**Chi Square Test**

Bibek Lamsal

CSE student from BVC Engineering College, Amalapuram, Andhra Pradesh

**CONTENTS:**

* **Introduction**
* **Important Terms**
* **Characteristics of the Test**
* **Application of Chi-Square Test**
* **Calculation of Chi-Square Test**
* **Example**
* **Limitation of the test**
* **Application of the test**

**Introduction:**

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Is was developed by Karl Pearson in1900. CHI SQUARE TEST is a non-parametric test not based on any assumption or distribution of any variable. This statistical test follows a specific distribution known as chi square distribution. In general, the test we use to measure the differences between what is observed and what is expected according to an assumed hypothesis is called the chi-square test.

**Important Terms:**

1. Parametric Test:

The test in which, the population constants like mean, std deviation, std error, correlation coefficient, proportion etc. and data tend to follow one assumed distribution such as normal, binomial, poisson etc.

1. Non-Parametric Test:

the test in which no constant of a population is used. Data do not follow any specific distribution and no assumption are made in these tests. E.g. to classify good, better and best we just allocate arbitrary numbers or marks to each category.

1. Hypothesis:

It is a definite statement about the population parameters.

1. Null-Hypothesis:

(H0) states that no association exists between the two cross-tabulated variables in the population, and therefore the variables are statistically independent. E.g. if we want to compare 2 methods method A and method B for its superiority, and if the assumption is that both methods are equally good, then this assumption is called as NULL HYPOTHESIS.

1. Alternative Hypothesis:

(H1) proposes that the two variables are related in the population. If we assume that from 2 methods, method A is superior than method B, then this assumption is called as Alternative Hypothesis.

1. Degree of Freedom:

It denotes the extent of independence (freedom) enjoyed by a given set of observed frequencies Suppose we are given a set of n observed frequencies which are subjected to k independent constraints(restrictions) then,

d.f. = (number of frequencies) – (number of independent constraints on them) In other terms,

df = (r – 1) . (c – 1)

where r = the number of rows c = the number of columns

1. CONTINGENCY TABLE:

When the table is prepared by enumeration of qualitative data by entering the actual frequencies, and if that table represents occurrence of two sets of events, that table is called the contingency table. It is also called as an association table.

**IMPORTANT CHARACTERISTICS OF A CHI SQUARE TEST**:

* This test (as a non-parametric test) is based on frequencies and not on the parameters like mean and standard deviation.
* The test is used for testing the hypothesis and is not useful for estimation.
* This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.
* This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

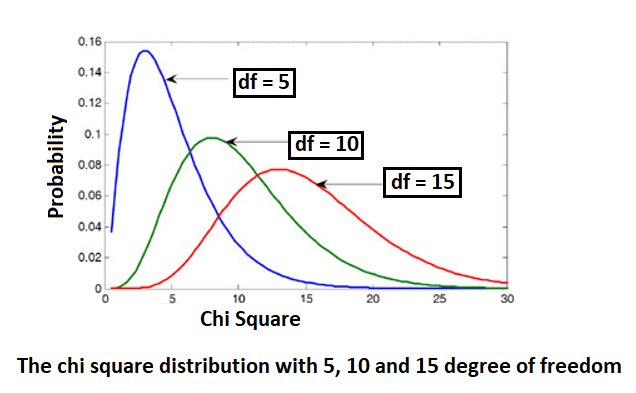
**Chi-Square Distribution**:

If X1, X2,….Xn are independent normal variates and each is distributed normally with mean zero and standard deviation unity, then

X12+X22+……+Xn2= ∑ Xi2

is distributed as chi square (X2) with n degrees of freedom (d.f.) where n is large.

The chi square curve for d.f. N=5,10 and 15 is as follows.

**If degree of freedom > 2:**

Distribution is bell shaped.

**If degree of freedom = 2:**

Distribution is L shaped with maximum ordinate at zero.

**If degree of freedom <2 (>0):**

Distribution L shaped with infinite ordinate at the origin.

**STEPS INVOLVED IN CALCULATING**:

1. Calculate the expected frequencies and the observed frequencies:

Expected frequencies fe:

the cell frequencies that would be expected in a contingency table if the two variables were statistically independent.

Observed frequencies fo:

the cell frequencies actually observed in a contingency table.

fe = (column total) . (row total) / N

To obtain the expected frequencies for any cell in any cross- tabulation in which the two variables are assumed independent, multiply the row and column totals for that cell and divide the product by the total number of cases in the table.

1. Then X2 is calculated.

**CALCULATION OF CHI SQUARE:**

X2 = ∑ (o − e)2 / e

Where, O = observed frequency,

E = expected frequency,

If two distributions (observed and theoretical) are exactly alike, X2= 0; (but generally due to sampling errors, X2 is not equal to zero)

**CONDITIONS FOR THE APPLICATION OF X2 TEST:**

The following conditions should be satisfied before X2 test can be applied:

1. The data must be in the form of frequencies

2) The frequency data must have a precise numerical value and must be organised into categories or groups.

3) Observations recorded and used are collected on a random basis.

4) All the items in the sample must be independent.

5) No group should contain very few items, say less than 10.

In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10.

6) The overall number of items must also be reasonably large. It should normally be at least 50.

**EXAMPLE:**

YATE’S CORRECTION

If in the 2\*2 contingency table, the expected frequencies are small say less than 5, then X2 test can’t be used. In that case, the direct formula of the chi square test is modified and given by Yate’scorrection for continuity R1 R2 C1 C2.

**LIMITATIONS OF A CHI SQUARE TEST:**

1) The data is from a random sample.

2) This test applied in a four fouled table, will not give a reliable result with one degree of freedom if the expected value in any cell is less than 5. in such case, Yate’s correction is necessary. i.e. reduction of the mode of (o – e) by half.

3) Even if Yate’s correction, the test may be misleading if any expected frequency is much below 5. in that case another appropriate test should be applied.

4) In contingency tables larger than 2\*2, Yate’s correction cannot be applied.

5) Interpret this test with caution if sample total or total of values in all the cells is less than 50.

6) This test tells the presence or absence of an association between the events but doesn’t measure the strength of association.

7) This test doesn’t indicate the cause and effect, it only tells the probability of occurrence of association by chance.

8) the test is to be applied only when the individual observations of sample are independent.

**APPLICATIONS OF A CHI SQUARE TEST:**

This test can be used in

1) Goodness of fit of distributions

2) test of independence of attributes

3) test of homogeneity.

**GitHub:**

<https://github.com/bibeklamsa1/Python-Machine-Learning-And-AI/tree/main/Chi%20Sqaure%20Test%20Paper>

**REFERENCE:**

<https://byjus.com/maths/chi-square-test/>

<https://www.mathsisfun.com/data/chi-square-test.html>

<https://www.slideshare.net/parth241989/chi-square-test-16093013>