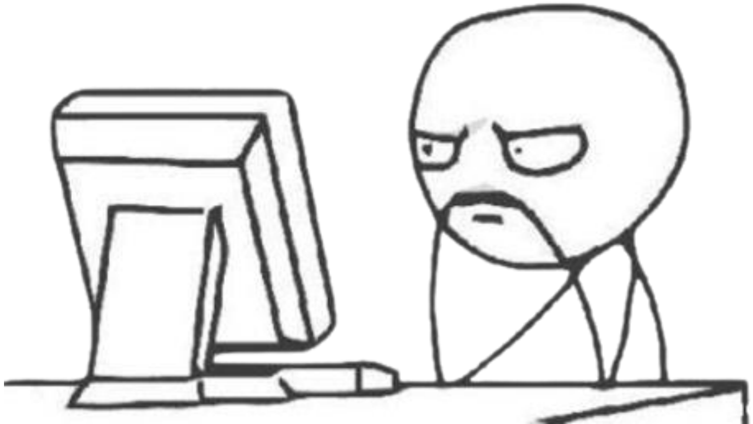
Hypothesis Testing

Chi-Square Test



Recap Test (Solve it and share your answers in the chat section only to me. Will discuss the solution at 07:08 AM)

Quiz: A group of 5 patients were treated with medicine A and another group of 7 patients with medicine B. Researchers claim that B is better than A in treating the disease. Test the claim at 5% significance level. (Note: In the below table, higher number is better)

A = [42, 39, 38, 60, 41]
B = [38, 42, 56, 64, 68, 69, 62]

(Two independent Sample T-Test)

$H_0: \mu_a \geq \mu_b$

$H_a: \mu_a < \mu_b$

① Sample size ($n < 30$) → T-Test

Sample mean with population mean

②

a) one sample T-Test: we are given with one group (sample) and a global/population mean. compare

b) Two independent sample T-Test: Two independent groups compare avg between the two groups

c) paired sample T-Test

→ one group but the data has been collected at two different time

	B·T	A·T
P-1		
P-2		
P-3		
⋮		

Z-Dist

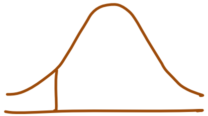
Z-Test
- one sample z-Test
- Two ind. sample z-Test

} average

- one sample z. prop. Test
- Two indp sample z. prop. Test

} proportion

Z-value



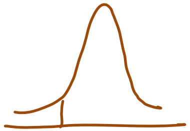
$N(\mu_p, \sigma/\sqrt{n})$

T-Dist

T-Test
- One sample T-Test
- Two indp sample T-Test
- Paired sample T-Test

} average

T-value



$N(\mu_p, \sigma/\sqrt{n}, dof)$

T-Dist is dependent on degree of freedom

1. Income between males and females
2. Test if pill improves the IQ level
3. Height of a group of people is less than 66 inches

(all target variables are continuous variable)

T- Test
Z- Test } continuous variable

chi-Square : Hypothesis testing on categorical data

① chi-square test of goodness of fit
② chi-square test of independence

chi-Square value

chi-Square dist

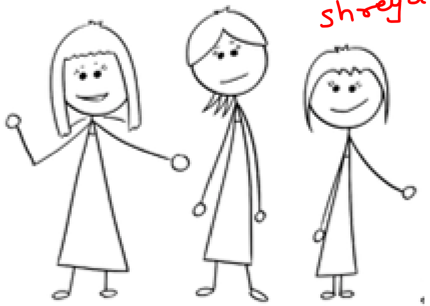
(dof)

Agenda:

- **Degree Of Freedom**
- **Chi-Square Motivations Using Toy Example**
- **Chi-Square Implementation In Business Case**

minimum info that we need to complete the data

Degree Of Freedom

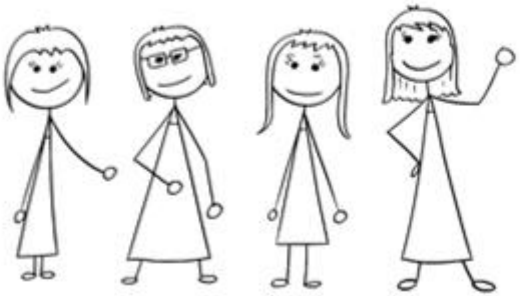


Sum = 5

$$[\underbrace{A,\ P}_{\text{these values are free to deviate}},\ S]$$

$\boxed{dof = (n-1)}$

when data is a linear series then degree of freedom $= (n-1)$



Total = 5

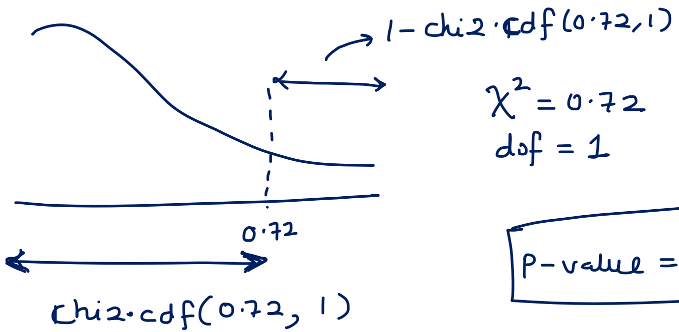
Ankita = 2
phani = 2
shreya =

Degree Of Freedom (Sample Standard-Deviation)

Dof

1. linear Series (1-Dimensional) = $(n-1)$
2. multi-dimensional (n-Dimensional) = $\sum_{i=1}^{d} (n-1)_i$
3. contengency table / cross-tab = $(r-1) \times (c-1)$

r = # of categories in rows
c = # of categories in columns



$\chi^2 = 0.72$
dof = 1

$$\boxed{\text{p-value} = 1 - \text{chi2.cdf}(\chi^2, \text{dof})}$$

Degree Of Freedom

S

S-1 Height	S-2 Weight
73	85
68	73
74	96
71	82
X	Y
AVG: 71	81.2

8

$$S\text{-}1 = (n-1)$$
$$S\text{-}2 = (n-1)$$

$$S = (n-1) + (n-1) = \boxed{\sum_{i=1}^{d} (n-1)}$$

$$(n-1) + (n-1) + (n-1) + \cdots + (n-1)_d$$

Age	income	height	weight

6

$$\underbrace{(n-1)}_{\text{age}} + \underbrace{(n-1)}_{\text{income}} + \underbrace{(n-1)}_{\text{height}} + \underbrace{(n-1)}_{\text{weight}}$$

$$(6-1) + (6-1) + (6-1) + (6-1)$$

$$\underbrace{5+5}_{10} + \underbrace{5+5}_{10}$$

$$10 + 10$$

20

Cross - tab

Degree Of Freedom

	India Win			
Sachin Century		False	True	Sum
	False	160	154	314
	True	16	30	46
	Sum	176	184	360

	India Win			
Sachin Century		False	True	Sum
	False	160	154	314
	True	16	30	46
	Sum	176	184	360

no. of categories in rows $\rightarrow r = 2$

$c = 2$

no. of categories in columns

Dof = 1

Contengency table / cross - tab

$Dof = (r-1) \times (c-1)$

$(2-1) \times (2-1)$

$= 1$

Degree Of Freedom

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

		Political Party				
Cities		A	B	C	D	Sum
	X	90	60	104	–	349
	Y	30	50	51	–	151
	Z	–	–	–	–	150
	Sum	150	150	200	150	650

$r = 3$
$c = 4$

$$dof = (r-1) \times (c-1) = 2 \times 3 = \underline{\underline{6}}$$

Chi-Square Test

1. Goodness Of Fit: If the data is uniformly distributed or not?
2. Test Of Independence: If two categorical variables are independent or not?

① Goodness of fit : we test if the data across various categories is uniformly distributed or not?

Toss of a fair coin (50)

	observed	Expected
Head	45	25
Tail	5	25

If this observed result is similar to the expected result or not?

we want to test if the observed data is uniformly dist. between the categories.

Test if the attendance across various stadiums is uniformly dist. or not or is there a particular stadium with higher or lower attendance than expected.

	observed attendance	expected attendance
Delhi	29K	25K
Chennai	30K	25K
Bangalore	45K	25K
Kolkata	12K	25K

categorical

(observed = expected)

Coin Toss

case-1

we fail to reject null

categorical

	Head	Tails
Expected	25	25
Actual	28	22

case-2

we reject null

	Head	Tails
Expected	25	25
Actual	45	05

H_0: Expected outcome = observed outcome

H_a: Expected outcome $\neq$ observed outcome

$$\chi^2 = \sum_{i=1}^{n} \frac{(O-E)^2}{E}$$

chi-square value / Statistic

Difference between actual and expected

Deviation of observed data from expected data

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

$$\chi^2 = 0.72$$

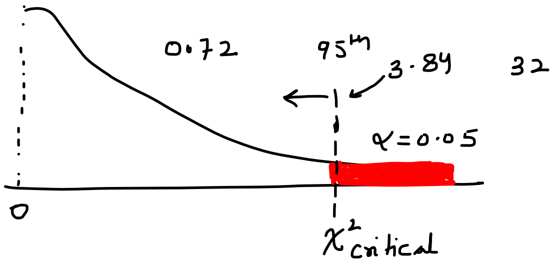
$$\chi^2 = [0, \infty]$$

$$\chi^2 = \left[\frac{(45-25)^2}{25}\right] + \left[\frac{(05-25)^2}{25}\right]$$

$$\chi^2 = 32$$

$$\chi^2_{critical} = (dof, 1-\alpha)$$

$$= (dof = 1, 1-\alpha = 0.95)$$



Example

observed

	Actual	Expected
Chennai	50	50
Bangalore	60	50
Delhi	40	50
Punjab	47	50
Kolkata	53	50
Total	250	250

Attendance at various stadiums hosting IPL is reported as given in the below table. Is the attendance uniform across all stadiums or there was stadium with significantly high attendance? $\alpha = 0.05$

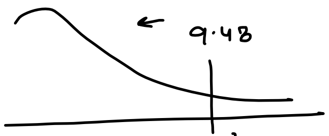
H_0 : Actual attendance = Expected attendance

H_a : Actual attendance $\neq$ expected attendance

$$\chi^2 = \sum_{i=1}^{n} \frac{(o-e)^2}{e}$$

$$= \frac{(50-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(47-50)^2}{50} + \frac{(53-50)^2}{50}$$

$$= \chi^2 = 4.36$$



we fail to reject null

(No such stadium where the attendance was signi higher or lower)

total attendance = 250

no. of stadiums = 5

what would be the uniform attendance = 250/5

assumption → all stadiums would receive the same attendance with no bias or preferences

chi-Square Test of independence (Two-categorical var.)

Objective → To test if two categorical feature are dependent or independent on each other?

Prod. cat	gender
A	M
A	F
B	M
A	M
C	F
B	F
A	M

Each sales happening is independent of the other

(A, B, C)

(m, F)

	m	f	
A	3	1	4
B	1	1	2
C	0	1	1
	4	3	7

(observed contengency table)

(Expected frequencies)

	m	f	
A	= 2	2	4
B	1	1	2
C	1	0	1
	$\boxed{4}$	3	7

$\left(\frac{\text{Row total}}{\text{Table total}}\right)$

colum total

4
- 0.57 × 4 → A ≃ 2.28 ≃ 2
- 0.28 × 4 → B = 1.12 ≃ 1
- 0.14 × 4 → C = 0.56 ≃ 1

3
- 0.57 × 3 → A = 1.74 ≃ 2
- 0.28 × 3 → B = 0.84 ≃ 1
- 0.14 × 3 → C = 0.42 ≃ 0

$\left(\frac{\text{Row total} \times \text{(column total)}}{\text{Table total}}\right)$

$= \frac{4 \times 4}{7}$

If we randomly select a customer what's the prob that customer will buy product A without looking & their gender =

$P(A) = 4/7 = 0.57$
$P(B) = 2/7 = 0.28$
$P(C) = 1/7 = 0.14$

irrespective of gender

Gender Impact on Offline/Online Purchase
Is the "Type of Purchase" influenced by "Gender"

$$\chi^2 = \frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + \frac{(206-242)^2}{242} + \frac{(102-57)^2}{57}$$

$\chi^2 =$

$dof = (r-1)\times(c-1)$
$= (2-1)\times(2-1)$
$= 1$
$=$

observed

	Gender			
Type Of Purchase		Male	Female	Sum
	Offline	527	72	599
	Online	206	102	308
	Sum	733	174	907

a customer can fall in one cell with no common elements

$P(\text{offline}) = 599/907 = 0.66$

$P(\text{online}) = 308/907 = 0.33$

expected

	Gender			
Type Of Purchase		Male	Female	Sum
	Offline	484	115	599
	Online	242	57	308
	Sum	733	174	907

expected values 75

733 → offline $= 0.66 \times 733 = 483.78 \simeq 484$

733 → online $= 0.33 \times 733 = 241.89 \simeq 242$

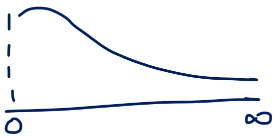
174 → offline $= 0.66 \times 174 = 114.84 \simeq 115$

174 → online $= 0.33 \times 174 = 57.42 \simeq 57$

Assumptions of Ch-2 Test

1. variables are categorical
2. observation are independent
3. each cell is mutually exclusive
4. Expected value in each cell > 5

chi-2 distribution



dependent on dof

1. 1-Dimension
   $dof = (n-1)$
2. n-Dimension
   $dof = \sum_{i=1}^{d} (n-1)$
3. cross-tab
   $dof = (r-1) \times (c-1)$

chi-Square test

1. chi-square goodness of fit → if the data across various categories is uniformly dist. or not?
2. chi-square test of independence → if two categorical variables are dependent or independent

functions

1. chi2
   - critical chi-square = chi2.ppf (dof, 1-α)
   - p-value = (1 - chi2.cdf (x, dof))
2. chisquare → (o, e)
3. chi2-contingency

$$\chi^2 = \sum_{i=1}^{n} \frac{(o-e)^2}{e}$$

→ deviation of the observed data from the expected data

χ^2 → chi-square statistics

Quiz:

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

   $dof = (n_1 - 1) + (n_2 - 1)$

1. In a chi-square test, what does the chi-statistic represent? → Diff. between observed and expected

1. 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? goodness of fit

car	240
bk	160
	100

   3

1. 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?

   $dof = (r-1) \times (c-1) = (2) \times (2) = 4$

1. 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?

   Test of independence

1. When testing the independence of two categorical variables, what are the assumptions of the chi-square test?