Chi-square distribution :-

* A chisquare (χ^2) is a statistic that checks for patterns or relationships in categorical Variables.

* Chi-squared distribution, denoted as χ^2 is related to the standard normal distribution because the chisquare dist is derived from the std normal distribution.

* Lets say Z denotes the standard normal distribution and $Z_1, Z_2 \dots Z_n$ is the independent random Variables drawn from Z which will have the normal distribution.

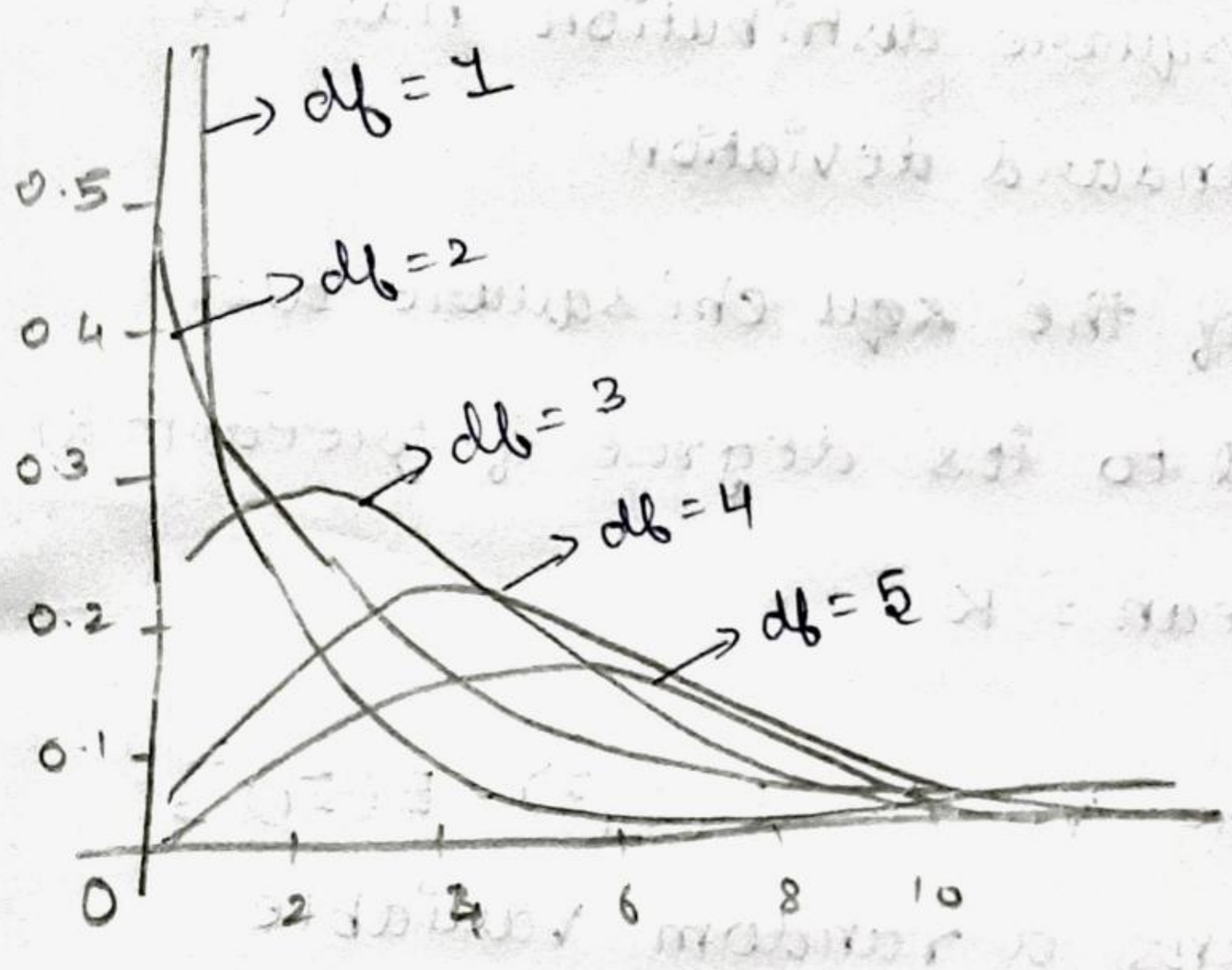
* Those random variables have mean = 0 and Std.dev = 1, which means they were std. normal dist.

* So, the sum of the squares of the random Variables say $(Z_1^2 + Z_2^2 \dots Z_n^2)$ follows the chisquare distribution.

* Those random Variables can be considered as the samples drawn from the main Z.

* The chisquare dist always assumes the Positive Values, since it is squared.

* The probabibility density function (PDF) of the chisquare dist will change if the degree of freedom changes.



* If the df increases the chisquare distribution approximates normal distributions.

* Basically above statements proves the central limit theorem, which is the sampling distribution approximates norm. dist.

* There can be a doubt that what is this random Vaniable means and how they are interpreted, this ~~can~~ will be explained in the chisquare tests.

* So the distribution can be generalized as

$$Q = \sum_{i=1}^{K} z_i^2 \sim \chi_K^2$$

where K is degree of freedom.

* The chi square distribution has its mean and standard deviation.

* The mean of the chi square dist. will be equal to its degree of freedom(K)

$$\text{mean} = K$$

* It is because $V(Z_1) = E(Z_1^2) - E(Z_1)^2$,

Here Z_1 means a random Variable.

$E(Z_1) = 0$ because its std. normal dist. But $E(Z_1^2)$ is a chisquare dist., so it will be equal to K.

* Its Variance is 2K.

* It is because, again

$$V(A) = E(A^2) - [E(A)]^2$$

$$V(\chi_1^2) = V(Z_1^2) = E(Z_1^4) - [E(Z_1^2)]^2$$

* $E(z,4)$ denotes the kurtosis, we all know that the chi square dist deals with the normal distribution. The kurtosis value for the normal distribution is 3. means it is the mesokurtic.

So, $V(\chi_1^2) = 3 - [E(z_1^2)]^2$

$= 3 - 1$

$= 2$

$$V(\chi_K^2) = V\left(\sum_{i=1}^{K} z_i^2\right) = V(z_1^2) + V(z_2^2) + \cdots + V(z_K)^2$$

$= 2 + 2 + 2 - \cdots$

$= 2K$

* So, mean $= K$, Variance $= 2K$.

Chi squared tests:-

* A chi square test (χ^2) is basically a data analysis on the basis of observations of a random set of variables

* It has 2 tests

  i) Test for independence

  ii) Goodness of fit.

Test for independence:-

* chisquare test for independence can be used to determine if there is an association between two categorical Variables.

* Basically it means whether 2 categorical features are independent or dependent on each other.

* The test statistic will be

$$\chi^2 = \sum \frac{(\text{Observed data} - \text{Expected data})^2}{\text{Expected data}}$$

* The observed data is the which we have Observed.

* The expected data or frequencies which States that the Variables are independent.

& which means our expectation over the theory that variables are independent.

Problem statement :-

120 people are surveyed for their preferred social media platform. Is there enough evidence to suggest social media preference is independent of gender?

H_0: Socialmedia preference independent of gender

H_1: Social medial Pref not independ of gender

Observed frequencies :-

	Male	Female	Total	
Facebook	15	20	35	marginal
Instagram	30	35	65	
Tik tok	5	15	20	
Total	50	70	120	

marginal

Contigency table

Now we have to calculate the expected frequency expecting the data to be independent.

In probability, if 2 random variable were independent

$$P(A \cap B) = P(A) \times P(B)$$

So here

$$E\left(\begin{matrix}\text{Joint}\\ \text{Probability}\end{matrix}\right) = \underset{\text{Prob}}{\text{Marginal}} \times \underset{\text{Prob}}{\text{Marginal}}$$

Expected frequencies:-

* Total or marginal values remains same as observed, because if the total changed then there is no meaning.

* we will find expected value for Joint Values.

$$E(\text{Male} \cap \text{Facebook}) = \frac{35 \times 50}{120} = 14.6$$

$$E(\text{Female} \cap \text{Facebook}) = \frac{35 \times 70}{120} = 20.4$$

* Like wise do for instagram & TikTok.

then the final contigency table for the expected frequency will be

* the joint values like Male ∩ Facebook. are the random variables. here in ~~indep~~ Test for independence.

	Male	Female	Total
Facebook	14·6	20·4	35
Instagram	27·1	37·9	65
Tiktok	8·3	11·7	20
Total	50	70	120

* The expected frequency in joint should always be greater than 5.

* It is because less than 5 frequencies make p-value inaccurate.

* If it occurs for example there are 3 categories young, middle, old. In this 3 for any 2 category, its coming less than 5 then we should merge middle, old and make as not young, by simply adding the Values.

* If frequency increase, it approximate Central limit theorem

Test statistic :-

$$\chi^2 = \sum \frac{(O-e)^2}{e} \sim \chi^2_2$$

$$df = (r-1)(c-1)$$

r = row
c = col

It means if i find male in insta & male in facebook, i can find remaining values just by subtracting with total so df = 2

$$\chi^2 = \frac{(15-14.6)^2}{14.6} + \frac{(30-27.1)^2}{27.1} + \frac{(5-8.3)^2}{8.3}$$

$$+ \frac{(20-20.4)^2}{20.4} + \frac{(35-37.9)^2}{37.9} + \frac{(15-11.7)^2}{11.7}$$

$$= 2.84$$

$$\boxed{\chi^2 = 2.84}$$

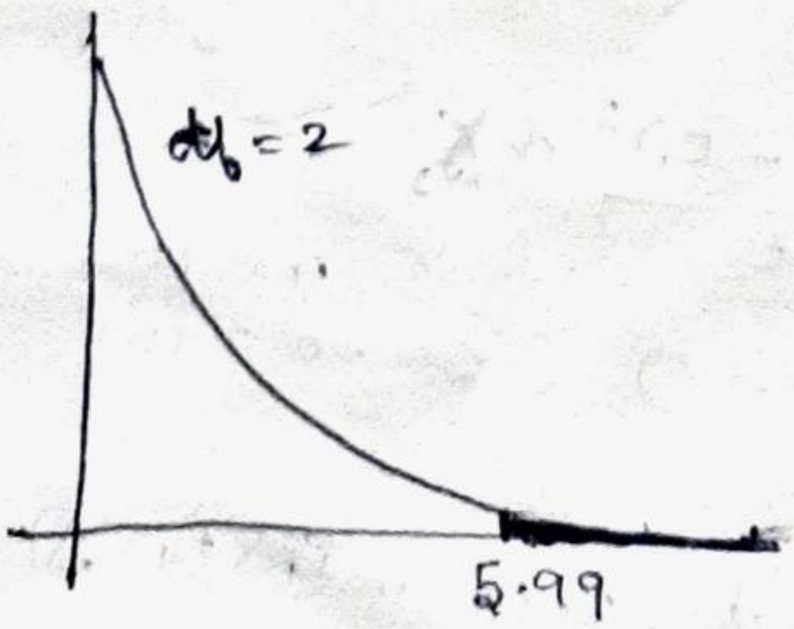
Decision rule

$\alpha = 0.05$

* Since Chi square distribution with degree of freedom, so we want to find area under right side of ~~curve~~ decision rule value using chi square table for $(0.05, 2) = 5.99$

(α ↓, df ↙)

* Reject the H_0 if $\chi^2 > 5.99$



* So 2.84 < 5.99

* So we dont have enough evidence to reject null hypothesis

* So Social media preference is independent of gender at 5-1. significance.

Goodness of fit :-

* The chi square goodness of fit is a statistical hypothesis test often used to evaluate whether sample data is the representative of the full population.

* As already said it is meant only for categorical Variables.

* In other words it can be said as how close the observed values to the expected value.

* For example i expect a population proportion to visit the shop ~~and~~, it is the expected value. And there will be observed values means how many people actually visited the shop.

* So goodness of fit shows how ~~the~~ good the observed data fits expected data.

Problem statement :-

1) In an art class of 75 students, 11 are left handed. Does this class fit the prevailing theory that 12·1· of people are left handed?

Population Proportion

H_0 : 12·1· of people are left handed (or) $\pi = 12$

H_1 : Prevailing theory is incorrect (or) $\pi \neq 12$

$\alpha = 0.05$ or 5·1·

	Observed	Expected (given) $q = (12\cdot1\cdot \times 75)$
Left handed	11	9
Right handed	64	66
Total	75	75

* The questions asks whether can we expect 12·1· of population is left handed.

Test statistic :-

$df = 1$ (because there is only 2 categories (left or right))

$$\chi_1^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \frac{(11-9)^2}{9} + \frac{(64-66)^2}{9}$$

$= 0.505$

$\chi_1^2 = 0.505$

decision rule:-

according to $\alpha = 0.05$

for ~~0.05 or~~ $P = 0.05$ & $df = 1$ the chi square value is 3.841

* So if $\chi_1^2 > 3.841$ then reject Ho

∴ Here $0.505 < 3.841$, So we didn't have enough evidence to reject the null hypothesis.

Result

The proportion of left handedness is 12.1.

2) Out of 600 throws in a scisor-Paper-Rock Competion there were 235 rocks, 194 scisors and 171 paperes thrown.

Is there evidence of a "weapon" Preference.

Ans:-

The question basically means whether ~~do~~ acc to question do we prefer 'on weapon rather than other.

$H_0: \pi_{Rock} = \pi_{Paper} = \pi_{Scissor}$

H_1: At least one should ~~not~~ differ $\pi_i \neq \pi_j$

	Obs	Exp
Rock	235	200
Paper:	194	200
Scissor	171	200
	600	600

$\alpha = 0.05$

Test statistic :-

$df = 2$

$$\chi_2^2 = \sum_{i=1}^{3} \frac{(O_i - e_i)^2}{e_i} \sim \chi_2^2$$

* Here we can have a doubt that why the expected values are same for all 3

∀ It is so because the H_0 is $\pi_{Rock} = \pi_{Paper} = \pi_{Scissor}$

$$\chi_2^2 = \frac{(35)^2}{200} + \frac{(6)^2}{200} + \frac{(29)^2}{200}$$

$$\chi_2^2 = 10.51$$

chi square critical value for $\alpha = 0.05$ and $df = 2$.

$\chi^2_{crit} = 5.99$

* If $\chi^2_2 > 5.99$ reject the Ho

* Here $10.51 > 5.99$

* So reject the Ho

Result :-

* So we have enough evidence to reject Ho

* And we can say there is weapon Preference as we all known more will people will prefer stone to gain points.