Test of Independence of Attributes

Another important application of the χ^2 – distribution is the testing of independence of attributes (attributes are characters which are non-measurable; for example Sex, Employment, Literacy etc, are all attributes). Suppose we want to test whether sex and employment are associated. In this case take a random sample from the population and classify the sample as given in the following table. The numbers in the table denote the frequencies (number of persons possessing the attributes).

	Male	Female	Total
Employed	50	20	70
Unemployed	15	15	30
Total	65	35	100

This type of table which has one basis of classification across column and another across row is known as a contingency table. The above table has 2 rows and 2 columns and hence it is called a 2×2 contingency table. A table which has r –rows and s–columns is called a $r \times s$ contingency table.

In testing the hypothesis the null hypothesis is taken as "employment independent of sex" whereas the alternative hypothesis is "the employment not independent of sex".

Then comes the question of determining the expected frequencies.

Assuming that H_0 is true, the totals are all kept the same.

For example the expected frequency for the 1st cell in the above table, determined by the formula:

$$\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}} = \frac{70 \times 65}{100} = 45.5$$

The other theoretical frequencies are determined on the same lines.

	Male	Female	Total
Employed	45.5	24.5	70
Unemployed	24.5	5.5	30
Total	65	35	100

It can be checked that by determining only one cell frequency the other expected frequencies can be easily obtained from the column and row totals. Thus in 2×2 contigency table the number of degrees of freedom is $(2-1)\times(2-1)=1$. In general in a $r\times s$ contigency table the number of degrees of freedom is $(r-1)(s-1)$.

Test Procedure:

STEP I : Write down the null hypothesis.

STEP II : Write down the alternative hypothesis.

STEP III : Calculate the theoretical frequencies for the contingency table.

STEP IV : Calculate $\chi^2 = \Sigma\left[\dfrac{(O-E)^2}{E}\right]$

STEP V : Write down the number of degrees of freedom.

STEP VI : Draw the conclusion on the hypothesis by comparing the calculated values of χ^2 with table value of χ^2.

Test of Independence of attributes

Example 1:

A random sample of employees of a large company was selected and the employees were asked to complete a questionnaire. One question asked was whether the employee was in favour of the introduction of flexible working hours. The following table classifies the employees by their response and by their area of work.

Response	Area of work	
	Production	Non-Production
In Favour	129	171
Not in Favour	31	69

Test whether there is evidence of a significant association between the response and the area of work.

Solution:

H_0 : There is no evidence of a significant association between the response and the area of work.

H_1 : There is an evidence of a significant association between the response and the area of work.

Now we have to calculate the expected frequencies to apply the χ^2 test.

On the assumption of, H_0, the expected frequency for the class " production and in favour" is given by

$$\frac{(A)\times(B)}{N} = \frac{160\times 300}{400} = 120$$

Similiarly, we can calculate the other expected frequencies

The other expected frequencies are

$$\frac{240\times 300}{400} = 180, \quad \frac{160\times 100}{400} = 40 \quad \frac{240\times 100}{400} = 60.$$

Table showing observed frequencies

Response	Production	Non-Production	Total
In Favour	129	171	300
Not in Favour	31	69	100
Total	160	240	400

Table showing expected frequencies

Response	Production	Non-Production	Total
In Favour	120	180	300
Non in Favour	40	60	100
Total	160	240	400

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
129	120	9	81	0.675
171	180	–9	81	0.450
31	40	–9	81	2.025
69	60	9	81	1.350
400	400			4.500

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 4.5$$

ndf = 1.

The table value of χ^2 for 1 df at 5% level = 3.81

Conclusion:

Since the calculated value of χ^2 is greater than the table value of χ^2, H_0 is rejected.

Hence there is evidence for a significant association between response and the area of work.

Example 2. *From the following table regarding the color of eyes of fathers and sons test whether the color of the son's eye is associated with that of the father.*

		Eye color of son	
		Light	*Not light*
Eye color of father	*Light*	*471*	*51*
	Not light	*148*	*230*

Sol. Null hypothesis H_0. The color of the son's eye is not associated with that of the father, *i.e.*, they are independent.

Under H_0, we calculate the expected frequency in each cell as

$$= \frac{\text{Product of column total and row total}}{\text{whole total}}$$

Expected frequencies are:

Eye color of son / *Eye color of father*	*Light*	*Not light*	*Total*
Light	$\dfrac{619\times 522}{900}=359.02$	$\dfrac{289\times 522}{900}=167.62$	522
Not light	$\dfrac{619\times 378}{900}=259.98$	$\dfrac{289\times 378}{900}=121.38$	378
Total	619	289	900

$$\chi^2=\frac{(471-359.02)^2}{359.02}+\frac{(51-167.62)^2}{167.62}+\frac{(148-259.98)^2}{259.98}+\frac{(230-121.38)^2}{121.38}$$
$$=261.498.$$

Conclusion. Tabulated value of χ^2 at 5% level for 1 d.f. is 3.841.

Since the calculated value of χ^2 > the tabulated value of χ^2, H_0 is rejected. They are dependent, *i.e.*, the color of the son's eye is associated with that of the father.

Example 3. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory:*

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Sol. Null hypothesis H_0. The production of bad parts is independent of the shift on which they were produced.

I.e., the two attributes, production and shifts, are independent.

Under H_0,
$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{\left[(A_iB_j)_0 - (A_iB_j)\right]^2}{(A_iB_j)_0}\right]$$

Calculation of expected frequencies

Let A and B be two attributes, namely, production and shifts. A is divided into two classes A_1, A_2, and B is divided into three classes B_1, B_2, B_3.

$$(A_1B_1)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850)\times(1000)}{2985} = 954.77$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850)\times(990)}{2985} = 945.226$$

$$(A_1B_3)_0 = \frac{(A_1)(B_3)}{N} = \frac{(2850)\times(995)}{2985} = 950$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{(135)\times(1000)}{2985} = 45.27$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{(135)\times(990)}{2985} = 44.773$$

$$(A_2B_3)_0 = \frac{(A_2)(B_3)}{N} = \frac{(135)\times(995)}{2985} = 45.$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
(A_1B_1)	960	954.77	27.3529	0.02864
(A_1B_2)	940	945.226	27.3110	0.02889
(A_1B_3)	950	950	0	0
(A_2B_1)	40	45.27	27.7729	0.61349
(A_2B_2)	50	44.773	27.3215	0.61022
(A_2B_3)	45	45	0	0
				1.28126

Conclusion. The tabulated value of χ^2 at 5% level of significance for 2 degrees of freedom $(r - 1)(s - 1)$ is 5.991. Since the calculated value of χ^2 is less than the tabulated value, we accept H_0, *i.e.*, the production of bad parts is independent of the shift on which they were produced.

Example 4. *From the following data, find whether hair color and sex are associated.*

Sex \ Color	Fair	Red	Medium	Dark	Black	Total
Boys	592	849	504	119	36	2100
Girls	544	677	451	97	14	1783
Total	1136	1526	955	216	50	3883

Sol. Null hypothesis H_0. The two attributes of hair color and sex are not associated, *i.e.*, they are independent.

Let A and B be the attributes of hair color and sex, respectively. A is divided into 5 classes ($r = 5$). B is divided into 2 classes ($s = 2$).

$$\therefore \quad \text{Degrees of freedom} = (r-1)(s-1) = (5-1)(2-1) = 4$$

$$\text{Under } H_0, \text{ we calculate } \chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{\left[(A_iB_j)_0 - (A_iB_j)\right]^2}{(A_iB_j)_0}$$

Calculate the expected frequency $(A_iB_j)_0$ as follows:

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{1136 \times 2100}{3883} = 614.37$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{1136 \times 1783}{3883} = 521.629$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1526 \times 2100}{3883} = 852.289$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1526 \times 1783}{3883} = 700.71$$

$$(A_3B_1)_0 = \frac{(A_3)(B_1)}{N} = \frac{955 \times 2100}{3883} = 516.482$$

$$(A_3B_2)_0 = \frac{(A_3)(B_2)}{N} = \frac{955 \times 1783}{3883} = 483.517$$

$$(A_4B_1)_0 = \frac{(A_4)(B_1)}{N} = \frac{216 \times 2100}{3883} = 116.816$$

$$(A_4B_2)_0 = \frac{(A_4)(B_2)}{N} = \frac{216 \times 1783}{3883} = 99.183$$

$$(A_5B_1)_0 = \frac{(A_5)(B_1)}{N} = \frac{50 \times 2100}{3883} = 27.04$$

$$(A_5B_2)_0 = \frac{(A_5)(B_2)}{N} = \frac{50 \times 1783}{3883} = 22.959$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
A_2B_1	592	614.37	500.416	0.8145
A_1B_2	544	521.629	500.462	0.959
A_2B_1	849	852.289	10.8175	0.0127
A_2B_2	677	700.71	562.1641	0.8023
A_3B_1	504	516.482	155.800	0.3016
A_3B_2	451	438.517	155.825	0.3553
A_4B_1	119	116.816	4.7698	0.0408
A_4B_2	97	99.183	4.7654	0.0480
A_5B_1	36	27.04	80.2816	2.9689
A_5B_2	14	22.959	80.2636	3.495
				9.79975

$\chi^2 = 9.799.$

Conclusion. Table of χ^2 at 5% level of significance for 4 d.f. is 9.488.

Since the calculated value of χ^2 < tabulated value H_0 is rejected, *i.e.*, the two attributes are not independent, *i.e.*, the hair color and sex are associated.

Example 11:

To test the efficiency of a new drug a controlled experiment was conducted wherein 300 patients were administered the new drug and 200 other patients were not given the drug. The patients were monitored and the results were obtained as follows:

	Cured	Condition Worsened	No Effect	Total
Given the drug	200	40	60	300
Not given the drug	120	30	50	200
Total	320	70	110	500

Use χ^2–test for finding the effect of the drug.

Solution:

H_0 : The drug is not effective.

H_1 : The drug is effective

Observed Frequencies

	Cured	Condition Worsened	No Effect	Total
Drug	200	40	60	300
No Drug	120	30	50	200
Total	320	70	110	500

Expected Frequencies

	Cured	Condition Worsened	No Effect	Total
Drug	$\frac{320 \times 300}{500} = 192$	42	66	300
No Drug	128	28	44	200
Total	320	70	110	500

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
200	192	8	64	0.3313
40	42	–2	4	0.0952
60	66	–6	36	0.5454
120	128	–8	64	0.5000
30	28	2	4	0.1429
50	44	6	36	0.8182
500	500			2.4330

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2.435$$

ndf $= (2-1)\times(3-1) = 2$

Table value for 2 df at 5% level $= 5.991$

Conclusion:

H_0 is accepted since the calculated value of χ^2 < the table values of χ^2.
Hence the drug is not effective.