# Learning the Chi-Square Goodness of Fit Test: A Step-by-Step Guide

Authored by

**Mohammed looti**

November 2, 2025

The **Chi-Square goodness of fit test** is a fundamental statistical tool used to assess how well an observed sample distribution aligns with a theoretical or hypothesized distribution. Essentially, it helps determine if there is a statistically significant difference between what we observe in the real world and what we would expect to see under a specific assumption.

While modern statistical software can perform this calculation instantaneously, understanding the underlying mechanism is crucial for interpreting the results correctly. The following detailed, step-by-step example walks you through executing a **Chi-Square goodness of fit test** entirely by hand, reinforcing the core concepts involved in this powerful analysis.

## The Statistical Scenario: Testing a Fair Dice

To illustrate the process, let us consider a common statistical challenge: determining if a standard six-sided dice is fair. If the dice is fair, we assume that the outcome of any single roll follows a uniform distribution--meaning the probability of landing on a 1, 2, 3, 4, 5, or 6 is exactly the same (1/6).

If we roll the dice a large number of times, the observed frequencies of each outcome should closely match the expected frequencies based on this assumption of fairness. Any significant deviation suggests the dice might be "loaded" or biased. To gather our data, we decide to roll the dice 60 times, recording the result of each roll. The collected observational data is summarized below:

**1**: 8 times
**2**: 12 times
**3**: 18 times
**4**: 9 times
**5**: 7 times
**6**: 6 times

Our objective now is to use the five defined steps of the Chi-Square goodness of fit test to rigorously determine if the observed pattern justifies rejecting the initial belief that the dice is fair.

## Step 1: Formulating the Null and Alternative Hypotheses

Every inferential statistical test begins with the establishment of two competing statements: the null hypothesis and the alternative hypothesis. These hypotheses formally define the statistical question we are attempting to answer.

The **null hypothesis** (H0) always represents the status quo or the statement of no effect or no difference. In the context of the goodness of fit test, H0 posits that the observed data fits the

theoretical distribution we are testing. Conversely, the **alternative hypothesis** (H1) states that there is a significant difference, meaning the observed data does not fit the theoretical distribution.

**H0** (Null Hypothesis): The distribution of the dice rolls is uniform. That is, the dice is equally likely to land on each number (1/6 probability for each side).
**H1** (Alternative Hypothesis): The distribution of the dice rolls is not uniform. The dice is biased, and the outcomes are not equally likely.

We are testing H0 under the assumption that it is true. We only reject H0 if the evidence from our observed data is so extreme that it is highly improbable if H0 were true.

## Step 2: Determining Observed and Expected Frequencies

The core of the Chi-Square test relies on comparing the actual counts we recorded (the **Observed Frequencies**, designated as O) against the counts we would theoretically expect to record if the null hypothesis were perfectly true (the **Expected Frequencies**, designated as E).

In our experiment, the total number of rolls (N) is 60. Since H0 states that the dice is fair, we expect each of the six outcomes to occur an equal number of times. Therefore, the expected frequency (E) for any single category is calculated by dividing the total number of trials by the number of categories (k):

$E = N / k = 60 / 6 = 10$

This means that, assuming the dice is perfectly fair, we would expect to see exactly 10 occurrences of the number 1, 10 occurrences of the number 2, and so forth. We can now construct a table comparing the Observed frequencies (O) to these Expected frequencies (E):

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **O** | 8 | 12 | 18 | 9 | 7 | 6 |
| **E** | 10 | 10 | 10 | 10 | 10 | 10 |

A crucial requirement for the validity of the Chi-Square test is that all expected frequencies (E) must be greater than or equal to 5. In this case, all our expected frequencies are 10, satisfying this necessary assumption for the test.

## Step 3: Calculating the Chi-Square Test Statistic

The next stage involves quantifying the magnitude of the difference between the observed data

and the expected data. This quantification is achieved through the calculation of the **Chi-Square test statistic**, denoted as $X^2$.

The formula for the Chi-Square statistic aggregates the standardized squared differences across all categories (k):

$X^2 = \Sigma$

The term (O - E) measures the raw difference between the observed and expected counts. This difference is then squared, ensuring that positive and negative deviations do not cancel each other out. Finally, dividing by E standardizes the difference relative to the size of the expected count. This adjustment is essential because a difference of 2 is far more significant if the expected count was 5 than if the expected count was 500.

We apply this calculation sequentially to each of the six categories and then sum the results to find the total $X^2$ value:

| | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| **O** | 8 | 12 | 18 | 9 | 7 | 6 | |
| **E** | 10 | 10 | 10 | 10 | 10 | 10 | |
| **$(O-E)^2 / E$** | 0.4 | 0.4 | 6.4 | 0.1 | 0.9 | 1.6 | $\Sigma$ = **9.8** |

As demonstrated in the calculation table, the individual contributions from each category are summed. The total Chi-Square test statistic ($X^2$) for this experiment turns out to be **9.8**. This value represents the total discrepancy between the observed results and the results expected if the dice were truly fair.

## Step 4: Finding the Critical Value and Degrees of Freedom

To determine if the calculated test statistic ($X^2 = 9.8$) is large enough to warrant rejecting the null hypothesis, we must compare it against a **critical value**. This critical value is drawn from the Chi-Square distribution table and depends on two parameters: the chosen significance level (alpha, $\alpha$) and the degrees of freedom (df).

The significance level ($\alpha$) is the probability threshold we set for rejecting H0 when it is actually true (Type I error). Conventionally, $\alpha$ is set at **0.05**, meaning we are willing to accept a 5% chance of incorrectly concluding the dice is biased when it is not.

The **degrees of freedom** (df) represent the number of independent values that can vary in an analysis. For the Chi-Square goodness of fit test, the degrees of freedom are calculated as the number of categories (k) minus one:

df = k - 1

Since there are six possible outcomes (categories 1 through 6) for the dice roll, we calculate the degrees of freedom as:

df = 6 - 1 = **5**

Using the Chi-Square distribution table, we locate the intersection of df = 5 and $\alpha = 0.05$. This intersection defines the critical value that sets the boundary between the region of "failure to reject H0" and the region of "rejection of H0."

| | P | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DF | 0.995 | 0.975 | 0.2 | 0.1 | 0.05 | 0.025 | 0.02 | 0.01 | 0.005 | 0.002 | 0.001 |
| 1 | .0004 | .00016 | 1.642 | 2.706 | 3.841 | 5.024 | 5.412 | 6.635 | 7.879 | 9.55 | 10.828 |
| 2 | 0.01 | 0.0506 | 3.219 | 4.605 | 5.991 | 7.378 | 7.824 | 9.21 | 10.597 | 12.429 | 13.816 |
| 3 | 0.0717 | 0.216 | 4.642 | 6.251 | 7.815 | 9.348 | 9.837 | 11.345 | 12.838 | 14.796 | 16.266 |
| 4 | 0.207 | 0.484 | 5.989 | 7.779 | 9.488 | 11.143 | 11.668 | 13.277 | 14.86 | 16.924 | 18.467 |
| 5 | 0.412 | 0.831 | 7.289 | 9.236 | 11.07 | 12.833 | 13.388 | 15.086 | 16.75 | 18.907 | 20.515 |
| 6 | 0.676 | 1.237 | 8.558 | 10.645 | 12.592 | 14.449 | 15.033 | 16.812 | 18.548 | 20.791 | 22.458 |
| 7 | 0.989 | 1.69 | 9.803 | 12.017 | 14.067 | 16.013 | 16.622 | 18.475 | 20.278 | 22.601 | 24.322 |
| 8 | 1.344 | 2.18 | 11.03 | 13.362 | 15.507 | 17.535 | 18.168 | 20.09 | 21.955 | 24.352 | 26.124 |
| 9 | 1.735 | 2.7 | 12.242 | 14.684 | 16.919 | 19.023 | 19.679 | 21.666 | 23.589 | 26.056 | 27.877 |
| 10 | 2.156 | 3.247 | 13.442 | 15.987 | 18.307 | 20.483 | 21.161 | 23.209 | 25.188 | 27.722 | 29.588 |
| 11 | 2.603 | 3.816 | 14.631 | 17.275 | 19.675 | 21.92 | 22.618 | 24.725 | 26.757 | 29.354 | 31.264 |
| 12 | 3.074 | 4.404 | 15.812 | 18.549 | 21.026 | 23.337 | 24.054 | 26.217 | 28.3 | 30.957 | 32.909 |
| 13 | 3.565 | 5.009 | 16.985 | 19.812 | 22.362 | 24.736 | 25.472 | 27.688 | 29.819 | 32.535 | 34.528 |
| 14 | 4.075 | 5.629 | 18.151 | 21.064 | 23.685 | 26.119 | 26.873 | 29.141 | 31.319 | 34.091 | 36.123 |
| 15 | 4.601 | 6.262 | 19.311 | 22.307 | 24.996 | 27.488 | 28.259 | 30.578 | 32.801 | 35.628 | 37.697 |
| 16 | 5.142 | 6.908 | 20.465 | 23.542 | 26.296 | 28.845 | 29.633 | 32 | 34.267 | 37.146 | 39.252 |
| 17 | 5.697 | 7.564 | 21.615 | 24.769 | 27.587 | 30.191 | 30.995 | 33.409 | 35.718 | 38.648 | 40.79 |
| 18 | 6.265 | 8.231 | 22.76 | 25.989 | 28.869 | 31.526 | 32.346 | 34.805 | 37.156 | 40.136 | 42.312 |

Consulting the table reveals that the critical value corresponding to df = 5 and $\alpha = 0.05$ is **11.07**. Any calculated $X^2$ value greater than 11.07 falls into the rejection region, indicating that the observed deviation is too large to be attributed to random chance.

## Step 5: Decision Making: Rejecting or Failing to Reject the Null Hypothesis

The final step involves comparing the calculated test statistic ($X^2$) from Step 3 with the critical value determined in Step 4. This comparison dictates our statistical decision regarding the initial hypotheses.

Our calculated test statistic is $X^2$ = **9.8**.

Our critical value is **11.07**.

The rule for the Chi-Square goodness of fit test states: If $X^2$ is greater than the critical value, we reject the null hypothesis (H0). If $X^2$ is less than or equal to the critical value, we fail to reject the null hypothesis.

Since 9.8 is less than 11.07, our result falls outside the critical rejection region. Therefore, we **fail to reject the null hypothesis**.

**Conclusion:** We do not have sufficient statistical evidence, based on the 60 rolls, to conclude that the dice is unfair or biased at the 0.05 significance level. The differences observed between the actual roll counts and the expected counts are likely due to random sampling variability rather than a systematic bias in the dice itself. The data is consistent with the hypothesis that the dice is equally likely to land on any of its six sides.

## Additional Resources and Next Steps

Mastering the Chi-Square goodness of fit test is a crucial skill for anyone engaging in data analysis or statistical inference. While this step-by-step example focused on a simple uniform distribution, the principles apply equally well to comparing observed data against any theoretical distribution (such as a normal or Poisson distribution).

For further exploration of the theoretical foundations and practical applications of this test, consult authoritative statistical textbooks and the following resources: