2 Representation Of Data

2.1 Introduction

Most scientific experiments are conducted in an attempt to answer some specific questions and generally result in the collection of data in the form of batches of numbers, usually referred to as sample. To extract information from the sample, there is need to organize and summarize the collected data. There are many ways of describing a sample. Commonly, we use either, a graph or a small set of numbers which summarize some properties of the sample such as its centre and spread.

2.2 Classification

The term classification is the process of arranging observations into different classes or categories according to some common characteristics.

The data may be classified by one or more characteristics at a time. When data is classified according to one characteristic, it is called one-way classification. When the data is classified by two characteristics at a time, it is called two-way classification. Similarly, the data classified by three characteristics is called three-way classification.

2.3 Tabulation

The process of making tables or arranging data into rows and columns is called tabulation. Tabulation may be simple, double, triple or complex depending upon the number of characteristics involved. Tables are the most common form of documentation used by the scientists.

2.3.1 Construction of tables

Following are the parts of table out of which first four are main part.

i) **Title**: A title is the heading at top of the table. The title should be brief and self explanatory. It describes the contents of the table.

ii) Column Captions and Box head: The headings for different columns are called column captions and this part of column captions is called Box-head. The column captions should be brief, clear and arranged in order of importance.

iii) Row Captions and Stub: The headings for different rows are called row captions and the part of the table containing row captions is called Stub. Row captions should be brief, clear and arranged in order of importance.

iv) Body of the Table and Arrangements of the data: The entries in different cells of column and rows in a table are called body of the table. It is the main part of the table. The data may be arranged qualitatively, quantitatively, chronologically, geographically or alphabetically.

v) Source Note: Source notes are given at the end of the table which indicate the compiling agency, publication, the data and page of the publication.

vi) Spacing and ruling: To enhance the effectiveness of a table, spacing and ruling is used. It is also used to separate certain items in the table. Thick or double lines or single lines are used to separate row captions and column captions. To indicate no entry in a cell of the table, dots (...) or dashes(---) are used. Zeroes are not used in a table for this purpose.

vii) Prefatory Notes and Footnotes: The prefatory note is given after the title of the table and the footnotes are given at the bottom of the table. Both are used to explain the contents of the table. The footnotes are usually indicated by* * *.

2.4 Frequency Distribution

To extract information from a data set, first and important step is to present it in a compact form. A frequency distribution is a compact form of data in a table which displays the categories of observations according to their magnitudes and frequencies such that the similar or identical numerical values are grouped together. The categories arc also known as groups, class intervals or classes. The number of values falling in a particular category is called the frequency of that category. It is usually denoted by *f*.

The relative frequency, denoted by r.f of a category is the proportion of observed frequency to the total frequency and is obtained by dividing observed frequency by the total frequency. The sum of the relative frequencies should be one (1) except for rounding error. The relative frequencies are important for making comparisons between two or more distributions. Otherwise, the different sample sizes of the data sets may distort comparisons.

The frequency distribution may be made for continuous data, discrete data and categorical data.

Following steps are taken into account while making a frequency distribution for continuous data:

i) Calculate range of the data, where

Range = maximum value in the data. – minimum value in the data

ii Decide about the number of classes. The minimum number of classes may be determined by the formula

Number of classes c = $1+ 3.3 \log(n)$ (2.1)

or $c = \sqrt{n}$ (approximately) (2.2)

where n is the total number of observations in the data.

This gives roughly the number of classes. There are certain other formulae suggested to decide the number of classes. Classes are the groups of data values constituting the frequency table. Usually, the classes of equal width are defined by the numerical limits or boundaries. Each class has a starting point called its lower limit and its end point called its upper limit.

The class limits are the end points of the class intervals both included in class interval. It is convenient to choose the end points of the class interval so that no observation falls on them. This can be obtained by expressing the end points to one more place of decimal than the observations themselves. For this purpose, class limits are usually converted to class boundaries to achieve continuity in the grouped data. This is done by expressing the upper limit of the first class to one more decimal place without changing the width of the class and starting the second class from the same value as is the end of the first class and so on. The upper values in the classes are included in the next class so that the classes do not overlap.

The number of classes are important. Neither we should make too few wide classes in which most of the variation in the data is lost nor we should have too many narrow classes in which the real values in the data are hardly grouped.

iii) Decide about width of the class. It is usually abbreviated by h and is obtained by the following relation:

$$h = \frac{\text{range}}{\text{number of classes}}$$

$$h = \frac{R}{c} \text{ (approximately)} \qquad (2.3)$$

It should be noted that always a convenient near number is chosen and it is not necessary to follow the rules of rounding because we are only grouping the data.

iv) The decision about the starting point of the first class is arbitrary usually, it is started before the minimum value in such a way that the mid point, the average of lower and upper class limits of the first class is properly placed.

v) Now, an observation is taken and a mark of vertical bar is made for a class it belongs. A running tally is kept till the last observation. The tally count 𝍸 indicates five.

Example 2.1: Student - Height Data

The height (in cms) of 30 students measured at the time of registration is given by

91, 89, 88, 87, 89, 91, 87, 92, 90, 98, 95, 97, 96, 100, 101, 96, 98, 99, 98, 100, 102, 99, 101, 105, 103, 107, 105, 106, 107, 112.

Make a suitable frequency distribution.

Solution: To construct a frequency distribution proceedas to follows:

i) Range = Maximum value minimum value (2.4)

In this data maximum value is 112 and minimum value is 87.

So, Range = 112 87 = 25

ii) Approximate number of classes or class intervals are number of classes c is given by

$= 1 + 3.3 \log(30)$

$= 1 + 3.3\ (1.4771)$

$= 5{:}87443$

$= 6$ (approximately)

iii) Width of the class interval (h) = range / number of classes

$$= \frac{25}{6}$$

$= 4.167$

$= 5$ (approximately)

5 is chosen for convenience, one may take 4 if he / she wishes so.

iv) Minimum value is 87, we start the first class from 86 with width of the class as 5, so, our first class is 86-90 with mid point 88, the average of lower and upper class limits i.e., (86+90)/2=88. Similarly, other classes are 91-95, 96-100, ..., 111-115. It is clear that maximum value 112 is included in the last class.

It is convenient to choose the end points of the class interval so that no observation falls on them. This can be obtained by expressing the end points to one more place of decimal than the observations themselves. Therefore, suitable class boundaries for this data would be 85.5 – 90.5, 90.5 – 95.5, . . ., 110.5 – 115.5. In the class boundaries, the upper values in the classes are included in the next class so that the classes are mutually exclusive i.e., 90.5 is the upper value of the first class and is lower value of the second class. In counting this would be included in the second class interval.

The class centres Y_i's are the middles of the classes. The class centres are also known as mid or middle points and are obtained either by averaging class limits or class boundaries i.e Y_i is the middle of the first class

$Y_1 = (85.5 + 90.5)/2 = 88$

The other mid points are 93,98, ... , 113 respectively.

v) Starting from first observation, all the 30 observations are assigned to the classes they belong. The first observation 87 falls in the first class 86-90, a tally mark is made in the tally column against this class. The second observation 90 belongs to the first class 86–90, a tally mark is made in tally column against this class and so on, the last observation 112 belongs to the last class 111–115. The number of tally marks in the tally column against each class gives the frequency of that class. The frequency distribution is given in Table 2.1

Table 2.1: Tally count and frequency distribution for the example 2.1.

Class limits	Class boundaries	y_i	Tally	Frequency
86-90	85.5-90.5	88	𝍸 |	6
91-95	90.5-95.5	93	||||	4
96-100	95.5-100.5	98	𝍸 𝍸	10
101-105	100.5-105.5	103	𝍸 |	6
106-110	105.5-110.5	108	|||	3
111-115	110.5-115.5	113	|	1
			Total	30

It is clear from the frequency table that 6 students have height between 85.5 and 90.5 cms, 4 students have height between 90.5 and 95.5 cms and so on and I student has height between 110.5 and 115.5 cms.

The relative frequency for a class can be computed by dividing its frequency by the total frequency. The frequency distribution with relative frequencies is given in Table 2.2.

Table 2.2: Frequency distribution with relative frequencies

Class boundaries	f	$r.f$
85.5-90.5	6	6/30 = 0.200
90.5-95.5	4	4/30 = 0.133
95.5-100.5	10	10/30 = 0.333
100.5-105.5	6	6/30 = 0.200
105.5-110.5	3	3/30 = 0.100
110.5-115.5	1	1/30 = 0.033
Total	30	1.000

It should be noted that the sum of the relative frequencies is one except for rounding error.

For discrete data: In case of discrete data, each observation is a whole number. So, while making a frequency distribution, the possible values are written in a column and a tally count of each value is made for the data. The number of tally count for each value is its frequency. The corresponding relative frequency is obtained by dividing each frequency by the total number of observations. The sum of the relative frequencies should be 1 except for rounding error.

For categorical data: In case of categorical data, the categories are placed in a column and a tally count is made for each category going through the data set which gives the frequency of each category.

Example 2.2: The observations about the number of rotten potatoes from twenty equal sized samples taken from a store are available as follows:

1, 2, 4, 3, 0, 1, 2, 3, 1, 1, 0, 2, 1, 0, 2, 3, 0, 0, 1, 3

Make a frequency table

Solution: The tally count and frequency table is made by going through each observation of the data and for each observation making a mark, vertical bar | against the appropriate value of the variable. In this data, the values of the variable vary from 0 to 4. These are written in a column and a tally count is kept going through the whole data. The resulting frequency distribution is given in Table 2.3.

Table 2.3: Tally count and frequency distribution for the example 2.2.

Number of rotten potatoes	Tally	*f*	r.f
0	𝍸	5	5/20 = 0.25
1	𝍸 𝍷	6	6/20 = 0.30
2	𝍷𝍷𝍷𝍷	4	4/50 = 0.20
3	𝍷𝍷𝍷𝍷	4	4/20 = 0.20
4	𝍷	1	1/20 = 0.05
Total		$\Sigma f = 20$	1.00

If the range of observations in the data is large, the same method is adopted as has been explained for the continuous data.

Open-end classes

In connection with the frequency tables, the term **open end classes** is sometimes used. It means that in a frequency table, either the lower limit of the Ist class or the upper limit of the last class is not a fixed number. It may happen that both of these are not fixed numbers. The frequency tables with open end classes are formed in some practical situations. The frequency table about the age of people in a certain locality is given in the adjacent table:

Age group	Frequency
Below 5	20
5 –14	37
15 – 24	67
25 – 34	90
35 – 44	87
45 – 54	60
55 – 64	55
65 – 74	45
75 and above	20

2.5 Cumulative Frequency Distribution

A cumulative frequency distribution is a table that displays class intervals and the corresponding cumulative frequencies. The cumulative frequency is denoted by c.f and for a class interval it is obtained by adding the frequencies of all the preceding classes including that class. It indicates the total number of values less than or equal to the upper limit of that class.

The relative frequencies, cumulative frequencies and cumulative relative frequencies for data for the example 2.1 are given in Table 2.4.

Table 2.4: Cumulative distribution for the example 2.1.

Class boundaries	*f*	*r.f*	*c.f*	c.r.f
85.5-90.5	6	6/30	6	6/30 = 0.200
90.5-95.5	4	4/30	6+4 =10	10/30 = 0.333
95.5-100.5	10	10/30	10+10=20	20/30 = 0.667
100.5-105.5	6	6/30	20+6 =26	26/30 = 0.867
105.5-110.5	3	3/30	26+3 =29	29/30 = 0.967
110.5-115.5	1	1/30	29+1 =30	30/30 = 1.000

As the cumulative frequency of a class indicates the total number of values less than or equal to the upper limit of that class, so, the cumulative frequency of 20 for a class 95.5 – 100.5 means that 20 values are less than 100.5 and similarly, the cumulative frequency of the last class 110.5 – 115.5 is 30 indicating that 30 values are less than 115.5.

If we want to compare two or more distributions, we compute relative cumulative frequencies or percentage cumulative frequencies because these would be comparable. Otherwise, the differences in sample sizes will distort comparisons.

The cumulative relative frequencies which are the proportions of the cumulative frequency, denoted by *c.r.f* are obtained by dividing the cumulative frequency by the total frequency. The *c.r.f* of a class can also be obtained by adding the relative frequencies of the preceding classes including that class. As cumulative relative frequencies are proportions, the multiplication by 100 gives corresponding percentage cumulative frequencies.

The relative cumulative frequencies are obtained by dividing the cumulative frequency by the total frequency i.e., for the first class interval it is 6/30 = 0.2, for the second class interval it is 10/30 = 0.33 and so on. The percentage cumulative frequency for each class can be obtained by multiplying its cumulative relative frequency by 100. The percentage cumulative frequency for 0.200 is (0.200) (100) = 20. The percentage cumulative frequency for 0.0333 is (0.333) (100) = 33.3 and so on, the percentage cumulative frequency for 1.000 is (1.000) (100) = 100.

2.5.1 Cumulative frequency distribution for discrete data

The cumulative frequency distribution for the discrete data is obtained in the same way as for the continuous data i.e., the cumulative frequency of a class is obtained simply by adding the preceding frequencies including the frequency for that class. The relative frequencies and the cumulative frequencies for the data of example 2.2. are given below in Table 2.5.

Table 2.5: Cumulative frequency distribution of the example 2.2.

Number of rotten potatoes	*f*	*r.f*	*c.f*	**c.r.f**
0	5	5/20	5	5/20
1	6	6/20	11	11/20
2	4	4/20	15	15/20
3	4	4/20	19	19/20
4	1	1/20	20	20/20
Total	20	1.00		

Example 2.4: Find out the relative frequency distribution for the following data. Where *x* denotes the number of hours worked in a day by a person in a locality of 265 people.

x	6	7	8	9	10	11	12	13
f	24	66	80	48	28	14	4	1

Solution:

x	**frequency**	**Relative frequency**
6	24	24/265 = 0.09
7	66	66/265 = 0.25
8	80	80/265 = 0.30
9	48	48/265 = 0.18
10	28	28/265 = 0.11
11	14	14/265 = 0.05
12	4	4/265 = 0.02
13	1	1/265 = 0.00
Total	265	1.00

Example 2.5: Find out the relative cumulative frequency distribution from the following data of example 2.4.

x	6	7	8	9	10	11	12	13
f	24	66	80	48	28	14	4	1

Solution:

x	*f*	**c.f**	**c.r.f.**
6	24	24	24/265 = 0.09
7	66	24 + 66 = 90	90/265 = 0.25
8	80	90 + 80 = 90	170/265 = 0.30
9	48	170 + 48 = 90	218/265 = 0.18
10	28	218 + 28 = 90	246/265 = 0.11
11	14	246 + 14 = 90	260/265 = 0.05
12	4	260 + 4 = 90	264/265 = 0.02
13	1	264 + 1 = 90	265/265 = 0.00
Total	265	---------	---------

2.6 Graphic representation of Data

There are reasons for drawing graphs. The most compelling being that one simple graph says more than twenty pages of prose. Many graphs just represent a summary of data that has been collected to support a particular theory. It is usually suggested that the graphic representation of the data should be looked at before proceeding for the format statistical analysis.

Common uses of graphs

i) Graphs are useful for checking assumptions made about the data, i.e., the probability distribution assumed.

ii) The graphs provide a useful subjective impression as to what the results of the format analysis should be. This serves as a check on calculations and statistical methodology used. Always believe your common sense before arithmetic calculations because for some problems calculations will be obvious from a graph.

iii) Graphs often suggest the form of a statistical analysis to be carried out, particularly, the graph of model fitted to the data.

iv) Graphs give a visual representation of the data or the results of statistical analysis to the reader which are usually easily understandable and more attractive.

v) Some graphs are useful for checking the variability in the observations and outliers can be easily detected.

Outliers are the data values which are highly inconsistent with the main body of the data. These may arise because of mistakes in copying, coding or may be some values that are different from the rest of the data just their own.

Important points for drawing graphs

There are a number of points worth keeping in mind when drawing graphs. The most important of these are:

i) clearly label axis with the names of the variables and units of measurement.

ii) keep the units along each axis uniform regardless of the scales chosen for axis.

iii) keep the diagram simple. Avoid any unnecessary details.

iv) a clear and concise title should be chosen to make the graph meaningful

v) if the data on different graphs are to be measured, always use identical scales.

vi) in the scatter plots, don't join up the dots. This makes it likely that you will see apparent patterns in any random scatter of points.

The general approach, which should be used to analyze the data, is as follows:

i) construct an appropriate diagram and summary of the data and come to an initial impression concerning the question posed. This is known as exploratory data analysis.

ii) follow this up with an appropriate formal analysis of data.

iii) compare the results of the formal analysis with your initial impression, and worry if they differ greatly.

The methods described here are appropriate for data on a **single variable**. Usually, the data will be measured on a continuous scale or at least if this is not the case, then the set of possible values will be reasonably large. The types of graphs commonly used are given below:

2.6.1 Simple Bar Diagram

To get an impression of the distribution of a discrete or categorical data set, it is usual to represent it by a bar diagram. To construct a bar diagram, the values of the variable or categories are taken along x-axis and a bar with height equal to its frequency is drawn on each category.

The first s.ep is to make a tally count of the data to help us to make a frequency distribution. The procedure is explained on the example 2.2. The frequency distribution is given in table 2.5 (a).

Table 2.5 (a): Frequency distribution for the data 2.2.

Number of rotten potatoes	Tally	Frequency
0	𝍸	5
1	𝍸 |	6
2	||||	4
3	||||	4
4	|	1
	Total	20

To construct a bar diagram, the number of rotten potatoes are taken along x-axis. The rotten potatoes vary from 0 to 4, so we mark the x-axis with 0,1,2,3 and 4. The value 0 has frequency 5, so a bar of height 5 is drawn along y-axis at point 0 on x-axis. Similarly a bar of height 6 is drawn along y-axis at point 1 on x-axis; a bar of height 4 is drawn along y-axis at point 2 and 3 on x-axis and finally a bar of height 1 is drawn on point 4. It is shown in figure 2.1.

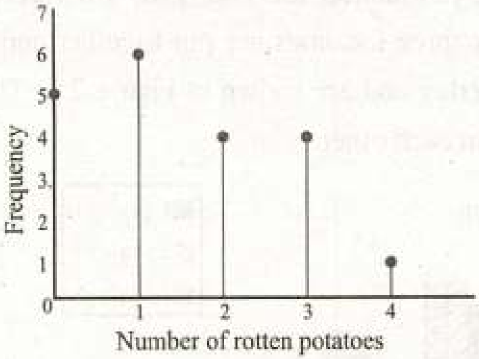

Figure 2.1: Bar diagram of rotten potatoes.

The gaps between the bars in the bar chart emphasize the gaps between the values that the discrete variable can take.

2.6.2 Multiple Bar Diagram

It is an extension of the simple bar diagram and is used to represent two or more related sets of data in the form of groups of simple bars. Its main purpose is to compare same characteristics of a variable.

Example 2.6: Following data is about the production of wheat in different localities of the Punjab for years 1987 to 1989.

Production in Kg. (thousands)			
Year	**1987**	**1988**	**1989**
Locality I	500	600	200
Locality II	600	700	400
Locality III	800	700	500

Draw an appropriate diagram for this data?

Solution: The appropriate diagram seems to be a multiple bar diagram because 3 bars, one of each locality, for each year will make the comparison between the production of three localities overtime.

To draw multiple bar diagram, the years are taken along x-axis and for each year three bars are drawn along y-axis, one for each locality to indicate the production. The bars showing production for year 1987 for three localities are put together, the bars for 1988 for three localities are put together and the bars for 1989 for three localities are put together and are shown in Figure 2.2. The bars are shaded individually to differentiate from each other.

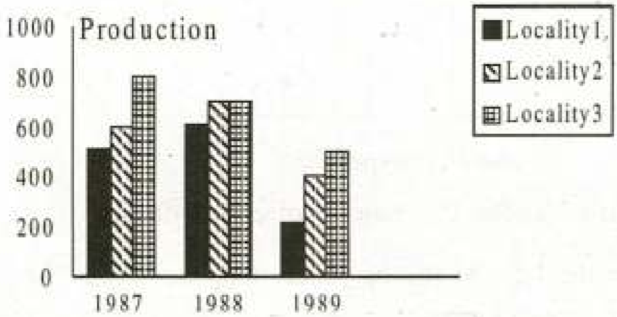

Figure 2.2: Multiple bar diagram of different localities.

2.6.3 Sub-divided Bar Diagram (Component Bar Diagram)

There are certain situations where the simple bar diagram represents the totals and it is possible to divide it further into different segments. For example, if the simple bar denotes the total population of insects caught in a field then it is possible to sub-divide it into male and female proportions.

Example 2.7: There were 500 people of blood group A (kind 1). 300 of blood group B (kind 2) and 400 of blood group O (kind 3). After classification, it was observed that for kind 1 there were 200 females, for kind 2 there were 100 females and for kind 3 there were 200 females.

Draw an appropriate diagram for this data?

Solution: The sub-divided bar diagram is useful in this situation to represent the number of males and females in each category. First construct simple bars and then divide it according to the number of males and females in each blood group category.

The simple bar and sub-divided bar diagrams are shown in Figure 2.3(a) and 2.3 (b) respectively.

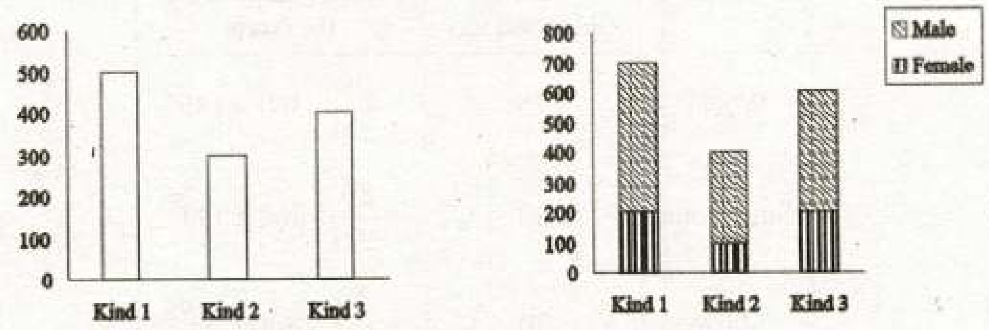

Figure 2.3(a)

Figure 2.3(b)

Figure 2.3: Simple and sub-divided bar diagrams

2.6.4 Pie Diagram

The Pie chart or Pie diagram is a division of a circular region into different sectors. It is constructed by dividing the total angle of a circle of 360 degrees into different components. The angle Q for each sector is obtained by the relation:

$$Q = \frac{\text{Component Part}}{\text{Total}} \times 360$$

Each sector is shaded with different colours or marks so that they look separate from each other.

It is a useful way of displaying the data where division of a whole into component parts needs to be presented. It can also be used to compare such divisions at different times.

Example 2.8: The data are available regarding total production of urea fertilizer and its use on different crops. Total production of urea is 200 thousand (kg) and its consumption for different crops wheat, sugarcane, maize and lentils is 75,80, 30 and 15 thousands (kg) respectively. Make an appropriate diagram to represent these data?

Solution: The appropriate diagram seems to be a pie chart because we have to present a whole into 4 component parts. To construct a pie chart, we calculate the proportionate arc of circle, i.e.,

$$\frac{75}{200} \times 360 = 135,\ \frac{80}{200} \times 360 = 144,\ \frac{30}{200} \times 360 = 54,\ \frac{15}{200} \times 360 = 27$$

Proportionate are of a circle (in degrees) f or different crops are in Table 2.6.

Table 2.6: Proportionate arc of a circle for crops.

Crops	**Fertilizer (thousand kg)**	**Proportionate arc of the circle**
Wheat	75	$\frac{75}{200}\times360^{\circ}=135^{\circ}$
Sugarcane	80	$\frac{80}{200}\times360^{\circ}=144^{\circ}$
Maize	30	$\frac{30}{200}\times360^{\circ}=54^{\circ}$
Lentils	15	$\frac{15}{200}\times360^{\circ}=27^{\circ}$
Total	200	360°

Draw a circle of an appropriate radius, make the angles clockwise or anti-clockwise with the help of protractor or any other device i.e., for wheat make an angle of 135 degrees, for sugarcane an angle of 144 degrees, for maize an angle of 54 degrees and for lentils an angle of 27 degrees and hence circular region is divided into 4 sectors. Shade each sector with different colours or marks so that they look separate from each other. The pie diagram is given in figure 2.4.

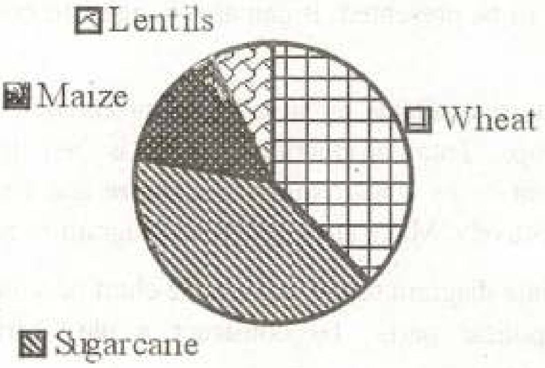

Figure 2.4

2.6.5 Histogram

A histogram is a useful graphic representation of data to get a visual ımpression about its distribution. It is constructed from the grouped data by taking class boundaries along x-axis and the corresponding frequencies along y-axis. If the data are in the ungrouped form, then first step is to arrange the data in the form of the grouped frequency distribution before making a histogram. The histogram may be constructed for the following two types of qualitative data.

i) Continuous grouped data ii) Discrete grouped data

Histogram: For Continuous grouped data

For the continuous grouped data the frequency distribution may be with equal class width or with unequal class width depending upon the nature of the data. To draw a histogram from the continuous grouped frequency distribution, the following steps are taken. The first two steps are common for equal / unequal class width but the third step is different.

i) Mark class boundaries of the classes along x-axis.

ii) Mark frequencies along y-axis.

iii) Draw a rectangle for each class such that the height of each rectangle is proportional to the frequency corresponding to that class. This is the case when classes are of equal width as they often are.

iv) If the classes are of unequal width, then the area instead of height of each rectangle is proportional to the frequency corresponding to that class and the height of each rectangle is obtained by dividing the frequency of the class by the width of that class.

Histogram: For Equal class Interval (data of table 2.1)

To construct a histogram, take following steps:

i) Mark the class boundaries 85.5 – 90.5, 90.5 – 95.5, ….., 110.5 – 115.5 on x-axis.

ii) Maximum frequency is 10, so label y-axis from 0 to 10.

iii) The frequency of the Ist class is 6, so the rectangle is raised uptill 6, the rectangle of the second class is raised uptill 4 and so on, the last rectangle is raised to height 1. The histogram is shown in Figure 2.5.

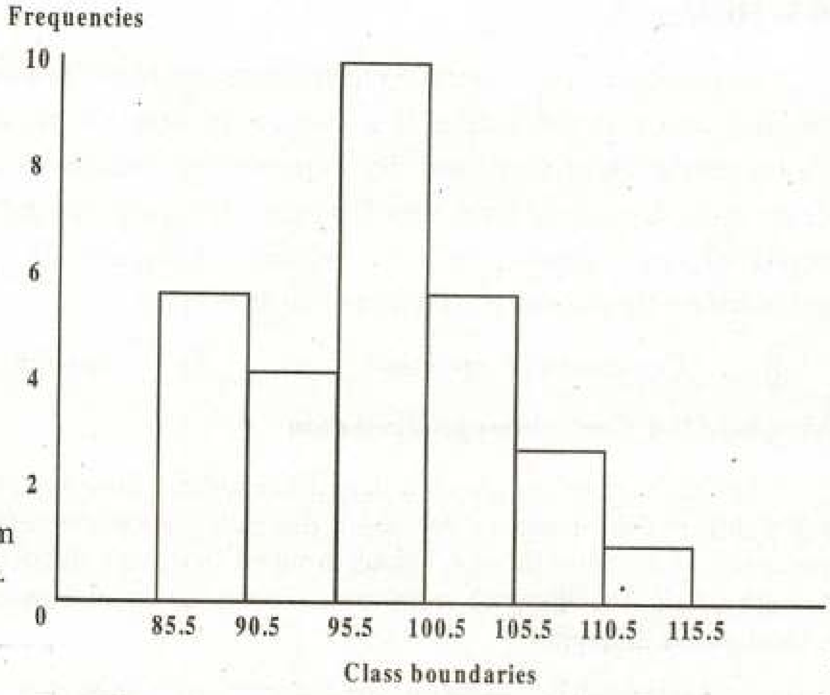

Figure 2.5: Histogram of student height data.

It may be noted that the area under a histogram can be calculated by adding up the areas of all the rectangles that constitute the histogram. The area of one rectangle is obtained by the multiplication of width of the class by the corresponding frequency i.e.,

Area of a single rectangle = width of the class × frequency of the class.

The area of above histogram is

$5(6) + 5(4) + 5(10) + 5(6) + 5(3) + 5(1) = 150$

Histogram: For Unequal Class Intervals

The principal reason for making histogram with equal class intervals is that the frequencies from class to class are directly comparable. However, there may be situations where unequal class intervals are appropriate. Firstly, in highly skewed distribution and secondly, the grouping of similar cases. In such situations while constructing a histogram, the width of the classes should be taken into account because the area of each rectangle is proportional to the frequency. This can be achieved by adjusting the heights of the rectangles. The height of each rectangle is obtained by dividing the frequency of the class by the width of that class.

Example 2.9: The frequency distribution of ages (in years) of 51 members of a locality is available adjacent table. Draw a histogram for this data?

Classes	Frequency
2 – 4	5
4 – 8	10
8 – 12	12
12 – 16	14
16 – 22	6
22 – 30	4

Solution: A look at this data, indicates that the width of the class intervals is not equal as first class has width 2; second, third, fourth classes have width 4, fifth has 6 and the last class has width 8 so, there is need to adjust the heights of the rectangles i.e., for the first class we have 2 as width of class and 5 as frequency, so height of the first class is 5/2=2.5. Similarly, for the others 10/4=2.5, 12/4=3, 14/4=3.5, 6/6=1.0, 4/8=0.5. These heights are also called adjusted frequencies. The width of the class and corresponding height of rectangles are in table 2.7

Table 2.7: Frequency distribution of the example 2.6 with adjusted heights.

Classes	Frequency	Width of class	Height of rectangle (adjusted frequency)
2 – 4	5	2	5/2 = 2.5
4 – 8	10	4	10/4 = 2.5
8 – 12	12	4	12/4 = 3.0
12 – 16	14	4	14/4 = 3.5
16 – 22	6	6	6/6 = 1.0
22 – 30	4	8	4/8 = 0.5

Taking class boundaries along *x*-axis and corresponding adjusted frequencies along *y*-axis, rectangles are drawn and the histogram is given in Figure 2.6.

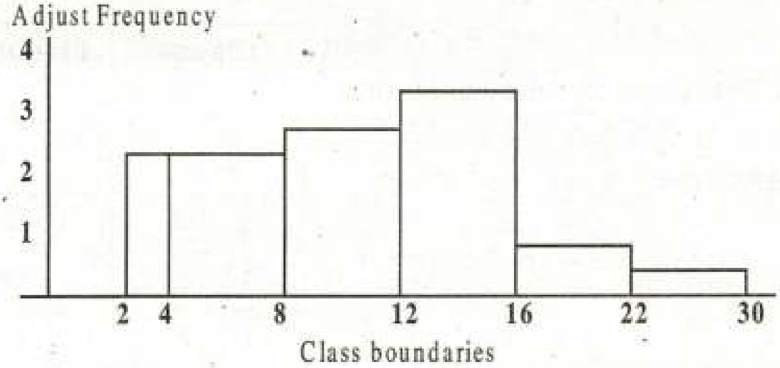

Figure 2.6: Histogram for unequal class intervals.

Histogram: discrete data

It should be noted that bar graphs are usually drawn for discrete and categorical data but there are some situations where there is need to make approximations, the histogram may be constructed.

To construct a histogram for discrete grouped data, following steps are taken:

i) mark possible values along *x*-axis.

ii) mark frequencies along *y*-axis.

iii) draw a rectangle centered on each value with equal width on each side possibly 0.5 to either side of the value.

The procedure is explained for the example 2.2.

The rotten potatoes vary from 0 to 4 so, *x*-axis is marked 0, 1, 2, 3, 4. The maximum frequency is 6 so, *y*-axis is marked from 0 to 6. A rectangle is drawn centered on each value whose height is equal to the corresponding frequency. The resulting diagram for the data is given in Figure 2.7.

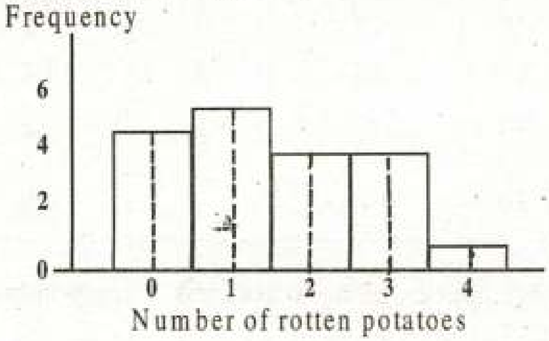

Figure 2.7: Histogram for rotten potatoes.

Advantages: The advantages of the histogram as compared to the unprocessed data are:

i) it gives range of the data.

ii) it gives location of the data.

iii) it gives clue about the skewness of the data.

iv) it gives information about the out of control situation.

2.6.6 Frequency polygon and frequency curve

A frequency polygon is a closed geometric figure used to display a frequency distribution graphically. Following steps are taken to make a frequency polygon from a frequency distribution.

i) Calculate mid values of the class boundaries.

ii) Mark these mid values along *x*-axis.

iii) Mark the frequencies along *y*-axis.

iv) Mark corresponding frequencies against each mid point, join them and extend it to *x*-axis.

It can also be obtained by joining the upper mid points of the rectangles of a histogram and extending ends to the *x*-axis. The distance from the *x*-axis to the plotted point corresponds to the frequency of the class. The frequency polygon smoothed is called frequency curve, which is useful to have a visual impression about the data i.e., it may help to know about the symmetry or skewness of the data. If we are interested to compare two distributions number of observations less than this is zero. For the grouped data of the example 2.1, it is shown in figure 2.8. It is clear that cumulative frequency polygon is an increasing function which starts from the lower class boundary of the first class at zero height and ends at the upper boundary of the last class with height equal to total frequency.

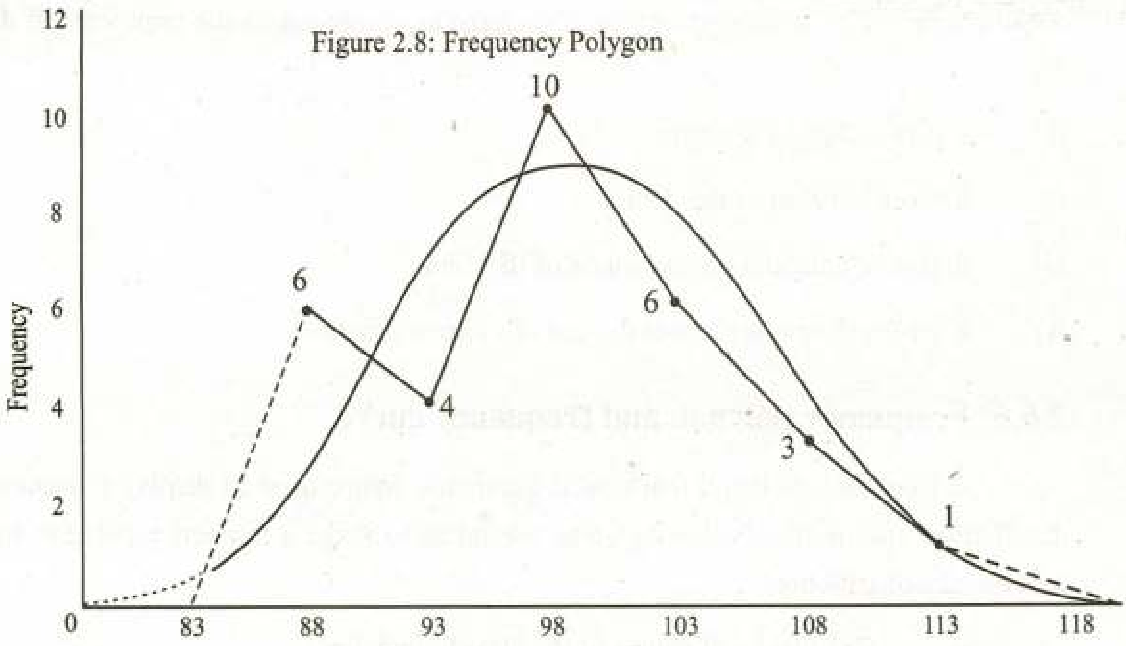

Figure 2.8: Frequency Polygon

This graph may be drawn using upper class boundaries and cumultative relative frequencies in which case it is called cumulative frequency function or polygon and can be used to locate certain values. It can be used to locate the quartiles or percentiles of the data. The figure 2.9 indicates the observation corresponding to the c.r.f. of 0.25

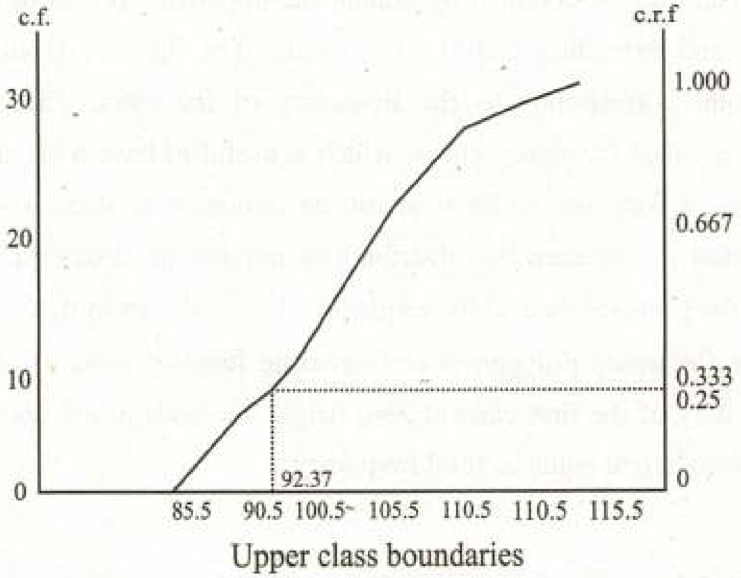

Figure 2.9: Cumulative frequency polygon of students height data.

We multiply the cumulative relative frequencies by hundred to get corresponding percentage cumulative frequencies. The c.r.f. polygon becomes the percentages along y-axis instead of c.r.f. The graph on the right hand side of figure 2.9 becomes percentage cumulative frequency polygon if we replace 0.200 by 20, the percentage cumulative frequency for 0.2000 as (0.200) (100) = 20: 0.333 by 33.3, the percentage cumulative frequency for 0.333 as (0.333) (100) = 33.3 and so on 1.00 by 100, the percentage cumulative frequency for 1.

Consider the following steps to draw a cumulative frequency polygon for discrete variable.

i) Choose horizontal axis on a graph paper and mark the data points from the smallest to the largest.

ii) Mark the vertical axis from zero to total frequency.

iii) Make a vertical jump of height equal to its frequency at the first point. Move horizontally from the top of this point until you are exactly above the second data point and make a jump equal to its frequency at the second point. Repeat this for all the data values. The cumulative frequency polygon for data of the example 2.2 is shown in Figure 2.10.

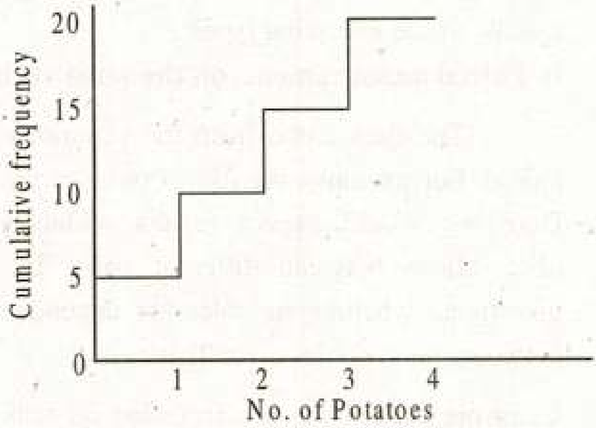

Figure 2.10: Cumulative frequency polygon.

It is clear from Figure 2.10 that there is a jump at each data value whose height is equal to its frequency. The cumulative frequency polygon is flat and horizontal between the data values. It starts from a height of zero on the left and goes to a height of total frequency at the right, being increasing function between smallest and largest values of the data set.

This graph may be drawn by taking data values along x-axis and the corresponding cumulative relative frequencies or percentage cumulative frequencies along y-axis, the relative frequencies being as heights at each point in such case it is called discrete cumulative frequency function or percentage cumulative frequency function or polygon.

2.6.7 Scatter Plots

Very often, many variables are measured on each individual. For example, we may consider two variables, height and weight of each individual in a class. Now, the resulting data set consists of n pairs of observation such as (x_i, y_i), $i = 1, 2, \ldots., n$; where each x_i denotes height and each y_i denotes weight. This is called a ***bivariate*** data set. A plot of two variables useful in such situations is scatter plot. It is obtained by taking one variable on x-axis and the other on y-axis. Each pair of values (x_i, y_i), $i = 1, 2, \ldots., n$; in the data set will contribute as a point in this bivariate plot and we usually put a cross (×) or dot (.) at the intersection of values.

A scatter plot is the best way of studying bivariate problems. The bivariate data are usually of the following types:

i) Paired measurements on the same variable

The data come from the situations where experimental units are deliberately paired. For example, the use of twins in the biological and psychological experiments. Here we would expect results within a pair of twins to be more alike than observations between different pairs. In such situations, the main interest is to investigate whether variables are dependent and if so what form of the relationship between the variables actually is.

Example 2.10: Data are recorded on milk yield of cows in the morning and in the evening.

Morning values:	4.5, 6.0, 5.5, 3.5, 4.5, 6.5, 7.0, 5.0, 4.5, 6.5
Evening values:	5.5, 6.5, 6.0, 5.5, 7.0, 5.5, 8.0, 6.0, 8.5, 7.0

The interest is whether the characteristic measured varies in any systematic pattern over the day.

Solution: As both the measurements are on same variable, the interest is therefore, not just in relationship between morning and evening measurements but also in comparing them. The line of equality is a useful visual aid for this type of data. For the scatter plot we take the morning values along x-axis and evening values along y-axis.

At the intersection of 4.5 and 5.5, we put a dot (.), similarly for 6.0, 6.5 and so on. The scatter plot is shown in Figure 2.11.

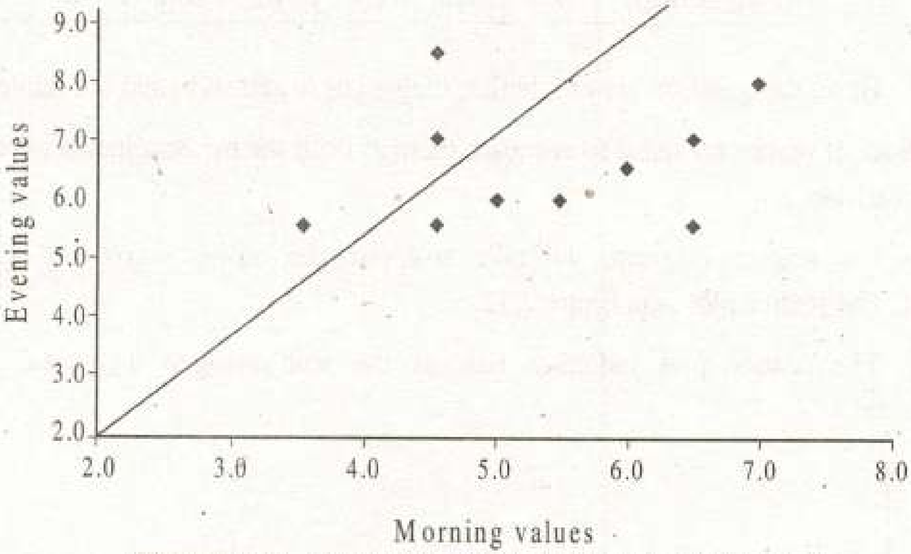

Figure: 2.11 Scatter plot of morning and evening values

Line of equality is at an angle of 45° as indicated in Figure 2.11 alongwith scatter plot. It is clear that most of the points are above the line of equality and so the milk yield in the morning and evening is not the same and more milk is obtained in the evening as compared with the morning.

With this data we can also look at the differences between morning and evening values and treat this as a one sample problem. However, we would no longer be able to see if the change was related to the initial value.

ii) Two related measurements

The pair of values may come from two variables which are related to each other. For example, samples of soil nitrogen and yield of a variety are taken in each of seven randomly selected agriculture locations. In such situations, it doesn't make any sense to compare them as both the measurements are not on the same variable.

Scatter plots are also drawn to examine the relationship between two related measurements.

Example 2.11: Samples of soil nitrogen and yield are taken in each of seven randomly selected agriculture locations. The soil nitrogen and yield are:

Location	1	2	3	4	5	6	7
Soil nitrogen	8.5	7.0	6.5	6.0	7.0	8.0	7.5
yield (kg)	9.0	7.5	6.0	4.5	6.0	7.0	6.0

Here, the question arises whether the soil nitrogen and yield are related?

Solution: It makes no sense to compare them as both the measurements are not on the same variable.

For scatter diagram, we take soil nitrogen along x-axis and yield along y-axis. The scatter plot is in figure 2.12.

The scatter plot indicates that as the soil nitrogen increases, yield also increases.

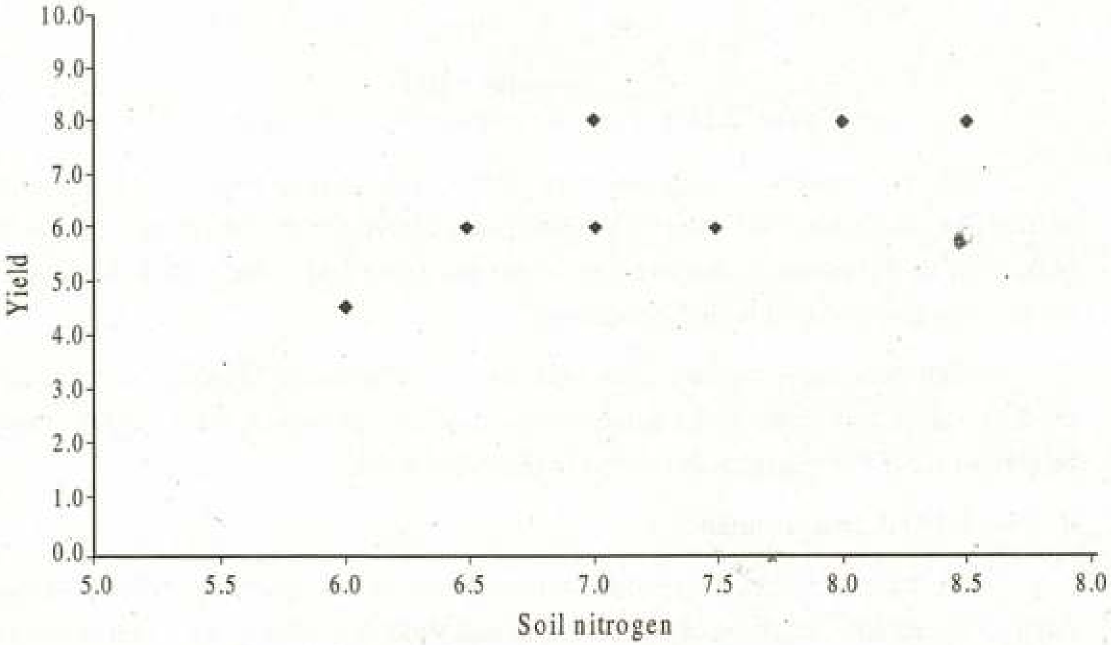

Figure 2.12: Scatter plot of soil nitrogen and yield.

2.7 Bivariate Frequency Distribution

The constructed frequency distribution, considering two variables at a time is called bivariate frequency distribution. The pairs of observations are taken into account while constructing a bivariate frequency distribution. The procedure to construct a bivariate frequency is the same except n pair of values (x_i, y_i), $i = 1, 2, \ldots., n$ are allocated at the intersection of classes of both the variables.

Example 2.12: A sample of 25 students was taken and their heights in feet (x) and weights in kilograms (y) were measured. The pairs below are (height, weight).

(5.5, 60),	(5.0, 55),	(4.3, 46),	(5.3, 67),	(4.9, 48),	(5.9, 69),
(5.4, 67),	(4.8, 55),	(5.3, 57),	(5.8, 67),	(5.3, 57),	(5.7, 65),
(5.8, 63),	(5.9, 65),	(4.8, 49),	(5.3, 55),	(5.1, 60),	(5.7, 65),
(4.7, 50),	(4.5, 50),	(5.3, 60),	(4.6, 53),	(5.4, 62),	(5.2, 59),
(4.7, 55).					

Solution: The minimum and maximum height is 4.3 and 5.9 feet respectively. The minimum and maximum weight is 46 and 69 kilogram respectively.

For height, we take 4 classes with an interval of 0.5. So, the class limits for height are 4.0-4.4, 4.5-4.9, 5.0-5.4, 5.5-5.9. The corresponding class boundaries are 3.95-4.45, 4.45-4.95, 4.95-5.45, 5.45-5.95. As usual, the class boundaries have been obtained by averaging the upper class limit of one class and the lower class limit of the next class i.e., the first class boundary is

$$(4.4+4.5)/2 = 4.45 \text{ and so on.}$$

For weight, we take 5 classes with an interval of 5.0. So, the class limits are 44-49, 50-54, 55-59, 60-64, 65-69. The corresponding class boundaries are 44.5-49.5, 49.5-54.5, 54.5-59.5, 59.5-64.5, 64.5-69.5.

Starting from first pair, all the 25 pairs are assigned to the classes they belong. The first pair (5.5,60) falls in the class with height (5.45-5.95) and weight (59.5-64.5), a tally mark is made in the table against their interaction. The second pair (5.0, 55) belongs to the class with height (4.95-5.45) and weight (54.5-59.5), a tally mark is made in table against their interaction and so on, the last pair (4.7, 55) belongs to the class with the height (4.45-4.95) and weight (54.5-59.5). The number of tally marks in each cell gives the frequency of the class with certain height and weight boundaries. The bivariate frequency distribution is given in Table 2.7. (a) and 2.7 (b).

Table 2.8 (a): Bivariate frequency table

	Height			
Weight	3.95-4.45	4.45-4.95	4.95-5.45	5.45-5.95
44.5-49.5	|	||		
49.5-54.5		|||		
54.5-59.5		||	𝍸	
59.5-64.5			|||	||
64.5-69.5			||	𝍸

Table 2.8 (a) is the bivariate frequency distribution (or table) after making a tally count.

Table 2.8 (a)

	Height			
Weight	3.95-4.45	4.45-4.95	4.95-5.45	5.45-5.95
44.5-49.5	1	2		
49.5-54.5		3		
54.5-59.5		2	5	
59.5-64.5			3	2
64.5-69.5			2	5

It is clear from the bivariate frequency table that there is one individual with weight between 44.5-49.5 kg and height between 3.95-4.45 feet. There are two individuals with weight between 44.5-49.5 kg and height between 4.45-4.95 feet and so on, there are 5 individuals with weight between 64.5-69.5 kg and height between 5.45-5.95 feet.

Exercise 2

Ans on Page 250

2.1 What are different methods of representation of statistical data.

2.2 Define the Histogram, the frequency polygon and the frequency curve.

2.3 What do you understand by classification and tabulation? Discuss their importance in a statistical analysis.

2.4 Distinguish between one-way and two-way tables. Illustrate your answers with examples. Also explain the following:

i) Classification according to attributes.
ii) Class limits.
iii) Length of class interval.
iv) Class frequency.

2.5 The following table gives the details of monthly budgets of two families. Represent these figures through a suitable diagram.

Items	**Family *A***	**Family *B***
Food	Rs. 600	Rs. 800
Clothing	Rs. 100	Rs. 100
House rent	Rs. 400	Rs. 500
Fuel and lighting	Rs. 100	Rs. 100
Miscellaneous	Rs. 300	Rs. 500
Total	**Rs. 1500**	**Rs. 2000**

2.6 Represent the following data through pie diagram.

Items of Expenditure	**Amount**
Food	4000
Clothing	1000
House Rent	2500
Education	1000
Fuel and light	600
Miscellaneous	2000

2.7 Define frequency Histogram. Draw a Histogram for the following frequency distribution giving the steps involved.

Mid values (*X*)	32	37	42	47	52	57	62	67
Frequency (*f*)	3	17	28	47	54	31	14	4

2.8 i. a) Write down the important points for drawing graphs?

b) In order to estimate the mean length of leaves from a certain tree, a sample of 100 leaves was chosen and their lengths are measured in millimeter. A grouped frequency table was set up and the results were as follows:

Mid value	2.2	2.7	3.2	3.7	4.2	4.7	5.2	5.7	6.2
Frequency	3	5	8	12	18	24	20	8	2

ii. a) Display the table in the form of a frequency polygon.

b) What are the boundaries of the interval whose mid point is 3.7 cm?

2.9 In a locality, total area is 500 acres where 250 acres are under sugarcane, 125 acres are under maize, 60 acres are under wheat and the remaining 65 acres are under other crops. Make a pie-diagram to represent the distribution of acreage under different crops.

2.10 A biologist was interested to know whether male spiders are longer or female spiders. He collected random samples of female and male green lynx spiders given below. Advise him?

Length of female (in mm)				**Length of male (in mm)**			
5.7	5.2	4.7	5.8	8.2	9.9	5.9	8.4
6.1	6.3	7.0	5.7	9.6	7.0	6,6	7.8
7.5	6.4	6.5	4.7	7.5	9.8	6.3	8.3
6.2	5.4	6.2	5.8	8.0	9.1	6.3	8.4
4.8	5.9	5.2	6.8	8.7	7.4	7.9	10.2
5.6	5.5						

2.11 i) What is meant by tabulation? Explain the main steps which are generally taken in tabulation?

ii) What is a frequency distribution? How is it constructed?

2.12 The following data gives the lifetime in minutes, recorded to the nearest tenth of a minute of 50 sprayed insects.

1.2	2.2	0.7	3.9	1.7	1.9	1.4	1.8	2.0	4.3
2.5	0.9	3.4	2.8	3.7	3.5	0.4	2.8	1.1	0.2
3.9	6.3	2.5	2.1	1.3	2.1	0.3	0.4	2.4	2.1
3.5	2.9	1.2	5.3	1.7	2.7	1.8	4.8	3.2	1.6
2.6	1.8	2.3	1.3	3.1	1.5	2.6	5.9	2.0	2.3

Using 8 intervals with the lowest starting at 0.1

i) Form a frequency distribution and a cumulative frequency distribution.

ii) Also draw Histogram and frequency polygon for the frequency distribution so formed.

2.13 i) What are the advantages of diagrammatic representation?

ii) Explain the following:

a) A Bar diagram b). Subdivided bar diagram

c) Multiple bar diagram.

2.14 The following data gives the record of a company's savings over the years. Draw a bar diagram to represent it:

Year	1950	1951	1952	1953	1954	1955	1956	1957
Rs.(000)	1010	2050	3458	1980	2300	1295	1520	1070

2.15 Draw a sub-divided bar diagram to represent the male and female population of four divisions of Punjab in 1961.

Division	**Male**	**Female**	**Both Sexes**
Lahore	35	30	65
Multan	35	31	66
Sargodha	32	28	60
Rawalpindi	21	19	40

2.16 The following information is available about the áge (in years) and their weights in kilograms (age, weight). Make a scatter plot and bivarite frequency distribution.

(20, 60), (15, 55), (14, 45), (17,60), (16, 48), (22, 70), (16, 63), (14, 55), (18, 57), (19, 67), (21, 67), (17, 65), (13, 60), (15, 60), (17, 49), (19, 65), (23, 73), (21, 65), (22, 70), (14, 50), (16, 60), (19, 59), (15, 62), (17, 59), (24, 75).

2.17 Fill in the blanks:

i) Classification is the __________ of arranging data according to some common characteristics.

ii) A table has at least __________ parts.

iii) In an open-end frequency distribution, either the __________ class limit of __________ group or upper limit of the __________ class are not given.

iv) In Histogram, with unequal class intervals, the area of each rectangle is __________ to class frequency.

v) An ogive is a __________ polygon.

vi) A frequency table can be represented graphically by a __________.

vii) A Histogram is a ____________ bar Chart with ______ space between its bars.

viii) The area of each bar is __________ to the frequency it represents.

ix) If mid-points of the tops of the consecutive bars in a Histogram are joined by straight lines, a ____________ is obtained.

2.18 Against each statement write T for true and F for false statement.

i) Grouped data and primary data are same.

ii) The class mark is also named as mid point.

iii) A table has at least three parts.

iv) The graph of a time series is called Histogram.

v) Cumulative frequencies are decreasing.

vi) The data 10, 5, 7, 6, 4 is the example of grouped data.

vii) The two fold division is also named as Dichotomy.

viii) Data can be presented by means of graph.

ix) A graph of cumulative frequency curve is called polygon.

x) In constructing a Histogram, midpoints are to be taken along *x*-axis.