Chi-Square

- A chi-square (χ^2) is a statistic that checks for patterns or relationships in a categorical variables.
- A categorical variable is a non-numeric characteristics like - Gender, Language spoken
- What kind of question can a chi-square can answer?

One way chi-Square - One variable → Is a six-sided die fair?

	1	2	3	4	5	6
Rolls	20	23	21	25	29	32

More than one variable → Is participating in a study group related to passing a exam?

$\chi^2 = 6.0, p < 0.5$, indicates significant relationship between groups.
$\chi^2 = 6.0, p > 0.5$ indicate, insignificant relationship between groups.

	Failed Exam	Passed Exam.
Not Study in Group	12	42
In Study Group	6	38

Q) Why not just look at the number count?

→ We can't be sure if it is reliable or created by chance. If we do again can we get same exact numbers. Use Inferential statistics and use chi-square.

Example - Does gender vary across educational majors?

	Engineering	Business	Psychology
Female	2	2	3
Male	3	2	2

If there is no relationship, we would expect gender to be evenly spread across majors.

If there is a relationship, we would expect gender to be unevenly spread across majors.

If there is No relationship. Tables are known as contigency table.

Even relationship Contingency table.

	Engineering	Business	Psychology
Female	50	50	50
Male	50	50	50

Uneven Relationship Contigency table.

	Engineering	Business	Psychology.
Female	40	55	55
Male	60	45	45.

$$\chi^2 = \frac{(\text{Expected} - \text{Observed})^2}{\text{Expected}}$$ ⇒ Calculate each group

i) Engg - Female $= \frac{(50-40)^2}{50} = \frac{100}{50} = 2$.

ii) Engg - Male $= \frac{(50-60)^2}{50} = \frac{100}{50} = 2$

So we will have 6 group.

$\chi^2 = 2+2+0.5+0.5+0.5+0.5$

$\chi^2 = 6$.

So, is 6 is a real difference? will it hold next time if new set of data comes?

- So, each chi-square has p values. The p-value tells us the likelihood that there is real difference.

If $p = 0.05$, 5% chance we would get these result with random data.

Accepted range normally is, $p \leq 0.5$ or else reject

It is square in formula because, we magnify effect of bigger differen and minimizes effect of smaller differences.

Sample size - Each group should have minimum count of 5.

Types of Chi-Square - i) Test of Independence ii) Goodness of fit.

i) Test of Independence - Test for a relationship between two categorical variables.
Eg - Testing relationship between gender & major.
It say any variable has a influence on other categorical variable.

ii) Goodness of fit - Compares categorical values in your sample to a known or hypothesized value. Eg - Comparing gender & major at your university to the gender & major distribution nationwide.

Limitations & Assumptions - i) It don't need normal distribution ii) Atleast one value in each cell of contigency table.
iii) It doesn't tell which levels of variable are driving the effect.
iv) All data should be independent. One variable should not affect other variable