Understanding the Chi-Square Test of Independence Using R: A Step-by-Step Guide with Examples

Authored by

Mohammed looti

November 7, 2025

The Chi-Square Test of Independence is a cornerstone statistical method utilized across various fields--from social science to market research--to rigorously assess whether an association exists between two categorical variables. This powerful technique is indispensable for analyzing frequency data, typically organized within a contingency table, enabling researchers to determine if the distribution of one characteristic is statistically independent of the distribution of the other.

Mastering this test is crucial for data analysis professionals. This comprehensive tutorial delivers a precise, step-by-step methodology for executing and accurately interpreting the Chi-Square Test of Independence using the leading statistical programming environment, **R**. We will walk through the theoretical foundation, data preparation, execution of the test, and the crucial final interpretation of the results.

The Core Objective and Underlying Assumptions

The fundamental objective of the Chi-Square Test is to statistically evaluate the consistency between the frequencies observed in a sample and the frequencies that would be mathematically **expected** if the two variables were entirely independent. When the observed counts substantially deviate from these calculated expected counts, it provides strong evidence supporting the existence of a true association between the variables in the population.

It is paramount to recognize that the validity of the Chi-Square test relies on strict underlying assumptions. First, the data must be derived from a **random sample** of the population of interest. Second, and critically, the expected frequency count for the vast majority of cells (typically 80% or more) in the contingency table should be five or greater. Violation of this minimum expected frequency assumption can lead to an unreliable approximation of the resulting p-value, potentially compromising the validity of the statistical conclusion.

Before proceeding to the computational steps in R, it is necessary to formally define the statistical framework, specifically establishing the null and alternative hypotheses that guide the entire analytical procedure.

Defining the Formal Statistical Hypotheses

In the context of statistical hypothesis testing, the Chi-Square test requires the clear articulation of two mutually exclusive statements concerning the population parameters. These statements define the relationship we are examining:

H0 (Null Hypothesis): The two categorical variables are statistically **independent**. This hypothesis posits that there is absolutely no association or relationship between them in the population from which the sample was drawn.
H1 (Alternative Hypothesis): The two categorical variables are **not independent** (i.e., they are

associated). Rejecting the null hypothesis implies that knowledge of one variable provides statistically significant predictive information about the other.

The ultimate determination of whether to reject or fail to reject the null hypothesis is determined by comparing the calculated **p-value** generated by the R function against our predetermined significance level.

Practical Application: Investigating Political Preference

To fully grasp the mechanics of the Chi-Square test, let us apply it to a practical, real-world scenario. Our objective is to determine whether a voter's **gender** exhibits a statistically significant association with their preferred political party affiliation. This investigation is representative of common analyses performed in political science, market segmentation, and sociological research.

For this analysis, we gathered a **simple random sample** comprising 500 registered voters. Each participant provided information regarding their gender (categorized as Male or Female) and their primary political affiliation (Republican, Democrat, or Independent). The resulting observed frequencies, which form the basis of our calculation, are systematically organized in the following **contingency table**:

	Republican	**Democrat**	**Independent**	**Total**
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

These specific cell counts represent the **observed frequencies**. Our subsequent steps in the **R** environment will utilize this quantitative data to compute the Chi-Square test statistic and determine if these counts differ significantly from what would be expected under the assumption of independence.

Step 1: Structuring Observed Frequencies in R

The initial and most critical step in R involves transforming our raw frequency counts into an appropriate data structure that the `chisq.test()` function can efficiently analyze. The preferred format for these analyses is an R **table** object, although we begin by creating a simple **matrix**.

We leverage the `matrix()` function to input the frequency data, specifying the number of columns (3) and utilizing the `byrow=TRUE` argument to ensure the data is read correctly, row by row (Male counts followed by Female counts). Crucially, we then assign explicit row and column names, which significantly enhances the readability and interpretability of the final output. The final

conversion to a formal R table object using `as.table()` prepares the data perfectly for statistical testing.

 # Create the data matrix containing observed frequencies (120, 90, 40 are Male counts)
data <- matrix(c(120, 90, 40, 110, 95, 45), ncol=3, byrow=TRUE)

# Assign meaningful column names (Political Preferences)
colnames(data) <- c("Rep","Dem","Ind")

# Assign meaningful row names (Gender Categories)
rownames(data) <- c("Male","Female")

# Convert the matrix into a formal R table object
data <- as.table(data)

View the final structured table
data

Rep Dem Ind
Male 120 90 40
Female 110 95 45

The resulting R object, named `data`, is now a perfectly structured representation of our original contingency table, containing all necessary observed frequencies to proceed with the rigorous hypothesis test.

Step 2: Running the `chisq.test()` Function

Once the frequency data is correctly organized in the `data` table, performing the Chi-Square Test of Independence is remarkably straightforward in R. We simply invoke the powerful, built-in function, **`chisq.test()`**, passing our prepared data object as the sole argument.

R automatically performs all required intermediate steps: calculating the expected frequencies for every cell based on marginal totals, computing the standardized differences (residuals), summing these differences to derive the Chi-Square statistic, and finally, using the appropriate degrees of freedom to determine the crucial p-value.

 # Perform the Chi-Square Test of Independence on the contingency table
chisq.test(data)

Pearson's Chi-squared test

```
data: data
X-squared = 0.86404, df = 2, p-value = 0.6492
```

The concise output displays the three most critical components needed for statistical inference: the test statistic (X-squared), the degrees of freedom (df), and the resultant p-value. The next step focuses on correctly interpreting these metrics.

Interpreting Statistical Results and Drawing a Conclusion

Accurately interpreting the output from the `chisq.test()` function is essential for translating statistical figures into a definitive conclusion about the population. The results provide a clear measure of the discrepancy between what we observed and what we expected under the condition of independence.

We must analyze the three key reported metrics:

Chi-Square Test Statistic (X-squared): 0.86404. This value represents the aggregate measure of the difference between the observed and expected frequencies. A higher X-squared value indicates a greater deviation from independence, suggesting stronger evidence against the null hypothesis.
Degrees of Freedom (df): 2. This parameter defines the specific shape of the Chi-Square distribution used for testing. It is mathematically derived as (Number of Rows - 1) multiplied by (Number of Columns - 1), yielding (2 - 1) × (3 - 1) = 2 in this example.
P-Value: 0.6492. This is the probability of obtaining a test statistic of 0.86404 (or something more extreme) strictly by random chance, assuming that the null hypothesis of independence is true.

The pivotal step in hypothesis testing is comparing the calculated **p-value** against a chosen **significance level**, denoted as alpha (α). Conventionally, researchers set this alpha level at 0.05, representing a 5% maximum threshold for committing a Type I error (incorrectly rejecting a true null hypothesis).

In this analysis, our calculated p-value (0.6492) is substantially larger than the standard significance threshold (0.05). Based on this comparison (0.6492 > 0.05), we conclude that we must **fail to reject the null hypothesis** (H0). The data gathered does not provide statistically compelling evidence to assert that a relationship exists between gender and political party preference in the broader population.

Therefore, we conclude that the observed minor differences in party preference between male and female voters are likely attributable purely to random sampling variability, and we cannot claim a statistically significant association between the two categorical variables.

Expanding Your Knowledge of Categorical Analysis

To further enhance your proficiency in analyzing categorical data and mastering the Chi-Square methodology within the R environment, consider exploring the following advanced tutorials and related informational resources:

A Comprehensive Introduction to the Chi-Square Test of Independence
Interactive Chi-Square Test of Independence Calculator Tool
Detailed Guide: How to Calculate the P-Value of a Chi-Square Statistic in R
Finding the Chi-Square Critical Value in R for Hypothesis Testing