Test of Hypotheses for independence of Attributes

Dr R Sinuvasan

Test of Hypotheses for independence of Attributes

Another important application of χ^2 test is the testing of independence of attributes.

Attributes: Attributes are qualitative characteristic such as levels of literacy, employment status, *etc.*, which are quantified in terms of levels/scores.

Contigency table: Independence of two attributes is an important statistical application in which the data pertaining to the attributes are cross classified in the form of a two – dimensional table. The levels of one attribute are arranged in rows and of the other in columns. Such an arrangement in the form of a table is called as a contingency table.

Step 1 : Framing the hypotheses

Null hypothesis H_0: The two attributes are independent

Alternative hypothesis H_1: The two attributes are not independent.

Step 2 : Data

The data set is given in the form of a contigency as under. Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N}; \quad i = 1, 2, \ldots m; \ j = 1, 2, \ldots n$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

The contingency table consisting of *m* rows and *n* columns.
The observed data is presented in the form of a contingency table :

		Attribute B						Total
		B_1	B_2	$\dots$	B_j	$\dots$	B_n	
Attribute A	A_1	O_{11}	O_{12}	$\dots$	O_{1j}	$\dots$	O_{1n}	R_1
	A_2	O_{21}	O_{22}	$\dots$	O_{2j}	$\dots$	O_{2n}	R_2
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_{i}	O_{i1}	O_{i2}	$\dots$	O_{ij}	$\dots$	O_{in}	R_i
	:	:	:	:	:	:	:	:
	.	.	.	.	.	.	.	.
	A_m	O_{m1}	O_{m2}	$\dots$	O_{mj}	$\dots$	O_{mn}	R_m
Total	C_1	C_2	$\dots$	C_j	$\dots$	C_{n}	$N = m\times n$	

Step 3 : Level of significance

Fix the desired level of significance α

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^{\ 2} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Step 5 : Critical value

The critical value is obtained from the table of χ^2 with $(m-1)(n-1)$ degrees of freedom at given level of significance, α as $\chi^2_{(m-1)(n-1),\, \alpha}$.

Step 6 : Decision

Decide on rejecting or not rejecting the null hypothesis by comparing the calculated value of the test statistic with the table value. If $\chi_0^{\ 2} \geq \chi^2_{(m-1)(n-1),\, \alpha}$ reject H_0.

Note:

- N, the total frequency should be reasonably large, say greater than 50.
- No theoretical cell-frequency should be less than 5. If cell frequencies are less than 5, then it should be grouped such that the total frequency is made greater than 5 with the preceding or succeeding cell.

Example

The following table gives the performance of 500 students classified according to age in a computer test. Test whether the attributes age and performance are independent at 5% of significance.

Performance	Below 20	21-30	Above 30	Total
Average	138	83	64	285
Good	64	67	84	215
Total	202	150	148	500

Solution:

Step 1 : **Null hypothesis** H_0: The attributes age and performance are independent.

Alternative hypothesis H_1: The attributes age and performance are not independent.

Step 2 : **Data**

Compute expected frequencies E_{ij} corresponding to each cell of the contingency table, using the formula

$$E_{ij} = \frac{R_i \times C_j}{N} \quad i = 1,2;\ j = 1,2,3$$

where,

N = Total sample size

R_i = Row sum corresponding to i^{th} row

C_j = Column sum corresponding to j^{th} column

Performance	Below average	Average	Above average	Total
Average	$\frac{285\times 202}{500}=115.14$	$\frac{285\times 150}{500}=85.5$	$\frac{285\times 148}{500}=84.36$	285
Good	$\frac{215\times 202}{500}=86.86$	$\frac{215\times 150}{500}=64.5$	$\frac{215\times 148}{500}=63.64$	215
Total	202	150	148	500

Step 3 : Level of significance $\alpha = 5\%$

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0{}^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

This chi-square test statistic is calculated as follows:

$$\chi_0^2 = \frac{(138-115.14)^2}{115.14} + \frac{(83-85.50)^2}{88.50} + \frac{(64-84.36)^2}{84.36} + \frac{(64-86.86)^2}{86.86} + \frac{(67-64.50)^2}{64.50} + \frac{(84-63.64)^2}{63.64}$$

$= 22.152$ with degrees of freedom $(3-1)(2-1) = 2$

Step 5 : Critical value

From the chi-square table the critical value at 5% level of significance is $\chi^2_{(2-1)(3-1),0.05} = \chi^2_{2,0.05} = 5.991$.

Step 6 : Decision

As the calculated value $\chi_0^2 = 22.152$ is greater than the critical value $\chi^2_{2,0.05} = 5.991$, the null hypothesis H_0 is rejected. Hence, the performance and age of students are not independent.

If the contigency table is 2 x 2 then the value of χ^2 can be calculated as given below:

	A	not A	Total
B	a	b	$a+b$
not B	c	d	$c+d$
Total	$a+c$	$b+d$	$N=a+b+c+d$

$$\chi_0^{\,2} = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_\alpha^{\,2}(1d.f)$$

Example

A survey was conducted with 500 female students of which 60% were intelligent, 40% had uneducated fathers, while 30 % of the not intelligent female students had educated fathers. Test the hypothesis that the education of fathers and intelligence of female students are independent.

Solution:

Step 1 : **Null hypothesis** H_0: The attributes are independent *i.e.* No association between education fathers and intelligence of female students

Alternative hypothesis H_1: The attributes are not independent *i.e* there is association between education of fathers and intelligence of female students

Step 2 : **Data**

The observed frequencies (O) has been computed from the given information as under.

	Intelligent females	Not intelligent females	Row total
Educated fathers	$300-120=180$	$\frac{30}{100}\times 200=60$	240
Uneducated fathers	$\frac{40}{100}\times 300=120$	$200-60=140$	260
Total	$\frac{60}{100}\times 500=300$	$500-300=200$	N= 500

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Calculation

Calculate the value of the test statistic as

$$\chi_0^{\,2}=\frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

where, a= 620, b = 380, c = 550, d = 450 and N = 2000

$$\chi_0^{\,2}=\frac{2000(620\times 450-380\times 550)^2}{(620+380)(550+450)(620+550)(380+450)}=10.092$$

Step 5 : Critical value

From chi-square table the critical value at 5% level of significance is $\chi^2_{1,0.05} = 3.841$

Step 6 : Decision

The calculated value $\chi_0^2 = 10.092$ is greater than the critical value $\chi^2_{1,0.05} = 3.841$, the null hypothesis H_0 is rejected. Hence, education of fathers and intelligence of female students are not independent.