Chi - Square test

→ Tests for categorical Variable → $\boxed{\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$

→ Take two categorical variable and check whether they share a relationship i.e whether they are independent or dependent
→ It is commonly used for Testing relationships between variables. The null hypothesis of the chi - square test is the no relationship exists on the categorical variables in the population, They are independent.

H_0 : They are independent
null hyp

H_1 : They are dependent
alt hyp

Problem : Chi- square test of Independence

	non smoker	smoker	
athlete	14	4	18
non - athlete	0	10	10
	14	14	

$P(S) = 1/2$, $P(S') = 1 - 1/2 = 1/2$

$P(NS) = 1/2$, $P(NS') = 1 - 1/2 = 1/2$

$P(A) = \frac{18}{28}$,

$P(NA) = \frac{10}{}$

$\overline{28}$

H_0 : S and A are independent

H_1 : S and A are dependent

Step 1 : Assume H_0 is correct

① $P(S) \cdot P(A) = \frac{1}{\cancel{2}} * \frac{\cancel{18}^{9}}{28}$

$= \frac{9}{28} \times 28 = 9$ people

② $P(S) \cdot P(NA) = \frac{1}{2} * \frac{10}{28} = \frac{5}{28}$

$= \frac{5}{28} \times 28 = 5$

③ $P(NS) \cdot P(A) = \frac{1}{\cancel{2}} * \frac{\cancel{18}^{9}}{28}$

$= \frac{9}{28} \times 28 = 9$ people

④ $P(NS) \cdot P(NA) = \frac{1}{\cancel{2}} * \frac{\cancel{10}^{5}}{28}$

$= \frac{5}{28} \times 28 = 5$

Calculated Value

	non smoker	smoker	
athlete	9	9	18
non - athlete	5	5	10
	14	14	

Step 2 : Convert this into chi- squared distribution

$$\boxed{\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

$$\Rightarrow \frac{(14-9)^2}{9} + \frac{(0-5)^2}{5} + \frac{(4-9)^2}{9} + \frac{(10-5)^2}{5}$$

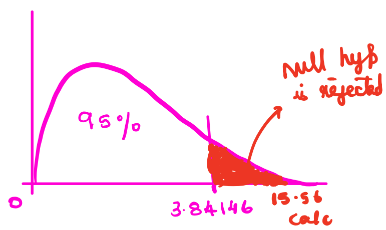
$= 2.78 + 2.78 + 5 + 5 = 15.56$ is **Not 0**

<u>Step 3</u> :

$df = (\text{no row} - 1) * (\text{no of colu} - 1) = 1$

$\alpha = 0.05$

$\therefore$ the critical value $= 3.84146$



Test statistics $= 15.56 >$ critical value $= 3.84146$

$\therefore$ Reject the null hypothesis

$\boxed{\text{Degree of freedom} = (\text{No of rows} - 1) * (\text{No of columns} - 1)}$

EM Algorithm $\Rightarrow$ Expectation Maximization Algorithm

Extremely Important

Normalization

Imp

Minmax Scaler

⇒ No matter the 'value' of the data after the application of Minmax Scaler the data is converted to [0, 1].

$$\boxed{\begin{array}{l} max = a \\ min = b \\ data = x_i \\ x_i \longrightarrow \dfrac{x_i - b}{a - b} \\ 0 < \dfrac{x_i - b}{a - b} < 1 \end{array}}$$

Standard Scaler

to make $x_1 \ldots x_n \longrightarrow N(\overset{\mu}{0}, \overset{\sigma^2}{1})$

$$\mu = \bar{x}$$

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n - 1}$$

$$\boxed{\text{normalization} \Rightarrow \begin{array}{l} x_1 \to \dfrac{x_1 - \mu}{\sigma} \\ x_2 \to \dfrac{x_2 - \mu}{\sigma} \\ \vdots \\ x_n \to \dfrac{x_n - \mu}{\sigma} \end{array}}$$

Robust Scaler

Change 'μ' mean by median

'σ' variance by IQR = Intra Quartile Range

i.e $\dfrac{x_i - (\text{median})}{IQR}$

Algorithms to study

1) Statistical Algorithm
2) ML Algorithms
3) Neural Networks / Deep Learning

1) Statistical algorithm

→ old
→ don't need computational efficiency
→ Pedantic ⇒ require many conditions to be satisfied

2) <u>ML Algorithm</u>

→ newer

→ need good - efficient computers

→ Need to satisfy minimal conditions

3) <u>DL Algorithms</u>

→ new

→ Extremely efficient machines

Statistical Algorithm

⇒ We try to bring out some sort of inference.

⇒ We make predictions.

Data → Continuous ⇒ Regression models

Data → Categorical ⇒ Classification models

Regression model

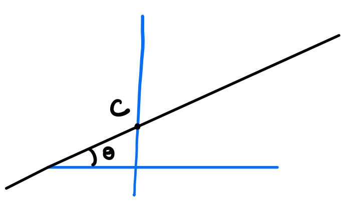
⇒ We will try to predict a continuous variable.

Linear model

first look at the plot

→ Linear model is essentially a straight line

Regression



$$y = mx + c$$

$$m = \tan\theta$$