Lecture 5: Chi_Square Test

#Agenda

1. Types of Test
2. Degrees of Freedom
3. Chi-squared Goodness of fit Test
4. Chi-squared Test for independence

Types of Test

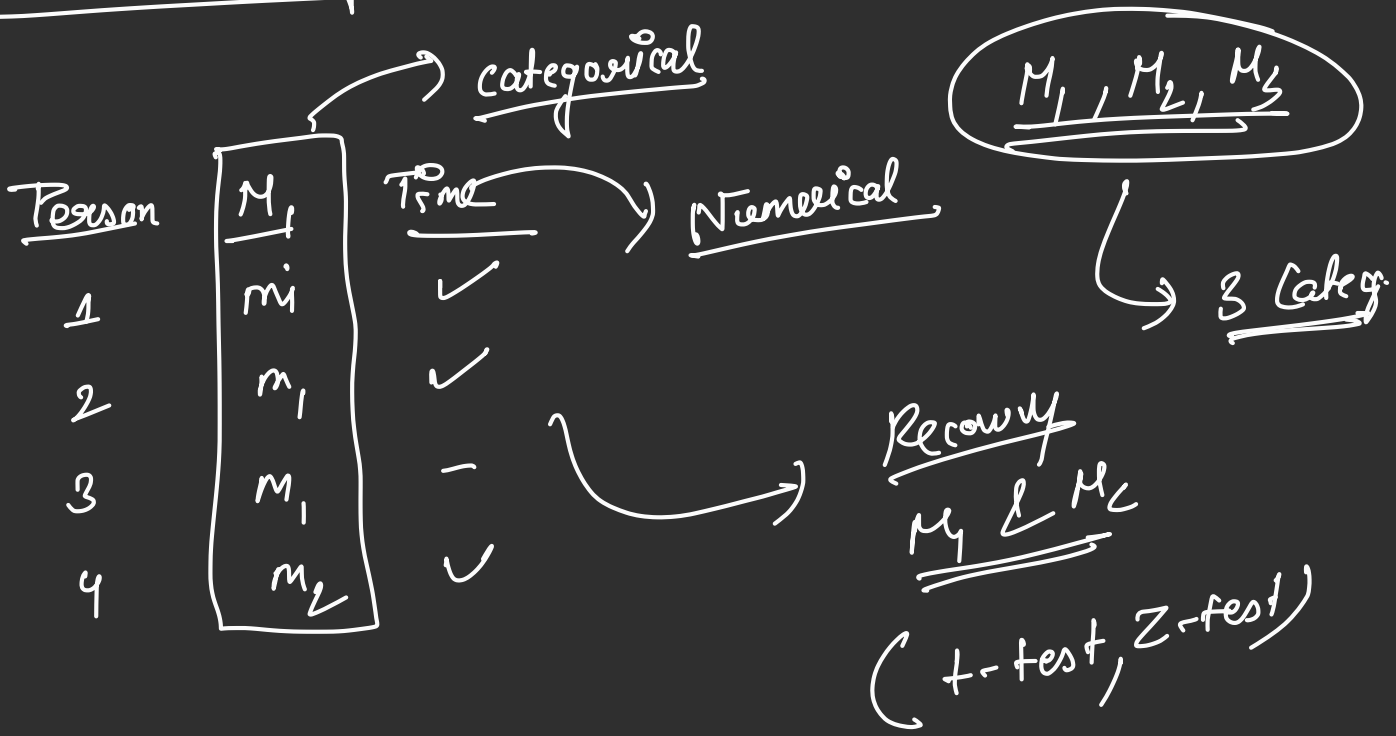
So far

Numeric Vs Categorical → 2 (T. test)
→ >2 (ANOVA)

Categorical Vs Categorical → Chi square test

Numeric Vs Numeric → Correlation

Numerical Vs Categorical



#Degrees of freedom (DOF)

① Setup I: Salary data

$P_1 \rightarrow 35L$

$P_2 \rightarrow ?\ y$

$P_3 \rightarrow ?\ (x)$

$Avg \rightarrow 35L$

$$\frac{x + 30 + 35}{3} = 35$$

$$x = 40$$

$$\frac{35 + x + y}{3} = 35$$

② Setup II : Salary data

$P_1 \rightarrow 35L$

$P_2 \rightarrow 36L$

$P_3 \rightarrow$? (x)

$P_4 \rightarrow 30L$

Avg $\rightarrow$ 37L

Q) What is the salary of the 3rd person?

$$\frac{35 + 36 + x + 30}{4} = 37$$

$x = 47$

$x = 47$ ✓

3 person salary should be present →

Generalise

n-1

$$\{DOF = n-1\}$$

③ Setup III : Height & Weight

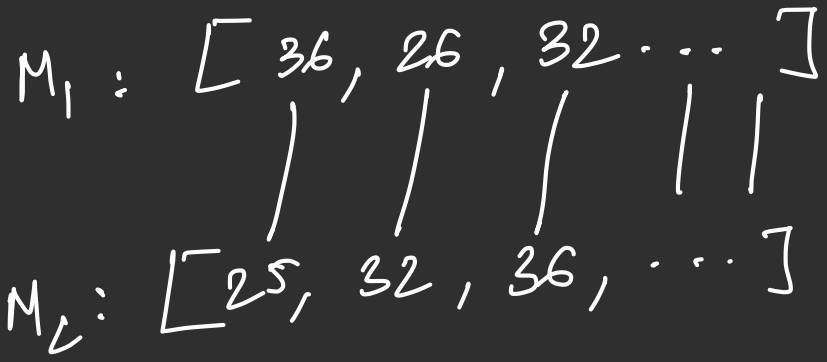
$n_1, n_2 \neq n$

H	W
73	85
68	73
74	
	82
62	70
Avg 71	81.2

$H \rightarrow n_1 - 1$
$W \rightarrow n_2 - 1$
} for ho $2(n-1)$ ✗

$(n_1 - 1) + (n_2 - 1)$
DOF_1 DOF_2

$DOF = (n_1 + n_2 - 2)$


M1 : [36, 26, 32 ...]
M2 : [25, 32, 36, ...]

Summary

For 1 sample: DOF $\Rightarrow (n-1)$

For 2 sample: DOF $\Rightarrow (n_1 + n_2 - 2)$

(4) Setup IV: Sachin (Century Vs Victory)

Win

Cent	F	T	
F	160		314
T			46
	176	184	360

→ row sum

→ grand Total

col sum

(DOF → 1)

$(1 \times 1) = \underline{\underline{1}}$

⑤ Setup V : Politicians

4 politicians : A, B, C, D
3 cities : X, Y, Z

$$(\#row - 1) * (\#cols - 1)$$

DOF ⇒ In this case

$(3-1) \times (4-1)$

$2 \times 3 = 6$

	A	B	C	D	
X	90	60	104		349
Y	30	50	51		151
Z					150
	150	150	200	150	650

DOF = 6

Two arrays

What will be the DOF?

Array 1 $\rightarrow (n_1 - 1)$

Array 2 $\rightarrow (n_2 - 1)$

$\Rightarrow (n_1 - 1) + (n_2 - 1)$ Ans

$\Rightarrow$

$M_1 = [62, \ldots]$ → Nume

$M_2 = [\ldots\ldots]$ → Nume

Medic	RoT
M_1	62
M_1	63
M_1	

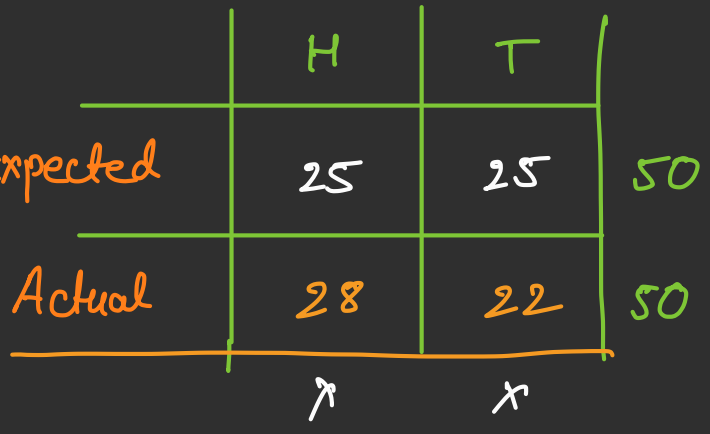
mean()

M_2 ( ) → mean()

#Coin Toss Example

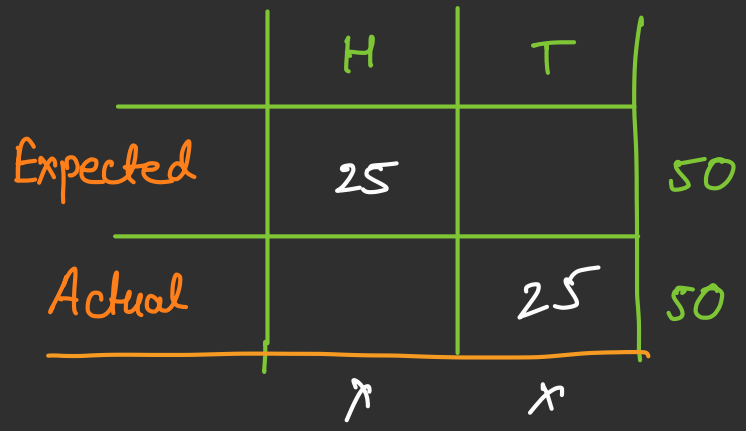
Objective: Check if a given coin is fair

No of tosses = 50 Times



Under H_0 (Coin is fair)

Guarantee !! → 25 H, 25 T



No of tosses = 50 Times

DOF = (#row - 1) × (#col - 1)

⇒ (2-1) × (2-1)

⇒ (1)

DOF = 1

DOF = 2

No of tosses = 50 Times

	H	T	
Expected	25	25	50
Actual	28	22	50

Under $H_0 \Rightarrow$

$(28 - 25) = +3$

$(22 - 25) = -3$

$+3 - 3 = 0$ ✓

Case I : $(28-25)+(22-25)$

$\Rightarrow +3+(-3)=0$ { the value becomes zero }

Case II : $(28-25)^2+(22-25)^2$

$\Rightarrow 9+9=\underline{18}$ (We solved the above problem, but it doesn't work when there are many trials/ sample)

Case III :

For $n=1000$,

$$(625-500)^2+(375-500)^2 = \checkmark$$

(This value is more as the trials keep increasing)

	H	T
Exp	500	500
Actual	650	350

$n = 1000$

$$(650 - 500)^2 + (350 - 500)^2 = \uparrow$$

Case IV (Final Solution)

	H	T
Expt	25	25
Actual	28	22

$$\Rightarrow \frac{(28-25)^2}{25} + \frac{(22-25)^2}{25} =$$

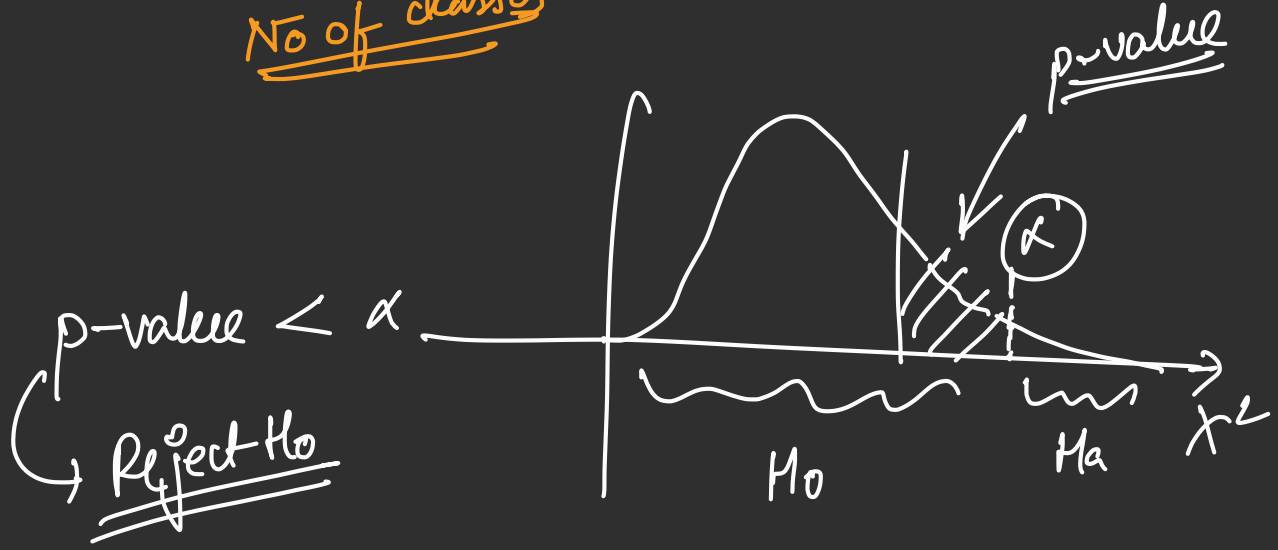
$$\text{chi}\,(\chi^2) = \sum_K \frac{(O-E)^2}{E}$$

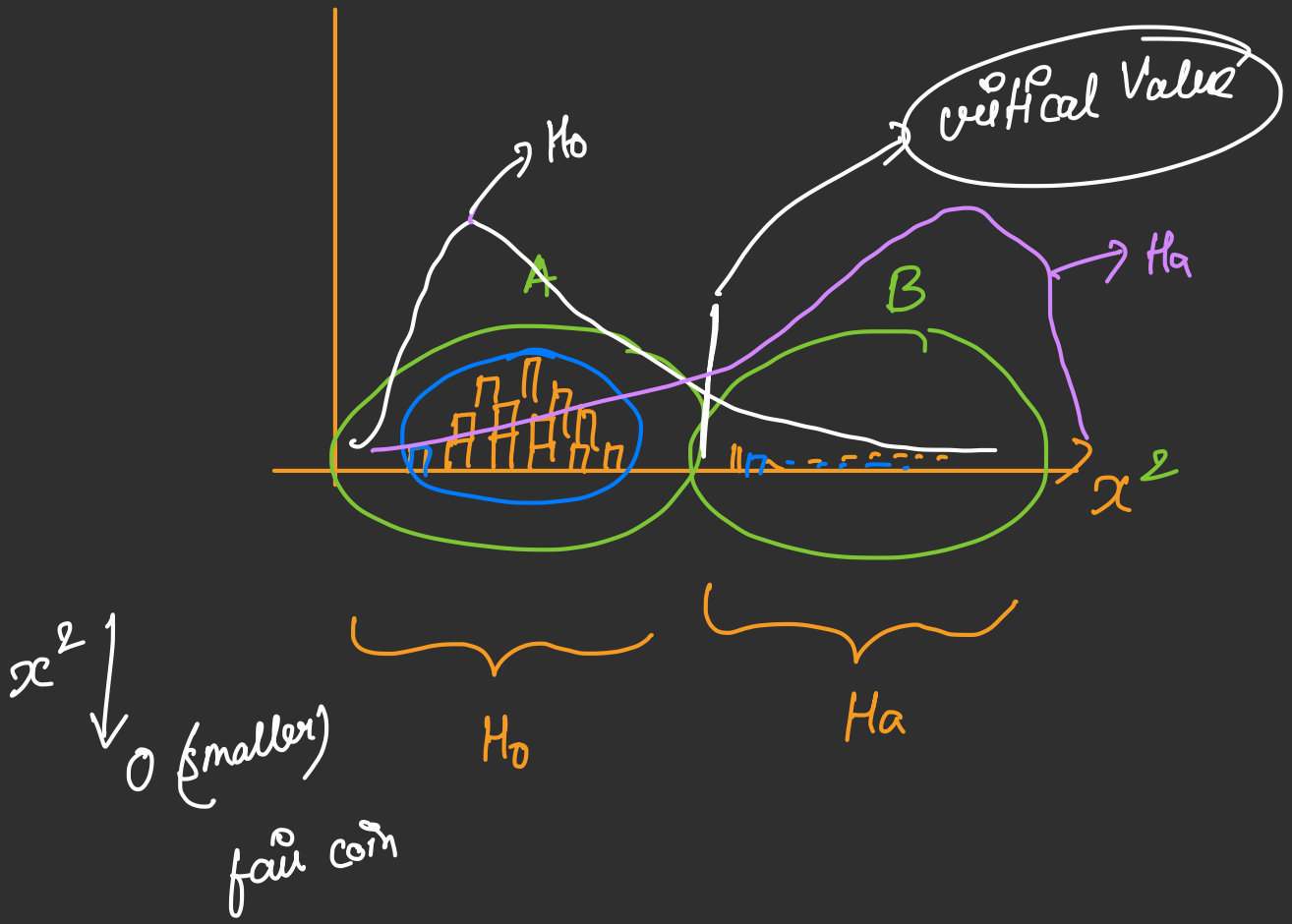
T-statistic

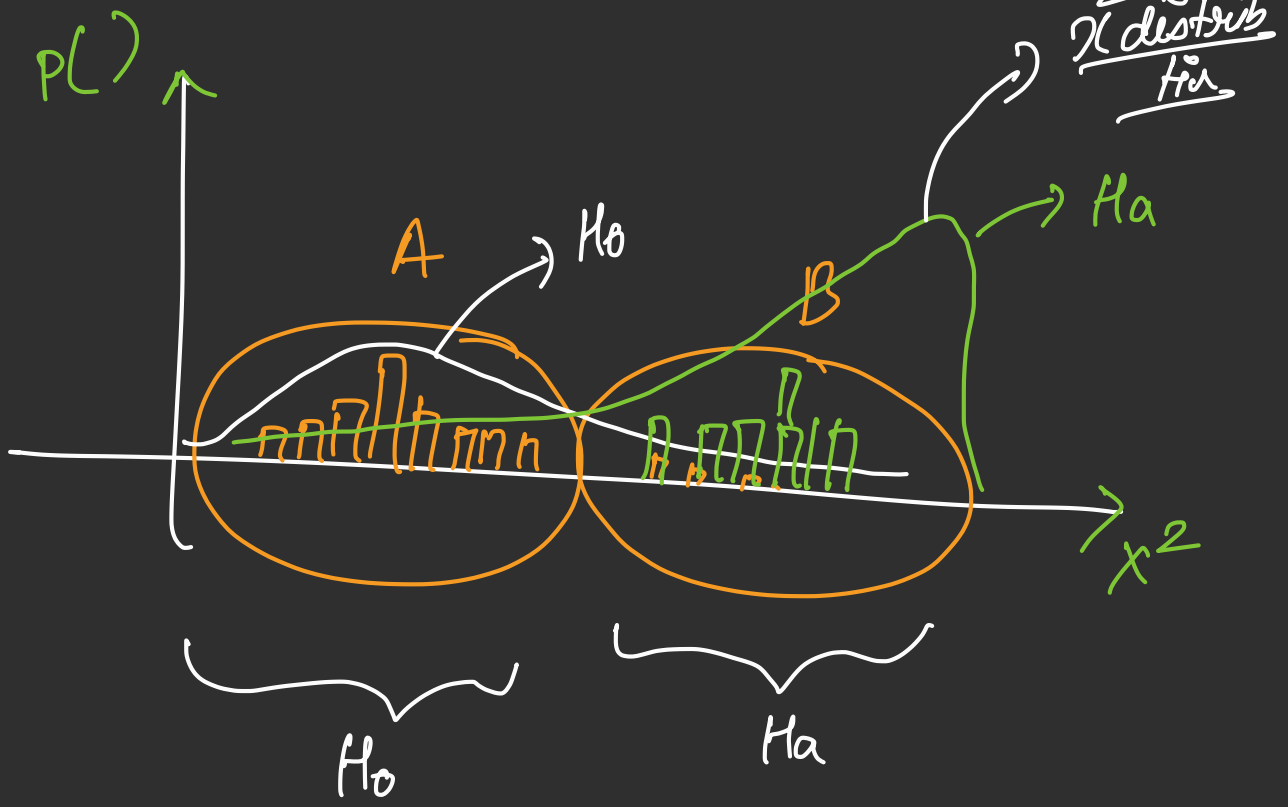
$$x^2 = \sum_{k} \frac{(O-E)^2}{E}$$

- O: observed value
- E: Expected

k: No of classes




Critical Value
H_0
A
B
H_a
χ^2
χ^2 ↓ 0 (smaller)
fair coin
H_0
H_a


P( )
A
Ho
B
Ha
χ^2 distribution
χ^2
Ho
Ha

Quiz

$$\chi^2 = \sum_{K} \frac{(O - E)^2}{E}$$

Observed

Expected

Question:

To assess whether a coin is fair or not, we need to compare the expected outcomes with the observed outcomes of tossing the coin.

The expected outcome for a fair coin toss is 50% heads and 50% tails, which corresponds to 25 heads and 25 tails in 50 tosses.

The observed outcomes from 50 coin tosses are 28 heads and 22 tails.

To determine whether the coin is fair, we perform a chi-square test to check if the observed results significantly deviate from the expected results. If the deviation is statistically significant, it may indicate that the coin is not fair.

Step 1: Assumption

Ho: Coin is fair

Ha: Coin is biased

Step 2: Distribution ?

Chi sq distribution

✓ Step 3: p-value

✓ Step 4: Comparing with α

50 Tosses

	H	T
Expect	25	25
Actual	28	22

$$\chi^2 = ?? \Rightarrow \left(\frac{(28-25)^2}{25} + \frac{(22-25)^2}{25} \right)$$

$$\Rightarrow \underline{0.72}$$

Ques 2

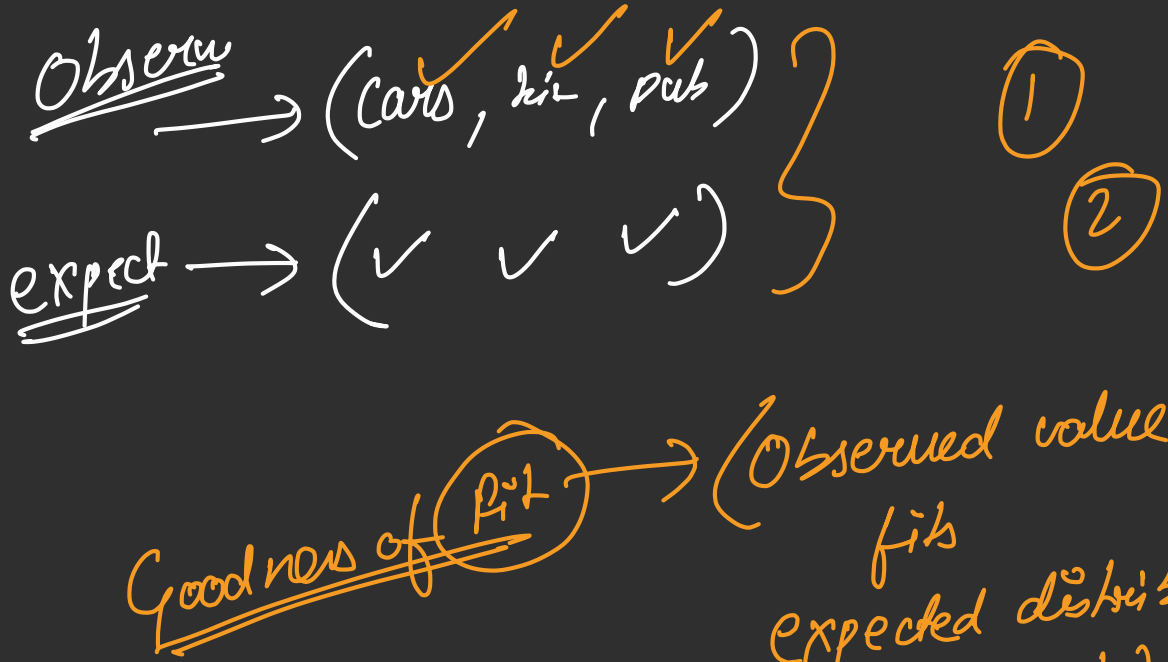
Question:

To assess whether a coin is fair or not, we need to compare the expected outcomes with the observed outcomes of tossing the coin.

The expected outcome for a fair coin toss is 50% heads and 50% tails, which corresponds to 25 heads and 25 tails in 50 tosses.

The observed outcomes from 50 coin tosses are 45 heads and 5 tails.

To determine whether the coin is fair, we perform a chi-square test to check if the observed results significantly deviate from the expected results. If the deviation is statistically significant, it may indicate that the coin is not fair.


Observed → (cars, bike, pub)
expect → (✓ ✓ ✓)
1
2
Goodness of fit → (Observed value fits expected distribution or not)

Test for Independence

Survey: Gender impacts online / offline purchases

↓

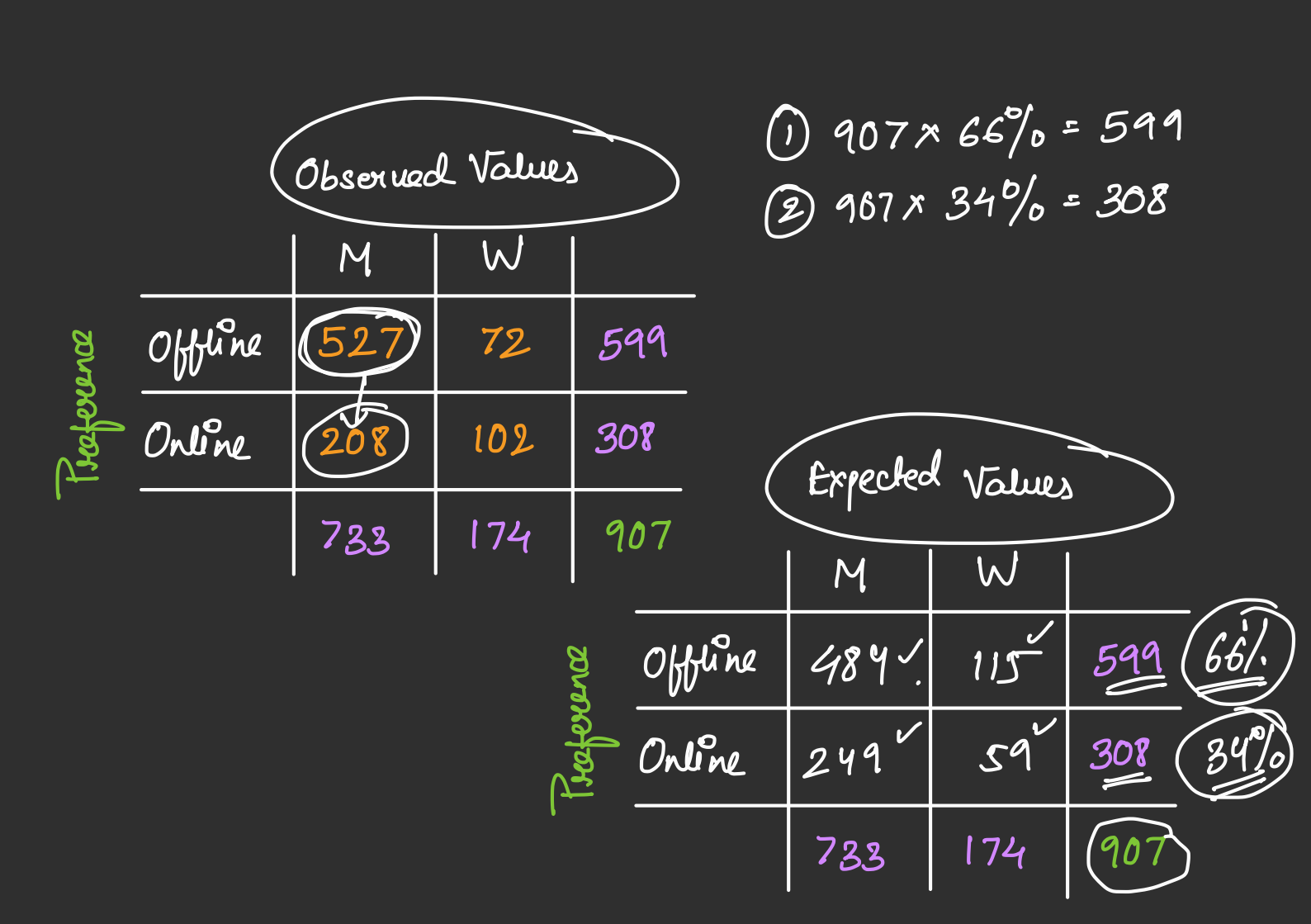
Preferences Vs Gender
(online/offline)

Observed Values

Preference	M	W	
Offline	527	72	599
Online	208	102	308
	733	174	907

H_0: Gender & preference are independent

H_a: Gender & preferences are dependent


Observed Values
M
W
Preference
Offline
527
72
599
Online
208
102
308
733
174
907
① 907 × 66% = 599
② 907 × 34% = 308
Expected Values
M
W
Preference
Offline
484
115
599
66%
Online
249
59
308
34%
733
174
907

$$\left\{ \chi^2 = \sum_K \frac{(O-E)^2}{E} \right\}$$

Quiz

	Online	print	TV
Buy			
Not Buy			

Test of Independence because it says, if there is any relationship b/w these categories.

Difference

① Coin toss → fit the expected distribution
(Goodness of fit) ✓
↳ when expected distribution is known

② preference Vs gender → Testing for independence
(Chi^2 contengency)

Assumptions

1. Variables are categorical
2. Observation are independent (Politician example)
3. Each cell is mutually exclusive (Only 1 choice example)
4. Expected value in each cell is