Statistics: Covariance and Correlation

Covariance

- Variance $= \frac{1}{N} * \sum_{i=1}^{N}(x_i - \bar{x})^2$
- Variance will explain how a data is varying
- Consider an age data of India, In India different age group people will be there
- If we want to how age in India is varying can be explained by Variance
- There is a situation we want to find out how an age related to income
- How an age related to income
- Here we are comparing two columns
- For that we have two methods:
    - Covariance
    - Correlation
- If we are working on single variable it is called as Variance
- If we work on two variables, then we will use Covariance (Two Columns)

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

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

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

$Age(x)$	$Income(y)$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
20	20000	-15	-15000	225000
25	25000	-10	-10000	100000
30	30000	-5	-5000	20000
35	35000	0	0	0
40	40000	5	5000	20000
45	45000	10	10000	100000
50	50000	15	15000	225000
$\bar{x} = 35$	$\bar{y} = 35000$			$= \frac{1}{N} * \sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})$ $= \frac{700000}{7}$ $= 100000$

Conclusion: Positive values indicates that there is a positive relation between age and income

Drawback:

- We can able to say, the features or columns are positively related or negatively related.
- But we cannot say, how much they are related

Covariance Matrix:

- Variables – Features – Columns – Input Variables – Input Variables – Prediction – Independent Variables
- All are same
- So, we have two columns:
  - Age
  - Income
- How many combinations possible? → 4 possible combinations
- 4 possible combinations are:

1. How age is varying w.r.t. income?
2. How income is varying w.r.t. age?
3. How an income is varying itself?
4. How an age is varying itself?

	Age(A)	Income(I)
Age(A)	$A - A(Variance)$	$A - I(CoVariance)$
Income(I)	$I - A(CoVariance)$	$I - I(Variance$

$$= \begin{bmatrix} V & CoV \\ CoV & V \end{bmatrix}$$

- If we have one more column as Education' '

	Age	Income	Education
Age	V	CoV	CoV
Income	CoV	V	CoV
Education	CoV	CoV	V

Correlation:

- We already know that; variance will provide is there relation between variables or not
- But it will not provide the amount of relation
- Correlation will provide amount of relation
- It will explain how two variables related with other; as well as how much they are related
- Correlation denoted by 'r'
- 'r' varies from -1 to +1
- -1 to 0: - Indicates negative relationship
- 0: - No relationship
- 0 to +1: - Indicates positive relationship

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

$$Var(x) = \frac{1}{N} * \sum_{i=1}^{N}(x_i - \bar{x})^2, standard\ deviiation = \sigma_x = \sqrt{Var(x)}$$

$$Var(y) = \frac{1}{N} * \sum_{i=1}^{N}(y_i - \bar{y})^2, standard\ deviiation = \sigma_y = \sqrt{Var(y)}$$

So, the formula for Correlation 'r' will be,

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})^2}}$$

$$r = \frac{Covariance(x,y)}{\sqrt{Var(x) * Var(y)}}$$

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

Example:

- r = 0.7, between age and income
- there is 70% positive relation between age and income

- r = -0.7, between age and income
- there is 70% negative relation between age and income

- r = 0, between age and income
- there is no relation between age and income

- **Independent of each other**
- **Perpendicular of each other**
- **90 degrees of phase shift**
- **Orthogonal of each other**

For Matrix:

$$\begin{bmatrix} 5 & -2 \\ -2 & 3 \end{bmatrix}$$

Negative

$$\begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix}$$

Positive

$$\begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix}$$

No relation