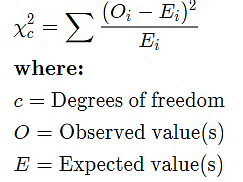
**The Anatomy of Chi Squared Test**

In many real-world problems, we arrive at a conjecture where quantifying an estimate from the data becomes arduous. A statistical test provides a system to make significant decisions about a process. The intent is to determine whether there is enough evidence to "reject" a hypothesis about the process.

In statistical jargon, chi-Squared test, also written as ***χ2*** test, is basically a hypothesis test to determine whether there lies any statistical significance between the expected frequencies and the perceived frequencies in one or more categorical features. As usual, there are two separate hypotheses involved; *null hypothesis* which states that there is no definite relationship between the two categorical features and *alternate hypothesis* which states that there is certainly a relationship between the categorical features. Chi-square test is primarily used in two cases *a) Compare the proportions or frequencies of categorical data (Goodness-of-Fit). b) Compare the independence of two characteristics*.

*Take for instance*, a house rent prediction use case where we need to predict the price based on certain features in the dataset. It consists of, let’s say, categorical features like *‘is cat allowed’*, *’is smoking allowed’*. These features will have binary or categorical values, and determining whether it affects the price of house becomes quite difficult. In such cases we can use this test to determine the significance of the features.

The formula of Chi-Squared test is:



*credit:* [*www.investopedia.com/terms/c/chi-square-statistic.asp*](https://www.investopedia.com/terms/c/chi-square-statistic.asp)

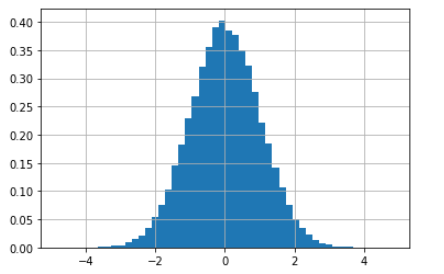
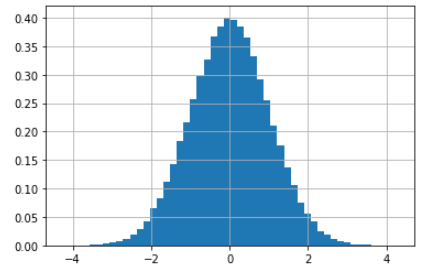
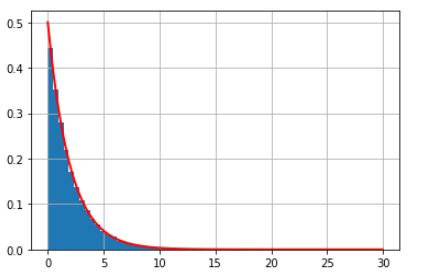
I will explain the significance of the terms in the formula with the help of an example after I give a brief explanation about the jargons involved in it. Ok, now we need to understand the important aspects involved in the test, namely *Chi squared distribution, degrees of freedom* and a *critical value*.

**Chi Squared Distribution:** Consider a *standard normal distribution (Z1 ~ N (0,1)).* If we take the square of this distribution, we eventually get a *chi-squared distribution* with 1 degrees of freedom. Now, if there are two different standard normal distributions, *Z1, Z2,* the chi-squared distribution will be the sum of squares of these standard normal distributions, i.e. ***χ22 ~ Z12 + Z22*** and the degrees of freedom will be 2. So to sum it up in statistical sense, *the chi-square distribution  with****k***[*degrees of freedom*](https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics))*is the distribution of a* ***sum of the squares*** *of k*[*independent*](https://en.wikipedia.org/wiki/Independence_(probability_theory))[*standard normal*](https://en.wikipedia.org/wiki/Standard_normal)*random variables*.

So, if *Z1, ..., Zk are*[*independent*](https://en.wikipedia.org/wiki/Independence_(probability_theory))*,*[*standard normal*](https://en.wikipedia.org/wiki/Standard_normal)*random variables*, then the sum of their squares,

Q= ~ χ***k2***

{\displaystyle Q\ =\sum \_{i=1}^{k}Z\_{i}^{2},}



**Z1**

**Z2**

***χ22***

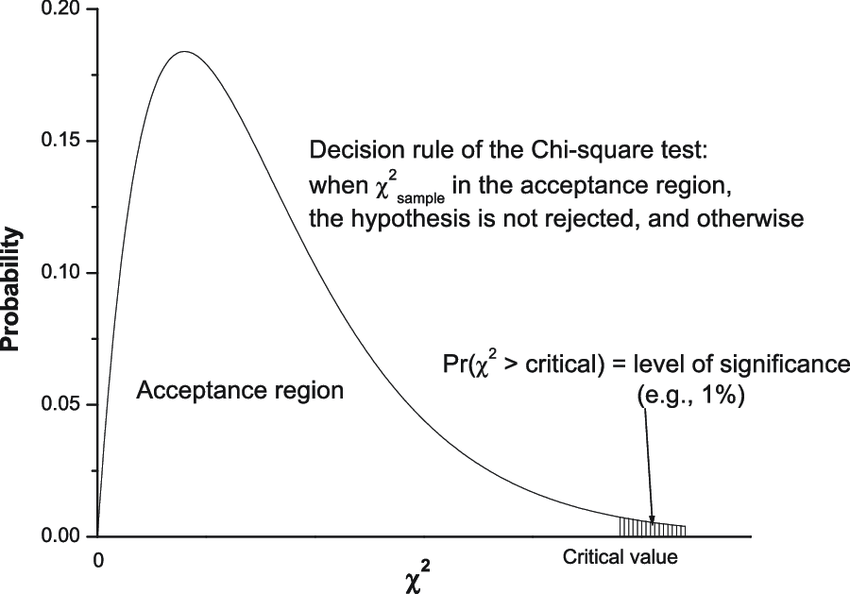
The pictures above describe the probability distribution functions of *two standard normal distributions* and *chi-squared distribution* with degree of freedom 2; formed with these normal distributions.

**Degrees of freedom**: In a random sample, *degrees of freedom* refer to the maximum number of independent values which help in calculating an estimate i.e. the values which are free to vary and does not depend on other variables in the sample. It is given by

*Degrees of Freedom* = (N – 1), where N is the number of items in the sample

The question arises that why are we subtracting 1 from the items in the sample. Consider an example where we have to select a set of numbers whose mean is 20. Ideally, we can select the set as *(10,20,30), (18,20,22)* and so on. So here as it is noticeable that when we fix the first two numbers there is only one way we arrive at a particular estimate (here mean of 20). So, the ***df*** *(degrees of freedom*) is *3-1=2*. On similar lines the ***df*** can also assume the value as *(r-1)\*(c-1)* where *r = number of rows* and *c = number of columns* when there are column level and row level estimates to be achieved.

**Critical Value:** In statistical jargon, a critical value in a hypothesis testing is a point on the test distribution that is compared to the test statistic to determine whether to reject the null hypothesis or accept it. So, in simple terms, in order to prove any kind of statistical significance, the test statistic (here chi-square test value) should be greater than the critical value. Determining the critical value is whole another post, let’s just see how critical value plays an important role in determining the significance of a hypothesis.



*Credits:* [*https://www.researchgate.net/figure/Schematic-diagram-showing-some-more-essence-of-the-chi-square-test-given-small\_fig2\_276964817*](https://www.researchgate.net/figure/Schematic-diagram-showing-some-more-essence-of-the-chi-square-test-given-small_fig2_276964817)

The above figure tells us that the main goal of the test value is to lie in the most significant area and not in the acceptable region so that we can reject the null hypothesis and accept the alternate hypothesis.

So, to sum it up, we basically first prepare the chi-square distribution based on the degree on freedom and then calculate the chi square test statistic. Next, we calculate the critical value based on the degrees of freedom(a precalculated table is available[*https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm*](https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)) and determine whether the critical value is less than the test statistic or not.

Let’s check out an example. As mentioned, Chi-square is basically used for two reasons, *to check the goodness of fit* and *to test the independence of variables*. In this post we will check the test of independence of variables where we deal with categorical data.

1. *Lay down the Hypotheses*:
2. X and Y are independent of each other(H0)
3. X and Y are dependent on each other(H1)
4. *Prepare the contingency table*- A contingency table is a special type of [frequency distribution table](https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/frequency-distribution-table/), where two or more variables are shown simultaneously.

Variables under consideration

Column totals

Row totals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Category 1 | Category 2 | Category 3 | Total |
| X | 100 | 300 | 400 | 800 |
| Y | 200 | 500 | 600 | 1300 |
| Total | 300 | 800 | 1000 | 2100 |

1. *Calculate the expected value and the test statistic*:

|  |  |  |  |
| --- | --- | --- | --- |
| (800\*300)/2100  Corresponding row total \* column total divided by the grand total | Category 1 | Category 2 | Category 3 |
| X | 114.28571 | 304.7619 | 380.95238 |
| Y | 185.71429 | 495.2381 | 619.04762 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Category 1  (100-114.285)2 / 114.285 | Category 2 | Category 3 |
| X | 1.7857143 | 0.0744048 | 0.952381 |
| Y | 1.0989011 | 0.0457875 | 0.5860806 |

*Plugging the values in the formula (O−E)2 /E**as discussed above.*

Adding all the values, we get ***χk2 = 4.53269.***

1. *Calculate the degrees of freedom* = *(rows-1) (columns-1)* = (2-1) (3-1) = **2**
2. *Check the critical vale based on the degrees of freedom* from the table. *(here it is* ***5.991****)*

So as ***χk2 < critical value, we fail to reject the null hypothesis and we can conclude that the variables are independent of each other.***

So, to sum it up, chi-square test is one of he most important test that helps is checking if the variables are independent of each other. Such a test plays a useful role when dealing with regression analysis where we need to check the dependency of two independent variables; so that maximum variance is captured by the model.

*Notes:*

*\*\* The chi-square distributions are plotted with matplotlib in python*

*\*\*The calculation shown above are calculated in excel.*

*\*\*special thanks to*[***Jason Brownlee***](https://machinelearningmastery.com/author/jasonb/)*and his blogs*