Despcriptive Statistics & Methods for Data science

Data Science:- Data science is the science which is used in computer science, statistics, machine learning visualization and Human computer interactions to collect, clean, integrate, Analyze, interact with data to create data products.

* Data science is a multi disciplinery field that uses scientific methods, processes, Algorithms and systems to extract knowledge from structure and unstructured data.

Eg:- Datascience and Machine learning can be used in cyber security programs to identity threats, attack, scam, malvare and also to prevent food.

* Data science provides meaningful information based on large amounts of Complex data (or) big data. Data science is about solving bussiness problems. Data science deals with enriching the data and making it better for their company. To Analyze the data and improving its quality

Statistics:- Statistic is a branch that deals with the study of collection, Analysis, interpretation, organization and presentation of data. Mathematically, statistics is defined as the set of equations which are used to analyze the science of data.

Eg:- statistics used in agriculture, biology, bussiness, hospitals, colleges e.t.c.

There are two types of statistics:

1. Descriptive
2. Inferential

Descriptive Statistics:- Descriptive Statistics Summa-rizes (or) Describes characteristics of a Data science. Descriptive Statistics Consists of two basic categories of meausres

Scanned with CamScanner

1. Meausres of central tendancy
2. Meausres of variability. (or) spread (or) Methods of Dispersion.

Meausres of central tendancy include the mean, median and mode. while meausres of variability include the standard deviation, variance, minimum, maximum variables and Skewness and kurtosis.

Infrential Statistics:- Infrential statistics that gives sample data to make decision (or) prediction. the most common methadologies are hypothesis test, confidence intervals and regression Analysis

population:- Population refers to the total set of observations that can be made. population includes all the elements from a set of data.

Eg:- Total students in a college.

sample:- A sample consists of one (or) more observations brought from the population.

statistical data:- A Sequence of observations made on a set of objects included in a sample drawn from population is known as statistical data.

i) ungrouped data:- The data which have been arranged in a systematic order is called ungrouped data (or) raw data

Eg:- 0,1,2,3---

ii) Grouped data:- Grouped data presented in the form of frequency distribution.

Eg:-

Classes	frequency
0-10	15
10-20	20
20-30	15

Collection of data:- the first step in an investigation is the collection of data, the data may be collected for the whole population (or) for the sample only. It is mostly collected on a sample basis

Types of data:- there are two types for the collection of data
i) primary data
ii) secondary data.

Scanned with CamScanner

i) Primary data:- Primary data is the first hand information which is directly collected from one source. It can be obtained from
* Direct personal Observation
* Direct / Indirect oral interviews
* Admistrative Questonaries

ii) Secondary data:- Secondary data is the Second hand Information which is already collected by an organisation for some purpose and are available for the present study. It can be obtained from
* Official [Applications, Agriculture, Industries]
* Semi-Official [Bank, railways e.t.c]
* Technical, ~~Genders~~, News paperals, Journals.

Type of variables:- i) Independent variable ii) dependent variable

i) Independent variable:- It is a variable that is the cause (or) reason of any situation which can be manipulated. this is also known as experimental (or) predictor variable.

ii) Dependent variable:- It is a variable something that depend on other factors. It is also known as outcome variable.

Eg:- $\underbrace{\text{Time spent on study}}_{\text{Independent}}$ causes a change in $\underbrace{\text{test mark}}_{\text{Dependent}}$.

Categorical variable:- Categorical variables represent the types of data. this is also known as discrete (or) qualitative variable.

Continous variable:- This variable is not restricted to particular values It is also known as Quantitative variable

Meausres of Central tendancy:- Central tendancy is that value which is most representative of that data science. It is a statistical meausre and calculates the location (or) position of a central point to explain the central tendancy of the whole Quantity of data. Meausres of central tendancy is also known as Meausre of central value (or) Meausre of location (or) Average of first order.

Meausres of central tendency are often called as averages.

Eg:- Kohli is representative of Cricket team of India.

The Three most common meausres of central tendancy are the mean, median and mode.

Scanned with CamScanner

Mean (Arthimetic mean):- Arthimetic mean of set of observa-tions is their sum divided by the No. of observations.

$$\bar{x} = \frac{\text{Sum of observations}}{\text{Total no. of observations}} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

Direct Method:- In case of frequency distribution

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum f_ix_i}{\sum f} = \frac{\sum f_ix_i}{N}$$

Shortcut Method:- If the values of x (or) f are large

$$\bar{x} = A + \frac{\sum f_id_i}{N}$$ where $d = x_i - A$, A = Assumed mean

Step-Deviation Method:-

$$\bar{x} = A + \frac{\sum f_iu_i}{N} \times h$$

where, h = length of interval

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

* If $\bar{x}_1, \bar{x}_2$ be the means of two samples of size n_1, n_2 then the mean $\bar{x}$ of the combined sample of size $n_1 + n_2$ is given by $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

Scanned with CamScanner

Calculate mean of the frequency distribution relating to the weight of 120 articles

Weight	0-10	10-20	20-30	30-40	40-50	50-60
No. of Articles	14	17	22	26	23	18

A) Direct method:- $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

class	Midvalue (x)	freq	$f_i x_i$	d_i $x_i - A$	$f_i d_i$	u	$f_i u_i$
0-10	5	14	70	-20	-280	-2	-28
10-20	15	17	255	-10	-170	-1	-17
20-30	25	22	550	0	0	0	0

Scanned with CamScanner

30 – 40	35	26	910	10	260	1	26
40 – 50	45	23	1035	20	460	2	46
50 – 60	55	18	990	30	540	3	54
		$N = \sum f_i = 120$	$\sum f_i x_i = 3810$		$\sum f_i d_i = 810$		81

$$= \frac{3810}{120} = 31.75.$$

Short cut method: $\bar{x} = A + \frac{\sum f_i d_i}{N}$ where $d_i = x_i - A$

$$= 25 + \frac{810}{120} = 31.75.$$

Step deviation method:– $\bar{x} = A + \frac{\sum f_i u_i}{N} \times h$ $\quad u = \frac{x_i - A}{h}$

$$= 25 + \frac{81}{120} \times 10 = 31.75.$$

Scanned with CamScanner

Median :- In a group of n observations arranged in ascending (or) descending order of magnitude then the middle value is called median. It is denoted by Me.

Note :- when we calculate median,

if n is even then median is $\dfrac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

If n is odd then $\frac{n+1}{2}$

Scanned with CamScanner

2. Find the median of discrete data

x	1	2	3	4	5	6
f	7	12	17	19	21	24

x	f	cf
1	7	7
2	12	19
3	17	36
4	19	55
5	21	76
6	24	100
	$\Sigma f = 100$	

Median = size of $\left(\frac{n+1}{2}\right)^{th}$

$= \left(\frac{6+1}{2}\right)^{th} = (3.5)^{th}$ item

$= 4.$

Continous data:- $l + \frac{\frac{N}{2} - c}{f} \times h$ = Median

Where, N = Total frequency

l = lower limit of Median class

f = frequency of median class.

c = cumilative frequency of the class preceeding to the median class.

h = class size.

Scanned with CamScanner

1. Find the median wage of the following distribution

wages	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No. of workers	3	5	20	10	5

C.I	f	c.f
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-7000	5	43
	$\Sigma f = 43$	

$$\frac{N}{2} = \frac{43}{2} = 21.5$$

$$4000 + \frac{21.5 - 8}{20} \times 1000$$

$$4000 + \frac{13.5}{20} \times 1000$$

$$4000 + 675$$

$$= 4675$$

Scanned with CamScanner

mode:- The value of the variable for which the frequency is maximum is called mode (or) modal value. It is denoted by z (or) M.

1. Find the mode of series:

19, 17, 16, 19, 7, 9, 8, 9, 7, 9, 19, 9.

A) 7, 7, 9, 9, 9, 9, 19, 19, 19, 16, 17.

mode = 4.

find the mode of discrete date

x	8	7	10	14	22	80
f	1	3	9	17	14	5

mode = 17.

continous data:- $z = l + \frac{f-f_1}{2f-f_1-f_2} \times h$

l = lower limit of modal data.

f = frequency of modal data.

f_1 = frequency of the class preceeding to the modal data.

f_2 = frequency of the class succeding to the modal data

h = size of class

1. Find mode of the following data

class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	4	13	21	44	33	22	7

Scanned with CamScanner

Class	frequency	cf
0-10	4	4
10-20	13	17
20-30	21	38
30-40	44	82
40-50	33	115
50-60	22	137
60-70	7	144
	$\Sigma f = 144$	

$$\text{mode} = l + \frac{f - f_1}{2f - f_1 - f_2} \times h$$

$$= 30 + \frac{44 - 21}{2(44) - 21 - 33} \times 10$$

$$= 30 + \frac{23}{34} \times 10$$

$$= 30 + 6.76$$

$$\boxed{\text{mode} = 36.76.}$$

Find the mode of the following distributions.

class	frequency
0-10	5
10-20	8
20-30	7
30-40	12
40-50	28
50-60	20
60-70	10
70-80	10

$$\text{mode} = 40 + \frac{28 - 12}{2(28) - 12 - 20} \times 10$$

$$40 + \frac{16}{24} \times 10$$

$$40 + 6.6$$

$$= 46.6.$$

find mean, median, mode
~~An incomplete~~ frequency distribution is given as below.

Grade	frequency	midvalue	$f_i x_i$	cf
40-49	3	44.5	133.5	3
50-59	5	54.5	272.5	8
60-69	6	64.5	387	14
70-79	9	74.5	670.5	23
80-89	8	84.5	676	31
90-99	7	94.5	661.5	38
	$\Sigma f_i = 38$		$\Sigma f_i x_i = 2801$	

$$\frac{N}{2} = \frac{\cancel{38}^{19}}{\cancel{2}} = 19$$

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{2801}{38} = 73.71$$

$$\text{median} = 70 + \frac{19 - 14}{9} \times 9 = 70 + 5 = 75$$

mode $= 70 + \dfrac{9-6}{2(9)-6-8} \times 9$

$= 70 + \dfrac{3}{4} \times 9 = 70 + \dfrac{27}{4} = 70 + 6.75$

$= 76.75$

Mode = 3 median − 2 mean.

Geometric mean:- Geometric mean of a set of n observations is nth roots of their product.

ungrouped data $G.M = \text{Antilog}\left(\frac{1}{N} \sum \log x_i\right)$

Grouped data = $G.M = \text{Antilog}\left(\frac{\sum f \log m}{N}\right)$

Where M is mid value.

1. Daily income of 10 families of a particular place is given below. Find G.M

x	$\log x_i$
85	1.9294
70	1.8450
15	1.1760
75	1.8750
500	2.6989
8	0.9030
45	1.6532
250	2.3979
40	1.6020
60	1.7781
	$\sum \log x = 17.8585$

$N = 10.$

$G.M = \text{Antilog}\left(\frac{1}{10} \times 17.8585\right)$

$= \text{Antilog}(1.78585)$

$= 61.073 \quad [10^{1.78585}]$

Scanned with CamScanner

3. Calculate G.M for the following data

Marks	freq	M	log mi	filogmi
4-8	6	6	0.7781	4.6686
8-12	10	10	1	10
12-16	18	14	1.1461	20.6298
16-20	30	18	1.2552	37.656
20-24	15	22	1.3424	20.136
24-28	12	26	1.41497	16.97964
28-32	10	30	1.47712	14.7712
	101			

$G.M = \text{Antilog}\left(\frac{\Sigma f \log m}{N}\right)$

$\Sigma f \log mi = 124.84124$

$G.M = \text{Antilog}\left(\frac{124.84124}{101}\right)$

$= \text{Antilog}(1.236051)$

$= 17.22070.$

Harmonic mean:- Harmonic mean is the reciprocal of the Average of the reciprocal values.

For raw data, $H.M = \frac{n}{\Sigma \frac{1}{x}}$

For ungrouped data, $H.M = \frac{N}{\Sigma \frac{fi}{xi}}$

For grouped data, $H.M = \frac{N}{\Sigma fi/mi}$

1. Find the H.M of the following

125, 130, 75, 10, 45, 5, 0.5, 0.4, 500, 150

A) $H.M = \frac{1}{x}$ = 0.008, 0.0076, 0.013, 0.1, 0.022, 0.2, 2, 2.5, 0.002, 0.0066

$H.M = \frac{10}{4.8592} = 2.05795$

2. Calculate H.M for the following

Marks	students	fi/xi
10	20	2
20	30	1.5

Scanned with CamScanner

25	50	2
40	15	0.375
50	5	0.1
	N= 120	5.975

$H.M = \frac{120}{5.975}$

$= 20.0836.$

3.

class	frequency	mi	fi/mi
10-20	4	15	0.266
20-30	6	25	0.24
30-40	10	35	0.2857
40-50	7	45	0.1555
50-60	3	55	0.05454
	30		1.00174

$H.M = \frac{30}{1.00174} = 29.947$

Meausres of variability / dispersion:- The meausre of the scatterdness of the marks of figures in a series about an average is called Meausre of variation/dispersion.

Dispersion can be classified in to two categories:

1. The Meausres which express the spread of observations in terms of distance b/w the values of selected Observations. Eg:- range and Inter quartile range.

2. The meausre which express the spread of observations in terms of the average of deviations of observations from some central value. Eg:- Mean deviations and standard deviations

Range:- The range is the difference between largest and smallest value in the series. It

∴ Range = largest value - smallest value

$\boxed{R = L - S}$

Co-efficient of Range:- $\frac{L-S}{L+S}$.

Quartile deviation:- Quartile are those value which divide the frequency in to four equal parts when the values are arranged in the Asscending order of magnitude. The lower quartile (Q_1) is ~~lil~~ mid way between the lower extreme and the median.

The upper quartile (Q_3) is Midway b/w median and the upper extreme.

$Q_3 - Q_1$ is called interquartile Range.

$$Q_1 = l + \frac{\frac{N}{4} - c}{f} \times h, \quad Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times h$$

Co-efficient of Quartile deviation / semi inter quartile

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Semi inter quartile range $Q = \frac{Q_3 - Q_1}{2}$.

Mean deviation:- Mean deviation is defined as the arthimetic Average of the absolute deviations of a series computed from any one of the Meausres of central tendancy.

$$M.D(\bar{x}) = \frac{\Sigma|x-\bar{x}|}{n} \quad \text{where } \bar{x} = \frac{\Sigma x}{n}$$

ungrouped, $M.D(\bar{x}) = \frac{\Sigma f|x-\bar{x}|}{N}$ where $\bar{x} = \frac{\Sigma f_i x_i}{N}$

grouped, $M.D(\bar{x}) = \frac{\Sigma f|m_i - \bar{x}|}{N}$, $\bar{x} = \frac{\Sigma f_i m_i}{N}$

step deviation:-

$$M.D = A + \frac{\Sigma f_i u_i}{N} \times h \quad \text{Where, } u_i = \frac{x_i - A}{h}$$

Co-efficient of mean deviation:-

$$\frac{M.D(\bar{x})}{\bar{x}}$$

Standard deviation:- (S.D) This meausre of dispersion was represented by karl pearson in 1893. SD is the Positive square root of the A.M of the squares of the deviations of the given values.

$$\sigma = \sqrt{\frac{1}{N}\Sigma f_i|x-\bar{x}|^2}$$

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

co-efficient of standard deviation:- $\frac{\sigma}{\bar{x}}$

shortcut Method:- $\sigma_{S.D} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$

where, $d = x_i - A$

Step deviation:-

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

Variance:- $\sigma^2 = (S.D)^2 \Rightarrow \sigma^2 = \frac{1}{N}\Sigma f_i |x_i - \bar{x}|^2$.

1. Find the range and co-efficient of range for the following.

x	f
10	2
12	4
13	6
9	4
15	6
20	8

Largest x value = 20

Smallest x value = 9

$R = L - S = 20 - 9 = 11$.

co-efficients of Range $= \frac{L-S}{L+S} = \frac{20-9}{20+9} = \frac{11}{29}$

$= 0.3793$

2.

class	frequency
2.5-7.5	4
7.5-12.5	8
12.5-17.5	3

Largest x value = 17.5

Smallest x value = 2.5

$R = L - S = 17.5 - 2.5 = 15$

$C.R = \frac{17.5 - 2.5}{17.5 + 2.5} = \frac{\cancel{15}\,3}{\cancel{20}\,4} = 0.75$.

3. calculate median lower, upper quartiles, from the following distribution obtained by 49 students in a class. find semi-interquartile range and mode

class	frequency	cf
5-10	5	5
10-15	6	11
15-20	15	26
20-25	10	36
25-30	5	41
30-35	4	45
35-40	2	47
40-45	2	49

Given $N = 49 \Rightarrow \frac{N}{2} = 24.5$

$Me = l + \frac{\frac{N}{2} - c}{f} \times h$

$= 15 + \frac{24.5 - 11}{\cancel{15}\,3} \times \cancel{5}$

$= 15 + 4.5 = 19.5$.

Lower quartile $= Q_1 = l + \frac{\frac{N}{4} - c}{f} \times h$ $\frac{N}{4} = \frac{49}{4} = 12.25$

$= 15 + \frac{12.25 - 11}{\cancel{15}\,3} \times \cancel{5}$

$= 15 + \frac{1.25}{3} = 15 + 0.41 = 15.041$

upper Quartile $Q_3 = l + \frac{\frac{3N}{4} - c}{f} \times h$ $\quad \frac{3 \times 49}{4} = 36.75$

$= 25 + \frac{36.75 - 36}{5} \times 5$

$= 25 + 0.75 = 25.75.$

Semi interquartile $= \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(25.75 - 15.04)$

$= 5.35$

Mode:- $l + \frac{f - f_1}{2f - f_1 - f_2} \times h$ Highest frequency = 15

$15 + \frac{15-6}{2(15)-6-10} \times 5 = 15 + \frac{9}{14} \times 5$

$= 15 + 3.214$

$= 18.214.$

Find the mean deviation and co-efficient of mean deviation from the mean of the following data.

x 38 70 48 40 42 55 63 46 54 44

Mean deviation $(\bar{x}) = \frac{\Sigma|x - \bar{x}|}{n}$

$\bar{x} = \frac{\Sigma x}{n} = \frac{50\cancel{0}}{1\cancel{0}} = 50$

$= \frac{12 + 20 + 2 + 10 + 8 + 5 + 13 + 4 + 4 + 6}{10}$

$= 8.4.$

C-officient of M.D $= \frac{M.D(\bar{x})}{\bar{x}} = \frac{8.4}{50} = 0.168$

2. Find the M.D from median for the data 34, 66, 30, 38, 44, 50, 40, 60, 42, 51.

A) 30, 34, 38, 40, 42, 44, 50, 51, 60, 66

$\frac{42+44}{2} = 43 = me$

$MD(Me) = \frac{\Sigma|x_i - me|}{n} = \frac{13 + 9 + 5 + 3 + 1 + 1 + 7 + 8 + 17 + 23}{10}$

$= \frac{87}{10} = 8.7$

calculate co-efficient of variation of the following data.

Item	freq	di	di^2	$fidi$	$fidi^2$
10	4	-6	36	-24	144
12	6	-4	16	-20	80
14	10	-2	4	-20	40
A 16	14	0	0	0	0
18	9	2	4	18	36
20	4	4	16	16	64
22	2	6	36	12	72
	48			-18	436

$\sigma = \sqrt{}$ $\bar{x} = A + \frac{\Sigma fidi}{N}$

$= 16 - \frac{18}{48} = 16 - 0.375 = 15.625.$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{436}{48} - \left(\frac{-18}{48}\right)^2}$$

$$= \sqrt{9.0833 - \frac{324}{2304}}$$

$$= \sqrt{9.0833 - 0.140625}$$

$$= \sqrt{8.942675} = 2.9904$$

co-efficient of variation $= \frac{\sigma}{\bar{x}} \times 100$

$$= \frac{2.9904}{15.625} = 19.138$$

Scanned with CamScanner

Skewness:- skewness is a meausre of symmetric in a statistical distribution in which the curve appears bend (or) skewed either to the left or right

mean ≠ median ≠ mode.

Types of skewness:-

Positive skewness:- If the distribution curve is stretched to wards right we say that their is +ve skewness in the data.

Negative skewness:- If the distribution curve is streched to ~~wards~~ wards left we say that their is -ve skewness in the data.

Meausres of skewness:-

(i) Karl pearson's co-efficient of skewness

(ii) Bowley's

(iii) Kelly's

Karl Pearson's co-efficient of skewness:- is widely used method

* It is denoted by Skp

$$Skp = \frac{\bar{x} - z}{\sigma}$$

$$Skp = \frac{3(\bar{x} - Me)}{\sigma}$$

Problem:-

1. calculate the co-efficient of skewness from the following data.

size	30	40	50	60	70	80	90	100
frequency	7	10	14	35	102	136	43	8

Scanned with CamScanner

A)

size	freq	$d_i = \frac{x_i - A}{h}$	fd	fd^2
30	7	-4	-28	112
40	10	-3	-30	90
50	14	-2	-28	56
60	35	-1	-35	35
(70) A	102	0	0	0
80	136	1	136	136
90	43	2	86	172
100	8	3	24	2
	335		125	673

mean $\bar{x} = A + \frac{\Sigma f_i d_i}{N} \times h$

$= 70 + \frac{125}{335} \times 10 = 70 + 3.731$

$\bar{x} = 73.731$

Mode Z = Highest frequency = 136.

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

$\sigma = \sqrt{\frac{673}{335} - \left(\frac{125}{335}\right)^2} \times 10$

$\sigma = \sqrt{2.008 - \frac{15625}{112225}} \times 10$

$\sigma = \sqrt{2.008 - 0.1392} \times 10$

$\sigma = \sqrt{1.868} \times 10$

$\sigma = 1.3666 \times 10 \quad \therefore \sigma = 13.6$

Karl pearson's co-efficient of skewness

$Sk_p = \frac{\bar{x} - Z}{\sigma} = \frac{73.731 - 80}{13.6} = -0.487 \sim -0.4609$

Scanned with CamScanner

kurtosis:-

moments:- moments are a set of statictical parame
to meausre a distribution

moments about mean:-

$d = x - A$

$$\mu_1' = \frac{\Sigma fd}{N}$$

$$\mu_2' = \frac{\Sigma fd^2}{N}$$

$$\mu_3' = \frac{\Sigma fd^3}{N}$$

$$\mu_4' = \frac{\Sigma fd^4}{N}$$

$d = \frac{x - A}{h}$

$$= \frac{\Sigma fd}{N} \times h$$

$$= \frac{\Sigma fd^2}{N} \times h^2$$

$$= \frac{\Sigma fd^3}{N} \times h^3$$

$$= \frac{\Sigma fd^4}{N} \times h^4$$

first moment:-

$$\mu_1 = \mu_1' - \mu_1' = 0$$

second moment: $\mu_2 = \mu_2' - \mu_1'^2$

third moment: $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

fourth moment: $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'$

Moment ratios:- Ratios in between moments are called moment ratio. we can meausre sckewness and kurtosis of the distribution

Skweness based on moments:-

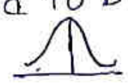
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}.$$

If $\beta_1 = 0$, then the distribution is symmetric.
If $\beta_1 > 0$, then the distribution is positively skewed.
If $\beta_1 < 0$, then the distribution is negatively skewed

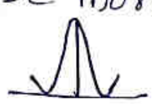
kurtosis:- kurtosis explain about the shape of a freaquency distribution.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

If $\beta_2 = 3$ then the distribution is said to be normal and the curve is Mesokurtic.



If $\beta_2 > 3$ then the distribution is said to be more peaked and the curve is lepto kurtic.



Scanned with CamScanner

If $\beta_2 < 3$ then the distribution is Platy kurtic 

1. calculate the first four moments of the following distributions about the mean and hence find β_1 and β_2.

x	f	$d = x - A$	fd	fd^2	fd^3	fd^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	0	512	0	2816

$\mu_1' = \frac{\Sigma fd}{N} \times h$

$\mu_2' = \frac{\Sigma fd^2}{N} \times h^2$

$\mu_3' = \frac{\Sigma fd^3}{N} \times h^3$

$\mu_4' = \frac{\Sigma fd^4}{N} \times h^4$

If we get big values then we add $d = \frac{x-A}{h} \times h$ else $d = x - A$

Moment about $x = 4$

$\mu_1' = \frac{\Sigma fd}{N} = 0$

$\mu_2' = \frac{\Sigma fd^2}{N} = \frac{512}{256} = 2$

$\mu_3' = \frac{\Sigma fd^3}{N} = 0$

$\mu_4' = \frac{\Sigma fd^4}{N} = \frac{2816}{256} = 11$

Moments about the mean:-

$\mu_1 = \mu_1' - \mu_1' = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 2$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$

Moment fact, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$, $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$

Scanned with CamScanner

the first four central moments of a distribution are 0, 2.5, 0.7, 18.75. Examine the kurtosis of the distribution.

A) $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$, $\mu_4 = 18.75$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0.49}{15.625} = 0.03136.$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.$$

Correlation :- correlation is a statistical tool used to meausre the relationship between two sets of variables and express each in a precise manner.

An Analysis of the covariance of two (or) more variable is usually called correlation.

Types of correlation :-

correlation

- on the basis of direction
  - -ve correlation
  - +ve correlation
- on the basis of no. of variables
  - simple correlation
  - multiple correlation
  - Partial correlation
- on the basis of ratio of change of direction
  - linear correlation
  - non-linear correlation

Negative correlation (Inverse correlation) :- Two variables are said to be correlated when both the variables vary in opposite direction.

Eg :- price and demand.

Positive correlation (Direct correlation) :- Two variables are said to be positively correlated when both the variables vary in the same direction

Eg :- Demand and supply

simple correlation :- It is a meausre used to determine the relationship between two variables.

Eg :- price & Demand, Demand and supply.

Multiple co-rrelation :- It is a meausre used to determine the relationship among several variables.

Eg :- rainfall, temperature, yield of crops.

Scanned with CamScanner

Partial correlation:- The study of variables, excluding some other variable is called partial correlation.

Eg:- Relation between study of two variables price and demand eliminating supply.

linear correlation:- If the ratio of change between two variables is uniform then there can be linear co-rrelation between them. such variables are plotted on a graph paper We get spreo straight line.

Non-linear relation:- (curvilinear):- The amount of change in one variable doesn't bear a constant ratio to the amount of change in the other variable. Then correlation is said to be curvilinear. If such variables plotted on a graph, the points would fall on a curve.

Karl pearson co-efficient of correlation:-

.When deviation is taken from A.M. the formula for coefficient of correlation is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \quad \text{where } x = x-\bar{x},\ y = y-\bar{y}$$

$$= \frac{\Sigma xy}{n\sigma_x \sigma_y} \quad \text{where, } \sigma_x = \text{S.D of } x\text{-series},\ \sigma_y = \text{S.D of } y\text{-series},\ n = \text{no. of observations}$$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\cdot\sqrt{\Sigma(y-\bar{y})^2}}$$

Scanned with CamScanner

A Psycological test of intelligence and of engineering ability were applied to 10 students. Here is a record of data showing intelligence ratio (Ir). and engineering ratio (ER). calculate the co-efficient of correlation.

Student	Intelligence ratio	ER	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
A	105	101	6	36	3	9	18
B	104	103	5	25	5	25	25
C	102	100	3	9	2	4	6
D	101	98	2	4	0	0	0
E	100	95	1	1	-3	9	-3
F	99	96	0	0	-2	4	0
G	98	104	-1	1	6	36	-6
H	96	92	-3	9	-6	36	-108
I	93	97	-6	36	-1	1	6
J	92	94	-7	49	-4	16	-112
	990	980		170		140	92

$$\bar{X} = \frac{\Sigma x}{n} = \frac{990}{10} = 99 \qquad \bar{Y} = \frac{\Sigma y}{n} = \frac{980}{10} = 98$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{92}{\sqrt{170}\sqrt{140}} = 0.596$$

When deviations are taken from Assumed mean

$$r = \frac{n\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{n\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{n\Sigma d_y^2 - (\Sigma d_y)^2}} \qquad \begin{array}{l} d_x = X - A_x \\ d_y = Y - A_y \end{array}$$

calculate co-efficient of correlation in the following

Height of father (x)	Height of son (y)	$d_x = x - 67$	$d_y = y - 68$	$d_x d_y$	d_x^2	d_y^2
65	67	-2	-1	2	4	1
66	68	-1	0	0	1	0
67	64	0	-4	0	0	16
A 67	A 68	0	0	0	0	0
68	72	1	4	4	1	16
69	70	2	2	4	4	4
71	69	4	1	4	16	1
73	70	6	2	12	36	4
		10	4	26	62	42

Scanned with CamScanner

$$r = \frac{8 \times 26 - 26}{\sqrt{8 \times 62 - \cancel{3844}_{160}} \times \sqrt{8 \times 42 - \cancel{1764}_{16}}}$$

$$= 0.471$$

Total sales turn over and net profit of 7 medium sized companies. calculate the karl pearson correlation

sales turn over	Net profit	$dx = x - 400$	$dy = y - 80$	$dxdy$	dx^2	dy^2
100	30	-300	-50	15000	90000	2500
200	50	-200	-30	6000	40000	900
300	60	-100	-20	2000	10000	400
400 A	A 80	0	0	0	0	0
500	100	100	20	2000	10000	400
600	110	200	30	6000	40000	900
700	130	300	50	15000	90000	2500
		0	0	46000	280000	7600

$$r = \frac{7 \times 46000 - \cancel{46000}\,0}{\sqrt{7 \times 280000 - 0}\ \sqrt{7 \times 7600 - 0}} = \frac{322000}{1400 \times 230.65}$$

$$= \frac{322000}{322911.75} = 0.9971.$$

variance - covariance method :- when co-variance and variance are given then

$$r = \frac{cov(xy)}{\sqrt{var(x)}\sqrt{var(y)}}$$

If covariance between x and y variables is 12.5 and variance of x and y are 16.4 and 13.8 respectively find the co-efficient of correlation between them.

A) $cov(x,y) = 12.5$ $var(x) = 16.4$ $var(y) = 13.8$

$$= \frac{12.5}{\sqrt{16.4 \times 13.8}} = 0.830$$

Spearman's rank correlation.

Charles edward spearman found the method of finding the coefficient of correlation by ranks. This method is useful in dealing with Qualitative characteristics such as character intelligence and beauty. The value of Rank correlation, coefficient always lies between -1 and 1.

Formula for rank correlation is

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

where, d = difference between two rands $(x-y)$

ρ = rank co-efficient of correlation.

n = no.of pair of observations.

1. when ranks are not given,

i) random sample of 5 college students selected and their grades in Mathematics and statistics are found to be. calculate spearmen's rank correlation coefficient

Here n=5

given data

Mathematics (x)	Rank (x)	statistiss (y)	Rank (y)
85	2	93	1
60	4	75	3
73	3	65	4
40	5	50	5
90	1	80	2

$d = x - y$ d^2

1   1   -1   0   -1   1   1   0   1 = 4

Scanned with CamScanner

spearman's rank correlation is $\rho = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$

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

$= 1 - 0.2 = 0.8.$

ii) calculate the co-efficient of correlation by rank Method

X	83	88	95	70	60	90	81	50
Y	120	134	130	115	110	140	142	100

X	Ran(x)	Y	Rank(y)	d = x - y	d^2
83	4	120	5	-1	1
88	3	134	4	-1	1
95	1	130	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
90	2	140	3	-1	1
81	5	142	2	3	9
50	8	100	8	0	0
					12

$\rho = 1 - \frac{6 \times 12}{8(8^2-1)} = 1 - \frac{72}{504} = 1 - 0.14 = 0.86$

Scanned with CamScanner

3) When ranks are repeated (or) equal :-

$$\rho = 1 - \frac{6\left[\sum d^2 + CF\right]}{n(n^2-1)}$$

where, $CF = \frac{\sum m^3 - m}{12}$

m = no. of times an item is repeated

Scanned with CamScanner

A sample of 12 fathers and their elder sons gave the following data about their elderson's calculate of rank correlation.

Father	Rank x	son	Rank y	$d=x-y$	d^2
65	9	68	5.5	3.5	12.25
63	11	66	9.5	1.5	2.25

Scanned with CamScanner

				$d=x-x$	
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
					72.5

$d = x - y$

In x-series, 67 & 68 repeated twice

$m = 2, 2. \quad C.F = \sum \frac{m^3 - m}{12} = \frac{2^3-2}{12} + \frac{2^3-2}{12}$

$= \frac{8-2+8-2}{12} = 1.$

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

$m = 4, 2, 2, \quad C.F = \frac{4^3-4}{12} + \frac{2^3-2}{12} + \frac{2^3-2}{12}$

$= \frac{60}{12} + 1 = 6.$

$C.F = 6 + 1 = 7.$

$\rho = 1 - \frac{6[\sum d^2 + C.F]}{n(n^2-1)} = 1 - \frac{6[72.5+6]}{12(12^2-1)}$

$= 1 - \frac{471}{1716}$ ~~$= 1 - 0.0458$~~

$= 1 - 0.274$ $= 0.9542.$

$= 0.726.$

Find the rank correlation ... of the following

Regression:- The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression.

Lines of regression:- The line described in the average relationship between two variables is known as line of regression.

Scanned with CamScanner

Regression line of y on x is $y-\bar{y} = b_{yx}(x-\bar{x})$

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

Here, $\bar{x} = \frac{\Sigma x}{n}$, $\bar{y} = \frac{\Sigma y}{n}$

$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$ (or) $\frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$

(or) $\frac{cov(xy)}{\sigma x}$

(or) $r\frac{\sigma x}{\sigma y}$

Assumed mean $b_{yx} = \frac{n\Sigma dx dy - \Sigma dx \Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2}$

$b_{xy} = \frac{n\Sigma dx dy - \Sigma dx\Sigma dy}{n\Sigma dy^2 - (\Sigma dy)^2}$

$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$ (or) $\frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$

(or) $\frac{cov(xy)}{\sigma y}$

(or) $r\frac{\sigma x}{\sigma y}$

formula

(The geometric mean b/w the regression coefficient $r_{xy} = \pm\sqrt{b_{xy} \cdot b_{yx}}$) for the following data for following regression equations

X	Y	XY	X^2	Y^2
1	15	15	1	225
2	25	50	4	625
3	35	105	9	1225
4	45	180	16	2025
5	55	275	25	3025
			55	7125

Scanned with CamScanner

$$b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5(625) - 15(175)}{5(55) - 15^2} = 10$$

$$b_{xy} = 0.1$$

Reg of X on Y is, $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$X - 3 = 0.1(Y - 35)$$

$$X - 3 = 0.1Y - 3.5$$

$$X = 0.1Y - 0.5$$

Reg of Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 35 = 10(X - 3) \qquad Y = 10X - 30 + 35$$

$$Y - 35 = 10X - 30 \qquad Y = 10X + 5$$

Using the following bivariant data i) find the two regression lines. ii) estimate x when $y = 7$. iii) estimate y when $x = 4$. iv) calculate r_{xy}

A)

X	Y	XY	X^2	Y^2
1	6	6	1	36
5	1	5	25	1
3	0	0	9	0
2	0	0	4	0
1	1	1	1	1
2	2	4	4	4
7	1	7	49	1
3	5	15	9	25
$\sum x = 24$	$\sum y = 16$	$\sum XY = 38$	102	68

$$\bar{X} = \frac{\sum X}{n} = \frac{24}{8} = 3, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{16}{8} = 2.$$

$$b_{xy} = \frac{n\sum XY - \sum X \sum Y}{n\sum Y^2 - (\sum Y)^2} = \frac{8 \times 38 - 24 \times 16}{8 \times 68 - (16)^2} = \frac{-80}{288}$$

$$= -0.278$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{8 \times 38 - 24 \times 16}{8 \times 102 - (24)^2}$$

$b_{yx} = -0.33.$

Regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$

$$x - 3 = -0.278\,(y - 2)$$

$$x = -0.278\,y + 3.556$$

Regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$y - 2 = -0.33\,(x - 3)$$

$$y = -0.33x + 2.99$$

ii) x when y = 7.

$$x = -0.278\,(7) + 3.556 = 1.61$$

iii) y when x = 4.

$$y = -0.33 \times 4 + 2.99 = 1.67$$

iv) $r_{xy} = \pm\sqrt{b_{xy} \cdot b_{yx}} = \sqrt{(-0.278)\,(-0.33)}$

$r_{xy} = -0.3029$ [$\because$ both the regression co-efficients b_{xy} and b_{yx} are -ve].

Scanned with CamScanner

from the following data write down two regression equations estimate the marks in x when $y = 70$.

	mean	S.D	
X	48.4	8.4	$r = 0.62$
Y	35.6	10.5	

A) Average, $\bar{X} = 48.4$ $\bar{Y} = 35.6$

$r = 0.62$

S.D, $\sigma_X = 8.4$, $\sigma_Y = 10.5$

Regression line of X on Y.

$$X - \bar{X} = 0.62 \times \frac{8.4}{10.5} (Y - \bar{Y})$$

$$X - 48.4 = 0.496 (Y - 35.6)$$

$$X - 48.4 = 0.496Y - 17.6576$$

$$X = 0.496Y - 17.6576 + 48.4$$

$$X = 0.496Y + 30.7424$$

when $y = 70$

$$X = 34.72 + 30.7424 = 65.4624$$

Scanned with CamScanner

Regression line of Y on X, $Y-\bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})$

$$Y-35.6 = 0.62 \times \frac{10.5}{8.4}(X-48.4)$$

$$Y-35.6 = 0.775(X-48.4)$$

$$Y = 0.775X - 37.5 + 35.6$$

$$\boxed{Y = 0.775X - 1.91}$$

Angle between two regression lines:-

Let θ be the angle between the regression lines

Regression line of Y on X is

$$Y-\bar{Y} = b_{YX}(X-\bar{X}) = r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})$$

Regression line of X on Y is

$$X-\bar{X} = b_{XY}(Y-\bar{Y}) = r\frac{\sigma_X}{\sigma_Y}(Y-\bar{Y})$$

Then, $\tan\theta = \left(\frac{1-r^2}{r}\right)\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$, θ is acute

$= \left(\frac{r^2-1}{r}\right)\frac{\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}$, θ is obtuse.

Note:- If $r=0$, $\tan\theta=\infty \Rightarrow \theta=\frac{\pi}{2}$

Hence, there is no relation between the two variables. i.e) they are independent.

If $r=\pm 1$, then $\tan\theta=0 \Rightarrow \theta=0$ or π

then the two regression lines are parallel (or) coincident.

1. If θ is the angle between two regression lines and S.D of y is twice the S.D of x and $r=0.25$. Then find $\tan\theta$.

A) Given, $r=0.25$

$$\sigma_y = 2\sigma_x$$

θ be the angle between two regression line,

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

$$= \left[\frac{1-(0.25)^2}{0.25}\right] \times \frac{2\sigma_x^2}{\sigma_x^2+(2\sigma_x)^2}$$

$$= 3.75 \times \frac{2\sigma_x^2}{\sigma_x^2+4\sigma_x^2}$$

$$= 3.75 \times \frac{2\cancel{\sigma_x^2}}{5\cancel{\sigma_x^2}}$$

$$= 3.75 \times 0.4 = 1.5$$

4) Test whether the equations $2x+3y=4$ and $x-y=5$ represent valid regression lines.

A) Let the regression line of X on Y is $2x+3y=4$

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

Let the regression line of Y on X is $x-y=5$

$$y = x-5 \quad \text{(2)}$$

comparing the lines with

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

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

i.e; $$r\frac{\sigma x}{\sigma y} = \frac{-3}{2} \quad \text{(3)}$$

$$r\frac{\sigma y}{\sigma x} = 1 \rightarrow \text{(4)}$$

(3) × (4) ⇒ $$r\frac{\cancel{\sigma x}}{\cancel{\sigma y}} \times r\frac{\cancel{\sigma y}}{\cancel{\sigma x}} = \frac{-3}{2}$$

$$r^2 = \frac{-3}{2}.$$

since, $-1 \le r \le 1$

∴ This two regression lines are valid equations

5) If $x = 2y+3$, $y = kx+6$ are the regression lines of x on y and y on x respectively. S.T ① $0 \le k \le \frac{1}{2}$

② If $k = \frac{1}{8}$ find r & $(\bar{x}, \bar{y})$.

A) Let the regression line of x on y is

$$x = 2y+3$$

Let the regression line of y on x is

$$y = kx+6.$$

$$b_{xy} = 2 \qquad b_{yx} = k$$

$$\therefore r = \pm\sqrt{b_{xy} \cdot b_{yx}} = \pm\sqrt{2k}$$

S.O.B.S

$$r^2 = 2k$$

$$-1 \le r \le 1$$

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

$$0 = k \Rightarrow k = 0$$

$$(1)^2 = 2k \Rightarrow k = \frac{1}{2}$$

$$\therefore 0 \le k \le \frac{1}{2}.$$

② $r^2 = 2k$

$r^2 = \cancel{2} \times \frac{1}{\cancel{8}4}$ when $k = \frac{1}{8}$

$$r^2 = \frac{1}{4}$$

$$r = \pm\frac{1}{2} = \frac{+1}{-2}.$$

Let $(\bar{x}, \bar{y})$ is passing through the equations then

$$\bar{x} = 2\bar{y}+3 \quad \text{—①}$$

$$\bar{y} = k\bar{x}+6 \Rightarrow \frac{\bar{x}}{8}+6. \quad ②$$

Solving ① & ②

$$\cancel{\bar{x}} - 2\bar{y} - 3 = 0$$

$$②\times 8 \quad \cancel{\bar{x}} - 8\bar{y} + 48 = 0$$

$$\underline{- \quad + \quad -}$$

$$6\bar{y} - 51 = 0$$

$$\bar{y} = \frac{51}{6} = 8.5$$

$$\bar{x} = 2 \times 8.5 + 3$$

$$= 17+3$$

$$\bar{x} = 20$$

Scanned with CamScanner

Method of least squares:— Method of least square is a device for finding the equation of a specified type of curve, which best fits for a given set of observations. This method ~~defines~~ depends on principle of least square.

"The sum of squares of difference between the observed and corresponding estimated values should be the minimum."

Fitting of a straight line:-

Let $x = A + By$ be the line of X on Y find a, b values with the following two normal equations to be solved. The normal equations are

$$\Sigma x = na + b\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Let $Y = A + BX$, be then the normal equations are

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

1. Consider the following data to obtain the two regression equations find a and b

X	6	2	10	4	8
Y	9	11	5	8	7

Scanned with CamScanner

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
Σx 30	$\Sigma y = 40$	220	340	214

Here, n=5

~~Normal~~ equations of a straight line $y = a + bx$ —

Normal equ's are $\Sigma y = na + b\Sigma x$ — ①

$\Sigma xy = a\Sigma x + b\Sigma x^2$ — ②

$214 = a(30) + b(220)$ — ③

$40 = 5a + 30b.$ — ④

solving ③ & ④

$a = 11.9 \quad b = -0.65.$

sub $y = a + bx$

$y = 11.9 - 0.65x$

$0.65x + y = 11.9.$

equation of a straight line. $x = a + by$ — ①

Normal equ are $\Sigma x = na + b\Sigma y$

$\Sigma xy = a\Sigma y + b\Sigma y^2$

$214 = 40a + b(340)$ — ②

$30 = 5a + b40$

$a = 16.4 \quad b = -1.3$

~~16.4 = a - b1.3~~ $x = 16.4 - 1.3y$

1. find the equation of regression line for y on x for the following data. Also estimate y if x=75

A)

x	y	x^2	xy
65	68	4225	4420
63	66	3969	4158
67	68	4489	4556
64	65	4096	4160
68	69	4624	4692

Scanned with CamScanner

62	66	3844	~~4~~ 4092
70	68	4900	4760
66	65	4356	4290
68	71	4624	4828
67	67	4489	4489
660	673	43616	44445

Eq. of a st line in $y = a + bx$

The normal eqns are $\Sigma y = na + b\Sigma x$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

$$673 = 10a + 660b$$

$$44445 = a\,660 + b\,43616$$

$$a = 35.48 \quad b = 0.4821$$

$$y = 35.48 + 0.4821x$$

If $x = 75$ $y = 35.48 + 0.4821 \times 75$

$$\boxed{y = 71.6375}$$

3) From a sample of 200 pairs of observation the following quantities were calculated $\Sigma x = 11.34$, $\Sigma y = 20.78$, $\Sigma x^2 = 12.16$, $\Sigma y^2 = 84.96$, $\Sigma xy = 22.13$. From the above data show how to compute the co-efficients of the equation $y = a + bx$.

A) $y = a + bx$

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

$$20.78 = 200a + 11.34b$$

$$22.13 = 11.34a + b\,12.16$$

$$a = 7.513 \times 10^{-4} \quad b = 1.82$$

$$= 0.00075$$

Scanned with CamScanner

4. Determine the equation of a st. line which best fits the Data

$a = 0.82 \quad b = 1.\ldots$

x	y	x^2	y^2	xy
10	10	100	100	100
12	22	144	484	264
13	24	169	576	312
16	27	256	729	432
17	29	289	841	493
20	33	400	1089	660
25	37	625	1369	925
113	182	1983	5188	3186

Eq of a st line in $y = a + bx$

The normal equations are $\Sigma y = na + b\Sigma x$

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$3186 = a113 + b1983$

$182 = 7a + b113$

$a = 0.799 \quad b = 1.56.$

Eq of a straight line in $x = a + by$

Normal equation are $\Sigma x = a\,\cancel{n} + b\Sigma y$

$\Sigma xy = a\Sigma y + b\Sigma y^2$

$3186 = a182 + b5188$

$113 = 7a + b182$

$a = 2.00 \quad b = 0.54$

Scanned with CamScanner