CHI SQUARE TEST

The chi square Test for Goodness of fit test claims about population proportions.

It is a non parametric test that is performed on categorical [ordinal and nominal] data.

There is a population of Male who likes different color bikes

		Theory	Sample
Yellow	Bike	1/3	22
Red	Bike	1/3	17

* Goodness of fit

Red Bike 1/3 17

Orange Bike 1/3 59

↳ Observed categorical distribution

⇩

Theory categorical distribution

Goodness Of fit test

In a science class of 75 students, 11 are left handed. Does th
fit the theory that 12% of people are left handed.

$\frac{3}{25} \times 75 = 9$

Untitled · 2/2 · 156%

Goodness Of fit test

In a science class of 75 students, 11 are left handed. Does this class fit the theory that 12% of people are left handed.

$3 \overline{)12} \vee 75^{3} = 9$

Ans)

	O	E
left handed	11	9
Right handed	64	66
	75	75

In 2010 Census of the city, the weight of the individuals in the city were found to be the following

<50kg	50 - 75	>75
20%	30%	50%

In 2020, weight of n=500 individuals were sampled. Below are the results

<50	50-75	>75
140	160	200

Using $\alpha = 0.05$, would you conclude the population difference of weights changed in the last 10 years?

Ans)

2010

Expected

<50kg	50 - 75	>75
20%	30%	50%

2020

n=500

Observed

<50	50-75	>75
140	160	200

Expected

<50	50-75	>75
0.2 x 500 = 100	0.3 x 500 = 150	0.5 x 500 = 250

① Null Hypothesis : H_0 : The data meets the expectation

Alternate Hyp : H_1 : The data does not meet the expectation

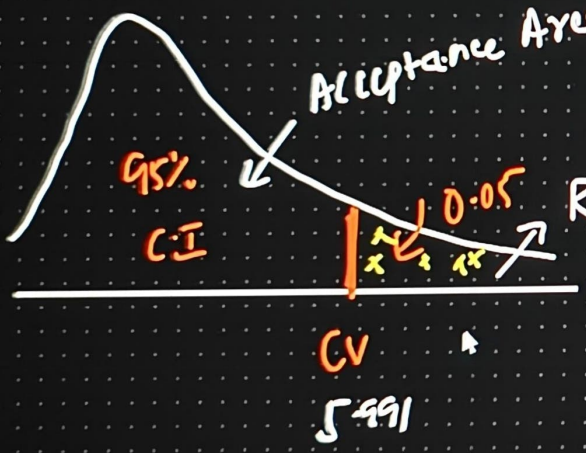
② $\alpha = 0.05$ (I: 95%

③ Degree of freedom

$df = K - 1 = 3 - 1 = 2$

$$df = K - 1 = 3 - 1 = 2$$

④ Decision Boundary



If χ^2 is greater than 5.99, Reject Ho

else

We fail to reject the Null Hypothesis

(5) Calculate Chi Square Test Statistics

2020

n=500

Observed

<50	50-75	>75
140	160	200

Expected

<50	50-75	>75
0.2 × 500 = 100	0.3 × 500 = 150	0.5 × 500 = 250

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

$$= \frac{(140-100)^2}{100} + \frac{(160-150)^2}{150} + \frac{(200-250)^2}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$= \frac{1600}{100} + \frac{100}{150} + \frac{2500}{250}$$

$$= 16 + 0.66 + 10$$

$$= 26.66$$

$$\chi^2 = 26.66$$

If χ^2 is greater than 5.99, Reject no
else
we fail to reject the Null Hypothesis

$\chi^2 = 26.66$

If χ^2 is greater than 5.99, Reject Ho

else

We fail to reject the Null Hypothesis

$26.66 > 5.99$, Reject Ho

Answer

The weights of 2020 population are different than those expected in the 2010 population

Chi-Square Test

Last Updated : 2 Aug, 2025

Chi-squared test indicates that there is a relationship between two entities. Handling data often involves testing hypotheses to extract useful information. In categorical analysis, chi-square tests are used to determine whether **observed frequencies differ significantly from expected frequencies** under a given hypothesis.

> *Chi-squared test, or χ^2 test, helps in determining whether these two variables are associated with each other.*

This test is widely used in market research, healthcare, social sciences, and more to analyze categorical relationships.

Formula For Chi-Square Test

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Symbols are broken down as follows:

- O_i: Observed frequency
- E_i: Expected frequency

Categorical Variables

Categorical variables classify data into distinct, non-numerical groups (e.g., colors, fruit types).

Key Characteristics:

1. **Distinct Groups:** No overlap (e.g., hair color: blonde, brunette).
2. **Non-Numerical:** No inherent order (e.g., "apple" ≠ "orange" numerically).
3. **Limited Options:** Fixed categories (e.g., traffic lights: red, yellow, green).

Example:

"Do you prefer tea, coffee, or juice?" → Categories: tea/coffee/juice.

Steps for Chi-Square Test

Steps and an illustration of an example of how sex influences which type of ice-cream a person will choose using a chi-square test are added below:

Step 1: Define Hypothesis

Step 1: Define Hypothesis

- **Null Hypothesis (H_0):** The observed frequencies match the expected distribution.
- **Alternative Hypothesis (H_1):** The observed frequencies do not match the expected distribution.

Step 2: Gather and Organize Data

Gather Information about the Two Category Variables:

Before performing a chi-square test, you should have on hand information about two categorical variables you wish to observe.

- You must collect details on people's sex (male or female) and their best flavors (e.g., chocolate, vanilla, strawberry).
- Once this information is collected, it can be inserted into a contingency table.

The hypothesis is that men prefer vanilla while women prefer chocolate. So we need to record how many have chosen vanilla among all male respondents versus the number who chose chocolate out of all female respondents.

Here's an example of what a contingency table might look like:

	Chocolate	Vanilla	Strawberry
Male	20	15	10
Female	25	20	30

Step 3: Calculate Expected Frequencies

- **Get Observed Frequency:** In any specific cell, the expected frequency can be described as the number of occurrences that would be expected if the two variables were independent.
- **Expected Frequency Calculation:** This involves multiplying the sums of rows and columns in proportion, then dividing by the total number of observations in a table.

Observed frequency is the table given above.

$$E_{ij} = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$$

- Male and chocolate: $\frac{45 \times 45}{120} = 16.875$
- Male and Vanilla: $\frac{45 \times 35}{120} = 13.125$

Summarizing,

- **Male:** Chocolate: 16.875, Vanilla: 13.125, Strawberry: 15.0,
- **Female:** Chocolate:12.125, Vanilla: 21.875, Strawberry: 25.0

Step 4: Perform Chi-Square Test

Use Chi-Square Formula:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(20-16.875)^2}{16.875} + \frac{(15-13.125)^2}{13.125} + \frac{(30-25)^2}{25} = 4.69$$

Step 5: Determine Degrees of Freedom (df)

df = (number of rows - 1) × (number of columns - 1)

$$df = (r-1)(c-1) = (2-1)(3-1) = 2$$

Step 6: Find p-value

- Compare χ^2 to the **Chi-Square Distribution Table** for the given df.

$\chi^2 = 4.69$ with df=2: Critical value at α=0.05 is **5.991**. Since 4.69 < 5.991, **p > 0.05**

Step 7: Interpret Results

- If the **p-value is less than a certain significance level (e.g., 0.05)**, then we **reject the null hypothesis**, which is commonly denoted by α. Thus, it means that category variables highly correlate with each other.
- When a **p-value is above α**, it implies that we **cannot reject the null hypothesis; hence, there is insufficient evidence for establishing** the relationship between these variables.

No significant evidence supports the claim that men prefer vanilla or women prefer chocolate (p>0.05).

Addressing Assumptions and Considerations

- Chi-square tests suppose that the observations are independent of one another; they are

Addressing Assumptions and Considerations

- Chi-square tests suppose that the observations are independent of one another; they are distinct.
- Each cell in the table should have a minimum of five values in it for better results. Otherwise, think about Fisher's exact test as an alternative measure if a table cell has fewer than five numbers in it.
- Chi-square tests do not indicate a causal relationship, but they identify an association between variables.

Goodness-Of-Fit

A goodness-of-fit test checks if a hypothesized model matches observed data. For example, testing whether urban residents are taller than rural ones by comparing actual height data to predictions.

Key Aspects:

1. **Purpose**: Validate if data fits an expected distribution.
2. **Data Types**: Works for both categorical (e.g., survey responses) and continuous (e.g., heights) data.
3. **Applications**: Compare observed vs. expected frequencies (e.g., Chi-Square test) and assess if data follows a specific distribution (e.g., normal distribution).
4. **Benefits**: Identifies model-data mismatch.

Applications of Chi-Square Test in Computer Science

A/B Testing & Feature Evaluation

- Compare user engagement (e.g., clicks, conversions) between two website versions (A vs. B).
- Chi-test is used to test if observed metrics (e.g., "Click" vs. "No Click") differ significantly between groups.

Applications of Chi-Square Test in Computer Science

A/B Testing & Feature Evaluation

- Compare user engagement (e.g., clicks, conversions) between two website versions (A vs. B).
- Chi-test is used to test if observed metrics (e.g., "Click" vs. "No Click") differ significantly between groups.
- Example: Observed: Version A: 120 clicks / 1,000 views; Version B: 150 clicks / 1,000 views. Chi-Square: Checks if the difference is statistically significant (not due to chance).

Machine Learning (Feature Selection)

- Identify categorical features correlated with target variables.
- Test if independence between features (e.g., "Browser Type" vs. "Purchase Decision") using the Chi-square test.
- Example: χ^2 p-value < 0.05 → "Browser Type" significantly affects purchases.

Database Query Optimization

- Assess if data is evenly distributed across partitions.
- Chi-square is used to test if actual row counts per partition match the expected uniform distribution.
- Example: Uneven distribution (χ^2 significance) suggests a poor sharding strategy.

Natural Language Processing (NLP)

- Evaluate word frequency distributions in texts.
- Compare observed word counts (e.g., "error" in logs) to the expected Poisson distribution.
- Example: Detects overused terms in spam emails (χ^2 highlights deviations from normal usage).

normal usage).

Solved Examples on Chi-Square Test

Example 1: A study investigates the relationship between eye color (blue, brown, green) and hair color (blonde, brunette, Redhead). The following data is collected:

Eye Color	Blonde	Brunette	Redhead	Total
Blue	35	52.5	12.5	100
Brown	28.1	42.1	9.8	80
Green	6.9	10.4	2.7	20

Solution:

Calculate the chi-square value for each cell in the contingency table using the formula

$\chi^2 = (O_i - E_i)^2 / E_i$

Solution:

Calculate the chi-square value for each cell in the contingency table using the formula

$\chi^2 = (O_i - E_i)^2 / E_i$

For instance, consider someone with brown hair and blue eyes:

$\chi^2 = (15 - 28.1)^2 / 28.1 \approx 6.07$.

To complete the total chi-square statistic, find each cell's chi-squared value and sum them up across all the nine cells in the table.

Degrees of Freedom (df):

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

$df = (3 - 1) \times (3 - 1)$

$df = 2 \times 2 = 4$

Finding p-value:

You may reference a chi-square distribution table to get an estimated chi-square stat of (χ^2) using the appropriate degrees of freedom. Look for the closest value and its corresponding p-value since most tables do not show precise numbers.

If your Chi-square value was 20.5, you would observe that the nearest number in the table for $df = 4$ is 14.88 with a p-value in 0.005; an illustration is.

Interpreting Results:

- Selecting a level of significance ($\alpha = 0.05$ is common)or than if the null hypothesis holds, the probability of either rejecting it at all is limited (Type I error).
- Compare the alpha value and p-value,
- When the p-value is less than the significance level, which in this case is written as p-value < 0.05, we can reject the null hypothesis, There is sufficient evidence to say that hair and eye color are related in one direction according to statistical terms, If the p-value is greater than the significance level it means that we cannot reject the null hypothesis therefore p-value ≥ 0.05,
- Based on the data at hand, we cannot say that there is a statistically significant correlation between eye and hair colors,