Chi squared test for independence in a contingency table

Two properties are associated if the probability of having one property affects the probability of having another. For example, the probability of passing an exam is increased by hard work. Also, it is generally agreed, as a matter of sociological observation, that the probability of having an income above the average is associated with your father's occupational background.

The opposite of association of properties is independence. Two properties are independent if having one property does not alter the probability of having the other. For example, sex and intelligence are thought to be independent. Being male does not increase the chance of being stupid!

Sometimes it is not known whether two properties are associated or not. What is required is a test of association, or, what is equivalent, a test of independence.

The χ^2 distribution can be used as a test of independence. We illustrate this by introducing an example.

Example (1)

A psychologist conducted a survey into the relationship between the way in which a calculator was held and the speed with which 10 arithmetical operations were performed. The calculator could be either placed on a table or held in the hand; the sums could be performed in either less than 2 minutes, between 2 and 3 minutes or more than 3 minutes. The following results were obtained for a sample of 150 children between 12 and 13 years old.

		Mode of Operation	
		On table	Hand held
Speed of computation	< 2	28	12
	2-3	25	35
	> 3	21	29

Test at the 5% significance level whether the mode of operation and speed of computation are associated.

This kind of table is called a contingency table. Since there are three rows and two columns it is a 3×2 contingency table.

H_0: Speed and mode are independent.

H_1: Speed and mode are associated.

In order to determine whether the two variables are associated it is necessary to calculate what the frequencies would be if there was absolutely no connection between them. In other words, we seek the expected frequencies. We can then compare these expected frequencies with the observed frequencies by means of a χ^2 test of goodness of fit.

To determine these expected frequencies, observe that the probability of being in a given row of the contingency table would be

$$P = \frac{\text{row total}}{\text{sample size}}$$

Likewise, the probability of being in a given column would be

$$P = \frac{\text{column total}}{\text{sample size}}$$

Thus the expected probability for a cell that is in the *i*th row and the *j*th column, assuming that the two probabilities are independent, is

$$\mathrm{P}\left(i\text{th}, j\text{th cell}\right) = \frac{\text{total for } i\text{th row}}{n} \cdot \frac{\text{total for } j\text{th column}}{n}$$

where n is the sample size. We calculate expected frequencies by multiplying the probability of being in a given cell by the total sample size, *n.*

Thus, the expected frequency for the cell in the *i*th row and the *j*th column is:

$$\text{Expected frequency} = \frac{i\text{th row total } \cdot j\text{th column total}}{\text{sample size}}$$

Example continued

For our example we find first the row and column totals:

		Table (T)	Hand (H)	Totals
Speed	< 2	28	12	40
	2-3	25	35	60
	> 3	21	29	50
	Totals	74	76	150

The expected frequencies are:

Copyright © Blacksacademy – July 2005

	T	H
<2	$\frac{40\cdot 74}{150}=19.73$	$\frac{40\cdot 76}{150}=20.27$
2 − 3	$\frac{60\cdot 74}{150}=29.60$	$\frac{60\cdot 76}{150}=30.40$
>3	$\frac{50\cdot 74}{150}=24.67$	$\frac{50\cdot 76}{150}=25.33$

Now we have to use the expected and observed frequencies to calculate a test statistic. The χ^2 test statistic is determined by

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

Let us make a table of the value of $\frac{(O-E)^2}{E}$ for each of the entries in the above contingency table. We call each entry a contribution. The contributions made to the χ^2 test statistic are, therefore

	T	H
<2	$\frac{(28-19.73)^2}{19.73}$ $=3.466$	$\frac{(12-20.27)^2}{20.27}$ $=3.374$
2–3	$\frac{(25-29.60)^2}{29.60}$ $=0.715$	$\frac{(35-30.40)^2}{30.40}$ $=0.696$
>3	$\frac{(21-24.67)^2}{24.67}$ $=0.546$	$\frac{(29-25.33)^2}{25.33}$ $=0.532$

The $\sum$ symbol indicates that we make a total of all these entries.

Hence, $\chi^2_{test}=3.466+3.374+0.715+0.696+0.546+0.532=9.329$

In order to compare this with a critical value, we need to know the degrees of freedom of statistic. The degrees of freedom of any row is constrained by one, because the last frequency is determined by the other entries in the row. Likewise the degrees of freedom of a column is the number of columns less one. Thus, in general

Copyright © Blacksacademy – July 2005

$\nu = \text{degrees of freedom} = (\text{row number} - 1)\cdot(\text{column number} - 1)$

Hence, in this example,

$\nu = (3-1)\cdot(2-1) = 2$

Hence, using tables

$\chi^2_{critical}(\nu = 2, p = 0.950) = 5.992$

Then

$\chi^2_{\text{test}} = 9.329 > \chi^2_{critical} = 5.992$

Therefore, we reject H_0 and accept H_1.

The result is significant at the 0.05 or 5% level. This means that there is a 5 in 100 probability that the difference between the two conditions could have arisen by chance.

According to these results the way you use your calculator does affect the speed with which you do a calculation.

In the case of 2×2 contingency tables you are advised to use Yates' correction, i.e. to use test statistic

$$\overline{\chi}^2_{test} = \sum \frac{\left(|O - \text{E}| - \frac{1}{2}\right)^2}{\text{E}}$$

instead of

$$\chi^2_{test} = \sum \frac{(O - \text{E})^2}{\text{E}}.$$

This is especially important with small samples when some or all expected frequencies are less than 5. For large samples, corrected and uncorrected chi-square values may be practically the same and as a consequence the correction factor may be ignored.

Copyright © Blacksacademy – July 2005

Example (2)

Each driver in a sample at size 50 was classified according to both seat-belt usage and sex to obtained the following 2×2 contingence table

		Seat-Belt Usage	
		Don't Use	Use
Sex	Male	6	11
	Female	8	25

Is there any associations between sex and seat-belt usage?

Test it at 10% level of significance

H_0 : Sex and seat-belt usage are independent,

H_1 : Sex and seat-belt usage are associated.

The rows and column totals:

	Don't Use (N)	Use (Y)	Totals
Male (M)	6	11	17
Female (F)	8	25	33
Totals	14	36	50

The expected frequencies:

	N	Y
M	$\frac{17 \cdot 14}{50} = 4.76$	$\frac{17 \cdot 36}{50} = 12.24$
F	$\frac{33 \cdot 14}{50} = 9.24$	$\frac{33 \cdot 36}{50} = 23.76$

$$\nu = (2-1) \cdot (2-1) = 1$$

The test statistic is $\overline{\chi}^2_{test} = \sum \frac{\left(|O - \mathrm{E}| - \frac{1}{2}\right)2}{\mathrm{E}}$.

Contributions made to the test statistic are:

Copyright © Blacksacademy – July 2005

	N	Y
M	$\left(16-4.76-\frac{1}{2}\right)^2$ 4.76 $=0.125$	$\left(11-12.24-\frac{1}{2}\right)^2$ 12.24 $=0.045$
F	$\left(18-9.24-\frac{1}{2}\right)^2$ 9.24 $=0.059$	$\left(25-23.76-\frac{1}{2}\right)^2$ 23.76 $=0.023$

Hence, $\chi^2_{test}=0.115+0.045+0.059+0.023=0.242$

$\chi^2_{critical}\left(\nu=1, p=0.100\right)=2.706$

Then

$\overline{\chi}^2_{\text{test}}<\chi^2_{critical}$

$\therefore$ Accept H_0, reject H_1

There is no evidence of association between sex and seat-belt usage.

Copyright © Blacksacademy – July 2005