Chi – Square Test

Course: ***M.A. Geography (Sem.-3)***

Paper: M-301(CC-10)

Quantitative Techniques and Research Methodology

By

Dr. Md. Nazim

Professor, Department of Geography

Patna College, Patna University

Lecture-3

Concept:

A chi-squared test, also written as χ^2 test, is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

In the standard applications of this test, the observations are classified into mutually exclusive classes. If the null hypothesis that there are no differences between the classes in the population is true, the test statistic computed from the observations follows a χ^2 frequency distribution. The purpose of the test is to evaluate how likely the observed frequencies would be assuming the null hypothesis is true.

Test statistics that follow a χ^2 distribution occur when the observations are independent and normally distributed, which assumptions are often justified under the central limit theorem. There are also χ^2 tests for testing the null hypothesis of independence of a pair of random variables based on observations of the pairs.

Chi-squared tests often refers to tests for which the distribution of the test statistic approaches the χ^2 distribution asymptotically, meaning that the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase.

The chi – square test is a useful measure of comparing experimentally obtained results with those of the expected theoretically. It is used as a test statistic in testing a hypothesis and provides a set of theoretical frequencies with which observed frequencies are compared. Chi- square test is applied to those problems in which we study whether the frequency with which a given event has occurred, is significantly different from the one as expected theoretically. This measure enables us to find out the degree of discrepancy between the observed and expected

frequencies. It determines whether the discrepancy so obtained is due to error of sampling or due to chance.

Chi (x) is a letter of Greek language. Helmert has invented x2 - distribution in 1875 and x2- test was first developed and used by kart Pearson in 1900.

It is defined as

$$X2 = \text{Summation } (Oi - Ei)\ 2 / Ei$$

Where, Oi = Observed frequency of ith event

Ei = Expected frequency of ith event

History:

In the 19th century, statistical analytical methods were mainly applied in biological data analysis and it was customary for researchers to assume that observations followed a normal distribution, such as Sir George Airy and Professor Merriman, whose works were criticized by Karl Pearson in his 1900 paper.

At the end of 19th century, Pearson noticed the existence of significant skewness within some biological observations. In order to model the observations regardless of being normal or skewed, Pearson, in a series of articles published from 1893 to 1916, devised the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of statistical analysis consisting of using the Pearson distribution to model the observation and performing a test of goodness of fit to determine how well the model really fits to the observations.

Pearson's chi-squared test:

In 1900, Pearson published a paper on the χ^2 test which is considered to be one of the foundations of modern statistics. In this paper; Pearson investigated a test of goodness of fit.

Suppose that n observations in a random sample from a population are classified into k mutually exclusive classes with respective observed

numbers x_i (for i = 1,2,…,k), and a null hypothesis gives the probability p_i that an observation falls into the ith class. So we have the expected numbers $m_i = np_i$.

Steps to Calculate Chi-squared Test: *The following steps are required to calculate the value of chi-square;*

1. Calculate all the expected frequencies i.e. Ei for all the values of i= 1, 2, 3,……n.

2. Take difference of each observed frequency (Oi) and the corresponding expected frequency (Ei) for each value of i i.e. find (Oi – Ei)

3. Square the difference for each value of i, i.e. calculate (Oi – Ei)2

4. Divide each square difference by corresponding expected frequency i.e. calculate (Oi – Ei)2 / Ei for all the values of i = 1,2,3…….n.

5. Add all these quotients obtained in step 4, then

X2 = Summation (Oi - Ei) 2 / Ei is the value of chi – square.

Properties of Chi-squared Test:

(a) Value of x2 is always positive as each pair is squared.

(b) x2 lies between 0 and infinity.

(c) Significance test on x2 is based on one-tailed test of the right hand side of standard normal curve.

(d) x2 is a statistic and not a parameter and hence it does not involve any assumption about the form of original distribution from which the observation has come.

Uses of Chi-squared Test:

Chi-squared test is a very powerful tool for testing hypothesis of a number of statistical problems;

1. Test of Goodness of Fit: - *It is used to test whether a frequency distribution fits the expected distribution? If the two curves –observed frequency curve and the expected frequency curve are drawn then the x2-statistic may be used to determine whether the two curves so drawn are fitted good or not. The term goodness of fit is used to test the concordance of of the fitness of these two curves. Under this test there is only one variable, so the degree of freedom (d.f.) = n- 1.*

The ***goodness of fit*** *of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov–Smirnov test), or whether outcome frequencies follow a specified distribution (see Pearson's chi-squared test). In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.*

2. Test of Independence of Attributes: - *Used to test if different populations have the same proportion of individuals with some characteristics. Chi- square test is used to see that the principles of classification of attributes are independent. Attributes are classified into a two – way table. The observed frequency in each cell (square) is known as cell frequency. Total frequency in each row or column of the two – way contingency table is known as marginal frequency.*

d.f. = (R- 1). (C – 1), where R = No. of Row and C = NO. of Column

This test discloses whether there is any association or relationship between two or more attributes?

3. Test of Homogeneity or Test of a Specified Standard Deviation: *Used to test the independence of two variables. It can determine whether the occurrence of one variable affects the occurrence of other variable? Chi- square test is used to test the homogeneity of attributes in respect of a particular characteristics or it may be used to test the population variance.*

X2 = (n – 1) s2 / s02,

Where, s2 = sample variance, s02 = hypothesized value of population variance.

More than two parameters:p1,p2,p3,p4,p5,□,pn

1) HypothesisH0:p1=p2=p3=□=pn

H1: The population's proportions are not all equal

2) Collect data

3) To find p-value: □2cdf□□2,∞,c−1□where□2=Σ□Oi−Ei□2Ei

4) Decision: RejectH0if p-value is less than or equal to□

Note: If we reject the null hypothesis, then we can conclude that not all populations' proportions are

Conditions for Applying Chi- square Test:

1. Each of the observation making up the sample of this test should be independent of each other.

2. The expected frequency of any item should not be less than 5.

3. Total number of observation used in this test must be large i.e. n> 30.

4. This test is used only for drawing inferences by testing the hypothesis. It cannot be used for estimation of parameters.

5. It is wholly dependent on degree of freedom.

6. Frequencies used in X2- test should be absolute and not relative in terms.

7. The observation collected for X2-test should be on the basis of random sampling.

Calculation:

Chi – square test is widely used to test the independence of attributes. It is applied to test the association between the attributes when the sample data is presented in the form of a contingency table with rows and columns.

Step –1: *Set up Null hypothesis H0: No association exists between the Attributes*

Alternative hypothesis H1: An association exists between the Attributes.

Step – 2: *Calculate the expected frequency*

Eij = Ri x Cj / n where, Ri = sum total of row in which Eij is lying

Cj = sum total of the column ……

n = total sample size

Step – 3: *Calculate X2*

Step – 4: *Find the table value of x2 for level of significance and degree of Freedom*

Step – 5: *Compare the Calculated X2 with the Tabulated X2;*

a. *If Calculated x2 < Tabulated x2, then accept the Null hypothesis*
b. *If Calculated x2 > Tabulated x2, then reject the Null hypothesis*

Example: ***One thousand students were graded according to their I.Q. and the Economic Conditions of their homes. Use X2 – test to find out whether there is any association between the Economic Conditions at home and I.Q. of the students?***

				C
Economic Condition	**I.Q. of the Student**			
	High	**Low**	**Total**	
Rich	**100**	**300**	**400**	
Poor	**350**	**250**	**600**	**R**
Total	**450**	**550**	**1000**	

d.f. =(r – 1) (c – 1) = 2-1x 2-1 = 1

Null hypothesis H0: There is no association

Alternative hypothesis H1: There is association

Calculate expected frequency E = Row Total x Column Total / Grand Total

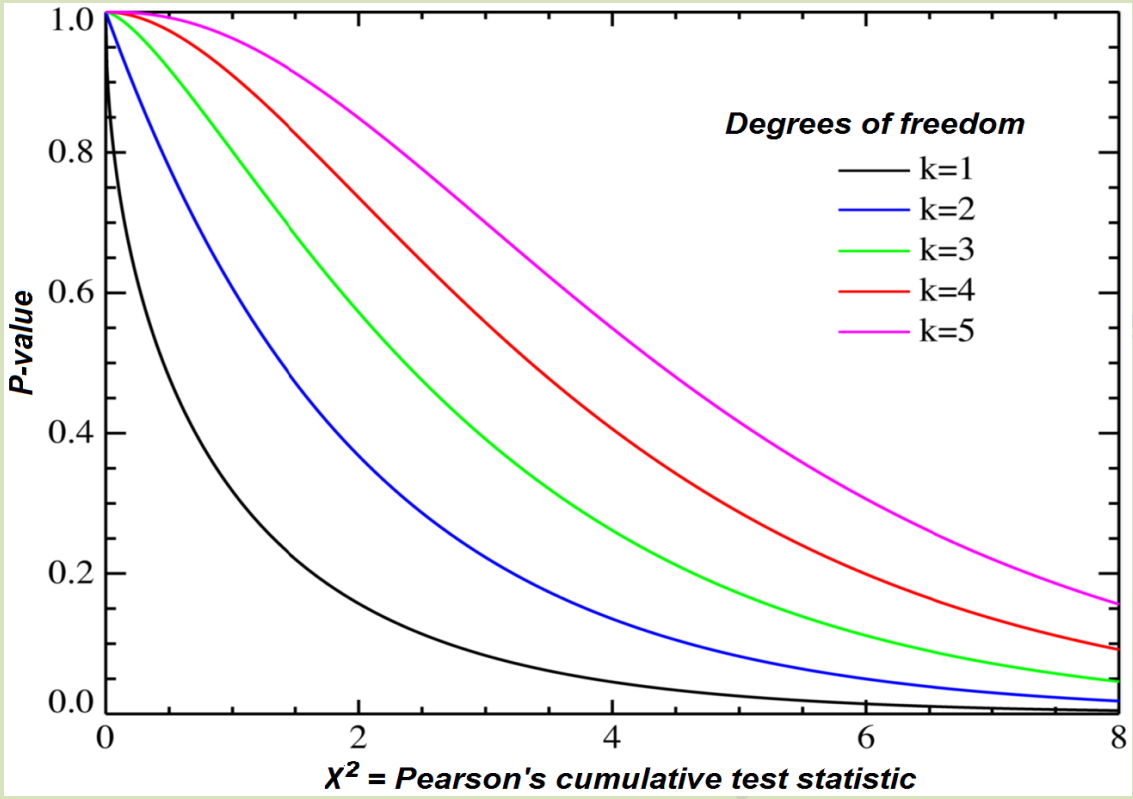
The new table will be

O	*E*	*(O – E)*	*(O – E)2*	*(O – E)2 /E*
100	*180*	*-80*	*6400*	*6400/180 =35.6*
350	*270*	*80*	*6400*	*6400/270 = 23.7*
300	*220*	*80*	*6400*	*6400/220 = 29.1*
250	*330*	*-80*	*6400*	*6400/330 = 19.4*
				100.78

Calculated X2 = (O – E)2/E = 100.78

Tabulated X2 = 3.84 at 1 d.f. and 0.05 significance level

Result – The Calculated X2 is greater than the Tabulated X2.Hence the hypothesis is rejected. Therefore, there is association between the Economic Condition and** the **I.Q. level of the students.


Degrees of freedom
k=1
k=2
k=3
k=4
k=5
P-value
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
X² = Pearson's cumulative test statistic