9 Chi-squared tests of categorical data

A statistician is someone who refuses to play the national lottery, but who does eat British beef. (anonymous)

9.1 Pearson's chi-squared statistic

Suppose, as in Section 8.6, that we observe $x_1, \dots, x_k$, the numbers of times that each of k possible outcomes occurs in n independent trials, and seek to make the **goodness-of-fit test** of

$$H_0: \; p_i = p_i(\theta) \text{ for } \theta \in \Theta_0 \quad \text{against} \quad H_1: \; p_i \text{ are unrestricted.}$$

Recall

$$2 \log L_x(H_0, H_1) = 2 \sum_{i=1}^k x_i \log \hat{p}_i - 2 \sum_{i=1}^k x_i \log p_i(\hat{\theta}) = 2 \sum_{i=1}^k x_i \log\bigl(\hat{p}_i / p_i(\hat{\theta})\bigr),$$

where $\hat{p}_i = x_i/n$ and $\hat{\theta}$ is the MLE of θ under H_0. Let $o_i = x_i$ denote the number of time that outcome i occurred and let $e_i = np_i(\hat{\theta})$ denote the expected number of times it would occur under H_0. It is usual to display the data in k cells, writing o_i in cell i. Let $\delta_i = o_i - e_i$. Then

$$\begin{aligned}
2 \log L_x(H_0, H_1) &= 2 \sum_{i=1}^k x_i \log\bigl((x_i/n)/p_i(\hat{\theta})\bigr) \\
&= 2 \sum_{i=1}^k o_i \log(o_i/e_i) \\
&= 2 \sum_{i=1}^k (\delta_i + e_i) \log(1 + \delta_i/e_i) \\
&= 2 \sum_{i=1}^k (\delta_i + e_i)(\delta_i/e_i - \delta_i^2/2e_i^2 + \cdots) \\
&\doteqdot \sum_{i=1}^k \delta_i^2/e_i \\
&= \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \qquad (1)
\end{aligned}$$

This is called the **Pearson chi-squared statistic**.

For H_0 we have to choose θ. Suppose the optimization over θ has p degrees of freedom. For H_1 we have $k-1$ parameters to choose. So the difference of these **degrees of freedom** is $k-p-1$. Thus, if H_0 is true the statistic (1) $\sim \chi^2_{k-p-1}$ approximately. A mnemonic for the d.f. is

$$\text{d.f.} = \#(\text{cells}) - \#(\text{parameters estimated}) - 1. \qquad (2)$$

Note that

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \left[\frac{o_i^2}{e_i} - 2o_i + e_i\right] = \sum_{i=1}^k \frac{o_i^2}{e_i} - 2n + n = \sum_{i=1}^k \frac{o_i^2}{e_i} - n\,. \qquad (3)$$

Sometimes (3) is easier to compute than (1).

Example 9.1 For the data from Mendel's experiment, the test statistic has the value 0.618. This is to be compared to χ^2_3, for which the 10% and 95% points are 0.584 and 7.81. Thus we certainly do not reject the theoretical model. Indeed, we would expect the observed counts to show even greater disparity from the theoretical model about 90% of the time.

Similar analysis has been made of many of Mendel's other experiments. The data and theory turn out to be too close for comfort. Current thinking is that Mendel's theory is right but that his data were massaged by somebody (Fisher thought it was Mendel's gardening assistant) to improve its agreement with the theory.

9.2 χ^2 test of homogeneity

Suppose we have a rectangular array of cells with m rows and n columns, with X_{ij} items in the (i,j) th cell of the array. Denote the row, column and overall sums by

$$X_{i\cdot} = \sum_{j=1}^n X_{ij}, \quad X_{\cdot j} = \sum_{i=1}^m X_{ij}, \quad X_{\cdot\cdot} = \sum_{i=1}^m \sum_{j=1}^n X_{ij}.$$

Suppose the row sums are fixed and the distribution of $(X_{i1}, \dots, X_{in})$ in row i is multinomial with probabilities $(p_{i1}, \dots, p_{in})$, independently of the other rows. We want to test the hypothesis that the distribution in each row is the same, i.e., $H_0: p_{ij}$ is the same for all i, $(= p_j)$ say, for each $j = 1, \dots, n$. The alternative hypothesis is $H_1: p_{ij}$ are unrestricted. We have

$$\log f(x) = \text{const} + \sum_i \sum_j x_{ij} \log p_{ij}\,, \quad \text{so that}$$

$$\sup_{H_1} \log f(x) = \text{const} + \sup \left\{ \sum_{i=1}^m \sum_{j=1}^n x_{ij} \log p_{ij} \;\middle|\; 0 \le p_{ij} \le 1, \; \sum_{j=1}^n p_{ij} = 1 \quad \forall i \right\}$$

Now, $\sum_j x_{ij} \log p_{ij}$ may be maximized subject to $\sum_j p_{ij} = 1$ by a Lagrangian technique. The maximum of $\sum_j x_{ij} \log p_{ij} + \lambda\left(1 - \sum_j p_{ij}\right)$ occurs when $x_{ij}/p_{ij} = \lambda$,

$\forall j$. Then the constraints give $\lambda = \sum_j x_{ij}$ and the corresponding maximizing p_{ij} is $\hat{p}_{ij} = x_{ij}/\sum_j x_{ij} = x_{ij}/x_{i\cdot}$. Hence,

$$\sup_{H_1} \log f(x) = \text{const} + \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}\log(x_{ij}/x_{i\cdot}).$$

Likewise,

$$\begin{aligned}\sup_{H_0} \log f(x) &= \text{const} + \sup\left\{\sum_i\sum_j x_{ij}\log p_j \,\middle|\, 0 \le p_j \le 1,\ \sum_j p_j = 1\right\},\\ &= \text{const} + \sum_i\sum_j x_{ij}\log(x_{\cdot j}/x_{\cdot\cdot}).\end{aligned}$$

Here $\hat{p}_j = x_{\cdot j}/x_{\cdot\cdot}$. Let $o_{ij} = x_{ij}$ and write $e_{ij} = \hat{p}_j x_{i\cdot} = (x_{\cdot j}/x_{\cdot\cdot})x_{i\cdot}$ for the expected number of items in position (i,j) under H_0. As before, let $\delta_{ij} = o_{ij} - e_{ij}$. Then,

$$\begin{aligned}2\log L_x(H_0,H_1) &= 2\sum_i\sum_j x_{ij}\log(x_{ij}x_{\cdot\cdot}/x_{i\cdot}x_{\cdot j})\\ &= 2\sum_i\sum_j o_{ij}\log(o_{ij}/e_{ij})\\ &= 2\sum_i\sum_j(\delta_{ij}+e_{ij})\log(1+\delta_{ij}/e_{ij})\\ &\doteq \sum_i\sum_j \delta_{ij}^2/e_{ij}\\ &= \sum_i\sum_j(o_{ij}-e_{ij})^2/e_{ij}\,. \qquad (4)\end{aligned}$$

For H_0, we have $(n-1)$ parameters to choose, for H_1 we have $m(n-1)$ parameters to choose, so the **degrees of freedom** is $(n-1)(m-1)$. Thus, if H_0 is true the statistic (4) $\sim \chi^2_{(n-1)(m-1)}$ approximately.

Example 9.2 The observed (and expected) counts for the study about aspirin and heart attacks described in Example 1.2 are

	Heart attack	**No heart attack**	**Total**
Aspirin	104 (146.52)	10,933 (10890.5)	11,037
Placebo	189 (146.48)	10,845 (10887.5)	11,034
Total	293	21,778	22,071

E.g., $e_{11} = \left(\frac{293}{22071}\right) 11037 = 146.52$. The χ^2 statistic is

$$\frac{(104-146.52)^2}{146.52} + \frac{(189-146.48)^2}{46.48} + \frac{(10933-10890.5)^2}{10890.5} + \frac{(10845-10887.5)^2}{10887.5} = 25.01\,.$$

The 95% point of χ_1^2 is 3.84. Since $25.01 > 3.84$, we reject the hypothesis that heart attack rate is independent of whether the subject did or did not take aspirin.

Note that if there had been only a tenth as many subjects, but the same percentages in each in cell, the statistic would have been 2.501 and not significant.

9.3 χ^2 test of row and column independence

This χ^2 test is similar to that of Section 9.2, but the hypotheses are different. Again, observations are classified into a $m \times n$ rectangular array of cells, commonly called a **contingency table**. The null hypothesis is that the row into which an observation falls is independent of the column into which it falls.

Example 9.3 *A researcher pretended to drop pencils in a lift and observed whether the other occupant helped to pick them up.*

	Helped	**Did not help**	**Total**
Men	370 (337.171)	950 (982.829)	1,320
Women	300 (332.829)	1,003 (970.171)	1,303
Total	670	1,953	2,623

To test the independence of rows and columns we take

$$H_0 : p_{ij} = p_i q_j \text{ with } 0 \le p_i, q_j \le 1,\ \sum_i p_i = 1,\ \sum_j q_j = 1\,;$$

$$H_1 : p_{ij} \text{ arbitrary s.t. } 0 \le p_{ij} \le 1,\ \sum_{i,j} p_{ij} = 1\,.$$

The same approach as previously gives MLEs under H_0 and H_1 of

$$\hat{p}_i = x_{i\cdot}/x_{\cdot\cdot},\quad \hat{q}_j = x_{\cdot j}/x_{\cdot\cdot},\quad e_{ij} = \hat{p}_i\hat{q}_j x_{\cdot\cdot} = (x_{i\cdot}x_{\cdot j}/x_{\cdot\cdot}),\quad \text{and}\quad \hat{p}_{ij} = x_{ij}/x_{\cdot\cdot}\,.$$

The test statistic can again be show to be about $\sum_{ij}(o_{ij}-e_{ij})^2/e_{ij}$. The e_{ij} are shown in parentheses in the table. E.g., $e_{11} = \hat{p}_1\hat{q}_1 n = \left(\frac{1320}{2623}\right)\left(\frac{670}{2623}\right) 2623 = 337.171$. The number of free parameters under H_1 and H_0 are $mn-1$ and $(m-1)+(n-1)$ respectively. The difference of these is $(m-1)(n-1)$, so the statistic is to be compared to $\chi^2_{(m-1)(n-1)}$. For the data above this is 8.642, which is significant compared to χ_1^2.

We have now seen Pearson χ^2 tests in three different settings. Such a test is appropriate whenever the data can be viewed as numbers of times that certain outcomes have occurred and we wish to test a hypothesis H_0 about the probabilities with which they occur. Any unknown parameter is estimated by maximizing the likelihood function that pertains under H_0 and e_i is computed as the expected number of times outcome i occurs if that parameter is replaced by this MLE value. The statistic is (1), where the sum is computed over all cells. The d.f. is given by (2).