

Lecture 8 : Chi-squared tests

Jonathan Marchini

Outline

Goodness-of-Fit Tests

Is a sample of data consistent with a given probability distribution ?

Association Tests

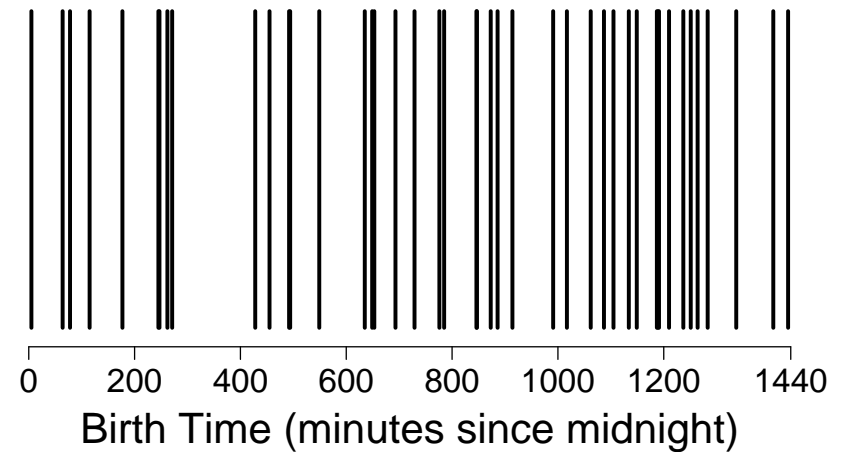
Are two categorical variables independent or are they associated/correlated?

Chi-squared Goodness-of-Fit Tests : Poisson

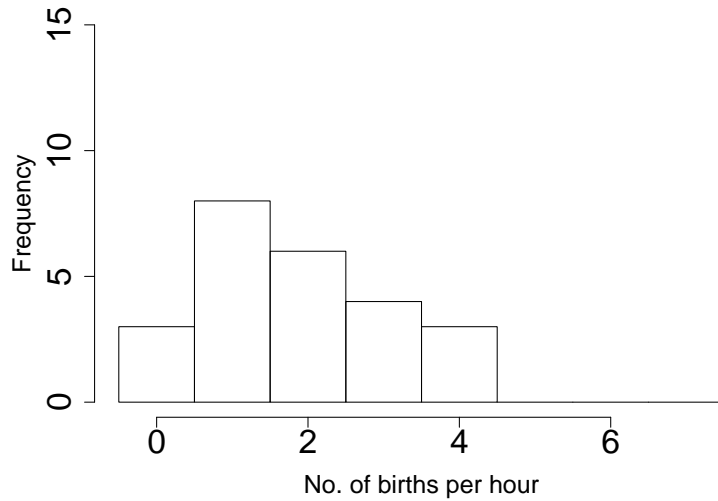
In Lecture 5 we saw how to fit a Poisson distribution to a sample of data.

We considered two sequences of birth times 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.

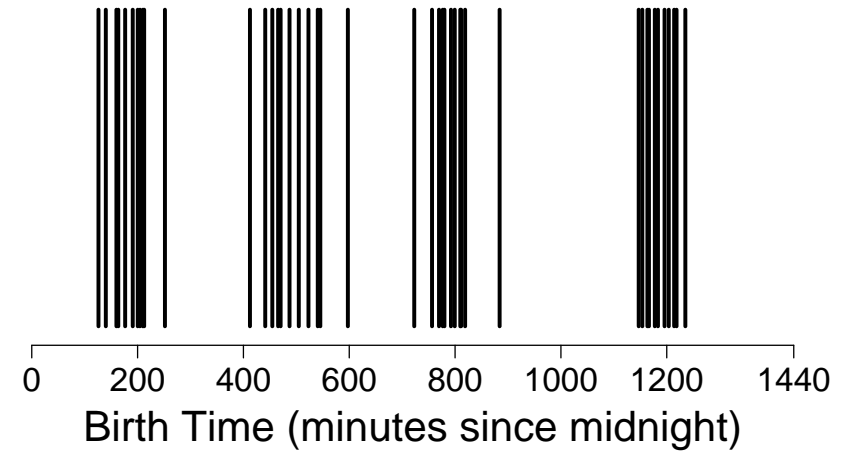
Random birth times



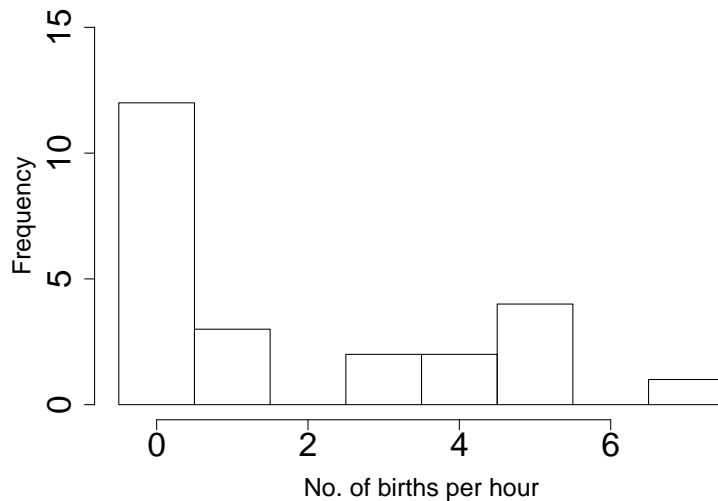
Histogram of random birth times



Non-random birth times



Histogram of non-random birth times



Fitting a Poisson Distribution

For the first sequence of birth times we calculated the sample mean of the data as

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

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 |

Chi-squared Goodness-of-Fit 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

H_1 : The data *do not* follow a Poisson distribution

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

For our example, this would be

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

Chi-squared Test Statistic

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 i th cell of the table.

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).

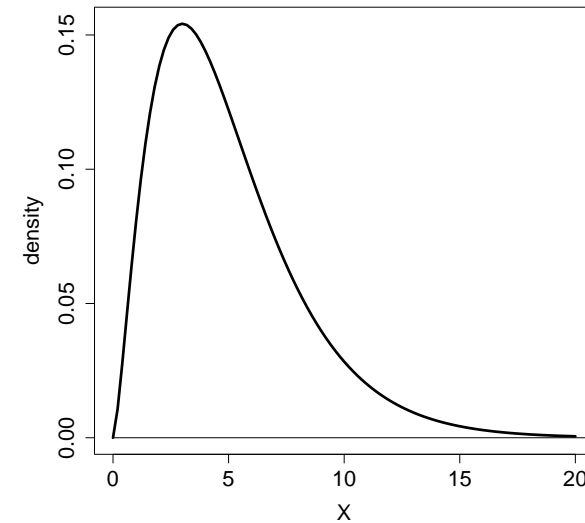
The Chi-squared Distribution

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).

The distribution exhibits positive skew and this is a general property of Chi-squared distributions.

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



Calculating the degrees of freedom

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

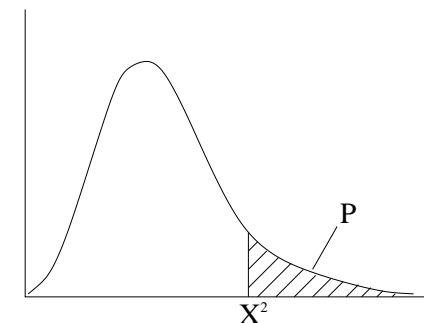
$$df = (k - 1) - p$$

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.

Determining the Critical Region

To calculate the critical region for the test we want to find a such that $P(\chi^2_5 > a) = 0.05$.



Chi-squared Tables

The formula book contains tables of critical values (Table 3 p.8)

| 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 | 11.07 | 15.09 |
| 6 | 12.59 | 16.81 |
| ⋮ | ⋮ | ⋮ |
| 60 | 79.08 | 88.38 |

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.

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

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

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.

Chi-squared Goodness-of-Fit Tests : Binomial

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 |

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

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

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

NB. A Binomial distribution is fitted in the same way as a Poisson distribution (see the Lecture 5 notes)

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.

| | | | | | | | | | | | | | |
|-----|-----|------|-------|-------|-------|--------|--------|--------|-------|-------|------|------|-----|
| 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 |

Chi-squared Tests of Association

The test statistic can then be calculated as

$$X^2 = \frac{(52 - 28.4)^2}{28.4} + \dots + \frac{(27 - 13.0)^2}{13.0}$$
$$= 105.95$$

The degrees of freedom are given by

$$df = (k - 1) - p = (11 - 1) - 1 = 9$$

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

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 (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, and each subject was given the choice of waiting alone or with others.

Contingency Tables

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.

Null and Alternative Hypotheses

The research hypothesis in this situation is whether anxiety is associated with a person's desire to be alone or in company.

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.

Calculating the Expected counts

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 to test distribution fits.

Under the assumption of independence we can calculate the probability of each cell of the table

To calculate the cell probabilities we multiply together the estimated probabilities of being in each row and column.

| | | | |
|----------|----------|----------|----|
| | T | A | |
| H | 12 | 5 | 17 |
| L | 4 | 9 | 13 |
| | 16 | 14 | 30 |

| | | | |
|----------|----------------------------------|----------------------------------|------------------------|
| | 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

| | | |
|----------------------|-------------|----------|
| | T | A |
| Observed Frequencies | H 12 | 5 |
| | L 4 | 9 |

| | | |
|----------------------|---------------|----------|
| | T | A |
| Expected Frequencies | H 9.07 | 7.93 |
| | L 6.93 | 6.07 |

Calculating the Chi-squared test statistic

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 i th 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

$$df = (r - 1)(c - 1)$$

In our example, we have a 2×2 table, so

$$df = (2 - 1)(2 - 1) = 1$$

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 |

This is an example of a 3×4 Contingency Table.

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 |
|-----------|--|--|--|--|
| MA | $249 \left(\frac{89}{249} \right) \left(\frac{28}{249} \right)$ | $249 \left(\frac{87}{249} \right) \left(\frac{28}{249} \right)$ | $249 \left(\frac{18}{249} \right) \left(\frac{28}{249} \right)$ | $249 \left(\frac{55}{249} \right) \left(\frac{28}{249} \right)$ |
| MI | $249 \left(\frac{89}{249} \right) \left(\frac{85}{249} \right)$ | $249 \left(\frac{87}{249} \right) \left(\frac{85}{249} \right)$ | $249 \left(\frac{18}{249} \right) \left(\frac{85}{249} \right)$ | $249 \left(\frac{55}{249} \right) \left(\frac{85}{249} \right)$ |
| NP | $249 \left(\frac{89}{249} \right) \left(\frac{136}{249} \right)$ | $249 \left(\frac{87}{249} \right) \left(\frac{136}{249} \right)$ | $249 \left(\frac{18}{249} \right) \left(\frac{136}{249} \right)$ | $249 \left(\frac{55}{249} \right) \left(\frac{136}{249} \right)$ |

⇒

| | 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

$$X^2 = \frac{(7 - 10.0)^2}{10.0} + \dots + \frac{(24 - 9.8)^2}{9.8}$$

$$= 15.37$$

For a 3×4 table the degrees of freedom are

$$df = (3 - 1)(4 - 1) = 6$$

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.

Next term

Lecture 1: Confidence Interval. Power of Test. t-test

Lecture 2: t-test. The Central Limit Theorem

Lecture 3: Analysis of Variance

Lecture 4: Linear Regression

Lecture 5: Non-parametric Statistics I

Lecture 6: Non-parametric Statistics II

Lecture 7: Revision on MT2004s lectures

Lecture 8: Revision on HT2005s lectures

webpage <http://www.stats.ox.ac.uk/~lim/Teaching.html>