

Statistics

{ 11:30 - 12pm }

- ① Covariance
- ② Pearson Correlation Coefficient
- ③ Spearman Rank Correlation Coefficient
- ④ CHI SQUARE TEST
- ⑤ ANNOVA (F-Test)

} practicals

Covariance

X ↑ Y ↑

X ↑ Y ↓

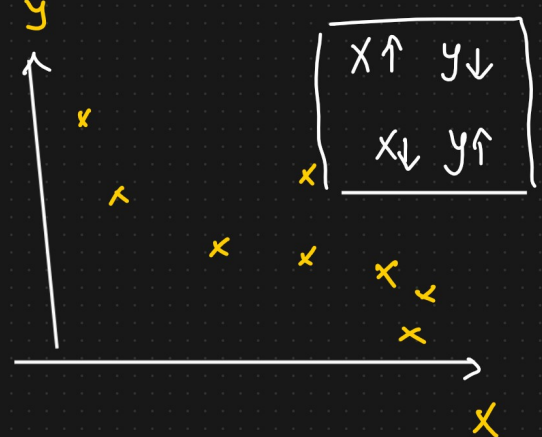
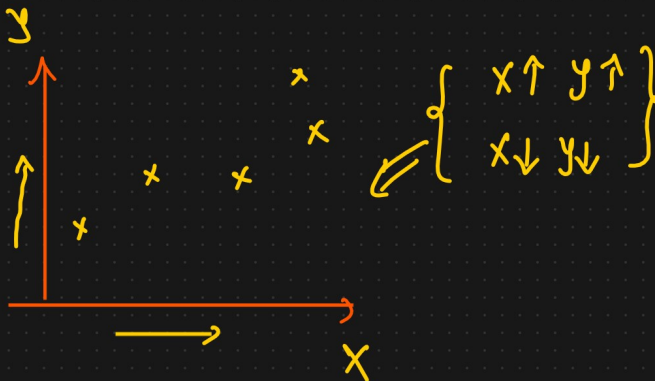
X ↓ Y ↑

X ↓ Y ↓

X =

Y =

{ quantity the relationship between X & Y }



$$Cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

↗
Cov(X, Y)

$$\Leftrightarrow Var(X) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

⇓

$$Cov(X, X) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$= \frac{\sum (x_i - \bar{x}) \times (x_i - \bar{x})}{N-1}$$

$$\text{Var}(X) = \sum_{i=1}^n \frac{(x - \bar{x})^2}{n-1} \Rightarrow \sum_{i=1}^n \frac{(x - \bar{x}) * (x - \bar{x})}{n-1}$$

↓

$$\underline{\underline{\text{Cov}(X, X)}} = \sum_{i=1}^n \frac{(x - \bar{x}) * (x - \bar{x})}{n-1}$$

+ve ↘ ↘ ↘ ↘

⇒ Positively Correlation

X ↑	Y ↑
X ↓	Y ↓

$$\text{Cov}(X, Y) = \left\{ \sum_{i=1}^n \frac{(x - \bar{x})(y - \bar{y})}{n-1} \right\}$$

$$\left\{ = \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{2} \right. \quad \bar{x} = 4 \quad \bar{y} = 5$$

$$= \frac{(-2)(-2) + 0 + (2)(2)}{2} = \frac{8}{2} = 4$$

X	Y
2	3
4	5
6	7

X ↑	Y ↓
X ↓	Y ↑

⇒ -ve Correlation ⇒ -ve value.

Disadvantage Covariance

$\text{Cov}(X, Y) \Rightarrow$ $\left. \begin{array}{l} \text{+ve value} \\ \text{or -ve value} \end{array} \right\}$

↓

Relationship

$[-1 \text{ to } 1]$

↑

$$\left. \begin{array}{l} \text{Cov}(X, Y) = 500 \\ \text{Cov}(Y, Z) = 600 \end{array} \right\}$$

Limit
+500 -400
-400 -300
+1000

∞

② Pearson Correlation Coefficient

$$r_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \quad [-1 \text{ to } 1]$$

The more the value towards 1 more the it is correlated

Dataset : 1000 features

X ~~X~~ Z A B C ~~X~~ O/P

Independent features

Dependent feature

True correlated

X, Y => 99%

90% 0.9
-ve correlation

③ Spearman Rank Correlation

$$r_s = \frac{\text{Cov}(R(x), R(y))}{\sigma_{R(x)} \sigma_{R(y)}}$$

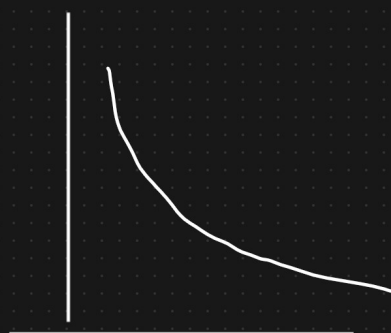
Marks



Spearman Rank
Cor = 1

X	Y	R _x	R _y
1	2	4	4
3	4	3	3
7	5	2	2
0	7	5	1
8	1	1	5

Keep it



-1

① Chi Square

The Chi Square Test Claims about population proportions.

It is a non parametric test that is performed on category (nominal or ordinal) data.

* In the 2000 U.S census, the ages of individuals in a small town were found to be the following.

<18	18-35	>35
20%	30%	50%

In 2010, ages of $n=500$ individuals were sampled. Below are the results

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the the population distribution of ages has ⁼ changed in the last 10 years?

Ans)	<18	18-35	>35
Expected	20%	30%	50%

$n=500$	Observed	121	288	91
	Expected	100	150	250

95% C.I

95% C.I

- ① H_0 = the data meets the expected distribution
 H_1 = the data do not meet the expected distribution

② State Alpha : $\alpha = 0.05$

③ Calculate the degree of freedom

$$df = n - 1 = 3 - 1 = 2 \Rightarrow 3 \text{ Categories.}$$

④ Decision Chi-Square Table.

If χ^2 is greater $\downarrow\downarrow$ 5.99 than, Reject H_0

5) Calculate Chi-square Test

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(121 - 100)^2}{100} + \frac{(288 - 150)^2}{150} + \frac{(91 - 250)^2}{250}$$
$$\chi^2 = \underline{\underline{232.494}}$$

$232.494 > 5.991$ Reject the Null Hypothesis.

- ⑤ A school principal would like to know which days of the week students are most likely to be absent. The principal expects the students will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which day of the week they had the highest number of

Student absences. The observed and expected results are shown in the table below. Based on these results, do the days for the highest number of absences occur with equal frequency (Use: 95% C.I)

	Monday	Tuesday	Wednesday	Thursday	FRIDAY
Observed	23	16	14	19	28
Expected	20	20	20	20	20.

Ans = 6.3 { Accept the Null Hypothesis }

{ Practicals + KDA + Feature Engineering }