CHI-SQUARED TEST

Data Science | CCDATSCL

The chi-squared is a statistical test that is used to determine **whether there is a relationship between two categorical variables.**

It is applied to sets of categorical data to evaluate **how likely it is that any observed difference between the sets arose by chance.**

Example: Littering Behavior Between Genders

1. Presume you observed 100 people to see who deposits garbage in the can and who litters. You want to see if there is a difference based on gender.A person can fall in one of for categories

Person	Deposit	Litter	Male	Female
John	Yes	no	Yes	No
Julia	No	Yes	No	Yes

	Deposit	Litter	
Female	18	7	25
Male	42	33	75
	60	40	100

Given this data, is there a significant difference in littering behavior between men and women?

Null Hypothesis H_0:

Males and females litter at the same rate.

Alternative Hypothesis H_1:

Males and females litter at different rates.

To answer this question, you have to figure out what numbers you might expect if everything were left to chance.

If $H_o$were true, that there is no difference based on gender.

	Deposit	Litter	
Female	18 15	7	25
Male	42	33	75
	60	40	100

Since 60 people deposited their garbage and 25% of them were female, you would expect 15 (25% of 60) females to be the value in the upper left cell, if there's an equal distribution – no effect on gender.

	Deposit	Litter	
Female	18 15	7	25
Male	42	33 30	75
	60	40	100

Since 40 people littered and 75% of them were male, you would expect **30** (25% of 60) males to be the value in the lower right cell, if there is no gender effect.

	Deposit	Litter	
Female	18 15	7 10	25
Male	42 45	33 30	75
	60	40	100

Working in a similar method, you can fill in all the expected values.

The further the observed values are from the expected values, the more likely that there really is a significant difference.

Chi-Squared Equation

$$\boxed{x^2 = \sum \frac{(O-E)^2}{E}}$$

Where:

O is the observed values of each cell

E is the expected values of each cell

	Deposit	Litter	
Female	18 15	7 10	25
Male	42 45	33 30	75
	60	40	100

$$x^2 = \frac{(18-15)^2}{15} + \frac{(7-10)^2}{10} + \frac{(42-45)^2}{45} + \frac{(33-30)^2}{30}$$

$$x^2 = \frac{9}{15} + \frac{9}{10} + \frac{9}{45} + \frac{9}{30}$$

$$x^2 = 0.6 + 0.9 + 0.2 + 0.3$$

$$\boxed{x^2 = 2.0}$$

Chi-squared Test: Degree of Freedom and Significance Level

$$\boxed{x^2 = 2.0}$$

$$\boxed{df = (number\ of\ rows\ - 1)\ (number\ of\ columns - 1)}$$

$$df = (2 - 1)\ (2 - 1)$$

$$\boxed{df = 1}$$

$$\boxed{significance\ level\ \sigma = 0.05}$$

Chi-squared test: Decision

if χ^2 is greater than 3.841 we reject the null hypothesis H_0

If χ^2 is less than 3.841, we fail to reject the null hypothesis H_0

Since the computed **chi-square value** $(\mathbf{2.0})$ **is less than the critical value** $(\mathbf{3.841})$**,** then we fail to reject the null hypothesis.

There is no statistically significant difference between gender and whether they litter (at $\alpha = 0.05$).