How to Easily Perform a Chi-Square Test by Hand: A Step-by-Step Guide

Authored by
stats writer

December 3, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Easily Perform a Chi-Square Test by Hand: A Step-by-Step Guide*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=104414

Overview of the Chi-Square Goodness of Fit Test

The manual execution of the Chi-Square test is a fundamental exercise in statistical analysis, allowing researchers to evaluate whether the frequency distribution of observed categorical data deviates significantly from a theoretical or hypothesized distribution. Specifically, we focus on the Goodness of Fit test, which determines if a sample distribution aligns with a claimed population distribution. This powerful non-parametric technique requires careful calculation of three key elements: the **Chi-Square statistic**, the degrees of freedom, and the critical value.

The overall procedure involves comparing the frequencies we actually measured (observed data) against the frequencies we would expect to see if the underlying assumptions were true (expected data). The larger the discrepancy between the observed and expected values, the greater the resulting Chi-Square statistic will be. When the calculated statistic exceeds the pre-determined critical value--a threshold based on the desired level of confidence--we conclude that the observed data pattern is unlikely to have occurred by random chance, leading to the rejection of the status quo assumption, also known as the Null Hypothesis.

Performing this test by hand provides deep insight into the mechanism of hypothesis testing. It involves a systematic series of arithmetic operations: first, establishing the theoretical expected values; second, quantifying the deviation between observed and expected outcomes; third, squaring these deviations and normalizing them by the expected values; and finally, summing these standardized differences to arrive at the final test statistic. Once the test statistic is computed and the degrees of freedom are established, comparison against a standard Chi-Square distribution table provides the necessary critical value for making a definitive statistical decision.

Setting the Stage: The Dice Fairness Example

To illustrate the step-by-step process of the Chi-Square Goodness of Fit test, we utilize a common scenario involving probability: testing whether a six-sided die is fair. A fair die is defined by the assumption that it is equally likely to land on any face (1, 2, 3, 4, 5, or 6) during any given roll. This constitutes our hypothesized distribution, meaning that the theoretical probability for each outcome is 1/6, establishing a uniform distribution expectation.

Our objective is to gather empirical evidence to challenge or support this assumption of fairness. We decide to roll the die a total of **60 times**, ensuring a sufficient sample size for the test to be statistically robust. We meticulously record the outcome of each roll, grouping the results into categories corresponding to the number shown on the face of the die. These counts form our crucial set of observed frequencies, which will be the basis for comparison against the theoretical expectation.

The empirical results collected after 60 rolls are summarized below, revealing the actual distribution of outcomes. Note that while the theoretical distribution suggests 10 occurrences for each number (60 total rolls divided by 6 sides equals 10), the observed frequencies show natural variation. This variation is precisely what the Chi-Square test quantifies to determine if it is merely random sampling fluctuation or evidence of inherent bias in the die.

Observed Frequencies from 60 Rolls

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

Step 1: Defining the Null and Alternative Hypotheses

The cornerstone of any inferential statistical test is the precise definition of the Null Hypothesis (H_0) and the Alternative Hypothesis (H_1). These statements represent the two competing claims about the population from which the data was drawn. The **Null Hypothesis** always represents the status quo, the assumption of no effect, or in this case, the assumption that the observed distribution is consistent with the hypothesized uniform distribution.

In the context of the die experiment, the hypotheses are formally stated as follows, establishing the framework within which our test statistic will be interpreted. We are essentially testing the proposition that all outcomes are equally probable versus the claim that some bias exists, causing unequal likelihoods among the outcomes. The test aims to find evidence **against** H_0.

H0 (Null Hypothesis): The die is fair; the probability of landing on any specific number (1 through 6) is equal ($P(1) = P(2) = dots = P(6) = 1/6$). The distribution of observed frequencies fits the hypothesized uniform distribution.
H1 (Alternative Hypothesis): The die is unfair; the probability distribution is not uniform, meaning the die is not equally likely to land on each number. The distribution of observed frequencies does not fit the hypothesized distribution.

The entire purpose of calculating the Chi-Square statistic is to determine if we have enough statistical evidence--meaning a large enough deviation from expectation--to reject H_0. If the evidence is weak (the calculated statistic is small), we fail to reject H_0, concluding that the observed deviations are likely due to random chance rather than inherent unfairness in the die.

Step 2: Calculating the Observed and Expected Frequencies

Before proceeding with the main calculation, we must organize and explicitly define our Observed Frequencies (O) and our Expected Frequencies (E). The Observed Frequencies are simply the counts recorded during the experiment, as listed previously. The **Expected Frequencies** represent the theoretical counts under the assumption that the Null Hypothesis (H_0) is perfectly true. Calculating these expected values accurately is critical for setting the baseline of comparison.

Since H_0 states that the die is perfectly fair and we rolled it 60 times, the expected count for any single outcome is calculated by multiplying the total number of trials by the theoretical probability for that outcome. For a fair six-sided die, the probability of rolling any single number is 1/6. Therefore, the expected frequency for each category is $mathbf{60 times (1/6) = 10}$.

We compile these figures into a comparative table, which is shown below. This visual organization is critical for the next step, as the Chi-Square formula relies on calculating the difference between the Observed and Expected counts for every category. Note: If we believe the dice is fair, this means we expect it to land on each number an equal amount of times--in this specific case, 10 times each.

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

Step 3: Calculation of the Chi-Square Test Statistic

The Chi-Square statistic, denoted as X^2 or chi^2, is the quantitative measure of the total disparity between the observed data and the expected model. It is calculated by summing the standardized, squared differences across all categories. The mathematical definition ensures that larger differences contribute more significantly to the final value, and the squaring ensures that deviations below the expectation do not cancel out deviations above the expectation.

The formal formula for the Chi-Square test statistic (X^2) for a Goodness of Fit test is:

$X^2 = Sigma frac{(O - E)^2}{E}$

We calculate this step-by-step for each category (side 1 through 6). This process involves three internal steps for each row: finding the raw difference $(O - E)$, squaring that difference $(O - E)^2$, and finally dividing the squared difference by the Expected frequency E. For instance, for the outcome '3' (Observed=18, Expected=10), the calculation is: $(18 - 10)^2 / 10 = (8)^2 / 10 = 64

/ 10 = 6.4$. We repeat this process for all six categories and then sum the resulting standardized values.

The table below details the calculation process, showing the difference, the squared difference, and the final standardized contribution $frac{(O - E)^2}{E}$ for each outcome. The summation of the final column yields our calculated test statistic.

	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	Σ = **9.8**

After summing the standardized differences across all categories, we find that the total Chi-Square test statistic, X^2, equals $mathbf{9.8}$. This calculated value quantifies the cumulative deviation observed in our 60 rolls relative to what was theoretically expected under the assumption of a fair die. This is the crucial number we must now compare against a theoretical distribution to determine its statistical significance.

Step 4: Determining Degrees of Freedom and the Critical Value

To properly interpret the calculated X^2 value, we must establish the appropriate critical threshold. This threshold depends on two parameters: the chosen significance level ($alpha$) and the degrees of freedom (df). The significance level, typically set at $alpha = 0.05$ (or 5%), represents the maximum probability of making a Type I error--the error of incorrectly rejecting the true Null Hypothesis--that we are willing to accept.

The degrees of freedom specify which specific Chi-Square distribution curve we must use for comparison. In a Goodness of Fit test, the df is calculated as the number of categories (k) minus one ($df = k - 1$). This formula is used because, given the fixed total sample size, once we know the counts of $k-1$ categories, the count of the final category is automatically determined. Since our experiment has 6 categories (the numbers 1 through 6), the calculation is $df = 6 - 1 = mathbf{5}$.

Using the standard Chi-Square distribution table, we look up the value corresponding to $df = 5$ and a significance level of $alpha = 0.05$. The critical value defines the boundary of the rejection region. If our calculated X^2 falls into this rejection region (i.e., is larger than the critical value), the result is deemed statistically significant. Consulting the table reveals that for $df=5$ and

$alpha=0.05$, the critical value is $mathbf{11.07}$.

	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

Step 5: Decision Making and Interpreting the Results

The final and most crucial step is comparing the calculated Test Statistic (X^2) with the Critical Value ($X^2_{text{critical}}$). This comparison dictates the statistical decision regarding the Null Hypothesis (H_0).

We found the following key values:

Calculated Test Statistic (X^2): $mathbf{9.8}$
Critical Value ($X^2_{text{critical}}$ for $df=5, alpha=0.05$): $mathbf{11.07}$

The decision rule is straightforward: If the calculated X^2 is greater than the critical value, we reject H_0. If the calculated X^2 is less than or equal to the critical value, we fail to reject H_0. In our case, 9.8 is clearly less than 11.07. Therefore, we **fail to reject the Null Hypothesis**.

Failing to reject the Null Hypothesis means that the difference observed between the actual observed frequencies and the expected frequencies is not statistically large enough to conclude, with 95% confidence, that the die is unfair. We conclude that we do not have sufficient evidence to definitively state that the die is biased or that its probability distribution is non-uniform. The minor fluctuations observed are likely attributable to the natural randomness inherent in the sampling process, rather than systematic unfairness.

Summary of the Chi-Square Goodness of Fit Test

The execution of the Chi-Square test by hand reinforces the understanding of how statistical significance is determined by linking observed reality to theoretical probability. The entire process ensures a rigorous quantitative assessment of whether categorical data aligns with a theoretical expectation. The calculated test statistic acts as a measure of accumulated deviation, while the critical value establishes the statistical threshold for making claims about the underlying population distribution.

In conclusion, the Chi-Square goodness of fit test is an invaluable tool for researchers across various disciplines seeking to compare frequency data. By following these methodical steps--from setting clear hypotheses and determining expected values to calculating the Chi-Square statistic and making a formal decision--one can draw reliable conclusions about the distribution of observational data relative to theory.