Chi-Square Test: Theory, Types, and Real-World Use Case

# What is Chi-Square Test?

The Chi-Square (χ²) test is a statistical method used to determine whether there is a significant association between two categorical variables or whether an observed distribution differs from an expected distribution. It is a non-parametric test, meaning it does not assume a normal distribution.

# Types of Chi-Square Tests

## 1. Chi-Square Test for Independence

This test determines whether two categorical variables are independent or related. It is applied to a contingency table to assess the association.

Example: Does gender affect product preference?

## 2. Chi-Square Goodness-of-Fit Test

This test evaluates whether a sample data matches a population distribution. It compares observed frequencies to expected frequencies of one categorical variable.

Example: Are die rolls fair (uniform distribution)?

# Real-World Use Case: Customer Preference by Region

Problem: A company wants to check whether product preference is independent of customer region.  
Variables:  
- Product Preference: A, B, C  
- Region: North, South, East  
  
Step 1: Create a contingency table of observed frequencies:  
| | A | B | C |  
|-----------|---|---|---|  
| North | 30| 20| 50|  
| South | 40| 30| 30|  
| East | 20| 25| 55|  
  
Step 2: Calculate expected frequencies using:  
Expected = (Row Total × Column Total) / Grand Total  
  
Step 3: Compute Chi-Square Statistic:  
χ² = Σ [(Observed - Expected)² / Expected] for each cell  
  
Step 4: Determine degrees of freedom:  
df = (rows - 1) × (columns - 1)  
  
Step 5: Compare χ² statistic with critical value or compute p-value  
- If χ² > critical value or p-value < α, reject null hypothesis  
- Conclusion: Product preference and region are not independent

# ML Use Cases

## a. Feature Selection

Chi-Square can help determine the relevance of categorical input features with respect to the target variable, especially in classification problems.

**🧪 What Does the Chi-Square Test Do Here?**

For each feature:

1. It creates a **contingency table** (cross-tab)  
   For example, let’s say your feature is "Color" (Red, Green, Blue) and your target is "Purchased" (Yes, No)

| **Color** | **Yes** | **No** |
| --- | --- | --- |
| Red | 20 | 10 |
| Green | 15 | 25 |
| Blue | 5 | 30 |

1. It checks:

**Are some target classes more common for certain categories of this feature?**

1. Chi-Square tells you:
   * If the observed distribution is **very different** from what would be expected **by random chance**
   * If **p-value is low**, it means the feature and target are **dependent** → 🎯 **feature is important**
   * If **p-value is high**, it means the feature and target are **independent** → 🚫 **feature is probably useless**

## b. Proportional Class Split Verification

Use Chi-Square to check if training and test sets have similar class proportions. This ensures unbiased model evaluation.

Here I can use value\_count or get This will give you the **observed proportions** — but:

❌ It **doesn’t tell** you if the difference is **statistically significant**  
✅ Chi-Square tells you **whether the difference is real or just due to chance**

Ho: Train and test sets have **no significant difference** in class proportions

H1: Train and test sets have **significantly different** class proportions

**📊 Interpretation**

* If **p-value < significance level** (e.g., 0.05), reject H₀

This means:

1. You may have **split the data badly** (e.g., forgot to stratify)
2. The **original data** itself may be **imbalanced**

* If p-value > significance level

1. It means difference in proportion is just random happened during splitting and it’s a one-time luck.

## c. Categorical Variable Relationship in EDA

During exploratory data analysis, Chi-Square can test if there's an association between two categorical features.