

Research Methodology
M.Com 2nd Semester
CHI SQUARE TEST


PRESENTED BY.
DR VANDANA PANDEY
ASSOCIATE PROFESSOR
DEPARTMENT OF COMMERCE
HCPGC VARANASI

CHI SQUARE TEST





CONTENTS:

- IMPORTANT TERMS**
 - INTRODUCTION**
 - CHARACTERISTICS OF THE TEST**
 - CHI SQUARE DISTRIBUTION**
 - APPLICATIONS OF CHI SQUARE TEST**
 - CALCULATION OF THE CHI SQUARE**
 - CONDITION FOR THE APPLICATION OF THE TEST**
 - EXAMPLE**
 - YATE'S CORRECTION FOR CONTINUITY**
 - LIMITATIONS OF THE 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 or established distribution such as normal, binomial, poisson etc.
- 2) **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.
- 3) **HYPOTHESIS**: It is a definite statement about the population parameters.



4) NULL HYPOTHESIS: (H_0) 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.

5) ALTERNATIVE HYPOTHESIS: (H_1) 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.



6) 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,

$$\text{d.f.} = (\text{number of frequencies}) - (\text{number of independent constraints on them})$$

In other terms,

$$\text{df} = (r - 1)(c - 1)$$

where

r = the number of rows

c = the number of columns

7) 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. (Latin, con- together, tangere- to touch). It is also called as an association table.



INTRODUCTION

- **The chi-square test is an important test amongst the several tests of significance developed by statisticians.**
- **It was developed by Karl Pearson in 1900.**
- **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**.**



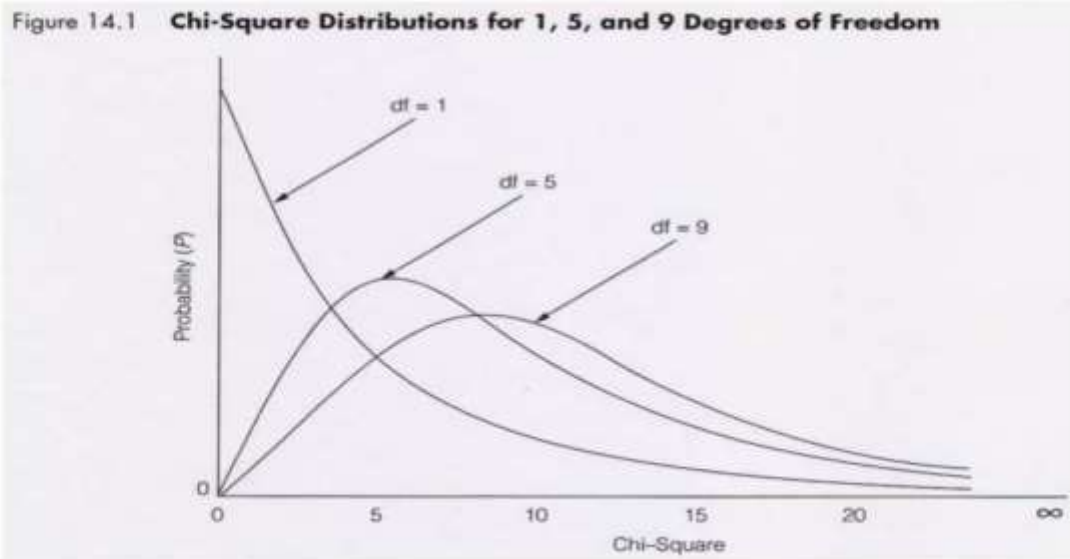
IMPORTANT CHARACTERISTICS OF A CHI SQUARE TEST

- **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 X_1, X_2, \dots, X_n are independent normal variates and each is distributed normally with mean zero and standard deviation unity, then $X_1^2 + X_2^2 + \dots + X_n^2 = \sum X_i^2$ is distributed as chi square (c^2) with n degrees of freedom (d.f.) where n is large. The chi square curve for d.f. $N=1, 5$ and 9 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.**



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.**



1) TEST OF GOODNESS OF FIT OF DISTRIBUTIONS:

➤ This test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data.

➤ The χ^2 test formula for goodness of fit is:

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Where,

o = observed frequency

e = expected frequency

➤ If χ^2 (calculated) > χ^2 (tabulated), with (n-1) d.f, then null hypothesis is rejected otherwise accepted.

➤ And if null hypothesis is accepted, then it can be concluded that the given distribution follows theoretical distribution.



2) TEST OF INDEPENDENCE OF ATTRIBUTES

- Test enables us to explain whether or not two attributes are associated.
- For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, χ^2 test is useful.
- In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever.
- χ^2 (calculated) $>$ χ^2 (tabulated) at a certain level of significance for given degrees of freedom, the null hypothesis is rejected, i.e. two variables are dependent. (i.e., the new medicine is effective in controlling the fever) and if, χ^2 (calculated) $<$ χ^2 (tabulated), the null hypothesis is accepted, i.e. 2 variables are independent. (i.e., the new medicine is not effective in controlling the fever).
- when null hypothesis is rejected, it can be concluded that there is a significant association between two attributes.

3) TEST OF HOMOGENITY

- **This test can also be used to test whether the occurrence of events follow uniformity or not e.g. the admission of patients in government hospital in all days of week is uniform or not can be tested with the help of chi square test.**
- **$\chi^2(\text{calculated}) < \chi^2(\text{tabulated})$, then null hypothesis is accepted, and it can be concluded that there is a uniformity in the occurrence of the events. (uniformity in the admission of patients through out the week)**



CALCULATION OF CHI SQUARE

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Where,

O = observed frequency

E = expected frequency

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; (but generally due to sampling errors, χ^2 is not equal to zero)



STEPS INVOLVED IN CALCULATING χ^2

- 1) Calculate the expected frequencies and the observed frequencies:

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

Observed frequencies f_o : the cell frequencies actually observed in a contingency table.

$$f_e = \frac{(\text{column total})(\text{row total})}{N}$$

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.


2) Then χ^2 is calculated as follows:

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$



CONDITIONS FOR THE APPLICATION OF χ^2 TEST

The following conditions should be satisfied before X^2 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 itmes 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. (Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.)**
 - 6) **The overall number of items must also be reasonably large. It should normally be at least 50.**
- 

EXAMPLE

A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is $1/6$ and as such the expected frequency of any one number coming upward is $132 \times 1/6 = 22$. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

Table 10.2

No. turned up	Observed frequency O_i	Expected frequency E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	-6	36	36/22
2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\therefore \sum [(O_i - E)^2 / E_i] = 9.$$

Hence, the calculated value of $\chi^2 = 9$.

\therefore Degrees of freedom in the given problem is

$$(n - 1) = (6 - 1) = 5.$$

The table value* of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.071. Comparing calculated and table values of χ^2 , we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

YATE'S CORRECTION

If in the 2*2 contingency table, the expected frequencies are small say less than 5, then χ^2 test can't be used. In that case, the direct formula of the chi square test is modified and given by Yate's correction for continuity




$$\chi^2(\text{corrected}) = \frac{N \cdot (|ad - bc| - 0.5N)^2}{R1R2C1C2}$$



LIMITATIONS OF A CHI SQUARE TEST

- 1) The data is from a random sample.
- 2) This test applied in a four fould 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 which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration.

THANK YOU