**A Gentle Introduction to the Chi-Squared Test for Machine Learning**

by [**Jason Brownlee**](https://machinelearningmastery.com/author/jasonb/) on June 15, 2018 in [**Statistics**](https://machinelearningmastery.com/category/statistical-methods/)

Tweet  **Share** **Share**

Last Updated on October 31, 2019

A common problem in applied machine learning is determining whether input features are relevant to the outcome to be predicted.

This is the problem of feature selection.

In the case of classification problems where input variables are also categorical, we can use statistical tests to determine whether the output variable is dependent or independent of the input variables. If independent, then the input variable is a candidate for a feature that may be irrelevant to the problem and removed from the dataset.

The Pearson’s chi-squared statistical hypothesis is an example of a test for independence between categorical variables.

In this tutorial, you will discover the chi-squared statistical hypothesis test for quantifying the independence of pairs of categorical variables.

After completing this tutorial, you will know:

* Pairs of categorical variables can be summarized using a contingency table.
* The chi-squared test can compare an observed contingency table to an expected table and determine if the categorical variables are independent.
* How to calculate and interpret the chi-squared test for categorical variables in Python.

Discover statistical hypothesis testing, resampling methods, estimation statistics and nonparametric methods [in my new book](https://machinelearningmastery.com/statistics_for_machine_learning/), with 29 step-by-step tutorials and full source code.

Let’s get started.

* **Update Jun/2018**: Minor typo fix in the interpretation of the critical values from the test (thanks Andrew).
* **Update Oct/2019**: Fixed language around factor/levels (thanks Marc)



A Gentle Introduction to the Chi-Squared Test for Machine Learning  
Photo by [NC Wetlands](https://www.flickr.com/photos/ncwetlands/38431877722/), some rights reserved

**Tutorial Overview**

This tutorial is divided into 3 parts; they are:

1. Contingency Table
2. Pearson’s Chi-Squared Test
3. Example Chi-Squared Test

**Need help with Statistics for Machine Learning?**

Take my free 7-day email crash course now (with sample code).

Click to sign-up and also get a free PDF Ebook version of the course.

[**Download Your FREE Mini-Course**](https://machinelearningmastery.lpages.co/leadbox/142f75173f72a2%3A164f8be4f346dc/5750943224168448/)

**Contingency Table**

A categorical variable is a variable that may take on one of a set of labels.

An example might be sex, which may be summarized as male or female. The variable or factor is ‘*sex*‘ and the labels or levels of the variable are ‘*male*‘ and ‘*female*‘ in this case.

We may wish to look at a summary of a categorical variable as it pertains to another categorical variable. For example, sex and interest, where interest may have the labels ‘*science*‘, ‘*math*‘, or ‘*art*‘. We can collect observations from people collected with regard to these two categorical variables; for example:



|  |  |
| --- | --- |
| 1  2  3  4  5  6 | Sex, Interest  Male, Art  Female, Math  Male, Science  Male, Math  ... |

We can summarize the collected observations in a table with one variable corresponding to columns and another variable corresponding to rows. Each cell in the table corresponds to the count or frequency of observations that correspond to the row and column categories.

Historically, a table summarization of two categorical variables in this form is called a [contingency table](https://en.wikipedia.org/wiki/Contingency_table).

For example, the *Sex=rows* and *Interest=columns* table with contrived counts might look as follows:



|  |  |
| --- | --- |
| 1  2  3 | Science, Math, Art  Male         20,      30,    15  Female       20,      15,    30 |

The table was called a contingency table, by Karl Pearson, because the intent is to help determine whether one variable is contingent upon or depends upon the other variable. For example, does an interest in math or science depend on gender, or are they independent?

This is challenging to determine from the table alone; instead, we can use a statistical method called the Pearson’s Chi-Squared test.

**Pearson’s Chi-Squared Test**

The Pearson’s Chi-Squared test, or just Chi-Squared test for short, is named for Karl Pearson, although there are variations on the test.

The Chi-Squared test is a [statistical hypothesis test](https://machinelearningmastery.com/statistical-hypothesis-tests/) that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.

Given the Sex/Interest example above, the number of observations for a category (such as male and female) may or may not the same. Nevertheless, we can calculate the expected frequency of observations in each Interest group and see whether the partitioning of interests by Sex results in similar or different frequencies.

The Chi-Squared test does this for a contingency table, first calculating the expected frequencies for the groups, then determining whether the division of the groups, called the observed frequencies, matches the expected frequencies.

The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same.

*When observed frequency is far from the expected frequency, the corresponding term in the sum is large; when the two are close, this term is small. Large values of X^2 indicate that observed and expected frequencies are far apart. Small values of X^2 mean the opposite: observeds are close to expecteds. So X^2 does give a measure of the distance between observed and expected frequencies.*

— Page 525, [Statistics](http://amzn.to/2u44zll), Fourth Edition, 2007.

The variables are considered independent if the observed and expected frequencies are similar, that the levels of the variables do not interact, are not dependent.

*The chi-square test of independence works by comparing the categorically coded data that you have collected (known as the observed frequencies) with the frequencies that you would expect to get in each cell of a table by chance alone (known as the expected frequencies).*

— Page 162, [Statistics in Plain English](http://amzn.to/2IFyS4P), Third Edition, 2010.

We can interpret the test statistic in the context of the chi-squared distribution with the requisite number of degress of freedom as follows:

* **If Statistic >= Critical Value**: significant result, reject null hypothesis (H0), dependent.
* **If Statistic < Critical Value**: not significant result, fail to reject null hypothesis (H0), independent.

The degrees of freedom for the chi-squared distribution is calculated based on the size of the contingency table as:



|  |  |
| --- | --- |
| 1 | degrees of freedom: (rows - 1) \* (cols - 1) |

In terms of a p-value and a chosen significance level (alpha), the test can be interpreted as follows:

* **If p-value <= alpha**: significant result, reject null hypothesis (H0), dependent.
* **If p-value > alpha**: not significant result, fail to reject null hypothesis (H0), independent.

For the test to be effective, at least five observations are required in each cell of the contingency table.

Next, let’s look at how we can calculate the chi-squared test.

**Example Chi-Squared Test**

The Pearson’s chi-squared test for independence can be calculated in Python using the [chi2\_contingency() SciPy function](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html).

The function takes an array as input representing the contingency table for the two categorical variables. It returns the calculated statistic and p-value for interpretation as well as the calculated degrees of freedom and table of expected frequencies.



|  |  |
| --- | --- |
| 1 | stat, p, dof, expected = chi2\_contingency(table) |

We can interpret the statistic by retrieving the [critical value](https://machinelearningmastery.com/critical-values-for-statistical-hypothesis-testing/) from the chi-squared distribution for the probability and number of degrees of freedom.

For example, a probability of 95% can be used, suggesting that the finding of the test is quite likely given the assumption of the test that the variable is independent. If the statistic is less than or equal to the critical value, we can fail to reject this assumption, otherwise it can be rejected.



|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | # interpret test-statistic  prob = 0.95  critical = chi2.ppf(prob, dof)  if abs(stat) >= critical:  print('Dependent (reject H0)')  else:  print('Independent (fail to reject H0)') |

We can also interpret the p-value by comparing it to a chosen significance level, which would be 5%, calculated by inverting the 95% probability used in the critical value interpretation.



|  |  |
| --- | --- |
| 1  2  3  4  5  6 | # interpret p-value  alpha = 1.0 - prob  if p <= alpha:  print('Dependent (reject H0)')  else:  print('Independent (fail to reject H0)') |

We can tie all of this together and demonstrate the chi-squared significance test using a contrived contingency table.

A contingency table is defined below that has a different number of observations for each population (row), but a similar proportion across each group (column). Given the similar proportions, we would expect the test to find that the groups are similar and that the variables are independent (fail to reject the null hypothesis, or H0).



|  |  |
| --- | --- |
| 1  2 | table = [ [10, 20, 30],  [6,  9,  17]] |

The complete example is listed below.



|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25 | # chi-squared test with similar proportions  from scipy.stats import chi2\_contingency  from scipy.stats import chi2  # contingency table  table = [ [10, 20, 30],  [6,  9,  17]]  print(table)  stat, p, dof, expected = chi2\_contingency(table)  print('dof=%d' % dof)  print(expected)  # interpret test-statistic  prob = 0.95  critical = chi2.ppf(prob, dof)  print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))  if abs(stat) >= critical:  print('Dependent (reject H0)')  else:  print('Independent (fail to reject H0)')  # interpret p-value  alpha = 1.0 - prob  print('significance=%.3f, p=%.3f' % (alpha, p))  if p <= alpha:  print('Dependent (reject H0)')  else:  print('Independent (fail to reject H0)') |

Running the example first prints the contingency table. The test is calculated and the degrees of freedom (*dof*) is reported as 2, which makes sense given:



|  |  |
| --- | --- |
| 1  2  3  4 | degrees of freedom: (rows - 1) \* (cols - 1)  degrees of freedom: (2 - 1) \* (3 - 1)  degrees of freedom: 1 \* 2  degrees of freedom: 2 |

Next, the calculated expected frequency table is printed and we can see that indeed the observed contingency table does appear to match via an eyeball check of the numbers.

The critical value is calculated and interpreted, finding that indeed the variables are independent (fail to reject H0). The interpretation of the p-value makes the same finding.



|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12 | [[10, 20, 30], [6, 9, 17]]    dof=2    [[10.43478261 18.91304348 30.65217391]  [ 5.56521739 10.08695652 16.34782609]]    probability=0.950, critical=5.991, stat=0.272  Independent (fail to reject H0)    significance=0.050, p=0.873  Independent (fail to reject H0) |

**Extensions**

This section lists some ideas for extending the tutorial that you may wish to explore.

* Update the chi-squared test to use your own contingency table.
* Write a function to report on the independence given observations from two categorical variables
* Load a standard machine learning dataset containing categorical variables and report on the independence of each.