UNIT-IV

Correlation

Introduction:

Correlation is the relationship that exists b/w two (or) more Variable i.e., if two variables are related one is changed the other will automatically be changed.

ex: Heights & weights, Demand & Supply, price & Demand Etc.,

Definition:

"Correlation analysis is a statistical technique used to measure the degree and direction of relationship between Variables".

Significance of Correlation:

→ Economic theory and business studies show relationship between variables.

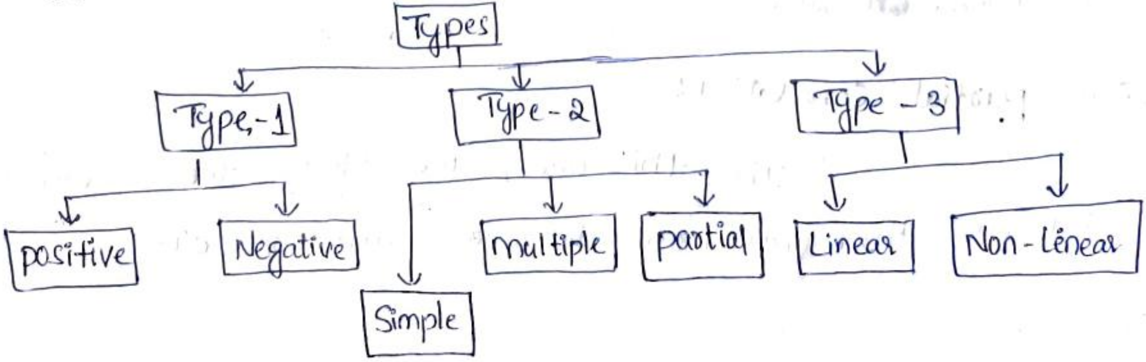
→ Correlation analysis helps in deriving exactly the degree and direction of such relationships.

→ Correlation analysis contributes to the understanding of Economic behaviour.

→ The measure of co-efficient of correlation is relative measure of change.

→ The Effect of correlation is to reduce range of uncertainity of our prediction.

Types of Correlation:



01. **positive correlation :**

If both the variables vary in same direction is called positive correlation i.e., if one variable increases, the other also increases (or) if one variable decreases the other also decrease.

Ex :-

x	10	20	30	40	50
y	1	2	3	4	5

02. **Negative correlation :**

If both the variables vary in opposite direction, i.e., called Negative Correlation i.e., if one variable increases, the other variable decreases. (or) if one variable decreases, the other variable increases.

Ex :-

x	10	20	30	40	50
y	10	8	6	4	2

03. **Simple correlation :**

When only two variables are studied i.e., the number of variables are given, choose relationship b/w only two variables.

04. **Multiple correlation :**

When three (or) more variables are studied i.e., n variables are given, relationship b/w more than two variables. etc.

05. **partial correlation :**

In this case, we study the relation between three (or) more variables and not all.

06. **Total multiple Correlation:**

In this Case, we study the relationship between three (or) more variables it may be all also.

07. **Linear correlation:**

If the amount of changes in One variable a constant ratio to the amaunt of changes in other Variable then the Correlation is said to be linear.

Ex:

x	10	20	30	40	50
y	2	4	6	8	10

08. **Non-Linear Correlation:**

If the amount of change in One Variable does not constant ratio to the amount of change the other variable then the correlation is said to be non-linear.

Ex:-

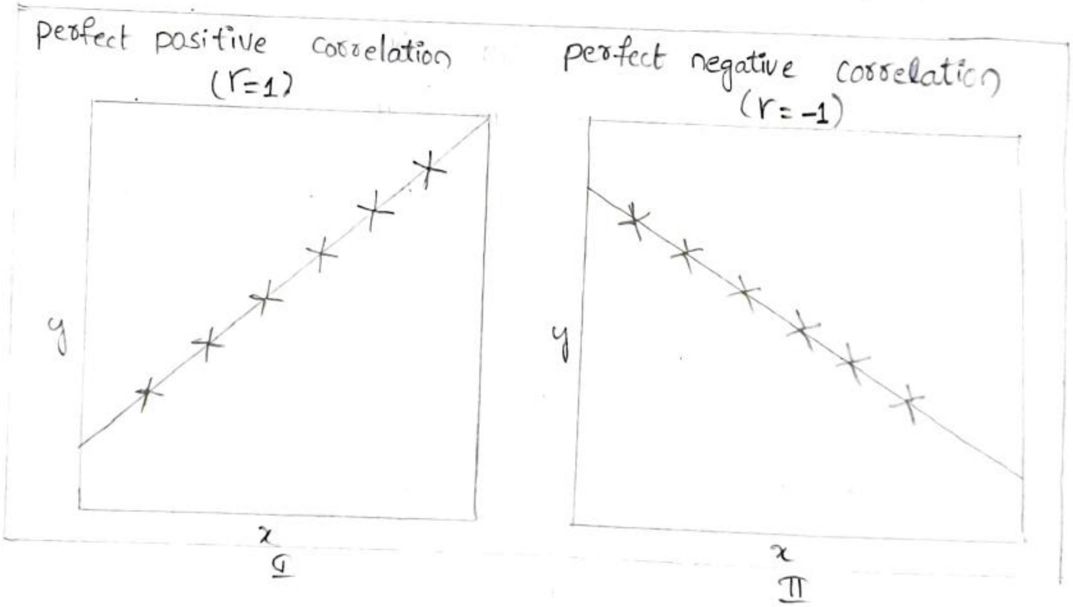
x	10	20	30	40	50
y	2	3	5	10	11

Methods of studying correlation :

The various methods of ascertaining whether two Variables are correlated (or) not are:

01. Scatter Diagram method.
02. Graphic method
03. Karl pearson's co-efficient of correlation
04. Concurrent Deviation method.
05. Method of Least Squares.

of these, The first two are based on the Knowledge of diagrams and graphs. whereas the others are the mathematical methods. Each of these methods shall be discussed in detail in the following pages.


perfect positive correlation (r=1)

perfect negative correlation (r=-1)

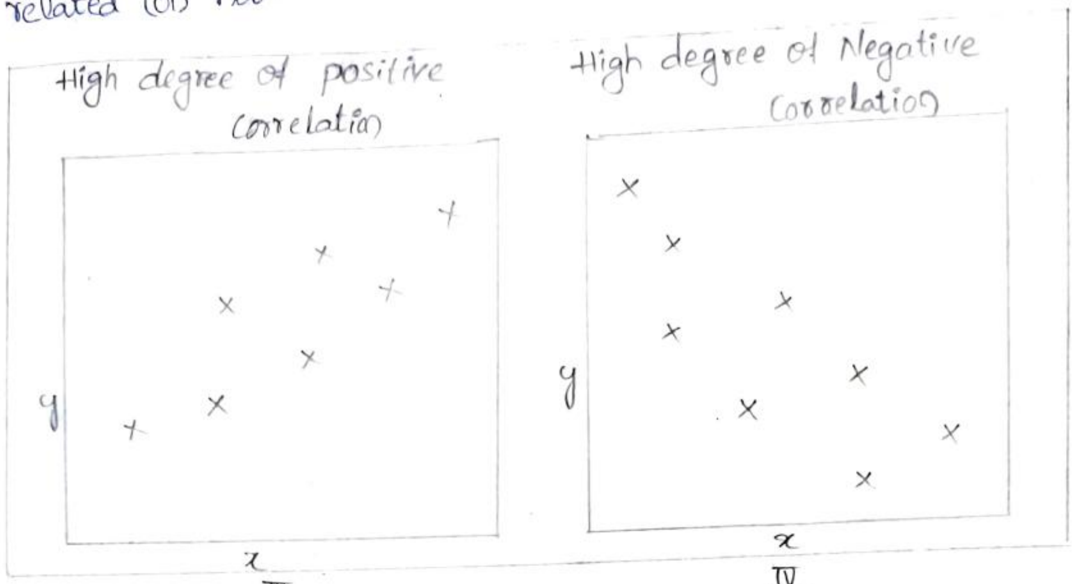
I

II

01. Scatter Diagram Method:

The Simplest device ascertaining whether two Variables are related is to prepare a dot chart called Scatter diagram. When this method is used the given data are plotted on a graph paper (or) simply Scatter plot in the form of dots. i.e., for each pair of X and Y values we put a dot and thus obtain as many points as the number of

Observations. By looking to the scatter of the various points we can form an idea as to whether the variables are - related (or) not.

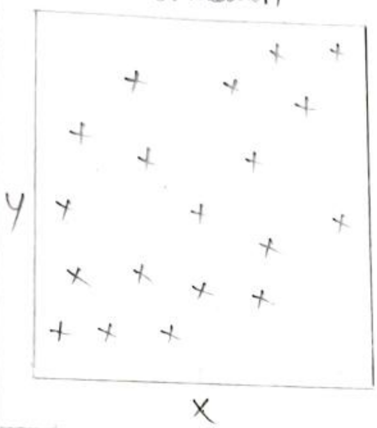


III

IV

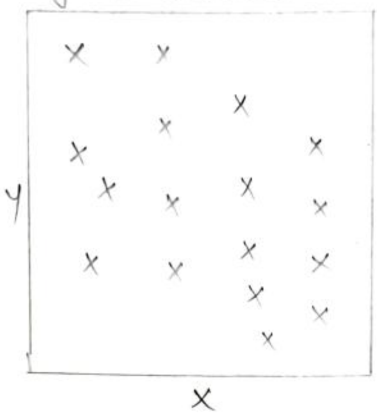
The greater scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line falling from the lower left-hand corner to the upper right-hand corner. Correlation is said to be perfectly positive ($r=+1$). On the other hand, If all the points are lying on a straight line rising from the upper left-hand corner to the lower right-hand corner of the diagram. Correlation is said to be perfectly negative ($r=-1$). If the plotted points fall in a narrow band there would be a high degree of correlation between the variables. Correlation shall be positive. If the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram-III) and negative if the points show a declining tendency from the upper left-hand corner to the lower right-hand corner of the diagram (diagram-IV). On the other hand, If the points are widely scattered over the diagram it indicates very little relationship between the variables - Correlation shall be positive if the points are rising from the lower left-hand corner to the upper right-hand corner (diagram(V)) and negative if the points are running from the upper left-hand side to the lower right-hand side of the diagram (diagram-VI). If the plotted points lie on a straight line parallel to the x-axis (or) in a haphazard manner. It shows absence of any relationship between the variables. ($r=0$) as shown in diagram VII.

Low degree of Positive Correlation



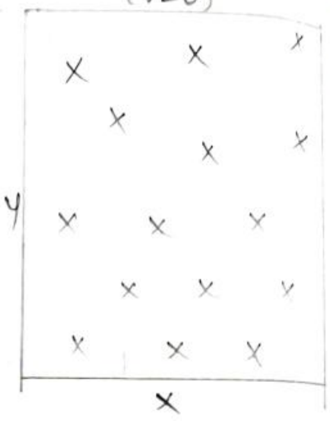
V

Low degree of Negative correlation



VI

No correlation ($r=0$)



VII

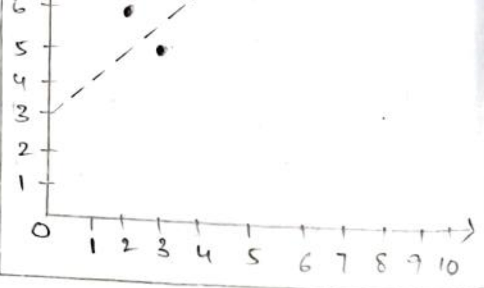
Illustration-1: Given the following pairs of values of the variables X and Y:

X:	2	3	5	6	8	9
Y:	6	5	7	8	12	11

a) Make a scatter diagram.

b) Is there any correlation between the variables X and Y?

c) By Graphic inspection, draw an estimating line.



Sol:- By looking at the scatter diagram we can say that the variables X and Y are correlated. Further, correlation is positive because the trend of the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is higher because the plotted points are near to the line which shows perfect relationship between the variables.

Merit and Demerits of the Method :

Merits : Following are the merits of scatter diagram method:

→ It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.

→ It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.

→ Making a scatter diagram usually is the ~~st~~ first step in investigating the relationship between two variables.

→ If the variables are related, we can see what kind of line, (or) estimating equation describes this relationship.

Demerits :

By applying this method we can get an idea about the direction of correlation and also whether it is high (or) low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

02. Graphic method :

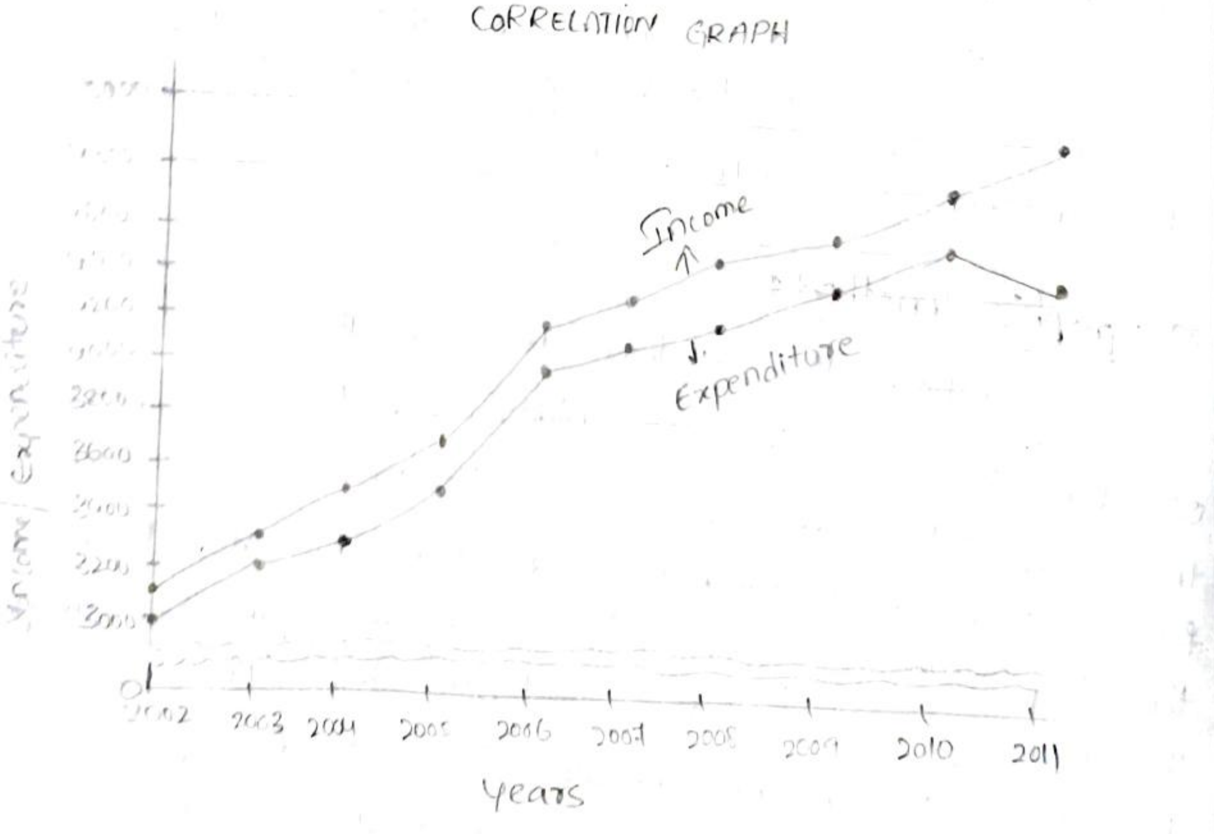
When this method is used the individual values of the two variables are plotted on the graph paper. We thus obtain two curves. One for X variable and another for Y variable. By examining the direction and closeness of the two curves so drawn we can infer whether (or) not the variables are related. If both the curves drawn on the graph are moving in the same direction (either upward (or) downward) correlation is said to be positive. On the other hand, If the curve are moving in the opposite directions correlation is said to be negative. The following example shall illustrate the method.

→ An estimating line (or) regression line is a line of average relationship. For details please see next chapter on 'Regression analysis'.

Illustration -2 : From the following data ascertain whether the income and expenditure of the 100 workers of a factory are correlated.

Year	Average income (in ₹)	Average Expenditure (in ₹)	Year	Average income (in ₹)	Average expenditure (in ₹)
2002	3100	3000	2007	4300	4100
2003	3320	3200	2008	4500	4200
2004	3500	3350	2009	4650	4400
2005	3700	3500	2010	4800	4650
2006	4200	4000	2011	5000	4500

Sol :- The following graph shows that the variables, income and expenditure, are closely related.



This method is normally used where we are given data over a period of time, i.e., in case of time series. However, as with the Scatter diagram method, in this method also we cannot get a numerical value describing the extent to which the variables are related.

03. Karl pearson's co-efficient of correlation:

Of the Several mathematical methods of measuring correlation, the karl pearson's method, popularly known as pearson's co-efficient of correlation, is the most widely used in practice. The pearson co-efficient of correlation is denoted by the symbol r. It is one of the very few symbols that are used universally for describing the degree of correlation between two Series. The formula for computing pearsonian r is:

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

where, $x = (X - \overline{X})$

$y = (Y - \overline{Y})$

σ_x = Standard deviation of Series X.

σ_y = Standard deviation of Series Y.

N = Number of pairs of observations

r = The (product moment) correlation co-efficient.

This method is to be applied only where deviations of items are taken from actual mean and not from assumed mean.

The value of co-efficient of correlation as obtained by the above formula shall always lie between 1k when r=+1, it means there is perfect positive correlation between the variables. when r=-1, It means there is perfect negative correlation between variables. when r=0, it means there is no relationship between the two variables. However, in practice such values of r as +1, -1 and 0 are rare. we normally get values which lie between

This method is normally used where we are given data over a period of time, i.e., in case of time series. However, as with the Scatter diagram method, in this method also we cannot get a numerical value describing the extent to which the variables are related.

03. Karl pearson's co-efficient of correlation:

Of the several mathematical methods of measuring correlation, the karl pearson's method, popularly known as pearson's co-efficient of correlation, is the most widely used in practice. The pearson co-efficient of correlation is denoted by the symbol r. It is one of the very few symbols that are used universally for describing the degree of correlation between two series. The formula for computing pearsonian r is:

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

where, $x = (X - \overline{X})$

$y = (Y - \overline{Y})$

σ_x = Standard deviation of series X.

σ_y = Standard deviation of series Y.

N = Number of pairs of observations

r = The (product moment) correlation co-efficient.

This method is to be applied only where deviations of items are taken from actual mean and not from assumed mean.

The value of co-efficient of correlation as obtained by the above formula shall always lie between 1 & when r=+1, it means there is perfect positive correlation between the variables. when r=-1, It means there is perfect negative correlation between variables. when r = 0, it means there is no relationship between the two variables. However, in practice such values of r as +1, -1 and 0 are rare. we normally get values which lie between +1 and -1 sunch as +0.8, -0.20, etc., The co-efficient of correlation describes not only the magnitude of correlation but also its direction. Thus, + 0.8 would mean that correlation is positive because the sign of r is + and the magnitude of correlation is 0.8 similarly −0.20 means low degree of negative correlation.

The above formula for computing pearson's co-efficient of correlation can be transformed to the following form which is easier to apply.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

where,
$$x = (X - \bar{X})$$
$$y = (Y - \bar{Y})$$

It is obvious that while applying this formula we have not to calculate seperately the standard deviation of X and Y series as is required by formula (i). This simplifies greatly the task of calculating correlation co-efficient.

Steps:

→ Take the deviations of X series from the mean of X and denote these deviations by x.

→ Square these deviations and obtain the total, i.e., Σx^2

→ Take the deviations of Y series from the mean of Y and denote these deviations by y.

→ Square these deviations and obtain the total, i.e., Σy^2.

→ multiply the deviations of X and Y series and obtain the total, i.e., Σxy.

→ Substitute the values of Σxy, Σx^2 and Σy^2 in the above formula.

The following examples will illustrate the procedure:

Illustration - 3: Calculate Karl pearson's co-efficient of correlation from the following data and interprete the values: 250%

→ The Co-efficient of correlation is said to be a measure of covariance between two series. The covariance of two series x and y is written as :

$$\text{Covariance} = \frac{\Sigma xy}{N}$$

where x and y stands for deviations of x and y series from their respective means.

In order to find out the value of correlation - Co-efficient, first we calculate covariance and then in order to convert it to a relative measure we divide the covariance by the standard deviation of the two series. the ratio so obtained is called Karl pearson's Co-efficient.

$$r = \frac{\Sigma xy}{N} \; \sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \; , \; \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$$

$$r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \times \frac{\Sigma y^2}{N}}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Rollno. of students	1	2	3	4	5
marks in accountancy(x)	48	35	17	23	47
marks in statistics(y)	45	20	40	25	45

Sol:- Let marks in Accountancy be denoted by x and marks in statistics by y.

Roll no.	X	$(X-\bar{X})$ x $(\bar{x}=34)$	x^2	Y	$(Y-\bar{Y})$ y $(\bar{Y}=35)$	y^2	xy
1	48	14	196	45	10	100	140
2	35	1	1	20	−15	225	−15
3	17	−17	289	40	5	25	−85
4	23	−11	121	25	−10	100	110
5	47	13	169	45	10	100	130
	$\Sigma X = 170$	$\Sigma x = 0$	$\Sigma x^2 = 776$	$\Sigma Y = 175$	$\Sigma y = 0$	$\Sigma y^2 = 550$	$\Sigma xy = 280$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{48+35+17+23+47}{5} = \frac{170}{5} = 34$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{45+20+40+25+45}{5} = \frac{175}{5} = 35$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{280}{\sqrt{776 \times 550}} = \frac{280}{653.299} = 0.429$$

Illustration : 4 : Making use of the data summarized below, Calculate the co-efficient of correlation.

Case	A	B	C	D	E	F	G	H
X_1	10	6	9	10	12	13	11	9
X_2	9	4	6	9	11	13	8	4

Sol:-

Calculation of coefficient of correlation.

Case	X_1	x_1 (X_1-10)	x_1^2	X_2	(X_2-8) x_2	x_2^2	x_1x_2
A	10	0	0	9	1	1	0
B	6	-4	16	4	-4	16	16
C	9	-1	1	6	-2	4	2
D	10	0	0	9	1	1	0
E	12	2	4	11	3	9	6
F	13	3	9	13	5	25	15
G	11	1	1	8	0	0	0
H	9	-1	1	4	-4	16	4
N=8	$\Sigma X_1=80$	$\Sigma x_1=0$	$\Sigma x_1^2=32$	$\Sigma X_2=64$	$\Sigma x_2=0$	$\Sigma x_2^2=72$	$\Sigma x_1x_2=43$

$$\bar{X} = \frac{\Sigma X_1}{N} = \frac{10+6+9+10+12+13+11+9}{8} = \frac{80}{8} = 10$$

$$\bar{X} = \frac{\Sigma X_2}{N} = \frac{64}{8} = 8$$

$$r = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}}$$

$$= \frac{43}{\sqrt{32 \times 72}} = \frac{43}{48} = 0.896$$

<u>Illustration-5</u>: The following table gives Indices of Industrial production of registered unemployed (in hundred thousand). Calculate the Value of the Co-efficient of correlation.

Year	2004	2005	2006	2007	2008	2009	2010	2011
Index of production	100	102	104	107	105	112	103	99
No. unemployed.	15	12	13	11	12	12	19	26

<u>Sol</u>:- Calculation of Karl pearson's Correlation of Co-efficient.

Year	X	x (X-104)	x^2	Y	y (Y-15)	y^2	xy
2004	100	-4	16	15	0	0	0
2005	102	-2	4	12	-3	9	6
2006	104	0	0	13	-2	4	0
2007	107	+3	9	11	-4	16	-12
2008	105	1	1	12	-3	9	-3
2009	112	8	64	12	-3	9	-24
2010	103	-1	1	19	+4	16	-4
2011	99	-5	25	26	11	121	-55
	$\Sigma X = 832$	$\Sigma x = 0$	$\Sigma x^2 = 120$	$\Sigma Y = 120$	$\Sigma y = 0$	$\Sigma y^2 = 184$	$\Sigma xy = -92$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{832}{8} = 104, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{120}{8} = 15$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-92}{\sqrt{120 \times 184}} = \frac{-92}{148.593} = -0.619$$

Direct method of finding out correlation Co-efficient:

Correlation Co-efficient can also be calculated without taking deviations of items either from actual mean (or) assumed mean. i.e., actual X and Y values.

The formula in such a case is:

$$r = \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2}\sqrt{N \Sigma Y^2 - (\Sigma Y)^2}}$$

The formula would give the same answer as we get when deviations of items are taken from actual mean (or) assumed mean. The following example shall illustrate the point.

Illustration - 6: Calculate Co-efficient of correlation from the data given below by the direct method. i.e., without taking the deviations of items from actual (or) assumed mean.

X	9	8	7	6	5	4	3	2	1
Y	15	16	14	13	11	12	10	8	9

Sol:- Calculation of co-efficient of correlation (Direct method)

X	X^2	Y	Y^2	XY
9	81	15	225	135
8	64	16	256	128
7	49	14	196	98
6	36	13	169	78
5	25	11	121	55
4	16	12	144	48
3	9	10	100	30
2	4	8	64	16
1	1	9	81	9
$\Sigma X = 45$	$\Sigma X^2 = 285$	$\Sigma Y = 108$	$\Sigma Y^2 = 1,356$	$\Sigma XY = 597$

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{N\sum x^2 - (\sum x)^2}\ \sqrt{N\sum y^2 - (\sum y)^2}}$$

$N = 9$
$\sum xy = 597$
$\sum x = 45$
$\sum y = 108$
$\sum x^2 = 285$
$\sum y^2 = 1356$

$$= \frac{9\times 597 - (45\times 108)}{\sqrt{9\times 285 - (45)^2}\ \sqrt{9\times 1356 - (108)^2}}$$

$$= \frac{5373 - 4860}{\sqrt{2565 - 2025}\ \sqrt{12204 - 11664}}$$

$$= \frac{513}{\sqrt{540}\times\sqrt{540}} = \frac{513}{23.23\times 23.23}$$

$$= \frac{513}{540} = 0.95$$

Calculation of correlation co-efficient when change of Scale and origin is made.

Since r is a pure number, shifting the origin and changing the scale of series does not affect its values.

Illustration-7: Calculate co-efficient of correlation from the following data.

X	100	200	300	400	500	600	700
Y	30	50	60	80	100	110	130

Sol:- To simplify calculation let every value of X be divided by 100 and every value of Y by 10 and denote these series by X' and Y'.

X	$X' = \left(\frac{X}{100}\right)$	x $(X'-4)$	x^2	Y	$Y' = \left(\frac{Y}{10}\right)$	y $(Y'-8)$	y^2	xy
100	1	−3	9	30	3	−5	25	15
200	2	−2	4	50	5	−3	9	6
300	3	−1	1	60	6	−2	4	2
400	4	0	0	80	8	0	0	0
500	5	+1	1	100	10	+2	4	2
600	6	+2	4	110	11	+3	9	6
700	7	+3	9	130	13	+5	25	15
	$\sum X' = 28$	$\sum x = 0$	$\sum x^2 = 28$		$\sum Y' = 56$	$\sum y = 0$	$\sum y^2 = 76$	$\sum xy = 46$

$$\bar{X} = \frac{\Sigma X'}{N} = \frac{28}{7} = 4, \quad \bar{Y} = \frac{\Sigma Y'}{N} = \frac{56}{7} = 8$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{46}{\sqrt{28 \times 76}} = 0.997$$

When Deviations are taken from an Assumed Mean:

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such a case, we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable.

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x) \times (\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

where, d_x refers to deviations of X series from an assumed mean, i.e., $(X - \bar{X})$

Similarly, d_y refers to deviations of Y series from an assumed mean i.e., $(Y - \bar{Y})$.

$\Sigma d_x d_y$ = Sum of the product of the deviations of X and Y series from their assumed means.

Σd_x^2 = Sum of the squares of the deviations of X series from an assumed mean.

Σd_y^2 = Sum of squares of the deviations of Y series from an assumed mean.

Σd_x = Sum of the deviations of X series from an assumed mean

Σd_y = Sum of the deviations of Y series from an assumed mean.

It may be pointed out that there are many variations of the above formula. For example, the above formula may be written as :

$$r = \frac{N\Sigma d_x d_y - \{(\Sigma d_x) \times (\Sigma d_y)\}}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

But the formula given above is the easiest to apply.

Steps :

→ Take the deviations of X series from an assumed mean and denote these deviations by d_x and obtain the total, i.e., Σd_x.

→ Take the deviations of Y series from an assumed mean and denote these deviations by d_y and obtain the total, i.e., Σd_y.

→ Square d_x and obtain the total Σd_x^2.

→ Square d_y and obtain the total Σd_y^2.

→ Multiply d_x and d_y and obtain the total $\Sigma d_x d_y$.

→ Substitute the value of $\Sigma d_x d_y$, Σd_x, Σd_y, Σd_x^2 and Σd_y^2 in the formula given above.

Assumption of the pearsonian co-efficient :

Karl pearson's co-efficient of correlation is based on the following assumptions:

→ There is linear relationship between the variables, i.e., when the two variables are plotted on a Scatter diagram a straight line will be formed by the points so plotted.

→ The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.

- → There is a Cause and effect relationship between the forces affecting the distribution of the items in the two Series. If Such a relationship is not formed between the Variables, i.e., If the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not Common.

Merits and limitations of the pearsonian Co-efficient:

Amongst the mathematical methods used for measuring the degree of relationship, Karl pearson's method is most popular. The correlation co-efficient summerizes in one figure not only the degree of correlation but also the direction, i.e., whether correlation is positive (or) Negative.

However, the utility of this co-efficient depends in part on a wide knowledge of the meaning of this 'yardstick' together with its limitations. The chief limitations of the method are:

- → The correlation co-efficient always assumes linear relationship regardless of the fact whether that assumptions is correct (or) not.
- → Great care must be exercised in interpreting the value of this co-efficient as very often the co-efficient is misinterpreted.
- → The value of the co-efficient is unduly affected by the extreme items.
- → As compared with other methods this method takes more time to compute the Value of correlation co-efficient.

Interpreting co-efficient of correlation:

The Co-efficient of Correlation measures the degree of relationship between two sets of figures. As the reliability of estimates depends upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of Co-efficient of Correlation, otherwise fallacious conclusions can be drawn.

Unfortunately, the interpretation of the co-efficient of correlation depends very much on experience. The full significance of r will only be (graped) grasped after working out a number of correlation problems and seeing the kinds of that give rise to various values of r. The investigator must know his data throughly in order to avoid errors of interpretation and emphasis. He must be familiar or become familiar with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general rules are given which would help in interpreting the value of r.

→ when r = +1. it means there is perfect positive relationship between the variables.

→ when r = -1. it means there is perfect negative relationship between the variables.

→ when r = 0. it means there is no relationship between the variables. i.e., the variables are uncorrelated.

→ The closer r is to +1 or -1. the closer the relationship between the variables and the closer r is to 0. the less close the relation.

Rank correlation co-efficient :

The karl pearson's method is based on the assumptions that the population being studied is normally distributed. when it is known that the population is not normal (or) when the shape of the distribution is not known. there is need for a measure of correlation that involves no assumption about the parameter of the population.

It is posible to avoid making any assumptions about the populations being studied by ranking the observations according to size and basing the calculations on the ranks rather than upon the original observations. It does not matter which way the items are ranked, item number one may be the largest (or) it may be the smallest. Using ranks rather than actual observations gives the Co-efficient of rank correlation.

→ For proof, please refer to next chapter on 'Regression Analysis'.

This method of finding out covariability (or) the lack of it between two variables was developed by the British psychologist charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability (or) the judgement of female beauty) Cannot be fixed, but the individual in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. Spearman's rank correlation co-efficient is defined as:

$$R = 1 - \frac{6 \Sigma D^2}{N(N^2-1)} \quad (or) \quad 1 - \frac{6 \Sigma D^2}{N^3 - N}$$

where, R denotes rank co-efficient of Correlation and D refers to the difference of rank between paired items in two series.

consider a set of n individuals, ranked according to two characters x and y.

Individual	$A_1 A_2 \dots A_1 \dots A_n$
Ranking (x)	$x_1, x_2 \dots x_i \dots x_n$
Ranking (y)	$y_1, y_2 \dots y_i \dots y_n$

$$\bar{x} = \frac{1}{N} \Sigma x_1 = \frac{1}{x} [1+2+3 \dots +N] = \frac{N+1}{2}$$

$$\bar{y} = \frac{1}{N} \Sigma y_1 = \frac{1}{y} [1+2+3 \dots +N] = \frac{N+1}{2}$$

$$\bar{x} = \bar{y}$$

$$\sigma_x^2 = \frac{N^2-1}{12} = \sigma_y^2 \text{ or } r = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

Now if di stands for the difference in ranks of ith individual. we have

$$d_i = x_i - y_i = (x_i - x) - (y_i - y)$$

$$= x_i' - y_i'$$

where x_i' and y_i' are deviations of x_i and y_i for the mean $\bar{x}$ & $\bar{y}$

$$\Sigma d_i^2 = \Sigma(x_i' - y_i')^2$$

$$= \Sigma x_i'^2 + \Sigma y_i'^2 - 2\Sigma x_i' y_i'$$

$$\Sigma d_i^2 = \frac{1}{N}\Sigma x_i'^2 + \frac{1}{N}\Sigma y_i'^2 - \frac{2\Sigma x_i' y_i'}{N}$$

$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

$$= 2\sigma_x^2 - 2r\sigma_x^2$$

$$= 2\sigma_x^2(1-r)$$

$$\frac{1}{N} = \Sigma d_i^2 = 2\sigma_x^2(1-r)$$

$$(1-r) = \frac{1}{N}\frac{\Sigma d_i^2}{2\sigma_x^2}$$

$$(1-r) = \frac{1}{N}\frac{\Sigma d_i^2}{N^2\sigma_x^2} = \frac{1}{N}\frac{6\Sigma d_i^2}{(N^2-1)}$$

$$R = 1 - \frac{6\Sigma d^2}{N^3 - N}$$

The value of this co-efficient interpreted in the same way as Karl pearson's correlation co-efficient. Ranges between +1 and -1. when r_2 is +1 there is complete agreement in the order of the ranks and the ranks are in the same direction. when r_2 is -1 there is complete agreement in the order of the ranks and they are in opposite directions. This shall be clear from the following.

R_1	R_2	D (R_1-R_2)	D^2	R_1	R_2	D (R_1-R_2)	D^2
1	1	0	0	1	3	-2	4
2	2	0	0	2	2	0	0
3	3	0	0	3	1	2	4
			$\Sigma D^2=0$				$\Sigma D^2=8$

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

$$=1-\frac{6\times 0}{3^3-2}=1-0=1$$

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

$$=1-\frac{6\times 8}{3^3-3}=1-2=-1$$

Features of Spearman's correlation Co-efficient:

→ The Sum of the differences of ranks between two Variables shall be zero. Symbolically. $\Sigma d=0$

→ Spearman's correlation coefficient is distribution-free (or) non-parametric because no strict assumptions are made about the form of population from which sample observations are drawn.

→ The Spearman's correlation co-efficient is nothing but karl pearsons's correlation co-efficient between the ranks. Hence, it can be interpreted in the same manner as pearsonian correlation coefficient.

In rank correlation we may have two types of problems:

→ where ranks are given.

→ where ranks are not given.

Where ranks are given:

Where actual ranks are given to us the steps required for computing rank correlation are:

<i> Take the difference of the two ranks i.e., $(R_1 - R_2)$ and denote these differences by D.

<ii> Square these differences and obtain the total $\sum D^2$.

<iii> Apply the formula : $R = 1 - \frac{6\sum D^2}{N^3 - N}$

problem:⑧ The ranking of 10 students in two subjects A and B are as follows.

A R_1	B R_2	A R_1	B R_2
6	3	4	6
5	8	9	10
3	4	7	7
10	9	8	5
2	1	1	2

Calculate rank correlation co-efficient.

Sol:- Calculation of rank correlation co-efficient

R_1	R_2	D $(R_1 - R_2)$	D^2
6	3	3	9
5	8	-3	9
3	4	-1	1
10	9	1	1
2	1	1	1
4	6	-2	4
9	10	-1	1
7	7	0	0
8	5	3	9
1	2	-1	1
N=10		$\sum D = 0$	$\sum D^2 = 36$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 36}{10^3 - 10}$$

$$= 1 - \frac{216}{1000 - 10}$$

$$= 1 - \frac{216}{990}$$

$$= 1 - 0.218$$

$$R = 0.782$$

problem : (9) Two ladies were asked to rank 7 different types of lipsticks. The sranks given by them are as follo

Lipsticks	A	B	C	D	E	F	G
Neelu (x)	2	1	4	3	5	7	6
Neena (y)	1	3	2	4	5	6	7

Calculate Spearman's rank correlation co-efficient.

X (R_1)	Y (R_2)	D (R_1-R_2)	D^2
2	1	1	1
1	3	-2	4
4	2	2	4
3	4	-1	1
5	5	0	0
7	6	1	1
6	7	-1	1
N=7		$\Sigma D=0$	$\Sigma D^2=12$

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6\times 12}{7^3-7}$$

$$= 1 - \frac{72}{343-7}$$

$$= 1 - \frac{72}{336}$$

$$= 1 - 0.214$$

$$\therefore R = 0.786$$

problem : (10) Ten competitors in a beauty contest are ranked by 3 judges in the following order.

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation co-efficient to determine which pair of judges has the nearest approach to common testes in beauty.

Sol :- In order to find out which pair of judges has the nearest approach to common testes in beauty. we compare rank correlation between the judgments of.

i) 1st judge and 2nd judge

ii) 2nd judge and 3rd judge

iii 1st judge and 3rd judge

1st judge (R_1)	2nd judge (R_2)	3rd judge. (R_3)	D^2 $(R_1-R_2)^2$	D^2 $(R_2-R_3)^2$	D^2 $(R_1-R_3)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
$N=10$	$N=10$	$N=10$	$\Sigma D^2=200$	$\Sigma D^2=214$	$\Sigma D^2=60$

1st and 2nd judge :

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6 \times 200}{10^3-10}$$

$$= 1 - \frac{1200}{990}$$

$$= 1 - 1.212$$

$$\boxed{R = -0.212}$$

2nd and 3rd judge:

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6 \times 214}{10^3-10}$$

$$= 1 - \frac{1284}{990}$$

$$= 1 - 1.297$$

$$\boxed{R = -0.297}$$

1st and 3rd judge:

$$R = 1 - \frac{6\Sigma D^2}{N^3-N}$$

$$= 1 - \frac{6 \times 60}{10^3-10}$$

$$= 1 - \frac{360}{990}$$

$$= 1 - 0.364$$

$$\boxed{R = +0.636}$$

∴ Since co-efficient of correlation is maximum in the judgments of the first and third judges we conclude that they have the nearest approach to common tastes in beauty.

Where ranks are not given :

When we are given the actual data and not the ranks. It will be necessary to assign the ranks. Ranks can be assigned by taking either highest value as 1 (or) the lowest ~~as~~ value as 1. But whether we start with the lowest value (or) the highest value we must follow the Same method in case of both the variables.

Problem (iii) Calculate Spearman's Co-efficient of correlation between marks assigned to ten students by judges X and Y in a certain competitive test as shown below.

S.No.	1	2	3	4	5	6	7	8	9	10
marks by judge X	52	53	42	60	45	41	37	38	25	27
marks by judge Y	65	68	43	38	77	48	35	30	25	50

Sol :- First assign ranks and then calculate rank correlation Co-efficient.

Calculation of Spearman's Co-efficient of correlation

X	R_1	Y	R_2	D (R_1-R_2)	D^2 $(R_1-R_2)^2$
52	8	65	8	0	0
53	9	68	9	0	0
42	6	43	5	1	1
60	10	38	4	6	36
45	7	77	10	-3	9
41	5	48	6	-1	1
37	3	35	3	0	0
38	4	30	2	2	4
25	1	25	1	0	0
27	2	50	7	-5	25
N=10				$\Sigma D=0$	$\Sigma D^2=76$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 76}{10^3 - 10}$$

$$= 1 - \frac{456}{990}$$

$$= 1 - 0.461$$

$$\boxed{R = 0.539}$$

Problem: (12) Quotations of index no. of Security prices of a certain joint stock company are given below.

year	1	2	3	4	5	6	7
Debenture price(x)	97.8	99.2	98.8	98.3	98.4	96.7	97.1
share price(y)	73.2	85.8	78.9	75.8	77.2	87.2	83.8

Using rank correlation method, determine the relationship between debenture prices and share prices.

Sol:- Calculation of Rank correlation co-efficient

X	R_1	Y	R_2	D^2 $(R_1 - R_2)^2$
97.8	3	73.2	1	4
99.2	7	85.8	6	1
98.8	6	78.9	4	4
98.3	4	75.8	2	4
98.4	5	77.2	3	4
96.7	1	87.2	7	36
97.1	2	83.8	5	9
				$\Sigma D^2 = 62$

$$R_\rho = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 62}{7^3 - 7}$$

$$= 1 - \frac{372}{343 - 7} = 1 - \frac{372}{336} = 1 - 1.107 = -0.107$$

Merits and Limitations of the Rank method :

Merits :

The merits of the Rank method can be discussed here.

→ This method is simpler to understand and easier to apply compared to the Karl pearson's method. The answers obtained by this method and the Karl pearson's method will be the same provided no value is repeated. i.e., all the items are different.

→ where the data are of a qualitative nature like honesty, efficiency, intelligence Etc., This method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying this method.

→ This is the only method that can be used where we are given the ranks and not the actual data.

→ Even where actual data are given, rank method can be applied for ascertaining correlation.

Limitations :

The method is however associated with a few limitations too.

→ This method cannot be used for finding out correlation in a grouped frequency distribution.

→ where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N exceeds 30 unless we are given the ranks and not the actual values of the variables.

Equal ranks :

In some cases it may be found necessary to rant two (or) more individuals (or) entries as equal. In such a case it is customary to give each individuals an average rank. Thus if two individuals are ranked equal at fifth place. They are each given the rank $\frac{5+6}{2}$. that is 5.5 while. if three are ranked equal at fifth place. they are given the rank $\frac{5+6+7}{3}=6$. In other words, where two (or) more items are to be ranked equal. the rank assigned for purposes of calculating co-efficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank co-efficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3-m)$ to the value of ΣD^2. where, m stands for the number of items whose ranks are common. If there are more than one such group of items with common rank. This value is added as many times the number of such groups. The formula can thus be written.

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3-m) + \frac{1}{12}(m^3-m) + \dots\right\}}{N^3-N}$$

problem: (13) Compute Spearman's rank correlation for the following observations:

Candidate	1	2	3	4	5	6	7	8
Judge X	20	22	28	23	30	30	23	24
Judge Y	28	24	24	25	26	27	32	30

Sol:-

Candidate	X	R_1	Y	R_2	D^2 $(R_1-R_2)^2$
1	20	1	28	6	25.00
2	22	2	24	1.5	0.25
3	28	6	24	1.5	20.25
4	23	3.5	25	3	0.25
5	30	7.5	26	4	12.25
6	30	7.5	27	5	6.25
7	23	3.5	32	8	20.25
8	24	5	30	7	4.00
					$\Sigma D^2 = 88.50$

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2 + \ldots\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left\{88.50 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right\}}{8^3-8}$$

$$= 1 - \frac{6\left[88.50 + 0.5 + 0.5 + 0.5\right]}{512-8}$$

$$= 1 - \frac{6 \times 90}{504}$$

$$= 1 - \frac{540}{504}$$

$$= 1 - 1.071 \quad 1.065$$

$$\boxed{R = -0.071} \quad -0.0654$$

When to use Rank correlation Co-efficient :

The Rank method has principal uses.

→ The initial data are in the form of ranks.

→ If N is fairly small (Say, not more than 25 (or) 30) rank method is sometimes applied to interval data as an approximation to the more time-consuming. r. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30. The labour required in ranking the scores becomes greater than is justified by the anticipated Saving of time through the rank formula.

04. Concurrent deviation method:

This method of studying correlation is the simplest of all the methods. The only thing that is required under this method is to find out the direction of change of X Variable and Y Variable. The formula applicable is:

$$r_c = \pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$$

where, r_c stands for co-efficients of correlation by the Concurrent method ; c stands for the number of concurrent deviations (or) the number of positive signs obtains after multiplying D_x with D_y.

n = Number of pairs of observations Compared.

problem: (14) Calculate the co-efficient of concurrent deviation from the following.

X	60	55	50	56	30	70	40	35	80	80	75
Y	65	40	35	75	63	80	35	20	80	60	50

Sol:- Calculation of co-efficient of concurrent deviation method.

	X	Dx	Y	Dy	DxDy
	60		65		
55-60	55	−	40	−	+
50-55	50	−	35	−	+
56-50	56	+	75	+	+
30-56	30	−	63	−	+
70-30	70	+	80	+	+
40-70	40	−	35	−	+
35-40	35	−	20	−	+
80-35	80	+	80	+	+
80-80	80	0	60	−	0
75-80	75	−	50	0 −	0 −
					c = 8

$$\therefore r_c = \pm\sqrt{\pm\frac{2c-n}{n}} = \pm\sqrt{\pm\frac{2\times 8-10}{10}} = \sqrt{\frac{6}{10}} = 0.774$$

problem: (15) Calculate co-efficient of concurrent deviation from the following data.

price	imports	price	imports
368	22	384	26
384	~~24~~21	395	24
385	24	403	29
381	20	400	28
347	22	385	27

Sol:- Calculation of co-efficient of concurrent deviation meth

price x	Direction of change of Valiable x Dx	imports y	Direction of change of Variable y Dy	$D_x D_y$
368		22		
384	+	21	−	−
385	~~1~~+	24	+	+
381	−	20	−	+
347	−	22	+	−
384	+	26	+	+
395	+	29	−	−
403	+	29	+	+
400	−	28	−	+
385	−	27	−	+
				c = 6

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

$$= \pm\sqrt{\pm\frac{2\times 6-9}{9}} = \pm\sqrt{\pm\frac{12-9}{9}} = \sqrt{\frac{3}{9}} = \sqrt{0.333}$$

$$\boxed{r_c = 0.577}$$

problem : (16) Calculate the co-efficient of correlation using the method of concurrent deviation between Supply and Demand of an item for a period of 10 years as given below:

year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Supply	125	160	164	174	155	170	165	162	172	175
Demand	115	125	192	190	165	174	124	127	152	169

Sol :- Calculation of correlation by concurrent Deviation method.

year	X (Supply)	Dx	y (Demand)	Dy	Dx Dy
2002	125		115		
2003	160	+	125	+	+
2004	164	+	192	+	+
2005	174	+	190	~~+~~ −	−
2006	155	−	165	−	+
2007	170	+	174	+	+
2008	165	−	124	−	+
2009	162	−	127	+	−
2010	172	+	152	+	+
2011	175	+	169	+	+
					C = 7

$$r_c = \pm \sqrt{\pm \frac{2c - n}{n}}$$

$$= \pm \sqrt{\pm \frac{2 \times 7 - 9}{9}}$$

$$= \pm \sqrt{\pm \frac{14 - 9}{9}}$$

$$= \pm \sqrt{\pm \frac{5}{9}}$$

$$= \pm \sqrt{\pm 0.555}$$

$$\boxed{r_c = 0.745}$$

Merits and Limitations of concurrent Deviation Method:

Merits:

The following are the basic advantages of this method.

→ It is simplest of all the methods

→ when the number of items is very large, this method may be used to form a quick idea about the degree of relationship before making use of more complicated methods.

Limitations:

This method is associated with the following limitations.

→ This method does not differentiate between small and big changes. For example, if x increases from 100 to 101 the sign will be plus and if y increases from 60 to 160. The sign will be plus. Thus, both get equal weight when they vary in the same direction.

→ The results obtained by this method are only a rough indicator of the presence (or) absense of correlation.

Regression Analysis :

Regression is the meacule of the average relationship between two (or) more Variables in terms of original units of the data.

It is mainly used to predict (estimate) the unknown Value of one Variable from the known Values of other Variable.

Types of regression :

1> Simple regression

2> multiple regression.

1> Simple regression :

If there are only two Variables under consideration, then the regression is called simple regression.

2> multiple regression :

If there are more than two Variables under consideration, then the regression is called multiple regression.

Lines of Regression (or) Linear Regression equations :

The lines which expresses the average relationship between two Variables are known as regression lines.

If x and y are two variables, there exists two equations for regreesion lines.

put all small letters x and y

Method of Least Squares :-

→ Regression equation of y on x :- The regression equation of y on x is given by $\boxed{y = a + bx}$

To determine the values of a and b, the normal equations to be solved are.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

→ Regression equation of x on y: The regression equation of x on y is given by $\boxed{x = a + by}$

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

$$\sum xy = a\sum y + b\sum y^2$$

Method of Regression Co-efficient:

The regression line of y on x is given by

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

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

where. b_{yx} is the regression co-efficient of y on x.

and $b_{yx} = r\,\dfrac{\sigma_y}{\sigma_x} = \dfrac{\sum xy}{\sum x^2}$

where, r = Correlation Coefficient

σ_y = S.D of y Series

σ_x = S.D of x Series

$x = x_i - \bar{x}$

$y = y_i - \bar{y}$

$\bar{x}$ = mean of x_i

$\bar{y}$ = mean of y_i

The regression line of x on y is given by

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

(or)

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

where, b_{xy} is the regression co-efficient of x on y

and $b_{xy} = r\,\dfrac{\sigma_x}{\sigma_y} = \dfrac{\sum xy}{\sum y^2}$

problem :- ① From the following data obtain the regression equations by the method of least Squares.

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

Sol :→ Regression equation of y on x :

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma x=30$	$\Sigma y=40$	$\Sigma xy=214$	$\Sigma x^2=220$	$\Sigma y^2=340$

$n=5$

Regression equation of y on x is $y = a+bx$

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

$40 = 5a + 30b$ —— (1)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$214 = 30a + 220b$ —— (2)

equation-(1) multiply with 6

$(1) \times 6 \Rightarrow 240 = 30\cancel{a} + 180b$

$(2) \Rightarrow \underset{(-)}{214} = \underset{(-)}{\cancel{30a}} + \underset{(-)}{220b}$

$26 = -40b$

$40b = -26$

$b = \frac{-26}{40}$

$\therefore \boxed{b = -0.65}$

'b' Value Substitute in equation-(1) $b = -0.65$

$$40 = 5a + 30b$$

$$40 = 5a + 30(-0.65)$$

$$40 = 5a - 19.5$$

$$40 + 19.5 = 5a$$

$$59.5 = 5a$$

$$\frac{59.5}{5} = a$$

$$\therefore \boxed{a = 11.9}$$

$\therefore$ Regression equation y on x = $\boxed{y = 11.9 - 0.65x}$

→ Regression equation x on y is $x = a + by$

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

$$30 = 5a + 40b \longrightarrow (1)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$214 = 40a + 340b \longrightarrow (2)$$

equation-(1) multiply with 8

$$(1) \times 8 \Rightarrow 240 = \cancel{40a} + 320b$$

$$(2) \Rightarrow \underset{(-)}{214} = \underset{(-)}{\cancel{40a}} + \underset{(-)}{340b}$$

$$26 = -20b$$

$$20b = -26$$

$$b = \frac{-26}{20}$$

$$\boxed{b = -1.3}$$

Substitute b Value in equation-1 :

$$30 = 5a + 40b$$

$$30 = 5a + 40(-1.3)$$

$$30 = 5a - 52$$

$$30 + 52 = 5a$$

$$82 = 5a$$

$$a = \frac{82}{5}$$

$$\boxed{a = 16.4}$$

Regression equation x on y = $\boxed{x = 16.4 - 1.3y}$

The regression Equations are

$$\boxed{\begin{array}{ll} y = a + bx & \text{is} \quad y = 11.9 - 0.65x \\ x = a + by & \text{is} \quad x = 16.4 - 1.3y \end{array}}$$

Problem: ② Find equation of regression line y on x for the following data.

x	65	63	67	64	68	62	70	66	68	67
y	68	66	68	65	69	66	68	65	71	67

Also estimate y if $x = 75$.

Sol:-

x	y	xy	x^2	y^2
65	68	4420	4225	4624
63	66	4158	3969	4356
67	68	4556	4489	4624
64	65	4160	4096	4225
68	69	4692	4624	4761
62	66	4092	3844	4356
70	68	4760	4900	4624
66	65	~~4828~~ 4290	4356	4225
68	71	4828	4624	5041
67	67	4489	4489	4489
$\Sigma x = 660$	$\Sigma y = 673$	$\Sigma xy = 44,445$	$\Sigma x^2 = 43616$	$\Sigma y^2 = 45325$

$n = 10$

Regression equation of y on x :

$$Y = a + bX$$

The normal equations are $\Sigma Y = na + b\Sigma X$

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

$$673 = 10a + 660b \quad \text{—— (1)}$$

$$44445 = 660a + 43616b \quad \text{—— (2)}$$

Equation-(1) multiply with 66

$$(1) \text{—} \times 66 \Rightarrow \quad 44418 = 660\cancel{a} + 43560b$$

$$(2) \qquad \Rightarrow \quad \underset{(-)}{44445} = \underset{(-)}{\cancel{660a}} + \underset{(-)}{43616b}$$

$$-27 = -56b$$

$$56b = 27$$

$$b = \frac{27}{56}$$

$$\boxed{b = 0.48}$$

Substitute b value in equation-(1):

$$673 = 10a + 660b$$

$$673 = 10a + 660(0.48)$$

$$673 = 10a + 316.8$$

$$673 - 316.8 = 10a$$

$$356.2 = 10a$$

$$a = \frac{356.2}{10}$$

$$\boxed{a = 35.62}$$

Regression equation of y on x is $Y = a + bX$

$$\boxed{Y = 35.62 + 0.48X}$$

if $x = 75$,,

$$y = 35.62 + 0.48(75)$$

$$= 35.62 + 36$$

$$\boxed{\therefore\ y = 71.62}$$

Regression equation of x on y :-

$$x = a + by$$

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

$$660 = 10a + 673b \quad \text{—— (1)}$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$44445 = 673a + 45325b \quad \text{—— (2)}$$

equation -(1) multiply with 67.3,,

$$(1) \times 67.3 \Rightarrow \quad 44418 = \cancel{673a} + 45292.9b$$

$$(2) \qquad \Rightarrow \quad 44445 = \cancel{673a} + 45325b$$

$$(-) \qquad (-) \qquad (-)$$

$$-27 = -32.1b$$

$$32.1b = 27$$

$$b = \frac{27}{32.1}$$

$$\boxed{b = 0.84}$$

Substitute b value in equation -(1):

$$660 = 10a + 673b$$

$$660 = 10a + 673(0.84)$$

$$660 = 10a + 565.32$$

$$660 - 565.32 = 10a$$

$$10a = 94.68$$

$$a = \frac{94.68}{10}$$

$$\boxed{a = 9.47}$$

Regression equation x on y is $x = a + by$

$$\boxed{x = 9.47 + 0.84y}$$

Another way of Calculating Regression equation:

problem: ① Calculate the Regression deviation of items from the mean of x and y Series.

X	6	2	10	4	8
Y	9	11	5	8	7

Sol:-

X	x $(X-\bar{X})$	x^2	Y	y $(Y-\bar{Y})$	y^2	xy
6	0	0	9	1	1	0
2	-4	16	11	3	9	-12
10	+4	16	5	-3	9	-12
4	-2	4	8	0	0	0
8	+2	4	7	-1	1	-2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$ $\bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$

Regression equation of x on y:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y} (Y - \bar{Y})$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

Hence , $x-6 = -1.3\ (y-8)$

$$x-6 = -1.3y + 10.4$$

$$x = -1.3y + 10.4 + 6$$

$$x = -1.3y + 16.4$$

$$\boxed{x = 16.4 - 1.3y}$$

Regression equation y on x :

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

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0.65$$

$$y-8 = -0.65\ (x-6)$$

$$y-8 = -0.65x + 3.9$$

$$y = -0.65x + 3.9 + 8$$

$$y = -0.65x + 11.9$$

$$\boxed{y = 11.9 - 0.65x}$$

~~problem~~:

Nill

Problem: ② Using the following ~~for~~ data obtain two Regression equations.

X	14	19	24	21	26	22	15	20	19
Y	31	36	48	37	50	45	33	41	39

Sol:-

X	x $(x-\bar{x})$	x^2	y	y $(y-\bar{y})$	y^2	xy
14	-6	36	31	-9	81	54
19	-1	1	36	-4	16	4
24	4	16	48	8	64	32
21	1	1	37	-3	9	-3
26	6	36	50	10	100	60
22	+2	4	45	5	25	10
15	-5	25	33	-7	49	35
20	0	0	41	1	1	0
19	-1	1	39	-1	1	1
$\Sigma x=180$	$\Sigma x=0$	$\Sigma x^2=120$	$\Sigma y=360$	$\Sigma y=0$	$\Sigma y^2=346$	$\Sigma xy=193$

$\bar{x} = \frac{\Sigma x}{N} = \frac{180}{9} = 20$, $\bar{y} = \frac{360}{9} = 40$

Regression equation x on y :

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

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{193}{346} = 0.557$$

$$X-20 = 0.557(Y-40)$$

$$X-20 = 0.557Y - 22.28$$

$$X = 0.557Y - 22.28 + 20$$

$$X = 0.557Y - 2.28$$

$$\therefore \boxed{X = -2.28 + 0.557Y}$$

Regression equation y on x:

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

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{193}{120} = 1.608$$

$$y-40 = 1.608(x-20)$$

$$y-40 = 1.608x - 32.16$$

$$y = 1.608x - 32.16+40$$

$$y = 1.608x + 7.84$$

$$\therefore \boxed{y = 7.84 + 1.608x}$$

problem: (3) The following data relate to the Scores obtained by a Sales man of a company in an intelligence test and their weekly Sales in Hundred rupees

Sales man Intelligence	A	B	C	D	E	F	G	H	I
Test Score(X)	50	60	50	60	80	50	80	40	70
weekly Sales(Y)	30	60	40	50	60	30	70	50	60

a) obtain the regression equation of Sales on intelligence test Scores of the Sales man.

b) If the intelligence test Score of a Sales man is 65. what would be his expected weekly Sales.

Sol :-

X	x $(x-\bar{x})$	x^2	Y	y $(y-\bar{y})$	y^2	xy
50	-10	100	30	-20	400	200
60	0	0	60	10	100	0
50	-10	100	40	-10	100	100
60	0	0	50	0	0	0
80	20	200	60	10	100	200
50	-10	100	30	-20	400	200
80	20	200	70	20	400	400
40	-20	200	50	0	0	0
70	10	100	60	10	100	100
$\Sigma X = 540$	$\Sigma x = 0$	$\Sigma x^2 = 1600$	$\Sigma Y = 450$	$\Sigma y = 0$	$\Sigma y^2 = 1600$	$\Sigma xy = 1200$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{540}{9} = 60, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{450}{9} = 50$$

Regression equation x on y :

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{1200}{1600} = \cancel{1.2}\ 0.75$$

$$X - 60 = 0.75(Y - 50)$$

$$X - 60 = 1.24 - 60$$

$$X = 1.24 - \cancel{60} + \cancel{60}$$

$$X = 1.24$$

$$\therefore \boxed{X = 1.24}$$

Regression equation y on x :

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1200}{1600} = 0.75$$

$$Y - 50 = 0.75(X - 60)$$

$$Y - 50 = 0.75X - 45$$

$y = 0.75x - 45 + 50$

$y = 0.75x + 5$

$\boxed{\cancel{y = -22 + 1.2x}}$

$\boxed{y = 0.75x + 5}$

if $x = 65$, $\cancel{y = -22 + 1.2(65)}$

$\cancel{= -22 + 78}$

$\boxed{\cancel{y = 56}}$

$y = 0.75x + 5$

$= 0.75(65) + 5$

$= 48.75 + 5$

$y = 53.75;$

problem: ④ Compute the two Regression co-efficient of equations on the basis of the following information.

	x	y
mean	40	45
standard deviation	10	9

The Karl pearson correlation co-efficient between x and y is 0.50. Also Estimate value of y for x = 48.

Sol :- Given values, $\bar{x} = 40$ $\bar{y} = 45$

$\sigma_x = 10$ $\sigma_y = 9$

$r = 0.50$

Regression equation of y on x :

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

$$y - 45 = 0.50 \times \frac{9}{10}(x - 40)$$

$$y - 45 = 0.45(x - 40)$$

$$y - 45 = 0.45x - 18$$

$$y = 0.45x - 18 + 45$$

$$y = 0.45x + 27$$

$$\therefore \boxed{y = 27 + 0.45x}$$

If $x = 48$, $y = 27 + 0.45x$

$$= 27 + 0.45 \times 48$$

$$= 27 + 21.6$$

$$\therefore \boxed{y = 48.6}$$

when $x = 48$, $y = 48.6$,

Regression equation of x on y:

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

$$x - 40 = 0.50 \times \frac{10}{9} (y - 45)$$

$$x - 40 = 0.556 (y - 45)$$

$$x - 40 = 0.556y - 25.02$$

$$x = 0.556y - 25.02 + 40$$

$$x = 0.556y + 15$$

$$\boxed{x = 15 + 0.556y}$$

Distinction between Correlation and Regression :

Correlation differs from Regression in the following respects:

Basis of Distinction	Correlation	Regression
1. What measures?	Correlation measures degree and direction of relationship between the variables.	Regression measures the nature and extent of average relationship between two (or) more variables in terms of the original units of the data.
2. whether relative (or) absolute measure.	It is a relative measure showing association between variables.	It is an absolute measure of relationship.
3. whether - independent of choice of both origin and Scale.	Correlation Co-efficient is independent of change of both origin and Scale.	Regression Co-efficient is independent of change of origin and not Scale.
4. whether independent of units of measurement.	correlation co-efficient is independent of units of measurement.	Regression Co-efficient is not independent of units of measurement.
5. Expression of relationship.	Expression of the relationship between the variables range from -1 to +1.	Expression of the relationship between the variables may be in any of the forms like– $Y = a+bx$ $Y = a+bx+cx^2$
6. whether a forecasting device?	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable.
7. Non-Sense	There may be non-sense correlation such as weight of girls and income of boys.	There is nothing like non-sense regression.