Chi-Square Tests in R: A Practical Guide to Analyzing Categorical Data

Authored by

Mohammed looti

November 12, 2025

Introduction to the Chi-Square Tests

The Chi-Square test is a fundamental tool in **inferential statistics**, primarily used when analyzing **categorical variables**. Contrary to popular belief, there are two distinct types of Chi-Square tests, each addressing a unique analytical question. Mastering both is essential for effective data analysis, especially when utilizing the powerful capabilities of the R programming language.

These statistical tests allow researchers to move beyond descriptive statistics to determine if observed data significantly deviates from what is expected under specific assumptions. Understanding the differences between these two tests--Goodness of Fit and Independence--is the first step toward accurate statistical modeling.

The two primary forms of the Chi-Square test are:

Chi-Square Goodness of Fit Test - Used to determine whether a single **categorical variable** follows a predetermined or **hypothesized distribution**. This test compares **observed frequencies** against theoretically expected frequencies.
Chi-Square Test of Independence - Used to assess whether there is a statistically **significant association** between two distinct **categorical variables**. This test is crucial for cross-tabulated data analysis.

This comprehensive guide will walk through practical, step-by-step examples demonstrating how to execute and correctly interpret the results of both tests using R.

Applying the Chi-Square Goodness of Fit Test (Example 1)

The **Chi-Square Goodness of Fit Test** is applied when we want to compare observed data against a theoretical model or a set of expected proportions. Consider a scenario where a store owner hypothesizes that customer traffic is evenly distributed throughout the working week (Monday through Friday). This implies that the expected proportion of customers is equal for each of the five days (20% each).

To challenge this null hypothesis, the owner records the actual number of customers entering the shop over a single week, yielding the following **observed frequencies**:

Monday: 50 customers
Tuesday: 60 customers
Wednesday: 40 customers
Thursday: 47 customers
Friday: 53 customers

We must now perform the Chi-Square goodness of fit test in R to statistically evaluate if these

observed customer counts are consistent with the store owner's initial claim of equal distribution. This test assesses the probability that the differences between the observed and expected values occurred purely by chance.

Executing the Goodness of Fit Test in R

The standard function for conducting this analysis in R is **chisq.test()**. This function requires two key arguments to execute the test accurately, allowing the software to calculate the test statistic and the associated **p-value**.

The general syntax for the function is:

chisq.test(x, p)

x: Represents a numerical vector containing the **observed frequencies** (the actual counts collected).
p: Represents a numerical vector detailing the **expected proportions** under the null hypothesis (which must sum to 1).

The following R code snippet demonstrates how we prepare the observed customer data and the expected proportions (0.2 for each of the five days) and then execute the test:

```
#create array of observed and expected frequencies
observed <- c(50, 60, 40, 47, 53)
expected <- c(.2, .2, .2, .2, .2)

#perform Chi-Square Goodness of Fit Test
chisq.test(x=observed, p=expected)
```

Chi-squared test for given probabilities

data: observed
X-squared = 4.36, df = 4, p-value = 0.3595

Interpreting the Goodness of Fit Results

Upon execution, the **chisq.test()** output provides the essential components needed for hypothesis testing: the calculated Chi-Square test statistic and the associated p-value. These values form the basis of our conclusion regarding the null hypothesis.

The calculated **Chi-Square test statistic** (X-squared) is **4.36**.

The corresponding **p-value** is **0.3595**.

To make a decision, we compare the p-value to a predetermined significance level (alpha), typically set at 0.05. Since the calculated p-value (0.3595) is significantly greater than the alpha level (0.05), the statistical rule dictates that **we fail to reject the null hypothesis**.

In practical terms, this outcome suggests that based on the observed data, there is **not sufficient statistical evidence** to conclude that the actual distribution of customer visits throughout the week differs significantly from the claimed equal distribution (20% per day). The variation observed (e.g., 60 customers on Tuesday vs. 40 on Wednesday) can reasonably be attributed to random chance, supporting the store owner's initial hypothesis.

Analyzing Association: The Chi-Square Test of Independence (Example 2)

The second crucial application of the Chi-Square framework is the **Test of Independence**, which determines if two **categorical variables** are statistically related or independent of one another. For this example, imagine researchers are investigating whether a person's **gender** is associated with their **political party preference**.

A simple random sample of 500 registered voters was surveyed, and the results were tabulated into the following contingency table, showing the joint frequencies of the two variables:

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

The null hypothesis in the Test of Independence is that the two variables (Gender and Party Preference) are independent. The purpose of the test is to statistically evaluate whether the observed differences across the table cells are large enough to suggest a genuine relationship between the two variables, or if the variations are merely due to sampling error.

Performing the Test of Independence in R

Unlike the Goodness of Fit test, the Test of Independence in R simply requires the contingency table (or matrix of observed counts) as its primary input. We use the **matrix()** function to structure the data correctly, ensuring the row and column names are defined for clarity, and then pass this structure to **chisq.test()**. Note that the total counts (margins) are excluded from the matrix input, as R calculates these internally.

```
#create table to hold survey data (inputting row by row: Male counts, then Female counts)
data <- matrix(c(120, 90, 40, 110, 95, 45), ncol=3, byrow=TRUE)
colnames(data) <- c("Rep","Dem","Ind")
rownames(data) <- c("Male","Female")
data <- as.table(data)

#perform Chi-Square Test of Independence
chisq.test(data)
```

Pearson's Chi-squared test

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

Analyzing the output from Pearson's Chi-squared test provides us with the necessary statistics for decision-making:

Chi-Square Test Statistic: **0.86404**
The corresponding **p-value**: **0.6492**

Since the resulting p-value (0.6492) is significantly higher than the standard 0.05 threshold, we must conclude that **we fail to reject the null hypothesis** of independence. This crucial finding implies that, based on this sample data, there is **no statistical evidence** to assert that gender and political party preference are associated. We treat the variables as statistically independent within this population.

Summary and Additional Resources

The Chi-Square family of tests provides robust methods for handling **categorical data**, whether testing against a theoretical distribution (Goodness of Fit) or assessing the association between two distinct variables (Independence). The R programming environment, using the intuitive **chisq.test()** function, makes both applications straightforward to execute and interpret once the **observed frequencies** are correctly structured.

For those interested in exploring related statistical concepts or performing other common tasks within R, the following resources offer valuable insight:

How to Perform a Chi-Square Test of Independence in R
How to Calculate the P-Value of a Chi-Square Statistic in R
How to Find the Chi-Square Critical Value in R