Chi-Square Goodness of Fit Test in Python: A Step-by-Step Guide

Authored by

Mohammed looti

November 8, 2025

The Chi-Square Goodness of Fit Test is an indispensable procedure in inferential statistics, serving as a powerful mechanism to validate fundamental assumptions about population distributions. This test is specifically engineered to determine if the distribution of counts for a categorical variable, collected empirically from a sample, significantly deviates from a known or hypothesized distribution. By comparing observed frequencies against expected frequencies, researchers can rigorously assess whether theoretical claims align with real-world data.

Crucially, the Goodness of Fit Test operates differently from standard tests that focus on means or variances; its entire scope is dedicated to analyzing proportions and frequencies across distinct categories. Its versatility makes it highly applicable across diverse domains, ranging from market research and biological studies to quality assurance in manufacturing, whenever the objective is to verify if observed outcomes conform to established probabilities. This comprehensive tutorial provides a structured, step-by-step guide on how to effectively execute, analyze, and interpret the results of the Chi-Square Goodness of Fit Test using the sophisticated statistical tools available within the SciPy library in Python.

Understanding the Core Statistical Hypotheses

Every reliable statistical analysis begins with the formal establishment of competing hypotheses. These statements--the null hypothesis and the alternative hypothesis--provide the framework necessary to interpret the calculated results and draw a statistically sound conclusion. For the Chi-Square Goodness of Fit Test, the hypotheses focus squarely on the relationship between the observed data and the hypothesized distribution.

The standard hypotheses used to frame this test are structured as follows, focusing on whether the population distribution conforms to the expected pattern:

H0 (Null Hypothesis): The distribution of the categorical variable in the population follows the specified hypothesized distribution. Any minor differences observed between the empirical data and the expected counts are attributed purely to random sampling variation.
H1 (Alternative Hypothesis): The distribution of the categorical variable in the population does not follow the hypothesized distribution. This suggests that the observed data pattern is significantly different from what was expected, pointing toward a genuine effect.

The primary goal of conducting the test is to quantify the evidence against the **Null Hypothesis** (H0). If the observed counts diverge significantly from the expected counts, the resulting test statistic will be large. A high test statistic typically corresponds to a small p-value, providing sufficient evidence to reject H0 in favor of the alternative.

Defining the Scenario and Collecting Observed Data

To demonstrate the practical utility of this test, let us examine a typical scenario faced in retail management. Imagine a shop owner who believes and asserts that customer traffic is equally distributed across the five standard weekdays (Monday through Friday). This claim of uniformity establishes our **hypothesized distribution**, where the expectation is that 20% of the total weekly traffic occurs on each day.

To verify this assertion empirically, a researcher conducts a study by meticulously recording the number of customers who enter the shop over a single work week. The results of this counting process yield the **observed counts**--the actual frequencies found in the sample data. This collection of raw data is essential for the subsequent statistical calculations.

The observed frequencies gathered over the week are summarized below:

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

The total number of observed customers for the week is 250 (50 + 60 + 40 + 47 + 53). Based on the shop owner's claim of an equal distribution across five days, the **expected count** for each individual day must be calculated as 250 total customers divided by 5 days, yielding 50 customers per day. We will now use these precisely defined observed and expected frequencies to calculate the Chi-Square statistic and formally test the owner's hypothesis.

Step 1: Preparing Data Arrays in Python

The initial and most critical practical step in implementing this test in **Python** involves structuring our data into appropriate numerical arrays. It is standard practice to utilize lists or NumPy arrays to securely store both the observed frequencies (f_{obs}) and the expected frequencies (f_{exp}). Ensuring that both arrays are perfectly ordered and aligned--meaning the first element of f_{obs} corresponds directly to the first element of f_{exp}, and so forth--is paramount for accurate calculation.

We define two separate arrays reflecting the observed data collected by the researcher and the expected counts derived directly from the shop owner's claim of a uniform distribution. This organization ensures that the statistical function correctly pairs the actual outcomes with the theoretical expectations:

expected =
observed =

In this arrangement, the lists are perfectly aligned by day of the week. For instance, the 50 observed customers on Monday are directly compared against the 50 expected customers for Monday, allowing the subsequent calculation to accurately measure the deviation for each category.

Step 2: Executing the Chi-Square Goodness of Fit Test

The SciPy library (3) is the backbone of statistical computing in Python, offering robust and optimized functions for hypothesis testing. To perform the Goodness of Fit Test, we utilize the `stats.chisquare` function. This function efficiently computes the Chi-Square test statistic and its corresponding p-value based solely on the provided frequency arrays. It is imperative to import the necessary statistical module from SciPy before execution.

The general syntax for the function is straightforward:

chisquare(f_obs, f_exp)

The function accepts two key parameters, which define the input data:

f_obs: An array containing the **observed frequency** counts for every category tested.
f_exp: An array containing the **expected frequency** counts for every category, as derived from the premise of the null hypothesis. Note that if this parameter is omitted, the function automatically assumes an equal probability (uniform distribution) across all categories.

The following Python code snippet demonstrates the execution of the test using our retail example data. We import the module and pass our aligned lists directly to the function:

import scipy.stats as stats

#perform Chi-Square Goodness of Fit Test
stats.chisquare(f_obs=observed, f_exp=expected)

(statistic=4.36, pvalue=0.35947)

Upon successful execution, the function returns a tuple containing the two results critical for our analysis: the calculated Chi-Square **test statistic** (4) and the associated **p-value** (4). For the customer traffic example, the Chi-Square statistic is calculated as **4.36**, and the corresponding p-value is determined to be **0.35947**.

Step 3: Interpreting the Statistical Results

The final and most determinative phase of the analysis involves interpreting the calculated values within the established framework of hypothesis testing. The calculated Chi-Square test statistic (5) serves as a quantitative measure of the overall deviation between the observed data and the data expected under the null hypothesis. Larger values of this statistic indicate a more significant discrepancy, suggesting that the observed data does not fit the null model well.

The final decision regarding the rejection or retention of the null hypothesis relies primarily on the p-value (5). The p-value represents the probability of obtaining sample data as extreme as, or more extreme than, the data observed in our study, assuming that the **Null Hypothesis** (H0) is actually true. This value is compared against a predetermined **significance level** ($alpha$), which is most commonly set at 0.05 (or 5%).

The decision rule is based on a straightforward comparison:

If the p-value leq $alpha$ (e.g., 0.05), we conclude that the evidence is strong enough to **reject** the null hypothesis.
If the p-value $>$ $alpha$ (e.g., 0.05), we conclude that there is insufficient evidence, and we **fail to reject** the null hypothesis.

In our specific example, the calculated p-value is 0.35947. Since this value (0.35947) is substantially greater than the conventional significance level of 0.05, we consequently must fail to reject the null hypothesis.

Conclusion: Relating the Test back to the Business Claim

The outcome of failing to reject the null hypothesis carries a crucial implication for the business scenario. It signifies that the discrepancies observed in the customer counts--such as the difference between 60 customers on Tuesday and 40 customers on Wednesday--are not statistically significant enough to conclude that the true underlying distribution is anything other than uniform. In statistical terminology, these variations could reasonably be attributed to random chance or expected sampling variability.

Therefore, based on the rigorous findings of the Chi-Square Goodness of Fit Test (3), we conclude that there is insufficient statistical evidence, at the 5% significance level, to assert that the true distribution of customer traffic across the weekdays deviates from a uniform distribution. The shop owner's original claim that traffic is equally distributed is statistically supported by the collected data.