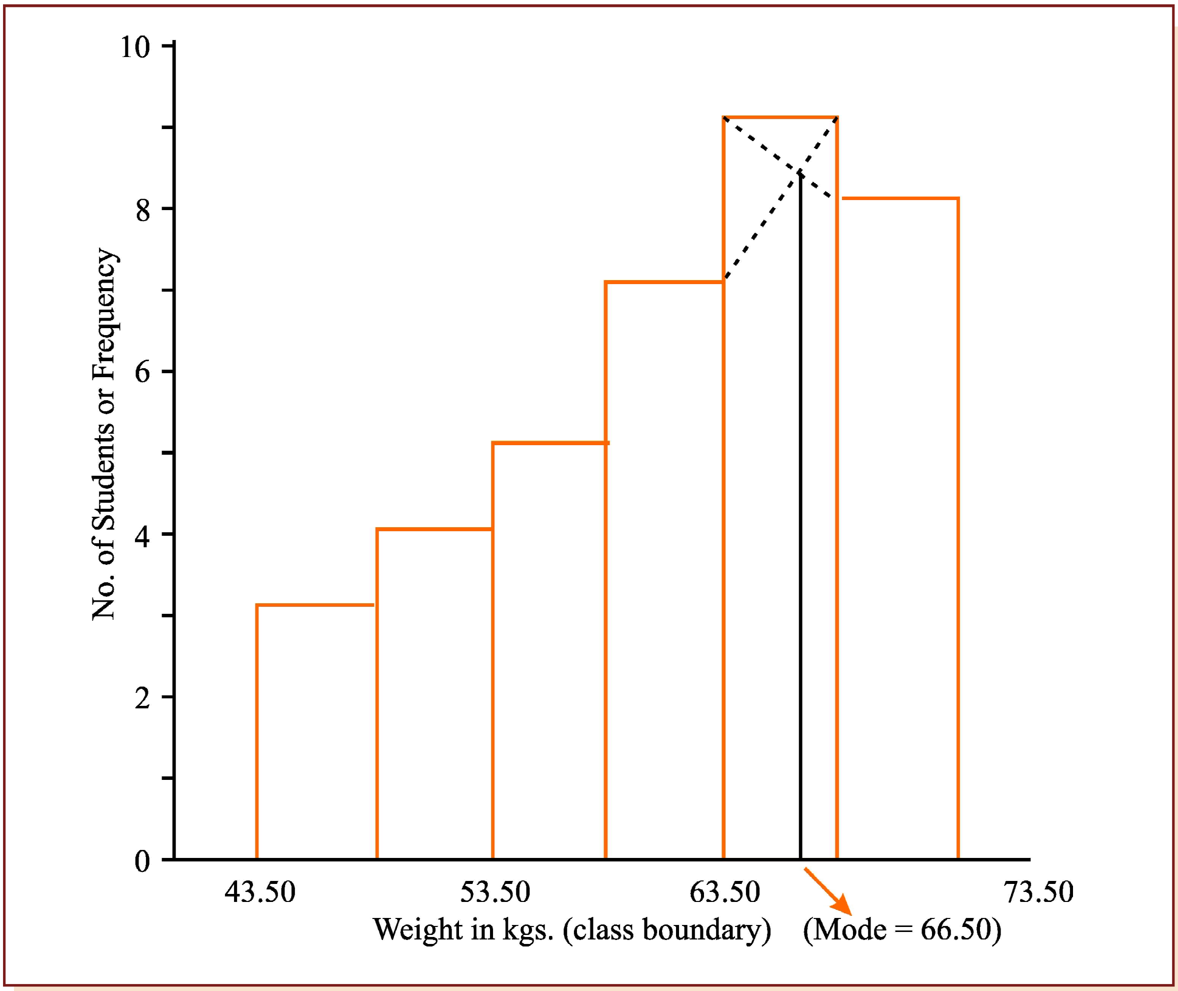


Figure 13.1.6

Showing histogram for the distribution of weight of 36 BBA students

(ii) Frequency Polygon

Usually frequency polygon is meant for single frequency distribution. However, we also apply it for grouped frequency distribution provided the width of the class intervals remains the same. A frequency curve can be regarded as a limiting form of frequency polygon. In order to draw a frequency polygon, we plot (x_i, f_i) for i = 1, 2, 3, ……….. n with x_i denoting the mid-point of the its class interval and f_i, the corresponding frequency, n being the number of class intervals. The plotted points are joined successively by line segments and the figure, so drawn, is given the shape of a polygon, a closed figure, by joining the two extreme ends of the drawn figure to two additional points (x_0,0) and (x_{n+1},0).

The frequency polygon for the distribution of weights of BBA students is shown in Figure 13.7. We can also obtain a frequency polygon starting with a histogram by adding the mid-points of the upper sides of the rectangles successively and then completing the figure by joining the two ends as before.

Mid-points	No. of Students (Frequency)
46	3
51	4
56	5
61	7
66	9
71	8

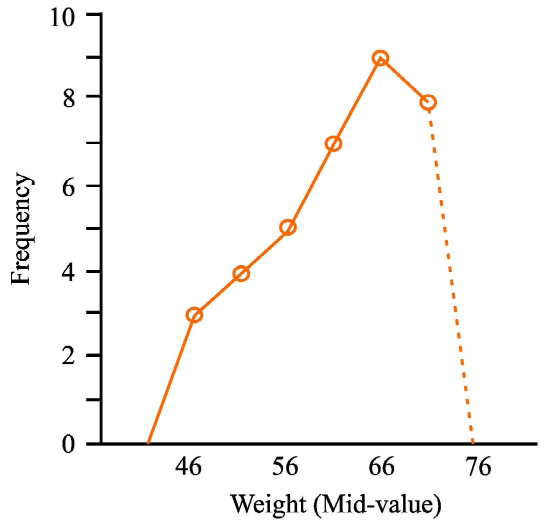


Figure 13.1.7

Showing frequency polygon for the distribution of height of 36 BBA students

(iii) Ogives or Cumulative Frequency Graph

By plotting cumulative frequency against the respective class boundary, we get ogives. As such there are two ogives – less than type ogives, obtained by taking less than cumulative frequency on the vertical axis and more than type ogives by plotting more than type cumulative frequency on the vertical axis and thereafter joining the plotted points successively by line segments. Ogives may be considered for obtaining quartiles graphically. If a perpendicular is drawn from the point of intersection of the two ogives on the horizontal axis, then the x-value of this point gives us the value of median, the second or middle quartile. Ogives further can be put into use for making short term projections.

Figure 13.8 depicts the ogives and the determination of the quartiles. This figure give us the following information.

1st quartile or lower quartile (Q_1)	=	55 kgs.
2nd quartile or median (Q_2 or Me)	=	62.50 kgs.
3rd quartile or upper quartile (Q_3)	=	68 kgs.

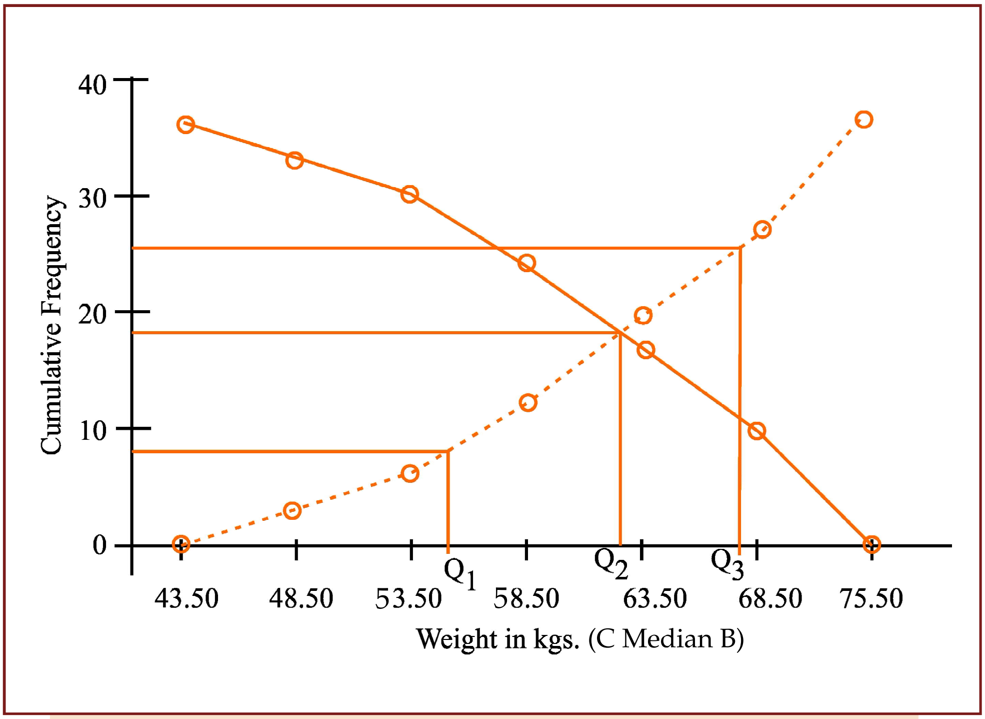


Figure 13.1.8

Showing the ogives for the distribution of weights of 36 BBA students

We find $Q_1 = 55$ kgs.

$Q_2 = Me = 62.50$ kgs.

$Q_3 = 68$ kgs.

Frequency Curve

A frequency curve is a smooth curve for which the total area is taken to be unity. It is a limiting form of a histogram or frequency polygon. The frequency curve for a distribution can be obtained by drawing a smooth and free hand curve through the mid-points of the upper sides of the rectangles forming the histogram.

There exist four types of frequency curves namely

(a) Bell-shaped curve;

(b) U-shaped curve;

(c) J-shaped curve;

(d) Mixed curve.

Most of the commonly used distributions provide bell-shaped curve, which, as suggested by the name, looks almost like a bell. The distribution of height, weight, mark, profit etc. usually belong to this category. On a bell-shaped curve, the frequency, starting from a rather low value, gradually reaches the maximum value, somewhere near the central part and then gradually decreases to reach its lowest value at the other extremity.

For a U-shaped curve, the frequency is minimum near the central part and the frequency slowly but steadily reaches its maximum at the two extremities. The distribution of Kolkata bound commuters belongs to this type of curve as there are maximum number of commuters during the peak hours in the morning and in the evening.

The J-shaped curve starts with a minimum frequency and then gradually reaches its maximum frequency at the other extremity. The distribution of commuters coming to Kolkata from the early morning hour to peak morning hour follows such a distribution. Sometimes, we may also come across an inverted J-shaped frequency curve.

Lastly, we may have a combination of these frequency curves, known as mixed curve. These are exhibited in the following figures.

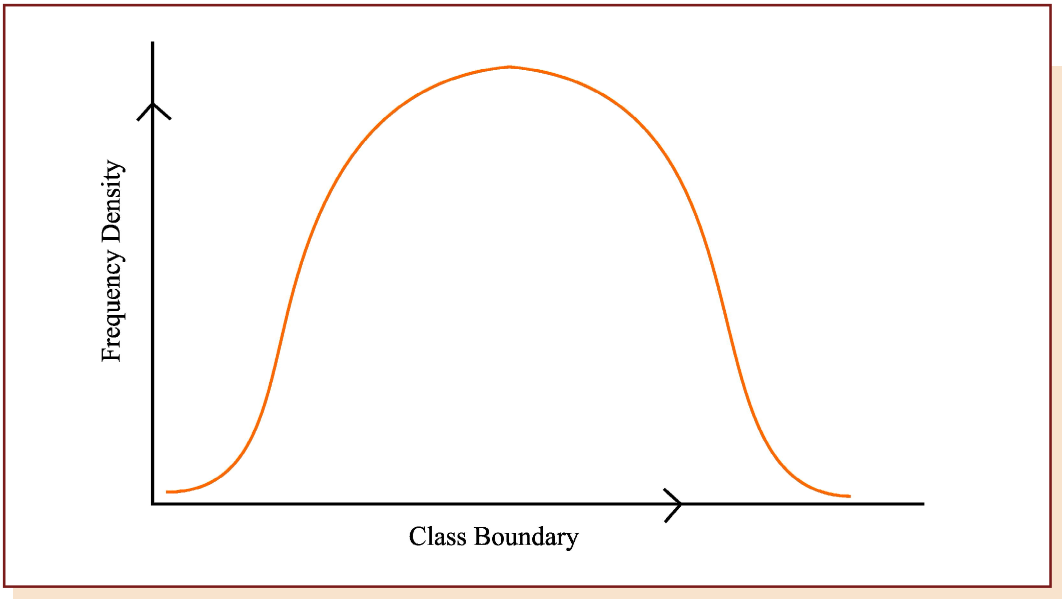


Figure 13.1.9

Bell-shaped curve

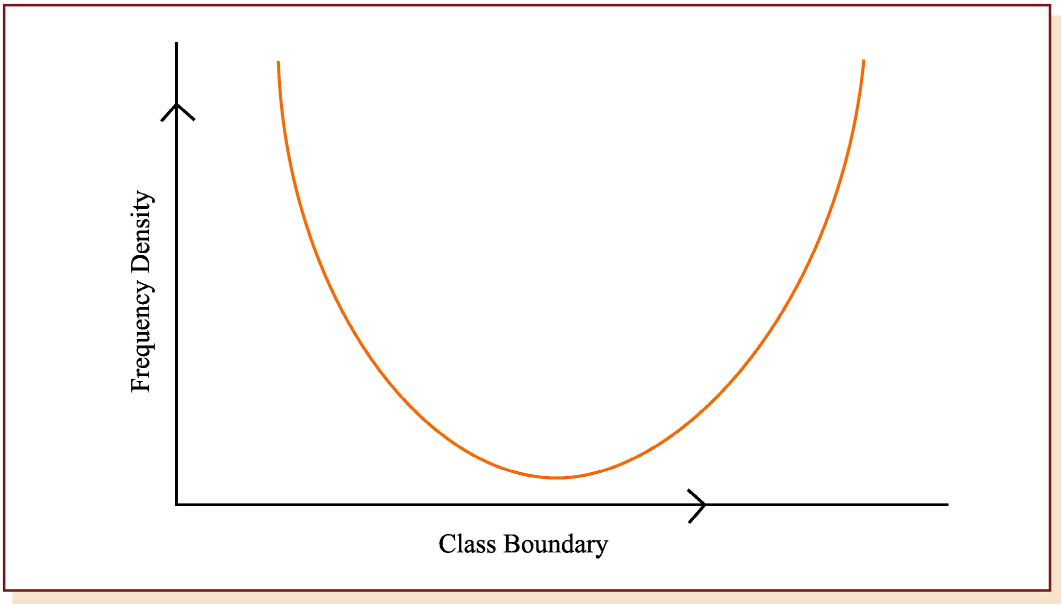


Figure 13.1.10

U-shaped curve

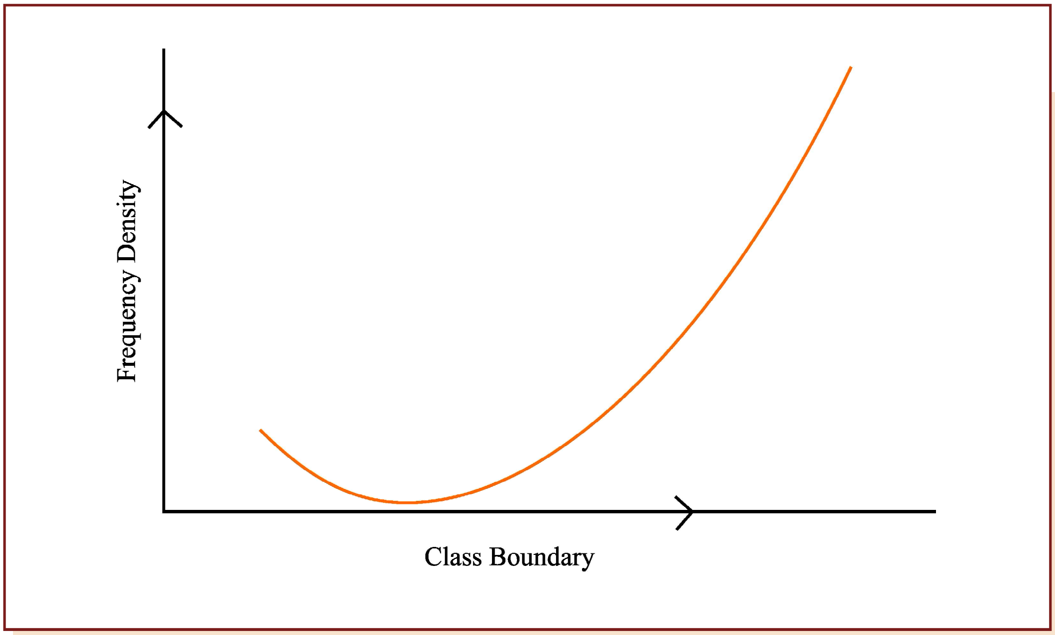


Figure 13.1.11

J-shaped curve

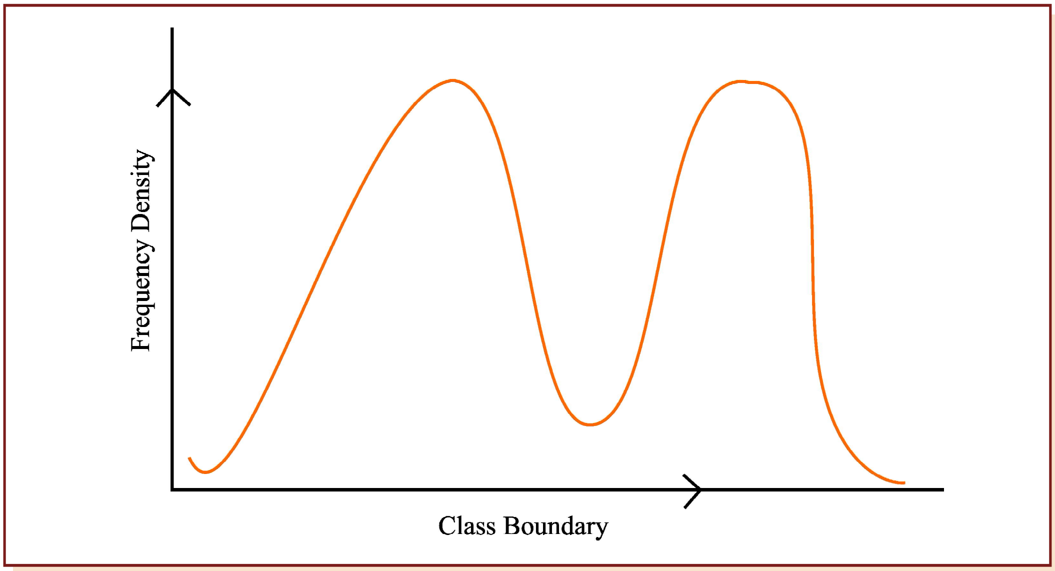


Figure 13.1.12

Mixed curve

SUMMARY

- Statistics deals with the aggregates. An individual, to a statistician has no significance except the fact that it is a part of the aggregate.
- Statistics is concerned with quantitative data. However, qualitative data also can be converted to quantitative data by providing a numerical description to the corresponding qualitative data.
- The theory of statistical inferences is built upon random sampling. If the rules for random sampling are not strictly adhered to, the conclusion drawn on the basis of these unrepresentative samples would be erroneous.
- We can broadly classify data as

  (a) Primary;

  (b) Secondary.
- Mode of Presentation of Data

  (a) Textual presentation;

  (b) Tabular presentation or Tabulation;

  (c) Diagrammatic representation.
- The types of diagrams:

  (a) Line diagram or Historiagram;

  (b) Bar diagram;

  (c) Pie chart.
- Frequency Distribution of a Variable

  (a) Find the largest and smallest observations and obtain the difference between them, known as Range, in case of a continuous variable.

  (b) Form a number of classes depending on the number of isolated values assumed by a discrete variable. In case of a continuous variable, find the number of class intervals using the relation, No. of class Interval × class length $\cong$ Range.

  (c) Present the class or class interval in a table known as frequency distribution table.

  (d) Apply 'tally mark' i.e. a stroke against the occurrence of a particulars value in a class or class interval.

  (e) Count the tally marks and present these numbers in the next column, known as frequency column, and finally check whether the total of all these class frequencies tally with the total number of observations.

UNIT I EXERCISE

Set A

Answer the following questions. Each question carries 1 mark.

1. Which of the following statements is false?
   (a) Statistics is derived from the Latin word 'Status'
   (b) Statistics is derived from the Italian word 'Statista'
   (c) Statistics is derived from the French word 'Statistik'
   (d) None of these.

2. Statistics is defined in terms of numerical data in the
   (a) Singular sense (b) Plural sense
   (c) Either (a) or (b) (d) Both (a) and (b).

3. Statistics is applied in
   (a) Economics (b) Business management
   (c) Commerce and industry (d) All these.

4. Statistics is concerned with
   (a) Qualitative information (b) Quantitative information
   (c) (a) or (b) (d) Both (a) and (b).

5. An attribute is
   (a) A qualitative characteristic (b) A quantitative characteristic
   (c) A measurable characteristic (d) All these.

6. Annual income of a person is
   (a) An attribute (b) A discrete variable
   (c) A continuous variable (d) (b) or (c).

7. Marks of a student is an example of
   (a) An attribute (b) A discrete variable
   (c) A continuous variable (d) None of these.

8. Nationality of a student is
   (a) An attribute (b) A continuous variable
   (c) A discrete variable (d) (a) or (c).

9. Drinking habit of a person is
   (a) An attribute (b) A variable
   (c) A discrete variable (d) A continuous variable.

10. Age of a person is

    (a) An attribute (b) A discrete variable

    (c) A continuous variable (d) A variable.

11. Data collected on religion from the census reports are

    (a) Primary data (b) Secondary data

    (c) Sample data (d) (a) or (b).

12. The data collected on the height of a group of students after recording their heights with a measuring tape are

    (a) Primary data (b) Secondary data

    (c) Discrete data (d) Continuous data.

13. The primary data are collected by

    (a) Interview method (b) Observation method

    (c) Questionnaire method (d) All these.

14. The quickest method to collect primary data is

    (a) Personal interview (b) Indirect interview

    (c) Telephone interview (d) By observation.

15. The best method to collect data, in case of a natural calamity, is

    (a) Personal interview (b) Indirect interview

    (c) Questionnaire method (d) Direct observation method.

16. In case of a rail accident, the appropriate method of data collection is by

    (a) Personal interview (b) Direct interview

    (c) Indirect interview (d) All these.

17. Which method of data collection covers the widest area?

    (a) Telephone interview method (b) Mailed questionnaire method

    (c) Direct interview method (d) All these.

18. The amount of non-responses is maximum in

    (a) Mailed questionnaire method (b) Interview method

    (c) Observation method (d) All these.

19. Some important sources of secondary data are

    (a) International and Government sources

    (b) International and primary sources

    (c) Private and primary sources

    (d) Government sources.

20. Internal consistency of the collected data can be checked when

    (a) Internal data are given (b) External data are given

    (c) Two or more series are given (d) A number of related series are given.

21. The accuracy and consistency of data can be verified by

    (a) Internal checking (b) External checking

    (c) Scrutiny (d) Both (a) and (b).

22. The mode of presentation of data are

    (a) Textual, tabulation and diagrammatic (b) Tabular, internal and external

    (c) Textual, tabular and internal (d) Tabular, textual and external.

23. The best method of presentation of data is

    (a) Textual (b) Tabular

    (c) Diagrammatic (d) (b) and (c).

24. The most attractive method of data presentation is

    (a) Tabular (b) Textual

    (c) Diagrammatic (d) (a) or (b).

25. For tabulation, 'caption' is

    (a) The upper part of the table (b) The lower part of the table

    (c) The main part of the table (d) The upper part of a table that describes the column and sub-column.

26. 'Stub' of a table is the

    (a) Left part of the table describing the columns

    (b) Right part of the table describing the columns

    (c) Right part of the table describing the rows

    (d) Left part of the table describing the rows.

27. The entire upper part of a table is known as

    (a) Caption (b) Stub

    (c) Box head (d) Body.

28. The unit of measurement in tabulation is shown in

    (a) Box head (b) Body

    (c) Caption (d) Stub.

29. In tabulation source of the data, if any, is shown in the

(a) Footnote
(b) Body
(c) Stub
(d) Caption.

30. Which of the following statements is untrue for tabulation?

(a) Statistical analysis of data requires tabulation
(b) It facilitates comparison between rows and not columns
(c) Complicated data can be presented
(d) Diagrammatic representation of data requires tabulation.

31. Hidden trend, if any, in the data can be noticed in

(a) Textual presentation
(b) Tabulation
(c) Diagrammatic representation
(d) All these.

32. Diagrammatic representation of data is done by

(a) Diagrams
(b) Charts
(c) Pictures
(d) All these.

33. The most accurate mode of data presentation is

(a) Diagrammatic method
(b) Tabulation
(c) Textual presentation
(d) None of these.

34. The chart that uses logarithm of the variable is known as

(a) Line chart
(b) Ratio chart
(c) Multiple line chart
(d) Component line chart.

35. Multiple line chart is applied for

(a) Showing multiple charts
(b) Two or more related time series when the variables are expressed in the same unit
(c) Two or more related time series when the variables are expressed in different unit
(d) Multiple variations in the time series.

36. Multiple axis line chart is considered when

(a) There is more than one time series
(b) The units of the variables are different
(c) (a) or (b)
(d) (a) and (b).

37. Horizontal bar diagram is used for

(a) Qualitative data
(b) Data varying over time
(c) Data varying over space
(d) (a) or (c).

38. Vertical bar diagram is applicable when
    (a) The data are qualitative
    (b) The data are quantitative
    (c) When the data vary over time
    (d) (b) or (c).

39. Divided bar chart is considered for
    (a) Comparing different components of a variable
    (b) The relation of different components to the table
    (c) (a) or (b)
    (d) (a) and (b).

40. In order to compare two or more related series, we consider
    (a) Multiple bar chart
    (b) Grouped bar chart
    (c) (a) or (b)
    (d) (a) and (b).

41. Pie-diagram is used for
    (a) Comparing different components and their relation to the total
    (b) Representing qualitative data in a circle
    (c) Representing quantitative data in circle
    (d) (b) or (c).

42. A frequency distribution
    (a) Arranges observations in an increasing order
    (b) Arranges observation in terms of a number of groups
    (c) Relates to a measurable characteristic
    (d) All these.

43. The frequency distribution of a continuous variable is known as
    (a) Grouped frequency distribution
    (b) Simple frequency distribution
    (c) (a) or (b)
    (d) (a) and (b).

44. The distribution of shares is an example of the frequency distribution of

   (a) A discrete variable

   (b) A continuous variable

   (c) An attribute

   (d) (a) or (c).

45. The distribution of profits of a blue-chip company relates to

   (a) Discrete variable

   (b) Continuous variable

   (c) Attributes

   (d) (a) or (b).

46. Mutually exclusive classification

   (a) Excludes both the class limits

   (b) Excludes the upper class limit but includes the lower class limit

   (c) Includes the upper class limit but excludes the upper class limit

   (d) Either (b) or (c).

47. Mutually inclusive classification is usually meant for

   (a) A discrete variable

   (b) A continuous variable

   (c) An attribute

   (d) All these.

48. Mutually exclusive classification is usually meant for

   (a) A discrete variable

   (b) A continuous variable

   (c) An attribute

   (d) Any of these.

49. The LCB is

   (a) An upper limit to LCL

   (b) A lower limit to LCL

   (c) (a) and (b)

   (d) (a) or (b).

50. The UCB is

(a) An upper limit to UCL (b) A lower limit to LCL

(c) Both (a) and (b) (d) (a) or (b).

51. length of a class is

(a) The difference between the UCB and LCB of that class

(b) The difference between the UCL and LCL of that class

(c) (a) or (b)

(d) Both (a) and (b).

52. For a particular class boundary, the less than cumulative frequency and more than cumulative frequency add up to

(a) Total frequency (b) Fifty per cent of the total frequency

(c) (a) or (b) (d) None of these.

53. Frequency density corresponding to a class interval is the ratio of

(a) Class frequency to the total frequency (b) Class frequency to the class length

(c) Class length to the class frequency (d) Class frequency to the cumulative frequency.

54. Relative frequency for a particular class

(a) Lies between 0 and 1 (b) Lies between 0 and 1, both inclusive

(c) Lies between –1 and 0 (d) Lies between –1 to 1.

55. Mode of a distribution can be obtained from

(a) Histogram (b) Less than type ogives

(c) More than type ogives (d) Frequency polygon.

56. Median of a distribution can be obtained from

(a) Frequency polygon (b) Histogram

(c) Less than type ogives (d) None of these.

57. A comparison among the class frequencies is possible only in

(a) Frequency polygon (b) Histogram

(c) Ogives (d) (a) or (b).

58. Frequency curve is a limiting form of

(a) Frequency polygon (b) Histogram

(c) (a) or (b) (d) (a) and (b).

59. Most of the commonly used frequency curves are

(a) Mixed (b) Inverted J-shaped

(c) U-shaped (d) Bell-shaped.

60. The distribution of profits of a company follows

(a) J-shaped frequency curve (b) U-shaped frequency curve

(c) Bell-shaped frequency curve (d) Any of these.

Set B

Answer the following questions. Each question carries 2 marks.

1. Out of 1000 persons, 25 per cent were industrial workers and the rest were agricultural workers. 300 persons enjoyed world cup matches on TV. 30 per cent of the people who had not watched world cup matches were industrial workers. What is the number of agricultural workers who had enjoyed world cup matches on TV?

(a) 260 (b) 240 (c) 230 (d) 250

2. A sample study of the people of an area revealed that total number of women were 40% and the percentage of coffee drinkers were 45 as a whole and the percentage of male coffee drinkers was 20. What was the percentage of female non-coffee drinkers?

(a) 10 (b) 15 (c) 18 (d) 20

3. Cost of sugar in a month under the heads raw materials, labour, direct production and others were 12, 20, 35 and 23 units respectively. What is the difference between the central angles for the largest and smallest components of the cost of sugar?

(a) 72^{o} (b) 48^{o} (c) 56^{o} (d) 92^{o}

4. The number of accidents for seven days in a locality are given below :

No. of accidents :	0	1	2	3	4	5	6
Frequency :	15	19	22	31	9	3	2

What is the number of cases when 3 or less accidents occurred?

(a) 56 (b) 6 (c) 68 (d) 87

5. The following data relate to the incomes of 86 persons :

Income in Rs. :	500–999	1000–1499	1500–1999	2000–2499
No. of persons :	15	28	36	7

What is the percentage of persons earning more than Rs. 1500?

(a) 50 (b) 45 (c) 40 (d) 60

6. The following data relate to the marks of a group of students:

Marks :	Below 10	Below 20	Below 30	Below 40	Below 50
No. of students :	15	38	65	84	100

How many students got marks more than 30?

(a) 65 (b) 50 (c) 35 (d) 43

7. Find the number of observations between 250 and 300 from the following data :

Value :	More than 200	More than 250	More than 300	More than 350
No. of observations :	56	38	15	0

(a) 56 (b) 23 (c) 15 (d) 8

Set C

Answer the following questions. Each question carries 5 marks.

1. In a study about the male and female students of commerce and science departments of a college in 5 years, the following datas were obtained :

1995	2000
70% male students	75% male students
65% read Commerce	40% read Science
20% of female students read Science	50% of male students read Commerce
3000 total No. of students	3600 total No. of students.

After combining 1995 and 2000 if x denotes the ratio of female commerce student to female Science student and y denotes the ratio of male commerce student to male Science student, then

(a) $x = y$ (b) $x > y$ (c) $x < y$ (d) $x \geq y$

2. In a study relating to the labourers of a jute mill in West Bengal, the following information was collected.

'Twenty per cent of the total employees were females and forty per cent of them were married. Thirty female workers were not members of Trade Union. Compared to this, out of 600 male workers 500 were members of Trade Union and fifty per cent of the male workers were married. The unmarried non-member male employees were 60 which formed ten per cent of the total male employees. The unmarried non-members of the employees were 80'. On the basis of this information, the ratio of married male non-members to the married female non-members is

(a) 1 : 3 (b) 3 : 1 (c) 4 : 1 (d) 5 : 1

3. The weight of 50 students in pounds are given below :

82,	95,	120,	174,	179,	176,	159,	91,	85,	175
88,	160,	97,	133,	159,	176,	151,	115,	105,	172
170,	128,	112,	101,	123,	117,	93,	117,	99,	90
113,	119,	129,	134,	178,	105,	147,	107,	155,	157
98,	117,	95,	135,	175,	97,	160,	168,	144,	175

If the data are arranged in the form of a frequency distribution with class intervals as 81-100, 101-120, 121-140, 141-160 and 161-180, then the frequencies for these 5 class intervals are

(a) 6, 9, 10, 11, 14 (b) 12, 8, 7, 11, 12 (c) 10, 12, 8, 11, 9 (d) 12, 12, 6, 9, 11

4. The following data relate to the marks of 48 students in statistics :

56, 10, 54, 38, 21, 43, 12, 22
48, 51, 39, 26, 12, 17, 36, 19
48, 36, 15, 33, 30, 62, 57, 17
5, 17, 45, 46, 43, 55, 57, 38
43, 28, 32, 35, 54, 27, 17, 16
11, 43, 45, 2, 16, 46, 28, 45

What are the frequency densities for the class intervals 30-39, 40-49 and 50-59

(a) 0.20, 0.50, 0.90

(b) 0.70, 0.90, 1.10

(c) 0.1875, 0.1667, 0.2083

(d) 0.90, 1.1, 0.7

5. The following information relates to the age of death of 50 persons in an area :

36, 48, 50, 45, 49, 31, 50, 48, 42, 57
43, 40, 32, 41, 39, 39, 43, 47, 45, 52
47, 48, 53, 37, 48, 50, 41, 49, 50, 53
38, 41, 49, 45, 36, 39, 31, 48, 59, 48
37, 49, 53, 51, 54, 59, 48, 38, 39, 45

If the class intervals are 31-33, 34-36, 37-39, …. Then the percentage frequencies for the last five class intervals are

(a) 18, 18, 10, 2 and 4. (b) 10, 15, 18, 4 and 2. (c) 14, 18, 20, 10 and 2.

(d) 10, 12, 16, 4 and 6.

ANSWERS

Set A

1.	(c)	**2.**	(b)	**3.**	(d)	**4.**	(d)	**5.**	(a)	**6.**	(b)
7.	(b)	**8.**	(a)	**9.**	(a)	**10.**	(c)	**11.**	(b)	**12.**	(a)
13.	(d)	**14.**	(c)	**15.**	(a)	**16.**	(c)	**17.**	(b)	**18.**	(a)
19.	(a)	**20.**	(d)	**21.**	(c)	**22.**	(a)	**23.**	(b)	**24.**	(c)

25.	(d)	**26.**	(d)	**27.**	(c)	**28.**	(a)	**29.**	(a)	**30.**	(b)
31.	(c)	**32.**	(d)	**33.**	(b)	**34.**	(b)	**35.**	(b)	**36.**	(d)
37.	(d)	**38.**	(d)	**39.**	(d)	**40.**	(c)	**41.**	(a)	**42.**	(d)
43.	(a)	**44.**	(a)	**45.**	(b)	**46.**	(b)	**47.**	(a)	**48.**	(d)
49.	(b)	**50.**	(a)	**51.**	(a)	**52.**	(a)	**53.**	(b)	**54.**	(a)
55.	(a)	**56.**	(c)	**57.**	(b)	**58.**	(c)	**59.**	(d)	**60.**	(c)

Set B

1.	(a)	**2.**	(b)	**3.**	(d)	**4.**	(d)	**5.**	(a)	**6.**	(c)
7.	(b)										

Set C

1.	(b)	**2.**	(c)	**3.**	(d)	**4.**	(d)	**5.**	(a)

ADDITIONAL QUESTION BANK

1. Graph is a

(a) Line diagram (b) Bar diagram (c) Pie diagram (d) Pictogram

2. Details are shown by

(a) Charts (b) Tabular presentation

(c) both (d) none

3. The relationship between two variables are shown in

(a) Pictogram (b) Histogram (c) Bar diagram (d) Line diagram

4. In general the number of types of tabulation are

(a) two (b) three (c) one (d) four

5. A table has

(a) four (b) two (c) five (d) none parts.

6. The number of errors in Statistics are

(a) one (b) two (c) three (d) four

7. The number of “Frequency distribution“ is

(a) two (b) one (c) five (d) four

8. (Class frequency)/(Width of the class) is defined as

(a) Frequency density (b) Frequency distribution

(c) both (d) none

9. Tally marks determines

(a) class width (b) class boundary (c) class limit (d) class frequency

10. Cumulative Frequency Distribution is a

(a) graph (b) frequency (c) Statistical Table (d) distribution

11. To find the number of observations less than any given value

(a) Single frequency distribution (b) Grouped frequency distribution

(c) Cumulative frequency distribution (d) None is used.

12. An area diagram is

(a) Histogram (b) Frequency Polygon

(c) Ogive (d) none

13. When all classes have a common width

(a) Pie Chart (b) Frequency Polygon

(c) both (d) none is used.

14. An approximate idea of the shape of frequency curve is given by

(a) Ogive (b) Frequency Polygon

(c) both (d) none

15. Ogive is a

(a) Line diagram (b) Bar diagram (c) both (d) none

16. Unequal widths of classes in the frequency distribution do not cause any difficulty in the construction of

(a) Ogive (b) Frequency Polygon

(c) Histogram (d) none

17. The graphical representation of a cumulative frequency distribution is called

(a) Histogram (b) Ogive (c) both (d) none.

18. The most common form of diagrammatic representation of a grouped frequency distribution is

(a) Ogive (b) Histogram (c) Frequency Polygon (d) none

19. Vertical bar chart may appear somewhat alike

(a) Histogram (b) Frequency Polygon

(c) both (d) none

20. The number of types of cumulative frequency is

(a) one (b) two (c) three (d) four

21. A representative value of the class interval for the calculation of mean, standard deviation, mean deviation etc. is

    (a) class interval (b) class limit (c) class mark (d) none

22. The number of observations falling within a class is called

    (a) density (b) frequency (c) both (d) none

23. Classes with zero frequencies are called

    (a) nill class (b) empty class (c) class (d) none

24. For determining the class frequencies it is necessary that these classes are

    (a) mutually exclusive (b) not mutually exclusive

    (c) independent (d) none

25. Most extreme values which would ever be included in a class interval are called

    (a) class limits (b) class interval (c) class boundaries (d) none

26. The value exactly at the middle of a class interval is called

    (a) class mark (b) mid value (c) both (d) none

27. Difference between the lower and the upper class boundaries is

    (a) width (b) size (c) both (d) none

28. In the construction of a frequency distribution, it is generally preferable to have classes of

    (a) equal width (b) unequal width (c) maximum (d) none

29. Frequency density is used in the construction of

    (a) Histogram (b) Ogive

    (c) Frequency Polygon (d) none when the classes are of unequal width.

30. "Cumulative Frequency" only refers to the

    (a) less-than type (b) more-than type (c) both (d) none

31. For the construction of a grouped frequency distribution

    (a) class boundaries (b) class limits (c) both (d) none are used.

32. In all Statistical calculations and diagrams involving end points of classes

    (a) class boundaries (b) class value (c) both (d) none are used.

33. Upper limit of any class is ________ from the lower limit of the next class

    (a) same (b) different

    (c) both (d) none

34. Upper boundary of any class coincides with the Lower boundary of the next class.

    (a) true (b) false (c) both (d) none.

35. Excepting the first and the last, all other class boundaries lie midway between the upper limit of a class and the lower limit of the next higher class.

(a) true (b) false (c) both (d) none

36. The lower extreme point of a class is called

(a) lower class limit (b) lower class boundary

(c) both (d) none

37. For the construction of grouped frequency distribution from ungrouped data we use

(a) class limits (b) class boundaries (c) class width (d) none

38. When one end of a class is not specified, the class is called

(a) closed- end class (b) open- end class (c) both (d) none

39. Class boundaries should be considered to be the real limits for the class interval.

(a) true (b) false (c) both (d) none

40. Difference between the maximum & minimum value of a given data is called

(a) width (b) size (c) range (d) none

41. In Histogram if the classes are of unequal width then the heights of the rectangles must be proportional to the frequency densities.

(a) true (b) false (c) both (d) none

42. When all classes have equal width, the heights of the rectangles in Histogram will be numerically equal to the

(a) class frequencies (b) class boundaries (c) both (d) none

43. Consecutive rectangles in a Histogram have no space in between

(a) true (b) false (c) both (d) none

44. Histogram emphasizes the widths of rectangles between the class boundaries.

(a) false (b) true (c) both (d) none

45. To find the mode graphically

(a) Ogive (b) Frequency Polygon

(c) Histogram (d) none may be used.

46. When the width of all classes is same, frequency polygon has not the same area as the Histogram.

(a) True (b) false (c) both (d) none

47. For obtaining frequency polygon we join the successive points whose abscissa represent the corresponding class frequency_____

(a) true (b) false (c) both (d) none

48. In representing simple frequency distributions of a discrete variable

(a) Ogive (b) Histogram (c) Frequency Polygon (d) both is useful.

49. Diagrammatic representation of the cumulative frequency distribution is

(a) Frequency Polygon (b) Ogive (c) Histogram (d) none

50. For the overlapping classes 0–10 , 10–20 , 20–30 etc.the class mark of the class 0–10 is

(a) 5 (b) 0 (c) 10 (d) none

51. For the non-overlapping classes 0–19 , 20–39 , 40–59 the class mark of the class 0–19 is

(a) 0 (b) 19 (c) 9.5 (d) none

52.

Class :	0–10	10–20	20–30	30–40	40–50
Frequency :	5	8	15	6	4

For the class 20–30 , cumulative frequency is

(a) 20 (b) 13 (c) 15 (d) 28

53. An Ogive can be prepared in _____________ different ways.

(a) 2 (b) 3 (c) 4 (d) none

54. The curve obtained by joining the points, whose x- coordinates are the upper limits of the class-intervals and y coordinates are corresponding cumulative frequencies is called

(a) Ogive (b) Histogram (c) Frequency Polygon (d) Frequency Curve

55. The breadth of the rectangle is equal to the length of the class-interval in

(a) Ogive (b) Histogram (c) both (d) none

56. In Histogram, the classes are taken

(a) overlapping (b) non-overlapping (c) both (d) none

57. For overlapping class-intervals the class limit & class boundary are

(a) same (b) not same (c) zero (d) none

58. Data classification is of________kinds

(a) four (b) Three (c) two (d) five

ANSWERS

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

26.	(c)	**27.**	(c)	**28.**	(a)	**29.**	(a)	**30.**	(a)
31.	(b)	**32.**	(a)	**33.**	(b)	**34.**	(a)	**35.**	(a)
36.	(b)	**37.**	(a)	**38.**	(b)	**39.**	(a)	**40.**	(c)
41.	(a)	**42.**	(a)	**43.**	(a)	**44.**	(b)	**45.**	(c)
46.	(b)	**47.**	(b)	**48.**	(c)	**49.**	(b)	**50.**	(a)
51.	(c)	**52.**	(d)	**53.**	(a)	**54.**	(a)	**55.**	(b)
56.	(a)	**57.**	(a)	**58.**	(a)				

UNIT 2 SAMPLING

LEARNING OBJECTIVES

After reading this unit a student will learn -

- Different procedure of sampling which will be the best representative of the population;

13.2.1 INTRODUCTION

There are situations when we would like to know about a vast, infinite universe or population. But some important factors like time, cost, efficiency, vastness of the population make it almost impossible to go for a complete enumeration of all the units constituting the population. Instead, we take recourse to selecting a representative part of the population and infer about the unknown universe on the basis of our knowledge from the known sample. A somewhat clear picture would emerge out if we consider the following cases.

In the first example let us share the problem faced by Mr. Basu. Mr. Basu would like to put a big order for electrical lamps produced by Mr. Ahuja's company "General Electricals". But before putting the order, he must know whether the claim made by Mr. Ahuja that the lamps of General Electricals last for at least 1500 hours is justified.

Miss Manju Bedi is a well-known social activist. Of late, she has noticed that the incidence of a particular disease in her area is on the rise. She claims that twenty per cent of the people in her town have been suffering from the disease.

In both the situations, we are faced with three different types of problems. The first problem is how to draw a representative sample from the population of electrical lamps in the first case and from the population of human beings in her town in the second case. The second problem is to estimate the population parameters i.e., the average life of all the bulbs produced by General Electricals and the proportion of people suffering form the disease in the first and second examples respectively on the basis of sample observations. The third problem relates to decision making i.e., is there enough evidence, once again on the basis of sample observations, to suggest that the claims made by Mr. Ahuja or Miss Bedi are justifiable so that Mr. Basu can take a decision about buying the lamps from General Electricals in the first case and some effective steps can be taken in the second example with a view to reducing the outbreak of the disease. We consider tests of significance or tests of hypothesis before decision making.

13.2.2 BASIC PRINCIPLES OF SAMPLE SURVEY

Sample Survey is the study of the unknown population on the basis of a proper representative sample drawn from it. How can a part of the universe reveal the characteristics of the unknown universe? The answer to this question lies in the basic principles of sample survey comprising the following components:

(a) Law of Statistical regularity

(b) Principle of Inertia

(c) Principle of Optimization

(d) Principle of Validity

(a) According to the law of statistical regularity, if a sample of fairly large size is drawn from the population under discussion at random, then on an average the sample would posses the characteristics of that population.

Thus the sample, to be taken from the population, should be moderately large. In fact larger the sample size, the better in revealing the identity of the population. The reliability of a statistic in estimating a population characteristics varies as the square root of the sample size. However, it is not always possible to increase the sample size as it would put an extra burden on the available resource. We make a compromise on the sample size in accordance with some factors like cost, time, efficiency etc.

Apart from the sample size, the sample should be drawn at random from the population which means that each and every unit of the population should have a pre-assigned probability to belong to the sample.

(b) The results derived from a sample, according to the principle of inertia of large numbers, are likely to be more reliable, accurate and precise as the sample size increases, provided other factors are kept constant. This is a direct consequence of the first principle.

(c) The principle of optimization ensures that an optimum level of efficiency at a minimum cost or the maximum efficiency at a given level of cost can be achieved with the selection of an appropriate sampling design.

(d) The principle of validity states that a sampling design is valid only if it is possible to obtain valid estimates and valid tests about population parameters. Only a probability sampling ensures this validity.

13.2.3 COMPARISON BETWEEN SAMPLE SURVEY AND COMPLETE ENUMERATION

When complete information is collected for all the units belonging to a population, it is defined as complete enumeration or census. In most cases, we prefer sample survey to complete enumeration due to the following factors:

(a) **Speed:** As compared to census, a sample survey could be conducted, usually, much more quickly simply because in sample survey, only a part of the vast population is enumerated.

(b) **Cost:** The cost of collection of data on each unit in case of sample survey is likely to be more as compared to census because better trained personnel are employed for conducting a sample survey. But when it comes to total cost, sample survey is likely to be less expensive as only some selected units are considered in a sample survey.

(c) **Reliability:** The data collected in a sample survey are likely to be more reliable than that in a complete enumeration because of trained enumerators better supervision and application of modern technique.

(d) **Accuracy:** Every sampling is subjected to what is known as sampling fluctuation which is termed as sampling error. It is obvious that complete enumeration is totally free from this sampling error. However, errors due to recording observations, biases on the part of the enumerators, wrong and faulty interpretation of data etc. are prevalent in both sampling and census and this type of error is termed as non-sampling errors. It may be noted that in sample survey, the sampling error can be reduced to a great extent by taking several steps like increasing the sample size, adhering to a probability sampling design strictly and so on. The non-sampling errors also can be contained to a desirable degree by a proper planning which is not possible or feasible in case of complete enumeration.

(e) **Necessity:** Sometimes, sampling becomes necessity. When it comes to destructive sampling where the items get exhausted like testing the length of life of electrical bulbs or sampling from a hypothetical population like coin tossing, there is no alternative to sample survey.

However, when it is necessary to get detailed information about each and every item constituting the population, we go for complete enumeration. If the population size is not large, there is hardly any merit to take recourse to sampling. If the occurrence of just one defect may lead to a complete destruction of the process as in an aircraft, we must go for complete enumeration.

13.2.4 ERRORS IN SAMPLE SURVEY

Errors or biases in a survey may be defined as the deviation between the value of population parameter as obtained from a sample and its observed value. Errors are of two types.

I. Sampling Errors

II. Non-Sampling Errors

Sampling Errors : Since only a part of the population is investigated in a sampling, every sampling design is subjected to this type of errors. The factors contributing to sampling errors are listed below:

(a) Errors arising out due to defective sampling design: Selection of a proper sampling design plays a crucial role in sampling. If a non- probabilistic sampling design is followed, the bias or prejudice of the sampler affects the sampling technique thereby resulting some kind of error.

(b) Errors arising out due to substitution: A very common practice among the enumerators is to replace a sampling unit by a suitable unit in accordance with their convenience when difficulty arises in getting information from the originally selected unit. Since the sampling design is not strictly adhered to, this results in some type of bias.

(c) Errors owing to faulty demarcation of units: It has its origin in faulty demarcation of sampling units. In case of an agricultural survey, the sampler has, usually, a tendency to underestimate or overestimate the character under consideration.

(d) Errors owing to wrong choice of statistic: One must be careful in selecting the proper statistic while estimating a population characteristic.

(e) Variability in the population: Errors may occur due to variability among population units beyond a degree. This could be reduced by following somewhat complicated sampling design like stratified sampling, Multistage sampling etc.

Non-sampling Errors

As discussed earlier, this type of errors happen both in sampling and complete enumeration. Some factors responsible for this particular kind of biases are lapse of memory, preference for certain digits, ignorance, psychological factors like vanity, non- responses on the part of the interviewees wrong measurements of the sampling units, communication gap between the interviewers and the interviewees, incomplete coverage etc. on the part of the enumerators also lead to non-sampling errors.

13.2.5 SOME IMPORTANT TERMS ASSOCIATED WITH SAMPLING

Population or Universe

It may be defined as the aggregate of all the units under consideration. All the lamps produced by "General Electricals" in our first example in the past, present and future constitute the population. In the second example, all the people living in the town of Miss Manju form the population. The number of units belonging to a population is known as population size. If there are one lakh people in her town then the population size, to be denoted by N, is 1 lakh.

A population may be finite or infinite. If a population comprises only a finite number of units, then it is known as a finite population. The population in the second example is obviously, finite. If the population contains an infinite or uncountable number of units, then it is known as an infinite population. The population of electrical lamps of General Electricals is infinite. Similarly, the population of stars, the population of mosquitoes in Kolkata, the population of flowers in Mumbai, the population of insects in Delhi etc. are infinite population.

Population may also be regarded as existent or hypothetical. A population consisting of real objects is known as an existent population. The population of the lamps produced by General Electricals and the population of Miss Manju's town are example of existent populations. A population that exists just hypothetically like the population of heads when a coin is tossed infinitely is known as a hypothetical or an imaginary population.

Sample

A sample may be defined as a part of a population so selected with a view to representing the population in all its characteristics selection of a proper representative sample is pretty important because statistical inferences about the population are drawn only on the basis of the sample observations. If a sample contains n units, then n is known as sample size. If a sample of 500 electrical lamps is taken from the production process of General Electricals, then n = 500. The units forming the sample are known as "Sampling Units". In the first example, the sampling unit is electrical lamp and in the second example, it is a human. A detailed and complete list of all the sampling units is known as a "Sampling Frame". Before drawing sample, it is a must to have a updated sampling frame complete in all respects before the samples are actually drawn.

Parameter

A parameter may be defined as a characteristic of a population based on all the units of the population. Statistical inferences are drawn about population parameters based on the sample observations drawn from that population. In the first example, we are interested about the parameter "Population Mean". If x a denotes the a th member of the population, then population mean m, which represents the average length of life of all the lamps produced by General Electricals is given by

$$\mu = \frac{\sum_{a=1}^{n} x_a}{N} \qquad (13.2.1)$$

Where N denotes the population size i.e. the total number of lamps produced by the company. In the second example, we are concerned about the population proportion P, representing the ratio of the people suffering from the disease to the total number of people in the town. Thus if there are X people possessing this attribute i.e. suffering from the disease, then we have

$$P = \frac{X}{N} \qquad (13.2.2)$$

Another important parameter namely the population variance, to be denoted by s^2 is given by

$$\sigma^2 = \frac{\sum (X_a - \mu)^2}{N} \qquad (13.2.3)$$

Also we have SD = $\sigma = \sqrt{\frac{\sum (X_a - \mu)^2}{N}}$ (13.2.4)

Statistics

A statistic may be defined as a statistical measure of sample observation and as such it is a function of sample observations. If the sample observations are denoted by $x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$, then a statistic T may be expressed as T = $f(x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n)$

A statistic is used to estimate a particular population parameter. The estimates of population mean, variance and population proportion are given by

$$\bar{x} = \hat{\mu} = \frac{\sum x_i}{n} \qquad (13.2.5)$$

$$S_2 = \hat{\sigma}^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n} \qquad (13.2.6)$$

and p = $p = \hat{P} = \frac{x}{n}$ (13.2.7)

Where x, in the last case, denotes the number of units in the sample in possession of the attribute under discussion.

Sampling Distribution and Standard Error of a Statistic

Starting with a population of N units, we can draw many a sample of a fixed size n. In case of sampling with replacement, the total number of samples that can be drawn is and when it comes to sampling without replacement of the sampling units, the total number of samples that can be drawn is ${}^{N}c_{n}$.

If we compute the value of a statistic, say mean, it is quite natural that the value of the sample mean may vary from sample to sample as the sampling units of one sample may be different from that of another sample. The variation in the values of a statistic is termed as "Sampling Fluctuations".

If it is possible to obtain the values of a statistic (T) from all the possible samples of a fixed sample size along with the corresponding probabilities, then we can arrange the values of the statistic, which is to be treated as a random variable, in the form of a probability distribution. Such a probability distribution is known as the sampling distribution of the statistic. The sampling distribution, just like a theoretical probability distribution possesses different characteristics. The mean of the statistic, as obtained from its sampling distribution, is known as "Expectation" and the standard deviation of the statistic T is known as the "Standard Error (SE)" of T. SE can be regarded as a measure of precision achieved by sampling. SE is inversely proportional to the square root of sample size. It can be shown that

$$SE\left(\bar{x}\right) = \frac{\sigma}{\sqrt{n}} \text{ for SRS WR}$$

$$= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad \text{for SRS WOR} \qquad (13.2.8)$$

Standard Error for Proporation

$$SE\,(p) = \sqrt{\frac{Pq}{n}} \qquad \text{for SRS WR}$$

$$\sqrt{\frac{Pq}{n}}\ \sqrt{\frac{N-n}{N-1}} \qquad \text{for SRS WOR} \qquad (13.2.9)$$

SRSWR and SRSWOR stand for simple random sampling with replacement and simple random sampling without replacement.

The factor $\sqrt{\frac{N-n}{N-1}}$ is known as finite population correction (fpc) or finite population multiplier and may be ignored as it tends to 1 if the sample size (n) is very large or the population under consideration is infinite when the parameters are unknown, they may be replaced by the corresponding statistic.

Illustrations

Example 13.2.1: A population comprises the following units: a, b, c, d, e. Draw all possible samples of size three without replacement.

Solution: Since in this case, sample size (n) = 3 and population size (N) = 5. the total number of possible samples without replacement = $^5c_3 = 10$

These are abc, abd, abe, acd, ace, ade, bcd, bce,bde,cde.

Example 13.2.2: A population comprises 3 member 1, 5, 3. Draw all possible samples of size two

(i) with replacement

(ii) without replacement

Find the sampling distribution of sample mean in both cases.

Solution: (i) With replacement :- Since n = 2 and N = 3, the total number of possible samples of size 2 with replacement = $3^2 = 9$.

These are exhibited along with the corresponding sample mean in table 15.1. Table 15.2 shows the sampling distribution of sample mean i.e., the probability distribution of $\bar{x}$.

Table 13.2.1

All possible samples of size 2 from a population comprising 3 units under WR scheme

Serial No.	Sample of size 2 with replacement	Sample mean ($\bar{x}$)
1	1, 1	1
2	1, 5	3
3	1, 3	2
4	5, 1	3
5	5, 5	5
6	5, 3	4
7	3, 1	2
8	3, 5	4
9	3, 3	3

Table 13.2.2

Sampling distribution of sample mean

$\bar{x}$	1	2	3	4	5	Total
P	1 / 9	2 / 9	3 / 9	2 / 9	1 / 9	1

(ii) without replacement: As N = 3 and n = 2, the total number of possible samples without replacement = $^NC_2 = {}^3C_2 = 3$.

Table 13.2.3

Possible samples of size 2 from a population of 3 units under WOR scheme

Serial No	Sample of size 2 without replacement	Sample mean ($\bar{x}$)
1	1 , 3	2
2	1 , 5	3
3	3 , 5	4

Table 13.2.4

Sampling distribution of mean

:	2	3	4	Total
P:	1 / 3	1/3	1/3	1

Example 13.2.3: Compute the standard deviation of sample mean for the last problem. Obtain the SE of sample mean applying 15.8 and show that they are equal.

Solution: We consider the following cases:

(i) with replacement :

Let U = $\bar{x}$ The sampling distribution of U is given by

U:	1	2	3	4	5
P:	1/9	2/9	3/9	2/9	1/9

$E(U) = \Sigma P_i U_i$

$= 1/9\times1 + 2/9\times2 + 3/9\times3 + 2/9\times4 + 1/9\times5 = 3$

$E(U^2) = \Sigma P_i U_i^2$

$= 1/9\times1^2 + 2/9\times2^2 + 3/9\times3^2 + 2/9\times4^2 + 1/9\times5^2$

$= 31/3$

$\therefore v(\bar{x}) = v(U) = E(U^2) - [E(U)]^2$

$= 31/3 - 3^2$

$= 4/3$

Hence $SE = \dfrac{2}{\sqrt{3}}$ (1)

Since the population comprises 3 units, namely 1, 5, and 3 we may take $X_1 = 1, X_2 = 5, X_3 = 3$

The population mean (m) is given by

$$\mu = \frac{\sum X_a}{N}$$

$$= \frac{1+5+3}{3} = 3$$

and the population variance $\sigma^2 = \frac{\sum (X_a - \mu)^2}{N}$

$$\frac{(1-3)^2 + (5-3)^2 + (3-3)^2}{3} = 8/3$$

Applying 15.8 we have, $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{3}} \times \frac{1}{\sqrt{2}} = \frac{2}{\sqrt{3}}$ (2)

Thus comparing (1) and (2), we are able to verify the validity of the formula.

(ii) without replacement :

In this case, the sampling distribution of V = is given by

V:	2	3	4
P:	1/3	1/3	1/3

E ($\bar{x}$) = E (V) = 1/3 × 2 +1/3 × 3 + 1/3 × 4

= 3

V ($\bar{x}$) = Var (V) = E (v^2) – $[E(v)]^2$

= $1/3 \times 2^2 + 1/3 \times 3^2 + 1/3 \times 4^2 - 3^2$

= 29/3 – 9

= 2/3

$$\therefore SE_{\bar{x}} = \frac{2}{\sqrt{3}}$$

Applying 13.2.8, we have

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{8}{\sqrt{3}} \times \frac{1}{\sqrt{2}} \times \frac{8}{\sqrt{3}} \times \frac{1}{\sqrt{2}} \times \sqrt{\frac{3-2}{3-1}} = \frac{2}{\sqrt{3}}$$

and thereby, we make the same conclusion as in the previous case.

13.2.6 TYPES OF SAMPLING

There are three different types of sampling which are

I. Probability Sampling

II. Non – Probability Sampling

III. Mixed Sampling

In the first type of sampling there is always a fixed, pre assigned probability for each member of the population to be a part of the sample taken from that population . When each member of the population has an equal chance to belong to the sample, the sampling scheme is known as Simple Random Sampling. Some important probability sampling other than simple random sampling are stratified sampling, Multi Stage sampling, Multi Phase Sampling, Cluster Sampling and so on. In non- probability sampling , no probability attached to the member of the population and as such it is based entirely on the judgement of the sampler. Non-probability sampling is also known as Purposive or Judgement Sampling. Mixed sampling is based partly on some probabilistic law and partly on some pre decided rule. Systematic sampling belongs to this category. Some important and commonly used sampling process are described now.

Simple Random Sampling (SRS)

When the units are selected independent of each other in such a way that each unit belonging to the population has an equal chance of being a part of the sample, the sampling is known as Simple random sampling or just random sampling. If the units are drawn one by one and each unit after selection is returned to the population before the next unit is being drawn so that the composition of the original population remains unchanged at any stage of the sampling then the sampling procedure is known as Simple Random Sampling with replacement. If, however, once the units selected from the population one by one are never returned to the population before the next drawing is made, then the sampling is known as sampling without replacement. The two sampling methods become almost identical if the population is infinite i.e. vary large or a very large sample is taken from the population. The best method of drawing simple random sample is to use random sampling numbers.

Simple random sampling is a very simple and effective method of drawing samples provided (i) the population is not very large (ii) the sample size is not very small and (iii) the population under consideration is not heterogeneous i.e. there is not much variability among the members forming the population. Simple random sampling is completely free from Sampler's biases. All the tests of significance are based on the concept of simple random sampling.

Stratified Sampling

If the population is large and heterogeneous, then we consider a somewhat, complicated sampling design known as stratified sampling which comprises dividing the population into a number of strata or sub-populations in such a way that there should be very little variations among the units comprising a stratum and maximum variation should occur among the different strata. The stratified sample consists of a number of sub samples, one from each stratum. Different sampling scheme may be applied to different strata and , in particular, if simple random sampling is applied for drawing units from all the strata, the sampling procedure is known as stratified random sampling. The purpose of stratified sampling are (i) to make representation of all the sub populations (ii) to provide an estimate of parameter not only for all the strata but also and overall estimate (iii) reduction of variability and thereby an increase in precision.

There are two types of allocation of sample size. When there is prior information that there is not much variation between the strata variances. We consider "Proportional allocation" or "Bowely's allocation where the sample sizes for different strata are taken as proportional to the population sizes. When the strata-variances differ significantly among themselves, we take recourse to "Neyman's allocation" where sample size vary jointly with population size and population standard deviation i.e. $n_i \mu N_iS_i$. Here n_i denotes the sample size for the i^{th} stratum, N_i and S_i being the corresponding population size and population standard deviation. In case of Bowley's allocation, we have $n_i \mu N_i$.

Stratified sampling is not advisable if (i) the population is not large (ii) some prior information is not available and (iii) there is not much heterogeneity among the units of population.

Multi Stage Sampling

In this type of complicated sampling, the population is supposed to compose of first stage sampling units, each of which in its turn is supposed to compose of second stage sampling units, each of which again in its turn is supposed to compose of third stage sampling units and so on till we reach the ultimate sampling unit.

Sampling also, in this type of sampling design, is carried out through stages. Firstly, only a number of first stage units is selected. For each of the selected first stage sampling units, a number of second stage sampling units is selected. The process is carried out until we select the ultimate sampling units. As an example of multi stage sampling, in order to find the extent of unemployment in India, we may take state, district, police station and household as the first stage, second stage, third stage and ultimate sampling units respectively.

The coverage in case of multistage sampling is quite large. It also saves computational labour and is cost-effective. It adds flexibility into the sampling process which is lacking in other sampling schemes. However, compared to stratified sampling, multistage sampling is likely to be less accurate.

Systematic Sampling

It refers to a sampling scheme where the units constituting the sample are selected at regular interval after selecting the very first unit at random i.e., with equal probability. Systematic sampling is partly probability sampling in the sense that the first unit of the systematic sample is selected probabilistically and partly non- probability sampling in the sense that the remaining units of the sample are selected according to a fixed rule which is non-probabilistic in nature.

If the population size N is a multiple of the sample size n i.e. N = nk, for a positive integer k which must be less than n, then the systematic sampling comprises selecting one of the first k units at random, usually by using random sampling number and thereby selecting every kth unit till the complete, adequate and updated sampling frame comprising all the members of the population is exhausted. This type of systematic sampling is known as "linear systematic sampling ". K is known as "sample interval".

However, if N is not a multiple of n, then we may write N = nk + p, p < k and as before, we select the first unit from 1 to k by using random sampling number and thereafter selecting every kth unit in a cyclic order till we get the sample of the required size n. This type of systematic sampling is known as "circular systematic sampling."

Systematic sampling is a very convenient method of sampling when a complete and updated sampling frame is available. It is less time consuming, less expensive and simple as compared to the other methods of sampling. However, systematic sampling has a severe drawback. If there is an unknown and undetected periodicity in the sampling frame and the sampling interval is a multiple of that period, then we are going to get a most biased sample, which, by no stretch of imagination, can represent the population under investigation. Furthermore, since it is not a probability sampling, no statistical inference can be drawn about population parameter.

Purposive or Judgement sampling

This type of sampling is dependent solely on the discretion of the sampler and he applies his own judgement based on his belief, prejudice, whims and interest to select the sample. Since this type of sampling is non-probabilistic, it is purely subjective and, as such, varies from person to person. No statistical hypothesis can be tested on the basis of a purposive sampling.

UNIT II EXERCISE

Set A

Answer the following questions. Each question carries one mark.

1. Sampling can be described as a statistical procedure

   (a) To infer about the unknown universe from a knowledge of any sample

   (b) To infer about the known universe from a knowledge of a sample drawn from it

   (c) To infer about the unknown universe from a knowledge of a random sample drawn from it

   (d) Both (a) and (b).

2. The Law of Statistical Regularity says that

   (a) Sample drawn from the population under discussion possesses the characteristics of the population

   (b) A large sample drawn at random from the population would posses the characteristics of the population

   (c) A large sample drawn at random from the population would possess the characteristics of the population on an average

   (d) An optimum level of efficiency can be attained at a minimum cost.

3. A sample survey is prone to

   (a) Sampling errors (b) Non-sampling errors

   (c) Either (a) or (b) (d) Both (a) and (b)

4. The population of roses in Salt Lake City is an example of

   (a) A finite population (b) An infinite population

   (c) A hypothetical population (d) An imaginary population.

5. Statistical decision about an unknown universe is taken on the basis of

   (a) Sample observations (b) A sampling frame

   (c) Sample survey (d) Complete enumeration

6. Random sampling implies

   (a) Haphazard sampling (b) Probability sampling

   (c) Systematic sampling (d) Sampling with the same probability for each unit.

7. A parameter is a characteristic of

   (a) Population (b) Sample

   (c) Both (a) and (b) (d) (a) or (b)

8. A statistic is

   (a) A function of sample observations (b) A function of population units

   (c) A characteristic of a population (d) A part of a population.

9. Sampling Fluctuations may be described as

   (a) The variation in the values of a statistic

   (b) The variation in the values of a sample

   (c) The differences in the values of a parameter

   (d) The variation in the values of observations.

10. The sampling distribution is

    (a) The distribution of sample observations

    (b) The distribution of random samples

    (c) The distribution of a parameter

    (d) The probability distribution of a statistic.

11. Standard error can be described as

    (a) The error committed in sampling

    (b) The error committed in sample survey

    (c) The error committed in estimating a parameter

    (d) Standard deviation of a statistic.

12. A measure of precision obtained by sampling is given by

    (a) Standard error (b) Sampling fluctuation

    (c) Sampling distribution (d) Expectation.

13. As the sample size increases, standard error

    (a) Increases (b) Decreases

    (c) Remains constant (d) Decreases proportionately.

14. If from a population with 25 members, a random sample without replacement of 2 members is taken, the number of all such samples is

    (a) 300 (b) 625

    (c) 50 (d) 600

15. A population comprises 5 members. The number of all possible samples of size 2 that can be drawn from it with replacement is

    (a) 100 (b) 15

    (c) 125 (d) 25

16. Simple random sampling is very effective if

    (a) The population is not very large

    (b) The population is not much heterogeneous

    (c) The population is partitioned into several sections.

    (d) Both (a) and (b)

17. Simple random sampling is

    (a) A probabilistic sampling (b) A non- probabilistic sampling

    (c) A mixed sampling (d) Both (b) and (c).

18. According to Neyman's allocation, in stratified sampling

    (a) Sample size is proportional to the population size

    (b) Sample size is proportional to the sample SD

    (c) Sample size is proportional to the sample variance

    (d) Population size is proportional to the sample variance.

19. Which sampling provides separate estimates for population means for different segments and also an over all estimate?

    (a) Multistage sampling (b) Stratified sampling

    (c) Simple random sampling (d) Systematic sampling

20. Which sampling adds flexibility to the sampling process?

    (a) Simple random sampling (b) Multistage sampling

    (c) Stratified sampling (d) Systematic sampling

21. Which sampling is affected most if the sampling frame contains an undetected periodicity?

    (a) Simple random sampling (b) Stratified sampling

    (c) Multistage sampling (d) Systematic sampling

22. Which sampling is subjected to the discretion of the sampler?

    (a) Systematic sampling (b) Simple random sampling

    (c) Purposive sampling (d) Quota sampling.

23. If a random sample of size 2 with replacement is taken from the population containing the units 3,6 and 1, then the samples would be

    (a) (3,6),(3,1),(6,1)

    (b) (3,3),(6,6),(1,1)

    (c) (3,3),(3,6),(3,1),(6,6),(6,3),(6,1),(1,1),(1,3),(1,6)

    (d) (1,1),(1,3),(1,6),(6,1),(6,2),(6,3),(6,6),(1,6),(1,1)

24. If a random sample of size two is taken without replacement from a population containing the units a,b,c and d then the possible samples are

    (a) (a, b),(a, c),(a, d) (b) (a, b),(b, c), (c, d)

    (c) (a, b), (b, a), (a, c),(c,a), (a, d), (d, a) (d) (a, b), (a, c), (a, d), (b, c), (b, d), (c,d)

ANSWERS

Set A

1.	(c)	2.	(c)	3.	(d)	4.	(b)	5.	(a)	6.	(d)
7.	(a)	8.	(a)	9.	(a)	10.	(d)	11.	(d)	12.	(a)
13.	(b)	14.	(a)	15.	(d)	16.	(b)	17.	(a)	18.	(a)
19.	(b)	20.	(d)	21.	(d)	22.	(c)	23.	(c)	24.	(d)