(*i*) Which firm, A or B, has a larger wage bill ?

(*ii*) In which firm, A or B, is there greater variability in individual wages ?

(*iii*) Calculate the variance of the distribution of wages of all the workers in the firms A and B taken together.

12. Find the co-efficient of skewness for the following distribution :

Class	*Frequency*	*Class*	*Frequency*
0—5	2	20—25	21
5—10	5	25—30	16
10—15	7	30—35	8
15—20	13	35—40	3

13. Calculate the quartile co-efficient of skewness for the following distribution :

x :	1—5	6—10	11—15	16—20	21—25	26—30	31—35
f :	3	4	68	30	10	6	2

14. Calculate the first four moments about the mean for the following data :

Variate :	1	2	3	4	5	6	7	8	9
Frequency :	1	6	13	25	30	22	9	5	2

15. The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40. Show that the mean is 3, variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

16. For a distribution, the mean is 10, variance is 16, γ_1 is + 1 and β_2 is 4. Find the first four moments about the origin.

17. The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Find moments about the mean.

18. Show that for a discrete distribution :

(*i*) $\beta_2 > 1$ (*ii*) $\beta_2 > \beta_1$.

Answers

1.	12.32 marks	**2.** 6.3	**3.** 9	**4.** 10.9 marks
5.	76.53, 9.87	**6.** 4, 7	**8.** 39.9, 4.9	**9.** A, B.
10.	Height	**11.** (*i*) B (*ii*) B (*iii*) ₹ 180, 121.36		
12.	– 1	**13.** 0.25	**14.** 0, 2.49, 0.68, 18.26	
16.	10, 116, 1544, 23184	**17.** 0, 16, – 64, 162.		

21.24. CORRELATION

In a bivariate distribution, if the change in one variable affects a change in the other variable, the variables are said to be *correlated*.

If the two variables deviate in the same direction *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct or positive*.

e.g., the correlation between income and expenditure is positive.

If the two variables deviate in opposite direction *i.e.*, if the increase (or decrease) in one results in a corresponding decrease (or increase) in the other, correlation is said to be *inverse or negative*.

e.g., the correlation between volume and the pressure of a perfect gas or the correlation between price and demand is negative.

Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding **proportional deviation** in the other.

21.25. SCATTER OR DOT DIAGRAMS

It is the simplest method of the diagrammatic representation of bivariate data. Let (x_i, y_i) $i = 1, 2, 3, \ldots\ldots, n$ be a bivariate distribution. Let the values of the variables x and y be plotted along the x-axis and y-axis on a suitable scale. Then corresponding to every ordered pair, there corresponds a point or dot in the xy-plane. The diagram of dots so obtained is called a *dot or scatter diagram*.

If the dots are vary close to each other and the number of observations is not vary large, a fairly good correlation is expected. If the dots are widely scattered, a poor correlation is expected.

21.26. KARL PEARSON'S CO-EFFICIENT OF CORRELATION (OR PRODUCT MOMENT CORRELATION CO-EFFICIENT)

Correlation co-efficient between two variables x and y, usually denoted by $r(x, y)$ or r_{xy} is a numerical measure of linear relationship between them and is defined as

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_1 - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\ \Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma\,(x_i - \bar{x})(y_1 - \bar{y})}{\sqrt{\frac{1}{n}\Sigma\,(x_i - \bar{x})^2 \,.\, \frac{1}{n}\Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma\,(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}.$$

Note. Correlation co-efficient is independent of change of origin and scale.

Let us define two new variables u and v as

$$u = \frac{x - a}{h}\,,\ v = \frac{y - b}{k} \quad \text{where } a, b, h, k \text{ are constants, then } r_{xy} = r_{uv}.$$

21.27. COMPUTATION OF CORRELATION CO-EFFICIENT

We know that $r_{xy} = \dfrac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$

Now $\dfrac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \dfrac{1}{n}\Sigma\,(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\,\bar{y})$

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

$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\ .\ \bar{x} - \bar{x}\ .\ \bar{y} + \bar{x}\ .\ \bar{y} = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,.\,\bar{y}$$

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

$$= \frac{1}{n}\Sigma x_i^2 - 2\bar{x}\ .\ \frac{1}{n}\Sigma x_i + \frac{1}{n}n\bar{x}^2 = \frac{1}{n}\Sigma x_i^2 - 2\bar{x}\,.\,\bar{x} + \bar{x}^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

Similarly, $\sigma_y{}^2 = \dfrac{1}{n}\,\Sigma y_i{}^2 - \overline{y}^2$

$$\therefore \qquad r_{xy} = \frac{\dfrac{1}{n}\,\Sigma x_i y_i - \overline{x}\,\overline{y}}{\sqrt{\left(\dfrac{1}{n}\,\Sigma x_i{}^2 - \overline{x}^2\right)\left(\dfrac{1}{n}\,\Sigma y_i{}^2 - \overline{y}^2\right)}} = \frac{n\,\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\,\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\,\Sigma y^2 - (\Sigma y)^2}}$$

If $\quad u = \dfrac{x-a}{h},\, v = \dfrac{y-b}{k} \quad$ then $\quad r_{xy} = r_{uv} = \dfrac{n\,\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\,\Sigma u^2 - (\Sigma u)^2}\,\sqrt{n\,\Sigma v^2 - (\Sigma v)^2}}$.

ILLUSTRATIVE EXAMPLES

Example 1. *Ten students got the following percentage of marks in Principles of Economics and Statistics* :

Roll Nos.	:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Marks in Economics	:	*78*	*36*	*98*	*25*	*75*	*82*	*90*	*62*	*65*	*39*
Marks in Statistics	:	*84*	*51*	*91*	*60*	*68*	*62*	*86*	*58*	*53*	*47*

Calculate the co-efficient of correlation.

Sol. Let the marks in the two subjects be denoted by x and y respectively.

x	y	$u = x - 65$	$v = y - 66$	u^2	v^2	uv
78	84	13	18	169	324	234
36	51	– 29	– 15	841	225	435
98	91	33	25	1089	625	825
25	60	– 40	– 6	1600	36	240
75	68	10	2	100	4	20
82	62	17	– 4	289	16	– 68
90	86	25	20	625	400	500
62	58	– 3	– 8	9	64	24
65	53	0	– 13	0	169	0
39	47	– 26	– 19	676	361	494
Total		0	0	5398	2224	2734

$$\overline{u} = \frac{1}{n}\,\Sigma u_i = 0,\ \overline{v} = \frac{1}{n}\,\Sigma v_i = 0$$

$$r_{uv} = \frac{\dfrac{1}{n}\,\Sigma u_i v_i - \overline{u}\,\overline{v}}{\sqrt{\left(\dfrac{1}{n}\,\Sigma u_i{}^2 - \overline{u}^2\right)\left(\dfrac{1}{n}\,\Sigma v_i{}^2 - \overline{v}^2\right)}} = \frac{\dfrac{1}{10}\,(2734)}{\sqrt{\dfrac{1}{10}\,(5398).\,\dfrac{1}{10}\,(2224)}} = 0.787$$

Hence $\quad r_{xy} = r_{uv} = 0.787.$

Example 2. *A computer while calculating correlation co-efficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$$n = 25, \qquad \Sigma X = 125, \qquad \Sigma X^2 = 650,$$
$$\Sigma Y = 100, \qquad \Sigma Y^2 = 460, \qquad \Sigma XY = 508.$$

It was, however, later discovered at the time of checking that he had copied down two pairs as *while the correct values were*

X	*Y*
6	*14*
8	*6*

X	*Y*
8	*12*
6	*8*

Obtain the correct value of correlation co-efficient.

Sol.

$$\left.\begin{aligned}
\text{Corrected } \Sigma X &= 125 - 6 - 8 + 8 + 6 = 125\\
\text{Corrected } \Sigma Y &= 100 - 14 - 6 + 12 + 8 = 100\\
\text{Corrected } \Sigma X^2 &= 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650\\
\text{Corrected } \Sigma Y^2 &= 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436\\
\text{Corrected } \Sigma XY &= 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520
\end{aligned}\right\}$$

(Subtract the incorrect values and add the corresponding correct values)

$$\overline{X} = \frac{1}{n}\,\Sigma X = \frac{1}{25} \times 125 = 5\ ; \quad \overline{Y} = \frac{1}{n}\,\Sigma Y = \frac{1}{25} \times 100 = 4$$

$$\text{Corrected } r_{xy} = \frac{\dfrac{1}{n}\Sigma XY - \overline{X}\,\overline{Y}}{\sqrt{\left(\dfrac{1}{n}\Sigma X^2 - \overline{X}^2\right)\left(\dfrac{1}{n}\Sigma Y^2 - \overline{Y}^2\right)}}$$

$$= \frac{\dfrac{1}{25} \times 520 - 5 \times 4}{\sqrt{\left(\dfrac{1}{25} \times 650 - 25\right)\left(\dfrac{1}{25} \times 436 - 16\right)}} = \frac{\dfrac{4}{5}}{\sqrt{(1)\left(\dfrac{36}{25}\right)}} = \frac{4}{5} \times \frac{5}{6} = \frac{2}{3} = 0.67.$$

Example 3. *If $z = ax + by$ and r is the correlation co-efficient between x and y, show that*

$$\sigma_z^{\,2} = a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2} + 2abr\ \sigma_x\,\sigma_y.$$

Sol. $z = ax + by$

$\Rightarrow$ $\quad \overline{z} = a\overline{x} + b\overline{y}, \qquad z_i = ax_i + by_i$

$$z_i - \overline{z} = a(x_i - \overline{x}) + b(y_i - \overline{y})$$

Now
$$\sigma_z^{\,2} = \frac{1}{n}\,\Sigma(z_i - \overline{z})^2 = \frac{1}{n}\,\Sigma[a(x_i - \overline{x}) + b(y_i - \overline{y})]^2$$
$$= \frac{1}{n}\,\Sigma[a^2(x_i - \overline{x})^2 + b^2(y_i - \overline{y})^2 + 2ab(x_i - \overline{x})(y_i - \overline{y})]$$
$$= a^2.\ \frac{1}{n}\,\Sigma(x_i - \overline{x})^2 + b^2.\ \frac{1}{n}\,\Sigma(y_i - \overline{y})^2 + 2ab.\ \frac{1}{n}\,\Sigma(x_i - \overline{x})(y_i - \overline{y})$$
$$= a^2\sigma_x^{\,2} + b^2\sigma_y^{\,2} + 2abr\ \sigma_x\sigma_y \qquad \Bigg|\ \because\ r = \frac{\dfrac{1}{n}\,\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x\sigma_y}$$

21.28. CALCULATION OF CO-EFFICIENT OF CORRELATION FOR A BIVARIATE FREQUENCY DISTRIBUTION

If the bivariate data on x and y is presented on a two way correlation table and f is the frequency of a particular rectangle in the correlation table, then

$$r_{xy} = \frac{\Sigma fxy - \frac{1}{n}\Sigma fx\ \Sigma fy}{\sqrt{\left[\Sigma fx^2 - \frac{1}{n}(\Sigma fx)^2\right]\left[\Sigma fy^2 - \frac{1}{n}(\Sigma fy)^2\right]}}$$

Since change of origin and scale do not affect the co-efficient of correlation,

$\therefore$ $r_{xy} = r_{uv}$ where the new variables u, v are properly chosen.

Example. *The following table gives according to age the frequency of marks obtained by 100 students in an intelligence test :*

Age (in years) *Marks*	*18*	*19*	*20*	*21*	*Total*
10—20	*4*	*2*	*2*		*8*
20—30	*5*	*4*	*6*	*4*	*19*
30—40	*6*	*8*	*10*	*11*	*35*
40—50	*4*	*4*	*6*	*8*	*22*
50—60		*2*	*4*	*4*	*10*
60—70		*2*	*3*	*1*	*6*
Total	*19*	*22*	*31*	*28*	*100*

Calculate the co-efficient of correlation between age and intelligence.

Sol. Let age and intelligence be denoted by x and y respectively.

Mid value	x y	*18*	*19*	*20*	*21*	f	u	fu	fu^2	fuv
15	10—20	4	2	2		8	– 3	24	72	30
25	20—30	5	4	6	4	19	– 2	– 38	76	20
35	30—40	6	8	10	11	35	– 1	– 35	35	9
45	40—50	4	4	6	8	22	0	0	0	0
55	50—60		2	4	4	10	1	10	10	2
65	60—70		2	3	1	6	2	12	24	– 2
	f	19	22	31	28	100	Totals	**– 75**	**217**	**59**
	v	2	– 1	0	1	Totals				
	fv	– 38	– 22	0	28	**– 32**				
	fv^2	76	22	0	28	**126**				
	fuv	56	16	0	13	**59**				

Let us define two new variables u and v as $u = \dfrac{y-45}{10}$, $v = x - 20$

$$r_{xy} = r_{uv} = \frac{\Sigma fuv - \frac{1}{n}\Sigma fu\,\Sigma fv}{\sqrt{\left[\Sigma fu^2 - \frac{1}{n}(\Sigma fu)^2\right]\left[\Sigma fv^2 - \frac{1}{n}(\Sigma fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-75)(-32)}{\sqrt{\left[217 - \frac{1}{100}(-75)^2\right]\left[126 - \frac{1}{100}(-32)^2\right]}} = \frac{59-24}{\sqrt{\frac{643}{4}\times\frac{2894}{25}}} = 0.25.$$

21.29. RANK CORRELATION

Sometimes we have to deal with problems in which data cannot be quantitatively measured but qualitative assessment is possible.

Let a group of n individuals be arranged in order of merit or proficiency in possession of two characteristics A and B. The ranks in the two characteristics are, in general, different. For example, if A stands for intelligence and B for beauty, it is not necessary that the most intelligent individual may be the most beautiful and *vice versa*. Thus an individual who is ranked at the top for the characteristic A *may be* ranked at the bottom for the characteristic B. Let (x_i, y_i), $i = 1, 2, \ldots\ldots, n$ be the ranks of the n individuals in the group for the characteristics A and B respectively. Pearsonian co-efficient of correlation between the ranks x_i's and y_i's is called the *rank correlation co-efficient* between the characteristics A and B for that group of individuals.

Thus rank correlation co-efficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\,\Sigma(y_i-\bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} \qquad \ldots(1)$$

Now x_i's and y_i's are merely the permutations of n numbers from 1 to n. *Assuming that no two individuals are bracketed or tied in either classification i.e., $(x_i, y_i) \neq (x_j, y_j)$ for $i \neq j$, both x and y take all integral values from 1 to n.*

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{n}(1+2+3+\ldots\ldots+n) = \frac{1}{n}\cdot\frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\Sigma x_i = 1+2+3+\ldots\ldots+n = \frac{n(n+1)}{2} = \Sigma y_i$$

$$\Sigma x_i^2 = 1^2 + 2^2 + \ldots\ldots + n^2 = \frac{n(n+1)(2n+1)}{6} = \Sigma y_i^2$$

If d_i denotes the difference in ranks of the i^{th} individual, then

$$d_i = x_i - y_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad [\because \ \bar{x} = \bar{y}]$$

$$\frac{1}{n}\Sigma d_i^2 = \frac{1}{n}\Sigma[(x_i-\bar{x})-(y_i-\bar{y})]^2$$

$$= \frac{1}{n}\Sigma(x_i-\bar{x})^2 + \frac{1}{n}\Sigma(y_i-\bar{y})^2 - 2.\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})$$

$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad ...(2) \quad [\text{Using } (1)]$$

But $$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 = \sigma_y^2$$

$\therefore$ From (2), $$\frac{1}{n}\Sigma d_i^2 = 2\sigma_x^2 - 2r\sigma_x^2 = 2(1-r)\,\sigma_x^2 = 2(1-r)\left[\frac{1}{n}\Sigma x_i^2 - \bar{x}^2\right]$$

$$= 2(1-r)\left[\frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\right]$$

$$= (1-r)\,(n+1)\left[\frac{4n+2-3n-3}{6}\right] = \frac{(1-r)(n^2-1)}{6} \quad \text{or} \quad 1-r = \frac{6\Sigma d_i^2}{n(n^2-1)}$$

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

Note. This is called *Spearman's Formula for* Rank Correlation.

$$\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = 0$$

always. This serves as a check on calculations.

Example . *The marks secured by recruits in the selection test (X) and in the proficiency test (Y) are given below* :

Serial No	:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
X	:	*10*	*15*	*12*	*17*	*13*	*16*	*24*	*14*	*22*
Y	:	*30*	*42*	*45*	*46*	*33*	*34*	*40*	*35*	*39*

Calculate the rank correlation co-efficient.

Sol. *Here the marks are given. Therefore, first of all, write down ranks.* In each series, the item with the largest size is ranked 1, next largest 2 and so on.

X	10	15	12	17	13	16	24	14	22	Total
Y	30	42	45	46	33	34	40	35	39	
Ranks in X (x)	9	5	8	3	7	4	1	6	2	
Ranks in Y (y)	9	3	2	1	8	7	4	6	5	
$d = x - y$	0	2	6	2	– 1	– 3	– 3	0	– 3	0
d^2	0	4	36	4	1	9	9	0	9	72

$\therefore$ $$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 72}{9 \times 80} = 1 - 0.6 = 0.4 \quad \text{Here } n = 9.$$

21.30. REPEATED RANKS

If any two or more individuals have same rank or the same value in the series of marks, then the above formula fails and requires an adjustment. In such cases, each individual is given an average rank. This common average rank is the average of the ranks which these individuals would have assumed if they were slightly different from each other. Thus, if two individual are ranked equal at the sixth place, they would have assumed the 6th and 7th ranks if they were ranked slightly different. Their common rank = $\frac{6+7}{2}$ = 6.5. If three individuals are ranked equal at fourth place, they would have assumed the 4th, 5th and 6th ranks if they were ranked slightly different. Their common rank = $\frac{4+5+6}{3} = 5$.

Adjustment. Add $\frac{1}{12}\, m(m^2 - 1)$ to Σd^2 where m stands for the number of times an item is repeated.

☞This adjustment factor is to be added for each repeated item.

Thus
$$r = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12} m_1(m_1^2 - 1) + \frac{1}{12} m_2(m_2^2 - 1) + \ldots\ldots\right\}}{n(n^2 - 1)}\ .$$

Example. *Obtain the rank correlation co-efficient for the following data :*

X :	*68*	*64*	*75*	*50*	*64*	*80*	*75*	*40*	*55*	*64*
Y :	*62*	*58*	*68*	*45*	*81*	*60*	*68*	*48*	*50*	*70.*

Sol. Here, marks are given, so write down the ranks.

X	68	64	75	50	64	80	75	40	55	64	Total
Y	62	58	68	45	81	60	68	48	50	70	
Ranks in X (x)	4	6	2.5	9	6	1	2.5	10	8	6	
Ranks in Y (y)	5	7	3.5	10	1	6	3.5	9	8	2	
$d = x - y$	– 1	– 1	– 1	– 1	5	– 5	– 1	1	0	4	0
d^2	1	1	1	1	25	25	1	1	0	16	72

In the X-series, the value 75 occurs twice. Had these values been slightly different, they would have been given the ranks 2 and 3. Therefore, the common rank given to them is $\frac{2+3}{2}$ = 2.5. The value 64 occurs thrice. Had these values been slightly different, they would have been given the ranks 5, 6 and 7. Therefore the common rank given to them is $\frac{5+6+7}{3} = 6$. Similarly, in the Y-series, the value 68 occurs twice. Had these values been slightly different, they would have been given the ranks 3 and 4. Therefore, the common rank given to them is $\frac{3+4}{2} = 3.5$.

Thus, m has the values 2, 3, 2.

$$\therefore \qquad r = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12} m(m^2 - 1) + \frac{1}{12} m(m^2 - 1) + \ldots\ldots\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left[72 + \frac{1}{12}\left\{2(2^2 - 1)\right\} + \frac{1}{12}\left\{3(3^2 - 1)\right\} + \frac{1}{12}\left\{2(2^2 - 1)\right\}\right]}{10(10^2 - 1)}$$

$$= 1 - \frac{6 \times 75}{990} = \frac{6}{11} = 0.545.$$

21.31. REGRESSION

Regression is the estimation or prediction of unknown values of one variable from known values of another variable.

After establishing the fact of correlation between two variables, it is natural curiosity to know the extent to which one variable varies in response to a given variation in the other variable *i.e.*, one is interested to know the nature of relationship between the two variables.

Regression measures the nature and extent of correlation.

21.32. LINEAR REGRESSION

If two variates x and y are correlated *i.e.*, there exists an association or relationship between them, then the scatter diagram will be more or less concentrated round a curve. This curve is called the *curve of regression* and the relationship is said to be expressed by means of *curvilinear regression.* In the particular case, when the curve is a straight line, it is called a *line of regression* and the regression is said to be *linear*.

A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimised [*See Fig. (a)*], it is called *the line of regression of y on x* and it gives *the best estimate of y for any given value of x.*

If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of x is minimised [*See Fig. (b)*], it is called *the line of regression of x on y* and it gives *the best estimate of x for any given value of y.*

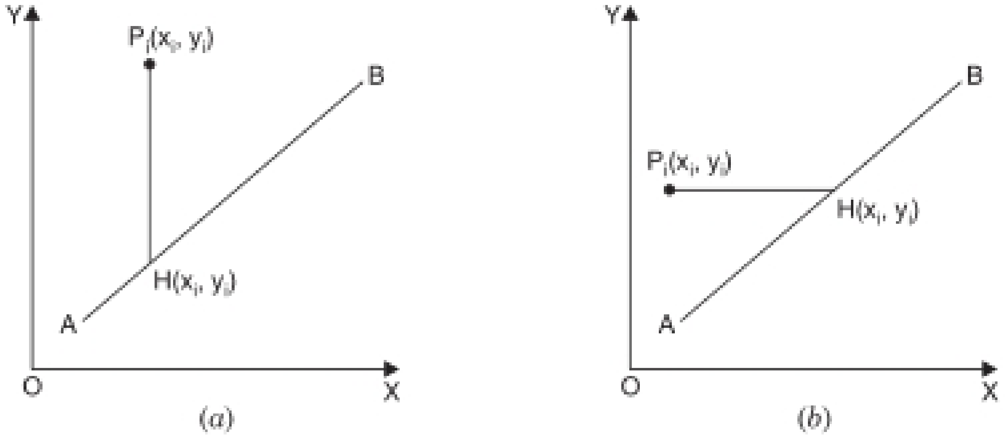

Fig.

21.33. LINES OF REGRESSION

Let the equation of line of regression of y on x be

$$y = a + bx \qquad ...(1)$$

Then
$$\overline{y} = a + b\overline{x} \qquad ...(2)$$

Subtracting (2) from (1), we have

$$y - \overline{y} = b(x - \overline{x}) \qquad ...(3)$$

The normal equations are
$$\Sigma y = na + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad ...(4)$$

Shifting the origin to $(\bar{x}, \bar{y})$, (4) becomes

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2 \qquad \text{...(5)}$$

Since $\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n\,\sigma_x\sigma_y} = r \qquad \therefore \quad \Sigma(x - \bar{x}) = 0$; and $\dfrac{1}{n}\,\Sigma(x - \bar{x})^2 = \sigma_x{}^2$

$\therefore$ From (5), $\quad nr\sigma_x\sigma_y = a.0 + b.n\sigma_x{}^2 \qquad \Rightarrow \qquad b = \dfrac{r\sigma_y}{\sigma_x}$

Hence, from (3), the line of regression of y on x is $\qquad y - \bar{y} = r\,\dfrac{\sigma_y}{\sigma_x}\,(x - \bar{x})$

Similarly, the line of regression of x on y is $\qquad x - \bar{x} = r\,\dfrac{\sigma_x}{\sigma_y}\,(y - \bar{y})$

$\dfrac{r\sigma_y}{\sigma_x}$ is called the regression co-efficient of y on x and is denoted by b_{yx}.

$\dfrac{r\sigma_x}{\sigma_y}$ is called the regression co-efficient of x on y and is denoted by b_{xy}.

Note. If $r = 0$, the two lines of regression become $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to X and Y axes respectively and passing through their means $\bar{y}$ and $\bar{x}$. They are mutually perpendicular. If $r = \pm 1$, the two lines of regression will coincide.

21.34. PROPERTIES OF REGRESSION CO-EFFICIENTS

Property I. *Correlation co-efficient is the geometric mean between the regression co-efficients.*

Proof. The co-efficients of regression are $\dfrac{r\sigma_y}{\sigma_x}$ and $\dfrac{r\sigma_x}{\sigma_y}$.

G.M. between them $\quad = \sqrt{\dfrac{r\sigma_y}{\sigma_x} \times \dfrac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r =$ co-efficient of correlation.

Property II. *If one of the regression co-efficients is greater than unity, the other must be less than unity.*

Proof. The two regression co-efficients are $\quad b_{yx} = \dfrac{r\sigma_y}{\sigma_x} \quad$ and $\quad b_{xy} = \dfrac{r\sigma_x}{\sigma_y}$.

Let $\quad b_{yx} > 1$, then $\dfrac{1}{b_{yx}} < 1 \qquad$ ...(1)

Since $\quad b_{yx} \,.\, b_{xy} = r^2 \le 1 \quad (\because \quad -1 \le r \le 1) \quad \therefore \quad b_{xy} \le \dfrac{1}{b_{yx}} < 1. \qquad$ | Using (1)

Similarly, if $\quad b_{xy} > 1$, then $b_{yx} < 1$.

Property III. *Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.*

Proof. We have to prove that $\dfrac{b_{yx}+b_{xy}}{2} > r$ or $\dfrac{\dfrac{r\sigma_y}{\sigma_x}+\dfrac{r\sigma_x}{\sigma_y}}{2} > r$

or $\sigma_y^2 + \sigma_x^2 > 2\sigma_x\sigma_y$ or $(\sigma_x - \sigma_y)^2 > 0$ which is true.

Property IV. *Regression co-efficients are independent of the origin but not of scale.*

Proof. Let $u = \dfrac{x-a}{h}$, $v = \dfrac{y-b}{k}$ where a, b, h and k are constants

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r \cdot \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h}\, b_{vu}$$

Similarly, $b_{xy} = \dfrac{h}{k}\, b_{uv}$.

Thus, b_{yx} and b_{xy} are both independent of a and b but not of h and k.

Property V. *The correlation co-efficient and the two regression co-efficients have same sign.*

Proof. Regression co-efficient of y on $x = b_{yx} = r\,\dfrac{\sigma_y}{\sigma_x}$

Regression co-efficient of x on $y = b_{xy} = r\,\dfrac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive, b_{yx}, b_{xy} and r have same sign.

21.35. ANGLE BETWEEN TWO LINES OF REGRESSION

If θ is the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta = \frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \quad \textit{where } r, \sigma_x, \sigma_y \textit{ have their usual meanings.}$$

Explain the significance of the formula when r = 0 and r = ± 1.

Proof. Equations to the lines of regression of y on x and x on y are

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

Their slopes are $m_1 = \dfrac{r\sigma_y}{\sigma_x}$ and $m_2 = \dfrac{\sigma_y}{r\sigma_x}$.

$$\therefore \qquad \tan\theta = \pm\frac{m_2 - m_1}{1 + m_2 m_1} = \pm\frac{\dfrac{\sigma_y}{r\sigma_x} - \dfrac{r\sigma_y}{\sigma_x}}{1 + \dfrac{\sigma_y^2}{\sigma_x^2}}$$

$$= \pm\frac{1-r^2}{r}\cdot\frac{\sigma_y}{\sigma_x}\cdot\frac{\sigma_x^2}{\sigma_x^2+\sigma_y^2} = \pm\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Since $r^2 \le 1$ and σ_x, σ_y are positive.

$\therefore$ +ve sign gives the acute angle between the lines.

Hence
$$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x\sigma_y}{\sigma_x{}^2+\sigma_y{}^2}$$

when $r = 0$, $\theta = \dfrac{\pi}{2}$

$\therefore$ The two lines of regression are perpendicular to each other.

Hence the estimated value of y is the same for all values of x and *vice versa*.

when $r = \pm 1$, $\tan\theta = 0$ so that, $\theta = 0$ or π.

Hence the lines of regression coincide and there is perfect correlation between the two variates x and y.

ILLUSTRATIVE EXAMPLES

Example 1. *Calculate the co-efficient of correlation and obtain the least square regression line of y on x for the following data :*

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

Also obtain an estimate of y which should correspond on the average to x = 6.2.

Sol.

x	y	$u = x - 5$	$v = y - 12$	u^2	v^2	uv
1	9	− 4	− 3	16	9	12
2	8	− 3	− 4	9	16	12
3	10	− 2	− 2	4	4	4
4	12	− 1	0	1	0	0
5	11	0	− 1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
Total		0	0	60	60	57

$$r_{xy} = r_{uv} = \frac{\frac{1}{n}\Sigma uv - \bar{u}\,\bar{v}}{\left(\frac{1}{n}\Sigma\bar{u}^2 - \bar{u}^2\right)\left(\frac{1}{n}\Sigma v^2 - \bar{v}^2\right)} = \frac{\frac{1}{9}(57) - 0}{\sqrt{\left[\frac{1}{9}(60) - 0\right]\left[\frac{1}{9}(60) - 0\right]}}$$

$$= \frac{19}{20} = 0.95$$

$$\frac{r\sigma_y}{\sigma_x} = \frac{r\sigma_v}{\sigma_u} = \frac{\frac{1}{n}\Sigma uv - \bar{u}\,\bar{v}}{\frac{1}{n}\Sigma u^2 - \bar{u}^2} = \frac{\frac{1}{9}(57) - 0}{\frac{1}{9}(60) - 0} = \frac{19}{20} = 0.95$$

Also $\bar{x} = 5 + \frac{1}{9}\Sigma u = 5, \bar{y} = 12 + \frac{1}{9}\Sigma v = 12$

Equation of the line of regression of y on x is

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

or $$y - 12 = 0.95\,(x - 5)$$

or $$y = 0.95x + 7.25$$

When $x = 6.2$, estimated value of $y = 0.95 \times 6.2 + 7.25 = 5.89 + 7.25 = 13.14$.

Example 2. *In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are eligible :*

Variance of $x = 9$

Regression equations : $8x - 10y + 66 = 0$, $40x - 18y = 214$.

What were (a) the mean values of x and y, (b) the standard deviation of y and (c) the co-efficient of correlation between x and y.

Sol. (*i*) **Since both the lines of regression pass through the point** $(\bar{x}, \bar{y})$ therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad ...(1)$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \quad ...(2)$$

Multiplying (1) by 5, $40\bar{x} - 50\bar{y} + 330 = 0$ ...(3)

Substracting (3) from (2), $32\bar{y} - 544 = 0 \quad \therefore \quad \bar{y} = 17$

$\therefore$ From (1), $8\bar{x} - 170 + 66 = 0$ or $8\bar{x} = 104 \quad \therefore \quad \bar{x} = 13$

Hence $\bar{x} = 13, \quad \bar{y} = 17$ ...(*a*)

(*ii*) Variance of $x = \sigma_x^2 = 9$ (given)

$\therefore \quad \sigma_x = 3$

The equations of lines of regression can be written as

$$y = .8x + 6.6 \quad \text{and} \quad x = .45y + 5.35$$

$\therefore$ The regression co-efficient of y on x is $\frac{r\sigma_y}{\sigma_x} = .8$ ...(4)

The regression co-efficient of x on y is $\frac{r\sigma_x}{\sigma_y} = .45$ ...(5)

Multiplying (4) and (5), $r^2 = .8 \times .45 = .36 \quad \therefore \quad r = 0.6$ ...(*b*)

(+ve sign with sq. root is taken because regression co-efficients are +ve).

From (4), $\sigma_y = \frac{.8\sigma_x}{r} = \frac{.8 \times 3}{0.6} = 4.$ ...(*c*)

TEST YOUR KNOWLEDGE

1. If the two regression co-efficients are 0.8 and 0.2, what would be the value of co-efficient of correlation.

2. Calculate the co-efficient of correlation for the following ages of husbands and wives :

Husbands's age x	:	23	27	28	28	29	30	31	33	35	36
Wife's age y	:	18	20	22	27	21	29	27	29	28	29

3. Establish the formula $\sigma^2_{x-y} = \sigma_x^{\,2} + \sigma_y^{\,2} - 2r\sigma_x\sigma_y$

where r is the correlation co-efficient between x and y.

4. Calculate the co-efficient of correlation for the following table :

x / y	*16–18*	*18–20*	*20–22*	*22–24*
10—20	2	1	1	
20—30	3	2	3	2
30—40	3	4	5	6
40—50	2	2	3	4
50—60		1	2	2
60—70		1	2	1

5. Ten students got the following percentage of marks in Chemistry and Physics :

Students	:	1	2	3	4	5	6	7	8	9	10
Marks in Chemistry	:	78	36	98	25	75	82	90	62	65	39
Marks in Physics	:	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation co-efficient.

6. Ten competitors in a musical test were ranked by the three judges x, y and z in the following order :

Ranks by x	:	1	6	5	10	3	2	4	9	7	8
Ranks by y	:	3	5	8	4	7	10	2	1	6	9
Ranks by z	:	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

7. A sample of 12 fathers and their eldest sons gave the following data about their heights in inches:

Father	:	65	63	67	64	68	62	70	66	68	67	69	71
Son	:	68	66	68	65	69	66	68	65	71	67	68	70

Calculate the co-efficient of rank correlation.

8. Find the correlation co-efficient between x and y for the given values. Find also the two regression lines.

x	:	1	2	3	4	5	6	7	8	9	10
y	:	10	12	16	28	25	36	41	49	40	50

9. In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively, $7x - 16y + 9 = 0$, $5y - 4x - 3 = 0$.

Calculate the co-efficient of correlation, $\overline{x}$ and $\overline{y}$.

10. Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation co-efficient between x and y.

11. The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (*i*) mean of x's, (*ii*) mean of y's and (*iii*) the correlation co-efficient between x and y.

Answers

1.	0.4	**2.** 0.82	**4.** 0.28
5.	0.84	**6.** x and z	**7.** 0.722
8.	$r = 0.96$; $y = 4.69x + 4.9$; $x = 0.2y - 0.64$		
9.	$r = 0.7395$; $\bar{x} = -0.1034$; $\bar{y} = 0.5172$.		
10.	$\bar{x} = 4$, $\bar{y} = 7$; $r = -0.5$		
11.	(*i*) 15.79	(*ii*) 3.74	(*iii*) – 0.6595

21.36. THEORY OF PROBABILITY

Here we define and explain certain terms which are used frequently.

(*a*) **Trial and event.** Let an experiment be repeated under essentially the same conditions and let it result in any one of the several possible outcomes. Then, the experiment is called a *trial* and the possible outcomes are known as *events* or *cases*.

For example : (*i*) Tossing of a coin is a trial and the turning up of head or tail is an event.

(*ii*) Throwing a die is a trial and getting 1 or 2 or 3 or 4 or 5 or 6 is an event.

(*b*) **Exhaustive events.** The total number of all possible outcomes in any trial is known as *exhaustive events* or *exhaustive cases*.

For example : (*i*) In tossing a coin, there are two exhaustive cases, head and tail.

(*ii*) In throwing a dice, there are 6 exhaustive cases, for any one of the six faces may turn up.

(*iii*) In throwing two dice, the exhaustive cases are $6 \times 6 = 6^2$ for any of the 6 numbers from 1 to 6 on one die can be associated with any of the 6 numbers on the other die.

In general, in throwing n dice, the exhaustive cases are 6^n.

(*c*) **Favourable events or cases.** The cases which entail the happening of an event are said to be *favourable* to the event. It is the total number of possible outcomes in which the specified event happens.

For example : (*i*) In throwing a die, the number of cases favourable to the appearance of a multiple of 3 are two *viz*. 3 and 6 while the number of cases favourable to the appearance of an even number are three, *viz*., 2, 4 and 6.

(*ii*) In a throw of two dice, the number of cases favourable to getting a sum 6 is 5, *viz*., (1, 5) ; (5, 1) ; (2, 4) ; (4, 2) ; (3, 3).

(*d*) **Mutually exclusive events.** Events are said to be *mutually exclusive* or *incompatible* if the happening of any one of them precludes (*i.e.,* rules out) the happening of all others, *i.e.*, if no two or more than two of them can happen simultaneously in the same trial.

For example : (*i*) In tossing a coin, the events head and tail are mutually exclusive, since if the outcome is head, the possibility of getting tail in the same trial is ruled out.

(*ii*) In throwing a die, all the six faces numbered, 1, 2, 3, 4, 5, 6 are mutually exclusive since any outcome rules out the possibility of getting any other.

(*e*) **Equally likely events.** Events are said to be *equally likely* if there is no reason to expect any one in preference to any other.

For example : (*i*) When a card is drawn from a well shuffled pack, any card may appear in the draw so that the 52 different cases are equally likely.