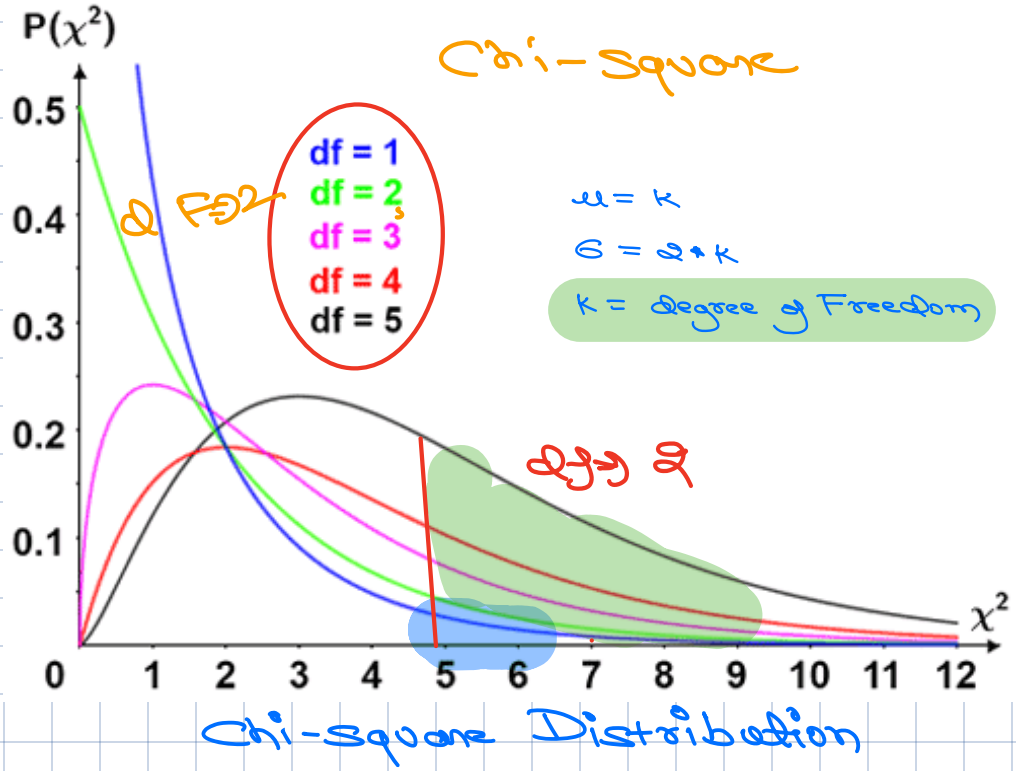
* Post Read



1. **Definition**: The chi-square distribution is defined as the distribution of a sum of the squares of k independent standard normal random variables (where k is the degrees of freedom).

2. **Degrees of Freedom**: The shape of the chi-square distribution depends on the degrees of freedom (df). As the degrees of freedom increase, the distribution becomes more symmetric and resembles a normal distribution.

3. **Properties**:
   - It is positively skewed, especially for lower degrees of freedom.
   - Its mean is equal to the degrees of freedom (k).
   - Its variance is $2k$.
   - The distribution is used only with positive values, as it represents squared values.

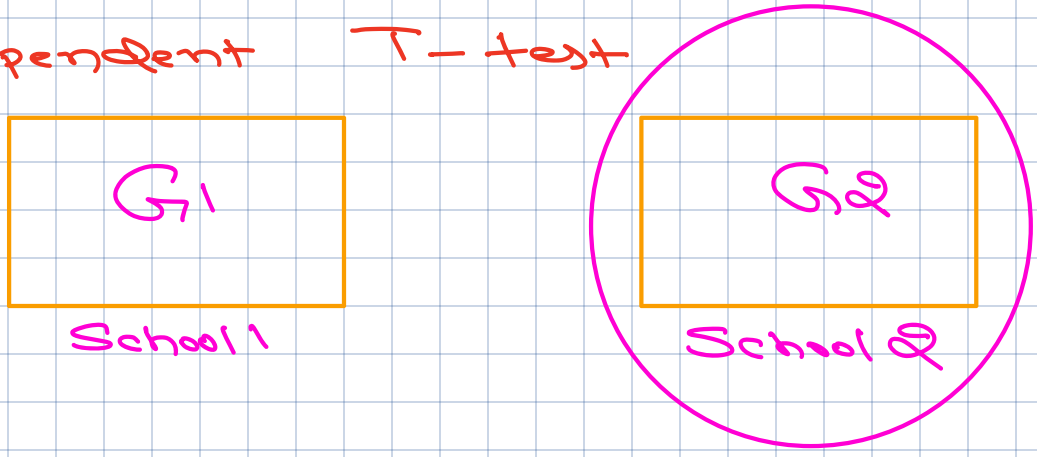
4. **Applications**:
   - **Goodness-of-fit tests**: To determine if an observed distribution matches an expected one.
   - **Test for independence**: In contingency tables, to see if two categorical variables are related.
   - **Variance estimation**: To test hypotheses about the variance of a normally distributed population.

Agenda

1. Paired T-test
2. DOF
3. Chi-square Test
   - Goodness of Fit
   - Test for independence
4. Assumption of Chi-square Test

Paired T-test

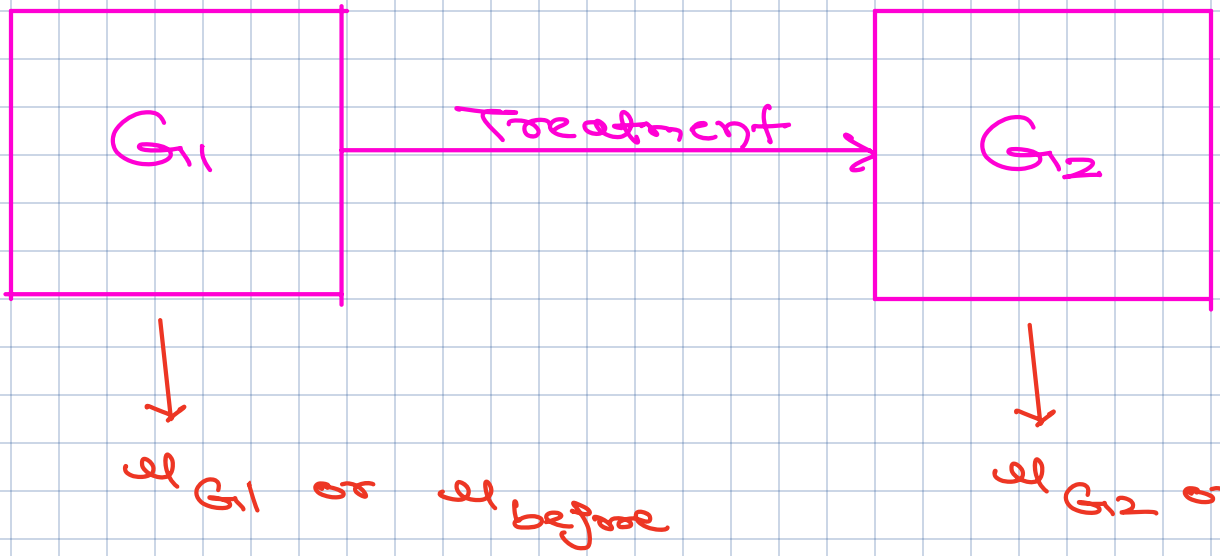
Independent T-test



H_0 : $\mu_{IQ\_G1} = \mu_{IQ\_G2}$

H_a : $\mu_{IQ\_G1} \neq \mu_{IQ\_G2}$

Paired T-test



μ_{G_1} or μ_{before} μ_{G_2} or μ_{after}

Diff = $\mu_{after} - \mu_{before}$

H_0 : $\mu_{before} = \mu_{After}$

H_a : $\mu_{before} \neq \mu_{After}$

Claim: Problem Solving Session helps student get better Marks

```
df_ps.head()
```

	id	test_1	test_2
0	0	40	38
1	1	49	44
2	2	65	69
3	3	59	63
4	4	44	43

before PS | After PS

H_0 : $\mu_{score\_t1} = \mu_{score\_t2}$

H_a : $\mu_{score\_t1} \neq \mu_{score\_t2}$ (Two tailed)

H_a : $\mu_{t1} < \mu_{t2}$

Recap

Setup for Relationship of variables :

1. Numerical vs Categorical (2 Categories)
   Z-test, t-test, proptions z-test
2. Numerical vs Categorical (> 2 Categories)
   Anova test (In upcoming Sessions)
3. Numerical vs Numerical
   Correlation (In upcoming Sessions)
4. Categorical vs Categorical
   Chi-Square Test

Chi-Square Test

Degree of Freedom

- Number of independent values that are free to vary when calculating a statistic
- It tells us about flexibility or choices we have when trying to make statistical Decision

Ex-1: 7 genre of movies for 7 weeks

DOF = ?

Week 1 : 7 choices

Week 2 : 6 choices

Week 3 : 5 choices

Week 4 : 4 choices

Week 5 : 3 choices

Week 6 : 2 choices

Week 7 : No choices → only 1 genre available

DF = 6

Ex2 : Numerical Data

Empid	Salary
1	36
2	35
3	?

mean = 35

DoF = 2

→ 34

Ex2 : Categorical Data

Empid	Gender
1	M
2	F
,,	,,
,,	,,

→

Empid	M	F
1	1	0
2	0	1
,,	,,	,,
,,	,,	,,

DOF ⇒ 1

Do we need both columns ⇒ ?

3 Categories → n − 1 ⇒ 3 − 1 ⇒ 2

HW:

Empid	Salary
1	36
2	35
3	?
4	30

mean = 37

DOF ⇒ 4 − 1 ⇒ 3

3rd Value ?

General Rule of DOF:

$\bar{x}$ = number

- Given a set of N numbers
- If you know average of n-numbers, what is minimum number of Data-points you need to know to Extract Full set

DOF = n-1

for 1-D array with length n

Ex-3

H	W
73	85
68	73
74	96
71	82
62	70
avg 71	81.2

n_1

n

$DoF = n_1 - 1 + n_2 - 1$

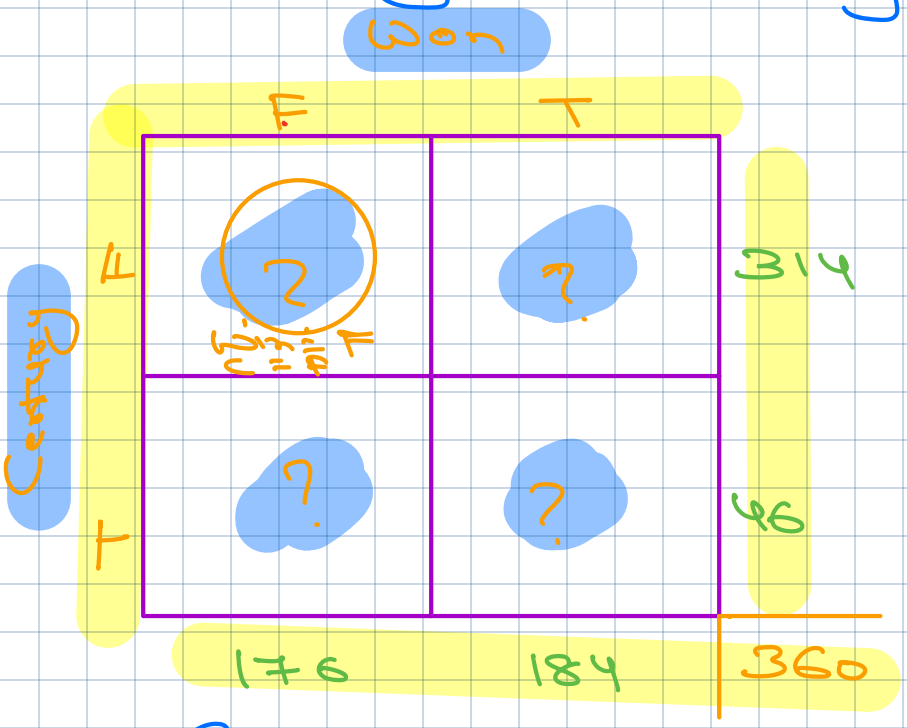
$= n_1 + n_2 - 2$

$= 2n - 2$

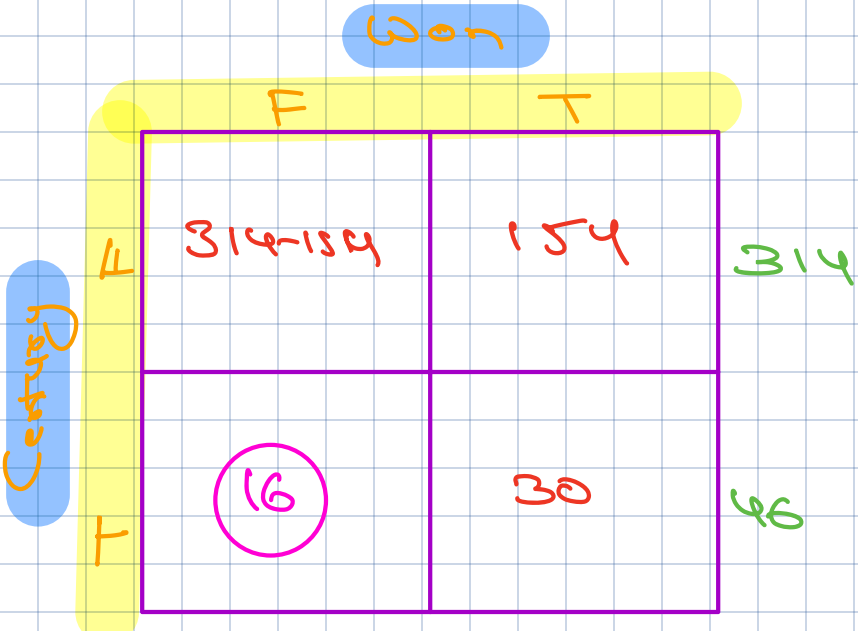
2 arrays with length n_1 and n_2 and means $\bar{x}_1$ and $\bar{x}_2$

Ex-4 **2D Contingency Table**

Sachin century and Victory



Contigency Table



DoF = ?

DoF = 1

$(2-1) * (2-1) = 1$

ind-1 ind-2

DoF $(n_1 - 1 + n_2 - 1)$

Ex-4: Given data of Regional Support for 4 Politicians A, B, C, D from survey of 3 cities X, Y, Z.

	A	B	C	D	Total
X	90	60	104	?	349
Y	30	42	51	?	151
Z	?	?	?	?	150
Total	150	150	200	150	650

rows = 3

cols = 4

DoF 6

(# row − 1) × # (cols − 1)

(3 − 1) × (4 − 1) ⇒ 2 × 3 = 6

Why D.F?

① DoF is crucial for chi-Square tests as it is important parameter in chi-square distribution i.e. Dist of Test Statistic under Null Hypothesis

② It controls Variability in data.

③ Higher DoF leads to Higher critical Values requiring large test statistics to reject Null Hypothesis for given (α).

Chi-Square Goodness of Fit Test

① Given a categorical Variable, you can test if there is any significant Difference b/w observed Values and Expected Values

Example: Coin Toss : 50 trials → 28 H, 22 T

H_0 : Fair Coin

H_1 : Biased Coin

	H	T
Observed	28	22
Expected	25	25

DOF ⇒ 1

For H_0 → Expected Heads ⇒ 25
→ Expected Tails ⇒ 25

Observation:

We can observe some deviation in Actual vs Expected

$$\frac{(O(H) - E(H))^2}{E(H)} + \frac{(O(T) - E(T))^2}{E(T)}$$

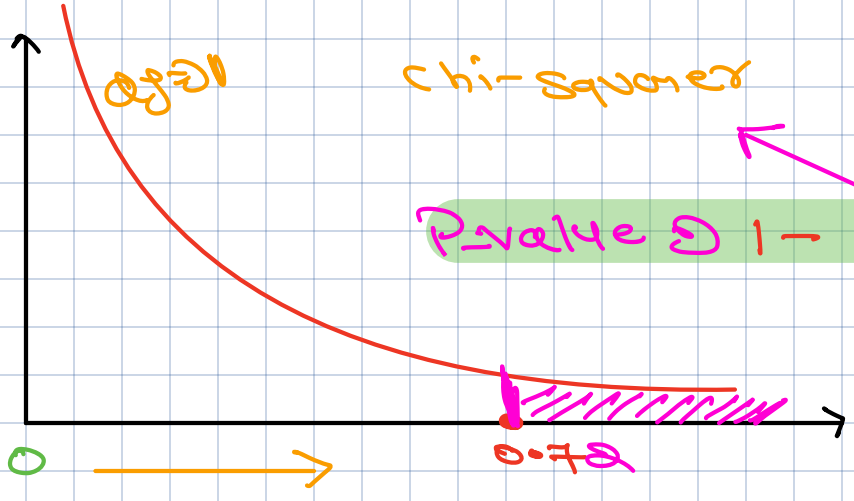
Diff b/w Heads Diff b/w Tails

$$\frac{(22-25)^2}{25} + \frac{(28-25)^2}{25} = 18/25$$

Chi-square Statistic = 18/25 = 0.72

DOF = 1

What would chi-square statistic look like when plotted under H_0 ?

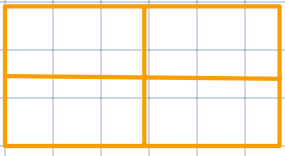


Chi-2 stat = χ^2

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

P-value = 1 − chi.cdf(chi2_stat, dof)

Q Can r change in Between?



No and DOF can't change b/w Experiment

Q How do we conclude the result i.e. Coin Biased or Not?

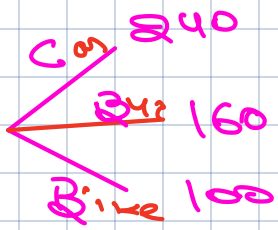
α = 0.05

If P-value < α :

Reject H_0

* To Do N.3

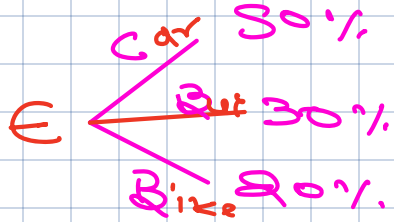
Python Example for Coin toss



Step 1 : Setup Hypothesis

H_0 :

H_a :



Step 2 : Pick distribution and α

Step 3 : Perform Test and calculate test statistics

Step 4 : Compare p-value and reject / fail to reject Null Hypothesis

Chi-Square Test of Independence

- Given two variables you want to test if they are related or independent

Marketing Firm Example

- Imagine you are running a marketing campaign.
- There are two modes of shopping
  1) Offline
  2) Online

Goal: Increase Online sales by running Targetted Campaign/ads

Consultancy Claim:

- Focus on Women
- Females have higher chance of shopping Online than males

In other words

There is relationship between 'Buying Pattern and Gender'

Buying Pattern: Offline, Online

Gender: M, F

Can we test and Evaluate this?

Gender vs Purchase

	M	W	
Offline	587	72	599
Online	206	102	308
	733	174	907

H_0 : Gender and preference are independent

H_a : Gender and preference are dependent

① Total % of population that prefers

① Offline → $\frac{599}{907}$ → 66%

② Online → $\frac{308}{907}$ → 34%

If gender has no impact

① Prefer offline

Expect men % → 66% → 66% 733
→ 484

Expect Women % → 66% → 66% 174
→ 115

② Prefer Online

Expect men % → 34% → 249

Expect Women % → 34% → 59

Chi-2 ~~stat~~ →

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

P-value → 1 − chi.cdf(chi2_stat, dof)

Chi-2 ~~stat~~ X^2 ⇒

$$\frac{(527-484)^2}{484} + \frac{(78-115)^2}{115} + \frac{(206-249)^2}{249} + \frac{(102-59)^2}{59}$$

* P_value ⇒ $1 - \text{Chi2. ~~stat~~}(X^2_{,1})$

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.

Var2 → advertising

Var1 → purchase decision

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?

Assumptions of Chi2 test

1. Random Sample
    - The data points for each group in your analysis must have come from a simple random sample.
    - This is important because if your groups were not randomly determined then your analysis will be incorrect.
    - In statistical terms this is called bias, or a tendency to have incorrect results because of bad data.
2. Variables are categorical
3. Mutually Exclusive Groups
    - The two groups of your categorical variable should be mutually exclusive.
    - For example, if your categorical variable is hungry (yes/no), then your groups are mutually exclusive, because one person cannot belong to both groups at once.
4. Observations are independent
    - Each of your observations (data points) should be independent.
    - This means that each value of your variables doesn't "depend" on any of the others.
    - For example, this assumption is usually violated when there are multiple data points over time from the same unit of observation (e.g. subject/customer/store).
    - Because the data points from the same unit of observation are likely to be related or affect one another.
    - A different test must be used if the researcher's data consists of paired samples, such as in studies in which a parent is paired with his or her child.
5. The value of the cell expected should be **5 or more** in at least 80% of the cells, and no cell should have an expected of less than one