**Chi-square Test — How to calculate Chi-square using Formula & Python Implementation**



Photo by [Kelly Sikkema](https://unsplash.com/@kellysikkema?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral" \t "_blank)

**1. Type of Test**

Chi-square is a non-parametric test, i.e., it does not require normal distribution or variance assumptions about the populations from which the samples are drawn.

**2. Purpose**

The general purpose of the Chi-square test is to compare discrete categorical data (count data). For example, a product may be categorized into two categories, such as defective/non-defective, or in more than two categories, such as excellent, good, fair, and poor. Chi-square tests are ideally suited to a data set in which both the variables to be compared are categorical. The Chi-square test compares observed values to theoretically expected values.

**3. Scope**

Non-parametric tests like the chi-square test are less powerful than parametric tests, i.e., they are less likely to reject the null hypothesis, especially when it is false. A few application areas include:

1. The chi-square test for checking the goodness of fit is utilized to check whether there are differences between the observed (experimental) value and the expected (theoretical) value. It establishes whether the distribution of the data remains similar when compared to the past
2. The Chi-square test for the independence of two attributes is used to check whether the two characteristics are independent. It is used to determine whether a categorical outcome variable (Y) is related or associated with another categorical predictor variable (X)

**4. Assumptions**

1. The sample is drawn randomly from the population. This is required to generalize the result to the entire population
2. Data is to be reported in raw frequencies (counts, not percentages)
3. Observations are independent
4. Variables are mutually exclusive (individuals cannot be assigned to more than one category) and exhaustive (include all possible contexts or categories)
5. Observed frequencies are not too small (n must be relatively large) in a 2 x 2 table; Chi-Square should not be used if n is less than 20 or any of the individual cell counts is less than 5

**5. Methodology**

Several types of chi-square tests depend on the data collection and tests required. We’ll look at the simplest example, which involves a 2 x 2 contingency table. Suppose you conducted a pharmaceutical trial on a group of patients suffering from Multiple Myeloma and hypothesized that the patients receiving the drug would show increased remission rates compared to those not. Let’s assume that the following data was collected.

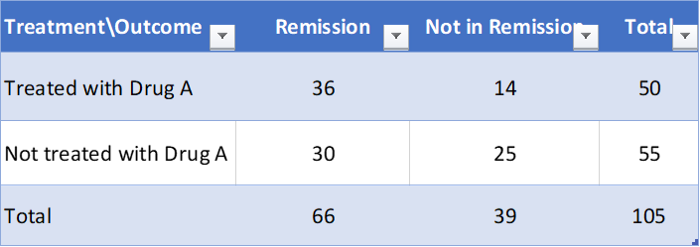


Figure 1. Contingency table using hypothetical data. Image — Prepared by the Autor using Excel.

**H0 (Null Hypothesis):** The distribution of observations across contexts is equal (observed frequencies do not depart from predicted frequencies), i.e., drug treatment does not affect the proportion of patients who went into remission.

**HA (Alternate Hypothesis):** The proportion of patients who went into remission is connected with drug treatment, demonstrating that observations are not evenly distributed over contexts (observed frequencies do depart from predicted frequencies).

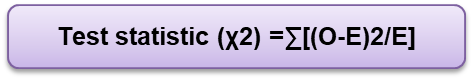


Figure 1. The formula for the test statistic. Image — Prepared by the Autor using Powerpoint.

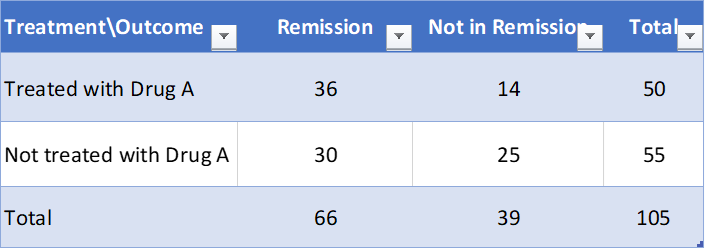
Where **O** is the *observed frequency* and **E** is the *expected frequency*.

**6. Calculating Chi-Square Statistics**

**6.1 Calculating Expected Frequency**

The Chi-square statistics for the table outlined below are calculated as follows:

1. Calculate the expected frequencies in each cell. Each cell’s expected frequency(E) is given by (row total x column total)/grand total
2. The row total represents the total number of observations for a given context (e.g., how many patients are treated with Drug A? — 50 patients)
3. The column total represents the total number of observations for a given dependent variable (e.g., how many patients are in remission? — 66 patients)
4. The grand total represents the total number of observations (e.g., the total number of patients participating in the Study — 105 patients)



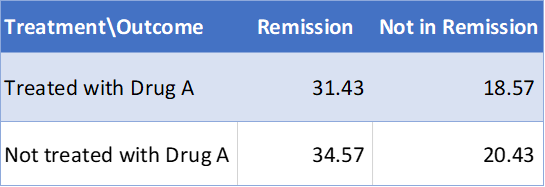


Figure 3. Contingency table and Expected Value using hypothetical data. Image — Prepared by the Autor using Excel.

The expected value of Treated with Drug A x Remission is calculated as (50 x 66)/ 105 = 31.43. 2. The observed frequency(O) is the data collected and reflects the actual number, i.e., for Treated with Drug A x Remission, it is 36.

**6.2 Calculating Test Statistics**

Calculate the statistic using the formula χ2=∑[(O-E)2/E].  
χ2=(36–31.42)²/31.42 + (14–18..57)²/18.57 + (30–34.57)²/34.57 + (25–20.43)²/20.43 = 3.418

**6.3 Calculating Degrees of Freedom**

Degrees of freedom = (# columns (or) dependent variable — 1)\*(# rows (or) categories — 1) =(2–1)(2–1)=1

**6.4 Odds Ratio**

It is another measure of association for 2 × 2 contingency tables. It occurs as a parameter in the most important model for categorical data. For a probability of success π, the odds of success are defined to be; odds = π/(1 − π).

The odds are nonnegative, with a value greater than 1.0 when success is more likely than failure. For instance, if π = 0.75, then the odds of success equal 0.75/0.25 = 3. When odds = 4.0, success is four times as likely as failure.

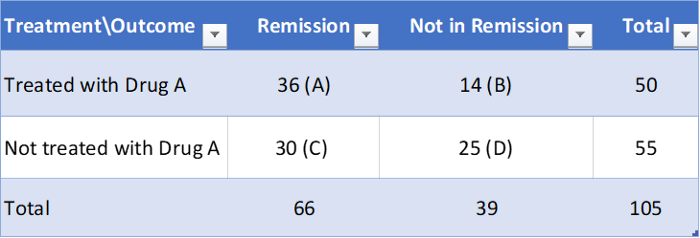


Figure 4. Reference to the Odds Ration calculation below. Image — Prepared by the Autor using Excel.

Therefore, Odds Ratio = (Odds of success in category A) / (Odds of success in category B). In the example, the Odds of remission for treated patients = 36/14 = 2.57, and the odds of remission for non-treated patients = 30/25 =1.2. The odds ratio is given by the ratio of these two odds, i.e., 2.57/1.2 = 2.14. This ratio signifies that a treated patient was 2.14 times more likely to find remission than a non-treated patient.  
  
**Shortcut to calculate to odds ratio:**  
Let us consider the above example, where we have to calculate the odds ratio of remission in the category treated over that of not-treated. In the above table, each cell value has been represented by an alphabet.  
Then, the Odds ratio will be (A\*D)/(B\*C). For our example, odds ratio= (36\*25)/(14\*30) = 2.14.

**6.5 Interpretation**

Compare the calculated χ2 statistic to a critical χ2 value to determine whether to reject the null hypothesis:

1. If calculated χ2 >critical χ2, p ≤0.05 — indicates that there is significant statistical evidence in support of rejecting the null hypothesis. There is less than or equal to a 5% probability that we could obtain this result by chance, which is an acceptable error level for experiments
2. If calculated χ2 ≤ critical χ2, p> 0.05 indicates no significant statistical evidence supporting rejecting the null hypothesis. There is greater than a 5% probability that we could obtain this result by chance, which exceeds the acceptable level of error for any experiment

In our example, we now have our chi-square statistic ( χ2 = 3.418), our predetermined alpha level of significance (0.05), and our degrees of freedom (df = 1). Entering the Chi-square distribution table with 1 degree of freedom and reading along the row, we find our value of χ2 (3.418) lies between 2.706 and 3.841. The corresponding probability for which is 0.07. The p-value hence is greater than the accepted significance level of 0.05 with a 95% Confidence Interval (CI) (i.e., p > 0.05); we don’t have enough evidence to reject the null hypothesis. In simpler terms, there is no significant difference between the proportions of patients in remission across the two treatment groups. The p value can be looked up using the link below.

**[Table: Chi-Square Probabilities](https://people.richland.edu/james/lecture/m170/tbl-chi.html" \t "_blank)**

[The areas given across the top are the areas to the right of the critical value. To look up an area on the left…](https://people.richland.edu/james/lecture/m170/tbl-chi.html" \t "_blank)

[people.richland.edu](https://people.richland.edu/james/lecture/m170/tbl-chi.html" \t "_blank)

**7. Implementation in Python**

Select two categorical columns from any dataset and then create a crosstab. The output of the cross tab is then used with chi2\_contingency() to get the p value for rejecting or accepting null hypothesis.

#----------step 1: Extract required columns, ideally it should be two columns at a timedata\_interim=df[['column1', 'column2']]#----------step 2: Create a cross tab which is a contigency tabledata\_cont=pd.crosstab(data\_interim['column1'], data\_interim['column2'])data\_contfrom scipy.stats import chi2\_contingencystat, p, dof, expected = chi2\_contingency(data\_cont)#-----------interpret p-value  
alpha = 0.05print("p value is " + str(p))if p <= alpha:  
   
 print('Dependent (reject H0)')  
   
else:  
 print('Independent (H0 holds true)')

**8. Conclusion**

In a pharmaceutical landscape, especially when working with patient data, physician data, or sales force data, there is a need to understand the correlation between different variables, specifically those categorical. Categorical attributes in the pharmaceutical landscape range between target segments, patient load, gender, demographics, call outcomes, call quality, discussion themes for a sales rep visit, and many more. When we examine fewer categories, the Chi-square test can allow us to establish relationships between these variables, which in return can be used to dive deep into variables or drivers of choice.