**Chi-square Analysis**

**Introduction**

In the previous chapter, our analyses are mainly based on continuous random variables. In this chapter, we shall discuss the analysis of discrete (otherwise called categorical data) variable of an elementary level which is within the scope of this course. Categorical variables may have categories which are naturally ordered. These are called ordinal variables, when people are asked to express their opinion on a certain issue and the choices and strongly oppose. These responses are naturally order hence the variable opinion is an ordinal variable.

Categorical data are where people or things are classified simultaneously by two or more attributes. The results of such a cross classification can be conveniently arranged as a table of courts known as a contingency Table. When only two classification variables are considered, the table is the 2x2 contingency or two-way tables

**Box 5.1: Chi-square Test**

 The **chi-square test**, written as -test is a useful measure of comparing experimentally obtained results with those expected theoretically and based on the hypothesis. It is used as a test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared.

The measure of chi-square enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies is due to error of sampling or due to a chance.

The 2—test was first used in testing statically hypothesis by Karl person in the year 1900 it is defined as: =

This is another distribution of considerable theoretical and practical importance. When researchers are making their guesses in an experimental situation, they may want to compare observed with theoretical frequencies. Information obtained empirically be directed observation in an experiment are referred to as observed frequencies. While the theoretical frequencies are generated on the basis of some hypothesis, which is different from the data obtained. The researcher however may want to find out whether the difference between the observed and theoretical frequencies are significant. It is the result that will then give him/her the opportunity to reject or accept his hypothesis.

**5.2 USES** **OF**  **2** **TEST**

The  2 test is a very powerful test for testing the hypothesis of a number of statistical problems. The important uses of  2 test are.

1. Test of goodness of fit.
2. Test of independence of Attributes
3. Test of Homogeneity or a test of a specific standard deviation.

**5.3 Condition for applying the chi-square test ( 2)**

1. Each of the observation constituting the sample for this test should be independent of each other.
2. The expected frequency of any item or cell should not be less than 5. Then frequencies from the adjacent items or cells are pooled together in order to make it 5 or more than 5.
3. The total number of observations used in this test must be large, i.e, n>30
4. This test is used only for drawing inferences by testing hypothesis. It cannot be used for estimation of parameter or learning order value.
5. It is wholly dependent on the degrees of freedom
6. The frequencies used in  2—test should be absolute and not relative in terms.
7. The observation collected for 2—test should be on random basis of sampling.

TABLE5.1: R x C Contingency Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ROW(R) | 1 | 2 | 3 | 4 | …….. | C |
|  | n11 | n12 | n13 | n14 |  | n1c |
| 2 | n21 | n22 | n23 | n24 |  | n2c |
| 3 | n31 | n32 | n33 | n34 |  | n3c |
| 4 | n41 | n42 | n43 | n44 |  | n4c |
| .. | .. | .. | .. | .. |  | .. |
| R | nr1 | nr2 | nr3 | nr4 |  | nrc |
| Total | n.1 | n.2 | n.3 | n.4 |  | n.c |

The total frequency in each row or column is called marginal frequency. This marginal frequency from R1 is n1 and that if column j called nj.

Note that Pij = nij

P(Ri) = = = Pi­

P(Ci) = = = Pi­j

n = =

= = 1

Given that Pij = P(Ri n Ci) under the null hypothesis, if the row and column classifications are independent, then

 Pij = P(Ri) P(Ci)

 =

 = PiPij

Therefore, the expected frequency for each cell is obtained as

 Eij = nPij2

 = n

 = (ni) (nj)

 n

=

For large ni the statistics

 = 

 =  where nij = 0

Is approximately distributed as chi-square with (r-1) (c-1) degree of freedom if the hypothesis is true. Hence, we would reject the null hypothesis of independence if the calculated value of the test statistic is greater than the tabulated value .

Note: That the chi-square as expressed above, can be written in any of the following equivalent form

 X2 = 

Or X2 = 

The Chi square test or test of goodness of fit is the statistics used to find out if there is a marked difference between the observed (O) or actual frequencies and the expected (E). If there is a marked difference between both the Chi square, i.e. X2 test will yield a numerical value large enough to be interpreted as statistically significant.

The formula is for Chi-square is given as:

 

Where, O represents Observed frequency

 E represents Expected frequency

 X2 = The calculated value of chi square

 X2 table means the table value of Chi-square.

We must point out that if X2 calculated ∑X2 table, then there is significant deviation. This means that if the calculated chi-square is equal to or greater than “table value” of chi-square, there is significant deviation.

Text question

1. A psychological skill was introduced to one or two groups of people who were suffering from a complaint (INSOMNIA). The numbers cured in each group are given in the table below. Test if the psychological skill has helped in curing the complaint (INSOMIA).

2. A random number of persons in two groups, were asked to test if they could tell the difference between two brands of butter, the results are given in the table below.

|  |  |
| --- | --- |
| Sex | Responses |
| Could tell | Couldn’t tell |
| MaleFemale | 520 | 1213 |

3.To determine whether the age of a driver age 21 years or older has any effect on the number of motor accidents he is involved in a survey which was conducted and the following information was obtained.

Test the hypothesis that the number of accidents is independent of the age of the driver at 5%.

Age of driver and number of accident

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of accident | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | Total |
| 0 | 148 | 221 | 186 | 120 | 72 | 747 |
| 1 | 44 | 30 | 21 | 36 | 20 | 151 |
| 2 | 19 | 13 | 10 | 4 | 3 | 49 |
| More than 2 | 4 | 5 | 2 | 1 | 3 | 15 |
| Total | 215 | 269 | 219 | 161 | 98 | 962 |

Text Answered

Step 1: To answer this question, you need to draw a 2 x 2 contingency table.

|  |  |  |
| --- | --- | --- |
| Group | Cured | Not Cured |
| III | 1911 | 614 |

**Hypothesis:**

 Ho: (Null hypothesis): psychological skills have helped in caring complaint (INSOMNIA)

H1: (Alternative Hypothesis): psychological skill has not helped in curing the complaint (INSOMIA).

Step 2: Find the row totals, column total and gland total.

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Cured | Not Cured | Total |
| III | 1911 | 614 | 2525 |
|  | 30 | 20 | 50 |

This is obtained through the following process:

 Row total = 19 + 6 = 25

 Row total = 11 + 14 = 25

 Column total = 19 + 11 = 30

 Column total = 6 + 14 = 20

 Grand total = 30 + 20 = 50 or 25 + 25 = 50

Step 3. Work out the expected frequency (E) for each of the cell separately using this formula

E = row total x column total

 grand total

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cured | Not Cured |  |
| Group I | C = 15A | O = 6B | 25 |
| Group II | O=11C | O = 14D | 25 |
|  | 30 | 20 | 50 |

For Cell A, E =

 

For Cell E = 

For Cell B, E = 

For Cell C, E = 

For Cell D, E = 

Therefore, the table will now read thus

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cured | Not Cured |  |
| Group I | O = 19E = 15 | O = 6E = 10 | 25 |
| Group II | O = 11E = 15 | O = 14E = 10 | 25 |
|  | 30 | 20 | 50 |

 Step 4: Calculate the X2 by working out the difference between O and E for each cell.





X2 = 1.06 + 1.6 + 1.06 + 1.6 = 5.32

X2 = 5.32

The degree of freedom (df) for any statistic is the number of component in its calculation that are free to vary.

df = (r-1) (c-1)

 = (2-1) (2-1)

df = 1

We observed from X2 table the following X2 at 5%, (1) = 3.84.

Since 5.32 > 3.84, the null hypothesis that the number cured is independent of the psychological skills is related at the 5 percent level of significance. Thus we conclude that the psychological skills have helped in caring complaint (INSOMNIA) and reject the alternative hypothesis that says psychological skill has not helped in curing the complaint (INSOMIA).

(2). Text Answered

Step 1: Let us find the row totals column totals and grand total.

|  |  |  |
| --- | --- | --- |
| Sex | Responses |  |
| Could tell | Couldn’t tell |  |
| MaleFemale | 5(a)20(c) | 12(b)13(d) | 1733 |
|  | 25 | 25 | 50 |

Step II: Work out the expected frequency (E) for each cell separately using this formula

E = Row total x Column total

 Grand total

 For Cell (a) E = 

For Cell (b) E = 

For Cell (c) E = 

For Cell (d), E = 

|  |  |  |
| --- | --- | --- |
| Sex | Responses |  |
| Could tell | Couldn’t tell |  |
| Male | O = 5E = 8.5 | O = 12E = 8.5 | 17 |
| Female | O = 20E = 16.5 | O = 13E = 16.5 | 33 |
|  | 25 | 25 | 50 |

Step 2: Calculate the X2 by working out the difference between (O) and (E) for each.





X2 = 1.441 + 1.441 + 0.7424 + 0.7424

X2 = 4.5668. df = (2-1) (2-1) = 1

 = 4.77 df = 1

We observed from the X2 table the following X2 at 5%(1) = 3.84

Since 4.37 > 3.84, the null hypothesis that the difference in the butter is independent of sex is rejected at the 5 percent level of significance.

(3). Text Answered

The expected value for each cell is obtained

E11 =

E12 =

E13 =

E14 =

The corresponding table for expected value is given below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Number of accident | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | Total |
| 0 | 167 | 208.9 | 170 | 125 | 76.1 | 747 |
| 1 | 33.7 | 42.2 | 34.4 | 25.3 | 15.4 | 151 |
| 2 | 11 | 13 | 10 | 4 | 3 | 49 |
| More than 2 | 4 | 5 | 2 | 1 | 3 | 15 |
| Total | 215 | 269 | 219 | 161 | 98 | 962 |

The null and alternative hypothesis are stated as

Ho: The age of a driver age 21 years or older has no effect on the number of motor accidents he is involved in

H1: The age actually has effect on the number of accidents

The chi-square statistics is used and computed as follows

 = 

 = + + + + + + +

= 21.6 + 0.0054 + 1.51 + 0.2 + 0.22 + 3.15 + 3.53 + 5.22 + 4.53 + 1.37 + 5.82 +

 0.036 + 0.13 + 2.15 + 0.8 + 1.06 + 1.52 + 0.58 + 0.961 + 1.5

= 36. 4524

**Decision Rule:** We Reject Ho if X2 >

**Decision:** since X2 = 36.45 > = 21.0 we reject Ho and conclude that the age of a driver aged 21 years or older has significant effect on the number of accident he is involved in. this conclusion simply means that the number of accidents a driver involved is depends on his age.

**5.4.Contingency coefficient**, Test C is given by C =  Where C = Contingency coefficient

chi-square

N=Grand total of subjects of cases

If C is near zero (or equal to zero) you can conclude that your variables are independent of each other, there is no association between them.

If C is away from zero there is some relationship, C can only take on positive values.

**5.4.1 Correlation of Attributes Test (Cramer’s V –Formula)**

The presentation in a contingency table often concern attributes of human beings or objects. The degree to which one of the attributes depend upon, is associated with or related to the other attribute is referred to as correlation of attributes. In a kxk contingency, the correlation of attributes, r is given as: r= (Cramer’s V –Formula) A measure that does indicate the strength of the association is **Cramer’s V defines as**

  = r

Text question

1.Used the parameters below to test for the contingency coefficient and Correlation of Attributes Test, where the chi-square value is 36.4525, Total number of observation is 962 and it dimension is 4

Contingency coefficient, C is  =0.19

And Correlation of Attributes Test

 = 0.11

|  |  |  |
| --- | --- | --- |
|  | HIV infection | NO HIV |
| Unprotected sex | 70 | 200 |
| No Sex | 15 | 20 |

1.A survey to investigate the relationship between exposure to unprotected sex and HIV infection and the information below was obtained

 Is HIV infection independent of exposure to sex? Take α = 1%

1b. Test the strength of the association between the two gender by using Cramer’s V and also find the contingency coefficient

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Not short-sighted | 23 | 16 |
| Short-sighted | 7 | 14 |