Lecture 8 : Chi-squared tests

Jonathan Marchini

November 26, 2004

Introduction

This Lecture is split into 3 section each of which involves or relates to hypthosis tests. In the first section we continue to build up our repertoire of statistical tests by studying Chi-squared tests. Chi-squared tests are a very useful class of tests that can be used to in two general situations. Firstly, Chi-squared tests can be used to test whether a sample of data is consistent with a specific theoretical distribution. For example, we might be interested in whether a sample of data is consistent with a Poisson distribution with some specified mean value λ. More importantly, Chi-squared tests can be used to test for an association between two categorical variables. Such "tests of association" are extremely useful and very common in both Psychology and Human Sciences. In the second section we consider the types of errors we can make when carrying out a statistical test and learn how to calculate the probabilities of making these errors. In the last section we study confidence intervals which are an alternative way to hypothesis tests of describing the result of an experiment.

1 Chi-squared Tests

1.1 Goodness-of-Fit Tests

1.1.1 The Poisson Distribution

In Lecture 5 we saw how to fit a Poisson distribution to a sample of data. We considered two sequences of birth times (Figures 1 3) and we were interested in testing whether each sequence was consistent with a hypothesis of randomly occurring birth times, i.e. we were interested in testing whether the counts of events within hour intervals was consistent with a Poisson distribution (Figures 2 4).

For the first sequence of birth times (Figures 1 and 2) we calculated the sample mean of the data as

$\bar{x} = \frac{44}{24} = 1.8333$

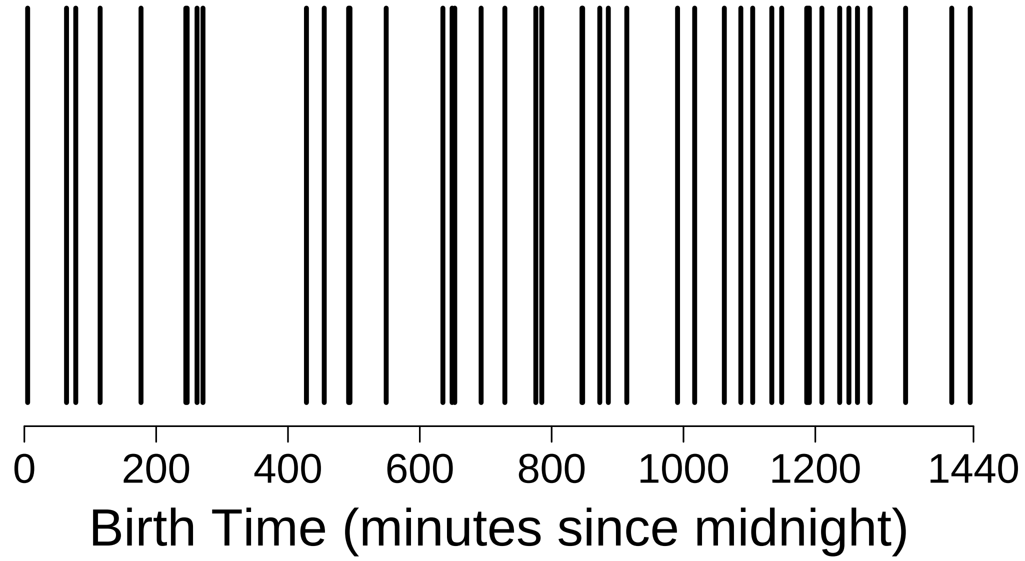


Figure 1: The birth times of the babies in the Babyboom dataset

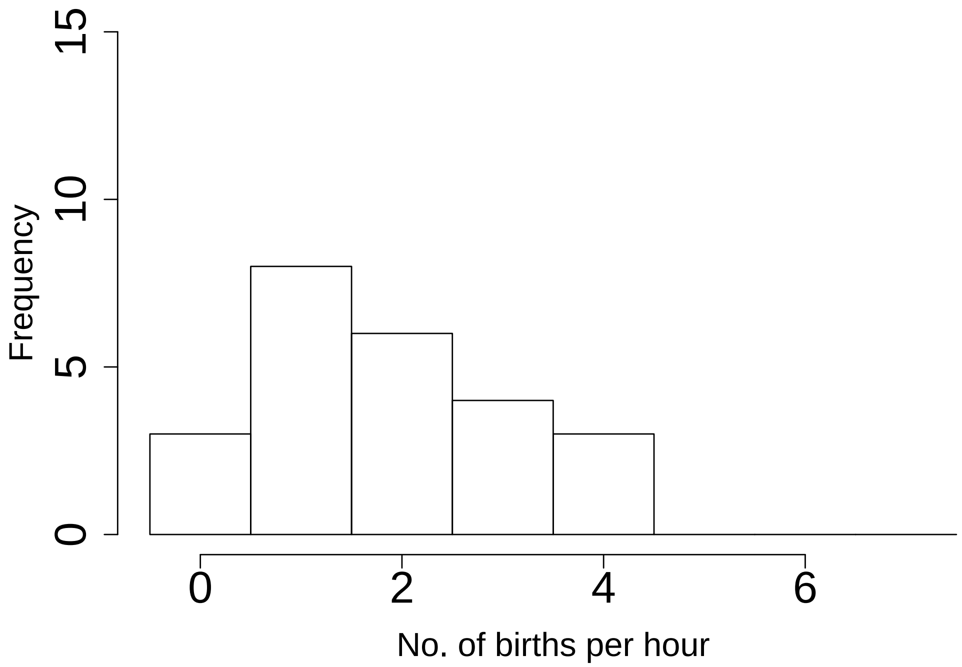


Figure 2: Histogram of birth times per hour of the babies in the Babyboom dataset

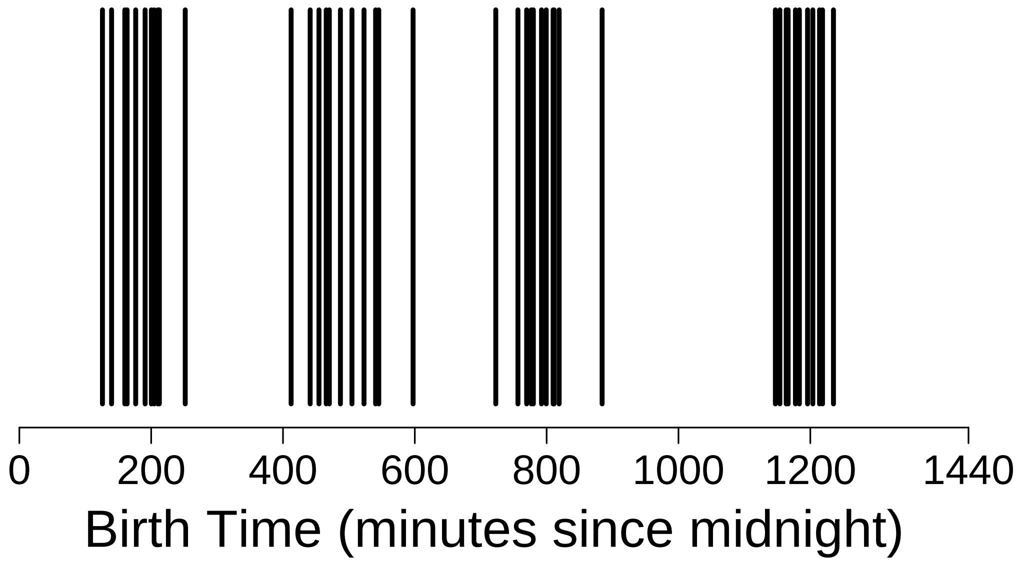


Figure 3: A sequence of non-random birth times

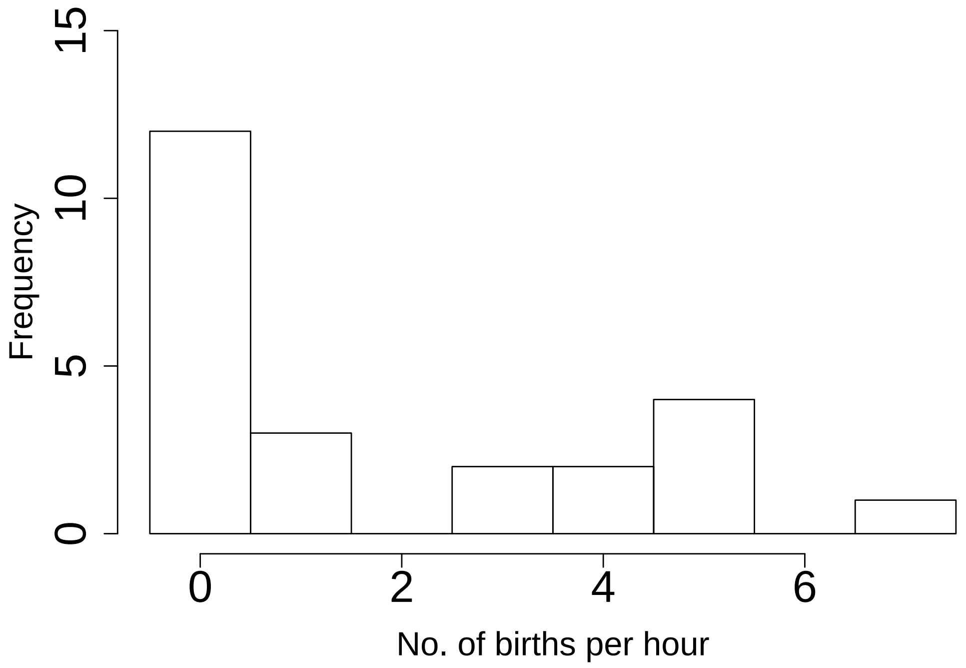


Figure 4: Histogram of birth times per hour of the non-random data.

We then fitted a Poisson distribution with this mean value ($\lambda = 1.8333$) by calculating the expected frequencies of the distribution.

x	0	1	2	3	4	5	≥ 6
Expected	3.84	7.04	6.45	3.94	1.81	0.66	0.27
Observed	3	8	6	4	3	0	0

We noted that the expected frequencies seemed to be quite close to the observed frequencies suggesting that a Poisson distribution might be a good model for the data. At the time we had no way of formally testing our hypothesis that the data are Poisson. Now that we have become familiar with the idea of a hypothesis test we can use a **Chi-squared Goodness-of-Fit Test** to test whether the Poisson distribution is a good fit.

As in the previous lecture the first thing we do is to write down the null and alternative hypotheses.

H_0 : The data follow a Poisson distribution with mean 1.8333
H_1 : The data *do not* follow a Poisson distribution with mean 1.8333

At this point we also decide upon a 5% significance level.

The test statistic used in a Chi-squared Goodness-of-Fit Test is

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

where O_i and E_i are the observed and expected frequencies for the ith cell of the table.

For our example, this would be

$$\begin{aligned} X^2 &= \frac{(3-3.84)^2}{3.84} + \frac{(8-7.04)^2}{7.04} + \frac{(6-6.45)^2}{6.45} \\ &\qquad + \frac{(4-3.94)^2}{3.94} + \frac{(3-1.81)^2}{1.81} + \frac{(0-0.66)^2}{0.66} + \frac{(0-0.27)^2}{0.27} \\ &= 2.06 \end{aligned}$$

The value of this statistic is small when the observed and expected frequencies are close and large when they are not close, thus large values of this test statistic indicate that null hypothesis may be false.

In order to calculate the p-value for the test or calculate the critical region for the test at a given level of significance we need to know the distribution of the test statistic under the assumptions of the null hypothesis. Under these assumptions it can be shown (not in this course) that the test statistic has a Chi-squared (χ^2) distribution (approximately).

A Chi-squared distribution is a continuous probability distribution defined on the range of positive values. The distribution has only one parameter, called the degrees of freedom (df). Figure 5 shows a Chi-squared distribution with 5 degrees of freedom, denoted χ^2_5. The distribution shown in the Figure exhibits positive skew and this is a general property of Chi-squared distributions.

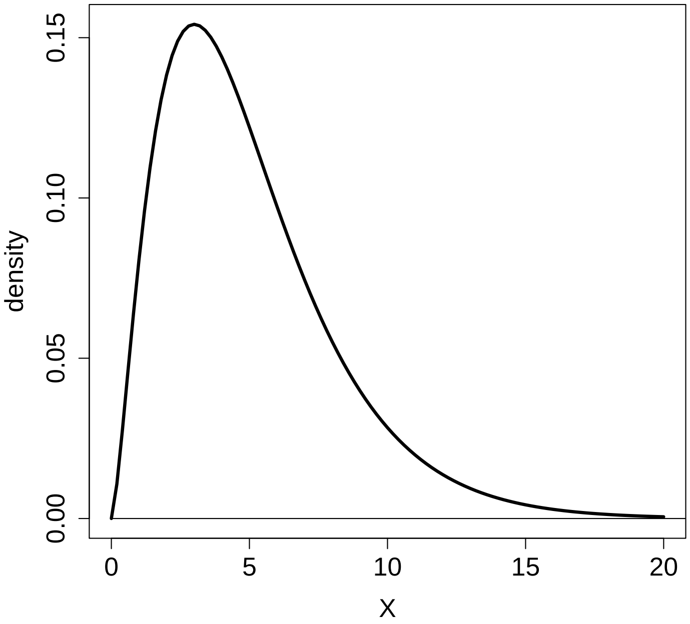


Figure 5: A Chi-squared distribution with 5 degrees of freedom (χ^2_5)

When carrying out a Chi-squared Goodness-of-Fit test the degrees of freedom are calculated as

$$\text{df} = (k-1) - p$$

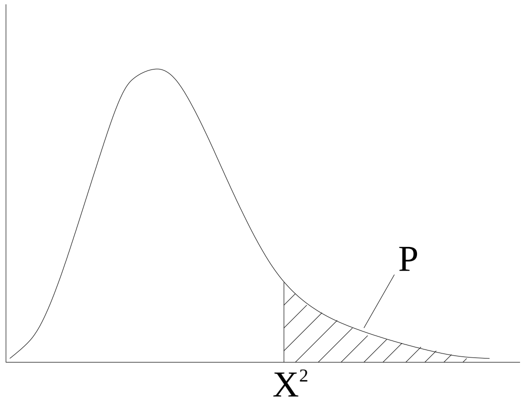
where k is the number of cells and p is the number of parameters estimated in order to fit the distribution.

In our example, df = 5 as $k = 7$ since there are 7 cells in our table and $p = 1$ since we estimated one parameter λ in order to fit the Poisson distribution.

As with the Normal distribution we use tables to look-up the probabilities we wish to calculate. In this example, we want to calculate the critical region for the test, i.e. we want to find a such that $P(\chi^2_5 > a) = 0.05$. The tables accompanying this course allow us to calculate probabilities of this type. The table is reproduced on the next page. The value we need from the table is highlighted with a box.

Thus the critical region for the test is $X^2 > 11.07$.

The test statistic does not lie in the critical region so we conclude that the evidence against the null hypothesis is not significant at the 5% level.



df	P = 0.05	P = 0.01
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	$\boxed{11.07}$	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89
40	55.76	63.69
60	79.08	88.38

Correcting small expected cell counts

We mentioned briefly before that the test statistic is approximately distributed as a χ^2 distribution. In general, this approximation is very good but it is not good when the values E_i fall below 5. To avoid this situation we group together cells so that all the expected counts are above 5. In our example the table would become

x	0-1	2	≥ 3
Expected	10.88	6.45	6.68
Observed	11	6	7

We then re-calculate the test statistic as

$$\begin{aligned} X^2 &= \frac{(11-10.88)^2}{10.88} + \frac{(6-6.45)^2}{6.45} + \frac{(7-6.68)^2}{6.68} \\ &= 0.048 \end{aligned}$$

This is very much smaller than the test statistic we calculated above. If we do not group cells with low expected counts then we may end up wrongly rejecting the null hypothesis.

As the size of the table has changed we need to re-calculate the degrees of freedom. In this case, $k = 3$ and $p = 1$ so that df = 1.

Using tables we can obtain the Critical Region as $X^2 > 3.84$.

The test statistic does not lie in the critical region so we conclude that the evidence against the null hypothesis is not significant at the 5$\%$ level.

Another example

The second sequence we considered in Lecture 5 was definitely not a random sequence (Figure 3 and 4). We estimated the mean of the distribution to be 1.8333. We can now test more formally whether the data represent a random sequence.

As before we start by writing down the null and alternative hypotheses.

H_0 : The data follow a Poisson distribution with mean 1.8333
H_1 : The data *do not* follow a Poisson distribution with mean 1.8333

At this point we also decide upon a 5% significance level.

Using the estimated value of λ (=1.8333) we can fit the distribution to obtain the following table of observed and expected frequencies.

x	0	1	2	3	4	5	≥ 6
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	12	3	0	2	2	4	1

Since some of the expected counts are less than 5 we need to group cells together to produce the table

x	0-1	2	≥ 3
Expected	10.88	6.45	6.68
Observed	15	0	11

We calculate the test statistic as

$$\begin{aligned} X^2 &= \frac{(15-10.88)^2}{10.88} + \frac{(0-6.45)^2}{6.45} + \frac{(11-6.68)^2}{6.68} \\ &= 10.8 \end{aligned}$$

The degrees of freedom are df = 1 and so as above the Critical Region is $X^2 > 3.84$.

In this example the test statistic does lie in the critical region so we conclude that there is significant evidence against the null hypothesis at the 5% level. The only remaining hypothesis we have is that the data does not follow a Poisson distribution with mean 1.8333.

1.1.2 The Binomial Distribution

In 1889 a researcher called Geissler studied hospital records and compiled data on the gender ratio. The table below shows the number of male children in 6115 families with 12 children. If the gender of successive children are independent and the probabilities remain constant over time, the number of males born to a particular family of 12 children should be a binomial random variable with 12 trials and an unknown probability p of success.

x	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	7	45	181	478	829	1112	1343	1033	670	286	104	24	3

We can use a Chi-squared test to test the hypothesis that the data follow a Binomial distribution.

H$_0$: The data follow a Binomial distribution
H$_1$: The data *do not* follow a Binomial distribution

At this point we also decide upon a 5$\%$ significance level.

From the data we know that $n = 6115$ and we can estimate p as

$$\hat{p} = \frac{\bar{x}}{12} = \frac{7(0) + 45(1) + \ldots + 3(12)}{12 \times 6115} = 0.480785$$

Thus we can fit a Bin(12, 0.480785) distribution to the data to obtain the expected frequencies (E) alongside the observed frequencies (O)[1]

x	0	1	2	3	4	5	6	7	8	9	10	11	12
E	2.3	26.1	132.8	410.0	854.2	1265.6	1367.3	1085.2	628.1	258.5	71.8	12.1	0.9
O	7	45	181	478	829	1112	1343	1033	670	286	104	24	3

We see that there are 2 cells with expected counts less than 5 so we group cells to obtain the table

x	0-1	2	3	4	5	6	7	8	9	10	11-12
E	28.4	132.8	410.0	854.2	1265.6	1367.3	1085.2	628.1	258.5	71.8	13.0
O	52	181	478	829	1112	1343	1033	670	286	104	27

The test statistic can then be calculated as

$$\begin{aligned} X^2 &= \frac{(52 - 28.4)^2}{28.4} + \ldots + \frac{(27 - 13.0)^2}{13.0} \\ &= 105.95 \end{aligned}$$

The degrees of freedom ar given by

$$\text{df} = (k-1) - p = (11-1) - 1 = 9$$

---

[1] A Binomial distribution is fitted in the same way as a Poisson distribution (see the Lecture 5 notes)

Thus, the Critical Region for the test is $X^2 > 16.92$.

The test statistics lies well within the Critical Region so we conclude that there is significant evidence against the null hypothesis at the 5% level

1.1.3 The Normal Distribution

The following table gives the heights in cm of 100 students. In such a situation we might be interested in testing whether the data follow a Normal distribution or not.

Height (cm)	155-160	161-166	167-172	173-178	179-184	185-190
Frequency	5	17	38	25	9	6

We can use a Chi-squared test to test the hypothesis that the data follow a Normal distribution.

H_0 : The data follow a Normal distribution
H_1 : The data *do not* follow a Normal distribution

At this point we also decide upon a 5% significance level.

From the data we can estimate the mean and standard deviation using the sample mean and standard deviation

$$\begin{aligned} \bar{x} &= 171.54 \\ s &= 7.1462 \end{aligned}$$

To fit a Normal distribution with this mean and variance we need to calculate the probability of each interval. This is done in four straightforward steps

(i) Calculate the upper end point of each interval (u)

(ii) Standardize the upper end points (z)

(iii) Calculate the probability $P(Z < z)$

(iv) Calculate the probability of each interval

(v) Calculate the expected cell counts

Height (cm)	155-160	161-166	167-172	173-178	179-184	185-190
Endpoint (u)	160.5	166.5	172.5	178.5	184.5	∞
Standardized (z)	-1.54	-0.71	0.13	0.97	1.81	∞
$P(Z < z)$	0.06	0.24	0.55	0.83	0.96	1.00
$P(a < Z < b)$	0.06	0.18	0.31	0.28	0.13	0.04
E	6	18	31	28	13	4

From this table we see that there is one cell with an expected count less than 5 so we group it together with the nearest cell

Height (cm)	155-160	161-166	167-172	173-178	179-190
E	6	18	31	28	17
O	5	17	38	25	15

We can then calculate the test statistic as

$$\begin{aligned} X^2 &= \frac{(5-6)^2}{6} + \ldots + \frac{(15-17)^2}{17} \\ &= 2.36 \end{aligned}$$

The degrees of freedom ar given by

$$\text{df} = (k-1) - p = (5-1) - 2 = 2$$

Thus, the Critical Region for the test is $X^2 > 5.99$.

The test statistics lies outside the Critical Region so we conclude that the evidence against the null hypothesis is not significant at the 5% level.

1.2 Chi-squared Tests of Association

This section develops a Chi-squared test that is very similar to the one of the preceding section but aimed at answering a slightly different question. To illustrate the test we use an example.

A psychology experiment was done to investigate the effect of anxiety on a person's desire to be alone or in company. A group of 30 subjects was randomly divided into two groups of sizes 13 and 17. One group, called the "high anxiety" group, was told that they would experience some painful electric shocks. The other, "low anxiety" group was told they would receive some electric shocks but that they would be mild and painless. Both groups were told that there would be a 10min wait before the experiment began, and each subject was given the choice of waiting alone or with the other subjects. The data from the experiment are presented in the table below.

	Wait Together (T)	Wait Alone (A)
High-Anxiety (H)	12	5
Low-Anxiety (L)	4	9

This table is an example of a **contingency table**, in which a sample of data is cross-classified in a table with r rows and c columns.

The research hypothesis in this situation is whether anxiety is associated with a person's desire to be alone or in company. For example, we might think that the proportion of people wanting to wait together would be higher in the high anxiety group than in the low anxiety group. This is in fact what we observe but we need to be able to test for the association in a more scientifically rigorous way than simply looking at the data. Thus, we want to perform an appropriate hypothesis test. In this situation, the null hypothesis would be that there is no association between the two variables. In other words, the null hypothesis is that the two variables are independent.

H_0 : The two variables are independent.
H_1 : The two variables are associated.

To carry out the test we calculate the expected cell frequencies under the assumption of independence and compare these to the observed frequencies using the same Chi-squared test statistic that we used in the previous section.

Under the assumption of independence the probability that person is both in the High Anxiety (H) group and chooses to Wait Together (T) is given by the Multiplication Law

$$P(H \cap T) = P(H)P(T)$$

This is the probability of being in the first cell (top left) of the table. As we don't know the probabilities we can estimate them from the data, i.e.

$$\begin{aligned} P(H) &= \frac{17}{30} \\ P(T) &= \frac{16}{30} \\ \Rightarrow P(H \cap T) &= \frac{17}{30}\frac{16}{30} = 0.30222 \end{aligned}$$

Similarly we can calculate the probabilities of each of the other cells.

	T	A	
H	$(\frac{17}{30})(\frac{16}{30})$	$(\frac{17}{30})(\frac{14}{30})$	$P(H) = \frac{17}{30}$
L	$(\frac{13}{30})(\frac{16}{30})$	$(\frac{13}{30})(\frac{14}{30})$	$P(L) = \frac{13}{30}$
	$P(T) = \frac{16}{30}$	$P(A) = \frac{14}{30}$	

Once we have calculated the probabilities for each cell under the assumption of independence we can calculate the expected cell counts by simply multiplying the probabilities by the number of observations in our data set, i.e. 30.

	T	A
H	$30 \times (\frac{17}{30})(\frac{16}{30}) = 9.07$	$30 \times (\frac{17}{30})(\frac{14}{30}) = 7.93$
L	$30 \times (\frac{13}{30})(\frac{16}{30}) = 6.93$	$30 \times (\frac{13}{30})(\frac{14}{30}) = 6.07$

Thus we have a table of observed frequencies and a table of expected frequencies

Expected Frequencies

	T	A
H	9.07	7.93
L	6.93	6.07

Observed Frequencies

	T	A
H	12	5
L	4	9

We can then calculate the Chi-squared statistic

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

where O_i and E_i are the observed and expected frequencies for the ith cell of the table.

For this example, we have

$$\begin{aligned} X^2 &= \frac{(12 - 9.07)^2}{9.07} + \frac{(5 - 7.93)^2}{7.93} + \frac{(4 - 6.93)^2}{6.93} + \frac{(9 - 6.07)^2}{6.07} \\ &= 4.6822 \end{aligned}$$

Under the null hypothesis, H_0, this test statistic has a Chi-squared distribution with the number of degrees of freedom given by

$$\text{df} = (r - 1)(c - 1)$$

In our example, we have a 2×2 table, so $r = 2$ and $c = 2$ and the number of degrees of freedom is

$$\text{df} = (2-1)(2-1) = 1$$

Thus the Critical Region of the test is $X^2 > 3.84$.

The test statistics lies inside the Critical Region so we conclude that there is significant evidence against the null hypothesis at the 5% level.

Another example

The following contingency table contains the data collected in a study to examine the relationship between the incidence of tuberculosis and the ABO blood group.

	O	A	AB	B
Moderate/Advanced	7	5	3	13
Minimal	27	32	8	18
Not present	55	50	7	24

We can use a Chi-squared test to test for an association between incidence of tuberculosis and the ABO blood group.

H_0 : Incidence of Tuberculosis and ABO blood group are independent.
H_1 : Incidence of Tuberculosis and ABO blood group are associated.

We set the significance level at 5%.

The first thing we need to do is calculate the marginal totals of the table

	O	A	AB	B	TOTAL
Moderate/Advanced	7	5	3	13	28
Minimal	27	32	8	18	85
Not present	55	50	7	24	136
TOTAL	89	87	18	55	249

From the marginal totals we can calculate the expected counts of all the cells

	O	A	AB	B
Moderate/Advanced	$249(\frac{89}{249})(\frac{28}{249})$	$249(\frac{87}{249})(\frac{28}{249})$	$249(\frac{18}{249})(\frac{28}{249})$	$249(\frac{55}{249})(\frac{28}{249})$
Minimal	$249(\frac{89}{249})(\frac{85}{249})$	$249(\frac{87}{249})(\frac{85}{249})$	$249(\frac{18}{249})(\frac{85}{249})$	$249(\frac{55}{249})(\frac{85}{249})$
Not present	$249(\frac{89}{249})(\frac{136}{249})$	$249(\frac{87}{249})(\frac{136}{249})$	$249(\frac{18}{249})(\frac{136}{249})$	$249(\frac{55}{249})(\frac{136}{249})$

$\Rightarrow$

	O	A	AB	B
Moderate/Advanced	10.0	9.8	2.0	6.2
Minimal	30.4	29.7	6.1	18.8
Not present	48.6	47.5	9.8	30.0

From the observed and expected counts we can calculate the test statistic as

$$\begin{aligned} X^2 &= \frac{(7-10.0)^2}{10.0} + \ldots + \frac{(24-9.8)^2}{9.8} \\ &= 15.37 \end{aligned}$$

For a 3×4 table the degrees of freedom are

$$\text{df} = (3-1)(4-1) = 6$$

Thus the Critical Region of the test is $X^2 > 12.59$.

The test statistics lies inside the Critical Region so we conclude that there is significant evidence against the null hypothesis at the 5% level.