

How to Calculate Expected Counts for Chi-Square Tests

Authored by
mohammed looti

January 8, 2026

RECOMMENDED CITATION

mohammed looti (2026). *How to Calculate Expected Counts for Chi-Square Tests*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=124975>

The calculation of expected counts is a foundational step in performing any Chi-Square test. These values are crucial because they establish a baseline expectation--the scenario that would exist if the underlying variables were completely independent or if the data perfectly matched a hypothesized distribution. Essentially, the expected count represents the theoretical frequency for a specific category or cell within a contingency table under the assumption that the null hypothesis is true.

To determine these figures, we calculate the number of observations that we would anticipate seeing in each cell based solely on the marginal totals (row totals and column totals) and the grand total of the sample. This process involves multiplying the overall sample size by the proportion designated for each category. Once calculated, these theoretical frequencies are rigorously compared against the actual observed counts collected from the sample data. The discrepancy between the expected and observed counts is quantified by the Chi-Square statistic, which ultimately determines the statistical significance and the degree of association between the variables being investigated.

The Role of Expected Counts in Statistical Inference

Expected counts are not merely calculated values; they are the statistical embodiment of the null hypothesis in a categorical analysis. When we conduct a Chi-Square analysis, we are testing a null hypothesis which typically states that there is no relationship or association between the categorical variables being studied. Therefore, the expected count for any given cell is the frequency that would be anticipated if this hypothesis of no association were perfectly true in the population.

The core mechanism of the Chi-Square test involves summarizing the difference between what we see (the **observed counts**) and what we expect to see under the null hypothesis (the **expected counts**). This summary is calculated using the formula: $\sum \frac{(O - E)^2}{E}$, where O is the observed frequency and E is the expected frequency. A large difference between the observed and expected counts leads to a larger Chi-Square test statistic, suggesting that the null hypothesis is implausible.

A fundamental requirement for the validity of the Chi-Square test is that the expected counts must meet certain minimum thresholds. Specifically, it is generally accepted that no more than 20% of the cells should have an expected count less than 5, and no cell should have an expected count of zero. If these assumptions are violated, the sampling distribution of the test statistic may be inaccurate, necessitating the use of alternative statistical methods like Fisher's exact test or combining categories.

Identifying the Two Primary Types of Chi-Square Tests

The method used to calculate the expected counts differs slightly depending on the specific application of the Chi-Square test being employed. Statisticians primarily recognize two main types, each serving a distinct purpose in analyzing categorical data.

The two common types of Chi-Square tests requiring the calculation of expected counts are:

Chi-Square Goodness of Fit Test: This test is employed to determine whether a single categorical variable follows a pre-specified or hypothesized distribution. For instance, testing if the distribution of colors sold matches the manufacturer's stated proportions, or if gender distribution in a class is uniform.

Chi-Square Test of Independence (or Association): This test is used to determine whether there is a statistically significant association or relationship between two independent categorical variables. For example, testing if political preference is independent of geographic region, or if product satisfaction is independent of the point of sale.

Understanding which test is appropriate for the research question is the first step toward correctly determining the expected frequencies, as the underlying formulas reflect different null hypotheses--one involving a comparison to theoretical proportions, and the other involving comparisons of proportions across different groups.

Example 1: Expected Counts for Chi-Square Goodness of Fit Test

The Chi-Square Goodness of Fit test requires us to define the hypothesized distribution before calculating the expected counts. The null hypothesis (H_0) for this test assumes that the observed data fits the specified theoretical distribution. The expected count calculation, therefore, is driven by the overall sample size and the proportion (or percentage) specified by the null hypothesis for each category.

Consider the scenario of a store owner who claims that customer traffic is equally distributed across the five weekdays (Monday through Friday). The null hypothesis states that the proportion of customers on Monday is 20%, Tuesday is 20%, and so on. If the total number of customers observed during the week is N , then the expected number of customers for any given day (E_i) must be N multiplied by the expected proportion (P_i).

The generic formula for the expected count in a Goodness of Fit test is:

$$\text{Expected count } (E_i) = \text{Expected proportion } (P_i) \times \text{Total count } (N)$$

This method ensures that the sum of all expected counts is equal to the total sample size, maintaining consistency with the overall data collected, while establishing the theoretical

frequencies under the condition of the hypothesized distribution.

Step-by-Step Goodness of Fit Calculation

Let us apply this formula to the store owner's claim. The owner observes the customer count over one full week and records the following observed counts:

Day	Customer Count
Monday	50
Tuesday	60
Wednesday	40
Thursday	47
Friday	53
Total	250

The total number of customers for the week (the total count, N) is $45 + 55 + 50 + 60 + 40 = \text{250}$.

Since the owner claims an equal number of customers each day, and there are five weekdays, the expected proportion (P_i) for each day is $\frac{1}{5}$, or 20% (0.20). Using the formula, we can calculate the expected count for any single day:

Expected count = Expected percentage \times Total count

Expected count = 0.20×250 total customers = 50

This means that if the store owner's claim (the null hypothesis) were perfectly true, we would expect to see exactly 50 customers each day. The expected count is uniform across all categories when the claim is one of equal distribution. The resulting table showing both the observed and expected counts looks like this:

Day	Customer Count	Expected Count
Monday	50	50
Tuesday	60	50
Wednesday	40	50
Thursday	47	50
Friday	53	50
Total	250	

With both the observed and expected counts established, the next stage of the analysis is to compute the Chi-Square test statistic. This calculated statistic, combined with the degrees of freedom, allows us to find the corresponding p-value, which dictates whether there is enough statistical evidence to reject the store owner's claim of uniform daily traffic. Detailed instructions on performing this exact analysis often require statistical software or specialized tools like Excel.

Example 2: Expected Counts for Chi-Square Test of Independence

When dealing with a Chi-Square Test of Independence, the structure changes from a single list of categories to a two-dimensional contingency table. Here, the null hypothesis (H_0) asserts that the two categorical variables are independent of one another. Therefore, the expected counts must reflect the assumption that the proportions observed in the total sample hold true across all subgroups defined by the other variable.

The method for calculating the expected count for any specific cell in a contingency table relies on the marginal totals--the row sums and the column sums. The calculation is based on the probability of two independent events occurring simultaneously. If two events (A and B) are independent, the probability of both occurring ($P(A \cap B)$) is the product of their individual probabilities ($P(A)$ times $P(B)$). Applying this to frequencies and the total sample size (N), we derive the standard formula:

$$\text{Expected count (E)} = (\text{Row Sum} \times \text{Column Sum}) / \text{Table Sum (N)}$$

This formula ensures that the expected frequency for a cell is proportional to the marginal probabilities of the category defined by that row and the category defined by that column, assuming no interaction between the two variables.

Applying the Formula: Test of Independence Walkthrough

Suppose a simple random sample of 500 voters is surveyed regarding their gender and political party preference. The initial survey results, or the observed counts, are displayed in the contingency table below:

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

To find the expected count for a specific cell, say "Male Republicans," we must first identify the relevant marginal totals. The row sum for Males is 250, the column sum for Republicans is 230, and the total sample size (N) is 500.

Using the formula for the Expected count:

Expected count (Male Republican) = (Row sum for Male \times Column sum for Republican) / Total sum

Expected count (Male Republican) = $(230 \times 250) / 500$

Expected count (Male Republican) = $57,500 / 500 = 115$

This value of 115 is the number of Male Republicans we would expect to see in the sample if gender and party preference were truly independent variables. We repeat this calculation for every cell in the table. For instance, the expected value for Female Democrats would be $(270 \times 185) / 500 = 135$:

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

Expected Male Republicans = $(230 \times 250) / 500 = 115$

By systematically applying this formula to all cells in the contingency table, we derive a complete set of expected counts. Notice that the marginal totals for the expected count table will match the marginal totals of the observed count table, a crucial internal check for calculation accuracy.

Observed Frequencies

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

Expected Frequencies

	Republican	Democrat	Independent	Total
Male	115	92.5	42.5	250
Female	115	92.5	42.5	250
Total	230	185	85	500

Interpreting the Discrepancy and Finalizing the Test

Once both the observed and the expected counts are available, the primary goal of the Chi-Square test can be executed: quantifying the total difference. The test statistic aggregates the squared deviations between O_{ij} and E_{ij} , weighted by the expected frequency itself. This weighting is important because large differences in cells with small expected frequencies contribute more heavily to the total statistic, reflecting a greater deviation from independence than the same numerical difference in a cell with a very large expected frequency.

For the Test of Independence example above, a large Chi-Square statistic would indicate that the observed distribution of voters across gender and political parties is significantly different from what we would expect if the two variables were independent. This allows us to reject the null hypothesis and conclude that there is a statistically significant association between gender and political party preference.

The final outcome of the test hinges on comparing the calculated Chi-Square value to a critical value from the Chi-Square distribution, or more commonly today, determining the corresponding p-

value. If the p-value is below the predetermined significance level (e.g., $\alpha = 0.05$), we conclude that the observed differences are unlikely to have occurred by random chance alone, confirming a genuine relationship between the variables.

Note: Specialized software is often used to automate the calculation of these expected counts and the subsequent Chi-Square statistic and p-value. This ensures accuracy, especially in large contingency tables. For instance, detailed guides explain how to perform this exact Chi-Square Test of Independence in Excel.

Practical Considerations and Assumptions

While the methodology for calculating expected counts is straightforward, the statistical interpretation relies on several key assumptions. The most critical, as mentioned earlier, relates to the magnitude of the expected frequencies. If too many cells have low expected counts (typically $E < 5$), the shape of the test statistic's distribution deviates from the theoretical Chi-Square distribution, potentially leading to incorrect inferences.

Another crucial assumption is that the data must be collected via a simple random sample and that the observations must be independent. For example, in the voter survey, each individual's preference must be independent of every other individual surveyed. If the sample were clustered or dependent (e.g., surveying members of the same family), the interpretation of the resulting p-value would be compromised.

Finally, the Chi-Square test only indicates whether an association exists; it does not measure the strength or direction of that association. Analysts often rely on secondary measures, such as Cramer's V or the Phi coefficient, to quantify the practical significance or magnitude of the relationship observed, complementing the hypothesis testing provided by the comparison of observed and expected counts.

Additional Resources for Chi-Square Analysis

The following resources provide additional information about Chi-Square tests and advanced techniques for handling categorical data:

Detailed statistical textbooks covering non-parametric statistics.

Official documentation for statistical software packages regarding contingency analysis.

Academic journals specializing in applied statistical methodology.