Subject	Statistics
Semester	04
Paper no	10
Paper Name	Testing of Hypothesis
Topic no	19
Topic name	Test for Independence of Two Attributes
SME	Ms Shubharekha
ID	Ms Varsha Shetty

E-learning Module on Test For Independence of Two Attributes

Learning Objectives

At the end of this session, you will be able to:

- Explain contingency tables
- Explain Chi square test for independence of attributes
- Describe the conditions and type of data required to apply the test
- Explain the applications of the test for 2 by 2 and 2 by k contingency tables

Introduction

In some situations the researcher classifies the experimental units according to two qualitative variables to generate bivariate data

- A defective piece of furniture is classified according to the type of defect and the production shift during which it was made
- A professor is classified by his professional rank and the type of the university (public or private) at which she works

Contingency Tables:

When two categorical variables are recorded we can summarize the data by counting the observed number of units that fall into each of the various intersection of category levels.

The resulting counts are displayed in an array called a contingency table.

Let the attributes under consideration be **A** and **B**

Let **A** be divided into **r** classes $\mathbf{A_1, A_2 \ldots A_r}$

Let **B** be divided into s classes say $\mathbf{B_1, B_2 \ldots B_s}$

Let the observations be classified according to their classes in attributes **A** and **B**

This classification can be represented in the form of a table called contingency table as shown below.

r x s Contingency Table

A \ B	B_1	B_2	.	.	B_s	M.T
A_1	O_{11}	O_{12}	.	.	01s	(A_1)
A_2	O_{21}	O_{22}	.	.	O_{2s}	(A_2)
.	.					
.	.					
A_r	O_{r1}	O_{r2}	.	.	O_{rs}	(A_r)
M. T	(B_1)	(B_2)			(B_s)	N

O_{ij} represents the number of individuals falling in class Ai and B_j of the attributes A and B

(A_i) represents the number of individuals in the class A_i and (B_j) represents the number of individuals in the class Bj

$$N = \Sigma(A_i) = \Sigma(B_j)$$

Chi Square Test of Independence of the Attributes A and B

The question of independence of the two methods of classification can be investigated using a test of hypothesis based on the Chi square statistic.

These are the hypothesis:

H_0: The attributes A and B are independent

H_1: The attributes A and B are dependent

The probabilities of an individual falling in different classes of attributes A and B under the null hypothesis are calculated as follows:

$P(A=A_i)= (A_i)/N$, probability that an individual falls in class Ai

$P(B=B_j)= (B_j)/N$, probability that an individual falls in class B_j

$$P(A=Ai \text{ and } B=Bj) = P(A=Ai)\, P(B=Bj)$$

probability that an individual falls in classes A_i and B_j according to the attributes A and B respectively.

Under the null hypothesis = $(A_i)/N\ (B_j)/N$

Let E_{ij} denotes the expected number of individuals with attributes A and B as A_i and B_j respectively.

Then

E_{ij}= P(A=A_i and B = B_j) N

$$= N\frac{(A_i)}{N}\frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N}$$; i=1,2…r; j=1,2…,s

If we knew the expected cell counts ($E_{ij} = n\ p_{ij}$) under the null hypothesis of independence, then we could use the chi square statistic to compare the observed and expected counts.

Now under the null hypothesis

$$\chi^2 = \frac{\sum_i \sum_j (Oij - Eij)^2}{E_{ij}}$$

is distributed as χ^2 variable.

Degrees of Freedom

There are r x s cells to be filled independently because of restrictions on the marginal and grand totals.

Therefore under the null hypothesis the above is distributed as a Chi square variable with (r-1)(s-1) degrees of freedom.

The d.f for a Chi-square grid are equal to the number of rows minus one times the number of columns minus one: that is, (r-1)(s-1)

In our simple 2x2 grid, the degrees of independence are therefore (2-1) (2-1), or **1**

The test procedure therefore is to reject the H_0 whenever the computed value of χ^2 is greater than the tabulated value $\chi_\alpha^2((r-1)(s-1))$

Assumptions

The chi-square test, when used with the standard approximation that a chi-square distribution is applicable, has the following assumptions:

- Simple Random Sample: The sample data is a random sampling from a fixed distribution or population where each member of the population has an equal probability of selection. Variants of the test have been developed for complex samples, such as where the data is weighted.

- Sample size (whole table) – A sample with a sufficiently large size is assumed.

If a chi square test is conducted on a sample with a smaller size, then the chi squared test will yield an inaccurate inference which might end up committing a Type 2 error.

- Expected Cell Count – Adequate expected cell counts. Some require 5 or more, and others require 10 or more.

A common rule is 5 or more in all cells of a 2 by 2 table, and 5 or more in 80 percent of cells in larger tables, but no cells with zero expected count. When this assumption is not met, Yate's correction is applied.

- Independence – The observations are always assumed to be independent of each other

This means that chi-square cannot be used to test correlated data. In those cases you might want to turn to McNemar's test.

A Special Case:

Consider a 2x2 contingency table given below. To test the null hypothesis that the attributes A and B are independent.

A \ B	B_1	B_2	M.T
A_1	a	b	a+b
A_2	c	d	c+d
M.T	a+c	b + d	a + b+ c + d=N

Under the null hypothesis the statistic Chi square is computed as follows:

The expected frequency of **a** is

$$E(a) = (a + b)(a + c) / N$$

$$a - E(a) = a- (a + b)(a + c) / N$$

$$=(ad- bc) / N$$

Similarly

$E(b) = (a + b)(b + d) / N$

$b - E(b) = (bc - ad) / N$

$E(c) = (c + d)(a + c) / N$

$c - E(c) = (bc - ad) / N$

$d - E(d) = (ad - bc) / N$

$$\chi^2 = \frac{(a-E(a))^2}{E(a)} + \frac{(b-E(b))^2}{E(b)} + \frac{(c-E(c))^2}{E(c)} + \frac{(d-E(d))^2}{E(d)}$$

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)}\right]$$

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{(a+c)(a+b)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(c+d)(a+c)} + \frac{1}{(c+d)(b+d)}\right]$$

$$\chi^2 = (ad-bc)^2\left[\frac{c+d+a+b}{(a+c)(a+b)(b+d)(c+d}\right]$$

$$= \left[\frac{N(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d}\right]$$

Under H_0 the above statistic is distributed as Chi square variable with 1 d.f

The test procedure is to compute χ^2 and to compare it with the tabulated value of $\chi_\alpha^2(1)$ and to reject the null hypothesis if the computed value exceeds the tabulated value.

Yate's Correction for Continuity in 2x2 Contingency Table:

The statistic χ^2 is distributed as a chi square variable only when all the cell theoretical frequencies are greater than or equal to 5.

If any one of the frequency is less than 5, the statistic χ^2 does not follow a Chi square distribution.

In such cases by applying the correction suggested by Yates, the statistic χ^2 can be regarded to follow Chi square distribution.

According to Yates in a 2x2 contingency table if any cell theoretical frequency is <5, 0.5 should be added to it and remaining frequencies should be adjusted such that the marginal frequencies remains unchanged.

The χ^2 statistic then is computed as usual.

After the correction chi square statistic is obtained as follows:

$$\chi^2 = \left[\frac{N\left| (a \pm \frac{1}{2})(d \pm \frac{1}{2}) - (b \mp \frac{1}{2})(c \mp \frac{1}{2}) \right|^2}{(a+c)(a+b)(b+d)(c+d} \right]$$

$$\chi^2 = \left[\frac{N\left[\left| ad - bc \right| - \frac{N}{2} \right]^2}{(a+c)(a+b)(b+d)(c+d} \right]$$

Brandt-Snedekor's Formula for 2x**k** Contingency Table

Let the observations be classified based on two attributes **A** and **B** into **2 x k** classes.

Let O_{ij} represents the number of individuals falling in class $\mathbf{A_i}$ and $\mathbf{B_j}$ of the attributes **A** and **B** i= 1,2 and j=1,2,...,k

The observations may be arranged in the form of **2 X k** contingency table as given below:

2 x k Contingency Table

A \ B	B_1	B_2	.....	B_k	M.T
A_1	O_{11}	O_{12}	.....	O_{1k}	m_1
A_2	O_{21}	O_{22}	.....	O_{2k}	m_2
M.T	n_1	n_2		n_k	N

H_0 : the attributes A and B are independent

Under H_0 the expected frequencies and the Chi square statistic are calculated as follows:

$$E_{1j} = E(O_{1j}) = m_1 n_j / N$$

$$E_{2j} = E(O_{2j}) = m_2 n_j / N$$

$$\chi^2 = \frac{\sum_{i=1}^{2} \sum_{j=1}^{k} (Oij - Eij)^2}{E_{ij}}$$

$$\chi^2 = \frac{\sum_{j=1}^{k}(O_{1j} - E_{1j})^2}{E_{1j}} + \frac{\sum_{j=1}^{k}(O_{2j} - E_{2j})^2}{E_{2j}}$$

$$\chi^2 = \frac{\sum_{j=1}^{k}(O_{1j} - \frac{m_1 n_j}{N})^2}{\frac{m_1 n_j}{N}} + \frac{\sum_{j=1}^{k}(O_{2j} - \frac{m_2 n_j}{N})^2}{\frac{m_2 n_j}{N}}$$

$$\chi^2 = \frac{N}{m_1}\sum_{j=1}^{k} n_j (\frac{O_{1j}}{n_j} - \frac{m_1}{N})^2 + \frac{N}{m_2}\sum_{j=1}^{k} n_j (\frac{O_{2j}}{n_j} - \frac{m_2}{N})^2$$

$$\chi^2 = \frac{1}{p}\sum_{j=1}^{k} n_j (p_j - p)^2 + \frac{1}{q}\sum_{j=1}^{k} n_j (q_j - q)^2$$

Where

$$p_j = \frac{O_{1j}}{n_j}$$

$$q_j = \frac{O_{2j}}{n_j}$$

$$p = \frac{m_1}{N}$$

$$q = \frac{m_2}{N}$$

It may be noted that

$$p_j + q_j = 1; p + q = 1$$

Hence

$$(q_j - q) = -(p_j - p) \Rightarrow (q_j - q)^2 = (p_j - p)^2$$

$$\chi^2 = \frac{1}{p}\sum_{j=1}^{k} n_j (p_j - p)^2 + \frac{1}{q}\sum_{j=1}^{k} n_j (p_j - p)^2$$

$$= (\frac{1}{p} + \frac{1}{q})\sum_{j=1}^{k} n_j (p_j - p)^2 = \frac{1}{pq}\sum_{j=1}^{k} n_j (p_j - p)^2$$

$$= \frac{1}{pq}\sum_{j=1}^{k} n_j (\frac{O_{1j}}{n_j} - \frac{m_1}{N})^2$$

$$= \frac{N^2}{m_1 m_2}(\sum_{j=1}^{k} \frac{O_{1j}^{\ 2}}{n_j} - \frac{m_1^{\ 2}}{N})$$

Under H_0 the statistic Chi square is distributed as a Chi square variable with (2-1)(k-1)=(k-1) degrees of freedom.

The test procedure in this case is to reject the null hypothesis whenever the computed value exceeds the tabulated value $\chi_\alpha^2(k-1)$

Type Of Data for Chi Square Test for Independence

A test of independence assesses whether paired observations on two variables, expressed in a contingency table are independent of each other.

A Chi-square test for independence of attributes is designed to analyze categorical data which are represented in the form of a contingency table.

That means that the data has been counted and divided into categories. It will not work with parametric or continuous data (such as height in inches).

For eg: if you want to test whether attending class influences how students perform on an exam, using test scores (0-100) as data would not be appropriate for a Chi-square test.

However, arranging students into the categories "Pass" and "Fail" would.

By dividing a class of 54 into groups according to whether they attended class and whether they passed the exam, you might construct a data set like this table

	Pass	Fail
Attended	25	6
Skipped	8	15

Problems:

The approximation to the chi-squared distribution breaks down if expected frequencies are too low.

It will normally be acceptable so long as no more than 20% of the events have expected frequencies below 5

In this case, a better approximation can be obtained by reducing the absolute value of each difference between observed and expected frequencies by half before squaring; this is called Yates's correction for continuity.

In cases where the expected value E, is found to be small the normal approximation of the multinomial distribution can fail, and in such cases it is found to be more appropriate to use the G-test, a likelihood ratio-based test statistic.

Where the total sample size is small, it is necessary to use an appropriate exact test, typically either the binomial test or (for contingency tables) Fisher's exact test; but note that this test assumes fixed and known marginal totals.