Chi-Square Test

1) Degree of freedom
2) Chi Square → Goodness of fit
   → Independence Test

---

① 2 Numerical Samples →

Num vs 2 Cat → Two Sample T Test

More than 2 Cat → Anova

② Cat vs Cat Comparisons →

⇒ IQ Score vs Schools (2 schools)

⇒ Gender vs Rating

⇒ Gender vs Preference online/offline

⇒ Is a coin fair?

⇒ Toss Multiple Times →

	H	T
# of	[50]	[50]

Counts → Category

⇒ Degree of freedom →

Q Lets say → we want to plan a movie week.
→ everyday we want to see a different genre.

7 genre → 7 days

(i) Choices for 1st night → 7
(ii) 2nd night → 6
(iii) 6th night → 2
(iv) 7th night → no choice

Why we need DOF → Chi-Square Test

⇒ 1) Salary of 3 individual

1st → 35 L
2nd → 36 L
3rd → ??

avg salary → 35 L

⇒ 2) 4 people →

35, 36, X, 37 → avg salary → 36

⇒ given you want to know n object, and you already know one aggregation value

How many objects should you already know to calculate all of them.

⇒ $n-1$ ← degree of freedom

Q

H	W
73	85
68	73
71	96
72	82
62	70
avg 69.2	81.2

SH, SW → 4H, 4W and avg | to calculate everything

→ 4+4

⇒ I need 8 values to predict remaining value.

minimum 8 values to predict →

minimum no. of variable needed to be known →

$$n_1 + n_2 - 2$$ ←

③ Sachin Century vs Victory

C \ Victory	Yes	No	
Yes			46
No			314
	184	176	360

Given margin aggregation are known

↓

only need 1 value

⇒ If both row & Column sums are known

↳ 1 value → dot = 1

⇒ ④ Elections → ↙ Politician

Region \	A	B	C	D	
X					349
Y					151
Z					150
	150	150	200	150	650

(# rows − 1) (# col − 1)

(3 − 1) (4 − 1)

2 × 3 = 6 values needs to be known.

⇒ Importance of DOF →

1) Chi-Square (Categorical) → Critical values
2) Based on DOF → distribution of data changes
3) Hypothesis Testing → find dist^n for H_0
   ↓
   Calculate p-value

If you have two arrays with lengths n1 and n2, what is the formula to calculate degrees of freedom for the chi-square test?

⇒ H = [ ] ← n1 length
W = [ ] ← n2 length.

$n_1 - 1$ $n_2 - 1$ ⇒ $(n_1 - 1) + (n_2 - 1)$.

⇒ [(x_1, y_1), (x_2, y_2) ... ( )]

$x_1 + y_1 =$ −

$x_2 + y_2 =$ −

⇒ Chi-Square Goodness of fit →

Q Is a Coin fair →

⇒ One categorical variable → 2 diff values.

⇒ Toss the Coin 50 times → 28 H, 22 T

⇒ ① H_0 → Coin is fair
H_a → Coin is unfair.

② If H_0 is true → 25 H, 25 T

H_0 is true ←

	H	T	
Exp	25	25	50
Actual	28	22	50

A

$dof = 1$

⇒ 500 tried

Exp	250	250	
Act	280	220	

B

Exp	250	250	
Act	253	247	

C

$$\chi^2 = \sum \frac{(Exp - Act)^2}{Act}$$

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

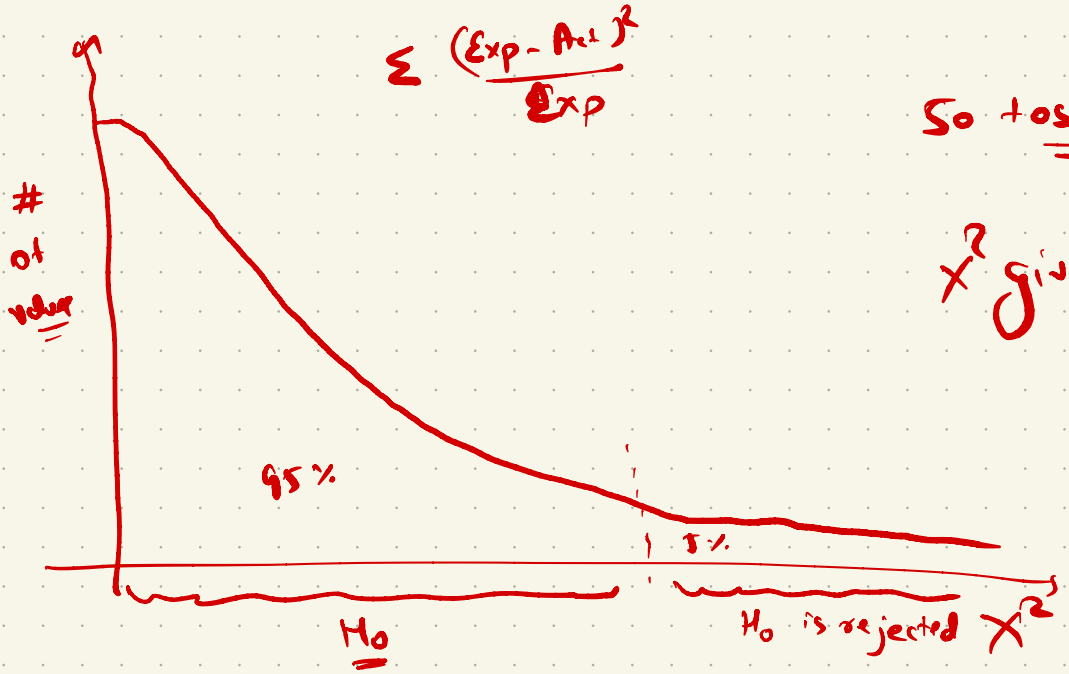
If I plot a graph of χ^2

If H_0 is true χ^2 → small ← H_0 is true

→ large ← H_0 is rejected

$\sum \frac{(Exp - Act)^2}{Exp}$

50 tosses



χ^2 given H_0 is true

χ^2 = Chi Statistic

Ls dist^n of χ^2 → Chi Square distribution

Based on $\alpha = 0.05$ → Critical χ^2 value

any test/experiment → if they have more extreme data → reject H_0.

$$\Rightarrow \chi^2 = \sum_{Exp} \frac{[Exp - Ac]^2}{Exp}$$

28 H
22 T

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

Actual ≡ observed

Exp ≡ Theoritical

$$\Rightarrow \chi^2 \propto \text{diff}$$

i.e. diff ↑ χ^2 ↑

Z score
T Stats
χ^2 Stats
} → Based on this we calculate p-value

Compare p-value against α.

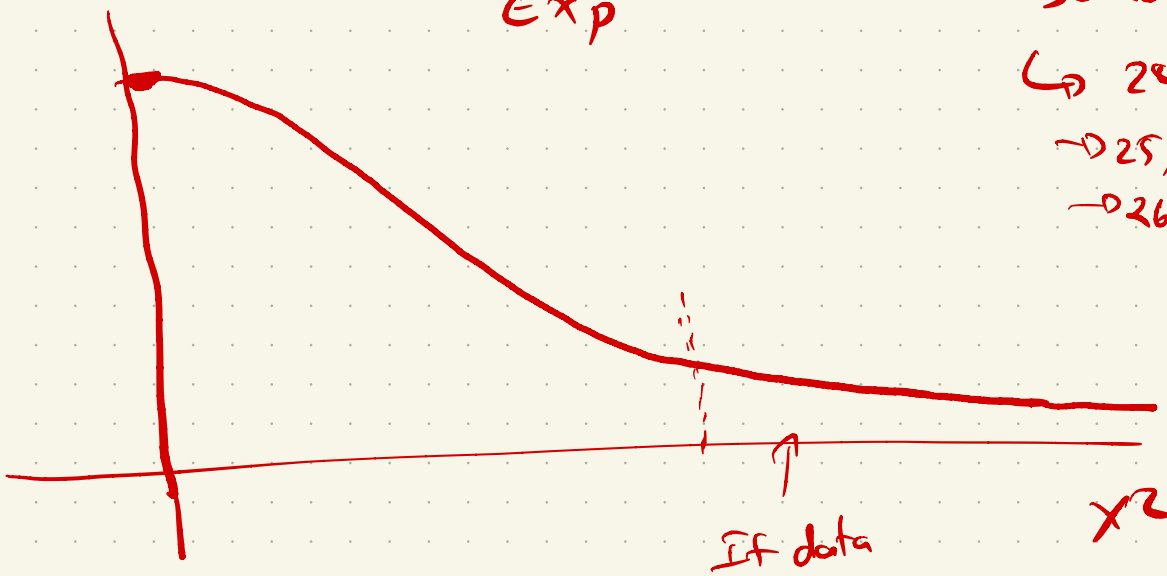
Z-Score → How many std dev away from mean the observed data is

χ^2 → How much diff^n from the expected data is the observed value.

p-value → 1 - norm.cdf (Z-score)

p-value → 1 - chi2.cdf (χ^2)

$$\Rightarrow \quad \chi^2 = \frac{\Sigma (Exp - Act)^2}{Exp}$$



50 coins

↳ 28,22 → –

→ 25,25 → ...

→ 26,24 → ...

If data lies in the extreme right → reject H_0

$\Rightarrow$

	H	T
Exp	25	25
Act	28	22

norm ↓ chi2

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

$$\chi^2 = \frac{18}{25} = 0.72$$

Corresponding p-value

$$1 - \text{chi2.cdf}(0.72, df = 1)$$

$$= p\text{-value}$$

A researcher is studying the preferences of people in a city for three different modes of transportation: car, bicycle, and public transit.
The researcher surveyed 500 individuals and found that 240 prefer cars, 160 prefer bicycles, and 100 prefer public transit.
The researcher wants to know if there is a significant difference between the observed preferences and the expected preferences based on historical data.
Which statistical test should the researcher use?

500

⇒	Obs	240	160	100
	Exp	200	200	100

Goodness of fit → How well does observed value fit the expected dist^{n}.

Good fit → Small χ^2 →

Chi-Square Test of Independence

⇒ 2 Categorial variable → Check if they are related.

Q = females are more likely to buy online clothes.

⇒ gender affects the preference of channel.

gender: M / F

channel: Offline / Online

Preference ↑	M	F	→ Gender
offline	527	72	599
online	206	102	308
	733	174	907

Observed data

⇒ H_0 → Gender has no relation with channel.

H_a →

⇒ 907 people surveyed

599 → offline

308 → online

733 M
174 F

$\Rightarrow \frac{599}{907} = 0.66$

$\Rightarrow \frac{308}{907} = 0.34$

→ first calculate population dist^n

M ⇒ 733 → 66% = 484
→ 34% = 115

F ⇒ 174 → 66% = 249
→ 34% = 59

Exp vdu

	M	W	
off	484	249	599
on	115	59	308
	733	174	907

Obs vdu

	M	W	
off	527	72	599
on	206	102	308
	733	174	907

Con
34%

$\Rightarrow$ 907 $\rightarrow$ 733 M
$\rightarrow$ 174 F

599 $\rightarrow$ $(599) \times \left(\frac{733}{907}\right) = 484$

$\rightarrow$

$$\chi^2 = \frac{(484-527)^2}{484} + \frac{(249-72)^2}{279} + \dots$$

$$1 - \text{Chi2.cdf}(\chi^2, df=1)$$

$$= \text{Pvalue}$$

A market researcher is exploring the connection between age group (under 25, 25-40, over 40) and smartphone brand preference (Brand A, Brand B, Brand C). The researcher collects data from 600 respondents and plans to perform a chi-square independence test.
How many degrees of freedom are associated with this test?

	A	B	C
<25			
25-40			
>40			

$= (\#\text{row} - 1)(\#\text{col} - 1)$

$= 2 \times 2$

$= 4$

⇒ Aerofit →

A marketing manager wants to determine if there is a relationship between the type of advertising (online, print, or TV) and the purchase decision (buy or not buy) of a product.
The manager collects data from 300 customers and records their advertising exposure and purchase decisions.
What statistical test should the manager use to analyze this data?

⇒ GOF → If we want to check how close to exp we are seeing

Indepence → Relationship b/w 2 Cat variable

⇒ Assumption for Chi-Square Test

1) Categorical variable
2) Each cell is mutually Exclusive ***
3) Observations are independent.
4) Each cell ≥ 5.

Collab Link : https://colab.research.google.com/drive/1unOC7zE0BH5VsYfOmQ-f-8ED-ZB1E2n4?usp=sharing

⇒ Rather than checking for actual data.

↓

we check dat of SM →

↳ Normal dist^{n}

What is the mean population.

⇒ So people Test → IQ

[100, 120, 130, ... 101, ...] So values.

So pts out of entire globe

⇒ So people → set 1 → avg of s_1 ←

So people → set 2 → avg of s_2 ←

1000s s_3 —

$\mu_1, \mu_2, \ldots \mu_{1000}$ → always follow normal dist^n

with overall μ = μ_{pop}

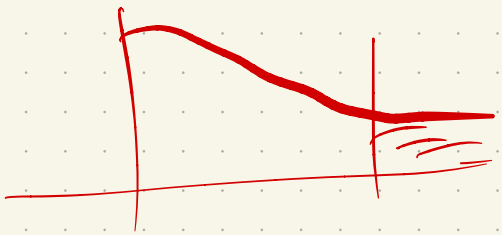
$\sigma = \sigma_{pop}/\sqrt{n}$

$\Rightarrow = \sum \frac{(Exp - obs)^2}{Ex_1}$

Exp → Counts

obs →

If H_0 true → Close to 2



⇒ chi Square

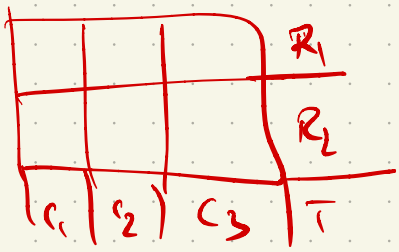
⇒ χ^2 ⟹ 1 − chi2.cdf(χ^2) = p_val

⇒ $\sum \frac{(Exp - Observed)^2}{exp}$ ✓ ( )

⇒ Chisquare → Just give array for Exp →
for obs →

⇒ Chi_Contigency →

Observed value + H_0 is that data is independ



⇒ Create a Expected table

↓

X^2

↓

get p-value

Collab : https://colab.research.google.com/drive/1unOC7zE0BH5VsYfOmQ-f-8ED-ZB1E2n4?usp=sharing