



MADURAI KAMARAJ UNIVERSITY



(University with Potential for Excellence)

DISTANCE EDUCATION

B.Sc. SECOND YEAR

MATHEMATICS

PART III

PAPER - IV

STATISTICS

UNITS : 1 to 5

Recognised by D.E.C.

www.mkudde.org

R 85

@All rights reserved

ACC-MEC
01179

STATISTICS SYLLABUS

- Unit 1** Measures of Central Tendencies - arithmetic mean - Partition values - Mode - Geometric mean and Harmonic mean - Measures of Dispersion - Range - quartile deviation - standard deviation - Mean deviation.
- Unit 2** Moments - skew ness - Kurtosis - Curve Fitting - Principle of least squares - Fitting straight line - a parabola - Fitting curves for 1. $Y = ab^x$ 2. $Y = ae^{bx}$ 3. $Y = bx^a$
- Unit 3** Correlation - Rank Correlation - Regression - Correlation coefficient for a bivariate frequency distribution.
- Unit 4** Interpolation - Finite differences - Newton's formula - lagrange's formula - theory of attributes - Consistency of Data - Independence and association of data.
- Unit 5** Index numbers - Average method - weighted index numbers - Consumer price index numbers - Analysis of time series.
- Unit 6** Probability - Conditional probability - Baye's theorem - Random Variables - Discrete random variables - Continuous random variables - Mathematical expectation - Moment generating function (Cumulant generating function and characteristic function - not included).
- Unit 7** Special distribution - Binomial, Poission and normal distribution Gamma - Chisquare - Student, Schedecor's - Fisher's distribution.
- Unit 8** Test of significance (large samples) sampling - sampling distribution - Testing of Hypothesis - Tests for proportion - mean - standard deviations - Correlation.
- Unit 9** Tests of significance (small samples) based on distribution - Test for significance based on F-test distribution - Tests for significance of an observed sample correlation.
- Unit 10** Test based on X^2 distribution - Test for population variance - To test the goodness of fit - Test for independence of attributes - Analysis of variance - one criterion of classification - Three criterion of classification - Latin squares.

Book for Reference

Statistics by S. Arumugam & Issac

SCHEME STATISTICS

Page No.

UNIT - 1

5

a) Measure of Central Tendencies

- 1) Arithmetic Mean
- 2) Median
- 3) Mode
- 4) Geometric Mean
- 5) Harmonic Mean

b) Measures of Dispersion

- 6) Range
- 7) Quartile Deviation
- 8) Mean Deviation
- 9) Standard Deviation

58

UNIT - 2

95

- 1) Moments
- 2) Skewness
- 3) Kurtosis
- 4) Curve Fitting

UNIT - 3

135

- 1) Correlation
- 2) Rank Correlation
- 3) Regression
- 4) Correlation Co-efficient for a bivariate frequency Distribution

UNIT - 4

173

- 1) Interpolation
- 2) Method of Finite Differences
- 3) Gregory - Newton Forward Interpolation Formulae
- 4) Lagrange's Interpolation Formulae
- 5) Theory of Attributes
- 6) Consistency of Data
- 7) Independence of Attributes
- 8) Co-efficient of Association

UNIT - 5

228

- 1) Index Numbers
- 2) Simple Index Numbers
- 3) Weighted Index Numbers
- 4) Cost of Living Index Numbers
- 5) Analysis of Time Series

UNIT - 6

1

- 1) Probability
- 2) Conditional Probability
- 3) Baye's Theorem
- 4) Random Variables
- 5) Mathematical Expectation
- 6) Moment Generating Function

UNIT - 7

34

Special Distributions

- 1) Binomial Distribution
- 2) Poission Distribution
- 3) Normal Distribution
- 4) Gamma Distribution
- 5) Chi-square Distribution
- 6) Student's t-distribution
- 7) Snedecor's Distribution
- 8) Fischer's z - distribution

UNIT - 8

70

Tests of significance (large samples)

- 1) Sampling
- 2) Sampling Distribution
- 3) Testing of Hypothesis
- 4) Test of significance for proporation or percentage
- 5) Test of significance for means
- 6) Test of significance for Standard Deviation
- 7) Test of significance for correlation co-efficient

UNIT - 9

95

Test of significance (small samples)

- 1) Test of significance based on mean
- 2) Test of significance based on F-test
- 3) Test for significance of an observed sample correlation

UNIT - 10

112

Test based on χ^2 - Distribution

- 1) χ^2 - test for population variance
- 2) χ^2 - test to test the goodness of fit
- 3) χ^2 - test to test the independence of attributes
- 4) Analysis of variance
- 5) Three criteria of classification - Latin square

Unit - 1

Measures of Central Tendency

The word average is very commonly used in day to day conversation. For example, we often talk of average boy in a class, average height of a girl, etc. When we say 'he is an average student' he is neither very good nor very bad, just a mediocre type of student. However, in statistics the term average has a different meaning.

Definition

One of the most important objectives of statistical analysis is to get one single value that describes the characteristic of the entire mass of unwidely data. Such a value is called the 'measures of central tendency' or an average or the expected value of the variable.

The author Clark said, "Average is an attempt to find one single figure to describe whole of figures". Since an average represents the entire data, its value lies somewhere in between the two extremes, ie., the largest and the smallest items.

There are two main objectives of the average

- 1) To get single value that describes the characteristic of the entire group.
- 2) To facillitate comparison.

Characteristic of an ideal average

- 1) It should be easy to understand.
- 2) It should be simple to compute.
- 3) It should be based on all the items.
- 4) It should not be unduly affected by extreme observations.
- 5) It should be rigidly defined.
- 6) It should be capable of further algebraic treatment.
- 7) It should be stable.

Types of averages

The following are the important types of averages

- 1) Arithmetic mean : (i) simple (ii) weighted
- 2) Median
- 3) Mode
- 4) Geometric mean
- 5) Harmonic mean

1.1 : Arithmetic mean

The most popular and widely used measure of representing the entire data by one value is what most laymen call an 'average' and what the statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by dividing this total by the number of items.

Calculation of simple arithmetic mean - Individual observations

The process of computing mean in case of individual observations is very simple. Add together the various values of the variable and divide the total by the number of items. In symbols,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{N} = \frac{\sum x}{N}$$

Here $\bar{x}$ = Arithmetic mean

$\sum x$ = sum of all the values of the variable x

ie $x_1, x_2, \dots, x_n$

N = Number of observations

Example 1 : The following values give the monthly income of 10 employees in an office :

Income (Rs)	1780	1760	1690	1750	1840	1920	1100
	1810	1050	1950				

Calculate the arithmetic mean of incomes

Solution : Let the income be denoted by the symbol x .

$$\bar{x} = \frac{\sum x}{N}$$

$$N = 10; \quad \sum x = \frac{x_1 + x_2 + \dots + x_{10}}{10}$$

$$\sum x = 16,650.$$

$$\therefore \bar{x} = \frac{\sum x}{N} = \frac{16650}{10} = 1665$$

So, the average income is Rs. 1665.

Short-cut method

- Steps. (1) Take an assumed mean ie A.
- (2) Take the deviations of items from the assumed mean and denote these deviations by d.
- (3) Obtain the sum of these deviations ie Σd
- (4) Apply the formula :
- $$\bar{x} = A + \frac{\Sigma d}{N}$$

Example 2 : From example 1 calculate (A.M.) arithmetic mean by taking the assumed mean.

Solution : Let us take 1800 as the assumed mean

Employee	Income (Rs.)	X - 1800 = d
1	1780	- 20
2	1760	- 40
3	1690	- 110
4	1750	- 50
5	1840	+ 40
6	1920	+ 120
7	1100	- 700
8	1810	10
9	1050	- 750
10	1950	+ 150
N = 10		$\Sigma d = -1350$

$$\bar{x} = A + \frac{\Sigma d}{N}$$

$$A = 1800, \quad \Sigma d = -1350, \quad N = 10$$

$$\bar{x} = 1800 - \frac{1350}{10} = 1800 - 135 = 1665$$

$\therefore$ The average income = Rs. 1665

Calculation of Arithmetic Mean - Discrete Series

In discrete series arithmetic mean may be computed by applying (i) Direct method or (ii) short cut method.

Direct Method : The formula for computing mean is $\bar{x} = \frac{\sum fx}{N}$ where f = Frequency

x = The variable. N = Total number of observations i.e. $\sum f$.

Steps : (i) Multiply the frequency of each row with the variable and obtain the total $\sum fx$

(ii) Divide the total obtained by step (i) by the number of observations ie total frequency.

Example 3 : From the following data of the marks obtained by 60 students of a class calculate the arithmetic mean.

Marks	No. of students
20	8
30	12
40	20
50	10
60	6
70	4

Solution : Let the marks be denoted by x and the number of students by f.

Marks x	No of students f	fx
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280

$$\sum N = 60 \quad \sum fx = 2460$$

$$\bar{x} = \frac{\sum fx}{N} = \frac{2460}{60} = 41$$

$\therefore$ The average mark is 41.

Short cut method

Steps : (i) Take an assumed mean (A)

(ii) Take the derivations of the variable x from the assumed mean and denote the deviations by d.

(iii) Multiply these deviations with the respective frequency and take the total fd.

(iv) Divide the total obtained in third step by the total frequency.

(v) Apply the formula :

$$\bar{x} = A + \frac{\sum fd}{N}$$

Example 4 : Using example 3 calculate arithmetic mean by the short-cut method.

Solution : Let us take 40 as the assumed mean

Marks x	No. of students f	(x - 40) d	fd
20	8	-20	-160
30	12	-10	-120
40	20	0	0
50	10	+10	+100
60	6	+20	+120
70	4	+30	+120
N = 60			$\sum fd = 60$

$$\bar{x} = \frac{A + \sum fd}{N}$$

$$= 40 + \frac{60}{60} = 40 + 1 = 41$$

The average mark is 41

Calculation of Arithmetic Mean - Continuous Series

In continuous series, arithmetic mean may be computed by applying the following methods. (i) Direct method or (ii) Short cut method.

Direct method

Steps : (i) Obtain the mid-point of each class and denote it by m.

(ii) Multiply these mid-points by the respective frequency of each class and obtain the total $\sum fm$.

(iii) Divide the total obtained in step (ii) by the sum of the frequency ie., N

So the formula is $\bar{x} = \frac{\sum fm}{N}$

Example 5 : From the following data compute arithmetic mean by direct method.

Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
No. of students	5	10	25	30	20	10

Solution :

Marks	Midpoint m	No. of students f	fm
0-10	5	5	25
10-20	15	10	150
20-30	25	25	625
30-40	35	30	1050
40-50	45	20	900
50-60	55	10	550

$$N = 100$$

$$\sum fm = 3,300$$

$$\bar{x} = \frac{\sum fm}{N} = \frac{3300}{100} = 33.$$

Short-cut Method : When the class intervals of the series are of uniform length;

Steps : (i) Take an assumed mean as A.

(ii) From the mid-point of each class deduct the assumed mean and divide by the width of the class interval. Call it as d. ie. Suppose m is the mid point of particular class and A is the assumed mean and i is the width of the class interval then

$$d = \frac{m - A}{i}$$

(iii) Multiply the respective frequencies of each class by these deviations and obtain the total $\sum fd$

(iv) Apply the formula : $\bar{x} = A + \frac{\sum fd}{N} \times i$

Example 6 : From the data of Example 4 compute arithmetic mean.

Solution : Take the assumed mean as 35.

Marks	Mid point m	No. of students f	m-A	(m-A)/10 (d)	fd
0-10	5	5	-30	-3	-15
10-20	15	10	-20	-2	-20
20-30	25	25	-10	-1	-25
30-40	35	30	0	0	0
40-50	45	20	10	1	20
50-60	55	10	20	2	20

$$N = 100$$

$$\Sigma fd = -20$$

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{N} \times i = 35 - \frac{20}{100} \times 10 \\ &= 35 - 2 = 33.\end{aligned}$$

Now we will see an example when the class intervals of the series are not of uniform length.

Example 7 : Calculate arithmetic mean from the following data.

Marks	0-10	10-30	30-60	60-100
No. of students	5	12	25	8

Solution : Since the class intervals are not of uniform length we can take 5 as the common factor and let us take 45 as the assumed mean.

Marks	Mid point m	f	(m-45)/5 (d)	fd
0-10	5	5	-8	-40
10-30	20	12	-5	-60
30-60	45	25	0	0
60-100	80	8	7	56

$$N = 50$$

$$\Sigma fd = -44$$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times C$$

$$= 45 - \frac{44}{50} \times 5 = 45 - 4.4 = 40.6$$

$$\therefore \bar{x} = 40.6$$

Frequency distribution with open intervals at the beginning or the end

Suppose a distribution has open interval at the beginning. Then we cannot calculate the midvalue of the first class nor the length of the class. So we assume that the length of the first class is equal to the length of the next class.

Similarly, if the last class is an open class, we take its length to be equal to the length of the immediately preceding class and thus find the upper boundary of the last class.

Example 8 : Calculate the mean of the following distribution.

Class interval	below 10	10-20	20-30	30-40	40-50
frequency	8	9	7	5	1

Solution :

Length of the second class = 10.

∴ Length of the first class = 10.

∴ True lower limit of the first class = 10 - 10 = 0.

Let us take 25 as the assumed mean

mid point m	f	d = (m-A)/i	fd
5	8	-2	-16
15	9	-1	-9
25	7	0	0
35	5	1	5
45	1	2	2

$$N = 30$$

$$\Sigma fd = -18$$

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

$$= 25 + \frac{-18}{30} \times 10$$

$$= 25 - 6 = 19$$

$$x = 19.$$

Example 9 : Calculate the mean of the following distribution.

Wages in Rs.	below 30	30-50	50-80	80-90	above 90
No. of workers	5	47	62	54	15

Solution : The width of the second class = $50 - 30 = 20$

$\therefore$ The width of the first class = 20.

The width of the 4th class = $90 - 80 = 10$.

$\therefore$ The width of the last class = 10.

The class intervals are not of uniform length.

Let us take 10 as the common factor and 65 as the assumed mean.

Class Interval	Mid point m	f	(m-A)/C d	fd
10-30	20	5	-4.5	-22.5
30-50	40	47	-2.5	-117.5
50-80	65	62	0	0
80-90	85	54	2	108
90-100	95	15	3	45

$N = 183$

$\Sigma fd = 13$

$$\bar{x} = A + \frac{\Sigma fd}{N} \times C$$

$$= 65 + \frac{13}{183} \times 10$$

$$= 65 + \frac{130}{183}$$

$$= 65 + 0.71$$

$$\bar{x} = 65.71$$

Correcting incorrect values

It sometimes happens that due to an oversight or mistake in copying, certain wrong items are taken while calculating mean. To find out the correct mean, we have to do a simple process. From incorrect Σx deduct wrong items and add correct items and then divide the correct Σx by the number of observations. The result so obtained will give the value of correct mean.

Example 10 : Mean of 100 observations is found to be 40. If at the time of computation two items are wrongly taken as 30 and 27 instead of 3 and 72. Find correct mean

Solution : $\bar{x} = \frac{\sum x}{N}$

$$\therefore \sum x = N\bar{x}$$

$$\bar{x} = 40, N = 100$$

$$\therefore \sum x = 100 \times 40 = 4000$$

$$\text{correct } \sum x = \text{Incorrect } \sum x - \text{wrong items} + \text{correct items}$$

$$= 4000 - (30 + 27) + 3 + 72$$

$$= 4000 - 57 + 75$$

$$= 4018$$

$$\text{Correct mean} = \frac{\sum x}{N} = \frac{4018}{100} = 40.18.$$

Example 11 : In the following distribution, the frequency of the fifth class is not known and the mean of the distribution is 33. Find the unknown frequency.

Class intervals	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	5	10	25	30	?	10

Solution : Let the frequency of the fifth class be f . Let us compute the mean by the usual method and equate it to 33 we get an equation in f . Solving this, we get the value of f . Let us take 35 as the assumed mean.

$$\text{Here } i = 10$$

class intervals	Mid point	f	$d = (m-A)/i$	fd
0-10	5	5	-3	-15
10-20	15	10	-2	-20
20-30	25	25	-1	-25
30-40	35	30	0	0
40-50	45	f	1	f
50-60	55	10	2	20
		$\sum f = 80 + f$	$\sum fd = f - 40$	

$$\bar{x} = A + \frac{\sum fd}{N} \times i$$

$$33 = 35 + \frac{f - 40}{80 + f} \times 10$$

$$33 - 35 = \frac{10(f-40)}{80 + f}$$

$$-2(80 + f) = 10(f-40)$$

$$80 + f = -5(f-40) = -5f + 200$$

$$f + 5f = 200 - 80 = 120$$

$$6f = 120$$

$$f = \frac{120}{6} = 20.$$

∴ Frequency of the fifth class = 20.

Given the means of two or more distributions to find the mean of the combined distribution :

We can find the mean of the combined set when the means of more than two sets are given. Suppose $\bar{x}_1$ is the mean of N_1 items, $\bar{x}_2$ is the mean of N_2 items, ..., $\bar{x}_k$ is the mean of N_k items. Then the mean $\bar{x}$ of the combined distribution is given by

$$\bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2 + \dots + N_k \bar{x}_k}{N_1 + N_2 + \dots + N_k}$$

Example 12 : The mean height of 25 male workers in a factory is 61 cms and the mean height of 35 female workers in the same factory is 58 cms. Find the combined mean height of 60 workers in the factory.

$$\text{Solution : } \bar{x} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$\text{Here } N_1 = 25, \bar{x}_1 = 61, N_2 = 35, \bar{x}_2 = 58$$

$$\therefore \bar{x} = \frac{25 \times 61 + 35 \times 58}{25 + 35}$$

$$= \frac{1525 + 2030}{60} = \frac{3555}{60} = 59.25$$

Thus the combined mean height of 60 workers is 59.25 cm.

Exercise

1. Compute the A.M. of the following data

Age in years	20	21	22	23	24	25
No. of students	5	9	12	7	6	1

2. Calculate the A.M. of the following data

Income in Rs.	100-150	150-200	200-250	250-300	300-350
No. of employees	6	8	13	11	7

3. From the data relating to two factories A and B, find the average of both factories put together.

	Factory A	Factory B
No. of workers	100	150
Mean Wages in Rs.	45	55

4. Find out the missing frequency of the following frequency distribution whose mean is 41.

Mid value	20	30	40	50	60	70
Frequency	8	?	20	10	6	4

5. Calculate the mean of the following distribution.

Size of the items	below 4	4-5	5-6	6-7	7-8	8-9	9-10	above 10
Frequency	3	8	28		66	2	6	3

(Hint : Since the frequency for 6-7 is not given take it as 0.)

Weighted Arithmetic Mean

There are cases where the relative importance of the different items is not the same. At that time, we compute weighted arithmetic mean. The term 'weight' stands for the relative importance of the different items. The formula for computing weighted arithmetic mean is $\bar{X}_w = \frac{\sum wx}{\sum w}$ where $\bar{X}_w$ represents the weighted arithmetic mean, x represents the variable values ie $x_1, x_2, \dots, x_n$.

In case of frequency distribution, ie $f_1, f_2, \dots, f_n$ are the frequencies of the variable values $x_1, x_2, \dots, x_n$ respectively then the weighted arithmetic mean is given by :

$$\bar{X}_w = \frac{\sum w (fx)}{\sum w}$$

The weighted arithmetic mean is specially useful in problems relating to :

- (i) Construction of index numbers.
- (ii) Standardized birth and death rates.

Example : A contractor employs three types of workers - male, female and children. To a male worker he pays Rs.40 per day, to a female worker Rs.32 per day and to a child worker Rs.15 per day. What is the average wage per day paid by the contractor?

Solution : The number of male, female and child workers need not be the same. So we cannot use the simple method. We have to calculate the weighted arithmetic mean. Let us assume that the number of male, female and child workers employed is 20, 15 and 5 respectively.

Wages per day (Rs)	No. of workers	
x	w	w x
40	20	800
32	15	480
15	5	75

$\Sigma w = 40$

$\Sigma wx = 1355$

$$\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{1355}{40} = 33.875$$

1.2 : Median

Next to mean, median has a wide usage in statistical works. Median is 'the middle most' or 'the central value' of the items. It signifies the location of the central values and thus a positional average.

Median is the value of the middle item of a given set of data arranged in ascending or in descending order of magnitude. ie., it divides the series of given data into two parts having the following properties.

- 1) The number of items in each part will be equal.
- 2) The values of items in one part will be greater than the value of the median and the values of items in the other part will be less than the value of the median.

For example if the income of five employees is Rs.900, 950, 1020, 1200 and 1280, then the median would be 1020. ie the value at middle position of the array.

In this example the calculation of median was simple because of odd number of observations. When an even number of observations are listed, the median is taken to be the arithmetic mean of two middlemost items. For example, if we are given the income of six employers as 900, 950, 1020, 1200, 1280 and 1300, then there are two middle position values. They are 1020 and 1200.

$$\text{So median} = \frac{1020 + 1200}{2} = \frac{2220}{2} = 1110.$$

Thus when N is odd, the median is an actual value of $\left(\frac{(N+1)}{2}\right)^{\text{th}}$ item. If N is even, the median is half the sum of the middle values.

Calculation of Median - Individual Observations

Steps : (i) Arrange the data in ascending or descending order of magnitude. (Usually we will arrange in ascending order)

(ii) If N is odd, median = $\frac{(N+1)}{2}^{\text{th}}$ item.

(iii) If N is even, median = half the sum of the middle values.

Example 1 : From the following data of the wages of 7 workers compute the median wage :

Wages (Rs)	1100	1150	1080	1120	1200	1160	1400
------------	------	------	------	------	------	------	------

Solution :

No.

Wages arranged in ascending order

1	1080
2	1100
3	1120
4	1150
5	1160
6	1200
7	1400

$$\text{Median} = \frac{N + 1^{\text{th}}}{2} \text{ item} = \frac{7 + 1}{2} = 4^{\text{th}} \text{ item.}$$

Size of 4th item = 1150

Hence the median wage = Rs. 1150.

Example 2 : Obtain the value of median from the following data :

391, 384, 591, 407, 672, 522, 777, 753, 2488, 1490.

Solution :

Sl. No.	Data arranged in ascending order
1	384
2	391
3	405
4	522
5	591
6	672
7	753
8	777
9	1490
10	2488

$$\text{Median} = \text{size of } \frac{N + 1}{2} \text{ item} = \frac{10 + 1}{2} = \frac{11}{2} = 5.5^{\text{th}} \text{ item}$$

$$\text{Size of } 5.5^{\text{th}} \text{ item} = \frac{5^{\text{th}} \text{ item} + 6^{\text{th}} \text{ item}}{2}$$

$$= \frac{591 + 672}{2} = \frac{1263}{2} = 631.5$$

Calculation of Median - Discrete Series

- Steps :**
- Arrange the data in ascending order of magnitude.
 - Find out the cumulative frequencies.
 - Apply the formula : Median = size of $\frac{N+1}{2}$ th item
 - Look at the cumulative frequency column and find that the total which is either equal to $\frac{(N+1)}{2}$ or next higher to that and determine the value of variable corresponding to it. That gives the value of median.

Example 3 : From the following data find the value of median :

Income (Rs)	1000	1500	800	2000	2500	1800
No. of persons	24	26	16	20	6	30

Solution :

Income arranged in ascending order	No. of persons	cumulative frequency
800	16	16
1000	24	40
1500	26	66
1800	30	96
2000	20	116
2500	6	122

$$\text{Median} = \text{size of } \frac{N + 1^{\text{th}}}{2} \text{ item}$$

$$= \frac{122 + 1}{2} = 61.5^{\text{th}} \text{ item}$$

Size of 61.5th item = 1500.

Calculation of Median - Continuous Series

Steps : Determine the particular class in which the value of median lies. Use $\frac{N}{2}$ as the rank of the median not $(N+1)/2$. In a continuous series we have to take $\frac{N}{2}$ not $(N+1)/2$. After ascertaining the class in which median lies, the following formula is used for determining the exact value of median.

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

L = Lower limit of the median class ie, the class in which the middle item of the distribution lies.

c.f. = cumulative frequency of the class preceding the median class or sum of the frequencies of all classes lower than the median class.

f = frequency of the median class.

i = class interval of the median class.

Example 4 : Calculate the median for the following frequency distribution :

Marks	No. of students
45-50	10
40-45	15
35-40	26
30-35	30
25-30	42
20-25	31
15-20	24
10-15	15
5-10	7

Solution : First arrange the data in ascending order and then find out median.

Marks	f	c.f
5-10	7	7
10-15	15	22
15-20	24	46
20-25	31	77
25-30	42	119
30-35	30	149
35-40	26	175
40-45	15	190
45-50	10	200

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item}$$

$$= \frac{200}{2} = 100^{\text{th}} \text{ item}$$

Median lies in the class 25-30

$$\text{Median} = L + \frac{N/2 - C.f.}{f} \times i$$

$$L = 25, \frac{N}{2} = 100, c.f = 77, f = 42, i = 5.$$

$$\therefore \text{Median} = 25 + \frac{100 - 77}{42} \times 5$$

$$= 25 + \frac{23}{42} \times 5 = 25 + 2.74 = 27.74.$$

Example 5 : Calculate the median from the following data :

Weight in gms	No. of Apples
410-419	14
420-429	20
430-439	42
440-449	54
450-459	45
460-469	18
470-479	7

Solution : Since the given series is not continuous we have to make it as a continuous one. So we deduct 0.5 from the lower limits and add 0.5 to the upper limits.

Weight in gms	f	c.f.
409.5-419.5	14	14
419.5-429.5	20	34
429.5-439.5	42	76
439.5-449.5	54	130
449.5-459.5	45	175
459.5-469.5	18	193
469.5-479.5	7	200

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item} = \frac{200}{2} = 100^{\text{th}} \text{ item}$$

Median lies in the class 439.5-449.5

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

$$L = 439.5, N/2 = 100, c.f = 76, f = 54, i = 10$$

$$\therefore \text{Median} = 439.5 + \frac{100 - 76}{54} \times 10$$

$$= 439.5 + \frac{24}{54} \times 10$$

$$= 439.5 + 4.44 = 443.94$$

$$\text{Median} = 443.94$$

Example 6 : Calculate median from the following data :

Value	Frequency
Less than 10	4
Less than 20	16
Less than 30	40
Less than 40	76
Less than 50	96
Less than 60	112
Less than 70	120
Less than 80	125

Solution : Since cumulative frequencies are given, first find frequencies.

Value	Frequency	c.f.
0-10	4	4
10-20	12	16
20-30	24	40
30-40	36	76
40-50	20	96
50-60	16	112
60-70	8	120
70-80	5	125

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item}$$

$$= \text{Size of } \frac{125}{2} = 62.5^{\text{th}} \text{ item}$$

Hence median lie in the class 30 - 40.

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

$$L = 30, N/2 = 62.5, c.f = 40, f = 36, i = 10.$$

$$\text{Median} = 30 + \frac{62.5 - 40}{36} \times 10$$

$$= 30 + 6.25 = 36.25$$

$$\text{Median} = 36.25$$

Example 7 : Compute median from the following data :

Mid value	Frequency
115	6
125	25
135	48
145	72
155	116
165	60
175	38
185	22
195	3.

Solution : Since we are given the mid-values, we should find out the upper and lower limits of the various classes.

Class	f	cf
110-120	6	6
120-130	25	31
130-140	48	79
140-150	72	151
150-160	116	267
160-170	60	327
170-180	38	365
180-190	22	387
190-200	3	390

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item.}$$

$$= \text{Size of } \frac{390}{2} = 195^{\text{th}} \text{ item.}$$

Median lies in the class 150-160.

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

$$L = 150, N/2 = 195, \text{c.f.} = 151, f = 116, i = 10.$$

$$\text{Median} = 150 + \frac{195 - 151}{116} \times 10$$

$$= 150 + 3.79 = 153.79.$$

$$\text{Median} = 153.79$$

Example 8 : Find the missing frequency in the following distribution if N is 100 and median is 30.

Marks	No. of students
0-10	10
10-20	?
20-30	25
30-40	30
40-50	?
50-60	10

Solution : Let the frequencies of the class 10-20 be x.

$$\text{Frequency of the class 40-50} = 100 - (10 + x + 25 + 30 + 10)$$

$$= 100 - (75 + x)$$

$$= 25 - x.$$

Marks	No. of students	c.f.
0-10	10	10
10-20	x	10 + x
20-30	25	35 + x
30-40	30	65 + x
40-50	25 - x	90
50-60	10	100

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

Since median is 30, it lies in the class 30 - 40. ($\because$ The class 20 - 30 would cover up to 29.999 and 30 would be covered in the class 30-40).

$$L = 30, \frac{N}{2} = 50, \text{c.f.} = 35 + x, f = 30.$$

$$30 = 30 + \frac{50 - (35 + x)}{30} \times 10$$

$$0 = \frac{15 - x}{3} \quad \therefore 15 - x = 0$$

$$\text{or } x = 15.$$

So frequency of class 10 - 20 = 15

and frequency of the class 40 - 50 = 25 - x

$$= 25 - 15 = 10.$$

Calculation of Median when class intervals are unequal

When the class intervals are unequal, the same formula can be applied. Since in the calculation of the median, we need the length of the median class only, we need not bother about the length.

Example 9 : Calculate median from the following data

Marks	0-10	10-30	30-60	60-80	80 -90
No. of students	5	15	30	8	2

Solution : Here the class intervals are unequal.

Marks	f	c.f.
0-10	5	5
10-30	15	20
30-60	30	50
60-80	8	58
80-90	2	60

$$\text{Median} = \text{size of } \frac{N}{2} \text{ th item} = \frac{60}{2} = 30^{\text{th}} \text{ item.}$$

So median lies in the class 30-60. $\therefore i = 30.$

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

$$L = 30, N/2 = 30, \text{c.f.} = 20, f = 30, i = 30.$$

$$\text{Median} = 30 + \frac{30 - 20}{30} \times 30 = 40.$$

$$\text{Median} = 40.$$

Graphic Method of locating the median for a continuous series

Among the frequency graphs, ogives give us the cumulative frequencies. Using these, we can find the median. We found the item corresponding to the less than cumulative frequency $\left(\frac{N}{2}\right)$. So the x-coordinate corresponding to the cumulative frequency $\left(\frac{N}{2}\right)$ in the ogive (less than or more than) is the median. Since the point whose x-coordinate is the median and y-coordinate is $\left(\frac{N}{2}\right)$ lies on both the ogives, the point of intersection of these two curves will also give the median. So we have two methods to calculate the median graphically.

Method - 1 : Draw 'the less than' or 'the more than' ogive. Draw a horizontal line through the point on the y-axis showing $\left(\frac{N}{2}\right)$. Let it meet the ogive at some point. To find its x-coordinate, draw the vertical through this point. Let it meet the x-axis at M. Find the value OM. That will be the median.

Method - 2 : Draw the two ogives. Find the x-coordinate of the point of intersection. That will be the median.

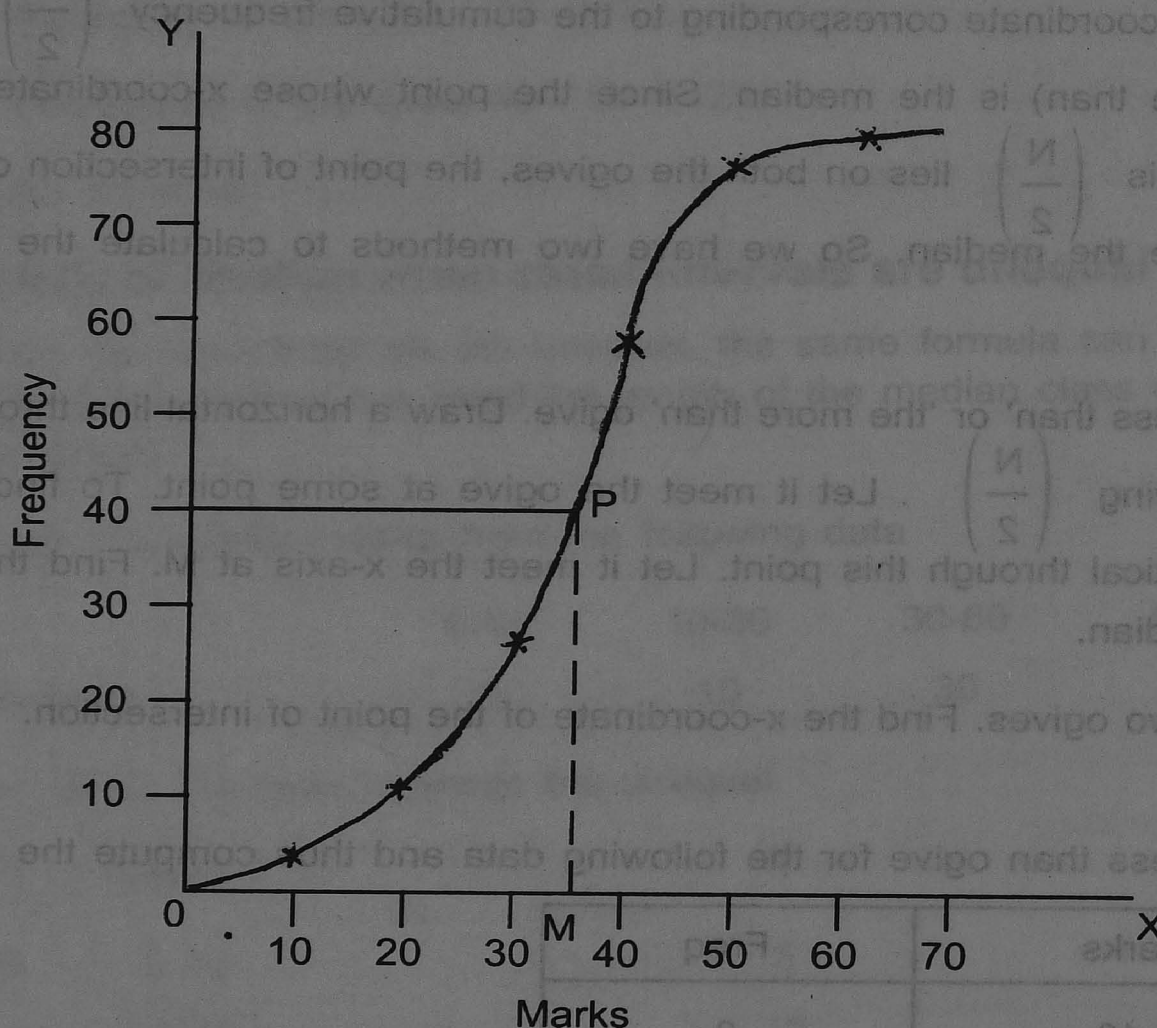
Example 1 : Draw the less than ogive for the following data and thus compute the median.

Marks	Freq
0-10	3
10-20	9
20-30	15
30-40	30
40-50	18
50-60	5

Step - 1 : Form 'the less than' cumulative frequency table

upper bound	Less than cum freq
10	3
20	12
30	27
40	57
50	75
60	80

Step - 2 : Draw the less than ogive



Step - 3 : Draw the horizontal line through 40

$$\left(\frac{N}{2} = \frac{80}{2} = 40 \right) \text{ on the y axis.}$$

Let it meet the curve at P. Draw a vertical through P. Let it meet the X axis at M.

The value corresponding to M = 34.

$\therefore$ Median = 34.

Example - 2 : Draw the two ogives for the following data and thus calculate the median.

Class	0-10	10-20	20-30	30-40	40-50
Freq	5	8	8	15	9

Step - 1 : We need both 'the less than cumulative frequency table' and the 'more than cumulative frequency table' to draw the two ogives. Let us first frame them.

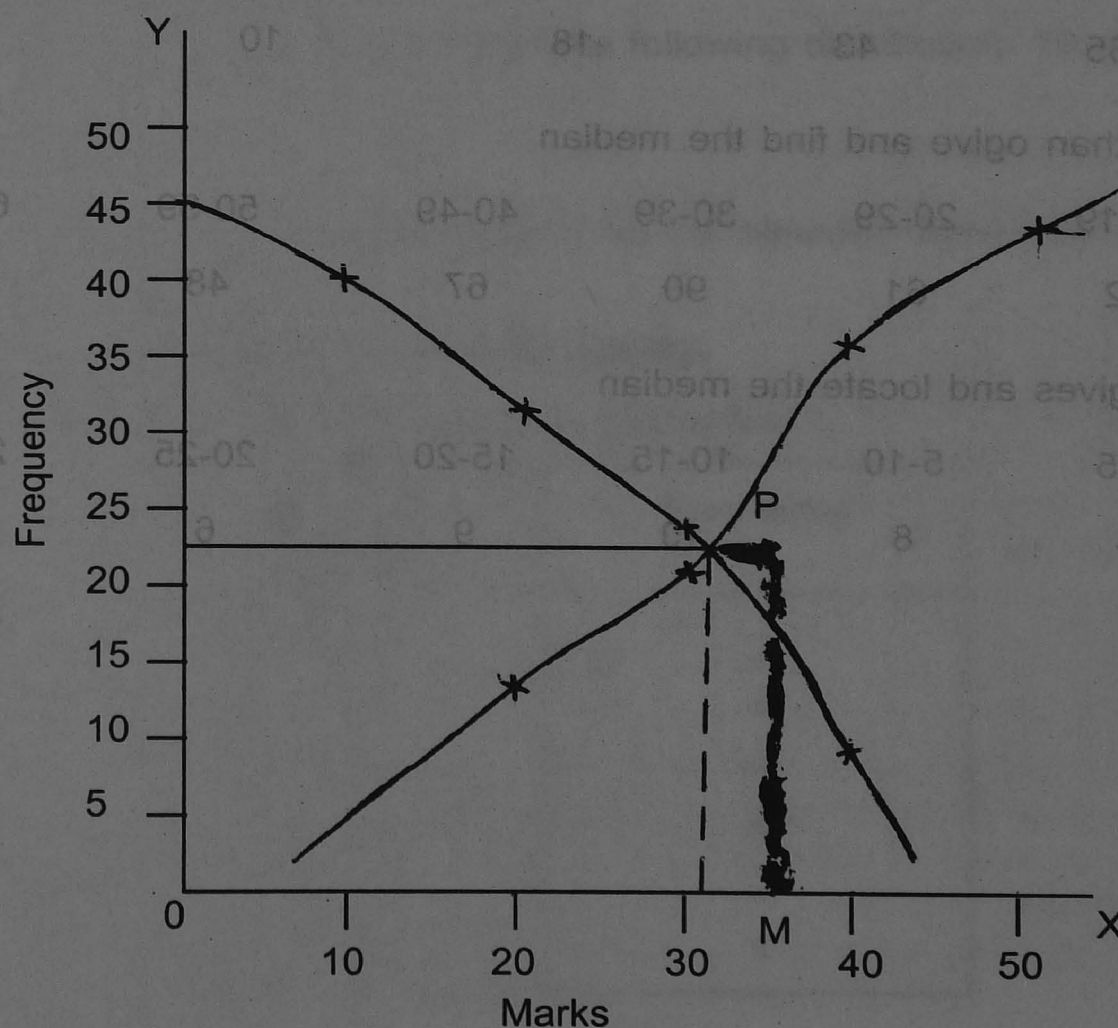
1) Less than cumulative frequency table

upper bound	Less than cum freq
10	5
20	13
30	21
40	36
50	45

2) More than cumulative frequency table :

Lower bound	More than cum. freq
0	45
10	40
20	32
30	24
40	9

Step - 2: Draw the two ogives



Exercise

1. Compute the median for the following ungrouped data.

57, 48, 49, 52, 45, 68, 71, 70, 49, 50, 62.

2. Compute the median for the following discrete series.

Marks	5	10	15	20	25	30	35	40	45	50
Freq	20	43	67	45	35	19	11	8	7	50

3. Compute the median for the following distribution :

(a) Height in cms	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Freq	15	46	75	53	40	18	3

(b) Mid value of the class	15	25	35	45	55	65	75
Freq	10	25	30	42	52	26	12

4. Compute the median graphically

- (a) Draw the less than ogive and find the median

Class	50-100	100-150	150-200	200-250	250-300
Freq	35	43	18	10	4

- (b) Draw the more than ogive and find the median

Class	10-19	20-29	30-39	40-49	50-59	60-69
Freq	52	61	90	67	48	40

- (c) Draw the two ogives and locate the median

Class	0-5	5-10	10-15	15-20	20-25	25-30
Freq	4	8	10	9	6	3

1.3 : Mode

The mode or the modal value is that value in a series of observations which occurs with the greatest frequency. For example, the mode of the series 3, 5, 8, 5, 4, 5, 9, 3 would be 5, since this value occurs more frequently than any of the others.

The mode is often said to be that value which occurs most often in the data, that is with the highest frequency. A set of data may have a single mode in which case it is said to be unimodal, it may have two modes which makes it bimodal or it may have several modes and be called multimodal.

Calculation of Mode

Let us first consider the ungrouped data. For example, consider the numbers.

1, 2, 7, 10, 11, 11, 12, 12, 12, 12, 13, 14, 17, 17, 17, 19, 23.

Here 12 occurs four times, 17 occurs three, 11 twice and other numbers only once. So 12 is taken as the mode.

So, given a set of numbers, we count the number of times each occurs i.e., we find the frequency of each and take that number having maximum frequency as the mode. If the total number of items is not large, it will be easy to count the frequencies, if we arrange the numbers in ascending (or descending) order of magnitudes. Otherwise we can form the frequency distribution table and find the mode.

Example 1 : Determine the mode of the following distribution. 19, 10, 10, 11, 12, 14, 14, 17, 19, 19, 21.

Mode = 19.

If 14 also occurred three times, it will be 'bimodal' distribution with modal values 14 and 19.

Calculation of Mode - Discrete series

Consider the following frequency distribution :

Size of the item	Frequency
10	3
12	4
14	12
16	5
18	2
20	1

Maximum frequency = 12.

Value of the corresponding item = 14.

∴ Mode = 14.

In the above distribution the frequencies of the item immediately above and below the mode are 4 and 5 respectively and they are much small compared to the maximum frequency 12. But in some distributions they may be closer to the maximum frequency i.e., the differences between the maximum frequency and the frequencies of the items just above and below it may be small. Then we use 'the method of grouping' to compute the mode.

Consider the following distribution

Value of the item	Frequency
1	2
2	3
3	2
4	6
5	10
6	12
7	11
8	3
9	1

The maximum frequency is 12 and the adjacent frequencies are 10 and 11, very close to 12. So we calculate the mode in the following way :

A grouping table has six columns. We group the members into 'twos' and find their totals and note the pair giving the maximum sum. Then we group them into 'threes' and do the same thing. Then we find the value of the item which occurs maximum number of times in the groups containing the maximum sums. That is taken as the mode.

We adopt the following method. First, we group the first two, third and fourth, fifth and sixth etc. and find their sum. If we have even number of items all the items will be included in the grouping. If odd, the last number will be omitted. Next, we omit the first number and group the second and the third, the fourth and the fifth etc and find their sums. As above, if we have odd number of items all the numbers from the second to the last will be included, otherwise the last will be included.

Next we group the first three namely, the first, second, third items, the fourth, fifth sixth items etc and find their sum. If the total number of items is a multiple of 3, all the items will be included in grouping. Otherwise, the last one or the last two items will be omitted. Next, we omit the first item and group the other items three by three and find their sum. Lastly we omit the first two and group the other items three by three and find their sum. If we want we could proceed further. But usually we stop with the grouping of three items. Then we note down the groups having the maximum sums and count the number of times each size occurs in these groups. The size occurring maximum number of times is taken as the mode.

Let us do the above problem using this method.

Grouping table

Value of the items	Frequencies and their sums							
	1	2	3	4	5	6		
1	2	} 5	} 5	} 7	} 11	} 18		
2	3							
3	2	} 8	} 16	} 28	} 33			
4	6							
5	10	} 22	} 23	} 15	} 33	} 26		
6	12							
7	11	} 14	} 4		} 15			
8	3							
9	1							

Analysis Table

Columns	Items corresponding to the max. freq								
	1	2	3	4	5	6	7	8	9
1						✓			
2					✓	✓			
3						✓	✓		
4				✓	✓	✓			
5					✓	✓	✓		
6						✓	✓	✓	
Total	-	-	-	1	3	6	3	1	

Six occurs maximum number (=6) of times and hence the mode is 6.

Example : Calculate the mode using the method of grouping.

Value of the items	Freq
2	12
4	14
6	15
8	18
10	20
12	19
14	12
16	8
18	4
20	3

Grouping Table

Value of the items	Frequencies					
	1	2	3	4	5	6
2	12	} 26	} 29	} 41	} 47	} 53
4	14					
6	15	} 33	} 38	} 57	} 51	
8	18					
10	20	} 39	} 31	} 24	} 15	} 39
12	19					
14	12	} 20	} 12	} 7	} 3	
16	8					
18	4	} 7	} 3	} 1	} 0	
20	3					

Analysis Table

Columns	Items corresponding to the max. freq				
	6	8	10	12	14
1			✓		
2			✓	✓	
3		✓	✓		
4		✓	✓	✓	
5			✓	✓	✓
6	✓	✓	✓		
Total	1	3	6	3	1

10 occurs maximum number (=6) of times and hence the mode = 10.

(Since in 16, 18, 20 the mode is not expected to lie these value of the items have been left out).

Calculation of Mode - Continuous Series : Method 1

Steps : (i) By preparing grouping table and analysis table or by inspection ascertain the modal class.

(ii) Determine the value of mode by applying the following formula :

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

where L = lower limit of the modal class. Δ_1 = the difference between the frequency of the modal class and the frequency of the pre-modal class, Δ_2 = the difference between the frequency of the modal class and the frequency of the post - model class, i = the class interval of the modal class.

Example 1 : Calculate mode from the following data :

Marks	No. of Students
Above 0	80
Above 10	77
" 20	72
" 30	65
" 40	55
" 50	43
" 60	28
" 70	16
" 80	10
" 90	8
" 100	0

Solution : Since this is cumulative frequency distribution, we first convert it into a simple frequency distribution.

Marks	No. of students
0-10	3
10-20	5
20-30	7
30-40	10
40-50	12
50-60	15
60-70	12
70-80	6
80-90	2
90-100	8

By inspection the modal class is 50 - 60.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 50, \Delta_1 = (15 - 12) = 3. \quad \Delta_2 = 15 - 12 = 3. \quad i = 10$$

$$M_0 = 50 + \frac{3}{3 + 3} \times 10 = 50 + 5 = 55$$

Example 2 : Find the value of mode from the data given below :

Weight (kg)	No. of students
93-97	2
98-102	5
103-107	12
108-112	17
113-117	14
118-122	6
123-127	3
128-132	1

Solution : By inspection mode lies in the class 108-112. But the real limits of this class are 107.5-112.5

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 107.5, \Delta_1 = 17-12 = 5, \Delta_2 = 17-14 = 3, i = 5$$

$$\therefore \text{Mode} = 107.5 + \frac{5}{5 + 3} \times 5 = 107.5 + 3.125 = 110.625$$

Thus the modal weight is 110.625kg.

Example 3 : From the following data of weight of 122 persons determine the modal weight :

Weight (in lbs)	No. of persons
100-110	4
110-120	6
120-130	20
130-140	32
140-150	33
150-160	17
160-170	8
170-180	2

Solution : By inspection it is difficult to say which is the modal class. Hence we prepare a grouping table and an analysis table.

Grouping table

Weight (in lbs)	No. of persons						
	1	2	3	4	5	6	
100 - 110	4	} 10	} 26	} 30	} 58	} 85	
110 - 120	6						
120 - 130	20	} 52	} 65	} 82	} 58		
130 - 140	32						
140 - 150	33	} 50	} 25			} 27	
150 - 160	17						
160 - 170	8	} 10					
170 - 180	2						

Analysis Table

Columns	Class in which mode is expected to lie		
	120 - 130	130 - 140	140 - 150
1			✓
2	✓	✓	✓
3		✓	
4		✓	✓
5	✓	✓	✓
6	✓	✓	✓
Total	3	5	5

This is a bi-modal series. Hence mode has to be determined by applying the formula named as **empirical formula** :

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Weight in lbs	m	No. of persons	c.f.	$\frac{(m-135)}{10}$ d	fd
100-110	105	4	4	-3	-12
110-120	115	6	10	-2	-12
120-130	125	20	30	-1	-20
130-140	135	32	62	0	0
140-150	145	33	95	1	33
150-160	155	17	112	2	34
160-170	165	8	120	3	24
170-180	175	2	122	4	8
		N = 122			$\Sigma fd = 55$

$$\bar{x} \text{ (Mean)} = A + \frac{\Sigma fd}{N} \times i$$

$$A = 135, \Sigma fd = 55, N = 122, i = 10.$$

$$\bar{x} = 135 + \frac{55}{122} \times 10 = 135 + 4.51 = 139.51$$

$$\text{Median} = \text{size of } \frac{N}{2} \text{th item}$$

$$= \text{Size of } \frac{122}{2} = 61 \text{st item.}$$

Hence median lies in the class 130 - 140.

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

$$L = 130, \frac{N}{2} = 61, \text{c.f.} = 30, f = 32, i = 10.$$

$$\text{Median} = 130 + \frac{61-30}{32} \times 10$$

$$= 130 + \frac{310}{32} = 139.69$$

$$\begin{aligned} \text{Mode} &= 3 \text{ Median} - 2 \text{ Mean} \\ &= (3 \times 139.69) - (2 \times 139.51) \\ &= 419.07 - 279.02 = 140.05. \end{aligned}$$

$\therefore$ Modal weight is 140.05lb.

Method 2 : Graphical Method

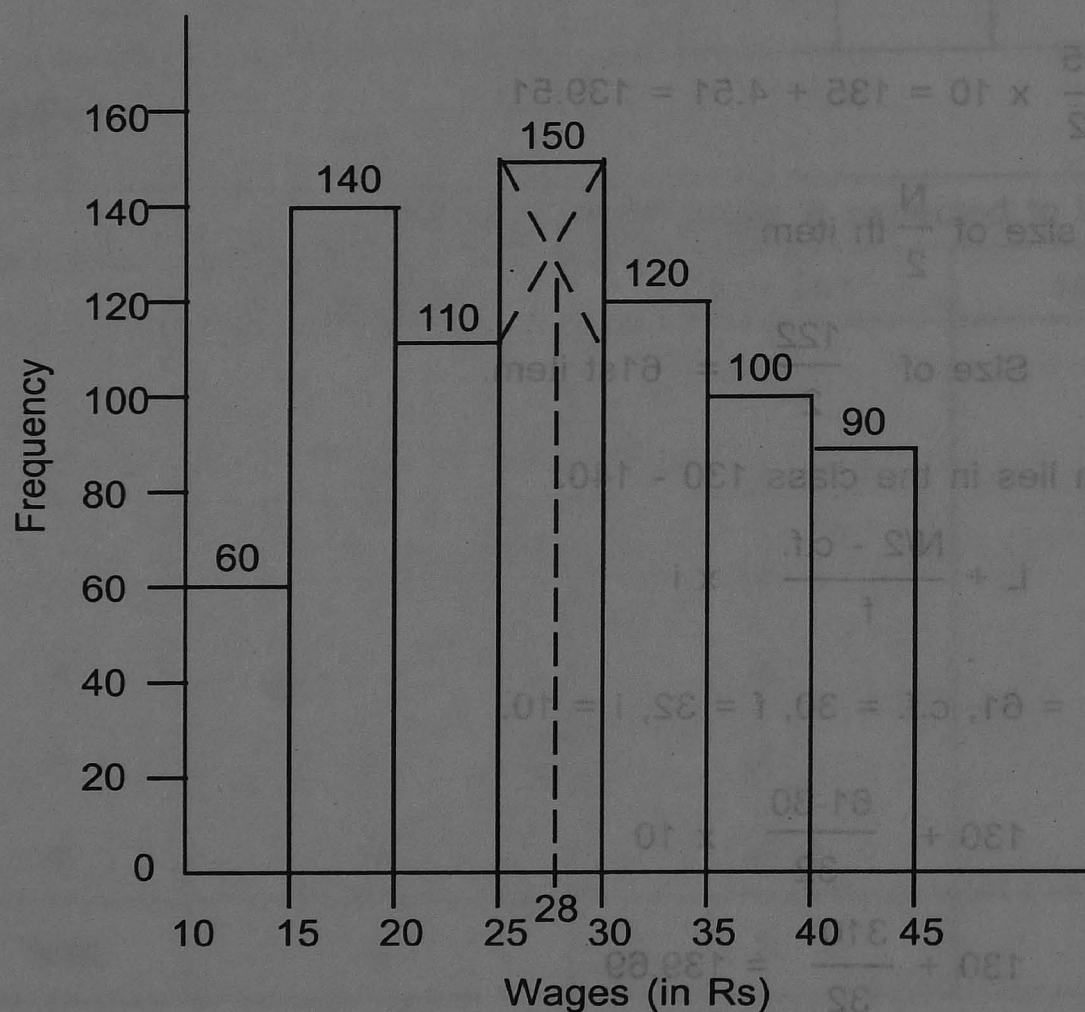
Steps : 1. Draw a histogram of the given data.

2. Draw two lines diagonally in the inside of the modal class for, starting from each upper corner of the bar to the upper corner of the adjacent bar.
3. Draw a perpendicular line from the intersection of the two diagonal lines to the x axis which gives us the modal value.

Example : Draw a histogram for the following distribution and find the model wage and check the value by direct calculation.

Wages (in Rs) :	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of Workers :	60	140	110	150	120	100	90

Solution



It is clear from the histogram that the modal value is 28.

Direct Calculation

Mode lies in the class 25-30.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 25, \Delta_1 = 150 - 110 = 40, \Delta_2 = 150 - 120 = 30, i = 5.$$

$$M_0 = 25 + \frac{40}{40 + 30} \times 5$$

$$= 25 + 2.86 = 27.86.$$

The slight difference in the two answers is due to the difficulty of reading very accurately from the graph.

Exercise

- Find the mode of the following ungrouped data 16, 17, 17, 17, 17, 18, 19, 20, 20, 20.
- Find the mode of the following discrete series using method of grouping.

X	95	105	115	125	135	145	155	165
F	10	37	65	80	51	35	18	4

- Find the mode of the following continuous series by the two methods. Calculate the mode using the empirical relation also.

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Freq	10	12	17	21	20	18	11	10

(ii) Marks : Above 0 Above 10 Above 20 Above 30 Above 40

No. of

Students : 80 77 72 65 43

Above 50 Above 60 Above 70 Above 80 Above 90 Above 100

55 28 16 10 8 0

1.4 Geometric Mean

Geometric mean is defined as the Nth root of the product of N items or values. If there are two items, we take the square root; if there are three items, the cube root and so on. In symbol, $G.M. = \sqrt[N]{(x_1)(x_2) \dots (x_n)}$ where $x_1, x_2, \dots, x_n$ refer to the various items of the series.

Example : Find the G.M. of 2, 3, 4.

$$G.M. = \sqrt[3]{2 \times 3 \times 4} = \sqrt[3]{24} = 2.885.$$

When the number of items is three or more the task of multiplying the numbers and of extracting the root becomes difficult.

To simplify calculations logarithms are used.

$$\therefore \log G.M = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{N}$$

$$\log G.M = \frac{\sum \log x}{N}$$

$$\therefore G.M = \text{Antilog} \left(\frac{\sum \log x}{N} \right)$$

Geometric Mean for Discrete Series

Consider the general table :

x	f
x_1	f_1
x_2	f_2
.	.
.	.
x_n	f_n
Total	N

clearly x_1 occurs f_1 times, x_2 occurs f_2 times, x_n occurs f_n times. So, the product

$$= x_1^{f_1} x_2^{f_2} \dots x_n^{f_n}$$

$$\therefore \text{G.M} = \left(\frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{N} \right)^{1/N}$$

$$\log \text{G.M} = \frac{1}{N} [f_1 \log x_1 + f_2 \log x_2 + \dots + f_n \log x_n]$$

$$= \frac{\sum f_i \log x_i}{N}$$

$$\text{G.M} = \text{Antilog} \left(\frac{\sum f_i \log x_i}{N} \right)$$

Example : Find the G.M. for the following data.

Marks	11	12	13	14	15
No. of students	3	7	8	5	2

Marks x	No. of Students	log x	f log x
11	3	1.0414	3.1242
12	7	1.0792	7.5544
13	8	1.1139	8.9112
14	5	1.1461	5.7305
15	2	1.1761	2.3522
	25		27.6725

$$\log \text{G.M.} = \frac{\sum f \log x}{N}$$

$$= \frac{27.6725}{25} = 1.1069$$

$$\text{G.M.} = \text{Antilog } 1.1069 = 12.79$$

Geometric Mean for Continuous Series

As in the case of the calculation of A.M., we assume that the items are concentrated at the midvalues and hence we get a discrete series. Then applying the formula,

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum_{i=1}^n f_i \log x_i}{N} \right) \quad \text{where } x_i \text{ is the midpoint of the } i\text{th interval, we get the geometric mean.}$$

Example : Find the G.M. of the following distribution

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	6	7	8	3	1

Midvalue	f	log x	f log x
5	6	0.6990	4.1940
15	7	1.1761	8.2327
25	8	1.3979	11.1832
35	3	1.5441	4.6323
45	1	1.6532	1.6532
	25		29.8954

$$\log \text{G.M.} = \frac{\sum f \log x}{N}$$

$$= \frac{29.8954}{25} = 1.1958$$

$$\text{G.M.} = \text{Antilog } 1.1958 = 15.7.$$

Properties of the Geometric Mean

Theorem 1 : If n items have a G.M. G_1 and m items have a G.M. G_2 and if G be the G.M. of the combined group then

$$\log G = \frac{n \log G_1 + m \log G_2}{n + m}$$

Proof : Let $x_1, x_2, \dots, x_n$ be the n items of the first group and $y_1, y_2, \dots, y_m$ be the m items of the second group.

$$G = \sqrt[n+m]{(x_1 x_2 \dots x_n y_1 y_2 \dots y_m)}$$

$$\log G = \frac{1}{n+m} [\log x_1 + \log x_2 + \dots + \log x_n + \log y_1 + \dots + \log y_m]$$

$$= \frac{1}{n+m} \left(\sum_{i=1}^n \log x_i + \sum_{j=1}^m \log y_j \right)$$

$$\log G_1 = \frac{1}{n} \sum_{i=1}^n \log x_i$$

$$\log G_2 = \frac{1}{m} \sum_{j=1}^m \log y_j$$

$$n \log G_1 = \sum_{i=1}^n \log x_i$$

$$\therefore m \log G_2 = \sum_{j=1}^m \log y_j$$

$$\log G = \frac{1}{n+m} [n \log G_1 + m \log G_2]$$

$$\text{ie, } \log G = \frac{n \log G_1 + m \log G_2}{n + m}$$

Theorem 2 : The geometric mean of a set of ratios is the quotient of the geometric mean of the numerators and geometric mean of the denominators.

Proof :

Let $\frac{a_1}{b_1}, \frac{a_2}{b_2}, \dots, \frac{a_n}{b_n}$ be the n ratios

$$\text{G.M.} = \left(\frac{a_1}{b_1}, \frac{a_2}{b_2}, \dots, \frac{a_n}{b_n} \right)^{1/n}$$

$$= \left(\frac{a_1}{b_1} \frac{a_2}{b_2} \frac{a_n}{b_n} \right)^{1/n}$$

$$= \frac{(a_1 a_2 \dots a_n)^{1/n}}{(b_1 b_2 \dots b_n)^{1/n}}$$

$$= \frac{\text{G.M. of the Numerators}}{\text{G.M. of the Denominators}}$$

Uses of Geometric Mean : The second theorem proved above makes the geometric mean a highly suitable average for use in index numbers which are really ratios. Generally, we use it with advantage in the cases of any set of ratios like production figures, sales figures, population figures etc.

For example, suppose we know the rates of increase of population over n periods say $r_1, r_2, \dots, r_n$ and r is the average rate of increase. Then r is calculated using the geometric mean.

Let p_0 be the population at the beginning of the period and $p_1, p_2, \dots, p_n$ the population of the succeeding periods and

$$\frac{p_1 - p_0}{p_0} = r_1$$

$$\therefore \frac{p_1}{p_0} = 1 + r_1$$

$$\text{Similarly, } \frac{p_2}{p_1} = 1 + r_2$$

$$\frac{p_5}{p_4} = 1 + r_5$$

$$\frac{p_n}{p_{n-1}} = 1 + r_n$$

$$\frac{p_n}{p_0} = \frac{p_n}{p_{n-1}} \cdot \frac{p_{n-1}}{p_{n-2}} \dots \frac{p_1}{p_0}$$

$$= (1 + r_n) (1 + r_{n-1}) \dots (1 + r_1)$$

If r is the average rate of increase, then $p_n = p_0 (1 + r)^n$

$$\therefore \frac{p_n}{p_0} = (1 + r)^n$$

$$\therefore (1 + r)^n = (1 + r_1) (1 + r_2) \dots (1 + r_n)$$

$$1 + r = \sqrt[n]{(1 + r_1) (1 + r_2) \dots (1 + r_n)}$$

$$r = \sqrt[n]{(1 + r_1) (1 + r_2) \dots (1 + r_n)} - 1$$

So, if $r_1, r_2, \dots, r_n$ are known, r is calculated using the formula 2. If the population p_n at the end and p_0 at the beginning are known then formula 1 will give the answer.

Example 1 : Find the average increase in the prices of three articles which increased in prices in a certain period by 80%, 40% and 20% respectively.

Solution :

$$r_1 = \frac{80}{100} = 0.8$$

$$r_2 = \frac{40}{100} = 0.4$$

$$r_3 = \frac{20}{100} = 0.2$$

Let r be the average rate of increase.

$$\text{Then } (1+r)^3 = (1+r_1)(1+r_2)(1+r_3)$$

$$= 1.8 \times 1.4 \times 1.2$$

$$1+r = \sqrt[3]{1.8 \times 1.4 \times 1.2} = 1.446$$

$$\therefore r = 0.446$$

$\therefore$ The average percent increase = 44.6%

Example 2 : The following table gives the population of a country in five decades. Find the average rate of increase and estimate the population for 1981 and 1966.

Years of Census	Population in millions
1931	283.9
1941	303.4
1951	305.7
1961	338.1
1971	388.8

Solution :

Let x be the average rate of increase

$$\therefore (1+r)^4 = \frac{388.8}{283.9}$$

$$1+r = \frac{\sqrt[4]{388.8}}{283.9}$$

$$= 1.081$$

$$\therefore r = 0.081$$

$$\text{Average rate of increase} = 8.1\%$$

$$\text{The estimated value for 1981} = 388.8 (1+r)$$

$$= 388.8 (1.081)$$

$$= 420.29 \text{ millions}$$

The population of 1966 is the G.M. of the population in 1961 and 1971. Hence it is equal to $\sqrt{338.1 \times 388.8} = 362.5 \text{ millions}$

Weighted Geometric Mean

We can calculate weighted geometric mean with the help of the following formula :

$$\begin{aligned} \text{G.M}_w &= \text{Anti log} \left(\frac{(\log x_1 \times w_1) + (\log x_2 \times w_2) + \dots + (\log x_n \times w_n)}{w_1 + w_2 + w_3 + \dots + w_n} \right) \\ &= \text{Antilog} \left(\frac{\sum (\log x \times w)}{\sum w} \right) \end{aligned}$$

$$\text{In symbol, } \text{G.M}_w = \sqrt[w_1]{x_1} \times \sqrt[w_2]{x_2} \times \sqrt[w_3]{x_3} \times \dots \times \sqrt[w_n]{x_n}$$

Example 1 : Find the weighted geometric mean from the following data :

Group	Index number	Weights
Food	260	46
Fuel and Lighting	180	10
Clothing	220	8
House Rent	230	20
Education	120	12
Misc	200	4

Solution

Group	Index No x	Weights w	Log x	w. Logx
Food	260	46	2.4150	111.0900
Fuel and Lighting	180	10	2.2553	22.5530
Clothing	220	8	2.3424	18.7392
House Rent	230	20	2.3617	47.2340
Education	120	12	2.0792	24.9504
Misc	200	4	2.3010	9.2040

$$\Sigma w = 100 \quad \Sigma w \log x = 233.7706$$

$$\begin{aligned} \text{G.M.}_w &= \text{Antilog} \left(\frac{\Sigma w \log x}{\Sigma w} \right) = \text{Antilog} \left(\frac{233.7706}{100} \right) \\ &= 217.6. \end{aligned}$$

Example 2 : The weighted geometric mean of the four numbers 8, 25, 17 and 30 is 15.3. If the weights of the first three numbers are 5, 3 and 4 respectively, find the weight of the fourth number.

Solution : Let the weight of the fourth number be w_1 .

x	w	log x	w log x
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
30	w_1	1.4771	$1.4771w_1$

$$\Sigma w = 12 + w_1$$

$$\Sigma w \log x = 13.6308 + 1.4771w_1$$

$$\log \text{G.M.} = \frac{\Sigma w \log x}{\Sigma w}$$

$$\log 15.3 = \frac{13.6308 + 1.4774w_1}{12 + w_1}$$

$$1.1847 (12 + w_1) = 13.6308 + 1.4771w_1$$

$$14.2164 + 1.1847w_1 = 13.6308 + 1.4771w_1$$

$$1.1847w_1 - 1.4771w_1 = 13.6308 - 14.2164 = -0.5856$$

$$w_1 = \frac{0.5856}{0.2922} = 2.003 \text{ or } 2 \text{ app.}$$

$\therefore$ The weight of the fourth number is 2.

1.5 - Harmonic Mean

Definition : The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocal of the arithmetic mean of the reciprocal of the individual observations.

$$\text{ie., H.M.} = \frac{N}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right)}$$

$$\text{In discrete series H.M.} = \frac{N}{\sum \left(f \times \frac{1}{x} \right)} = \frac{N}{\sum \left(\frac{f}{x} \right)}$$

$$\text{In continuous series H.M.} = \frac{N}{\sum \left(f \times \frac{1}{m} \right)} = \frac{N}{\sum \left(\frac{f}{m} \right)}$$

where m is the mid point

Calculation of Harmonic Mean - Individual Observations

Example 1 : Find the harmonic mean from the following : 5, 10, 20

$$\begin{aligned} \text{Harmonic Mean} &= \frac{3}{\frac{1}{5} + \frac{1}{10} + \frac{1}{20}} \\ &= \frac{3}{0.2 + 0.1 + 0.05} \\ &= \frac{3}{0.35} \end{aligned}$$

$$\text{Harmonic Mean} = 8.57$$

Calculation of Harmonic Mean - Discrete Series

Example 2 : From the following data compute the value of harmonic mean :

Marks	10	20	25	40	50
No. of students	20	30	50	15	5

Solution

Marks x	f	f/x
10	20	2.000
20	30	1.500
25	50	2.000
40	15	0.375
50	5	0.100
N = 120		$\Sigma (f/x) = 5.975$

$$\text{H.M.} = \frac{N}{\Sigma (f/x)} = \frac{120}{5.975} = 20.08$$

Calculation of Harmonic Mean - Continuous Series

Example 3 : From the following data compute the value of harmonic mean :

Class Interval	10-20	20-30	30-40	40-50	50-60
Frequency	4	6	10	7	3

Solution :

Class Interval	Midpoint m	Frequency f	f / m
10-20	15	4	0.267
20-30	25	6	0.240
30-40	35	10	0.286
40-50	45	7	0.156
50-60	55	3	0.055
		N = 30	$\Sigma (f/m) = 1.004$

$$\text{H.M.} = \frac{N}{\Sigma (f/m)} = \frac{30}{1.004} = 29.88$$

Example 4 : An automobile driver travels from plain to hill station 100 km distance at an average speed of 30 km per hour. He then makes the return trip at average speed of 20 km per hour. What is his average speed over the entire distance (200 km)?

Solution

Harmonic mean of 30 and 20 is

$$\text{H.M.} = \frac{2}{\frac{1}{20} + \frac{1}{30}} = \frac{2}{\frac{10}{120}} = \frac{2 \times 120}{10} = 24 \text{ km/hr}$$

$$\text{H.M.} = 24 \text{ km p. h.}$$

Example 5 : An aeroplane covers the four sides of a square at speeds of 1,000, 2,000, 3,000 and 4,000 k.m. per hour respectively. What is the average speed of the plane in its flight around the square?

Solution

$$\text{H.M.} = \frac{4}{\frac{1}{1000} + \frac{1}{2000} + \frac{1}{3000} + \frac{1}{4000}}$$

$$= \frac{4}{\frac{12 + 6 + 4 + 3}{12000}} = \frac{4}{\frac{25}{12000}}$$

$$= \frac{4 \times 12000}{25} = 1920 \text{ km per hour}$$

$$\text{H.M.} = 1920 \text{ km per hour.}$$

Weighted Harmonic Mean : It is calculated with the help of the following formula :

$$\begin{aligned} \text{H.M}_w &= \frac{\sum w}{\left(\frac{1}{a} \times w_1\right) + \left(\frac{1}{b} \times w_2\right) + \left(\frac{1}{c} \times w_3\right)} \\ &= \frac{\sum w}{\sum (w/x)} \end{aligned}$$

Example 6 : A cyclist covers his first five km at an average speed of 10 km p h another 3 km at 8 km p.h. and the last two km at 5 km p.h. Find the average speed of the entire journey and verify your answer.

Solution

$$\text{H.M}_w = \frac{\sum w}{\left(\frac{1}{a} \times w_1\right) + \left(\frac{1}{b} \times w_2\right) + \left(\frac{1}{c} \times w_3\right)}$$

The various speeds are 10 km p.h., 8 km p.h., and 5 km p.h. $\therefore$ a, b and c are 10, 8 and 5 respectively. The distances covered are 5, 3 and 2 km respectively. Hence w_1 , w_2 and w_3 are 5, 3 and 2.

$$\begin{aligned} \text{H.M.}_w &= \frac{10}{\left(\frac{1}{10} \times 5\right) + \left(\frac{1}{8} \times 3\right) + \left(\frac{1}{5} \times 2\right)} \\ &= \frac{10}{\frac{1}{2} + \frac{3}{8} + \frac{2}{5}} = \frac{10}{\frac{51}{40}} \\ &= \frac{10 \times 40}{51} = 7.84 \text{ km p.h.} \end{aligned}$$

$\therefore$ The average speed for the entire journey is 7.84 km p.h.

Verification

Speed (km p.h.)	Distance (km)	Time taken (in minutes)
10	5	30.0
8	3	22.5
5	2	24.0
Total	10	76.5

In 76.5 minutes he covers 10 km.

$\therefore$ In 60 minutes he would cover $\frac{10 \times 60}{76.5} = 7.84 \text{ km}$

Average speed = 7.84 km.

Properties of various averages

Properties of the A.M.

1) The algebraic sum of the deviations of all the items from the mean is zero.

Proof

$$\text{We know that } \bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$$

$$\therefore N \bar{x} = \sum f_i x_i \quad 1$$

Sum of the algebraic deviations of the items from the mean

$$\begin{aligned}
 &= \sum f_i (x_i - \bar{x}) \\
 &= \sum f_i \bar{x} - \sum f_i \bar{x} \\
 &= N \bar{x} - \bar{x} \sum f_i \\
 &= N \bar{x} - N \bar{x} = 0.
 \end{aligned}$$

2. The sum of the squared deviations about the mean is minimum.

Proof : Let A be any origin

Sum of the squared deviations from A = $\sum f_i (x_i - A)^2$

$$\begin{aligned}
 &= \sum f_i [(x_i - \bar{x}) + (\bar{x} - A)]^2 \\
 &= \sum f_i [(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - A) + (\bar{x} - A)^2] \\
 &= \sum f_i (x_i - \bar{x})^2 + 2(\bar{x} - A) \sum f_i (x_i - \bar{x}) + (\bar{x} - A)^2 \sum f_i \\
 &= \sum f_i (x_i - \bar{x})^2 + N(\bar{x} - A)^2 \text{ (by the first property)}
 \end{aligned}$$

$$\therefore \sum f_i (x_i - A)^2 = \sum f_i (x_i - \bar{x})^2 + N(\bar{x} - A)^2$$

$$(\bar{x} - A)^2 = 0 \text{ for } \bar{x} = A \text{ and}$$

$$> 0 \text{ for } \bar{x} \neq A$$

$\therefore \sum f_i (x_i - A)^2$ is minimum when $\bar{x} = A$.

3. If $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$ are two series of observations and $z_i = x_i + y_i$ then $\bar{z} = \bar{x} + \bar{y}$.

Proof : By definition $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

$$\bar{z} = \frac{(x_1 + y_1) + (x_2 + y_2) + \dots + (x_n + y_n)}{n}$$

$$= \frac{(x_1 + x_2 + \dots + x_n) + (y_1 + y_2 + \dots + y_n)}{n}$$

$$= \bar{x} + \bar{y}.$$

Property of Median

The sum of the deviations of the items from median, ignoring signs is the least. For example, the median of 4, 6, 8, 10, 12 is 8. The deviations from 8 ignoring signs are 4, 2, 0, 2, 4 and the total is 12. This total is smaller than the one obtained if deviations are taken from any other value. Thus if deviations are taken from 7, values ignoring signs would be 3, 1, 1, 3, 5 and the total is 13.

Merits and Demerits of various averages

1. Arithmetic Mean

Merits

1. It is the simplest average to understand and easiest to compute.
2. It is affected by the value of every item in the series.
3. It is defined by a rigid mathematical formula with the result that everyone who computes the average gets the same answer.
4. Being determined by a rigid formula, it lends itself to subsequent algebraic treatment better than the median or mode.
5. It is relatively reliable.
6. The mean is typical in the sense that it is the centre of gravity, balancing the values on either side of it.
7. It is a calculated value, and not based on position in the series.

Demerits

1. Since the value of mean depends upon each and every item of the series, extreme items, i.e., very small and very large items, affect the value of the average.
2. In a distribution with open-end classes the value of mean cannot be computed without making assumption regarding the size of the class interval of the open-end classes.
3. The arithmetic mean is not always a good measure of central tendency.

Median

Merits

1. The median has a well defined formula.
2. The calculations involved in finding the median are simple and readily understood particularly in the cases of ungrouped data and discrete frequency distribution.
3. In most cases, it is a proper representative of the data on hand and its meaning is easily understood.
4. Median is not affected by the extreme values.
5. It can be calculated without a knowledge of the values of the extreme items provided the total number of items is known.

Demerits

1. In cases, where extreme items are to be given importance, median is less useful.
2. It requires the given data, if ungrouped to be arrayed before it can be determined and that involves considerable work.
3. Given the median of two or more distributions we cannot obtain the median of the combined distribution.

Mode

Merits

1. Mode is easily understood.
2. It can be easily located if the frequency curve is available.
3. The mode is not affected by extreme values.
4. Its value can be determined in open-end distributions without ascertaining the class limits.
5. It can be used to describe qualitative phenomenon.

Demerits

1. The value of mode cannot always be determined. In some cases we may have a bimodal series.
2. It is not capable of algebraic manipulations.
3. The value of mode is not based on each and every item of the series.
4. It is not a rigidly defined measure.
5. Mode is quite unstable.

Geometric Mean

Merits

1. It is based on each and every item of the series.
2. It is rigidly defined.
3. It is useful in average ratios and percentages and in determining rates of increase and decrease.
4. It gives less weight to large items and more to small ones than does the arithmetic average.
5. It is capable of algebraic manipulation.

Demerits

1. It is difficult to understand
2. It is difficult to compute and to interpret and so has restricted application.
3. It cannot be computed when there are both negative and positive values in a series or one or more of the values are zero.

Harmonic Mean

Merits

1. Its value is based on every item of the series.
2. It tends itself to algebraic manipulation.
3. In problems relating to time and rates it gives better results than other averages.

Demerits

1. It is not easily understood.
2. It is difficult to compute.
3. It gives largest weight to smallest items.

The above discussions of merits and demerits of various averages show that each average has its own field of importance and no single average is appropriate for all kinds of data. The suitability of different averages depends upon the nature of data and the statistical problem on hand.

Arithmetic mean is the most suitable average in many social and economic studies especially in the calculation of average output, average import and export, average cost of production etc. In general when we have a fairly symmetrical distribution we use A.M.

Geometric Mean is the only suitable average in the calculation of index numbers of prices and periodic increase in population where the rates of change are of major concern and where equal relative changes should be regarded as of equal importance.

Harmonic mean is a type of average capable of application only within restricted field. In finding the average of time rates and price movements, it is more suitable.

Median is useful when the given data are not directly capable of numerical measurement. Also, when the data are incomplete but the total number of items and accurate information about the central items are known, median is the only average that could be used.

In business and commerce, mode is of very great value. When the emphasis is on the most typical or the most common value mode is the suitable one.

To conclude, the A.M. is highly useful and is to be used when we require a fairly reliable average or when we have further calculations to be made.

Exercise

1. Describe the different measures of central tendency used in statistics.
2. Calculate the G.M. and H.M. for the following data

Marks	11	12	13	14	16
No. of students	3	7	8	5	2

3. An aeroplane flies along the four sides of a square at speeds 500 kmph, 600 kmph, 700 kmph and 800 kmph. Find its average speed.

Measures of Dispersion

The various measures of central value discussed above give us one single figure that represents the entire data. But the averages alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of the distribution. Measures of dispersion help us in studying this important characteristic of a distribution.

Let us consider the following three data containing five values each representing marks obtained by students :

Data I 60, 60, 60, 60, 60

Data II 56, 60, 64, 62, 58

Data III 0, 80, 100, 30, 90.

The arithmetic mean of each set is 60. Yet there are differences between the three data. The data themselves are not identical. A noticeable difference between these three data is the difference in the degrees of concentration between them. In data I, all the values are equal to the A.M. In data II the values are scattered slightly but on the whole they lie fairly close to the A.M. In data III, the items do not fall close to each other and show comparatively greater difference since one value is as low as zero and another value is as high as 100.

From the above observations that averages are not sufficient. They must be supported and supplemented with other measures. A measure of variation or dispersion is one that measures the extent to which there are differences between individual observations and some central or average value.

Significance of measuring variation

Measures of variation are needed for four basic purposes :

- (i) To determine the reliability of an average.
- (ii) To serve as a basis for the control of the variability.
- (iii) To compare two or more series with regard to their variability.
- (iv) To facilitate the use of other statistical measures.

Properties of a good measure of variation

- (i) It should be simple to understand
- (ii) It should be easy to compute
- (iii) It should be rigidly defined
- (iv) It should be based on each and every item of the distribution.
- (v) It should be amenable to further algebraic treatment
- (vi) It should have sampling stability
- (vii) It should not be unduly affected by extreme items.

Methods of studying variation

The following are the important methods of studying variation :

1. The Range
2. The Interquartile Range and the Quartile Deviation (Q.D.)
3. The Mean Deviation or Average Deviation (M.D.)
4. The Standard Deviation (S.D.)

Of these the first two, namely, the range and quartile deviation are positional measures because they depend on the values at a particular position in the distribution. The other two, the average deviation and the standard deviation, are called calculation measures of deviation because all of the values are employed in their calculation.

Absolute and Relative Measures of Variation

Measures of dispersion may be either absolute or relative. Absolute measures of dispersion are expressed in the same statistical unit in which the original data are given such as rupees, kilograms, tonnes, etc. These values may be used to compare the variation in two distributions provided the variables are expressed in the same units and of the same average size. In case, the two sets of data are expressed in different units, such as quintals of sugar

versus tonnes of sugarcane, or if the average size is very different such as manager's salary versus workers' salary, the absolute measures of dispersion are not comparable. In such cases measures of relative dispersion should be used.

A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It should be remembered that while computing the relative dispersion the average used as base should be the same one from which the absolute deviations were measured. This means that the arithmetic mean should be used with the standard deviation, and either the arithmetic mean or median with the mean deviation.

1.6 - Range

Range is the simplest method of studying dispersion. It is defined as the difference between the value of the smallest item and the value of the largest item included in the distribution.

In symbol, Range = $L - S$ Where L = Largest item and S = smallest item.

The relative measure corresponding to range, called the coefficient of range, is obtained of applying the following formula :

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

If the averages of the two distributions are about the same, a comparison of the range indicates that the distribution with the smaller range has less dispersion, and the average of that distribution is more typical of the group.

Example 1 : The following are the prices of shares of ABCO. Ltd from Monday to Saturday.

Day	Price (Rs.)
Monday	200
Tuesday	210
Wednesday	208
Thursday	160
Friday	220
Saturday	250

Calculate range and its coefficient.

Solution :

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

$$L = 250 \text{ and } S = 160$$

$$\therefore \text{Range} = 250 - 160 = \text{Rs. } 90$$

$$\begin{aligned} \text{Coefficient of Range} &= \frac{L - S}{L + S} \\ &= \frac{250 - 160}{250 + 160} = \frac{90}{410} = 0.22 \end{aligned}$$

There are two methods of determining the range from data grouped into a frequency distribution. The first method is to find the difference between the upper limit of the highest wage class and the lower limit of the lowest wage class. The other method is to subtract the mid point of the lowest wage class from the mid point of the highest wage class. In practice, both the methods are used.

Example 2 : Calculate coefficient of Range from the following data :

Marks	No. of students
10-20	8
20-30	10
30-40	12
40-50	8
50-60	4

Solution

$$\text{Range} = L - S = 60 - 10 = 50 \text{ (By first method)}$$

$$\begin{aligned} \text{Coefficient of range} &= \frac{L - S}{L + S} \\ &= \frac{60 - 10}{60 + 10} = \frac{50}{70} = 0.714 \end{aligned}$$

$$\text{Range} = 30 - 5 = 25 \text{ (By the second method)}$$

$$\begin{aligned} \text{Coefficient of range} &= \frac{25}{30 + 5} = \frac{25}{35} = \frac{5}{7} \\ &= 0.714. \end{aligned}$$

Merits of Range

1. It is the simplest method to understand.
2. It is the easiest to compute.
3. It takes the minimum time to calculate the value of range.

Limitations

1. Range is not based on each and every item of the distribution.
2. It is subject to fluctuations of considerable magnitude from sample to sample.
3. Range cannot be computed in case of open-end distributions.
4. Range is most unreliable.

Uses of Range

1. When statistical methods of quality control are used, control charts are prepared and in preparing these charts range plays a very important role.
2. Range is useful in studying the variations in the prices of stocks and shares and other commodities that are sensitive to price changes from one period to another.
3. The meteorological department does make use of the range in determining the difference between the minimum temperature and the maximum temperature.
4. The range is a most commonly used measure of dispersion in everyday life.

1.7 - The interquartile range (or) The Quartile Deviation (QD)

The unstable character of the range, as a measure of dispersion, can be remedied to a certain extent by leaving out the items at either end of the distribution and considering the difference between two other more suitable items known as Quartiles.

We define the first quartile (the lower quartile) Q_1 to be the value of the variable which is such that one quarter of the observations lie below Q_1 . Similarly, we define the third quartile (or the upper quartile) Q_3 to be that value of the variable which is such that three quarters of the observations lie below Q_3 . Thus if M is the median, Q_1 , M and Q_3 divide the distribution into four equal parts.

The interquartile range represents the difference between the third quartile and the first quartile. In symbol,

$$\text{Interquartile range} = Q_3 - Q_1$$

The interquartile range is reduced to the form of the semi-interquartile range or quartile deviation by dividing it by 2.

$$\text{In symbol, Quartile Deviation (Q.D.)} = \frac{Q_3 - Q_1}{2}$$

Quartile deviation gives the average amount by which the two quartiles differ from the median. Quartile deviation is an absolute measure of dispersion. The relative measure corresponding to this measure, called the coefficient of quartile deviation, is calculated as follows.

$$\text{Coefficient of Q.D.} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Coefficient of quartile deviation can be used to compare the degree of variation in different distributions.

Computation of Quartile Deviation

The process of computing quartile deviation is very simple, we have just to compute the values of the upper and lower quartiles.

Where Q_1 = size of $\left(\frac{N+1}{4}\right)^{\text{th}}$ item

and Q_3 = size of $\left(\frac{3(N+1)}{4}\right)^{\text{th}}$ item

when the items are in ascending order of magnitude.

Individual Observations

Example 1 : Find out the value of quartile deviation and its coefficient from the following data :

Roll No :	1	2	3	4	5	6	7
Marks :	20	28	40	12	30	15	50

Solution

Marks arranged in ascending order :

12 15 20 28 30 40 50

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{size of } \left(\frac{7+1}{4} \right) = 2^{\text{nd}} \text{ item}$$

Size of 2nd item is 15. $\therefore Q_1 = 15$.

$$Q_3 = \text{Size of } 3 \left(\frac{N+1}{4} \right) \text{ th item}$$

$$= \text{Size of } \frac{3 \times 8}{4} = 6^{\text{th}} \text{ item}$$

Size of 6th item is 40. $\therefore Q_3 = 40$.

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{40 - 15}{2} = 12.5$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40 - 15}{40 + 15} = \frac{25}{55} = 0.455$$

Discrete Series

Example 2 : Compute coefficient of quartile deviation from the following data:

Marks	10	20	30	40	50	80
No. of students	4	7	15	8	7	2

Solution

Marks	Frequency	c.f.
10	4	4
20	7	11
30	15	26
40	8	34
50	7	41
80	2	43

$$Q_1 = \text{Size of } \frac{N + 1}{4} \text{ th item}$$

$$= \text{Size of } \frac{43 + 1}{4} \text{ th item} = 11^{\text{th}} \text{ item}$$

Size of 11th item is 20. $\therefore Q_1 = 20$.

$$Q_3 = \text{Size of } 3 \left(\frac{N + 1}{4} \right) \text{th item}$$

$$= \text{Size of } 3 \left(\frac{43 + 1}{4} \right) \text{th item} = 33\text{rd item}$$

Size of 33rd item is 40.

$$\therefore Q_3 = 40$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{40 - 20}{2} = 10$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40 - 20}{40 + 20}$$

$$= \frac{20}{60} = 0.333$$

Example 1

Continuous Series

Example 3 : Compute Quartile Deviation and its coefficient from the following data :

x :	10-20	20-30	30-40	40-60	60-70	70-80
f :	12	19	5	10	9	6

Solution

x	f	c.f.
10-20	12	12
20-30	19	31
30-40	5	36
40-60	10	46
60-70	9	55
70-80	6	61

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$Q_1 = \text{Size of } \frac{N}{4} \text{ th item} = \frac{61}{4} = 15.25 \text{th item}$$

Q_1 lies in the class 20-30

$$Q_1 = L + \frac{N/4 - \text{c.f.}}{f} \times i$$

$$L = 20, \quad \frac{N}{4} = 15.25, \text{ c.f.} = 12, f = 19, i = 10$$

$$\therefore Q_1 = 20 + \frac{15.25 - 12}{19} \times 10 = 20 + 1.71 = 21.71$$

$$Q_3 = \text{Size of } \frac{3N}{4} \text{ th item} = \frac{3 \times 61}{4} = 45.75 \text{th item}$$

Q_3 lies in the class 40 - 60

$$Q_3 = L + \frac{3N/4 - \text{c.f.}}{f} \times i$$

$$L = 40, \quad \frac{3N}{4} = 45.75, \text{ c.f.} = 36, f = 10, i = 20$$

$$\therefore Q_3 = 40 + \frac{45.75 - 36}{10} \times 20 = 40 + 19.5 = 59.5$$

$$\text{Q.D.} = \frac{59.5 - 21.71}{2} = \frac{37.79}{2} = 18.895$$

$$\begin{aligned} \text{Co-efficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{59.5 - 21.71}{59.5 + 21.71} \\ &= \frac{37.79}{81.21} = 0.465 \end{aligned}$$

Deciles and Percentiles

We know that a given distribution is divided into four equal parts by Q_1 , M and Q_3 . In the same way, given series may be divided into ten or hundred equal parts. The nine points which divide the distribution into ten equal parts are called the deciles and they are denoted by $D_1, D_2, \dots, D_9$. The ninety nine points which divide the distribution into 100 equal parts are

called the percentiles and they are denoted by $P_1, P_2, \dots, P_{99}$. In other words D_i is that value

of the variable such that $\left(\frac{i}{10}\right)$ th of the observations lie below it and P_j is that value of the variable such that $\left(\frac{j}{100}\right)$ th of the observations lie below it.

In calculation,

Example 2: Compute D_4 , P_{11} for the following data

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	2	21	23	10	10

$$D_i = \text{size of } \left[i \left(\frac{(N+1)}{10} \right) \right] \text{th item for ungrouped data and discrete series.}$$

$$P_j = \text{size of } \left[j \left(\frac{(N+1)}{100} \right) \right] \text{th item for ungrouped data and discrete series}$$

$$D_i = \frac{L + i \left(\frac{N}{10} \right) - \text{c.f.}}{f} \times i \text{ for continuous series}$$

$$P_j = \frac{L + j \left(\frac{N}{100} \right) - \text{c.f.}}{f} \times i \text{ for continuous series}$$

Example 1

Calculate D_3 , D_7 , P_9 , P_{34} for the following data.

27, 35, 48, 50, 60, 63, 65, 80, 86, 90, 92, 93, 94, 96, 98

Solution

The numbers are in ascending order of magnitude.

$$N = 15$$

$$D_3 = \text{size of } 3 \left(\frac{15+1}{10} \right) \text{th item}$$

$$= \text{size of } 3 \left(\frac{15+1}{10} \right) \text{th item}$$

$$= \text{size of 4.8th item}$$

$$= \text{size of 4th item} + 0.8 (\text{size of 5th item} - \text{size of 4th item})$$

$$= 50 + 0.8 (60 - 50)$$

$$= 50 + 8 = 58$$

$$D_7 = \text{Size of } 7 \left(\frac{N+1}{10} \right) \text{th item}$$

$$= \text{size of } 7 \left(\frac{15+1}{10} \right) \text{th item}$$

$$= \text{size of 11.2th item}$$

$$= \text{size of 11th item} + 0.2 (\text{size of 12th item} - \text{size of 11th item})$$

$$= 92 + 0.2 (93 - 92)$$

$$= 92.2$$

$$P_9 = \text{size of } 9 \left(\frac{N+1}{100} \right) \text{th item}$$

$$= \text{size of } 9 \left(\frac{16}{100} \right) \text{th item}$$

$$= \text{size of 1.44th item}$$

$$= \text{size of first item} + 0.44 (\text{size of 2nd item} - \text{size of 1st item})$$

$$= 27 + 0.44 (35 - 27)$$

$$= 27 + 0.44(8)$$

$$= 27 + 3.52$$

$$= 30.52$$

$$P_{34} = \text{size of } 34 \left(\frac{N+1}{100} \right) \text{th item}$$

$$= \text{size of } \left(\frac{34 \times 16}{100} \right) \text{th item}$$

$$= \text{size of 5.44th item}$$

$$= \text{size of 5th item} + 0.44 (\text{size of 6th item} - \text{size of 5th item})$$

$$= 60 + 0.44 (63 - 60)$$

$$= 60 + 1.32$$

$$= 61.32$$

Example 2 : Compute D_4 , P_{71} for the following data

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	2	21	23	34	10

Solution

class	f	c.f
0-10	2	2
10-20	21	23
20-30	23	46
30-40	34	80
40-50	10	90

1. To compute D_4

$$N = 90$$

$$\frac{4N}{10} = \frac{4 \times 90}{10} = 36$$

4th Decile class is 20 - 30

$$D_4 = L + \frac{\frac{4N}{10} - c.f}{f} \times i$$

$$L = 20, \frac{4N}{10} = 36, c.f = 23, f = 23, i = 10$$

$$\therefore D_4 = 20 + \frac{36 - 23}{23} \times 10$$

$$= 20 + \frac{13}{23} \times 10 = 20 + 5.65 = 25.65$$

2. To compute P_{71}

$$N = 90$$

$$\frac{71N}{100} = \frac{71 \times 90}{100} = 63.9$$

71st Percentile class is 30 - 40

$$P_{71} = L + \frac{\frac{71N}{100} - c.f}{f} \times i$$

$$L = 30, \quad \frac{71N}{100} = 63.9, \quad c.f = 46, \quad f = 34, \quad i = 10$$

$$p_{71} = 30 + \frac{63.9 - 46}{34} \times 10$$

$$= 30 + \frac{17.9}{34} \times 10 = 30 + \frac{179}{34} = 30 + 5.26 = 35.26$$

Exercise

1. Calculate the range for the following data :

Rs. 4000, 4200, 4400, 4600, 4800

2. Find the Q.D. of the following ungrouped data :

40, 95, 72, 80, 50

3. Calculate the range and Q.D. for the following data :

x	10	15	40	50	55
f	6	10	18	7	4

4. Determine the range and Q.D. for the following distribution

x	80-89	90-99	100-109	110-119	120-129	130-139	140-149	150-159
f	15	44	133	150	125	82	35	16

5. Find 3rd, 8th, 9th deciles and 11th, 19th, 27th and 71st percentiles

Class	140-150	150-160	160-170	170-180	180-190	190-200
f	4	6	10	18	9	3

1.8 Mean Deviation

The range and the Q.D. give only the general idea about the scatter, ie., how far the items are spread. To study the formation of a distribution we should take the deviations from an average. The average deviation and the standard deviation help us in achieving this goal.

The mean deviation is also known as the average deviation. It is the average difference between the items in a distribution and the median or mean of that series.

Computation of Mean Deviation - Individual Observations

If $x_1, x_2, x_3, \dots, x_N$ are N given observations then the deviation about an average A is given by

$$\begin{aligned} \text{M.D.} &= \frac{1}{N} \sum |x - A| \\ &= \frac{1}{N} \sum |D| \quad \text{or} \quad \frac{\sum |D|}{N} \end{aligned}$$

Where $|D| = |x - A|$ is the modulus value or absolute value of the deviation ignoring plus and minus signs.

The relative measure corresponding to the mean deviation, called the coefficient of mean deviation is obtained by dividing mean deviation by the particular average used in computing mean deviation. Thus if mean deviation has been computed from median, coefficient of

$$\text{M.D.} = \frac{\text{M.D.}}{\text{Median}}$$

If mean has been used while calculating the value of M.D., coefficient of M.D. = $\frac{\text{M.D.}}{\text{Mean}}$

Example 1 : Calculate the mean deviation and its coefficient of the income group of five members given below.

(Rs) : 4000 4200 4400 4600 1800

Solution

$$\begin{aligned} \text{Median} &= \text{Size of } \frac{N+1}{2} \text{ th item} \\ &= \frac{5+1}{2} = 3\text{rd item} \end{aligned}$$

Size of 3rd item is 4400.

Income	Deviation from median 4400 D
4000	400
4200	200
4400	0
4600	200
4800	400
N = 5	$\Sigma D = 1200$

$$\begin{aligned}\text{Mean Deviation} &= \frac{\Sigma | D |}{N} \\ &= \frac{1200}{5} = 240\end{aligned}$$

This means that the average deviation of the individual incomes from the median income is Rs. 240.

$$\begin{aligned}\text{Coefficient of M.D.} &= \frac{\text{M.D.}}{\text{Median}} \\ &= \frac{240}{4400} = 0.054\end{aligned}$$

Calculation of Mean Deviation - Discrete Series

Steps : (i) Calculate the median of the series.

- (ii) Take the deviations of the items from median ignoring signs and denote them by | D |.
- (iii) Multiply these deviations by the respective frequencies and obtain the total $\Sigma f | D |$.
- (iv) Divide the total obtained in step (ii) by the number of observations. This gives us the value of mean deviation.

The formula for calculating mean deviation is

$$\text{M.D.} = \frac{\Sigma f | D |}{N}$$

Example 2

Calculate the mean deviation from the following series :

x	10	11	12	13	14
f	3	12	18	12	3

Solution

Median = Size of $\frac{N+1}{2}$ th item

$$= \frac{48+1}{2} = 24.5\text{th item}$$

Size of 24.5th item is 12 ($\because$ c.f = 33 when f = 18).

$\therefore$ Median = 12.

x	f	D	f D	c.f.
10	3	2	6	3
11	12	1	12	15
12	18	0	0	33
13	12	1	12	45
14	3	2	6	48

$$N = 48$$

$$\Sigma f|D| = 36$$

$$\text{M.D.} = \frac{\Sigma f|D|}{N}$$

$$= \frac{36}{48} = 0.75$$

Example 3

Calculate the mean deviation from the mean for the following data :

Size	2	4	6	8	10	12	14	16
Frequency	2	2	4	5	3	2	1	1

Solution

x	f	fx	$ x - 8 $ $ D $	f D
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8

$$N = 20$$

$$\Sigma fx = 160$$

$$\Sigma f (|D|) = 56$$

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{160}{20} = 8$$

$$M.D. = \frac{\Sigma f |D|}{N} = \frac{56}{20} = 2.8$$

Calculation of Mean Deviation - Continuous Series

For calculating mean deviation in continuous series the procedure remains the same as discussed above. The only difference is that here we have to obtain the mid-point of the various classes and take deviations of these points from median. The formula is same ie.

$$M.D. = \frac{\Sigma f |D|}{N}$$

Example 4 : Find the median and mean deviation of the following data :

Size	Frequency
0-10	7
10-20	12
20-30	18
30-40	25
40-50	16
50-60	14
60-70	8

Solution

Size	f	c.f.	m	m - 35.2	f D
0-10	7	7	5	30.2	211.4
10-20	12	19	15	20.2	242.4
20-30	18	37	25	10.2	183.6
30-40	25	62	35	0.2	5.0
40-50	16	78	45	9.8	156.8
50-60	14	92	55	19.8	277.2
60-70	8	100	65	29.8	238.4
N = 100				$\Sigma f D = 1314.8$	

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item} = \frac{100}{2} = 50\text{th item}$$

Median lies in the class 30 - 40

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

$$L = 30, \frac{N}{2} = 50, \text{c.f.} = 37, f = 25, i = 10$$

$$\therefore \text{Median} = 30 + \frac{50-37}{25} \times 10 = 30 + 5.2 = 35.2$$

$$\text{M.D.} = \frac{\Sigma f |D|}{N} = \frac{1314}{100} = 13.148$$

Example 5 : Calculate the mean deviation and its coefficient from the following data :

Class	Frequency
0-10	5
10-20	8
20-30	12
30-40	15
40-50	20
50-60	14
60-70	12
70-80	6

Solution

Class	Frequency	c.f	m	m - 43	f D
0-10	5	5	5	38	190
10-20	8	13	15	28	224
20-30	12	25	25	18	216
30-40	15	40	35	8	120
40-50	20	60	45	2	40
50-60	14	74	55	12	168
60-70	12	86	65	22	264
70-80	6	92	75	32	192

$$N = 92$$

$$\Sigma f |D| = 1414$$

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ th item} = \frac{92}{2} = 46\text{th item}$$

Median lies in the class 40-50

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

$$L = 40, \frac{N}{2} = 46, \text{c.f} = 40, f = 20, h = 10$$

$$\therefore \text{Median} = 40 + \frac{46-40}{20} \times 10 = 40+3 = 43$$

$$\text{M.D.} = \frac{\Sigma f |D|}{N} = \frac{1414}{92} = 15.37$$

$$\text{Coefficient of M.D.} = \frac{\text{M.D.}}{\text{Median}} = \frac{15.37}{43} = 0.357$$

1.9 The Standard Deviation (S.D.)

The standard deviation concept was introduced by Karl Pearson in 1823. It is by far the most important and widely used measure of studying dispersion. Standard deviation is known as root mean square deviation for the reason that it is the square root of the mean of the squared deviations from the arithmetic mean. It is denoted by the small Greek letter σ (read as sigma).

The standard deviation measures the absolute dispersion. If we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.

Difference between Mean Deviation and Standard deviation

- (i) Algebraic signs are ignored while calculating mean deviation whereas in the calculation of standard deviation signs are taken into account.
- (ii) Mean deviation can be computed either from median or mean. The standard deviation is always computed from the arithmetic mean because the sum of the squares of the deviation of items from arithmetic mean is the least.

Calculation of S.D. - Individual Observations

There are two methods to calculate S.D. by individual observations.

Method 1 : Deviations taken from Actual Mean

Steps : (i) Calculate the actual mean of the series i.e., $\bar{x}$.

- (ii) Take the deviations of the items from the mean, i.e., find $(x - \bar{x})$. Denote these deviations by x .
- (iii) Square these deviations and obtain the total Σx^2 .
- (iv) Divide Σx^2 by the total number of observations i.e., N and extract the square-root. This gives us the value of standard deviation. x^2

In symbol, $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ where $x = x - \bar{x}$.

Example 1 : Calculate S.D. for the following items

30, 90, 20, 10, 80, 70

Solution :

$$\bar{x} = \frac{30 + 90 + 20 + 10 + 80 + 70}{6}$$
$$= \frac{300}{6} = 50$$

x	$x - \bar{x}$ ie., $x - 50 = x$	x^2
30	-20	400
90	40	1600
20	-30	900
10	-40	1600
80	30	900
70	20	400
		$x^2 = 5800$

$$\sigma^2 = \frac{5800}{6} = \frac{2900}{3}$$

$$\sigma = \sqrt{\frac{2900}{3}} = \sqrt{966.7} = 31.1$$

Method 2 : Deviations taken from Assumed Mean

Sometimes the actual mean is in fractions. In such case the deviations be taken from an assumed mean and the necessary adjustment made in the value of the standard deviation.

Steps : (i) Take the deviations of the items from an assumed mean ie., obtain $(x - A)$. Denote these deviations by d . Take the total of these deviations ie., obtain Σd .

(ii) Square these deviations and obtain the total Σd^2

(iii) Substitute the values of Σd^2 , Σd and N in the formula

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

Example 2 : Blood serum cholesterol levels of 10 persons are as under :

240, 260, 290, 245, 255, 288, 272, 263, 277, 251

Calculate standard deviation with the help of assumed mean.

Solution : Here $\Sigma x = 2641$ $\therefore \bar{x} = 264.1$ $\therefore$ We shall take A.M. = 264

x	$(x - 264)$ d	d^2
240	-24	576
260	-4	16
290	26	676
245	-19	361
255	-9	81
288	24	576
272	8	64
263	-1	1
277	13	169
251	-13	169
$\Sigma x = 2641$	$\Sigma d = 1$	$\Sigma d^2 = 2689$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$$\Sigma d^2 = 2689, \Sigma d = 1, N = 10$$

$$\sigma = \sqrt{\frac{2689}{10} - \left(\frac{1}{10}\right)^2} = 268.9 - 0.01$$

$$\sigma = 16.398$$

Calculation of Standard Deviation - Discrete Series

Steps : (i) Take the deviations of the items from an assumed mean and denote these deviations by d .

- (ii) Multiply these deviations by the respective frequencies and obtain the total Σfd
- (iii) Obtain the squares of the deviations, ie., calculate d^2
- (iv) Multiply the squared deviations by the respective frequencies and obtain the total Σfd^2 .
- (v) Substitute the values in the formula

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \text{ where } d = x - A$$

Example 3 : Calculate the S.D. from the data given below :

Size of item	Frequency
3.5	3
4.5	7
5.5	22
6.5	60
7.5	85
8.5	32
9.5	8

Solution

x	f	(x - 6.5) d	fd	fd ²
3.5	3	-3	-9	27
4.5	7	-2	-14	28
5.5	22	-1	-22	22
6.5	60	0	0	0
7.5	85	1	85	85
8.5	32	2	64	128
9.5	8	3	24	78
N = 217			Σ fd = 128	Σ fd ² = 362

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$\Sigma fd^2 = 362, \quad \Sigma fd = 128, \quad N = 217.$$

$$\sigma = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = \sqrt{1.668 - 0.348} = 1.149$$

Calculation of S.D. - Continuous Series

Steps : (i) Find the mid points of various classes.

(ii) Take the deviations of these mid-points from an assumed mean and denote these deviations divide by the class interval by d.

(iii) Multiply the frequencies of each class with these deviations and obtain Σfd .

(iv) Square the deviations and multiply them with the respective frequencies of each class and obtain Σfd^2 .

Example 4 : Calculate mean and S.D. of the following frequency distribution of marks :

Marks	No. of students
0-10	5
10-20	12
20-30	30
30-40	45
40-50	50
50-60	37
60-70	21

Solution

Let us take assumed mean = 35

Marks	m.p m	f	(m-35)/10 d	fd	fd ²
0-10	5	5	-3	-15	45
10-20	15	12	-2	-24	48
20-30	25	30	-1	-30	30
30-40	35	45	0	0	0
40-50	45	50	1	50	50
50-60	55	37	2	74	148
60-70	65	21	3	63	189
N = 200			$\Sigma fd = 118$		$\Sigma fd^2 = 510$

$$\bar{x} = A + \frac{\sum fd}{N} \times i = 35 + \frac{118}{200} \times 10$$

$$= 35 + 5.9 = 40.9$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2 \times 10} = \sqrt{\frac{510}{200} - \left(\frac{118}{200}\right)^2 \times 10}$$

$$= \sqrt{2.55 - 0.3481} \times 10 = 1.4839 \times 10 = 14.839$$

Coefficient of variation (C.V.)

The relative measure of dispersion corresponding to standard deviation called the Coefficient of Variation (C.V.) is defined as

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100$$

Coefficient of Variation has much practical utility. When two groups of data are given the group of data for which C.V. is less, is considered less variable, more stable, more uniform or more consistent.

Example 5

The means and standard deviations of the marks obtained by two students x and y are given below. Which of the two candidates is more consistent in his performance?

Solution

	x	y
Mean	72	78
S.D.	8	6

$$\text{C.V. for x} = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{8}{72} \times 100 = 11.1\%$$

$$\text{C.V. for y} = \frac{6}{78} \times 100 = 7.7\%$$

$\therefore$ y is more consistent in his performance.

Example 6 : Find the S.D. and C.V. for the following distribution.

Class	Freq
0-19	12
20-39	28
40-59	52
60-79	6
80-99	2

Solution

Let the assumed mean = 49.5

m	f	(x - 49.5) / 20 d	fd	fd ²
9.5	12	-2	-24	48
29.5	28	-1	-28	28
49.5	52	0	0	0
69.5	6	1	6	6
89.5	2	2	4	8
N = 100			$\Sigma fd = -42$	$\Sigma fd^2 = 90$

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{N} \times i \\ &= 49.5 + \left(\frac{-42}{100} \right) \times 20 \\ &= 49.5 - 8.4 = 41.1\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2 \times i} \\ &= \sqrt{\frac{90}{100} - \left(\frac{-42}{100} \right)^2 \times 20} \\ &= \sqrt{0.9 - 0.1764 \times 20} = \sqrt{0.7236} \times 20 = 17.01\end{aligned}$$

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{17.01}{41.1} \times 100 = \frac{1701}{41.1} = 41.4\%$$

Properties of Standard Deviation

1. If σ_1 and σ_2 are the standard deviations of two groups of sizes N_1 and N_2 respectively, then the S.D. of the combined group is given by

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

where σ_{12} = combined standard deviation

σ_1 = S.D. of first group

σ_2 = S.D. of second group

$d_1 = (\bar{x}_1 - \bar{x}_{12}); d_2 = (\bar{x}_2 - \bar{x}_{12})$

2. S.D. is greater than M.D. from mean
3. The relations between S.D., M.D., and Q.D.

$$(i) \quad \text{Q.D.} = \frac{2}{3} \text{ S.D.}$$

$$(ii) \quad \text{M.D.} = \frac{4}{5} \text{ S.D.}$$

Example 7 : Find the mean and S.D. of the combined distribution of the following two distributions

	Group A
No. of items	100
Mean	50
S.D.	10

	Group B
No. of items	500
Mean	60
S.D.	11

Solution

$$N_1 = 100, \quad \bar{X}_1 = 50, \quad \sigma_1 = 10$$

$$N_2 = 500, \quad \bar{X}_2 = 60, \quad \sigma_2 = 11.$$

Let $\bar{x}_{12}$ and σ_{12} denote the mean and S.D. of the combined distribution.

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$= \frac{100 (50) + 500 (60)}{100 + 500}$$

$$= \frac{5000 + 30000}{600}$$

$$= \frac{35000}{600}$$

$$= \frac{175}{3}$$

$$= 58.3$$

$$d_1 = \bar{x}_1 - \bar{x}_{12}$$

$$= 50 - 58.3 = -8.3$$

$$d_2 = \bar{x}_2 - \bar{x}_{12}$$

$$= 60 - 58.3 = 1.7$$

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times 100 + 500 \times 121 + 100 (8.3)^2 + 500 (1.7)^2}{100 + 500}}$$

$$= \sqrt{\frac{10000 + 60500 + 100 (68.89) + 500(2.89)}{600}}$$

$$= \sqrt{\frac{78834}{600}}$$

$$= \sqrt{131.39} = 11.47$$

$$\bar{x}_{12} = 58.3; \quad \sigma_{12} = 11.47$$

Example 8 : Following are the marks obtained by two students A and B in 10 sets of examination :

Sets :	1	2	3	4	5	6	7	8	9	10
Marks obtained by A :	44	80	76	48	52	72	68	56	60	64
Marks obtained by B :	48	75	54	60	63	69	72	51	57	56

If the consistency of performance is the criterion for awarding the prize, who should get the prize?

Solution : For finding out consistency we have to compare the coefficient of variation of A and B.

x	(x - 62) x	x ²	y	(y - 60.5) y	y ²
44	-18	324	48	-12.5	156.25
80	18	324	75	14.5	210.25
76	14	196	54	-6.5	42.25
48	-14	196	60	-0.5	0.25
52	-10	100	63	2.5	6.25
72	10	100	69	8.5	72.25
68	6	36	72	11.5	132.25
56	-6	36	51	-9.5	90.25
60	-2	4	57	-3.5	12.25
64	2	4	56	-4.5	20.25
$\Sigma x = 620$	$\Sigma x = 0$	$\Sigma x^2 = 1320$	$\Sigma y = 605$	$y = 0$	$\Sigma y^2 = 742.5$

Coefficient of variation A :

$$\bar{x} = \frac{\Sigma x}{N} = \frac{620}{10} = 62$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{1320}{10}} = \sqrt{132} = 11.489$$

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{11.489}{62} \times 100 = 18.53$$

Coefficient of variation B :

$$\bar{y} = \frac{\Sigma y}{N} = \frac{605}{10} = 60.5$$

$$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{742.5}{10}} = 8.62$$

$$\text{C.V.} = \frac{\sigma}{\bar{y}} \times 100 = \frac{8.62}{60.5} \times 100 = 14.25$$

Since coefficient of variation is less in case of student B as compared to A, B should get the prize.

Example 9 : From the data given below state which series is more consistent :

Variable	Series A	Series B
10-20	20	13
20-30	18	22
30-40	32	40
40-50	40	32
50-60	22	18
60-70	18	10

Solution : In order to find out which series is more consistent, we shall compare the coefficient of variation.

Series A

Variable	m	f	(m-45)/10 d	fd	fd ²
10-20	15	10	-3	-30	90
20-30	25	18	-2	-36	72
30-40	35	32	-1	-32	32
40-50	45	40	0	0	0
50-60	55	22	1	22	22
60-70	65	18	2	36	72
N = 140			$\Sigma fd = -40$		$\Sigma fd^2 = 288$

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{N} \times i \\ &= 45 - \frac{40}{140} \times 10 = 45 - 2.86 = 42.14\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times 10 \\ &= \sqrt{\frac{288}{140} - \left(\frac{-40}{140}\right)^2} \times 10 \\ &= \sqrt{2.057 - 0.082} \times 10\end{aligned}$$

$$= \sqrt{1.975} \times 10 = 1.405 \times 10 = 14.05$$

$$C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{14.05}{42.14} \times 100 = 33.34 \text{ percent}$$

Series B

Variable	m	f	(m-45)/10 d	fd	fd ²
10-20	15	18	-3	-54	162
20-30	25	22	-2	-44	88
30-40	35	40	-1	-40	40
40-50	45	32	0	0	0
50-60	55	18	1	18	18
60-70	65	10	2	20	40
N = 140			$\Sigma fd = -100$		$\Sigma fd^2 = 348$

$$\begin{aligned}\bar{x} &= A + \frac{\Sigma fd}{N} \times i \\ &= 45 - \frac{100}{140} \times 10 = 45 - 7.14 = 37.86\end{aligned}$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times 10$$

$$= \sqrt{\frac{348}{140} - \left(\frac{-100^2}{140}\right)} \times 10$$

$$= \sqrt{2.486 - 0.510} \times 10$$

$$= \sqrt{1.976} \times 10 = 1.406 \times 10 = 14.06$$

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{14.06}{37.86} \times 100 = 37.14 \text{ percent}$$

Since coefficient of variation is less for series A, series A is more consistent.

Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in their original series. Variance is nothing but the square of the standard deviation. ie., variance = σ^2

$$\text{or } \sigma = \sqrt{\text{variance}}$$

Example 10 : After calculation of the mean and S.D. of ten observations, a copying mistake was detected. When copying the values, 236 was written for 326 through oversight. If the incorrect mean and S.D. calculated are 260.1 and 5.6 respectively, find the correct values of the mean and S.D.

Solution : Let the correct values of mean and S.D. be $\bar{x}$, σ and calculated incorrect values be $\bar{x}_1$, σ_1

$$\therefore \bar{x}_1 = 260.1, \sigma_1 = 5.6$$

Let $x_{10} = 326$ (correct value) and

$$x_{10}^1 = 236 \text{ (incorrect value)}$$

$$\therefore \bar{x} = \frac{x_1 + x_2 + \dots + x_{10}}{10}$$

$$\bar{x}_1 = \frac{x_1 + x_2 + \dots + x_{10}^1}{10}$$

$$\bar{x} - \bar{x}_1 = \frac{x_{10} - x_{10}^1}{10} = \frac{326 - 236}{10} = \frac{90}{10} = 9$$

$$\therefore \bar{x} = \bar{x}_1 + 9$$

$$= 260.1 + 9 = 269.1$$

$$\sigma^2 = \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N} \right)^2$$

$$= \frac{x_1^2 + x_2^2 + \dots + x_{10}^2}{10} - (\bar{x})^2$$

$$\sigma_1^2 = \frac{x_1^2 + x_2^2 + \dots + x_{10}^2}{10} - (\bar{x}_1)^2$$

$$\sigma^2 - \sigma_1^2 = \frac{x_{10}^2 - x_1^2}{10} - [\bar{x}^2 - \bar{x}_1^2]$$

$$\sigma^2 - \sigma_1^2 = \frac{(326)^2 - (236)^2}{10} - [(269.1)^2 - (260.1)^2]$$

$$= \frac{(326 + 236)(326 - 236)}{10} - [529.2 \times 9]$$

$$= (562 \times 9) - (529.2 \times 9)$$

$$= (562 - 529.2) \times 9$$

$$= 32.8 \times 9$$

$$= 295.2$$

$$\sigma^2 = \sigma_1^2 + 295.2$$

$$= (5.6)^2 + 295.2$$

$$= 31.36 + 295.2$$

$$\sigma^2 = 326.56$$

$$\therefore \sigma = \sqrt{326.56} = 18.07$$

Comparison of various measures of dispersion

Range

Merits

- 1) Range is easily calculated
- 2) It is easy to understand

Demerits

- 1) It is not based on each and every item of the distribution.
- 2) It is easily affected by fluctuations of sampling.
- 3) It does not give any idea about the items between the two extremes.
- 4) It cannot be computed in the case of open end distributions.
- 5) It fails to consider the central tendency of the data.

Uses

- 1) It is useful in studying the variations in prices of stocks and shares and other commodities. By computing the range, we get an idea about the variation in the price level.
- 2) In quality control, the range is used to study the variation in the quality of the units manufactured.
- 3) A doctor is interested in knowing the range of the varying temperatures of his patients.
- 4) The information about the minimum and the maximum temperatures and their difference is essential for the public.

Quartile Deviation

Merits

- 1) It is easily understood
- 2) It is easily calculated
- 3) Though it is not based on all values of the items, all the items play an indirect role in its determination.
- 4) The extreme values do not affect its values and so, it can be calculated in case of open end distribution.

Demerits

- 1) It is not capable of further algebraic treatment.
- 2) It is affected by the fluctuations of sampling and hence not stable.

Uses

It is used in calculating skewness and kurtosis of a frequency distribution.

Mean Deviation

Merits

- 1) Unlike Q.D. and range, it is a measure based on the values of all the items.
- 2) It is easily understood.

Demerits

- 1) It is not suitable for further algebraic treatment.
- 2) It is less stable than the S.D.

Uses

It is mostly used only in economic statistics. It is greatly used in situation where the degrees of dispersion of a distribution is to be shown to the general public.

The National Bureau of Economic Research of the U.S.A. has found it to be the most practical measure of dispersion that can be used in its work of forecasting cycles.

Standard Deviation

Merits

- 1) It is based on all the items.
- 2) It is rigidly defined.
- 3) It lends itself for further algebraic treatment.
- 4) It is not affected much by fluctuations of sampling.

Demerits

- 1) It is not easy to understand.
- 2) It is difficult to calculate.
- 3) It gives more weight to extreme items and less weight to items which are nearer to the mean since we are squaring the deviations.

Uses

- 1) It is used in further statistical work, for example in computing skewness, correlation etc.
- 2) It is of greatest value in testing the reliability of measures computed from samples.
- 3) It is considered to be the best measure of dispersion and hence it is very commonly used.

Choice of a Measure of Dispersion

Range can be used only as a rough and ready measure of spread or when the data are incomplete or when the variations in the values of the items are very small.

Quartile Deviation is a better measure than the range sine it is not affected by the values of the extreme items. It is used only when M.D. or S.D. cannot be easily calculated.

Mean deviation has an advantage over S.D. as it can be easily calculated and easily understandable. The economists and businessmen prefer this measure because of the ease in calculation.

Standard deviation is the most useful among all the measures of dispersion as it possesses most of the properties required by an ideal measure. For all practical purposes requiring greater accuracy and precision, S.D. should be preferred.

To conclude, it can be stated that if statistical reliability is the prime factor to be considered, the order of preference is S.D., M.D., Q.D and the range. But in case quickness alone is to be considered the order is just the reverse.

Exercise

1. Calculate the mean deviation about mean, median and S.D. for the following sets of values.
 - (a) 16, 3, 15, 5, 14, 13
 - (b) 20, 28, 40, 70, 76, 60, 58
2. Calculate the mean deviation about mean and median and the mean coefficient of dispersion for the following data

Class	80-89	90-99	100-109	110-119	120-129
Freq	7	13	18	7	5

3. The following are the scores of two batsmen A and B in ten innings

A	12	15	6	73	7	19	19	36	84	29
B	47	12	76	42	4	51	37	48	13	0

Who is a better run scorer? Who is the more consistent player?

4. Give the characteristics of an ideal measure of dispersion.

5. Calculate the S.D. for the following distribution

Ht in cms	140-145	145-150	150-155	155-160	160-165
No. of boys	1	18	50	142	219

6. The mean height and S.D. of a sample of 400 individuals are found to be 165 cms and 5.8 cms respectively. In another sample of 600 individuals the mean height is found to be 165.8 cms and S.D. is 6.1 cms. If the two samples are combined, find the mean height and the S.D. of the combined distribution.
7. The mean and S.D. based on 60 observations were calculated as 276 mm and 22 mm respectively. But on subsequent scrutiny it was found that through oversight one observation was recorded wrongly as 372 mm although its correct value was 327 mm. Find the corrected values of the mean and S.D.

Exercise

1. Calculate the mean deviation about mean, median and S.D. for the following sets of values.
- (a) 16, 3, 15, 5, 14, 13
- (b) 20, 28, 40, 70, 76, 60, 58
2. Calculate the mean deviation about mean and median and the mean coefficient of dispersion for the following data

Class	80-89	90-99	100-109	110-119	120-129
Freq	7	13	18	7	5

3. The following are the scores of two batsmen A and B in ten innings

	A	12	15	8	73	7	19	19	18	18
B	47	12	76	42	4	51	37	48	13	0

4. Give the characteristics of an ideal measure of dispersion.
- Who is a better run scorer? Who is the more consistent player?

UNIT - II

2.1 Moments

Definition : If $x_1, x_2, \dots, x_n$ are the values in a set and A , any number then the r^{th} moment about A or the r^{th} raw moment, written as μ_r' is defined as the quantity

$$\mu_r' = \frac{1}{N} \sum_{i=1}^N (x_i - A)^r \text{ for } r = 0, 1, 2, 3, \dots$$

When $A = \bar{x}$, the arithmetic mean, it is called the r^{th} central moment and is written as μ_r .

$$\text{So, } \mu_r = \frac{1}{N} \sum_{i=1}^N f_i (x_i - \bar{x})^r$$

When the given data are represented as frequency distribution we have the modified formula

$$\mu_r' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - \bar{A})^r$$

$$\text{and } \mu_r = \frac{1}{N} \sum_{i=1}^N f_i (x_i - \bar{x})^r \text{ where } f_i \text{ are the frequencies, } N = \sum_{i=1}^N f_i \text{ and } x_i \text{ are the}$$

values of the different items in the case of discrete series and midvalues of the different classes in the case of continuous series.

Let us write the formula for first few moments.

$$\mu_0' = \frac{1}{N} \sum_{i=1}^N (x_i - A)^0$$

$$= \frac{1}{N} (N) = 1$$

Similarly, $\mu_0 = 1$

$$\mu_1' = \frac{1}{N} \sum_{i=1}^N (x_i - A)$$

$$= \frac{1}{N} \left[\sum_{i=1}^N x_i - \sum_{i=1}^N A \right]$$

$$= \frac{1}{N} [N\bar{x} - AN] \quad \left(\because \frac{x_1 + x_2 + \dots + x_N}{N} \right) = \bar{x}$$

$$\mu_1' = \bar{x} - A$$

$$= d$$

$$\mu_1 = \frac{1}{N} \sum_1^N (x_i - \bar{x}) = \frac{1}{N} \left(\sum_1^N x_i - \sum_1^N \bar{x} \right)$$

$$= \frac{1}{N} (N\bar{x} - N\bar{x})$$

$$= 0$$

$$\mu_1' = \frac{1}{N} (x_i - A)^2 = \text{Mean square deviation from } A$$

$$\mu_2 = \frac{1}{N} \sum (x_i - \bar{x})^2 = \sigma^2 - \text{The variance}$$

Example 1: Find the first, second and third moments about $A = 80$ of the following data.
58,62,48,44,60,50,55,45,42,56,

Solution : Let us take $A = 50$

x_i	$(x_i - A)$ $x_i - 50$	$(x_i - A)^2$	$(x_i - A)^3$
58	8	64	512
62	12	144	1728
48	-2	4	-8
44	-6	36	-216
60	10	100	1000
50	0	0	0
55	5	25	125
45	-5	25	-125
42	-8	64	-512
56	6	36	216
$\sum (x_i - A) = 20$		$\sum (x_i - A)^2 = 498$	$\sum (x_i - A)^3 = 2720$

$$\mu_1' = \frac{1}{N} \sum_{i=1}^N (x_i - A)^1 = \frac{20}{10} = 2$$

$$\mu_2' = \frac{1}{N} \sum_{i=1}^N (x_i - A)^2 = \frac{498}{10} = 49.8$$

$$\mu_3' = \frac{1}{N} \sum_{i=1}^N (x_i - A)^3 = \frac{2720}{10} = 272$$

Example 2 : Calculate the first three central moments for the above data.

Solution

$$\bar{x} = \frac{58 + 62 + 48 + 44 + 60 + 50 + 55 + 45 + 42 + 56}{10} = \frac{520}{10} = 52$$

x	(x - $\bar{x}$) x - 52	(x - $\bar{x}$) ²	(x - $\bar{x}$) ³
58	6	36	216
62	10	100	1000
48	-4	16	-64
44	-8	64	-512
60	8	64	512
50	-2	4	-8
55	3	9	27
45	-7	49	-343
42	-10	100	-1000
56	4	16	64
$\Sigma (x - \bar{x}) = 0$		$\Sigma (x - \bar{x})^2 = 458$	$\Sigma (x - \bar{x})^3 = -108$

$$\mu_1 = \frac{1}{N} \Sigma (x - \bar{x}) = 0$$

$$\mu_2 = \frac{1}{N} \Sigma (x - \bar{x})^2 = \frac{458}{10} = 45.8$$

$$\mu_3 = \frac{1}{N} \Sigma (x - \bar{x})^3 = \frac{-108}{10} = -10.8$$

Relation between the Central Moments and Raw Moments

$$\mu_0' = \mu_0 = 1$$

$$\mu_1' = d = \bar{x} - A \quad \text{I}$$

$$\mu_1 = 0 \quad \text{II}$$

$$\mu_2 = \frac{1}{N} \sum_1^N (x - \bar{x})^2$$

$$= \frac{1}{N} \sum_1^N [(x - A) - (\bar{x} - A)]^2$$

$$= \frac{1}{N} \sum_1^N [(x - A)^2 - 2d(x - A) + d^2]$$

$$\text{using } d = \bar{x} - A$$

$$= \frac{1}{N} \sum_1^N (x - A)^2 - 2d \sum_1^N (x - A) + \frac{1}{N} (Nd^2)$$

$$= \mu_2' - 2d^2 + d^2 \quad \text{since } \sum_1^N (x - A) = N\bar{x} - NA = Nd$$

$$\mu_2 = \mu_2' - d^2 \quad \text{III}$$

$$\mu_3 = \frac{1}{N} \sum_1^N (x - \bar{x})^3$$

$$= \frac{1}{N} \sum_1^N [(x - A) - (\bar{x} - A)]^3$$

$$= \frac{1}{N} \sum_1^N [(x - A)^3 - 3d(x - A)^2 + 3d^2(x - A) - d^3]$$

$$\text{where } d = \bar{x} - A$$

$$= \frac{1}{N} \sum_1^N (x - A)^3 - 3d \frac{1}{N} \sum_1^N (x - A)^2 + 3d^2 \frac{1}{N} \sum_1^N (x - A) - \frac{1}{N} \sum_1^N d^3$$

$$= \mu_3' - 3d \mu_2' + 3d^2 (d) - d^3$$

$$= \mu_3' - 3d \mu_2' + 3d^3 - d^3$$

$$\mu_3 = \mu_3' - 3d \mu_2' + 2d^3 \quad \text{IV}$$

Let us prove the general formula :

$$\begin{aligned}
 \mu_r &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^r \\
 &= \frac{1}{N} \sum_{i=1}^N [(x_i - A) - (\bar{x} - A)]^r \\
 &= \frac{1}{N} \sum_{i=1}^N [(x_i - A)^r - rc_1 d (x_i - A)^{r-1} + rc_2 d^2 (x_i - A)^{r-2} - \dots + (-1)^r d^r] \\
 &= \frac{1}{N} \sum_{i=1}^N [(x_i - A)^r - rc_1 d \frac{1}{N} \sum_{i=1}^N (x_i - A)^{r-1} + rc_2 d^2 \frac{1}{N} \sum_{i=1}^N (x_i - A)^{r-2} - \dots + (-1)^r d^r] \\
 \therefore \mu_r &= \mu_r' - rc_1 d \mu_{r-1}' + rc_2 d^2 \mu_{r-2}' - \dots + (-1)^r d^r \mu_0'
 \end{aligned}$$

In particular

$$\begin{aligned}
 \mu_4 &= \mu_4' - 4d \mu_3' + 6d^2 \mu_2' - 4d^3 \mu_1' + d^4 \mu_0' \\
 &= \mu_4' - 4d \mu_3' + 6d^2 \mu_2' - 3d^3 (\because \mu_1' = d) \quad \text{VI}
 \end{aligned}$$

In the case of continuous series with uniform class intervals we could take the deviations of the various items in class units and thus simplify the calculation still further.

Let $d_i = \frac{x_i - A}{i}$ where A is any origin and i length of the class interval.

$$\therefore x_i = A + id_i$$

$$\mu_r' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - A)^r$$

$$= \frac{1}{N} \sum_{i=1}^N f_i (id_i)^r$$

$$\mu_r' = i^r \frac{1}{N} \sum_{i=1}^N f_i d_i^r \quad \text{where } d_i = \frac{x_i - A}{i} \quad \text{VII}$$

Example 1 : Calculate the first four central moments for the following distribution.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Solution : Let us take $A = 4$

x_i	f_i	$x_i - 4$	$f_i (x_i - 4)$	$f_i (x_i - 4)^2$	$f_i (x_i - 4)^3$	$f_i (x_i - 4)^4$
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
	256		0	512	0	2816

$$\mu_1' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - 4) = 0$$

$$\mu_2' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - 4)^2 = \frac{512}{256} = 2$$

$$\mu_3' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - 4)^3 = 0$$

$$\mu_4' = \frac{1}{N} \sum_{i=1}^N f_i (x_i - 4)^4 = \frac{2816}{256} = 11$$

$$\mu_1 = 0$$

$$\mu_2 = d = 0$$

$$\therefore \bar{x} - 4 = 0 \text{ ie., } \bar{x} = 4 = A$$

$\therefore$ The Central moment are $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 0$, $\mu_4 = 11$

Example 2 : Calculate the four central moments of the following frequency distribution.

Class	16-20	20-24	24-28	28-32	32-36	Total
f	5	10	24	18	3	60

Solution : Here the class intervals have equal length = 4. So, let us calculate the raw moments about some A using the formula (VII). Let us take A = 26.

x	f	$d_i = \frac{x_i - A}{i} = \frac{x_i - 26}{4}$	fd_i	fd_i^2	fd_i^3	fd_i^4
18	5	-2	-10	20	-8	80
22	10	-1	-10	10	-10	10
26	24	0	0	0	0	0
30	18	1	18	18	18	18
34	3	2	6	12	24	48
	60		4	60	-8	156

$$\mu'_r = i^r \cdot \frac{1}{N} \sum_{i=1}^n f_i d_i^r \text{ where } d_i = \frac{x_i - 26}{4} \text{ and } i = 4$$

$$\mu'_1 = 4 \cdot \frac{1}{60} \cdot 4 = \frac{4}{15} = 0.27$$

$$\mu'_2 = 4^2 \cdot \frac{1}{60} \cdot 60 = 16$$

$$\mu'_3 = 4^3 \cdot \frac{1}{60} \cdot (-8) = \frac{-128}{15} = -8.53$$

$$\mu'_4 = 4^4 \cdot \frac{1}{60} \cdot (156) = \frac{6656}{10} = 665.6$$

To find the central movement

$$\mu_1' = d = 0.27$$

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - d^2 = 16 - (0.27)^2$$

$$= 16 - 0.0729$$

$$= 15.9271$$

$$= 15.93$$

$$\mu_3 = \mu_3' - 3d \mu_2' + 2d^3$$

$$= -8.53 - 3(0.27)(16) + 2(0.27)^3$$

$$= -8.53 - 12.96 + 0.039366$$

$$= -21.49 + 0.04$$

$$= -21.45$$

$$\mu_4 = \mu_4' - 4d \mu_3' + 6d^2 \mu_2' - 3d^4$$

$$= 665.6 - 4(0.27)(-8.53) + 6(0.27)^2(16) - 3(0.27)^4$$

$$= 665.6 + 9.2124 + 6.9984 - 0.01594323$$

$$= 665.60 + 9.21 + 7.00 - 0.02$$

$$= 681.79$$

Example 3 : The first three moments of a distribution about the value 2 are 1, 16 and -40. Find the first three central moments.

Solution : $A = 2$, $\mu_1' = 1$, $\mu_2' = 16$, $\mu_3' = -40$

$$\mu_1 = 0 \text{ for all distributions}$$

$$\mu_1' = 1 = d \quad (= \bar{x} - 2)$$

$$\mu_2 = \mu_2' - d^2 = 16 - 1 = 15$$

$$\mu_3 = \mu_3' - 3d \mu_2' + 2d^3$$

$$= -40 - 3(16) + 2$$

$$= -40 - 48 + 2$$

$$= -86$$

Example 4 : The first four moments of a distribution about the value 5 are 2, 20, 40 and 250 respectively. Find the mean, variance, μ_3 and μ_4 .

Solution : $A = 5$, $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$, $\mu_4' = 250$

$$\mu_1' = 2 = d = \bar{x} - A$$

$$\therefore \bar{x} - 5 = 2$$

$$\bar{x} = 7$$

$$\mu_2 = \mu_2' - d^2$$

$$= 20 - 4 = 16$$

$$\text{ie } \sigma^2 = 16$$

$$\mu_3 = \mu_3' - 3d \mu_2' + 2d^3$$

$$= 40 - 3(2)(20) + 2(2^3)$$

$$= 40 - 120 + 6$$

$$= -64$$

$$\mu_4 = \mu_4' - 4d \mu_3' + 6d^2 \mu_2' - 3d^4$$

$$= 250 - 4(2)(40) + 6(2^2)(20) - 3(2^4)$$

$$= 250 - 320 + 480 - 48$$

$$\mu_4 = 362$$

Example 5 : The first three moments about the origin are given by

$$\mu_1' = \frac{n+1}{2}, \quad \mu_2' = \frac{(n+1)(2n+1)}{6}, \quad \mu_3' = \frac{n(n+1)^2}{4}$$

Find the first three central moments.

Solution :

$$\mu_1 = 0$$

$$\mu_1' = \frac{n+1}{2} = d$$

$$\mu_2 = \mu_2' - d^2$$

$$\begin{aligned}
 \mu_2 &= \frac{(n+1)(2n+1)}{6} - \left[\frac{n+1}{2} \right]^2 \\
 &= (n+1) \left[\frac{2n+1}{6} - \frac{n+1}{4} \right] \\
 &= \frac{n+1}{12} [2(2n+1) - 3(n+1)] \\
 &= \frac{n+1}{12} [4n+2-3n-3] = \frac{(n+1)(n-1)}{12} \\
 &= \frac{n^2-1}{12}
 \end{aligned}$$

$$\begin{aligned}
 \mu_3 &= \mu_3' - 3d\mu_2' + 2d^3 \\
 &= \frac{n(n+1)^2}{4} - 3 \frac{(n+1)}{2} \frac{(n+1)(2n+1)}{6} + 2 \left[\frac{(n+1)}{2} \right]^3 \\
 &= \frac{(n+1)^2}{4} [n - (2n+1) + (n+1)]
 \end{aligned}$$

$$\mu_3 = 0$$

β and r coefficients

Karl Pearson defined the following four coefficients, based upon the first four central moments.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$r_1 = \sqrt{\beta_1} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\sqrt{\sigma^6}} = \frac{\mu_3}{\sqrt{\sigma^3}}$$

$$\therefore r_1 = \frac{\mu_3}{\sigma^3}$$

$$r_2 = \beta_2 - 3$$

Example 6 : Calculate β and r coefficients for $\mu_2 = 16$, $\mu_3 = -64$, $\mu_4 = 362$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = 1$$

$$\therefore r_1 = \sqrt{\beta_1} = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{362}{(16)^2} = \frac{362}{256} = 1.41$$

$$r_2 = \beta_2 - 3 = 1.41 - 3 = -1.59$$

Example 7 : For a distribution, the mean is 10, the variance is 16, $r_1 = 1$ and $\beta_2 = 4$. Find the first four central moments.

Solution :

$$\mu_1 = 0$$

$$\mu_2 = \sigma^2 = \text{variance} = 16$$

$$\therefore \sigma = 4$$

$$r_1 = \frac{\mu_3}{\sigma^3}$$

$$1 = \frac{\mu_3}{4^3}$$

$$\therefore \mu_3 = 64$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{16^2}$$

$$\mu_4 = 4 \times 256 = 1024$$

$$\mu_4 = 1024$$

Sheppard's correction for grouping errors

While calculating moments it is assumed that all the values of a variable in a class interval are concentrated at the centre of that interval i.e., mid point. However, in practice it is not so the assumption is an approximation to facilitate calculations and it introduces some error which is known as grouping error. If the distribution is symmetrical or almost symmetrical the items will be found on either side of the mean more or less symmetrically and hence the errors due to the assumption will be compensated. But in the calculation of S.D. we square the deviations and hence errors are cumulative and so a correction has to be applied for grouping. Sheppard has proved the following results for a frequency distribution which is continuous and bell shaped.

$$\begin{aligned}\mu_2 (\text{corrected}) &= \mu_2 (\text{calculated}) - \frac{i^2}{12} \\ \mu_4 (\text{corrected}) &= \mu_4 (\text{calculated}) - \frac{i^2}{2} \mu_2 (\text{calculated}) \\ &+ \frac{7}{240} i^4 \text{ where 'i' is the width of the class interval.}\end{aligned}$$

The first and third moments need no correction.

Example 8 : Apply shepperd's corrections to the central moments calculated as $\mu_2 = 15.93$, $\mu_4 = 681.79$ and $i = 4$.

$$\begin{aligned}\text{Solution : } \mu_2 (\text{corrected}) &= \mu_2 (\text{calculated}) - \frac{i^2}{12} \\ &= 15.93 - \frac{4^2}{12} \\ &= 15.93 - \frac{16}{12} = 15.93 - 1.33 \\ &= 14.60 \\ \mu_4 (\text{Corrected}) &= \mu_4 (\text{calculated}) - \frac{i^2}{2} \mu_2 (\text{Calculated}) + \frac{7}{240} i^4 \\ &= 681.79 - \frac{16}{2} (15.93) + \frac{7}{240} (256) \\ &= 681.79 - 127.44 + 7.47 \\ \mu_4 (\text{corrected}) &= 561.82\end{aligned}$$

Exercise

- Find the first three moments about 4 of the values 21, 10, 9, 3.
- Find the first four central moments and apply sheppard's correction.

Class	3-7	8-12	13-17	18-22	23-27	28-32	33-37	38-42
Freq	2	108	580	175	80	32	18	5

- The first three moments of a distribution about the origin 3 or 2, 10 and -30. Find the first three central moments. Also calculate the mean, the variance, β and r .
- The first four moments of a distribution about 4 are 1, 4, 10, 45. Find the four central moments, mean, variance, β_1 , β_2 , r_1 and r_2 .

Moments of a continuous distributions

Suppose we consider the histogram of a frequency distribution. If we go on decreasing the widths of the class intervals indefinitely the different histograms will approach a smooth curve and this curve will represent the frequency curve of the distribution. In general for every frequency distribution, we can associate a continuous curve. The study of these curves will enable us to know the distribution.

Let us assume that $y = f(x)$ represents the curve where x is an element of the population and $f(x)$ a real number. Since $f(x)$ is associated with the frequency, and the frequency, is always non negative, we assume $f(x) \geq 0$ for all x .

The curve $y = f(x)$ is the limit of histogram and we know that area of the histogram is the total frequency N . So the area between the curve and the x - axis is equal to the total frequency N . So, if the minimum and maximum value of x are a and b , then we have

$$N = \int_a^b f(x) dx \quad \text{-----} \quad 1$$

But if the curve extends from $-\infty$ to ∞ ,

$$\text{then } N = \int_{-\infty}^{\infty} f(x) dx \quad \text{-----} \quad 2$$

The area between the curve and the x -axis from $x = C$ to $x = d$

ie., $\int_c^d f(x) dx$ will give the total frequency in the interval $[c, d]$. So, if δx is the width

of an elementary interval then the elementary area $f(x) \delta(x)$ will be the frequency of the class of width δx .

Now, let us try to calculate $\bar{x}$, σ and moments of a frequency distribution using the frequency curve. Let the curve lie in the interval $[a, b]$.

$$1) \quad A.M. = \bar{x}$$

Frequency of each elementary interval = $f(x) \delta(x)$

$$\therefore \bar{x} = \frac{\sum x f(x) \delta(x)}{N}$$

$$= \frac{1}{N} \int_a^b x f(x) dx \quad (\text{Taking the limit of the sum as } \delta x \rightarrow 0)$$

$$\begin{aligned}
 2) \quad \mu_1 &= \frac{\sum (x - \bar{x}) f(x) \delta x}{N} \\
 &= \frac{1}{N} \int_a^b (x - \bar{x}) f(x) dx \\
 &= \frac{1}{N} \int_a^b x f(x) dx - \bar{x} \frac{1}{N} \int_a^b f(x) dx \\
 &= \frac{1}{N} \int_a^b x f(x) dx - \bar{x} \frac{1}{N} N = 0
 \end{aligned}$$

$$\begin{aligned}
 3) \quad \mu_2' \text{ about the origin zero} &= \frac{1}{N} \sum x^2 f(x) \delta x \\
 &= \frac{1}{N} \int_a^b x^2 f(x) dx
 \end{aligned}$$

$$4) \quad \mu_2 = \frac{1}{N} \int_a^b (x - \bar{x})^2 f(x) dx$$

$$5) \quad \mu_2' \text{ about the origin } A = \frac{1}{N} \int_a^b (x - A)^2 f(x) dx$$

$$6) \quad \mu_k' \text{ about the origin zero} = \frac{1}{N} \int_a^b x^k f(x) dx$$

$$7) \quad \mu_k' \text{ about } A = \frac{1}{N} \int_a^b (x - A)^k f(x) dx$$

$$8) \quad \mu_k = \frac{1}{N} \int_a^b (x - \bar{x})^k f(x) dx$$

$$9) \quad \text{Suppose } F(x) \text{ gives the less than cumulative frequency upto } x, \text{ then } F(x) = \int_a^x f(t) dt$$

So, the value of x such that

$$F(x) = \int_a^x f(t) dt = \frac{N}{2} \text{ gives the median.}$$

Note : The curve $y = F(x)$ denotes the less than Ogive.

10) At the mode, the curve $y = f(x)$ has a maximum and so $f'(x) = 0$

$$\text{Now } F(x) = \int_a^x f(t) dt$$

$$\text{Hence } F'(x) = f(x)$$

$$\text{and } F''(x) = f'(x)$$

$$\therefore f'(x) = 0 \Rightarrow F''(x) = 0$$

But $F''(x) = 0$ gives point of inflection of the curve $y = F(x)$

Hence the abscissa of the point of inflection of the Ogive gives the mode of the frequency curve.

11) The mean deviation about $\bar{x}$

$$= \frac{1}{N} \int_a^b f(x) |x - \bar{x}| dx$$

12) The quartiles are given by Q_1 and Q_3 such that $\int_a^{Q_1} f(x) dx = \int_{Q_3}^b f(x) dx = \frac{1}{4} N$

Example 1 : x is a variate following a uniform distribution in the interval $[0, a]$, ie., $f(x) = k$ a constant for all x in $[0, a]$ and zero elsewhere. Find (i) its S.D (ii) the average deviation and (iii) the r^{th} moment about the mean.

Solution :

$$\text{Total frequency in } [0, a] = N = \int_0^a k dx = [kx]_0^a = ka$$

$$\text{Mean} = \frac{1}{N} \int_0^a x f(x) dx = \frac{1}{N} \int_0^a kx dx$$

$$\bar{x} = \frac{k}{ka} \left[\frac{x^2}{2} \right]_0^a = \frac{a}{2}$$

$$r^{\text{th}} \text{ moment about the mean} = \mu_r$$

$$= \frac{1}{N} \int_0^a (x - \bar{x})^r f(x) dx$$

$$= \frac{1}{ka} \int_0^a \left(x - \frac{a}{2}\right)^r k dx$$

$$= \frac{1}{a} \left[\frac{\left(x - \frac{a}{2}\right)^{r+1}}{r+1} \right]_0^a$$

$$= \frac{1}{a(r+1)} \left[\left(a - \frac{a}{2}\right)^{r+1} - \left(0 - \frac{a}{2}\right)^{r+1} \right]$$

$$= \frac{1}{a(r+1)} \left[\left(\frac{a}{2}\right)^{r+1} - \left(-\frac{a}{2}\right)^{r+1} \right]$$

$$\text{When } r \text{ is even } \mu_r = \frac{1}{a(r+1)} \left[2 \left(\frac{a}{2}\right)^{r+1} \right]$$

$$= \frac{a^{r+1}}{a(r+1) 2^r} = \frac{a^r}{(r+1) 2^r}$$

$$\text{When } r \text{ is odd, } \mu_r = 0$$

ie., all the odd moments are zero.

$$\text{When } r = 2, \mu_2 = \frac{a^2}{3 \cdot 2^2} = \frac{a^2}{12}$$

$$\text{ie., } \sigma^2 = \frac{a^2}{12}$$

$$\sigma = \sqrt{\frac{a^2}{12}} = \frac{a}{2\sqrt{3}}$$

Average deviation about the mean :

$$\frac{1}{N} \int_0^a |x - \bar{x}| f(x) dx$$

$$= \frac{1}{ka} \int_0^a \left| x - \frac{a}{2} \right| k dx$$

$$\text{for } 0 \leq x \leq \frac{a}{2}, \quad \left| x - \frac{a}{2} \right| = \frac{a}{2} - x$$

$$\text{and for } \frac{a}{2} \leq x \leq a, \quad \left| x - \frac{a}{2} \right| = x - \frac{a}{2}$$

$$\therefore \text{Average deviation} = \frac{k}{ka} \left[\int_0^{a/2} \left(\frac{a}{2} - x \right) dx + \int_{a/2}^a \left(x - \frac{a}{2} \right) dx \right]$$

$$= \frac{1}{a} \left[\left(\frac{a}{2} x - \frac{x^2}{2} \right) \Big|_0^{a/2} + \left(\frac{x^2}{2} - \frac{a}{2} x \right) \Big|_{a/2}^a \right]$$

$$= \frac{1}{a} \left[\left\{ \frac{a}{2} \cdot \frac{a}{2} - \frac{1}{2} \left(\frac{a}{2} \right)^2 \right\} + \left\{ \left(\frac{a^2}{2} - \frac{a}{2} a \right) - \left(\frac{1}{2} \left(\frac{a}{2} \right)^2 - \frac{a}{2} \cdot \frac{a}{2} \right) \right\} \right]$$

$$= \frac{1}{a} \left[\frac{a^2}{4} - \frac{a^2}{8} - \frac{a^2}{8} + \frac{a^2}{4} \right] = \frac{a}{4}$$

$$\therefore \text{S.D.} = \frac{a}{2\sqrt{3}}$$

$$\text{Average deviation} = \frac{a}{4}$$

$$\mu_r = 0 \text{ if } r \text{ is odd}$$

$$= \frac{ar}{(r+1) 2^r} \text{ if } r \text{ is even.}$$

Example 2 : For the distribution $f(x) = \frac{2}{\pi} \frac{1}{1+x^2}$

defined in the interval $[0, \infty]$, find the median.

$$\begin{aligned} \text{Solution : } N &= \int_0^{\infty} \frac{2}{\pi} \frac{1}{1+x^2} dx \\ &= \frac{2}{\pi} \left[\tan^{-1} x \right]_0^{\infty} = \frac{2}{\pi} \cdot \frac{\pi}{2} = 1 \end{aligned}$$

If M is the median, then

$$\int_0^M f(x) dx = \frac{1}{2}$$

$$\text{i.e., } \int_0^M \frac{2}{\pi} \frac{1}{1+x^2} dx = \frac{1}{2}$$

$$= \frac{2}{\pi} \left[\tan^{-1} x \right]_0^M = \frac{1}{2}$$

$$\tan^{-1} M = \frac{\pi}{4}$$

$$\therefore M = \tan \left(\frac{\pi}{4} \right) = 1$$

ie., median = 1.

Exercise 1. A certain continuous distribution is given by the equation $f(x) = x^2$ when $0 \leq x \leq 1$ and $f(x) = (2-x)^3$ when $1 \leq x \leq 2$ calculate $\bar{x}$ and σ .

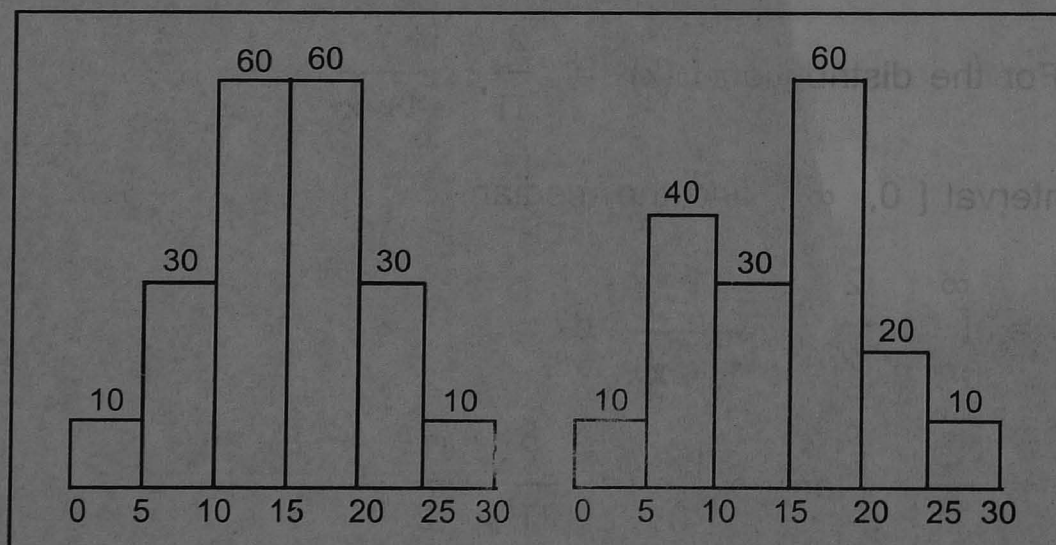
2. The frequency function of a variate x following a triangular distribution is given as follows.

$$f(x) = x \quad (0 \leq x \leq 1) \text{ and } f(x) = \frac{3-x}{2} \quad (1 \leq x \leq 3)$$

find the mean, variance and r^{th} moment.

2.2 Skewness

Skewness helps us to understand a distribution. Two distributions may have the same mean and standard deviation but may differ widely in their overall appearance as can be seen from the following:

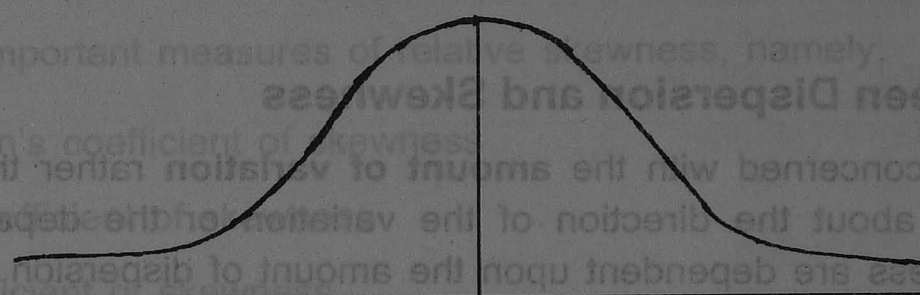


In both these distributions the value of mean and standard deviation is the same $\bar{x} = 15$, $\sigma = 5$. But it does not imply that the distributions are alike in nature. The distribution on the left - hand side is a symmetrical one whereas the distribution on the right - hand is asymmetrical or skewed. Measures of skewness help us to distinguish between different types of distributions.

The term 'skewness' refers to lack of symmetry i.e., when a distribution is not symmetrical it is called a skewed distribution. If the curve is normal, spread will be the same on both sides of the centre point and the mean, median and mode will all have the same value. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution.

These are two types of distribution. They are :

(a) Symmetrical Distribution



$$\bar{x} = \text{Median} = \text{Mode}$$

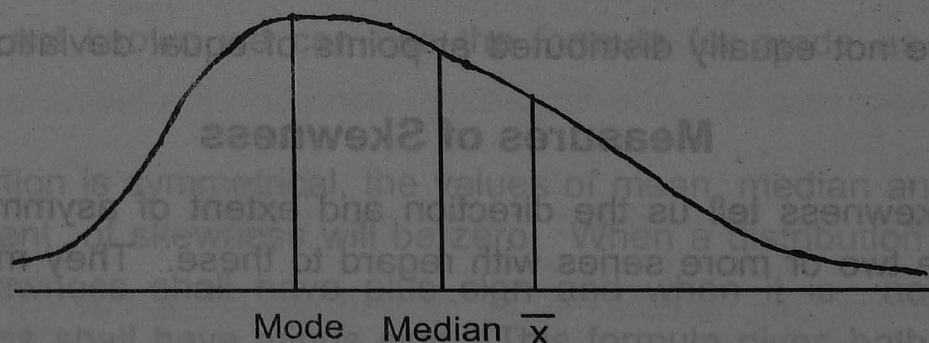
It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.

(b) A Skewed Distribution

A distribution which is not symmetrical is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

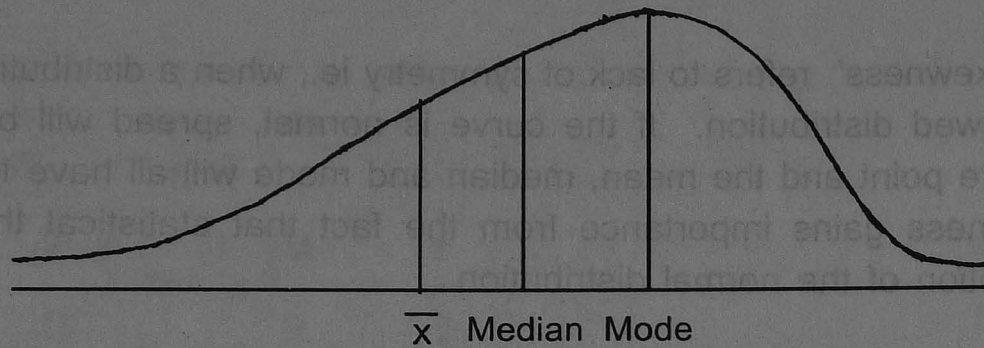
(i) Positively skewed distribution

In the positively skewed distribution, the value of the mean is maximum and that of mode least - the median lies in between the two.



(ii) Negatively skewed distribution

In a negatively skewed distribution the value of mode is maximum and that of mean least - the median lies in between the two.



In the positively skewed distribution the frequencies are spread out over a greater range of values on the high value end of the curve than they are on the low value end. In the negatively skewed distribution the position is reversed i.e., the excess tail is on the left-hand side.

Difference between Dispersion and Skewness

Dispersion is concerned with the **amount of variation** rather than with its **direction**. Skewness tells us about the direction of the variation or the departure from symmetry. Measures of skewness are dependent upon the amount of dispersion.

Although skewness is an important characteristic for defining the precise pattern of a distribution, it is rarely calculated in business and economic series. Variation is the most important characteristic of a distribution.

Tests of skewness

Skewness is present if :

- (i) The values of mean, median and mode do not coincide.
- (ii) When the data are plotted on a graph they do not give the normal bell - shaped form.
- (iii) The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
- (iv) Quartiles are not equidistant from the median.
- (v) Frequencies are not equally distributed at points of equal deviation from the mode.

Measures of Skewness

Measures of skewness tell us the direction and extent of asymmetry in a series, and permit us to compare two or more series with regard to these. They may either be absolute or relative.

Absolute measures of skewness

Skewness can be measured in absolute terms by taking the difference between mean and mode. Symbolically,

$$\text{Absolute skewness} = \bar{x} - \text{Mode}$$

If the value of mean is greater than mode skewness will be positive, i.e., we shall get a plus sign in the above formula. Conversely, if the value of mode is greater than mean, we shall get a minus sign meaning thereby that the distribution is negatively skewed.

If the absolute differences were expressed in relation to some measure of spread of values in their respective distributions, the measure would be relative and used directly for comparison.

Relative Measures of Skewness

There are four important measures of relative skewness, namely,

1. The Karl Pearson's coefficient of skewness
2. The Bowley's coefficient of skewness
3. The Kelly's coefficient of skewness
4. Measure of skewness based on moments.

Requisites of a good measure of skewness

1. It should be a pure number
2. It should have a zero value, when the distribution is symmetrical
3. It should have some meaningful scale of measure so that we could easily interpret the measured value.

1. Karl Pearson's coefficient of Skewness

The formula is

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

When Mode is not known we can use the formula for mode as $\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$.

When a distribution is symmetrical, the values of mean, median and mode coincide and therefore, the coefficient of skewness will be zero. When a distribution is positively skewed, the coefficient of skewness shall have plus sign and when it is negatively skewed, the coefficient of skewness shall have minus sign. This formula gives both the direction as well as the extent of skewness.

Theoretically, the value of this coefficient varies between ± 3 ; however, in practice it is rare that the coefficient of skewness obtained exceeds ± 1 .

2) Bowley's coefficient of Skewness

Bowley's measure is based on quartiles. In a symmetrical distribution first and third quartiles are equidistant from the median. In an asymmetrical distribution the third quartile is the same distance over the median as the first quartile is below it. i.e.,

$$Q_3 - \text{Median} = \text{Median} - Q_1$$

$$\text{or } Q_3 + Q_1 - 2 \text{ Median} = 0$$

If this distribution is positively skewed the top 25 per cent of the values will tend to be farther from median than the bottom 25 per cent and the reverse for negative skewness. Hence a possible measure is coefficient of Skewness

$$= \frac{(Q_3 - \text{Median}) - (\text{Median} - Q_1)}{(Q_3 - \text{Median}) + (\text{Median} - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

It must be remembered that the results obtained by these two measures are not to be compared with one another. Especially the numerical values are not related to one another since the Bowley's measure is limited to values between -1 and +1, while Pearson's measure has no such limits.

3) Kelly's Coefficient of Skewness

Kelly has suggested the following formula for measuring skewness upon the 10th and the 90th percentiles or the first and ninth deciles.

$$\text{Coefficient of Skewness} = \frac{P_{10} + P_{90} - 2 \text{ median}}{P_{90} - P_{10}}$$

$$\text{also} = \frac{D_1 + D_9 - 2 \text{ Median}}{D_9 - D_1}$$

This measure of skewness has one theoretical attraction if skewness is to be based on percentiles.

4) Measure of skewness based on the third moment

A measure of skewness may be obtained by making use of the second and third moments about the mean. β_1 is used as relative measure of skewness and we know $\beta_1 = \frac{\mu_3^2}{\mu_2^3} \ln$

a symmetrical distribution β_1 shall be zero. The greater the value of β_1 the more skewed the distribution. But β_1 cannot tell us about the direction of skewness, i.e., whether it is positive or negative, because μ_2 being the sum of the cubes of the deviations from the mean may be positive or negative but μ_3^2 is always positive. Since μ_2 is the variance it is always positive. $\therefore \beta_1$ is always positive. This drawback is removed if we calculate Karl Pearson's r_1 which is the square root of β_1 . $\therefore$ The sign of skewness would depend upon the value of μ_3 . If μ_3 is positive we will have positive skewness and if μ_3 is negative we will have negative skewness.

Example 1 : Calculate Karl Pearson's Coefficient of skewness :

x :	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5
f :	28	42	54	108	129	61	45	33

Solution :

x	f	(x - 27.5) / 5 d	fd	fd ²
12.5	28	-3	-84	252
17.5	42	-2	-84	168
22.5	54	-1	-54	54
27.5	108	0	0	0
32.5	129	1	129	129
37.5	61	2	122	244
42.5	45	3	135	405
47.5	33	4	132	528

$$N = 500$$

$$\Sigma fd = 296$$

$$\Sigma fd^2 = 1780$$

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$\text{Mean} = \bar{x} = A + \frac{\Sigma fd}{N} \times i$$

$$A = 27.5, \quad \Sigma fd = 296, \quad N = 500, \quad i = 5$$

$$\bar{x} = 27.5 + \frac{296}{500} \times 5 = 30.46$$

Mode : Since the maximum frequency is 129, 32.5 is the modal value

$$S.D = \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

$$\sum fd^2 = 1780, N = 500, \sum fd = 296, i = 5$$

$$\sigma = \sqrt{\frac{1780}{500} - \frac{296^2}{500^2}} \times 5 = \sqrt{3.56 - 0.35} \times 5$$

$$\sigma = 8.96$$

$$\text{Coefficient of skewness} = \frac{30.46 - 32.5}{8.96} = \frac{-2.04}{8.96}$$

$$= -0.228$$

Example 2 : Find Bowley's coefficient of skewness for the following frequency distribution.

No. of children per family	0	1	2	3	4	5	6
No of families	7	10	16	25	18	11	8

Solution :

No of Children per family x	No of families f	c.f
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95

$$N = 95$$

$$\text{Skewn ess} = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

$$Q_1 = \text{size of } \frac{N+1}{4} \text{ th item}$$

$$= \frac{95+1}{4} = 24\text{th item} \quad \therefore Q_1 = 2$$

$$Q_3 = \text{size of } \frac{3(N+1)}{4} \text{ th item}$$

$$= \frac{3 \times 96}{4} = 72\text{th item} \quad \therefore Q_3 = 4$$

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{ th item} = \frac{96}{2} = 48 \text{ th item}$$

$$\therefore \text{Median} = 3$$

$$\therefore \text{Coefficient of Skewness} = \frac{4+2 - 2(3)}{4 - 2} = \frac{0}{2} = 0$$

Example 3 : Calculate coefficient of skewness based on quartiles and median from the following data

Variable	Frequency
0-10	12
10-20	16
20-30	26
30-40	38
40-50	22
50-60	15
60-70	7
70-80	4

Solution : The Bowley's coefficient of skewness is given by $\frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$

$\therefore$ we have to find this coefficient of skewness.

Variable	f	c.f
0-10	12	12
10-20	16	28
20-30	26	54
30-40	38	92
40-50	22	114
50-60	15	129
60-70	7	136
70-80	4	140

$$Q_1 = \text{size of } \frac{N}{4} \text{ th item} = \frac{140}{4} = 35\text{th item.}$$

$\therefore$ It lies in the class 20-30

$$Q_1 = L + \frac{\frac{N}{4} - \text{c.f}}{f} \times i$$

$$L = 20, \quad \frac{N}{4} = 35, \quad \text{c.f} = 28, \quad f = 26, \quad i = 10$$

$$\therefore Q_1 = 20 + \frac{35-28}{26} \times 10 = 20 + 2.69 = 22.69$$

$$Q_3 = \text{size of } \frac{3N}{4} \text{ th item} = \frac{3 \times 140}{4} = 105 \text{ th item}$$

It lies in the class 40-50

$$Q_3 = L + \frac{\frac{3N}{4} - \text{c.f}}{f} \times i$$

$$L = 40, \quad \frac{3N}{4} = 105, \quad \text{c.f.} = 92, \quad f = 22, \quad i = 10$$

$$Q_3 = 40 + \frac{105-92}{22} \times 10 = 40 + 5.91 = 45.91$$

$$\text{Median} = \text{size of } \frac{N}{2} \text{ th item} = \frac{140}{2} = 70\text{th item}$$

It lies in the class 30-40

$$\text{Median} = \frac{L + \frac{\frac{N}{2} - c.f.}{f} \times i}{1}$$

$$L = 30, \frac{N}{2} = 70, c.f. = 54, f = 38, i = 10$$

$$\text{Median} = 30 + \frac{70 - 54}{38} \times 10 = 30 + 4.21 = 34.21$$

$$\text{Coefficient of skewness} = \frac{45.91 + 22.69 - 2(34.21)}{45.91 - 22.69}$$

$$= \frac{68.6 - 68.42}{23.22} = 0.008$$

Exercise : 1) The daily expenditure of 100 families is given below :

Daily Expenditure	0-20	20-40	40-60	60-80	80-100
No. of families	13	?	27	7	16

If the mode of the distribution is 44, Calculate the Karl Pearson coefficient of skewness

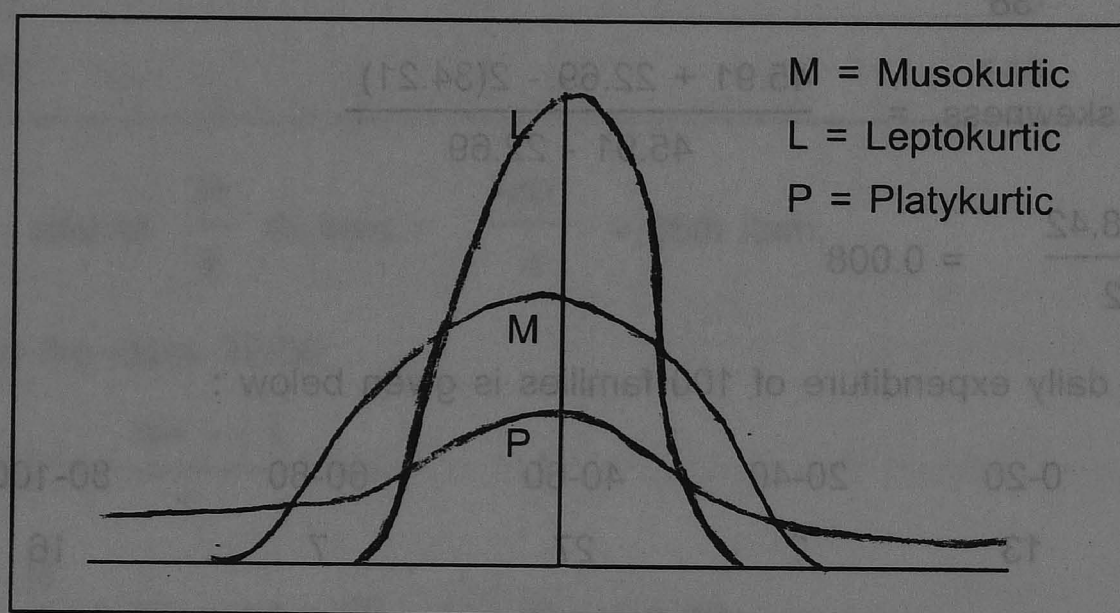
2) Calculate quarters and median and hence find out coefficient of skewness for the following data :

Daily Wage (Rs)	No. of workers
Less than 10	12
10-20	28
20-30	50
30-40	66
40-50	18
50-60	16
More than 60	10

2.3 - Kurtosis

Kurtosis in Greek means "bulginess". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. If a curve is more peaked than the normal curve, it is called 'leptokurtic'. In such a case items are more closely bunched around the mode. On the other hand, if a curve is more flat - topped than the normal curve, it is called 'Platykurtic'. The normal curve itself is known as 'mesokurtic'. The condition of peakedness or flat - toppedness itself is known as kurtosis of excess.

The following diagram illustrates the shape of three different curves mentioned above.



Measures of Kurtosis

The most important measure of kurtosis is the value of the coefficient β_2 . It is defined as $\beta_2 = \frac{\mu_4}{\mu_2^2}$ where $\mu_4 = 4^{\text{th}}$ moment and $\mu_2 = 2^{\text{nd}}$ moment. The greater the value of β_2 , the more peaked the distribution.

For a normal curve, the value of $\beta_2 = 3$. When the value of β_2 is greater than 3, the curve is more peaked than the normal curve, i.e., leptokurtic. When the value of β_2 is less than 3 the curve is less peaked than the normal curve i.e., platykurtic. The normal curve and other curves with $\beta_2 = 3$ are called mesokurtic.

Sometimes, r_2 , the derivative of β_2 , is used as a measure of kurtosis. r_2 is defined as

$$r_2 = \beta_2 - 3$$

For a normal distribution $r_2 = 0$. If r_2 is positive, the curve is leptokurtic and if r_2 is negative, the curve is platykurtic.

Example : Compute kurtosis based on the following data :

x :	4.5	14.5	24.5	34.5	44.5	54.5	64.5	74.5	84.5	94.5
f :	1	5	12	22	17	9	4	3	1	1

Solution : For kurtosis we have to compute β_2

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

x	f	(x - 44.5) / 10 d	fd	fd ²	fd ³	fd ⁴
4.5	1	-4	-4	16	-64	256
14.5	5	-3	-15	45	-135	405
24.5	12	-2	-24	48	-96	192
34.5	22	-1	-22	22	-22	22
44.5	17	0	0	0	0	0
54.5	9	1	9	9	9	9
64.5	4	2	8	16	32	64
74.5	3	3	9	27	81	243
84.5	1	4	4	16	64	256
94.5	1	5	5	25	125	625
N = 75			$\Sigma fd = -30$	224	-6	2072

$$\mu_1' = \frac{\Sigma fd}{N} = \frac{-30}{75} = -0.4$$

$$\mu_2' = \frac{\Sigma fd^2}{N} = \frac{224}{75} = 2.99$$

$$\mu_3' = \frac{\Sigma fd^3}{N} = \frac{-6}{75} = -0.08$$

$$\mu_4' = \frac{\Sigma fd^4}{N} = \frac{2072}{75} = 27.63$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 2.99 - (-0.4)^2 = 2.99 - 0.16 = 2.83$$

$$\mu_3 = \mu_3' - 3 \mu_1' \mu_2' + 2 (\mu_1')^2 \mu_2' = -0.08 - 3 (-0.4) (2.99) + 2 (-0.4)^2$$

$$= -0.08 + 3.588 - 0.128 = 3.38$$

$$\mu_4 = \mu_4' - 4 \mu_1' \mu_3' + 6 \mu_1'^2 \mu_2' - 3 (\mu_1')^4$$

$$= 27.63 - 4 (-0.4) (-0.08) + 6 (-0.4)^2 (2.99) - 3 (-0.4)^4$$

$$= 27.63 - 0.128 + 2.87 - 0.077$$

$$= 30.295$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{30.295}{(2.83)^2} = \frac{30.295}{8.01} = 3.782$$

Since the value of β_2 is greater than 3, the curve is more peaked than the normal curve i.e., leptokurtic.

Exercise

1. The first four central - moments of distribution are 0, 2.5, 0.7 and 18.75. Compute kurtosis of the distribution.
2. The first four moments of a distribution about the value 5 are 2, 20, 40 and 50. Compute kurtosis.

2.4 - Curve Fitting

Introduction : Statistical constants like averages measures of dispersion and skewness do not furnish us with all the information about the data. It may be necessary for us to preserve the data as a whole. To achieve this purpose, we try to find out a functional relationship between the variables comprising the data. The functional relationship between two variables gives us the expected values of a variable corresponding to the given values of the other variable. The differences between the observed and the expected values of the variable are accounted for as due to fluctuations inherent in the sample. The method of curve fitting enables us to find out the probable functional relationship between the variables that occur in the data.

Scatter diagram : Let x and y be the two variables comprising the data. Choose any two perpendicular lines as coordinate axes to represent the variables x and y . Plot the points corresponding to the different values (x, y) . These points are said to form a scatter diagram.

Types of curve fitting : Only the quantitative data lends itself to curve fitting. If the points of the scatter diagram cluster round a straight line a linear type of fit is suggested. If the points of a scatter diagram cluster round a curve, the corresponding curve is fitted to the data.

Some of the types of curves that can be fitted to any data are :

- (i) The straight line (linear type), $y = ax + b$
- (ii) A parabolic curve of the form $y = ax^2 + bx + c$
- (iii) A logarithmic fit - curve $y = ae^{bx}$ and the curve $y = ax^b$. For frequency distributions, we can fit a normal curve or binomial curve depending on the nature of the distributions.

Principle of least squares : When the type of curve that best fit the data is determined from other conditions, the functional relationship between the variable comprising the data can be easily determined by estimating the constants that occur in the equation to the curve. These constants known as parameters are determined by the method of least squares. The method of least squares states that given a set of measured value of a variable, that values is most likely to be the true estimate which makes the sum of the squares of the residual. If $x_1, x_2, \dots, x_n$ are the n measured values of the variable whose true value is x , then the residuals of $x_1, x_2, \dots, x_n$ from x are $(x - x_1), (x - x_2), \dots, (x - x_n)$. The principle of least squares implies that x is the best value of the variable under consideration only if

$$(x - x_1)^2 + (x - x_2)^2 + \dots + (x - x_n)^2 \text{ is a minimum .}$$

Normal equations : By studying the scatter diatgram, we obtain an appropriate form of the curve to be fitted to the data. By assuming the equation to such a curve we make use of the principle of least squares to etimate the best value of the parameters that occur in the equatilon to the curve. The principle that the sum of the squares of the residuals should be a minimum, gives us certain equations involving the unknown parameters. The number of such equations is always equal to the number of parameters assumed in the equation to the curve. These equations are known as the normal equations. Normal equations, when solved gives the best value of the parameters that occur in the equation to the curve.

Fitting a straight line : Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the observed set of values. Since the two variables x and y are related linearly, the functional relationship between them is of the form $y = ax + b$.

$\therefore$ The residual = $y - ax - b$. Here a and b are the parameters which are to be found out.

The principle of least square states that

$$\Sigma [y_i - (ax_i + b)]^2 = f(a, b) \text{ should be minimum.}$$

By differential calculus, we get

$$\frac{\partial f}{\partial a} = 0 = \frac{\partial f}{\partial b} \text{ ————— } 1$$

$$\left. \begin{aligned} \text{Here } \frac{\partial f}{\partial a} &= -2 \sum [y_i - (ax_i + b)] x_i \\ \text{and } \frac{\partial f}{\partial b} &= -2 \sum [y_i - (ax_i + b)] \end{aligned} \right\} 2$$

∴ (1) and (2) lead us to the following equations

$$\sum y_i = a \sum x_i + n b$$

$$\sum x_i y_i = a \sum x_i^2 + b \sum x_i$$

These equations can also be written as

$$\sum y = a \sum x + n b$$

$\sum xy = a \sum x^2 + b \sum x$ and these are called normal equations and from these equations we can determine the values of parameters a and b.

Example 1 : Find by the method of least squares, the straight line that best fits the following data :

Depth of water applied in inches (x)	0	12	24	36	48
Average field (in tons per acre) (y)	35	55	65	80	70

Solution : Let the best fitting line be

$$y = ax + b$$

∴ The normal equations are $\sum y = a \sum x + nb$

$$\sum xy = a \sum x^2 + b \sum x$$

To find a and b let us form the table first:

Let A = 24 and i = 12

x	y	$x = \frac{x - 24}{12}$	x^2	xy
0	35	-2	4	0
12	55	-1	1	-55
24	65	0	0	0
36	80	1	1	80
48	70	2	4	140
	305	0	10	95

In terms of the new variable x, y the normal equations are

$$\Sigma xy = a \Sigma x^2 + b \Sigma x$$

$$\Sigma y = a \Sigma x + n b$$

$$\begin{aligned} 95 &= 10a \\ 305 &= 5b \end{aligned} \quad \Rightarrow \quad \begin{aligned} a &= 9.5 \\ b &= 61 \end{aligned}$$

$\therefore$ The line of best fit is $y = 9.5x + 61$.

In terms of the original variables x and y, the line is $y = 9.5 \frac{(x - 24)}{12} + 61$

$$\text{ie., } y = 0.8x + 42$$

2) Fitting a parabola : Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be the observed set of values and $y = ax^2 + bx + c$ be the equation to the parabola of best fit.

The residual corresponding to any point (x, y) on the curve is $y - ax^2 - bx - c$.

In this equation, a, b, c are the parameters which are to be found out.

The normal equations involving these parameters are

$$\Sigma y = a \Sigma x^2 + b \Sigma x + nc$$

$$\Sigma xy = a \Sigma x^3 + b \Sigma x^2 + c \Sigma x$$

$$\Sigma x^2 y = a \Sigma x^4 + b \Sigma x^3 + c \Sigma x^2$$

These equations can be derived either by using differential calculus method or by using the ordinary method as in the case of a linear.

Solving these 3 equations we get the values of a, b, c and hence the parabola

$$y = ax^2 + bx + c$$

Example 2 : Fit a curve of the form $y = ax^2 + bx + c$ to the following data.

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

Solution : Let us take the origin for x at 2.5 and the new variable is $X = 2(x - 2.5)$

If $y = ax^2 + bx + c$ is the best fit for the given data, the normal equations are

$$\Sigma y = a \Sigma x^2 + b \Sigma x + nc$$

$$\Sigma xy = a \Sigma x^3 + b \Sigma x^2 + c \Sigma x$$

$$\Sigma x^2y = a \Sigma x^4 + b \Sigma x^3 + c \Sigma x^2$$

Let us form the table

x	y	$x = 2(x - 2.5)$	x^2	x^3	x^4	xy	x^2y
1.0	1.1	-3	9	-27	81	-3.3	9.9
1.5	1.3	-2	4	-8	16	-2.6	5.2
2.0	1.6	-1	1	-1	1	-1.6	1.6
2.5	2	0	0	0	0	0	0
3	2.7	1	1	1	1	2.7	2.7
3.5	3.4	2	4	8	16	6.8	13.6
4	4.1	3	9	27	81	12.3	36.9
	16.2	0	28	0	196	14.3	69.9

Substituting in the normal equations, we get

$$16.2 = 28a + 7c \quad \text{-----} \quad (1)$$

$$14.3 = 28b \quad \text{-----} \quad (2)$$

$$69.9 = 196a + 28c \quad \text{-----} \quad (3)$$

From (2), $b = 0.51$

From (1) and (3), we get $64.8 = 112a + 28c$

and $69.9 = 196a + 28c$

subtracting we get $84a = 5.1$

$$\therefore a = 0.06$$

substituting this value of a in (1) we get,

$$C = 2.07$$

∴ The equation of the curve is

$$y = 0.06 x^2 + 0.51 x \cdot 2.07$$

∴ The equation in terms of the original variables x, y is

$$y = (0.06) [2(x - 2.5)]^2 + 0.51 [2(x - 2.5)] + 2.07$$

$$= 0.24x^2 - 1.2x + 1.5 + 1.02x - 2.55 + 2.07$$

$$= 0.24x^2 - 0.18x + 1.02$$

The equation of the best fitting parabola is

$$y = 0.24x^2 - 0.18x + 1.02$$

(3) Logarithmic Fit : In certain type of data, the scatter diagram cannot be easily drawn. We come across such data in Biology, Banking, Business etc. The equations to such curves, fitting the data would be of the form $y = ae^{bx}$ (called as exponential curve) or $y = ab^x$ where a and b are certain constants whose values are to be determined.

If we assume the equation to the curve in the form $y = ae^{bx}$ the increase in y is very large as compared to an increase in x. (provided a and b are + ve) and as such an ordinary graph sheet will not be sufficient to draw the scatter diagram of the points (x, y). In such cases, we reduce the equation of the curve to a linear type, by a suitable transformations of the variable. Taking logarithm (to the base 10) on both sides of $y = ae^{bx}$ we get

$\log_{10} y = \log_{10} a + bx \log_{10} e$ By putting $Y = \log_{10} y$, $A = \log_{10} a$ and $B = b \log_{10} e$ we get the equation as $y = A + Bx$ which is a linear type. Here the normal equations can be written as usually.

Example 3 : Fit a curve of the form $y = Ae^{Bx}$ to the data given below :

x	0	1	2	3
y	2	27	310	2500

Solution : Taking log on both sides $y = Ae^{Bx}$ we get, $\log_{10} y = \log_{10} A + Bx \log_{10} e$

Put $y = \log_{10} y$, $a = \log_{10} A$ and $b = B \log_{10} e$

Then the equation can be rewritten as

$$y = a + bx$$

The normal equations in this case are

$$\Sigma y = na + b \Sigma x$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2$$

1

x	y	Y = log y	xY	x ²
0	2	0.3010	0	0
1	27	1.4314	1.4314	1
2	320	2.4914	4.9828	4
3	2500	3.3979	10.1937	9
6		7.6217	16.6079	14

Substituting the values in I, we get

$$7.6217 = 4a + 6b$$

$$16.6079 = 6a + 14b$$

Solving these we get $a = 0.3528$, $b = 1.03507$

Hence the straight line fit for (x, y) is

$$Y = 0.3528 + 1.03507 x \quad \text{----- II}$$

$$Y = \log_{10} y \Rightarrow y = 10^Y$$

$\therefore$ The equation of the exponential curve fitted to this is $y = 10^{0.3528 + 1.03507 x}$
 $= 10^{0.3528} \times 10^{1.03507 x}$

$$\text{ie., } y = 2.253 \times e^{2.383x} \text{ as } 10 = e^{2.3829}$$

(4) Fitting a curve of the form $y = ab^x$

This equation can be brought to linear type by taking logarithm on both sides, as done in fitting exponential curve $y = ab^x$ becomes $\log_{10} y = \log_{10} a + x \log_{10} b$

By putting $Y = \log_{10} y$, $A = \log_{10} a$, $B = \log_{10} b$ we get the linear equation as $Y = A + Bx$ which can be fitted very easily.

Example 4 : The population of a country at ten year intervals is given below :

Year	1881	1891	1901	1911	1921	1931	1941
Population in millions	3.9	5.3	7.3	9.6	12.9	17.1	23.2

By fitting the curve of the form $y = ab^x$ to this data estimate the population for 1951.

Solution : $y = ab^x$ is the equation of the curve to be fitted.

$$\therefore \log_{10} y = \log_{10} a + x \log_{10} b$$

$$\text{Put } \log_{10} y = Y, \log_{10} a = A \text{ and } \log_{10} b = B$$

$$\text{Now the equation is } Y = A + Bx$$

$$\text{Put } X = \frac{x - 1911}{10}$$

The the equation of the curve to be fitted in terms of the new variables X and Y will be $Y = A + B X$

The normal equation are :

$$\Sigma Y = nA + B \Sigma X$$

$$\Sigma XY = A \Sigma X + B \Sigma X^2$$

Year x	Population y	$X = \frac{x - 1911}{10}$	$Y = \log_{10} y$	X^2	XY
1881	3.9	-3	0.5911	9	-1.7733
1891	5.3	-2	0.7243	4	-1.4486
1901	7.3	-1	0.8633	1	-0.8633
1911	9.6	0	0.9823	0	0
1921	12.9	1	1.1106	1	1.1106
1931	17.1	2	1.2330	4	2.4660
1941	23.2	3	1.3655	9	4.0965
		$\Sigma X = 0$	6.8701	28	2.5879

Substituting in the normal equations, we get

Year	1881	1891	1901	1911	1921	1931	1941
Population in millions	3.9	5.3	7.3	9.8	12.9	17.1	23.5

$$\therefore A = 0.9814, B = 0.1281$$

$$\log_{10} a = A \Rightarrow a = 10^A = 10^{0.9814} = 9.581$$

$$\log_{10} b = B \Rightarrow b = 10^B = 10^{0.1281} = 1.343$$

$\therefore$ The required equation is $y = 9.581 (1.343)^x$

To estimate the population for 1951:

$$x = 1951 \Rightarrow x = \frac{1951-1911}{10} = 4$$

$$\therefore y = (9.581) (1.343)^4 = 31.18$$

$\therefore$ Population in 1951 is 31.18 million.

Fitting the curve of the form $y = ax^m$

$y = ax^m$ implies $\log_{10} y = \log a + m \log x$ which assumes the form $y = mX + c$ where $X = \log_{10} x$, $Y = \log_{10} y$, $c = \log_{10} a$. So we fit a straight line to the data (x, y) i.e., to $(\log_{10} x, \log_{10} y)$. The normal equations are as usual

$$\Sigma y = m \Sigma x + nc$$

$$\text{and } \Sigma xy = m \Sigma x^2 + \Sigma x$$

Example 5 : The pressure of a gas and its volume are known to be related by an equation of the form $pv^m = k$. In a certain experiment the following volumes of a quantity of gas were observed for the pressures specified. Find the value of m by fitting a straight line.

p kg per sq.m.	0.5	1	1.5	2.0	2.5	3
v in C.C	1620	1000	750	620	520	460

Solution : $pv^m = k \Rightarrow \log p + m \log v = \log k$

Let $x = \log_{10} p$, $y = \log_{10} v$

i.e., $x + m y = c$ where $c = \log k$.

The normal equations are

$$\Sigma x + m \Sigma y = nc$$

$$\text{and } \Sigma x^2 + m \Sigma xy = c \Sigma x$$

p	v	x = log p	y = log v	x ²	xy
0.5	1620	T.6990	3.2095	0.09061	-0.5061
1	1000	0.0000	3.000	0	0
1.5	750	0.1761	2.8751	0.03100	0.5063
2	620	0.3010	2.7924	0.09061	0.8405
2.5	520	0.3979	2.7160	0.12800	0.1081
3	460	0.4771	2.6628	0.2276	0.1271
		1.0511	17.2558	0.6159	0.59812

Substituting in the normal equations, we get

$$1.0511 + 17.2558 m = 6c$$

$$0.6159 + 0.59812 m = 1.0511 c$$

Solving these equations, we get

$$m = 1.7 \text{ and } c = 4.93$$

$$\therefore k = \text{antilog}(c) = 85110$$

$$\therefore \text{The required gas equation is } pv^{1.7} = 85110.$$

Exercise

- The working funds of a Bank during seven consecutive years from 1961 are given below :

1961	1962	1963	1964	1965	1966	1967
147	191	213	254	296	393	418

Fit a straight line to the above data use the result to estimate the working funds during 1968.

- Fit a parabola of second degree to the following data :

x	0	1	2	3	4	5	6
y	14	18	23	29	36	40	46

3. Fit a curve of the form $y = A e^{Bx}$ to the observation given below :

x	1	2	3	4	5	6
y	14	27	40	55	68	300

4. The population of a state at 10 yearly intervals is given below :

year (x)	1881	1891	1901	1911	1921	1931	1941	1951
Population in millions (y)	3.9	5.3	7.3	9.6	12.9	13.1	23.2	30.9

Fit a curve of the form $y = ab^x$ and estimate the population of 1947.

Exercise

1. The working funds of a Bank during seven consecutive years from 1961 are given below:

1961	1962	1963	1964	1965	1966	1967
147	191	213	234	266	393	418

- Fit a straight line to the above data use the result to estimate the working funds during

2. Fit a parabola of second degree to the following data:

x	0	1	2	3	4	5	6
y	14	18	23	29	36	40	46

Unit - 3

3.1 - Correlation

So far, we have confined ourselves to the distributions involving only one variable i.e., univariate distributions. We may come across certain series where each item of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons or if we write down the marks in two or more subjects obtained by students in an examination we get distributions involving several variables. In such distributions we may be interested to find out if there is any correlation among the variables under study. Let us consider only the distributions having two variables i.e., bivariate distributions.

In a bivariate distribution, if the change in one variable results in a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, i.e., if the increase in one results in a corresponding increase in the other and also the decrease in one corresponds to a decrease in the other, correlation is said to be direct or positive. But if they deviate in the opposite direction, i.e., if the increase in one results in a corresponding decrease in the other and vice versa., correlation is said to be inverse or negative.

For example the correlation between

- (i) the heights and weights of a group of persons
- (ii) the income and expenditure
- (iii) records of rainfall and yield of crops
- (iv) demand and price of a commodity
- (v) radius and area of a circle are positive

The correlation between

- (i) volume and pressure of a perfect gas
- (ii) supply and price of a commodity
- (iii) length and breadth of rectangles of constant area are negative.

Degree of correlation

Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other. In general, the degree of correlation may be of any one of the following :

- (a) Perfect positive
- (b) Limited degree of positive

- (c) No Correlation
- (d) Limited degree of negative
- (e) Perfect negative

Measurement of Correlation

Correlation can be studied by

(i) Graphic methods

- (a) Correlation graphs
- (b) Scatter diagrams

ii) Mathematical methods

- (a) Concurrent Deviation method
- (b) Product moment method
- (c) Difference method
- (d) Ranking method
- (e) Regression equation method

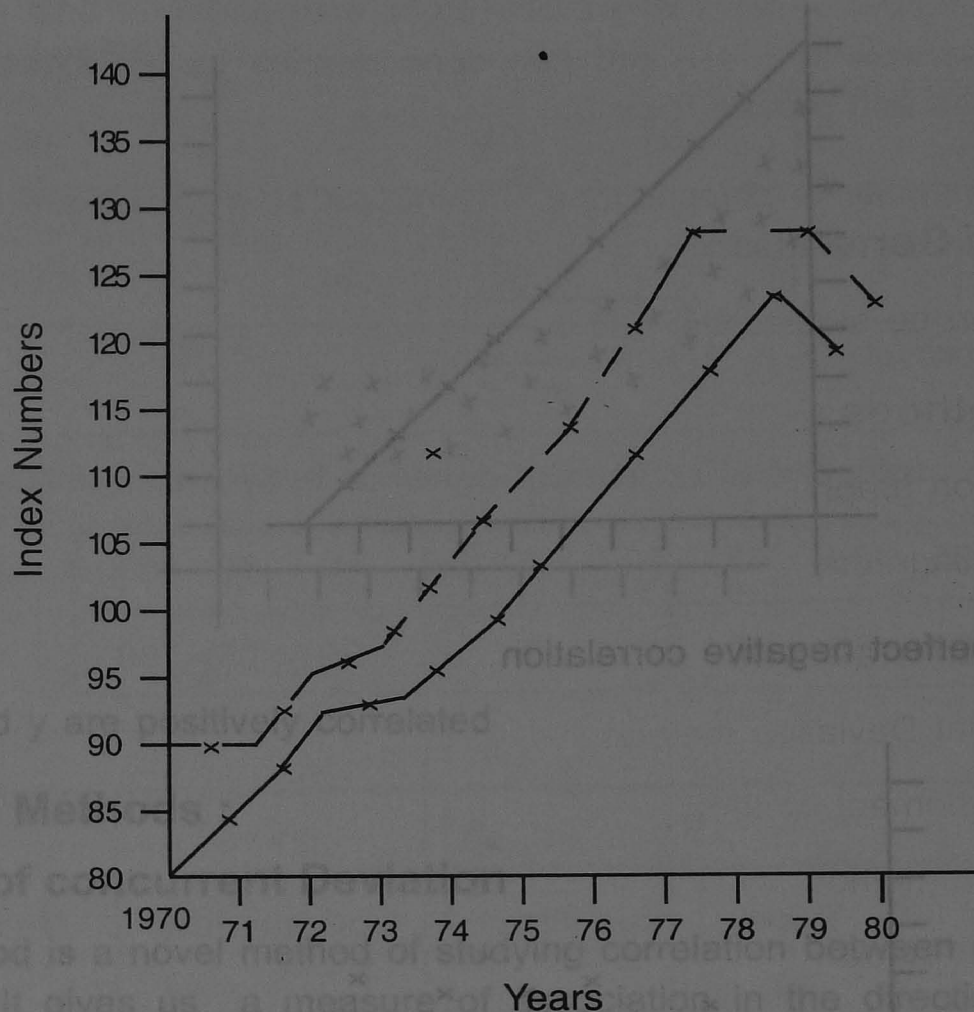
(ii) Graphic Methods

(a) Correlation Graphs

To get a rough idea about correlation graphs may be used. The two series may be plotted on a graph paper and their directions and closeness are noted. This is a usual method of finding out correlation. If the two series show similar tendency, it is an indication of positive correlation. On the other hand, if the two graphs move in opposite directions, correlation is negative.

Example : Plot the following two series on a graph paper and study their correlation

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Cost of Living index	80	85	90	100	100	105	108	120	125	130	125
Wage Index	90	90	100	105	109	115	120	130	140	140	130

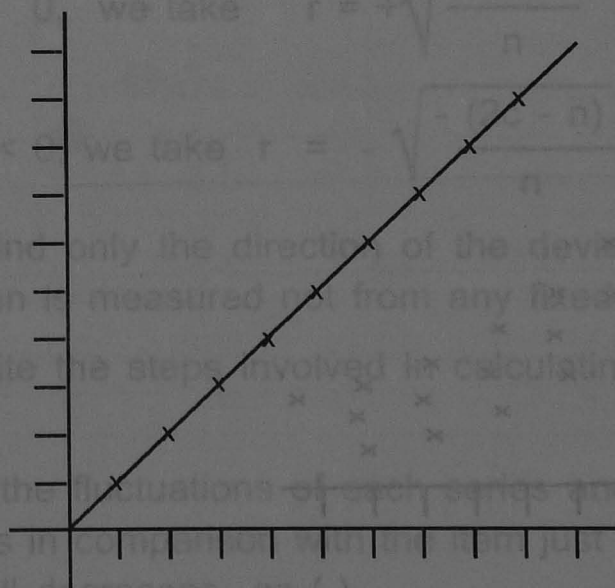


The two graphs move in the same direction and hence the two items are positively correlated.

(b) Scatter diagram

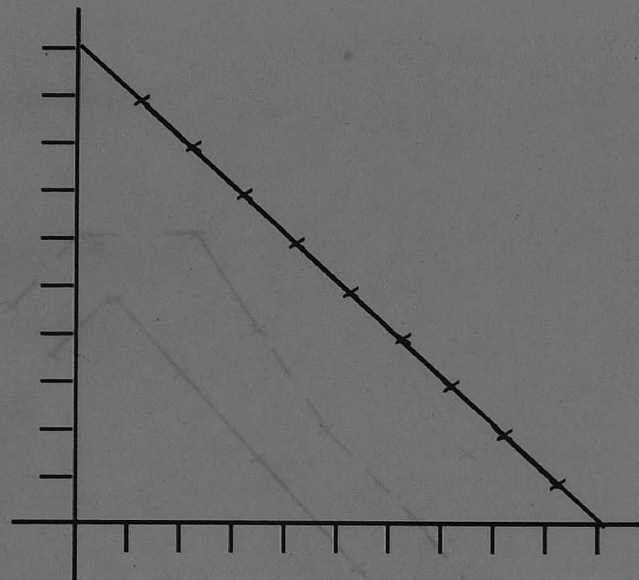
A scatter diagram is simply a graph on which points are plotted to represent pairs of values of the two variables. Taking one of the variables as x and the other as y , we represent them along x and y axis respectively and plot (x_i, y_i) . From this diagram we get the idea about the degree of correlation. eg. if the points are dense, i.e., very close to each other we expect a fairly high amount of correlation between the variables and if the points are widely scattered, we expect a poor correlation.

Example 1 :



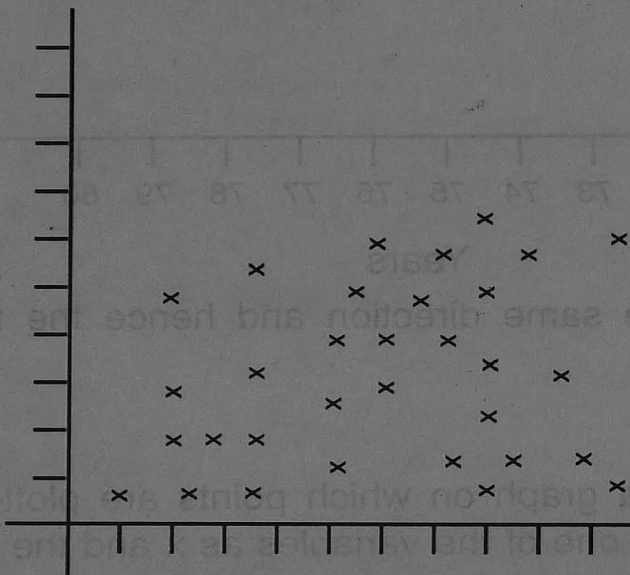
Here there is a perfect positive correlation between x and y .

Example 2 :



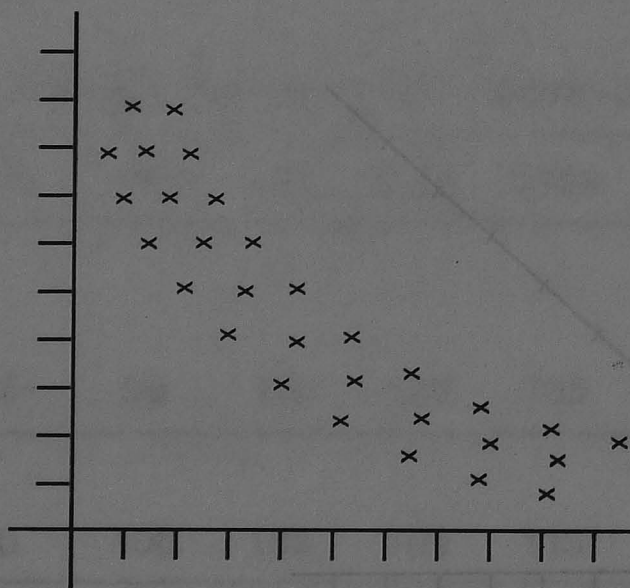
Here there is a perfect negative correlation

Example 3 :



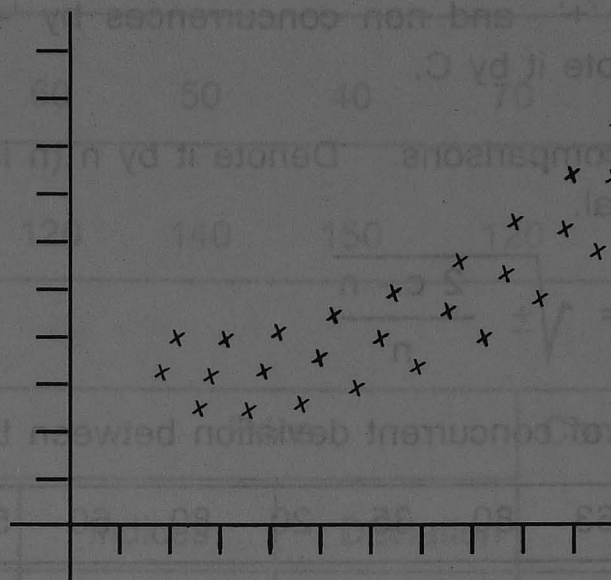
There is no correlation.

Example 4 :



Here x and y are negatively correlated

Example 5 :



Here x and y are positively correlated

Mathematical Methods :

(a) Method of concurrent Deviation

This method is a novel method of studying correlation between short term fluctuations in two series. It gives us a measure of association in the direction of change. In this method, we note the direction of deviation of each item from the preceding one and count the number of items where increase or decrease occur concurrently in the two variables and denote it by 'C'. Denoting by n, the number of comparisons, we write the coefficient of concurrent deviation by

$$r = \pm \sqrt{\pm \frac{2c - n}{n}}$$

The use of sign depends upon the sign of $\sqrt{\frac{2c - n}{n}}$

$$\text{If } \frac{2c - n}{n} > 0, \text{ we take } r = +\sqrt{\frac{2c - n}{n}}$$

$$\text{If } \frac{2c - n}{n} < 0, \text{ we take } r = -\sqrt{\frac{-(2c - n)}{n}}$$

Note that we find only the direction of the deviation and not the amount of deviation and also the deviation is measured not from any fixed value but from the preceding item.

Now, let us write the steps involved in calculating the coefficient of correlation by this method.

Step 1 : Examine the fluctuations of each series and find whether each item increases or decreases in comparison with the item just preceding. All increases are noted as (+) and all decreases as (-)

Step 2 : Count the number of items where both signs are positive or both negative. Denote the concurrences by '+' and non concurrences by '-' Count the number of concurrences and denote it by C.

Step 3 : Count the number of comparisons. Denote it by n (n is equal to one less than the total number of total).

Step 4 : Apply the formula $r = \sqrt{\pm \frac{2c - n}{n}}$

Example 1 : Find the coefficient of concurrent deviation between the following series.

A	60	40	35	75	63	80	35	20	80	60	50
B	60	55	50	56	30	70	40	35	80	75	80

Solution :

A		B		Concurrences	Non Concurrence
Value	Deviation	Value	Deviation		
60		60			
40	-	55	-	+	
35	-	50	-	+	
75	+	56	+	+	
63	-	30	-	+	
80	+	70	+	+	
35	-	40	-	+	
20	-	35	+	+	
80	+	80	+	+	
60	-	75	-	+	
50	-	80	+		
				9	

$$C = 9, n = 10$$

$$2c - n = (2 \times 9) - 10 = 18 - 10 = 8$$

$$r = + \sqrt{\frac{2c - n}{N}} = \sqrt{\frac{8}{10}} = \sqrt{0.8} = 0.89$$

Example 2 : Find the coefficient of concurrent deviation between supply and price.

Supply indices	80	60	50	40	70	60	70	80	90
Price indices	100	120	140	150	120	140	130	120	100

Solution :

Supply		Price		Concurrences	Non - Concurrences
Indices	Deviation	Indices	Deviation		
80		100			
60	-	120	+		-
50	-	140	+		-
40	-	150	+		-
70	+	120	-		-
60	-	140	+		-
70	+	130	-		-
80	+	120	-		-
90	+	100	-		-
				0	

$$C = 0, \quad n = 8$$

$$2C - n = -8$$

$$\therefore r = -\sqrt{\frac{-(-8)}{8}} = -\sqrt{1} = -1$$

Exercise : Compute the coefficient of concurrent deviations for the following data.

(i)

A	8.5	9.2	9.3	8.5	7.2	5.9	5.1	6.6	7.9	7.6
B	60	65	61	74	92	157	130	106	58	80

(ii)

A	60	80	66	29	51	72	63	55	43	41
B	55	45	62	71	58	47	50	56	63	64

(b) Pearson's Product Moment Method

Let x and y be the variables under study and let their values be denoted by $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$

Let $\bar{x} = \frac{\sum x_i}{n}$ and $\bar{y} = \frac{\sum y_i}{n}$ be their arithmetic mean.

Then the quantity $p = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$ is defined as the product moment of x and y

$(x_i - \bar{x})$ and $(y_i - \bar{y})$ are deviations of the variables from the fixed quantities $\bar{x}$ and $\bar{y}$. If the positive value of $(x - \bar{x})$ correspond to the positive value of $(y - \bar{y})$ or if the negative value of one correspond to the negative value of the other, the product $(x - \bar{x})(y - \bar{y})$ will be positive. On the other hand, if the positive value of $(x - \bar{x})$ correspond to the negative value of $(y - \bar{y})$ or vice versa, the product will be negative. Hence the mean of these values will give us a measure of correlation. Note that here we take the magnitude of the deviations also into consideration unlike the previous method of concurrent deviation.

We take the ratios $\frac{x - \bar{x}}{\sigma_x}$ and $\frac{y - \bar{y}}{\sigma_y}$ and find the mean of the product to measure

correlation where σ_x, σ_y denote the S.D of x and y respectively. Thus the coefficient of correlation 'r' is given by

$$r = \frac{1}{n} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y} \quad (1)$$

$$r = \frac{p}{\sigma_x \sigma_y} \quad (2)$$

This is known as Pearson's Product Moment Correlation Coefficient.

To compute, we have to calculate

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}, \quad \sigma_x = \frac{\sum (x_i - \bar{x})^2}{n}, \quad \sigma_y = \frac{\sum (y_i - \bar{y})^2}{n}$$

To minimise the difficulties, in computations, we choose the origin suitably as in the case of computations of $\bar{x}$ and σ .

Let x be measured from A and y be measured from B .

$$\text{Let } X = x - \bar{x}, Y = y - \bar{y}, d_1 = \bar{x} - A, d_2 = \bar{y} - B$$

$$x - A = x - \bar{x} + \bar{x} - A = X + d_1$$

$$y - B = y - \bar{y} + \bar{y} - B = Y + d_2$$

$$\begin{aligned} (x - A)(y - B) &= (X + d_1)(Y + d_2) \\ &= XY + d_1Y + d_2X + d_1d_2 \end{aligned}$$

$$\begin{aligned} \Sigma(x - A)(y - B) &= \Sigma(XY + d_1Y + d_2X + d_1d_2) \\ &= \Sigma XY + d_1 \Sigma Y + d_2 \Sigma X + nd_1d_2 \end{aligned}$$

$$= \Sigma(x - \bar{x})(y - \bar{y}) + nd_1d_2$$

$$[\therefore \Sigma x = \Sigma(x - \bar{x}) = 0 \text{ and}$$

$$\Sigma Y = \Sigma(y - \bar{y}) = 0]$$

$$\therefore (x - \bar{x})(y - \bar{y}) = \Sigma(x - A)(y - B) - nd_1d_2$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y} = \frac{\Sigma(x - A)(y - B) - nd_1d_2}{n \sigma_x \sigma_y}$$

$$\text{ie., } r = \frac{\Sigma(x - A)(y - B) - n(\bar{x} - A)(\bar{y} - B)}{n \sigma_x \sigma_y} \quad (3)$$

$$\text{If } A = 0, B = 0 \text{ then } r = \frac{\Sigma xy - n \bar{x} \bar{y}}{n \sigma_x \sigma_y} \quad (4)$$

Theorem : To prove that $-1 \leq r \leq 1$.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

$$\sigma_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} \text{ and } \sigma_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n}$$

$$r^2 = \frac{[\Sigma(x_i - \bar{x})(y_i - \bar{y})]^2}{n^2 \sigma_x^2 \sigma_y^2}$$

$$= \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{[\sum (x_i - \bar{x})^2][\sum (y_i - \bar{y})^2]}$$

$$x_i - \bar{x} = a_i \text{ and } y_i - \bar{y} = b_i$$

$$\therefore r^2 = \frac{(\sum a_i b_i)^2}{(\sum a_i^2)(\sum b_i^2)}$$

We know that the Schwartz inequality states that if a_i, b_i are n real quantities, then $(\sum a_i b_i)^2 \leq \sum a_i^2 \sum b_i^2$

$$\therefore r^2 \leq 1$$

$$\therefore |r| \leq 1$$

$$\text{ie., } -1 \leq r \leq 1$$

ie., r always lies between -1 and $+1$

Note : If $r = +1$, the correlation is perfect positive and if $r = -1$, the correlation is perfect negative. If $r = 0$, the variables are not correlated.

Theorem 2 : Correlation coefficient r is independent of change of scale and origin.

$$\text{By definition, } r_{x,y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y}$$

$$\text{Let } u = \frac{x - A}{h_1} \text{ and } v = \frac{y - B}{h_2} \text{ where } A \text{ and } B \text{ are chosen origin for } x \text{ and } y.$$

$$\text{Then } x = uh_1 + A \quad \text{————— (1)}$$

$$y = vh_2 + B \quad \text{————— (2)}$$

$$\therefore \bar{x} = \frac{\sum x}{n} = h_1 \frac{\sum u}{n} + \frac{\sum A}{n}$$

$$\text{ie., } \bar{x} = h_1 \bar{u} + A \quad \text{————— (3)}$$

$$\text{and } \bar{y} = \frac{\sum y}{n} = h_2 \bar{v} + B \quad \text{————— (4)}$$

(where $\bar{x}, \bar{y}, \bar{u}$ and $\bar{v}$ represent the corresponding arithmetic means)

$$r_{u,v} = \frac{\sum (u - \bar{u})(v - \bar{v})}{n \sigma_u \sigma_v}$$

Our claim is to prove $r_{x,y} = r_{u,v}$

Let us find the relations between (i) σ_x and σ_u

(ii) σ_y and σ_v

$$\begin{aligned}\sigma_x^2 &= \frac{\sum (x - \bar{x})^2}{n} \\ &= \frac{\sum [(uh_1 + A - (\bar{u}h_1 + A))^2]}{n} \quad (\text{by (1) and (3)}) \\ &= \frac{\sum h_1^2 (u - \bar{u})^2}{n} = h_1^2 \sigma_u^2 \quad (5)\end{aligned}$$

$$\text{Similarly } \sigma_y^2 = h_2^2 \sigma_v^2 \quad (6)$$

$$\begin{aligned}\sum (x - \bar{x})(y - \bar{y}) &= \sum (uh_1 + A - \bar{u}h_1 - A)(vh_2 + B - \bar{v}h_2 - B) \\ &= \sum h_1 h_2 (u - \bar{u})(v - \bar{v}) \\ &= h_1 h_2 \sum (u - \bar{u})(v - \bar{v}) \quad (7)\end{aligned}$$

By (5), (6) and (7)

$$\begin{aligned}r_{x,y} &= \frac{h_1 h_2 \sum (u - \bar{u})(v - \bar{v})}{n (h_1 \sigma_u) (h_2 \sigma_v)} \\ &= \frac{\sum (u - \bar{u})(v - \bar{v})}{n \sigma_u \sigma_v} = r_{u,v}\end{aligned}$$

Example 1 : Find the correlation coefficient r for the following data.

x	60	80	46	29	51	72	63	55	43	41
y	54	45	62	71	58	47	50	56	63	64

Solution : Let $A = 50$ and $B = 55$

(Choose A and B as numbers close to the mid value of the maximum and the minimum values of the series)

x	y	x - A	y - B	(x - A) ²	(y - B) ²	(x - A) (y - B)
60	54	10	-1	100	1	-10
80	45	30	-10	900	100	-300
46	62	-4	7	16	49	-28
29	71	-21	16	441	256	-336
51	58	1	3	1	9	3
72	47	22	-8	484	64	-176
63	50	13	-5	169	25	-65
55	56	5	1	25	1	5
43	63	-7	-8	49	64	-56
41	64	-9	9	81	81	-81
		40	20	2266	650	-1044

$$\bar{x} = A + \frac{\Sigma (x - A)}{n}, \quad n = 10, A = 50$$

$$\therefore \bar{x} - A = \frac{\Sigma (x - A)}{n} = \frac{40}{10} = 4$$

$$\text{ie., } d_1 = 4$$

$$\bar{y} = B + \frac{\Sigma (y - B)}{n}$$

$$\therefore d_2 = \bar{y} - B = \frac{\Sigma (y - B)}{n} = \frac{20}{10} = 20$$

$$\begin{aligned} \sigma_x^2 &= \frac{\Sigma (x - A)^2}{n} - d_1^2 \\ &= \frac{2266}{10} - 4^2 = 226.6 - 16 = 210.6 \end{aligned}$$

$$\sigma_x = \sqrt{210.6} = 14.51$$

$$\begin{aligned} \sigma_y^2 &= \frac{\Sigma (y - B)^2}{n} - d_2^2 \\ &= \frac{650}{10} - 2^2 = 65 - 4 = 61 \end{aligned}$$

$$\sigma_y = \sqrt{61} = 7.81$$

$$\begin{aligned}
 r &= \frac{\sum (x - A) (y - B) - n d_1 d_2}{n \sigma_x \sigma_y} \\
 &= \frac{-1044 - 10 (4) (2)}{10 (14.51) (7.81)} \\
 &= \frac{-1124}{(145.1) (7.81)} = -0.992
 \end{aligned}$$

Example 2: For 100 values in x and y, the following totals were got:

$$\sum x = 15,000, \quad \sum x^2 = 22,72,500$$

$$\sum y = 6,800, \quad \sum y^2 = 4,63,025$$

$$\sum xy = 10,22,250$$

Find pearson's correlation coefficient

solution

Taking $A = 0, B = 0$, r is given by

$$r = \frac{\sum xy - n \bar{x} \bar{y}}{n \sigma_x \sigma_y}$$

$$\text{where } \bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}$$

$$\sigma_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

$$\sigma_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2$$

$$\bar{x} = \frac{15000}{100} \quad (\because n = 100)$$

$$\bar{x} = 150$$

$$\bar{y} = \frac{6800}{100} = 68$$

$$\sigma_x^2 = \frac{2272500}{100} - (150)^2$$

$$= 22725 - 22500 = 225$$

$$\sigma_x = 15$$

$$\sigma_y^2 = \frac{463025}{100} - (68)^2 = 4630.25 - 4624 = 6.25$$

$$\sigma_y = 2.5$$

$$\therefore r = \frac{1022250 - 100 (150) (68)}{100 (15) (2.5)}$$

$$= \frac{1022250 - 1020000}{3750} = \frac{2250}{3750} = 0.6$$

Example 3 : Prove that $r_{x,y} = r_{u,v}$ if $x = ku + a$ and $y = mv + b$ where k, a, m, b are constants.

Solution

$$r_{x,y} = \frac{\sum (x - \bar{x}) (y - \bar{y})}{n \sigma_x \sigma_y}$$

$$r_{u,v} = \frac{\sum (u - \bar{u}) (v - \bar{v})}{n \sigma_u \sigma_v}$$

$$x = ku + a$$

$$\therefore \bar{x} = k\bar{u} + a$$

$$\sigma_x^2 = \frac{\sum (x - \bar{x})^2}{n}$$

$$= \frac{\sum (ku + a - k\bar{u} - a)^2}{n}$$

$$= \frac{k^2 \sum (u - \bar{u})^2}{n} = k^2 \sigma_u^2$$

$$\sigma_x = k \sigma_u$$

$$y = mv + b$$

$$\bar{y} = m\bar{v} + b$$

$$\therefore \sigma_y = m \sigma_v$$

$$\begin{aligned}\Sigma (x - \bar{x}) (y - \bar{y}) &= \Sigma [k (u - \bar{u}) m (v - \bar{v})] \\ &= km \Sigma (u - \bar{u}) (v - \bar{v})\end{aligned}$$

I	92	89	87	86	83
II	86	83	81	77	68

$$r_{x,y} = \frac{\Sigma (x - \bar{x}) (y - \bar{y})}{n \sigma_x \sigma_y}$$

$$= \frac{km \Sigma (u - \bar{u}) (v - \bar{v})}{n k \sigma_u m \sigma_v}$$

$$= \frac{\Sigma (u - \bar{u}) (v - \bar{v})}{n \sigma_u \sigma_v}$$

$$= r_{u,v}$$

Note :

From this we see that the correlation coefficient is independent of scale and origin.

(c) Difference Method

Given the values of x and y , we deduce the following formula for 'r' from Pearson's coefficient by taking the difference $(x - y)$

$$\text{ie., } r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{(x-y)}^2}{2\sigma_x \sigma_y}$$

Proof : Let $z = x - y$

$$\therefore \bar{z} = \frac{\Sigma z}{n} = \frac{\Sigma (x - y)}{n} = \frac{\Sigma x}{n} - \frac{\Sigma y}{n} = \bar{x} - \bar{y}$$

$$\Sigma (z - \bar{z}) = \Sigma [(x - y) - (\bar{x} - \bar{y})]^2$$

$$= \Sigma [x - \bar{x} - (y - \bar{y})]^2$$

$$= \Sigma [(x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})]$$

$$\therefore \sigma_z^2 = \frac{\Sigma (z - \bar{z})^2}{n} = \frac{1}{n} \Sigma (x - \bar{x})^2 + \frac{1}{n} \Sigma (y - \bar{y})^2 - \frac{2}{n} \Sigma (x - \bar{x})(y - \bar{y})$$

$$\text{ie., } \sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r \sigma_x \sigma_y$$

$$[\therefore r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y}]$$

$$\therefore r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{(x-y)}^2}{2 \sigma_x \sigma_y}$$

Example : Find the correlation coefficient of marks obtained by 10 students in two different subjects using difference method.

I	92	89	87	86	83	77	71	63	53	50
II	86	83	91	77	68	85	52	82	37	57

Solution : To calculate σ_x and σ_y we need A, B, d_1 and d_2 . Let us take A = 71 and B = 68.

x	y	x - A	y - B	(x - A) ²	(y - B) ²	(x - A) (y - B)	x - y	(x - y) ²
92	86	21	18	441	324	378	6	36
89	83	18	15	324	225	270	6	36
87	91	16	23	256	529	368	-4	16
86	77	15	9	225	81	135	9	81
83	68	12	0	144	0	0	15	225
77	85	6	17	36	289	102	-8	64
71	52	0	-16	0	256	0	19	361
63	82	-8	14	64	196	-112	-19	361
53	37	-18	-31	324	961	558	16	256
50	57	-21	-11	441	121	231	-7	49
		41	38	2255	2982	1930	33	1485

$$d_1 = \bar{x} - A = \frac{\Sigma x - A}{n} = \frac{41}{10} = 4.1$$

$$d_2 = \bar{y} - B = \frac{\Sigma y - B}{n} = \frac{38}{10} = 3.8$$

$$\sigma_x^2 = \frac{\Sigma (x - A)^2}{n} - d_1^2 = \frac{2255}{10} - (4.1)^2$$

$$= 225.5 - 16.81 = 208.69$$

$$\sigma_x = \sqrt{208.69} = 14.4$$

$$\sigma_y^2 = \frac{\Sigma (y - B)^2}{n} - d_2^2 = \frac{2982}{10} - (3.8)^2$$

$$= 298.2 - 14.44 = 283.76$$

$$\sigma_y = \sqrt{283.76} = 16.8$$

$$\sigma_{x-y}^2 = \frac{\sum (x - y)^2}{n} - \left(\frac{\sum (x - y)}{n} \right)^2$$

$$= \frac{1485}{10} - \left(\frac{33}{10} \right)^2$$

$$= 148.5 - 10.89 = 137.61$$

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2 \sigma_x \sigma_y}$$

$$= \frac{208.69 + 283.76 - 137.61}{2 (14.4) (16.8)}$$

$$= \frac{354.84}{483.84} = 0.73$$

3.2 Rank Correlation

When the individual values are not available but their order is determined, it is possible to determine the correlation between the ranks and this will reflect to some extent, the correlation between actual values. Here, the variates are merely the set of numbers 1, 2, ..., n in two different orders. Hence

$$\bar{x} = \bar{y} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

$$\sigma_x^2 = \sigma_y^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{\sum y^2}{n} - (\bar{y})^2$$

$$= \frac{1^2 + 2^2 + \dots + n^2}{n} - \frac{(n+1)^2}{2}$$

$$= \frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}$$

$$= \frac{n+1}{12} [2(2n+1) - 3(n+1)]$$

$$= \frac{(n+1)(n-1)}{12} = \frac{n^2 - 1}{12}$$

$$\sigma_{x-y}^2 = \frac{\sum (x-y)^2}{n} - (\bar{x} - \bar{y})^2$$

$$\begin{aligned}
 &= \frac{\sum (x-y)^2}{n} - 0 \quad [\because \bar{x} - \bar{y} = 0] \\
 \therefore r &= \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2 \sigma_x \sigma_y} \\
 &= \frac{2(n^2-1)}{12} - \frac{\sum (x-y)^2}{n} \\
 &\quad 2 \left(\frac{n^2-1}{12} \right)
 \end{aligned}$$

$$= 1 - \frac{6 \sum (x - y)^2}{n(n^2 - 1)}$$

This formula is usually denoted by ρ (read as 'rho')

$$\rho = 1 - \frac{6 \sum (x - y)^2}{n(n^2 - 1)}$$

$$\text{or } \rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \quad \text{where } d = x - y$$

Example 1: Ten competitors in a beauty contest are ranked by two judges in the following order.

x	1	6	5	10	3	2	4	9	7	8
y	6	4	9	8	1	2	3	10	5	7

On the basis of this taking, examine if the judges have common tastes in beauty.

Solution :

x	y	x-y	(x-y) ²
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			60

$$\rho = 1 - \frac{6 \sum (x-y)^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 60}{10(100-1)} = 1 - \frac{36}{99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$= 0.64$$

Since $\rho = 0.64$, x and y are positively correlated i.e., the two judges have common tastes for beauty.

Example 2 : Calculate the coefficient of rank correlation between the marks in subjects A and B given below:

A	30	50	25	30	60	70	80	65	75	98
B	50	60	20	40	70	40	90	60	40	80

Solution : Since there are ties, let us first assign the ranks to the above data.

Subject A

Rank 1 : 98

Rank 2 : 80

Rank 3 : 75

Rank 4 : 70

Rank 5 : 65

Rank 6 : 60

Rank 7 : 50

Two students secured 30 marks each. If they had secured different marks, we would have assigned the ranks 8 and 9 to them. So we assign the rank $\frac{8+9}{2} = 8.5$ to both. Next rank 10 is 25 marks.

Subject B

Rank 1 : 90

Rank 2 : 80

Rank 3 : 70

Two students secured 60 marks each. So we assign the rank $\frac{4+5}{2} = 4.5$ to both. Next mark is assigned the rank 6.

Rank 6 = 50 marks.

Next 3 students got 40 marks each. So we assign the rank $\frac{7+8+9}{2} = 8$ to all the 3 students and we assign the rank 10 to the next mark.

Rank 10 = 20 marks

Let us tabulate these values.

x	y	Rank in x	Rank in y	d = x-y	d ²
30	50	8.5	6	2.5	6.25
50	60	7	4.5	2.5	6.25
25	20	10	10	0	0
30	40	8.5	8	0.5	0.25
60	70	6	3	3	9
70	40	4	8	-4	16
80	90	2	1	1	1
65	60	5	4.5	0.5	0.25
75	40	3	8	-5	25
98	80	1	2	-1	1
					65

$$r = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(65)}{10(100-1)} = 1 - \frac{6 \times 65}{10 \times 99} = 1 - \frac{13}{33}$$

$$= \frac{20}{33} = 0.61$$

Exercise

- The following data give the prices of commodity A(x) and the prices of commodity B (y) on 10 days.

x	50	68	27	70	42	74	63	66	55	65
y	100	72	212	65	125	60	72	55	85	55

Determine the correlation coefficient r.

2. The following data give index numbers of supply X and index numbers of price Y.

X	128	115	124	134	120	130	128	116	115	132
Y	125	130	127	132	128	130	126	127	130	127

Find the rank correlation coefficient.

3. For two variates x and y the following totals were obtained

$$\Sigma(x-A)^2 = 2568 \text{ where } A = 50$$

$$\Sigma(y-B)^2 = 35868 \text{ where } B = 1130$$

$$\Sigma x = 580$$

$$\Sigma y = 900$$

$$\Sigma(x-A)(y-B) = -9179$$

Find Pearson's Correlation Coefficient.

4. The following data give the marks in Maths and the marks in English obtained by 10 students in an Examination. Determine the coefficient of correlation between ranks in the two subjects.

x	99	57	76	47	68	30	61	58	82	44
y	73	40	72	61	70	78	39	65	66	62

3.3 - Regression

The term 'regression' literally means 'stepping back towards the average'. Galton found that the off spring of abnormally tall or short parents tend to 'regress' or 'step back' to the average population height. But the term 'regression' as now used in statistics is only a convenient term referring to a measure of average relationship between the variables. With its help, we can know the average probable change in one series given a certain amount of change in the other.

If the variables in a bivariate distribution - a distribution involving two variables are related, we know that the points in the scatter diagram will cluster round some curve called the curve of regression. We try to find a functional relationship between the given variables. This functional relationship will help us to determine the probable value of one variable corresponding to a given value of the other.

If this curve of regression is a straight line, it is called a regression line and there is said to be a linear regression between the variables otherwise, regression is said to be curvi-linear. Usually we assume the regression to be linear and find the regression line using the principle of least squares.

Assuming x to be the independent variable any y , dependent we get one regression line. This line is called the regression line of y on x . Hence, if we are given any value of x we can predict the expected value of y from the above equation. Similarly, if y is independent and x is dependent on y , we get another line called regression line of x on y . In some problems, x and y may be so interrelated that due to several factors they are dependent on each other without any one being independent. In that case, we derive both equations and find the value of y for a given x from the first and find the value of x for a given y from the second equation. For example, consider I.Q. and scholastic achievement.

Regression Line of y and x

Suppose $y = ax + b$ be the equation. Using principle of least squares, we get the normal equations.

$$\sum y_i = a \sum x_i + nb \quad \text{--- (1)}$$

$$\sum x_i y_i = a \sum x_i^2 + b \sum x_i \quad \text{--- (2)}$$

$$\text{Dividing (1) by } n, \text{ we get } \frac{\sum y_i}{n} = a \frac{\sum x_i}{n} + b$$

$$\text{ie., } \bar{y} = a\bar{x} + b \quad \text{--- (3)}$$

This shows that the line of regression passes through $(\bar{x}, \bar{y})$

$$\text{We know } \sigma^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2$$

$$\therefore \frac{\sum x_i^2}{n} = \sigma^2 + \bar{x}^2 \quad \text{--- (4)}$$

$$\text{Also } r = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{n \sigma_x \sigma_y}$$

$$\text{ie., } r \sigma_x \sigma_y = \frac{\sum x_i y_i}{n} - \bar{x} \bar{y}$$

$$\text{ie., } \frac{\sum x_i y_i}{n} = r \sigma_x \sigma_y + \bar{x} \bar{y} \quad \text{--- (5)}$$

Dividing (2) by n , we get

$$\frac{\sum x_i y_i}{n} = \frac{a \sum x_i^2}{n} + \frac{b \sum x_i}{n}$$

$$\text{ie } r \sigma_x \sigma_y + \bar{x} \bar{y} = a (\sigma_x^2 + \bar{x}^2) + b \bar{x} \text{ using (4) and (5)}$$

using (3), we get

$$r \sigma_x \sigma_y + \bar{x} \bar{y} = a \sigma_x^2 + a \bar{x}^2 + (\bar{y} - a\bar{x}) \bar{x}$$

$$r \sigma_x \sigma_y + \bar{x} \bar{y} = a \sigma_x^2 + a \bar{x}^2 + \bar{x} \bar{y} - a\bar{x}^2$$

$$\therefore r \sigma_x \sigma_y = a \sigma_x^2$$

$$\text{ie } a = \frac{r \sigma_y}{\sigma_x}$$

$$b = \bar{y} - a \bar{x}$$

$$= \bar{y} - \frac{r \sigma_y}{\sigma_x} \bar{x}$$

Hence the equation $y = ax + b$ becomes

$$y = \frac{r \sigma_y}{\sigma_x} x + \bar{y} - \frac{r \sigma_y}{\sigma_x} \bar{x}$$

$$\text{ie., } y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

This equation is called the regression equation of y on x in the standard form.

Note 1 : $r \frac{\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x is denoted as b_{yx} .

Note 2 : Using $r = \frac{p}{\sigma_x \sigma_y}$ where p is the product moment, we get

$$r \frac{\sigma_y}{\sigma_x} = \frac{p \sigma_y}{\sigma_x \sigma_y \sigma_x} = \frac{p}{\sigma_x^2}$$

$$\text{ie., } b_{yx} = \frac{r \sigma_y}{\sigma_x} = \frac{p}{\sigma_x^2}$$

Note 3 : Introducing new variables $x = \frac{x - \bar{x}}{\sigma_x}$ and $y = \frac{y - \bar{y}}{\sigma_y}$, the above equation reduces to $Y = rX$. The variables X, Y are called standard normal variables. Hence in terms of the standard normal variables X, Y , the regression equation of Y on X is $Y = r X$.

Regression Line of x on y

Assuming the curve to be of the form $x = g(y)$ and in particular $x = cy + d$ we get another line. This line is called the regression line of x on y. Interchanging x and y in the above equation, we get the regression equation of x on y as

$$x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})$$

In terms of the standard normal variables, the above equation reduces to $X = r Y$.

Note 1 : $r \frac{\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y and written as b_{xy} .

Note 2 : $b_{xy} = \frac{r \sigma_x}{\sigma_y} = \frac{p}{\sigma_y^2}$

When $r = \pm 1$, the two regression equations represent the same line. For all other values of r, ie., for $-1 < r < 1$, the two lines are distinct. The first line gives the average rate of change in y for unit change in x. So, we can estimate the value of y for a given value of x using the first equation. Similarly, we get the probable value of x for a given value of y from the second equation. Since $(\bar{x}, \bar{y})$ lies on both the equations, it is the point of intersection of the two regression lines.

Example 1 : Given the following sets of data, determine the equations of the lines of regression. Find y when $x = 164$ and x when $y = 170$.

Father's ht (x)	158	160	163	165	167	170	167	172	177	181
Son's ht (y)	163	158	167	170	160	180	170	175	172	175

Solution : The regression line of y on x is

$$Y = r X$$

$$\text{ie., } \frac{y - \bar{y}}{\sigma_y} = r \frac{x - \bar{x}}{\sigma_x}$$

$$\text{ie., } y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$$

Similarly, we get the regression line of x on y as

$$x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})$$

To determine these lines we require $\bar{x}$, $\bar{y}$, σ_x , σ_y and r.

Let us calculate these values first.

Let A = 170, B = 170

x	y	x - A	y - B	(x - A) ²	(y - B) ²	(x - A)(y - B)
158	163	-12	-7	144	49	84
160	158	-10	-12	100	144	120
163	167	-7	-3	49	9	21
165	170	-5	0	25	0	0
167	160	-3	-10	9	100	30
170	180	0	10	0	100	0
167	170	-3	0	9	0	0
172	175	2	5	4	25	10
177	172	7	2	49	4	14
181	175	11	5	121	25	55
1680	1690	-20	-10	510	456	334

$$\bar{x} = \frac{1680}{10} = 168, d_1 = \bar{x} - A = 168 - 170 = -2$$

$$\bar{y} = \frac{1690}{10} = 169, d_2 = \bar{y} - B = 169 - 170 = -1$$

$$\sigma_x^2 = \frac{\sum (x-A)^2}{n} - d_1^2 = \frac{510}{10} - 4 = 47$$

$$\therefore \sigma_x = \sqrt{47} = 6.856$$

$$\sigma_y^2 = \frac{\sum (y-B)^2}{n} - d_2^2 = \frac{456}{10} - (-1)^2 = 45.6 - 1 = 44.6$$

$$\sigma_y = \sqrt{44.6} = 6.678$$

$$r = \frac{\sum (x-A)(y-B) - n d_1 d_2}{n \sigma_x \sigma_y}$$

$$r = \frac{334 - 10(-2)(-1)}{10(6.856)(6.676)}$$

$$= \frac{314}{10(6.856)(6.676)} = 0.6858$$

$$\begin{aligned} \text{Regression coefficient of } y \text{ on } x &= \frac{r \sigma_y}{\sigma_x} \\ &= \frac{(0.6858)(6.678)}{6.856} = 0.668 \end{aligned}$$

$$\begin{aligned} \text{Regression coefficient of } x \text{ on } y &= \frac{r \sigma_x}{\sigma_y} \\ &= \frac{(0.6858)(6.856)}{6.678} = 0.704 \end{aligned}$$

$$\text{Regression line of } y \text{ on } x \text{ is } y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$$

$$\begin{aligned} \text{ie., } y - 169 &= 0.668 (x - 168) \\ &= 0.668x - 112.2 \end{aligned}$$

$$\therefore y = 0.668x + 56.8$$

_____ 1

$$\text{Regression line of } x \text{ on } y \text{ is } x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})$$

$$\begin{aligned} \text{ie., } x - 168 &= 0.704 (y - 169) \\ &= 0.704y - 119 \end{aligned}$$

$$\therefore x = 0.704y + 49$$

$$\text{Given } x = 164$$

$$\begin{aligned} y &= 0.668x + 56.8 \\ &= 0.668(164) + 56.8 \\ &= 109.5 + 56.8 \\ &= 166.3 \text{ cms.} \end{aligned}$$

$$\text{Given } y = 170$$

$$\begin{aligned} x &= 0.704y + 49 \\ &= 0.704(170) + 49 \\ &= 119.68 + 49 \\ &= 168.68 \end{aligned}$$

Example 2: The following data relate to x (the amount of rainfall in inches) and y (the yield of paddy on tons per acre)

	x	y
A.M.	25"	40 tons
S.D.	3"	6 tons

$$r = 0.8$$

Determine the regression lines and estimate the value of y when $x = 22$.

Solution : Given $\bar{x} = 25"$

$$\bar{y} = 40 \text{ tons}$$

$$\sigma_x = 3"$$

$$\sigma_y = 6 \text{ tons}$$

$$r = 0.8$$

Regression coefficient of y on x is $\frac{r \sigma_y}{\sigma_x} = \frac{0.8(6)}{3} = 1.6$

Regression coefficient of x on y is $\frac{r \sigma_x}{\sigma_y} = \frac{0.8(3)}{6} = 0.4$

Regression line of y on x is $y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$

$$\begin{aligned} \text{ie., } y - 40 &= 1.6 (x - 25) \\ &= 1.6x - 40 \end{aligned}$$

$$\therefore y = 1.6x$$

Regression line of x on y is $x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})$

$$\text{ie., } x - 25 = 0.4 (y - 40)$$

$$= 0.4y - 16$$

$$\therefore x = 0.4y + 9$$

Given $x = 22$ ". To find the estimated value of y, we note that y is dependent on x. So, we consider regression equation of y on x. ie., $y = 1.6x$.

When $x = 22$, $y = (1.6) 22 = 35.2$ tons per acre.

Example 3 : For a certain bivariate distribution of 400 items, it is known that $\Sigma x = 1183$, $\Sigma y = 3748$, $\Sigma xy = 39626$, $\Sigma x^2 = 13948$, $\Sigma y^2 = 140103$. Find the regression lines.

Solution : Given that $n = 400$

$$\Sigma x = 1183, \Sigma y = 3748, \Sigma xy = 39626, \Sigma x^2 = 13948, \Sigma y^2 = 140103$$

We should calculate $\bar{x}$, $\bar{y}$, σ_x , σ_y , p to get the regression line.

$$\bar{x} = \frac{\Sigma x}{n}$$

$$\bar{x} = \frac{1183}{400} = 2.96$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{3748}{400} = 9.37$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{13948}{400} - (2.96)^2 = 34.87 - 8.76 = 26.11$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - (\bar{y})^2 = \frac{140103}{400} - (9.37)^2 = 350.26 - 87.8 = 262.46$$

$$p = \frac{\Sigma xy}{n} - \bar{x} \bar{y} = \frac{39626}{400} - (2.96)(9.37) = 99.07 - 27.74 = 71.33$$

$$\therefore b_{yx} = \frac{r \sigma_y}{\sigma_x} = \frac{p}{\sigma_x^2} = \frac{71.33}{26.11} = 2.73$$

$$b_{xy} = \frac{r \sigma_x}{\sigma_y} = \frac{p}{\sigma_y^2} = \frac{71.33}{262.46} = 0.272$$

$\therefore$ The regression lines are

$$y - 9.37 = 2.73 (x - 2.96)$$

$$x - 2.96 = 0.272 (y - 9.37)$$

Properties of the Regression Coefficients

Property 1: Both regression lines $y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$ and $x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})$ pass through the point $(\bar{x}, \bar{y})$

This property is useful in the sense that given the equations of the regression lines, we will be able to calculate the mean values of the variables.

Property 2 : 'r', the correlation coefficient is the geometric mean of the regression coefficients.

Proof : G.M. of the regression coefficients

$$= \sqrt{\frac{r \sigma_y}{\sigma_x} \times \frac{r \sigma_x}{\sigma_y}} = \sqrt{r^2} = r$$

= the correlation coefficient .

Corollary : If one of the regression coefficient is greater than unity, then the other must be less than unity.

Proof : $b_{yx} b_{xy} = r^2 \leq 1$. Hence the result.

Property 3 : The angle θ between the regression lines is given by

$$\tan \theta = \pm \frac{1 - r^2}{r} \left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \right)$$

Proof : The equation of the regression lines are

$$y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x}) \quad \text{--- (1) and}$$

$$x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y}) \quad \text{--- (2)}$$

$$\text{Slope of (1)} = \frac{r \sigma_y}{\sigma_x} = m_1 \text{ (say)}$$

$$\text{Slope of (2)} = \frac{\sigma_y}{r \sigma_x} = m_2 \text{ (say)}$$

We know from analytical geometry that the angle θ between two straight lines whose slopes are m_1, m_2 is given by

$$\tan \theta = \pm \frac{m_2 - m_1}{1 + m_1 m_2}$$

$$\begin{aligned} \therefore \tan \theta &= \frac{\pm \frac{\sigma y}{r \sigma x} - \frac{r \sigma y}{\sigma x}}{1 + \frac{\sigma y}{r \sigma x} \frac{r \sigma y}{\sigma x}} \\ &= \pm \frac{\frac{\sigma y}{\sigma x} \left(\frac{1}{r} - r \right)}{1 + \left(\frac{\sigma y}{\sigma x} \right)^2} \\ &= \pm (1 - r^2) \frac{\sigma y}{\sigma x} \frac{r \left(\frac{\sigma x^2 + \sigma y^2}{\sigma x^2} \right)}{r \left(\frac{\sigma x^2 + \sigma y^2}{\sigma x^2} \right)} \\ &= \pm \frac{1 - r^2}{r} \left(\frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2} \right) \end{aligned}$$

Corollary 1 : Suppose $r = \pm 1$

$$\text{Then } \tan \theta = 0$$

$$\therefore \theta = 0$$

So, the lines are parallel. But both pass through $(\bar{x}, \bar{y})$. Hence they coincide.

This means that when the correlation between the variables is perfect positive or perfect negative the regression lines coincide.

Corollary 2 : Suppose $r = 0$ then $\tan \theta = \infty$ i.e., $\theta = 90^\circ$

So the lines are perpendicular to each other i.e., when the correlation between the variables is zero the regression lines are perpendicular to each other.

Property 4 : The regression coefficients are both positive or both negative according as the correlation coefficient is positive or negative $b_{xy} = \frac{r \sigma y}{\sigma x}$ and $b_{yx} = \frac{r \sigma x}{\sigma y}$. $\sigma x, \sigma y$ being standard deviations, are always taken as positive, numbers. Hence the result.

Standard error of Estimate

The standard error of estimate of y on x , is defined as the square root of the arithmetic mean of the squares of the residuals from the regression line of y on x . It is denoted by S_y .

Regression line of y on x is

$$y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$$

$$\text{ie., } y = \bar{y} + \frac{r \sigma_y}{\sigma_x} (x - \bar{x})$$

$$S_y^2 = \frac{1}{n} \sum_i \left\{ y_i - \left(\bar{y} + \frac{r \sigma_y}{\sigma_x} (x_i - \bar{x}) \right) \right\}^2$$

$$= \frac{1}{n} \sum_i \left\{ (y_i - \bar{y}) - r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \right\}^2$$

$$= \frac{1}{n} \sum \left\{ (y_i - \bar{y})^2 + r^2 \frac{\sigma_y^2}{\sigma_x^2} (x_i - \bar{x})^2 - 2r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) (y_i - \bar{y}) \right\}$$

$$= \frac{1}{n} \sum (y_i - \bar{y})^2 + r^2 \frac{\sigma_y^2}{\sigma_x^2} \frac{1}{n} \sum (x_i - \bar{x})^2 - 2r \frac{\sigma_y}{\sigma_x} \frac{1}{n} \sum (x_i - \bar{x}) (y_i - \bar{y})$$

$$= \sigma_y^2 + r^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 - 2r \frac{\sigma_y}{\sigma_x} r \sigma_x \sigma_y$$

$$= \sigma_y^2 + r^2 \sigma_y^2 - 2r^2 \sigma_y^2 = \sigma_y^2 - r^2 \sigma_y^2$$

$$\therefore S_y^2 = \sigma_y^2 (1 - r^2)$$

Similarly, denoting by S_x , the standard error of estimate of x on y , we get the formula.

$$S_x^2 = \sigma_x^2 (1 - r^2)$$

Corollary : $|r| \leq 1$

Proof : $S_y^2 \geq 0$

$$\text{ie., } \sigma_y^2 (1 - r^2) \geq 0$$

But since $\sigma_y^2 \geq 0$ always, $(1 - r^2) \geq 0$

$$\text{ie., } r^2 \leq 1.$$

$$\text{ie., } |r| \leq 1.$$

Example 1 : Two lines of regression are $x + 2y = 5$ and $2x + 3y = 8$. Find $\bar{x}$, $\bar{y}$ and r .

Solution : $(\bar{x}, \bar{y})$ is the point of intersection of the two regression lines. Hence solving these equations, we get $(\bar{x}, \bar{y})$

$$x + 2y = 5 \quad \text{--- (1)}$$

$$2x + 3y = 8 \quad \text{--- (2)}$$

$$(1) \times 2 \quad 2x + 4y = 10$$

$$(2) \quad 2x + 3y = 8$$

Subtracting, $y = 2$

$$x + 2y = 5 \Rightarrow x + 4 = 5$$

$$\therefore x = 1$$

$$\therefore \bar{x} = 1, \bar{y} = 2$$

Of the two equations given, we do not know which one is the regression line of y on x and which is the other line. So, we rewrite the equations in such a way that the product of the regression coefficient is less than one.

$x + 2y = 5$ is rewritten as

$$2y = -x + 5$$

$$y = -\frac{1}{2}x + \frac{5}{2} \quad \text{--- (3)}$$

$2x + 3y = 8$ is rewritten as

$$2x = -3y + 8$$

$$x = -\frac{3}{2}y + 4 \quad \text{--- (4)}$$

If these two expressions represent the two lines, then the regression coefficients will be

$(-\frac{1}{2})$ and $(-\frac{3}{2})$ and the product $\frac{3}{4} < 1$. Hence we have the regression coefficient as

$(-\frac{1}{2})$ and $(-\frac{3}{2})$.

[Writing the two equations as $x = -2y + 5$ and $y = -\frac{2}{3}x + \frac{8}{3}$ we see that the product of the coefficients $= (-2)(-\frac{2}{3}) = \frac{4}{3} > 1$]

We know $r^2 =$ product of regression coefficients

$$= (-\frac{1}{2})(-\frac{3}{2}) = \frac{3}{4}$$

$$\therefore r = \pm \frac{\sqrt{3}}{2}$$

Since the regression coefficients and r have the same sign $r = -\frac{\sqrt{3}}{2} = -0.866$

Example 2 : In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible : Variance of $x = 9$. The regression equations are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$. What are the mean values of x and y , S.D. of y and the coefficient of correlation. Find also the standard errors of estimate.

Proof : Regression lines are

$$8x - 10y + 66 = 0$$

$$40x - 18y - 214 = 0$$

$$4x - 5y + 33 = 0 \quad \text{———— (1)}$$

$$20x - 9y - 107 = 0 \quad \text{———— (2)}$$

$$(1) \times 5 \quad 20x - 25y + 165 = 0$$

$$(2) \quad \underline{20x - 9y - 107 = 0}$$

$$\text{Subtracting} \quad -16y + 272 = 0$$

$$16y = 272$$

$$y = 17$$

$y = 17$ Substituting in (1) we get

$$4x - 5(17) + 33 = 0$$

$$4x = 85 - 33 = 52$$

$$x = 13$$

$$\therefore \bar{x} = 13, \quad \bar{y} = 17$$

Rewriting the equations, so that the product of the regression coefficients is less than or equal to 1, we have the regression line of y on x as

$$y = \frac{8}{10}x + \frac{66}{10} \text{ and the regression line of x on y as } x = \frac{18}{40}y + \frac{214}{40}$$

$$\therefore b_{yx} = \frac{8}{10} = \frac{4}{5} = r \frac{\sigma_y}{\sigma_x}$$

$$\text{and } b_{xy} = \frac{18}{40} = \frac{9}{20} = r \frac{\sigma_x}{\sigma_y}$$

$$r^2 = b_{yx} b_{xy} = \frac{4}{5} \cdot \frac{9}{20} = \frac{36}{100} = 0.36$$

$$r^2 = \pm 0.6$$

Since the regression coefficient are positive $r = 0.6$

$$\text{Variance of } x = \sigma_x^2 = 9 \text{ (given)}$$

$$\therefore \sigma_x = 3$$

$$\frac{r \sigma_y}{\sigma_x} = \frac{4}{5}$$

$$\therefore 0.6 \frac{\sigma_y}{3} = \frac{4}{5}$$

$$\sigma_y = \frac{4}{5} \times \frac{3}{0.6} = 4$$

Standard error of estimate of y on x = S_y .

$$\begin{aligned} \text{where } S_y^2 &= \sigma_y^2 (1-r^2) \\ &= 16(1 - 0.36) = 16(0.64) \end{aligned}$$

$$\therefore S_y = 4(0.8) = 3.2$$

Standard error of estimate of x on y = S_x

$$\begin{aligned} \text{where } S_x^2 &= \sigma_x^2 (1-r^2) \\ &= 9(1-0.36) = 9(0.64) \end{aligned}$$

$$\therefore S_x = 3(0.8) = 2.4$$

Exercise

1. Find the two regression lines for the following data.

x	22	24	26	26	27	28	29	30	31	32	34	35
y	18	20	20	24	22	27	21	29	27	27	27	31

2. For the heights (y) and the weight (x) of 100 students the following results were found.
 $\bar{x} = 150$ lbs, $\bar{y} = 68$ inches $\sigma_x = 20$ lbs, $\sigma_y = 2.5$ inches, $r = 0.6$. John weights 200 lbs and Roe is 5 feet tall. Estimate the height of John from his weight and the weight of Roe from his height.
3. The regression lines have equations $5x = 6y + 24$ and $1000y = 768x - 3608$. Find the correlation coefficient.
4. The two regression equations are $3x + 2y - 26 = 0$ and $6x + y - 21 = 0$. Find $\bar{x}$, $\bar{y}$, r . If $\sigma_x^2 = 25$ find σ_y . Find also the standard errors of estimate.
5. The two lines of regression are given by $3x - 10y + 40 = 0$; $10x - 27y + 82 = 0$ and the variance of x is 9, Obtain the means of x and y, variance of y and r.
6. The regression lines of x and y are given by $2y - x = 50$ and $3y - 2x = 10$. Find their respective means and correlation coefficient. If $\sigma_x = 3$, find σ_y also.
7. Two variables give the following data.
 $\bar{x} = 20$, $\bar{y} = 15$, $\sigma_x = 4$, $\sigma_y = 3$, $r = +0.7$. Write the two regression equations and obtain the most likely value of y when $x = 24$.

Test marks	18	19	20	21
200-250	4	4	2	1
250-300	3	5	4	2
300-350	2	6	8	5
350-400	1	4	6	10
Total	10	19	20	18

3.4 - Correlation Coefficient for a bivariate frequency distribution

When the number of observations is larger the data are often classified into two way frequency distribution called a correlation table.

The formula for calculating the coefficient of correlation is :

$$r = \frac{N \sum f dx dy - \sum f dx \sum f dy}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \sqrt{N \sum f dy^2 - (\sum f dy)^2}}$$

- Steps:**
- Take the step deviations of variable X and denote these deviations by dx.
 - Take the step deviations of the variable Y and denote these deviations by dy.
 - Multiply dx dy and the respective frequency of each cell and write the figure obtained in right hand upper corner of each cell.
 - Add together all the cornered values as calculated in step (iii) and obtain the total $\sum f dx dy$.
 - Multiply the frequencies of the variable X by the deviations of X and obtain the total $\sum f dx$.
 - Take the squares of the deviations of variable X and multiply them by the respective frequencies and obtain $\sum f dx^2$.
 - Multiply the frequencies of the variable Y by the deviations of Y and obtain the total $\sum f dy$.
 - Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and obtain $\sum f dy^2$.
 - Substitute the values of $\sum f dx dy$, $\sum f dx$, $\sum f dx^2$, $\sum f dy$, and $\sum f dy^2$ in the above formula and obtain the value of r.

Example: The following table gives the frequency according to groups of marks obtained by 67 students in an intelligence test. Measure the degree of relationship between age and intelligence test :

	Age in years				Total
Test marks	18	19	20	21	
200-250	4	4	2	1	11
250-300	3	5	4	2	14
300-350	2	6	8	5	21
350-400	1	4	6	10	21
Total	10	19	20	18	67

Solution:

X Y	Age in Years				f	fdy	fdy ²	fdxdy
	dy	dx	18	19	20	21		
			-1	0	1	2		
200-250 -1		+4	4	0	2	1	11	-11
250-300 0		0	3	5	4	2	14	0
300-350 1		-2	2	6	8	5	21	21
350-400 2		-2	1	4	6	10	21	42
Total			10	19	20	18	N = 67	Σfdy = 52
fdx			-10	0	20	36	Σfdx = 46	
fdx ²			10	0	20	72	Σfdx ² = 102	
fdxdy			0	0	18	48	Σfdxdy = 66	

$$r = \frac{N \sum fdx dy - \sum fdx \sum fdy}{\sqrt{N \sum fdx^2 - (\sum fdx)^2} \sqrt{N \sum fdy^2 - (\sum fdy)^2}}$$

$$= \frac{67 \times 66 - 46 \times 52}{\sqrt{67 \times 102 - (46)^2} \sqrt{67 \times 116 - (52)^2}} = \frac{4422 - 2392}{\sqrt{4718} \sqrt{5068}}$$

$$= \frac{2030}{68.688 \times 71.19} = 0.415$$

Probable error or correlation coefficient

The probable error of the coefficient of correlation helps in interpreting its value. If r is the correlation coefficient in a sample of n pairs of observations, then its standard error is given by

$$\text{S.E. } (r) = \frac{1 - r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is given by $\text{P.E. } (r) = 0.6745 \times \text{S.E.}$

The reason for taking the factor 0.6745 is that in a normal distribution, the range $\mu \pm 0.6745\sigma$ covers 50% of the total area.

If $r < \text{P.E. } (r)$, correlation is not at all significant. If $r > 6 \times \text{P.E. } (r)$ it is definitely significant.

Exercise: 1. Calculate the coefficient of correlation from the following data using 44 and 26 as the origin of X and Y respectively.

X : 43 44 46 40 44 42 45 42 38 40 42 57

Y : 29 31 19 18 19 27 27 29 41 30 26 10

2) Calculate the coefficient of correlating from the following data using correlation table.

X \ Y	200-300	300-400	400-500	500-600	600-700
10-15	—	—	—	3	7
15-20	—	4	9	4	3
20-25	7	6	12	5	—
25-30	3	10	19	8	—

Unit 4

4.1 - Interpolation

Interpolation is a fundamental operation in Mathematics. Suppose the following table represents a set of corresponding values of two quantities x and y .

$x :$	x_0	x_1	x_2	x_3		x_n
$y :$	y_0	y_1	y_2	y_3		y_n

We may require the value of $y = y_r$ corresponding to a value $x = x_r$ between x_0 and x_n . Thus interpolation denotes the process of computing the values of a function for any value of the independent variable within an interval for which some values are given. The term extrapolation is used to denote the process of finding the values outside this interval. However, the term interpolation is applied to both processes.

Let $y = f(x)$ be a function taking exactly the values $y_0, y_1, \dots, y_n$ corresponding to the values $x_0, x_1, \dots, x_n$ of the independent variable x . If the function $f(x)$ is determined, then the value of y corresponding to any x can be readily calculated from $y = f(x)$. But in most practical problems, it will be very difficult to determine the function $y = f(x)$ in an exact form, with the help of the given values of x and y . In such cases, we replace $f(x)$ by a simpler function which assumes the same values of $f(x)$ and from which the other values may be computed to the desired degree of accuracy. This process of replacing or representing a given function by a simpler one constitutes interpolation in a broader sense of the term.

Thus, let $\phi(x)$ denote an arbitrary simpler function so constructed that it has the same values as $f(x)$ for $x_0, x_1, \dots, x_n$. Then if $f(x)$ is replaced by $\phi(x)$ over a given interval, the process constitutes interpolation. The function $\phi(x)$ is said to be an interpolating formula or interpolating function.

Methods of Interpolation

There are two methods of interpolation (i) Graphic method (ii) Algebraic method

Graphic Method

In this method, each observation (x, y) is represented by a point and all such points are plotted on a graph sheet. These points are connected by a smooth curve. To find the value of y corresponding to any value of x , we draw an ordinate through the point $(x, 0)$ and find the y coordinate of the point where this line meets the curve. This value of y gives an estimate of the interpolated value.

Algebraic method

Various algebraic methods have been developed by which interpolation of figures can be done. The important algebraic methods used, can be grouped into the following three categories.

- (i) Linear interpolation and Parabola interpolation
- (ii) Methods of Finite differences (Newton's formula)
- (iii) Methods applicable in case of unequal intervals of the independent variable. (Lagrange's formula)

(i) Linear interpolation: The simplest of all interpolation is the case in which the interpolating formula is linear. In linear interpolation we are approximating the graph of $y = f(x)$ by a straight line passing through two successive points where $f(x)$ is known.

If only two observations are given by (x_1, y_1) and (x_2, y_2) we can easily find the value of y corresponding to any value of x in the interval (x_1, x_2) using the equation to the line through (x_1, y_1) , (x_2, y_2) .

The equation to the line joining two points is given by
$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}$$

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} (x - x_1) \quad (1)$$

For any value of x , we can get the corresponding value of y from (1)

Example 1: In an examination, 245 persons got more than 35% and 193 got more than 48 %. Estimate as to have many got more than 40%.

Solution: $x_1 = 35, y_1 = 245$

$x_2 = 48, y_2 = 193$

$x = 40, y = ?$

$$\begin{aligned} y &= y_1 + \frac{y_2 - y_1}{x_2 - x_1} (x - x_1) \\ &= 245 + \frac{193 - 245}{48 - 35} (x - 35) \end{aligned}$$

$$= 245 - \frac{52}{13} (x - 35)$$

$$= 245 - 4(x - 35)$$

$$= 245 - 4x + 140$$

$y = 385 - 4x$. This is the formula for interpolation.

$$\begin{aligned} \text{When } x &= 40, y = 385 - 4(40) \\ &= 385 - 160 = 225 \end{aligned}$$

$\therefore$ 225 persons got more than 40%

- (ii) If suppose $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ are three known observations, then we can find the equation to the second degree parabola

$$y = a_0 + a_1x + a_2x^2 \text{ passing through these three points } y = a_0 + a_1x + a_2x^2 \quad (2)$$

Substituting the co-ordinates of the points in (2), we get three simultaneous equations in the unknowns a_0, a_1, a_2 .

$$y_1 = a_0 + a_1x_1 + a_2x_1^2$$

$$y_2 = a_0 + a_1x_2 + a_2x_2^2$$

$$y_3 = a_0 + a_1x_3 + a_2x_3^2$$

Solving these equations, we get for any value of x we can find y .

Similarly, if four pairs of observations are given, we fit a cubic parabola.

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 \text{ which passes through these points.}$$

In general, given $(n+1)$ pairs of observations, we fit an n^{th} degree polynomial

$$y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n \text{ passing through these points.}$$

Using this equation, we can find the value of y , for any given value of x .

Example 2: Complete the record given below of the export of a certain commodity during five years.

year	1977	1978	1979	1980	1981
Export (in tons)	443	384	?	397	467

Since 4 pairs of values are known, we fit a cubic parabola $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$.

To simplify the calculations, we take the x value corresponding to the missing value as zero. ie., we take 1979 as origin and take the x values for other years as $-2, -1, 1, 2$ respectively. So the observations are $(-2, 443), (-1, 384), (0, y_0), (1, 397), (2, 467)$

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

Substituting these values we get,

$$a_0 - 2a_1 + 4a_2 - 8a_3 = 443 \quad (1)$$

$$a_0 - a_1 + a_2 - a_3 = 384 \quad (2)$$

$$a_0 = y_0 \quad (3)$$

$$a_0 + a_1 + a_2 + a_3 = 397 \quad (4)$$

$$a_0 + 2a_1 + 4a_2 + 8a_3 = 467 \quad (5)$$

$$(1) + (5) \Rightarrow 2a_0 + 8a_2 = 910 \quad (6)$$

$$(2) + (4) \Rightarrow 2a_0 + 2a_2 = 781 \quad (7)$$

$$(6) - (7) \Rightarrow 6a_2 = 129$$

$$a_2 = \frac{129}{6} = \frac{43}{2}$$

Substituting in (7) we get $2a_0 + 43 = 781$

$$2a_0 = 781 - 43 = 738$$

$$a_0 = 369$$

Since we need only y_0 and not the equation we need not solve for a_1, a_2 .

$\therefore$ The export for 1979 is 369 tons.

4.2 - Method of finite differences

We use a simpler method to find the approximating polynomial and interpolate. For explaining this method, we define two operators Δ and E .

Generally, if x is an independent variable and y is the corresponding dependent variable, we denote the relation as $y = f(x)$ or $\phi(x)$. Let us use the notation $y = u_x$.

Let $a, a+h, a+2h, \dots, a+nh$ be the $(n+1)$ values taken by x and $u_a, u_{a+h}, \dots, u_{a+nh}$ be the corresponding $(n+1)$ values of y . Then the differences $(u_{a+h} - u_a), (u_{a+2h} - u_{a+h})$ etc are termed as the first differences of u_x and are denoted as $\Delta u_a, \Delta u_{a+h}$ etc. In general $\Delta u_x = u_{x+h} - u_x$. Here Δ is called as the first difference operator and h is the interval of difference.

The first difference of u_x namely Δu_x can be subjected to the same process again. So, we get the 2nd differences of u_x .

$$\begin{aligned}\text{Symbolically, } \Delta^2 u_x &= \Delta (\Delta u_x) = \Delta (u_{x+h} - u_x) \\ &= (u_{x+2h} - u_{x+h}) - (u_{x+h} - u_x)\end{aligned}$$

In a similar way we can get the third, fourth etc differences of u_x and they are respectively denoted as $\Delta^3 u_x, \Delta^4 u_x$ etc.

Example 1: Find $\Delta u_x, \Delta^2 u_x$ where $u_x = x^2$

Solution: $\Delta u_x = u_{x+h} - u_x$

$$= (x+h)^2 - x^2 = h^2 + 2hx$$

$$\Delta^2 u_x = \Delta (\Delta u_x) = \Delta (h^2 + 2hx)$$

$$\text{Put } h^2 + 2hx = V_x$$

$$\begin{aligned}\therefore \Delta V_x &= V_{x+h} - V_x = h^2 + 2h(x+h) - (h^2 + 2hx) \\ &= 2hx + 2h^2 - 2hx = 2h^2.\end{aligned}$$

Example 2: Find Δu_x when (i) $u_x = x(x-1)(x-2)(x-3)$ and (ii) $u_x = \frac{1}{x(x+1)(x+2)}$ the interval of difference being unity in each case.

Solution: (i) Here $h = 1$

$$\text{and } u_x = x(x-1)(x-2)(x-3)$$

$$\Delta u_x = u_{x+h} - u_x = u_{x+1} - u_x$$

$$= (x-1)(x+1-1)(x+1-2)(x+1-3) - x(x-1)(x-2)(x-3)$$

$$= (x+1)x(x-1)(x-2) - x(x-1)(x-2)(x-3)$$

$$= x(x-1)(x-2)[x+1 - x+3]$$

$$= 4x(x-1)(x-2)$$

$$(ii) \quad u_x = \frac{1}{x(x+1)(x+2)}$$

$$h=1$$

$$\therefore U_{x+1} = \frac{1}{(x+1)(x+2)(x+3)}$$

$$\Delta U_x = U_{x+1} - U_x$$

$$= \frac{1}{(x+1)(x+2)(x+3)} - \frac{1}{x(x+1)(x+2)}$$

$$= \frac{1}{(x+1)(x+2)} \left[\frac{1}{x+3} - \frac{1}{x} \right]$$

$$= \frac{-3}{x(x+1)(x+2)(x+3)}$$

Properties of the operator Δ

The difference operator Δ satisfies the following properties.

$$(i) \quad \Delta(u_x \pm v_x) = \Delta u_x \pm \Delta v_x$$

$$(ii) \quad \Delta(ku_x) = k\Delta U_x$$

$$(iii) \quad \Delta^m \Delta^n U_x = \Delta^n \Delta^m U_x = \Delta^{m+n} U_x$$

Proof: (i) Let $w_x = u_x + v_x$

$$\begin{aligned}\therefore \Delta(u_x + v_x) &= w_{x+h} - w_x \\ &= (u_{x+h} + v_{x+h}) - (u_x + v_x) \\ &= (u_{x+h} - u_x) + (v_{x+h} - v_x) \\ &= \Delta u_x + \Delta v_x\end{aligned}$$

In a similar way we can prove $\Delta(u_x - v_x) = \Delta u_x - \Delta v_x$

(ii) Let $w_x = ku_x$

$$\begin{aligned}\therefore \Delta(ku_x) &= \Delta w_x \\ &= w_{x+h} - w_x \\ &= ku_{x+h} - ku_x = k(u_{x+h} - u_x) \\ &= k(\Delta u_x)\end{aligned}$$

(iii) $\Delta^m \Delta^n U_x = \Delta^m (\Delta^n U_x)$

$$\begin{aligned}&= \Delta^m (\underbrace{\Delta, \Delta, \dots, \Delta}_{n \text{ times}} U_x) \\ &= (\underbrace{\Delta, \Delta, \dots, \Delta}_{n \text{ times}}) (\underbrace{\Delta, \Delta, \dots, \Delta}_{n \text{ times}} U_x)\end{aligned}$$

Example 1: Given $U_0 = 2, U_1 = 6, U_2 = 14, U_3 = 30, U_4 = 50, U_5 = 76, U_6 = 110, U_7 = 154, U_8 = 210, U_9 = 280, U_{10} = 366, U_{11} = 470, U_{12} = 594, U_{13} = 740, U_{14} = 910, U_{15} = 1106, U_{16} = 1340, U_{17} = 1614, U_{18} = 1930, U_{19} = 2298, U_{20} = 2720, U_{21} = 3206, U_{22} = 3758, U_{23} = 4378, U_{24} = 5068, U_{25} = 5830, U_{26} = 6676, U_{27} = 7608, U_{28} = 8638, U_{29} = 9768, U_{30} = 11000, U_{31} = 12346, U_{32} = 13808, U_{33} = 15488, U_{34} = 17390, U_{35} = 19526, U_{36} = 21900, U_{37} = 24526, U_{38} = 27408, U_{39} = 30550, U_{40} = 33966, U_{41} = 37660, U_{42} = 41646, U_{43} = 45928, U_{44} = 50510, U_{45} = 55406, U_{46} = 60630, U_{47} = 66186, U_{48} = 72078, U_{49} = 78320, U_{50} = 84926, U_{51} = 91910, U_{52} = 99286, U_{53} = 107068, U_{54} = 115270, U_{55} = 123916, U_{56} = 133020, U_{57} = 142606, U_{58} = 152690, U_{59} = 163298, U_{60} = 174446, U_{61} = 186150, U_{62} = 198426, U_{63} = 211290, U_{64} = 224766, U_{65} = 238870, U_{66} = 253626, U_{67} = 269050, U_{68} = 285166, U_{69} = 301990, U_{70} = 319546, U_{71} = 337860, U_{72} = 356958, U_{73} = 376866, U_{74} = 397610, U_{75} = 419216, U_{76} = 441706, U_{77} = 465106, U_{78} = 489440, U_{79} = 514734, U_{80} = 541014, U_{81} = 568306, U_{82} = 596636, U_{83} = 626030, U_{84} = 656514, U_{85} = 688114, U_{86} = 720866, U_{87} = 754796, U_{88} = 789930, U_{89} = 826296, U_{90} = 863920, U_{91} = 902830, U_{92} = 943062, U_{93} = 984654, U_{94} = 1027634, U_{95} = 1072030, U_{96} = 1117870, U_{97} = 1165182, U_{98} = 1213994, U_{99} = 1264346, U_{100} = 1316270, U_{101} = 1369796, U_{102} = 1424954, U_{103} = 1481782, U_{104} = 1540310, U_{105} = 1600578, U_{106} = 1662616, U_{107} = 1726454, U_{108} = 1792122, U_{109} = 1859660, U_{110} = 1929106, U_{111} = 1999502, U_{112} = 2070890, U_{113} = 2143310, U_{114} = 2216802, U_{115} = 2291406, U_{116} = 2367162, U_{117} = 2444110, U_{118} = 2522290, U_{119} = 2601742, U_{120} = 2682506, U_{121} = 2764622, U_{122} = 2848130, U_{123} = 2933070, U_{124} = 3019482, U_{125} = 3107406, U_{126} = 3196882, U_{127} = 3287950, U_{128} = 3380642, U_{129} = 3474990, U_{130} = 3571034, U_{131} = 3668816, U_{132} = 3768386, U_{133} = 3869786, U_{134} = 3973058, U_{135} = 4078242, U_{136} = 4185382, U_{137} = 4294522, U_{138} = 4405706, U_{139} = 4518986, U_{140} = 4634406, U_{141} = 4751918, U_{142} = 4871566, U_{143} = 4993402, U_{144} = 5117470, U_{145} = 5243802, U_{146} = 5372442, U_{147} = 5503430, U_{148} = 5636810, U_{149} = 5772626, U_{150} = 5910922, U_{151} = 6051750, U_{152} = 6195152, U_{153} = 6341182, U_{154} = 6489890, U_{155} = 6641322, U_{156} = 6795530, U_{157} = 6952566, U_{158} = 7112482, U_{159} = 7275330, U_{160} = 7441162, U_{161} = 7609930, U_{162} = 7781696, U_{163} = 7956512, U_{164} = 8134430, U_{165} = 8315502, U_{166} = 8499782, U_{167} = 8687322, U_{168} = 8878186, U_{169} = 9072426, U_{170} = 9269996, U_{171} = 9470950, U_{172} = 9675342, U_{173} = 9883226, U_{174} = 10094656, U_{175} = 10309686, U_{176} = 10528370, U_{177} = 10750762, U_{178} = 10976916, U_{179} = 11206896, U_{180} = 11440766, U_{181} = 11678582, U_{182} = 11920402, U_{183} = 12166290, U_{184} = 12416310, U_{185} = 12670536, U_{186} = 12929034, U_{187} = 13191870, U_{188} = 13459110, U_{189} = 13730826, U_{190} = 14007086, U_{191} = 14287966, U_{192} = 14573542, U_{193} = 14863890, U_{194} = 15159096, U_{195} = 15459246, U_{196} = 15764426, U_{197} = 16074722, U_{198} = 16390220, U_{199} = 16710996, U_{200} = 17037136, U_{201} = 17368726, U_{202} = 17704852, U_{203} = 18045600, U_{204} = 18391056, U_{205} = 18741306, U_{206} = 19096446, U_{207} = 19456572, U_{208} = 19821780, U_{209} = 20192156, U_{210} = 20567786, U_{211} = 20948656, U_{212} = 21334862, U_{213} = 21726400, U_{214} = 22123356, U_{215} = 22525726, U_{216} = 22933606, U_{217} = 23347082, U_{218} = 23766250, U_{219} = 24191196, U_{220} = 24622006, U_{221} = 25058776, U_{222} = 25501602, U_{223} = 25950580, U_{224} = 26405806, U_{225} = 26867376, U_{226} = 27335386, U_{227} = 27809932, U_{228} = 28291110, U_{229} = 28779016, U_{230} = 29273746, U_{231} = 29775396, U_{232} = 30284062, U_{233} = 30799840, U_{234} = 31322826, U_{235} = 31853116, U_{236} = 32390806, U_{237} = 32935992, U_{238} = 33488670, U_{239} = 34048936, U_{240} = 34616886, U_{241} = 35192616, U_{242} = 35776222, U_{243} = 36367800, U_{244} = 36967446, U_{245} = 37575256, U_{246} = 38191326, U_{247} = 38815752, U_{248} = 39448630, U_{249} = 40089956, U_{250} = 40739826, U_{251} = 41398336, U_{252} = 42065582, U_{253} = 42741660, U_{254} = 43426676, U_{255} = 44120726, U_{256} = 44823916, U_{257} = 45536342, U_{258} = 46258100, U_{259} = 46989286, U_{260} = 47729996, U_{261} = 48480326, U_{262} = 49240372, U_{263} = 50010230, U_{264} = 50790006, U_{265} = 51579796, U_{266} = 52379606, U_{267} = 53189532, U_{268} = 53999670, U_{269} = 54819916, U_{270} = 55640366, U_{271} = 56471016, U_{272} = 57311872, U_{273} = 58162940, U_{274} = 59024316, U_{275} = 59896096, U_{276} = 60778276, U_{277} = 61670952, U_{278} = 62574220, U_{279} = 63488176, U_{280} = 64412916, U_{281} = 65348536, U_{282} = 66295042, U_{283} = 67252540, U_{284} = 68221126, U_{285} = 69200906, U_{286} = 70191886, U_{287} = 71194062, U_{288} = 72207440, U_{289} = 73232116, U_{290} = 74268186, U_{291} = 75315746, U_{292} = 76374892, U_{293} = 77445720, U_{294} = 78528336, U_{295} = 79622746, U_{296} = 80729056, U_{297} = 81847272, U_{298} = 82977490, U_{299} = 84119806, U_{300} = 85284326, U_{301} = 86461146, U_{302} = 87650372, U_{303} = 88852000, U_{304} = 90066126, U_{305} = 91292846, U_{306} = 92532256, U_{307} = 93784462, U_{308} = 95049470, U_{309} = 96327386, U_{310} = 97618306, U_{311} = 98922326, U_{312} = 100239542, U_{313} = 101569960, U_{314} = 102913676, U_{315} = 104270686, U_{316} = 105641096, U_{317} = 107025012, U_{318} = 108422530, U_{319} = 109833756, U_{320} = 111258686, U_{321} = 112697416, U_{322} = 114150042, U_{323} = 115616660, U_{324} = 117097376, U_{325} = 118592186, U_{326} = 120101196, U_{327} = 121624512, U_{328} = 123162230, U_{329} = 124714446, U_{330} = 126281266, U_{331} = 127862686, U_{332} = 129458702, U_{333} = 131069420, U_{334} = 132694836, U_{335} = 134335046, U_{336} = 135990156, U_{337} = 137660272, U_{338} = 139345490, U_{339} = 141045906, U_{340} = 142761626, U_{341} = 144492746, U_{342} = 146239362, U_{343} = 148001580, U_{344} = 149779496, U_{345} = 151573116, U_{346} = 153382536, U_{347} = 155207852, U_{348} = 157049170, U_{349} = 158906586, U_{350} = 160780106, U_{351} = 162669826, U_{352} = 164575842, U_{353} = 166498260, U_{354} = 168437176, U_{355} = 170392696, U_{356} = 172364916, U_{357} = 174353932, U_{358} = 176359850, U_{359} = 178382766, U_{360} = 180422786, U_{361} = 182480006, U_{362} = 184554526, U_{363} = 186646442, U_{364} = 188755860, U_{365} = 190882876, U_{366} = 193027596, U_{367} = 195189916, U_{368} = 197369936, U_{369} = 199567752, U_{370} = 201783470, U_{371} = 204017186, U_{372} = 206268902, U_{373} = 208538720, U_{374} = 210826736, U_{375} = 213133056, U_{376} = 215457776, U_{377} = 217799996, U_{378} = 220159816, U_{379} = 222537436, U_{380} = 224932956, U_{381} = 227346476, U_{382} = 229778096, U_{383} = 232227916, U_{384} = 234696136, U_{385} = 237182856, U_{386} = 239688176, U_{387} = 242212196, U_{388} = 244754916, U_{389} = 247316436, U_{390} = 249896856, U_{391} = 252496276, U_{392} = 255114796, U_{393} = 257752416, U_{394} = 260409236, U_{395} = 263085356, U_{396} = 265780876, U_{397} = 268495896, U_{398} = 271230516, U_{399} = 273984836, U_{400} = 276758956, U_{401} = 279552976, U_{402} = 282366996, U_{403} = 285191116, U_{404} = 288035436, U_{405} = 290899956, U_{406} = 293784776, U_{407} = 296689896, U_{408} = 299615416, U_{409} = 302561536, U_{410} = 305528356, U_{411} = 308515976, U_{412} = 311524496, U_{413} = 314553916, U_{414} = 317604336, U_{415} = 320675756, U_{416} = 323768176, U_{417} = 326881596, U_{418} = 329916016, U_{419} = 332971436, U_{420} = 336047856, U_{421} = 339145276, U_{422} = 342263696, U_{423} = 345403116, U_{424} = 348563536, U_{425} = 351744956, U_{426} = 354947376, U_{427} = 358170796, U_{428} = 361415216, U_{429} = 364680636, U_{430} = 367967056, U_{431} = 371274476, U_{432} = 374602896, U_{433} = 377952316, U_{434} = 381322736, U_{435} = 384714156, U_{436} = 388126576, U_{437} = 391559996, U_{438} = 395014416, U_{439} = 398489836, U_{440} = 401986256, U_{441} = 405503676, U_{442} = 409042096, U_{443} = 412601516, U_{444} = 416181936, U_{445} = 419783356, U_{446} = 423405776, U_{447} = 427049196, U_{448} = 430713616, U_{449} = 434399036, U_{450} = 438105456, U_{451} = 441832876, U_{452} = 445581296, U_{453} = 449350716, U_{454} = 453141136, U_{455} = 456952556, U_{456} = 460784976, U_{457} = 464638396, U_{458} = 468512816, U_{459} = 472408236, U_{460} = 476324656, U_{461} = 480262076, U_{462} = 484220496, U_{463} = 488199916, U_{464} = 492199336, U_{465} = 496218756, U_{466} = 500258176, U_{467} = 504317596, U_{468} = 508397016, U_{469} = 512496436, U_{470} = 516615856, U_{471} = 520755276, U_{472} = 524914696, U_{473} = 529094116, U_{474} = 533293536, U_{475} = 537512956, U_{476} = 541752376, U_{477} = 546011796, U_{478} = 550291216, U_{479} = 554590636, U_{480} = 558910056, U_{481} = 563249476, U_{482} = 567608896, U_{483} = 571988316, U_{484} = 576387736, U_{485} = 580807156, U_{486} = 585246576, U_{487} = 589705996, U_{488} = 594185416, U_{489} = 598684836, U_{490} = 603204256, U_{491} = 607743676, U_{492} = 612303096, U_{493} = 616882516, U_{494} = 621481936, U_{495} = 626101356, U_{496} = 630740776, U_{497} = 635400196, U_{498} = 640079616, U_{499} = 644779036, U_{500} = 649498456, U_{501} = 654237876, U_{502} = 658997296, U_{503} = 663776716, U_{504} = 668576136, U_{505} = 673395556, U_{506} = 678234976, U_{507} = 683094396, U_{508} = 687973816, U_{509} = 692873236, U_{510} = 697792656, U_{511} = 702732076, U_{512} = 707691496, U_{513} = 712670916, U_{514} = 717670336, U_{515} = 722689756, U_{516} = 727729176, U_{517} = 732788596, U_{518} = 737868016, U_{519} = 742967436, U_{520} = 748086856, U_{521} = 753226276, U_{522} = 758385696, U_{523} = 763565116, U_{524} = 768764536, U_{525} = 773983956, U_{526} = 779223376, U_{527} = 784482796, U_{528} = 789762216, U_{529} = 795061636, U_{530} = 800381056, U_{531} = 805720476, U_{532} = 811079896, U_{533} = 816459316, U_{534} = 821858736, U_{535} = 827278156, U_{536} = 832717576, U_{537} = 838176996, U_{538} = 843656416, U_{539} = 849155836, U_{540} = 854675256, U_{541} = 860214676, U_{542} = 865774096, U_{543} = 871353516, U_{544} = 876952936, U_{545} = 882572356, U_{546} = 888211776, U_{547} = 893871196, U_{548} = 899550616, U_{549} = 905250036, U_{550} = 910969456, U_{551} = 916708876, U_{552} = 922468296, U_{553} = 928247716, U_{554} = 934047136, U_{555} = 939866556, U_{556} = 945705976, U_{557} = 951565396, U_{558} = 957444816, U_{559} = 963344236, U_{560} = 969263656, U_{561} = 975203076, U_{562} = 981162496, U_{563} = 987141916, U_{564} = 993141336, U_{565} = 999160756, U_{566} = 1005200176, U_{567} = 1011259596, U_{568} = 1017339016, U_{569} = 1023438436, U_{570} = 1029557856, U_{571} = 1035697276, U_{572} = 1041856696, U_{573} = 1048036116, U_{574} = 1054235536, U_{575} = 1060454956, U_{576} = 1066694376, U_{577} = 1072953796, U_{578} = 1079233216, U_{579} = 1085532636, U_{580} = 1091852056, U_{581} = 1098191476, U_{582} = 1104550896, U_{583} = 1110930316, U_{584} = 1117329736, U_{585} = 1123749156, U_{586} = 1130188576, U_{587} = 1136647996, U_{588} = 1143127416, U_{589} = 1149626836, U_{590} = 1156146256, U_{591} = 1162685676, U_{592} = 1169245096, U_{593} = 1175824516, U_{594} = 1182423936, U_{595} = 1189043356, U_{596} = 1195682776, U_{597} = 1202342196, U_{598} = 1209021616, U_{599} = 1215721036, U_{600} = 1222440456, U_{601} = 1229179876, U_{602} = 1235939296, U_{603} = 1242718716, U_{604} = 1249518136, U_{605} = 1256337556, U_{606} = 1263176976, U_{607} = 1270036396, U_{608} = 1276915816, U_{609} = 1283815236, U_{610} = 1290734656, U_{611} = 1297674076, U_{612} = 1304633496, U_{613} = 1311612916, U_{614} = 1318612336, U_{615} = 1325631756, U_{616} = 1332671176, U_{617} = 1339730596, U_{618} = 1346809016, U_{619} = 1353907436, U_{620} = 1361025856, U_{621} = 1368164276, U_{622} = 1375322696, U_{623} = 1382501116, U_{624} = 1389699536, U_{625} = 1396917956, U_{626} = 1404156376, U_{627} = 1411414796, U_{628} = 1418693216, U_{629} = 1425991636, U_{630} = 1433310056, U_{631} = 1440648476, U_{632} = 1448006896, U_{633} = 1455385316, U_{634} = 1462783736, U_{635} = 1470202156, U_{636} = 1477640576, U_{637} = 1485098996, U_{638} = 1492577416, U_{639} = 1500075836, U_{640} = 1507594256, U_{641} = 1515132676, U_{642} = 1522691096, U_{643} = 1530269516, U_{644} = 1537867936, U_{645} = 1545486356, U_{646} = 1553124776, U_{647} = 1560783196, U_{648} = 1568461616, U_{649} = 1576160036, U_{650} = 1583878456, U_{651} = 1591616876, U_{652} = 1599375296, U_{653} = 1607153716, U_{654} = 1614952136, U_{655} = 1622770556, U_{656} = 1630608976, U_{657} = 1638467396, U_{658} = 1646345816, U_{659} = 1654244236, U_{660} = 1662162656, U_{661} = 1670101076, U_{662} = 1678059496, U_{663} = 1686037916, U_{664} = 1694036336, U_{665} = 1702054756, U_{666} = 1710093176, U_{667} = 1718151596, U_{668} = 1726229016, U_{669} = 1734326436, U_{670} = 1742443856, U_{671} = 1750581276, U_{672} = 1758738696, U_{673} = 1766916116, U_{674} = 1775113536, U_{675} = 1783330956, U_{676} = 1791568376, U_{677} = 1800000000, U_{678} = 1808446000, U_{679} = 1816906400, U_{680} = 1825381200, U_{681} = 1833870400, U_{682} = 1842374000, U_{683} = 1850892000, U_{684} = 1859424400, U_{685} = 1867971200, U_{686} = 1876532400, U_{687} = 1885108000, U_{688} = 1893698000, U_{689} = 1$

$$E^2 u_x = E(Eu_x)$$

$$= E(u_{x+h}) = u_{x+2h}$$

$$E^3 u_x = u_{x+3h} \text{ and so on.}$$

$$\text{In general } E^n U_x = U_{x+nh}$$

The operator E also satisfies certain properties. They are

$$(i) \quad E(U_x \pm V_x) = EU_x \pm EV_x$$

$$(ii) \quad E(kU_x) = kEU_x$$

$$(iii) \quad E^m E^n (U_x) = E^n E^m (U_x) = E^{m+n} (U_x)$$

Proof: (i) Let $W_x = U_x + V_x$

$$\therefore W_{x+h} = U_{x+h} + V_{x+h}$$

$$\therefore E(U_x + V_x) = E(W_x) = W_{x+h} = U_{x+h} + V_{x+h}$$

$$= EU_x + EV_x$$

Similarly, we can prove $E(U_x - V_x) = E(U_x) - E(V_x)$

(ii) Let $kU_x = W_x$

$$\therefore W_{x+h} = kU_{x+h}$$

$$E(kU_x) = E(W_x) = kU_{x+h} = k(E(U_x))$$

$$\therefore E(kU_x) = kE(U_x)$$

(iii) $E^m E^n (U_x) = E^m (EE \dots n \text{ times}) U_x$

$$= (EE \dots m \text{ times}) (EE \dots n \text{ times}) U_x$$

$$= (EE \dots m+n \text{ times}) U_x$$

$$= E^{m+n} (U_x)$$

$$= E^{n+m} (U_x) = E^n E^m (U_x)$$

The relation between E and Δ

We know $\Delta U_x = U_{x+h} - U_x$

$$\text{and } EU_x = U_{x+h}$$

$$\therefore \Delta U_x = EU_x - U_x = (E-1) U_x$$

$$\therefore \Delta = E-1 \text{ or } E = 1 + \Delta$$

We have seen earlier, when $U_x = x^2$, a polynomial of degree 2 then $\Delta U_x = 2x$ a polynomial of degree 1. $\Delta^2 U_x = 2$ constant and all other higher differences from Δ^3 is zero.

In fact, we can generalize this result for any polynomial of degree n . If the degree of a polynomial is n then the n th order difference will be a constant and higher order differences will be zero. $\therefore \Delta^{n+1} U_x = 0$.

We know $\Delta = E - 1$

$$\therefore (E-1)^{n+1} U_x = 0$$

Expanding this expression using binomial theorem and using the result $E^k U_x = U_{x+kh}$ we get the required results.

We usually take $x=0$ and $h=1$

$$\therefore E^k U_0 = U_k$$

Example 1: Given $U_0 = -8, U_1 = -2, U_2 = 6, U_3 = 40, U_4 = 148$ find $\Delta^4 U_0$ and $\Delta^3 U_1$.

Solution: $\Delta^4 U_0 = (E-1)^4 U_0$

$$= (E^4 - 4E^3 + 6E^2 - 4E + 1) U_0$$

$$= E^4 U_0 - 4E^3 U_0 + 6E^2 U_0 - 4E U_0 + U_0$$

$$= U_4 - 4U_3 + 6U_2 - 4U_1 + U_0$$

$$= 148 - 4(40) + 6(6) - 4(-2) - 8$$

$$= 148 - 160 + 36 + 8 - 8$$

$$= 24$$

$$\Delta^3 U_1 = (E-1)^3 U_1$$

$$= (E^3 - 3E^2 + 3E - 1) U_1$$

$$= E^3 U_1 - 3E^2 U_1 + 3E U_1 - U_1$$

$$= U_4 - 3U_3 + 3U_2 - U_1$$

$$= 148 - 120 + 18 + 2$$

$$= 48.$$

Example 2: Find $\Delta^4 U_0$ and $\Delta^2 U_2$ given $U_0 = 6$, $U_1 = 7$, $U_2 = 16$, $U_3 = 39$, $U_4 = 82$

Solution: $\Delta^4 U_0 = (E-1)^4 U_0$

$$= (E^4 - 4E^3 + 6E^2 - 4E + 1) U_0$$

$$= E^4 U_0 - 4E^3 U_0 + 6E^2 U_0 - 4E U_0 + U_0$$

$$= U_4 - 4U_3 + 6U_2 - 4U_1 + U_0$$

$$= 82 - 4(39) + 96 - 28 + 6$$

$$= 82 - 156 + 96 - 28 + 6$$

$$= 0.$$

$$\Delta^2 U_2 = (E-1)^2 U_2$$

$$= (E^2 - 2E + 1) U_2$$

$$= E^2 U_2 - 2E U_2 + U_2$$

$$= U_4 - 2U_3 + U_2$$

$$= 82 - 78 + 16$$

$$= 20.$$

Example 3: Given $U_0 = 2$, $U_2 = 4$, $U_3 = 10$, $U_4 = 22$. Find U_1 .

Solution: Since 4 values are given, we assume u_x to be a polynomial of degree 3.

$$\therefore \Delta^4 U_x = 0$$

$$\therefore \Delta^4 U_0 = 0.$$

$$\text{ie., } (E - 1)^4 U_0 = 0$$

$$(E^4 - 4E^3 + 6E^2 - 4E + 1)U_0 = 0$$

$$U_4 - 4U_3 + 6U_2 - 4U_1 + U_0 = 0$$

$$22 - 40 + 24 - 4U_1 + 2 = 0$$

$$4U_1 = 8$$

$$\therefore U_1 = 2.$$

Example 4: Given $U_0 = -4$, $U_1 = -2$, $U_4 = 220$, $U_5 = 546$, $U_6 = 1148$ find U_2 and U_3 .

Solution: Since we are given five values, we can assume U_x to be a polynomial of degree 4.

$$\therefore \Delta^5 U_x = 0 \text{ for all } x.$$

$$\therefore \Delta^5 U_0 = 0, \Delta^5 U_1 = 0$$

$$\Delta^5 U_0 = (E - 1)^5 U_0 = 0.$$

$$\text{ie., } (E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)U_0 = 0$$

$$E^5 U_0 - 5E^4 U_0 + 10E^3 U_0 - 10E^2 U_0 + 5E U_0 - U_0 = 0$$

$$\text{ie., } U_5 - 5U_4 + 10U_3 - 10U_2 + 5U_1 - U_0 = 0$$

$$546 - 1100 + 10U_3 - 10U_2 - 10 + 4 = 0.$$

$$\text{ie., } 10U_3 - 10U_2 = 560$$

$$\Delta^5 U_1 = (E - 1)^5 U_1$$

$$\text{ie., } E^5 U_1 - 5E^4 U_1 + 10E^3 U_1 - 10E^2 U_1 + 5E U_1 - U_1 = 0.$$

$$U_6 - 5U_5 + 10U_4 - 10U_3 + 5U_2 - U_1 = 0.$$

$$1148 - 5(546) + 10(220) - 10U_3 + 5U_2 + 2 = 0.$$

$$-10U_3 + 5U_2 = -620 \quad (2)$$

$$(1) + (2) \Rightarrow -5U_2 = -60.$$

$$U_2 = 12$$

$$10U_3 - 120 = 560$$

$$10U_3 = 680$$

$$U_3 = 68$$

$$\therefore U_2 = 12, \quad U_3 = 68.$$

Example 5: Find the figure of sales for the year 1983 from the data given below:

Year	1977	1978	1979	1980	1981	1982
Sales (Thousand Units)	120	150	160	180	200	225

Solution: Let $U_0, U_1, U_2, U_3, U_4, U_5$ denote 120, 150, 160, 180, 200, 225 respectively.

Since 6 values are given, we assume U_x to be a polynomial of degree 6.

$$\therefore \Delta^6 U_x = 0$$

$$\therefore \Delta^6 U_0 = 0$$

$$\text{ie., } (E - 1)^6 U_0 = 0.$$

$$\text{ie., } E^6 U_0 - 6E^5 U_0 + 15E^4 U_0 - 20E^3 U_0 + 15E^2 U_0 - 6EU_0 + U_0 = 0.$$

$$\text{ie., } U_6 - 6U_5 + 15U_4 - 20U_3 + 15U_2 - 6U_1 + U_0 = 0$$

$$U_6 - 6(225) + 15(200) - 20(180) + 15(160) - 6(150) + 120 = 0.$$

$$\text{ie., } U_6 - 1350 + 3000 - 3600 + 2400 - 900 + 120 = 0.$$

$$U_6 = (1350 + 3600 + 900) - (3000 + 2400 + 120)$$

$$= 5850 - 5520 = 330.$$

$\therefore$ Expected sales for the year 1983 = 330 Thousand units.

Exercise

1. Given U_x find ΔU_x , $\Delta^2 U_x$ for $h = 1$.

(i) $U_x = x^3 + 3x^2$

(ii) $U_x = x(x+2)(x+3)$

2. (i) Find the missing figure;

Year	1950	1955	1960	1965	1970	1975	1980
(Production)	100	120	150	?	210	?	320

(ii)

x	1966	1968	1970	1972	1974
y	85	?	151	264	388

3. The annual sales of a concern are given below

Year	1975	1977	1979	1981	1983
Sales	125	163	204	238	282

Extrapolate the sales for the year 1985

4.3 = Gregory - Newton Forward Interpolation Formula

Let $Y_x = f(x)$ be a function which takes the values $y_0, y_1, \dots, y_n$ corresponding to the values $x_0, x_1, x_2, \dots, x_n$ of the independent variable x .

Let the values of x be at equidistant intervals. i.e., Let $x_i = x_0 + ih$ ($i = 0, 1, 2, \dots, n$) Let $y(x)$ be a polynomial of the n th degree in x , taking the same values as y_x corresponding to $x = x_0, x_1, x_2, \dots, x_n$.

i.e., $y(x_0) = y_0, y(x_1) = y_1, \dots$

We can assume for convenience, that

$$\begin{aligned}
 y(x) = & a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) \\
 & + a_3(x - x_0)(x - x_1)(x - x_2) + \dots \\
 & + a_n(x - x_0)(x - x_0)(x - x_1)(x - x_2), \dots (x - x_{n-1})
 \end{aligned} \tag{1}$$

This polynomial contains the $n+1$ constants $a_0, a_1, \dots, a_n$ which can be determined as follows:

Putting $x = x_0, x_1, \dots, x_n$ successively in (1), we get $y_0 = y(x_0) = a_0$ (i)

$$y_1 = y(x_1) = a_0 + a_1(x_1 - x_0) \quad \text{(ii)}$$

$$y_2 = y(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) \quad \text{(iii)}$$

and so on. From these, the values of $a_0, a_1, a_2, \dots$ can be found in terms of y_0 and its forward differences.

$$a_0 = y_0 \text{ from (i)}$$

$$\text{We know } \Delta y_0 = y_1 - y_0 = a_1(x_1 - x_0) = a_1 h$$

$$\therefore a_1 = \frac{\Delta y_0}{h}$$

$$\Delta y_1 = y_2 - y_1 = a_1(x_2 - x_1) + a_2(x_2 - x_0)(x_2 - x_1)$$

$$= a_1 h + a_2 \cdot 2h \cdot h$$

$$\Delta y_1 = \Delta y_0 + 2a_2 h^2$$

$$\therefore h_2 = \frac{\Delta y_1 - \Delta y_0}{2h^2} = \frac{\Delta^2 y_0}{2! h^2}$$

$$\Delta^3 y_0$$

Similarly $a_3 = \frac{\Delta^3 y_0}{3! h^3}$ and so on.

Putting these values in (1) we get,

$$\begin{aligned} y(x) = y_0 &+ \frac{\Delta y_0}{h} (x - x_0) + \frac{\Delta^2 y_0}{2! h^2} (x - x_0)(x - x_1) \\ &+ \frac{\Delta^3 y_0}{3! h^3} (x - x_0)(x - x_1)(x - x_2) + \dots \end{aligned} \quad (2)$$

In (2), let us put $x - x_0 = ph$

$$\text{Then } x - x_1 = x - x_0 + x_0 - x_1$$

$$= ph - (x_1 - x_0) = ph - h = (p - 1)h$$

$$x - x_2 = x - x_1 + x_1 - x_2 = x - x_1 - (x_2 - x_1)$$

$$= (p - 1)h - h = (p - 2)h \text{ and so on.}$$

Hence (2) becomes

$$y(x_0 + p_h) = y_0 + \frac{\Delta y_0}{h} ph + \frac{\Delta^2 y_0}{2! h^2} ph(p - 1)h$$

$$+ \frac{\Delta^3 y_0}{3! h^3} ph(p - 1)h(p - 2)h + \dots$$

$$\text{ie., } y_p = y_0 + p \Delta y_0 + \frac{p(p-1)}{2!} \Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!} \Delta^3 y_0 + \dots$$

$$+ \frac{p(p-1)(p-2) \dots (p - (n - 1))}{n!} \Delta^n y_0 \quad (3)$$

(3) is known as Gregory - Newton forward difference formula

Example: The following are data from the steam table :

Temp° C:	140	150	160	170	180
Pressure:	3.685	4.864	6.302	8.076	10.225
kgf / Om ²¹					

Using Newton's formula, find the pressure of the steam for a temperature of 142°.

Solution: 142° is near the beginning of the table. Hence we use Newton's forward formula. The forward differences are

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
140	3.685				
		1.169			
150	4.854		0.279		
		1.448		0.047	
160	6.302		0.326		0.002
		1.774		0.049	
170	8.076		0.375		
		2.149			
180	10.225				

$$y = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!} \Delta^3 y_0 + \frac{p(p-1)(p-2)(p-3)}{4!} \Delta^4 y_0$$

$$x_0 = 140, y_0 = 3.685, \Delta y_0 = 1.169, \Delta^2 y_0 = 0.279, \Delta^3 y_0 = 0.047, \Delta^4 y_0 = 0.002.$$

Let y_p be the value of y when $x = 142$.

$$\therefore p = \frac{x - x_0}{h} = \frac{142 - 140}{10} = \frac{2}{10} = 0.2$$

$$\therefore y_p = 3.685 + 0.2 \times 1.169 + \frac{0.2(0.2-1)}{2} \times 0.279$$

$$+ \frac{0.2(0.2-1)(0.2-2)}{1.2.3} \times 0.047$$

$$+ \frac{0.2(0.2-1)(0.2-2)(0.2-3)}{1.2.3.4} \times 0.002$$

$$= 3.685 + 0.2338 - 0.02232 + 0.002256 - 0.0000672$$

$$= 3.899$$

$$y_p = 3.899$$

Note:

The initial value y_0 can be taken anywhere in the table.

If we take $y_0 = 4.854$, then

$$\Delta y_0 = 1.448, \quad \Delta^2 y_0 = 0.326, \quad \Delta^3 y_0 = 0.049 \quad \text{and} \quad x_0 = 150$$

$$p = \frac{x - x_0}{h} = \frac{142 - 150}{10} = \frac{-8}{10} = -0.8$$

Gregory - Newton Backward Interpolation Formula

Let $y_x = f(x)$ be a function which takes the values $y_0, y_1, \dots, y_n$ corresponding to the values $x_0, x_1, \dots, x_n$ of the independent variable x . Let the values of x be at equidistant intervals. i.e., let $x_i = x_0 + ih$ ($i = 0, 1, 2, \dots, n$). Let $y(x)$ be a polynomial of the n^{th} degree in x , taking the same values as y_x corresponding to $x = x_0, x_1, \dots, x_n$. i.e., $y(x_0) = y_0, y(x_1) = y_1, \dots, y(x_n) = y_n$. We can assume, for convenience, that $y(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \dots + a_n(x - x_n)(x - x_{n-1}) \dots (x - x_1)$ (1)

This polynomial contains the $n+1$ constants $a_0, a_1, \dots, a_n$ which can be determined as follows.

$$\text{Putting } x = x_n, x_{n-1}, \dots, x_0 \text{ successively in (1), we get } y_n = y(x_n) = a_0 \quad (\text{i})$$

$$y_{n-1} = y(x_{n-1}) = a_0 + a_1(x_{n-1} - x_n) \quad (\text{ii})$$

$$y_{n-2} = y(x_{n-2}) = a_0 + a_1(x_{n-2} - x_n) + a_2(x_{n-2} - x_n)(x_{n-2} - x_{n-1}) \quad (\text{iii})$$

and so on.

From these, the values of $a_0, a_1, a_2, \dots$, can be found in terms of y_n and its backward differences.

$$a_0 = y_n \text{ from (i)}$$

$$\nabla y_n = y_n - y_{n-1} = -a_1(x_{n-1} - x_n)$$

(where ∇ is the operator which denotes the backward difference)

$$\nabla y_n = a_1 (x_n - x_{n-1}) = a_1 h$$

$$\therefore a_1 = \frac{\nabla y_n}{h}$$

$$\nabla y_{n-1} = y_{n-1} - y_{n-2}$$

$$= a_1 (x_{n-1} - x_{n-2}) - a_2 (x_{n-2} - x_n) (x_{n-2} - x_{n-1})$$

$$= a_1 h - a_2 (-2h) (-h)$$

$$= \nabla y_n - 2h^2 a_2$$

$$\therefore a_2 = \frac{\nabla y_n - \nabla y_{n-1}}{2h^2} = \frac{\nabla^2 y_n}{2! h^2}$$

$$\nabla^3 y_n$$

Similarly $a_3 = \frac{\nabla^3 y_n}{3! h^3}$ and so on.

Putting these values in (1), we get $y(x) = y_n + \frac{\nabla y_n}{h} (x - x_n) + \frac{\nabla^2 y_n}{2! h^2} (x - x_n)(x - x_{n-1}) + \frac{\nabla^3 y_n}{3! h^3} (x - x_n)(x - x_{n-1})(x - x_{n-2}) + \dots \dots \dots (2)$

In (2) let us put $x - x_n = ph$

$$\text{Then } x - x_{n-1} = x - x_n + x_n - x_{n-1} = ph + h = (p+1)h$$

$$x - x_{n-2} = x - x_{n-1} + x_{n-1} - x_{n-2}$$

$$= (p+1)h + h = (p+2)h \text{ and so on.}$$

Hence (2) becomes

$$y(x_n + ph) = y_n + \frac{\nabla y_n}{h} ph + \frac{\nabla^2 y_n}{2! h^2} ph (p+1)h$$

$$+ \frac{\nabla^3 y_n}{3! h^3} p h (p+1) h (p+2) h + \dots$$

$$\text{ie., } y_p = y_n + p \nabla y_n + \frac{p(p+1)}{2!} \nabla^2 y_n + \frac{p(p+1)(p+2)}{3!} \nabla^3 y_n$$

$$+ \frac{p(p+1)(p+2) \dots (p+n-1)}{n!} \nabla^n y_n \quad (3)$$

(3) is known as Gregory - Newton backward difference formula.

Example:

The following data gives the melting point of an alloy of lead and zinc, where t is temperature in deg C and p is the % of lead in the alloy.

p : 40 50 60 70 80 90

t: 184 204 226 250 276 304

Using Newton's interpolation formula, find the melting point of the alloy containing 84% of lead.

Solution: Since 84 is near the end of the table, we use Newton's backward formula

The backward difference are :

p	t	∇t	$\nabla^2 t$	$\nabla^3 t$	$\nabla^4 t$	$\nabla^5 t$
40	184	20				
50	204	22	2			
60	226	24	2	0		
70	250	26	2	0	0	
80	276	28	2	0	0	0
90	304					

$$t = t_n + k \nabla t_n + \frac{k(k+1)}{1.2} \nabla^2 t_n$$

$$p_n = 90, \quad t_n = 304, \quad \nabla t_n = 28, \quad \nabla^2 t_n = 2.$$

Let t_{84} be the value of t when $p = 84$.

$$k = \frac{p - p_n}{h} = \frac{84 - 90}{10} = \frac{-6}{10} = -0.6$$

$$t_{84} = 304 - 0.6(28) + \frac{(-0.6)(-0.6+1)}{1.2} \times 2$$

$$= 304 - 16.8 - 0.24$$

$$t_{84} = 287 \text{ app.}$$

Interpolation with unequal intervals Divided differences

Let the function $y = f(x)$ take the values $f(x_0), f(x_1), f(x_2), \dots, f(x_n)$ corresponding to the values $x_0, x_1, x_2, \dots, x_n$ of the argument x where $x_1 - x_0, x_2 - x_1, \dots, x_n - x_{n-1}$ need not be necessarily equal.

The first divided difference of $f(x)$ for the arguments x_0, x_1 , is defined as $\frac{f(x_1) - f(x_0)}{x_1 - x_0}$

and is denoted by $f[x_0, x_1]$ or by $\Delta_{x_1} f(x_0)$ or by $[x_0, x_1]$

$$\therefore f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \quad (i)$$

The second divided difference of $f(x)$ for three arguments x_0, x_1, x_2 , is defined as

$$f[x_0, x_1, x_2] = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0} \quad (ii)$$

This is also denoted by $\Delta_{x_1, x_2}^2 f(x_0)$. The third divided difference of $f(x)$ for the four arguments x_0, x_1, x_2, x_3 is defined as

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$$

The quantities in (i), (ii), and (iii) are called divided differences of orders 1, 2, 3 respectively. We can represent them in a divided difference table as follows.

Argument	Entry	1st divided diff	2nd divided diff	3rd divided diff
x	$f[.]$	$f[.,.]$	$f[.,.,.]$	$f[.,.,.,.]$
x_0	$f[x_0]$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
x_1	$f[x_1]$	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	$f[x_1, x_2, x_3, x_4]$
x_2	$f[x_2]$	$f[x_2, x_3]$	$f[x_2, x_3, x_4]$	
x_3	$f[x_3]$	$f[x_3, x_4]$		
x_4	$f[x_4]$			

Example 1: Find out the divided difference of y_x given that

$x :$	1	2	4	7	12
$y_x :$	22	30	82	106	206

Solution: The divided difference table is

x	y _x	[.]	[.,.]	[.,.,.]	[.,.,.,.]
1	22	$\frac{30-22}{2-1} = 8$	$\frac{26-8}{4-1} = 6$	$\frac{-3.6-6}{7-1} = -1.6$	
2	30	$\frac{82-30}{4-2} = 26$		$\frac{0.51+1.6}{12-1} = 0.19$	
4	82		$\frac{8-26}{7-2} = -3.6$		
7	106	$\frac{106-82}{7-4} = 8$		$\frac{1.5+3.6}{12-2} = 0.51$	
12	206	$\frac{206-106}{12-7} = 20$	$\frac{20-8}{12-4} = 1.5$		

Example 2: Prove that the third divided differences with arguments a, b, c, d of the function $\frac{1}{x}$ is equal to $-\frac{1}{abcd}$

Solution: Here $x_0 = a$, $x_1 = b$, $x_2 = c$, $x_3 = d$ $f(x) = \frac{1}{x}$.

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Example: Given the data

$$= \frac{\frac{1}{b} - \frac{1}{a}}{b - a} = -\frac{1}{ab}$$

$$f[x_0, x_1, x_2] = \frac{f[x_0, x_1] - f[x_1, x_2]}{x_0 - x_2}$$

Solution: The divided difference table is

$$= \frac{\frac{1}{ab} - \frac{1}{bc}}{a - c} = \frac{a - c}{abc(a - c)} = \frac{1}{abc}$$

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$$

$$= \frac{\frac{1}{bcd} - \frac{1}{abc}}{d - a}$$

$$= \frac{a - d}{abcd(d - a)} = -\frac{1}{abcd}$$

Newton's interpolation formula for unequal intervals

Let $f(x_0), f(x_1), \dots, f(x_n)$ be the values of $f(x)$ corresponding to the arguments $x_0, x_1, \dots, x_n$ not necessarily equally spaced. From the definition of divided differences, we have

$$f[x, x_0] = \frac{f(x) - f(x_0)}{x - x_0}$$

$$\therefore f(x) = f(x_0) + (x - x_0) f[x, x_0] \quad (1)$$

$$f[x, x_0, x_1] = \frac{f[x, x_0] - f[x_0, x_1]}{x - x_1}$$

$$\therefore f[x, x_0] = f[x_0, x_1] + (x - x_1) f[x, x_0, x_1]$$

Substituting this value of $f[x, x_0]$ in (1) we get

$$f(x) = f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x, x_0, x_1] \quad (2)$$

$$f[x, x_0, x_1, x_2] = \frac{f[x, x_0, x_1] - f[x_0, x_1, x_2]}{x - x_2}$$

$$\therefore f[x, x_0, x_1] = f[x_0, x_1, x_2] + (x - x_2) f[x, x_0, x_1, x_2]$$

Substituting this value of $f[x, x_0, x_1]$ in (2) we have

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f(x_0, x_1) + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &\quad + (x - x_0)(x - x_1)(x - x_2) f(x, x_0, x_1, x_2) \end{aligned} \quad (3)$$

Proceeding in this manner, we get,

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &\quad + (x - x_0)(x - x_1)(x - x_2) f[x_0, x_1, x_2, x_3] + \dots \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x_0, x_1, \dots, x_n] \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_n) f[x, x_0, x_1, \dots, x_n] \end{aligned} \quad (4)$$

If the function $f(x)$ is a polynomial of degree n , the $(n+1)^{\text{th}}$ divided difference will be zero. So, the last term in (4) i.e.,

$$f[x_0, x_1, x_2, \dots, x_n] = 0$$

Hence (4) becomes

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &\quad + \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x_0, x_1, \dots, x_n] \end{aligned} \quad (5)$$

The formula (5) is called Newton's divided difference interpolation formula :

Example: Given the data

x	0	1	2	5
f(x)	2	3	12	147

find the cubic function of x.

Solution: The divided difference table is

x	f	[,]	[,,]
0	2		
		3-2	
		1-0	
			1

Example: Given the values

x	1	3	14	17	31	35
---	---	---	----	----	----	----

f(x)	68.7	64.0	44.0	39.1		
			12-3			
			2-1			
				9		

$$\frac{9-1}{2-0} = 4$$

$$\frac{9-4}{5-0} = 1$$

$$\frac{45-9}{5-2} = 9$$

2	12
---	----

$$f(x) = \frac{147 - 12}{5 - 2} = 45$$

5	147
---	-----

Newton's divided difference interpolation formula is

$$f(x) = f(x_0) + (x-x_0) f[x_0, x_1] + (x-x_0)(x-x_1) f[x_0, x_1, x_2] + (x-x_0)(x-x_1)(x-x_2) f[x_0, x_1, x_2, x_3]$$

Here $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $f(x_0) = 2$,

$$f[x_0, x_1] = 1, f[x_0, x_1, x_2] = 4, f[x_0, x_1, x_2, x_3] = 1$$

Substituting these values in (1),

$$f(x) = 2 + (x-0) 1 + (x-0)(x-1) 4 + (x-0)(x-1)(x-2) 1$$

$$= x^3 + x^2 - x + 2$$

4.4- Lagrange's Interpolation formula

Let $y = f(x)$ be a function which takes the values $y_0, y_1, \dots, y_n$ corresponding to $x = x_0, x_1, \dots, x_n$. Since there are $n+1$ values of y corresponding to $n+1$ values of x , we can represent the function $f(x)$ by a polynomial in x of degree n .

$$\text{Let } f(x) = A_0 x^n + A_1 x^{n-1} + \dots + A_n$$

We can choose the above polynomial as follows.

$$\begin{aligned} y = f(x) = & a_0 (x-x_1)(x-x_2)\dots(x-x_n) \\ & + a_1 (x-x_0)(x-x_2)\dots(x-x_n) \\ & + a_2 (x-x_0)(x-x_1)(x-x_3)\dots(x-x_n) \\ & + \dots + a_n (x-x_0)(x-x_1)\dots(x-x_{n-1}) \end{aligned} \quad (1)$$

This is true for all values of x involved

$$\text{When } x = x_0, \quad y = y_0$$

Hence substituting in (1) we get,

$$\begin{aligned} y_0 &= a_0 (x_0-x_1)(x_0-x_2)\dots(x_0-x_n) \\ \therefore a_0 &= \frac{y_0}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)} \end{aligned}$$

Similarly putting $x = x_1, y = y_1$, in (1) we have

$$a_1 = \frac{y_1}{(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)}$$

$$a_2 = \frac{y_2}{(x_2-x_0)(x_2-x_1)\dots(x_2-x_n)}$$

$$a_n = \frac{y_n}{(x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1})}$$

Substituting the values of $a_0, a_1, a_2, \dots, a_n$ in (1)

$$\text{we get, } y = f(x) = \frac{(x-x_1)(x-x_2)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)} y_0 + \frac{(x-x_0)(x-x_2)\dots(x-x_n)}{(x_1-x_0)(x_1-x_2)\dots(x_1-x_n)} y_1 + \dots + \frac{(x-x_0)(x-x_1)\dots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\dots(x_n-x_{n-1})} y_n \quad (2)$$

Equation (2) is lagrange's interpolation formula. It can be used whether the values $x_0, x_1, \dots, x_n$ are equally spaced or not and whether the difference of y become smaller or not.

Example: Given the values

$x : 14 \quad 17 \quad 31 \quad 35$

$f(x) : 68.7 \quad 64.0 \quad 44.0 \quad 39.1$

find the value of $f(x)$ corresponding to $x=27$

Solution: Since there are only four corresponding pairs of values given, the polynomial representing the data is

$$y = f(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} y_0 + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} y_1 + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} y_2 + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} y_3$$

Here $x_0 = 14, x_1 = 17, x_2 = 31, x_3 = 35$

$y_0 = 68.7, y_1 = 64.0, y_2 = 44.0$ and $y_3 = 39.1$

$$\therefore y = \frac{(x-17)(x-31)(x-35)}{(14-17)(14-31)(14-35)} \quad (68.7)$$

$$+ \frac{(x-14)(x-31)(x-35)}{(17-14)(17-31)(17-35)} \quad (64.0)$$

$$+ \frac{(x-14)(x-17)(x-35)}{(31-14)(31-17)(31-35)} \quad (44.0)$$

$$+ \frac{(x-14)(x-17)(x-31)}{(35-14)(35-17)(35-31)} \quad (39.1)$$

$$\text{ie., } y = \frac{(x-17)(x-31)(x-35)}{(-3)(-17)(-21)} \quad (68.7)$$

$$+ \frac{(x-14)(x-31)(x-35)}{(3)(-14)(-18)} \quad (64.0)$$

$$+ \frac{(x-14)(x-17)(x-35)}{(17)(14)(-4)} \quad (44.0)$$

$$+ \frac{(x-14)(x-17)(x-31)}{(21)(18)(4)} \quad (39.1)$$

Let y_{27} be the value corresponding to $x=27$.

Putting $x=27$ in (1)

$$y_{27} = \frac{(27-17)(27-31)(27-35)}{(-3)(-17)(-21)} \quad (68.7)$$

$$+ \frac{(27-14)(27-31)(27-35)}{(3)(-14)(-18)} \quad (64.0)$$

$$+ \frac{(27-14)(27-17)(27-35)}{(17)(14)(-4)} (44.0)$$

$$+ \frac{(27-14)(27-31)(27-17)}{(21)(18)(4)} (39.1)$$

$$= \frac{10(-4)(-8)}{(-3)(-17)(-21)} (68.7) + \frac{13(-4)(-8)}{3(-14)(-18)} (64)$$

$$+ \frac{13(10)(-8)}{17(14)(-4)} (44) + \frac{13(10)(-4)}{(21)(18)(4)} (39.1)$$

$$= -20.52 + 35.22 + 48.07 - 13.45$$

$$= 49.3$$

Exercise

- Using Newton's forward interpolation formula, find the value of y when x=21 from the following tabulated values of the function :

x:	20	23	26	29
y:	0.3420	0.3907	0.4384	0.4848

Ans (0.3583)

- Using Newton's formula, find y when x = 27 from the following data:

x :	10	15	20	25	30
y :	35.4	32.2	29.1	26.0	23.1

Ans (24.8)

- Use Lagrange's formula to find the value of y at x= 6 from the following data.

x :	3	7	9	10
y :	168	120	72	63

Ans (147)

Theory of Attributes

Classification of statistical data are of two kinds - quantitative and qualitative. For example, the classification of a population according to different age groups, according to heights and weights of the individuals or the classification of commodities according to prices are quantitative classification. On the other hand, if the population is divided into different religions groups like Hindus, Muslims, Christians etc or into literates and illiterates or into employed and unemployed we have qualitative classification. Qualitative classification is usually termed as 'attributive classification' and the qualities are called 'attributes'.

When we divide the statistical population according to one attribute, we can either divide it into two classes - those possessing the attribute and those not possessing it or divide it into more than two classes. For example, the population of a locality may be either divided into two classes as Hindus and non Hindus or into more than two as Hindus, Christians, Muslims and others. The first type of classification is called 'twofold' or 'dichotomous' and the second type, 'manifold'.

First, we shall consider the dichotomous classification of attributes. Given two attributes, we want to know whether they are dependent or not i.e., whether the possession of one affects that of the other or not. In case they are dependent, we say the attributes are associated and we define a coefficient to measure the extent of association.

Notations

Generally we use capital letters A,B,C etc to denote the presence of attributes, and the Greek letters α , β , r etc to denote the absence of attributes. In the place of Greek letters small letters a,b,c etc., are also used. Thus if A represents the attribute of being deaf then α denotes absence of deafness or not being deaf.

A,B,C.... are called positive attributes and the corresponding α , β , r or a, b, c,..... are called negative attributes.

Combination of attributes is denoted by combination of letters. Thus if A denotes the attribute of being educated, and B denotes the attribute of being employed then the combination of attributes is represented by grouping together the letters (capital and Greek) that indicate the attributes concerned.

Thus AB stands for educated and employed

$A\beta$ stand for educated and unemployed

αB stands for uneducated and employed

$\alpha \beta$ stands for uneducated and unemployed

A third-attribute namely 'being married' can also be added to this. In that case ABC will denote educated, employed and married $A\beta r$ will denote educated, unemployed and unmarried etc.

Class Frequencies

The number of observations in any class is called the class frequency. For example in the above case $(AB) = 250$ means that there are two hundred and fifty individuals in the population who are both educated and employed at the same time. To denote this class frequency, we enclose the class - symbol within brackets. Thus (AB) represents the number of members possessing the attribute AB.

Order of Classes

A class specified by n attributes give us a class of the n th order and its frequency as a frequency of the n^{th} order. Suppose we have two attributes A and B, then we talk of classes of the first order and of the second order.

Thus A, B, α , β are classes of the first order and AB, $A\beta$, αB , $\alpha\beta$ are classes of the second order. In the case of two attributes, there can be no further classes of higher orders. So classes of the second order are called 'classes of the ultimate order'. In the case of the population with three attributes, classes of the third order ie, the highest order classes are called 'classes of the ultimate order'. The total number of observations is denoted by N and is called the class of the zero order.

Suppose we have a dichotomous classification with respect to n attributes. Then the number of class frequencies of order r is $(n C_r 2^r)$ since r attributes can be chosen in $n C_r$ ways and each of the r attributes contributes two classes. Thus the total number of class frequencies of all orders

$$\text{for } n \text{ attributes is } \sum_{r=0}^n n C_r 2^r = (1+2)^n = 3^n.$$

Also, the total number of classes of ultimate order is determined by the formula 2^n .

Now we shall show that any class frequency can be expressed as a sum of class frequencies of higher order.

Let A stand for male. α stand for female. B stand for married and β stand for unmarried C stand for employed and r stand for unemployed and N stand for total population.

The total population is equal to the total number of males plus the total number of females.

$$\text{So } N = (A) + (\alpha) \quad (1)$$

$$\text{Similarly } N = (B) + (\beta) \quad (2)$$

$$\text{and } N = (C) + (r) \quad (3)$$

Since males may be married or unmarried total number of males will be equal to the total number of married males plus the total number of unmarried males.

$$\text{i.e, } (A) = (AB) + (A\beta) \quad (4)$$

Similarly, total number of married = total number of married males + total number of married females is denoted by $(B) = (AB) + (\alpha B)$ (5)

$$\text{Also } (\alpha) = (\alpha B) + (\alpha\beta) \quad (6)$$

$$\text{and } (\beta) = (A\beta) + (\alpha\beta) \quad (7)$$

The above relation shows a first order class frequency in terms of 2nd order.

Substituting (4) and (6) in (1) we get

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta) \quad (8)$$

Now married males (AB) may be employed or unemployed and therefore the number of married males (AB) will be equal to the sum of employed married males and unemployed married males. So $(AB) = (ABC) + (ABr)$

$$\text{Similarly } (A\beta) = (A\beta C) + (A\beta r)$$

$$(\alpha B) = (\alpha BC) + (\alpha Br)$$

$$(\alpha\beta) = (\alpha\beta C) + (\alpha\beta r)$$

It can be found that a second order class frequency can be expressed as the sum of third order class frequencies. Substituting these in (8) we get $N = (ABC) + (ABr) + (A\beta C) + (A\beta r) + (\alpha BC) + (\alpha Br) + (\alpha\beta C) + (\alpha\beta r)$ showing a zero order in terms of third order class frequency.

Similarly (4) can be written as

$$(A) = (ABC) + (ABr) + (A\beta C) + (A\beta r)$$

showing a first order in terms of third order class frequency.

Thus we see that all the class frequencies can be expressed as an aggregate of some of the class frequencies of highest order or ultimate order. So in the case of two attributes or three attributes, given a few class frequencies we can find the remaining class frequencies.

In case of two attributes, the frequencies are grouped into a table with two rows and two columns known as 2 x 2 contingency table.

Attributes	A	α	Total
B	(AB)	(αB)	(B)
β	(A β)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

Thus from the table, $(AB) + (A\beta) = (A)$ and so $(A\beta) = (A) - (AB)$

Example 1: From the following data, calculate the remaining class frequencies.

$$N = 100, (A) = 50, (B) = 70 \text{ and } (AB) = 30.$$

Solution:

The remaining classes are α , β , AB, αB , $\alpha\beta$.

We know that $N = (A) + (\alpha)$

$$\therefore (\alpha) = N - (A) = 100 - 50 = 50.$$

Similarly $N = (B) + (\beta)$

$$\therefore (\beta) = N - (B) = 100 - 70 = 30$$

From the table, we see that $(AB) + (A\beta) = (A)$

$$\therefore (A\beta) = (A) - (AB) = 50 - 30 = 20$$

$$(B) = (AB) + (\alpha B)$$

$$\therefore (\alpha B) = (B) - (AB) = 70 - 30 = 40$$

$$(\alpha) = (\alpha B) + (\alpha\beta)$$

$$\therefore (\alpha\beta) = (\alpha) - (\alpha B) = 50 - 40 = 10.$$

Example 2:

Given the following set of class frequencies find the remaining class frequencies:

$$(\alpha\beta r) = 45, (\alpha BC) = 37, (\alpha Br) = 28, (ABC) = 20,$$

$$(ABr) = 27, (A\beta C) = 39, (A\beta r) = 46, (\alpha\beta C) = 33$$

Solution: In the case of 3 attributes the highest order is the third order. There are 2^3 classes of highest order and all of them are given in the problem. So, we have to find the class frequencies of the 0th order, 1st order and 2nd order.

$$(AB) = (ABC) + (ABr) = 20 + 27 = 47$$

$$(AC) = (ABC) + (A\beta C) = 20 + 39 = 59$$

$$(BC) = (ABC) + (\alpha BC) = 20 + 37 = 57$$

$$(A\beta) = (A\beta C) + (A\beta r) = 39 + 46 = 85$$

$$(Ar) = (ABr) + (\alpha Br) = 27 + 28 = 55$$

$$(\alpha B) = (\alpha BC) + (\alpha Br) = 37 + 28 = 65$$

$$(Br) = (ABr) + (\alpha Br) = 27 + 28 = 55$$

$$(\alpha C) = (\alpha BC) + (\alpha\beta C) = 37 + 33 = 70$$

$$(\beta C) = (A\beta C) + (\alpha\beta C) = 39 + 33 = 72$$

$$(\alpha\beta) = (\alpha\beta C) + (\alpha\beta r) = 33 + 45 = 78$$

$$(\beta r) = (A\beta r) + (\alpha\beta r) = 46 + 45 = 91$$

$$(\alpha r) = (\alpha Br) + (\alpha\beta r) = 28 + 45 = 73.$$

We have found all the second order class frequencies. Now we shall find all first order frequencies.

$$(A) = (AB) + (A\beta) = 47 + 85 = 132$$

$$(B) = (BC) + (Br) = 57 + 55 = 112$$

$$(C) = (BC) + (\beta C) = 57 + 72 = 129$$

$$(\alpha) = (\alpha B) + (\alpha\beta) = 65 + 78 = 143$$

$$(\beta) = (A\beta) + (\alpha\beta) = 85 + 78 = 163$$

$$(r) = (Ar) + (\alpha r) = 55 + 73 = 128.$$

$$N = (A) + (\alpha) = (B) + (\beta) + (C) + (r)$$

$$= 132 + 143 = 112 + 163 = 129 + 146 = 276.$$

Now we shall write any class frequency in terms of the positive frequency. It will help us to find all the class frequencies, when we know the positive frequencies.

Given the attribute A, we regard it as an operator and write $A.N = (A)$

Similarly, $\alpha N = (\alpha)$

So, $N = (A) + (\alpha) = A.N + \alpha.N = (A + \alpha) N$.

So in a symbolic expression, we write

$$A + \alpha = 1$$

$$\alpha = 1 - A \text{ and } A = 1 - \alpha$$

Similarly $\beta = 1 - B, r = 1 - C$

$$AB.N = (AB)$$

$$AN.B = (AB) \text{ etc.}$$

Example 1:

$$A\beta = A(1-B)$$

i.e. to find $(A\beta) = (A(1-B))$

$$\therefore (A\beta) = A(1-B)N = (A-AB)N = (A) - (AB)$$

So, we have to find $(A\beta)$ and (AB)

$$(\alpha\beta) = (1-A)(1-B)N$$

$$(A\beta) = A(1-B)N$$

$$= (1-A-B+AB)N$$

$$= N - (A) - (B) + (AB)$$

$$(ABr) = AB(1-C)N$$

$$= (AB - ABC)N$$

$$= (AB) - (ABC)$$

$$(\alpha\beta r) = (1-A)(1-B)(1-C)N$$

$$= N - (A) - (B) - (C) + (AB) + (BC) + (AC) - (ABC)$$

Example 2

Given the following positive frequencies, find out the frequencies of the ultimate classes.

$N = 50,000$, $(A) = 45000$, $(B) = 5700$, $(C) = 1500$, $(AB) = 1700$, $(AC) = 750$, $(BC) = 700$, $(ABC) = 300$.

Solution: (ABC) , $(A\beta r)$, $(A\beta C)$, (ABr) , (αBC) , (αBr) , $(\alpha\beta C)$, $(\alpha\beta r)$ are the ultimate class frequencies. The first one is given. We have to find the remaining seven.

$$\begin{aligned}(ABr) &= AB(1-C)N \\ &= (AB - ABC)N \\ &= (AB) - (ABC) \\ &= 1700 - 300 = 1400\end{aligned}$$

$$\begin{aligned}(A\beta C) &= A(1-B)C N \\ &= (AC - ABC)N \\ &= (AC) - (ABC) \\ &= 750 - 300 = 450\end{aligned}$$

$$\begin{aligned}(A\beta r) &= A(1-B)(1-C)N \\ &= (A - AB - AC + ABC)N \\ &= (A) - (AB) - (AC) + (ABC) \\ &= 4500 - 1700 - 750 + 300 \\ &= 2350\end{aligned}$$

$$\begin{aligned}(\alpha BC) &= (1-A)BC N \\ &= (BC) - (ABC) \\ &= 700 - 300 = 400\end{aligned}$$

$$\begin{aligned}(\alpha Br) &= (1-A)B(1-C)N \\ &= (B) - (AB) - (BC) + (ABC) \\ &= 5700 - 1700 - 700 + 300 \\ &= 3600\end{aligned}$$

$$(\alpha\beta C) = (1-A)(1-B)CN$$

$$= (C) - (AC) - (BC) + (ABC)$$

$$= 1500 - 750 - 700 + 300 = 350$$

$$(\alpha\beta r) = (1-A)(1-B)(1-C)N$$

$$= N - (A) - (B) - (C) + (AB) + (BC) + (AC) - (ABC)$$

$$= 50000 - 4500 - 5700 - 1500 + 1700 + 700 + 750 - 300$$

$$= 41150$$

Example 3: 100 children took three examinations. 40 passed the first, 39 passed the second and 48 passed the third. 10 passed all three, 21 failed all three, 9 passed the first two and failed the third, 19 failed the first two and passed the third. Find how many children passed at least two examinations show that for the question asked, certain of the given frequencies are not necessary. Which are they?

Solution: Let A, B, C denote the passing of the first, second and third examinations respectively. Then $N = 100$, $(A) = 40$, $(B) = 39$, $(C) = 48$, $(ABC) = 10$, $(\alpha\beta r) = 21$, $(ABr) = 9$, $(\alpha\beta C) = 19$.

To find the number of children who have passed at least two:

i.e, to find $(ABC) + (ABr) + (A\beta C) + (\alpha\beta C)$

So, we have to find $(A\beta C)$ and $(\alpha\beta C)$

$$(A\beta C) = A(1-B)CN$$

$$= (AC) - (ABC)$$

$$(\alpha\beta C) = \alpha(1-B)CN$$

$$= (\alpha C) - (\alpha\beta C)$$

$$\therefore (ABC) + (ABr) + (A\beta C) + (\alpha\beta C)$$

$$= (ABC) + (ABr) + (AC) - (ABC) + (\alpha C) - (\alpha\beta C)$$

$$= (ABr) + (C) - (\alpha\beta C)$$

$$= 9 + 48 - 19 = 38.$$

We have used only (C) , (ACr) , $(\alpha\beta C)$. The other frequencies are not necessary.

4.6 Consistency of Data

The class frequencies are obtained by counting the number of items possessing or not possessing one or more attributes. So, no class frequency can be negative. Hence, if any frequency has a negative value, the data are said to be inconsistent.

Example 1: $N = 50$, $(AB) = 30$, $(A\beta) = 35$, $(\alpha B) = 10$ $(\alpha\beta) = 20$.

Solution: $(A) = (AB) + (A\beta) = 30 + 35 = 65$

$$(\alpha) = N - (A) = 50 - 65 = -15.$$

Hence, the above data are inconsistent.

Example 2: Test for consistency the following data.

$$N = 100, (A) = 80, (B) = 55, (AB) = 20.$$

Solution: $(\alpha\beta) = (1-A)(1-B)N$

$$= N - (A) - (B) + (AB)$$

$$= 100 - 80 - 55 + 20 = -15$$

$\therefore$ The given data are inconsistent.

Example 3: Verify the consistency of the following data $N = 1000$, $(A) = 500$, $(B) = 550$, $(C) = 450$, $(AB) = 200$, $(BC) = 250$, $(AC) = 150$, $(ABC) = 120$.

Solution: $(\alpha\beta\gamma) = (1-A)(1-B)(1-C)N$

$$= N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$$

$$= 1000 - 500 - 550 - 450 + 200 + 250 + 150 - 120$$

$$= 1600 - 1620 = -20.$$

$\therefore$ The data are inconsistent

In order to test the consistency of data, we have to check whether any class frequency is negative or not. To check this, we can also apply the following tests.

Case 1: For a single attribute A, we have conditions of consistency as follows.

$$(i) (A) \geq 0 \\ N - (\alpha) \geq 0 \quad \therefore (\alpha) \leq N$$

$$(ii) (\alpha) \geq 0 \\ N - (A) \geq 0 \quad \therefore (A) \leq N$$

Case 2: For two attributes A, B the conditions of consistency are:

$$(i) (AB) \geq 0$$

$$(ii) (A\beta) \geq 0$$

$$\text{i.e., } A(1-B) \geq 0$$

$$(A) - (AB) \geq 0$$

$$\text{i.e., } (AB) \leq (A)$$

Similarly (iii) $(AB) \leq (B)$

$$(iv) (\alpha\beta) \geq 0$$

$$(1-A)(1-B)N \geq 0$$

$$N - (A) - (B) + (AB) \geq 0$$

$$\text{i.e., } (AB) \geq (A) + (B) - N$$

Case 3: For three attributes A, B, C the conditions of consistency are:

$$(1) (ABC) \geq 0$$

$$(2) (ABr) \geq 0$$

$$(AB) - (ABC) \geq 0$$

$$(ABC) \leq (AB)$$

Similarly, (3) $(ABC) \leq (BC)$

$$(4) (ABC) \leq (CA)$$

$$(5) (A\beta r) \geq 0$$

$$A(1-B)(1-C)N \geq 0$$

$$(A) - (AB) - (AC) + (ABC) \geq 0$$

$$\therefore (ABC) \geq (AB) + (AC) - (A)$$

Similarly, (6) $(ABC) \geq (BC) + (BA) - (B)$

$$(7) (ABC) \geq (CA) + (CB) - (C)$$

$$(8) (\alpha\beta r) \geq 0$$

$$\begin{aligned} \text{i.e., } (1-A)(1-B)(1-C)N &\geq 0 \\ N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) &\geq 0 \\ \therefore (ABC) &\leq (AB) + (BC) + (AC) - (A) - (B) - (C) + N \end{aligned}$$

$$\begin{aligned} (9) \text{ From (1) and (8) we get} \\ (AB) + (BC) + (CA) &\geq (A) + (B) + (C) - N \end{aligned}$$

$$\begin{aligned} (10) \text{ From (2) and (7) we get} \\ (AC) + (BC) - (C) &\leq (ABC) \leq (AB) \\ \therefore (AC) + (BC) - (AB) &\leq (C) \end{aligned}$$

$$\begin{aligned} (11) \text{ From (4) and (6) we get} \\ (AB) + (BC) - (AC) &\leq (B) \end{aligned}$$

$$\begin{aligned} (12) \text{ From (3) and (5) we get} \\ (AB) + (AC) - (BC) &\leq (A) \end{aligned}$$

We also note the following inequalities.

$$(i) (AB) \geq (A) + (B) - N.$$

Given 3 attributes A, B, C let $T = AB$.

$$\begin{aligned} (ABC) = (TC) &\geq (T) + (C) - N \\ &\geq (AB) + (C) - N \\ &\geq (A) + (B) - N + (C) - N \\ &\geq (A) + (B) + (C) - 2N \end{aligned}$$

$$\therefore (ii) (ABC) \geq (A) + (B) + (C) - 2N$$

Given 4 attributes A, B, C, D let $T = ABC$

$$\begin{aligned} (ABCD) = (TD) &\geq (T) + (D) - N \\ &\geq (ABC) + (D) - N \\ &\geq (A) + (B) + (C) + (D) - 3N \end{aligned}$$

So, in general, if there are n attributes $A_1, A_2, \dots, A_n$ then
 $(A_1, A_2, \dots, A_n) \geq (A_1) + (A_2) + \dots + (A_n) - (n-1)N.$

Example 1: A survey of 1000 companies gives the following data:

- (i) Companies with capital of more than Rs.10 lakhs 510.
- (ii) Companies making profit 490.
- (iii) Companies under managing agents 427
- (iv) Companies with capital more than 10 lakhs making profit 189

- (v) Companies with capital more than 10 lakhs of under managing agents 140
- (vi) Companies making profit and under managing agents 85 show that the information as it stands must be incorrect.

Solution: Let A denote companies with capital more than 10 lakhs, B companies making profit and C companies under managing agents.

$$N = 1000, (A) = 510, (B) = 490, (C) = 427$$

$$(AB) = 189, (AC) = 140, (BC) = 85.$$

One of the inequalities to be satisfied for consistency in the case of three sets.

$$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$$

$$\text{L.H.S} = 189 + 140 + 85 = 414$$

$$\text{R.H.S} = 510 + 490 + 427 - 1000$$

$$= 1427 - 1000 = 427.$$

$$\text{L.H.S.} \leq \text{R.H.S}$$

Hence the information is incorrect.

Example 2: If in a collection of houses actually invaded by small pox 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest percentage of the vaccinated that must have been attacked?

Solution: Let A denote the inhabitants that are attacked and B vaccinated

$$\text{Then } N = 100, (A) = 70, (B) = 85$$

We have to find the least value of (AB)

$$(AB) \geq (A) + (B) - N$$

$$\geq 70 + 85 - 100 = 55.$$

$\therefore$ 55 is the required lowest percentage.

Example 3: In a university examination 70% of the candidates passed English, 95% passed in second language and 65% passed in the optionals. Find how many atleast should have passed the whole examination.

Solution: Let A stand for passing in English, B for passing in the second language and C for passing in the optionals. So, we have

$$N = 100, (A) = 70, (B) = 95, (C) = 65$$

$$(ABC) \geq (A) + (B) + (C) - 2N$$

$$\geq 70 + 95 + 65 - 200$$

$$\geq 30$$

$\therefore$ At least 30% should have passed the whole examination.

Example 4: In a hotly fought battle, 70% at least lost an eye, 75% at least lost an ear, 80% at least lost a leg and 85% at least lost an arm. Find what % atleast lost all four.

Solution: Let A, B, C, D denote respectively losing an eye, ear, leg or arm. Then $N = 100, (A) \geq 70, (B) \geq 75, (C) \geq 80, (D) \geq 85$.

$$(ABCD) \geq (A) + (B) + (C) + (D) - 3N$$

$$\geq 70 + 75 + 80 + 85 - 300$$

$$\geq 10$$

$\therefore$ At least 10% lost all four.

Example 5: If $(A) = 50, (B) = 60, (C) = 50, (A\beta) = 5, (Ar) = 20, N = 100$, find the greatest and least possible values of (BC)

Solution: $N = 100, (A) = 50, (B) = 60, (C) = 50, (A\beta) = 5, (Ar) = 20$

$$(AB) = (A) - (A\beta) = 50 - 5 = 45$$

$$(AC) = (A) - (Ar) = 50 - 20 = 30.$$

$$(AB) + (BC) + (AC) \geq (A) + (B) + (C) - N$$

$$45 + (BC) + 30 \leq 50 + 60 + 50 - 100$$

$$(BC) \leq 60 - 75 = -15.$$

We know $(BC) \geq 0$ also the above inequality does not give any special information of what we require.

$$(AB) + (AC) - (BC) \leq (A)$$

$$45 + 30 - (BC) \leq 50$$

$$75 - (BC) \leq 50$$

$$-(BC) \leq 50 - 75 = -25$$

$$(BC) \geq 25$$

(1)

$$(AB) + (BC) - (AC) \leq (B)$$

$$45 + (BC) - 30 \leq 60$$

$$(BC) \leq 60 + 30 - 45 = 45$$

$$(BC) \leq 45$$

(2)

$$(AC) + (BC) - (AB) \leq (C)$$

$$30 + (BC) - 45 \leq 50$$

$$(BC) \leq 50 + 45 - 30 = 65$$

$$(BC) \leq 65$$

(3)

From (1), (2) and (3) we get

$$25 \leq (BC) \leq 45.$$

Example 6: Given $N = 200$, $(A) = 100$, $(B) = 120$, $(C) = 160$, $(AB) = 70$, $(AC) = 90$, $(BC) = 84$, find the least and greatest values of (ABC)

Solution: $(ABC) \leq (AB)$

$$\leq 70$$

(1)

$$(ABC) \leq (BC)$$

$$\leq 84$$

(2)

$$(ABC) \leq (AC)$$

$$\leq 90$$

(3)

$$(ABC) \leq (AB) + (BC) + (AC) - (A) - (B) - (C) + N$$

$$\leq 70 + 84 + 90 - 100 - 120 - 160 + 200$$

$$\leq 64$$

(4)

(1), (2), (3) and (4) will be satisfied if

$$(ABC) \leq 64$$

$$(ABC) \geq (A) + (B) + (C) - 2N$$

$$\geq 100 + 120 + 160 - 400$$

$$\geq -20.$$

Since $(ABC) \geq -20$, the above inequality is satisfied.

$$(ABC) \geq (AB) + (AC) - (A)$$

$$\geq 70 + 90 - 100$$

$$\geq 60$$

(5)

$$\begin{aligned}(ABC) &\geq (BC) + (AB) - (B) \\ &\geq 84 + 70 - 120 \\ &\geq 34\end{aligned}$$

$$\begin{aligned}(ABC) &\geq (AC) + (BC) - (C) \\ &\geq 90 + 84 - 160 \\ &\geq 14\end{aligned}$$

(5), (6), (7) will be satisfied if $(ABC) \geq 60$.

The least value of $(ABC) = 60$

The greatest value of $(ABC) = 64$.

Exercise:

1. Given $N = 200$, $(A) = 76$, $(B) = 170$, $(C) = 96$, $(AB) = 60$, $(BC) = 70$, $(AC) = 39$, $(ABC) = 25$, find the ultimate class frequencies.
2. In a report, the following data are given 50% items have characters A and B, 35% have A but not B, 25% have B but not A. Show that there must be misprint in the data.
3. If $N = 120$, $(A) = 60$, $(B) = 90$, $(C) = 30$, $(BC) = 15$, $(AC) = 15$, find the limits between which (AB) must be.
4. In an examination which was very tough at least 65% failed in Mathematics, 75% at least failed in Physics and 85% at least failed in Chemistry. How many at least failed in all the three?

4.7 Independence of attributes

Two attributes which have absolutely no relationship between them are said to be independent. If two attributes A and B are independent, we expect that the proportion of A amongst B's is same as proportion of A amongst not B's i.e, amongst β 's.

$$\text{or } \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad (1)$$

This means that the presence or absence of B is nothing to do with the occurrence of A.

$$\text{From (1) } 1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$$

$$\text{i.e, } \frac{(B) - (AB)}{(B)} = \frac{(\beta) - (A\beta)}{(\beta)}$$

$$\text{i.e, } \frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$$

So, the proportion of the α 's among the B's is the same as the proportion of the α 's among the β 's.

$$\text{Also } \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{N} \quad (2)$$

i.e, the proportion of A's amongst the B's is the same as the proportion of A's in the population

$$\text{From (2) we get } \frac{(AB)}{(A)} = \frac{(B)}{N} = \frac{(B) - (AB)}{N - (A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\text{So we have } \frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

$$\text{Also we have } \frac{(AB)}{N} = \frac{(A)(B)}{N} \quad (3)$$

$$\therefore \frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N} \quad (4)$$

Since $\frac{(A)}{N}$ gives the probability of the occurrence of the attribute A in the population and $\frac{(B)}{N}$ is the probability of the occurrence of the attribute B, (4) shows that the probability of the occurrence of A and B together is equal to the product of the probability of the occurrence of A and the probability of the occurrence of B if A and B are independent.

Association:

$(AB) = \frac{(A)(B)}{N}$ gives the values of (AB) if A and B are independent.

If $(AB) \neq \frac{(A)(B)}{N}$ we say that A and B are associated.

If $(AB) > \frac{(A)(B)}{N}$ we say that A and B are positively associated.

If $(AB) < \frac{(A)(B)}{N}$ we say that A and B are negatively associated.

Example 1: From the data given below, find out if the attributes A and B are independent, positively associated or negatively associated.

(i) $(A) = 60, (B) = 43, (AB) = 45, N = 80$

(ii) $(A) = 480, (AB) = 294, (\alpha) = 580, (\alpha B) = 380$

(iii) $(AB) = 256, (\alpha B) = 768, (A\beta) = 48, (\alpha\beta) = 144.$

Solution: (i) $\frac{(A)(B)}{N} = \frac{60 \times 43}{80} = \frac{129}{4} = 32.25$

But $(AB) = 45$

$\therefore (AB) > \frac{(A)(B)}{N}$ and hence A and B are positively associated.

(ii) $(B) = (AB) = (\alpha B)$
 $= 294 + 380 = 674$

$N = (A) + (\alpha)$
 $= 480 + 580 = 1060$

$$\frac{(A)(B)}{N} = \frac{480 \times 674}{1060} = 305.2 > (AB)$$

$$\text{i.e., } (AB) < \frac{(A)(B)}{N}$$

Hence A and B are negatively associated.

$$(iii) (A) = (AB) + (A\beta) = 256 + 48 = 304$$

$$(B) = (AB) + (\alpha B) = 256 + 768 = 1024$$

$$N = (A) + (\alpha)$$

$$= (A) + (\alpha\beta) + (\alpha B)$$

$$= 304 + 768 + 144 = 1216$$

$$\frac{(A)(B)}{N} = \frac{304 \times 1024}{1216} = 256 = (AB)$$

$\therefore$ A and B are independent

Thus the following table gives the criteria for finding whether two attributes are independent or positively or negatively associated or all possible combinations

Attribute	Indepence or No association	Positive Association	Negative Association
A and B	$(AB) = \frac{(A)(B)}{N}$	$(AB) > \frac{(A)(B)}{N}$	$(AB) < \frac{(A)(B)}{N}$
A and β	$(A\beta) = \frac{(A)(\beta)}{N}$	$(A\beta) > \frac{(A)(\beta)}{N}$	$(A\beta) < \frac{(A)(\beta)}{N}$
α and B	$(\alpha B) = \frac{(\alpha)(B)}{N}$	$(\alpha B) > \frac{(\alpha)(B)}{N}$	$(\alpha B) < \frac{(\alpha)(B)}{N}$
α and β	$(\alpha\beta) = \frac{(\alpha)(\beta)}{N}$	$(\alpha\beta) > \frac{(\alpha)(\beta)}{N}$	$(\alpha\beta) < \frac{(\alpha)(\beta)}{N}$

Example 2: Of 1600 persons in a locality exposed to small - pox 300 in all were attacked of 1600 persons 400 had been vaccinated and of these, only 30 were attacked. Can vaccination be regarded as a preventive measure for small-pox.

Solution: Let A denote the persons attacked and B denote the persons vaccinated.

In symbols, the given data are $N = 1600$, $(A) = 300$ $(B) = 400$ and $(AB) = 30$.

$$\frac{(A)(B)}{N} = \frac{300 \times 400}{1600} = 75$$

$$(AB) = 30$$

Since $(AB) < \frac{(A)(B)}{N}$, there is negative association between A and B.

i.e, there is negative association between the attacked and the vaccinated

⇒ there is positive association between the not attacked and the vaccinated.

⇒ vaccination is effective in preventing small pox.

Example 3: Given that $(A) = 60$, $(AB) = 55$, $(\alpha) = 40$ and $(\alpha\beta) = 10$ examine if A and B are associated.

Solution: We can compare $\frac{(AB)}{(A)}$ with $\frac{(\alpha\beta)}{(\alpha)}$ and see whether A and B are associated or not.

(AB)	55	11	
$\frac{(AB)}{(A)} = \frac{55}{60} = \frac{11}{12}$			
(A)	60	12	
$(\alpha\beta)$	10	1	3
$\frac{(\alpha\beta)}{(\alpha)} = \frac{10}{40} = \frac{1}{4}$			
(α)	40	4	12
$\therefore \frac{(AB)}{(A)} > \frac{(\alpha\beta)}{(\alpha)}$			

i.e, A and B are positively associated.

Exercise

1. Out of 70,000 literates in a particular district of India the number of criminals was 500. Out of 930, 050 illiterates in the same district the number of criminals was 15,000. On the basis of these figures, do you find any association between illiteracy and criminality?
2. In a sample study of 100 items two characters A and B were noted. Following were the various class frequencies?
 $N = 100$, $(A) = 80$, $(B) = 70$ and $(AB) = 62$.
Is there any association between A and B?

4.8 Coefficient of Association

The above methods only reveal us whether two attributes are associated or not. They do not reveal the degree or the extent of association between two attributes. To measure the intensity of association between two attributes A and B, Andrew Yule has given the formula of co-efficient of association denoted by the symbol Q, as follows:

$$Q = \frac{(AB) (\alpha\beta) - (A\beta) (\alpha B)}{(AB) (\alpha\beta) + (A\beta) (\alpha B)} \quad (1)$$

When there is a complete association between A and B every B will be A, and $(AB) = (A)$.
 $\therefore (A\beta) = 0$

$$\text{So (1) becomes } Q = \frac{(A) (\alpha\beta) - 0}{(A) (\alpha B) + 0} = 1$$

When there is complete disassociation or negative association no B will be A.
i.e, $(AB) = 0$ and $(A\beta) = (A)$

$$\therefore (1) \text{ becomes } Q = \frac{0 - (A) (\alpha B)}{0 + (A) (\alpha B)} = -1$$

When A and B are independent,

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \text{ and } \frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\alpha)}$$

$$\text{From these we get, } \frac{(AB)}{(\alpha B)} = \frac{(A\beta)}{(\alpha\beta)}$$

$$\therefore (AB) (\alpha\beta) - (A\beta) (\alpha B) = 0$$

From (1), we get $Q = 0$ when A and B are independent. Thus we find that Q ranges from -1 to 1.

Another measure which serves the same purpose as Yule's coefficient of Association is given below. this is called the coefficient of Colligation.

It is defined by
$$y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$y = \frac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}} \quad (2)$$

It can be shown that $y = +1$ when there is complete association, $y = -1$ when there is complete disassociation and $y = 0$ when there is independence between the attributes A and B. Thus y ranges from -1 to +1.

$$\begin{aligned} \text{From (2), } \frac{y+1}{y-1} &= \frac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)} + \sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)} - \sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}} \\ &= \frac{2\sqrt{(AB)(\alpha\beta)}}{-2\sqrt{(A\beta)(\alpha B)}} = \frac{-\sqrt{(AB)(\alpha\beta)}}{\sqrt{(A\beta)(\alpha B)}} \end{aligned}$$

Squairing both sides,

$$\frac{(y+1)^2}{(y-1)^2} = \frac{(AB)(\alpha\beta)}{(A\beta)(\alpha B)}$$

Again

$$\frac{(y+1)^2 - (y-1)^2}{(y+1)^2 + (y-1)^2} = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$\text{i.e., } \frac{4y}{2(1+y^2)} = Q$$

$$\text{i.e., } Q = \frac{2y}{(1+y^2)}$$

Example 4: The accompanying table summarises public opinion on the issue of third terms of the Presidents of the United States. Obtain Yule's Co-efficient of association and colligation for the table.

	Rich	Poor	Total
Attitude favourable	508	1559	2067
Attitude unfavourable	905	1114	2019
Total	1413	2673	4086

Solution:

Denote rich by A and Attitude favourable by B. The table is now:

	A	α	Total
B	(AB)=508	(α B) = 1559	2067
β	(A β)=905	($\alpha\beta$) = 1114	2069
Total	1413	2673	N=4086

Yule's coefficient of Association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{508 \times 1114 - 905 \times 1559}{508 \times 1114 + 905 \times 1559}$$

$$= \frac{-844983}{1976807} = -0.4275$$

Coefficient of Colligation is

$$y = 1 - \frac{\sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)}}$$

$$1 + \frac{\sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)}}$$

$$y = 1 - \frac{\sqrt{905 \times 1559}}{\sqrt{508 \times 1114}}$$

$$1 + \frac{\sqrt{905 \times 1559}}{\sqrt{508 \times 1114}}$$

$$= \frac{1-1.579}{1+1.579} = \frac{-0.579}{2.579} = -0.2249$$

Partial Association

In a study of two attributes A (literacy) and B (employment), we might get positive association thus concluding that a larger proportion of A's are B's. It may be that most of those who are literate are males and males are employed in a greater numbers we have obtained a positive association. i.e, the positive association between A and B may be due to the association of both with C - males.

So, to get a correct idea of the dependence between the two attributes A and B, we find another association with the categories of the third attribute C and r. Such associations in the sub-population are called Partial Association. We use the notations $Q_{AB|C}$, $Q_{AB|r}$ to denote the coefficient of Partial Association in the sub populations C and r respectively.

Coefficient of Total Association between A and B in the given population.

$$Q_{AB} = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Coefficient of Partial Association between A and B in the sub population C. = $(Q_{AB.C})$

$$Q_{AB.C} = \frac{(ABC)(\alpha\beta C) - (A\beta C)(\alpha BC)}{(ABC)(\alpha\beta C) + (A\beta C)(\alpha BC)}$$

In the sub population r:

$$Q_{ABr} = \frac{(ABr)(\alpha\beta r) - (A\beta r)(\alpha Br)}{(ABr)(\alpha\beta r) + (A\beta r)(\alpha Br)}$$

Example 1: Find the coefficient of total and partial associations between education and employment in a group of 100 persons divided according to sex.

$$(A) = 60, (B) = 80, (AB) = 48, (C) = 48, (ABC) = 40, (AC) = 42, (BC) = 44$$

Solution:

$$Q_{AB} = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$(A\beta) = (A) - (AB) = 60 - 48 = 12$$

$$(\alpha B) = (B) - (AB) = 80 - 48 = 32$$

$$\begin{aligned} (\alpha\beta) &= (\alpha) - (\alpha\beta) = N - (A) - (\alpha B) \\ &= 100 - 60 - 32 = 8 \end{aligned}$$

$$\therefore Q_{AB} = \frac{48(8) - 12(32)}{48(8) + 12(32)} = \frac{384 - 384}{384 + 384} = 0$$

Attributes A and B are independent.

$$Q_{ABC} = \frac{(ABC)(\alpha\beta C) - (A\beta C)(\alpha BC)}{(ABC)(\alpha\beta C) + (A\beta C)(\alpha BC)}$$

$$(ABC) = (AC) - (A\beta C)$$

$$= 42 - 40 = 2$$

$$(\alpha BC) = (BC) - (ABC)$$

$$= 44 - 40 = 4$$

$$(\alpha\beta C) = (1 - A) (1 - B) C N$$

$$= (C) - (AC) - (BC) + (ABC)$$

$$= 48 - 42 - 44 + 40 = 2$$

$$Q_{ABC} = \frac{40 \times 2 - 2 \times 4}{40 \times 2 + 2 \times 4} = \frac{80 - 8}{80 + 8} = \frac{72}{88} = \frac{9}{11} = 0.82$$

$$Q_{ABr} = \frac{(ABr) (\alpha\beta r) - (A\beta r) (\alpha Br)}{(ABr) (\alpha\beta r) + (A\beta r) (\alpha Br)}$$

$$\begin{aligned} (ABr) &= (AB) - (ABC) \\ &= 48 - 40 = 8 \end{aligned}$$

$$\begin{aligned} (\alpha\beta r) &= (1-A) (1-B) (1-C) N \\ &= N - (A) - (B) - (C) + (AB) + (BC) + (CA) - (ABC) \\ &= 100 - 60 - 80 - 48 + 48 + 44 + 42 - 40 \\ &= 6. \end{aligned}$$

$$\begin{aligned} (A\beta r) &= A(1-B) (1-C) N \\ &= (A) - (AB) - (AC) + (ABC) \\ &= 60 - 48 - 42 + 40 \\ &= 10. \end{aligned}$$

$$\begin{aligned} (\alpha Br) &= (1 - A) B (1 - C) N \\ &= (B) - (AB) - (BC) + (ABC) \\ &= 80 - 48 - 44 + 40 \\ &= 28 \end{aligned}$$

$$Q_{ABr} = \frac{8 \times 6 - 10 \times 28}{8 \times 6 + 10 \times 28} = \frac{48 - 280}{48 + 280} = \frac{-232}{328} = \frac{-29}{41} = -0.7$$

Thus, there is a positive association between literacy and employment among males and negative association among females. So, the apparent independence in the total population may be due to their compensating effect in the subuniverse.

Exercise

- Find the coefficient of association between training and success in job in a set of 200 workers given $(A) = 100$, $(AB) = 58$.
- Given $(ABC) = 30$, $(ABr) = 2$, $(\alpha BC) = 15$, $(A\beta C) = 4$, $(A\beta r) = 5$, $(\alpha Br) = 8$, $(\alpha\beta C) = 2$, $(\alpha\beta r) = 20$. Find the total and partial association between A and B
- Trace the strength of sanity and deafness from the following data:

	Deaf	Not deaf	Total
Sane	40	100	140
Insane	80	50	130
Total	120	150	270

UNIT - 5

5.1 Index Numbers

In economics and business, we have a few variables like wages, prices of goods, industrial production, agricultural production which change with respect to place and time. If all these variables change exactly in the same direction, i.e., if all of them increase or decrease simultaneously, there will be no difficulty in predicting the future course of action and also, we can easily know how much they have changed during the last interval. But unfortunately these variables do not behave like this. They move in a diverse manner which creates problems of estimation. And for knowing the economic conditions of a particular place, or over a period, it would be better if we have a single value which indicates such change. Statisticians have defined such a single value and called it 'the Index Number'.

Historically the first index was constructed in 1764 to compare the Italian price index in 1750 with the price level in 1500. Though originally developed for measuring the effect of change in prices, index numbers have become today one of the most widely used statistical devices and there is hardly any field where they are not used.

An index number may be defined as a specialized average designed to measure the change in a group of related variables over a period of time.

Infact, index numbers are barameters of economic activity i.e., if one wants to get an idea as to what is happening to an economy, he should look to important indices such as the index number of industrial production, agricultural production and business activity.

Index numbers in effect, relate a variable or variables in a given period to the same variable or variables in another period, called the base period. An index, the simplified name for index numbers, which is computed from a single variable is called a univariate index whereas an index which is constructed from a group of variables is considered a composite index.

Uses of index numbers

1. They help in framing suitable policies.
2. They reveal trends and tendencies.
3. They are important in forecasting future economic activity
4. They are used to adjust the original data for price changes or to adjust wages for cost of living changes and thus transform nominal wages into real wages.

Classification of Index Number

Index numbers are classified in terms of what they measure. In economics and business the classifications are: (1) price (2) quantity (3) value and (4) special purpose.

Index numbers may be constructed for a single commodity or for a group of commodities.

We will be discussing in detail only price index numbers.

Before constructing index numbers a careful thought must be given to the following problems.

1. Purpose of the Index number

The purpose for which an index number is constructed must be clearly stated, since the selection of items depends on the purpose. There is no such index number which can serve as an all-purpose index. Every index is of a limited and particular use. For example, if we want to construct the cost of living index we have to consider the retail prices of commodities of daily use.

2. Selection of a base period

Since index numbers measure change in the value of one or more variables over a period of time, they are constructed with references to some base period. The base period is a period with which comparisons of relative changes are made. It may be a year, a month or a day. The index for the base period is always taken as 100. Though the selection of the base period depends mainly upon the object of the index it would be better to consider the following points also, when selecting the base period.

(i) The base period should be normal one

The period that is selected as base should be normal. i.e., it should be free from abnormalities like famine, boom, depression, draught flood etc. But at times it may be difficult to select a year which is normal in all respects - a year which is normal in one respect may be abnormal in another. To solve such a problem, an average of a number of years 3 or 4 may be taken as the base. This process will reduce the effect of extremes.

(ii) The base period should not be too distant in the past

Since index numbers are helpful in decision making and economic policies are often a matter of short period, we should not select a base period that is too distant. In dearness allowance at present, there is no advantage in taking 1970 or 1975 as base. The comparison should be with the preceding year or the month after which dearness allowances has not been revived.

(iii) Fixed base or chain base:

While selecting the base, a decision has to be made as to whether the base shall remain fixed or not i.e., whether we want a fixed base index or chain base index.

In the fixed base method, the year or period to which all other prices are related is constant.

In the chain base method, the prices of a year (or period) are compared with those of the preceding year (or period) and not with the fixed year (or period)

The chain base method gives a better picture. Also, it is possible to include new items and delete old items. The only drawback is that comparisons cannot be made over a long period.

(3) Choice of an Average

Since index numbers are specialized averages a decision has to be made as to which particular averages - arithmetic mean, geometric mean, harmonic mean, median and mode - should be used for constructing the index. Theoretically, geometric mean is the best average in the construction of index numbers because of the following reasons: (i) geometric mean gives equal weights to equal ratio of change. (ii) geometric mean is less susceptible to major variations as a result of violent fluctuations in the values of the individual items (iii) index numbers calculated by using this average are reversible i.e., these indices satisfy 'Time Reversal Test' and hence base shifting is easily possible.

But arithmetic mean is more popularly used because it is easy to compute. However, wherever possible, geometric means should be preferred.

(4) Selection of appropriate weights

The term 'weight' refers to the relative importance of the different items in the construction of the index. All items are not of equal importance and hence it is necessary to devise some suitable method whereby the varying importance of the different items is taken into account. This is done by allocating weights. Thus we have broadly two types of indices - unweighted and weighted indices.

There are two methods of assigning weights - (i) implicit and (ii) explicit.

In the implicit weighting, a commodity is included in the construction of the index number a number of times. If rice is to be given in an index three times as much weight as wheat, then three varieties of rice or the same variety three times may be included as against one variety of wheat.

In explicit weighting, some outward evidence of importance of the various items in the index is given. We use two types of weights while constructing a price index number (i) quantity weights of the amount of quantity produced, distributed or consumed in a particular period and (ii) value weight $p \times q$ stands for the price and q for quantity.

(5) Selection of an appropriate formula

A number of formulae has been derived for constructing the index and the problem is to select the most appropriate formula. The choice depends not only on the purpose of the index but also on the data available. No particular formula can be regarded as the best under all circumstances. On the basis of ones knowledge of the characteristics of different formulae, available data and the purpose of the index we should choose the formula.

5.2 Simple Index numbers

Unweighted simple Index Number of prices

- (i) The simple Aggregative method.
- (ii) The Average of price relatives method.

Simple Aggregative method

This is the simplest method of constructing index numbers. In this method, the aggregate of current year prices for the various commodities is divided by the aggregate of base year. Prices and the quotient is expressed as a percentage i.e., it is multiplied by 100.

Suppose p_0 represents the base year price and p_1 , the current year price, the price index p_{01} is given by,

$$p_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Example 1

Construct an index for 1985 taking 1983 as base for the following data.

Commodities	A	B	C	D	E	F
Price in 1983 (Rs.) p_0	40	25	50	8.62	24.60	15
Price in 1985 (Rs.) p_1	60	31.25	37.50	8.62	18.45	11.25

$$\begin{aligned}\sum p_0 &= 40 + 25 + 50 + 8.62 + 24.60 + 15 \\ &= 163.22\end{aligned}$$

$$\begin{aligned}\sum p_1 &= 60 + 31.25 + 37.50 + 8.62 + 18.45 + 11.25 \\ &= 167.07\end{aligned}$$

$$p_{01} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{167.07}{163.22} \times 100 = 102.3$$

This means that as compared to 1983, there is a net increase in the prices of commodities included in the index is the extent of 2.3% in 1985.

Limitations of this index

This above method of constructing this index is based on the assumption that for various commodities, their price rates are expressed in the same unit. If the commodities are expressed in different units, then this index is really meaningless.

For example, if the price of rice is given for a quintal, instead of a kg, then there will be a sharp rise in the index.

Also, the above method does not give any consideration to the relative importance of the commodities.

For example, if the price of 1kg of rice and 1kg of ghee are given for the two years, the index number will not indicate the real change in the cost of living. Instead, the price of 1 quintal of rice and 1kg of ghee may be given. It is even better to omit the item ghee if we are interested in the calculation of cost of living index for middle class families.

Hence this index is useful only if all the articles included are known to be of equal importance.

1. Average of Price Relatives Method

A price relative is the ratio of the price of a variable in a period to its value in the base period.

We have two methods of calculating the relatives - Fixed base method and chain base method.

Fixed base method

In this method, either (i) the average price of some arbitrarily chosen period is taken as the base or (ii) the average of the prices of a number of period is taken as the base. The base, so chosen is used for quite some time.

The average price of the base chosen is taken as 100 and the price in each of the other year is expressed as a percentage of this amount. These numbers are called Fixed base relatives.

$$\text{i.e., Fixed base relatives} = \frac{\text{Price of the commodity for the current year}}{\text{Price of the commodity for the base year}} \times 100$$

$$= \frac{p_1}{p_0} \times 100$$

Chain Base method

In this method, the base year is not fixed but it is taken as the previous one.

$$\text{i.e., chain base relative} = \frac{\text{Price of the commodity for the current year}}{\text{Price of the commodity for the previous year}} \times 100$$

These, numbers are also called 'link relatives'.

If we know one of these sets of relatives fixed base relatives and chain base relatives, we can easily find the other set.

Suppose $p_0, p_1, p_2, \dots, p_k$ are the prices of a commodity in $(k+1)$ years, then $\frac{p_1}{p_0}$,

$\frac{p_2}{p_0}, \dots, \frac{p_k}{p_0}$ are the fixed base relatives (by taking 0th year as fixed base), and

$\frac{p_1}{p_0}, \frac{p_2}{p_1}, \dots, \frac{p_k}{p_{k-1}}$ are the chain base relatives

Clearly,

$$\frac{p_k}{p_0} = \frac{p_k}{p_{k-1}} \cdot \frac{p_{k-1}}{p_0}$$

$$= \frac{p_k}{p_{k-1}} \cdot \frac{p_{k-1}}{p_{k-2}} \cdot \dots \cdot \frac{p_1}{p_0}$$

$$\text{Also } \frac{p_k}{p_{k-1}} = \frac{p_k}{p_0} \bigg/ \frac{p_{k-1}}{p_0}$$

i.e, the link relative of one year is of quotient of the fixed base relatives of that year and the preceding year.

Years :	1978	79	80	81	82	83	84	85
Price in (Rs) :	75	50	65	60	72	79	69	75

(a) Fixed base relatives with 1978 as base

Year	Price	Price relatives : $\frac{p_1}{p_0} \times 100$	
1978	75	$\frac{75}{75} \times 100$	100
1979	50	$\frac{50}{75} \times 100$	67
1980	65	$\frac{65}{75} \times 100$	87
1981	60	$\frac{60}{75} \times 100$	80
1982	72	$\frac{72}{75} \times 100$	96
1983	70	$\frac{70}{75} \times 100$	93
1984	69	$\frac{69}{75} \times 100$	92
1985	75	$\frac{75}{75} \times 100$	100

Chain base relatives

Year	Price	Price relatives : $\frac{p_k}{p_{k-1}} \times 100$	
1978	75	-	100
1979	50	$\frac{50}{75} \times 100$	67
1980	65	$\frac{65}{50} \times 100$	130
1981	60	$\frac{60}{65} \times 100$	92
1982	72	$\frac{72}{60} \times 100$	120
1983	70	$\frac{70}{72} \times 100$	97
1984	69	$\frac{69}{70} \times 100$	99
1985	75	$\frac{75}{69} \times 100$	109

Example 3 : Find the chain base relatives from the fixed base relative with the base year 1982.

Year	1982	83	84	85	86	87
Fixed base Relatives	94	98	102	95	98	100

Solution : Chain base relative of a year = $\frac{p_k}{p_{k-1}} = \frac{p_k}{p_0} \div \frac{p_{k-1}}{p_0}$

= $\frac{\text{Fixed base relative of the year}}{\text{fixed base relative of the previous year}}$

Year	Fixed base relatives	Chain base relatives	
1981	100	-	100
1982	94	$\frac{94}{100} \times 100$	94
1983	98	$\frac{98}{94} \times 100$	104.3
1984	102	$\frac{102}{98} \times 100$	104.1
1985	95	$\frac{95}{102} \times 100$	93.1
1986	98	$\frac{98}{95} \times 100$	103.2
1987	100	$\frac{100}{98} \times 100$	102.0

Example 4 : From the chain base index number given below calculate the fixed base index numbers.

Year	1978	79	80	81	82	83
Chain base index	105	75	71	105	92	90

Solution : Fixed base relative of a year = $\frac{p_k}{p_0}$

$$= \frac{p_k}{p_{k-1}} \cdot \frac{p_{k-1}}{p_{k-2}} \cdot \dots \cdot \frac{p_1}{p_0}$$

= product of the chain base relative of the year and those of the previous years.

Year	Chain base relative	Fixed base relative	
1978	105	-	105
1979	75	$\frac{105}{100} \times \frac{75}{100} \times 100$	76
1980	71	$\frac{105}{100} \times \frac{75}{100} \times \frac{71}{100} \times 100$	56
1981	105	$\frac{105}{100} \times \frac{75}{100} \times \frac{71}{100} \times \frac{105}{100} \times 100$	59
1982	95	$\frac{105}{100} \times \frac{75}{100} \times \frac{71}{100} \times \frac{105}{100} \times \frac{95}{100} \times 100$	56
1983	90	$\frac{105}{100} \times \frac{75}{100} \times \frac{71}{100} \times \frac{105}{100} \times \frac{95}{100} \times \frac{90}{100} \times 100$	51

Note : Just as we defined price relatives, we can define quantity relative and other relatives.

$$\begin{aligned}
 \text{Quantity relative} &= \frac{\text{Number of units of a commodity in the current year}}{\text{Number of units of a commodity in the base year}} \times 100 \\
 &= \frac{q_1}{q_0} \times 100
 \end{aligned}$$

Similarly, we can define sales relative, output relative etc. Having found the relatives, we take the average of those relatives as index numbers. If the arithmetic mean is used to find the average, we have to define it as

$$p_{01} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right)}{n} \text{ where } n \text{ denotes the number of commodities.}$$

If Geometric mean is used, the formula to compute the index is:

$$p_{01} = \left[\pi \left(\frac{p_1}{p_0} \times 100 \right) \right]^{1/n} \text{ where } \pi \text{ denotes the product.}$$

$$\text{ie., } p_{01} = \text{antilog} \left[\frac{\sum \log \left(\frac{p_1}{p_0} \times 100 \right)}{n} \right]$$

$$= \text{antilog} \frac{\sum \log p}{n} \quad \text{where } p = \frac{p_1}{p_0} \times 100$$

This is called as geometric mean index numbers.

We know that geometric mean is useful in averaging rates, ratios and percentages. It is particularly suitable for the construction of index numbers because index numbers show percentage change rather than absolute amounts of change.

If Harmonic mean is used, the formula computing the index is

$$p_{01} = \frac{n}{\sum \left(\frac{p_0}{p_1 \times 100} \right)}$$

Example 5 :

Compute price index for the following data by the average of price relative method by using A.M, G.M., and H.M.

Commodity	Price in 1980 (Rs)	Price in 1985 (Rs)
A	20	25
B	30	30
C	10	15
D	25	35
E	40	45
F	50	55

Solution :

1) Index number using Arithmetic Mean of Price Relatives

Commodities	1980 (p_0) Rs	1985 (p_1) Rs	$p = \frac{p_1}{p_0} \times 100$
A	20	25	$\frac{25}{20} \times 100 = 125$
B	30	30	$\frac{30}{30} \times 100 = 100$
C	10	15	$\frac{15}{10} \times 100 = 150$
D	25	35	$\frac{35}{25} \times 100 = 140$
E	40	45	$\frac{45}{40} \times 100 = 112.50$
F	50	55	$\frac{55}{50} \times 100 = 110.00$
			737.50

$$p_{01} = \frac{\sum p}{n} = \frac{737.50}{6} = 122.92$$

2) Index number using Geometric mean of price relatives

Commodities	Price in 1980 (p_0) Rs	Price in 1985 (p_1) Rs	Price relatives $P = \frac{p_1}{p_0} \times 100$	log p
A	20	25	125.00	2.0966
B	30	30	100.00	2.0000
C	10	15	150.00	2.1761
D	25	35	140.00	2.1461
E	40	45	112.50	2.0512
F	50	55	110.00	2.0414
				12.5117

$$p_{01} = \text{Antilog} \left(\frac{\sum (\log p)}{n} \right)$$

$$= \text{Antilog} \left(\frac{12.5117}{6} \right) = \text{Antilog} (2.0853) = 121.7$$

Using H.M. $p_{01} = \frac{6}{\frac{1}{125} + \frac{1}{100} + \frac{1}{150} + \frac{1}{140} + \frac{1}{112.5} + \frac{1}{110}}$

$$= \frac{6}{0.050} = 120$$

Exercise

1. From the following data construct the fixed base relative for 1982 taking 1980 as base.

Commodities	Price in 1980 (Rs)	Price in 1982(Rs)
A	60	70
B	40	60
C	80	90
D	110	120
E	30	20

2. From the following data construct an index for 1970 taking 1962 as base by the price relative method using a) arithmetic mean b) geometric mean c) H.M. for averaging relatives.

Commodities	Price in 1968(Rs)	Price in 1970(Rs)
A	65	65
B	60	60
C	30	80
D	100	100
E	20	20

3. From the link relatives given below calculate the fixed base relatives

1975	1976	1977	1978
94	104	104	93

5.3 Weighted Index Numbers

All the commodities that enter into the construction of an index need not be of the same importance. Their relative importance will depend on the purpose for which index number is constructed to serve. For example, among the commodities that are considered in the construction of a cost of living index some may be regarded as luxuries and some necessities. Besides, the commodities in the different groups like food, clothing etc have varying importance according as the index is meant for a worker, middle class, or a villager. Hence to get a fair measure of the cost of living index, the price relatives have to be weighted in proportion to their importance.

Different schemes of weighing are proposed by different statisticians and hence we have a number of formulae for constructing index number. For a general price index, the importance of a commodity is measured by its money value in exchange. i.e., $p \times q$ where p is the average price of the commodity and q , the quantity of the commodity exchanged.

Some of the important formulae are the following :

1) Laspeyre's Price Index Number

Let p_0 , q_0 be the base year price and base year quantity and p_k , q_k the current year price and quantity. Taking $p_0 q_0$ as the weight, the weighted arithmetic mean of the price relatives

$$\left(\frac{p_k}{p_0} \right) \text{ is given by } \frac{\sum (p_0 q_0) \frac{p_k}{p_0}}{\sum (p_0 q_0)} = \frac{\sum p_k q_0}{\sum p_0 q_0}$$

This index number $\frac{\sum p_k q_0}{\sum p_0 q_0} \times 100$ ----- (1) is known as Laspeyre's index number.

(2) Paasche's Price Index number

The weighted harmonic mean of the price relatives $\left(\frac{p_k}{p_0} \right)$ with weights $(p_k q_k)$ is known as Paasche's price index number.

This index is given by the formula

$$\frac{\sum p_k q_k}{\sum p_k q_k \left(\frac{p_0}{p_k} \right)} = \frac{\sum p_k q_k}{\sum p_0 p_k} \times 100 \text{ (In percentage)} \text{ ----- (2)}$$

(3) Fisher's (Ideal) Price Index Number

It is the geometric mean of the indices of Laspeyre and Paache. This formula is defined by Prof. Fisher. It is given by

$$\sqrt{\frac{\sum p_k q_0 \sum p_k q_k}{\sum p_0 q_0 \sum p_0 q_k}} \times 100 \text{ ---- (3)}$$

(4) Edgeworth and Marshall's Price Index Number

It is given by the formula
$$\frac{\sum p_k (q_0 + q_k)}{\sum p_0 (q_0 + q_k)} \times 100 \quad \text{--- (4)}$$

(5) In general if w_k are the weights, the index number is given by

$$\left[\frac{\sum w_k \left(\frac{p_k}{p_0} \right)}{\sum w_k} \right] \times 100 \quad \text{--- (5)}$$

Example 6 : The following are group index numbers and the group weights of an average working class family's budget. Construct the index number for the given weights.

Group	Price Relative	Weights
Food	352	40
Fuel and Lighting	220	12
Clothing	230	10
Rent	160	3
Misc	190	15

Solution

We find the weighted arithmetic mean of the relatives $\frac{\sum w_k x_k}{\sum w_k}$ where x_k is the relative

x_k	w_k	$w_k x_k$
352	40	14080
220	12	2640
230	10	2300
160	13	2080
190	15	2850
	90	23950

$$\frac{\sum w_k x_k}{\sum w_k} = \frac{23950}{90} = 266.1$$

Example 7 : Calculate the price index number for 1981 by (1) Laspeyres's method (2) Paasche's method (3) Marshall and Edgeworth's method and (4) Fisher's method from the following data.

Commodity	1980		1981	
	Price	Quantity	Price	Quantity
A	4	50	10	40
B	3	10	9	2
C	2	5	4	2

Solution : We require $p_0 \cdot q_0$, $p_k q_0$, $p_0 q_k$ and $p_k q_k$

Commodities	p_0	q_0	p_k	q_k	$p_0 q_0$	$p_k q_0$	$p_0 q_k$	$p_k q_k$
A	4	50	10	40	200	500	400	160
B	3	10	9	2	30	90	18	6
C	2	5	4	2	10	20	8	4
					240	610	426	170

Index number of prices

$$\begin{aligned}
 (1) \quad \text{Laspeyre's Index} &= \frac{\sum p_k q_0}{\sum p_0 q_0} \times 100 \\
 &= \frac{610}{240} \times 100 = 254.2
 \end{aligned}$$

$$\begin{aligned}
 (2) \quad \text{Paasche's Index} &= \frac{\sum p_k q_k}{\sum p_0 q_k} \times 100 \\
 &= \frac{426}{170} \times 100 = 250.6
 \end{aligned}$$

$$\begin{aligned}
 (3) \quad \text{Marshall and Edgeworth Index} &= \frac{\sum p_k (q_0 + q_k)}{\sum p_0 (q_0 + q_k)} \times 100 \\
 &= \frac{\sum p_k q_0 + \sum p_k q_k}{\sum p_0 q_0 + \sum p_0 q_k} \times 100 \\
 &= \frac{610 + 426}{240 + 170} \times 100 \\
 &= \frac{1036}{410} = 252.7
 \end{aligned}$$

(4) Fisher's Index

$$= \sqrt{\frac{\sum p_k q_0}{\sum p_0 q_0} \times \frac{\sum p_k q_k}{\sum p_0 q_k}}$$

$$= \sqrt{\frac{610}{240} \times \frac{420}{170}} \times 100 = 252.4$$

Example 8: From the data given below, compute the index number of price by an appropriate method:

Commodity	Base year		Current year	
	Price	Total value	Price	Total value
A	40 units	240 units	30 units	210 units
B	45 "	180 "	50 "	250 "
C	90 "	45 "	40 "	60 "

Solution: Here we are given the total value itself. i.e, (pq) The appropriate index is Fisher's ideal index. We know that price x quantity = total value

$$\therefore \text{quantity} = \frac{\text{total value}}{\text{price}}$$

so the required table is

Commodity	p_0	q_0	p_k	q_k	$p_0 q_0$	$p_0 q_k$	$p_k q_k$	$p_k q_0$
A	40	6	30	7	240	280	210	180
B	45	4	50	5	180	225	50	200
C	90	0.5	40	1.5	45	135	260	20
			Total		465	640	520	400

Fisher's ideal index

$$= \sqrt{\frac{\sum p_k q_0}{\sum p_0 q_0} \times \frac{\sum p_k q_k}{\sum p_0 q_k}} \times 100$$

$$= \sqrt{\frac{400}{465} \times \frac{520}{640}} \times 100 = 83.60$$

Quantity Index Numbers

Price index numbers measure the changes in the price of certain goods. Similarly, quantity index numbers measure the changes in the physical volume of certain goods produced or distributed or consumed. These indices are useful as they indicate the level of output. In constructing quantity index numbers, we find simple averages of quantity relatives or using prices or values as weights, find weighted averages of quantity relatives.

Quantity indices can be easily obtained from price indices by interchanging p and q .

- 1) Laspeyre's Quantity index number $= \frac{\sum p_0 q_k}{\sum p_0 q_0} \times 100$
- 2) Paasche's Quantity index number $= \frac{\sum p_k q_k}{\sum p_0 q_0} \times 100$
- 3) Fisher's (ideal) Quantity index number $= \sqrt{\frac{\sum p_0 q_k}{\sum p_0 q_0} \times \frac{\sum p_k q_k}{\sum p_k q_0}} \times 100$
- 4) Edgeworth and Marshall's Quantity index number $= \frac{\sum (p_0 + p_k) q_k}{\sum (p_0 + p_k) q_0} \times 100$

Value index numbers

The value of a single commodity is the product of its price and quantity. Thus a value index is the ratio of the sum of the values of the goods in a year to the sum of their values in the base year.

$$\text{i.e., value index no: } \frac{\sum p_k q_k}{\sum p_0 q_0} \times 100$$

Since the weights are inherent in the value figures, only the aggregate index is taken.

Example: Compute Laspeyre's quantity index for the following data. Also compute the value index.

Commodity	Units consumed		Price per unit in Rs	
	1982	1985	1982	1985
A	20	16	1.2	1.5
B	35	38	2.1	2.7
C	10	9	3.0	3.5
D	45	50	0.8	1.2

Solution: The required table is as follows:

Commodities	q_0	q_k	p_0	p_k	$q_k p_0$	$q_0 p_0$	$q_k p_k$
A	20	16	1.2	1.6	19.2	24	24
B	35	38	2.1	2.7	79.8	73.5	102.6
C	10	9	3.0	3.5	27	30	31.5
D	45	50	0.8	1.2	40	36	60
					166	163	218.1

$$\text{Laspeyre's Quantity index} = \frac{\sum q_k p_0}{\sum q_0 p_0} \times 100$$

$$= \frac{166}{163.5} \times 100 = 101.50$$

$$\text{Value index} = \frac{\sum p_k q_k}{\sum p_0 q_0} \times 100 = \frac{218.1}{163.5} \times 100 = 133.5$$

Exercise

1) Construct Fisher's ideal price and Quantity Index Numbers for the following data. Also, find the value index number

Item No.	Price per unit in (Rs)		Quantities (No. of units)	
	1975	1978	1975	1978
1	9.2	15	5	5
2	8	12	10	11
3	4	5	6	6
4	1	1.25	4	8

2. Construct index numbers of price from the following data by (i) Laspeyre's method (ii) Paasche's method (iii) Marshall and Edgeworth's method (iv) and Fisher's ideal method. Construct also that corresponding quantity index numbers and value index.

Commodities	1968		1975	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

(Ans 125, 126.2, 125.5 and 125.5)

Tests on Index Numbers

A good index number should satisfy the following tests:

- (i) The commodity Reversal Test
- (ii) The Time Reversal Test
- (iii) The Factor Reversal Test
- (iv) Circular Test.

(i) The commodity Reversal Test

It requires that the index number should be independent of the order in which the different commodities are considered. This is satisfied by all the index numbers since any average simple or weighted - does not depend on the order of the constituent items.

(ii) The Time Reversal Test

If the same base year and the current year are interchanged, the index number of the latter should be the reciprocal of the former, i.e., their product should be equal to 1.

The simple aggregate index satisfies this test for

$$\frac{\sum p_1}{\sum p_0} \times \frac{\sum p_0}{\sum p_1} = 1$$

The Arithmetic Mean and Harmonic mean index of price relatives do not satisfy this test.

$$\text{ie., } \frac{1}{n} \sum \left(\frac{p_k}{p_o} \right) \times \frac{1}{n} \sum \left(\frac{p_o}{p_k} \right) \text{ need not be equal to 1.}$$

$$\text{Also } \left[\frac{n}{\sum \left(\frac{p_o}{p_k} \right)} \times \frac{n}{\sum \left(\frac{p_k}{p_o} \right)} \right] \text{ need not be equal to 1.}$$

But the Geometric Mean index satisfies the time reversal test for

$$\left(\pi \frac{p_k}{p_o} \right)^{\frac{1}{n}} \times \left(\pi \frac{p_o}{p_k} \right)^{\frac{1}{n}} = 1$$

where π represents the product.

Bias in Index Numbers

We know that $A.M. \geq G.M. \geq H.M$ So if I_A, I_G, I_H denote the A.M., G.M., and H.M., of relatives then, $I_A \geq I_G \geq I_H$ (1)

Let I'_A, I'_G, I'_H be the corresponding indices after interchanging the base and the current year. But we have $I'_A \geq I'_G \geq I'_H$ (2)

From (1) and (2), we have

$$I_A I'_A \geq I_G I'_G \geq I_H I'_H$$

$$\text{But } I_G I'_G = 1.$$

$$\therefore I_A I'_A \geq 1 \text{ and } I_H I'_H \leq 1.$$

So, Arithmetic mean index is said to have an upward bias and the Harmonic mean index is said to have a downward bias. This bias can be rectified by taking the geometric mean of one A.M. index and one H.M. index. This process of getting index number is known as "the method of crossing."

In weighted index numbers, clearly Laspeyre's index number and Paasche's index number do not satisfy this test. But Fisher's index number got by the method of crossing from Laspeyre's and Paasche's index number satisfies this test for

$$I_{ok} = \sqrt{\frac{\sum p_k q_o}{\sum p_o q_o} \times \frac{\sum p_o q_k}{\sum p_k q_k}} \text{ and}$$

$$I_{ko} = \sqrt{\frac{\sum p_o q_k}{\sum p_k q_k} \times \frac{\sum p_k q_o}{\sum p_o q_o}} \text{ (got by interchanging the current and base year)}$$

$$I_{ok} \times I_{ko} = 1$$

Also Marshall Edgeworth index number satisfies this test for

$$I_{ok} = \frac{\sum p_k (q_o + q_k)}{\sum p_o (q_o + q_k)}, \quad I_{ko} = \frac{\sum p_o (q_o + q_k)}{\sum p_k (q_o + q_k)}$$

$$\text{and } I_{ok} \times I_{ko} = 1.$$

(iii) The Factor Reversal Test

This test is applicable only to weighted index numbers. This test implies that an index of prices should serve equally well as an index of quantities. In the index number formula, if p and q be interchanged the product of the two indices so formed should be equal to the ratio of the aggregate values in the two years. i.e., if $I(p, q)$ is a given weighted index number and $I(q, p)$ is the one got from it by interchanging p and q we should have

$$I_{(p,q)} \times I_{(q,p)} = \frac{\sum p_k q_k}{\sum p_o q_o}$$

Only Fisher's index number satisfies this test for

$$I_{(p, q)} = \sqrt{\frac{\sum p_k q_0}{\sum p_0 q_0} \times \frac{\sum p_k q_k}{\sum p_0 q_k}} \quad \text{and}$$

$$I_{(q, p)} = \sqrt{\frac{\sum p_0 q_k}{\sum p_0 q_0} \times \frac{\sum p_k q_k}{\sum p_k q_0}} \quad \text{and so}$$

$$I_{(p, q)} \times I_{(q, p)} = \sqrt{\frac{(\sum p_k q_k)^2}{(\sum p_0 q_0)^2}} = \frac{\sum p_k q_k}{\sum p_0 q_0}$$

An index number is defined as an **ideal index number** if it satisfies all the three tests namely commodity reversal test, time reversal test and factor reversal test.

Example: Since only Fisher's index number satisfies all these three tests it is called an ideal index.

(iv) Circular test

This test is an extension of the time reversal test. The test requires that if an index is constructed for the year 1 on base year 0 and for the year 2 on base year 1 we should get the same result as if we calculated directly the index for the year 2 on base year 0 without going through year 1 as intermediary.

Symbolically, if these three index numbers are P_{01} , P_{12} and P_{02} the circular test is satisfied if $P_{01} \times P_{12} = P_{02}$

i.e $P_{01} \times P_{12} \times P_{20} = 1$ if it satisfies the time reversal test.

We can show that none of the four weighted index numbers including Fisher's ideal index satisfies this test. But the simple aggregate index satisfies this test for

$$\frac{\sum P_1}{\sum P_0} \times \frac{\sum P_2}{\sum P_1} \times \frac{\sum P_0}{\sum P_2} = 1$$

5.4 Cost of Living index numbers

The cost of living index numbers we designed to measure the change in the cost of maintaining the same standard of living as in the base year.

Since different classes of people consume different types of commodities, the changes in the prices affect them differently. Also, the pattern of consumption changes from region to region. So, the same index number cannot serve the needs of all classes and all regions. So the country is divided into different homogeneous regions and in each region, the index numbers are computed separately for the different classes of people.

Next, the commodities consumed and the weights attached to these, should be decided. This is done by conducting sample family budget enquiries. This enquiry should supply information regarding the nature, quality and the quantities of the commodities consumed by the people so that we know their relative importance. Then retail prices of the different commodities should be obtained. Then we find an appropriate index number.

It should be carefully noted that cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level. At present the term 'cost of living index' is replaced by 'price of living index', or 'cost of living price index' or 'consumer price index'.

Uses of consumer Price Indices

- 1) They provide guidelines to the government for deciding dearness allowance and general economic policies.
- 2) They are used to measure changing purchasing power of the currency, real income etc.
- 3) They are also used for analysing markets for particular kinds of goods and services.

Construction of a Consumer Price Index

1. Decisions

Decision about the class of people for whom the index is meant. It is to be decided first whether the index is meant for industrial workers, government employees, teachers, etc., The scope of the index as well as the geographical area to be covered should also be decided. It is necessary that the selected class should form a homogeneous group of people with respect to income.

2) Conducting Family Budget Enquiry

The Second step is to conduct a sample 'family budget enquiry' using the technique of random sampling. The enquiry should cover reasonably adequate number of families and should be conducted during a normal period. This would provide us with the information about the amount, an average family spends on different items of consumption. This would enable us to determine the weights for computing the index. Only those items which are used by majority of the class of people are to be included in the index. These items may be classified into certain major groups such as food, cloth, fuel and light, house rent and miscellaneous. Each of these groups may further be divided into subgroups eg., food may be subdivided into wheat, rice, pulses, ghee, sugar etc., The miscellaneous includes items such as medicine, education, gifts, newspapers, dhobi etc.

3) Obtaining Price Quotations

It is not so easy to collect retail prices because such prices may vary from place to place, shop to shop and person to person. Price quotations should be obtained from the localities in which the class of people concerned reside or from where they usually make their purchases. In this connection the following points are worth considering:

- (a) The retail prices should relate to a fixed list of items and for each item the quality should be fixed by means of suitable specifications.
- (b) Retail prices should be those actually charged to consumers or cash sales.
- (c) The retail prices of the commodities should be collected at regular intervals of time.

Cost of living index numbers are generally constructed for each week. The average of the weekly index number is taken as the index number for a month. The average of monthly index numbers gives the cost of living index for the whole year.

4) Methods of Construction

There are two methods of constructing the cost of living index number. They are:

- i) The Aggregate Expenditure Method
- ii) The family budget method.

i) The Aggregate Expenditure Method

In this method, we assume that the quantities consumed in the base year are also being consumed in the current year. So the index number is a ratio between the current year aggregate expenditure and the base year aggregate expenditure.

$$\therefore \text{Consumer Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Note: Actually it is the same as Laspeyre's Weighted Aggregate index Number.

(ii) The Family Budget Method

Under this method the following steps are to be taken for the construction of a Consumer price index number.

- 1) Carefully study the family budgets of a large number of people for whom the index is meant.
- 2) Estimate the aggregate expenditure of an average family on various items consumed by the particular group of people.
- 3) The aggregate expenditure as found out in (2) on different items shall constitute the weights that is $v = p_0 q_0$.
- 4) Obtain the price relatives for each commodity i.e, $p = \frac{p_k}{p_0} \times 100$
- 5) Multiply the price relatives by the value weights and get p_v .

6) PV is divided by the sum of the weights, v symbolically, Consumer Price Index $= \frac{\sum pv}{\sum v}$

$$= \frac{\sum \frac{p_k}{p_0} (p_0 q_0)}{\sum p_0 q_0} \times 100$$

$$= \frac{\sum p_k q_0}{\sum p_0 q_0} \times 100$$

Example 1

An enquiry into the budgets of the middle class families in Bombay gave the following information:

Expenses on	Food	Rent	Clothing	Fuel	Misc
	35%	15%	20%	10%	20%
Price(1982) Rs.	150	50	100	20	60
Price (1983) Rs	174	60	125	25	90

What changes in the cost of living figure of 1983 have taken place is compared to 1982?

Solution: The percentage under each item of consumption such as food, fuel etc shows the proportionate expenses incurred on those items out of the total expense of 100%. Hence these are to be taken as weight. The required table is:

Items	P_0	P_1	V	$P = \frac{P_1}{P_0} \times 100$	PV
Food	150	174	35	$\frac{174}{150} \times 100 = 116$	4060
Rent	50	60	15	$\frac{60}{50} \times 100 = 120$	1800
Clothing	100	125	20	$\frac{125}{100} \times 100 = 125$	2500
Fuel	20	25	10	$\frac{25}{20} \times 100 = 125$	1250
Misc	60	90	20	$\frac{90}{60} \times 100 = 150$	3000
			$\sum V = 100$		$\sum PV = 12610$

$$\text{Cost of Living index} = \frac{\sum PV}{\sum V} = \frac{12610}{100} = 126.1$$

This indicates that in Bombay the cost of living of the middle class families had gone up by 26.1% in 1983 compared to the cost of living in 1982.

Example 2

Construct the cost of living index for 1986 based on 1980 from the following data.

Group	Food	Housing	Clothing	Fuel and Light	Misc
Group Index No for 1986 based on 1980	122	140	112	116	106
Weights	32	10	10	6	42

Solution: The required table is

Group	Index Number P	Weights V	PV
Food	122	32	3904
Housing	140	10	1400
Clothing	112	10	1120
Fuel and Light	116	6	696
Misc	106	42	4452
		$\sum V = 100$	$\sum PV = 11572$

$$\text{Cost of living index} = \frac{\sum PV}{\sum V} = \frac{11572}{100} = 115.72$$

Example 3

Calculate the index number using the aggregate Expenditure method for the 1982 with 1980 as base year from the following data

Commodity	Quantity in units in 1980	Price per unit in 1980 Rs	Price per unit in 1982 Rs
A	100	8	12
B	10	5	5.25
C	20	48	52
D	65	15	16.50
E	80	19	27

Solution

Commodity	q_0	p_0	p_1	$p_1 q_0$	$p_0 q_0$
A	100	8	12	1200	800
B	10	5	5.25	52.5	50
C	20	48	52	1040	960
D	65	15	16.50	1072.5	975
E	30	9	27	810	570
				4175	3355

$$\text{Consumer Price Index} = \frac{4175}{3355} \times 100$$

$$= 124.5$$

Exercise

- Construct the Cost of living index number for 1985 on the basis of 1983 from the following data using (i) the aggregate expenditure method (ii) the family budget method

Commodities	Quantity in 1983	Price in 1983		Price in 1985	
		Rs.	Ps.	Rs.	Ps.
A	6	5-75		9-00	
B	6	5-00		8-00	
C	1	6-00		9-00	
D	6	8-00		10-00	
E	4	2-00		1-50	
F	1	20-00		15-00	

- What is meant by value weights? Give an example
- What is meant by bias in index numbers?
- Prove using the following data that the Factor Reversal test and Index reversal test are satisfied by Fisher's ideal formula for index number.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	12	60
D	10	30	6	24

5.5 Analysis of Time Series

One of the most important tasks before economists and businessmen these days is to make estimates for the future. The first step in making estimates for the future consists of gathering information from the past. In this connection one usually deals with statistical data which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as 'time series'. Thus when we observe numerical data at different points of time the set of observations is known as time series. For example, if we observe production, sales, population, imports, exports, etc., at different points of time, say over the last 5 or 10 years, the set of observations formed shall constitute time series. Hence, in the analysis of time series, time is the most important factor because the variable is related to time which may be either year, month, week, day, hour or even minutes or seconds.

Utility of time series analysis

1. It helps in understanding past behaviour
2. It helps in planning future operations.
3. It helps in evaluating current accomplishments.
4. It facilitates comparison.

Components of time series

There are four types of patterns, movements or components of a time series. They are:

- 1) Secular trend
- 2) Seasonal Variations
- 3) Cyclical Variations
- 4) Irregular Variations.

It may be noted that any or all of these components may be present in any particular series.

Let us have a detailed explanation of each of the four components.

1. Secular Trends

The component of a time series which reveals the general tendency of the data over a long period is called the secular trend. The general tendency may be to increase or decrease but it cannot be both ways. All this increase or decrease is not sudden. It is very slow and gradual. And the factors for this gradual change are changes in population or tastes and habits of people, technological development, improvements in technical and productivity etc., secular trend is also called 'long time trend'.

Trend values are measured by anyone of the following methods:

- (i) 'Free hand' Method
- (ii) Semi-Average Method
- (iii) Moving Average Method
- (iv) Method of Least Square

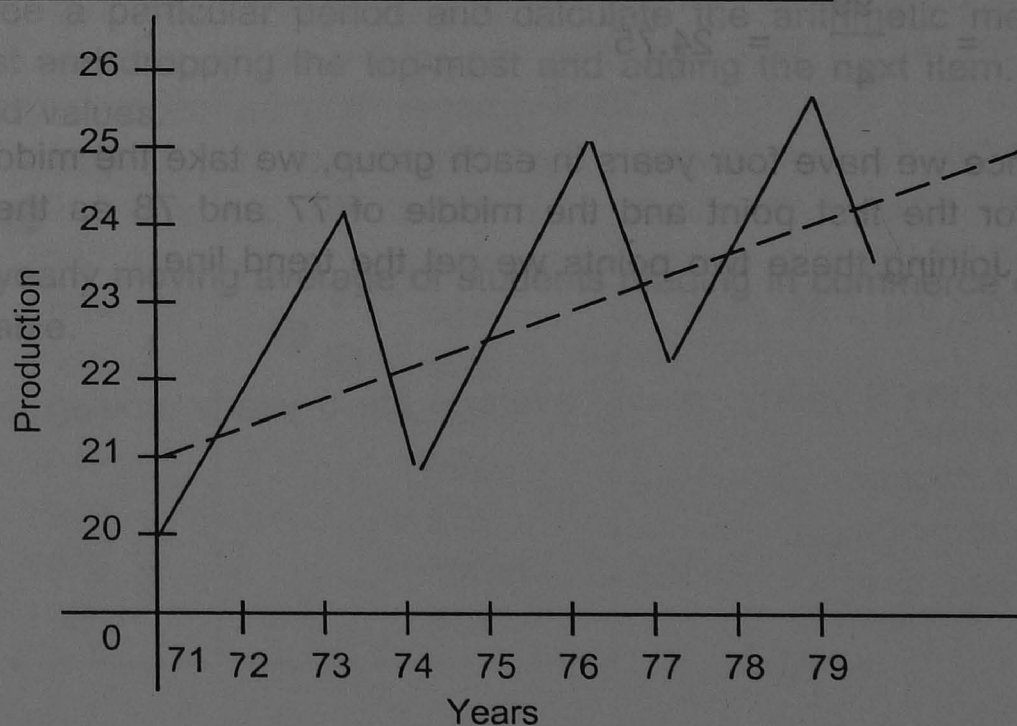
(i) Free hand Method

This is the simplest method of studying trend. In this method we first draw the graphs of the series by taking the time along the x-axis and the values of the given variable along the y-axis and then draw a straight line which will best fit the data. The line will give the trend values. There is no specific rule for drawing such a line. We draw it in such a way that it passes between the plotted points and that are fluctuations on one side are approximately equal to those in the other side and it shows the general movements.

Example 1: Fit a trend line to the following data by the free hand method.

Year	Production of steel in mil. tones
1971	20
1972	22
1973	24
1974	21
1975	23
1976	25
1977	23
1978	26
1979	25

Solution



This is the simplest method of estimating trend. If a person has experience in computing trend and has knowledge of the economic history of the concern of industry, the trend line drawn by him will be much better than the line got by a rigid mathematical formula.

But this method is highly subjective because the trend line depends on the personal judgement of investigator and so, different persons may draw different trend lines for the same set of data.

(ii) Semi-Average Method

The given data are divided into two equal parts. If there are odd number of years like 7, 9, etc., the middle year is left and then the two equal parts are formed. An average of each part is calculated and two points are plotted with the middle period of the two parts and the corresponding averages as coordinates. A straight line is drawn through these two points. This line is taken as the trend line.

Example 2

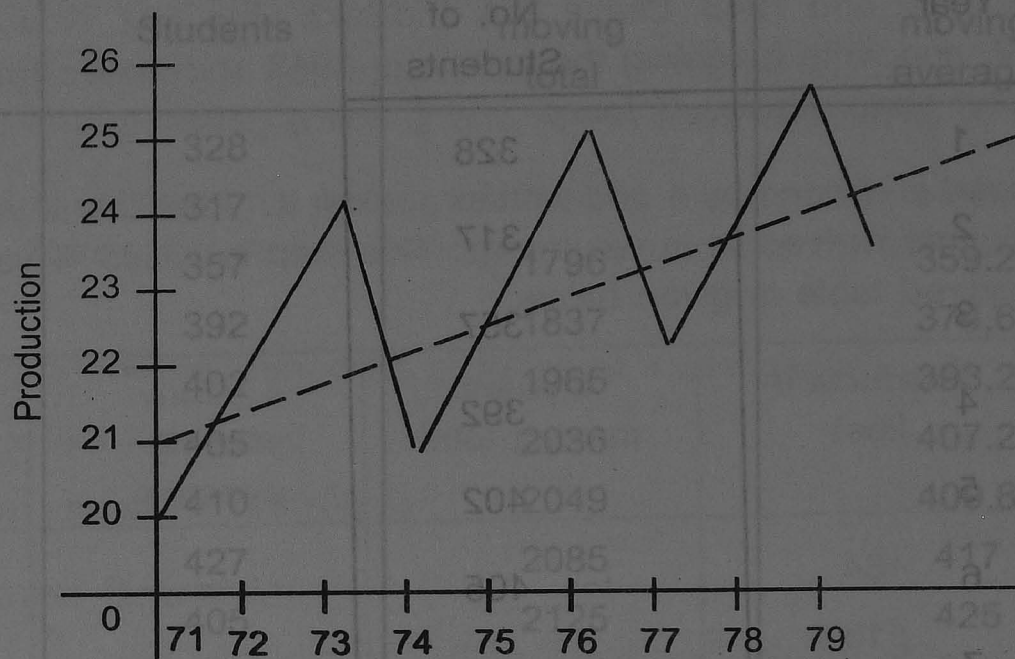
Fit a trend line to the data in Example:1 by the method of semi-averages.

Solution: Since nine years are given, we omit the middle year and find the average of the first four and the last four years

$$\begin{aligned}\text{Average of the first four years} &= \frac{20 + 22 + 24 + 21}{4} \\ &= \frac{87}{4} = 21.75.\end{aligned}$$

$$\begin{aligned}\text{Average of the second four years} &= \frac{25 + 23 + 26 + 25}{4} \\ &= \frac{99}{4} = 24.75\end{aligned}$$

Again, since we have four years in each group, we take the middle of 72 and 73 as the x co-ordinate for the first point and the middle of 77 and 78 as the x-co ordinate for the second point. Joining these two points we get the trend line.



(iii) Moving Average Method

The moving average method is a simple process of trend measurement which is quite accurate under certain conditions. The main purpose of this method is to obtain the trend values so that all types of fluctuations are eliminated, or reduced to a minimum. We choose a period and find the average of first three years and next three years and so on. Thus we get a series of averages by dropping the top item of each group and adding the next item. These averages, called the moving averages give the trend values since all other fluctuations are almost wiped out by taking the average.

A long period moving average is good from the point of view of reducing irregular fluctuations. But longer the period, the further away are the trend values from the original ones. Hence the period of the moving average should neither be too long to distort the trend values nor be too short to make it impossible to reduce irregular fluctuations.

So, we first see a particular period and calculate the arithmetic mean of this period starting from the first and dropping the top-most and adding the next item. These averages will give us the trend values.

Example 2

Calculate five yearly moving average of students reading in commerce colleges as shown from the following table.

1945	502
1946	505
1947	557
1948	571
1949	606
1950	612

Year	No. of Students
1	328
2	317
3	357
4	392
5	402
6	405
7	410
8	427
9	405
10	428
11	445
12	447
13	480
14	482

In the above, after writing the necessary data the first two columns, we introduce a third column for 'moving total' for convenience. Thus the first five years total i.e., $328+317+357+392+402 = 1796$ is written against the middle of these 5 years namely, the 3rd year. Then the first year figure 328 is dropped and the sixth year figure 405 is taken up along with the other 4 years figures. Then the total of 317, 357, 392, 402 and 405 is found out and is written against the middle of 2nd, 3rd, 4th, 5th and 6th years - namely the 4th. This is continued till the total of last 5 years figures namely 438, 445, 447, 480, 482 is found and written against the 12th year.

Then to find the 5 yearly moving average the 5 yearly moving totals are to be divided by 5.

Year	No. of Students	5 yearly moving total	5 yearly moving average
1	328		
2	317		
3	357	1796	359.2
4	392	1837	374.6
5	402	1965	393.2
6	405	2036	407.2
7	410	2049	409.8
8	427	2085	417
9	405	2125	425
10	428	2162	432.4
11	445	2215	443
12	447	2292	458.4
13	480		
14	482		

Since the period was odd, it enabled us to put the moving total and hence moving average against the middle year. Suppose, the period is even, the moving average in this case does not correspond to any original time period. This difficulty is overcome by a method known as centering. This is best explained in the following example.

Example 4

Assuming a 4 yearly cycle calculate the trend by the method of moving averages from the following data relating to the production of tea in India.

Year	Production (lbs)
1941	464
1942	515
1943	518
1944	467
1945	502
1946	540
1947	557
1948	571
1949	586
1950	612

Solution: Find out the sum of the first four years and write it against the middle of the 2nd and 3rd year i.e, 1942 and 1943. Leave 464 and take upto 502. Now add the figures 515, 518, 467 and 502, and write it against the middle of 1943 and 1944. Proceed like this till the sum of the last four years figures is found

Next each total is divided by 4 and written against it. After this 2 yearly moving average is found for the figures arrived at in column 4. This step is called '4 yearly moving average centred'. The required table is given below:

Year	Production (lbs)	4 years moving total	4 yearly moving average	4 yearly moving average
1941	464			
1942	515			
1943	518	1964	491.00	495.7
1944	467	2002	500.50	503.6
1945	502	2027	506.75	511.6
1946	540	2066	516.50	529.5
1947	557	2170	542.50	553.0
1948	571	2254	563.50	572.5
1949	586	2326	581.50	
1950	612			

The figures under the last column are called trend values.

Merits of Moving Average Method

- (i) It is a simple method
- (ii) This method is not subject to personal prejudice and bias of the estimator.
- (iii) If the period of moving average corresponds with the period of cycle, the cyclical variations can be completely eliminated.
- (iv) It is a flexible method of measuring trend for the reason that if a few more figure are added to the data, the entire calculations are not changed. We get some more trend values.

Limitations of the Moving Average Method

- 1) There are no set rules available for the choice of the period. Hence great care has to be exercised in selecting the period of moving average.

- 2) Trend values cannot be computed for all the years.
- 3) Since the moving average is not represented by a mathematical functions, this method cannot be used in forecasting which is one of the main objectives of trend analysis.
- 4) When the trend situation is not linear, the moving average lies either above or below the true value. Hence the moving average is appropriate for trend computations only when
 - a) forecasting is not the immediate purpose of investigation.
 - b) the trend is not linear and
 - c) the cyclical variations are regular.

(iv) Method of Least Squares

One of the best ways of obtaining trend values is the method of least squares. It is a mathematical method and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied.

- i) The sum of deviations of the actual values of the variable and its computed values is zero.
- ii) The sum of the squares of the deviations of the actual and computed values is the least.

The line obtained by this method is called the line of best fit. We know that a straight line is represented by $y = a + bx$ ----- (1) where y represents the estimated values of the trend and x represents the deviation in some convenient time periods. Once the unknowns a and b are found by using the given data, we can find the equation to the straight line trend.

Using the principle of least squares, we get the normal equations in a and b as

$$na + b \sum x = \sum y \quad \text{---- (2)}$$

$$a \sum x + b \sum x^2 = \sum xy \quad \text{---- (3) where } n = \text{the number of pairs of } x \text{ and } y \text{ values.}$$

Let $y_1, y_2, \dots, y_n$ and $x_1, x_2, \dots, x_n$ represent the variables and let $y_1 = a + bx_1, y_2 = a + bx_2, \dots, y_n = a + bx_n$. Adding these, we get $\sum y = na + b\sum x$ which is (2). Now multiply by $x_1, x_2, \dots, x_n$ respectively the equations $y_1 = a + bx_1, y_2 = a + bx_2, \dots, y_n = a + bx_n$.

$$\text{We get } x_1 y_1 = ax_1 + bx_1^2$$

$$x_2 y_2 = ax_2 + bx_2^2$$

$$\vdots$$

$$x_n y_n = ax_n + bx_n^2$$

Adding we get $\Sigma xy = a \Sigma x + b \Sigma x^2$ which is (3)

Choose the mid point in time as the origin so that the negative values in the first half of the series balance out the positive values in the second half i.e, $\Sigma x = 0$. This means that the time variable is measured as a deviation from its mean. Since $\Sigma x = 0$, (2) and (3) become

$$na = \Sigma y$$

$$\text{and } b \Sigma x^2 = \Sigma xy$$

$$\therefore a = \frac{\Sigma y}{n} \quad \text{and } b = \frac{\Sigma xy}{\Sigma x^2} \quad \text{----(4)}$$

Note: a refers to the intercept and it is the arithmetic mean of y. b indicates the slope of the line which gives the change in y for unit change in x.

Example 5 : Fit a straight line trend to the following data by the method of least squares. Compute the trend values for all the years from 1974 to 1978. Estimate the sales for 1979. Obtain also the short-term oscillations.

Year:	1974	1975	1976	1977	1978
Sales:	35	56	79	80	40
(in 1000 Rs)					

Solution: For the application of the formulae

$$a = \frac{\Sigma y}{n} \quad \text{and } b = \frac{\Sigma xy}{\Sigma x^2} \quad \text{we have to the origin at the middle of the}$$

years given.

We take 1976 as O and first fill up the first five columns.

Years	X	Sales (in 1000 Rs) Y	XY	X ²	Trend Values Y _c	Short term Oscillations
1974	-2	35	-70	4	51.2	-16.2
1975	-1	56	-56	1	54.6	+1.4
1976	0	79	0	0	58.0	+21.0
1977	1	80	80	1	61.4	+18.6
1978	2	40	80	4	64.8	-24.6
Total	0	290	34	10		

After completing all the first five columns in the table above we determine a and b.

$$a = \frac{\Sigma y}{n} = \frac{290}{5} = 58$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{34}{10} = 3.4$$

$\therefore$ The trend equation is $y_c = 58 + 3.4x$. This is the straight line trend required. y_c denotes calculated trends. For 1974 to 1978, y_c 's are calculated as follows.

Start from 1976. Substituting $x = 0$

$$y_c = 58 + (3.4 \times 0) = 58$$

$\therefore$ For 1976 the trend value is 58.

For 1977, $x = 1$

$$\therefore y_c = 58 + (3.4 \times 1)$$

$= 58 + 3.4 = 61.4$. This is got by adding the b value 3.4 with the previous year trend value. So for 1978, the trend value will be $61.4 + 3.4 = 64.8$. For 1975, we have to subtract 3.4 from 58. We get 54.6 for 1974, $54.6 - 3.4 = 51.2$

The estimated sales for 1979 is $64.8 + 3.4 = 68.2$. y is the observed value. y_c / calculated value. The difference $y - y_c$ gives the short term oscillations.

Example 6

Fit a straight line trend equation by the method of least squares and estimate the trend values:

Year:	1961	1962	1963	1964	1965	1966	1967	1968
Value:	80	90	92	83	94	99	92	104

Solution :

Here we have even number of years. The formulae for 'a' and 'b' are applicable only when $\Sigma x = 0$. So the origin is to be taken at the middle of 1964 and 1965. In that case $x = -\frac{1}{2}$ for 1964 and $x = \frac{1}{2}$ for 1965. To avoid fraction, we multiply the deviation by 2.

Original number	Deviations from 1964-65	Deviations multiplied by 2 = x	Y (Value)	XY	X ²	Trend Values y _c
1961	-3.5	-7	80	-560	49	83.0
1962	-2.5	-5	90	-450	25	85.5
1963	-1.5	-3	92	-276	9	88.0
1964	-0.5	-1	83	-83	1	90.5
	0.0					
1965	0.5	1	94	94	1	93.0
1966	1.5	3	99	297	9	95.5
1967	2.5	5	92	460	25	98.0
1968	3.5	7	104	728	49	100.5
Total		0	734	210	168	

$$a = \frac{\Sigma y}{n} = \frac{734}{8} = 91.75$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{210}{168} = 1.25$$

∴ The trend equation is $y_c = a + bx$

$$= 91.75 + 1.25x$$

For 1965, $y_c = 91.75 + 1.25 \times 1$

$$= 91.75 + 1.25 = 93.0$$

For 1966, $y_c = 91.75 + 1.25 \times 3$

$$= 91.75 + 3.75 = 95.5$$

For 1967, $y_c = 95.5 + 2.50 = 98.0$

For 1964, $y_c = 93 - 2.50 = 90.50$

For 1963, $y_c = 90.50 - 2.50 = 88.0$

For 1962, $y_c = 88 - 2.50 = 85.5$

For 1961, $y_c = 85.5 - 2.50 = 83.0$

For 1968, $y_c = 98.0 + 2.50 = 100.50$

We calculate the trend values mainly for the following two reasons:

- (1) To study the manner in which the value of a variable behaves in a long time. This study is possible only if we can eliminate the other components of the time series.
- (2) To study both the regular and irregular fluctuations, we need the trend value.

Exercise

1. Calculate by (i) freehand method (ii) by method of semi-average and (iii) five yearly moving average the trend of the following time series.

Year	No. of failures
1960	26
1961	28
1962	32
1963	20
1964	12
1965	12
1966	10
1967	9
1968	13
1969	11
1970	14

2. Determine the 3 yearly moving average for the following data

Year	1972	1973	1974	1975	1976	1977	1978	1979	1980
Sales in lakh units	65	95	80	115	105	135	125	150	140

3. Fit a straight line trend to the following data.

Year :	1959	1960	1961	1962	1963	1964	1965	1966
Sales in :	38	40	65	72	69	60	87	96

- (i) Compute the trend values for all these years.
- (ii) Estimate the sales for 1967.
- (iii) Obtain the short-term fluctuations.
4. What is a time series? How will you analyse the time series?
5. Write short notes on moving average.

(2) Seasonal Variations

Seasonal variations refer to the identical or almost identical pattern which a time series appears to follow during corresponding periods of successive years. Such movements are due to recurring events which take place annually. Because of natural factors such as climate and man-made conventions such as festivals, most business or economic activities vary in some systematic pattern during the seasons of the year-hence the term seasonal variations.

Changes in climate and weather have profound influence in causing seasonal variations. Climatic conditions including variations in rainfall, sunshine heat and wind produce variations in agricultural productions. The consumption of electricity is greater in summer than in winter.

Customs, habits and traditions also have a considerable effect on time series. For instance during Deepavali and Christmas days there will be a marked peak in retail sales of clothes. During the first week of every month bank deposits and daily sales in retail shops are at the highest level.

We measure the seasonal variations using one of the following methods.

- (i) Method of simple averages.
- (ii) Ratio to moving average method.

(i) Method of simple averages

This is the simplest method of obtaining a seasonal index. Here we collect the monthly (or quarterly or weekly or any period given) data over a number of years given and find the averages for each month. These averages are termed 'monthly averages'. Then the average of these monthly averages is computed and the ratio:

$$\frac{\text{Monthly average for a month}}{\text{Average of monthly averages}} \times 100 \text{ is calculated}$$

These ratios are the required seasonal indices. The steps in calculating the seasonal indices are

- i) Tabulate the given data in the following way - months or quarters or weeks in the first column, years or any or any other periods given in the first row and fill up the body of the table.
- 2) Find the total of each month -ie, find the row-wise total and then the mean by dividing the total by the number of years given. These are the monthly averages.
- 3) Find the total of these monthly averages and find the mean of these figures. This is the general average.

4) Compute seasonal index for month

$$= \frac{\text{Monthly average for that month}}{\text{General Average}} \times 100$$

Example 1

Find the seasonal indices, assuming that the data do not exhibit any trend or cycle.

Month	Sales in thousand units					
	1977	1978	1979	1980	Monthly total	Monthly average
Jan	10	11	10	12	43	10.75
Feb	12	11	12	13	48	12.00
March	13	12	11	13	49	12.25
April	15	13	12	15	55	13.75
May	16	14	13	16	59	14.75
June	16	14	15	18	63	15.75
July	17	15	15	20	67	16.75
Aug	18	15	17	20	70	17.50
Sep	18	15	18	21	72	18.00
Oct	19	16	20	22	77	19.25
Nov	22	18	22	24	86	21.50
Dec	22	10	24	25	81	20.25
					192.50	

$$\text{General Average} = \frac{192.50}{12} = 16.04 = 16$$

$$\text{Seasonal index for Jan} = \frac{\text{Monthly average for Jan}}{\text{General average}} \times 100$$

$$= \frac{10.75}{16} \times 100 = 167.2$$

Similarly we calculate for other months, Tabulating these values we have

Month	Seasonal Index	
Jan	$\frac{10.75}{16} \times 100$	167.2
Feb	$\frac{12}{16} \times 100$	75.0
March	$\frac{12.25}{16} \times 100$	76.6
April	$\frac{13.75}{16} \times 100$	85
May	$\frac{14.75}{16} \times 100$	92.2
June	$\frac{15.75}{16} \times 100$	98.4
July	$\frac{16.75}{16} \times 100$	104.7
August	$\frac{17.50}{16} \times 100$	109.4
Sep	$\frac{18}{16} \times 100$	112.5
Oct	$\frac{19.25}{16} \times 100$	120.3
Nov	$\frac{21.5}{16} \times 100$	134.4
Dec	$\frac{20.25}{16} \times 100$	126.6

Since this method ignores the trend factor, the indices obtained by this method may not represent the seasonal variations. Most of the economic series have trends and therefore the seasonal index computed by this method is actually an index of trends and seasonals. So, this index serves the purpose only when no definite trend exists.

(ii) Ratio to moving Average Method

This is the most widely used method of measuring seasonal variations. Here we first eliminate seasonal variations from the data by smoothening the original data. Since seasonal variations recur every year i.e., these variations have a time span of 12 months (or 4 quarters) a centered 12 month moving average (or a centered 4 quarterly moving average) eliminates seasonal and irregular fluctuations. Assuming the multiplicative model, we remove the trend and cyclic fluctuations by dividing the original data by these moving averages (ie y/TC) and express the quotient in percentages. Next we find the seasonal index using the first method. This will eliminate the I factor.

Before look into the example let us see the method of analysis of time series.

The analysis of time series consists of separating or decomposing each of the four components present in the data. There are two methods available to study the decomposition of a time series. (a) The Additive Model (b) The Multiplicative Model.

According to the additive model,

$y = T + S + C + I$. where Y stands for the original data and T, S, C, I for trend, seasonal variations, cyclical fluctuations and irregular fluctuations respectively.

In this model, all the components are expressed in the units in which the given series is. All the components are treated as residuals.

$$\text{i.e., } S + C + I = Y - T$$

$$C + I = Y - T - S$$

$$I = Y - T - S - C$$

For the multiplicative model, it is assumed that the output of an economy is the product of various forces operating on one another. i.e. $y = T \times S \times C \times I$.

In this model, trend alone is in the units of the original series. Other components are expressed as relatives or percentages.

Multiplicative model is mostly used in practice rather than the additive model.

Example : Calculate the seasonal indices by the ratio to moving average method for the following data:

Year	I Quarter	IIQ	IIIQ	IVQ
1971	68	62	51	63
1972	65	58	66	61
1973	68	63	63	67

Solution

Year	Quarter	Given Figures	4 Quarterly moving total	4 Quarterly moving average	4Qly moving average centred	Given figure as % of moving average
1971	I	68				
	II	62				
	III	61	254	63.50	63.125	96.63
	IV	63	251	62.75	62.250	101.20
1972	I	65	247	61.75	62.375	104.21
	II	58	252	63.00	62.750	92.43
	III	66	250	62.50	62.875	104.97
	IV	61	253	63.25	62.875	95.50
1973	I	68	258	64.50	64.125	106.04
	II	63	255	68.75	64.500	97.67
	III	63	261	63.25		
	IV	67				

The last column is obtained in the following way

$$1971 - \text{III Q} = \frac{61}{63.125} \times 100$$

$$1971 - \text{IV Q} = \frac{63}{62.250} \times 100 \text{ etc}$$

To calculate seasonal index:

Quarter	1971	1972	1973	Qly Total	Qly Av	Seasonal index	
I		104.2	106.04	210.25	105.125	$\frac{105.125}{99.8} \times 100$	105.30
II		92.43	97.67	120.04	95.05	$\frac{95.05}{99.8} \times 100$	95.25
III	96.63	104.97		201.60	100.80	$\frac{100.80}{99.8} \times 100$	100.97
IV	101.20	95.90		196.70	98.35	$\frac{98.35}{99.8} \times 100$	98.52
					399.325		

$$\text{General Average} = \frac{399.325}{4} = 99.831 = 99.8$$

Assuming the multiplicative model, we calculated the seasonal variations. We can also use the additive model and calculate the seasonal variations. Here, we have to subtract the moving average from the given data and then find the indices using the method of simple averages

Exercise

Find the seasonal variations using the two methods

Years	Quarters			
	I	II	III	IV
1974	60	80	72	68
1975	68	114	100	88
1976	80	116	108	96
1977	108	152	136	124
1978	160	184	172	164

(3) Cyclical Variations

Cyclical variations are also regular like seasonal variations. But they are different from the seasonal variations in the following aspects.

- (i) The period of seasonal variations is less than one year. But the period of cyclical variations is more than one year. Usually, the period ranges between five years to ten years.
- (ii) Seasonal variations recur at regular intervals of time whereas cyclical variations recur but not at regular intervals.

Also, cyclical variations are different from secular trend in the following aspects:

- (i) Secular trend denotes movements over a long period. But cyclical fluctuations represent movements over a period ranging between five and ten years.
- (ii) Secular trends tell us continuous movements in the same direction. But cyclical fluctuations do not give movements in the same direction instead, they give the alterations of uptrend and downtrend in economic activity.

These cyclical variations are generally associated with the trade cycle. A trade cycle consists of four phases. (i) prosperity (ii) recession (iii) depression and (iv) recovery. The period when the movement of business activities is upward and in the peak is known as 'prosperity' or 'boom'. This is followed by a period of contraction ending in a 'recession' or 'depression'. Following that recovery or expansion ensures and eventually develops into a period of prosperity. These trade cycles are generally found to be irregular and uneven in duration.

We use the method of moving averages to calculate this component.

4) Irregular Fluctuations

Irregular or random or erratic variations refer to those variations in business activity which are generally non-recurring and hence completely unpredictable. Irregular movements are generally fire strike and other unseen and natural catostrophies.

Year	1974	1975	1976	1977	1978
60	80	114	118	108	160
55.25	68	108	108	108	175
169.00	88	108	108	108	184
68	88	108	108	108	184

(3) Cyclical Variations

Cyclical variations are also regular like seasonal variations. But they are different from the seasonal variations in the following aspects.

- The period of seasonal variations is less than one year. But the period of cyclical variations is more than one year. Usually, the period ranges between five years to ten years.
- Seasonal variations recur at regular intervals of time whereas cyclical variations recur but not at regular intervals.
- Also, cyclical variations are different from secular trend in the following aspects:
 - Secular trend denotes movements over a long period. But cyclical fluctuations represent movements over a period ranging between five and ten years.
 - Secular trends tell us continuous movements in the same direction. But cyclical fluctuations do not give movements in the same direction instead they give the alterations of upward and downward in economic activity.