DNYANSAGAR ARTS AND COMMERCE COLLEGE, BALEWADI, PUNE – 45

Subject : Statistics (sub code CA-105 CBCS 2019 Pattern)

Class : F.Y. BBA(CA)

Prof . S. B. Potadar

www.dacc.edu.in

Origin and Growth of Statistics

ORIGIN OF STATISTICS

The word 'Statistics' seems to have been derived from the Latin word 'STATUS' or the Italian word '*statista*' or the German word '*statistik*' each of which means a 'political state'.

Sir Ronald A Fisher (1890-1962)

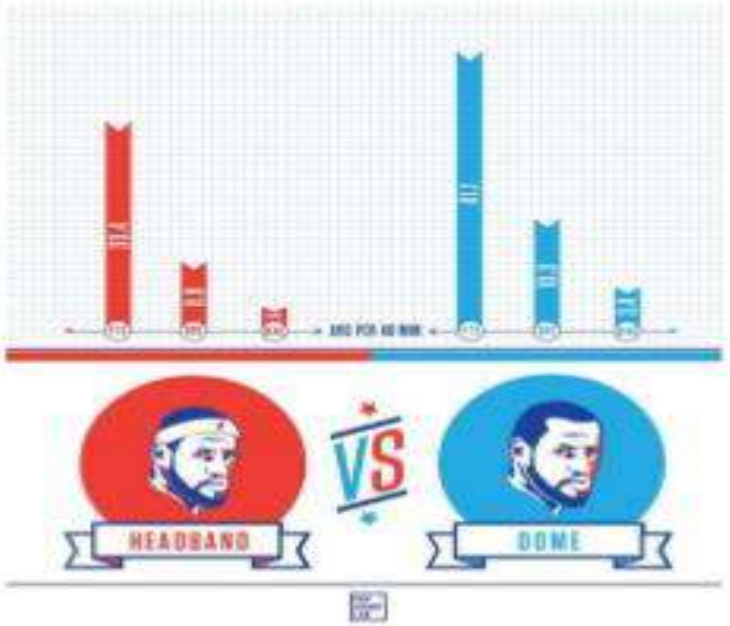
is the father of Statistics

Definition by Croxton and Cowden

- "Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis."
  - 1. Collection of Data
  - 2. Presentation of data
  - 3. Analysis of data
  - 4. Interpretation of data

Functions of Statistics

- 1. Condensation
- 2. Comparison
- 3. Forecasting
- 4. Estimation



Scope of Statistics

- 1. Statistics and Industry
- 2. Statistics and Commerce
- 3. Statistics and Agriculture
- 4. Statistics and Economics
- 5. Statistics and Education
- 6. Statistics and Planning
- 7. Statistics and Medicine
- 8. Statistics and Modern applications

Types of data

i) primary data
ii) Secondary data

primary data :- The data which are collected from the field under the control and supervision of an investigator.
For eg: Your own questionnaire

Secondary data :- Data gathered and recorded by someelse prior to and for a purpose other than the current project.

While studying any phenomeon we come across two types of characteristics
i) Costant
ii) variable

Constant :-
The characteristic which does not change its value or nature is considered as Constant.
For eg :- Height of a person after 25 years of age.

Types of characteristics :

i) Attributes
ii) Variables

Attribute :- A qualitative characteristics like sex, nationality, religion, blood group, beauty is called attribute.

Variables :- A quantitative characteristics like weight of person, examination marks, population of country is called variable.

It can be clearly noticed that variables can be measured by numbers.

Further the variables can be divided into two categories

* Discrete variable : A variable taking only particular values or isolated values is called as discrete variable.

For eg :- i) Number of students in a class.
ii) population of a contry.
iii) Number of workers in a factory

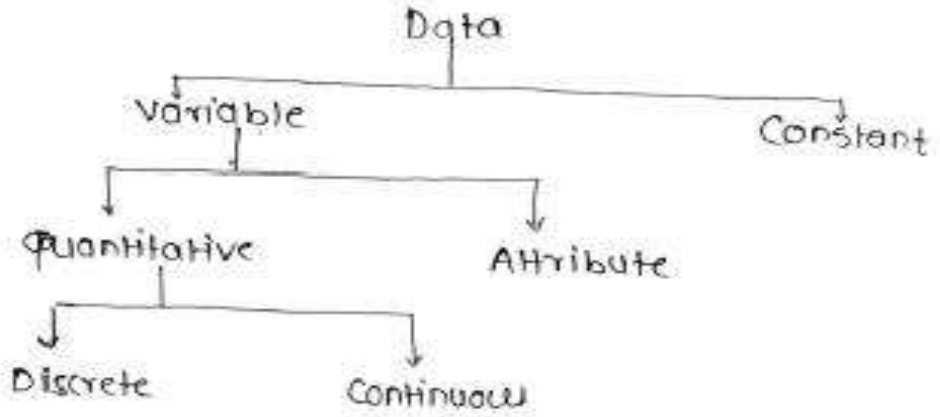
etc.

Continuous variable :- A variable taking all possible values in a certain range is called Continuous variable.

eg :- i) weight of a person
ii) length of a screw
iii) speed of a vehicle.

etc.

The following diagram summaries the various types of data :



Classification :-

In order to study a characteristic or a group of characteristics of any type, the first phase is to collect the data.

Raw data :- The unprocessed data in terms of individual observations is called as raw data.

For further statistical analysis, the data items are arranged in increasing or decreasing order. However if there is a huge amount of observations merely order arrangement is not enough. It does not furnish much useful information nor does it reduce the bulk of data. Data in this form are difficult to comprehend, analyse and interpret.

The entire process of making homogeneous and non-overlaping groups of observations according to similarities is called as classification.

Methods of classification :-

There are two methods of classification

i) Inclusive method
ii) Exclusive method

i) Inclusive method :-

The classification in which both upper and lower limit are included in the same class is known as inclusive method of classification.

For eg :-

Monthly income
0 - 2000
2001 - 4000
4001 - 6000
6001 - 8000

ii) Exclusive method :-

When upper limit of each class is excluded from the respective class and is included in the immediate next class then the method of classification is known as exclusive method of classification.

foreg:-

Monthy income
0 - 2000
2000 - 4000
4000 - 6000
6000 - 8000

Class-limits :-

The two numbers designating the class-interval are called as class-limits.

class ⇒ 20 - 40

Here 20 and 40 are called class limits

smaller number 20 is the lower limit

larger number 40 is the upper limit of the class.

Class boundaries :-

The class boundaries are the numbers upto which the actual magnitude of observation in the class can extend. The class boundaries are also called as actual limits or extended limits.

Class limits	Class boundaries
10 - 19	9.5 – 19.5
20 - 29	19.5 - 29.5
30 - 39	29.5 - 39.5

Class width or class interval :-

The size or width or class interval is the difference between lower and upper class boundaries

i.e width of a class interval = upper class boundary - lower class boundary.

Class-mark or Mid values :-

It is the mid-point of class interval and the same can be obtained as follows:

$$\text{Mid value} = \frac{\text{upper limit} + \text{lower limit}}{2}$$

open end class:-

A class in which one of the limits is not specified is called an open end class.

for eg :- In the following frequency distribution

Daily Sales in ₹

open end class

- below 2000
- 2000 - 3000
- 3000 - 4000
- 4000 and above

Cumulative Frequencies :-

In many situations it is required to find the number of observations below or above a certain value.

For example :- In case of a frequency distribution of income, the number of persons below poverty line or in case of frequency distribution of examination of marks, number of candidates above 60 etc.

In this case cumulative frequencies are much useful.

There are two types of cumulative frequencies.

i) Less than type cumulative frequencies.

ii) more than type cumulative frequency.

i) Less than type cumulative frequency :-

Here we add frequencies from top (lowest-Class) to bottom (highest class) and write them at each stage against the upper boundary of the corresponding class.

ii) More than type Cumulative frequency :-

Here we add frequencies from bottom (highest-class) to top(lowest Class) and

Example :

Marks	Frequency	L.C.f	M.C.F
0-10	5	5	4+4+15+12+5 = 40
10-20	12	5+12=17	4+4+15+12 = 35
20-30	15	5+12+15=32	4+4+15=23
30-40	4	5+12+15+4=36	4+4=8
40-50	4	5+12+15+4+4=40	4
Total	40		

Graphical Representation :-

various graphs associated with frequency distribution. Generally, graphs are used to represent mathematical relationship between two variables.

i) Histogram :-

It is one of the most important and simple method of presenting a frequency distribution.

for drawing histogram there are two cases.

i) Classes with equal class width
ii) classes with unequal class width

i) Classes with equal class width :-

To draw histogram in this case we perform the following steps :-

Step 1: If classes are inclusive type then by computing class boundaries, convert it to exclusive type. And go to step 2. Otherwise directly go to step 2

Step 2: Take class boundaries on X-axis and frequencies on Y-axis.

Step 3: For each class, draw a rectangle of height proportional to the corresponding class frequency. The graph obtained by these continuous rectangle is called histogram.

Example :-

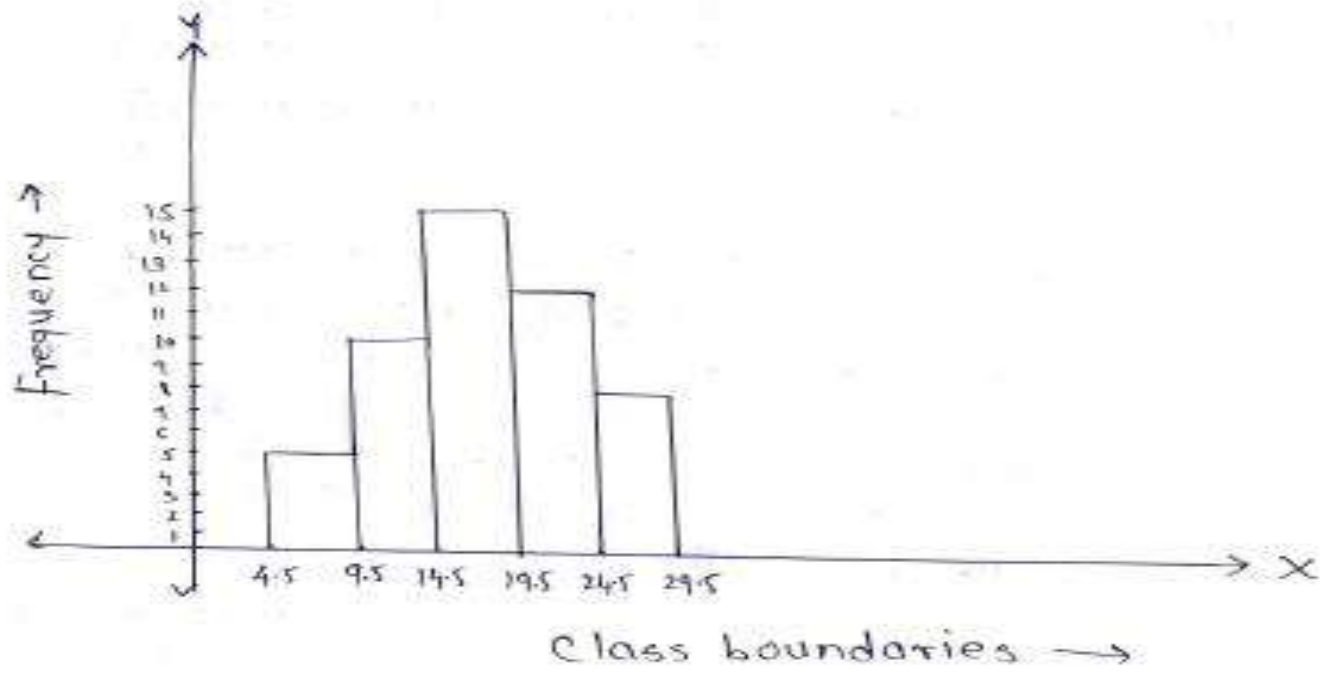
1) Draw the histogram for the following frequency distribution.

Class	5-9	10-14	15-19	20-24	25-29
Frequency	5	10	15	12	8

→ Here the classes are inclusive type so convert it to exclusive type

Class limits	Class boundaries	Frequency
5-9	4.5 - 9.5	5
10-14	9.5 - 14.5	10
15-19	14.5 - 19.5	15
20-24	19.5 - 24.5	12
25-29	24.5 - 29.5	8

Since here class width equal we draw histogram as below.



ii) Classes with unequal class width :-

To draw histogram in this case we perform the following steps :-

Step 1 :- If classes are of inclusive type then by computing class boundaries, convert it to exclusive type, and go to step 2. otherwise directly go to step 2.

Step 2 :- Since classes having unequal classwidth, find frequency density for each class. where,

$$\text{Frequency density} = \frac{\text{frequency of the class.}}{\text{classwidth of same class.}}$$

Step 3 :- Consider class boundaries on X-axis and frequency density on Y-axis.

step 4 : For each class interval, draw a rectangle of height proportional to the corresponding freq. density

The graph obtained by these continuous rectangles is called histogram.

Example :-

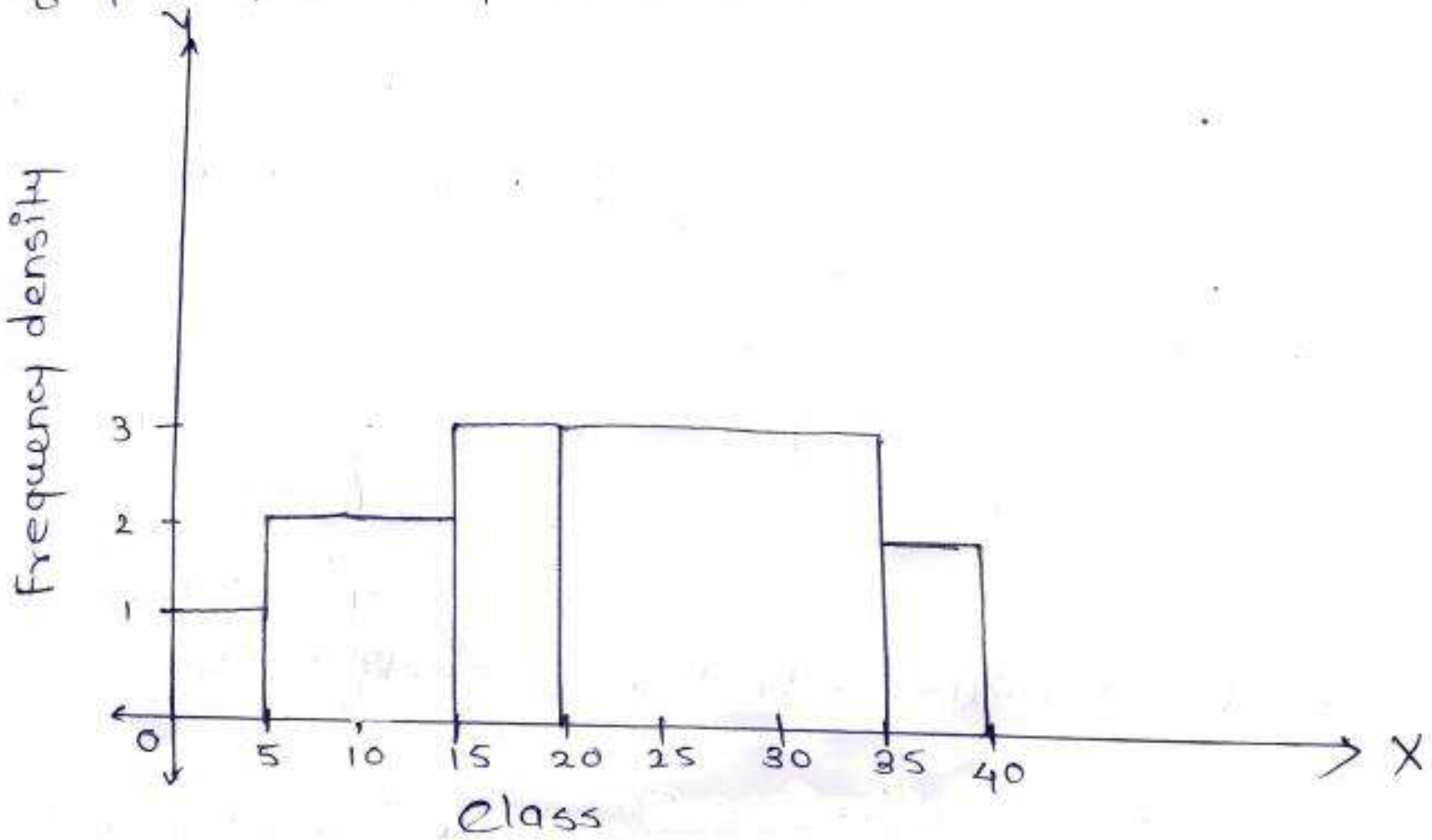
Draw the histogram for the following frequency distribution.

Class	0-5	5-15	15-20	20-35	35-40
Frequency	5	20	15	36	8.

→ Here classes are exclusive type. In the given example the classes having unequal class width so we find frequency density as below

Class	0-5	5-15	15-20	20-35	35-40
Frequency	5	20	15	45	10
Frequency density	$\frac{5}{5}=1$	$\frac{20}{10}=2$	$\frac{15}{5}=3$	$\frac{45}{15}=3$	$\frac{10}{5}=2$

→ We draw histogram by taking class boundaries on X-axis and frequency density on y-axis as below



Frequency polygon :-

A graph is expected to be in the form of a smooth curve. Histogram does not fulfill this requirement Therefore, another way of presentation of frequency distribution is frequency polygon.

This type of graph enables us to understand the pattern in the data more clearly.

To draw frequency polygon following steps are

step 1: Find mid values of each class.

step 2: Mid values are taken x axis and frequencies on Y-axis to draw the graph. using suitable scale.

step 3: Using the mid points for the x values and the frequencies as the y values plot the points.

Step 4: Connect adjacent points with line segments. Draw a line back to the x-axis at the beginning and end of the graph.

Frequency Curve :-

There is little difference in frequency polygon and frequency curve. If the points are joined by a smooth curves instead of straight lines we get a closed figure called frequency curve.

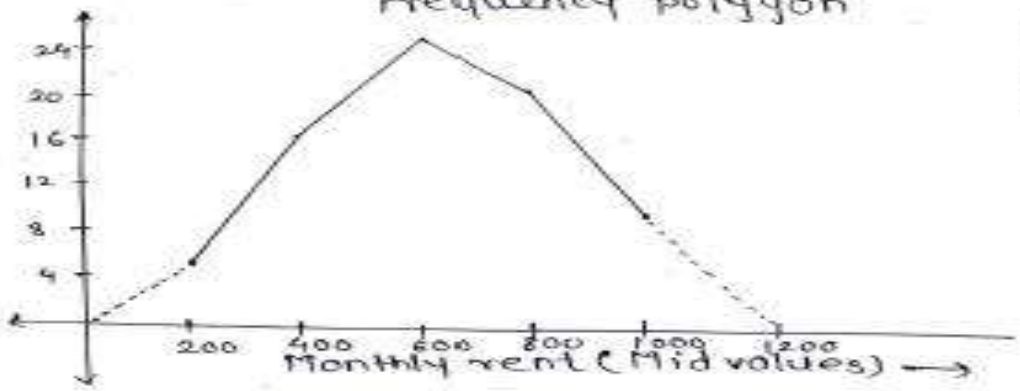
Ex 1. Draw a frequency polygon and a frequency curve for the following data :

Monthly house rent	100-300	300-500	500-700	700-900	900-1100
No. of families.	6	16	24	20	10

→ Mid values of classes are taken on x axis and frequency on y-axis.

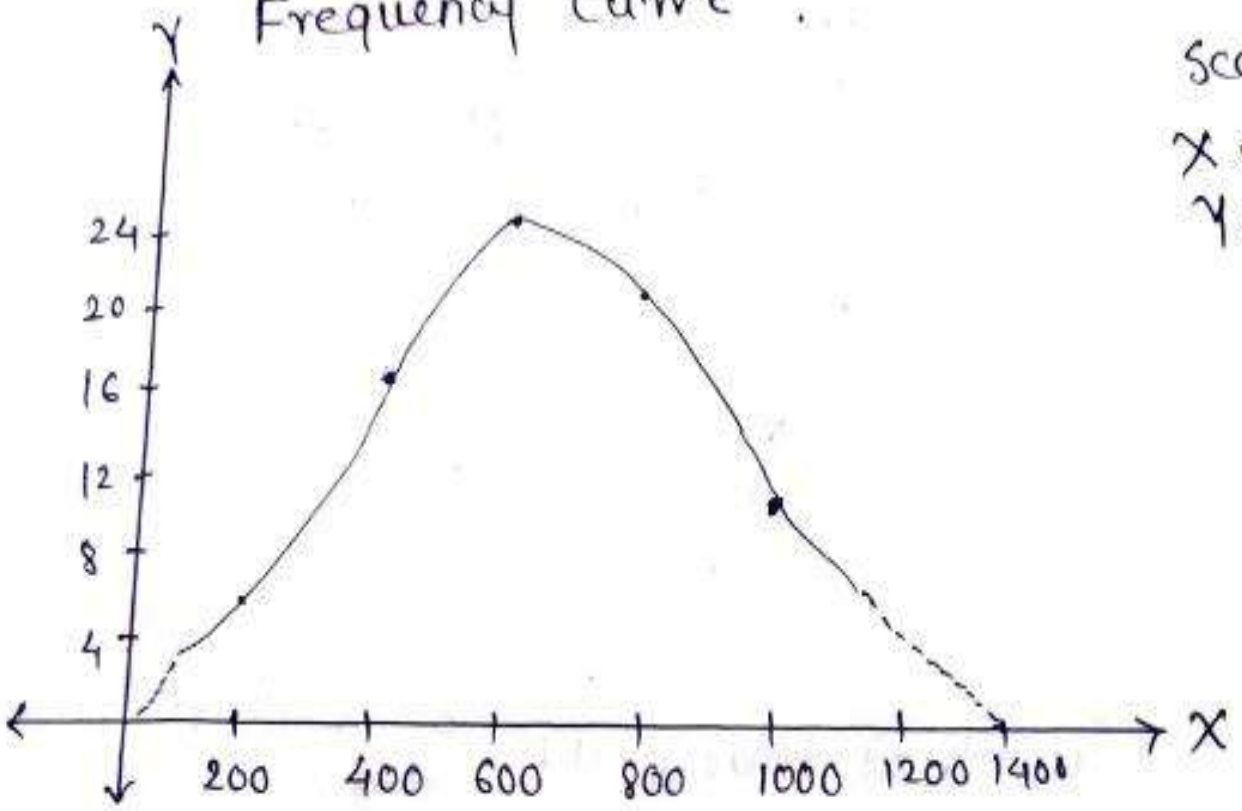
Monthly house Rent	100-300	300-500	500-700	700-900	900-1100
Mid values	200	400	600	800	1000
No. of families	6	16	24	20	10

Frequency polygon



Scale
x axis 1 unit = ₹200
y axis 1 Unit = 4 units.

Frequency Curve.



Scale

X axis 1 unit = ₹200

Y axis 1 unit = 4 units.

Cumulative Frequency Curve or ogive curve

Cumulative frequency distribution is represented by cumulative frequency curve or ogive curve.

There are two types of cumulative frequencies, hence there are two types of cumulative frequency curves.

i) Less than ogive curve
ii) More than ogive curve.

i) Less than ogive curve :- To draw less than ogive curve we perform following steps:

Step 1: If classes are inclusive type then make it exclusive and go to step 2 otherwise go to step 2 directly.

Step 2: Find less than cumulative frequencies.

Step 3: Consider upper class limit/ boundary on X axis and l.c.f on Y-axis.

Step 4 : Plot the corresponding pairs on XY plane.

Step 5 : Through all the points draw a smooth curve. Join the curve on x-axis by taking one additional point at lower with 0 frequency.

The curve obtained is called less than ogive curve.

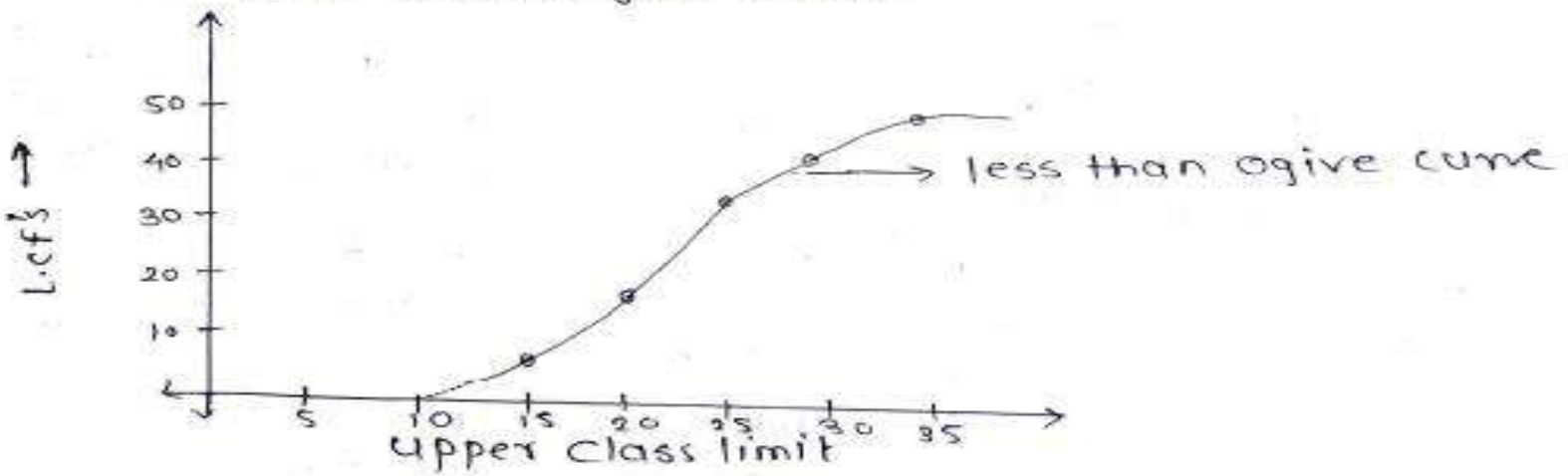
Example :- From the following frequency distribution of weight of 50 students, draw less than ogive curve.

Weight (in kg)	10-15	15-20	20-25	25-30	30-35
No. of students	5	12	15	10	8

→ Now, calculate less than cumulative frequencies.

Weight (in kg)	No. of students	upper limit	L.C.f
10 - 15	5	15	5
15 - 20	12	20	17
20 - 25	15	25	32
25 - 30	10	30	42
30 - 35	8	35	50

less than ogive curve



ii) More than ogive curve :- To draw more than ogive curve. we perform following steps.

Step 1: If classes are inclusive type than make it exclusive and go to step 2.

Step 2: Find more than cumulative frequencies (m.c.f)

Step 3: Consider lower limit / boundary on x axis and m.c.f on Y axis.

Step 4: plot the corresponding pairs of points on (X,Y) plane.

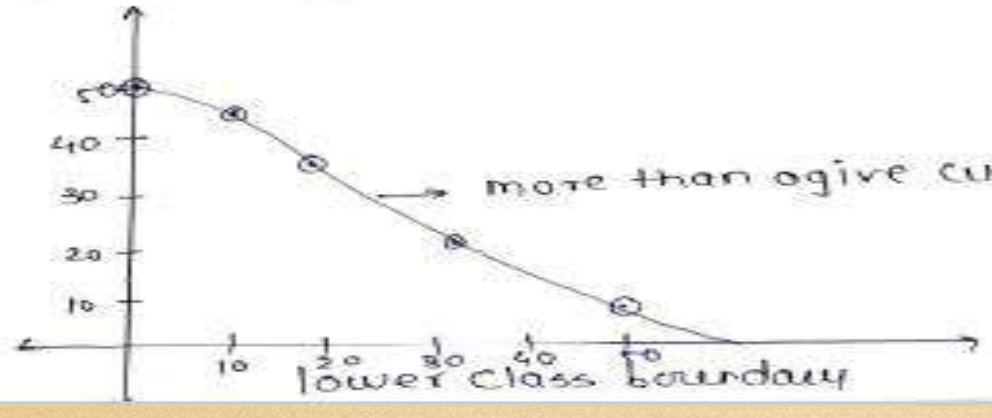
Step 5: Through all the points draw a smooth curve. Join the curve on x axis by taking one additional point at upper end with 0 frequency.

Example : From the following frequency distribution draw more than ogive curve.

Class	0 - 10	10-20	20-30	30-40	40 - 50
Frequency	5	10	15	12	8

→ Now, calculate more than cumulative frequencies.

Class	Frequency	lower limit	m.c.f
0-10	5	0	50
10-20	10	10	45
20-30	15	20	35
30-40	12	30	20
40-50	8	40	8

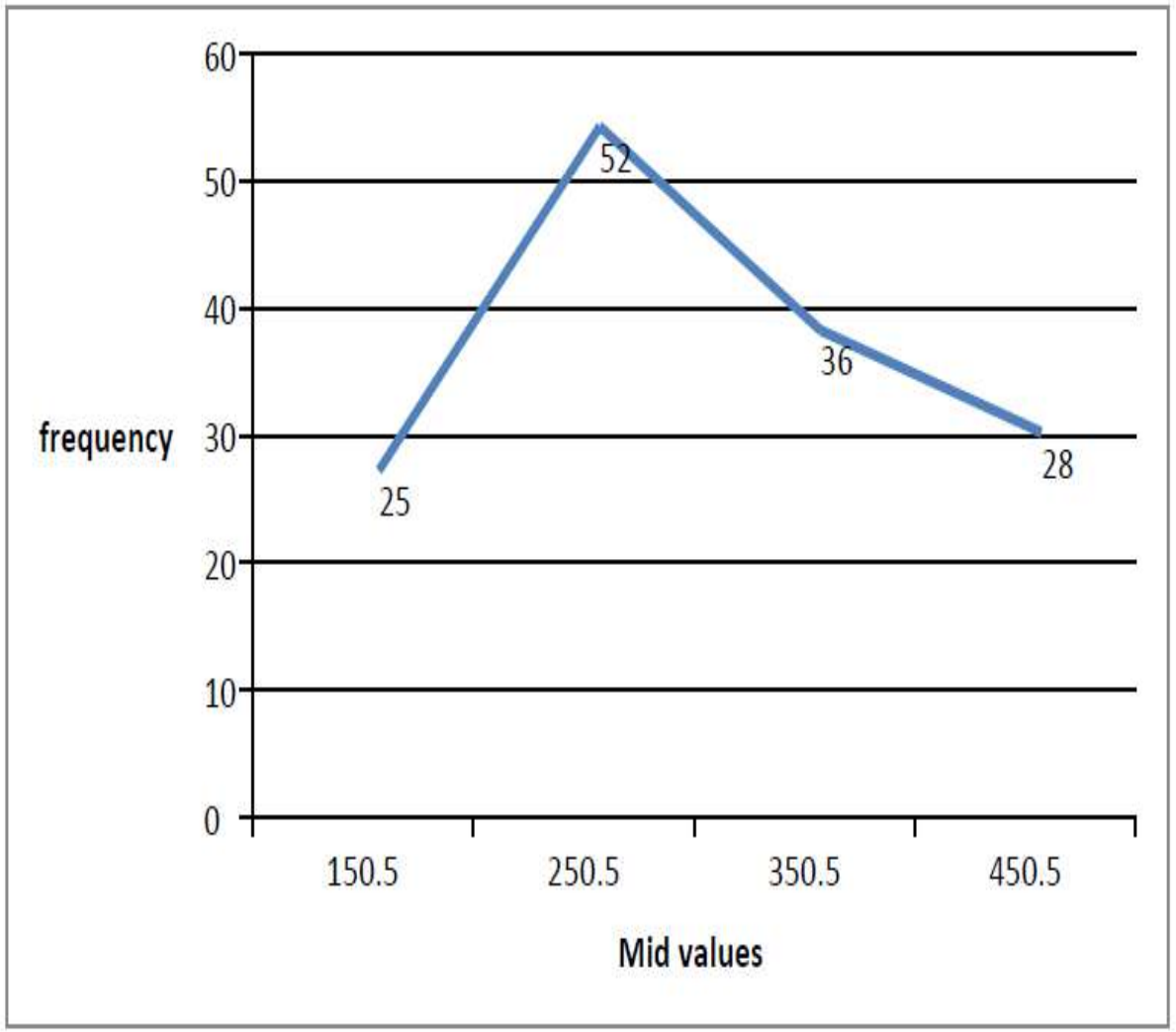


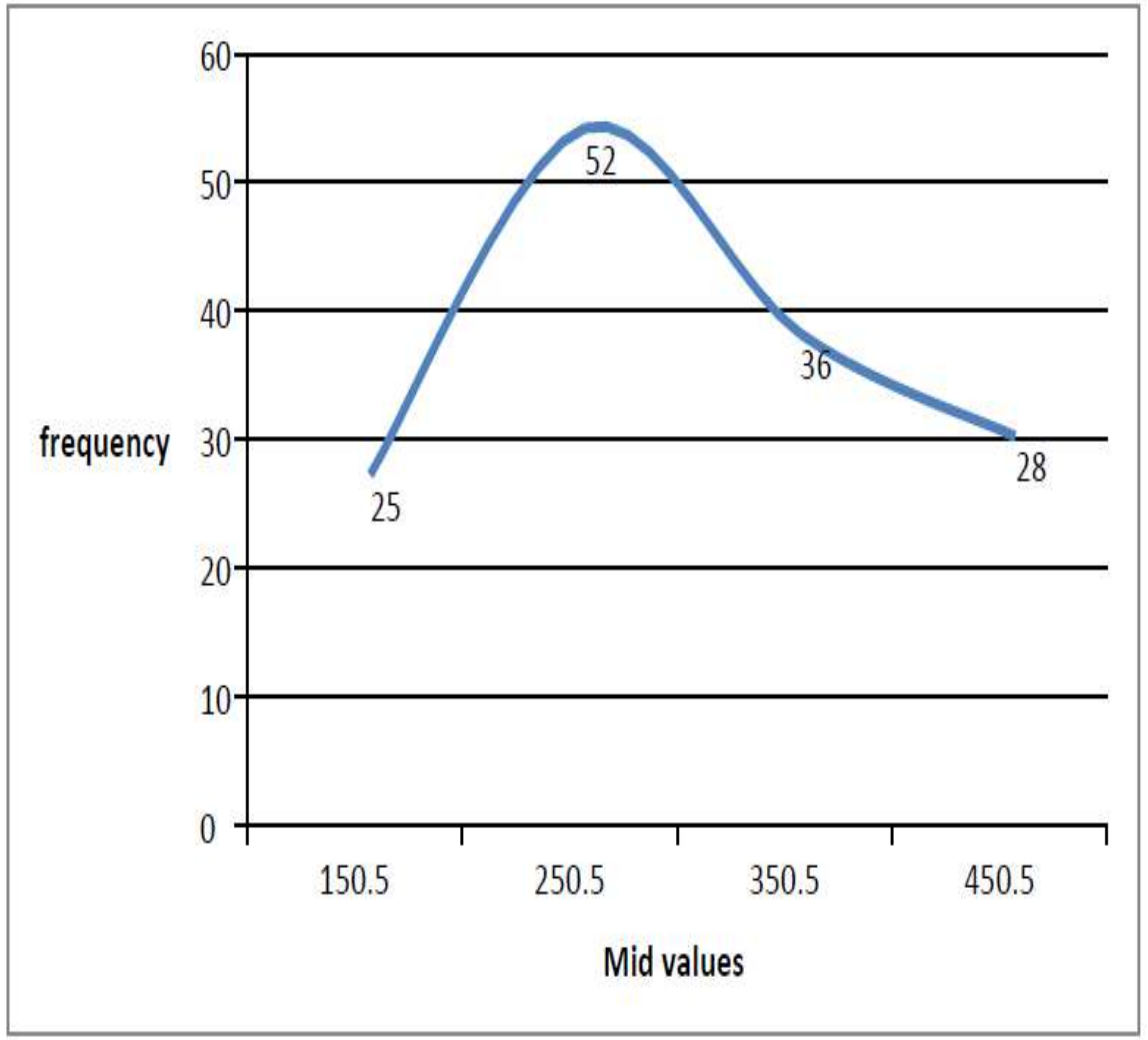
1. From the following data construct frequency curve and frequency polygon.

Marks	101-200	201-300	301-400	401-500
No.of students	25	52	36	28

Solution : Here classes are inclusive type so convert it to exclusive type.

Marks	No.of students	Class boundaries	Mid value
101-200	25	100..5 - 200.5	150.5
201-300	52	200.5 - 300.5	250.5
301-400	36	300.5 - 400.5	350.5
401-500	28	400.5 - 500.5	450.5


60
50
40
30
20
10
0
frequency
52
36
28
25
150.5
250.5
350.5
450.5
Mid values

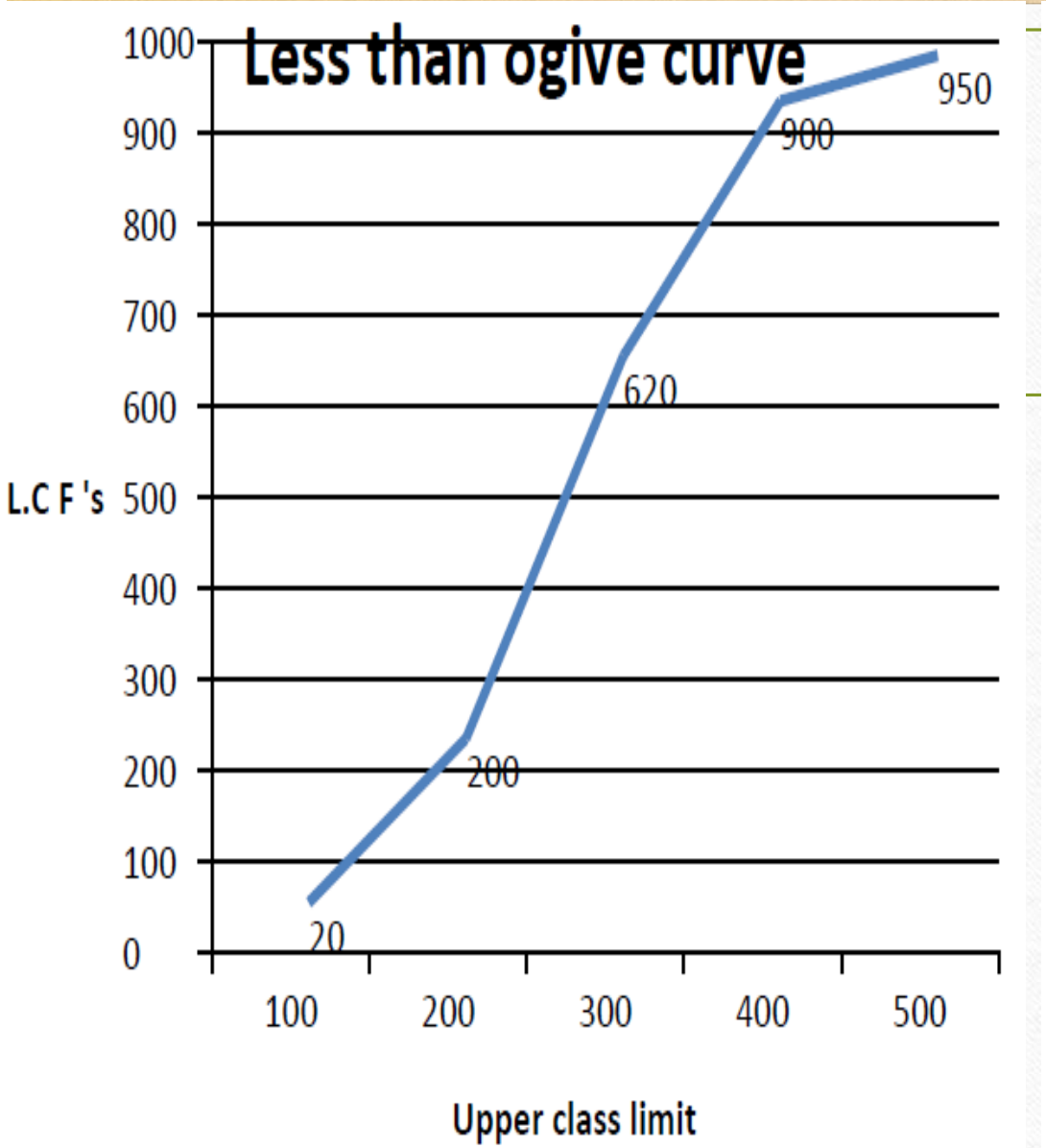

60
50
40
30
20
10
0
frequency
52
36
28
25
150.5
250.5
350.5
450.5
Mid values

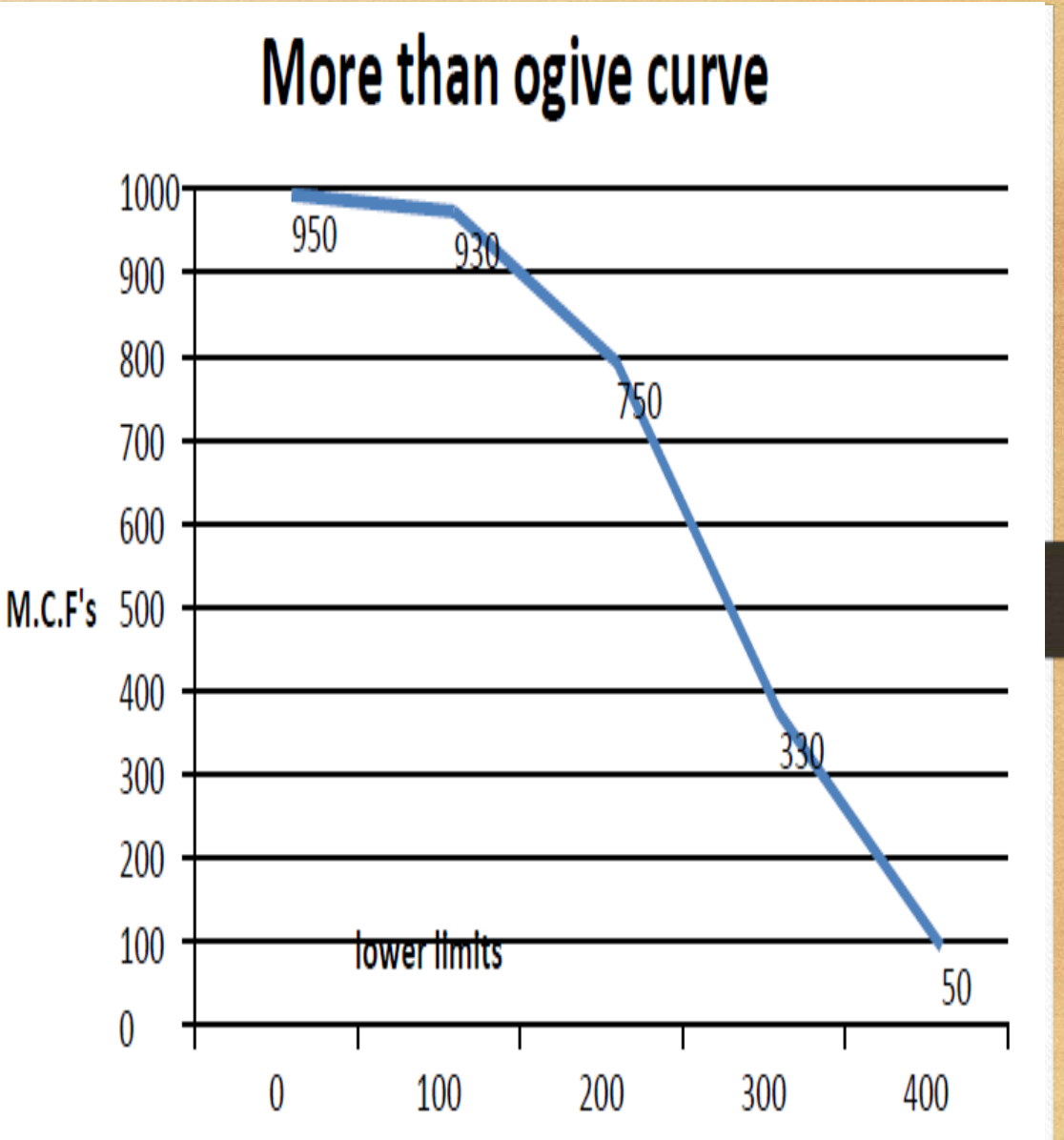
2. From the following data construct ogive curves.

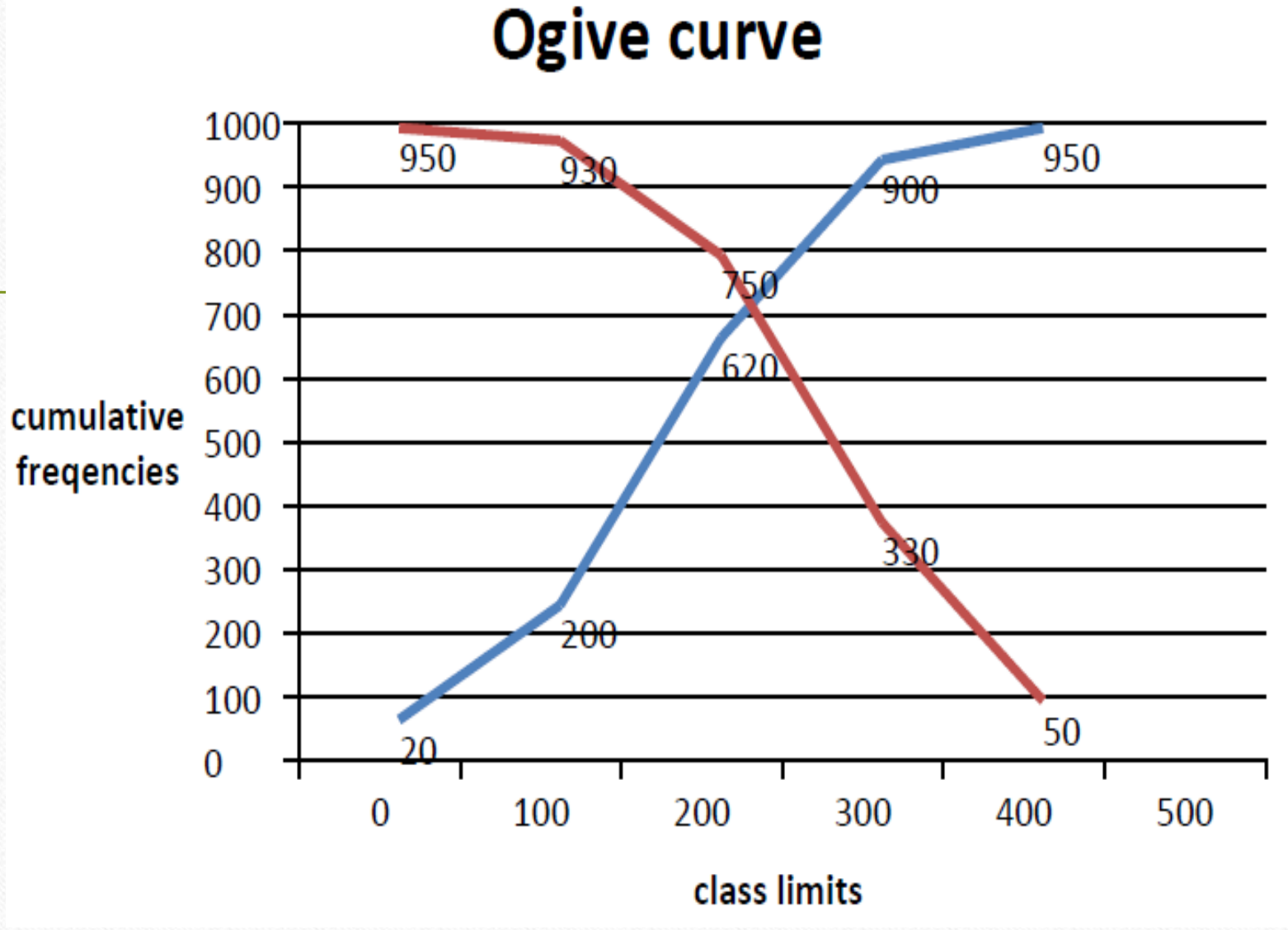
class	0-100	100-200	200-300	300-400	400-500
frequency	20	180	420	280	50

Solution :

Class	LL	UL	Freq	lcf	mcf
0-100	0	100	20	20	**950**
100-200	100	200	180	200	930
200-300	200	300	420	620	750
300-400	300	400	280	900	330
400-500	400	500	50	**950**	50
Total			**950**		


Less than ogive curve
1000
900
800
700
600
500
400
300
200
100
0
L.C F 's
950
900
620
200
20
100
200
300
400
500
Upper class limit


More than ogive curve
1000
900
800
700
600
500
400
300
200
100
0
M.C.F's
950
930
750
330
50
lower limits
0
100
200
300
400


Ogive curve
cumulative freqencies
1000
900
800
700
600
500
400
300
200
100
0
950
930
750
620
330
50
20
200
900
950
0
100
200
300
400
500
class limits

Frequency Distribution

Introduction

- **Frequency** is how often something occurs.

Example: Sam played football on

- Saturday Morning,
- Saturday Afternoon
- Thursday Afternoon

The frequency was 2 on Saturday, 1 on Thursday and 3 for the whole week.

Frequency Distribution

frequency distribution table

A data table that lists a set of scores and their frequency.

score	tally	frequency (f)
1	||||	4
2	𝍸 ||||	9
3	𝍸 |	6
4	𝍸 ||	7
5	|||	3
6	||	2

Frequency Distribution

A frequency distribution is constructed for three main reasons:

1. To facilitate the analysis of data.
2. To estimate frequencies of the unknown population distribution from the distribution of sample data and

4.2 Raw data:

The statistical data collected are generally raw data or ungrouped data. Let us consider the daily wages (in Rs) of 30 labourers in a factory.

80	70	55	50	60	65	40	30	80	90
75	45	35	65	70	80	82	55	65	80
60	55	38	65	75	85	90	65	45	75

Arrangement of data in ascending order

30	35	38	40	45	45	50	55	55	55
60	60	65	65	65	65	65	65	70	70
75	75	75	80	80	80	80	85	90	90

Discrete (or) Ungrouped frequency distribution:

Example 1:

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

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

Solution:

Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0	|||	3
1	𝍸 ||	7
2	𝍸 𝍸	10
3	𝍸 |||	8
4	𝍸 |	6
5	||||	4
6	||	2
	Total	40

Exercise - I

The following data gives the number of children in 50 families. Construct a discrete frequency table.

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

Continuous frequency distribution: In this form of distribution refers to groups of values.

Wage distribution of 100 employees

Weekly wages (Rs)	Number of employees
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
350-400	5
Total	100

Inclusive Method:

When the upper limit of the class is added or included in the same class is called as inclusive method

Raw Data

185, 303, 198, 201, 207,
213, 215, 218, 313, 226,
229, 231, 325, 236, 239,
241, 244, 248, 252, 256,
259, 262, 324, 340, 266,
180, 269, 271, 278, 280,
282, 285, 287, 290, 294,
295, 300, 190, 306, 308,
223, 317, 320, 321, 232,
326, 328, 332, 335, 338.

Tally Mark Method

Class	Tally Mark	Frequency
180 – 219	𝍸 ||||	9
220 – 259	𝍸 𝍸 ||||	14
260 – 299	𝍸 𝍸 ||	12
300 – 340	𝍸 𝍸 𝍸	15

Example 2:

Form a grouped frequency distribution from the following data by inclusive method taking 4 as the magnitude of class intervals.

31,23,19,29,22,20,16,10,13,34
38,33,28,21,15,18,36,24,18,15
12,30,27,23,20,17,14,32,26,25
18,29,24,19,16,11,22,15,17,10

Exclusive Method:

When the upper limit of the class is included in next class is called as exclusive method

Ex 14.2, 4 teachoo.com

The heights of 50 students, measured to the nearest centimetres, have been found to be as follows:

161	150	154	165	168	161	154	162	150	151
162	164	171	165	158	154	156	172	160	170
153	159	161	170	162	165	166	168	165	164
154	152	153	156	158	162	160	161	173	166
161	159	162	167	168	159	158	153	154	159

(i) Represent the data given above by a grouped frequency distribution table, taking the class intervals as 160 - 165, 165 - 170, etc.

We see that

Minimum = 150

& Maximum = 173

So, we take intervals as

150 – 155, 155 – 160,

..... 170 – 175

Height	Tally	No. of Students (frequency)
150 - 155	𝍸 𝍸 II	12
155 - 160	𝍸 IIII	9
160 - 165	𝍸 𝍸 IIII	14
165 - 170	𝍸 𝍸	10
170 - 175	𝍸	5
Total		50

Class Limit

- The class limits are the lowest and the highest values that can be included in the class. For example, take the class 30-40. The lowest value of the class is 30 and highest class is 40.
- In statistical calculations, lower class limit is denoted by L and upper class limit by U.

Lower Class Limits

are the smallest numbers that can actually belong to different classes

Lower Class Limits

Rating	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Upper Class Limits

are the largest numbers that can actually belong to different classes

Upper Class Limits

Rating	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Class Interval

- The class interval may be defined as the size of each grouping of data. For example, 50-75, 75-100, 100-125... are class intervals.

Class (Rs.)	Tally Marks	Frequency Students
20 - 30	𝍸	5
30 - 40	𝍸 III	8
40 - 50	𝍸 IIII	9
50 - 60	𝍸 𝍸	10
60 - 70	𝍸 I	6
70 - 80	II	2
Total		40

Class Interval
is the range into which data are divided

Frequency (f)
number of data values that fall in the range

Width or size of the class interval

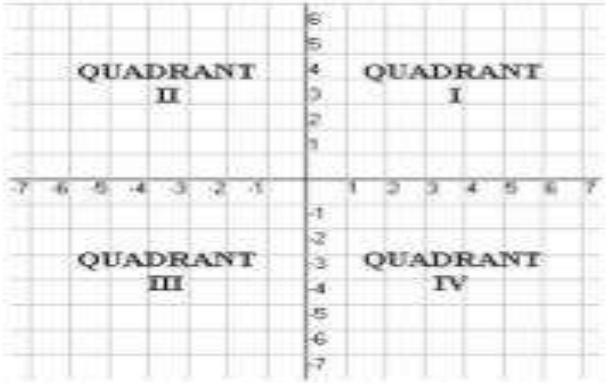
- The difference between the lower and upper class limits is called Width or size of class interval and is denoted by ' C' .

Class	Frequency, *f*
1 – 4	4
5 – 8	5
9 – 12	3
13 – 16	4
17 – 20	2

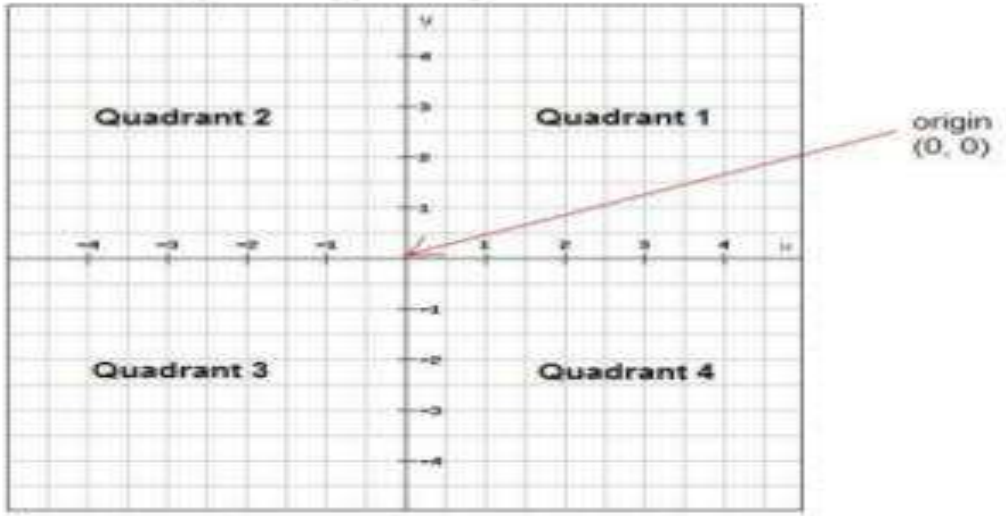
4 – 1 = 3
8 – 5 = 3
12 – 9 = 3
13-16=3

The class width is 3.

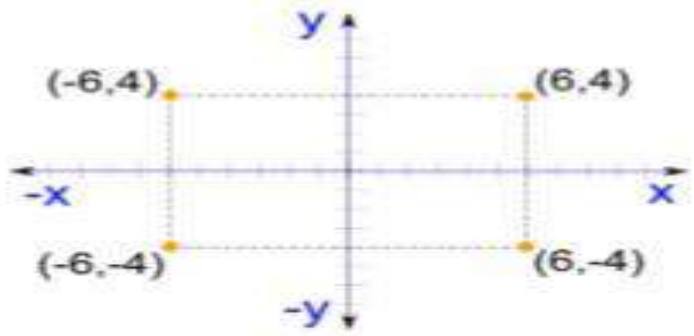
- **What is a graph?**
- It is a diagram that exhibits a relationship, often functional, between two sets of numbers as a set of points having coordinates determined by the relationship.



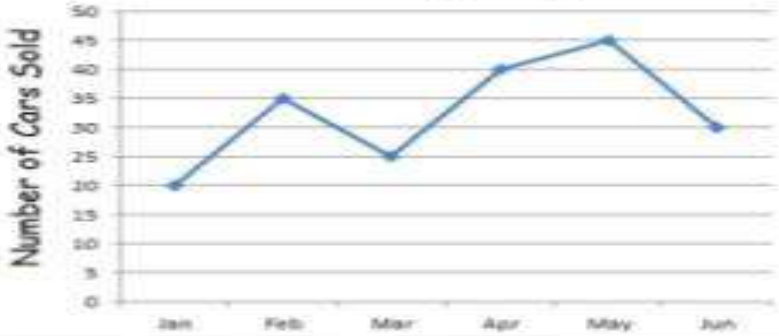
- **Origin (0,0)**



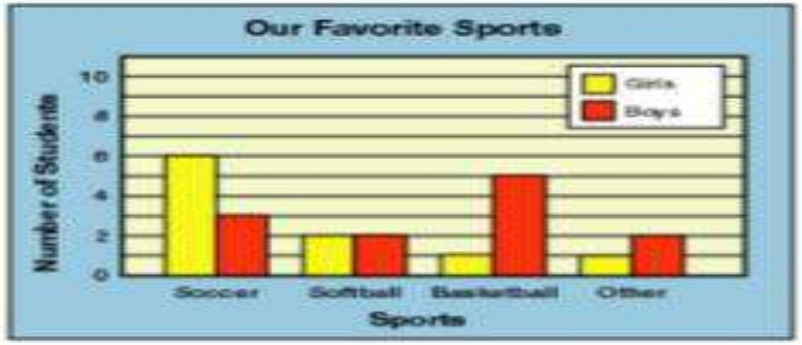
Example



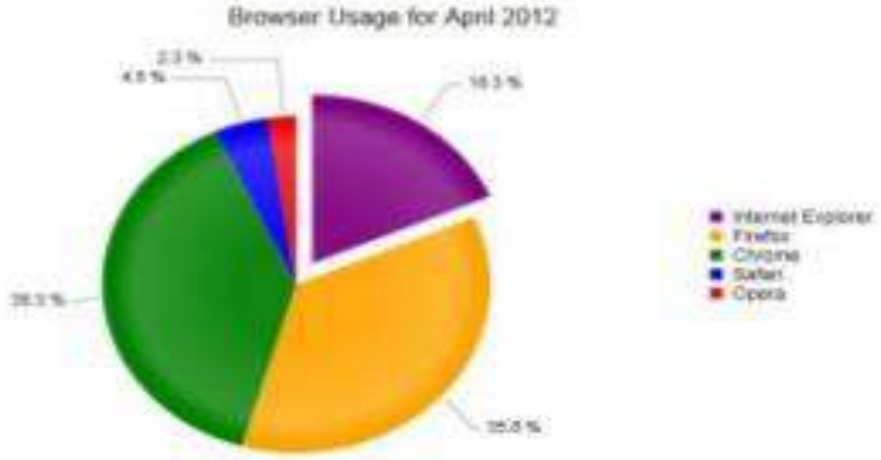
- **Line Graph**
- A *line graph* is useful in displaying data or information that changes continuously over time.
- The points on a *line graph* are connected by a *line*.
- Another name for a *line graph* is a *line* chart.



- **Bar Graph**
- A *bar graph* is a chart that uses either horizontal or vertical *bars* to show comparisons among categories.



- **Pie Chart**
- A *pie chart* (or a circle *chart*) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.



Unit 2 :

Measures Of Central Tendency

CENTRAL TENDENCY

A single number used to represent the data is referred as an **average**. The average will be the central value of the data and so averages are known as measures, of central tendency, By the term "central tendency" of a given data, we mean that the central value of the data about which the observations are concentrated.

For Eg. Average Marks, Average Profit, Average run rate of team in one day

Objectives of average :

1. To obtain a single representative quantity for the entire data.
2. To facilitate comparison.

There are several averages in use, hence it is necessary to discuss the requisites of good or ideal average.

Requisites of ideal average

1. It should be simple to understand and easy to calculate.
2. It should be rigidly defined.
3. It should be based on all observations in the data.
4. It should be capable of further mathematical treatment.
5. It should be least affected by extreme observations.
6. It should possess sampling stability.

Types of Averages

The three commonly used measures of central tendency are

(i) Arithmetic mean

(ii) Median

(iii) Mode.

ARITHMETIC MEAN

Definition:

Arithmetic mean (A. M.) or simply mean is defined as the ratio of total of all the observations to the number observations.

$$\text{A.M.} = \frac{\text{Total of all the observations}}{\text{No.of observations}}$$

CALCULATION OF ARITHMETIC MEAN :

Case 1 :

UNGROUPED DATA OR INDIVIDUAL OBSERVATIONS

Suppose there are n observations $X_1, X_2, \ldots, X_n$. The arithmetic mean of these observations is denoted by $\bar{X}$ and is calculated as ,

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sigma X}{n}$$

Example :

The monthly income (in Rs.) of 10 families in a village is as follows

1200, 1000, 1100, 1250, 950, 1300, 1150, 1200, 1050

Find the average income of these families.

Sol. Here, $n = 10$

$\sum X = 1200 + 1000 + \ldots\ldots + 1050 = 11550$

Arithmetic Mean $= \frac{\sigma X}{}$

$= \frac{11550}{10}$

$=$ Rs. 1155

CASE 2 :

DISCRETE FREQUENCY DISTRIBUTION

Suppose there are n observations $x_1, x_2, \ldots\ldots x_n$ with frequencies $f_1, f_2, \ldots\ldots f_n$ respectively.

In tabular form they are,

Values	$x1$ $x2$ ..... xn
Frequencies :	$f1$ $f1$ ..... fn

The arithmetic mean of such data is given by,

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 \ldots\ldots + f_n x_n}{f_1 + f_2 + \ldots\ldots f_n}$$

$$= \frac{\sigma f x}{N}$$

Where $N = \sigma f$

EXAMPLE Calculate arithmetic mean for the following frequency distribution

X	5	6	7	8	9	10
F	6	10	9	6	5	4

Sol. To calculate arithmetic mean, let us prepare following table :

X	f	fx
5	6	30
6	10	60
7	9	63
8	6	48
9	5	45
10	4	40
Total	40	286

$N = \sum f = 40, \sum f X = 286$

$\therefore$ Arithmetic mean $= \bar{X} = \frac{\sum f X}{\sum f} = \frac{286}{40} = 7.15$

CASE 3 :

CONTINUOUS FREQUENCY DISTRIBUTION

In continuous frequency distribution, frequency is associated with class and not a single value. For calculation purpose we assume that the frequency is associated with mid-value of class.

Let $X_1, X_2, \ldots\ldots X_n$ be the mid – value of n classes with frequencies as $f_1, f_2, \ldots\ldots f_n$ respectively. The arithmetic mean of such a data is given by,

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \ldots\ldots + f_n x_n}{f_1 + f_2 + \ldots\ldots f_n}$$

$$= \frac{\Sigma f x}{N}$$

Where $= \Sigma f$

EXAMPLE

Compute arithmetic mean for the following frequency distribution

Profit (Rs.) Per shop.	**0-100**	**100-200**	**200-300**	**300-400**	**400-500**
No. of shops	10	18	27	13	12

Solution

Classes	Mid-Values (X)	Frequencies (f)	fx.
0 – 100	50	10	500
100 – 200	150	18	2700
200 – 300	250	27	6750
300 – 400	350	13	4550
400 – 500	450	12	5400
	Total	80	19900

$$N = \sum f = 80, \sum fx = 19900$$

$$\text{Arithmetic mean} = \bar{X} = \frac{\sum fx}{N}$$

$$= \frac{19900}{80}$$

$$= \text{Rs. } 248.75$$

Example 1:

The following is a distribution of weekly salaries of the employees of a firm.

Salary in ₹	No. of employees
0 – 500	2
500 – 1000	8
1000 – 1500	12
1500 – 2000	23
2000 – 2500	25
2500 – 3000	20
3000 – 3500	9
3500 – 4000	1

Compute arithmetic mean of salaries.

Example 2 : *A variable takes values $a, a+d, a+2d, \ldots, a+(n-1)d$. Find its arithmetic mean.*

Solution : Here $x_1 = a,\ x_2 = a + d, \ldots, x_i = a + (i-1)\,d.$

$$\therefore \quad \bar{x} = \frac{\sum x_i}{n} = \left(\begin{array}{c}\text{Sum of n terms in Arithmetic}\\ \text{progression with first term a and}\\ \text{common difference d}\end{array}\right) \div n$$

$$= \frac{S_n}{n} = \frac{na + \frac{n(n-1)}{2}d}{n} = a + \frac{(n-1)\,d}{2}$$

Example 3 : *A variable takes values 1, 2, ..., n with frequencies 1, 2, 3, ..., n respectively. Find its arithmetic mean.*

Solution :

x_i	1	2	3	.....	n	Total
f_i	1	2	3	.....	n	$n(n+1)/2$
$f_i x_i$	1^2	2^2	3^2	.....	n^2	$n(n+1)(2n+1)/6$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{n(n+1)(2n+1)}{6} \div \frac{n(n+1)}{2} = \frac{2n+1}{3}$$

Example 4 : *Arithmetic mean of weight of 100 boys is 50 kg and the arithmetic mean of 50 girls is 45 kg. Calculate the arithmetic mean of combined group of boys and girls.*

Solution : Let $\bar{x}_1$ and n_1 be the mean and size of group of boys and $\bar{x}_2$ and n_2 be the mean and size of group of girls. So $n_1 = 100$, $\bar{x}_1 = 50$, $n_2 = 50$, $\bar{x}_2 = 45$. Hence, combined mean is

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(100 \times 50) + (50 \times 45)}{100 + 50} = \frac{7250}{150} = 48.3333$$

Example 5 : *Find the arithmetic mean given that* $\sum(x - 10) = 230$ *and* $n = 50$

Solution : Let $u = x - 10$, hence $\sum u = 230$

$$\therefore \quad \bar{x} = 10 + \bar{u} = 10 + \frac{230}{50} = 14.6$$

Example 6 : *Arithmetic mean of 50 items is 104. While checking, it was noticed that observation 98 was misread as 89. Find the correct value of mean.*

Solution : Incorrect mean $= 104 = \dfrac{\text{Incorrect sum}}{n}$

$\therefore$ Incorrect sum $= 104 \times 50 = 5200$

Correct sum = Incorrect sum + Correct observation – Incorrect observation.

$= 5200 + 98 - 89 = 5209$

$\therefore$ Correct mean $= \dfrac{\text{Correct sum}}{n} = \dfrac{5209}{50} = 104.18$

Example

From the following data find the missing frequencies, it is given that mean is 15.3818 and total frequency is 55

Class	9-11	11-13	13-15	15-17	17-19	19-21
Frequency	3	7	---	20	---	5

Solution: Let the missing frequencies be a and b

Class	Mid value (xi)	Frequency (fi)	fixi
9-11	10	3	30
11-13	12	7	84
13-15	14	a	14a
15-17	16	20	320
17-19	18	b	18b
19-21	20	5	100
	Total	35+a+b =N=55	534+14a+18b =∑fixi

We get two equations from the given information

i.e. $35+a+b = 55$

Therefore $a+b= 20$ ..... (1)

$$\bar{X} = \frac{\Sigma fx}{N}$$

$$15.3818 = \frac{534 + 14a + 18b}{55}$$

$$845.999 = 534+14a+18b$$

$$14a+18b = 311.999 \quad \text{.....(2)}$$

Solving (1) and (2) we get, $a= 12.0002$, $b= 7.9998$

After rounding off the values, $a = 12$ and $b=8$

Thus, frequency of the class 13-15 is 12 and that of 17-19 is 8.

Find the arithmetic mean using shortcut method

Wages (X)	3-6	6-9	9-12	12-15	15-18
Number of workers (f)	10	20	30	40	50

Wages	Midpoint of X	Number of workers (F)	A= 10.5 d = x - A	Fd
3-6	4.5	10	4.5 – 10.5 = -6	-60
6-9	7.5	20	7.5 – 10.5 = -3	-60
9-12	10.5 = A	30	10.5 – 10.5 = 0	0
12-15	13.5	40	13.5 – 10.5 = 3	120
15-18	16.5	50	16.5 – 10.5 = 6	300
		$\sum f = 150$		$\sum fd = 300$

Arithmetic mean using shortcut method, $X = A + (\sum Fd / \sum F)$

A= assumed mean in the above series

= 10.5+(300/150)

=10.5+2

=12.5

Example 2 : Find the mean of the following frequency distribution :

Class interval	0-10	10-20	20-30	30-40	40-50
Number of workers (f)	7	10	15	8	10

Class interval	Class mark (xi)	Frequency (fi)	di = xi - 25	fi di
0 - 10	5	7	-20	-140
10 - 20	15	10	-10	-100
20 - 30	25	15	0	0
30 - 40	35	8	10	80
40 - 50	45	10	20	200
		Σ fi = 50		40

Assumed mean A = 25;
N = 50
Σfi di = 40

$$\overline{x} = A + h\left[\frac{1}{N}\sum fi\ ui\right]$$

⇒ Mean = 25 + 40 / 50
⇒ Mean = 25.8

(iii) Step-deviation Method

It is the most simplified method, with short calculations. A common factor is taken and all the deviations dx are divided by that factor, say i; and then the following formula is applied.

$$\bar{X} = A + \frac{\Sigma f d'x}{N} \times i$$

Where A is Assumed mean ; f the frequency :

$N = \Sigma f$, i is common factor such that $d'x = \frac{dx}{i}$ ...(i)

Example 3. Use step deviation method to find $\bar{X}$ for the data given in example 1.

Solution Let Assumed Mean (A) be = 45

Income X	Mid value m	f	$d'x = \frac{m-A}{i}$	fdx'	
10–20	15	4	$\frac{15-45}{10} = -3$	−12	
20–30	25	7	$\frac{25-45}{10} = -2$	−14	−42
30–40	35	16	$\frac{35-45}{10} = -1$	−16	
40–50	45	20	$\frac{45-45}{10} = 0$	0	
50–60	55	15	$\frac{55-45}{10} = 1$	15	31
60–70	65	8	$\frac{65-45}{10} = 2$	16	
		N = 70		$\Sigma f d'x = -11$	

$$\bar{X} = A + \frac{\Sigma f d'x}{N} \times i$$

$$\therefore \bar{X} = 45 + \frac{-11}{70} \times 10$$

$$= 45 - \frac{11}{7} = 45 - 1.57 = \mathbf{43.43.}$$

Note : If length of class intervals is equal then we can take $d'x$ direclyas in this problem or in the next problems.

Weighted Arithmetic Mean

Symbilically ; $\bar{X}_w = \frac{\Sigma WX}{\Sigma W}$ Where X is variable, W is assigned weight and $\bar{X}_w$ is weighted Arithmetic Mean.

Very Important Note :

To achieve $\bar{X}_w$ we use the same procedure as is used to find $\bar{X}$. In weighted A.M., W is taken instead of f. Formula is $\frac{\Sigma WX}{\Sigma W}$ instead of $\frac{\Sigma fX}{\Sigma f}$.

Example 1. Following are given the marks obtained and weights assigned to the subjects of a student. Calculate Weighted A.M.

Subject	English	Punjabi	Economics	Mathematics	Accounts
Marks :	60	65	53	50	40
Weight :	1	1	2	3	4

Solution

Subject	Marks X	Weights (W)	Product (XW)
English	60	1	60
Punjabi	65	1	65
Ecnomics	53	2	106
Mathematics	50	3	150
Accounts	40	4	160
		$\Sigma W = 11$	$\Sigma XW = 541$

$\Sigma XW = 541$; $\Sigma W = 11$

$\therefore \bar{X}_W = \frac{\Sigma XW}{\Sigma W} = \frac{541}{11}$

$= 49.18.$

(a) *Inclusive Series :*

Series such as 5–9, 10–14, 15– 19, 20–24, ... is known as inclusive series.

Example 1. Find A.M. for the data given below (Inclusive series).

Class Interval :	4–6	7–9	10–12	13–15	16–18	19–21	22–24
Frequency :	1	3	7	15	11	3	2

Solution

As the difference between upper limit of one interval and lower limit of next interval is one, we will deduct and add .5 from lower limit and to upper limit of every interval to make intervals exclusive.

Let Assumed mean (A) be = 14 ; $i = 3$

C.I.	C.I. Improved	Mid Points m	Frequency f	Step deviation $d'x$	$fd'x$	
4–6	3.5–6.5	5	1	–3	–3	
7–9	6.5–9.5	8	3	–2	–6	–16
10–12	9.5–12.5	11	7	–1	–7	
13–15	12.5–15.5	14	15	0	0	
16–18	15.5–18.5	17	11	1	11	
19–21	18.5–21.5	20	3	2	6	+23
22–24	21.5–24.5	23	2	3	6	
			N = 42		$\Sigma fdx = 7$	

As $\bar{X} = A + \frac{\Sigma fd'x}{N} \times i$

$\therefore \bar{X} = 14 + \frac{7}{42} \times 3$

$= 14 + .5 = 14.5$

(b) *Open End Intervals*

These are those intervals or classes, where either the lower limit of first interval or the upper limit of last interval or these both are not given.

Example 2. Find A.M. for the following data (Open end Interval).

Class Intervals :	Below 20	20–30	30–40	40–50	50–60	60–70	Above 70
Frequency :	6	10	24	28	14	5	3

Solution

As each class interval has length of 10 units. Hence lowest class interval will be 10–20 and highest as 70–80 and Let Assumed Mean (A) be = 45 ; $i = 10$

C.I.	M.P. m	f	d 'x	fd 'x	
10–20	15	6	–3	–18	
20–30	25	10	–2	–20	–62
30–40	35	24	–1	–24	
40–50	45	28	0	0	
50–60	55	14	1	14	
60–70	65	5	2	10	33
70–80	75	3	3	9	
		N = 90		$\Sigma fd'x = -29$	

$$\text{As } \bar{X} = A + \frac{\Sigma fdx'}{N} \times i$$

$$\therefore \bar{X} = 45 + \frac{-29}{90} \times 10$$

$$= 45 + \frac{-29}{9}$$

$$= 45 - 3.22 = \mathbf{41.78.}$$

(c) *Cumulative series*

These series are of two types e.g.; (*i*) Less than (*ii*) More than OR (*i*) Not above (*ii*) Not below :

Example 3. Find A.M. for following data (Cumulative Series)

Class Interval :	Below 50	Below 60	Below 70	Below 80	Below 90	Below 100
Frequency :	3	11	34	59	72	80

Solution

Let Assumed Mean (A) be = 75 ; $i = 10$

C.I.	Mid point m	f	$d'x$	$fd'x$	
40–50	45	= 3	–3	–9	
50–60	55	11–3 = 8	–2	–16	–48
60–70	65	34–11 = 23	–1	–23	
70–80	75	59–34 = 25	0	0	
80–90	85	72–59 = 13	1	13	+29
90–100	95	80–72 = 8	2	16	
		N = 80		$\Sigma fd'x = -19$	

As $\bar{X} = A + \frac{\Sigma fd'x}{N} \times i$

$\therefore \bar{X} = 75 + \frac{-19}{80} \times 10$

$= 75 - \frac{19}{8}$

$= 75 - 2.375 = \mathbf{72.625.}$

(**Note.** These class intervals will also be obtained if given class intervals and frequencies are as given below.)

Intervals :	Above 40	Above 50	Above 60	Above 70	Above 80	Above 90
Frequency :	80	77	69	46	21	8

In Exclusive Form

Class Intervals :	40–50	50–60	60–70	70–80	80–90	90–100
Frequency :	(80–77)=3	(77–69)=8	(69–46)=23	(46–21)=25	(21–8)=13	(8–0) = 8

(d) *When Middle Points are given*

When middle points are given, we convert it into exclusive series noting the difference between each mid point, we get the length of each interval as follows.

Given :

Example 4. For the following data, calculate $\bar{X}$ (Mid points given)

Mid points :	6	10	14	18	22	26	30
Frequency :	2	7	18	29	17	11	6

Solution

Difference between each mid point = 4.

$\therefore$ Deducting and adding 2 = $\left(\frac{4}{2}\right)$ to each mid point ; We get the intervals as below.

C.I.	Mid Point m	Frequency f	Step-deviation d 'x	fd 'x	
4–8	6	2	–3	–6	
8–12	10	7	–2	–14	–38
12–16	14	18	–1	–18	
16–20	18	29	0	0	
20–24	22	17	1	17	
24–28	26	11	2	22	+57
28–32	30	6	3	18	
		N = 90		Σfd 'x = 19	

Let Assumed Mean (A) be = 18 ; $i = 4$.

$$\bar{X} = A + \frac{\Sigma fd'x}{N} \times i$$

$$= 18 + \frac{19}{90} \times 4 = 18 + \frac{76}{90}$$

$$= 18 + .84 = \mathbf{18.84.}$$

Note. It is not necessary to obtain class intervals, but it is necessary to do if lower limit is needed to obtain Median Decile, Percentile or Mode.

(e) Case of Unequal Intervals

Example 5. Calculate $\bar{X}$ for following series.

Class Interval :	4–8	8–20	20–28	28–44	44–68	68–80
Frequency :	3	8	12	21	10	6

Solution

Let Assumed mean (A) be = 36 ; And Class Interval (i) = 2 (For IInd Method)

C.I.	M.P. m	f	fm	m–A =dx	fdx		$d'x = \frac{dx}{2}$	$fd'x$	
4–8	6	3	18	–30	–90		–15	–45	
8–20	14	8	112	–22	–176	–410	11	–88	–205
20–28	24	12	288	–12	–144		–6	–72	
28–44	36	21	756	0	0		0	0	
44–68	56	10	560	20	+200	+428	10	100	214
68–80	74	6	444	38	+228		19	114	
		N = 60	Σfm =2178		Σfdx = 18		$\Sigma fd'x$ =9	$\Sigma fd'x$ =9	

Merits and Demerits of Arithmetic mean

Merits :

1. It is easy to calculate and simple to follow.
2. It is based on all observations.
3. It is rigidly defined.
4. It possesses sampling stability.
5. It is capable of further mathematical treatment. Given the means and sizes of two or more groups, we can find mean of combined group.

Demerits :

1. It is applicable only for quantitative data.
2. It is unduly affected by extreme observations.
3. It cannot be computed for frequency distribution with open end class. (For an open end class we cannot find mid point).
4. It cannot be determined graphically.
5. Sometimes arithmetic mean may not be an actual observation in a data.

MEDIAN

The median is that value of the data which divides the data in two equal groups, one group consists of all the observations greater than median and the other group consists of all the observations smaller than median.

In other words if the observations are arranged in increasing or decreasing order, then the middle observation will be the median.

COMPUTATION OF MEDIAN

Case I :

UNGROUPED DATA OR INDIVIDUAL OBSERVATIONS

In case of ungrouped data, to find median firstly arrange the observation in increasing or decreasing order, Let there be n observations.

Case :

I .If *n* is **odd** number, then the median is $\left(\frac{n+}{1_2}\right)$ observation.

II. If *n* is **even** number, then the median is

$$\frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

Example : Compute the median for the following data:

3000, 3500, 2500, 3050, 3200, 2800, 2900

Solution :

Firstly arrange the observations in increasing order.

2500, 2800, 2900, 3000, 3050, 3200, 3500

Here $n = 7$, a odd number

$$\therefore \text{Median} = \left(\frac{n}{2} + 1\right)^{th} \text{ observation}$$

$$= 4^{\text{th}} \text{ observation}$$

Median = 3000

Example: Following are the temperatures recorded in a city during first 10 days of a month. Obtain the median temperature,

40, 42, 38, 37, 41, 40, 37, 38, 35, 42

Solution: The Observations in increasing order are :

35, 37, 37, 38, 38, 40, 40, 41, 42, 42

Here $n = 10$, a even number

$$\therefore \text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{obs} + \left(\frac{n}{2}+1\right)^{th}\text{obs}}{2}$$

$$= \frac{5^{th}\text{obs} + 6^{th}\text{obs}}{2}$$

$$= \frac{38 + 40}{2}$$

Median = 39.

Case 2 :

DISCRETE FREQUENCY DISTRIBUTION

In case of discrete frequency distribution, the median is calculated using following steps :

Step 1 : Obtain less than cumulative frequency distribution.

Step 2 : Obtain N, the total of all frequency.

Step 3 : If N is odd, median will be the value of $\left(\frac{N}{2}+1\right)^{th}$ observation

If N is even, median will be the mean of $\left(\frac{N}{2}\right)^{th}$ and $\left(\frac{N}{2}+1\right)^{th}$ observation

Example: For the following frequency distribution,

size of item	2	4	6	8	10	12
frequency	6	10	20	24	12	8

Compute Median.

Solution: Firstly we will obtain the less than cumulative frequency distribution.

size of item (xi)	frequency (fi)	LCF
2	6	6
4	10	16
6	20	36
8	24	60
10	12	72
12	8	80
	N= **80**	

Here, $N = 80$ $\therefore \frac{N}{2} = 40$

C.f. just greater than 40 is 60, which corresponds to 8.

Median $= 8$

Case 3 :

CONTINUOUS FREQUENCY DISTRIBUTION

Step 1 : Obtain the class boundaries.

Step 2 : Obtain less than cumulative frequencies.

Step 3 : Locate the median class. Median class is the class in which median i.e. $\left(\frac{N}{2}\right)^{th}$ observation falls. In other words, it is in a class where less than cumulative frequency is equal to or exceeds N/2 for the first time.

Step 4 : Apply the formula and find the median.

$$\text{Median} = l + \frac{N/2 - \text{c.f.}}{f} \times h$$

where, l = lower boundary of the median class

N = total frequency

c.f. = less than cumulative frequency of the class just preceding the median class.

f = frequency of median class

h = class width

Example : For the following frequency distribution find the median

Class	0-100	100-200	200-300	300-400	400-500	500-600	600-700
Frequency	9	15	18	21	18	14	5

Solution : Firstly we will obtain less than cumulative frequency distribution

Classes	Frequencies	Lcf
0 —100	9	9
100-—200	15	24
200 —300	18	42
300 —400	21	63
400 —500	18	81
500 —600	14	95
600— 700	5	100

Here $N = \Sigma f = 100 \quad \therefore \frac{N}{2} = 50$

Median class is the class for which the less than cum. Frequency-is just greater than 50.

$\therefore$ Median class = 300 – 400

$$\text{Median} = l + \frac{N/2 - \text{c.f.}}{f} \times h$$

L = 300, *h*= 100, *f* = 21, c.f. = 42

$$\text{Median} = 300 + \frac{50 - 42}{21} \times 100$$

= 300 + 38.0952

= 338.095238

Example : Find the missing frequency from the following data, given that the median mark is 23.

Marks	0-10	10-20	20-30	30-40	40-50
No. of Students	5	8	?	6	3

Solution : Let the missing frequency be x.

Let us obtain less than cumulative frequencies for the given data.

Marks	Frequencies	Less than cum frequency
0 – 10	5	5
10 – 20	8	13
20 – 30	X	$13 + x$
30 – 40	6	$19 + x$
40 – 50	3	$22 + x$

Since We have given that median = 23, which lies in the class 20 – 30.

$\therefore$ Median class = 20 – 30

$\therefore l = 20, b = 10, f = \text{x}$, c.f. = 13, N = Median $= l + \dfrac{\text{N/2} - \text{c.f.}}{\text{f}} \times \text{h}$

$$\therefore 23 = 20 + \frac{10}{x}\left[\frac{22 + x}{2} - 13\right]$$

$$\therefore 23 - 20 = \frac{10}{x}\left[\frac{22 + x - 26}{2}\right]$$

$$\therefore 3 = \frac{10}{x}\left(\frac{x-4}{2}\right)$$

$$\therefore \frac{3}{1} = \frac{5}{1}\left(\frac{x-4}{2}\right)$$

$$\therefore 3\,(x) = 1 \times 5\,(x - 4)$$

$$\therefore 3x = 5x - 20$$

$$\therefore 5x - 3x = 20$$

$$\therefore 2x = 20$$

$$\therefore x = 10$$

$\therefore$ missing frequency is $= 10$

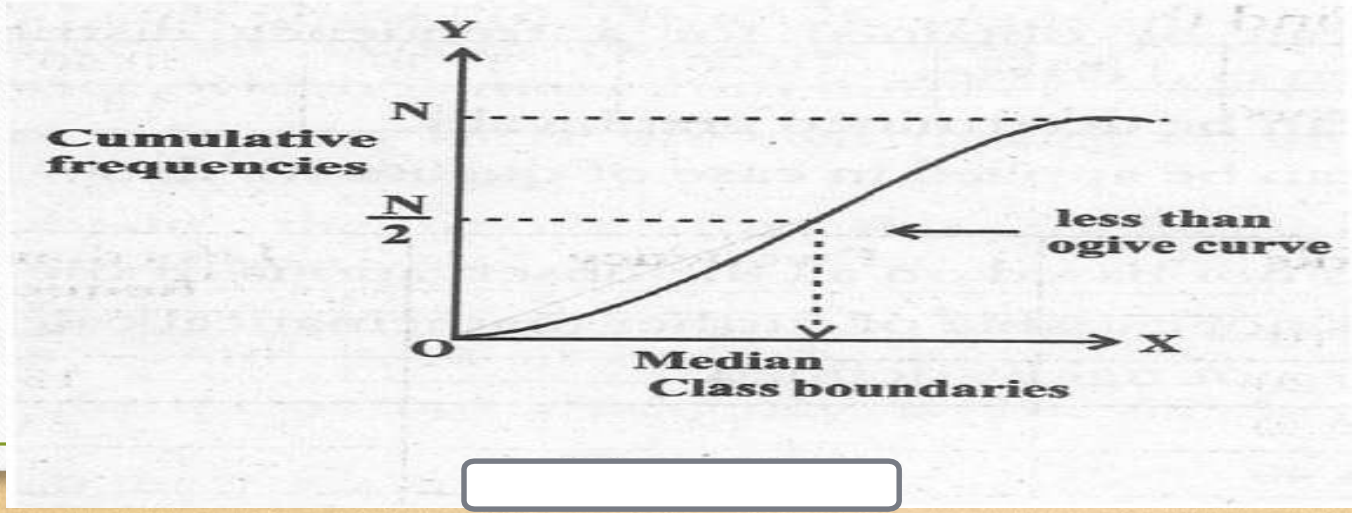
Computation of median –By Graphical Method

Median can be obtained graphically by means of ogive curve.

Plot less than cumulative frequency curve taking upper boundaries on X axis, and less than cumulative frequency on Y- axis Draw a line parallel to X axis

passing through the point ~~on~~ N/2 Y-axis. From the point of

- intersection of this line and ogive curve, draw a
- perpendicular to X-axis. The value at the foot of this perpendicular (on X-axis) is the value of median

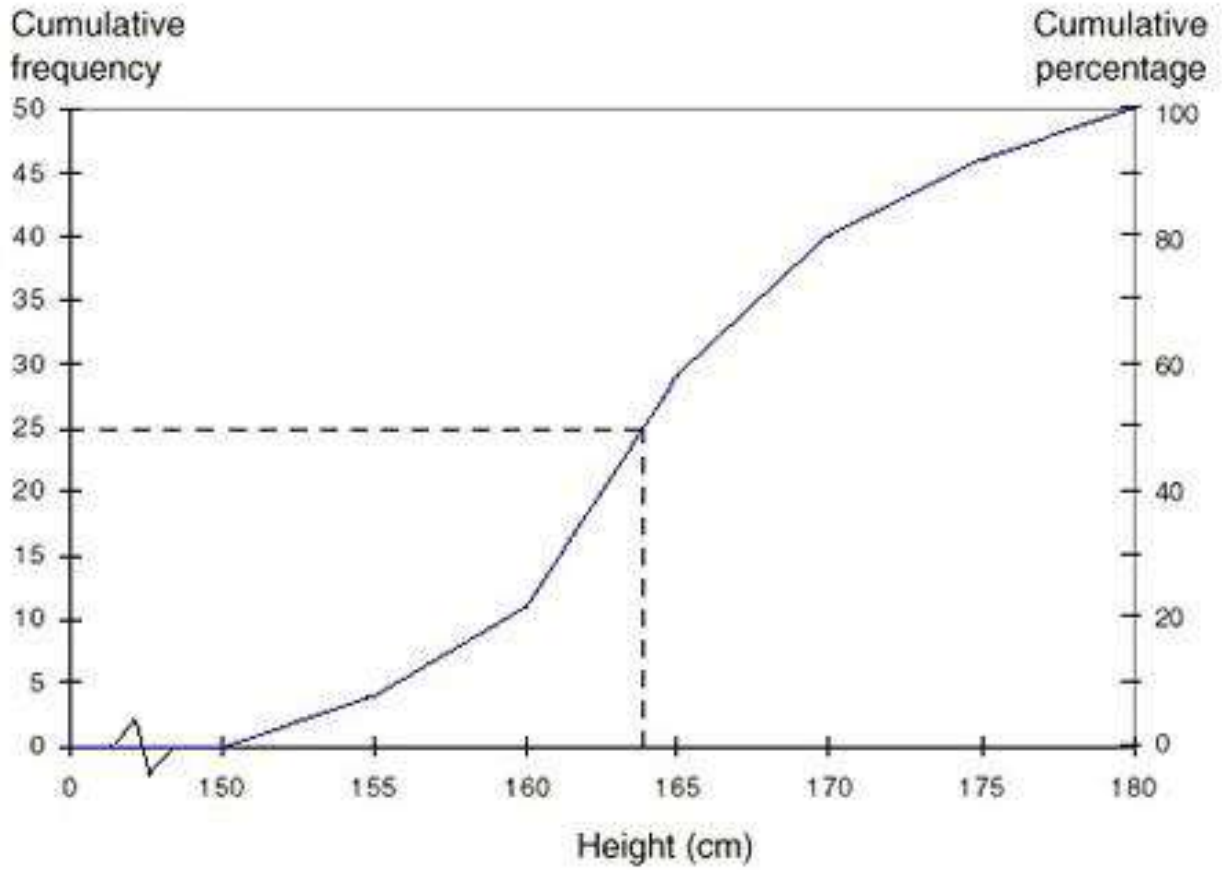


Example : Find the median graphically using following data :

Height (cm)	Frequency
150-155	4
155-160	7
160-165	18
165-170	11
170-175	6
175-180	4

Solution: Here N = 50 N/2 = 25

Height (cm)	Frequency	L.C.F
150-155	4	4
155-160	7	11
160-165	18	29
165-170	11	40
170-175	6	46
175-180	4	50



By just looking at the graph, you can try to find the median value. The median is the point where the x-axis (Height) intersects with the midpoint (25) of the y-axis (Cumulative frequency). You will see that the median value is approximately 164 cm. Using mathematical calculations, you can find out that the value is actually 163.9 cm.

MERITS AND DEMERITS OF MEDIAN

Merits and Demerits of Median :

Merits :

1. It is easy to understand and easy to calculate.
2. It is not affected due to extreme observations.
3. It can be computed for a distribution with open end classes.
4. It can be determined graphically.
5. It is applicable to qualitative data also. In this case observations are arranged in order according to the quality and the middle most observation is obtained. The quality of this item is taken to be average quality or median quality.

Demerits :

1. It is not based on all the observations, hence it is not proper representative.
2. It is not capable of further mathematical treatment.
3. It is not as rigidly defined as the arithmetic mean.

MODE

Mode is the value of the data which occurs most frequently in the data.

It is the most repeated observation of the data.

Example :

In election results, a party with largest votes is considered to be a representative of that area.

Here mode is appropriate average. Similarly modal shoe size is which demanded by the largest number of people, modal wage is the wage that is earned by more workers than any other wage.

COMPUTATION OF MODE

Case I :

UNGROUPED DATA AND DISCRETE FREQUENCY DISTRIBUTION

For ungrouped data as well as for discrete frequency distribution, mode is that value which occurs maximum number of times i.e. it is the value with highest frequency in the data.

Example 1 : Find the mode of the following data :

31, 35, 40, 31, 30, 35, 38, 30, 31, 32.

Solution : Here 31 repeats 3 times which is highest number of times :

∴ Mode = 31.

Example 2 : Find the mode of the following data :

40, 42, 35, 40, 38, 42, 37, 32

Solution : Here 40 occurs twice as well as 42 also occurs twice, which is maximum number of times.

∴ for this data there are two modes 40 and 42. Thus data are bimodal.

Example 3 : Obtain the mode for the following frequency distribution

X	5	6	7	8	9	10	11
Y	8	10	18	14	11	7	3

Solution : Here maximum frequency is 18 which occurs for X = 7,

∴ Mode = 7

Case II : CONTINUOUS FREQUENCY DISTRIBUTION

Step 1 : Obtain the class – boundaries.

Step 2 : Locate the modal class. Modal class is class in which mode lies or a class with the largest frequency.

Step 3 : Apply the formula and find the mode.

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

where,

l = Lower boundary of modal class.

f_m = Frequency of modal class.

f_1 = Frequency of pre-modal class.

f_2 = Frequency of post-modal class.

h = Width of modal class.

Example : *Calculate modal income from the following income distribution :*

Daily income (Rs.)	30 and below	31-60	61-90	91-120	121-150	above 150
No. of Persons	22	198	110	95	42	33

Solution :

Class boundaries	Frequency
below 30.5	$f_1 = 22$
30.5 – 60.5	$f_m = 198$ Modal class
60.5 – 90.5	$f_2 = 110$
90.5 – 120.5	95
120.5 – 150.5	42
above 150.5	33

Modal class is 31–60.

Here we get $l = 30.5$, $f_m = 198$, $f_1 = 22$, $f_2 = 110$, $h = 30$

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

$$= 30.5 + \frac{198 - 22}{2 \times 198 - 22 - 110} \times 30 = 50.5$$

Note :

1. If the maximum frequency is repeated, to find the mode uniquely, a method of grouping is adopted and a modal class is determined. The method of grouping is beyond the scope of book.
2. Mode cannot be determined if modal class is at the extreme. (i.e. the maximum frequency occurs at the beginning or at the end of the frequency distribution.)
3. Modal, pre-modal and post-modal classes should be of the same width.
4. If $f_1 = f_2$ then mode is the mid-point of modal class.

Computation of mode – by Empirical relation : Arithmetic mean, mode and median are averages, hence we expect that those should be identical in value. However, this is true only in ideal situation. It is true whenever the frequency curve is perfectly symmetric and bell-shaped. For a moderately asymmetric unimodal frequency distribution the following empirical relationship holds approximately.

$$\text{Mean} - \text{Mode} \cong 3\,(\text{Mean} - \text{Median}) \quad \ldots (1)$$

In some situations mode is ill-defined (see notes 1, 2 stated above). To overcome this difficulty in computing mode, the empirical relation (1) is used. If any two averages included in (1) are known, the remaining third can be computed. Therefore, if mean and median are known, then mode can be determined.

The empirical relation cannot be theoretically proved. Karl Pearson has stated it on the basis of vast experience. This relationship is observed to be valid for number of data sets after actual computations.

Example: Following is an incomplete distribution having modal marks as 44.

Marks	0-20	20-40	40-60	60-80	80-100
No. of Students	5	18	?	12	5

Find the missing frequency.

Solution : Let the missing frequency be x.

Now, since mode = 44, which lies in the class 40 – 60.

$\therefore$ Modal Class $= 40 - 60$

$f_m = x, f_1 = 18, f_2 = 12, l = 40, h = 20$

$\therefore$ Mode $= l + \left(\frac{fm - f1}{2fm - f1 - f2}\right) \times h$

$\therefore 44 = 40 + \left[\frac{(x-18)}{2x-18-12}\right] \times 20$

$\therefore 4 = \frac{x-18}{2x-30} \times 20$

$\therefore \frac{x-18}{2x-30} = \frac{4}{20} = \frac{1}{5}$

$\therefore 5(x-18) = 1(2x-30)$

$\therefore 5x - 90 = 2x - 30$

$\therefore 5x - 2x = -30 + 90$

$\therefore 3x = 60$

$\therefore x = \frac{60}{3} = 20$

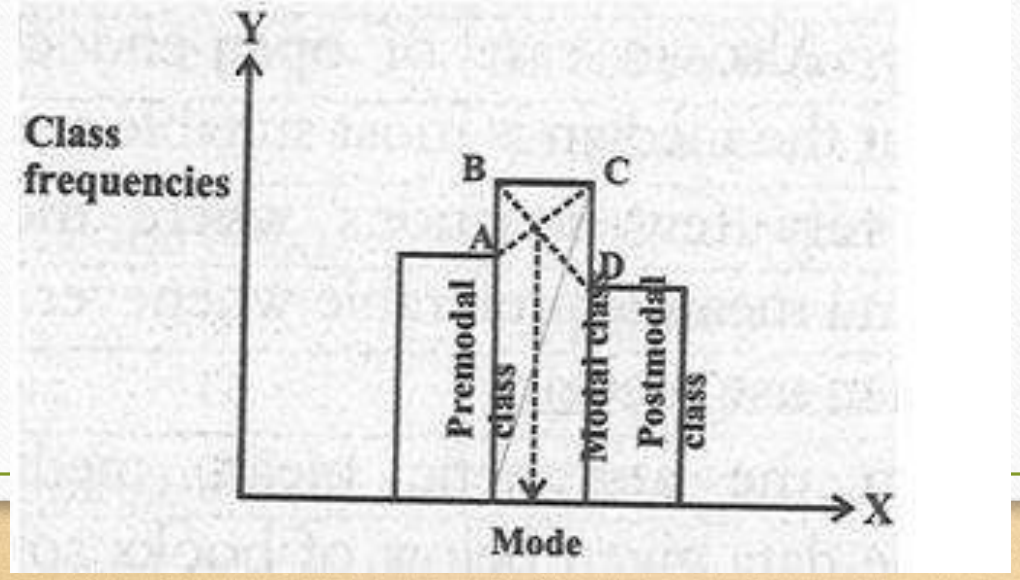
$\therefore$ Missing frequency is 20.

Computation of Mode – By Graphical Method

The Mode can be determined graphically by plotting histogram of the given distribution.

Consider the pre-modal, modal and post modal classes of the histogram as shown in given figure

From the point of the intersection of AC and BD draw perpendicular to X-axis. This foot of perpendicular represents the mode.



MERITS AND DEMERITS OF MODE

Merits :

1. It is simple to understand and easy to calculate.
2. It is applicable to quantitative as well as qualitative data.
3. It is not affected by extreme values in the data.
4. It can be determined graphically.
5. It can be obtained for a frequency distribution with open end classes.

Demerits :

1. It is not based on all the observations of the data.
2. It is not suitable when the no. of observations in the data is very small.
3. It is indeterminate when maximum frequency is at one-end of the distribution.
4. It is not capable of further mathematical calculations.

Partition Values

- Quartiles
- Percentiles
- Deciles

Quartiles, Deciles and Percentiles : Earlier we have seen that median divides the total number of observations into two equal parts. Similarly in order to make four equal parts we use quartiles, for making 10 equals parts we use deciles and for making 100 equal parts we use percentiles, when the observations are ordered.

DEFINITIONS

Quartiles:

The observations Q_1, Q_2, Q_3 which divide the total number of observations into 4 equal parts are called quartiles.

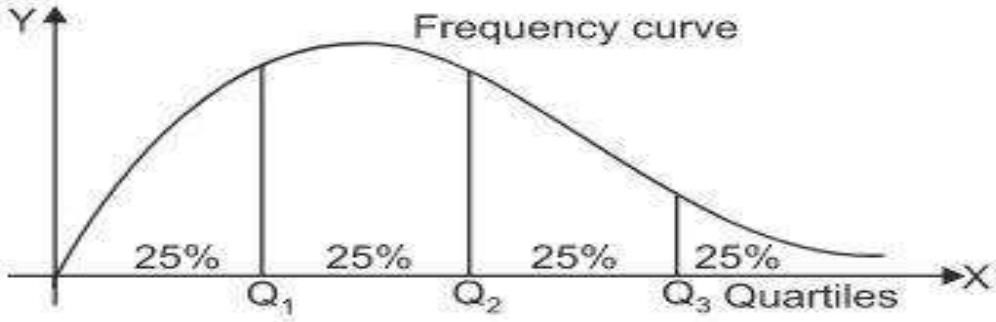
Deciles :

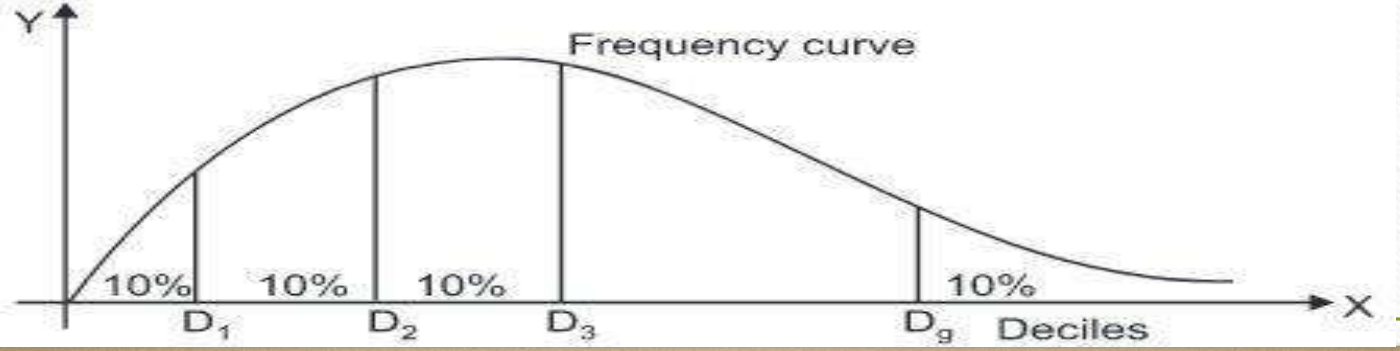
The observations D_1, D_2,….., D_9 which divide the total number of observations into 10 equal parts are called deciles.

Percentiles :

The observations P_1, P_2,….. ,P_{99} which divide the total number of observations into 100 equal parts are called percentiles.

Median, quartiles, deciles and percentiles are called partition values in common. The procedure of obtaining median is used to compute other partition values with appropriate changes. To obtain the partition values of series of individual observations, many calculations or formulae are not required. However, to compute partition values of a continuous frequency distribution, corresponding formula of median can be suitably modified. In this case, first of all less than cumulative frequency is determined. Using these cumulative frequencies a class in which partition value lies is decided and then using the formula, partition value is determined.





Unit 3 :

Measures Of Dispersion

The measures of central tendency or averages gives us an idea of the value about which the observations are concentrated. But we cannot get complete idea about the data by averages only. There may be a number of case where averages are same but data differ widely from each other. To illustrate this point, consider the following, example :

The runs scored by two batsmen in 5 one-day matches as follows :

Match No.	:	1	2	3	4	5	Total	Mean
Batsman A	:	40		45	35	30	50	200	40
Batsman B	:	120		00	60	00	20	200	40

In the above example we see that on an average both the batsmen have scored same 40 runs. But if we observe each observation, then it can be seen that batsmen A is more consistent than batsman B, because the observation of A are near about its mean 40, while for B the observations are far away from its mean 40. Thus there is less variation in the observations of A than that of B.

By dispersion we mean the variations in the observations from average. Average will be a good representative if dispersion is less. Thus the average of 6 data will be more reliable or representative if the data have less variations.

Requisites for an Measures of Dispersion

1. It should be simple to understand and easy to calculate.
2. It should be rigidly defined.
3. It should be based on all observations in the data.
4. It should be capable of further mathematical treatment.
5. It should be least affected by extreme observations.

	Absolute Measures	Relative Measures
1.	Absolute measures of dispersion are expressed in the same measuring units in which the original data are given such as Rupees, cm, kgs etc.	A measures of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is also known as coefficient of dispersion.
2.	The absolute measures of dispersion can be used for comparison only when the variables are expressed in the same units and the distributions have the same average size.	Variability of all kinds of distributions can be compared using relative measures of dispersion
3.	Range , quartile deviation , mean deviation , standard deviation , variance are examples of absolute measures of dispersion.	Coefficient of range, coefficient of quartile deviation, Coefficient , mean deviation ,, mean variation are some examples of relative measures of dispersion.

The various measures of dispersion are :

- Range and coefficient of range

- Variance
- Standard deviation
- Coefficient variation.

Range and Coefficient of Range

Definition : If L is the largest observation and S is the smallest observation then range is the difference between L and S. Thus,

$$\text{Range} = L - S$$

and the corresponding relative measure is

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

Example

: *Compute (1) range and coefficient of range the following data : 100, 24, 14, 105, 21, 35, 106, 16, 100, 72, 68, 103, 61, 90, 20, 15, 102, 104.*

Solution : (i) Here, Smallest observation (S) = 14

Largest observation (L) = 106

$$\text{Range} = L - S = 106 - 14 = 92$$

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{92}{106+14} = \frac{92}{120} = 0.7667$$

Example

The prices of shares of a certain company from Monday to Friday are as follows :

Days	: Mon	Tues	Wed	Thu	Fri
Price (in Rs.)	: 524	502	544	519	558

Calculate the range.

Sol. Here,

Largest Observation(L) = 558

Smallest Observation(S) = 502

Range = L-S

= 558 – 502 = 56

Example : A collor manufacturer is considering the production of a new style of collar to attract college students. The wing following statistics of neck circumferences are available based upon measurement of typical group of college students

Circumference of a Collar (in inches)	Number of Student
12.5	8
13.0	38
13.5	60
14.0	126
14.5	132
15.0	58
15.5	36
16.0	2
16.5	2

Calculate the range.

Sol. Here,

Largest Observation(L) $= 16.5$

Smallest Observation(S) $= 12.5$

$$\text{Range} = H - L$$
$$= 16.5 - 12.5$$
$$= 4$$

Example :The following is the distribution of weight of 300 persons.

Weight (in lbs)	Number of Persons
80-90	15
90-100	40
100-110	50
110-120	55
120-130	45
130-140	35
140-150	30
150-160	24
160-170	6
Total	300

Calculate the range and coefficient of range.

Sol. Largest observation (L) $= 170$

Smallest observation (S) $= 80$

Range $=$ L - S

$= 170 - 80$

$= 90$

$$\text{coefficient of range} = \frac{L-S}{L+S}$$

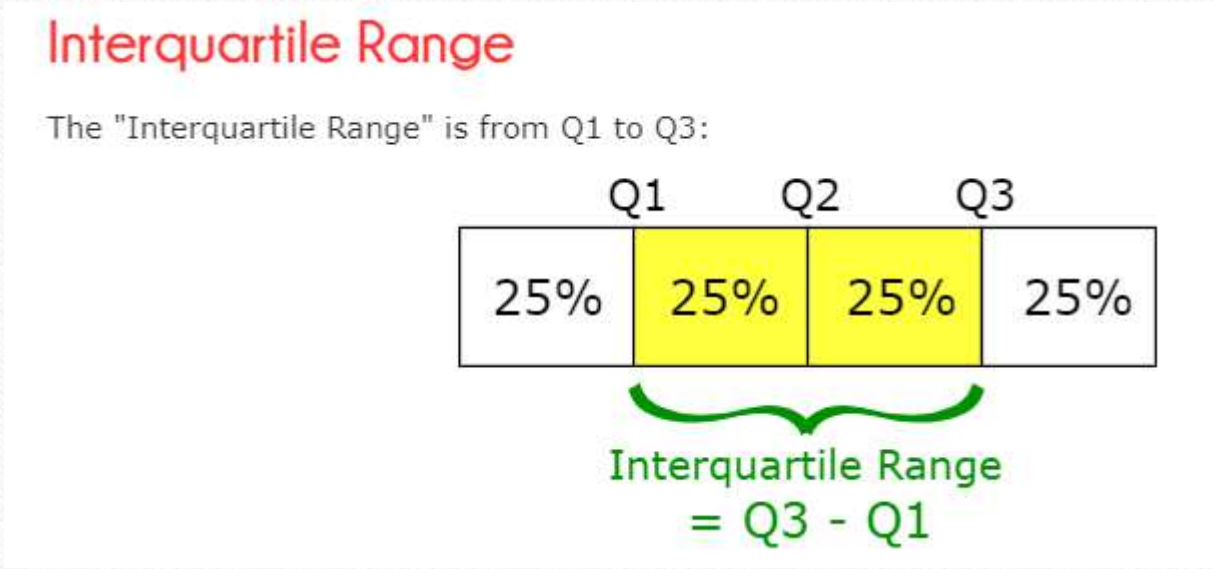
$$= \frac{170-80}{170+80}$$

$$= 0.36$$

Quartile Deviation

It is a measure of dispersion based on the upper quartile Q_3 and the lower quartile Q_1. It is also known as semi- inter quartile range and is defined as

$$\text{Quartile Deviation} = \frac{(Q_3 - Q_1)}{2}$$



HOW TO COMPUTE

Individual Series

Example 1. From the following data compute inter quartile range, quartile Deviation and Coefficient of Quartile Deviation.

24	7	11	9	17	3	20	14	4	22	27

Solution

Arranging the series in Ascending Order; N = 11

3	4	7	9	11	14	17	20	22	24	27

$$Q_1 = \text{Size of } \left(\frac{11+1}{4}\right) \text{th term} = \text{Size of 3rd term} = 7$$

$$Q_3 = \text{Size of } \frac{3(11+1)}{4} \text{ th term} = \text{Size of 9th term} = 22.$$

$$\therefore \quad \text{Inter Quartile Range} = Q_3 - Q_1 = 22 - 7 = 15$$

$$\left.\begin{array}{c}\text{Semi–Inter Quartile Range} \\ \text{or} \\ \text{Quartile Deviation (Q.D.)}\end{array}\right] = \frac{Q_3 - Q_1}{2} = \frac{15}{2} = 7.5$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{22-7}{22+7} = \frac{15}{19} = .52 \text{ (App.)}$$

Example 2. Compute co-efficient of Q.D.

34	24	11	33	4	8	17	7	14	21	13	25	29	27

Solution

N = 14; Arranging the terms in Ascending Order.

4	7	8	11	13	14	17	21	24	25	27	29	33	34

N = 14.

$$Q_1 = \text{Size } \frac{(14+1)}{4} \text{ th term} = \text{Size of } 3.75 \text{ th term}$$

= 3rd term + .75 (4th term–3rd term)

= 8 + .75 (11–8) = 8 + .75 × 3 = 8 + 2.25 = **10.25.**

$$Q_3 = \text{Size of } \frac{3\,(14+1)}{4} \text{ th term} = \text{Size of } 11.25 \text{ th term}$$

= 11th term + .25 (12th term–11th term)

= 27 + .25 (29–27) = 27 + .25 × 2 = 27 + .5 = **27.5**

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{27.5 - 10.25}{27.5 + 10.25} = \frac{17.25}{37.75} = .46$$

Discrete Series

Example 1. Compute Inter Quartile Range, Coefficient of Quartile Deviation for following data.

X :	4	8	12	16	20	24	28	32
f :	4	9	17	40	53	37	24	16

Solution

X	f	C.f.
4	4	4
8	9	13
12	17	30
16	40	70
20	53	123
24	37	160
28	24	184
32	16	200
	N = 200	

Q_1 = Size of $\frac{(200+1)}{4}$ th term = Size of 50.25th term = 16

Q_3 = Size of $\frac{3(200+1)}{4}$ th term = Size of 150.75th term = 24

Inter Quartile Range = $Q_3 - Q_1 = 24 - 16 = 8$

Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{24-16}{24+16} = \frac{8}{40} = .2$

3.4.3.3. Computation of Quartile Deviation - Continuous Series (Exclusive)

Following are the formulae for calculating the Q_1 and Q_3 in continuous series:

$$Q_1 = L_1 + \frac{\frac{N}{4} - c}{f} \times i \quad \text{and} \quad Q_3 = L_1 + \frac{\frac{3N}{4} - c}{f} \times i$$

Where, L_1 = Lower limit;
N = Number of observation,
c = cumulative frequency of class preceding to quartile class;
i = class interval

Following example explains the calculation of quartile deviation in Continuous series.

Example 7: Calculate Quartile Deviation from the following data:

Population (in thousand)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Localities	4	7	15	12	3	9	6

Solution:

Population	No. of Localities (f)	Cumulative Frequency
10-20	4	4
20-30	7	11
30-40	15	26
40-50	12	38
50-60	3	41
60-70	9	50
70-80	6	56

Q_1 = Size of $\left(\frac{N}{4}\right)^{th}$ item $= \left(\frac{56}{4}\right)^{th}$ item $= 14^{th}$ item which lies in class interval 30-40.

Here, $L_1 = 30$, $c = 11$, $f = 15$ and $i = 10$.

So, $$Q_1 = L_1 + \frac{\frac{N}{4} - c}{f} \times i = 30 + \frac{14 - 11}{15} \times 10 = 30 + \frac{30}{15} = 32$$

Q_3 = Size of $\left(\frac{3N}{4}\right)^{th}$ item $= \left(\frac{3 \times 56}{4}\right)^{th}$ item $= 42^{th}$ item which lies in class interval 60-70.

Here, $L_1 = 60$, $c = 41$, $f = 9$ and $i = 10$.

(C) Continuous Series

Example 1. Compute Coefficient of Quartile Deviation for following data.

Class Interval :	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency :	8	16	22	30	24	12	6

Solution

C.I.	f	C.f.
0–10	8	8
10–20	16	24
20–30	22	46
30–40	30	76
40–50	24	100
50–60	12	112
60–70	6	118
	N = 118	

N_1 for $Q_1 = \frac{N}{4} = \frac{118}{4} = 29.5$ and $C.f. = 24$

$L = 20$; $f = 22$; $i = 10$

$$Q_1 = L + \frac{N_1 - C_f}{f} \times i = 20 + \frac{29.5 - 24}{22} \times 10$$

$$= 20 + \frac{55}{22} = 20 + 2.5 = \mathbf{22.5}$$

$$N_1 \text{ for } Q_3 = \frac{3N}{4} = \frac{3 \times 118}{4} = 88.5 \; ; \; C_f = 76 \; ; L = 40 \; ; \; f = 24 \; ; \; i = 10$$

$$Q_3 = L + \frac{N_1 - C_f}{f} \times i = 40 + \frac{88.5 - 76}{24} \times 10$$

$$= 40 + \frac{125}{24} = 40 + 5.2 = \mathbf{45.2}$$

$$\text{I.Q.R.} = 45.2 - 22.5 = 22.7$$

$$\text{Q.D.} = \frac{45.2 - 22.5}{2}$$

$$= \frac{22.7}{2} = 11.35$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{45.2 - 22.5}{45.2 + 22.5} = \frac{22.7}{67.7} = .335.$$

Variance . Standard Deviation and tion.

Definition : The arithmetic mean of squares of deviations taken from arithmetic mean is called as **variance.**

Clearly, Variance $= \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ for individual observations

$= \dfrac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}$ for frequency distribution

Note : Symbolically we write variance of x as Var (x). The term variance is suggested by Prof. R. A. Fisher.

Remark : The units of original items and that of the variance are not same.

For example : If items are measured in cm., then the variance will be expressed in $(cm)^2$. Therefore we take positive square root of variance. It is called as standard deviation or least root mean square deviation.

Definition : The positive square root of mean of squares of the deviations taken from arithmetic mean is called as **standard deviation** (S.D.).

It is denoted by σ (read as sigma, a lower case Greek letter).

Therefore, $$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$ for individual observations

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}}$$ for frequency distribution

Merits and Demerits of Range

- Merits:
- Range is the simple to understand and easy to calculate.
- It should be rigidly defined.
- It requires less time to calculate.

- Demerits
- Range is not based on all the observations of data
- Range is very affected by fluctuations of sampling.
- Range cannot be computed for the distributions with open end classes.
- Range is not suitable for further mathematical treatments.

Merits and Demerits of Quartile deviation

- Merits:

- It is the simple to understand and easy to calculate.
- It can be computed for the distributions with open end classes.
- It is not affected by the presence of extreme values.

- Demerits
- It not based on all the observations of data
- It is affected by fluctuations of sampling.
- It is not suitable for further mathematical treatments.

Coefficient of variation

: Prof. Karl Pearson suggested the relative measure of standard deviation. It is called as coefficient of variation (C.V.)

It is given by, $C.V. = \frac{S.D}{|A.M.|} \times 100 = \frac{\sigma}{|\bar{x}|} \times 100\%$ ... (1)

Coefficient of variation is always expressed in percentage.

Remarks : 1. R.H.S. of (1) includes the multiplier 100, because $\frac{\sigma}{|\bar{x}|}$ is too small in many cases. Thus, for convenience it is multiplied by 100.

2. Frequently we need to compare dispersions of two or more groups. If the values in data set are large in magnitude, naturally variation among them will be proportionately larger.

For example : Standard deviation of weights of a group of elephants will be larger than that of a group of humanbeings. Suppose standard deviation of weights of a group of elephants is 15 kg and that of humanbeings is also 15 kg. In this case we cannot say, both the groups have identical variation. This is because average weight of a group of elephants is larger than that of the average weight of a group of persons. Therefore for comparing variations between two different data sets, a measure based on the ratio of σ and $\bar{x}$ would be appropriate. This is achieved in coefficient of variation. It measures variation in all data sets using a common yard stick; moreover it is free from units.

3. According to Prof. Karl Pearson coefficient of variation is the percentage variation in mean whereas S.D. gives the total variation in the mean.

Properties of Variance and Standard deviation

1. Effect of change of origin : Variance (standard deviation) is invariant to the change of origin. In other words, if a constant is added to (or subtracted from) each item, the variance (standard deviation) remains same.

2. Effect of change of origin and scale :

If $u = \frac{(x - a)}{h}$, a and h being constants, then var (U) = $\frac{1}{h^2}$ var (x) or

$\sigma_u = \frac{\sigma_x}{h}$.

3. Combined Variance and Standard Deviation :

Suppose there are two groups. First is of size n_1 with arithmetic mean $\bar{x}_1$ and variance σ_1^2. Second group is of size n_2 with arithmetic mean $\bar{x}_2$ and variance σ_2^2. Then the variance of combined group of size $n_1 + n_2$ is given by;

$$\sigma_c^2 = \frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}$$

where, $d_1 = \bar{x}_1 - \bar{x}_c$, $d_2 = \bar{x}_2 - \bar{x}_c$, and $\bar{x}_c$ is combined arithmetic mean.

Generalisation : Let there be k groups ($k \geq 2$) with size of i^{th} group as n_i, arithmetic mean $\bar{x}_i$ and variance σ_i^2, i = 1, 2, 3, ..., k. The combined variance of k groups is given by

$$\sigma_c^2 = \frac{\sum_{i=1}^{k} n_i\left(\sigma_i^2 + d_i^2\right)}{\sum_{i=1}^{k} n_i}$$

where, $d_i = \bar{x}_i - \bar{x}_c$, and $\bar{x}_c$ = Combined Arithmetic Mean.

Merits of standard deviation :

1. It is based on all observations.
2. It is rigidly defined.
3. It is capable of further mathematical treatment.
4. It does not ignore algebraic signs of deviations.
5. It is not much affected by sampling fluctuations.

Demerits of standard deviation :

1. It is difficult to understand and to calculate.
2. It cannot be computed for a distribution with open-end class.
3. It is unduly affected due to extreme deviations.
4. It cannot be calculated for qualitative data.

Mean Deviation - Individual Series

Example 8: The following are the monthly expenditure of six families. Calculate mean deviation from mean and mean deviation from median.

Expenditure (₹)	4,260	4,980	8,460	5,240	4,780	6,480

Solution: First we arrange the given data in ascending order:

Expenditure (₹)	4,260	4,780	4,980	5,240	6,480	8,460

Here, N = 6 which is even, so

$$\text{Median} = \frac{\left(\frac{N}{2}\right)^{th} \text{ value} + \left(\frac{N}{2}+1\right)^{th} \text{ value}}{2} = \frac{\left(\frac{6}{2}\right)^{th} \text{ value} + \left(\frac{6}{2}+1\right)^{th} \text{ value}}{2}$$

$$= \frac{3^{rd} \text{ value} + 3.5^{th} \text{ value}}{2} = \frac{4980 + 5240}{2} = 5110$$

$$\text{A.M.} = \frac{\sum X}{N} = \frac{1}{6} \times 34200 = 5700.$$

About Mean

Serial No.	X (₹)	Deviation from A.M. ignoring ± sign $\lvert d_x \rvert$
1	4260	1440
2	4780	920
3	4980	720
4	5240	460
5	6480	780
6	8460	2760
Total	**ΣX = 34200**	**$\Sigma \lvert d_x \rvert$ = 7080**

About Median

Serial No.	X (₹)	Deviation from Median ignoring ± sign $\lvert d_x \rvert$
1	4260	850
2	4780	330
3	4980	130
4	5240	130
5	6480	1370
6	8460	3350
Total	**ΣX = 34200**	**$\Sigma \lvert d_x \rvert$ = 6160**

$$\text{Mean deviation (about mean)} = \frac{\sum |d_x|}{N} = \frac{7080}{6} = 1180$$

$$\text{Mean deviation (about median)} = \frac{\sum |d_x|}{N} = \frac{6160}{6} = 1026.6$$

Mean Deviation - Discrete Series

N = total frequency

Example 9: Calculate the Mean-deviation from the following data:

Quantity Demanded (Units)	10	20	30	40	50	60	70	80	90	10
frequency	7	13	16	6	14	19	28	17	21	9

Solution:

Quantity Demanded (X)	frequency (f)	fX	$\|d_x\| = (X - \bar{X})$ ignoring ± sign	$f\|d_x\|$
10	7	70	50	350
20	13	260	40	520
30	16	480	30	480
40	6	240	20	120
50	14	700	10	140
60	19	1140	0	0
70	28	1960	10	280
80	17	1360	20	340
90	21	1890	30	630
100	9	900	40	360
Total	**$\Sigma f = 150$**	**$\Sigma fX = 9000$**		**3220**

$$\text{Arithmetic Mean } (\bar{X}) = \frac{\sum fX}{\sum f} = \frac{9000}{150} = 60$$

$$\text{Mean Deviation} = \frac{\Sigma f|d_x|}{\Sigma f} = \frac{3220}{150} = 21.5$$

Mean Deviation - Continuous Series

Example 10: Calculate Mean deviation from the mean for the following data:

Sales (₹ in thousand)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of days	5	6	8	15	17	7	9	3

Solution:

Calculation of Mean Deviation

Sales	Mid Value(X)	No. of days (f)	fX	$\lvert d_x \rvert = \lvert X - \bar{X} \rvert$	$f\lvert d_x \rvert$
0 – 10	5	5	25	35	175
10 – 20	15	6	90	25	150
20 – 30	25	8	200	15	120
30 – 40	35	15	525	5	75
40 – 50	45	17	765	5	85
50 – 60	55	7	385	15	105
60 – 70	65	9	585	25	225
70 – 80	75	3	225	35	105
		$\Sigma f = 70$	$\Sigma fX = 2800$		$\Sigma f\lvert d_x \rvert = 1040$

$$\text{Mean}(\bar{X}) = \frac{\sum fX}{\sum f} = \frac{2800}{70} = 40$$

$$\text{Mean Deviation (M.D.)} = \frac{\sum f\lvert d_x \rvert}{\sum f} = \frac{1040}{70} = 14.86$$

Example 11: Find S.D of (₹) 8, 10, 15, 24, 28.

Solution: Let the assumed mean (A) = 16.

Calculation from Arithmetic Mean

X	$x = (X - \bar{X})$	x^2
8	-9	81
10	-7	49
15	-2	4
24	7	49
28	11	121
$\Sigma X = 85$		$\Sigma x^2 = 304$

Calculation from Assumed Mean

X	$d_x = (X - A)$	d_x^2
8	-8	64
10	-6	36
15	-1	1
24	8	64
28	12	144
	$\Sigma dx = 5$	$\Sigma d_x^2 = 309$

From Arithmetic Mean

$$\bar{X} = \frac{\sum X}{N} = \frac{85}{5} = 17$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \frac{\sum (X - \bar{X})^2}{N} = \sqrt{\frac{1}{5} \times 304} = \sqrt{60.8} = 7.8$$

From Assumed Mean: Let A (assumed mean) = 16

$$\sigma = \sqrt{\frac{\sum d_x^2}{N} - \left(\frac{\sum d_x}{N}\right)^2} = \sqrt{\frac{309}{5} - \left(\frac{5}{5}\right)^2} = \sqrt{61.8 - (1)^2} = \sqrt{60.8} = 7.8$$

Note: If the actual mean is in fraction, then it is better to take deviations from an assumed mean for avoiding too much calculation.

Example 12: Calculate standard deviation from the following series:

Age (in years)	15	25	35	45	55	65
No. of Persons	7	25	20	16	11	6

Solution:

Calculation of Standard Deviations

Age (X)	No. of Persons (f)	fX	$d_x = (X - \bar{X})$	fd_x^2
15	7	105	– 22 (15-37)	3388
25	25	625	– 12	3600
35	20	700	– 2	80
45	16	720	8	1024
55	11	605	18	3564
65	6	390	28	4704
	$\Sigma f = 85$	$\Sigma fX = 3145$		$\Sigma fd_x^2 = 16360$

$\bar{x} = \frac{\Sigma f}{6} = \frac{85}{6} = 14.16$ $\quad \bar{x} = \frac{\Sigma f}{\Sigma fx} = \frac{3145}{85} = 37$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{3145}{85} = 37$$

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{\sum f(d_x)^2}{\sum f}} = \sqrt{\frac{16360}{85}} = \sqrt{192.5} = 13.87$$

Example 13: Find the Standard Deviation of the following series:

X	10	11	12	13	14	Total
f	4	16	22	14	6	62

Solution: Let the assumed mean (A) = 12.

Calculation of Standard Deviations

X	f	$d_x = (X - A)$	d_x^2	fd_x	fd_x^2
10	4	–2	4	– 8	16
11	16	–1	1	– 16	16
12	22	0	0	0	0
13	14	1	1	14	14
14	6	2	4	12	24
Total	$\Sigma f = 62$			$\Sigma fd_x = 2$	$\Sigma fd_x^2 = 70$

$4 \times (-2)^2 = 4 \times 4 = 16$

$16 \times (-1)^2 = 16 \times 1 = 16$

$$\sigma = \sqrt{\frac{\sum f(d_x)^2}{\sum f} - \left(\frac{\sum fd_x}{\sum f}\right)^2} = \sqrt{\frac{70}{62} - \left(\frac{2}{62}\right)^2}$$

$$= \sqrt{1.13 - 0.001} = 1.06$$

Example 15: Find the S.D. from the following figures:

Height (Inches)	44-46	46-48	48-50	50-52	52-54	Total
No. of Children	5	25	28	22	5	85

Solution: Let A (assumed mean) = 49

Height (Inches)	Mid-point (X)	No. of Children (f)	$d_x = X - 49$ $x - \bar{x}$	$d_x' = d/2$ $dx/2$	fd'_x	fd'^2_x
44-46	45	5	− 4	− 2	− 10	20
46-48	47	25	− 2	− 1	− 25	25
48-50	49	28	0	0	0	0
50-52	51	22	2	1	22	22
52-54	53	5	4	2	10	20
Total		$\Sigma f = 85$			$\Sigma fd_x' = -3$	$\Sigma fd_x'^2 = 87$

$$\sigma = \sqrt{\left\{\frac{\sum f(d_x')^2}{\sum f} - \left(\frac{\sum d_x'}{\sum f}\right)^2\right\}} \times i = \sqrt{\frac{87}{85} - \left(\frac{-3}{85}\right)^2} \times 2 = 1.01 \times 2 = 2.02$$

UNIT 4 : Correlation and Regression(for ungrouped data)

Karl Pearson formulated perhaps the greatest formula to find the degree of correlation.
He being a reputed, well known statistician, worked very hard on the theory of correlation. This formula was established in 1896.

Merits and Demerits of Pearson's method of studying correlation:

Merits:

1. This method indicates the presence or absence of correlation between two variables and gives the exact degree of their correlation.
2. In this method, we can also ascertain the direction of the correlation; positive, or negative.
3. This method has many algebraic properties for which the calculation of co-efficient of correlation, and other related factors, are made easy

Demerits:

1. It is more difficult to calculate than other methods of calculations.
2. It is much affected by the values of the extreme items.
3. It is based on a many assumptions, such as: linear relationship, cause and effect relationship etc. which may not always hold good.

Assumptions:

Karl Pearson based his formula on following basic assumptions:

(A) Two variables are affected by many independent causes and form a normal distribution.
(B) The cause and effect relationship exists between two variables.
(C) The relationship between two variables is linear. It is often denoted by r.

A. Direct Method

Type I : This method is used when given variables are small in magnitude.

Formula : $$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Example 1. Calculate Karl Pearson's coefficient of correlation between the age and weight of the children :

Age (years) :	1	2	3	4	5
Weight (kg.) :	3	4	6	7	12

Solution : $\Sigma X = 15$; $\Sigma Y = 32$; $\Sigma X^2 = 55$; $\Sigma Y^2 = 254$; $\Sigma XY = 117$

Age (X)	Weight (Y)	X^2	Y^2	XY
1	3	1	9	3
2	4	4	16	8
3	6	9	36	18
4	7	16	49	28
5	12	29	144	60
15	32	55	254	117

As $$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$\therefore\ r = \frac{5 \times 117 - 15 \times 32}{\sqrt{5 \times 55 - (15)^2}\ \sqrt{5 \times 254 - (32)^2}}$$

$$= \frac{585 - 480}{\sqrt{275 - 225}\ \sqrt{1270 - 1024}} = \frac{105}{\sqrt{50 \times 246}} = \frac{105}{\sqrt{12300}} = \frac{105}{110.90} = 0.9467 \text{ Ans.}$$

Type II : It is direct formula to find r. This formula can effectively be used where $\bar{X}$ and $\bar{Y}$ is not in fractions. The formula is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$$; where dx is the deviation of X variable from its $\bar{X}$.

y is the deviation of Y variable from its $\bar{Y}$. ; xy is the product of the two above

dx^2 is the square of x ; y^2 is the square of dy.

Example 2. Calculate coefficient of correlation between death and birth rate for the following data.

Birth Rate	24	26	32	33	35	30
Death Rate	15	20	22	24	27	24

Solution

Birth Rate X	Death Rate Y	$(X-\bar{X}) = x$	$(Y-\bar{Y}) = y$	$(X-\bar{X})^2 = x^2$	$(Y-\bar{Y})^2 = y^2$	$(X-\bar{X})(Y-\bar{Y}) = xy$
24	15	–6	–7	36	49	42
26	20	–4	–2	16	4	8
32	22	2	0	4	0	0
33	24	3	2	9	4	6
35	27	5	5	25	25	25
30	24.	0	2	0	4	0
$\Sigma X = 180$	$\Sigma Y = 132$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 90$	$\Sigma y^2 = 86$	$\Sigma xy = 81$
$\bar{X} = \frac{180}{6} = 30$	$\bar{Y} = \frac{132}{6} = 22$					

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{(81)}{\sqrt{90 \times 86}} = \frac{81}{\sqrt{7740}} = \frac{81}{87.98} = .92$$

There is no effect of any of the two types of change. We may add or subtract or multiply or divide each term by certain content, there is no effect on coefficient of correlation.

Example 4. Compute coefficient of correlation by Karl Pearson Method for the following data

X :	1800	1900	2000	2100	2200	2300	2400	2500	2600
f :	5	5	6	9	7	8	6	8	9

Solution

Let the A.M.s A_x and A_y be 2200 and 6 for X and Y series respectively

X	Y	dx	(i=100) dx	dy	dx^2	dy^2	dxdy
1800	5	–400	–4	–1	16	1	4
1900	5	–300	–3	–1	9	1	3
2000	6	–200	–2	0	4	0	0
2100	9	–100	–1	3	1	9	–3
2200	7	0	0	1	0	1	0
2300	8	100	1	2	1	4	2
2400	6	200	2	0	4	0	0
2500	8	300	3	2	9	4	6
2600	9	400	4	3	16	9	12
N = 9			$\Sigma dx = 0$	$\Sigma dy=9$	$\Sigma dx^2 = 60$	$\Sigma dy^2 = 29$	$\Sigma dx\,dy = 24$

$$r = \frac{(9)(24) - (0)(9)}{\sqrt{(9)(60) - (0)^2}\,\sqrt{(9)(29) - (9)^2}} = \frac{216}{\sqrt{97200}} = .69$$

(Note : We can also proceed dividing X by 100)