18. Find the coefficient 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

19. Calculate the quartile coefficient 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

20. 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

21. 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.

22. 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.

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

24. Show that for a discrete distribution:

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

Answers

1.	12.32	**2.**	6.3	**3.**	9	**4.**	10.9
5.	75.53, 9.87	**6.**	9, 1.61	**7.**	32, 32.6, 12.4	**8.**	$31.35, $16.64
9.	(*i*) 4, 7	**10.**	10.9, 15.26	**11.**	$17.10	**12.**	14.24, 0.72, 0.52
13.	39.9, 4.9	**14.**	A	**15.**	Height	**16.**	A, B
17.	(*i*) B (*ii*) B (*iii*) $180, 121.36	**18.**	–1	**19.**	0.25	**20.**	0, 2.49, 0.68, 18.26
		22.	10, 116, 1544, 23184	**23.**	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, the 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 directions, *i.e.*, if the increase (or decrease) in one results in a corresponding decrease (or increase) in the other, the 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

This is the simplest method of the diagrammatic representation of bivariate data. Let (x_i, y_i) $i = 1, 2, 3, \ldots, n$ 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 very close to each other and the number of observations is not very large, a fairly good correlation is expected. If the dots are widely scattered, a poor correlation is expected.

21.26 KARL PEARSON'S COEFFICIENT OF CORRELATION (OR PRODUCT MOMENT CORRELATION COEFFICIENT)

The correlation coefficient between two variables x and y, usually denoted by $r(x, y)$ or r_{xy} is a numerical measure of the linear relationship between them and is defined as

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

Note. The correlation coefficient is independent of change of origin and scale.

Let us define two new variables u and v as

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

21.27 COMPUTATION OF THE CORRELATION COEFFICIENT

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

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

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

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

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

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

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

$\therefore$

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

If $u=\dfrac{x-a}{h}, v=\dfrac{y-b}{k}$ then $r_{xy}=r_{uv}=\dfrac{\frac{1}{n}\Sigma u_i v_i-\bar{u}\,\bar{v}}{\sqrt{\left(\frac{1}{n}\Sigma u_i^2-\bar{u}^2\right)\left(\frac{1}{n}\Sigma v_i^2-\bar{v}^2\right)}}$.

ILLUSTRATIVE EXAMPLES

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

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

Calculate the coefficient of correlation.

Sol. Let the grades 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

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

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

$$=\frac{2734}{\sqrt{5398\times 2224}}=0.787$$

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

Example 2. *Find the coefficient of correlation for the following table*:

x :	*10*	*14*	*18*	*22*	*26*	*30*
y :	*18*	*12*	*24*	*6*	*30*	*36*

Sol. Let $u = \dfrac{x-22}{4},\ v = \dfrac{y-24}{6}$.

x	y	u	v	u^2	v^2	uv
10	18	– 3	– 1	9	1	3
14	12	– 2	– 2	4	4	4
18	24	– 1	0	1	0	0
22	6	0	– 3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		–3	–3	19	19	12

$$\bar{u} = \frac{1}{n}\Sigma u_i = \frac{1}{6}(-3) = -\frac{1}{2};\ \bar{v} = \frac{1}{n}\Sigma v_i = \frac{1}{6}(-3) = -\frac{1}{2}$$

$$r_{uv} = \frac{\frac{1}{n}\Sigma u_i v_i - \bar{u}\,\bar{v}}{\sqrt{\left(\frac{1}{n}\Sigma u_i^2 - \bar{u}^2\right)\left(\frac{1}{n}\Sigma v_i^2 - \bar{v}^2\right)}} = \frac{\frac{1}{6}(12) - \frac{1}{4}}{\sqrt{\left[\frac{1}{6}(19) - \frac{1}{4}\right]\left[\frac{1}{6}(19) - \frac{1}{4}\right]}} = 0.6$$

Hence $r_{xy} = r_{uv} = 0.6$.

Example 3. *A computer, while calculating the correlation coefficient between two variables X and Y from 25 pairs of observations, obtained the following results*:

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

It was, however, later discovered at the time of checking that two pairs had been copied incorrectly as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of the correlation coefficient.

Sol.

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

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

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

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

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

Example 4. *If $z = ax + by$ and r is the correlation coefficient 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. $z = ax + by$

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

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

Now $$\sigma_z^2 = \frac{1}{n}\Sigma(z_i - \bar{z})^2 = \frac{1}{n}\Sigma[a(x_i - \bar{x}) + b(y_i - \bar{y})]^2$$

$$= \frac{1}{n}\Sigma\left[a^2(x_i - \bar{x})^2 + b^2(y_i - \bar{y})^2 + 2ab(x_i - \bar{x})(y_i - \bar{y})\right]$$

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

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma y \qquad \left| \because\ r = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y}\right.$$

21.28 CALCULATION OF THE COEFFICIENT 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 the change of origin and scale do not affect the coefficient 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 grades obtained by 100 students in an intelligence test:*

Age (in years) / Grades	*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 coefficient of correlation between age and intelligence.

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

Mid value	y \ x	*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), $i = 1, 2, \ldots, n$ be the ranks of the n individuals in the group for the characteristics A and B respectively. The Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the *rank correlation coefficient* between the characteristics A and B for that group of individuals.

Thus the rank correlation coefficient

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

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.,* $(x_i, y_i) \neq (x_j, y_y)$ *for* $i \neq j$, *both x and y take all integral values from 1 to n.*

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

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

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

If d_i denotes the difference in ranks of the ith individual, then

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

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

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

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

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

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

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

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 grades 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 coefficient.

Sol. *Here the grades 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

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

21.30 REPEATED RANKS

If any two or more individuals have the same rank or the same value in the series of grades, 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 that these individuals would have assumed if they were slightly different from each other. Thus, if two individuals are ranked equal at the sixth place, they would have assumed the 6th and 7th ranks if they were ranked slightly differently. Their common rank = $\frac{6+7}{2}$ = 6.5. If three individuals are ranked equal in fourth place, they would have assumed the 4th, 5th, and 6th ranks if they were ranked slightly differently. 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
$$r = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12} m(m^2 - 1) + \frac{1}{12} m(m^2 - 1) + \ldots\right\}}{n(n^2 - 1)}$$

Example. *Obtain the rank correlation coefficient 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, grades 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 three times. 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.

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

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

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

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 to want to know the extent to which one variable varies in response to a given variation in the other variable; one is interested to know the nature of the relationship between the two variables.

Regression measures the nature and extent of correlation.

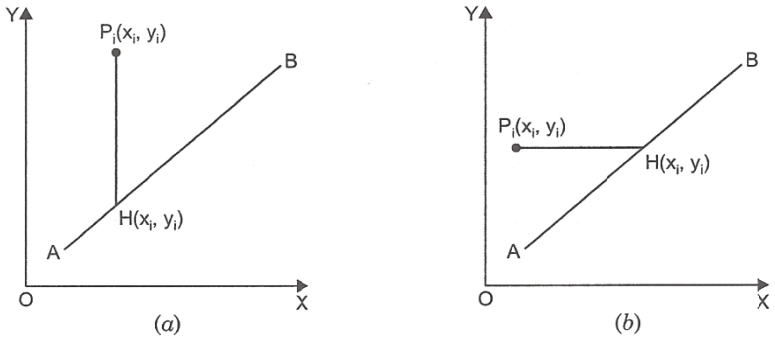
21.32 LINEAR REGRESSION

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 around 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 that 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 minimized [*See* part (*a*) of the figure on the next page], 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 deviation parallel to the axis of *x* is minimized [*See* part (*b*) of the figure on the next page], 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.*



21.33 LINES OF REGRESSION

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

$$y = a + bx \quad \ldots (1)$$

Then $$\bar{y} = a + b\bar{x} \quad \ldots (2)$$

Subtracting (2) from (1), we have

$$y - \bar{y} = b(x - \bar{x}) \quad \ldots (3)$$

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

$$\Sigma yx = a\Sigma x + b\Sigma x^2 \quad \ldots (4)$$

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

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2 \quad \ldots (5)$$

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

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

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

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

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

$\dfrac{r\sigma_x}{\sigma_y}$ is called the regression coefficient 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 the 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

Property I. *The correlation coefficient is the geometric mean between the regression coefficients.*

Proof. The coefficients of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

G.M. between them $= \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r =$ coefficient of correlation.

Property II. *If one of the regression coefficients is greater than 1, the other must be less than 1.*

Proof. The two regression coefficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Let $b_{yx} > 1$, then $\frac{1}{b_{yx}} < 1$ . . . (1)

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

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

Property III. *The arithmetic mean of regression coefficients is greater than the correlation coefficient.*

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

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 coefficients are independent of the origin but not of scale.*

Proof. Let $u = \frac{x-a}{h}, v = \frac{y=b}{k}$ 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, $b_{xy} = \frac{h}{k} b_{uv}$.

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

Property V. *The correlation coefficient and the two regression coefficients have the same sign.*

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

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

Since σ_x and σ_y are both positive, $b_{yx}, b_{xy},$ and r have the 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 of 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 $m_2 = \dfrac{\sigma_y}{r\sigma_x}$.

$\therefore$
$$\tan\theta = \pm\frac{m_2-m_1}{1+m_2m_1} = \pm\frac{\dfrac{\sigma_y}{r\sigma_x}-\dfrac{r\sigma_y}{\sigma_x}}{1+\dfrac{\sigma_y^2}{\sigma_x^2}}$$

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

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

$\therefore$ Positive 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
$$r = 0,\ \theta = \frac{\pi}{2}$$

$\therefore$ 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 a perfect correlation between the two variates x and y.

Note.
$$\frac{r\sigma_x}{\sigma_y} = \frac{\frac{1}{n}\Sigma xy-\bar{x}\,\bar{y}}{\sigma_x\sigma_y}\cdot\frac{\sigma_x}{\sigma_y} = \frac{\frac{1}{n}\Sigma xy-\bar{x}\,\bar{y}}{\sigma_y^2} = \frac{\frac{1}{n}\Sigma xy-\bar{x}\,\bar{y}}{\frac{1}{n}\Sigma y^2-\bar{y}^2}$$

Similarly,
$$\frac{r\sigma_y}{\sigma_x} = \frac{\frac{1}{n}\Sigma xy-\bar{x}\,\bar{y}}{\frac{1}{n}\Sigma x^2-\bar{x}^2}.$$

ILLUSTRATIVE EXAMPLES

Example 1. *Calculate the coefficient 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 that should correspond on the average to x = 6.2.

Sol.

x	y	$u = x - 5$	$u = y - 12$	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 $x = 6.2$, the 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, only the following results are legible*:

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 coefficient 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

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

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

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

Subtracting (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 $\bar{x} = 13, \quad \bar{y} = 17 \quad \ldots (a)$

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

$\therefore$ $\sigma_x = 3$

The equations of the lines of regression can be written as

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

$\therefore$ The regression coefficient of y on x is $\dfrac{r\sigma_y}{\sigma_x} = .8 \quad \ldots (4)$

The regression coefficient of x on y is $\dfrac{r\sigma_x}{\sigma_y} = .45 \quad \ldots (5)$

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

(Positive sign with square root is taken because regression coefficients are positive.)

From (4), $\sigma_y = \dfrac{.8\sigma_x}{r} = \dfrac{.8 \times 3}{0.6} = 4. \quad \ldots (c)$

TEST YOUR KNOWLEDGE

1. (*a*) Calculate the correlation coefficient for the following heights in inches of fathers (X) and their sons (Y).

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

(*b*) Find the correlation coefficient between x and y from the given data:

x:	78	89	97	69	59	79	68	57
y:	125	137	156	112	107	138	123	108

(*c*) Find the correlation coefficient from the following data:

x:	92	89	87	86	83	77	71	63	53	50
y:	86	88	91	77	68	85	52	82	37	57