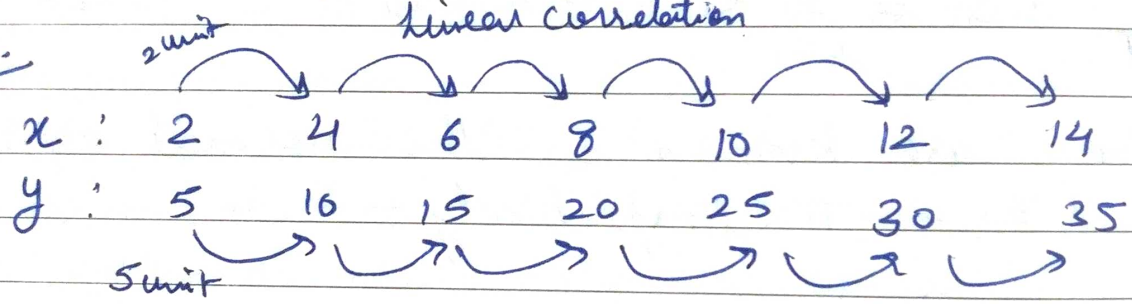
UNIT-II :-

Linear Correlation :-

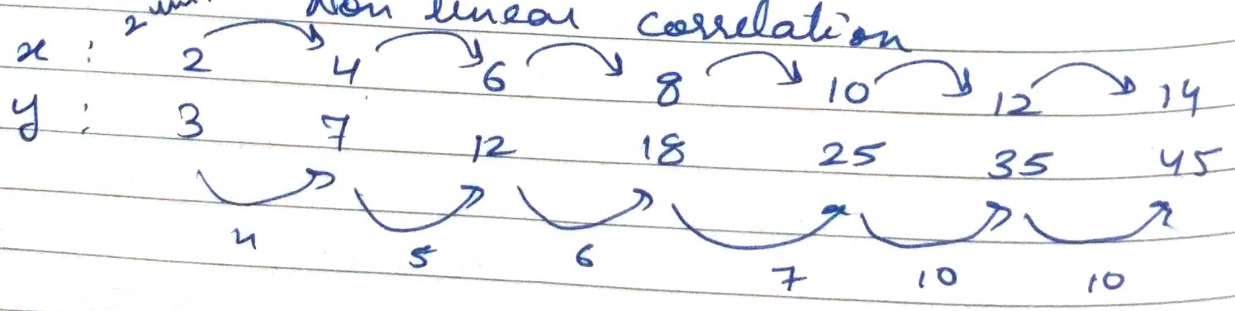
When two variable change in a constant proportion, it is called linear correlation. If the two sets of data bearing fixed proportion to each other are shown on a graph paper, their relationship will be indicated by a straight line Thus, linear correlation implies a straight line relationship.

eg:



Non-linear correlation :-

When the two variables do not change in any constant proportion, the relationship is said to be non-linear. Such a relationship does not form a straight line relationship.

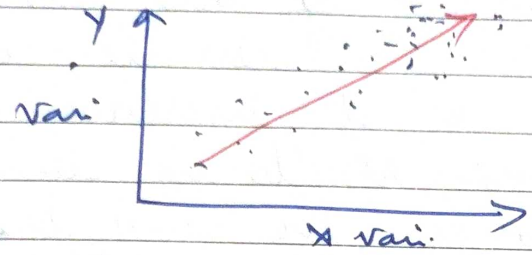


Correlation :- The term correlation is combination of two words.

Co - "Together" relation - "connection"

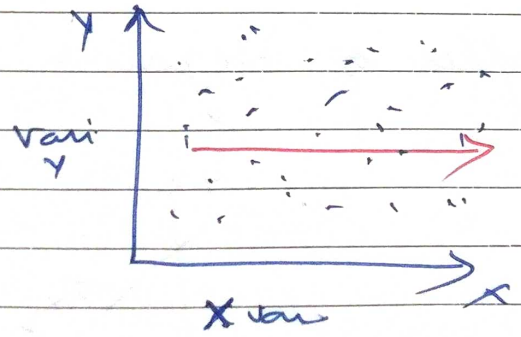
between two quantity

A unit change in one variable is reacted by another variable equivalently, directly or indirectly is known as or correlation between two variables.



A unit.

Or else a unit change in one variable does not show change in another variable or movement in one variable does not show movement in another one we say they are uncorrelated



Generally correlations are of two types

1) Positive Correlation & 2) Negative Correlation

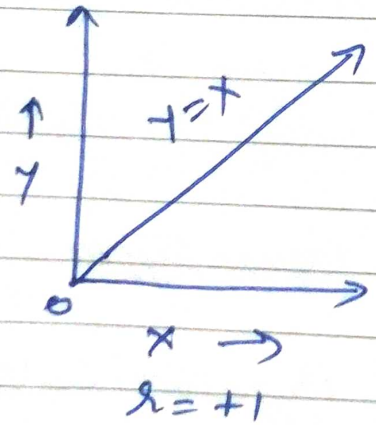
1) Positive correlation :- If two variables deviates in same direction i.e. if increase or decrease in one results in a corresponding increase or decrease in the other

we say that correlation is direct or positive

eg: i) the heights & weights of a group of persons
ii) the income and expenditure is positive and the correlation

2) Negative correlation :- If two variables deviate in opposite direction i.e. if increase or decrease in one results in decrease or increase in other, we say that the correlation is indirect or negative

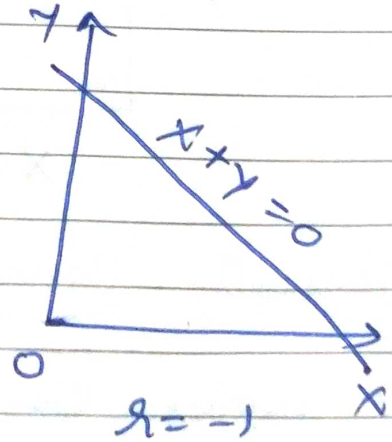
eg: i) Profit & loss of a commodity
~~ii) The income & expenditure is~~
ii) The price & demand of a commodity
iii) The volume & pressure of atmosphere.
all are examples of negative correlation



$r = +1$

Positive correlation

X = height, Y = weight
X = income, Y = Expenditure



$r = -1$

Negative correlation

X = Price, Y = demand
X = volume, Y = pressure

Karl Pearson Coefficient of correlation :—

To measure linear relationship between two variables Karl Pearson (1867 - 1936) a British Biomatrician developed a formula call correlation coefficient.
Correlation coefficient between two random variables X and Y, usually denoted by $r(X,Y)$ or simply r_{xy} is a numerical measure of linear relationship between them and is defined as

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

If (x_i, y_i) ; $i = 1, 2, \ldots, n$ is the bivariate distribution, then

$$cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

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

$$= \mu_{11}$$

$$\sigma_x^2 = E[\{X - E(X)\}^2] = \frac{1}{n}\Sigma (x_i - \bar{x})^2$$

$$\sigma_y^2 = E[\{Y - E(Y)\}^2] = \frac{1}{n}\Sigma (y_i - \bar{y})^2$$

Also

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

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

$$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 \quad ; \quad \sigma_y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2$$

Remark or Properties :-

1) Limits for correlation coefficient :-
Pearson corr. coeff. can not exceed 1 numerically. In other words it lies between -1 to +1.
Symbollically

$$-1 \leq r \leq +1$$

Proof :- Let us consider the sum of squares

$$\sum_{i=1}^{n} \left[\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y} \right]^2$$, which is always non-negative

$$\sum_{i=1}^{n} \left[\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y} \right]^2 \geq 0$$

$$\sum_{i=1}^{n} \left[\left(\frac{x-\bar{x}}{\sigma_x}\right)^2 + \left(\frac{y-\bar{y}}{\sigma_y}\right)^2 \pm 2\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right) \right] \geq 0$$

$$\Rightarrow \left[\sum_{i=1}^{n} \frac{(x-\bar{x})^2}{\sigma_x^2} + \sum_{i=1}^{n} \frac{(y-\bar{y})^2}{\sigma_y^2} \pm 2 \sum_{i=1}^{n} \frac{(x-\bar{x})}{\sigma_x}\frac{(y-\bar{y})}{\sigma_y} \right] \geq 0$$

Dividing whole eq^{n}. by n

$$\Rightarrow \left[\frac{1}{\sigma_x^2} \frac{1}{n} \sum_{i=1}^{n} (x-\bar{x})^2 + \frac{1}{\sigma_y^2} \frac{1}{n} \sum_{i=1}^{n} (y-\bar{y})^2 + \frac{2}{\sigma_x \sigma_y} \frac{1}{n} \sum_{i=1}^{n} (x-\bar{x})(y-\bar{y}) \right] \geq 0$$

$$\Rightarrow \left[\frac{1}{\sigma_x^2}\cdot\sigma_x^2 + \frac{1}{\sigma_y^2}\cdot\sigma_y^2 \pm \frac{2}{\sigma_x\sigma_y}\,\mathrm{cov}(x,y)\right] \geq 0$$

$$\Rightarrow \left[1 + 1 \pm \frac{2}{\sigma_x\sigma_y}\cdot r_{xy}\,\sigma_x\sigma_y\right] \geq 0$$

$$\Rightarrow \quad 2 \pm 2r_{xy} \geq 0$$

$$\Rightarrow \quad \pm r_{xy} \geq -1$$

$$\Rightarrow \quad -r \leq 1$$

$$\text{or} \quad r \geq 1$$

$$-1 \leq r \leq 1$$

Hence proved.

2) Correlation coefficient is independent of change of origin and scale.

If x and y are given variable and they are transformed to u and v by change of origin & scale viz

$$u = \frac{x-A}{h} \quad \& \quad v = \frac{y-B}{k} \quad ; \; h>0 \; k>0 \qquad \longrightarrow (1)$$

where A, B, h and k are constants $h>0, k>0$ then the correlation coefficient between x and y is same as the correlation coefficient between u and v. i.e.

$$\rho(x,y) = \rho(u,v)$$

$$\Rightarrow \quad \rho_{xy} = \rho_{uv}.$$

So $u = \frac{x-A}{h} \Rightarrow x = A + hu$

& $v = \frac{y-B}{k} \Rightarrow y = B + kv$ — (2)

Summing both sides and dividing by n, we get

$$\frac{\Sigma x}{n} = \frac{\Sigma(A+hu)}{n} \quad \& \quad \frac{\Sigma y}{n} = \frac{\Sigma(B+kv)}{n}$$

$$\bar{x} = \frac{nA}{n} + h\bar{u} \quad , \quad \bar{y} = \frac{nB}{n} + k\bar{v} \qquad \text{— (3)}$$

Subtracting (3) from (2) we get

$$x - \bar{x} = A + hu - A - h\bar{u}$$

$$y - \bar{y} = B + kv - B - k\bar{v}$$

$$(x - \bar{x}) = h(u - \bar{u})$$

$$(y - \bar{y}) = k(v - \bar{v})$$

$$\rho_{xy} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

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

$$r(x,y) = \frac{\Sigma h(u-\bar{u})\,k(v-\bar{v})}{\sqrt{\Sigma h^2(u-\bar{u})^2}\cdot\sqrt{\Sigma k^2(v-\bar{v})^2}}$$

$$= \frac{\Sigma (u-\bar{u})(v-\bar{v})}{\sqrt{\Sigma(u-\bar{u})^2}\sqrt{\Sigma(v-\bar{v})^2}}$$

$$r_{xy} = r_{uv}$$

3) Two independent variables are uncorrelated but the converse is not true.

Proof:- If X and Y are indept. variables, then

$$cov(X,Y) = 0$$

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = 0$$

Hence two indept. variables are uncorrelated, but the converse of the theorem i.e. two uncorrelated variables may not be independent may be understand by the following example

We have

							Total
X	-3	-2	-1	1	2	3	$\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	-27	-8	-1	1	8	27	$\Sigma XY = 0$

$$\therefore \bar{X} = \frac{1}{n}\Sigma X = 0, \quad cov(X,Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = 0$$

$$\therefore r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = 0$$

In above example we see that two variables are uncorrelated but not independent.

Correlation coefficient of Bivariate Frequency Distribution

When the data given are too large we arrange them in a two-way table.
If there are n classes for x and m classes for y, there will be in all $m \times n$ cells in the two way table. By going through the pairs of values of x and y we can find the frequency for each cell. The whole set of cell frequencies will define a bivariate frequency distribution.
The column totals and row totals will give us the marginal distributions of x and y.
A particular column or row will be called the conditional distribution of y for given x or of x for given y respectively.

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

X series → / Y series ↓	Classes Mid Points x_1 x_2 ..... x_i ..... x_m	Total of frequencies of Y $g(y)$
y_1		
y_2		
y_j	$f(x,y)$	$g(y) = \sum_x f(x,y)$
y_n		
Total of frequencies of X. $f(x)$	$f(x) = \sum_y f(x,y)$	$N \to \sum_x \sum_y f(x,y)$ $\downarrow$ $\sum_y \sum_x f(x,y)$

In the given table a bivariate frequency table is shown for x and y. There are m classes of y placed along the horizontal line and n classes of x along a vertical line and f_{ij} is the freq^{n}. of individuals lying in the $(i,j)^{th}$ cell.

Here

$$\sum_x f(x,y) = g(y)$$

is the sum of the frequencies along any row and

$$\sum_y f(x,y) = f(x)$$

is the sum of the frequencies along any column.

We observe that

$$\sum_x \sum_y f(x,y) = \sum_y \sum_x f(x,y) = \sum_x f(x) = \sum_y g(y) = N$$

Thus

$$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x,y) = \frac{1}{N} \left[\sum_x x \sum_y f(x,y) \right]$$

$$\bar{x} = \frac{1}{N} \sum_x x f(x)$$

Similarly

$$\bar{y} = \frac{1}{N} \sum_x \sum_y y f(x,y) = \frac{1}{N} \left[\sum_y y \sum_x f(x,y) \right]$$

$$\bar{y} = \frac{1}{N} \sum_y y \cdot g(y)$$

$$\sigma_x^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x,y) - \bar{x}^2$$

$$= \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

& $$\sigma_y^2 = \frac{1}{N} \sum_x \sum_y y^2 f(x,y) - \bar{y}^2$$

$$= \frac{1}{N} \sum_y y^2 g(y) - \bar{y}^2$$

$$cov(x,y) = \frac{1}{N} \sum_x \sum_y xy f(x,y) - \bar{x}\bar{y}$$

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