UNIT 12: CHI-SQUARE ANALYSIS

UNIT STRUCTURE

12.1 Learning Objectives

12.2 Introduction

12.3 Chi square test

12.4 Chi square test for the Goodness of Fit

12.5 Chi square test for the independence of variables

12.6 Chi square test for the equality of more than two population proportions

12.7 Let Us Sum Up

12.8 Further Reading

12.9 Answers to Check Your Progress

12.10 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- discuss the non-parametric test
- discuss the Independence of attributes
- describe the degrees of freedom
- dicuss chi- Square Distribution
- discuss the Goodness of Fit

12.2 INTRODUCTION

In this Unit we are going to discuss the various concepts of Chi-Square Analaysis.Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. This unit will give us a fair idea about the Chi square test for the Goodness of Fit, Chi square test for the independence of variables and Chi square test for the equality of more than two population proportions.

Let us dicuss this concepts in the following sections.

12.3 CHI SQUARE ($\div^2$) TEST

Chi square test is applies 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. The measure of Chi square enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained between observed frequencies and theoretical frequencies is due to error of sampling or due to chance

The Chi-square is computed on the basis of frequencies in a sample and thus the value of Chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a non-parametric test.The Chi Square (χ^2) Test is defined as

Degrees of Freedom (df)

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by ν.

Case I If the data is given in the form of a series of variables in a row or column, then the degrees of freedom = (number of items in the series) - 1, i.e., = n - 1.

Case II When the number of frequencies are put in cells in a contingency table, the degrees of freedom will be the product of (number of rows less one) and the (number of columns less one) i.e., = (R - 1)(C-1), where R is the number of rows and C is the number of columns.

Uses of Chi Square ($\div^2$) Test

The $\div^2$ test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of $\div^2$ test are:

12.4 CHI SQUARE TEST FOR GOODNESS OF FIT

$\div^2$ test is a measure of probabilities of association between the attributes. It gives us an idea about the divergence between the observed and expected frequencies. Thus the test is also described as the test of "Goodness of Fit". If the curves of these two distributions, when superimposed do not coincide or appear to diverge much we say the fit is poor. On the other hand if they do not diverge much, then the fit is less poor.

Example: A survey with 320 families with 5 children each revealed the

No. of Boys:	5	4	5	2	1	0
No. of Girls:	0	1	2	3	4	5
No. of Families:	14	56	110	88	40	12

Is the result consistent with the hypothesis that the male and female births are equally probable?

$$\alpha \; 320\left[\frac{1}{2}+\frac{1}{2}\right]^5 = 320\left[\frac{1}{32},\frac{5}{32},\frac{10}{32},\frac{10}{32},\frac{5}{32},\frac{1}{32}\right] = 10,50,100,100,50,10$$

Solution:

Null Hypothesis H_0 : Male and female births are equaly probable

Alternative Hypothesis H_1 : Male and female births are not equaly probable

Calculation of $\div^2$: Based on the null hypothesis the expected frequencies are given by

Table for $\div^2$

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
14	10	4	16	1.60
56	50	6	36	1.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	10	100	2.00
12	10	2	4	0.40
				$\chi^2 = 7.16$

Level of Significance: = 0.05

Critical Value: The table value of $\div^2$ at 5% level of significance with 6 - 1 = 5 df is 11.07.

Decision: Since the calculated value of $\div^2$ is less than table value of $\div^2$ so null hypothesis is accepted and can be concluded that the male and female births are equally probable.

12.5 CHI SQUARE TEST FOR THE INDEPENDENCE OF VARIABLES

$\div^2$ test can be applied to test, the independence between various attributes when the sample data is presented in the form of a contingency table with any number of rows 'r' and columns 'c' the null hypothesis and alternative hypothesis are set as follows:

Null Hypothesis H_0 : The attributes are independent

Alternative Hypothesis H_1 : The attributes are not independent

Example: Samples of 400 students of undergraduate and 400 postgraduate classes were taken to know their opinion about autonomous college. 290 undergraduate and 310 postgraduate students favoured the autonomous status. Test whether the opinion about the autonomous status of students independent of their education level or not.

Solution: The contingency table of observed frequencies is

Table of expected frequencies will be

Level of Education	Favour	Not in Favour
Undergraduate	(600 x 400)/800 = 300	(600 x 400)/800 = 300
Post graduate	(200 x 400)/800 = 100	(200 x 400)/800 = 100

Null Hypothesis H_0: Opinion regarding an autonomous status are independent of level of education.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(290-300)^2}{300} + \frac{(310-100)^2}{100} + \frac{(110-300)^2}{300} + \frac{(90-100)^2}{300} = 556$$

df = (2-1) (2-1) = 1

Critical Value: Table value of $\div^2$ at 5% level of significance with 1 df is 3.84.

Decision: Since calculated $\div^2$ is greater than table value of $\div^2$, so H_0 is rejected and can be concluded that opinion about autonomous is not independent of level of education.



CHECK YOUR PROGRESS

Q.1: Write short notes about Chi-square test. Why it is called a non parametric test?

.....

.....

Q.2. Define degrees of freedom.

.....

.....

Q.3. Give the test statistic of chi-square test.

.....

.....

Q.4. What are the uses of chi-square test?

.....

.....

12.6 CHI SQUARE TEST FOR THE EQUALITY OF MORE THAN TWO POPULATION PROPORTIONS

Here, the null hypothesis is that the proportion in each the response variable categories is the same for all treatment groups. The alternate hypothesis is that the set of proportions differs somewhere among the treatment groups. It is possible to write the hypotheses in terms of population parameters. Let δ_{ij} be the proportion of treatment population i that is found in response category j. There are G treatment groups and K response categories. The null hypothesis is:

H : $\delta_{11} = \delta_{21} = \delta_{31} = \ldots = \delta_{G1}$

The basic summary statistic in a chi-square analysis is the contingency table which summarizes the number of observations for each combination of the explanatory and response variable. Sample percentages are often computed (denoted as p_{ij} to distinguish them from the δ_{ij} in the population).

Example: Suppose assembly elections are announced in three major states. A major political party wishes to test if the proportions of its supporters in the three states are the same or not. The party conducts a sample survey of 1000 people in each state and finds that there are 300, 350 and 425 supporters in the sample in the surveyed state. Using these data test if the three proportions are the same or not.

Solution: Here, we wish to test

H_0: $Đ_1 = Đ_2 = Đ_3$ (All the proportions are the same)

H_1: Not all $Đ_1 = Đ_2 = Đ_3$ are equal

Note that, we only test if all the proportions are the same or not. In case we conclude that all the proportions are not the same, we cannot find which one is different.

We first make the 2X3 contingency table of the given data as below:

Supporter	State 1	State 2	State 3	Total
Yes	300	350	425	1075
No	700	650	575	1925
Total	1000	1000	1000	3000

The expected cell frequencies will be as follows:

Supporter	State 1	State 2	State 3
Yes	358.33	358.33	358.33
No	641.67	641.67	641.67

Test Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(300-358.33)^2}{358.33} + \frac{(350-358.33)^2}{358.33} + \frac{(425-358.33)^2}{358.33}$$

$$+ \frac{(700-641.67)^2}{641.67} + \frac{(650-641.67)^2}{641.67} + \frac{(575-641.67)^2}{641.67}$$

$$= 9.495 + 0.194 + 12.404 + 5.302 + 0.108 + 6.927$$

$$= 34.43$$

The test statistic follows chi-square distribution with (2-1)(3-1)=2 d.f. Therefore critical value at 5% level of significance is 5.991. Since computed test statistic is larger than the critical value, we reject the null hypothesis at 5% level of significance and conclude that all proportions of supporters of the political party in three states are not the same.



12.7 LET US SUM UP

Chi square test is computed on the basis of frequencies in a sample and thus the value of chi square so obtained is a statistic.

Chi-square is not a parameter as its value is not derived from the observations in a population. Hence chi-square test is a non parametric test.

It is not concerned with any population distribution and its observation.

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by ν.

Chi square test can be applied for

- Test of Goodness of Fit
- Test for the independence of variables
- Test for the equality of more than two population proportions



12.8 FURTHER READING

1. Kothari, C.R. and Garg, G.(2014), 'Research Methodology - Methods and Techniques', New Age International Publishers
2. Arora, P.N. and Arora, S.(2009), ' Statistics for Management', S. Chand



12.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q.1. The Chi-square is computed on the basis of frequencies in a sample and thus the value of Chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a non-parametric test.

Ans to Q.2. The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by .

Ans to Q.3. The χ^2 test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of χ^2 test are:

- Test of Goodness of Fit
- Test for the independence of variables
- Test for the equality of more than two population proportions



12.10 MODEL QUESTIONS

Q.1. A sample of 300 students of undergraduate and 300 students of Postgraduate classes of a University were asked to give their opinion toward the autonomous colleges. 190 of the undergraduate and 210 of the Postgraduate students favoured the autonomous status.

Test at 5% level of significance whether the opinion of students towards the status are independent of the classes or not.

Q.2. In 60 throws of a dice, face one turned up 6 times, face two or three 18 times, face four or five 24 times and face six 12 times. Test at 10% significance level whether the dice is unbiased?

Q.3. One thousand girls in a college were graded according to their I.Q. and the economic conditions of their homes. Use chi-square test to find out whether there is any association between economic conditions at home and I.Q. of girls.

Q.4. From the table given below, test whether the colour of son's eye is associated with that of father's eye.

Eye Colour of Father	EYE COLOUR OF SONS	
	Light	Dark
Light	471	151
Dark	148	230

*****