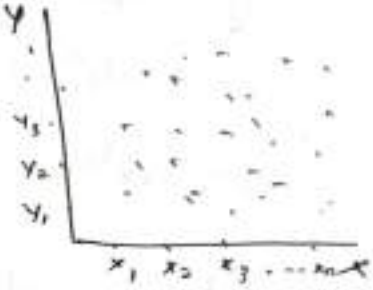
Curve fitting :- The method of finding an equation of the curve that approximates the given set of 'n' data pts is called as curve fitting.

Scattered diagram :- To find the mathematical relationship betwee the two variables 'x' & 'y' plot the given set of 'n' values (x_1, y_1) (x_2, y_2) ..... (x_n, y_n) in xy plane, then the resulting set of pts forms a 'scattered' diagram'



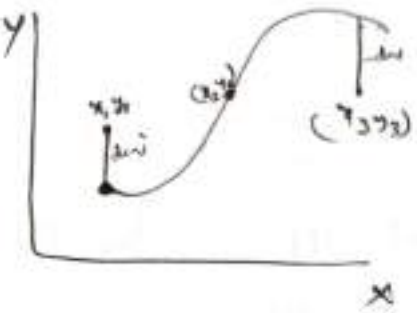
Method of least squares :-

The procedure of finding an eqn of a curve or a st. line that best fits the ~~pts~~ given set of 'n' data pts is called as method of least squares.

Let $(x_1,y_1)(x_2,y_2)$ ..... (x_n,y_n) be the set of 'n' data pts satisfying the eqn $y = f(x) = a_0 + a_1 x$, then $d_i = (y_i - (a_0 + a_1 x_i))$ denotes the diff b/w observed & expected ~~for~~ values.

These d_i's are called as 'deviations" or "errors" or residuals" and it may be +ve or -ve or zero.

The method of least squares criteria (or) principle states that of all the curves, approximating the given set of data, the curve having least or min. sum of the squares of the deviations is called as "best fitting curve or "least squares curve".



$y = a_0 + a_1 x$ → st. line.

$y = a_0 + a_1 x + a_2 x^2$ — 2nd degree polynomial or parabola.

$y = ae^{bx}$
$= ab^x$ } exponential curve.

$y = ax^b$ — Geometric curve.

Normal eqns of a st. line :- Let $y = f(x) = a_0 + a_1 x$ be a st. line then $y_i = f(x_i) = a_0 + a_1 x_i$ represents the family of st. lines.

$$d_i = (y_i - (a_0 + a_1 x_i))$$

$$\Sigma d_i^2 = \Sigma (y_i - a_0 - a_1 x_i)^2 \quad \text{—(i)}$$

By the method of least squares, sum of the squares of deviations should be min.

To find min, diff partially eq (i) w.r.t a_0

$$\frac{\partial}{\partial a_0} \Sigma d_i^2 = 2 \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))(-1)$$

$$0 = -2 \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a_0 + a_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow \Sigma y_i = n a_0 + a_1 \Sigma x_i \quad \text{—(A)}$$

diff (i) partially w.r.t a_1.

$$\frac{\partial \Sigma d_i^2}{\partial a_1} = \frac{\partial}{\partial a_1} \Sigma (y_i - (a_0 + a_1 x_i))^2.$$

$0 = 2\Sigma(y_i - (a_0 + a_1 x_i)(-x_i))$

$\Rightarrow \Sigma(x_i y_i - (a_0 x_i + a_1 x_i^2)) = 0$

$\rightarrow \Sigma x_i y_i = a_0 \Sigma x_i + a_1 \Sigma x_i^2$ ... (B).

Eq (A) & (B) are called normal eqns of a st line.

Normal eqns of a parabola.

$y = a_0 + a_1 x + a_2 x^2$.

$d_i = (y_i - (a_0 + a_1 x_i + a_2 x_i^2))$.

$\Sigma d_i^2 = \Sigma (y_i - (a_0 + a_1 x_i + a_2 x_i^2))^2$ — (1).

Diff (1) w.r.t a_0 partially v get.

$\Sigma y_i = n a_0 + a_1 \Sigma x_i + a_2 \Sigma x_i^2$ — (A)

Diff (1) partially w.r a_1, v get

$\Sigma x_i y_i = a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3$ — (B)

Diff (1) partially w.r a_2, v get

$\Sigma x_i^2 y_i = a_0 \Sigma x_i^2 + a_1 \Sigma x_i^3 + a_2 \Sigma x_i^4$ — (C).

(A), (B), (C) represents normal eqns of parabola.

Normal eqns of exponential curve :-

$y = a e^{bx}$ — exponential curve.

Apply $\log_e$ b.s.

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

$\Rightarrow Y = A + bx$. is a st line.

where $Y = \log_e y$, $A = \log_e a$

& normal eqns are $\Sigma Y_i = nA + b\Sigma x_i$

$\Sigma x_i Y_i = A\Sigma x_i + b\Sigma x_i^2$.

1). The following data pertain to the no. of jobs per day and the cpu system time reqd.

No. of jobs: (x_i)	1	2	3	4	5
Cpu time (y_i)	2	5	4	9	10.

Fit a st. line, estimate the mean cpu time at $x=35$.

Sol) Normal eqns of st. line are

$$\Sigma y_i = na_0 + a_1 \Sigma x_i$$
$$\Sigma x_i y_i = a_0 \Sigma x_i + a_1 \Sigma x_i^2$$

$$30 = 5a_0 + 15a_1 \quad \text{—} (1)$$
$$110 = 15a_0 + 55a_1 \quad \text{—} (2).$$

Solving (1) & (2) we get

$$a_1 = 2, \quad a_0 = 0.$$

∴ The reqd st line to be fitted is $y = 0 + 2x$.

$$y(x) = 2x.$$
$$y(3.5) = 2(3.5) = 7.$$

Note:- Whenever 'x' & 'y' are very large & equidistant replace 'x' by $X = \frac{x - \text{middle term}}{h}$ if n is odd.

$$X = \frac{x - \text{mean of middle terms}}{h/2} \quad \text{if } n \text{ is even.}$$

2) The following are the measurements of the air velocity & the evaporation coefficient of burning fuel droplets in air impulse engine.

Air vel (x)	20	60	100	140	180	220	260	300	340	3
Evaporation coeff (y)	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1

x	y	X	xy	x^2
0	0.18	-9	-162	81
60	0.37	-7	-2.59	49
100	0.35	-5	-1.75	25
140	0.78	-3	-2.34	9
180	0.56	-1	-0.56	1
220	0.75	1	0.75	91
260	1.18	3	3.54	9
300	1.36	5	6.8	25
340	1.17	7	8.19	49
380	1.65	9	14.85	81
	$\Sigma y_i = 8.35$	$\Sigma X_i = 0$	$\Sigma x_i y_i = 25.27$	$\Sigma X_i^2 = 330$

$$\Sigma y_i = n a_0 + a_1 \Sigma X_i$$

$$\Sigma x_i y_i = a_0 \Sigma X_i + a_1 \Sigma X_i^2$$

$$8.35 = 10 a_0 + a_1(0)$$

$$25.27 = 0(a_0) + 330 a_1$$

$$a_0 = 0.835$$

$$a_1 = 0.0765$$

The reqd st. line to be fitted is

$$y = (0.835) + (0.0765) X.$$

$$y = 0.835 + (0.0765)\left(\frac{x - 200}{20}\right)$$

$$y = 0.069 + (0.003825)x.$$

3). Fit a parabola $y = a + bx + cx^2$

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Sol). Normal eqns of a parabola are

$$\Sigma y_i = n a_0 + a_1 \Sigma x_i + a_2 \Sigma x_i^2$$

$$\Sigma x_i y_i = a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3 .$$

$$\Sigma x_i^2 y_i = a_0 \Sigma x_i^2 + a_1 \Sigma x_i^3 + a_3 \Sigma x_i^4 .$$

x_i	y_i	$x_i y_i$	x_i^2	x_i^3	x_i^4	$x_i^2 y_i$
1	2	2	1	1	1	2
2	6	12	4	8	16	24
3	7	21	9	27	81	63
4	8	32	16	64	256	128
5	10	50	25	125	625	250
6	11	66	36	216	1296	396
7	11	77	49	343	2401	539
8	10	80	64	512	4096	640
9	9	81	81	729	6561	729 .
45	74	421	285	2025	15333	2771

on solving we get $a_0 = -0.928$, $a_1 = 3.523$, $a_2 = -0.267$.

$$y = -0.928 + 3.523x - 0.267x^2 .$$

4). Fit a curve of the form $y = ae^{bx}$ for the foll. data

x	0.0	0.5	1.0	1.5	2.0	2.5
y	0.1	0.45	2.15	9.15	40.35	180.75

$y = ae^{bx}$

Applying $\log_e$ on b.s.

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

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

$Y = A + bx$.

where $Y = \log_e y$ & $A = \log_e a$.

Then normal eqns are

$$\Sigma Y_i = nA + b\Sigma x_i$$

$$\Sigma x_i Y_i = A\Sigma x_i + b\Sigma x_i^2$$

x_i	$Y_i = \log_e y$	x_i^2	$x_i Y_i$
0	-2.302	0	0
0.5	-0.798	0.25	-0.399
1	0.765	1	0.765
1.5	2.213	2.25	3.3195
2	3.697	4	7.394
2.5	5.197	6.25	12.992
7.5	8.76	13.75	24.035

$8.76 = 6A + 7.5b$.

$24.035 = 7.5A + 13.75b$.

$A = -2.278$, $b = 2.9908$.

$Y = -2.278 + (2.9908)x$

$a = \text{Antilog}(A) = \text{Antilog}(-2.27)$. $= 0.103$.

∴ The reqd exponential curve is

$y = ae^{bx} = (0.103)e^{2.99x}$.

Geometrical curve:- $y = ax^b$.

Applying $\log_{10}$ on b.s.

$$\log_{10} y = \log_{10} a + b \log_{10} x.$$

$$Y = A + bX.$$

where $Y = \log_{10} y$, $A = \log_{10} a$, $X = \log_{10} x$.

$$\Sigma y_i = nA + b\Sigma x_i$$

$$\Sigma x_i y_i = A\Sigma x_i + b\Sigma x_i^2$$

4)

x	1	2	3	4	5
y	0.5	0.2	4.5	8	12.5.

Sol $y = ax^b$

$$\log_{10} y = \log_{10} a + b \log_{10} x.$$

$$Y = A + bX.$$

& the normal eqns are

$$\Sigma Y_i = nA + b\Sigma X_i$$

$$\Sigma X_i Y_i = A\Sigma X_i + b\Sigma X_i^2$$

$X_i = \log_{10} x_i$	$Y_i = \log_{10} y_i$	$X_i Y_i$	X_i^2
0	-0.3010	0	0
0.3010	-0.698	-0.210	0.0906
0.477	0.653	0.311	0.2275
0.6020	0.9030	0.543	0.3624.
0.698	1.096	0.765	0.4872
2.078	1.653	1.409	1.167.

$$1.653 = 5A + 2.078b.$$

$$1.409 = 2.078A + 1.167b.$$

$$A = -0.658, \quad b = 2.379.$$

$= 0.546$.

$$y = -0.658 + (2.379)x .$$

$$a = 0.2197 .$$

$$y = (0.2197)x^{(2.379)} .$$

5). Fit the model $y = ax^b$ to the following data.

x	1	2	3	4	5	6
y	2.96	4.26	5.21	6.10	680	7.50 .

Sol $y = ax^b$

$$\log_{10} y = \log_{10} a + b \log_{10} x .$$

$$Y = A + bX .$$

$$\Sigma Y_i = nA + b\Sigma X_i$$

$$\Sigma X_i Y_i = A\Sigma X_i + b\Sigma X_i^2 .$$

$X_i = \log_{10} x_i$	$Y_i = \log_{10} y_i$	$X_i Y_i$	X_i^2
0	0.4742	0	0
0.3010	0.6294	0.1894	0.0906
0.4771	0.7168	0.3419	0.2276
0.6020	0.7853	0.4727	0.3624
0.698	0.8325	0.581	0.4872
0.778	0.8750	0.6807	0.6052
2.8561	4.3132	2.2657	1.773 .

$$4.3132 = 6A + 2.8561b .$$

$$2.2657 = 2.8561A + 1.773b .$$

$$A = 0.4741 \Rightarrow a = 2.979 .$$

$$b = 0.514 .$$

$$y = 0.4741 + (0.514)x .$$

Correlation:- If a change in one variable results a change in another variable, then those two variables are said to be correlated.

If the 2 random variables 'x' & 'y' move or deviate in the same direction, then they are directly correlated or +vely correlated.

If the 2 random variables move in opp. direction, then they are said to be inversely or -vely correlated.

Note:- The coefficient of correlation is denoted by 'r' and

$-1 \le r \le 1$.

1) If $r=0$, then there is no linear correlation between 2 variables
2) If $r=1$, then the correlation is perfect and +ve.
3) If $r=-1$, then the correlation is -vely correlated & perfect
4) If $0<r<1$, then they are partially correlated & +ve.
5) If $-1<r<0$, they are partially correlated -ve

Karl pearson's correlation formula:- ~~It is a linear relation~~

If 'x' & 'y' are two r.v.s then correlation coefficient is given by $r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$.

or $r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$

$\therefore cov(X,Y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n}$.

or $r = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\sqrt{\Sigma y_i^2}}$

or $r = \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2}\sqrt{\Sigma Y_i^2}}$

where $X_i = (x_i - \bar{x})$

$Y_i = (y_i - \bar{y})$

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

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

r is a measure of linear relationship between 'x' & 'y'

Correlation :- If a change in one variable results a change in another variable, then those two variables are said to be correlated.

If the 2 r.v's 'x' & 'y' move in the same direction, then they are directly +vely correlated.

If the 2 r.v's 'x' & 'y' move in opp dir, then they are said to be inversely or -vely correlated.

If only 2 variables are considered for correlation analysis, it is called simple correlation. When three or more variables are considered, then it is a problem of multiple correlation.

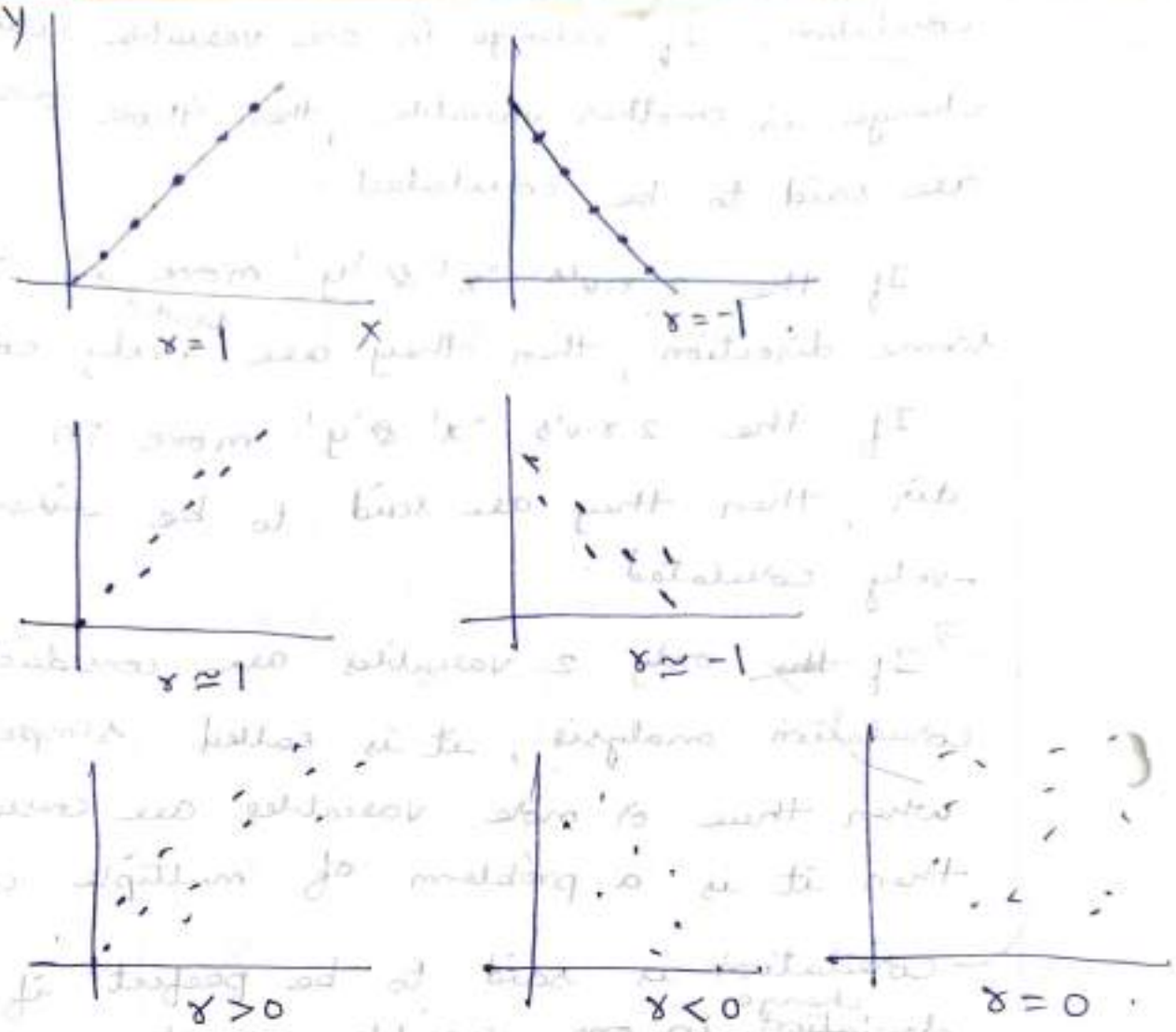
Correlation is said to be perfect if the change in one variable results in the same amt of change in another variable.

If the amt of change in one variable tend to bear a const. ratio to the amt of change in the other variable, then the correlation is said to be linear otherwise non-linear or curvilinear.

The measure of correlation is called as correlation coefficient or correlation Index. * See back Contd

Methods of Studying correlation:-

1) Scatter diagram Method
2) Karlpearson Method
3) Rank correlation Method
4) Method of least Squares ...



* The coeff of correlation is denoted by r and $-1 \le r \le 1$

(i) If $r=0$, then there is no linear correlation betwn 2 variables.

(ii) If $r=1$, then correlatin is perfect & +ve

(iii) If $r=-1$, " " " " -ve

(iv) If $0<r<1$, then partial & +ve

(v) If $-1<r<0$, then partial & -ve.

1). The following data relate to the marks of 10 students in the internal test and the university examination for the max of 50 marks. (6)

Internal marks (x)	25	28	30	32	35	36	38	39	42	45
University marks (y)	20	26	29	30	25	18	26	35	35	46

Sol) $\bar{x} = \frac{\sum x_i}{n} = 35$ $\because n = 10$.

$\bar{y} = \frac{\sum y_i}{n} = 29$.

x_i	y_i	$x_i = x_i - \bar{x}$	$y_i = y_i - \bar{y}$	$\tilde{x_i}$	$\tilde{y_i}$	$x_i y_i$
25	20	-10	-9	100	81	90
28	26	-7	-3	49	9	21
30	29	-5	0	25	0	0
32	30	-3	1	9	1	-3
35	25	0	-4	0	16	0
36	18	1	-11	1	121	-11
38	26	3	-3	9	9	-9
39	35	4	6	16	36	24
42	35	7	6	49	36	42
45	46	10	17	100	289	170
		$\sum x_i = 0$	$\sum y_i = 0$.	$\sum \tilde{x_i} = 358$	$\sum \tilde{y_i} = 598$	$\sum x_i y_i = 324$.

$$\gamma = \frac{324}{\sqrt{358 \times 598}} = \frac{324}{\sqrt{214084}}$$

$= 0.7002$. partially correlated & +ve.

Spearman's Rank correlation formula:-

$$\gamma = 1 - 6\frac{\sum d_i^2}{n(n^2-1)}.$$

Pf:- Let a gp. of 'n' individuals be arranged in order of merit in possession of 2 characteristics 'A' & 'B'.

Rank of a person in possession of 2 characteristics 'A' & 'B' may or may not be the same and no 2 individuals have same rank in either 'A' or 'B'. Then the random variables 'X' & 'Y' takes the ranks X_i, Y_i where $i = 1, 2, 3, \ldots n$, where X_i, Y_i takes the values $1, 2, 3, \ldots n$

Then $\bar{X} = \bar{Y} = \frac{1+2+3+\cdots+n}{n}$

$$= \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

$$\Rightarrow \boxed{\bar{X} = \bar{Y} = \frac{n+1}{2}}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

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

$$= \frac{1^2+2^2+3^2+\cdots+n^2}{n} - \left(\frac{n+1}{2}\right)^2.$$

$$\frac{n(n+1)(2n+1)}{6n} - \frac{(n^2+2n+1)}{4}$$

$$\Rightarrow \frac{4n^2+6n+2-3n^2-6n-3}{12}$$

$$\sigma_y^2 = \sigma_x^2 = \frac{n^2-1}{12}$$

Let d_i be the deviation and $d_i = x_i - y_i$ where $x_i \& y_i$.

$\sum d_i = (x_i - \bar{x}) - (y_i - \bar{y})$. add & sub $\bar{x}$.

$d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \quad \because (\bar{x} = \bar{y})$.

$d_i^2 = \left((x_i - \bar{x}) - (y_i - \bar{y})\right)^2$

$\Sigma d_i^2 = \Sigma\left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 - 2(x_i - \bar{x})(y_i - \bar{y})\right]$

$= \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$

divide each term by 'n'

$$\frac{\Sigma d_i^2}{n} = \frac{\Sigma(x_i - \bar{x})^2}{n} + \frac{\Sigma(y_i - \bar{y})^2}{n} - 2\frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

$$\frac{\Sigma d_i^2}{n} = \sigma_x^2 + \sigma_y^2 - 2\,\mathrm{cov}(x, y).$$

$$\frac{\Sigma d_i^2}{n} = \sigma_x^2 + \sigma_y^2 - 2\gamma\sigma_x\sigma_y \qquad \left(\because \gamma = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}\right)$$

$$\frac{\Sigma d_i^2}{n} = 2\sigma_x^2 - 2\gamma\sigma_x^2 \qquad (\because \sigma_x = \sigma_y)$$

$$\frac{\Sigma d_i^2}{n} = 2\sigma_x^2(1 - \gamma).$$

$$\frac{\Sigma d_i^2}{n} = 2\frac{(n^2 - 1)}{12}(1 - \gamma).$$

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

$$\Rightarrow \boxed{\gamma = 1 - 6\frac{\Sigma d_i^2}{n(n^2 - 1)}.}$$

Note:- If 2 individuals or more have the same rank, then add $\frac{m(m^2 - 1)}{12}$ to Σd_i^2 where 'm' represents no. of individuals having the same ranks.

Prob 1) The following are the marks obtained by 8 students in economics & statistics.

Economics (x)	78	56	36	66	25	75	82	62
Statistics (y)	84	44	51	58	60	68	62	58

Compute the spearman correlation coefficient between 'x' & 'y'

x	y	Rank of x = X	Rank of y = Y	$d_i = x - y$	$d_i^2 = (x_i - y_i)^2$
78	84	2	1	1	1
56	44	6	8	-2	4
36	51	7	7	0	0
66	58	4	5.5	-1.5	2.2
25	60	8	4	4	16
75	68	3	2	1	1
82	62	1	3	-2	4
62	58	5	5.5	-0.5	0.2

d_i is repeated 2 times, $m = 2$

$$= \frac{m(m^2-1)}{12}$$

$$\frac{2(3)}{12} = 0.5 \quad \& \quad \Sigma d_i^2 = 28.5$$

Add correlation factor to Σd_i^2, then

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

$$1 - \frac{6(29)}{504}$$

$$= 1 - 0.345$$

$$= 0.655 \Rightarrow$$ It is +ve & partially correlated

Regression:- The statistical method that helps us to estimate the unknown value of one variable with the help of related variable is called regression. (8)

If there is any relationship b/w 2 variables 'x' & 'y', then the pts in the scattered diagram concentrate or cluster around some curve and that curve is called as "curve of Regression".

If that curve is a st. line, then that line is called as "line of regression".

In simple regression or linear regression, we have only 2 variables, one independent and the other as dependent variable, whereas in multiple regression, we have one dependent variable and more than one independent variables

If 'y' is dependent on 'x' then line of regression of 'y' on 'x' is $y = a_0 + a_1 x$. by using method of least squares.

If 'x' is dependent on 'y' then line of regression of 'x' on 'y' is $x = b_0 + b_1 y$.

These two are graphical methods

The line of regression of 'y' on 'x' is

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

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

Here $b_{yx} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$ where $x_i = (x - \bar{x})$

The slope of regression line of 'y' on 'x' is also called as Regression coefficients.

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

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

$b_{xy} = \frac{\Sigma x_i y_i}{\Sigma y_i^2}$ where $(y_i = y_i - \bar{y})$

These two are algebraic methods

The slope of regression line of 'x' on 'y' is also called as regression coefficient.

Note :- 1) The regression lines passes through the pt $(\bar{x}, \bar{y})$. At the pt of intersection of these 2 lines, we can get $\bar{x}$ & $\bar{y}$.

2). If $r = \pm 1$, then the 2 lines coincide & we get only one line.

3). product of slopes $\left(r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}\right) = r^2$.

(a) If the regression coefficients are both +ve, then r is also +ve.

(b) If the regression coefficients are both -ve, then r is also -ve.

(c) If one regression coeff is less than one, then the other regression coeff must be greater than one.

Regression coeff is independent of origin.

Multiple Regression :- If the dependent variable y is a function of more than one independent variable i.e.

$y = f(x_1, x_2 \cdots x_n)$ & if 'f' is linear, then

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$.

Suppose we have 2 variables.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. where $\beta_0, \beta_1, \beta_2$ are pop. parameters

$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$ where b_0, b_1, b_2 are estimates

$$d_i = (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}))$$

$$\sum d_i^2 = \sum (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}))^2 \quad \text{--- (1)}$$

Normal eqns are given by w.r.t b_0 b_1 b_2 and equate them to zero, we get

$$\sum y_i = n b_0 + b_1 \sum x_{1i} + b_2 \sum x_{2i}$$

$$\sum x_{1i} y_i = b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i}$$

$$\sum x_{2i} y_i = b_0 \sum x_{2i} + b_1 \sum x_{1i} x_{2i} + b_2 \sum x_{2i}^2$$

*) Obtain the rank correlation coefficient for the following data (9)

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Sol)

x	y	Rank (x) = X	Rank (y) = Y	$d_i = X_i - Y_i$	$d_i^2 = (x_i - y_i)^2$
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16

$\Sigma d_i^2 = 72$

75 is repeated twice, $m = 2$.

64 is repeated thrice, $m = 3$

68 is repeated twice, $m = 2$

Add correlation factor to Σd_i^2, then

$m = 2 \Rightarrow 6/12 = 0.5$, $m = 3 \Rightarrow 24/12 = 2$.

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

$$= 1 - \frac{6(75)}{990}$$

$$= 1 - 0.454$$

$$= 0.546.$$

2). Two independent variables 'x' & 'y' have means 5 & 10 and variances 4 & 9 respectively. find the correlation coeff b/w. 'u' & 'v' where (i) $u = 3x+4y$, $v = 3x-y$.

(ii) If 'x' & 'y' are not independent, $r = 0.5$, $u = x+y$, $v = x-y$.

Sol)

$u = 3x+4y$, $v = 3x-y$.

Given 'x' & 'y' are independent

$$r_{uv} = \frac{\sigma_{uv}^2 - \sigma_u^2 - \sigma_v^2}{2\sigma_u\sigma_v} \quad \text{---(1)}$$

Given $\bar{x} = 5$, $\bar{y} = 10$. $\sigma_x^2 = 4$, $\sigma_y^2 = 9$.

$$\sigma_u^2 = \sigma_{3x+4y}^2 = \Sigma\left(\frac{(3x+4y) - \overline{(3x+4y)}}{n}\right)^2$$

$$= \Sigma\left(\frac{3(x-\bar{x}) + 4(y-\bar{y})}{n}\right)^2$$

$$= \Sigma\frac{9(x-\bar{x})^2}{n} + 16\Sigma\frac{(y-\bar{y})^2}{n} + 24\Sigma\frac{(x-\bar{x})(y-\bar{y})}{n}$$

$\because$ Indpnt $\Sigma(x-\bar{x})(y-\bar{y}) = 0$

$$\sigma_u^2 = 9\sigma_x^2 + 16\sigma_y^2$$

$$\sigma_u^2 = 180.$$

Similarly $\sigma_{3x-y}^2 = 9\sigma_x^2 + \sigma_y^2$

$$\sigma_v^2 = 36+9 = 45.$$

$$\sigma_{u+v}^2 = \sigma_{6x+3y}^2 = 36\sigma_x^2 + 9\sigma_y^2 \qquad \begin{matrix} 3x+4y+3x-y \\ 6x+3y \end{matrix}$$

$$= 36\times 4 + 9(9)$$

$$= 144 + 81$$

3) The eqns of 2 regression lines obtained in correlation analysis are $3x+12y=19$ & $3y+9x=46$. find (10)

(i) correlation coeff r. (ii) Mean values of 'x' & 'y'

(iii) the ratio of coeff of variability of x to that of y

Sol (i) $y = \frac{19-3x}{12}$ — ① $\quad x = \frac{46-3y}{9}$ — ②

here $b_{xy} = 0.25$

$b_{yx} = -0.33$.

$r^2 = b_{xy} \times b_{yx}$

$= -0.33 \times -0.25$

$r = -0.287$.

(ii) W.K.T. the 2 lines of regression passes through $\bar{x}$ & $\bar{y}$

$3\bar{x} + 12\bar{y} = 19$

$3\bar{y} + 9\bar{x} = 46$

$\Rightarrow \bar{x} = 5$

$\bar{y} = 0.3$.

(iii) $\frac{\sigma_x^2}{\sigma_y^2} = \frac{1}{r^2}\left(r\frac{\sigma_x}{\sigma_y}\right)^2$

$= \frac{1}{r^2}(b_{xy})^2$

$= 1.389$.

$= \frac{4}{3}$.

correlation coefficient for Bivariate frequency distribution (Compound)

Large grouped data is arranged into a bivariate frequency table for Bivariate frequency distribution correlation coefficient is given by

$$r = \frac{N\Sigma f_{xy}U_xU_y - \Sigma f_xU_x \Sigma f_yU_y}{\sqrt{\left[N\Sigma f_xU_x^2 - (\Sigma f_xU_x)^2\right]\left[N\Sigma f_yU_y^2 - (\Sigma f_yU_y)^2\right]}}$$

where N – total frequency, $f_{(xy)}$ – cell frequencies.

f_x, f_y are marginal frequencies.

$$U_x = \frac{X-A}{C_1}, \quad U_y = \frac{Y-B}{C_2}$$

Let X & Y be the mid values.
C_1 & C_2 are class intervals of X & Y and in general

$$C_1 = C_2$$

A & B are assumed values for X & Y.

Consider the frequency distribution of final grades of 100 students in mathematics & statistics.

Statistics grades	Mathematics grades: 40-49	50-59	60-69	70-79	80-89	90-99	Total
90-99				2	4	4	10
80-89			1	4	6	5	16
70-79			5	10	8	1	24
60-69	1	4	9	5	2		21
50-59	3	6	6	2			17
40-49	3	5	4				12
Total	7	15	25	23	20	10	100

Calculate the correlation coefficient.

Set - 4 5(a) 5(b) Dec - 2011.

$n = 10 \quad \sigma_x = 5.4, \quad \sigma_y = 6.2 \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 66$

1) $r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$.

$= \frac{66}{10 \times 5.4 \times 6.2}$

$=$

2) Y on X $\quad a_1 x + b_1 y + c_1 = 0$

$\Rightarrow y = -\frac{c_1}{b_1} - \frac{a_1}{b_1} x$

x on y $\quad a_2 x + b_2 y + c_2 = 0$

$x = -\frac{c_2}{a_2} - \frac{b_2}{a_2} y$.

$b_{yx} = -\frac{a_1}{b_1} \qquad b_{xy} = -\frac{b_2}{a_2}$

$b_{yx} \times b_{xy} = r^2$

$\left(-\frac{a_1}{b_1}\right)\left(-\frac{b_2}{a_2}\right) = r^2$

$\frac{a_1 b_2}{a_2 b_1} = r^2$

$\therefore \quad 0 \le r^2 \le 1$

$0 \le \frac{a_1 b_2}{a_2 b_1} \le 1$

$\Rightarrow \boxed{a_1 b_2 \le a_2 b_1}$

$m_r = \frac{1}{r} \frac{\sigma_y}{\sigma_x}$

If $x = 2y+3$ & $y = Kx+6$. are the regression lines of x on y & y on x respectively then (i) S.T. $0 \le K \le \frac{1}{2}$

(ii) If $K = \frac{1}{8}$ find r & $(\bar{x}, \bar{y})$

Soln $b_{xy} = 2$, $b_{yx} = K$.

(i) $b_{xy} \times b_{yx} = r^2 \Rightarrow 2K = r^2$ and w.k.T. $|r| \le 1$ and $r^2 \ge 0$

$\Rightarrow r^2 \le 1$ & $r^2 \ge 0$

$\Rightarrow 0 \le r^2 \le 1$

$\Rightarrow 0 \le 2K \le 1$

$\Rightarrow 0 \le K \le \frac{1}{2}$.

(ii) If $K = \frac{1}{8}$, then $2\left(\frac{1}{8}\right) = r^2 \Rightarrow r^2 = \frac{1}{4} \Rightarrow r = \pm\frac{1}{2}$

$x = 2y+3$, $y = \frac{1}{8}x + 6$ $\Rightarrow 8y = x + 48$

$\Rightarrow$ $\because$ Regression lines passes then their means

$\bar{x} - 2\bar{y} = 3$ — (1) & $-\bar{x} + 8\bar{y} = 48$ — (2)

on solv (1) & (2)

we get $(\bar{x}, \bar{y}) = (20, 8.5)$

→. If θ is the angle between two regression lines of y on x & x on y. then p.T. $\tan\theta = \frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$

Pf:- Regression line of y on x is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(\bar{x} - \bar{x})$ — (1)

Regression line of x on y is $(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(\bar{y} - \bar{y})$ — (2)

eq (2) can be written as $(y - \bar{y}) = \frac{1}{r}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$.

Slope of this line is $\frac{1}{r}\frac{\sigma_y}{\sigma_x} = m_1$

Slope of 1st line is $m_2 = r\frac{\sigma_y}{\sigma_x}$ & second line $m_1 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$

Then θ is the angle betwn two regression lines

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{1}{r}\frac{\sigma_y}{\sigma_x}\cdot\frac{r\sigma_y}{\sigma_x}}$$

$$\tan(m_1 - m_2)$$

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

If $0 < \theta < \pi/2$, $\tan\theta$ is +ve.

If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \pi/2$

∴ when $\theta = \pi/2$ there is no relationship ∴ the lines are ⊥

→ Find the most plausible values of 'x' & 'y' from foll eqns.

$$x + y = 6, \quad 2x + y = 2, \quad 2x + 5y = 7, \quad 3x + 4y = -4$$

Soln: Most plausible or reasonable values are given by solⁿ of eqns obtained after diff sum of the squares of given expressions w.r 'x' & 'y' partially and equating them to zero.

Let

$$U = (x+y-6)^2 + (2x+y-2)^2 + (2x+5y-7)^2 + (3x+4y+4)^2$$

$$\frac{\partial U}{\partial x} =$$

$$\frac{\partial U}{\partial y} =$$

Then solve to get $(x, y) = (\quad)$

(1) (i) If θ is the angle betwn two regression lines and S.D. of Y is twice the S.D. of X and $r = 0.25$ find $\tan\theta$.

(ii) If $\sigma_x = \sigma_y = \sigma$ and the angle betwn the regression lines is $\tan^{-1}\frac{4}{3}$ find r.

correlation coefficient is given by

$$\gamma = \frac{N\Sigma f U_x U_y - (\Sigma f_x U_x)(\Sigma f_y U_y)}{\sqrt{\left[N\Sigma f_x U_x^2 - (\Sigma f_x U_x)^2\right]\left[N\Sigma f_y U_y^2 - (\Sigma f_y U_y)^2\right]}}$$

- Here N is the total frequency. Each square in a table is called as a cell and the no. indicated in the cell is called the cell frequency.
- The totals indicated in the last row & last colm are called marginal totals or marginal frequencies.
- Here X is grouped into K classes & Y is grouped into m classes.
- f_{ij} or $f_{\bullet}$ is the cell frequency of i^{th} X-class interval & j^{th} Y-class interval.

X_{L_1}, X_{U_1} denotes lower & upper limits of 1st class, Y_{L_m}, Y_{U_m} denotes the lower & upper limits of m^{th} class.

- Blank cell denotes zero cell frequency.

$$N = \sum_{j=1}^{K} \sum_{i=1}^{m} f_{ij}$$

- Let X_i be the mid value or class mark of the i^{th} X-class & Y_j be the mid value of j^{th} Y class.

$$U_x = \frac{X-A}{C_1} \quad U_y = \frac{Y-B}{C_2}, \text{ where}$$

C_1 = class size of X-intervals,
C_2 = class size of Y intervals.
A - assumed classmark for X-classes.
B - assumed class mark for Y-classes.
f_x : marginal frequencies of X (colm sums of f_{ij}'s)
f_y : marginal frequencies of Y (row sums of f_{ij}'s)

Also $\Sigma f_x = \Sigma f_y = N$.

fU_xU_y is denoted by a no. in corner of each cell.

In general $C_1 = C_2$, $U_x U_y$ can be 0, ±1, ±2, ±3, ...

& $\Sigma f U_x U_y = \Sigma f U_y U_x$.

Maths Grades X

Statistics Grades Y	class mark X → / U_y ↓ / U_x →	44.5	54.5	64.5	74.5	84.5	94.5	f_y	$f_y U_y$	$f_y U_y^2$	Sum of corner nos in each row.
	class mark Y ↓	-2	-1	0	1	2	3				
	94.5 / 2				2 [4]	4 [16]	4 [24]	10	20	40	44
	84.5 / 1			1 [0]	4 [4]	6 [12]	5 [15]	16	16	16	31
	74.5 / 0			5 [0]	10 [0]	8 [0]	1 [0]	24	0	0	0
	64.5 / -1	1 [2]	4 [4]	9 [0]	5 [-5]	2 [-4]		21	-21	21	-3
	54.5 / -2	3 [12]	6 [12]	6 [0]	2 [-4]			17	-34	68	20
	44.5 / -3	3 [18]	5 [15]	4 [0]				12.	-36	108	33
	f_x	7	15	25	23	20	10	$\Sigma f_x = \Sigma f_y = N = 100$	$\Sigma f_y U_y = -55$	$\Sigma f_y U_y^2 = 253$	$\Sigma f U_y U_x = 125$
	$f_x U_x$	-14	-15	0	23	40	30	$\Sigma f_x U_x = 64$			
	$f_x U_x^2$	28	15	0	23	40	30	$\Sigma f_x U_x^2 = 236$			
	Sum of corner nos in each colm.	32	31	0	-1	24	39	$\Sigma f U_x U_y = 125$			

check

$$r = \frac{100(125) - 64(-55)}{\sqrt{(100 \times 236 - 64^2)\,(100(253) - (-52)^2)}}$$

$$= 0.7686 .$$