6th DAY LIVE SESSION

1. CHI SQUARE ✓
2. Covariance ✓
3. Pearson Correlation Coefficient ✓
4. Spearman Rank Correlation ✓
5. Practical Implementation
   Z-test, t-test, chi square test
6. F Test (ANOVA)

CHI SQUARE TEST

1. Chi Square Test claims about population proportions

(non-parametric test usually occurs wrt populatin proportion, when we are given some proportions of data, we can't specifically use parametric test)

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

→ why chi square test is used

Q) In the 2000 Indian Census, the age of the individual in a small town were found to be the following:

less than 18	18-35	>35
20%	30%	50%

In 2010, age 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 population distribution of ages has changed in the last 10 years?

Ans)

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

{Population} 2000

Expected

<18	18-35	>35
121	288	91
100	500 x 0.3	500 x 0.5

n=500

Observed

Expected

(We are multiplying as we want to find expected here wrt above 2000 population)

500*20/100

150 250

<18	18-35	>35
121	288	91
100	150	250

Observation

Expected

{Chi Square table}

① H_0 = The data meets the distribution 2000 census (null hypothesis)

H_1 = The data does not meet " " " (alternate hypothesis)

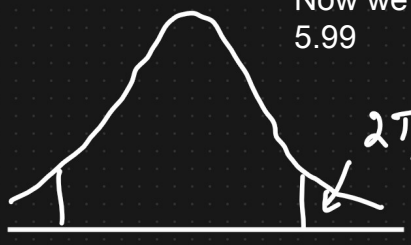
② $\alpha = 0.05$ (95% C.I)

df: 2, $\alpha = 0.05$

(degree of freedom)

③ Degree of freedom = n-1 = 3-1 = 2 // n = no. of categories, here 3(<18, 18-35, >35)

④ Decision Boundary



Now we will look in chi square table for given df = 2 and α = 0.05, we get 5.99

2 Tail Test

It is a 2 tail test because data may be less than our distribution or it may be more

If χ^2 is greater than 5.99 reject H_0

(Now we will find chi square and we will reject null hypothesis, if it's value is greater than 5.99)

5) Calculate Test Statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

// f0: observed, fe: expected

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

$$= 232.94$$

$\chi^2 = 232.494 > 5.99$ { Reject the Null Hypothesis }

$\alpha = 0.05$ ↙ 0.01 / 0.10 } Domain.

$0.002 < 0.05$ { Reject the null hypothesis }

If we get 0.11 so

$0.11 > 0.05$ ⇓ Accept the Null

0.11 ✗ 0.05 ⇓ Reject the Hypothesis

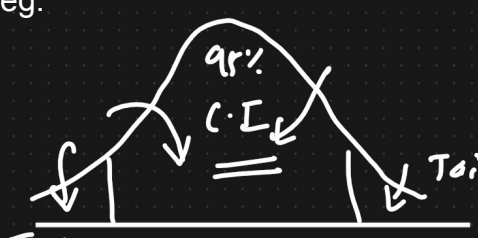
It is decided by domain expert

I Was Correct Here

⇓

If { p-value < Significance value ⇓ Reject the Null Hypothesis. } else { Accept the Null Hypothesis }

eg:



p 0.11 > 0.05 ⇓ Accept

$p = 0.002 < 0.05$ ⇓ Reject the null hypothesis

② Covariance

Eg 1:

X	Y
Weight	Height
50	160
60	170
70	180
75	181

X↑ Y↑

X↓ Y↓

here when x is increasing y is also increasing and when x is decreasing then y is decreasing

Eg 2:

No. of hour study	play
2	6
3	4
4	3

X↑ Y↓

X↓ Y↑

here when x is increasing y is decreasing and when x is decreasing then y is increasing

Quantify relationship between X & Y

How we can quantify the relationship using numbers: We use covariance

Covariance Formula:

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

+ve → positive Correlation

If X↑ Y↑ / X↓ Y↓

-ve → negative Correlation

If X↓ Y↑ / X↑ Y↓

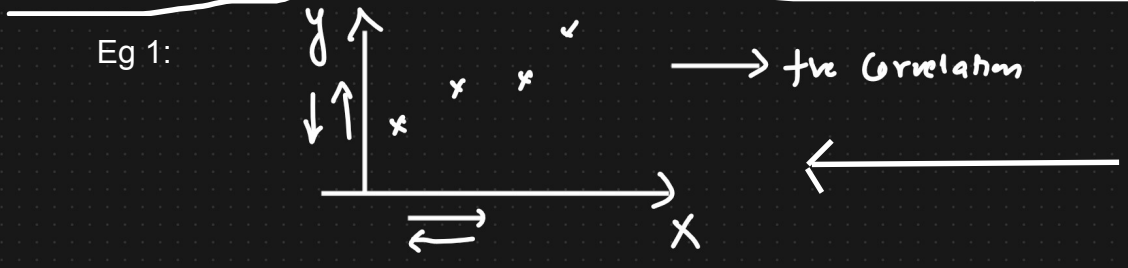
= +ve or -ve

or

0

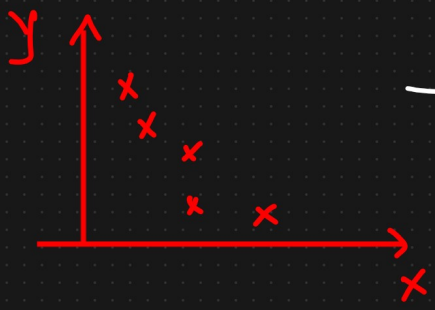
X↑ & Y

Eg 1:



Suppose we have a dataset which looks like this, this will have +ve correlation as here x is increasing then y is also increasing and when x decr then y decr. Even if we apply above formula we will get +ve value i.e +ve correlation

eg 2:

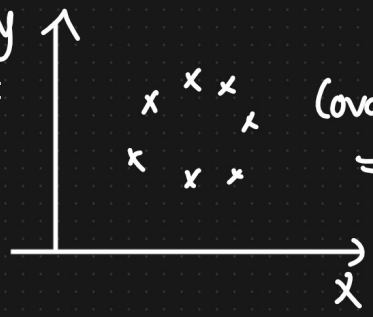


→ -ve Correlation

⇓

-ve Covariance

eg 3:



Covariance = 0

Disadvantage of Covariance

① Positive or Negative ✓

+100

-200

-2000

+1000 {Direction}

It can be +ve or -ve but it doesn't tell difference between 100 and 1000, i.e magintude can't be measured

Covariance can only measure the directional relationship between two assets. It cannot show the strength of the relationship between assets. The correlation coefficient is a better measure of that strength.

② Pearson Correlation Coefficient

(it restricts the value between -1 to 1)

(-1 to 1)

The more towards +1 more positively correlation

The more towards -1 more negatively correlated

Formula:

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \rightarrow \{-1 \text{ to } 1\}$$

Pearson coefficient is good for linear graphs, i.e when all the point lie on same line, it will give 1 or -1 as correlation

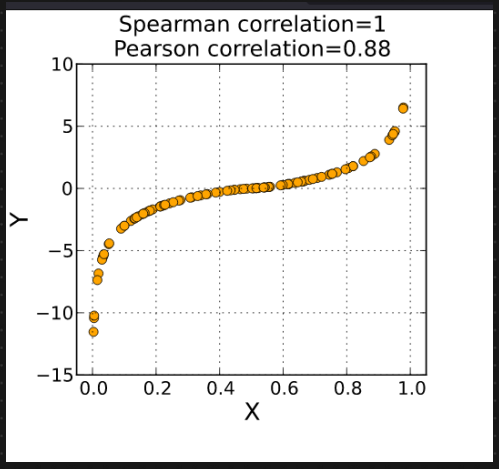
Spearman Correlation:

$$Spear_{(x,y)} = \frac{Cov(R(x), R(y))}{R\sigma_x \times R\sigma_y}$$

But when graph is non-linear and all point are increaseing, pearson coefficient will give close to 1 but not 1, here spearman works well and can give +1 correlation:

X Height	y Weight	Rank R(x)	Rank R(y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

180 has rank 1, 170 has rank 2 and so on



Non linear properties are captured by spearman rank correlation, that's why we use it

If $P_{value} \leq 0.05$ → Reject the NULL Hypothesis

Coding

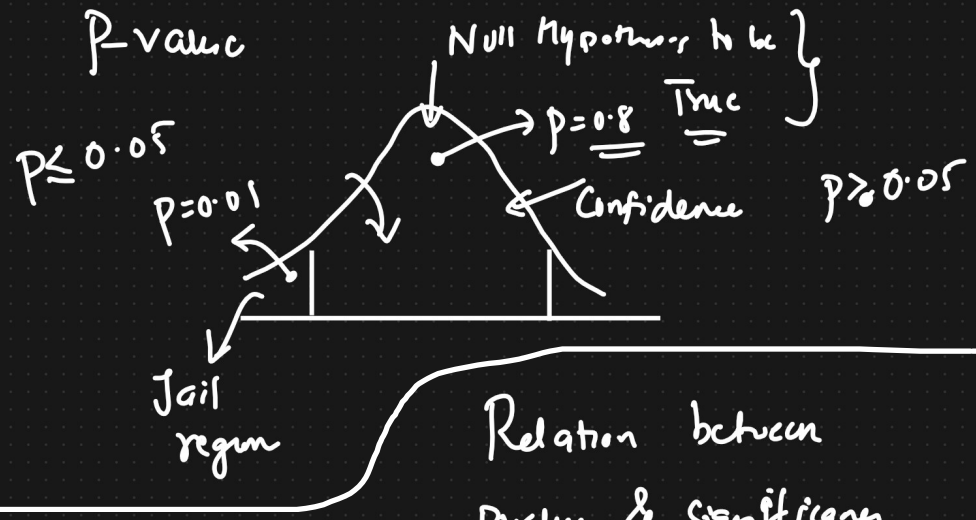
probability

↓

5% probability the null Hypothesis is Correct

$\alpha = 0.05$

$P \geq 0.05$ → Accept the NULL Hypothesis



Relation between pvalue & significance

P-value ≤ Significance [C.I]

↳ Reject the Null Hypothesis

p≥α

↳ Accept the Null

(To understand better check next pdf)

① P Value and Significance value