# Statistics

{ 11:30 - 12pm }

1. Covariance
2. Pearson Correlation Coefficient
3. Spearman Rank Correlation Coefficient.

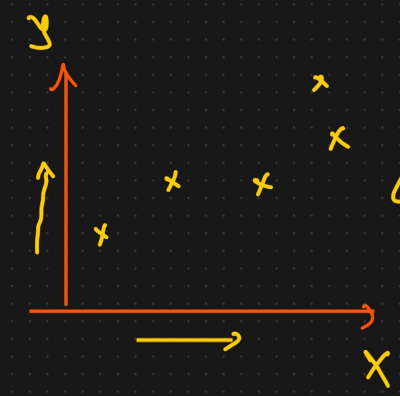
} ✓ ✓ ✓ practicals

4. CHI SQUARE TEST
5. ANNOVA (F-Test)

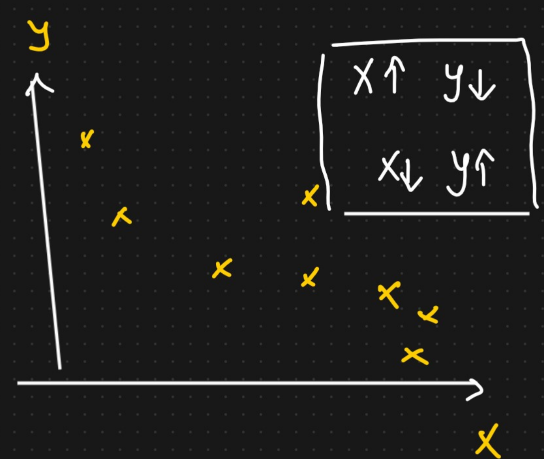
## Covariance

| | | X | y |
|---|---|---|---|
| X↑ | y↑ | = | = |
| X↑ | y↓ | - | - |
| X↓ | y↑ | - | - |
| X↓ | y↓ | - | - |

{quantify the relationship between X & Y}




$\left\{ \begin{array}{l} X\uparrow\ y\uparrow \\ X\downarrow\ y\downarrow \end{array} \right\}$




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

Cov(X, Y)

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

$\Downarrow$

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

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

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

$\Downarrow$

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

+ve

$$\boxed{\begin{matrix} X\uparrow & Y\uparrow \\ X\downarrow & Y\downarrow \end{matrix}} \Rightarrow \text{Positively Correlation}$$

| X | Y |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 6 | 7 |
| $\bar{x} = 4$ | $\bar{y} = 5$ |

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

$$= \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4) * (7-5)}{2}$$

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

$$\boxed{\begin{matrix} X\uparrow & Y\downarrow \\ X\downarrow & Y\uparrow \end{matrix}} \Rightarrow \text{-ve Correlation} \Rightarrow \text{-ve value.}$$

Disadvantage Covariance

$Cov(X,Y) = 500$

$Cov(Y,Z) = 600$

$Cov(X,Y) \Rightarrow$ +ve value or -ve value $\Rightarrow$ Limit

+500   -400   -300   -400   +1000

$\boxed{\infty}$

$\Downarrow$

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

$\Uparrow$

② Pearson Correlation Coefficient

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

[-1 to 1]

The more the value towards 1 more +ve it is correlated

Dataset : 1000 features

Independent features

Dependent feature

$\cancel{X}$ $\cancel{Y}$ Z A B C $\rightarrow$ $\cancel{O/P}$

+ve correlated

X, Y => 99%

90% 0.9

-ve correlation

⇓

Keep it

④ Spearman Rank Correlation

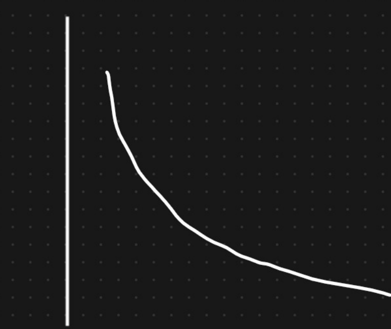
$$r_S = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \sigma_{R(y)}}$$

Marks

| ~~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 |

Spearman Rank

Corr = 1



-1

## (f) Chi Square

The Chi Square Test claims about population proportions.

It is a non parametric test that is performed on categorical (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% |

95% C.I

n=500

| | <18 | 18-35 | >35 |
|---|---|---|---|
| Observed | 121 | 288 | 91 |
| Expected | 100 | 150 | 250 |

① $H_0$ = the data meets the expected distribution

$H_1$ = the data donot meet the expected distn

② State Alpha : $\alpha = 0.05$

③ Calculate the degree of freedom

$df = n-1 = 3-1 = 2$ ⇒ 3 categories.

④ Decision

Chi Square Table.

⇓

If $X^2$ is greater 5.99 than, Reject $H_0$

5) Calculate Chi square Test

$$X^2 = \sum \frac{(fo - fe)^2}{fe} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$X^2 = 232.494$$

232.494 > 5.991 Reject the Null Hypothesis.

④ A school principal would like to know which days of the week students are most likely to be absent. The principal expect 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 absenses. The observed and expected results are shown in the table below. Based on these results, do the days for the highest number of absences occurs with equal frequencies (use 95% C.I)

| | Monday | Tuesday | Wednes | Thursday | FRIDAY |
|---|---|---|---|---|---|
| Observed | 23 | 16 | 14 | 19 | 28 |
| Expected | 20 | 20 | 20 | 20 | 20. |

Ans = 6.3 {Accept the Null Hypothesis}

{Practicals + EDA + Feature Engineering}