Chi-Square Tests (12 pages; 19/4/24)

Contents

(A) Contingency Tables

O_i

	voting				
	Con	Lab	Lib	oth	
Male	45	37	12	6	100
Female	24	40	12	9	85
	69	77	24	15	185

(1) Contingency Tables are used to perform a hypothesis test about possible association between two factors, such as gender and voting.

We start with the table of observed frequencies, O_i (where, in this table, O_1 would be 45 and O_2 could be 37, with O_8 being 9; though O_2 could equally well have been 24).

This table is used to generate a table of expected frequencies, E_i.

These are calculated on the basis that the null hypothesis (no association) is true. We reason that, as $\frac{69}{185}$ of the voters are Conservatives, we would expect there to be $\frac{69}{185} \times 100$ male Conservative voters, assuming that there is no association between gender and voting.

Another way of performing the calculation is to say that, of the total of 185 voters, $\frac{69}{185}$ of these are expected to be Conservatives, and of these $\frac{100}{185}$ are expected to be male, so that the expected number of male Conservative voters is $185 \times \frac{69}{185} \times \frac{100}{185}$ (this approach has the advantage that each E_i has the same form). [Note that we could alternatively have taken the total of Conservative voters and multiplied by the proportion of male voters, to give $\frac{100}{185} \times 69$]

Applying this to each of the O_i, we thus obtain the table of expected frequencies:

E_i voting

	Con	Lab	Lib	oth	
male	37.297	41.622	12.973	8.108	100
Female	31.703	35.378	11.027	6.892	85
	69	77	24	15	185

(2) The next step is to measure the deviation of the $O_i s$ from the $E_i s$. In order to avoid positive and negative deviations cancelling out, the $O_i - E_i$ are squared. It is shown in (C) that by forming $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$, we obtain a test statistic that doesn't depend (significantly at least) on the total number of observations ($N = \sum_{i=1}^{n} O_i$).

As discussed in (C), when there is no association between the factors, this test statistic has been found to have an approximate χ^2 distribution, with a certain number of degrees of freedom (discussed below).

Thus, for the example above: $\frac{(O_1 - E_1)^2}{E_1} = \frac{(45 - 37.297)^2}{37.297} = 1.5909$

and $\sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} = 5.9314$, as shown below.

i	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	45	37.297	1.5909
2	37	41.622	0.5133
3	12	12.973	0.0730
4	6	8.108	0.5481
5	24	31.703	1.8716
6	40	35.378	0.6038
7	12	11.027	0.0859
8	9	6.892	0.6448
			5.9314

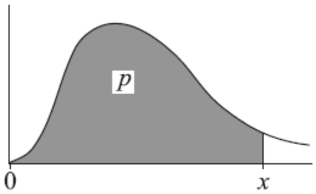
Note: Another implication of including $(O_i - E_i)^2$ is that large differences are given a bigger weighting.

(3) When there are just two rows or two columns (or both), there is a case for having a separate column for $(O_i - E_i)^2$, as a check, as this value will be the same for the two cells in a particular column (row), where there are two rows (columns): if one of the O_i is larger than the corresponding E_i, then the other O_i in the same column must be smaller by the same amount, in order for the column total of the O_i to be the same as that of the E_i.

(4) From the χ^2 table (see below), we can find the 'critical value' of X^2; eg the value of X^2 that will only be exceeded in 5% of cases, when there is no association between the factors. [Note that, confusingly, the table refers to the χ^2 variable as X, rather than X^2.] The degrees of freedom are explained below.

TABLE 6 PERCENTAGE POINTS OF THE χ^2 DISTRIBUTION

The table gives the values of x satisfying $P(X \leqslant x) = p$, where X is a random variable having the χ^2 distribution with ν degrees of freedom.



p / ν	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	p / ν
1	0.00004	0.0002	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	5
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188	10

(5) Degrees of freedom

Various statistical distributions depend on a parameter known as the degrees of freedom.

Clearly when the observed and expected frequencies are being compared in a Contingency table, they are not entirely independent, as the row and column totals of the observed frequencies have been used to determine the expected frequencies. The degrees of freedom measure the extent of

independence. Had the expected frequencies been determined without reference to the observed frequencies, then, in the case of an $m \times n$ table, all mn terms of X^2 would be independent.
To obtain the degrees of freedom (d.f. - denoted by ν [nu]), we deduct 1 from mn for each constraint.
The deductions are as follows:
grand total fixed: -1
row totals fixed $-(m-1)$ [last one covered by grand total]
column totals fixed $-(n-1)$ [last one covered by grand total]

This gives $\nu = mn - 1 - (m-1) - (n-1)$
$= mn - m - n + 1 = (m-1)(n-1)$

(6) Hypothesis Test

For the above example:
H_0: there is no association between gender and voting habits
H_1: there is some association
significance level: eg 5%
Reject H_0 if $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} > {\chi^2}_\nu$
In the example above, $\nu = (2-1)(4-1) = 3$,
so that the critical value, ${\chi^2}_\nu$ (at the 5% level) is 7.815
As $X^2 = 5.9314 < 7.815$, we accept H_0, and conclude that there isn't sufficient evidence of an association between gender and voting habits.

(7) The χ^2 test is one-tailed. The right-hand critical values are the ones that are used when performing hypothesis tests. However, a very low value of X^2 could be regarded as suspicious (eg the data might have been rigged). In theory, this could be tested for using the left-hand critical values from the table.

(8) Special Situations

(i) Small expected frequencies

The test can be shown to be unreliable when any of the E_i are less than 5.

To get round this, factors (of similar type) need to be grouped together. Thus, in the above example, had the "Female/Others" cell had an expected frequency less than 5, then we might group the Liberal and Others factors together. There would be no justification, however, for combining Conservatives and Others (as the Conservatives are not a minor party).

(ii) Yates' Correction for a 2×2 table

The χ^2 model can be shown to be less accurate for 2×2 tables, and the following adjusted value for X^2 has been found to be more appropriate: $\sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$

Note that, in the case of a 2×2 table, all the values of $(|O_i - E_i| - 0.5)^2$ will be the same.

(9) Notes

(i) The sample of observed frequencies must be random.

(ii) Once H_1 has been accepted, values of $\frac{(O_i-E_i)^2}{E_i}$ for particular cells may suggest a theory as to the nature of an association between factors.

(iii) Association does not imply "cause & effect". For example, level of education and future wealth: although there is highly likely to be cause and effect here, there could also be indirect factors connected with genetics or upbringing, which have an impact on both level of education and future wealth.

(B) Goodness of Fit

χ^2 tables are used in a similar way to Contingency tables.

Instead of calculating expected frequencies from the row and column

totals of the Contingency table, they are obtained from a supposed model distribution.

Example: To test whether a die is biased, we could roll it 600 times, and observe the numbers of 1s, 2s etc. There would be 6 cells or 'classes', and the expected frequency in each case would be 100. Then $X^2 = \sum_{i=1}^{6} \frac{(O_i-E_i)^2}{E_i}$

In general, the E_i are obtained by multiplying the total observed frequency by the probability for class i, according to the model

distribution; ie $E_i = p_i N$, where $N = \sum_{i=1}^{n} O_i$
(so that $\sum_{i=1}^{n} E_i = \sum_{i=1}^{n} p_i N = N \sum_{i=1}^{n} p_i = N$)

The number of degrees of freedom is the number of free variables, which is the number of classes less the number of restrictions. Here there is one restriction: the total of the observed frequencies has to equal 600. Thus, $\nu = 6 - 1 = 5$
If, in addition, the observed frequencies are used to estimate parameters of the model distribution, then the degrees of freedom will be given by:
$\nu = \text{number of classes} - 1 - \text{number of parameters estimated}$

In general, the null and alternative hypotheses would be:
H_0: the data are drawn from the model population
H_1: this is not the case

Notes

(i) If a model has been based on data, then fresh data are needed when carrying out the Goodness of Fit test.
(ii) For continuous data, observations would need to be grouped into suitable classes. Discrete data may also be grouped.
(iii) One use of the Goodness of Fit test is in establishing whether data come from a Normal distribution (for example, when deciding whether a t-test is appropriate for a small sample). The mean and variance of the supposed Normal distribution are

estimated from the data.

However, if a t-test is subsequently carried out, it will be necessary to obtain another sample, as the original sample is not entirely random: it is a sample that passes the χ^2 test for Normality.

(iv) Class widths need not be the same. As a rule of thumb, it is best to choose them in such a way that the expected frequencies are in the range $8 - 12$. Thus the class width would be narrower where the model distribution has the greatest density.

(v) In the case of the Binomial distribution $B(n, p)$, n doesn't count as a parameter.

(C) Justification for the test statistic

(1) At first glance, the test statistic $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ would appear to suffer from the following drawback: If a different scale were adopted, so that (for example) the O_i were all multiplied by 10, then the E_i would also be expected to be multiplied by 10. Then, with the same n (the number of classes), X^2 would be multiplied by $\frac{10^2}{10} = 10$. So the test statistic is greater relative to the critical value (which remains the same). This apparent drawback is resolved in (2).

(2) Let $N = \sum_{i=1}^{n} O_i$, where there are n cells in the table (for either a Contingency Table or one for Goodness of Fit). Then

$\sum_{i=1}^{n} E_i = N$ also.

Then $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{n} \frac{{O_i}^2}{E_i} - 2\sum_{i=1}^{n} O_i + \sum_{i=1}^{n} E_i$

$= (\sum_{i=1}^{n} \frac{{O_i}^2}{E_i}) - 2N + N$

$= (\sum_{i=1}^{n} \frac{{O_i}^2}{E_i}) - N$

Then each $\frac{{O_i}^2}{E_i}$ term is of the same order of magnitude as O_i (or E_i), and so X^2 is of the same order of magnitude as $(\sum_{i=1}^{n} O_i) - N$ $= N - N = 0$; ie X^2 will be a small number (not dependent on N).

[The argument in (1) doesn't take account of the fact that the test statistic can be written as the difference of two items of the same order of magnitude. (Had it been the sum, then the argument in (1) would have been valid.)]

(3) The statistician Karl Pearson demonstrated (in 1900) that, when the the null hypothesis of the test is correct (ie when there is no association between the factors), and when the E_i are obtained from $E_i = p_i N$, where the p_i are known (and not based on the observed frequencies O_i), then the distribution of the test statistic $X^2 = (\sum_{i=1}^{n} \frac{{O_i}^2}{E_i}) - N$ tends to that of the ${\chi_{n-1}}^2$ random variable $\sum_{i=1}^{n-1} {Z_i}^2$, as $N \to \infty$, where the Z_i are independent standardised Normal random variables ($n - 1$ being the degrees of freedom).

This corresponds to a Goodness of Fit test where the model is not based on the observed frequencies (eg rolling a die, where the $p_i = \frac{1}{6}$), and the degrees of freedom is the number of cells n, less one (due to the restriction that $\sum_{i=1}^{n} E_i = \sum_{i=1}^{n} O_i$).

(4) Pearson also demonstrated that, when the p_i are estimated from the observed frequencies, $X^2 = (\sum_{i=1}^{n} \frac{{O_i}^2}{E_i}) - N$ is still approximately distributed as χ^2 for large enough N, with a reduction of 1 in the degrees of freedom for each estimation of a parameter of the model (leading to the p_i).

This situation also covers Contingency Tables, where the E_i are derived from the observed frequencies; again, with an appropriate reduction in the degrees of freedom.