Correlation

Definition:- The study of existance magnitude and direction of variation between two or more variables is called correlation

Types of correlation:-

i) Positive and Negative correlation

If both variables change in the same direction (i.e. both increases or decreases) the correlation is called positive

e.g. advertising and sales. If one variable increases other decreases or vice versa then it is negative corelation e.g. T.V. registration and cinema attendance

ii) Linear and non-linear correlation:

If the graph of two variables is straight line it is linear. If graph is not straight line but a curve it is non linear

Definition: Co-variance between x & y

$$\therefore \operatorname{cov}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n}$$

$$\bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}$$

$$\operatorname{cov}(x,y) = \frac{\sum xy}{n} - \bar{x}\,\bar{y}$$

Definition : correlation coefficient

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

where $\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$, $\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$

$$\therefore \boxed{r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n \cdot \sigma_x \sigma_y}}$$

OR

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sigma_x \cdot \sigma_y}$$

OR

$$r = \frac{\frac{\Sigma xy}{N} - \frac{\Sigma x}{N} \cdot \frac{\Sigma y}{N}}{\sqrt{\frac{\Sigma x}{N} - \left(\frac{\Sigma x}{N}\right)^2} \sqrt{\frac{\Sigma y}{N} - \left(\frac{\Sigma y}{N}\right)^2}}$$

where x & y are actual values

$N \rightarrow$ No. of observations.

Remark ① Correlation coefficient is independent of change of origin & change of scale

② Correlation coefficient is denoted by r and it always lies betn -1 to $+1$

$$-1 \leq r \leq 1$$

Ex. ① Find co-rrelation coefficient between two variables for the following data.

x:	24	27	32	40	45	48	52
Y:	30	26	23	14	11	09	07

Sol^n.

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

where $\sigma_x = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$, $\sigma_y = \sqrt{\frac{\Sigma (y-\bar{y})^2}{n}}$

Now prepare table

x	Y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
24	30	-14.29	204.2	12.86	165.38	-183.77
27	26	-11.29	127.46	8.86	78.49	-100.03
32	23	-6.29	39.56	5.86	34.34	-36.86
40	14	1.71	2.93	-3.14	9.86	-5.37
45	11	6.71	45.02	-6.14	37.69	-41.19
48	9	9.71	94.28	-8.14	66.26	-79.04
52	7	13.71	187.96	-10.14	102.82	-139.02

$\bar{x} = \frac{\Sigma x}{n} = \frac{268}{7} = 38.29$

$\bar{y} = \frac{\Sigma y}{n} = 17.14$

$\Sigma(x-\bar{x})^2 = 701.41$

$\Sigma(y-\bar{y})^2 = 494.84$

$\Sigma(x-\bar{x})(y-\bar{y}) = -585.28$

$$\therefore r = \frac{-585.28}{7 \cdot \sqrt{\frac{701.41}{7}} \sqrt{\frac{494.84}{7}}} = -0.99$$

-ve correlation

Ex. ② obtain coefficient of correlation for the following data

x : 12 17 22 27 32
y : 113 119 117 115 121

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
12	113	-10	100	-4	16	40
17	119	-5	25	2	4	-10
22	117	0	0	0	0	0
27	115	5	25	-2	4	-10
32	121	10	100	4	16	40

$\bar{x} = \frac{\Sigma x}{n} = \frac{110}{5} = 22$ $\bar{y} = \frac{\Sigma y}{n} = \frac{586}{n} = 117$

$\Sigma(x-\bar{x})^2 = 250$, $\Sigma(y-\bar{y})^2 = 40$

$\Sigma(x-\bar{x})\cdot(y-\bar{y}) = 60$

coefficient of correlation $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y}$

$= \frac{60}{5\sqrt{\frac{250}{5}}\sqrt{\frac{40}{5}}}$

$= 0.6$

conclusion: Two variables are correlated with degree 0.6

Ex ③ calculate the coefficient of correlation for the following data

x: 105, 104, 102, 101, 100, 99, 98, 96, 93, 92
y: 101, 103, 100, 98, 95, 96, 104, 92, 97, 94

sol^n: We know that, coefficient is given by

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

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})\cdot(y-\bar{y})$
105	101	6	36	3	9	18
104	103	5	25	5	25	25
102	100	3	9	2	4	6
101	98	2	4	0	0	0
100	95	1	1	-3	9	-3
99	96	0	0	-2	4	0
98	104	-1	1	6	36	-6
96	92	-3	9	-6	36	18
93	97	-6	36	-1	1	6
92	94	-7	49	-4	16	28

$\bar{x} = 99, \quad \bar{y} = 98 \qquad \sum(x-\bar{x})^2 = 170$

$\sum(y-\bar{y})^2 = 140 \qquad \sum(x-\bar{x})(y-\bar{y}) = 92$

$\sigma_x = \sqrt{\frac{170}{10}} = 4.12, \quad \sigma_y = \sqrt{\frac{140}{10}} = 3.74$

$$r = \frac{92}{10 \times 4.12 \times 3.74} = 0.59$$

Ex. A computer while calculating the correlation coefficient between the x and gave the following results.

$n = 30$, $\Sigma x = 120$, $\Sigma y = 90$, $\Sigma x^2 = 600$, $\Sigma y^2 = ...$

$\Sigma xy = 356$. It was later found that the computer had copied down two pairs of observations as.

x	y		x	y
8	10	instead of	8	12
12	7		10	8

Find the correct value of coefficient of correlation r

Solⁿ:- Σx = 120 − (sum of incorrect values) + (sum of correct values)

$= 120 - (20) + 18 = 118$

Correct value of $\bar{x} = \frac{\Sigma x}{n} = \frac{118}{30} = 3.93$

Correct value of $\Sigma y = 90 - (10+7) + (12+8)$
$= 93$

∴ Correct value of $\bar{y} = \frac{\Sigma y}{n} = \frac{93}{30} = 3.1$

Correct value of $\Sigma xy =$

$= 356 - [8 \times 10 + 12 \times 7] + [8 \times 12 + 10 \times 8]$
$= 368$

correct Value of $\Sigma x^2 = 600 - [8^2 + 12^2] + [8^2 + 10^2]$

$= 556$

correct Value of $\Sigma y^2 = 250 - (10^2 + 7^2) + (12^2 + 8^2)$

$= 309$

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

$$= \sqrt{\frac{556}{30} - (3.93)^2} = \sqrt{3.068}$$
$$= 1.76$$

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

$$= \sqrt{\frac{309}{30} - (3.1)^2} = \sqrt{0.69} = 0.8506$$

$$r = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

$$= \frac{\frac{368}{30} - (3.93)(3.1)}{1.76 \times 0.8506}$$

$$= \frac{\cancel{0}\ 12.26 - 12.183}{1.4970} = \cancel{0.051}$$