Chi-Squared Test

- Recap framework, Z & T Test
- Degrees of Freedom
- Chi Square Test
  1. Coin Toss
  2. Shopping Example
  3. Aerofit Case Study

Framework

1. H_0 Vs H_a → Burden of proof (solid evidence)

   ↳ default assumption (in the absence of data)

2. Test stat (from observation)
3. R Vs L Vs Two-Tailed
4. p-value → Prob of seeing data assuming H_0 was true
5. Compare p-value with α → significance level

Chance Vs Signifance

(avg IQ of 10 student → 101.5)

Recovery after taking drug 1 Vs drug 2 } ttest_ind

Two sets of samples

About population Std dev "σ" & no. of samples "n"

$Z\_stat = \dfrac{X-\mu}{\sigma/\sqrt{n}}$

T_stat $\dfrac{X-\mu}{s/\sqrt{n}}$

2 possibilities

→ we know σ — Z test

→ do not know σ
"s" sample std dev

→ n is small (n < 30) — T test

→ n is large (n > 30) — T test (can also use Z Test)
$s \simeq \sigma$

By default → use T test

Degree of freedom

Salary ① $35L, 36L, \underbrace{?}_{34L} \xrightarrow{\text{avg}} 35L$

From these 3 nos, we only need 2 (if we know avg)

② $35L, 36L, \underbrace{?}_{\downarrow\, 39L}, 38L \xrightarrow{\text{avg}} 37L$

From 3 out of 4, we can find missing no.
if avg is known

Q If "n" nos are given with their mean,
how many will I need to know?

$\underline{\underline{n-1}}$ "degrees of freedom"

H & W

	Inch	Kg
Df=8	73	85
	68	73
	74	96
	71	82
	X	X
Avg	71	81.2

n_1 no of height values

n_2 no of weight values

& their avg

$n_1 - 1 + n_2 - 1$

$\boxed{n_1 + n_2 - 2} \longrightarrow$ degrees of Freedom

Sachin Century Vs Victory

Win

Century	False	True	
False	160	154	314
True	16	30	46
	176	184	360

df = 1

Regional support for Politician

4 politicians → A , B , C , D

3 Cities → X , Y , Z

$(r-1)(c-1)$

$2 \times 3 = 6$

	A	B	C	D	
X	90	60	104	95	349
Y	30	50	51	20	151
Z	30	40	45	35	150
	150	150	200	150	650

df = 6

2×3

Coin Toss : fair or Biased

50 tosses

H_0: fair coin

H_a: Biased coin

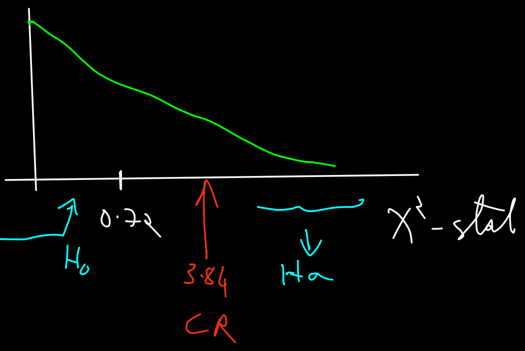
	Heads	Tails
Expected	25	25
Observed	28	22

$\rightarrow df = 1$

$$\chi^2 \text{ statistic} : \frac{(28-25)^2}{25} + \frac{(22-25)^2}{25} = 0.72$$

Under H_0, will this stat be low or high

Right tailed because H_a is on right side



Online Vs Offline → does this preference depend on Gender

Survey observation

	M	W		
Offline	527	72	599	66%
Online	206	102	308	34%
	733	174	907	

Expected under H_0

	M	W
Offline	484	115
Online	249	59

H_0: Offline/online is independent of Gender

H_a: depends on gender

$$\frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + (\quad) + (\quad) \simeq 59$$

Suppose H_0 is true, 66% prefer offline,

Among 733 men, how many are expected to prefer offline?

→ 66% of 733 → 484

Among 174 women, how many are expected to prefer offline?

→ 66% of 174 → 115

$$\chi^2 : \frac{(527-484)^2}{484} +$$

$$\sum_{i=1}^{4} \frac{(O_i - e_i)^2}{e_i}$$

Assumptions of Chi Squared

① Variables are Categorical

② Observations are independent

③ Each cell is mutually exclusive

④ Expected value in each cell should be greater than 5 (atleast 80% of the cells)