Unit 1:-

Diccriptive statistics and Methods for data Science

data science:-

* data science is the study of data it involves developing methods of collecting data.
* Computer science involves creating programmes and algorithm's while data science covers any type of data which may or may not use computer.
* Data science includes the collection, organisation, analysis and presentation of data.
* Company may use data science to develop effective ways to store, manage and analyse the data.
* Data science is a data application which acquire it values from the data itself and create more data as result.
* Data science is not only a application, it is a data product, so, data science enables the creation of product.

Eg:-

for example:-

google is master creating data product through the data what google having it creates effective ways for google search through the data sciences.

Eg:- google page rank algorithm, spell checking, voice regonition

Statistics:-

Introduction:-

* Statistics is a form of mathematical analysis dealing with the collection, analysis, interpretation and presentation of data numerically.
* It is a collection of quantitative data.
* It is used to estimate the population parameter

for example:-

i) Average no of students taking admission into college every year.

ii) the details of a people suffering from dengue in the country.

iii) the average no of people affected by the road accidents in the state everyday

iv) the central budget for the year 2020 and the distribution of rupees for various products.

→ Statistics are two types:

i) Discriptive statistics.

ii) Inferential statistics

i) Discriptive statistics: It is summarized the data from the sample using measures such as

* Mean
* Median
* standard deviation.

ii) Inferential statistics:- It is obtains conclusions from the data that are subject to the random variation.

* The measures of statistical data are of

four types

* Measures of frequency (count, percent etc)
* Measures of central tendency (Mean, mode, median)
* Measures of variation (arrange, variance, standard deviation)
* Measures of position (Ranks, Percentiles)

Population:-

A population includes all the elements from a set of given data.

Sample:- A sample consists 1 or more observations obtained from the population.

* depending on the sample method a sample can have fewer observation than the population the same no of observation or more observation
* Sample set is the subset of the population set
* A measurable characteristics of a population such as mean (or) standard deviation is called parameter but a measurable characteristics of sample is called statistics.
* The size of the sample is always less than the size of population.
* Generally, population parameter is denoted by P (for large population) (or) p (for small population) and sample parameter is denoted by 's' (for large sample) or 's' (for small sample)
* In a sampling process we may choose two different samples from the same population (or) two different samples from the two different

populations.

Eg:- In a computer boards manufacturing company, the company manufacture thousand boards per a day. A distributer choosed 50 boards for testing.

hence, the population $p = 1000$

$S = 50$.

***

Collection of data:-

A systematic plan and meaningful way of collecting information is known as collection of data.

* the methods of collection of data dependson various aspects such as
  * Objective
  * Scope and
  * Nature of problems.
* the data may be collected from two main sources
  a) primary data
  b) Secondary data

a) Primary data:- A statistical data 1st time which are collected originally in the nature is called primary data

* Primary data are collected directly by the authority who are require to collect them
* The source from which the primary data is collected is called primary source.
* The primary data is collected by workers, investigators and enumerators

for example:- In India the sources of primary data are the census of India

published by the government.

* the primary data may be collected by any one of the following methods

i) Direct personal interview
ii) In direct personal interview
iii) Mail questioned method
iv) Information from local agents.

b) Secondary data :-

the data collected by some earlier agency but it used and analysed by any others is called Secondary data.

* There are several sources of secondary data which may use to analyse the data.

* the secondary data can be categorized into two sources.

* published sources
* Unpublished sources

* Published Sources :-

Some of published sources are government publications, international publications, and unofficial publication, private publication.

* Unpublished Sources :-

the information take up from the sources like letters, diaries, autobiographics.. etc, are called unpublished sources.

Classification of data:

Arranging the collected data into different groups or classes on the bases of their similarities is called classification of data.

* The classification of data separating the characters having similarities and dissimilarities.

* A group or class has to be determine as the bases of the nature and the purpose on which it is to be used.

* It provides a bases for tabulation and analysis of the data.

* It helps in presenting the original data in a simple form.

There are two methods of classification

i) Classification according to attributes

ii) Classification according to variables.

i) Classification according to attributes:-

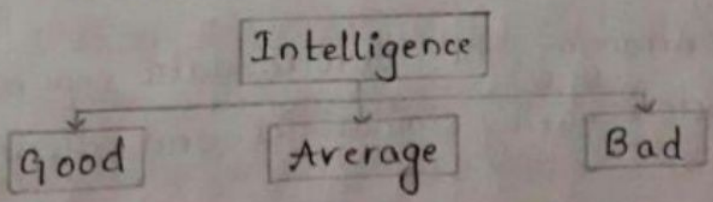
* It is a qualitative characteristics. it can be measured quantitatively as presence (or) absence of the characteristics.

Ex:- Intelligence, beauty so....on

There are two types of classification of attributes.

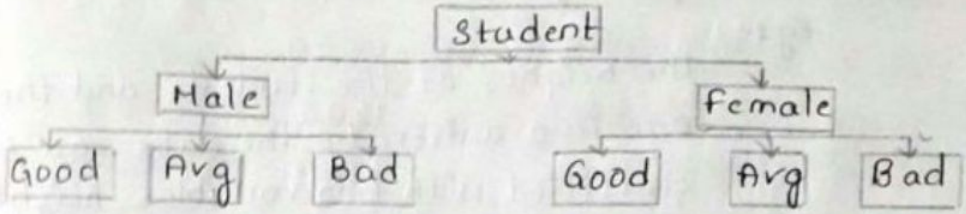
* Simple classification:

Here the data is classified only one attributes.



Manifold classification:-

* Here the data is classified more than one attributes.



ii) Classification of data according to variable

* It is a 45 above data which is measured numerically. Here, the data is shown in the form of frequency distribution

for example:-

height, weight, mass etc.

Based on the no of variables there are 3 types of frequency distribution

i) Unvarient frequency distribution: the frequency distribution is formed with one variable is called the frequency distribution.

Eg:-

the marks of 30 students are

52 72 58 55 75 53
54 64 70 74 56 51
50 80 85 92 68 96
76 95 65 75 98 72
88 77 99 62 73 79

Marks class interval	Tally Marks	No of students frequency `f'
50 - 60	𝍸 III	8
60 - 70	IIII	4
70 - 80	𝍸 𝍸	10
80 - 90	IIII	4
90 - 100	𝍸	5
		$\Sigma f = 30$

ii) Bivarient frequency distribution: the frequency distribution is formed with two variable is called bivarient frequency distribution.

Eg:-

the heights of the students and their age in a university the data can be classified with two variables height and age.

Height (ft)	Age			
	15-20	20-25	25-30	30-35
4-5				
5-6				
6-7				

iii) Multi varient frequency distribution:

The frequency distribution formed by mo than two is called multivarient frequency distribution.

Example: the student data in a university in a three variables as age, height and gender

Height (ft)	Age							
	15-20		20-25		25-30		30-35	
	M	f	M	f	M	f	M	f
4-5								
5-6								
6-7								

Fabulation of data: the fabulation of data may be defined as a logical and systematic organisation of statistical data in rows and columns.

The main objective of fabulation of data are:

* Systematic presentation of data
* to compare the data.
* identification of required values.
* For detective of errors.
* Economy of space for reference.

In classification the data is divided on the bases of similarity where as tabulation is the process of classified facts in closed columns.

* Classification provides a bases for tabular presentation the statistical tables may be classified in the following ways.
  * Simple and complex tables
  * General purpose or referential table.
  * Special purpose or summary table

* Simple and complex tables:-

Simple table:

* In a simple table, the data are classified w.r.t to single characteristic.
* It is also called as one-way table.

Example:

the data of students studied in different years of a college is given below.

Academic years	No of students
1998 - 1999	1900
1999 - 2000	1500
2001 - 2002	1600
2002 - 2003	1700
2003 - 2004	1800
2004 - 2005	1750

Complex table:

The complex table is the data (or) grouped into different classes w.r.t two (or) more characteristics simultaneously.

* If the data are classified with two characteri that table is called two-way table.
* If it is expressed with three characteristics then it is called three way table.

Example:

The number of students in college according to sex and martial status during 1985 - 1986 to 1988 - 1999

academic year	Male		Female	
	Martial	unmartial	Martial	unmartia
1985 - 86	42	30	28	32
1986 - 87	26	40	33	29
1987 - 88	35	28	25	31
1988 - 89	40	36	31	24

ii) General purpose (or) reference table:

* This type of table are prepared to store information and they contain void range of information relating to a specified subjected.

This tables are prepared in a systematic manner.

Example: The table appended to the censes reports, and RBI bulletens etc.

Special purpose (or) summary tables:

* this tables are constructed in a specific points and are very useful for the purpose of comparision.
* This tables are also called text tables
* Generally this table indicates rates, percentage averages etc.

Example:

The number of sales in two company of same product displayed in the summary table.

Year	Sales in company in A (lakhs in Rs)	Sales in company in B (lakhs in Rs)
2008	78	62
2009	90	86
2010	84	95
2011	76	81
2012	85	90

*** Measures of central tendency:-

Measures of central tendency gives an idea about the concentration of the values in the central path of the distribution. A measure of central tendency is a statistical average or single

value which represents entire distribution.

* The following are important measures of central tendency.
  * Arthimetic mean
  * Median
  * Mode
  * geometric mean
  * Harmonic mean

* Some of the characteristics given below to be satisfied by an ideal measure of central tendency
  * It should be based on all observation.
  * It should be easy to calculate.
  * It shouldn't be easy to calculate and be effected by extreme observations.
  * It should be capable of further mathematical treatment.

* **Arithmetic mean:-**

Arithmetic mean of set of observations is the sum of all observation divided by number of observation. It is denoted by $\bar{x}$.

* Mean for ungrouped data:

If $x_1, x_2, x_3, \ldots\ldots, x_n$ are observations then the mean is defined as $\bar{x} = \frac{\sum x_i}{n} = \frac{x_1+x_2+x_3\ldots+x_n}{n}$

where n = no of observations.

* Mean for grouped data:

$f_1, f_2, f_3 \ldots f_n$ are the frequencies of the variable $x_1, x_2, x_3 \ldots x_n$ then the mean $\bar{x} = \frac{\sum f_i x_i}{N}$

where, N = sum of all frequencies

Weighted Mean:

If $w_1, w_2, \ldots w_n$ are the weights of the observations $x_1, x_2, x_3 \ldots x_n$ then the mean $\bar{x}$ is defined as

$$\text{weighted mean } \bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

Merits and demerits of mean:

Merits:

* It is easy to calculate.
* It is based on observation
* It is accurate and reliable.
* It is suitable for further mathematical treatment.

Demirts:

* It cannot be dealed with qualitative factors. Ex: intelligence, beauty etc
* It is effected by eutreme values.
* It cannot be calculated for open-end classes.

Problems:-

Q) The weights of the 6 competetors in a game are given below: 58kg, 62kg, 56kg, 63kg, 55kg, 61kg find Arithmetic mean of weight of the competetors.

(A)

Arithmetic mean $\bar{x} = \frac{58+62+56+63+55+61}{6}$

$$\bar{x} = \frac{355}{6}$$

$$\bar{x} = 59.16$$

Q) Find the arithmetic mean of the following frequency distribution.

x	1	2	3	4	5	6	7
f	5	9	12	17	14	10	6

(A)

Arithmetic mean $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$

x_i	f_i	$f_i x_i$
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	$\Sigma f_i = 73$	$\Sigma f_i x_i = 299$

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\bar{x} = \frac{299}{73}$$

$$\bar{x} = 4.095$$

Note: Deviation method

If for the class intervals, frequencies are given, we can find A.M by deviation method.

Here, AM $\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i} \times h$

f_i = frequency

$d_i = \frac{x_i - A}{h}$

A - assumed mean.

Problems:-

7) Calculate the AM of the marks from the table.

Marks	0-10	10-20	20-30	30-40	40-50	50-6
No of students	12	18	27	20	17	6

8)

Interval	f_i	x_i (mid value)	$f_i x_i$
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
	$\Sigma f_i = 100$		$\Sigma f_i x_i = 2800$

Arithmetic mean

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

$$\bar{x} = \frac{2800}{100}$$

$$\bar{x} = 28,$$

9) Calculate the mean for the following frequency distribution by deviation method.

class interval	0-8	8-16	16-24	24-32	32-40	40-48
frequency	8	7	16	24	15	7

C.I	x_i (mid values)	f_i	$\frac{x_i - A}{h} = \frac{x_i - 28}{8}$	$f_i d_i$
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16
24-32	28 (A)	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
		$\Sigma f_i = 77$		$\Sigma f_i d_i = -25$

$$\text{AM } \bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i} \times h$$

$$= 28 + \frac{(-25)}{77} \times 8$$

$$= 25.402$$

Median :-

Median of a distribution is a value of the variable which divides it into 2 equal parts. Median is the middle value of the distribution.

* For the ungrouped data, if the number of observations is odd then the median is the middle value after the values have been arranged in ascending (or) descending order.

* If the number of observation is even then the median of the distribution is arthimetic mean of the middle two terms.

* For a discrete frequency distribution the cummulative frequency just greater than $N/2$, the corresponding value of x is median.

* for a continuous frequency distribution the median is obtained by the following formula.

$$\text{Median} = l + \frac{\left(\frac{N}{2} - c\right)}{f} \times h$$

where,

l = Lower limit of median class

f = frequency of median class

$N = \sum f_i$ (sum of frequency)

h = width of the class interval.

c = cumulative frequency preceding to the median.

Problems:-

Q) Find median of the values 25, 30, 35, 15, 10.

A) Ascending order : 10, 15, 25, 30, 35

Median = middle term

= 25

Q) Find the median for the following data

x	5	8	11	14	17	20	23
f	2	8	12	20	10	6	3

(A)

x	f	Cummulative frequency
5	2	2
8	8	10
11	12	22
14	20	42
17	10	52
20	6	58
23	3	61
	$\sum f_i$ = 61	

$\frac{N}{2} = \frac{61}{2} = 30.5$

The next greater cummulative frequency of $\frac{N}{2} = 30.5$ is 42

$\therefore$ Median = 19//

8) Find the median to the following data

C.I	40-50	50-60	60-70	70-80	80-90
frequency	5	12	23	8	2

9)

C.I	f	Cummulative frequency
40-50	5	5
50-60	12	17(c)
(60)-70	(23)f	40
70-80	8	48
80-90	2	50
	$\Sigma f = 50$	

$\frac{N}{2} = 25$

Greater cummulative frequency = 40

$l = 60$

$f = 23$

$c = 17$

$h = 10$

$$\text{Median} = l + \frac{\left(\frac{N}{2} - c\right)}{f} \times h$$

$$= 60 + \frac{(25-17)}{23} \times 10$$

$$= 63.478$$

6) Find the median of the following data

C.I	20-30	30-40	40-50	50-60	60-70
f	3	5	20	10	5

(A)

c I	f	c.f
20-30	3	3
30-40	5	8 (c)
(L)40-50	20(f)	28
50-60	10	38
60-70	5	33
	$\Sigma f_i = 43$	

Greater cummulative frequency c = 8

$l = 40$

$f = 20$

$h = 10$, $N/2 = \frac{43}{2} = 21.5$

$$\text{Median} = 40 + \frac{(21.5 - 8)}{20} \times 10$$

$$= 40 + \frac{13.5}{20_2} \times 10$$

$$= 40 + 6.75$$

Median = 46.75 //

Mode :- Mode is defined as the value which occurs most frequently in a set of observation.

* For an ungrouped data:

* the mode is the value which repeats maximum number of times

* for a descrete frequency distribution:

first find the maximum frequency and its corresponding 'x' value is the mode.

* for a continuous frequency distribution

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] \times h$$

where,

l = lower limit of a model class

f_1 = frequency of model class

f_0 = frequency preceding the model class

f_2 = frequency succeeding model class

h = width of C.I

Problems:

Q) Obtain the mode of the value: 10, 12, 15, 20, 12, 16, 18, 15, 12, 10, 16, 20, 12, 24.

A) Here 12 repeated more times

i.e., 4 times

∴ Mode = 12

Q) Find the mode of the following distribution.

C.I	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	4	3

(A)

The maximum frequency is 25

Its corresponding value is '4'

∴ mode = 4

Q) find mode of the corresponding data

C.I	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-
f	5	7	10	18	20	12	8	2

A)

C.I	f
0-5	5
5-10	7
10-15	10
15-20	18 (f_0)
(l) 20-25	20 (f_1)
25-30	12 (f_2)
30-35	8
35-40	2

$$\text{Mode} = l + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h$$

$$= 20 + \left(\frac{(20-18)}{40-18-12}\right) \times 5$$

$$= 20 + \left[\frac{2}{\frac{10}{8}}\right] \times 5$$

$$= 20 + 1$$

$$\text{Mode} = 20$$

Geometric Mean (G.M) :-

Geometric mean of a set of "n" observations is the n^{th} root of their product. Thus the geometric mean (G.M) of "n" observations $x_1, x_2, \ldots, x_n$ is

$$G.M = \sqrt[n]{x_1 x_2 x_3 \cdots x_n}$$

$$= (x_1 x_2 x_3 \cdots x_n)^{1/n}$$

Calculation of Geometric Mean (G.M) :-

(i) For an ungrouped data

$$G.M = \text{Antilog}\left[\frac{1}{n}\sum \log x_i\right]$$

where "n" is number of values

(ii) For an grouped data (Discrete data) :-

for the grouped data the geometric mean is

$$G.M = \text{Antilog}\left[\frac{1}{N}\sum f_i \log x_i\right]$$

where $N = \sum f_i$

(iii) For the continuous data :-

for the continuous data Geometric mean is

$$G.M = \text{Antilog}\left[\frac{1}{N}\sum f_i \log x_i\right]$$

where $N = \sum f_i$ and x_i is the mid value of the interval

Problems :-

① Find the Geometric mean of monthly income of ten families of a particular place is given below

85, 70, 15, 75, 500, 8, 45, 250, 40, 36

Sol:- Given ungrouped data

85, 70, 15, 75, 500, 8, 45, 250, 40, 36

we know that the Geometric mean of the ungrouped data is

$$G.M = \text{Antilog}\left[\frac{1}{n}\sum \log x_i\right]$$

where "n" is no. of observations.

Here $n = 10$

x	$\log x_i$
70	1.8451
15	1.1761
75	1.8751
500	2.6990
8	0.9031
45	1.6532
250	2.3979
40	1.6021
36	1.5563
	$\Sigma \log x_i = 17.6373$

Geometric Mean

$$G.M = \text{Antilog}\left[\frac{1}{n}\Sigma \log x_i\right]$$

$$= \text{Antilog}\left[\frac{1}{10}\,17.6373\right]$$

$$= \text{Antilog}[1.7637]$$

$$= 10^{1.7637}$$

$$G.M = 58.0363$$

H.W

(2) Find the Geometric mean of the following data

15, 25, 35, 45, 55

(3) Find the geometric mean for the data given below.

x	10	15	25	40	50
f	4	6	10	7	3

Sol :- Given grouped data (Discrete data) is

x	10	15	25	40	50
f	4	6	10	7	3

The geometric mean for the grouped data is

$$G.M = \text{Antilog}\left[\frac{1}{N}\Sigma f_i \log x_i\right], \text{ where } N = \Sigma f_i$$

x	Frequency f	$\log x$	$f_i \log x$
10	4	1	4
15	6	1.1761	7.0566
25	10	1.3979	13.9790
40	7	1.6021	11.2147
50	3	1.6990	5.0970
	$N = \Sigma f_i = 30$		$\Sigma f_i \log x = 41.3473$

The Geometric mean of the grouped data is

$$G.M = \text{Antilog}\left[\frac{1}{N}\Sigma f_i \log x\right]$$

$$= \text{Antilog}\left[\frac{1}{30}(41.3473)\right]$$

$$= \text{Antilog}(1.3782)$$

$$G.M = 10^{1.3782} = 23.8891$$

H.W

(4) Find the geometric mean of the following data

x	8	25	17	30
f	5	3	4	2

(5) Compute the geometric mean from the following data

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	5	7	15	25	8

Sol:- Given continuous data

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	5	7	15	25	8

We know that the geometric mean for the continuous grouped data is

$$G.M = \text{Antilog}\left[\frac{1}{N}\Sigma f_i \log x\right]$$

where $N = \Sigma f_i$

and x is Mid value of the interval

Marks	No. of students f	Mid value x	$\log x$	$f_i \log x$
0 – 10	5	5	0.6990	3.495
10 – 20	7	15	1.1761	8.2327
20 – 30	15	25	1.3979	20.9685
30 – 40	25	35	1.5441	38.6025
40 – 50	8	45	1.6532	13.2256
	$N = \Sigma f_i = 60$			$\Sigma f_i \log x = 84.5243$

The geometric mean

$$G.M = \text{Antilog}\left[\frac{1}{N}\Sigma f_i \log x\right]$$

$= \text{Antilog}\left[\frac{1}{60}(84.5243)\right]$

$= \text{Antilog}[1.4087]$ $\quad [\because \text{Antilog}\, x = 10^x]$

$= 10^{1.4087}$

H.W $\quad G.M = 25.6271$

⑥ Find Geometric mean for the following data

C.I	10-20	20-30	30-40	40-50	50-60
f	4	6	10	7	3

Harmonic Mean (HM) :-

Harmonic Mean (HM) is the reciprocal of the arithmetic mean of the reciprocals of the given values. and it is denoted by HM.

$$\text{i.e } H.M = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$$

calculation of Harmonic Mean :-

(i) For an ungroped data :-

For the ungrouped data Harmonic Mean

$$HM = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$$ where "n" is no. of observations

(ii) For the grouped data (discrete data) :-

For the grouped data Harmonic mean is

$$H.M = \frac{1}{\frac{1}{N}\sum \frac{f_i}{x_i}} = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_i}}$$

where $N = \sum f_i$

(iii) For the continuous data :-

For the continuous data the Harmonic mean is

$$HM = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_i}}$$

where $N = \sum f_i$ and x_i is the Mid value of the interval

Problems :-

① Find the harmonic mean for the following data

6, 15, 35, 40, 900, 520, 300, 400, 1800, 2000

Sol :- Given ungrouped data values

6, 15, 35, 40, 900, 520, 300, 400, 1800, 2000

The Harmonic mean for the ungrouped data is

$$HM = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$$ where "n" is no. of observations

Here $n = 10$

x_i	$\frac{1}{x_i}$
6	0.1667
15	0.0667
35	0.0286
40	0.0250
900	0.0011
520	0.0019
300	0.0033
400	0.0025
1800	0.0006
200	0.0005
	$\sum \frac{1}{x_i} = 0.2969$

The harmonic mean

$$HM = \frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$$

$$= \frac{1}{\frac{1}{10} 0.2969}$$

$$= \frac{1}{0.0297}$$

$$HM = 33.6700$$

HW

② Find the harmonic mean for the following values

10, 20, 25, 40, 50

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

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

Sol) :- Given grouped data

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

The harmonic mean of the grouped data is

$$HM = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_i}} \quad \text{where } N = \sum f_i$$

Marks x_i	No. of students f_i	$\frac{1}{x_i}$	$f_i \frac{1}{x_i}$
10	20	0.10	2.000
20	30	0.05	1.5000
25	50	0.04	2.0000
40	15	0.025	0.375
50	5	0.02	0.1000
	$N = \sum f_p = 120$		$\sum f_i \frac{1}{x_p} = 5.975$

The harmonic mean is

$$HM = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_p}}$$

$$= \frac{1}{\frac{1}{120}(5.975)}$$

H.W

$$HM = \frac{1}{0.0498} = 20.0803$$

(4) Find Harmonic Mean for the following data.

x	10	15	25	40	50
f	4	6	10	7	3

(5) Find Harmonic mean for the following data

C.I	10-20	20-30	30-40	40-50	50-60
f	4	6	10	7	3

Sol :- Given continuous data

C.I	10-20	20-30	30-40	40-50	50-60
f	4	6	10	7	3

we know that the Harmonic Mean for the continuous data is

$$H.M = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_i}} \quad \text{where } N = \sum f_i .$$

Here x_i is the mid value of the interval

class Interval	frequency f_i	Mid value x_i	$\frac{1}{x_i}$	$f_i \frac{1}{x_i}$
10-20	4	15	0.0667	0.2668
20-30	6	25	0.04	0.24
30-40	10	35	0.0286	0.286
40-50	7	45	0.0222	0.1554
50-60	3	55	0.0182	0.0546
	$N = \sum f_i = 30$			$\sum f_i \frac{1}{x_i} = 1.0028$

The Harmonic Mean

$$HM = \frac{1}{\frac{1}{N}\sum f_i \frac{1}{x_i}}$$

$$= \frac{1}{\frac{1}{30}(1.0028)}$$

$$= \frac{1}{0.0334}$$

$$HM = 29.9401$$

H.W

⑥ Find the HM for the following data

C.I	0-10	10-20	20-30	30-40	40-50
f	5	7	15	25	8

Mean Deviation (M.D) :- The mean deviation is an average of deviations from an average of given data. Thus it is also known as "Average deviation".

calculation of mean deviation :-

(i) For the ungrouped data :-

For the ungrouped data mean deviation is

$$M.D = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

where $A = \bar{x} = \frac{\sum x_i}{n}$ = mean and 'n' is no. of observations

(ii) For the grouped data :-

For the grouped data the mean deviation is given by

$$M.D = \frac{\sum_{i=1}^{n} |x_i - A| f_i}{N}$$

where $N = \Sigma f_i$

and $A = \bar{x} = \frac{\Sigma x_i f_i}{N}$

Note :-

$$\text{coefficient of } M.D = \frac{\text{Mean deviation}}{\text{Average of which it is calculated}}$$

Problems :-

(1) Calculate the mean deviation for the following data

15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12

Sol :- Given ungrouped data

15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12

we know that the mean deviation for the ungrouped data is

$$M.D = \frac{\Sigma |x_i - \bar{x}|}{n}$$

Here No. of observations $n = 10$

x_i	$x_i - \bar{x} = x_i - 15$	$\lvert x - \bar{x} \rvert$
15	0	0
20	5	5
17	2	2
19	4	4
21	6	6
13	-2	2
12	-3	3
10	-5	5
17	2	2
9	-6	6
12	-3	3
$\Sigma x_i = 165$		$\Sigma \lvert x_i - \bar{x} \rvert = 38$

$$\text{Mean} = A = \bar{x} = \frac{\Sigma x_i}{n} = \frac{165}{11}$$

$$= 15$$

Mean Deviation

$$M.D = \frac{\Sigma |x_i - \bar{x}|}{n}$$

$$= \frac{38}{11}$$

$$= 3.4545$$

H.W

(2) calculate the mean deviation and coefficient of MD for the following data

70, 65, 68, 70, 75, 73, 80, 70, 83, 86

Ans :- M.D = 5.6 coefficient of MD = 0.0757

(3) Find the M.D and coefficient of M.D for the following data

x	2	4	7	8	10	12	15	16
f	2	4	10	20	32	18	10	4

Sol :- Given grouped data

x	2	4	7	8	10	12	15	16
f	2	4	10	20	32	18	10	4

we know that the mean deviation for the grouped data is

$$M.D = \frac{\Sigma |x_i - \bar{x}| f_i}{N} \quad \text{where} \quad \bar{x} = \frac{\Sigma x_i f_i}{N}$$

$$N = \Sigma f_i$$

Here $n = 8$

$N = \Sigma f_i = 2+4+10+20+32+18+10+4 = 100$

x	f	xf	$x_i - \bar{x} = x_i - 10$	$d_i = \lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert f_i$
2	2	4	−8	8	16
4	4	16	−6	6	24
7	10	70	−3	3	30
8	20	160	−2	2	40
10	32	320	0	0	0
12	18	216	2	2	36
15	10	150	5	5	50
16	4	64	6	6	24
	$\Sigma f = 100$	$\Sigma fx = 1000$			$\Sigma \lvert x_i - \bar{x} \rvert f_i = 220$

Mean $\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f} = \frac{1000}{100} = 10$

Thus mean deviation is

$$MD = \frac{\Sigma |x_i - \bar{x}| f_i}{N}$$

$= \frac{220}{100}$

M.D $= 2.2$

coefficient of mean deviation $= \frac{\text{Mean deviation}}{\text{mean}}$

$= \frac{2.2}{10} = 0.22$

Standard deviation:-

The standard deviation of a distribution is denoted by σ and is defined as

i) for an ungrouped data:

Variance = $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$ or Standard deviation = $\sigma = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$

ii) for a grouped data:

$$\sigma^2 = \frac{1}{N}\sum f_i (x_i - \bar{x})^2$$

(or)

$$\sigma = \sqrt{\frac{1}{N}\sum f_i (x_i - \bar{x})^2}$$

Here, $\bar{x}$ - the arithmetic mean

a) Compute the standard deviation of the values 10, 6, 8, 12, 20.

(A)

$$\bar{x} = \frac{10+6+8+12+20}{5} = \frac{56}{5} = 11.2$$

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{5}\left[(10-11.2)^2 + (6-11.2)^2 + (8-11.2)^2 + (12-11.2)^2 + (20-11.2)^2\right]$$

$$\sigma^2 = \frac{1}{5}\left[1.44 + 27.04 + 10.24 + 0.64 + 77.44\right]$$

$$\sigma^2 = \frac{1}{5}\left[116.8\right]$$

$$\sigma^2 = 23.36$$

$$\sigma = \sqrt{23.36}$$

$$\sigma = 4.833 //$$

a) Compute standard deviation of the following data

C.I	0-10	10-20	20-30	30-40	40-50	50-60
f	2	8	16	28	12	4

(A)

C.I	f_i	x_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
0-10	2	5	10	-27.4	750.76	1501.52
10-20	8	15	120	-17.4	302.76	2422.08
20-30	16	25	400	-7.4	54.76	876.16
30-40	28	35	980	2.6	6.76	189.28
40-50	12	45	540	12.6	158.76	1905.12
50-60	4	55	220	22.6	510.76	2043.04
			$\Sigma f_i x_i = 2270$			$\Sigma f_i(x_i - \bar{x})^2 = 8937.2$

Mean $\bar{x} = \frac{1}{N} \Sigma f_i x_i$

$= \frac{1}{70}(2270)$

$\bar{x} = 32.4$

Variance= $\sigma^2 = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^2$

$\sigma^2 = \frac{1}{70}(8937.2)$

$\sigma^2 = 127.67$

Standard Deviation= $\sigma = \sqrt{127.67}$

$\sigma = 11.299$

Skewness :-

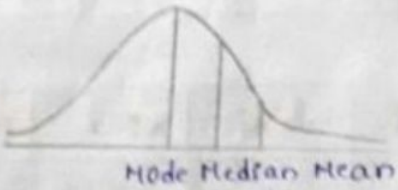
A distribution is said to be skewed if the mean, median, mode are fall at different points.

i.e., mean ≠ median ≠ mode.

* Skewness means lack of symmetry. It gives an idea about the shape of the curve of the given data.

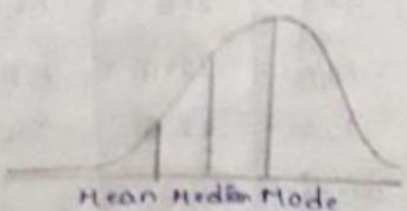
The skewed curve is not symmetrical but stretched more to oneside than to the other.

i) Stretched right



Mode < Median < Mean

ii) Stretched left



Mean < Median < mode.

Generally the skewness is denoted by 'S_K' and is defined as skewness

S_K = Mean - Median

(or)

S_K = Mean - Mode

Karl pearson's coefficient of skewness:

The karl pearson's coefficient of skewness is defined as

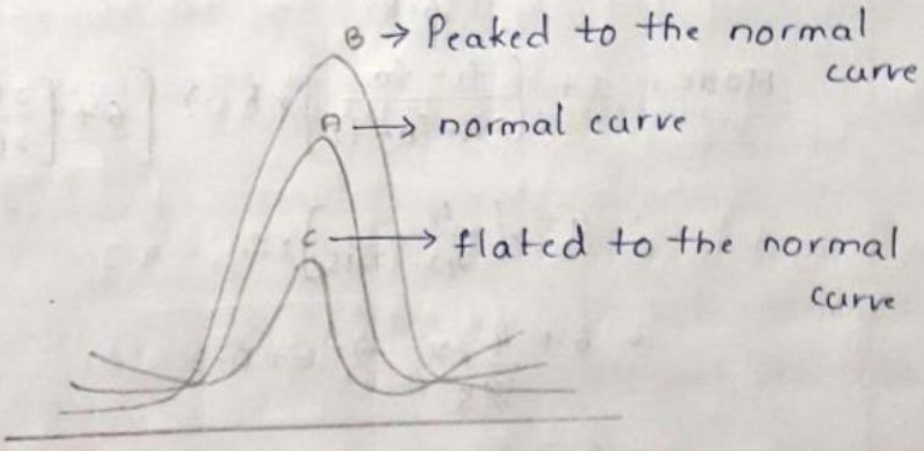
$$\therefore Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Kurtosis:—

* The Measures of central tendency and skewness maynot give a complete idea about distribution.

* There is an another statistical measures namely kurtosis which gives an idea about flatness or peakness of the curve.

* If two or more distribution have average and skewness the another measures which is used to compare the distributions is called kurtosis.



Problem:

Q) Find Karl pearson's coefficient of skewness to the following data.

C.I	0-2	2-4	4-6	6-8	8-10	10-12	12-14
f	6	8	17	21	15	11	2

C.I	f_i	x_i	$f_i x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
0-2	6	1	6	-5.8	33.64	201.84
2-4	8	3	24	-3.8	14.44	43.32
4-6	17 (f_0)	5	85	-1.8	3.24	55.08
(26-8	21 (f_1)	7	147	0.2	0.04	0.84
8-10	15 (f_2)	9	135	2.2	4.84	72.6
10-12	11	11	121	4.2	17.64	194.04
12-14	2	13	26	6.2	38.44	76.88
	$\sum f_i$ = 80		$\sum f_i u_i$ = 544			$\sum f_i(x_i - \bar{x})^2$ = 644.6

$$\text{Mean} = \frac{1}{N} \sum f_i x_i$$

$$= \frac{544}{80}$$

$$= 6.8$$

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] \times h \Rightarrow \left[6 + \left[\frac{21 - 17}{2(21) - 17 - 15}\right] \times 2\right]$$

$$= 6 + \left[\frac{4}{42 - 17 - 15}\right] \times 2$$

$$= 6 + \frac{4}{10} \times 5 \Rightarrow 6 + 0.8$$

$$= 6.8$$

Variance= $\sigma^2 = \frac{1}{N} \sum f_i (x_i - \bar{x})^2$

$$\sigma^2 = \frac{1}{8} \times 644.6$$

$$\sigma^2 = 80.575$$

S.D $\sigma = \sqrt{80.575}$ (Since Standard Deviation= $\sqrt{\text{Variance}}$)

$$\sigma = 8.976$$

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.D}}$$

$$S_k = \frac{6.8 - 6.8}{8.976} = 0$$

***

Correlation :- The relation between two random variables is called correlation. There are two types of correlations

(i) positive correlation.
(ii) Negative correlation.
(iii) Linear and non-linear correlation.

(i) **Positive correlation :-** If the values of the two variables deviate in the same direction.

i.e If the increase in the values of one variable results, on an average, in a corresponding increase in the values of the other variable (or) If the decrease in the values of one variable results on an average in a corresponding decrease in the value of the other variable then the correlation is called positive correlation.

Ex :- 1. Heights and weights of the individuals.
2. The family income and expenditure on luxury items
3. Amount of rainfall and yield of the crop.

(ii) **Negative correlation :-**

If the increase (decrease) in one variable results on an average in a corresponding decrease (increase) in the values of the other variable then the correlation is called negative correlation.

Ex :-
1. price and demand of a commodity.
2. sale of woollen garments and the day temparature
3. volume and pressure of a perfect gas.

(iii) **Linear and non-linear correlation :-**

→ The correlation between two variables is said to be linear. If corresponding to a unit change in one variable, there is a constant change in other variable over the entire range of the values. (or) Two variables x and y

are said to be linearly related if there exists a relationship of the form $y = a + bx$

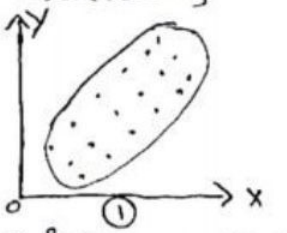
→ The correlation between two variables is said to be non-linear (or) curve linear if corresponding to a unit change in one variable, the other variable doesn't change at a constant rate but at fluctuating rate
(or) Two variables x and y are said to be non linearly (or) curve linearly related if there exists relationship of the forms $y = a + bx + cx^2$, $y = ae^{bx}$

Methods of studying correlation:-

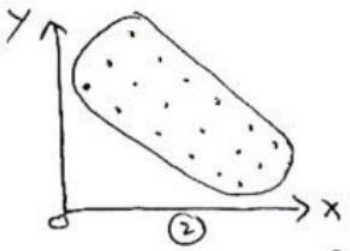
1. Scatter diagram method.
2. Karl pearsons correlation coefficient method.
3. Rank method.

Scatter diagram method:-

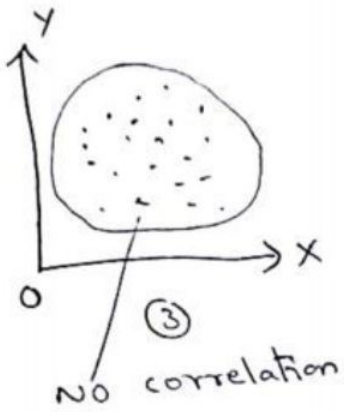
Scatter diagram method is the simplest way of diagrammatic representation of bi-variate data. In this method if "n" pair of values for two variables x and y are given as $(x_1, y_1), (x_2, y_2) \dots\dots (x_n, y_n)$ Then these points are plotted on the x-axis and y-axis in the xy-plane. The diagram of dots so obtained is known as scatter diagram. This method gives a fairly good, rough idea about the relationship (correlation) between two variables



Here x increases as well as y increases. so it is a positive correlation.



Here x increases and y decreases. so it is a negative correlation.



***

Karl pearson's correlation coefficient (or) coefficient of correlation :- Karl pearson suggested a mathematical method for measuring the magnitude of linear relationship between two variables. This is known as Karl pearson's correlation coefficient (or) product - moment correlation coefficient.

Karl pearsons correlation coefficient between two variables x and y is denoted by r (or) r_{xy} and is defined as

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where $cov(x,y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} = \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$

means $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$

standard deviations $\sigma_x = \sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}$

$\sigma_y = \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}$

Properties of correlation coefficient :-

1. The correlation coefficient r lies between -1 and 1 i.e $-1 \le r \le 1$
2. If there is no relation between any two variables then they are uncorrelated.
3. Two independent variables are uncorrelated. ie If x and y are independent variables then $r_{xy} = 0$

*** ① Calculate the correlation coefficient (r) for the following heights (inches) of fathers (x) and their sons (y)

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Sol :- By the def. of correlation coefficient, we have

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$

means $\bar{x} = \frac{\Sigma x_i}{n}$ $\bar{y} = \frac{\Sigma y_i}{n}$

S.D $\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\Sigma x_i = 544$	$\Sigma y_i = 552$	$\Sigma x_i y_i = 37560$	$\Sigma x_i^2 = 37028$	$\Sigma y_i^2 = 38132$

Here no. of entries (items) $n = 8$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{544}{8} = 68$ $\bar{y} = \frac{\Sigma y_i}{n} = \frac{552}{8} = 69$

covariance of x and y is $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\cdot\bar{y}$

$= \frac{1}{8}(37560) - 68(69)$

$= 4695 - 4692$

$= 3$

Standard deviation of x is

$\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$= \sqrt{\frac{1}{8}(37028) - (68)^2}$

$= \sqrt{4.5} = 2.1213$

Standard deviation of y is

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

$= \sqrt{\frac{1}{8}(38132) - (69)^2}$

$= \sqrt{5.5} = 2.3452$

The correlation coefficient is

$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$= \frac{3}{2.1213(2.3452)} = 0.6030$

② calculate the Karl pearson's correlation coefficient for the following data

Marks in statistics	30	60	30	66	72	24	18	12	42	06
Marks in accounts	06	36	12	48	30	06	24	36	30	12

Sol:- By def of Karl pearson's correlation coefficient we have

$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

where $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y}$

$\bar{x} = \frac{\Sigma x_i}{n}$ $\bar{y} = \frac{\Sigma y_i}{n}$

S.D $\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

Marks in statistics (x_i)	Marks in accounts (y_i)	$x_i y_i$	x_i^2	y_i^2
30	06	180	900	36
60	36	2160	3600	1296
30	12	360	900	144
66	48	3168	4356	2304
72	30	2160	5184	900
24	06	144	576	36
18	24	432	324	576
12	36	432	144	1296
42	30	1260	1764	900
06	12	72	36	144
$\Sigma x_i = 360$	$\Sigma y_i = 240$	$\Sigma x_i y_i = 10368$	$\Sigma x_i^2 = 17784$	$\Sigma y_i^2 = 7632$

Here number of entries $n = 10$

Mean $\bar{x} = \frac{\Sigma x_i}{n}$

$= \frac{360}{10} = 36$

$\bar{y} = \frac{\Sigma y_i}{n}$

$= \frac{240}{10} = 24$

Covariance $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\cdot\bar{y}$

$= \frac{1}{10}(10368) - 36(24)$

$= 172.8$

S.D of x is $\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$= \sqrt{\frac{1}{10}(17784) - (36)^2}$

$= \sqrt{482.4} = 21.9636$

S.D of y is $\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

$= \sqrt{\frac{1}{10}(7632) - (24)^2}$

$= \sqrt{187.2} = 13.6821$

The Karl pearsons correlation coefficient is

$$r = r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{172.8}{21.9636(13.6821)}$$

$r = 0.5750$ [Here clearly "r" is lies b/w -1 & 1]

(P3) A computer while calculating correlation coefficient between two variables x and y from 25 pairs of observations obtained the following results $n = 25$, $\Sigma X = 125$, $\Sigma X^2 = 650$, $\Sigma Y = 100$ $\Sigma Y^2 = 460$, $\Sigma XY = 508$. However, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values are

X	Y
8	12
6	8

Then obtain the correct value of correlation coefficient

Sol :-

The corrected value of $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

The corrected value of $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

" " $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

" " $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

" " $\Sigma XY = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8) = 520$

$n = 25$

Mean $\bar{X} = \frac{\Sigma X}{n} = \frac{125}{25} = 5$ $\quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{100}{25} = 4$

covariance $cov(xy) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y}$

$= \frac{1}{25}(520) - 5(4) = 0.8$

S.D of X $= \sigma_x = \sqrt{\frac{1}{n}\Sigma X^2 - \bar{X}^2} = \sqrt{\frac{1}{25}(650) - 5^2} = \sqrt{1} = 1$

S.D of Y $= \sigma_y = \sqrt{\frac{1}{n}\Sigma Y^2 - \bar{Y}^2} = \sqrt{\frac{1}{25}(436) - 4^2} = \sqrt{1.44} = 1.2$

The corrected correlation coefficient is

$$r = \frac{cov(xy)}{\sigma_x \sigma_y} = \frac{0.8}{1(1.2)} = 0.6667$$

(4) Given $n = 10$, $\sigma_x = 5.4$, $\sigma_y = 6.2$ and sum of product of deviation from the mean of x ce y is 66. Find the correlation coefficient.

Sol:- Given

$n = 10$

$\sigma_x = 5.4$

$\sigma_y = 6.2$

Also given sum of the product of deviation from the mean of x ce y is

$$\Sigma (x - \bar{x})(y - \bar{y}) = 66$$

we know that-

$$cov(x, y) = \frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{10}(66)$$

$$= 6.6$$

The karl pearson's correlation coefficient

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{6.6}{5.4\,(6.2)}$$

$$= \frac{6.6}{33.48} = 0.1971$$

***Rank correlation coefficient (or) spearman's rank correlation coefficient :- The correlation coefficient is obtained from the ranks of two individual characteristics (or) variables x and y is called rank correlation coefficient (or) spearman's rank correlation coefficient.

The rank correlation coefficient is denoted by ρ and is defined as

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

where d_i = difference of corresponding ranks of x and y

n = no. of terms in the series

properties :-

1. $-1 \leq \rho \leq 1$ i.e rank correlation is lies between -1 and 1.
2. If $\rho = 1$ then there is complete agreement in the order of the ranks and they are in same direction.
3. If $\rho = -1$ then there is complete disagreement in the order of the ranks and they are in opposite direction

problems when ranks are given :-

1. The following are the ranks obtained by 10 students in two subjects, statistics and maths. To what extent the knowledge of the students in two subjects is related

Statistics	1	2	3	4	5	6	7	8	9	10
Maths	2	4	1	5	3	9	7	10	6	8

Sol:- Given $n = 10$

let x_i be ranks in statistics

y_i be ranks in Maths

statistics x_i	Maths y_i	$d_i = x_i - y_i$	d_i^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			$\Sigma d_i^2 = 40$

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(40)}{10((10)^2-1)}$$

$$= 1 - \frac{6(40)^4}{10(100-1)}$$

$$= 1 - \frac{24}{99}$$

$$\rho = \frac{75}{99} = 0.7575$$

(2) A random sample of 5 students are selected and their grades in Maths and statistics are found to be

Maths	85	60	73	40	90
statistics	93	75	65	50	80

Then find rank correlation coefficient.

Sol :-

Here $n = 5$

Let x_i be ranks in Maths

y_i be ranks in statistics

[Here give "ranks" based on their marks in decending order]

Marks in Maths(x)	Rank x_i	Marks in statistics(y)	Rank y_i	$d_i = x_i - y_i$	d_i^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	60	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					$\Sigma d_i^2 = 4$

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(4)}{5(5^2-1)}$$

$$= 1 - \frac{6(4)}{5(24)}$$

$$= 1 - \frac{1}{5}$$

$$= \frac{5-1}{5}$$

$$= \frac{4}{5}$$

$$\rho = 0.8$$

(3) compute spearman's rank correlation coefficient to following data

Marks in subject A	20	14	36	29	5	11
Marks in subject B	19	9	25	10	2	6

Sol :-

No. of entries $n = 6$

Let x_i be the ranks in subject A

y_i be the ranks in subject B

Here give "ranks" based on their marks in decending order.

Marks in subject A x	Rank x_i	Marks in subject B y	Rank y_i	$d_i = x_i - y_i$	d_i^2
20	3	19	2	1	1
14	4	9	4	0	0
36	1	25	1	0	0
29	2	10	3	-1	1
5	6	2	6	0	0
11	5	6	5	0	0
					$\Sigma d_i^2 = 2$

spearman's rank correlation coefficient

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(2)}{6(6^2-1)}$$

$$= 1 - \frac{12}{6(35)}$$

$$\rho = \frac{35-2}{35} = \frac{33}{35} = 0.9429$$

Equal ranks (or) Repeated ranks :-

If any two (or) more items are with same value in that case common ranks are given to repeated items. The common rank is the average of the ranks which these items would have assumed if they were different from each other and the next item will get next rank to ranks already used in computing the common rank.

The rank correlation coefficient when repeated ranks are exist is

$$\rho = 1 - \frac{6\left(\sum d_i^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \cdots\right)}{n(n^2-1)}$$

where n = number of observations in series

m = number of times rank is repeated

d_i = difference of corresponding ranks of x and y

Ex:-

1. An item is repeated at rank "4" twice then the common rank $= \frac{4+5}{2} = \frac{9}{2} = 4.5$

2. An item is repeated thrice at rank 3 then the common rank $= \frac{3+4+5}{3} = \frac{12}{3} = 4$ and next rank will be "6"

Prob (1) :- The following table gives the score obtained by 11 students in English and Telugu. Then find the rank correlation coefficient.

Scores of English	40	46	54	60	70	80	82	85	85	90	95
Scores of Telugu	45	45	50	43	40	75	55	72	65	42	70

Sol :- Number of observations $n = 11$

Let x_i be the ranks in English

y_i be the ranks in Telugu.

Marks in English(x)	Ranks x_i	Marks in Telugu(y)	Ranks y_i	$d_i = x_i - y_i$	d_i^2
40	11	45	7.5 (m=2)	3.5	12.25
46	10	45	7.5	2.5	6.25
54	9	50	6	3	9
60	8	43	9	−1	1
70	7	40	11	−4	16
80	6	75	1	5	25
82	5	55	5	0	0
85	3.5 (m=2)	72	2	1.5	2.25
85	3.5	65	4	−0.5	0.25
90	2	42	10	−8	64
95	1	70	3	−2	4
					$\Sigma d_i^2 = 140$

$\frac{3+4}{2} = 3.5$

$\frac{7+8}{2} = 7.5$

In x (English) series 85 repeated 2 times so $m = 2$

In y (Telugu) series 45 repeated 2 times so $m = 2$

The rank correlation coefficient for repeated ranks is given by

$$\rho = 1 - \frac{6\left\{\Sigma d_i^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{140 + \frac{1}{12}\times 2(2^2-1) + \frac{1}{12}\times 2(2^2-1)\right\}}{11(11^2-1)}$$

$$= 1 - \frac{6\left\{140 + \frac{1}{12}(3) + \frac{1}{12}(3)\right\}}{11(120)}$$

$$= 1 - \frac{6\left\{140 + \frac{1}{2} + \frac{1}{2}\right\}}{11(120)}$$

$$= 1 - \frac{141}{220}$$

$$\rho = \frac{220-141}{220} = \frac{79}{220} = 0.3591$$

Prob (2) :- A sample of 12 fathers and their elder sons gave the following data. Calculate the spearman's rank correlation coefficient.

Fathers	65	63	67	64	68	62	70	66	68	67	69	71
sons	68	66	68	65	69	66	68	65	71	67	68	70

Sol :-

Number of observations in the series $n = 12$

Let x_i be the ranks of the fathers

y_i be the ranks of the sons

Fathers x	Ranks x_i	Sons y	Ranks y_i	$d_i = x_i - y_i$	d_i^2
65	9	68	5.5	3.5	12.25
63	11	66	9.5	1.5	2.25
67	6.5	68	5.5	1	1
64	10	65	11.5	-1.5	2.25
68	4.5	69	3	1.5	2.25
62	12	66	9.5	2.5	6.25
70	2	68	5.5	-3.5	12.25
66	8	65	11.5	-3.5	12.25
68	4.5	71	1	3.5	12.25
67	6.5	67	8	-1.5	2.25
69	3	68	5.5	-2.5	6.25
71	1	70	2	-1	1
					$\sum d_i^2 = 72.5$

$\frac{4+5}{2} = 4.5$

$\frac{6+7}{2} = 6.5$

$\frac{4+5+6+7}{4} = 5.5$

$\frac{9+10}{2} = 9.5$

$\frac{11+12}{2} = 11.5$

In x series 68 repeated 2 times, 67 repeated 2 times so $m = 2, 2$

In y series 68 repeated 4 times, 65 repeated 2 times and 66 repeated 2 times

so $m = 4, 2, 2$

The spearman's rank correlation coefficient is given by

$$\rho = 1 - \frac{6\left\{\sum d_i^2 + \left(\frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1)\right) + \left(\frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1)\right)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{72.5 + \left(\frac{1}{12}2(2^2-1)+\frac{1}{12}2(2^2-1)\right)+\left(\frac{1}{12}4(4^2-1)+\frac{1}{12}2(2^2-1)+\frac{1}{12}2(2^2-1)\right)\right\}}{12(12^2-1)}$$

$$= 1 - \frac{6\left\{72.5 + \frac{4}{12}2(2^2-1) + \frac{1}{12}4(4^2-1)\right\}}{12(143)}$$

$$= 1 - \frac{6\left\{72.5 + \frac{8}{12}(3) + \frac{4}{12}(+5)\right\}}{12(143)}$$

$$= 1 - \frac{6\left\{72.5 + 2 + 5\right\}}{12(143)}$$

$$= 1 - \frac{79.5}{286}$$

$$= 1 - 0.2780$$

$$\rho = 0.7220$$

H.W

Prob (3) :- obtain rank correlation coefficient for the following data

x	50	55	65	50	55	60	50	65	70	75
y	110	110	115	125	140	115	130	120	115	160

Prob (4) :- The coefficient of rank correlation of the marks data given by 10 students in statistics and Engineering Maths has found to be 0.5. It was later discover that the difference in the ranks of two subjects obtain by one of the students was wrongly taken as 3 instead of 7. Find the corrected rank correlation coefficient.

Sol :- Given

Number of students $n = 10$

rank correlation coefficient $\rho = 0.5$

we know that

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$0.5 = 1 - \frac{6\Sigma d_i^2}{10(10^2-1)}$$

$$0.5 = 1 - \frac{6\Sigma d_i^2}{10(99)}$$

$$\frac{6\Sigma d_i^2}{990} = 1 - 0.5$$

$$6\Sigma d_i^2 = 990(0.5)$$

$$6\Sigma d_i^2 = 495$$

$$\Sigma d_i^2 = \frac{495}{6} = 82.5$$

Also given that wrongly difference of ranks 3 is taken, instead of 7

so corrected $\Sigma d_i^2 = 82.5 - 3^2 + 7^2$

$= 82.5 - 9 + 49$

$= 122.5$

The corrected rank correlation coefficient

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

$$= 1 - \frac{6(122.5)}{10(10^2-1)}$$

$$= 1 - \frac{6(122.5)}{10(99)}$$

$$= 1 - 0.7424$$

$$\rho = 0.2576$$

H.W

Prob ⑤ :- The ranks of the 15 students in two subjects A and B are given below, the two numbers with in the brackets denoting the ranks of the students in A and B respectively. (1,10), (2,7), (3,2), (4,6), (5,4), (6,8), (7,3), (8,1), (9,11) (10,15), (11,9), (12,5), (13,14), (14,12) and (15,13). Use spearman's formula find rank correlation coefficient.

Sol:- Here $n = 15$ Ans:- 0.514

Principle of least squares :-

Principle of least square consists of determing the values of the unknown parameters by minimize the sum of its squares of the errors

Method of least squares :-

Let $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ be the "n" observed set of values of an experiment and $y = f(x)$ the curve to be fitted for the given data.

At $x = x_i$ the observed value of y is y_i and expected value for the fitted curve is $f(x_i)$ and e_i is the error approximation. $e_i = y_i - f(x_i)$

$S = e_1^2 + e_2^2 + \cdots + e_n^2$ is the sum of the squares of the errors of the curve for given set of points, If 'S' is minimum then it is said to be best fitted curve. This method is known as the method of least squares. The method of least squares consisting of minimizing "S"

***

Fitting of a straight line :-

Let $(x_1, y_1), (x_2, y_2) \cdots\cdots (x_n, y_n)$ be the "n" observed set of values of an experiment and

Let $y = a + bx$ be the straight line to be fitted — (1)
for the given data

By method of least squares normal equations of straight line are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad \text{— (2)}$$ where "n" is no. of observations

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad \text{— (3)}$$

Solving equations (2) & (3) after substituting the values Σx_i, Σy_i, Σx_i^2 and $\Sigma x_i y_i$. we get the values of a and b. Let it be $\hat{a}$ and $\hat{b}$

substituting $\hat{a}$ and $\hat{b}$ in eqn (1), we have

$$y = \hat{a} + \hat{b}x$$

which is the best fitted curve for the given data

Problems

5M *** (1) Fit a straight line $y = a + bx$ for the following data

x	0	1	2	3	4
y	1.0	1.8	3.3	4.5	6.3

Sol:- Given data

x	0	1	2	3	4
y	1.0	1.8	3.3	4.5	6.3

Let $y = a + bx$ be the straight line to be fitted for the given data — (1)

Normal equations of the straight line are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad — (2)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad — (3)$$

Here $n = 5$

x_i	y_i	$x_i y_i$	x_i^2
0	1.0	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x_i = 10$	$\Sigma y_i = 16.9$	$\Sigma x_i y_i = 47.1$	$\Sigma x_i^2 = 30$

Substituting the tabular values in eqn's (2) & (3) we have

$$16.9 = 5a + 10b \quad — (4) \qquad [\because n = 5$$

$$47.1 = 10a + 30b \quad — (5)$$

Solving eqn's (4) & (5)

$2 \times (4) - (5) \Rightarrow$ $33.8 = 10a + 20b$

$$\underline{47.1 = 10a + 30b}$$

$$-13.3 = -10b$$

$$13.3 = 10b$$

$$b = \frac{13.3}{10} = 1.33$$

Substituting b value in eqn (4)

$$16.9 = 5a + 10(1.33)$$

$$16.9 = 5a + 13.3$$

$$5a = 16.9 - 13.3$$

$5a = 3.6$

$$a = \frac{3.6}{5} = 0.72$$

Substituting values of a and b in eqn (1) we have

$$y = 0.72 + 1.33x$$

which is the required best fitted straight line for the given data

---

(2) 5M Fit a straight line for the given data using method of least squares. Also estimate the value of y when $x = 10$

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Sol :- Given data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Let $y = a + bx$ be straight to be fitted for the given data — (1)

By method of least squares normal equations of the straight line are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad — (2)$$

where "n" is no. of observations

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad — (3)$$

x_i	y_i	$x_i y_i$	x_i^2
1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	63	81
11	8	88	121
14	9	126	196
$\Sigma x_i = 56$	$\Sigma y_i = 40$	$\Sigma x_i y_i = 364$	$\Sigma x_i^2 = 524$

substituting the tabular values in eqns (2) & (3) we have

$$40 = 8a + 56b \quad -(4)$$

$$364 = 56a + 524b \quad -(5)$$

solving eqns (4) & (5)

$7 \times (4) - (5) \Rightarrow$

$$280 = 56a + 392b$$
$$364 = 56a + 524b$$
$$-84 = -132b$$

$$132b = 84$$

$$b = \frac{84}{132} = 0.6364$$

Put 'b' value in eqn (4), we have

$$40 = 8a + 56(0.6364)$$
$$40 = 8a + 35.6384$$
$$8a = 40 - 35.6384$$
$$8a = 4.3616$$
$$a = \frac{4.3616}{8} = 0.5452$$

substituting values of a and b in eqn (1) we have

$$y = 0.5452 + 0.6364x \quad -(6)$$

which is the required best fitted straight line for the given data

Deduction:- put $x = 10$ in equation (6)

$$y = 0.5452 + 0.6364(10)$$
$$= 0.5452 + 6.3640$$
$$= 6.9092$$

∴ when $x = 10$, the value of y is 6.9092

(3) Fit a straight line $y = a + bx$ to the following data by the method of least squares.

x	0	1	3	6	8
y	1	3	2	5	4

Ans :- $y = 1.646 + (0.3761)x$

Fitting of a parabola (or) second degree polynomial :-

Let $(x_1, y_1), (x_2, y_2) \dots\dots (x_n, y_n)$ be the "n" observed set of values of an experiment.

Let $y = a + bx + cx^2$ be the second degree polynomial to be fitted for the given data. — (1)

By the method of least squares normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad — (2)$$

where "n" is no. of observations

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad — (3)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad — (4)$$

solving equations (2), (3) & (4) after substituting the values Σx_i, Σy_i, $\Sigma x_i y_i$, Σx_i^2, Σx_i^3, Σx_i^4, $\Sigma x_i^2 y_i$, we get values a, b & c. Let it be $\hat{a}$, $\hat{b}$ & $\hat{c}$

substituting the values $\hat{a}$, $\hat{b}$ and $\hat{c}$ in equation (1) we get

$$y = \hat{a} + \hat{b}x + \hat{c}x^2$$

which is the best-fitted curve for the given data.

Problems :-

(1) Fit a second degree polynomial to the following data

x	0	1	2	3	4
y	1.0	1.8	1.3	2.5	6.3

Sol :- Given data

x	0	1	2	3	4
y	1.0	1.8	1.3	2.5	6.3

Let $y = a + bx + cx^2$ — (1) be the second degree polynomial to be fitted for the given data

By the method of least squares the normal equations of the second degree polynomial are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad -(2)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad -(3)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad -(4)$$

Here $n = 5$

x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y_i$	x_i^3	x_i^4
0	1.0	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	6.3	25.2	16	100.8	64	256
$\Sigma x_i = 10$	$\Sigma y_i = 12.9$	$\Sigma x_i y_i = 37.1$	$\Sigma x_i^2 = 30$	$\Sigma x_i^2 y_i = 130.3$	$\Sigma x_i^3 = 100$	$\Sigma x_i^4 = 354$

Substituting the tabular values in eqn's (2), (3) & (4) we have

$$12.9 = 5a + 10b + 30c \quad -(5)$$

$$37.1 = 10a + 30b + 100c \quad -(6)$$

$$130.3 = 30a + 100b + 354c \quad -(7)$$

Solving eqn's (5) & (6)

$2 \times (5) - (6) \Rightarrow$

$$10a + 20b + 60c = 25.8$$
$$10a + 30b + 100c = 37.1$$
$$-10b - 40c = -11.3$$
$$10b + 40c = 11.3$$
$$b + 4c = 1.13 \quad -(8)$$

Solving eqn's ⑥ & ⑦, we have

$3\times⑥ - ⑦ \Rightarrow$ $30a + 90b + 300c = 111.3$

$-30a - 100b - 354c = -130.3$

$-10b - 54c = -19.0$

$10b + 54c = 19.0 \quad -⑨$

Solving eqn's ⑧ & ⑨, we have

$10\times⑧ - ⑨ \Rightarrow$ $10b + 40c = 11.3$

$-10b - 54c = -19.0$

$-14c = -7.7$

$14c = 7.7$

$c = \frac{7.7}{14} = 0.55$

Put $c = 0.55$ in eqn ⑧

$b + 4(0.55) = 1.13$

$b = 1.13 - 4(0.55)$

$= -1.07$

Put $b = -1.07$ & $c = 0.55$ in eqn ⑤ we have

$12.9 = 5a + 10(-1.07) + 30(0.55)$

$12.9 = 5a + 5.8$

$5a = 12.9 - 5.8$

$5a = 7.1$

$a = \frac{7.1}{5} = 1.42$

Substituting the values of a, b and c in eqn ① we have

$y = 1.42 + (-1.07)x + (0.55)x^2$

$y = 1.42 - 1.07x + 0.55x^2$

which is required best fitted second degree polynomial curve for the given data.

(2) Fit a parabola (second degree polynomial) to the data points given in the following table

x	0	1.0	2.0
y	1.0	6.0	17.0

Sol :- Give data

x	0	1.0	2.0
y	1.0	6.0	17.0

Let $y = a + bx + cx^2$ — (1) be the parabola to be fitted for the given data

By the method of atleast squares the normal equations of the parabola are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad -(2)$$

where "n" is no. of observations

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad -(3)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad -(4)$$

Here $n = 3$

x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y_i$	x_i^3	x_i^4
0	1.0	0	0	0	0	0
1.0	6.0	6.0	1	6	1	1
2.0	17.0	34.0	4	68	8	16
$\Sigma x_i = 3$	$\Sigma y_i = 24$	$\Sigma x_i y_i = 40$	$\Sigma x_i^2 = 5$	$\Sigma x_i^2 y_i = 74$	$\Sigma x_i^3 = 9$	$\Sigma x_i^4 = 17$

Substituting the tabular values in eqn's (2), (3) & (4) we have

$$24 = 3a + 3b + 5c \quad -(5)$$

$$40 = 3a + 5b + 9c \quad -(6)$$

$$74 = 5a + 9b + 17c \quad -(7)$$

Solving eqn's (5) & (6)

(5) − (6) ⇒ $3a + 3b + 5c = 24$

$\underline{-3a + 5b + 9c = 40}$

$-2b - 4c = -16$

$2b + 4c = 16$

$b + 2c = 8$ — (8)

Solving eqn's (6) and (7), we have

$5 \times (6) - 3 \times (7) \Rightarrow 15a + 25b + 45c = 200$

$\underline{-15a + 27b + 51c = 222}$

$-2b - 6c = -22$

$2b + 6c = 22$

$b + 3c = 11$ — (9)

Solving eqn's (8) and (9), we have

(8) − (9) ⇒ $b + 2c = 8$

$\underline{-b + 3c = 11}$

$-c = -3$

$c = 3$

Put $c = 3$ in eqn (8)

$b + 2(3) = 8$

$b = 8 - 6$

$b = 2$

Put $b = 2$ and $c = 3$ in eqn (5) we have

$3a + 3(2) + 5(3) = 24$

$3a + 6 + 15 = 24$

$3a = 24 - 21$

$3a = 3$

$a = \frac{3}{3} = 1$

Substituting values of a, b and c in eqn (1) we have

$y = 1 + 2x + 3x^2$

which is the required best fitted second degree polynomial (parabola) curve for the given data.

(3) using least squares method, fit a second degree polynomial to the following data

x	0	1	2	3	4	5	6	7	8
y	12.0	10.5	10.0	8.0	7.0	8.0	7.5	8.5	9.0

Ans :- $y = 12.2 - 1.85x + 0.183x^2$

Fitting of exponential curve $y = ae^{bx}$:-

Let $(x_1, y_1), (x_2, y_2) \cdots\cdots (x_n, y_n)$ be the "n" observed set of values of an experiment.

Let $y = ae^{bx}$ ① be the exponential curve to be fitted for the given data.

Taking $\log_e$ on both sides in eqn ①

$$\log_e y = \log_e (ae^{bx})$$

$$= \log_e a + \log_e e^{bx}$$

$$= \log_e a + bx \log_e e$$

$$\log_e y = \log_e a + bx \quad (1) \qquad \because \log_e e = \ln(e)$$

Now Let $Y = \log_e y$, $X = x$, $A = \log_e a$ then we have

$$Y = A + bX$$

which is a straight line eqn.

By the method of least squares, normal equations of the straight line are

$$\sum_{i=1}^{n} Y_i = nA + b \sum_{i=1}^{n} X \quad — (2)$$

$$\sum_{i=1}^{n} X_i Y_i = A \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 \quad — (3)$$

Solving these two normal equations after substituting values of ΣX_i, ΣY_i, $\Sigma X_i Y_i$, ΣX_i^2 we will get the values of A and

we have

$$A = \log_e a$$

$$a = e^A$$

substituting values of a and b in eqn (1) we get the best fitted exponential curve to the given data

Problems :-

(1) using method of least squares fit a exponential curve of form $y = ae^{bx}$ to the following data

x	0	1	2	3	4	5	6	7	8
y	20	30	52	77	135	211	326	550	1052

Sol Given data

x	0	1	2	3	4	5	6	7	8
y	20	30	52	77	135	211	326	550	1052

Let $y = ae^{bx}$ — (1) be the exponential curve to be fitted for the given data.

taking $\log_e$ on both sides in eqn (1)

$$\log_e y = \log_e (ae^{bx})$$

$$= \log_e a + \log_e e^{bx}$$

$$= \log_e a + bx \log_e e$$

$$\log_e y = \log_e a + bx \qquad [\because \log_e e = \ln(e) = 1]$$

let $Y = \log_e y$ $X = x$ $A = \log_e a$ then we have

$$Y = A + bX$$

which is a straight line equation.

By method of least squares normal equations of the straight line are

$$\sum_{i=1}^{n} Y_i = nA + b\sum_{i=1}^{n} X_i \quad — (2)$$

$$\sum_{i=1}^{n} X_i Y_i = A\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \quad — (3)$$

Here $n = 9$

x	y	$X_i = x_i$	$Y_i = \log_e y = \ln(y)$	$X_i Y_i$	X_i^2
0	20	0	2.9957	0	0
1	30	1	3.4012	3.4012	1
2	52	2	3.9512	7.9024	4
3	77	3	4.3438	13.0314	9
4	135	4	4.9053	19.6212	16
5	211	5	5.3519	26.7595	25
6	326	6	5.7869	34.7214	36
7	550	7	6.3099	44.1693	49
8	1052	8	6.9584	55.6672	64
		$\Sigma X_i = 36$	$\Sigma Y_i = 44.0043$	$\Sigma X_i Y_i = 205.2736$	$\Sigma X_i^2 = 204$

substituting the tabular values in eqns (2) & (3) we have

$$44.0043 = 9A + 36b \quad -(4)$$

$$205.2736 = 36A + 204b \quad -(5)$$

Solving eqn's (4) & (5)

$4 \times (4) - (5) \Rightarrow 36A + 144b = 176.0172$

$$36A + 204b = 205.2736$$

$$-60b = -29.2564$$

$$b = \frac{29.2564}{60}$$

$$= 0.4876$$

put $b = 0.4876$ in eqn (4) we have

$$44.0043 = 9A + 36(0.4876)$$

$$9A = 44.0043 - 17.5536$$

$$9A = 26.4507$$

$$A = \frac{26.4507}{9} = 2.9390$$

we have

$$A = \log_e a \Rightarrow a = e^A$$

$$= e^{2.9390} = 18.8969$$

substituting values of a and b in eqn (1)

we have $y = (18.8969)\, e^{0.4876x}$

which is the required best fitted exponential curve for the given data.

(2) Fit a curve of the form $y = ae^{bx}$ to the following data by least squares method.

x	1	5	7	9	12
y	10	15	12	15	21

Sol:- Given data

x	1	5	7	9	12
y	10	15	12	15	21

Let $y = ae^{bx}$ — (1) be the exponential curve to be fitted for the given data

Taking $\log_e$ on both sides in eqn (1)

$$\log_e y = \log_e ae^{bx}$$

$$= \log_e a + \log_e e^{bx}$$

$$\log_e y = \log_e a + bx \log_e e$$

$$\log_e y = \log_e a + bx \qquad \because \log_e e = 1$$

let $Y = \log_e y$, $A = \log_e a$, $X = x$ then we have

$$Y = A + bX$$

which is a straight line equation. By the method of least squares normal equations of the straight line are

$$\Sigma Y_i = nA + b\Sigma X_i \quad — (2)$$

$$\Sigma X_i Y_i = A\Sigma X_i + b\Sigma X_i^2 \quad — (3)$$

x	y	$X_i = x_i$	$Y_i = \log_e y_i = \ln y_i$	$X_i Y_i$	X_i^2
1	10	1	2.3026	2.3026	1
5	15	5	2.7080	13.54	25
7	12	7	2.4850	17.395	49
9	15	9	2.7080	24.372	81
12	21	12	3.0445	36.534	144
		$\Sigma X_i = 34$	$\Sigma Y_i = 13.2481$	$\Sigma X_i Y_i = 94.1436$	$\Sigma X_i^2 = 300$

Substituting the above tabular values in eqns (2) & (3) we have

$$13.2481 = 5A + 34b \quad -(4)$$

$$94.1436 = 34A + 300b \quad -(5)$$

Solving eqn's (4) & (5) we have

$34 \times (4) - 5 \times (5) \Rightarrow$ $170A + 1156b = 450.4354$

$\underline{-170A \pm 1500b = \underline{4}70.718}$

$$-344b = -20.2826$$

$$b = \frac{20.2826}{344}$$

$$b = 0.0590$$

put $b = 0.0590$ in eqn (4), we have

$$13.2481 = 5A + 34(0.0590)$$

$$5A = 13.2481 - 2.006$$

$$5A = 11.2421$$

$$A = \frac{11.2421}{5} = 2.2484$$

we have $A = \log_e a$

$$\Rightarrow a = e^A = e^{2.2484} = 9.4726$$

Substituting values of a and b in eqn (1), we have

$$y = 9.4726\, e^{0.059x}$$

which is the required best fitted exponential curve for the given data

(3) fit a exponential curve for the following data

x	2	3	4	5	6
y	8.3	15.4	33.1	64.2	127.4

Regression Analysis :- Regression is the measure of average relationship between two or more variables of the given data. In regression analysis there are two types of variables.

The variable which is influenced is called **dependent variable**. The variable which influence the value is called **independent variable**

Types of Regression :-

Some important types of regression are

1. **Simple regression :-**

The regression analysis for studing of only two variables at a time is called the "simple regression"

2. **Linear Regression :-**

If the regression curve is a straight line then there is a linear regression between the variably under study.

3. **Non-linear Regression :-** If the regression is not a straight line then it is called a non-linear (or) curvilinear regression.

Ex :- parabola, exponential curves

Regression coefficient :-

Let "b" be the slope of the line of regression of y on x, is called the regression coefficient of y on x. It represents the increment in the value of dependent variable y corresponding to a unit change in the value of independent variable x. we write the regression coffiicient of y on x is b_{yx}. It can be represented as

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

"y the regression coefficient of x ony indicates the change in the value of variable x corresponding to a unit change in the value of variable y and we write regression coefficient of x on y is b_{xy} It can be represented as

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

Properties of Regression coefficient :-

1. correlation coefficient is the geometric mean b/w the regression coefficients

$$\text{i.e } r = \pm\sqrt{b_{yx} \times b_{xy}} \Rightarrow r^2 = b_{yx} \times b_{xy}$$

2. If one of the regression coefficient is greater than unity, then the other must be less than unity

If $b_{yx} > 1$ then $b_{xy} < 1$

3. The modulus value of the arithmetic mean of the regression coefficient is not less than (greater than) the modulus value of the correlation coefficient "r"

$$\text{ie If } \left|\frac{1}{2}(b_{yx} + b_{xy})\right| > |r| \text{ then } (\sigma_y - \sigma_x)^2 > 0$$

4. Regression coefficients are independent of the change of origin but not of scale ie $b_{yx} \neq b_{xy}$

5. Both regression coefficients will have the same sign

6. The angle b/w two regression lines is

$$\tan\theta = \left(\frac{1 - r^2}{|r|}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

***

Difference between correlation and Regression :-

Correlation	Regression
1. correlation is a measure of degree of covariability between two variables.	1. Regression establishes the functional relationship between dependent and independent variables.
2. In correlation both the variables are random variables.	2. In Regression one variable is dependent variable and other one is independent variable
3. The correlation coefficient is a relative measure	3. Regression coefficient is an absolute measure
4. In correlation $r_{xy} = r_{yx}$	4. In Regression $b_{xy} \neq b_{yx}$

Regression equation (or) Regression line :-

(i) Regression line of y on x :-

It is the line which gives the best estimate for the value of y for a specified value of x and is given by

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad (\because b_{yx} = r\frac{\sigma_y}{\sigma_x})$$

$$(y - \bar{y}) = \frac{cov(x,y)}{\sigma_x \sigma_y}\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad (\because r = \frac{cov(x,y)}{\sigma_x \sigma_y})$$

It can also be convert in to the form of $y = a + bx$

where $\bar{x}$ = mean of x series $= \frac{\Sigma x_i}{n}$

$\bar{y}$ = mean of y series $= \frac{\Sigma y_i}{n}$

r = correlation coefficient of x and y

i.e $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$$cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$$

S.D of $x = \sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

S.D of $y = \sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

(ii) Regression line of x on y :-

It is the line which gives the best estimate for the value of x for a specified value of y.

It is given by

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad (\because b_{xy} = r\frac{\sigma_x}{\sigma_y})$$

$$(x - \bar{x}) = \frac{cov(x,y)}{\sigma_x \sigma_y}(y - \bar{y}) \quad (\because r = \frac{cov(x,y)}{\sigma_x \sigma_y})$$

It can also be convert in to the form of $x = a + by$

Note :- Regression lines always passes through $(\bar{x}, \bar{y})$

Problems :-

1. Find the most likely production corresponding to a rain fall 40 from the following data

	Rainfall x	Production y
Average	30	500 kgs
S.D	5	100 kgs
r	0.8	

Sol :- Given data

	Rainfall x	production y
Average	30	500 kgs
S.D	5	100 kgs
r	0.8	

Here we have to find production y value at rain fall $x = 40$

So we have to find the Regression eqn of y on x

Given $\bar{x} = 30$, $\sigma_x = 5$, $\bar{y} = 500$, $\sigma_y = 100$ and $r = 0.8$

The Regression equation of y on x is

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$(y - 500) = 0.8 \left(\frac{100}{5}\right)(x - 30)$$

$$y - 500 = 16(x - 30)$$

$$y - 500 = 16x - 480$$

$$y = 16x - 480 + 500$$

$$y = 16x + 20$$

which is the regression eqn of y on x

$\therefore$ put $x = 40$ in above eqn

$$y = 16(40) + 20$$

$$= 640 + 20$$

$$y = 660$$

$\therefore$ production corresponding to rainfall $x = 40$ is $y = 660$.

(2) construct two regression lines for the following data and estimate (a) profit if sales are Rs 12000 crs
(b) sales if profit is Rs 200 crs

	Sales (Rs. in crores)	Profits Rs. in crores
Average	1500	130
Variance	144	81

correlation coefficient is 0.65

Sol :- Let x is sales & y is profit

From the given data

Average (or) mean of sales x is $\bar{x} = 1500$

Average (or) mean of profit y is $\bar{y} = 130$

Variance of sales x is $\sigma_x^2 = 144$

$\Rightarrow$ standard deviation of x is $\sigma_x = \sqrt{\text{variance}}$
$= \sqrt{144} = 12$

variance of profit y is $\sigma_y^2 = 81$

$\Rightarrow$ standard deviation of y is $\sigma_y = \sqrt{\text{variance}}$
$= \sqrt{81} = 9$

Correlation coefficient $\rho = 0.65$

Regression coefficient of x on y is $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$b_{xy} = 0.65\left(\frac{12}{9}\right)$$
$$= 0.8667$$

Regression coefficient of y on x is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$
$$= 0.65\left(\frac{9}{12}\right) = 0.4875$$

The Regression line of y on x is

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$
$$(y - 130) = 0.4875(x - 1500)$$
$$y - 130 = 0.4875x - 1500(0.4875)$$
$$y = 0.4875x - 1500(0.4875) + 130$$
$$y = 0.4875x - 601.25 \quad -(1)$$

The Regression line of x on y is

$$(x-\bar{x}) = b_{xy}(y-\bar{y})$$

$$(x-1500) = 0.8667\,(y-130)$$

$$x-1500 = 0.8667y - 130(0.8667)$$

$$x = 0.8667y - 130(0.8667) + 1500$$

$$x = 0.8667y + 1387.329 \quad \text{—— (2)}$$

(a) Given sales $x = 2000$

Now we have to find profit $y = ?$

Put $x = 2000$ in eqn (1)

$$y = 0.4875(2000) - 601.25$$

$$y = 373.25 \text{ Crs}$$

(b) Given profit $y = 200$

Now we have to find sales $x = ?$

Put $y = 200$ in eqn (2)

$$x = 0.8667(200) + 1387.329$$

$$x = 1560.669$$

H.W

Prob (3) :- Find two regression lines on basis of the following data

	x	y
Mean	40	45
S.D	10	9

Karl pearsons correlation coefficient is 0.5 Also estimate the value of y for $x = 48$

Ans :- $y = 0.45x + 27$

$x = 0.5556y - 14.998$

y value when $x = 48$ is 48.6

Prob (4) :- From a sample of 200 pairs of observations the following quantities were calculated

$\Sigma x = 11.34$, $\Sigma y = 20.78$, $\Sigma x^2 = 12.16$, $\Sigma y^2 = 84.96$

$\Sigma xy = 22.13$ then find Regression eqn of y on x and compute the correlation coefficient.

Sol :- Given no. of observations $n = 200$

Also given $\Sigma x = 11.34$, $\Sigma y = 20.78$, $\Sigma x^2 = 12.16$

$\Sigma y^2 = 84.96$ and $\Sigma xy = 22.13$

Mean (or) Average of x is $\bar{x} = \frac{\Sigma x_i}{n} = \frac{11.34}{200} = 0.0567$

Mean of y is $\bar{y} = \frac{\Sigma y_i}{n} = \frac{20.78}{200} = 0.1039$

By def of covariance, we have

$$cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y}$$

$$= \frac{1}{200}(22.13) - 0.0567(0.1039)$$

$$= 0.1107 - 0.0059$$

$$= 0.1048$$

Standard deviation of x is $\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$$= \sqrt{\frac{12.16}{200} - (0.0567)^2}$$

$$= \sqrt{0.0608 - 0.0032}$$

$$= \sqrt{0.0576}$$

$$\sigma_x = 0.24$$

Standard deviation of y is $\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

$$= \sqrt{\frac{84.96}{200} - (0.1039)^2}$$

$$= \sqrt{0.4248 - 0.0108}$$

$$\sigma_y = 0.6434$$

correlation coefficient

$$\gamma = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{0.1048}{0.24(0.6434)} = 0.6787$$

Regression coefficient of y on x is

$$b_{yx} = \gamma \cdot \frac{\sigma_y}{\sigma_x}$$

$$= 0.6787\left(\frac{0.6434}{0.24}\right)$$

$$b_{yx} = 0.1048$$

Regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$(y - 0.1039) = 0.1048(x - 0.0567)$$

$$y - 0.1039 = 0.1048x - 0.1048(0.0567)$$

$$y = 0.1048x - 0.0059 + 0.1039$$

$$y = 0.1048x + 0.098$$

Prob (5) :- From the following data obtain the two Regression equations and calculate the correlation coefficient.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Also estimate the value of y which should corresponding to $x = 6.2$

Sol :- By def of correlation coefficient, we have

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$

mean $\bar{x} = \frac{\Sigma x_i}{n}$, $\bar{y} = \frac{\Sigma y_i}{n}$

standard deviation SD $\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	9	9	1	81
2	8	16	4	64
3	10	30	9	100
4	12	48	16	144
5	11	55	25	121
6	13	78	36	169
7	14	98	49	196
8	16	128	64	256
9	15	135	81	225
$\Sigma x_i = 45$	$\Sigma y_i = 108$	$\Sigma x_i y_i = 597$	$\Sigma x_i^2 = 285$	$\Sigma y_i^2 = 1356$

Here $n = 9$

Mean $\bar{x} = \frac{\Sigma x_i}{n} = \frac{45}{9} = 5$

Mean $\bar{y} = \frac{\Sigma y_i}{n} = \frac{108}{9} = 12$

covariance $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$

$= \frac{1}{9}(597) - 5(12)$

$= 6.3333$

Standard Deviation $SD = \sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$= \sqrt{\frac{285}{9} - (5)^2} = 2.5820$

Standard Deviation of y is $SD = \sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

$= \sqrt{\frac{1356}{9} - (12)^2}$

$= 2.5820$

correlation coefficient $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$= \frac{6.3333}{2.5820(2.5820)}$

$= 0.95$

Regression coefficient of y on x is

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$= 0.95\left(\frac{2.5820}{2.5820}\right) = 0.95$

Regression coefficient of x on y is

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$= 0.95\left(\frac{2.5820}{2.5820}\right) = 0.95$

Regression line of y on x is

$(y - \bar{y}) = b_{yx}(x - \bar{x})$

$(y - 12) = 0.95(x - 5)$

$y = 0.95x - 4.75 + 12$

$y = 0.95x + 7.25$ — (1)

Regression line of x on y is

$(x - \bar{x}) = b_{xy}(y - \bar{y})$

$(x - 5) = 0.95(y - 12)$

$x = 0.95y - 11.4 + 5$

$$x = 0.95y - 6.4 \quad -(2)$$

y when $x = 6.2$:-

Put $x = 6.2$ in eqn (1)

$$y = 0.95(6.2) + 7.25$$

$$y = 13.14$$

Prob (6) :- Find two Regression lines of the following data

x	10	15	12	17	13	16	24	14	22	20
y	30	42	45	46	33	34	40	35	39	38

Sol:-

By def of correlation coefficient, we have

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}$

mean $\bar{x} = \frac{\Sigma x_i}{n}$ $\bar{y} = \frac{\Sigma y_i}{n}$

$$S.D = \sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	30	300	100	900
15	42	630	225	1764
12	45	540	144	2025
17	46	782	289	2116
13	33	429	169	1089
16	34	544	256	1156
24	40	960	576	1600
14	35	490	196	1225
22	39	858	484	1521
20	38	760	400	1444
$\Sigma x_i = 163$	$\Sigma y_i = 382$	$\Sigma x_i y_i = 6293$	$\Sigma x_i^2 = 2839$	$\Sigma y_i^2 = 14840$

Mean $\bar{x} = \frac{\Sigma x_i}{n} = \frac{163}{10} = 16.3$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{382}{10} = 38.2$$

covariance $cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y}$

$$= \frac{1}{10}(6293) - 16.3(38.2)$$

$$cov(x,y) = 6.64$$

Standard Deviation of $x = \sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \bar{x}^2}$

$$= \sqrt{\frac{2839}{10} - (16.3)^2}$$

$$= \sqrt{283.9 - 265.69}$$

$$\sigma_x = 4.2673$$

Standard Deviation of y is $\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \bar{y}^2}$

$$= \sqrt{\frac{14840}{10} - (38.2)^2}$$

$$= \sqrt{1484 - 1459.24}$$

$$= 4.9759$$

By def of correlation coefficient, we have

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$= \frac{6.64}{(4.2673)(4.9759)} = 0.3127$$

Regression coefficient of y on x is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$= 0.3127\left(\frac{4.9759}{4.2673}\right) = 0.3646$$

Regression coefficient of x on y is

$$b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right)$$

$$= 0.3127\left(\frac{4.2673}{4.9759}\right) = 0.2682$$

Regression line of y on x is

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

$$(y - 38.2) = 0.3646(x - 16.3)$$

$$y = 0.3646x - 5.9430 + 38.2$$

$$y = 0.3646x + 32.257$$

The Regression line of x on y is

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$(x - 16.3) = 0.2682(y - 38.2)$$

$$x = 0.2682y - 10.2452 + 16.3$$

$$x = 0.2682y + 6.0548$$

***

Prob (7) :- The equations two regression lines obtained in a correlation analysis are the following $2x = 8 - 3y$ and $2y = 5 - x$ then find the value of the correlation coefficient.

Sol :- Given two Regression lines

$$2x = 8 - 3y$$

$$2y = 5 - x$$

we know that Regression lines passes through the point $(\bar{x}, \bar{y})$

so, we have

$$2\bar{x} = 8 - 3\bar{y} \text{ and } 2\bar{y} = 5 - \bar{x}$$

$$\bar{x} = \frac{8 - 3\bar{y}}{2} \qquad \bar{y} = \frac{5 - \bar{x}}{2}$$

$$\bar{x} = \frac{8}{2} - \frac{3}{2}\bar{y} \qquad \bar{y} = \frac{5}{2} - \frac{1}{2}\bar{x}$$

$$\bar{x} = 4 - \frac{3}{2}\bar{y}$$

This is in the form of $\bar{y} = a + b\bar{x}$

This is in its form of $\bar{x} = a + b\bar{y}$

Here Regression coefficient $b_{yx} = b = -\frac{1}{2}$

Here Regression coefficient $b_{xy} = b = -\frac{3}{2}$

we know that correlation coefficient

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

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

$$= \pm\sqrt{\frac{3}{4}}$$

$$r = \pm 0.866$$

Prob (5) :- In a partially destroyed laboratory record of an analysis of correlation data the following results are legible. Variance of x is 9, Regression equations are $8x - 10y + 66 = 0$, $40x - 18y = 214$. Who are

(i) Mean values of x and y

(ii) correlation coefficient between two variable x and y

(iii) standard deviation of y

Sol :- Given that

variance of x is $\sigma_x^2 = 9 \Rightarrow \sigma_x = \sqrt{9}$

S.D $\Rightarrow \sigma_x = 3$

Regression lines are

$$8x - 10y + 66 = 0$$

$$40x - 18y = 214$$

(i) we know that Regression lines are passes through the point $(\bar{x}, \bar{y})$

So, we have

$$8\bar{x} - 10\bar{y} = -66 \quad -(1)$$

$$40\bar{x} - 18\bar{y} = 214 \quad -(2)$$

Solving eqns (1) & (2)

$5 \times (1) - (2) \Rightarrow$

$$40\bar{x} - 50\bar{y} = -330$$

$$-40\bar{x} + 18\bar{y} = -214$$

$$-32\bar{y} = -544$$

$$\bar{y} = \frac{544}{32} = 17$$

Put $\bar{y} = 17$ in equation (1), we have

$$8\bar{x} - 10(17) = -66$$

$$8\bar{x} = -66 + 170$$

$$\bar{x} = \frac{104}{8} = 13$$

$\therefore$ Means of x and y are $\bar{x} = 13$ and $\bar{y} = 17$

(ii) from eqn (1), we have

$$8\bar{x} - 10\bar{y} = -66$$

$$10\bar{y} = 8\bar{x} + 66$$

$$\bar{y} = \frac{8}{10}\bar{x} + \frac{66}{10}$$

$$\bar{y} = \frac{4}{5}\bar{x} + \frac{33}{5}$$

from eqn (2) we have

$$40\bar{x} - 18\bar{y} = 214$$

$$40\bar{x} = 214 + 18\bar{y}$$

$$\bar{x} = \frac{214 + 18\bar{y}}{40}$$

$$\bar{y} = \frac{33}{5} + \frac{4}{5}\bar{x}$$

This is in the form of

$$\bar{y} = a + b\bar{x}$$

Regression coefficient

$$b_{yx} = b = \frac{4}{5}$$

$$\bar{x} = \frac{214}{40} + \frac{18}{40}\bar{y}$$

$$\bar{x} = \frac{107}{20} + \frac{9}{20}\bar{y}$$

This is in the form of

$$\bar{x} = a + b\bar{y}$$

Regression coefficient

$$b_{xy} = b = \frac{9}{20}$$

The correlation coefficient of x and y is

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

$$= \pm\sqrt{\frac{4}{5} \times \frac{9}{20}}$$

$$= \pm\sqrt{\frac{9}{25}}$$

$$r = \pm \frac{3}{5} = \pm 0.6$$

(iii) we know that the regression coefficient of y on x is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$\frac{4}{5} = 0.6\,\frac{\sigma_y}{3}$$

$$\frac{4}{5} \times \frac{3}{0.6} = \sigma_y$$

$$4 = \sigma_y$$

$$\therefore \sigma_y = 4$$

H.W The standard deviation of y is $\sigma_y = 4$

Prob ⑨ :- Two random variables have the regression eqns $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Then find the mean values and the correlation coefficient of x and y. If the variance of x is 25 then find the standard deviation of y from the above data.

Ans:- $\bar{x} = 4$, $\bar{y} = 7$, $r = \pm 0.5$ and $\sigma_y = 15$