# 

Contents

[Chi – square test 1](#_Toc518388172)

[Chi-square goodness-of-fit test (also known as the Pearson's chi-squared test or c2 test) 1](#_Toc518388173)

[Chi-Square Test for homogeneity 2](#_Toc518388174)

[Chi-Square Test for Independence 2](#_Toc518388175)

# Chi – square test

* Goodness of fit
* Homogeneity
* independence

Use this analysis to test how well a sample of categorical data fits a theoretical distribution.

Chi-square tests of association and independence

The calculations for these tests are the same, but the question you're trying to answer may be different.

Test of **association**: Use a test of association to determine whether one variable is associated with a different variable. For example, determine whether the sales for different colors of cars depends on the city where they are sold.

Test of **independence**: Use a test of independence to determine whether the observed value of one variable depends on the observed value of a different variable. For example, determine whether the candidate that a person votes for is independent of the voter's gender.

# Assumptions

**1. Sample size assumption:**

The chi-square test can be used to determine differences in proportions using a two-by-two contingency table. It is however important to understand that the chi-square tests yields only an approximate p-value, on which a correction factor is then applied. This only works well when your datasets are large enough. When sample sizes are small, as indicated by more than 20% of the contingency cells having expected values ­< 5 a **Fisher's exact test**  maybe more appropriate. This test is one of a class of [“exact tests](http://en.wikipedia.org/wiki/Exact_test)”, because the significance of the deviation from a “[null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis)” can be calculated exactly, rather than relying on an approximation.

**2. Independence assumption:**

Secondly, the chi-square test cannot be used on correlated data. When you are looking to test differences in proportions among matched pairs in a before/after scenario, an appropriate choice would be the **McNemar's**test. In essence, it is a chi-square goodness of fit test on the two discordant cells, with a null hypothesis stating that 50% of the changes (agreements or disagreements) go in each direction. This test requires the same subjects to be included in the before and after measurements i.e. the pairs should be matched one-on-one.

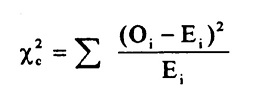
# When to use chi square test

[Market researchers](https://www.surveygizmo.com/market-research-survey-solutions/) use the Chi-Square test when they find themselves in one of the following situations:

They need to estimate how closely an observed distribution matches an expected distribution. This is referred to as a “goodness-of-fit” test.

They need to estimate whether two random variables are independent.

# Chi square formula



0 = observed

E = expected

# Chi-square goodness-of-fit test (also known as the Pearson's chi-squared test or c2 test)

It allows you to compare categorical data with a theoretical distribution.

The test is applied when you have one [categorical variable](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

It has the null hypothesis that the data follows the specified distribution, and the alternative hypothesis that it does not.

it tests whether the distribution of sample categorical data matches an expected distribution. For example, you could use a chi-squared goodness-of-fit test to check whether the race demographics of members at your church or school match that of the entire U.S. population or whether the computer browser preferences of your friends match those of Internet uses as a whole.

The test is only suitable if sufficient data is available, which is commonly defined as each category having an expected frequency (under the null hypothesis) of at least five. The test should not be confused with the chi-square test of association (see the next section in this chapter), which helps to determine whether two or more categorical variables are associated.

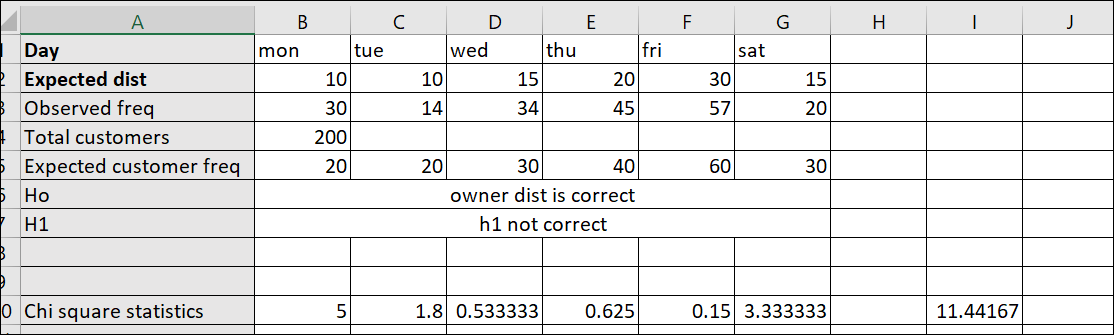
The chi-square goodness-of-fit test and the chi-square test of association are both forms of the Pearson chi-square test, but they answer different questions about the data.

To illustrate the difference, if you had recorded the results of a series of six-sided die rolls

and wanted to use this data to determine whether your die was fair, you would use the chi-square goodness-of-fit test.

The null hypothesis for the test would be that each of the six sides is equally likely to be rolled, with probability 1/6. However, if you had recorded the name of the person rolling the die as well as the result of the die roll and you wanted to determine whether there was any relationship between the result of the roll and the person that rolled the die, you would use the chi-square test of association.

**Example**



If this has a chi square distribution probability of getting a result this extreme is p value . If the result that we obtained is more extreme than our critical value then we would rejct the ull hypothesis

degree of freedom here is 5 since if we have all the information for 5 then we can figureout degree of freedom for the last

Then we will calculate critical chi square level

with degree of freedom =5 and significance level of 5%

critical chi square value was 11.07(which

Result we got with our statistics has less probability than the significant level then we will reject the null hypothesis

**Example 2**

 you could use a chi-squared goodness-of-fit test to check whether the race demographics of members at your church or school match that of the entire U.S. population or whether the computer browser preferences of your friends match those of Internet uses as a whole.

Let's generate some fake demographic data for U.S. and Minnesota and walk through the chi-square goodness of fit test to check whether they are different:

|  |
| --- |
| national = pd.DataFrame(["white"]\*100000 + ["hispanic"]\*60000 +\  ["black"]\*50000 + ["asian"]\*15000 + ["other"]\*35000)    minnesota = pd.DataFrame(["white"]\*600 + ["hispanic"]\*300 + \  ["black"]\*250 +["asian"]\*75 + ["other"]\*150)  national\_table = pd.crosstab(index=national[0], columns="count")  minnesota\_table = pd.crosstab(index=minnesota[0], columns="count")  print( "National")  print(national\_table)  print(" ")  print( "Minnesota")  print(minnesota\_table) |

\*Note: The chi-squared test assumes none of the expected counts are less than 5.

Similar to the t-test where we compared the t-test statistic to a critical value based on the t-distribution to determine whether the result is significant, in the chi-square test we compare the chi-square test statistic to a critical value based on the [chi-square distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution). The scipy library shorthand for the chi-square distribution is chi2. Let's use this knowledge to find the critical value for 95% confidence level and check the p-value of our result:

|  |
| --- |
| crit = stats.chi2.ppf(q = 0.95, *# Find the critical value for 95% confidence\**  df = 4) *# Df = number of variable categories - 1*  print("Critical value")  print(crit)  p\_value = 1 - stats.chi2.cdf(x=chi\_squared\_stat, *# Find the p-value*  df=4)  print("P value")  print(p\_value) |

In [4]:

observed = minnesota\_table

national\_ratios = national\_table/len(national) *# Get population ratios*

expected = national\_ratios \* len(minnesota) *# Get expected counts*

chi\_squared\_stat = (((observed-expected)\*\*2)/expected).sum()

print(chi\_squared\_stat)

# Chi-Square Test for homogeneity

This lesson explains how to conduct a **chi-square test of homogeneity**. The test is applied to a single [categorical variable](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from two different populations. It is used to determine whether frequency counts are distributed identically across different populations.

For example, in a survey of TV viewing preferences, we might ask respondents to identify their favorite program. We might ask the same question of two different populations, such as males and females. We could use a chi-square test for homogeneity to determine whether male viewing preferences differed significantly from female viewing preferences

# Chi-Square Test for Independence

The test is applied when you have two [categorical variables](http://stattrek.com/Help/Glossary.aspx?Target=Categorical%20variable) from a single population. It is used to determine whether there is a significant association between the two variables.

The chi-square independence test is a procedure for testing if two categorical variables are related in some population.

The [null hypothesis](https://www.spss-tutorials.com/null-hypothesis/) for a chi-square independence test is thattwo categorical variables are independent in some population.

**statistical independence** means that the frequency distribution of a variable is the same for all levels of some other variable.

# Disadvantages

* Can't use percentages
* Data must be numerical
* Categories of 2 are not good to compare
* The number of observations must be 20+
* The test becomes invalid if any of the expected values are below 5
* Quite complicated to get right - difficult formula