UNIT-IV

CORRELATION

Concept of correlation

In a bivariate distribution, If the change in one variable affects a change in the other variable, the variables are said to be correlated.

The relationship between two variables such that a change in one variable given in a positive or negative change in the other variable is known as correlation.

If the two variable deviate in the same direction (ie) if the increase or decrease in the one variable results in a corresponding increase or decrease in the other correlation is said to be direct or positive.

But if they constantly deviate in the opposite direction (ie) if increase or decrease in one gives a corresponding decrease or increase in the other variable correlation is said to be inverse or negative.

Eg. For positive correlation

(i) heights and weights of a group of persons.

(ii) The price and supply of a commodity.

Eg. For negative correlation
(i) Price and demand of a commodity
(ii) The volume and pressure of a perfect gas.

correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Methods of correlation:-

For the bivariate distribution (x_i, y_i) $(i = 1, 2, \ldots n)$ if the values of the variables X and Y be plotted along the x axis and Y axis the diagram of dots obtained is scatter diagram. we can get a fairly good idea whether the variables are correlated and or not if the points are very close to other. we expect a fairly good amount of correlation between the variables if the points are widely scattered a poor correlation is expetes.

I Karl pearson's coefficient of correlation

correlation co-efficient between two random variables x and y is denoted by $\gamma(X,Y)$ or γ_{xy} and defined as

$$\gamma_{xy} = \frac{cov(x,y)}{\sigma x \, \sigma y} \rightarrow (1)$$

For deviation method:

$$cov(x,y) = \frac{1}{n} \Sigma (xi - \bar{x})(yi - \bar{y})$$

$$\sigma x^2 = \frac{1}{n} \Sigma (xi - \bar{x})^2$$

$$\sigma y^2 = \frac{1}{n} \Sigma (yi - \bar{y})^2$$

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

$\bar{x} = \frac{\Sigma xi}{n}$

$n\bar{x} = \Sigma xi$

$n\bar{y} = \Sigma yi$

$$cov(x,y) = \frac{1}{n} \Sigma (xi - \bar{x})(yi - \bar{y})$$

$$= \frac{1}{n} \Sigma [xi\,yi - xi\bar{y} - \bar{x}\,yi + \bar{x}\bar{y}]$$

$$= \frac{1}{n} \Sigma xi\,yi - \frac{1}{n} \Sigma xi\bar{y} - \bar{x}\frac{1}{n}\Sigma yi + \frac{n\bar{x}}{n}\bar{y}$$

$$= \frac{1}{n} \Sigma x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \frac{1}{n} \Sigma x_i y_i - \bar{x}\bar{y}$$

$\frac{1}{n}\Sigma(\bar{x}) = \frac{n\bar{x}}{n} = \bar{x}$

$$\sigma x^2 = \frac{1}{n} \Sigma (xi - \bar{x})^2$$

$$= \frac{1}{n} \Sigma (xi - \bar{x})(xi - \bar{x})$$

$$= \frac{\Sigma xi^2}{n} - 2\frac{\Sigma xi}{n}\bar{x} + (\bar{x})^2$$

$$= \frac{1}{n} \Sigma [xi^2 - xi\bar{x} - \bar{x}\,xi + (\bar{x})^2]$$

$$= \frac{1}{n} \Sigma x_i^2 - 2\bar{x}\bar{x} + (\bar{x})^2$$

$$= \frac{\Sigma xi^2}{n} - (\bar{x})^2 = \frac{1}{n} \Sigma xi^2 - 2\bar{x}\frac{\Sigma xi}{n} + \bar{x}^2$$

$$= \frac{\Sigma xi^2}{n} - (\bar{x})^2 = \frac{1}{n} \Sigma x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2$$

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

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

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

$$r_{xy} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - (\bar{x})^2\right)\left(\frac{1}{n}\sum y_i^2 - (\bar{y})^2\right)}}$$

correlation co-efficient cannot exceed unity. It always lies between +1 and -1. If $r=+1$ correlation is perfect and positive. If $r=-1$ correlation is perfect and negative

If the variable are independent $r=0$, but the converse is not true.

Theorem:-

correlation coefficient is independent of change of origin and scale.

prove:- $\frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{Cov(u,v)}{\sigma_u \sigma_v}$.

Proof:

Let $u = \frac{x-a}{h}$ & $v = \frac{y-b}{k}$

$\therefore x = a + hu$; $y = b + kv$

$r_{xy} = r_{uv}$

we have to prove that $u_{xy} = v_{uv}$

$x = a + hu$ $\quad$ $y = b + kv$

Taking Expectation; $E(x) = a + hE(u)$ $\quad$ $E(y) = b + kE(v)$

$x - E(x) = (a + u.h) - (a + h.E(u))$ $\quad$ $y - E(y) = (b + kv) - (b + k.E(v))$

$= u.h - h.E(u)$ $\quad$ $= kv - k.E(v)$

$= h(u - E(u))$ $\quad$ $= k(v - E(v))$

$x - E(x) = h\{U - E(U)\}$; $y - E(y) = K\{V - E(V)\}$

$\text{cov}(x, y) = E[\{x - E(x)\}\{y - E(y)\}]$

$= E[h\{U - E(U)\} K\{V - E(V)\}]$

$= hK \cdot E[\{U - E(U)\} \cdot \{V - E(V)\}]$

$= hK \cdot \text{cov}(U, V)$

$\sigma x^2 = E[x - E(x)]^2$

$\sigma y^2 = E[y - E(y)]^2$

$\sigma x^2 = E(h\{U - E(U)\})^2$

$= E(h^2\{U - E(U)\}^2$

$= h^2 E[U - E(U)]^2$

$\sigma x^2 = h^2 \sigma U^2$

$\sigma x = h \sigma U$

$\sigma y^2 = E[y - E(y)]^2$

$= E(K\{V - E(V)\})^2$

$= E(K^2\{V - E(V)\}^2]$

$= K^2 \cdot E[V - E(V)]^2$

$\sigma y^2 = K^2 \cdot \sigma V^2$

$\sigma y = K \sigma V$

correlation coefficient between x and y is

$\gamma xy = \frac{\text{cov}(x, y)}{\sigma x \, \sigma y}$

$$= \frac{kx \cdot cov(u,v)}{k \cdot \sigma_u \cdot k \cdot \sigma_v}$$

correlation.

$\gamma_{xy} = \gamma_{uv}$.

Regression

$\therefore \gamma_{xy} = \gamma_{uv}$

$$= \frac{cov(u,v)}{\sigma_u \cdot \sigma_v}$$

$\therefore \gamma_{xy} = \gamma_{uv}$

Theorem

Two independent variable are uncorrelated

$r \leq 0$.

$$Cov(x,y) = E[\{x - E(x)\}\{y - E(y)\}]$$
$$= E[xy - x \cdot E(y) - E(x) \cdot y + E(y)E(y)]$$
$$= E(xy) - E(x) \cdot E(y) - E(x)E(y) + E(x)E(y)$$
$$= E(xy) - E(x) \cdot E(y)$$

If x and y are independent,

$$E(xy) = E(x) \cdot E(y)$$

← For independent we can write like this

$$\therefore Cov(x,y) = E(x)E(y) - E(x) \cdot E(y) = 0$$

Hence

$$\gamma_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.$$

(i.e) the two independent variables are uncorrelated. but the converse is not true.

Two independent variables are uncorrelated

(i.e) Two uncorrelated variables may not be independent.

Example :- 1

a) If $Z = aX + bY$ and r is the correlation coefficient between X and Y. Show that

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,r.\sigma_X\sigma_Y$$

b) Show that $r = \sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2 / 2\sigma_X\sigma_Y$ where $\sigma_X.\sigma_Y$ and σ_{X-Y} are the standard derivations of X, Y and $X-Y$.

Soln:-

$Z = aX + bY$ — change into variable → $[Z - E(Z)]$

$E(Z) = aE(X) + bE(Y)$

$Z - E(Z) = a\{X - E(X)\} + b\{Y - E(Y)\}$

$\sigma_Z^2 = \text{Var}(Z) = E(Z - E(Z))^2$

$= E[a\{X - E(X)\} + b\{Y - E(Y)\}]^2$

$= E[a^2\{X - E(X)\}^2 + b^2\{Y - E(Y)\}^2 + 2ab\{X - E(X)\}\{Y - E(Y)\}]$

$\sigma_X^2 = E(X - E(X))^2$

$\sigma_Y^2 = E(Y - E(Y))^2 = a^2E[(X - E(X))]^2 + b^2E[(Y - E(Y))]^2 + 2abE[\{X - E(X)\}\{Y - E(Y)\}]$

$r = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$

$\text{Cov}(X,Y) = r.\sigma_X\sigma_Y$

$= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\text{Cov}(X,Y)$

$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,r\sigma_X\sigma_Y$

Putting $a = 1, b = -1$ we get $Z = X - Y$

$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$

$2r\sigma_X\sigma_Y = \sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2$

$r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$

$\text{Var}(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\,\text{Cov}(X,Y)$

$$Var(aX-bY) = a^2V(X) + b^2V(Y) - 2ab\,Cov(X,Y)$$

Probable error of correlation coefficient

(If 'r' is the correlation coefficient then its standard error (S.E) $= \frac{1-r^2}{\sqrt{n}}$

Probable error (P.E) $= 0.6745 \times S.E$

$= 0.6745 \times \left(\frac{1-r^2}{\sqrt{n}}\right)$)

Probable error is helpful in testing the reliability of an observed correlation coefficient.

If $r > P.E$ correlation is not at all significant.

If $r > 6\,P.E$ it is definitely significant probable error also helps us to find the limits with in which the correlation co-efficient can be expected to vary the limits are $r \pm P.E$

Limits for correlation coefficient:-

$$r_{xy} = \frac{Cov(X,Y)}{\sigma x\, \sigma y} = \frac{1/n \sum (x_i - \bar{x})(y_i - \bar{y})}{\left\{\frac{1}{n}\sum (x_i - \bar{x})^2 \frac{1}{n}(y_i - \bar{y})^2\right\}^{1/2}}$$

Squaring on both sides,

$$r_{xy}^2 = \frac{1/n^2 [\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{1/n \sum (x_i - \bar{x})^2 \frac{1}{n} \sum (y_i - \bar{y})^2}$$

$= (\sum_{i=1}^{n} a_i . b_i)^2$

$\sum a_i^2 \sum b_i^2$

where $a_i = (x_i - \bar{x})$; $b_i = (y_i - \bar{y})$

Schwarz inequality states that, if a_i, b_i $(i = 1, 2 \ldots n)$ are real quantities then,

$(\sum a_i b_i)^2 \leq (\sum a_i^2)(\sum b_i)^2$

the sign of equality holding if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \frac{a_n}{b_n}$

using this we get $r_{xy}^2 \leq 1 = r^2 \leq 1$

$|r_{xy}| \leq 1$

$-1 \leq r_{xy} \leq 1$

(ie) correlation co-efficient cannot exceed unity. Hence, correlation coefficient lies between -1 and +1.

Rank correlation

Let (x_i, y_i) $i = 1, 2 \ldots n$ be the ranks of the i^{th} individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i, y_i is called the rank correlation.

Assume that no two students get the same rank in either classification.

X & Y takes the values 1, 2 ... n.

$$\therefore \bar{X} = \bar{Y} = \frac{1}{n}(1+2+3 \ldots +n)$$

$$= \frac{1}{n}\left[\frac{n(n+1)}{2}\right]$$

$$= \frac{n+1}{2}$$

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

$$= \frac{1}{n}\left[n\left[\frac{(n+1)(2n+1)}{6}\right] - \frac{(n+1)^2}{4}\right]$$

$$= \frac{(n+1)}{12}[2(2n+1) - 3(n+1)]$$

$$= \frac{n+1}{12}[4n+2-3n-3]$$

$$= \frac{n+1}{12}[n-1]$$

$$= \frac{n^2-1}{12}$$

$\because \bar{x} = \bar{y}, \sigma_x^2 = \sigma_y^2$

$$\therefore \sigma_x^2 = \sigma_y^2 = \frac{n^2-1}{12}$$

$x_i \neq y_i$

Let $d_i = x_i - y_i$

$= (x_i - \bar{x}) - (y_i - \bar{y})$

$= x_i - \bar{x} + \bar{x} - \bar{y} + \bar{y} - y_i$

$$\Sigma d_i^2 = \Sigma \{(x_i - \bar{x})(y_i - \bar{y})\}^2$$

$$= \Sigma (x_i - \bar{x})^2 + \Sigma (y_i - \bar{y})^2 - 2\Sigma (x_i - \bar{x})(y_i - \bar{y}).$$

Dividing by 'n'

$$\frac{1}{n}\Sigma d_i^2 = \frac{1}{n}\Sigma (x_i - \bar{x})^2 + \frac{1}{n}\Sigma (y_i - \bar{y})^2 - \frac{2}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sigma x^2 + \sigma y^2 - 2 \text{ cov}(X, Y)$$

$$\frac{1}{n}\Sigma d_i^2 = 2\sigma x^2 - 2r\sigma x \sigma y$$

$\therefore \sigma x = \sigma y$

$$= 2\sigma x^2 - 2r \cdot \sigma x^2$$

$$\frac{1}{n}\Sigma d_i^2 = 2\sigma x^2 (1 - r)$$

$$\Sigma d_i^2 = n \cdot 2\sigma x^2 (1 - r)$$

$$\frac{\Sigma d_i^2}{2\sigma x^2 n} = 1 - r$$

$$r = 1 - \frac{\Sigma d_i^2}{2n(\sigma x^2)}$$

$$= 1 - \frac{\Sigma d_i^2}{2n\left(\frac{n^2 - 1}{12}\right)}$$

$$= 1 - \frac{\Sigma d_i^2}{n^2\left(\frac{n^2-1}{6}\right)}$$

$$r = 1 - \left\{\frac{6\Sigma d_i^2}{n(n^2-1)}\right\}$$

This is known as Spearman's formula for the rank correlation co-efficient.

Note:-

For Rank correlation Let us use the symbol ρ for r

$$\therefore \rho = 1 - \left\{\frac{6\Sigma d_i^2}{n(n^2-1)}\right\}$$

Repeated Rank:-

If any two or more students get the same rank then the spearman's formula for calculating the rank correlation co-efficient breaks down.

In this case common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed and the next items would have assumed if they were slightly different from each other and the next item will get the rank next to the ranks already assumed. As a

result of this we add the correlation factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is the number of times an item has repeated value. This correlation factor is to be added for each repeated value.

$$\therefore \rho = 1 - \left\{ \frac{6\{\Sigma d_i^2 + \text{correlation factor}\}}{n(n^2-1)} \right\}$$

Limits for rank correlation co-efficient

Spearman's rank correlation coefficient is

$$\rho = 1 - \left[\frac{6\Sigma d_i^2}{n(n^2-1)} \right]$$

If

$x_i = y_i$ then $d_i = 0 \therefore \rho = +1$

(ie) If the ranks are equal then the maximum value of ρ is ± 1

If the ranks are in the opposite direction the $\rho = -1$

(ie) x : 1 2 3 . . . (n-1) n

y : n n-1 n-2 . . . . . 2 1

$$\Sigma d^2 = (1-n)^2 + \{2-(n-1)\}^2 + \{3-(n-2)\}^2 + \ldots \{(n-1)-2\}^2 + (n-1)^2$$

$$= (1-n)^2 + (3-n)^2 + (5-n)^2 + \ldots + (n-3)^2 + (n-1)^2$$

$= (1-n)^n + (n-3)^2 + \ldots$ up to $\frac{1}{6}n^3$ terms.

$$\Sigma d^2 = \sum_{r=1}^{n} [n-(2r-1)]^2$$

$$= \sum_{r=1}^{n} [(n+1) - 2r]^2$$

$$= \sum_{r=1}^{n} [(n+1)^2 - 4r(n+1) + 4r^2]$$

$$= n(n+1)^2 - 4(n+1)\Sigma r + 4\Sigma r^2$$

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

$$= n(n+1)^2 - 2n(n+1)(n+1) + \frac{2}{3}n(n+1)(2n+1)$$

$$\Sigma d^2 = n(n+1)\left[(n+1) - 2(n+1) + \frac{2(2n+1)}{3}\right]$$

$$= n(n+1)\left[n+1-2n-2+\frac{4n+2}{3}\right]$$

$$\Sigma d^2 = \frac{n(n+1)}{3}\left[3n+3-6n-6+4n+2\right]$$

$$\Sigma d^2 = \frac{n(n+1)}{6}\{n-1\}$$

$$= \frac{n(n^2-1)}{3}$$

$$P = 1 - \left\{\frac{6\Sigma d^2}{n(n^2-1)}\right\}$$

$$= 1 - \frac{6n(n^2-1)/3}{n(n^2-1)}$$

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

The limits for rank correlation coefficient is $-1 \leq P \leq 1$.

REGRESSION

pefinition :-

Regression analysis is a mathematical measure of the average relationship between two (or) more variables in terms of the original best fits of the data.

equation :-

Regression means stepping back towards average :-

The line of regression is the line of "best fit" and is obtained by the principle of last squares.

Let the regression line of y on x is

$Y = a + bx_i$. x is independent variable Y is dependent variable.

According to the principle of least squares the constants a and b are to be determined such that

$E = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ is minimum

$\frac{\partial E}{\partial a} = 2\Sigma(y - a - bx_i)(-1)$

$\frac{\partial E}{\partial a} = 0 \Rightarrow -2\Sigma(y_i - a - bx_i) = 0$.

$\Sigma y_i = na - b\Sigma x_i = 0$

$\Sigma y_i = na + b\Sigma x_i \rightarrow (1)$

$\frac{\partial E}{\partial b} = 0$

$$-2\sum x_i (y_i - a - bx_i) = 0$$

$$\sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad \rightarrow (2)$$

These two are the normal equations for finding a and b. dividing (1) by n we get $\bar{y} = a + b\bar{x}$ $\rightarrow$ (3)

(ie) the line of regression of y on x passes through $(\bar{x}, \bar{y})$

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

$$\frac{1}{n}\sum x_i y_i = cov(xy) + (\bar{x})(\bar{y}) \rightarrow (4)$$

Also,

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

$$\frac{1}{n}\sum x_i^2 = \sigma_x^2 + (\bar{x})^2$$

Dividing (2) by n $\cdot \frac{\sum x_i y_i}{n} = a\frac{\sum x_i}{n} + \frac{b\sum x_i^2}{n}$

Using (4) & 4:a we get

$$cov(x, y) + \bar{x}\bar{y} = a\bar{x} + b[\sigma_x^2 + (\bar{x})^2]$$

Multiplying (3) by $\bar{x}$

$$\bar{x}\bar{y} = a\bar{y} + b(\bar{x})^2$$

Subtracting

$$cov(x, y) = b \cdot \sigma_x^2$$

$$b_{yx} = \frac{cov(x,y)}{\sigma x^2}$$

b is the slope, $(\bar{x}, \bar{y})$ is one points.

∴ The equation is $y - \bar{y} = b(x - \bar{x})$

$$y - \bar{y} = \frac{cov(x,y)}{\sigma x^2}(x - \bar{x})$$

$$y - \bar{y} = \frac{r \cdot \sigma x \sigma y}{\sigma x^2}(x - \bar{x}) \qquad r = \frac{cov(x,y)}{\sigma x \sigma y}$$

$$y - \bar{y} = r \cdot \frac{\sigma y}{\sigma x}(x - \bar{x})$$

Starting with the equation $x = a + by$ and proceeding similarly we get the line of regression of x on y as:

$$x - \bar{x} = r \cdot \frac{\sigma x}{\sigma y}(y - \bar{y})$$

Always we have two lines of regression one is y on x and the other is x on y.

Line of regression of Y on x is used to find y for any given x regression line of x on y is used to find x for a given y.

where $r = \pm 1$, both the lines reduce to

$$\frac{y - \bar{y}}{\sigma y} = \pm \frac{x - \bar{x}}{\sigma x}$$

(ie) the two lines of regression coincide and we have only one lin

Regression coefficients

The slope of the line 'b' is known as regression co-efficient.

Regression co-efficient of y on x is

$$b_{yx} = \frac{cov(x,y)}{\sigma_y^2}$$

$$= r \cdot \frac{\sigma_y}{\sigma_x}$$

Regression coefficient of x on y is

$$b_{xy} = \frac{cov(x,y)}{\sigma_y^2}$$

$$= r \cdot \frac{\sigma_x}{\sigma_y}$$

Properties of regression coefficients

1. Correlation co-efficients is the geometric mean between the regression co-efficients.

$$b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y} \text{ and } b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

$$b_{xy} \times b_{yx} = \left\{ r \cdot \frac{\sigma_x}{\sigma_y} \right\} \left\{ r \cdot \frac{\sigma_y}{\sigma_x} \right\} = r^2$$

$$\therefore r = \pm \sqrt{b_{xy} \cdot b_{yx}}$$

If the regression coefficients are positive correlation coefficients is positive.
If the regression coefficients are negative

$|r|$ is negative.

2. If one of the regression co-efficient is greater than unity the other must be less than unity

(ie) If by $x > 1$ then $b_{xy} < 1$

Let by $x > 1$

$$\frac{1}{b_{yx}} < 1$$

$$r^2 \leq 1 \Rightarrow b_{yx} \times b_{xy} \leq 1$$

Dividing by (b_{yx})

$$b_{xy} \leq \frac{1}{b_{yx}} < 1$$

(ie) $b_{yx} > 1 \Rightarrow b_{xy} < 1$

3. Arithmetic mean of the regression coefficient is greater than the correlation coefficient we have to prove that

$$\frac{1}{2}(b_{yx} + b_{xy}) \geq r$$

$$\Rightarrow b_{yx} + b_{xy} \geq 2r$$

$$\Rightarrow b_{yx} + b_{xy} \geq 2(\pm\sqrt{b_{xy} + b_{yx}})$$

$$b_{yx} + b_{xy} \pm 2\sqrt{b_{xy} b_{yx}} \geq 0$$

$$\left[\sqrt{b_{yx}} \pm \sqrt{b_{xy}}\right]^2 \geq 0$$ which is always so since the square of a real quantity is always non negative

4. Regression coefficients are independent of the origin but not scale.

Let $U = \frac{x-a}{h}$ and $V = \frac{y-b}{k}$

$x = a + hU$ and $y = b + kV$

a, b, h, k are constant

$E(x) = a + hE(U)$ and $E(y) = b + kE(V)$

$x - E(x) = h\{U - E(U)\}$ and $y = E(y) = k\{V - E(V)\}$

$\sigma x^2 = E(x - E(x))^2$ and $\sigma y^2 = E[y - E(y)]^2$

$= E[h\{U - E(U)\}]^2 = E[k\{V - E(V)\}]^2$

$= h^2 E[U - E(U)]^2 = k^2 - E\{V - E(V)\}^2$

$= h^2 . \sigma V^2$

$cov(x, y) = E[\{x - E(x)\}\{y - E(y)\}]$

$= E[h\{U - E(U)\} k\{V - E(V)\}]$

$= E[hk\{U - E(U)\}\{V - E(V)\}]$

$= hk . E[\{U - E(U)\}\{V - E(V)\}]$

$= hk . cov(U, V)$

$b_{yx} = \frac{cov(x, y)}{\sigma y^2} = \frac{hk\, cov(U, V)}{k^2 \sigma V^2}$

$= \frac{h}{k} \frac{cov(U, V)}{\sigma V^2} = \frac{h}{k} b_{UV}$

Angles between two lines of regression

Regression line of y on x is

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$ slope of this line is $r \cdot \frac{\sigma_y}{\sigma_x}$

Regression line of x on y is

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

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

Slope is $\frac{\sigma_y}{r \cdot \sigma_x}$

If θ is the acute angle between the two regression lines, then

$$\tan\theta = \frac{r \cdot \frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r \cdot \sigma_x}}{1 + r\frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_y}{r \cdot \sigma_x}}$$

$$= \frac{r^2\sigma_y\sigma_x - \sigma_x\sigma_y}{r\sigma_x^2} \Big/ \frac{1 + \frac{\sigma_y^2}{\sigma_x^2}}{}$$

$$= \frac{\sigma_x\sigma_y}{r\sigma_x^2}(r^2 - 1) \Big/ \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}$$

$$= \frac{\sigma_x\sigma_y(1 - r^2)}{r} \Big/ \sigma_x^2 + \sigma_y^2$$

$$= \frac{1 - r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

$$\therefore \theta = \tan^{-1}\left[\frac{1-r^2}{r}\left\{\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right\}\right].$$

If $r=0$, $\tan\theta = \infty \Rightarrow \theta = \pi/2$

(i.e). If the two variable are uncorrelation the regression lines are perpendicular to each other.

If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ (or) $\theta = \pi$.

(i.e) The two regression lines either co-incide or they are parallel to each other. But since both the lines of regression pass through the point $(\bar{x}, \bar{y})$ they cannot be parallel. Hence in the case of perfect correlation positive or negative the two lines of regression coincide.

Define correlation:-

The term correlation refers to the relationship between the variables.

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

Types of correlation:-

* positive and negative
* simple (or) Partial (or) multiple
* Linear (or) non-linear (or) No correlati

Positive correlation:-

When the value of the two variables changes in the same direction there is a positive correlation between

the two variables.

Negative correlation:-
When the value of the two variable changes in the opposite direction, there is negative correlation between the two variables.

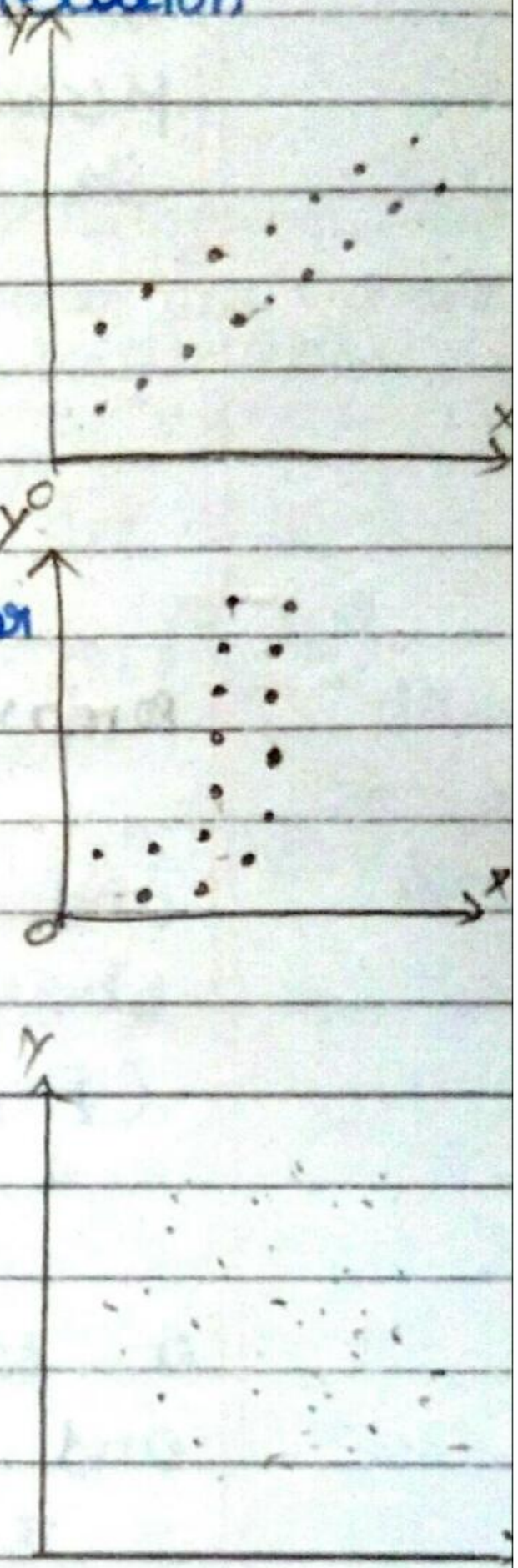
Simple correlation:-
When only two variables are considered as under positive or negative correlation the correlation between them is called simple correlation.

Partial correlation:-
When more than two variables are considered, the correlation between two of them when all other variables are held constants (i.e) when the linear effects of all other variables on them are removed is called partial correlation.

Multiple correlation:-
When more than two variable are considered, the correlation between one of them and its estimate based on the group consisting of the other variable is called Multiple correlation.

Linear correlation:-

When all the points lie on a line or scattered dots around a line, there is a linear correlation between the two variables.



Non-linear correlation:-

When all the points lie exactly on a curve or scattered around the curve, there is a non-linear correlation between the two variables.

No correlation:-

When the points are scattered neither around a line nor around a curve, there is no correlation between the two variables.