STATISTICS

(SESSION-6)

COVARIANCE :

- Co means two.

- In generally variance means ,How data varies

➢ **Variance =** $\frac{1}{N} \times \sum_{i=1}^{N}(x_i - \overline{x})^2$

- Variance Explained about data variance in column itself.

- For example , we have age data available.

How age is varying is explained by age variance.

AGE	SALARY
20	20000
25	25000
30	30000
35	35000
40	40000
45	45000
50	50000

Here , we observe that how a variable is changing means varying according to the another variable.

- In data we have columns , This columns also called as Features.

Variable = Columns = Features

- For Example **,** We have age Vs salary

How salary is varying according to the age is , This is given by **Covariance.**

- Variance = single column
- Covariance = two columns

Variance (x) = $\frac{1}{N}\times\sum_{i=1}^{N}(x_i - \bar{x})^2$

CoVariance (x,x) = $\frac{1}{N}\times\sum_{i=1}^{N}(x_i - \bar{x}) * (x_i - \bar{x})$

CoVariance (x,y) = $\frac{1}{N}\times\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})$

Step – 1 : Calculate the mean of Age ($\bar{x}$) and Income ($\bar{y}$)

Step – 2 : $(x_i - \bar{x})$ Subtract each value of Age from mean of Age

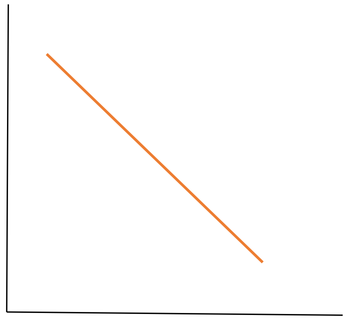
Step – 3 : $(y_i - \overline{y)}$ Subtract each value of Income from mean of Income

Step – 4 : Multiply $(x_i - \bar{x})(y_i - \bar{y})$

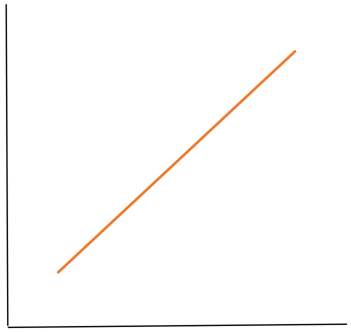
Step – 5 : Addition

Age	Income	$(x_i - \bar{x})$	$(y_i - \overline{y)}$	$(x_i - \bar{x})(y_i - \bar{y})$
20	20000	20-35 = -15	-15000	225000
25	25000	25-35 = -10	-10000	100000
30	30000	30-35 = -5	-5000	25000
35	35000	35-35 = 0	0	0
40	40000	40-35 = 5	5000	25000
45	45000	45-35 = 10	10000	100000
50	50000	50-35 = 15	15000	225000
Age ($\bar{x}$) = 35	Income ($\bar{y}$) = 35000			$\frac{700000}{7} = 100000$

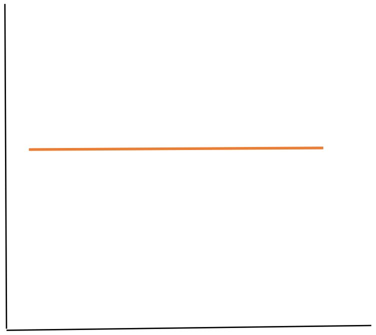
By seeing the covariance values we can say that both are positively related or negative



Negative correlation



positive correlation



no relation

Plot name : Scatter plot

- Graph between two numerical variables

COVARIANCE MATRIX :

Let us Assume ,

We have Age and Salary

$$\begin{array}{cc} A & S \\ \left[\begin{array}{c} v(A) \\ cov(S,A) \end{array}\right. & \left.\begin{array}{c} cov(A,S) \\ v(S) \end{array}\right] \end{array}$$

- There are two columns :

1. Age

2. Salary

- let us know ,

How many combinations possible

0 and 1

00 01 10 11

Age and Salary has 4 Combinations.

- Age with Salary : Covariance
- Age with Age : Variance
- Salary with Age : Covariance
- Salary with Salary : Variance

Variance(x) = covariance(x,x) = 1

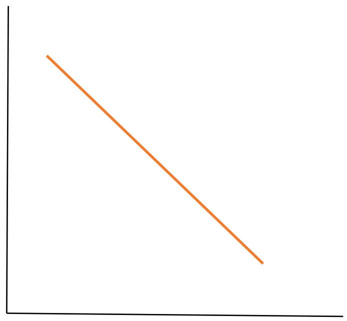
Covariance(x,y) = Covariance(y,x)

- If covariance has positive value , then that means the two variables are positively correlated.
- If covariance has negative value , then that means the two variables are negatively correlated.
- If covariance has zero value , then that means the two variables has no relation.
- Covariance Provides only whether the variables has relation or not.
- It will not provide how much percentage the variables are related with each other.
- Covariance can be anything from information to information.

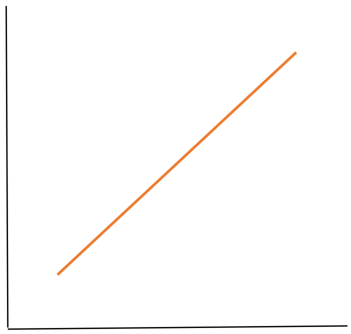
$\begin{bmatrix}25 & 5\\ 5 & 30\end{bmatrix}$ → covariance value is Positive So, Positive relation.

$\begin{bmatrix}25 & -5\\ -5 & 30\end{bmatrix}$ → covariance value is Negative So, Negative relation.

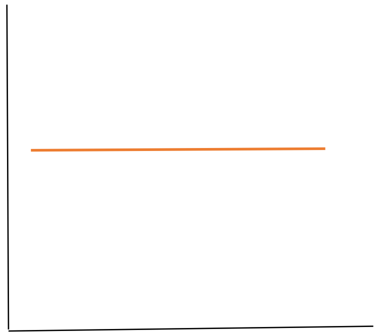
$\begin{bmatrix}25 & 0\\ 0 & 30\end{bmatrix}$ → covariance value is Zero So , No relation.



Negative correlation



Positive correlation



No correlation

Plot name : Scatter plot

- Graph between two numerical variables.

CORRELATION COEFFICIENT (r) :

We already know that , Covariance will provide that there is a relation or not , But it will not provide the amount of relation.

- Correlation will provide the Amount of relation.
- It will Explain how two variables related to each other.

In order to get the percentage of relationship we will use :

Pearson correlation coefficient

- It is denoted by ‘r’
- r varies from -1 to 1
- r = -1 to 0 means Negative relation
- r = 0 means no relation
- r = 1 means Positive relation

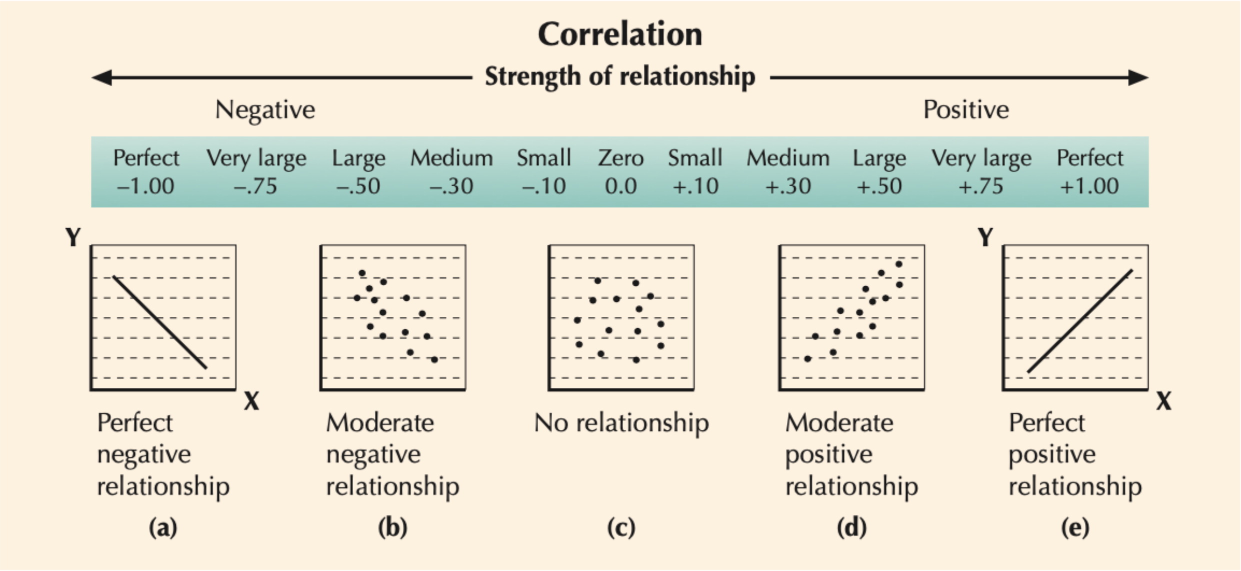
For Example :

- If r = 0.7
  - There is a 70% Positive relation between Age and Salary
- If r = -0.7
  - There is a 70% Negative relation between Age and Salary
- If r = 0
  - There is a no relation between Age and Salary

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

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

- Features or Variables are having No relation.
- No relation also called as Independent each other.
- Independence also called as Perpendicular each other.
- Perpendicular is also called as Orthogonal each other.
- Orthogonal is also called as 90 Degrees phase shift.



Perfect	Very large	Large	Medium	Small	Zero	Small	Medium	Large	Very large	Perfect
−1.00	−.75	−.50	−.30	−.10	0.0	+.10	+.30	+.50	+.75	+1.00

Perfect negative relationship (a)

Moderate negative relationship (b)

No relationship (c)

Moderate positive relationship (d)

Perfect positive relationship (e)

$\begin{bmatrix} 5 & 20 \\ 20 & 10 \end{bmatrix}$	$\begin{bmatrix} 5 & -20 \\ -20 & 10 \end{bmatrix}$	$\begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}$
(1)	(2)	(3)
Positive	negative	no relation