Geo Factsheet

www.curriculumpress.co.uk Number 156

The Chi-SquaredTest for Association

The chi-squared test for association (or *association index* or *chi-squared contingency tables*) is commonly used in projects to measure whether two factors are *associated* (e.g. whether greater numbers of a certain plant species occur in areas with high rainfall). It can also be used to compare, for example, population data from two villages. This Factsheet explains how to use the test.

What does the test do?

There are two ways of using this test - testing for association and testing for a difference in structures.

Testing for	Example	Hypotheses	It it's significant it tells you...
Association	Is there any association between type of industry and distance from the town centre?	H_0: There is no association between the two variables H_1: There is association between the two variables	There is some association between the variables - although not what kind! (eg - it doesn't tell you *which* industries are located near the town centre)
Difference in structures	Is there any difference in the population structures of two villages?	H_0: There is no difference between the structures H_1: There is a difference between the structures	There is a difference between the two structures - although not what the difference is (eg one village could have more old people, or more young people - the test doesn't say which)

How does it work?

The test is used to compare **observed frequencies** (what is produced from the investigation) with **expected frequencies** (what you'd expect from the null hypothesis). The closer the two are, the more likely it is the null hypothesis is true.

For example, suppose we were doing an investigation into the affect of air pollution on asthma. Fifty students in a polluted area and fifty students in an unpolluted area were asked if they suffered from asthma. Here's one set of possible results:

	Polluted area	Unpolluted area
Asthma	50	0
No asthma	0	50

If we got these results, we'd be pretty certain pollution affected asthma.

Now we'll look at another possible set:

	Polluted area	Unpolluted area
Asthma	26	24
No asthma	24	26

With these results, we'd probably think pollution didn't make much difference.

What about these results?

	Polluted area	Unpolluted area
Asthma	32	22
No asthma	18	28

Do we think this shows "enough" difference to say whether pollution affects asthma?

The chi-squared test for association gives us a way of deciding what constitutes "enough" difference. The method of calculation is **exactly the same** for both ways of using it.

Warning!

Any chi-squared test can **only** be used with **frequencies** (that is, numbers of things/ people in particular categories). You cannot use it on data that are, for example, lengths, areas, percentages....

If you have data that are, for example, lengths, you may be able to **convert** them into frequencies by noting, for example, the number of leaves less than 10cm in length, between 10 and 20 cm etc.

Your **expected values** must be at least five. You can think of this as having an **average** of at least five items in each category. This doesn't mean that each category actually has to have more than 5, but the majority of them should. If you do the calculations for the test, and find the expected frequencies are less than five - you need to go back and get more data.

Investigations using this test

- Whether different types of industry are found at different distances from the city centre.
- Whether different sectors of the rural population experience different accessibility problems
- Whether socio-economic group affects recycling behaviour
- Whether there is a difference in population age/sex structure between two villages
- Whether there is a difference in employment by sector between two regions
- Whether there is a difference in population structure by socio-economic class between two areas in a town

Worked Example 1
In an investigation into distances of industries from the city centre, the following results were obtained

	Number of industries at		
	Less than 1km	1 - 3 km	Over 3km
Service industries	15	5	2
Hi-tech industries	2	6	8
Other industries	5	6	7

Step 1: Write down the **hypotheses**

H_0: Type of industry and distance from the city centre are not associated
H_1: Type of industry and distance from city centre are associated

Step 2: Work out the **row**, **column** and **overall totals** for the original data

	< 1km	1 - 3 km	> 3km	**Row Totals**
Service industries	15	5	2	**22**
Hi-tech industries	2	6	8	**16**
Other industries	5	6	7	**18**
Column Totals	**22**	**17**	**17**	**54**

overall total

Step 3: Calculate the **expected frequencies** for each category using the formula

$$\frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

Put the values in a a table.

	< 1km	1 - 3 km	> 3km
Service industries	22×22/54 = **8.96**	22×17/54 = **6.93**	22×17/54 = **6.93**
Hi-tech industries	16×22/54 = **6.52**	16×17/54 = **5.03**	16×17/54 = **5.03**
Other industries	18×22/54 = **7.33**	18×17/54 = **5.67**	18×17/54 = **5.67**

Exam Hint: *- Don't worry if your expected frequencies are not whole numbers. They don't have to be!* ***Do not*** *round them to the nearest whole number - this will make your test less accurate.*

Step 4: For each of your categories, work out

$$\frac{(O - E)^2}{E}$$

(O = **o**bserved values, from the experiment
E = **e**xpected values, from step 3)
Put the values in a a table.

eg for Hi-tech, 1-3km: $\frac{(6 - 5.03)^2}{5.03} = 0.19$

	< 1km	1 - 3 km	> 3km
Service industries	4.07	0.54	3.51
Hi-tech industries	3.13	0.19	1.75
Other industries	0.74	0.02	0.31

Step 5: **Add up** all these values.
This gives the **chi-squared value**

chi-squared value =4.07 + 0.54 + 3.51 + 3.13 + 0.19 + 1.75 + 0.74 + 0.02 + 0.31 = **14.26**

Step 6: Work out the **degrees of freedom** using the formula:
(no. of rows - 1)(no. of columns - 1)

Degrees of freedom = (3 - 1)×(3 - 1) = 4

Exam Hint: *- Don't worry about what degrees of freedom means! Unless you want to study Statistics as a subject, you don't need to know!*

Step 7: Get a chi-squared table and **look up the value** for the appropriate significance level (usually 5%) and the degrees of freedom.

Tables value for 4 df and 5% significance is **9.49**

df	.10	**.05**	.025	.01	.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	**9.49**	11.14	13.23	14.86

Step 8: **Make a decision** - if your chi-squared value is **bigger** than the one from the tables, you can **reject** the null hypothesis. Otherwise you have to accept it.

Our value (14.26) is larger than the tables value (9.49), so we **reject** the null hypothesis.

Step 9: **Write down your conclusion.**

At the 5% level of significance, we can conclude that industry type and distance from the town centre are associated.

Worked Example 2
A student wished to compare the employment by sector in her home town (A) with that of a neighbouring town (B)

Step 1: Write down the **hypotheses**

H_0: There is no difference in employment by sector between towns A and B
H_1: There is a difference in employment by sector between towns A and B

Step 2: Work out the **row**, **column** and **overall totals** for the original data

Employment Sector	Town A	Town B	Row Total
Forestry/Fishing	40	21	**61**
Energy & Water Supply	203	125	**328**
Manufacturing	1243	400	**1643**
Construction	400	321	**721**
Distribution etc	2051	1500	**3551**
Transport & Communication	462	350	**812**
Finance, Real Estate etc	1450	1563	**3013**
Public Administration etc	206	37	**243**
Education, Health & Social Work	1700	928	**2628**
Other Services	245	230	**475**
Column Total	**8000**	**5475**	**13475**

Step 3: Calculate the **expected frequencies** for each category using the formula

$$\frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

Put the values in a table

eg for Construction, Town B: $\frac{721 \times 5475}{13475} = 292.9$

Employment Sector	Town A	Town B
Forestry/Fishing	36.2	24.8
Energy & Water Supply	194.7	133.3
Manufacturing	975.4	667.6
Construction	428.1	292.9
Distribution etc	2108.2	1442.8
Transport & Communication	482.1	329.9
Finance, Real Estate etc	1788.8	1224.2
Public Administration etc	144.3	98.7
Education, Health & Social Work	1560.2	1067.8
Other Services	282.0	193.0

Step 4: For each of your categories, work out

$$\frac{(\mathbf{O} - \mathbf{E})^2}{\mathbf{E}}$$

(O = **o**bserved values, from the investigation
E = **e**xpected values, from step 3)

eg for Construction, Town B: $\frac{(321 - 292.95)^2}{292.95} = 2.686$

Employment Sector	Town A	Town B
Forestry/Fishing	0.40	0.58
Energy & Water Supply	0.35	0.52
Manufacturing	73.42	107.26
Construction	1.84	2.70
Distribution etc	1.55	2.27
Transport & Communication	0.84	1.22
Finance, Real Estate etc	64.17	93.76
Public Administration etc	26.38	38.57
Education, Health & Social Work	12.53	18.30
Other Services	4.85	7.09

Step 5: **Add up** all these values.
This gives the **chi-squared value**

Chi-squared value =0.3955 + 0.5780 + 0.3511 + ... + 4.8555+ 7.0948 = **458.5863**

Step 6: Work out the **degrees of freedom** using the formula: **(rows - 1)(columns - 1)**

Degrees of freedom = (10 - 1)×(2 - 1) = 9

Step 7: Get a chi-squared table and **look up the value** for the appropriate significance level (usually 5%) and the degrees of freedom.

Tables value for 9 df and 5% significance is **16.92**

df	.10	**.05**	.025	.01	.005
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.54	20.09	21.96
9	14.68	**16.92**	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19

Step 8: **Make a decision** - if your chi-squared value is **bigger** than the one from the tables, you can **reject** the null hypothesis. Otherwise you have to accept it.

Our value (458.5863) is larger than the tables value (16.92) , so we **reject** the null hypothesis - there is a difference in employment by sector between towns A and B

Question

A student decides to investigate whether socio-economic group affects recycling practises within his/her home area.
S/he decides to concentrate on paper recycling.

(a) State suitable null and alternative hypotheses for this investigation [1]

The student collects the following data, and decides to carry out a chi-squared test:

	Number of Households		
socio-economic group	Recycle paper regularly	Recycle paper occasionally	Never recycle paper
A, B, C1	24	31	18
C2, D, E	6	5	12

(b) Calculate the expected frequencies, based on the null hypothesis given in (a) [5]

(c) Calculate $\sum \frac{(O - E)^2}{E}$ [3]

(d) State the number of degrees of freedom [1]

(e) Use chi-squared tables to decide whether the null hypothesis should be accepted or rejected, using a 5% significance level. [2]

(f) The student says: "This shows that people in the higher socioeconomic groups are more likely to recycle".
Explain why this conclusion should not be drawn from the results of the test alone. [1]

(g) Comment on the categories used by the student, as displayed in the results table above. [1]

14

Answer

(a) H_0: Social class and approach to paper recycling are not associated/ are independent/are not linked
H_1: Social class and approach to paper recycling are associated/dependent/linked;

(b) Row totals: 73, 23. Column totals 30, 36, 30. Overall total 96;
Expected frequencies: (1 mark for correct method, 3 marks for all answers correct)

socio-economic group	Recycle paper regularly	Recycle paper occasionally	Never recycle paper
A B C1	73 ×30/96 =22.8	73 ×36/96 =27.4	73 ×30/96 =22.8
C2 D E	23 ×30/96 =7.2	23 ×36/96 =8.6	23 ×30/96 =7.2

(c) $(24 - 22.8)^2/22.8 + (31 - 27.4)^2/27.4 + (18 - 22.8)^2/22.8 + (6 - 7.2)^2/7.2 + (5 - 8.6)^2/8.6 + (12 - 7.2)^2/7.2$;
= 0.063 + 0.473 + 1.011 + 0.200 + 1.507 + 3.200;
= 6.454;

(d) $(2 - 1)(3 - 1) = 2$;

(e) Tables value at 5% significance and 4 degrees of freedom is 5.99;
Chi-squared value is greater than that, so reject the null hypothesis;

(f) The test only tells us there is an association, not what the relationship between the two is;

(g) Comment such as: If these were the categories used in the questionnaire, they are open to interpretation - what counts as "regularly" or "occasionally";

Acknowledgements: *This Geo Factsheet was researched and written by Cath Brown.* ***Curriculum Press.*** *Bank House, 105 King Street, Wellington, TF1 1NU.* ***ISSN 1351-5136***