Linear correlation and regression

Correlation is a statistical tool which studies the relationship between two variables.

Correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

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

Positive correlation

If the increase in the values of one variable results, on an average, in ~~the~~ a corresponding increase in the values of the other variable or if a decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable correlation is said to be positive or direct.

eg. (1) Heights & weights.

(2) The family income and expenditure on luxury items.

Negative correlation

The correlation is said to be negative, or inverse if the ~~[illegible]~~ increase (decrease) in the values of ~~the~~ variable results, on an average, in a corresponding decrease (increase) in the values of the other variable

(i) Price & demand of a commodity.

(ii) Volume & Pressure of a perfect gas.

<u>Linear and Non linear Correlation</u>

The correlation between two variables, is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values.

for eg. Consider the data,

x :	1	2	3	4	5
y :	5	7	9	11	13.

Here for a unit change in the value of x, there is a constant change say 2 in the corresponding values of y. This can be mathematically expressed as,

$$y = 2x + 3.$$

In general 2 variables x & y are said to be linearly related, if there exists a relationship of the form,

$$y = a + bx$$ between them.

ie This is the equation to a straight line if we are plotting the points in the xy-plane.

<u>Non linear or curvilinear correlation</u>

The relationship between two variables is said to be non-linear or curvilinear

If corresponding to a unit change in one variable, the other variable does not change at a constant rate.

In such cases if ~~the~~ the data are plotted on the xy-plane, we do not get a straight line curve.

Methods of studying correlation

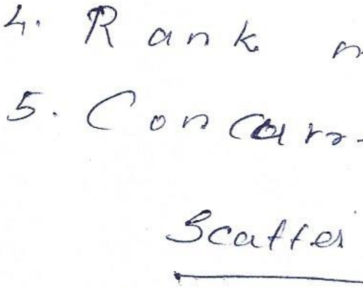
The commonly used methods for studying the correlation between two variables (linear correlation) are,

1. Scatter diagram method
2. Karl Pearson's coefficient of correlation.
3. Two-way frequency table
4. Rank method.
5. Concurrent deviations method.

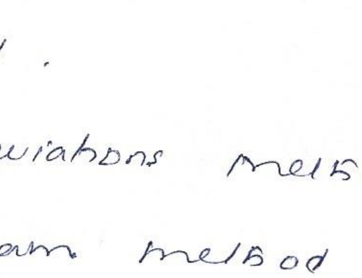
Scatter diagram method

From scatter diagrams we can form a rough idea about the relationships between the two variables.

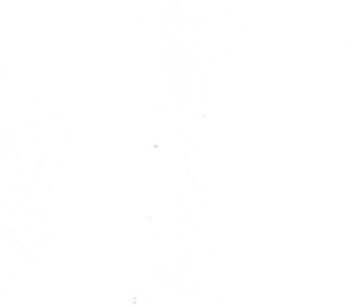
The following diagrams of the scattered data depict different forms of correlation.



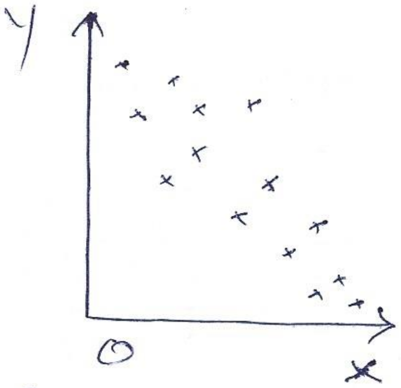
Perfect positive correlation.



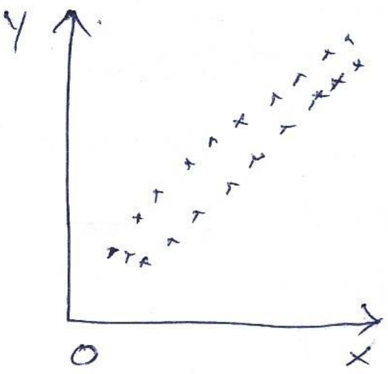
Perfect negative Correlation.



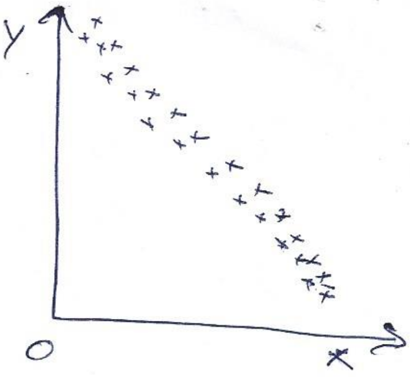
Low degree of positive correlation



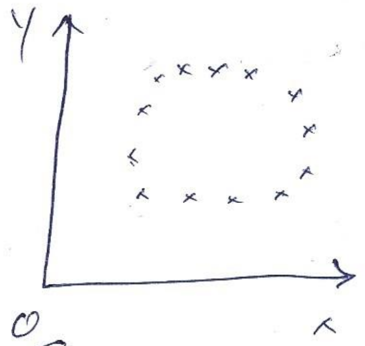
(Low degree ob negative correlation)



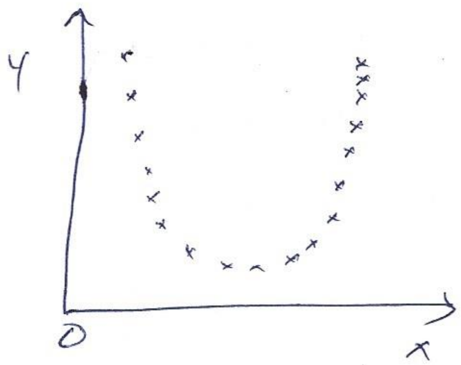
(Hight degree of Positive correlation)



(High degree of negative correlation)



(NO correlation).



(NO correlation)

The method of scatter diagram enables us to form a rough idea of the nature of the relationship between the two variables merely by inspection of the graph.

This method is not abbected by extreme observations, where as all mathematical formulae ob ascertaining correlation b/w 2 variables are abbected by extreme observations.

But this method is not suitable ib the no. of observations is large.

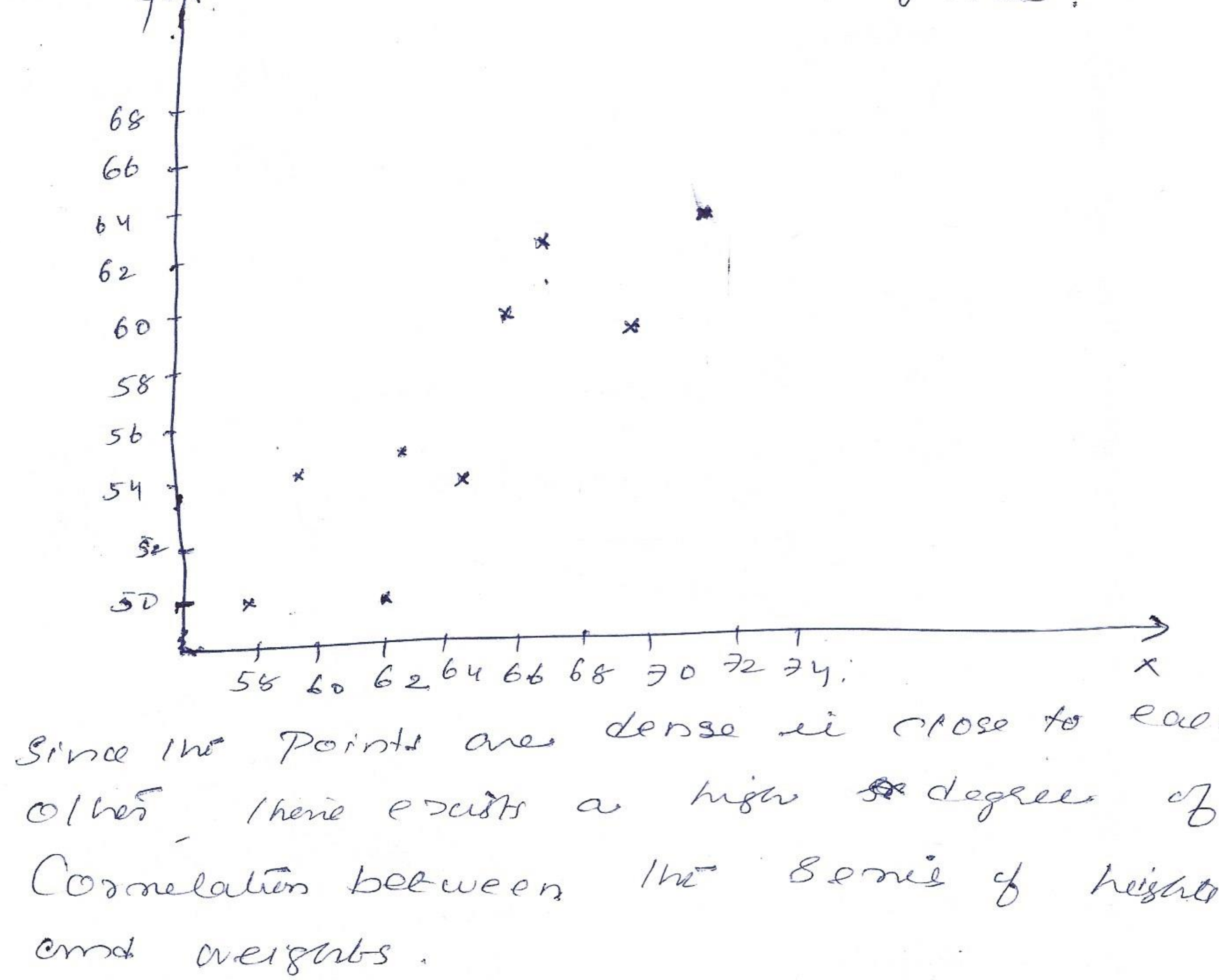
The scatter diagram method only tells us about the nature of the relationship. whether it is +ve or -ve and whether it is high or low. It ~~does not~~ does not

Provide an exact measure of the relationship b/w the 2 variables.

The scatter diagram enables us to obtain an approximate estimating line or line of best fit by free hand method.

eg: The following are the heights & weights of 10 students in a class.

Draw a scatter diagram and indicate whether correlation is positive or negative.



Since the points are dense i.e close to each other, there exists a high ~~se~~ degree of correlation between the series of heights and weights.

Also the points reveal an upward trend starting from left bottom and going up ~~up~~ towards the right top, the correlation is positive.

So there is a fairly high degree of positive correlation b/w the heights & weights of students in a class.

Karl Pearson's Coefficient of Correlation

The statistician karl Pearson suggested a mathematical method for measuring the magnitude of linear relationship b/w two variables.

Karl Pearson's or pearsons correlation coefficient b/w 2 variables X and Y denoted by $r(X,Y)$ or r_{xy} or r is a numerical relationship b/w and is defined as,

$$r = \frac{Cov(x,y)}{\sigma x \cdot \sigma y} \quad \text{———} (1)$$

where $Cov(x,y)$ is the covariance b/w X & Y

σx, the standard deviation of X

σy, the standard deviation of Y.

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

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

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

where $(x_1, y_1)(x_2, y_2)\ldots\ldots(x_n\, y_n)$ are n Pairs of observations of the ~~the~~ variables x and y in a bivariate distribution.

Substituting $cov(x,y)$, σx & σy in (1) we get,

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

Another formula

$$r = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left(n\Sigma x^2 - (\Sigma x)^2\right)\left(n\Sigma y^2 - (\Sigma y)^2\right)}}$$

Problem ① Calculate Karl Pearson's coefficient of Correlation between expenditure and sales from the data given below.

~~expenses: 39 65 62 90 82 75 25~~
~~Sales~~ :

Expenses:	39	65	62	90	82	75	25	98	36	78
Sales:	47	53	58	86	62	68	60	91	51	84

Let expenses be denoted by x & sales by y.

x	y	$x-\bar{x}=x-65$	$y-\bar{y}=y-66$	$(x-65)^2$	$(y-66)^2$	$(x-65)(y-66)$
39	47	−26	−19	676	361	494
65	53	0	−13	0	169	0
62	58	−3	−8	9	64	24
90	86	25	20	625	400	500
82	62	17	−4	289	16	−68
75	68	10	2	100	4	20
25	60	−40	−6	1600	36	240
98	91	33	25	1089	625	825
36	51	−29	−15	841	225	435
78	84	13	18	169	324	234
$\Sigma x = 650$	$\Sigma y = 660$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 5398$	$\Sigma(y-\bar{y})^2 = 2224$	$\Sigma(x-\bar{x})(y-\bar{y}) = 2704$

$\bar{x} = \frac{\Sigma x}{n} = \frac{650}{10} = 65$; $\bar{y} = \frac{\Sigma y}{n} = \frac{660}{10} = 66$.

Corrn coefft, $r = \frac{\Sigma[(x-\bar{x})(y-\bar{y})]}{\sqrt{\Sigma(x-\bar{x})^2 \, \Sigma(y-\bar{y})^2}} = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{\sqrt{12005152}} = \frac{2704}{3464.8451} = \underline{0.7804}$

Prob(2) Calculate the coefficient of correlation between x and y from the following data.

	x	y
no. of Pairs of observations	15	15
A. M.	25	18
Standard deviation	3.01	3.03
Sum of Squares of deviations from means	136	138.

Summation of products deviations of x & y from their arithmetic means = 122.

Solution

Using the notations,

$n = 15;\ \bar{x} = 25,\ \bar{y} = 18,\ \sigma x = 3.01,\ \sigma y = 3.03$

$\Sigma(x-\bar{x})^2 = 136;\ \Sigma(y-\bar{y})^2 = 138$

$\Sigma(x-\bar{x})(y-\bar{y}) = 122.$

Karl Pearson's Coefficient of Correlation $\Big\} = \dfrac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sigma x\ \sigma y}$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n.\sigma x\, \sigma y}$$

$$= \frac{122}{15 \times 3.01 \times 3.03} = \frac{122}{136.8045}$$

$$= 0.8917$$

Prob:(3) Given the following information,

$r_{xy} = 0.8;\ \Sigma xy = 60,\ \sigma y = 2.5,\ \Sigma x^2 = 90$

where x & y are the deviations from respective means, find the no. of items (n) here $x = X-\bar{X}$ & $y = Y-\bar{Y}$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n\sigma_x\sigma_y} = \frac{\Sigma xy}{n\sigma_x\sigma_y}$$

ie $0.8 = \dfrac{60}{n\times\sqrt{\frac{90}{n}}\times 2.5} = \dfrac{60}{\sqrt{n}\times\sqrt{90}\times 2.5}$

$$\left(\sigma x^2 = \frac{1}{n}\Sigma(X-\bar{X})^2 = \frac{1}{n}(\Sigma x^2) = \frac{90}{n}\right)$$

ie $\sqrt{n}\times\sqrt{90}\times 2.5\times 0.8 = 60$

$\sqrt{n}\times\sqrt{90} = \frac{60}{2} = 30.$

Squering $90n = 30^2$

$n = \frac{30\times 30}{90} = 10.$

ie no. of items $= \underline{\underline{10}}$

From the following table calculate the coefficient of correlation by karl Pearson's method.

X :	6	2	10	4	8
Y :	9	11	?	8	7

Arithmetic means of X & Y are 6 & 8 respectively.

Solution

firstly find the missing value.

It is given that $\bar{y} = 8$.

ie $\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{9+11+a+8+7}{5} = \dfrac{35+a}{5} = 8$

ie $35 + a = 40$

$\therefore a = 40 - 35 = \underline{5}.$

x	y	$x-6$	$y-8$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
6	9	0	1	0	1	0
12	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
$\Sigma x=30$	$\Sigma y=40$	0	0	$(\Sigma(x-\bar{x})^2=40$	$\Sigma(y-\bar{y})^2 = 20$	$\Sigma(x-\bar{x})(y-\bar{y}) = -26.$

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

$$= \cancel{\$}\ \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{\sqrt{800}} = \frac{26}{28.2843}$$

$$= -0.9192 = \underline{0.92}$$

Properties of correlation coefficient (r)

1. r always lies b/w -1 & $+1$

i.e. $-1 \leq r \leq 1$

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

3. Two independent variables are uncorrelated i.e. the ~~[illegible]~~ correlation b/w x & y is 0, if x & y are independent variables.

4. Coefficient of correlation between x & y is same as that b/w y & x.

5. Correlation coefficient has a well defined formula

~~6.~~

H.W ① Compute karl Pearsons Coefficient of Correlation

Price :	11	12	13	14	15	16	17	18	19	20
demand :	30	29	29	25	24	24	24	21	18	15

2. Find the coefficient of correlation between x & y and interpret the result.

x:	1.2	1.1	1.9	1.8	1.0	0.9
y:	15	10	20	10	10	5

[If you multiply x values by 10 & divide y values by 5, the result will not change. Then the new x & y values will be obtained)

Probable error

Probable error of the coefficient of correlation is a statistical measure which measures reliability and dependability of the value of coefficient of correlation.

If probable error is added to or subtracted from the coefficient of correlation, it would give two such limits within which we can expect the value of coefficient of correlation to vary.

$$\text{Probable error} = \frac{0.6745\,(1-r^2)}{\sqrt{n}}$$

where r is the coefficient of correlation and n is the number of pairs of observation.

The quantity $\frac{1-r^2}{\sqrt{n}}$ is called standard error of correlation.

Prob. If r = 0.6 & n = 64; find probable error & standard error.

$$\text{Probable error} = \frac{0.6745 \times (1-r^2)}{\sqrt{n}} = \frac{0.6745 \times (1-0.36)}{\sqrt{64}}$$

$$= \underline{\underline{0.54}}$$

$$\text{standard error} = \frac{1-r^2}{\sqrt{n}} = \frac{1-0.36}{\sqrt{64}} = \underline{\underline{0.08}}$$

Interpretation of coefficient of correlation on the basis of Probable error

1. If the coefft of correlation is less than its Probable error, correlation is not at all significant.

2. If correlation coefft is more than 6 times its Probable error, it is significant.

3. If the Probable error is not much and if the coefft of correlation is 0.5 or more it is generally considered to be significant.

In the Previous Problem, Probable error is very small & the correlation is significant