

Health Sciences M.Sc. Programme

Applied Biostatistics

Week 7: Chi-squared tests

The chi-squared test for association

Table 1 shows the relationship between housing tenure for a sample of pregnant women and whether they had a premature delivery. We have the outcome of pregnancy tabulated by type of housing. This kind of cross-tabulation of frequencies is also called a **contingency table** or **cross classification**.

Each entry in the table is a frequency, the number of individuals having these characteristics. It can be quite difficult to measure the strength of the association between two qualitative variables, but it is easy to test the null hypothesis that there is no relationship or association between the two variables. If the sample is large, we do this by a chi-squared test.

The chi-squared test for association in a contingency table works like this. The null hypothesis is that there is no association between the two variables, the alternative being that there is an association of any kind. We find for each cell of the table the frequency which we would expect if the null hypothesis were true. To do this we use the row and column totals, so we are finding the expected frequencies for tables with these totals, called the **marginal** totals.

There are 1443 women, of whom 899 were owner occupiers, a proportion $899/1443$. If there were no relationship between time of delivery and housing tenure, we would expect each column of the table to have the same proportion, $899/1443$, of its members in the first row. Thus the 99 patients in the first column would be expected to have $99 \times 899/1443 = 61.7$ in the first row. (By 'expected' we mean the average frequency we would get in the long run. We could not actually observe 61.7 subjects). The 1344 patients in the second column would be expected to have $1344 \times 899/1443 = 837.3$ in the first row. The sum of these two expected frequencies is 899, the row total. Similarly, there are 258 patients in the second row and so we would expect $99 \times 258/1443 = 17.7$ in the second row, first column and $1344 \times 258/1443 = 240.3$ in the second row, second column. We calculate the expected frequency for each row and column combination, or cell. The 10 cells of Table 1 give us the expected frequencies shown in Table 2. Notice that the row and column totals are the same as in Table 1.

In general, the expected frequency for a cell of the contingency table is found by

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

It does not matter which variable is the row and which the column.

We now compare the observed and expected frequencies. If the two variables are not associated, the observed and expected frequencies should be close together, any discrepancy being due to random variation. We need a test statistic which measures this. The differences between observed and expected frequencies are a good place to start. We cannot simply sum them as the sum would be zero, both observed and expected frequencies having the same grand total, 1443. We can resolve this as we resolved a similar problem with differences from the mean, by squaring them. The size of the difference will also depend in some way on the number of patients. When the row and column totals are small, the difference between observed and expected is forced to be small. It turns out that the best test statistic is found by adding together for each cell of the table the observed frequency minus the expected frequency all squared, divided by the expected frequency.

Table 1. Contingency table showing time of delivery by housing tenure

Housing tenure	Premature delivery	Term delivery	Total
Owner-occupier	50	849	899
Council tenant	29	229	258
Private tenant	11	164	175
Lives with parents	6	66	72
Other	3	36	39
Total	99	1344	1443

Table 2. Expected frequencies under the null hypothesis for Table 1

Housing tenure	Premature delivery	Term delivery	Total
Owner-occupier	61.7	837.3	899.0
Council tenant	17.7	240.3	258.0
Private tenant	12.0	163.0	175.0
Lives with parents	4.9	67.1	72.0
Other	2.7	36.3	39.0
Total	99.0	1344.0	1443.0

This is often written as

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

We call this the **chi-squared statistic**. For Table 1 this comes to 10.5.

The distribution of this test statistic when the null hypothesis is true and the sample is large enough is the **Chi-squared distribution**., also written ' **χ^2 distribution**', using Greek letter χ (chi, pronounced 'ki' as in 'kite'). We shall discuss what is meant by 'large enough' later. The Chi-squared distribution is a family of distributions, like the Normal and the t distributions. It has one parameter, called the **degrees of freedom**. As you might guess, it is closely related to the t distribution. It is a skew distribution that is always positive. Figure 1 shows the Chi-squared distribution with 4 degrees of freedom. The 5% point is shown on the graph.

Like the Normal and Student's t distributions, we use a table of the probability points, but computers can calculate the required numbers every time. Table 3 shows some percentage points of the Chi-squared Distribution, the value which will be exceeded with given probabilities for different degrees of freedom.

Which member of the Chi-squared family do we need? For a contingency table, the degrees of freedom is given by

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

For Table 1 we have $(5 - 1) \times (2 - 1) = 4$ degrees of freedom. We see from Table 3 that for 4 degrees of freedom the 5% point is 9.49 and 1% point is 13.28, so our observed value of 10.5 has probability between 1% and 5%, or 0.01 and 0.05. If we use a computer program which prints out the actual probability, we find $P = 0.03$. The data are not consistent with the null hypothesis and we can conclude that there is good evidence of a relationship between housing and time of delivery.

Figure 1. The Chi-squared distribution with 4 degrees of freedom, showing the upper 5% point

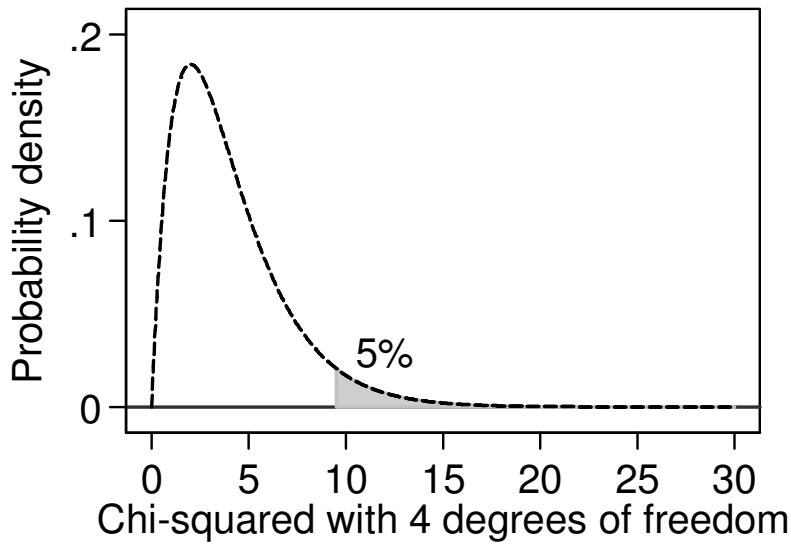


Table 3. Percentage points of the Chi-squared Distribution

Degrees of freedom	Probability that the tabulated value is exceeded							
	0.10	10%	0.05	5%	0.01	1%	0.001	0.1%
1		2.71		3.84		6.63		10.83
2		4.61		5.99		9.21		13.82
3		6.25		7.81		11.34		16.27
4		7.78		9.49		13.28		18.47
5		9.24		11.07		15.09		20.52
6		10.64		12.59		16.81		22.46
7		12.02		14.07		18.48		24.32
8		13.36		15.51		20.09		26.13
9		14.68		16.92		21.67		27.88
10		15.99		18.31		23.21		29.59
11		17.28		19.68		24.73		31.26
12		18.55		21.03		26.22		32.91
13		19.81		22.36		27.69		34.53
14		21.06		23.68		29.14		36.12
15		22.31		25.00		30.58		37.70
16		23.54		26.30		32.00		39.25
17		24.77		27.59		33.41		40.79
18		25.99		28.87		34.81		42.31
19		27.20		30.14		36.19		43.82
20		28.41		31.41		37.57		45.32

The chi-squared statistic is not an index of the strength of the association. If we double the frequencies in Table 1, this will double chi-squared, but the strength of the association is unchanged. We can only use the chi-squared test when the numbers in the cells are frequencies, not when they are percentages, proportions or measurements.

The Chi-squared test for a contingency table is known as **Pearson's Chi-squared test** or simply as the Chi-squared test.

Validity of the Chi-squared test for small samples

When the null hypothesis is true, the chi-squared statistic follows the Chi-squared distribution provided the expected values are large enough. This is a large sample test. The smaller the expected values become, the more dubious will be the test.

The conventional criterion for the test to be valid is usually attributed to the great statistician W. G. Cochran. The rule is this: the chi-squared test is valid if at least 80% of the expected frequencies exceed 5 and all the expected frequencies exceed 1. We can see that Table 2 satisfies this requirement, since only 2 out of 10 expected frequencies, 20%, are less than 5 and none are less than 1. Note that this condition applies to the expected frequencies, not the observed frequencies. It is quite acceptable for an observed frequency to be 0, provided the expected frequencies meet the criterion.

This criterion is open to question. Simulation studies appear to suggest that the condition may be too conservative and that the chi-squared approximation works for smaller expected values, especially for larger numbers of rows and columns. At the time of writing the analysis of tables based on small sample sizes, particularly 2 by 2 tables, is the subject of hot dispute among statisticians. As yet, no-one has succeeded in devising a better rule than Cochran's, so I would recommend keeping to it until the theoretical questions are resolved. Any chi-squared test which does not satisfy the criterion is always open to the charge that its validity is in doubt.

The alternative is **Fisher's exact test**, also known as the **Fisher-Irwin exact test**. This works for any sample size. In the past, it was used only for small samples in 2 by 2 tables, because of computing problems. The calculations are very time-consuming and prone to error. We calculate the probability of every possible table with the given row and column totals. When the table has more than two rows or columns or large frequencies, this can be a very large number of tables indeed. We then sum the probabilities for all the tables as or less probable than the observed. With modern computers, this is not so difficult as in the past and we can do Fisher's exact test for any table. For Table 1, $X^2 = 10.5$, with 4 d.f. Using a computer, the P value for this is 0.033. Fisher's exact test gives $P = 0.034$, so they are very similar.

Things can be different for small samples. Consider Table 4, which shows the relationship between acute renal failure and death in a group of patients with peritonitis. At least one expected frequency must be less than five, because the first column total is less than 10. In fact, for row 1, column 1 the expected frequency is 2.9. Chi-squared = 5.51, df = 1, $P = 0.019$, whereas Fisher's exact test gives $P = 0.049$, which is much bigger.

For two by two tables, Yates produced a correction to the chi-squared statistic which makes it smaller. For Table 4, Yates' chi-squared = 3.86, df = 1, $P = 0.0495$. This is very close to the Fisher probability.

Fisher's exact test is not as exact as its name suggests. There are two different versions for two by two tables and different programs may produce different answers and its validity is hotly debated. I think it is valid and we should always use it. However, for tables with many rows and columns the calculations become impossible for current computers and we have to use an approximation, which SPSS calls the Monte-Carlo method.

Table 4. Acute renal failure and death in patients with peritonitis (Jonathan Fennell, unpublished)

Status after 3 months	Renal failure		Total
	Yes	No	
Dead	6	31	37
Alive	2	62	64
Total	8	93	101

Table 5. Assessment of radiological appearance at six months as compared with appearance on admission in the MRC streptomycin trial (MRC 1948)

Radiological assessment	Streptomycin	Control
Considerable improvement	28	4
Moderate or slight improvement	10	13
No material change	2	3
Moderate or slight deterioration	5	12
Considerable deterioration	6	6
Deaths	4	14
Total	55	52

Tests for linear association

Table 5 shows the main results of the first real randomised controlled clinical trial, the MRC trial of streptomycin for the treatment of pulmonary tuberculosis (MRC 1948). Patients' chest X-ray images were graded at the start and after six months by three observers blind to treatment. The assessment is clearly ordered, from considerable improvement to considerable deterioration to death.

We can do a chi-squared test for association on this table, which gives $\chi^2 = 26.97$, 5 d.f., $P = 0.0001$. This tests the null hypothesis of no association against the alternative hypothesis of an association of some sort. This association could be that patients given streptomycin are more likely to be in the improvement categories and less likely to be in the deterioration or death categories than control patients, but it could also be that patients given streptomycin are more likely to be in the considerable improvement or considerable deterioration or death categories and less likely to be in the mild/moderate categories than control patients, a kill or cure model. Any association would be against the null hypothesis. If we shuffle the rows round, we get the same chi-squared test, which does not take the order of the rows into account.

Clearly in Table 5, the row categories do have an important order and we would like a test which takes the ordering of the categories into account. Several tests do this, including the Armitage chi-squared test for trend, the Mantel-Haenszel linear-by-linear association, and Kendall's rank correlation tau b. SPSS does the Mantel-Haenszel linear-by-linear association chi-squared test, whether you want it or not. This is almost identical to the Armitage test, which is the one described in *An Introduction to Medical Statistics*. It gives a chi-squared statistic with one degree of freedom which tests the null hypothesis that there is no linear relationship between the row and column variables against the alternative that there is no linear relationship, although there may be a relationship of some other kind.

What do we mean by a linear relationship? First, we assign numerical values to the categories. Usually this is 1, 2, 3, 4, etc. Here Considerable improvement =1, Moderate or slight improvement =2, No material change = 3, Moderate or slight deterioration =4, Considerable deterioration =5, Death =6, and Streptomycin = 1, Control =2. We then say, given these numerical scales, is there a relationship of the form

$$\text{improvement} = \text{constant} + \text{another constant} \times \text{treatment}$$

Obviously, this cannot predict the improvement exactly, but we find the constants which give the best prediction for these data and then test whether they predict the improvement better than we would get by chance. All we are interested in is the test statistic, and we do not even see the two constants. We shall discuss linear trends more in Weeks 9 and 10.

For the data of Table 5, the Mantel-Haenszel linear-by-linear chi-squared = 17.76, 1 d.f., $P < 0.0001$. (For comparison, Armitage chi-squared test for trend gives chi-squared = 17.93, 1 d.f., $P < 0.0001$, which is very similar.)

These trend tests should be valid even when the contingency chi-squared test is not, provided we have at least 30 observations. They can be significant even when the contingency chi-squared is not. They give a more powerful test against a more restricted null hypothesis.

J. M. Bland,
17 August 2006

MRC (1948) Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal* **2**, 769-782.