Q: Find if there is any significant correlation between the heights and weights given below.

Height in inches	57	59	62	63	64	65	55	58	57
Weight in lbs	113	117	126	126	130	129	111	116	112

Solution :

Height in inches x	Deviation from Mean (60) $X = x - \overline{x}$	Square of deviations X^2	Weight in lbs y	Deviations from Mean $Y = y - \overline{y}$	Square of deviations Y^2	Product of deviations of X and Y series (XY)
57	−3	9	113	−7	49	21
59	−1	1	117	−3	9	3
62	2	4	126	6	36	12
63	3	9	126	6	36	18
64	4	16	130	10	100	40
65	5	25	129	9	81	45
55	−5	25	111	−9	81	45
58	−2	4	116	−4	16	8
57	−3	9	112	−8	64	24
540	0	102	1080	0	472	216

Coefficient of correlation $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \times \Sigma Y^2}}$

$$\therefore \quad r = \frac{216}{\sqrt{102 \times 471}} = 0.98$$

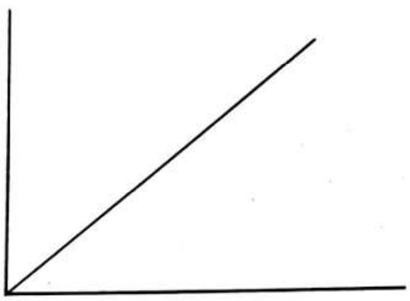
Note : Limits for correlation coefficient are $-1 \le r \le 1$.

Hence correlation coefficient can not exceed one numerically.

If $r = 1$ correlation is perfect and positive.

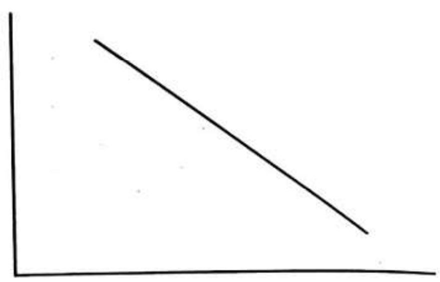
If $r = -1$ correlation is perfect and negative. If $r = 0$, then there is no relationship between the variables.

Perfect positive correlation



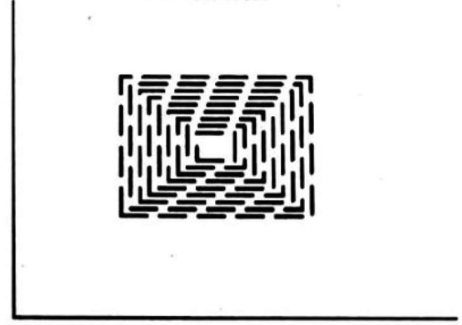
$r = +1$

Perfect Negative correlation

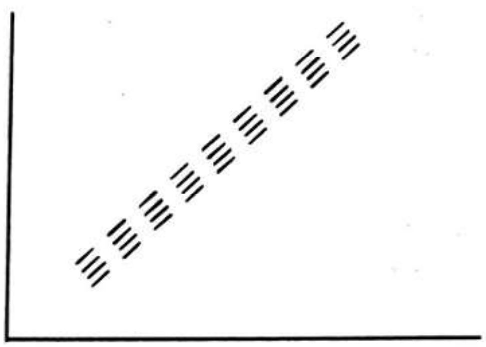


$r = -1$

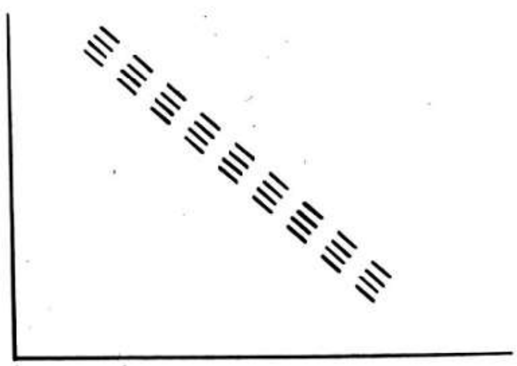
No correlation



High degree of positive correlation



High degree of negitive correlation



RANK CORRELATION COEFFICIENT

A British Psychologist Charles Adward Spearman found out the method of finding the coefficient of correlation by ranks. This method is based on rank and is useful in dealing with qualitative characteristics such as morality, character, intelligence and beauty. It can not be measured quantitatively as in the case of Pearson's coefficient of correlation. It is based on the ranks given to the observations. Rank correlation is applicable only to the individual observations. The formula for Spearman's rank correlation is given by

$$\rho = 1 - \frac{6\,\Sigma D^2}{N(N^2 - 1)}$$

where $\rho \rightarrow$ Rank coefficient of correlation

$D^2 \rightarrow$ Sum of the squares of the differences of two ranks.

$N \rightarrow$ Number of paired observations

PROPERTIES OF RANK CORRELATION COEFFICIENT

1. The value of ρ lies between +1 and −1
2. If $\rho = 1$, there is complete agreement in the order if the ranks and the direction of the rank is same.
3. If $\rho = -1$ then there is complete disagreement in the order of the ranks and they are in opposite directions.

Procedure to solve problems :

1. When the ranks are given.

Step. 1. Compute the difference of two ranks and denote it by D.

Step. 2. Square D and get ΣD^2

Step. 3. Obtain ρ by substituting the figures in the formula.

2. When the ranks are not given, but actual data are given, then we must give ranks. We can give ranks by taking the highest as 1 or the lowest value as 1, next to the highest (lowest) as 2 and follow the same procedure for both the variables.

Q: Following are the rank obtained by 10 students in two subjects, Statistics and Mathematics. To what extent the knowledge of the students in two subjects is related ?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Solution :

Rank in statistics (x)	Rank in Mathematics (y)	$D = (x - y)$	D^2
1	2	−1	1
2	4	−2	4
3	1	+2	4
4	5	−1	1
5	3	+2	4
6	9	−3	9
7	7	0	0
8	10	−2	4
9	6	+3	9
10	8	+2	4
			$\Sigma D^2 = 40$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6 \times 40}{10(10^2-1)} = 1 - \frac{240}{10(100-1)} = 1 - \frac{240}{990} = 1 - 0.24 = 0.76$$

Q: From the following data calculate the rank correlation coefficient after making adjustment for tied ranks.

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution : First we have to assign ranks to the variables.

X	Rank (x)	Y	Rank (y)	$D = x - y$	D^2
48	3 8	13	5.5 5.5	2.5	6.25
33	5 6	13	5.5 5.5	0.5	0.25
40	4 7	24	1 10	−3	9.00
9	10 1	6	8.5 2.5	−1.5	2.25
16	8 3	15	4 7	4	16.00
16	8 3	4	10 1	2	4.00
65	1 10	20	2 9	1	1.00
24	6 5	9	7 4	1	1.00
16	8 3	6	8.5 2.5	5	0.25
57	2 9	19	3 8	1	1.00
					$\Sigma D^2 = 41$

16 is repeated 3 times in X items hence $m = 3$. Since 13 and 6 are repeated twice in Y items; hence $m = 2$.

$$\therefore \rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m^2 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right]}{N^3 - N}$$

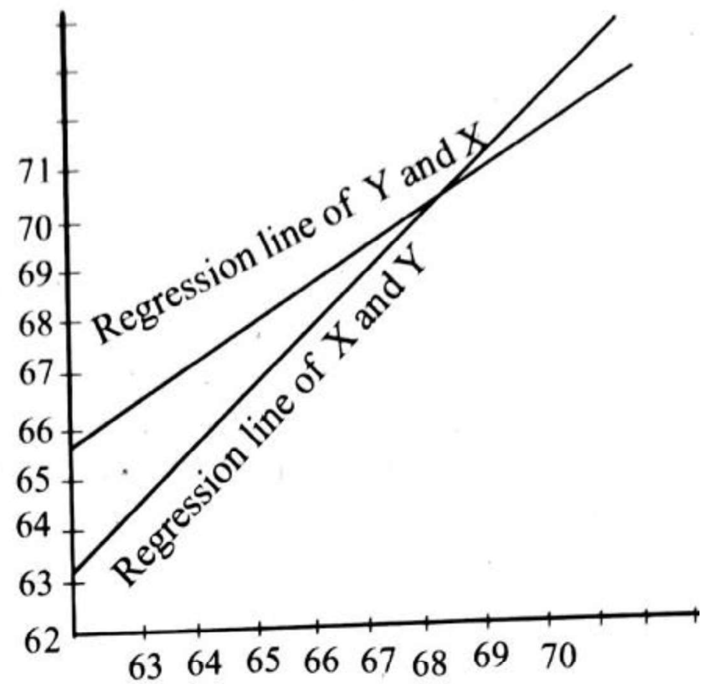
$$= 1 - \frac{6\left[41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{990}$$

$$= +0.733$$

REGRESSION

e.g. Fit a regression line on the scatter diagram for the following data.

X	Y
65	68
67	68
62	66
70	68
67	67
69	68
71	70



Regression Line : A regression line is a straight line fitted to the data by the method of least squares. It indicates the best possible mean value of one variable corresponding to the mean value of the other. There are always two regression lines constructed for the relationship between two variables X and Y. Thus one regression line shows the regression of X upon Y and the other shows the regression of Y on X.

Regression equation of Y on X :

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\,\Sigma X + b\,\Sigma X^2$$

Regression equation of X on Y :

Normal equations are

$$\Sigma X = Na + b\,\Sigma Y$$

$$\Sigma XY = a\,\Sigma Y + b\,\Sigma Y^2$$

Q: Determine the equation of a staight line which best fits the data.

X:	10	12	13	16	17	20	25
Y:	10	22	24	27	29	33	37

Solution : Straight line is $Y = a + bx$ The two normal equations are $\Sigma Y = b\,\Sigma X + N\,a$

$$\Sigma XY = b\,\Sigma X^2 + a\,\Sigma X$$

X	X^2	Y	XY
10	100	10	100
12	144	22	264
13	169	24	312
16	256	27	432
17	289	29	493
20	400	33	660
25	625	37	925
$\Sigma X = 113$	$\Sigma X^2 = 1938$	$\Sigma Y = 182$	$\Sigma XY = 3186$

Substituting the values, we get

$$113b + 7a = 182 \qquad \ldots (1)$$

$$1983b + 113a = 3186 \qquad \ldots. (2)$$

Solving (1) and (2) , we get $a = 0.82, b = 1.56$

Thus the equation of the straight line is $\mathrm{Y} = a + b\mathrm{X}$

$$\therefore \ \mathrm{Y} = 0.82 + 1.56\ \mathrm{X}$$

This is called the regression equation of Y on X.

THANK YOU