Correlation

Introduction :-

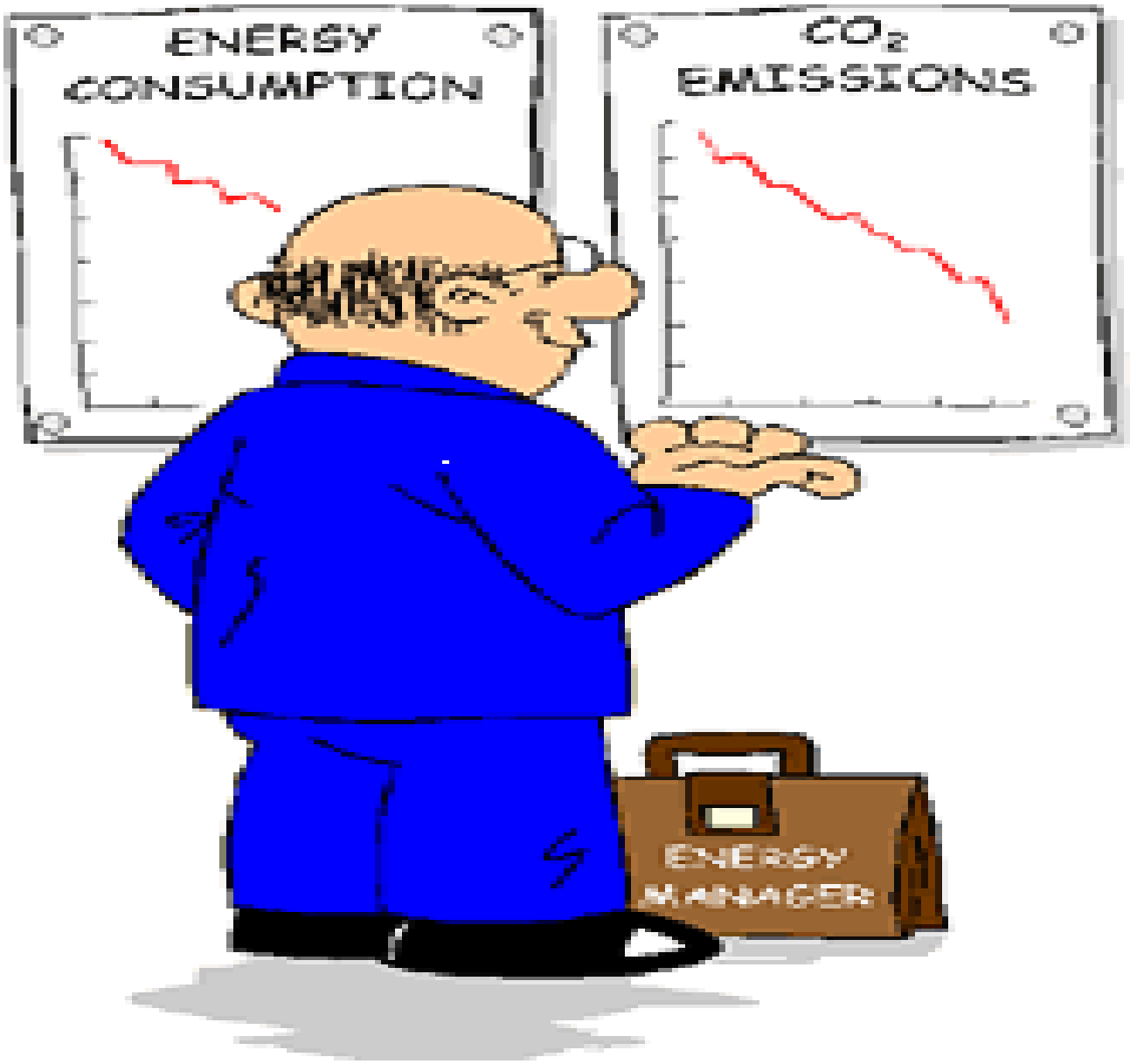
Univariate distribution are the distributions where unit can take only one variable value. In Bivariate distribution units can take two variable values and the distribution where units can take more than two variable values are known as Multivariate distributions.

In bivariate distributions we may be interested to know:

1. Any relationship between the variables under study.
2. The effect of one variable on other.
3. Their moment togetherness.

Correlation is a statistical tool which studies the relationship between the variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between two variables.

The variables are said to be correlated if the change in one variable results in a corresponding change in other variable.


ENERGY CONSUMPTION
CO_2 EMISSIONS
ENERGY MANAGER

Study of correlation deals with the degree (strength) of mutual statistical relationship between two or more variables. i.e., correlation studies the correspondence of movement (going togetherness) between two variables or series of paired items.
For example :

1. If the price increases the demand decreases.
2. Cases of lung cancer may increase if the smoking habit increases.
3. Sale of woolen garments increases as the temperature decreases.

In the above said examples the two variables move together in same direction or in opposite direction. But there are cases when two variables move independently and there is no tendency of 'going togetherness' between them.

In correlation we do not deal with one series but rather with the association or relationship between two series,

and we do not measure variation with one series but rather compare variation in two or more series.

The two series may vary together in the same direction; or
They may vary together in opposite directions; or
They do not vary together at all

To measure of association of series through correlation we must have sufficient number of items in the series.

If we have only two or three pair of values, we can not generalize concerning the way in which they vary together.

Moreover there must not be a blank in one series where there is a value in the other series; there must be a pairing throughout.

Definitions :

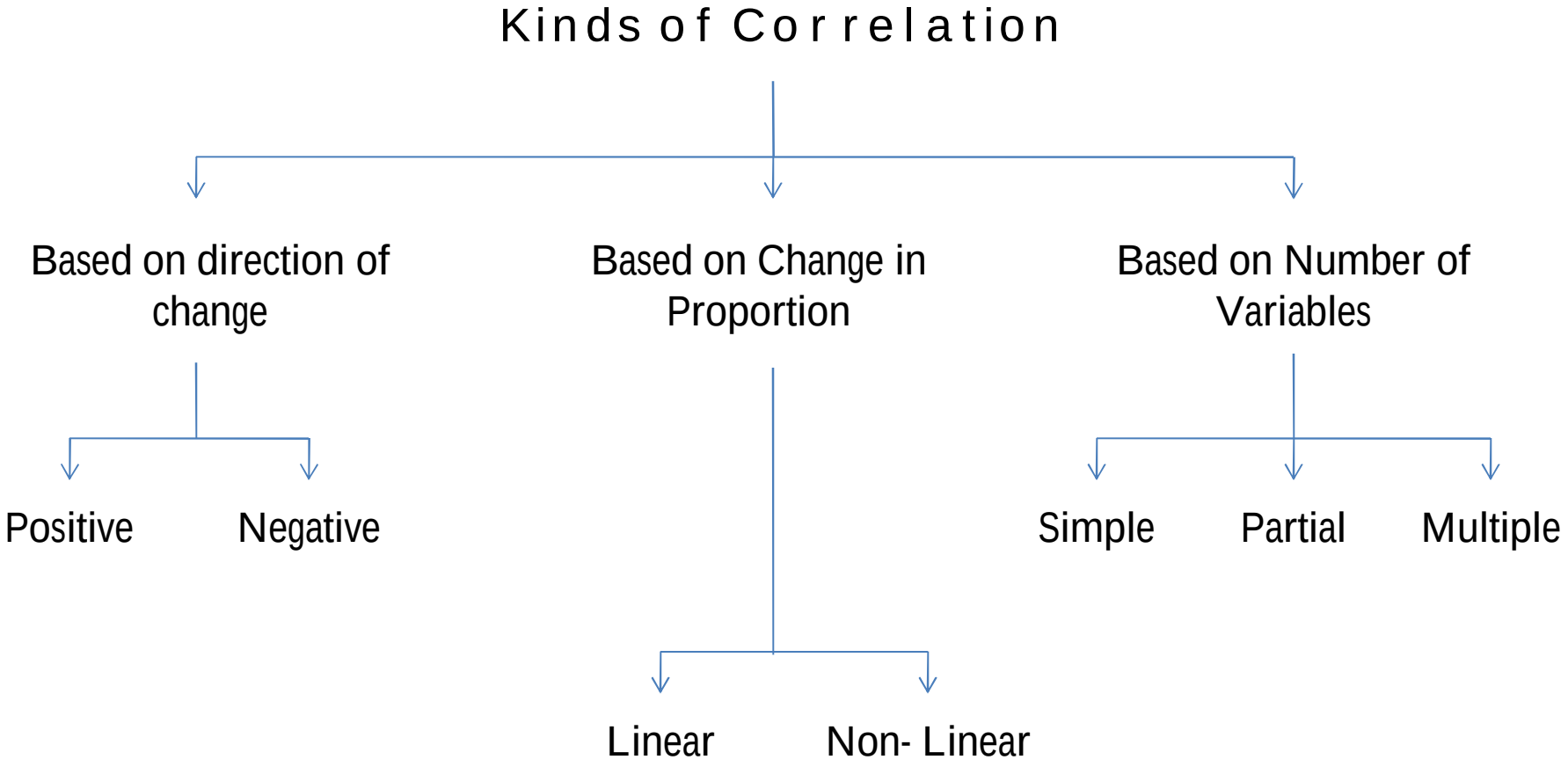
Correlation has been defined in different ways. Some of main definitions are as under.

1. Correlation measures the closeness of relationship between two variables, more exactly of the closeness of the linear relationship.
2. According to the words of Bodington ; “Whenever some definite connection exists between the two or more groups, classes or series or data there is said to be a correlation”.

Importance and Utility of Correlation :

Correlation has been defined in different ways. Some of main definitions are as under.

1. The coefficient of correlation helps in measuring the extent of relationship between two variables in one figure only.
2. Existence of relationship between two or more variables enables us to predict what will happen in the future, e.g., if the production of wheat increases and all the other factors are constant there may be a downfall in the price of wheat.
3. If the two variables are closely related we can estimate the value of one variable given the value of other variable.
4. Correlation facilitates decision-making in business organizations. Expectations about the behavior of certain variables are also on correlation analysis.


Kinds of Correlation
Based on direction of change
Positive
Negative
Based on Change in Proportion
Linear
Non- Linear
Based on Number of Variables
Simple
Partial
Multiple

Positive and Negative Correlation :

If two variables move together in same direction, the correlation between them is said to be *Positive*. If two variables move in opposite directions, the correlation between them is said to be *Negative*. If they do not move together at all there is *No Correlation* between them.

Example :

1. Since the price and demand move in opposite direction, the correlation between them is negative.
2. Smoking habit and cases of lung cancer move in the same direction, correlation between them is positive.

Liner and Non - linear Correlation :

If there is a proportionate change in the value of two variables the correlation is known as *Linear* . If the change in the value of two variables is not proportionate the correlation is known as *Non – Linear.*

Example :

1. The law of demand says other factor remaining constant, increase in price of commodity is followed by a decrease in its demand, but we can not find any proportionality relationship between them.
2. A proportionate change can be observed between consumption of coffee and number of employees.

Example :

1. | x | 1 | 2 | 3 | 4 | 5 |
   |---|---|---|---|---|---|
   | y | 2 | 4 | 6 | 8 | 10 |

   Linear Correlation

2. | x | 1 | 2 | 3 | 4 |
   |---|---|---|---|---|
   | y | 3 | 5 | 8 | 15 |

   Non - Linear Correlation

Correlation Based on Number of Variables :

When only two variables are involved and the relationship is studied between those two variables the correlation is known as *Simple Correlation*. When more than two variables are involved but the relationship is studied between two variables only, keeping other variables as constant then the correlation is known as *Partial Correlation*. But if more than two variables are involved and the relationship is studied between all of them. then the correlation is known as *Multiple Correlation.*

Some Important Points :

1. There should be sufficient number of items in the series.
2. In correlation analysis we do not deal with one series only but the association or relationship between two or more series.
3. We do not measure the variation in one series only rather we compare variation in two or more series.
4. We study only Linear Correlation.
5. Correlation does not necessarily mean cause and effect relationship.
6. The sign of ‘r’ indicates the type of linear relationship whether positive or negative.

Measure of Correlation :

1. Scatter Diagrams.
2. Karl Pearson's coefficient for measuring linear correlation.
3. Method of Rank Differences (Spearman's Rank Correlation Coefficient).

Scatter Diagram :

Scatter diagram or dot diagram is a graphical representation of pair of numerical values of the two variables. Each pair of values is represented by a dot on the graph. The scatter of points and the direction of the scatter diagram revels the nature and degree of correlation between two variables.

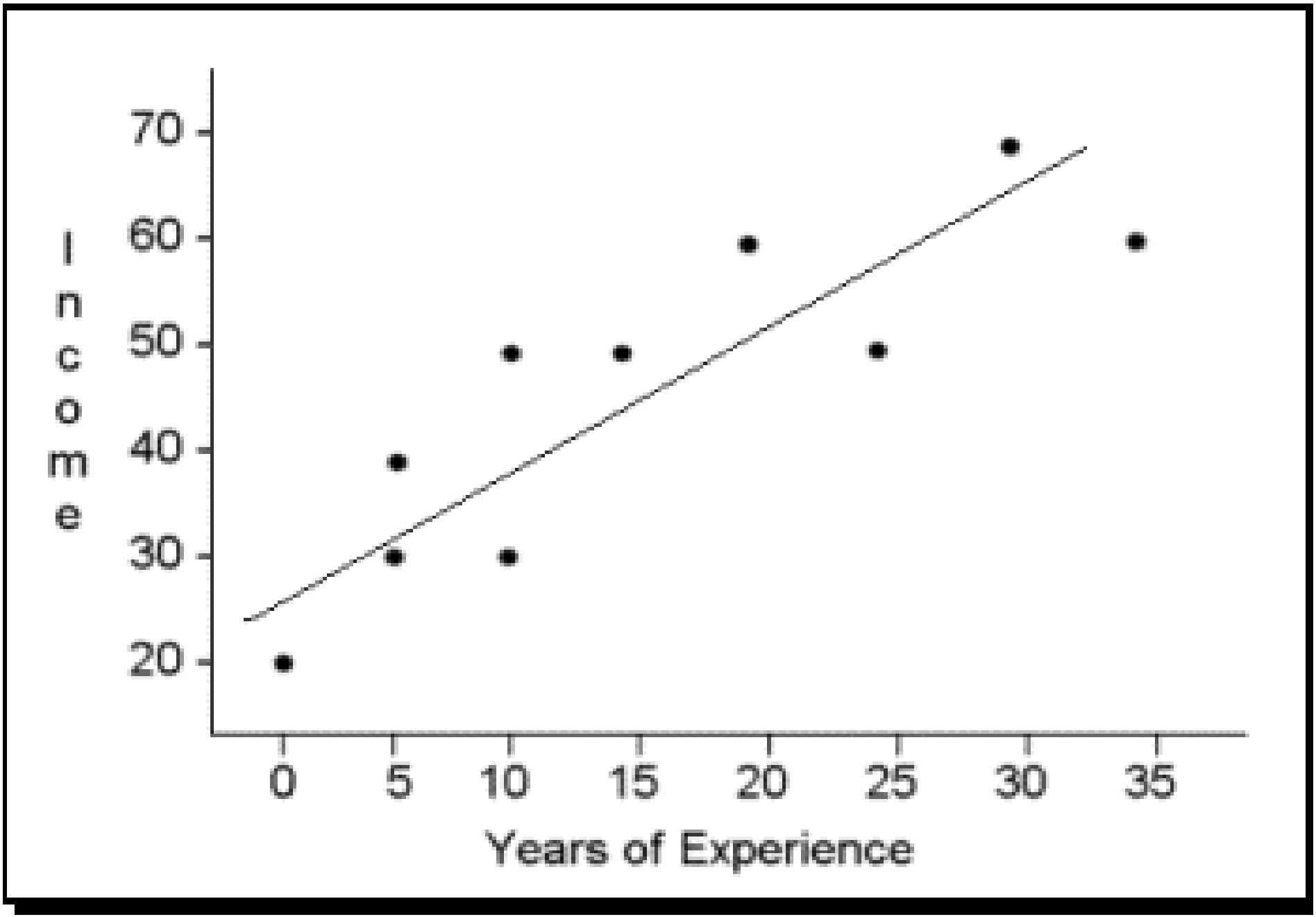
If all the points lie on a straight line having positive slope (i.e. rising line) the correlation is said to be perfect positive. In this case coefficient of correlation 'r = + 1'.

If all the points lie on the line having negative slope the correlation is known as perfect negative. In this case coefficient of correlation 'r = - 1.

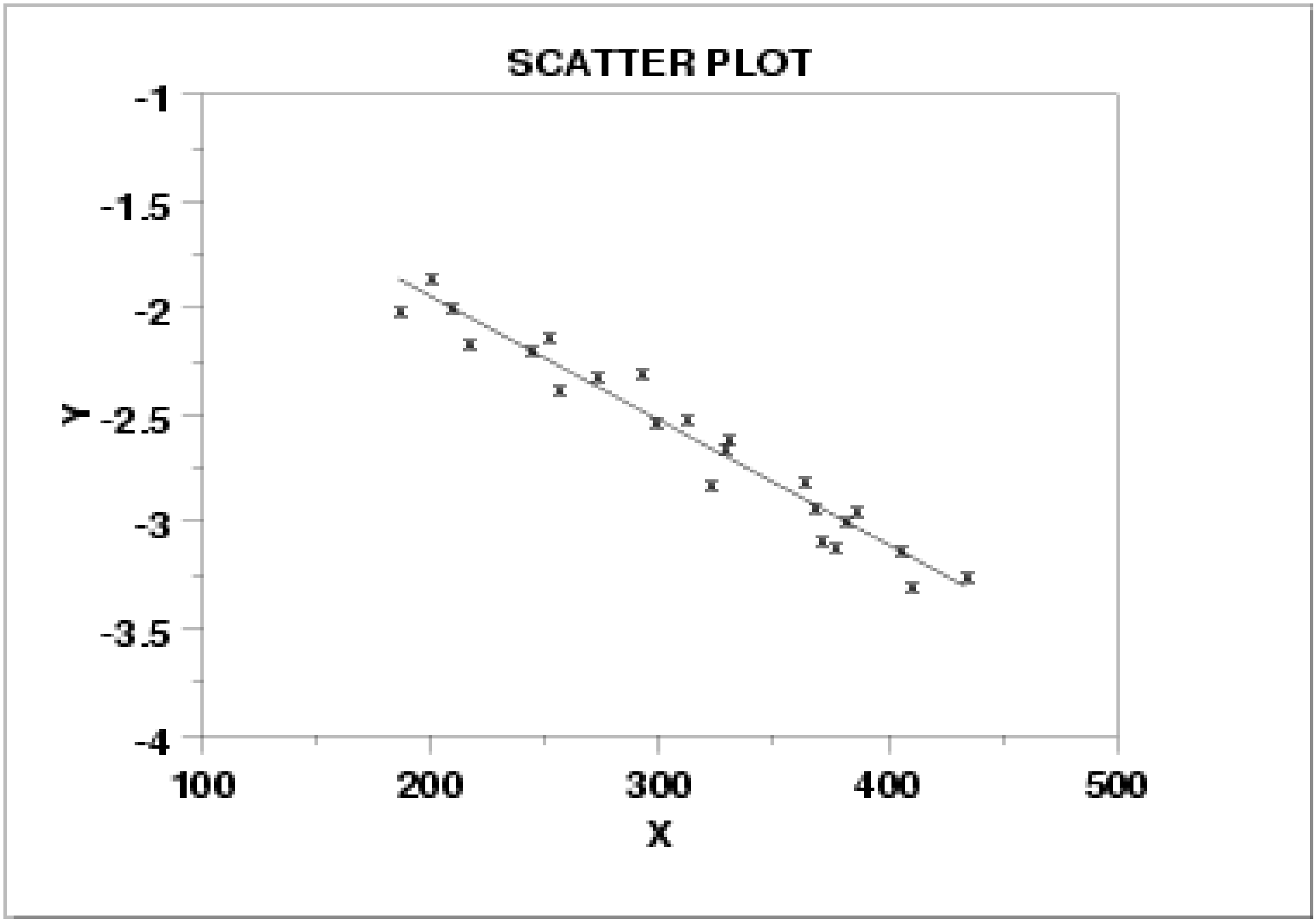
In general if low values of one variables go with the low values of other variable and high value of one variable goes with the high value of other variable, the path traced by these points runs roughly from lower left to upper right corner, the relationship is Direct and Positive.

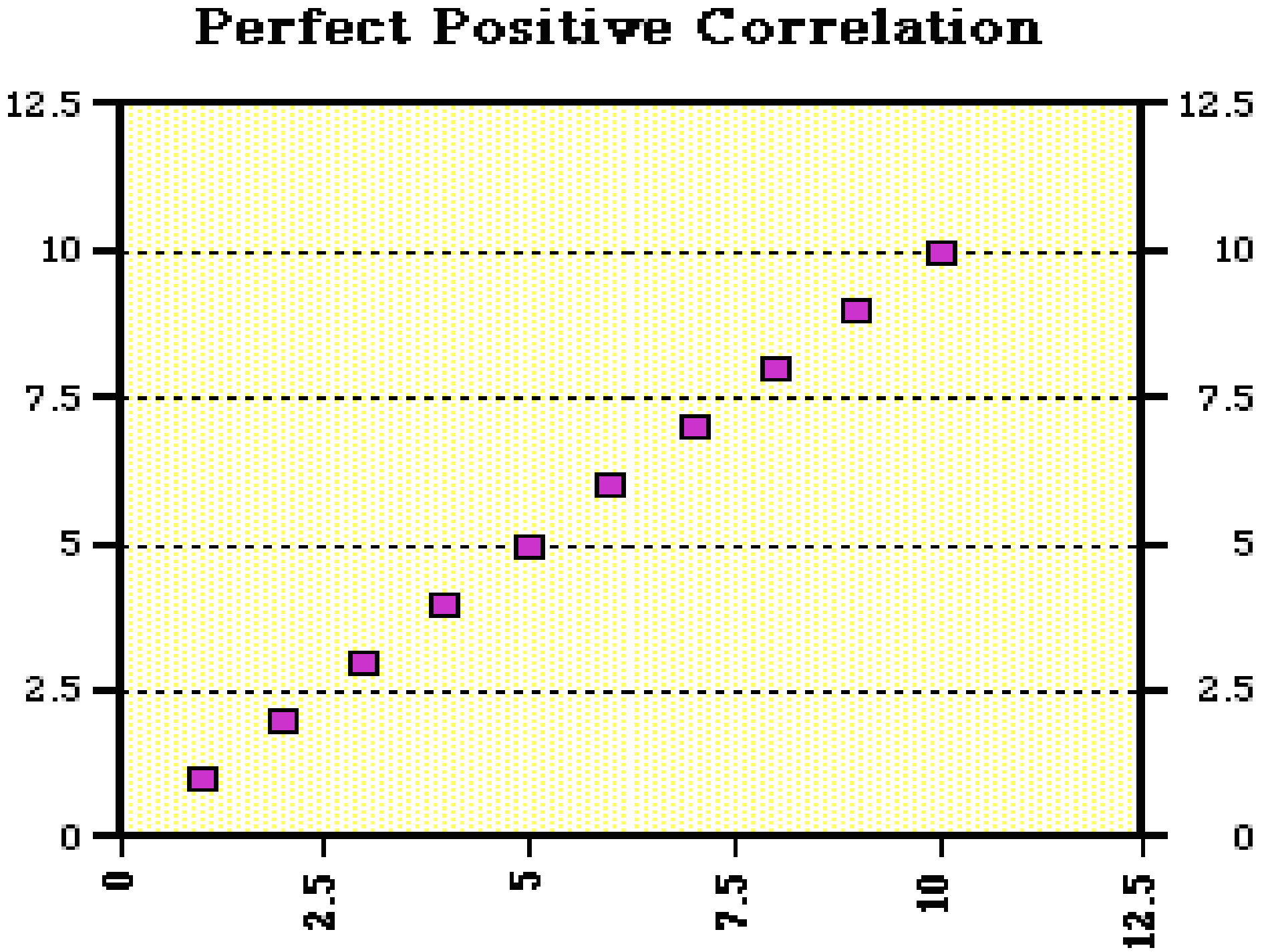
And low values of one variables go with the high values of other variable, while high value of one variable goes with the low values of other variable, the path traced by these points roughly from upper corner to the lower right corner, relationship is inverse and called negative.

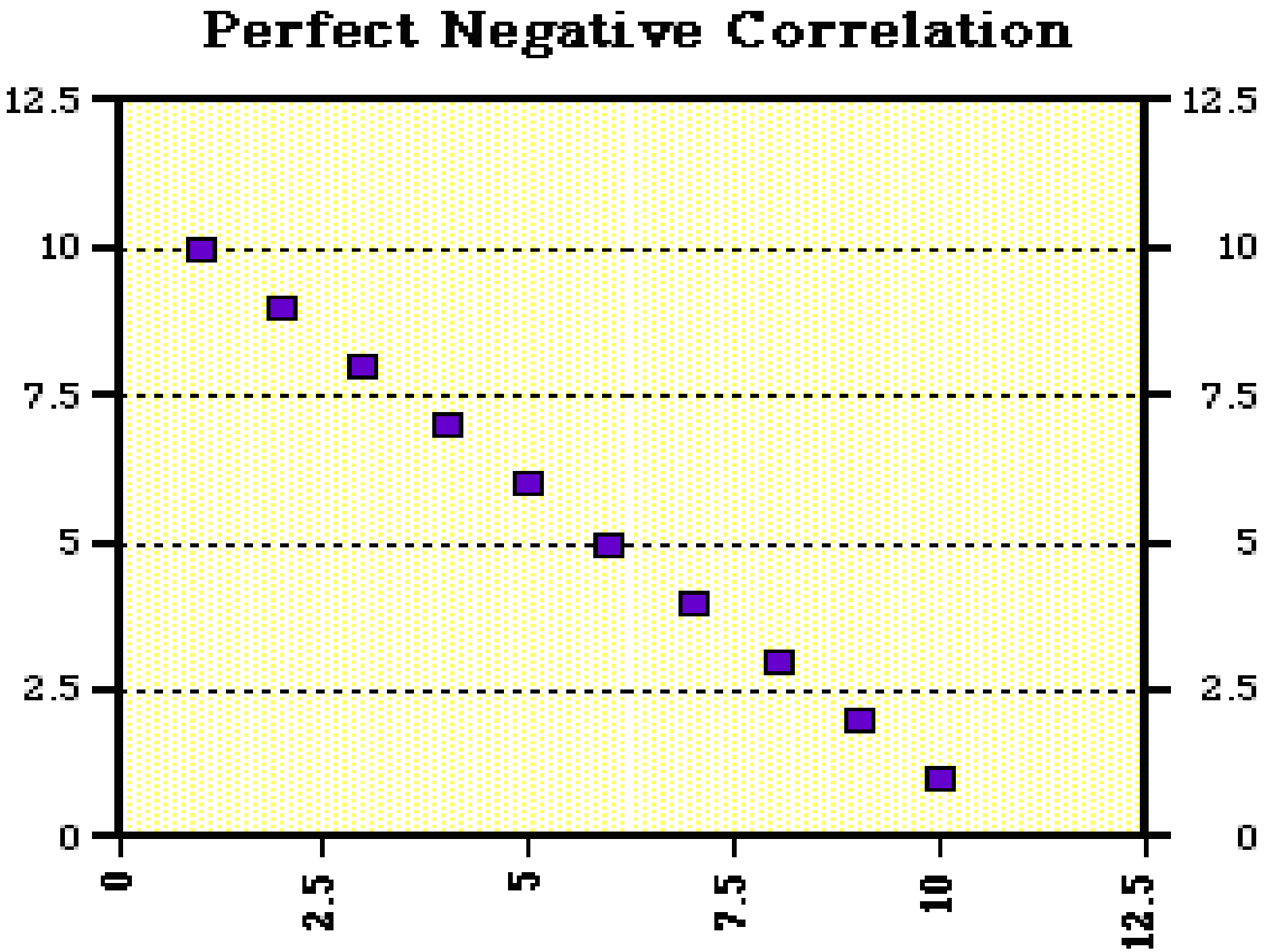
Positive Correlation

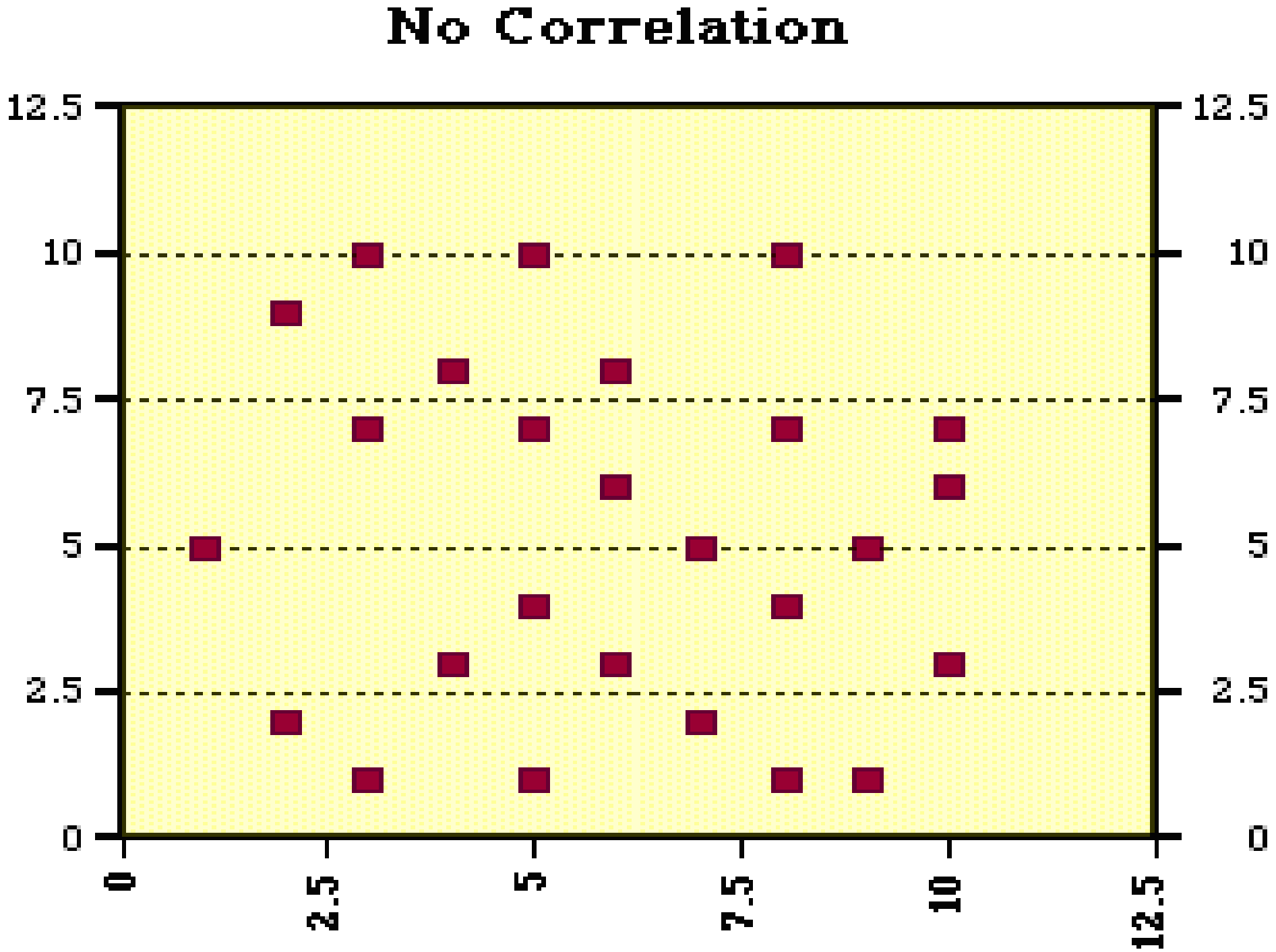


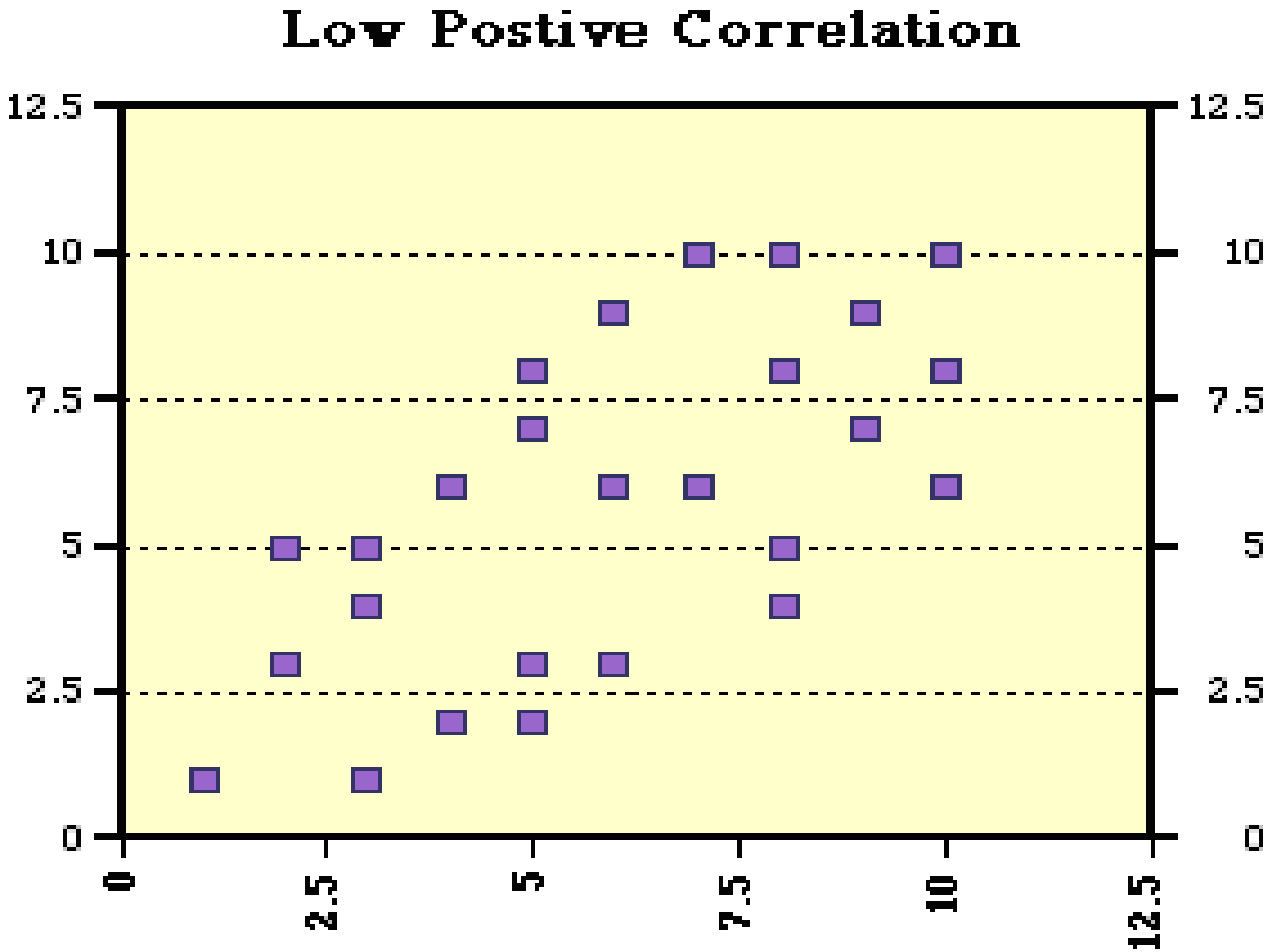
Negative Correlation

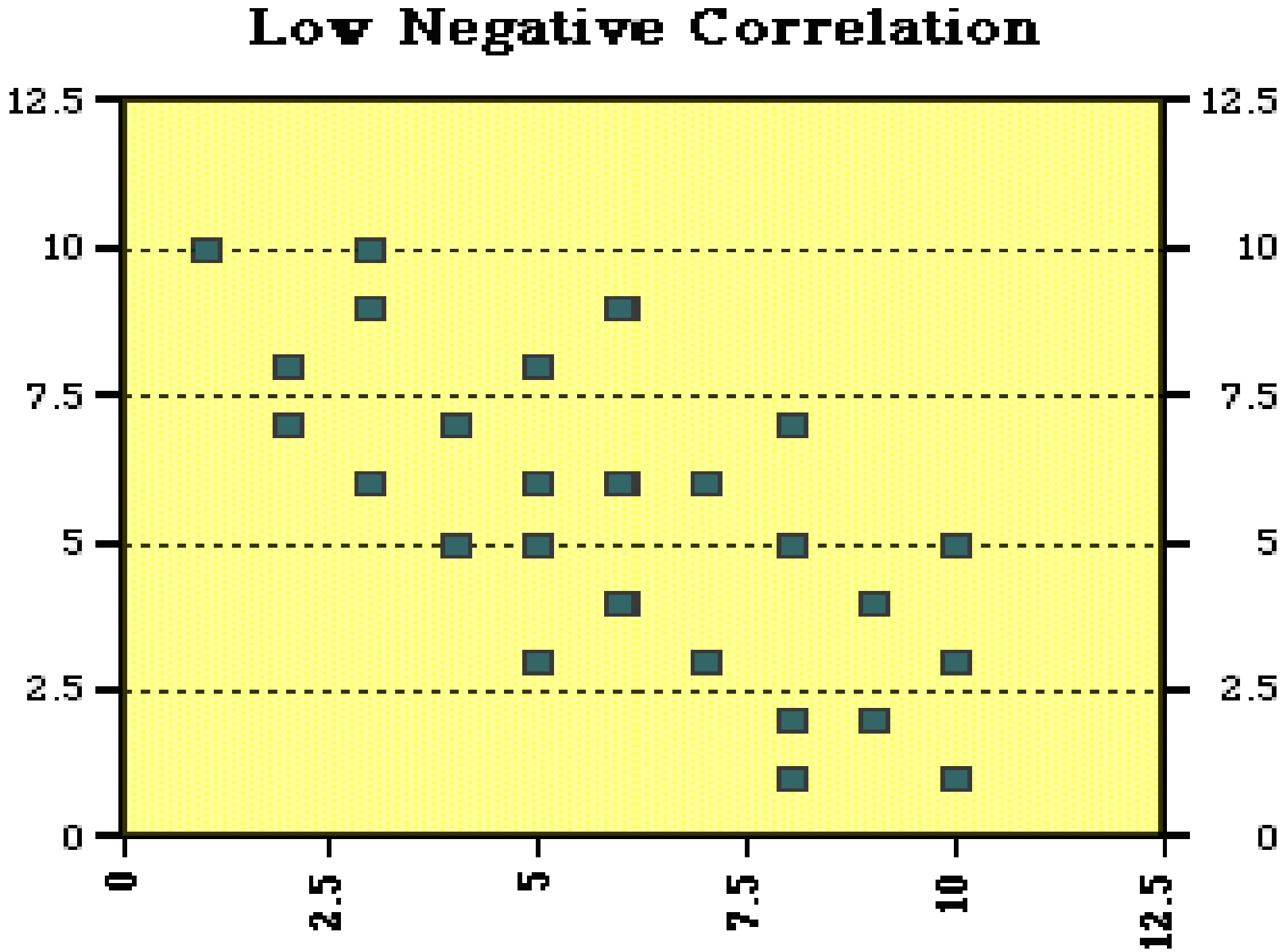



Perfect Positive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Perfect Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


No Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Postive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

Merits and Limitations of the Scatter – Diagram Method :

1. It is a non – mathematical and easy way to find the correlation between two variables.
2. By drawing a line of best fit by free hand method through the plotted dots, the method can be used for estimating the missing value of the dependent variable for a given value of independent variable.
3. The shape of scatter – diagram reveals whether the correlation is Linear or Non – linear which enables us to know the pattern of relationship existing between two variables. Scatter diagrams gives us an idea whether correlation is positive or negative.
4. The values of extreme observations do not affect the method.

Demerits :

It gives only rough idea how the two variables are related. The methods gives an idea about the direction of correlation and also whether it is how or low. But this method does not give any quantitative measure of the degree or the extend of correlation.

Interpretation of correlation coefficient

Karl Pearson Coefficient of Correlation:

A mathematical method of measuring the intensity or the magnitude of linear relationship between two variable series was suggested by Karl Pearson (1867 – 1936), a great British Bio – metrician and Statistician and by far the most widely used method in practice.

Karl Pearson's measure is known as Pearson's correlation coefficient between two variables (series) X and Y, usually denoted by 'r (X, Y)' or 'r_{xy}' or simply 'r', is a numerical measure of linear relationship between them.

Calculation of Correlation Coefficient :

For ungrouped data. Karl Pearson's coefficient of correlation can be obtained by using any of the following three methods :

(i) Actual Mean Method
(ii) Direct Method
(iii) Short – Cut Method

Actual Mean Method :

$$r = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{n\left(\sigma_x \sigma_y\right)}$$

$$= \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\left[\sum \left(X - \overline{X}\right)^2 \sum \left(Y - \overline{Y}\right)^2\right]}}$$

$$= \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$$

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

$y = Y - \overline{Y}$

Example:

From the following table calculate the Karl Pearson's coefficient of correlation:

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

Arithmetic mean of y is 8.

Solution:

$$\bar{y} = \frac{\sum y}{n} = \frac{35 + ?}{5} = 8 \Rightarrow ? = 5$$

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$$

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

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{-26}{\sqrt{40 \times 20}} = -0.92$$

Direct Method :

In case mean values of the two series in a bivariate data are fractional values and number of observations their volume in the two series is not very large, the following simplified form of formula may be used for calculating the value of ‘r’.

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

$$= \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]}\sqrt{\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Short – Cut Method :

When mean values are fractional and the number of paired observations is large, and the observations has large values, computing of 'r' can be simplified by using the deviations of the of the observations from some suitably chosen constant or constants. The constants for deviations of X and Y can be either same or different. The formula for computing correlation coefficient based on deviations is as under :-

$$r = \frac{N\sum d_x d_y - \left(\sum d_x\right)\left(\sum d_y\right)}{\sqrt{\left[N\sum d_x^2 - \left(\sum d_x\right)^2\right]}\sqrt{\left[N\sum d_y^2 - \left(\sum d_y\right)^2\right]}}$$

$$= \frac{\frac{\sum d_x d_y}{N} - \frac{\sum d_x}{N}\frac{\sum d_y}{N}}{\sqrt{\frac{\sum d_x^2}{N} - \left(\frac{\sum d_x}{N}\right)^2}\sqrt{\frac{\sum d_y^2}{N} - \left(\frac{\sum d_y}{N}\right)^2}}$$

$$= \frac{\frac{\sum d_x d_y}{N} - \frac{\sum (X - A)}{N}\frac{\sum (Y - B)}{N}}{\sigma_x \sigma_y}$$

$$= \frac{\frac{\sum d_x d_y}{N} - \left(\frac{\sum X}{N} - \frac{NA}{N}\right)\left(\frac{\sum Y}{N} - \frac{NB}{N}\right)}{\sigma_x \sigma_y}$$

$$= \frac{\frac{\sum d_x d_y}{N} - \left(\overline{X} - A\right)\left(\overline{Y} - B\right)}{\sigma_x \sigma_y} = \frac{\sum d_x d_y - N\left(\overline{X} - A\right)\left(\overline{Y} - B\right)}{N\sigma_x \sigma_y}$$

Assumptions of Karl Pearson's Coefficient :

Karl Pearson's coefficient of correlation as based on the following assumptions :-

(i) Linear Relationship :

In this method a linear relationship between two variables is **assumed**. In such case, the paired observations on the two variables plotted on a scatter – diagram cluster around a straight line.

(ii) Causal Relationship :

In studying correlation, we expect a cause and effect relationship between the forces affecting the values in the two series.

Merits of Karl Pearson's Coefficient of Correlation :

1. It is important and popular method of measuring the relationship between two variables. It gives a precise and quantitative value indicating the degree of relationship existing between the two variables. The value of 'r' is easily interpretable.
2. It measures the direction as well as the relationship between the two variables.

Demerits of Karl Pearson's Coefficient of Correlation :

1. The value of the coefficient is affected by the extreme values.
2. Its computational procedure is difficult as compared to other methods.
3. It assumes the **Linear Relationship** between the two variables.

Example 1 :

Calculate the correlation coefficient between the height of father and height of son from the given data :

Table: 1(Heights of Father's and Son's)

Height of Father (in inches)	64	65	66	67	68	69	70
Height of Sun (in inches)	66	67	65	68	70	68	72

Table: 2(Calculation for ‘r’)

Height of Father (X)	Height of Son (Y)	(X - Mean) X – 67 = x	(Y- Mean) Y – 68 = y	x^2	y^2	xy
64	66	-3	-2	9	4	6
65	67	- 2	- 1	4	1	2
66	65	- 1	- 3	1	9	3
67	68	0	0	0	0	0
68	70	1	2	1	4	2
69	68	2	0	4	0	0
70	72	3	4	9	16	12
$\Sigma X = 469$	$\Sigma Y = 476$			$\Sigma x^2 = 28$	$\Sigma y^2 = 34$	$\Sigma xy = 25$

$$\overline{X} = \frac{\sum X}{N} = 67$$

$$\overline{Y} = \frac{\sum Y}{N} = 68$$

Since the actual Means of X and Y are whole numbers, we can use actual mean method of computing ‘r’.

$$r = \frac{\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\left[\sum \left(X - \overline{X}\right)^2 - \sum \left(Y - \overline{Y}\right)^2\right]}}$$

$$= \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$$

$$= \frac{25}{\sqrt{[28 \times 34]}} = 0.81$$

Case :

Table 3 shows the sales revenue and advertisement expenses of a company for past 10 months. Find the coefficient of correlation between the sales and advertisement.

Table 3: Sales and Advertisements for 10 months

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Ad (000 INR)	10	11	12	13	11	10	9	10	11	14
Sales (000 INR)	110	120	115	128	137	145	150	130	120	115

r= - 0.51

TABLE 4

Calculation of correlation coefficient between sales and advertisement

Month	*Sales (x)*	*Advertisement (y)*	*xy*	x^2	y^2
Jan	110	10	1100	12,100	100
Feb	120	11	1320	14,400	121
Mar	115	12	1380	13,225	144
Apr	128	13	1664	16,384	169
May	137	11	1507	18,769	121
June	145	10	1450	21,025	100
July	150	9	1350	22,500	81
Aug	130	10	1300	16,900	100
Sept	120	11	1320	14,400	121
Oct	115	14	1610	13,225	196
Sum	1270	111	14,001	162,928	1253

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} = \frac{10 \times 14{,}001 - (111) \times (1270)}{\sqrt{10 \times (162{,}928) - (1270)^2}\sqrt{10 \times (1253) - (111)^2}}$$

$$= \frac{140010 - 140970}{\sqrt{1{,}629{,}280 - 1{,}612{,}900} \times \sqrt{12{,}530 - 12{,}321}} = \frac{-960}{\sqrt{16{,}380} \times \sqrt{209}} = \frac{-960}{127.9843 \times 14.4568} = \frac{-960}{1850.2434}$$

$$= -0.51$$

Case :

A Computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results:

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

It was, however discovered at the time of checking that two pairs of observations were interpreted by a computer bug wrong. They were taken as (6, 14) and (8, 6) while correct values were (8, 12) and (6, 8). Prove that the correct value of correlation coefficient should be 2/3.

Solution:

Correct $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Correct $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Correct $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Correct $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Correct $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$

Corrected value of r is given by:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right] \times \left[n\sum Y^2 - (\sum Y)^2\right]}}$$

$$= \frac{250 \times 520 - 125 \times 100}{\sqrt{[25 \times 650 - (125)^2]} \times \sqrt{[25 \times 436 - (100)^2]}}$$

$$= \frac{2}{3}$$

Calculate the coefficient of correlation from the following data:

Age of husband	23	27	28	29	30	31	33	35	36	39
Age of wife	18	22	23	24	25	26	28	29	30	32

r = 0.9956

Find Karl Pearson's coefficient of correlation between sales and expenses of the following ten firms:

Firm	1	2	3	4	5	6	7	8	9	10
Sales (000 units)	50	50	55	60	65	65	65	60	60	50
Expenses (000 INR)	11	13	14	16	16	15	15	14	13	13

r = 0.7866

X	Y	$x = X - 31.1$	$y = Y - 25.7$	x^2	y^2	xy
23	18	-8.1	-7.7	65.61	59.29	62.37
27	22	-4.1	-3.7	16.81	13.69	15.17
28	23	-3.1	-2.7	9.61	7.29	8.37
29	24	-2.1	-1.7	4.41	2.89	3.57
30	25	-1.1	-0.7	1.21	0.49	0.77
31	26	-0.1	0.3	0.01	0.09	-0.03
33	28	1.9	2.3	3.61	5.29	4.37
35	29	3.9	3.3	15.21	10.89	12.87
36	30	4.9	4.3	24.01	18.49	21.07
39	32	7.9	6.3	62.41	39.69	49.77
$\Sigma X = 311$	$\Sigma Y = 257$			$\Sigma x^2 = 202.9$	$\Sigma y^2 = 158.1$	$\Sigma xy = 178.3$

Calculation of Coefficient of Correlation for Grouped Data:

The formula of calculating the correlation coefficient is:

$$r = \frac{\sum f d_x d_y - n\left(\frac{\sum f d_x}{n}\right)\left(\frac{\sum f d_y}{n}\right)}{\sqrt{n\frac{\sum f d_x}{n} - \left(\frac{\sum f d_x}{n}\right)^2}\sqrt{\frac{\sum f d_y}{n} - \left(\frac{\sum f d_y}{n}\right)^2}}$$

Case:

Calculate the coefficient of correlation from the following data:

Marks in Statistics	*Marks in Finance*					
	10	20	30	40	50	*Total*
5	2	4	1	4	1	12
10	8	2	5	1	×	16
15	×	3	2	1	×	6
20	×	1	3	2	4	10
25	×	×	4	2	×	6
Total	10	10	15	10	5	

	X	10	20	30	40	50	f	fd_y	fd_y^2	fd_xd_y
	d_x	-2	-1	0	+1	+2				
y	d_y									
5	-2	2(+8)	4(+8)	1(0)	4(-8)	1(-4)	12	-24	48	+4
10	-1	8(+16)	2(+2)	5(0)	1(-1)	×	16	-16	16	+17
15	0	×	3(0)	2(0)	1(0)	×	6	0	0	0
20	+1	×	1(-1)	3(0)	2(+2)	4(+8)	10	+10	10	+9
25	+2	×	×	4(0)	2(+4)	×	6	+12	24	+4
Total		10	10	15	10	5	50	-18	98	34
	fd_x	-20	-10	0	+10	+10	-10			
	fd_x^2	40	10	0	10	20	80			
	fd_xd_y	+24	+9	0	-3	+4	34			

$$r = \frac{N\sum f d_x d_y - \sum f d_x \sum f\, d_y}{\sqrt{N\sum f d_x{}^2 - (\sum f d_x)^2}\sqrt{N\sum f d_y{}^2 - (\sum f d_y)^2}}$$

$$= \frac{50 \times 34 - (10)(-18)}{\sqrt{50 \times 80 - (-10)^2}\sqrt{98 \times 50 - (-18)^2}}$$

$$= 0.36$$

Probable Error:

By using probable error we can find the extent to which it is dependable. It is denoted by P.E.(r) its an old measure of using the reliability of an observed value of correlation coefficient in so far as it depends on the conditions of random sampling.

If r is the observed correlation coefficient in a sample of n pairs of observations then its standard error, usually denoted by S.E.(r) is given by;

$$S.E.(r) = \frac{1-r^2}{\sqrt{n}}$$

Probable error of the coefficient of correlation is given by;

$$P.E.(r) = 0.6745 \times S.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

The reason behind taking the range 0.6745 is that 50% of the observations lie between $\mu \pm 0.6745\,\sigma$, where μ is the mean and σ is standard deviation.

x
2
3
4
5
6
7
8
9
10

Mean $\mu = 6$

S.D. $\sigma = 2.5$

Mean $\mu \pm \sigma$ will contain almost 65.8% values of the observations. For the given observations almost 6 values

Mean $\mu \pm \sigma$ = [6 – 2.5, 6 + 2.5] = [3.5, 8.5] = [4, 5, 6, 7, 8]

Mean $\mu \pm 2\sigma$ will contain almost 95% values of the observations. For the given observations almost 9 values

Mean $\mu \pm 2\sigma$ = [6 – 5, 6 + 5] = [1, 11] = [2, 3, 4, 5, 6, 7, 8, 9, 10]

Mean $\mu \pm 3\sigma$ will contain almost 99% values of the observations. For the given observations almost 9 values

Mean $\mu \pm 3\sigma$ = [6 – 7.5, 6 + 7.5] = [- 1.5, 13.5]

Area wheat	*Area other*					
	0 –	500 -	1000 -	1500 -	2000 – 2500	*Total*
0 – 200	12	6	-	-	-	18
200 – 400	2	18	4	2	1	27
400 – 600	×	4	7	3	-	14
600 – 800	×	1	-	2	1	4
800 – 1000	×	×	-	1	2	3
Total	14	29	11	8	4	66

- 0.746

Spearman's Rank Correlation Coefficient :

Rank Correlation Coefficient permits us to correlate two sets of positive of qualitative observations which are subject to ranking such as qualitative productivity ratings (poor, fair, good, very good, etc.) for a group of workers by two independent observers. This will also give an idea whether the two observers have common or different tastes likings in a particular attribute or characteristics. Ranks can be assigned either by two persons (called judges) to a single characteristics, say, beauty, honesty, intelligence, etc., or by a single person or two characteristics. When the marks are assigned by two persons to a single characteristics, the correlation is found between the opinion or tastes of the two persons. High positive correlation indicates that the two persons have the same taste in that characteristic. If two characteristics are judged by the same person, e.g., marks obtained in training and quantum of sales, then correlation is found between two characteristics.

To calculate the Rank Correlation Coefficient :

1. We first rank the two series say X's and Y's individually among themselves, giving rank 1 to the largest (or smallest) value, rank 2 to the second largest (second smallest) and so on in each series separately.
2. Find the differences 'D' of the corresponding Ranks of X and Y.
3. Sequence these differences and find the sum of the squares of these differences, i.e., ΣD^2.
4. Calculate rank correlation coefficient by using the formula :

$$R = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)}$$

Where, 'N' denotes the number of paired values.

The above formula is applicable when no value in any of the two series is repeated. (Repeated values are known as tied values and are given the same Rank). When there are ties, we assign to each of the observations the mean of the ranks which they jointly occupy.

For Example:

If the third and fourth largest values of a variable are the same, we assign to each values, the rank = (3 + 4)/2 = 3.5 and if the fifth, sixth and seventh largest values of a variable are the same, we assign to each rank = (5 + 6 + 7)/3 = 6.

When some of the values are repeated and average ranks are assigned, the following formula is used to calculate rank correlation coefficient,

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

Where m = number of times a particular value is repeated. Repetition of values can be one series or both the series. Repetition can be in one value or more than one value.

Ex:

From following data, find out coefficient of rank correlation between price and supply.

Price	4	6	8	10	12	14	16	18
Supply	10	15	20	25	30	35	40	45

Solution :

Price	Rank (R_1)	Supply	Rank (R_2)	D = ($R_2 - R_1$)	D^2
4	8	10	8	0	0
6	7	15	7	0	0
8	6	20	6	0	0
10	5	25	5	0	0
12	4	30	4	0	0
14	3	35	3	0	0
16	2	40	2	0	0
18	1	45	1	0	0

$$R = 1 - \frac{6\sum D^2}{N\left(N^2 - 1\right)} = 1 - \frac{0}{8\left(8^2 - 1\right)} = 1$$

Ex:

From following data, find out coefficient of rank correlation between price and supply.

x	50	33	40	10	15	15	65	24	15	57
y	12	12	24	6	15	4	20	9	6	18

Solution :

x	Rank (R_1)	y	Rank (R_2)	D = ($R_2 - R_1$)	D^2
50	3	12	5.5	2.5	6.25
33	5	12	5.5	0.5	0.25
40	4	24	1	+ 3.0	9.00
10	10	6	8.5	+ 1.5	2.25
15	8	15	4	+ 4.0	16.00
15	8	4	10	2.0	4.00
65	1	20	2	1.0	1.00
24	6	9	7	1.0	1.00
15	8	6	8.5	0.5	0.25
57	2	18	3	1.0	1.00

Here in the first series, i.e., X series value 15 is repeated 3 times, in the Y series, the values 12 and 6 are each repeated twice.

$\therefore$ Rank correlation coefficient

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

$$= 1 - \frac{6\left[\sum D^2 + \sum \frac{m_1\left(m_1^{\,2} - 1\right) + m_2\left(m_2^{\,2} - 1\right) + m_3\left(m_3^{\,2} - 1\right)}{12}\right]}{N\left(N^2 - 1\right)}$$

$$= 1 - \frac{6\left[41 + \frac{3(9-1) + 2(4-1) + 2(4-1)}{12}\right]}{10 \times (100 - 1)}$$

$$= 1 - \frac{6\left[41 + \frac{24 + 6 + 6}{12}\right]}{990}$$

$$= 1 - \frac{44 \times 6}{990} = 0.733$$

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

R = 0.7333

S. No	1	2	3	4	5	6	7	8	9	10
X	12	18	32	18	25	24	25	40	38	22
y	16	15	28	16	24	22	28	36	34	19

R = 0.95

Marks in Statistics	30	38	28	27	28	23	30	33	28	35
Marks in Mathematics	29	27	22	29	20	29	18	21	27	22

R = - 0.3515

Twelve entries in painting competition were ranked by two judges as shown below:

Entry	A	B	C	D	E	F	G	H	I	J	K	L
Judge 1	5	2	3	4	1	6	8	7	10	9	12	11
Judge 2	4	5	2	1	6	7	10	9	11	12	3	8

What degree of agreement between two judges?

R = 0.46

Probable Error:

Probable error is the coefficient of correlation is given by:

$$P.E. = 0.6745\left(\frac{1-r^2}{\sqrt{n}}\right) = 0.6745\,(S.E)\; where, S.E. = \left(\frac{1-r^2}{\sqrt{n}}\right)$$

"The probable error of the correlation coefficient is an amount which if added to or subtracted from the average correlation coefficient produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall."

Regression Analysis

Introduction:

Regression means stepping back towards the average. In statistics regression analysis is applicable to all those fields where two or more related variables have the tendency to go back to mean.

According to Blair "Regression is the measure of average relationship between two or more variables in terms of original units of data."

The chief objective of Regression analysis is to know the nature of relationship between two variables and to use it for predicting the most likely value of the dependent variable corresponding to a given known value of the independent variable. However it may be noted that the regression relation is not reversible, i.e.

The regression equation used to predict the value of Y from a given value of X can not be used to predict the value of X from a given value of Y.

So, the regression relation is average, irreversible and functional relation.

There are two regression equations used:

We use regression equation of Y on X to predict the value of Y from the given value of X

And,

Regression equation of X on Y is used to predict the value of X from the given value of Y.

The Irreversible Relation:

1. The increment in family income shows an increment in expenditure but the increment in the expense of the family does not show the increment in family income.

2. If the rainfall is timely and good the crop will be good but if the crop is good there is not guarantee that the rainfall is timely and good.

Difference between Correlation and Regression:

Correlation	Regression
It is merely concerned with determining how strongly the two variables are linearly related.	It precedes correlation.
Not able to solve the prediction problems.	It solves the prediction problems.
Coefficient of correlation is independent of the change of the origin and scale.	Coefficient of regression is independent of the change of origin only.
Coefficient of correlation satisfies the relation $-1 \leq r \leq +1$.	Coefficient of regression satisfies the relation $0 \leq r^2 \leq 1$.

Regression Lines:

"The device used for estimating the value of one variable from the value of other consist of a line through the points drawn in such a manner as to represent the average relationship between the two variables. Such a line is called the line of regression".

J R Stockton

As per the method of least squares, two regression lines are:

$$Y - \overline{Y} = b_{YX}\left(X - \overline{X}\right) \text{.....}(1)$$

and

$$X - \overline{X} = b_{XY}\left(Y - \overline{Y}\right) \text{.....}(2)$$

where ,

$\overline{X}$ = Mean of series X.

$\overline{Y}$ = Mean of series Y.

$b_{YX} = r\dfrac{\sigma_Y}{\sigma_X} =$ Regression coefficient of Y on X.

$b_{XY} = r\dfrac{\sigma_X}{\sigma_Y} =$ Regression coefficient of X on Y.

Why do we need two regression lines to find the value of two variables X and Y

Since the regression relation is irreversible, one equation is not sufficient to predict the value of two variables X and Y. Moreover two regression equations are derived under different sets of assumptions, therefore one equation is not sufficient to find X and Y.

Properties of Regression Coefficients:

1. Both the regression coefficients should be of same sign.

$$(b_{YX}) \times (b_{XY}) = r^2$$

2. Correlation coefficient is the G.M. of two regression coefficients.

$$r = \sqrt{(b_{YX})(b_{XY})}$$

3. Both regression coefficients can not be more than 1.

4. Regression coefficients denote the rate of change.

Properties of Regression Lines:

1. The A.M. of X and A.M. of Y lies on the regression lines.

2. If r = 0, two regression lines are perpendicular to each other.

3. If two regression lines are identical, the correlation between the variables is perfect.

4. Angles between the regression lines can be given by

$$\tan\ \theta\ =\ \left[\ \frac{\sigma_X \sigma_Y}{\sigma_X^2\ +\ \sigma_Y^2}\left(\ \frac{1-r^2}{r}\right)\right]$$

Example:

A survey was conducted to study the relationship between expenditure on accommodation (x) and expenditure on the entertainment (y) and the following results were obtained:

Expenditure on	Mean	S.D.
Accommodation	173	66
Entertainment	47.8	22
Correlation coefficient	0.57	

Estimate the expenditure on entertainment if the expenditure on accommodation is 200.

Solution:

$Here\ ,$

$\overline{X} = 173$

$\overline{Y} = 47.58$

$\sigma_x = 66 \quad \sigma_y = 22 \quad r = 0.57$

$\therefore$

$$b_{yx} = r\frac{\sigma_x}{\sigma_y} = 0.57 \times \frac{22}{66} = 0.19$$

$$Y - \overline{Y} = b_{YX}\left(X - \overline{X}\right)$$

$$Y = 14.71 + 0.19\,X$$

$for \quad X = 200$

$$Y = 14.71 + 0.19 \times 200 = 52.71$$

Meaning of Regression

- Meaning of Regression is an act of returning or going back.

- It was first used in 1877 by "Sir Fransis Galton". While studying the relationship between the height of father and sons.

- The statistical tool with the help of which we are in a position to estimate(predict) the unknown values of one variable from known values of another variable is called **Regression.**

Significance of Regression Analysis

It can be expressed under following heads:

1. The relationship of cause and effect between two or more variables can be analyzed with the help of regression analysis.

2. The change in the value of one variable can be determined from regression coefficient if there is change of a unit in the value of other variable.

3. It provides estimates of values of the dependent variable on the basis of values of the independent variable in the areas of social, economic and business activities.

4. In the field of business, regression is very useful because with the help of it a businessman can predicting future production, consumption, investment, prices, profits, sales, etc.

Types of Regression

1. Simple regression :

If regression analysis is based only on two variable, is called simple regression. A simple regression is one which is confined to only two variables say, X and Y. Here 'X' is a independent variable and 'Y' is a dependent variable. The functional relationship between X and Y;
i.e. , $Y = f(X)$

2. Multiple regression:

If more than two variables are studied at a time in regression analysis, it is called multiple regression. A multiple regression analysis is one which is made among more than two related variables at a time say X,Y and Z. The functional relationship in such case is expressed as under;

$Y = f(X,Z)$, *or* $X = f(Y,Z)$, *or* $Z = f(X,Y)$.

3. **Linear regression**: If the regression line is in the form of a straight line, it indicates linear regression between the variables under the study in case of linear regression the values of the dependent variable changes at a constant rate for a unit change in the value of the independent variable. This constant change may be in terms of absolute amount or percentage.

4. **Curvi-linear or non-linear regression**: If the regression line is not a straight line but a smoothed curve, regression is termed as curvi-linear or non-linear.

Regression Equations

X on Y

- This equation describes the variation in the values of X for the given changes in Y.
- It estimates the value of X for the given value of Y.

$$\left(X-\overline{X}\right)=r\times\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

X=Value of X variable to be predicted

$\overline{X}$ =Arithmetic Mean of X series

r=Correlation Coefficient of X and Y series

σ_x =Standard deviation of X series

σ_y = Standard deviation of Y series

Y=That value of Y variable, corresponding to which the vale of X variable is to be predicted

$\overline{Y}$ =Arithmetic Mean of Y series

Y on X

- This equation describes the variation in the values of Y for the given changes in X.
- It estimates the value of Y for the given value of X.

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

$\overline{Y}$ =Value of Y variable to be predicted

$\overline{X}$ =Arithmetic Mean of X series

r=Correlation Coefficient of XandY series

σ_x =Standard deviation of X series

σ_y = Standard deviation of Yseries

Y=That value of X variable, corresponding to which the vale of $\overline{Y}$ variable is to be predicted

$\overline{Y}$ =Arithmetic Mean of Y series

EXAMPLE 1

- The following information are given to you

	Husband Age	Wife Age
MEAN	25 years	22 years
STANDARD DEVIATION	4 years	5 years

- Coefficient of correlation between ages of husband and wives =+0.8
- Find the expected age of husband when wives age is 12 years &expected age of wife when husband age is 33 years.

Given that: $\overline{X} = 25, \overline{Y} = 22, \sigma_x = 4, \sigma_y = 5, r = 0.8$

Regression equation of X on Y

$$\left(X - \overline{X}\right) = r \times \frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$$\left(X - 25\right) = 0.8 \times \frac{4}{5}\left(Y - 22\right)$$

$$\left(X - 25\right) = 0.64(Y - 22)$$

$$X = 0.64Y - 14.08 + 25$$

$$X = 0.64Y + 10.92$$

if the age of wife is 12 then the age of husband is :-

X=0.64*12+10.92

=7.68+10.92

=18.60 years

Regression equation of Y on X

$$\left(Y - \overline{Y}\right) = r \times \frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

$$\left(Y - 22\right) = 0.8 \times \frac{5}{4}\left(X - 25\right)$$

$$Y - 22 = 1.0 \times \left(X - 25\right)$$

$$Y = X - 25 + 22$$

$$Y = X - 3$$

If the age of husband is 33 years then the age of wife is :-

Y= 33-3

Wife age =30 Years

REGRESSION COEFFICIENTS

This coefficient indicates that if there is a unit change in the value of one variable, than what will be the average change in the value of other variable. Since there are two regression equation, therefore, there are two regression coefficient-regression coefficient of X on Y and regression coefficient of Y on X.

- Regression coefficient of X on Y (bxy)
- This coefficient represents the change in the value of X for a unit change in the value of the variable Y
- When X and Y series are given and deviations have been taken from assumed mean in one or in both series

$$bxy = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sum d^2 y \times N - \left(\sum dy\right)^2}$$

- Regression coefficient of Y on X (byx)
- This coefficient represents the change in the value of Y for a unit change in the value of variable X
- When X and Y series are given and deviations have been taken from assumed mean in one or both series

$$byx = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sum d^2 x \times N - \left(\sum dx\right)^2}$$

Example 2

From the following data calculate:

(a) the two regression coefficients

(b) the two regression equation

Population(in thousands)	18	19	20	21	22	23	24	25	26	27
No.of TV sets demanded	14	16	16	18	18	19	20	20	21	21

X	dx from 23	d^2x	Y	dy from 18	d^2y	dxdy
18	-5	25	14	-4	16	20
19	-4	16	16	-2	4	8
20	-3	9	16	-2	4	6
21	-2	4	18	0	0	0
22	-1	1	18=A	0	0	0
23=A	0	0	19	1	1	0
24	1	1	20	2	4	2
25	2	4	20	2	4	4
26	3	9	21	3	9	9
27	4	16	21	3	9	12
$\sum X = 225$	$\sum dx = -5$	$\sum d^2x = 85$	$\sum Y = 183$	$\sum dy$=3	$\sum d^2y = 51$	$\sum dxdy = 61$

REGRESSION COEFFICIENT

X on Y

$$bxy = \frac{\sum dxdy \times N - \left(\sum dx \times \sum dy\right)}{\sum d^2 y \times N - (\sum dy)^2}$$

$$= \frac{61 \times 10 - (-5 \times 3)}{51 \times 10 - (3)^2}$$

$$= \frac{610 + 15}{510 - 9}$$

$$= \frac{625}{501}$$

$$= 1.247$$

Y on X

$$byx = \frac{\sum dxdy \times N - (\sum dx \times \sum dy)}{\sum d^2 x \times N - (\sum dx)^2}$$

$$= \frac{61 \times 10 - (-5 \times 3)}{85 \times 10 - (-5)^2}$$

$$= \frac{610 + 15}{850 - 25}$$

$$= \frac{625}{825}$$

$$= .758$$

(b) Regression Equations

X on Y

Regression Equation of X on Y:This equation describes the variation in the values of X for the given changes in Y.

$$X - \bar{X} = bxy\ (Y - \bar{Y})$$

$$X - 22.5 = 1.247\ (Y - 18.3)$$

$$X - 22.5 = 1.247\ Y - 22.82$$

$$X = 1.247\ Y - 22.82 + 22.5$$

$$X = 1.247\ Y - 0.32$$

Y on X

Regression Equations of Y on X:

This equation describes the variation in the

Values of Y for the given changes in X

$$Y - \bar{Y} = byx\ (X - \bar{X})$$

$$Y - 18.3 = 0.758\ (X - 22.5)$$

$$Y - 18.3 = 0.758\ X - 17.055$$

$$Y = 0.758\ X - 17.055 = 18.3$$

$$Y = 0.758\ X + 1.245$$

REGRESSION LINES

- Regression lines are the lines of best fit expressing mutual average relationship between two series. These lines give the best estimate of one variable for any given value of other variable.
- If we take the case of two variable X and Y we shall have two regression line as X on Y and Y on X .

- Obtain the regression equation of Y on X and X on Y from the following table giving the Sale of goods “X” and goods “Y”.

Sale of goods “Y”	Sale of goods “x”				
(in UNITS)	**5-15**	**15-25**	**25-35**	**35-45**	**total**
0-10	1	1	-	-	2
10-20	3	6	5	1	15
20-30	1	8	9	2	20
30-40	-	3	9	3	15
40-50	-	-	4	4	8
Total	5	18	27	10	60

X / Y		dx / dy	5-15	15-25	25-35	35-45	f	fd_y	fd_y^2	fd_xd_y
		M.P.	10	20	30	40				
			-1	0	1	2				
	M.P.									
0-10	5	-2	1 2	1 0	_	_	2	-4	8	2
10-20	15	-1	3 3	6 0	5 -5	1 -2	15	-15	15	-4
20-30	25	0	1 0	8 0	9 0	2 0	20	0	0	0
30-40	35	1	-	3 0	9 9	3 6	15	15	15	15
40-50	45	2	-		4 8	4 16	8	16	32	24
f			5	18	27	10	N=60	$\sum fd_y$ =12	$\sum fd_y^2$ =70	$\sum fd_xd_y$ = 37
fd_x			-5	0	27	20	$\sum fd_x$ =42			
fd_x^2			5	0	27	40	$\sum fd_x^2 = 72$			
fd_xd_y			5	0	12	20	$\sum fd_xd_y = 37$			

Regression Equation

- Y on X

$$Y - \overline{Y} = r \frac{\sigma_x}{\sigma_y}\left(X - \overline{X}\right)$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{N \sum fdxdy - \left(\sum fdx \sum fdy\right)}{N \sum fd_x^2 - \left(\sum fd_x\right)^2} \times \frac{i_y}{i_x}$$

$$= \frac{60\ (37) - (42)(12)}{60\ (72) - (42)^2} \times \frac{10}{10} = \frac{2220 - 504}{4320 - 1764} = \frac{1716}{2556} = 0.67$$

$$\overline{Y} = A + \frac{\sum fd_y}{N} \times i$$

$$\overline{Y} = 25 + \frac{12}{60} \times 10 = 27$$

$$\overline{X} = A + \frac{\sum fd_x}{N} \times i$$

$$\overline{X} = 20 + \frac{42}{60} \times 10 = 27$$

$$Y - 27 = 0.67\ (X - 27) = 0.67\ X - 18.09$$

$$Y - 27 = 0.67\ X - 18.09$$

$$Y = 8.91 + .67\ X$$

– X on Y

$$X - \overline{X} = r \times \frac{\sigma_x}{\sigma_y}\left(Y - \overline{Y}\right)$$

$$r\frac{\sigma_x}{\sigma_y} = \frac{N\sum fd_x d_y - \sum fd_x \sum fd_y}{N\sum fd_y^2 - \left(\sum fd_y\right)^2} \times \frac{i_x}{i_y}$$

$$\frac{60\left(37\right) - 42 \times 12}{60\left(70\right) - \left(12\right)^2} = \frac{2220 - 504}{4200 - 144} = \frac{1716}{4056} = 0.423$$

$$X - 27 = 0.423\left(Y - 27\right) = 0.423\,Y - 11.42$$

$$X = 15.58 + 0.423\,Y$$

REFERENCES

- *BUSINESS STATISTICS BY: S.P.GUPTA & M.P.GUPTA*

- *PRINCIPLE OF STATISTICS BY :*

  Dr. S.M.SHUKLA & Dr S. P. SAHAI

- *FUNDAMENTAL OF STATISTICS BY:*

  B.M.AGARWAL

THANK YOU