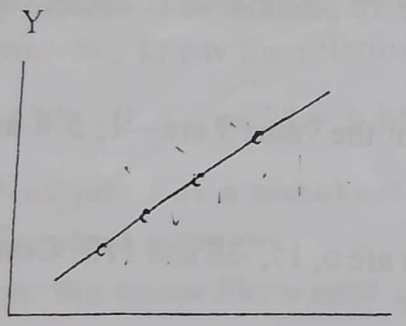
Correlation and Regression

9.1 Bivariate distribution:

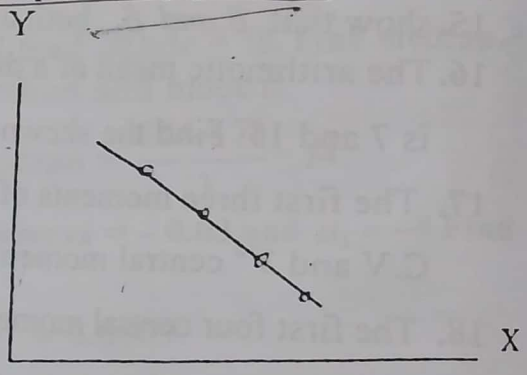
So far we have confined ourselves to univariate distribution involving only one variable. We may however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of person, we shall get what is known as Bivariate distribution – one variable relating to height and other variable relating to weight.

9.2 Scatter diagram and it's interpretation:

Scatter diagram: Two related variables when plotted in a graph paper in the form of dots is called dotogram or scatter diagram. In scatter diagram there are two axis one is X and other is Y. for each pair of X and Y values we put dots and thus obtain as many points as the number of observation. By looking to the scatter of the various points we may conform an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the lesser is the degree of relationship in between the two variables. The more nearly the points come to the line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive.

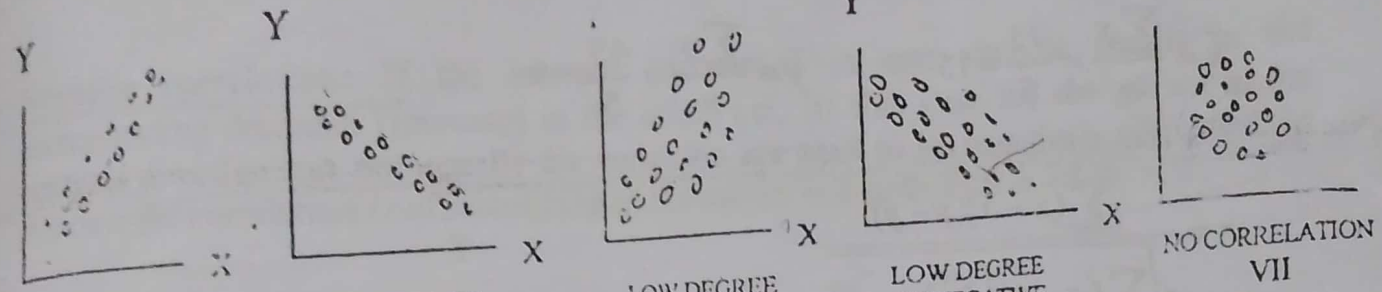


PERFECT POSITIVE CORRELATION
I



PERFECT NEGATIVE CORRELATION
II

(i.e, r = + 1) (diagram I). 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 right-hand corner of the diagram, correlation is said to be perfectly negative (i.e, r = -1) (diagram II). 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 upper left-hand corner to the lower left hand corner of the diagram (diagram IV). On the other hand if the points are widely scattered over the diagrams it indicates very low degree of 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 running from the upper left-hand side to the lower right hand side to 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 the absence of any relationship between the variables (i.e, r = 0) as shown by (diagram VII).



9.3 Correlation:

Correlation is defined as a connection or relationship between two or more statistical series. If the change in one variable effects a change in the other variable, the variable are said to be correlated. If the increase (decrease) in one variable result in the corresponding increase (decrease) in the others, i.e., if the changes are in the same direction, the variable are positively correlated. For example, the height and weifgts of a group of persons is positively correlated. If the increase (decrease) in one variable results in the corresponding decrease (increase) in the other i.e., in this the change are in the opposite direction the variables are said to be negatively correlated. For example, the volume and pressure of a perfect gas is negatively correlated. if the changes do not depict any of the above two types. The variables not corre...

9.4 Correlation-coefficient:

The co-efficient of correlation measures the degree of linear relationship between two variables or determines the degree of correlation between two related variables. Karl Pearson a great British Bio-metrician defined correlation co-efficient between x and y in 1980 as,

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

Direct method of computation for ungrouped data

Calculation of co-efficient of correlation from the following data by direct method.

Stock in the shop in taka (lakh) (x)	Profit in taka (lakh) (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
5	3	-10	-6	60	100	36
10	4	-5	-5	25	25	25
15	8	0	-1	0	0	1
20	12	5	3	15	25	9
25	18	10	9	90	100	81
$\sum x = 75$	$\sum y = 45$			$\sum(x-\bar{x})(y-\bar{y}) = 190$	$\sum(x-\bar{x})^2 = 250$	$\sum(y-\bar{y})^2 = 152$

$$\bar{x} = \frac{\sum x}{n} = \frac{75}{5} = 15 \qquad \bar{y} = \frac{\sum y}{n} = \frac{45}{5} = 9$$

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

or $r = \frac{190}{\sqrt{250 \times 152}} = \frac{190}{38000} = 0.9764$

9.5 Importance of correlation co-efficient:

Karl Pearson's correlation co-efficient method is most popular. The correlation coefficient summarizes in one figure not only the degree of correlation but also the degree, whether correlation is positive or negative.

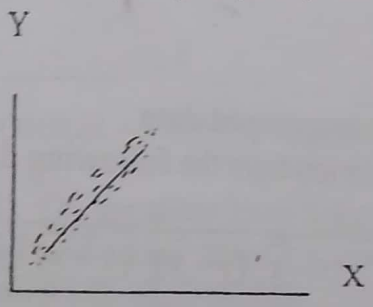
The following general guidelines are given which would help in interpreting the value of r.

1. When r = + 1 means there is a perfect positive correlation between the variables.
2. When r = -1 it means there is a perfect negative correlation between the variables.
3. When r = 0 it means there is no correlation between the variables. That is the variables are not correlated.

9.6 Classification of simple correlation:

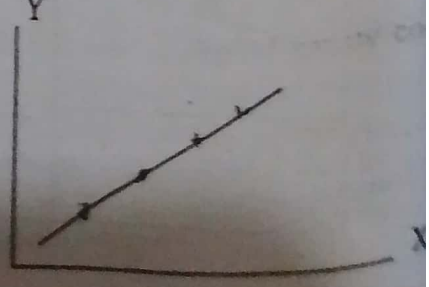
Correlation is classified in several different ways. Which are given below.

(a) **Positive correlation:** in the change in one variable effect a change in the other variable is said to be correlated. If the increase (decrease)in one variable results in the corresponding increase (decrease) in the others i.e., if the changes are in the same direction, but not equal, the variables are positively correlated. In this case the correlation coefficient between two variables exist in the range 0 to 1, i.e., $0<x<1$. In positive correlation r = 0.1, 0.3, 0.5, 0.8,.....

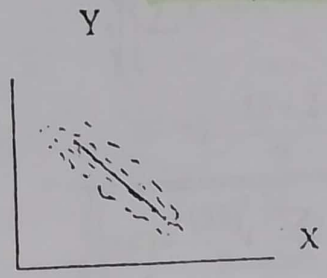


$0 < r < 1$

Perfect positive correlation: If the increase (decrease) in one variable result in the corresponding increase (decrease) in the others i.e., if the changes are in the same direction and equally, the variables are perfects positively correlated. In the case the correlation coefficient between two variables is 1.

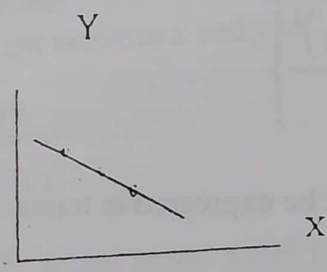


Negative correlation: If the increase (decrease) in one variable results in the corresponding decrease (increase) in the other i.e., in this case the change are in the opposite direction but not equally the variables are said to be negatively correlated. In this case the correlation coefficient exist in the range -1 to 0. That is $-1 < r < 0$.



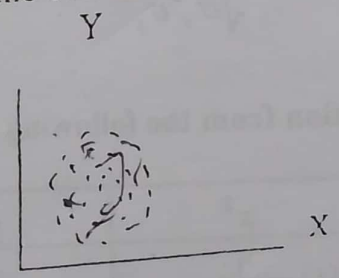
$-1 < r < 0$

Perfect Negative correlation: If the increase (decrease) in one variable results in the corresponding decrease (increase) in the other i.e., in this case the changes are equally in the opposite direction, then the variables are said to be perfect negatively correlated. In this the correlation coefficient $r = -1$

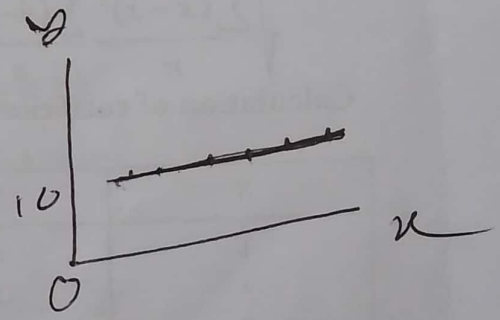


$r = -1$

Zero correlation: If x and y are independent variable then changes of one don't effect the others. In this case the correlation coefficient $r = 0$



$r = 0$



9.7 Karl Pearson's correlation coefficient can simplify in this following way

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

$$r_{xy} = \frac{\sum(xy - \bar{x}y - x\bar{y} + \bar{x}\bar{y})}{\sqrt{\sum(x^2 - 2x\bar{x} + \bar{x}^2)\sum(y^2 - 2y\bar{y} + \bar{y}^2)}}$$

$$r_{xy}=\frac{\sum(xy)-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}}{\sqrt{(\sum x^2-2n\bar{x}^2+n\bar{x}^2)(\sum y^2-2n\bar{y}^2+n\bar{y}^2)}}$$

$$r_{xy}=\frac{\sum(xy)-n\bar{x}\bar{y}}{\sqrt{(\sum x^2-n\bar{x}^2)(\sum y^2-n\bar{y}^2)}}$$

$$r_{xy}=\frac{\sum(xy)-n\frac{\sum x}{n}\frac{\sum y}{n}}{\sqrt{\left\{\sum x^2-\frac{(\sum x)^2}{n}\right\}\left\{\sum y^2-\frac{(\sum y)^2}{n}\right\}}}$$

$$r_{xy}=\frac{\sum(xy)-\frac{\sum x\sum y}{n}}{\sqrt{\left\{\sum x^2-\frac{(\sum x)^2}{n}\right\}\left\{\sum y^2-\frac{(\sum y)^2}{n}\right\}}}$$

The coefficient of correlation can also be expressed in terms of covariance as given below

$$r_{xy}=\frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}\frac{\sum(y-\bar{y})^2}{n}}} \quad \text{or} \quad r_{xy}=\frac{\text{cov}(x,y)}{\sqrt{\sigma_x^2\sigma_y^2}} \quad \text{or} \quad r_{xy}=\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$$

Calculation of coefficient of correlation from the following data

x	y	x^2	y^2	xy
1	-5	1	25	-5
4	2	16	4	8
7	-2	49	4	-14
8	-3	64	9	-24
10	6	100	36	60
12	1	144	1	12
13	3	169	9	39
$\sum x=55$	$\sum y=2$	$\sum x^2=543$	$\sum y^2=88$	$\sum xy=76$

$$\text{Co-efficient of correlation } r_{xy} = \frac{\sum(xy) - \frac{\sum x \sum y}{n}}{\sqrt{\left\{\sum x^2 - \frac{(\sum x)^2}{n}\right\}\left\{\sum y^2 - \frac{(\sum y)^2}{n}\right\}}}$$

$$= \frac{76 - \frac{55 \times 2}{7}}{\sqrt{\left\{543 - \frac{(55)^2}{7}\right\}\left\{88 - \frac{(2)^2}{7}\right\}}} = \frac{60.29}{\sqrt{110.86 \times 87.43}} = 0.61$$

Short-cut method of computation for ungrouped data

Calculation of co-efficient of correlation from the following data.

Production in mds. (x)	Pricr in tk. (y)	u = x-36	v = y - 35	uv	u^2	v^2
25	30	-11	-5	55	121	25
27	35	-9	0	0	81	0
36	34	0	-1	0	0	1
42	40	6	5	30	36	25
55	38	9	3	27	81	9
Total		-5	2	112	319	60

Co-efficient of correlation

$$r = \frac{\sum(uv) - \frac{\sum u \sum v}{n}}{\sqrt{\left\{\sum u^2 - \frac{(\sum u)^2}{n}\right\}\left\{\sum v^2 - \frac{(\sum v)^2}{n}\right\}}} = \frac{112 - \frac{-5\times 2}{5}}{\sqrt{\left\{319 - \frac{(-5)^2}{5}\right\}\left\{60 - \frac{(2)^2}{5}\right\}}} = \frac{114}{\sqrt{314\times 59.2}} = 0.84$$

2. Correlation co-efficient between x and y takes values from -1 to $+1$, i.e. $-1 \le r \le 1$

Proof: Let us consider

$$\left\{\frac{(x_i-\bar{x})}{\sigma_x} \pm \frac{(y_i-\bar{y})}{\sigma_y}\right\}^2 \ge 0$$

$$\Rightarrow \frac{(x_i-\bar{x})^2}{\sigma_x^2} \pm \frac{2(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} + \frac{(y_i-\bar{y})^2}{\sigma_y^2} \ge 0$$

$$\Rightarrow \frac{\sum(x_i-\bar{x})^2}{\sigma_x^2} \pm \frac{2\sum(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} + \frac{\sum(y_i-\bar{y})^2}{\sigma_y^2} \ge 0$$

$$\Rightarrow \frac{n\sigma_x^2}{\sigma_x^2} \pm \frac{2n\,\mathrm{cov}(x,y)}{\sigma_x\sigma_y} + \frac{n\sigma_y^2}{\sigma_y^2} \ge 0$$

$$\Rightarrow n \pm 2nr_{xy} + n \ge 0 \Rightarrow 2n \pm 2nr_{xy} \ge 0$$

$$\Rightarrow 1 \pm r_{xy} \ge 0 \Rightarrow 2n(1 \pm r_{xy}) \ge 0$$

$$\Rightarrow 1 + r_{xy} \ge 0, 1 - r_{xy} \ge 0 \Rightarrow 1 \pm r_{xy} \ge 0$$

$$\Rightarrow r_{xy} \ge -1, r_{xy} \le 1$$

$$\therefore -1 \le r_{xy} \le 1$$

$1 - r_{xy} \ge 0 \Rightarrow -r_{xy} \ge -1 \Rightarrow r_{xy} \le 1$

$1 + r_{xy} \ge 0 \Rightarrow r_{xy} \ge -1$

Alternative Method

Proof: $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$

Let $a = \dfrac{x-\bar{x}}{\sqrt{\sum(x-\bar{x})^2}}$, $b = \dfrac{y-\bar{y}}{\sqrt{\sum(y-\bar{y})^2}}$

and

$$\sum a^2 = \frac{\sum(x-\bar{x})^2}{\sum(x-\bar{x})^2} = 1, \sum b^2 = \frac{\sum(y-\bar{y})^2}{\sum(y-\bar{y})^2}$$

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

$$\sum(a+b)^2 = \sum a^2 + 2\sum ab + \sum b^2$$

$$= 1 + 2r + 1 = 2(1+r) \ge 0$$

$$\text{or } 1 + r \ge 0 \quad \text{or} \quad r \ge -1 \quad \ldots\ldots(i)$$

Similarly

Then $\sum(a-b)^2 = \sum a^2 - 2\sum ab + \sum b^2$

$$= 1 - 2r + 1 = 2(1-r) \ge 0$$

$$\text{or } 1 - r \ge 0 \quad \text{or} \quad r \le 1 \quad \ldots\ldots(ii)$$

from (i) and (ii) we get $-1 \le r \le$

3.The correlation co-efficient 'r' is symmetrical with respect to x and y. i.e., $r_{xy} = r_{yx}$

4. The correlation coefficient 'r'is the geometric mean between two regression co-efficient, i.e., $r = \sqrt{b_{xy} \times b_{yx}}$.

5.For perfect positive correlation or negative correlation, there regression equation is identical .

Proof: For perfect correlation we have $r = \pm 1 \quad \Rightarrow r^2 = 1$

We know $r = \sqrt{b_{xy} \times b_{yx}}$. $\Rightarrow r^2 = b_{xy} b_{yx} \quad \Rightarrow b_{xy} b_{yx} = 1 \quad \therefore b_{yx} = \frac{1}{b_{xy}}$.

Now the regression line of y on x is $Y - \bar{y} = b_{yx}(x - \bar{x}) \cdots\cdots\cdots (i)$

Putting the value of $b_{yx} = \frac{1}{b_{xy}}$. in (i) we get $Y - \bar{y} = \frac{1}{b_{xy}}(x - \bar{x}) \Rightarrow X - \bar{x} = b_{xy}(y - \bar{y})$

Which indicates the regression line of x on y: So the given condition the above two-regression lines are identical.

9.9 Assumption of Pearsonian's coefficient

The Karl Pearson's coefficient of correlation is based on the following assumptions.

1. There are 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 .
2. 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.
3. 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.

9.10 Limitation of the Pearsonian coefficient

1. The correlation coefficient always assumes linear relationship regardless of the fact whether the assumption is true or not.
2. Great care must be exercised in interpreting the value of this coefficient, as very often the coefficient is misinterpreted.
3. The value of the coefficient is unduly affected by the extreme values.
4. As compared to other methods of finding correlation, this method is more time consuming.

9.11 Uses of correlation:

There are importance uses of correlation as follows:

1. Use in social, business and economic field.
2. To find the relationship between two variables.
3. To find the relationship between different variables and combined influence of a group of independent variables.

9.12 Make comments: r = 0, r > 1, r <1

There are three parts of the above statement. The first statement is true and other two statements are false.

r = o, means there are no relation between the two variables, i.e., correlation coefficient is zero. This statement is completely true for independent variables.

The second statement r > 1, indicates that coefficient of correlation is greater then one. Which is not possible, because $-1 \le x \le 1$

The third statement r < -1 is not true, because $-1 \le x \le 1$

9.13 Prove that when x and y are independent then co-efficient of correlation r = 0

Proof: Since x and y are independent then cov (x, y) = 0

$$\Rightarrow \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = 0$$

Now co-efficient of correlation x and y ,

r

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{0}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = 0 \quad \textbf{proved}$$

9.14 Problem: If y = mx + c, find the coefficient of correlation between x and y

Given that

$$y = mx + c$$

$$\bar{y} = m\bar{x} + c$$

We know that

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

$$r_{xy} = \frac{\sum(x-\bar{x})(mx+c-m\bar{x}-c)}{\sqrt{\sum(x-\bar{x})^2\sum(mx+c-m\bar{x}-c)^2}}$$

$$r_{xy} = \frac{m\sum(x-\bar{x})(x-\bar{x})}{m\sqrt{\sum(x-\bar{x})^2\sum(x-\bar{x})^2}}$$

$$= \frac{\sum(x-\bar{x})^2}{\sum(x-\bar{x})^2} = 1$$

$$\therefore r_{xy} = 1$$

D-2009

9.15 Problem: If $y = -\frac{x}{2}$, find r_{xy} and comment.

Solution:

$$y = -\frac{x}{2}$$

$$\therefore x = -2y$$

$$\therefore \bar{x} = -2\bar{y}$$

Now

$$r_{xy} = \frac{\sum(-2y + 2\bar{y})(y - \bar{y})}{\sqrt{\sum(-2y + 2\bar{y})^2 \sum(y - \bar{y})^2}}$$

$$= \frac{-2\sum(y - \bar{y})^2}{2\sum(y - \bar{y})^2} = -1$$

Comment: since r = -1, so the correlation is said to be perfectly negative

D-2010

9.16 Problem: If y = 3x+2, find the correlation coefficient between x and y.

Solution: Let the correlation between x and y is r_{xy}. now

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

$y = 3x + 2$

$\bar{y} = 3\bar{x} + 2$

Here $y = 3x+2$, $\sum y = 3\sum x + 2$

$$\therefore \bar{y} = 3\bar{x} + 2$$

$$\therefore r_{xy} = \frac{\sum(x - \bar{x})(3x + 2 - 3\bar{x} - 2)}{\sqrt{\sum(x - \bar{x})^2 \sum(3x + 2 - 3\bar{x} - 2)^2}}$$

$$r_{xy} = \frac{3\sum(x - \bar{x})(x - \bar{x})}{3\sqrt{\sum(x - \bar{x})^2 \sum(x - \bar{x})^2}}$$

$$= \frac{\sum(x - \bar{x})^2}{\sum(x - \bar{x})^2} = 1$$

$\sqrt{\sum(x_i - \bar{x})^2}$

Comment: Since r = 1, so the correlation is said to be perfectly positive.

D-09,10

9.17 Problem: If u = 3x+5, v = 7-5y, find r_{uv} given that $r_{xy} = 0.86$

Solution: Given that $r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = 0.86$ given that u = 3x+5

$\therefore \bar{u} = 3\bar{x} + 5$

v = 7-5y $\quad \therefore \bar{v} = 7 - 5\bar{y}$

Correlation coefficient between u and v

$$r_{uv} = \frac{\sum(u-\bar{u})(v-\bar{v})}{\sqrt{\sum(u-\bar{u})^2\sum(v-\bar{v})^2}}$$

$$r_{uv} = \frac{\sum(3x+5-3\bar{x}-5)(7-5y-7+5\bar{y})}{\sqrt{\sum(3x+5-3\bar{x}-5)^2\sum(7-5y-7+5\bar{y})^2}} = \frac{\sum(3x-3\bar{x})(-5y+5\bar{y})}{\sqrt{\sum(3x-3\bar{x})^2\sum(-5y+5\bar{y})^2}}$$

$$= \frac{-3\times5\sum(x-\bar{x})(y-\bar{y})}{3\times5\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{3\sum(x-\bar{x})(-5)(y-\bar{y})}{\sqrt{3^2\sum(x-\bar{x})^2\times5^2\sum(y-\bar{y})^2}}$$

$= -r_{xy} = -0.86$

9.18 Problem: If x and y are independent and σ_x^2, σ_y^2 are the variance of x and y respectively. Find the co-efficient of correlation x+y and x-y.

Solution: Let us suppose that n pair of values of (x,y) are

$(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots\ldots (x_n, y_n)$

Arithmetic mean of x and y are $\bar{x}$ and $\bar{y}$ respectively then

$$\sigma_x^2 = \frac{\sum(x_i-\bar{x})^2}{n} \text{ and } \sigma_y^2 = \frac{\sum(y_i-\bar{y})^2}{n}$$

Let, $u_i = x_i + y_i or \sum u_i = \sum x_i + \sum y_i \Rightarrow \frac{\sum u_i}{n} = \frac{\sum x_i}{n} + \frac{y_i}{n} \therefore \bar{u} = \bar{x} + \bar{y}$

Similarly $v_i = x_i - y_i and \bar{v} = \bar{x} - \bar{y}$

Now $\sigma_u^2 = v(x+y) = v(x) + v(y) = \sigma_x^2 + \sigma_y^2$ [since x,y are independent ,so cov(x,y) = 0]

$$\sigma_v^2 = v(x-y) = v(x) + v(y) = \sigma_x^2 + \sigma_y^2$$

$$\text{cov}(x,y) = \frac{1}{n}\sum(u_i-\bar{u})(v_i-\bar{v})$$

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

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

$$= \frac{1}{n}\sum(x_i - \bar{x})^2 - \frac{1}{n}\sum(y_i - \bar{y})^2 = \sigma_x^2 - \sigma_y^2$$

Correlation coefficient of u and v is

$$r_{uv} = \frac{\text{cov}(x,y)}{\sqrt{\sigma_u^2 \sigma_v^2}} = \frac{\sigma_x^2 - \sigma_y^2}{\sqrt{(\sigma_x^2 + \sigma_y^2)(\sigma_x^2 + \sigma_y^2)}} = \frac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}$$

9.20 Rank correlation

For some situation it is difficult to measure the values of the variables from bivariate distribution numerically but they can be ranked. The correlation coefficient between the two ranks is usually called rank correlation co-efficient. The British psychologist Charles Edward Spearman developed this correlation co-efficient in 1904.

For rank correlation show that $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$

Proof : Let (x_i, y_i) i = 1,2,3,4,5,6,.....n. denote the ranks of the ith individul of two characteristics A and B respectively. Each of the variables x and y takes the values 1,2,3,4,5.....n.

Hence $\sum x_i = 1+2+3+4+5+\cdots\cdots+n = \frac{n(n+1)}{2}$

$$\sum y_i = 1+2+3+4+5+\cdots\cdots+n = \frac{n(n+1)}{2}$$

$$\sum x_i^2 = \sum y_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + \cdots\cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Hence $d_i^2 = \sum(x_i - y_i)^2 = \sum x_i^2 - 2\sum x_i y_i + \sum y_i^2$

or $2\sum x_i y_i = \sum x_i^2 + \sum y_i^2 - \sum d_i^2$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)(2n+1)}{6} - \sum d_i^2$$

$$\sum x_i y_i = \frac{n(n+1)(2n+6)}{6} - \frac{\sum d_i^2}{2}$$

The correlation between x and y is $\rho = \dfrac{\sum(x_i y_i) - \dfrac{\sum x_i \sum y_i}{n}}{\sqrt{\left\{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}\right\}\left\{\sum y_i^2 - \dfrac{(\sum y_i)^2}{n}\right\}}}$

$$= \frac{\dfrac{n(n+1)(2n+1)}{6} - \dfrac{\sum d_i^2}{2} - \dfrac{\dfrac{n(n+1)}{2}\dfrac{n(n+1)}{2}}{n}}{\sqrt{\left\{\dfrac{n(n+1)(2n+1)}{6} - \dfrac{\left\{\dfrac{n(n+1)}{2}\right\}^2}{n}\right\}^2}}$$

$$= \frac{\dfrac{n(n+1)}{2}\left\{\dfrac{2n+1}{3} - \dfrac{n+1}{2}\right\} - \dfrac{\sum d_i^2}{2}}{\dfrac{n(n+1)(2n+1)}{6} - \dfrac{n^2(n+1)^2}{4n}}$$

$$\rho = \frac{\dfrac{n(n+1)}{2}\left\{\dfrac{4n+2-3n-3}{6}\right\} - \dfrac{\sum d_i^2}{2}}{\dfrac{n(n+1)}{12}\{4n+2-3n-3\}} = \frac{\dfrac{n(n+1)(n-1)}{12} - \dfrac{\sum d_i^2}{2}}{\dfrac{n(n+1)(n-1)}{12}}$$

$$= 1 - \frac{\sum d_i^2}{2} \times \frac{12}{n(n-1)}$$

$$= 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

9.20 (a) **Example :** The ranks of same 16 students in mathematics and physics are as follows. Two numbers within brackets denote the ranks of the students in mathematics and physics.

(1,1) (2,10) (3,3) (4,4) (5,5) (6, 7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (14,12) (15,16) (16,13)

Calculate the rank correlation coefficient for proficiencies of this group in mathematics and physics.

Solution.

Ranks in Maths. (X)	Rank in Physics (Y)	d =x-y	d^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
Total		0	136

Rank correlation coefficient is given by , $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 136}{16\times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0.8$

9.20(b)Example : A Group of students are ranked by two teachers on the basis of their merit. Computation of the rank correlation co-efficient between the two sets of rankings.

Solution.

Teacher A	Teacher B	d =x-y	d^2
1	2	-1	1
2	3	-1	1
3	4	-1	1
4	1	3	9
5	5	0	0
6	7	-1	1
7	8	-1	1
8	10	-2	4
9	6	3	9
10	9	1	1
Total			28

Rank correlation coefficient is given by, $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 28}{10\times 99} = 1 - 0.17 = 0.83$

9.21 Regression:

The term "regression" literally means, "stepping back towards the average". It was first used by a British Biometerician Sir Francis Galton (1822 –1911), in connection with the inheritance of stature. Galton found that the offspring of abnormally tall or short parents tend to "regress"or "step back" to the average population height. But the term "regression" as now used in statistics is only a convenient term without having any reference to biometry.

Definition: The statistical tool with the help of which we are in a position to estimate the unknown values of one variable from known values of another variable is called regression. With the help of regression analysis, we are in a position to find out the average probable change in one variable given a certain amount of change in another. In regression analysis there are two types of variables. The variables whose value is to be predicted is called dependent variable and the variable which is used for prediction, is called independent variable. In regression analysis independent variable is also known as regressor or predictor variable while the dependent variable is also known as regressed or explained variable.9.22

9.22 Francis Galton's concept about fathers and sons :

The term "regression" was first used in 1877 by Francis Galton while studying the relationship between the height of fathers and sons. He study of height of about one thousand fathers and sons revealed a very interesting relationship, i.e., tall fathers tend to have tall sons and short fathers short sons; but the average height of the sons of a group of tall fathers is less than that of the tall fathers and the average height of the sons of a group of short fathers is greater than that of the short fathers.

9.23 Line of Regression

If the variables in a bivariate distribution are related, we will find the points in the scatter diagram will cluster round some curve called the "curve of regression". If the curve is a straight line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear. The line of regression is the line which give the best estimate to the value of one variable for any specific value of the other variable. Thus the line of Regression is the line of "best fit" and is obtained by the principle of least squares

Let us suppose that in the bivariate distribution (x_i, y_i); I =1,2,3,.....n.

Y is dependent variable and X is independent variable. The regression line of Y on X will be $Y = a + bX$.

Regression equation

Regression lines are fitted to the observed data by means of regression equation. The two regression equation for two straight regression lines are.

i. $Y = a_1 + b_1 x$ (Regression equation of yon x)

ii. $X = a_2 + b_2 y$ (Regression equation of x on y)

Co-efficient of regression:

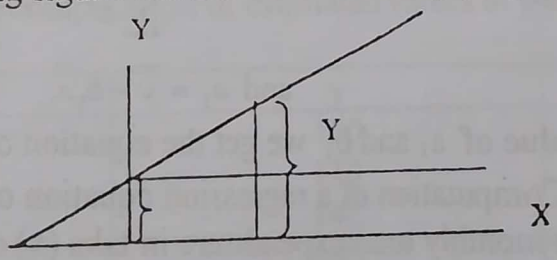
Consider two regression equations are

i. $Y = a_1 + b_1 x$ (Regression equation of yon x)

iii. $X = a_2 + b_2 y$ (Regression equation of x on y)

In this equations a_1 , a_2 and b_1 , b_2 are constants. b_1 gives the slope of the regression line of y on x to the x – axis and is called co-efficient of regression of y on x . Thus the coefficient of regression of y on x is the amount of change in y for an unit change in x. Similarly, b_2 is the slope of the regression line of x on y to the y-axis and thus give the amount of change in x for a unit change in y. This b_2 is called the co-efficient of regression of x on y .

The following figure will show these clearly (Regression equation of x on y)



9.24 Least square method

The basis assumption of least square method is that the sum of the squares of the deviation or distances of individual points from the fitted line should be minimum. In fitting the regression line of y on x we minimize the sum of squares of the deviation along the vertical axis and in case of fitting the regression line of x on y the distances to be made minimum are taken along the horizontal axis (x-axis) by minimizing the distances of the point from the line we try to minimize the error involved in our prediction. In other words, the deviations of the points from the line give us the measure of the error of estimate or prediction. We minimize the

sum of squares of deviations and not the sum of deviations simply because sum of deviation may be zero (since some of the deviations are positive while others are negative)

The figure (page - 120) will show the process of measuring deviation of the points from the line of best fit.

Let the deviations of the points from the line be denoted by d. If y_1, y_2, y_3……,etc., are the observed values of variable and Y_1, Y_2, Y_3 ……,etc., are the corresponding estimated values which are on the line of best fit then we may denoted d_1, d_2, d_3, ……,etc., as follows; $d_1 = y_1 - Y_1$, $d_2 = y_2 - Y_2$, $d_3 = y_3 - Y_3$ ……, etc., If we now substitute the values of Y's from the equation $Y = a_1 + b_1x$ we get the deviation as

$$d_1 = y_1 - a_1 - b_1x_1$$

$$d_2 = y_2 - a_1 - b_1x_2$$

$$d_3 = y_3 - a_1 - b_1x_3$$

.

.

.

$$d_n = y_n - a_1 - b_1x_n$$

so the sum of the squares of the deviation is

$$\sum d^2 = \sum (y - a_1 - b_1x)^2$$

By using the maximum and minimum formula of differential calculus we shall get the value of

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and $a_1 = \bar{y} - b_1\bar{x}$.

Substituting the value of a_1 and b_1 we get the equation of y on x is $Y = a_1 + b_1x$.

9.24.a Example : Computation of a regression equation of monthly expenditure on food in taka (y) on monthly total expenditure in taka (x) of 10 person

y	120	110	140	150	180	125	190	160	140	140
x	215	225	225	275	275	290	290	315	350	400

The equation of linear regression of y on x may be written as $Y = a_1 + b_1x$. the value of a_1 and b_1 can be estimated from the data by using the formula given by least square method.

The formula is = $$\frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}$$

Table for calculation of values of a_1 and b_1

y	x	$d_y = y-140$	$d_x = x-275$	$d_x d_y$	$(d_x)^2$
120	215	-20	-60	1200	3600
110	225	-30	-50	1500	2500
140	225	0	-50	0	2500
150	275	10	0	0	0
180	275	40	0	0	0
125	290	-15	15	-225	225
190	290	50	15	750	225
160	315	20	40	800	1600
140	350	0	75	0	5625
140	400	0	125	0	15625
Total		55	110	4025	31900

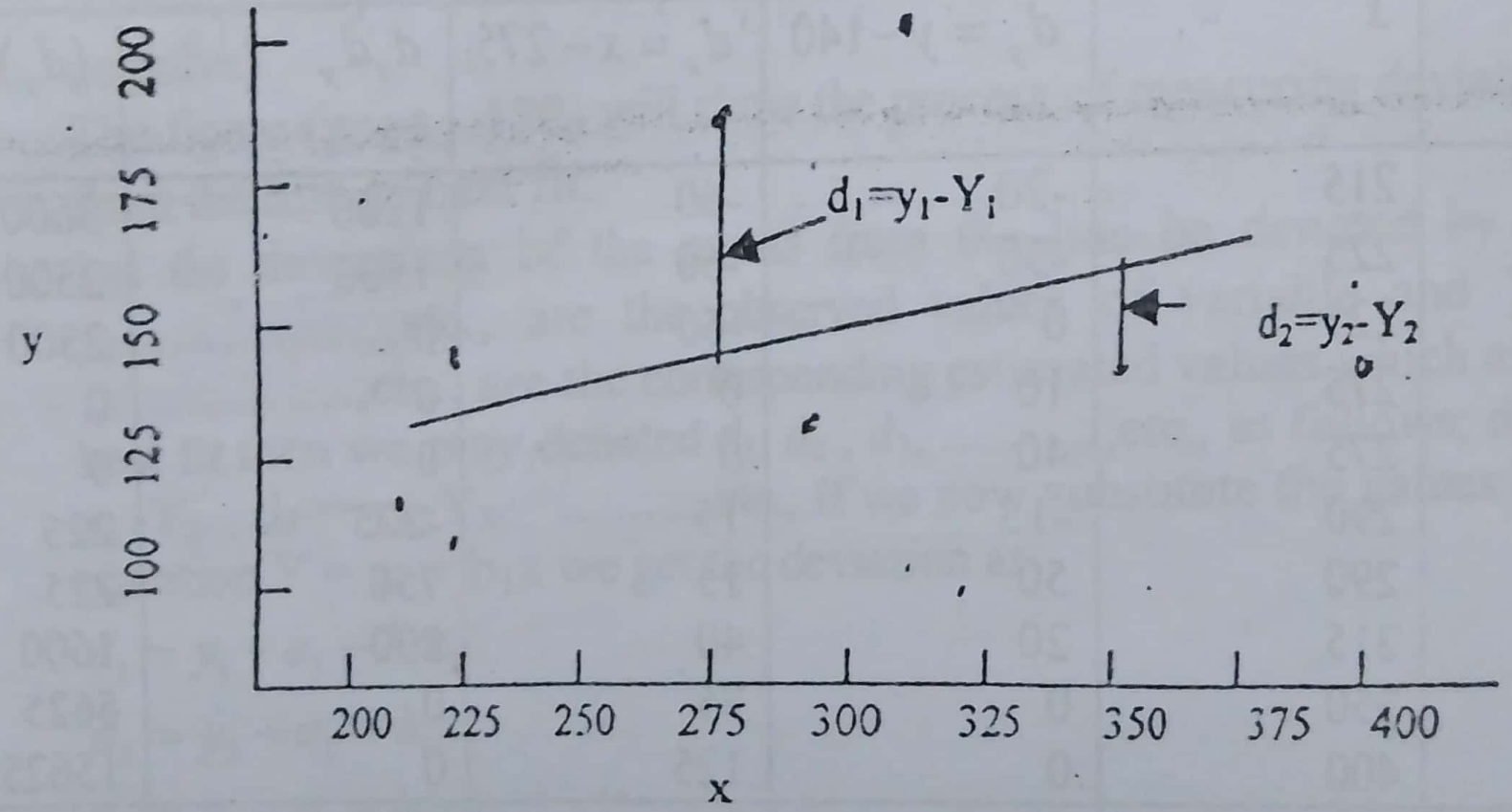
$$b_1 = \frac{4025 - \frac{55\times110}{10}}{31900 - \frac{(110)^2}{10}} = \cdot11, \bar{x} = 275 + \frac{110}{10} = 286, \bar{y} = 140 + \frac{55}{10} = 145\cdot5, a_1 = 145\cdot5 - \cdot11\times286 = 114\cdot$$

Therefor the equation of y on x is $Y = 114\cdot04 + \cdot11x$

Now we are in a position to find out the expected values of y for the observed values of x by applying the equation $Y = a_1 + b_1x$ i.e., $Y = 114\cdot04 + \cdot11x$ in the following figure. The expected values (Y's) are then plotted against the corresponding observed values of x on the scatter diagram in plotting the pairs of values normal rules as applicable to graphs are followed. The plotted points will form a straight line which is the regression line of y on x. This is illustrated in the following figure.

Table showing computation of estimated values of y-variable

X	y	Estimated Y
215	120	137.69
225	110	138.79
225	140	138.79
275	150	144.29
275	180	144.29
290	125	145.94
290	190	145.94
315	160	148.69
350	140	152.54
400	140	158.04
Total	1455	1455.00



Regression line of y on x

9.27.a Problem: Calculate the regression equations of x on y and y on x from the following data. Represent them on the same graph paper . For $x = 50$, estimate the value of y. Also calculate the value of correlation coefficient.

x : 20 22 21 24 22 25 28 27 19 18

y : 18 20 18 21 19 20 23 22 16 17

Solution : Calculation of regression equations

x	Y	x^2	y^2	xy
20	18	400	324	360
22	20	484	400	440
21	18	441	324	378
24	21	576	441	504
22	19	484	361	418
25	20	625	400	500
28	23	784	529	644
27	22	729	484	594
19	16	361	256	304
18	17	324	289	306
$\sum x = 226$	$\sum y = 194$	$\sum x^2 = 5028$	$\sum y^2 = 3808$	$\sum xy = 4448$

Regression equation of x on y is given by $X = a + by$, here $b = \dfrac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2}$

$$b = \frac{10 \times 4448 - 226 \times 194}{10 \times 3808 - (194)^2} = \frac{636}{444} = 1 \cdot 432,$$

$$\bar{x} = \frac{\sum x}{n} = \frac{266}{10} = 22 \cdot 6 \quad \text{and} \quad \bar{x} = \frac{\sum y}{n} = \frac{194}{10} = 19 \cdot 4$$

$$\therefore a = \bar{x} - b\bar{y} = 22 \cdot 6 - 1 \cdot 432 \times 19 \cdot 4 = -5 \cdot 1808$$

Hence the required regression equation of x on y is given by

$X = -5.18 + 1.432\ y$ .....(1)

Again the regression equation of y on x is: $Y = a + bx$

Here $b = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \dfrac{10\times 4448 - 226\times 194}{10\times 5208 - (226)^2} = \dfrac{636}{1004} = 0.633$

$\therefore a = \bar{y} - b\bar{x} = 19.4 - 0.633\times 22.6 = 5\cdot 08$

Hence the required regression equation of y on x is given by

$Y = 5.08 + 0.633x$ .....(2)

Alternate method (b)

$$\text{Cov}(x,y) = \frac{\sum xy}{n} - (\bar{x})(\bar{y}) = \frac{4448}{10} - (22.6)(19.4) = 6.36$$

$$\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x} = \frac{5208}{10} - (22.6)^2 = 10.04$$

$$\sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y} = \frac{3808}{10} - (19.4)^2 = 4.44$$

Regression equation of y on x is

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

Here $\bar{y} = 19.4, \bar{x} = 22.6$ and $b_{yx} = \dfrac{\text{cov}(x,y)}{\sigma_x^2} = \dfrac{6.36}{10.04} = \cdot 633$

$\therefore Y - 19.4 = \cdot 633(x - 22\cdot 6)$

$Y = 5.08 + 0.633x$

Regression equation of x on y is $X - \bar{x} = b_{xy}(y - \bar{y})$ here $b_{xy} = \dfrac{\text{cov}(x,y)}{\sigma_y^2} = \dfrac{6.36}{4.44} = 1.432$

$\Rightarrow X - 22.6 = 1.432(y - 19.4)$

$\Rightarrow X = -5.18 + 1.432y$

Alternative method (a):

When deviation taken from arithmetic means of x and y

X	Y	$X - \bar{X}$ (x)	$Y - \bar{Y}$ (y)	x^2	y^2	xy
20	18	-2.6	-1.4	6.76	1.96	3.64
22	20	-0.6	0.6	0.36	.36	-0.36
21	18	-1.6	-1.4	2.56	1.96	2.24
24	21	1.4	1.6	1.96	2.56	2.24
22	19	-0.6	-0.4	0.36	0.16	0.24
25	20	2.4	0.6	5.76	0.36	1.44
28	23	5.4	3.6	29.16	12.96	19.44
27	22	4.4	2.6	19.36	6.76	11.44
19	16	-3.6	-3.4	12.96	11.56	12.24
18	17	-4.6	-2.4	21.16	5.76	11.04
$\sum X = 226$	$\sum Y = 194$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 1000.4$	$\sum y^2 = 44.4$	$\sum xy = 63.59$

$\bar{x} = \frac{\sum X}{n} = \frac{226}{10} = 22.6$ and $\bar{y} = \frac{\sum Y}{n} = \frac{194}{10} = 19.4$

Regression equation of x on y is $X - \bar{x} = b_{xy}(y - \bar{y})$

$$b_{xy} = \frac{\sum xy}{\sum y^2}$$

$$= \frac{63.59}{44.4} = 1.433$$

$\Rightarrow$ X - 22.6 = 1.43(y – 19.4)

$\therefore$ **X = - 5.18 + 1.432 y**

Regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

$$= \frac{63.59}{100.4} = 0.633$$

$Y - 19.4 = 0.633(x - 22.6)$

$\therefore$ **Y =5.08 + 0.633x**

Alternative method (b)

When deviation taken from Assumed Means

x	y	x – 22 d_x	y- 20 d_y	d_x^2	d_y^2	$d_x d_y$
20	18	-2	-2	4	4	4
22	20	0	0	0	0	0
21	18	-1	-2	1	4	2
24	21	2	1	4	1	2
22	19	0	-1	0	1	0
25	20	3	0	9	0	0
28	23	6	3	36	9	18
27	22	5	2	25	4	10
19	16	-3	-4	9	16	12
18	17	-4	-3	16	9	12
$\sum x$=226	$\sum y = 194$	$\sum d_x$=6	$\sum d_y$ =-6	$\sum d_x^2$ =10 4	$\sum d_y^2$ =4 8	$\sum d_x d_y$ =60

Regression equation of x on y is $X - \bar{x} = b_{xy}(y - \bar{y})$

Here $\bar{x} = \frac{\sum X}{n} = \frac{226}{10} = 22.6$ and $\bar{y} = \frac{\sum Y}{n} = \frac{194}{10} = 19.4$

And

$$b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_y^2 - (\sum d_y)^2} = \frac{10 \times 60 - 6(-6)}{10 \times 48 - (-6)^2} = \frac{636}{444} = 1 \cdot 432$$

$\therefore X - 22.6 = 1.432\,(y - 19.4) \Rightarrow X = -5.18 + 1.432\,y$

Regression equation of y on x is $Y - \bar{y} = b_{yx}(x - \bar{x})$

Here $b_{yx} = \dfrac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_x^2 - \left(\sum d_x\right)^2} = \dfrac{10\times 60 - 6(-6)}{10\times 104 - (6)^2} = \dfrac{636}{1004} = 0.633$

$\therefore \quad Y - 19.4 = 0.633\,(x - 22.6)$

$\Rightarrow \quad \mathbf{Y = 5.08 + 0.633x}$

Regression equation of y on x is $Y = 5.08 + 0.633x$ When $x = 50$ then $Y = 5.08 + 0.633(50)$

$\therefore Y = 36.73$

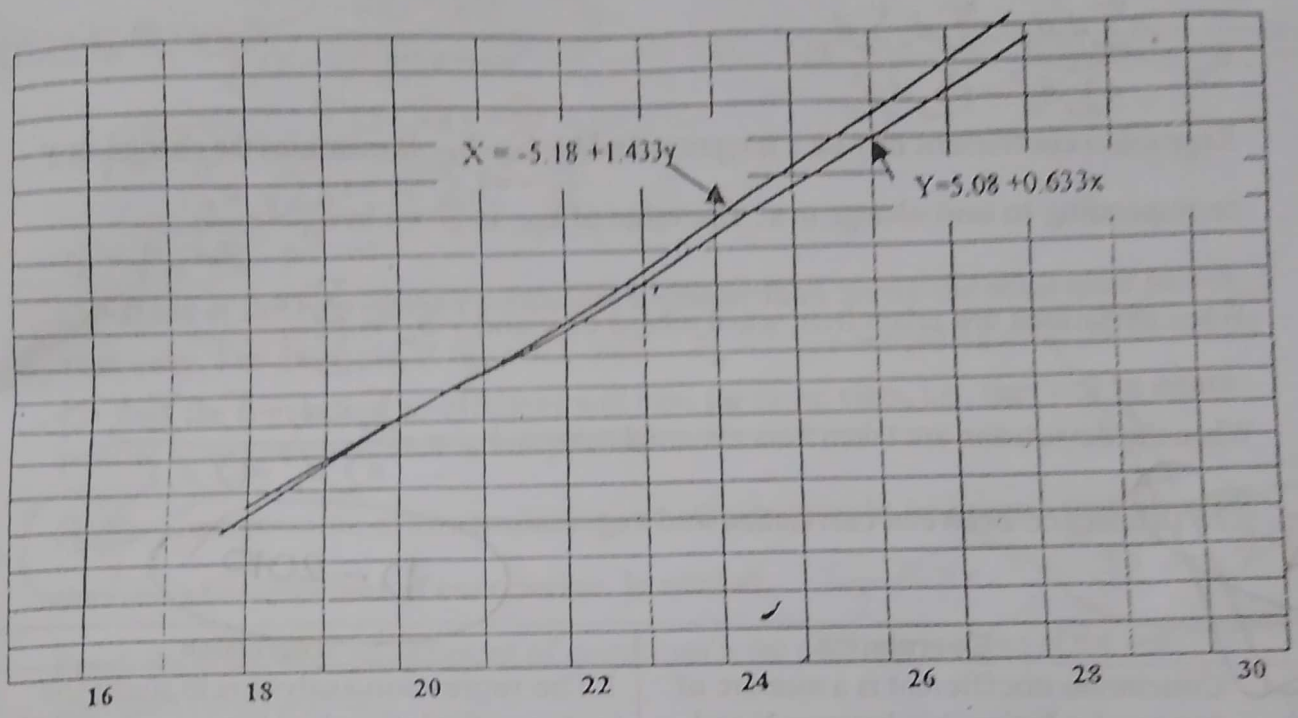
Graphing regression equation

From the regression equation of y on x we can estimate the most probable value of y for various values of x and from the regression equation of x on y we can estimate the most probable values of x for various values of y

	(Estimated values of Y) Y=5.08 + 0.633x
x=20	Y =5.08 +0.633(20) = 17.74
x=22	Y =5.08 +0.633(22) = 19.00
x=21	Y =5.08 +0.633(21) = 18.37
x=24	Y =5.08 +0.633(24) = 20.27
x=22	Y =5.08 +0.633(22) = 19.00
x=25	Y =5.08 +0.633(25) = 20.91
x=28	Y =5.08 +0.633(28) = 22.80
x=27	Y =5.08 +0.633(27) = 22.17
x=19	Y =5.08 +0.633(19) = 17.10
x=18	Y =5.08 +0.633(18) = 16.47

To plot the regression line of y on x , we will take the actual values of x and estimated values of y

	(Estimated values of X) X= -5.18 + 1.432y
y = 18	X = -5.18 +1.432 (18) = 20.596
y = 20	X = -5.18 +1.432 (20) = 23.46
y = 18	X = -5.18 +1.432 (18) = 20.596
y = 21	X = -5.18 +1.432 (21) = 24.892
y = 19	X = -5.18 +1.432 (19) = 22.028
y = 20	X = -5.18 +1.432 (20) = 23.46
y = 23	X = -5.18 +1.432 (23) = 27.75
y = 22	X = -5.18 +1.432 (22) = 26.32
y = 16	X = -5.18 +1.432 (16) = 17.73
y = 17	X = -5.18 +1.432 (17) = 19.16



Coefficient of correlation:

We are given $b_{xy} = 1.432$, $b_{yx} = 0.633$, Now

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{1.432 \times 0.633} = \sqrt{0.9064} = 0.95$$

9.28 Regression coefficients

The quantity b in the regression equations is called the "regression coefficient" or "slope coefficient". Since there are two regression equations therefore, there are two regression coefficients-regression coefficient of x on y and regression coefficient of y on x.

Regression coefficient of x on y

The regression coefficient of x on y is represented by the symbol b_{xy} or b_1. It means the change in x corresponding to a unit change in y. The regression coefficient of x on y is given by $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

When deviation are taken from the means of x and y, the regression coefficient is obtained by $b_{xy} = \frac{\sum xy}{\sum y^2}$

When deviations are taken from assumed means, the value of b_{xy} or b_1 is obtain as follows:

$$b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_y^2 - (\sum d_y)^2}$$

Regression coefficient of y on x is represented by b_{xy} b_1. It measures the change in y corresponding to unit change in x. The value of b_{yx} is given by $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

When deviations are taken from actual means of x and y $b_{yx} = \frac{\sum xy}{\sum x^2}$

When the deviations are taken from assumed means $b_{xy} = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{n\sum d_x^2 - (\sum d_x)^2}$

9.28 Difference between correlation and regression:

D - 2010

Correlation	Regression
1. Correlation coefficient is a measure of degree of relationship between X and Y.	1. The regression analysis is to study the 'nature of relationship' between X and Y.
2. The cause and effect relation is not clearly indicated through correlation co-efficient.	2. the cause and effect relation is clearly indicated Through regression analysis.
3. In correlation co -efficient there are no concept about independent and dependent variables.	3. In regression analysis, we find the clear concept about independent and dependent variable.
4. For X and Y variables $r_{xy} = r_{yx}$	4. For X and Y variables, $b_{xy} \neq b_{yx}$.
5. The co efficient of correlation lies between -1 and $+1$. Symbolically $-1 \le r \le 1$	5. The regression coefficient lies between $-\infty$ to ∞. Symbolically $-\infty < b_{xy} < \infty$ or $-\infty < b_{yx} < \infty$
6. The coefficient of correlation is independent of change of origin and scale.	6. The regression coefficient is independent of origin but dependent on scale

D-09

9.29 Properties of regression coefficients

(1) The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically $r = \sqrt{b_{xy} \times b_{yx}}$

Proof : We know that the coefficient of regression of x on y is $b_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2}$

Again the co-efficient of regression of y on x is $b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

$$\therefore b_{yx} \times b_{xy} = \frac{\{\sum(x-\bar{x})(y-\bar{y})\}^2}{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}$$

$$\text{or } \sqrt{b_{xy} \times b_{yx}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} = r_{xy}$$

$r_{xy} = \sqrt{b_{xy} \times b_{yx}}$ proved

(2) If one of the regression co-efficient is greater than unity the other must be less than unity. For example, if $b_{yx} = 1.2$ and $b_{xy} = 1.4$, r would be $= \sqrt{1.2 \times 1.4} = 1.29$

(3) Both the regression co-efficient will have the same sign, i.e., they will be either positive or negative . which is not possible.

(4) The average value of two regression co-efficient would be greater than or equal the value of co-efficient of correlation. In symbols $\frac{b_{xy} + b_{yx}}{2} \geq r$.

D — 2009

Proof: We know that co-efficient of correlation is the geometric mean of the two regression co-efficients

Such that $r_{xy} = \sqrt{b_{xy} \times b_{yx}}$

Now the arithmetic means of b_{yx} and b_{xy} is $\frac{b_{xy} + b_{yx}}{2}$ and G.M is $\sqrt{b_{xy} \times b_{yx}}$

We know that $A.M \geq G.M$

$$\frac{b_{xy} + b_{yx}}{2} \geq \sqrt{b_{xy} \times b_{yx}}$$

$$\therefore \frac{b_{xy} + b_{yx}}{2} \geq r \text{ proved}$$

(5) Regression co-efficient are independent of change of origin but not scale.

Proof: We know the regression co-efficient of y on x is

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \quad \ldots\ldots\ldots\ldots (I)$$

Let us now change the origin and scales. Deduct a fixed quantity 'a' from x and 'b' from y. Also divide x and y series by a fixed value 'h' and 'k'.
Let the new values be denoted by u and v.

$$u = \frac{x-a}{h}$$

$$x = a + hu$$

$$\bar{x} = h\bar{u} + a$$

$$v = \frac{y-b}{k}$$

$$y = kv + b$$

$$\bar{y} = k\bar{v} + b$$

Putting the value of x, $\bar{x}$ and y, $\bar{y}$ in equation (1) we get

$$b_{yx}=\frac{\sum(a+hu-a-h\bar{u})(kv+b-k\bar{v}-b)}{\sum(a+uh-a-h\bar{u})^2}$$

$$b_{yx}=\frac{hk\sum(u-\bar{u})(v-\bar{v})}{h^2\sum(u-\bar{u})^2}=\frac{k}{h}b_{uv}$$

Proceeding in the above way we get $b_{xy}=\frac{h}{k}b_{uv}$.

Which shows that the regression co-efficient are independent on change of origin but not of scale.

9.30 Problem: Given that $\bar{x}=20$, $\bar{y}=15$, $\sigma_x=4, \sigma_y=3, r=\cdot 7$. Determine the two regression equations. When x=24 find the value of y.

Solution: We know that regression equation y on x is y = a+bx, Here a = $\bar{y}-b\bar{x}$ and

$$b=\frac{\text{cov}(x,y)}{\sigma_x^2}$$

Again we know that $r=\frac{\text{cov}(x,y)}{\sigma_x\,\sigma_y}$

or $\text{cov}(x,y)=r\sigma_x\sigma_y=(\cdot 7)\times(4)(3)=8\cdot 4 \quad \therefore b=\frac{8\cdot 4}{4^2}=\cdot 525$ and

$a=15-(\cdot 525\times 20)=4\cdot 5$

So the regression equation of y on x is $y=4\cdot 5+\cdot 525x$

Again the regression equation of x on y is $x=a_1+b_1y$ here $b_1=\frac{\text{cov}(x,y)}{\sigma_y^2}=\frac{8\cdot 4}{3^2}=\cdot 93$

and $a_1=\bar{x}-b_1\bar{y}=20-\cdot 93\times 15=6\cdot 05$

Therefor the regression equation of x on y is $x=6\cdot 05+\cdot 93y$. When x is 24 then the estimated value of y is $y=4\cdot 5+\cdot 525\times 24=17\cdot 1$

9.31 Problem: Given that $\bar{x}=20$, $\bar{y}=15$, $\sigma_x=4, \sigma_y=3, r=\cdot 7$. Then show that $r=\sqrt{b_1b_2}$

Solution: We know that regression equation y on x is $y=a_1+b_1x$,

Here $a_1=\bar{y}-b_1\bar{x}$ and $b_1=\frac{\text{cov}(x,y)}{\sigma_x^2}$.

Again we know that $r=\frac{\text{cov}(x,y)}{\sigma_x\,\sigma_y}$ or $\text{cov}(x,y)=r\sigma_x\sigma_y=(\cdot 7)\times(4)(3)=8\cdot 4$

$\therefore b_1=\frac{8\cdot 4}{4^2}=\cdot 525$ and $b_2=\frac{\text{cov}(x,y)}{\sigma_y^2}=\frac{8\cdot 4}{3^2}=\cdot 93 \quad \therefore b_1.b_2=\cdot 525\times\cdot 93 \quad \therefore\sqrt{b_1.b_2}=0.7$

$\therefore r=\sqrt{b_1b_2}$ showed

Mathematical problems

17. If the relation between x and y is denoted by the equation $y = mx + c$, find the coefficient of correlation between them and comment.

18. If $y = -\frac{x}{2}$, find r_{xy} and comment.

19. If $y = 3x + 2$, find the correlation coefficient and comment.

20. If x and y are two independent random variables find the correlation coefficient between x+y and x-y.

21. Given that

$\sum xy = 200, \bar{x} = 5, \bar{y} = 4, \sigma_x = \sqrt{10}, \sigma_y = 3, u = 3x + 4y$ *and* $v = 3x - y$ find r_{uv}, when bivariate (x,y) has 8 pairs of values.

22. If $y - 4x - 2 = 0$, then find the coefficient of correlation between x and y.

23. If $r_{xy} = 0.5$, then find the coefficient of correlation between $a + 3x$ and $b + 5y$

24. Find the regression line of x on y and that of y on x. Represent them on the same graph paper. For $x = 85$, estimate the value of y.

X	63	71	70	65	63	65	68	67	67	66
y	67	70	68	66	66	68	69	68	67	65