Scatter diagram:-

A scatter diagram is a diagrammatic representation of bivariate data. Suppose we are given n pairs of values of variables x & y. Taking two mutually perpendicular st. lines as axes of reference for x and y, each pair of given values can be plotted as a point in the graph paper. The figure obtained, when all the n pairs of values have been plotted, is called a scatter diagram or a dot diagram. By a scatter diagram one can know the nature and the intensity of association between the variables under study.

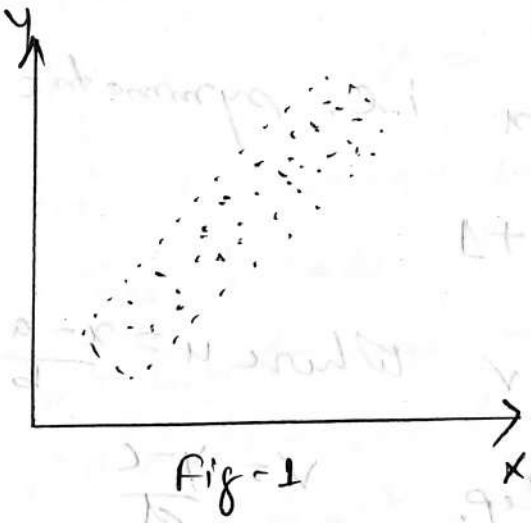

Fig-1

Fig-1 x & y are positively correlated (i.e $x\uparrow, y\uparrow$ / $x\downarrow, y\downarrow$)

Ex:- height and weight of a male. etc.

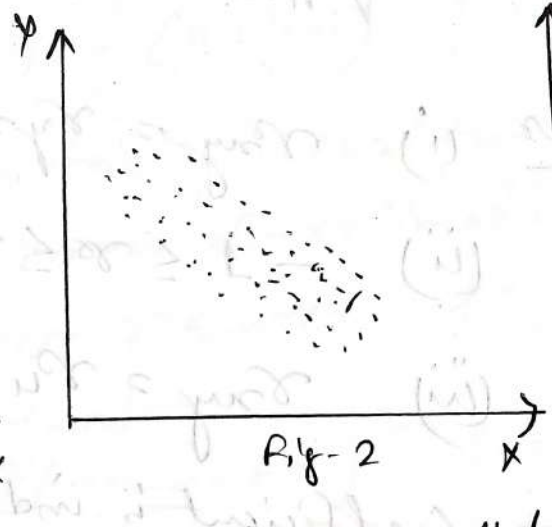

Fig-2

Fig2:- x & y are negatively correlated. (i.e $x\uparrow, y\downarrow$ / $x\downarrow, y\uparrow$)

Ex:- the price and demand of a commodity etc.

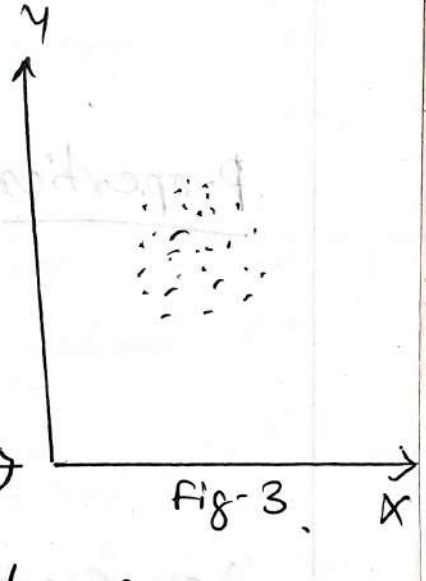

Fig-3

Fig-3:- Zero correlation.

Ex:- quality like affection, kindness is in most cases non-correlated with the academic achievements etc.

Correlation Coefficient :- Correlation Coefficient is a measure of degree or extent of linear relationship between two variables X & Y. The popln corr. coeff. is denoted by ρ and its estimate by r.

Let (x_i, y_i) be the n paired sample values. The formula for the sample corr. coefficient r is

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

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

Properties (i) $r_{xy} = r_{yx}$ i.e. symmetric.

(ii) $-1 \le r \le +1$.

(iii) $r_{xy} = r_{uv}$ where $u = \frac{x-a}{b}$, $v = \frac{y-c}{d}$

i.e. Corr. Coefficient is indep. of change of origin and change of scale also.

a, b, c, d are arbitary const. $a, b, c, d > 0$.

(1) St. Line. Curve fitting

Suppose we have n pairs of values $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ of the variables x and y.

Let the St. line eqn of y on x be.

$$y = a + bx.$$

Since we would like to use this equation for prediction purposes, the constant a and b have to be estimated on the basis of observed values of x and y. From among different methods that are available for determination of a and b, we use here the method of least squares. Which has many desirable properties.

When $x = x_i$, the observed value of y is y_i and the predicted value of y is $a + bx_i$. So

$$e_i = y_i - (a + bx_i) \quad [\text{observed} = \text{True value} + \text{error}]$$

is the error in taking $a + bx_i$ for y_i. $i = 1(1)n$.

This is called error of estimation. The method of least squares requires a & b should be so determined that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$ is a minimum with respect to a and b. The equations for the determination of a and b are

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n} e_i^2\right) = 0 \quad \text{and} \quad \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} e_i^2\right) = 0.$$

Hence we get

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \quad \text{---} (1).$$

and $$\sum_{i=1}^{n} (y_i - a - bx_i) \cdot x_i = 0.$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \quad \text{---} (2).$$

From the above two normal equations (1) & (2) we have to find the values of a and b. Let $\hat{a}$ and $\hat{b}$ are the estimated values of a and b.

So the fitted st. line eqn is

$$y = \hat{a} + \hat{b} x$$

parabolic eqn:-

Suppose we have n pairs of values $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ of the variables x & y.

Let the parabolic eqn be is,

$$y = a + bx + cx^2$$

Here the constants a, b and c have to be estimated on the basis of observed values of x and y. We use here the method of least squares

when $x = x_i$, the observed value of y is y_i and the predicted value of y is $a + bx_i + cx_i^2$, so

$e_i = y_i - (a + bx_i + cx_i^2)$ is the error in taking $a + bx_i + cx_i^2$ for y_i. $i = 1(1)n$. This is called error of estimation. The method of least sq. requires a, b & c should be so determined that $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$ is a minimum with respect to a, b & c. The equations for the determination of a, b & c are.

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n} e_i^2\right) = 0, \quad \frac{\partial}{\partial b}\left(\sum_{i=1}^{n} e_i^2\right) = 0, \quad \frac{\partial}{\partial c}\left(\sum_{i=1}^{n} e_i^2\right) = 0.$$

Hence we get,

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \quad \text{—— (1)}.$$

Again, $\sum_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0.$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \quad \text{—— (2)}$$

and, $\sum_{i=1}^{n} (y_i - a - bx_i) \cdot x_i^2 = 0$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^2 = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i^3 + c \sum_{i=1}^{n} x_i^4 \quad \text{—— (3)}$$

From the above three normal equations (1), (2) & (3) we have to find the values of a, b and c. Let $\hat{a}$, $\hat{b}$ and $\hat{c}$ are the estimated value of a, b and c.

So the fitted parabolic eqn is

$$y = \hat{a} + \hat{b}x + \hat{c}x^2$$

Exponential eqn :-

Let the exponential eqn is

$$y = a \cdot b^x$$

taking log in both side we get

$$\log y = \log a + x \log b.$$

or $Y = A + Bx$ where $Y = \log y$

$A = \log a$

$B = \log b.$

By the method of least square we get two normal equations.

$$\sum_{i=1}^{n} y_i = nA + B\sum x_i. \quad \text{---} (1)$$

$$\sum_{i=1}^{n} x_i y_i = A\sum x + B\sum x_i^2 \quad \text{---} (2).$$

From the normal equation we have to find the value of A & B. Let $\hat{A}$ & $\hat{B}$ are the estimated value of A & B.

$\therefore \hat{A} = \log a$ & $\hat{B} = \log b$

$\Rightarrow a = \text{antilog of } \hat{A}$ $\quad b = \text{antilog } \hat{B}$

So the fitted exponential eqn is

$$\therefore Y = \hat{a} \cdot (\hat{b})^x.$$

Goodness of Fit! - This is a very well-known test designed by Karl Pearson in 1900. Whether there is a discrepancy between the theory and practice of an exponential data can be checked through this test.

Let the null hypothesis H_0 be such that there is no significant difference between the values obtained through practice and that expected from the theoretical standpoint. The alternative hypothesis be such that the difference of values obtained through the theory and practice is significant.

We choose the test statistic as

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Where $O_i = (i = 1(1)n)$ is the ith observed value or frequency and $E_i (i = 1(1)n)$ be the i^{th} expected frequency.

The above is popularly known as chi-sq (χ^2) statistic, which is a χ^2 distn with $(n-1)$ degrees of freedom.

When the null hypothesis is true

$$\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$$

Which is the only constraint for which the degrees of freedom is reduced by one. When the calculated value of χ^2 is greater than the tabulated value of χ^2, the null hypothesis is to be rejected otherwise accepted.