SLC

Teaching & Learning Services

Te Rātonga Ako

CHI SQUARE

DISTRIBUTION

Lincoln University

Te Whare Wānaka o Aoraki

Introduction to the Chi Square Test of Independence

This test is used to analyse the relationship between *two* sets of discrete data.

Contingency tables are used to examine the relationship between subjects' scores on two or more qualitative or categorical variables. For example, consider the hypothetical experiment on the effect of smoking on divorce to find if there is any relationship between them. In a sample of 167, 73 from 85 of the smokers had been divorced. Of the non-smokers, only 43 of 82 had been divorced. These data are depicted in the contingency table shown above.

	Divorced	Not divorced	Total
Smoker	73	12	85
Non-smoker	43	39	82
Total	116	51	167

The cell entries are cell frequencies.
The top left cell with a "73" in it means: 73 subjects in the group who smoked were divorced.

From the table, it looks as if the smokers of the group are more likely to divorce than the non-smokers.

Thus, the column a subject is in (divorced or not divorced) is contingent upon (depends on) the row the subject is in (smoker or non-smoker).

If the columns are not contingent on the rows, then the rows and column frequencies are independent. The test of whether the columns are contingent on the rows is called the ***chi square test of independence.***

The ***null hypothesis*** is the hypothesis that there is no relationship between row and column frequencies.

Parameters and Symbols

H_o	the null hypothesis
H_A *or* H_1	the alternative hypothesis
O	observed count
E	expected count
χ^2_{test}	critical chi-square test value
χ^2_{table}	chi-square value from the table

Formulae

For each cell in the contingency table: $$E = \frac{(row\ total) \times (column\ total)}{(grand\ total)}$$

To calculate the critical value: $$\chi^2_{test} = \Sigma \frac{(O-E)^2}{E}$$

Degrees of freedom $$df = (no.\ of\ rows - 1) \times (no.\ of\ columns - 1)$$

Steps To Follow to Perform the Chi-Squared Test:

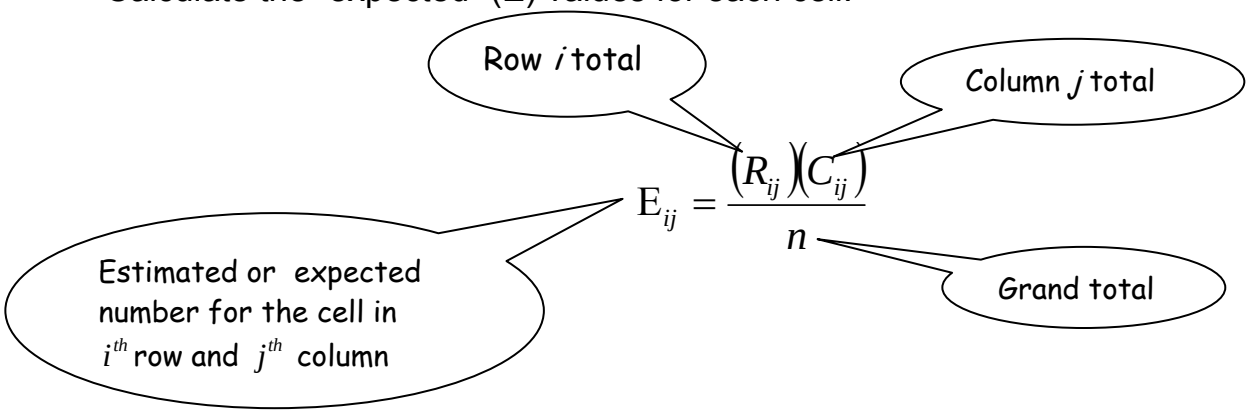
1 State the **null hypothesis**.
This states that the variables in the contingency table are independent (or the classification of one variable does not affect the classification of the other).
eg H_o*: Divorce is not related to being a smoker.*

2 State the **alternative hypothesis.**
This states that the variables are dependent, or that the direction of one affects the direction of the other.

eg H_A*: Divorce is related to being a smoker.*

3 Calculate the **expected values**.

- Record the experimental results (observed values, O) in a contingency table of *r* rows and *c* columns.
- Calculate the "expected" (E) values for each cell:



- Record these values in the contingency table.

	Divorced		**Not divorced**		
	Observed	Expected	Observed	Expected	Total
Smoker	73	$\frac{85\times116}{167}=59$	12	$\frac{85\times51}{167}=26$	85
Non Smoker	43	$\frac{82\times116}{167}=57$	39	$\frac{82\times51}{167}=25$	82
Total	116		51		167

- Calculate the chi-square test value:

observed number of responses for cell in i^{th} row and j^{th} column

$$\chi^2_{test} = \sum_{ij}\left[\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}\right]$$

sum for every cell in each row and each column

$$\chi^2 = \frac{(73-59)^2}{59} + \frac{(12-26)^2}{26} + \frac{(43-57)^2}{57} + \frac{(39-25)^2}{25} (= 3.32 + 7.54 + 3.44 + 7.84) = 22.14$$

NOTE: ***It is not necessary to write these individual calculations separately.***
It is quicker to put numbers in your calculator exactly as is.

4 Find the **critical value** from chi-squared tables.

To use these tables, two pieces of information are required:
- the level of significance – usually at the 5% level (written as $\alpha = 0.05$)
- degrees of freedom = df = (no. of rows minus 1) $\times$ (no. of columns -1) = $(r-1)(c-1)$

With 1 degree of freedom and $\alpha = 0.05$, from tables, $\chi^2 = 3.841$

5 Compare the test statistic with table statistic.

If the **calculated test statistic** $\left(\chi^2_{test}\right)$ **> the critical value from tables** χ^2_{table}, then the null hypothesis is rejected and the alternative hypothesis is favoured.

Otherwise, we conclude that there is " not enough evidence to reject H_o."

In our example, test value of $\chi^2 = 22.14$. This is greater than the table value of 3.841. Hence we reject H_O in favour of H_A, that is that divorce and smoking are related.

The chi-squared test is **valid** if **all** $E_{ij} \geq 1$ and no more than 20% of $E_{ij} \leq 5$;

ie χ^2 is most accurate when the expected frequency in each cell $\geq$ 5.

Example questions

1. EXAM QUESTION 2001 Semester 2

In a recent trial, researchers gave subjects either aspirin or a placebo (an inactive substance) to determine whether aspirin reduced the risk of a heart attack.
The results of this study are summarised in the following 2×2 table.

	Heart Attack			
	Yes		**No**	
	observed	Expected	Observed	Expected
Aspirin	104		10933	
Placebo	189		10845	

(a) State the alternative hypothesis for a chi-square test of this data. (1)
(b) Calculate the expected values for this data. (2)
(c) Calculate the Chi-square statistic for this data. (3)
(d) At the $\alpha = 0.05$ level of significance, draw an appropriate conclusion from this test. Also, write a sentence explaining your conclusion, so that it would be understood by someone with no previous training in statistics. (4)

2 Test 2 Summer School 2004

Use the following information to answer Questions 8 - 12

The Director of Transportation of a large company is interested in the pattern of usage of her van pool. She considers her routes to be divided into local and non-local. She is particularly interested in learning if there is a difference in the proportion of males and females who use the local routes. She takes a sample of a day's riders and finds the following:

	Male	Female	Total
Local	20	40	60
Non-local	30	10	40
Total	50	50	100

The Director uses this information to perform a Chi-Square test using a level of significance of 0.05.

(8) Referring to the above Table, what are the degrees of freedom of the test?

(9) Referring to the above Table, what is the expected cell frequency in the Male/Local cell?

(10) Referring to the above Table, what is the calculated value of the Chi-Square test statistic?

(11) What is the critical Chi-Square value for the above Table for a level of significance of 0.05?

(12) What conclusion from the Chi-Square test for the above data can be made?

3. Test question

A study was conducted to determine whether brand awareness of female TV viewers and the gender of the spokesperson are independent. Each, in a sample of 300 female TV viewers, was asked to identify a product advertised by a celebrity spokesperson. The gender of the spokesperson and whether or not the viewer could identify the product was recorded. The numbers in each category are given below.

	Male spokesperson	Female spokesperson
Identified product	41	61
Could not identify	109	89

a. Which of the following would be an appropriate alternative hypothesis?

A Brand awareness of female TV viewers and the gender of spokesperson are independent.

B Brand awareness of female TV viewers and the gender of spokesperson are not independent.

C Brand awareness of female TV viewers and the products they identified are not independent.

D The gender of the spokesperson and the products identified are related.

b. What are he degrees of freedom of the test statistic?

c. Referring to Table 1, what is the calculated χ^2 test statistic?

d. Referring to χ^2 Table, at 5% level of significance, what is the critical value of the test statistic?

Answers

1.

	Heart Attack				
	Yes		**No**		Total
	Observed	Expected	Observed	Expected	
Aspirin	104	**147**	10933	**10890**	11037
Placebo	189	**146**	10845	**10888**	11034
Totals	293		21778		22071

(a) H_2: That taking aspirin affects the incidence of heart attack
(b) As in table above.
(c) Chi-square (test) = 25.58
(d) Chi-square (table) = 3.84

"So since the test value is greater than the table value, we can reject the null hypothesis in favour of the alternative ie that taking aspirin will lessen your chances of having a heart attack."

2. (8) 1 (9) 30 (10) 16.67 (11) 3.841
(12) Since, calculated test statistic > critical value, reject the null hypothesis

3. a. **B**: Brand awareness of female TV viewers and the gender of spokesperson are not independent.
b. 1 c. 5.9418 d. 3.8415

Further practice questions from past papers

1. The following is a printout produced from part of a study of the relationship between family income and choice of shopping mall.

Shopping mall

Income	1	2	3	All
1	60 65.74	25 25.20	14 8.09	99 99.00
2	66 71.05	32 27.24	9 8.71	107 107.00
3	127 116.21	40 44.55	8 14.24	175 175.00
All	253	97	31	381
	253.00	97.00	31.00	381.00

CHI-SQUARE = 10.292 WITH D.F. = 4

(a) For α = 0.05, what conclusion would you come to?

(b) Show how the result of 44.55 was calculated for 'shopping mall=2, income=3".

(c) What contribution to the overall chi-square value do the 'shopping mall=2, income=3" make?

2. A class of students was asked if they drank alcohol. Of those that did, a sample was asked more detail about their drinking. Each was asked for their preferred drink : [1 = wine, 2 = beer, 3 = spirits] and for the amount they drank each week on a scale of 1 to 4:[1= "a little" to 4 = "a lot"]. The table below shows the results.

Amount

Preference	1	2	3	4	ALL
1	8 5.72	4 2.60	1 1.82	0 2.86	13 13
2	5 8.80	3 4.00	3 2.80	9 4.40	20 20.00
3	9 7.48	3 3.40	3 2.38	2 3.74	17 17.00
ALL	22 22.00	10 10.00	7 7.00	11 11.00	50 50.00

CHI-SQUARE = 12.933 with d. f. = 6

(a) What is the *critical value* for the chi-square statistic (α = 0.05)?

(b) Give your conclusion: what does the chi-squared test suggest?

(c) Show how the value of 7.48 [row 3, column 1] is calculated.

(d) What contribution to the overall chi-squared value do the wine drinkers make?

3. A random sample of 300 students were asked their income (low to high in four categories 1 to 4) and whether they owned a car (Yes or No). The results obtained are in the table below:

Income category

Owns a car	1	2	3	4	Total
No	65 60.19	48 **	5 6.51	4 6.51	122 122
Yes	83 87.81	72 71.20	11 9.49	12 9.49	178 178
Total	148	120	16	16	300

(a) Given that the chi-squared test = 2.886 with d.f. = 3, at □ = 0.05, what is the correct conclusion to draw from this test?

(b) What is the expected value for the cell corresponding to "Does not own car, Income = 2" (ie the missing value)?

(c) What is the contribution to the overall Chi-squared value of 2.886 from the cell corresponding to "owns car, Income = 4"?

4 At the 0.05 significance level, use the data in the following table to test the claim that the sentence is independent of the category of crime.

	Embezzlement	**Fraud**	**Forgery**	**Total**
Sent to jail	22	130	20	172
Not sent to jail	57	146	25	228
Total	79	276	45	400

Answers:

1. a) $\chi^2 = 9.488$ from tables. Since this is less than test value, we reject the null hypothesis. That is, choice of shopping mall is dependent on family income.

b) $E = \frac{175 \times 97}{381} = 44.55$ c) Contribution = $\frac{(40-44.55)^2}{44.55} = 0.4647$

2. a) $\chi^2 = 12.592$ b) Since test value of 12.933 is greater than table value, we can reject the null hypothesis, in favour of the alternative – that amount of alcohol consumed *does* depend on the type preferred.

c) $E_{31} = \frac{17 \times 22}{50} = 7.48$

d) $contribution\ of\ wine\ drinkers: \chi^2 = \frac{(8-5.52)^2}{5.72} + \frac{(4-2.60)^2}{2.60} + \frac{(1-1.82)^2}{1.82} + \frac{(0-2.86)^2}{2.86}$
=5.097

3. (a) χ^2 from tables =7.815. Since χ^2 test value of 2.886 is less than table value, there is not enough evidence to reject H_O.

(b) Expected value = $\frac{122 \times 120}{300} = 48.8$

(c) Contribution from the cell corresponding to "owns car, Income = 4" = $\frac{(12-9.49)^2}{9.49} = 0.664$

4

Observed	Expected	Observed	Expected	Observed	Expected
22	$\frac{172 \times 79}{400} = 34$	130	$\frac{172 \times 276}{400} = 119$	20	$\frac{172 \times 55}{400} = 24$
57	$\frac{228 \times 79}{400} = 45$	146	$\frac{228 \times 276}{400} = 157$	25	$\frac{228 \times 45}{400} = 26$

$$\chi^2 = \frac{(22-34)^2}{34} + \frac{(130-119)^2}{119} + \frac{(20-24)^2}{24} + \frac{(57-45)^2}{45} + \frac{(146-157)^2}{157} + \frac{(25-26)^2}{26}$$

$$= 4.2 + 1.02 + 0.67 + 3.2 + 0.77 + 0.038 = 9.9$$

χ^2 from tables (with 2 degrees of freedom and α = 0.05) = 5.991. Since test value is greater than table value, we reject H_O in favour of H_A - that type of sentence depends on type of crime.