Lecture 9. Tests of goodness-of-fit and independence

9.1. Goodness-of-fit of a fully-specified null distribution

Suppose the observation space $\mathcal{X}$ is partitioned into $k$ sets, and let $p_i$ be the probability that an observation is in set $i$, $i = 1, \ldots, k$. Consider testing $H_0$: the $p_i$'s arise from a fully specified model against $H_1$: the $p_i$'s are unrestricted (but we must still have $p_i \geq 0$, $\sum p_i = 1$). This is a goodness-of-fit test.

Example 9.1

Birth month of admissions to Oxford and Cambridge in 2012

<table>
<thead>
<tr>
<th>Month</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>470</td>
<td>515</td>
<td>470</td>
<td>457</td>
<td>473</td>
<td>381</td>
<td>466</td>
<td>457</td>
<td>437</td>
<td>396</td>
<td>384</td>
<td>394</td>
</tr>
</tbody>
</table>

Are these compatible with a uniform distribution over the year? □

Example 9.1 continued:

Under $H_0$ (no effect of month of birth), $\hat{p}_i$ is the proportion of births in month $i$ in 1993/1994 - this is not simply proportional to number of days in month, as there is for example an excess of September births (the ‘Christmas effect’).

Suppose the observation space $\mathcal{X}$ is partitioned into $k$ sets, and let $p_i$ be the probability that an observation is in set $i$, $i = 1, \ldots, k$.

For a generalised likelihood ratio test of $H_k: \sum n_i \log p_i = 0$, we need to find the maximised likelihood under $H_0$ and $H_1$.

Under $H_1$: the loglikelihood is

$$L = \text{constant} + \sum n_i \log p_i.$$ We want to maximise this subject to $\sum p_i = 1$.

By considering the Lagrangian $\mathcal{L} = \sum n_i \log p_i - \lambda (\sum p_i - 1)$, we find mle's $\hat{p}_i = n_i/n$. Also $|\Theta_k| = k - 1$.

Under $H_0$: $H_0$ specifies the values of the $p_i$'s completely, $p_i = \hat{p}_i$; say, so $|\Theta_0| = 0$.

Putting these two together, we find

$$2 \log \Lambda = 2 \sum n_i \log \left( \frac{\hat{p}_i}{p_i^\alpha} \right) = 2 \sum n_i \log \left( \frac{n_i}{\hat{p}_i} \right), \quad (1)$$

Here $|\Theta_k| - |\Theta_0| = k - 1$, so we reject $H_0$ if $2 \log \Lambda > \chi^2_{k-1}(\alpha)$ for an approximate size $\alpha$ test.
Likelihood ratio tests

A similar common situation has $H_0: p_i = p_i(\theta)$ for some parameter $\theta$ and $H_1$ as before. Now $|\Theta_0|$ is the number of independent parameters to be estimated under $H_0$.

Under $H_0$: we find mle $\hat{\theta}$ by maximising $\sum n_i \log p_i(\theta)$, and then

$$2 \log \Lambda = 2 \log \left( \frac{\hat{p}_1^{n_1} \cdots \hat{p}_k^{n_k}}{p_1(\theta)^{n_1} \cdots p_k(\theta)^{n_k}} \right) = 2 \sum n_i \log \left( \frac{n_i}{np_i(\hat{\theta})} \right).$$

(2)

Now the degrees of freedom are $k - 1 - |\Theta_0|$, and we reject $H_0$ if $2 \log \Lambda > \chi^2_{k-1-|\Theta_0|}(\alpha)$.

Example 9.1 continued using R:

```r
chisq.test(n, p=ptilde)
data:  n
X-squared = 44.6912, df = 11, p-value = 5.498e-06
```

Pearson’s Chi-squared tests

Notice that (??) and (??) are of the same form.

Let $o_i = n_i$ (the observed number in $i$th set) and let $e_i = n \tilde{p}_i$ in (??) or $np_i(\hat{\theta})$ in (??). Let $\delta_i = o_i - e_i$. Then

$$2 \log \Lambda = 2 \sum o_i \log \left( \frac{o_i}{e_i} \right)$$

$$= 2 \sum (e_i + \delta_i) \log \left( 1 + \frac{\delta_i}{e_i} \right)$$

$$\approx 2 \sum \left( \frac{\delta_i^2}{e_i} - \frac{\delta_i^3}{2e_i^2} \right)$$

$$= \sum \frac{\delta_i^2}{e_i} = \sum \frac{(o_i - e_i)^2}{e_i}.$$

where we have assumed $\log \left( 1 + \frac{\delta_i}{e_i} \right) \approx \frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2}$, ignored terms in $\delta_i^3$, and used that $\sum \delta_i = 0$ (check).

This is Pearson’s chi-squared statistic; we refer it to $\chi^2_{k-1-|\Theta_0|}$.

Example 9.2

Mendel crossed 556 smooth yellow male peas with wrinkled green female peas. From the progeny let

- $N_1$ be the number of smooth yellow peas,
- $N_2$ be the number of smooth green peas,
- $N_3$ be the number of wrinkled yellow peas,
- $N_4$ be the number of wrinkled green peas.

We wish to test the goodness of fit of the model $H_0: (p_1, p_2, p_3, p_4) = (9/16, 3/16, 3/16, 1/16)$, the proportions predicted by Mendel’s theory.

Suppose we observe $(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$. We find $(e_1, e_2, e_3, e_4) = (312.75, 104.25, 104.25, 34.75)$, $2 \log \Lambda = 0.618$ and $\sum \frac{(o_i - e_i)^2}{e_i} = 0.604$.

Here $|\Theta_0| = 0$ and $|\Theta_1| = 4 - 1 = 3$, so we refer our test statistics to $\chi^2_3$.

Since $\chi^2_3(0.05) = 7.815$ we see that neither value is significant at 5% level, so there is no evidence against Mendel’s theory.

In fact the $p$-value is approximately $P(\chi^2_3 > 0.6) \approx 0.96$. □

NB So in fact could be considered as a suspiciously good fit.
Example 9.3

In a genetics problem, each individual has one of three possible genotypes, with probabilities $p_1, p_2, p_3$. Suppose that we wish to test $H_0 : p_i = p_i(\theta)$ for $i = 1, 2, 3$, where $p_1(\theta) = \theta^2$, $p_2(\theta) = 2\theta(1 - \theta)$, and $p_3(\theta) = (1 - \theta)^2$, for some $\theta \in (0, 1)$.

We observe $N_i = n_i$, $i = 1, 2, 3$ ($\sum N_i = n$).

Under $H_0$, the mle $\hat{\theta}$ is found by maximising

$$\sum n_i \log p_i(\theta) = 2n_1 \log \theta + n_2 \log(2\theta(1 - \theta)) + 2n_3 \log(1 - \theta).$$

We find that $\hat{\theta} = (2n_1 + n_2)/(2n)$ (check). Also $|\Theta_0| = 1$ and $|\Theta_1| = 2$.

Now substitute $p_i(\hat{\theta})$ into (2), or find the corresponding Pearson’s chi-squared statistic, and refer to $\chi^2_{1}$. □

Example 9.4

Consider a two-way contingency table with $r$ rows and $c$ columns.

- For $i = 1, \ldots, r$ and $j = 1, \ldots, c$ let $p_{ij}$ be the probability that an individual selected at random from the population under consideration is classified in row $i$ and column $j$ (ie in the $(i,j)$ cell of the table).
- Let $p_{i+} = \sum_j p_{ij} = \bar{y}(\text{in row } i)$, and $p_{+j} = \sum_i p_{ij} = \bar{x}(\text{in column } j)$.
- We must have $p_{i+} = \sum_j p_{ij} = 1$, ie $\sum_i p_{ij} = 1$.
- Suppose a random sample of $n$ individuals is taken, and let $n_{ij}$ be the number of these classified in the $(i,j)$ cell of the table.
- Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$, so $n_{++} = n$.
- We have

$$(N_{11}, N_{12}, \ldots, N_{1c}, N_{21}, \ldots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, p_{12}, \ldots, p_{1c}, p_{21}, \ldots, p_{rc})$$

We may be interested in testing the null hypothesis that the two classifications are independent, so test $H_0 : p_{ij} = p_{i+}p_{+j}$ for $i = 1, \ldots, r, j = 1, \ldots, c$ (with $\sum_i p_{i+} = 1 = \sum_j p_{+j}$, $p_{i+}p_{+j} \geq 0$), $H_1 : p_{ij}$'s unrestricted (but as usual need $p_{i+} = 1$, $p_{+j} \geq 0$).

- Under $H_0$, the mle’s are $\hat{p}_{ij} = n_{ij}/n$.
- Under $H_0$, using Lagrangian methods, the mle’s are $\hat{\theta}_{i+} = n_{i+}/n$ and $\hat{\theta}_{+j} = n_{+j}/n$.
- Write $o_{ij}$ for $n_{ij}$ and let $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n_{i+}n_{+j}/n$.
- Then

$$2\log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} \log \left( \frac{o_{ij}}{e_{ij}} \right) \approx \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

using the same approximating steps as for Pearson’s Chi-squared test.

- We have $|\Theta_0| = rc - 1$, because under $H_1$ the $p_{ij}$‘s sum to one.
- Further, $|\Theta_1| = (r - 1) + (c - 1)$, because $p_{i+}, \ldots, p_{+c}$ must satisfy $\sum_j p_{ij} = 1$ and $p_{+1}, \ldots, p_{+c}$ must satisfy $\sum_i p_{ij} = 1$.
- So $|\Theta_1| - |\Theta_0| = rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$. 

A table in which observations or individuals are classified according to two or more criteria is called a contingency table.

Example 9.4

500 people with recent car changes were asked about their previous and new cars.

<table>
<thead>
<tr>
<th>New car</th>
<th>Previous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large 56</td>
</tr>
<tr>
<td></td>
<td>Medium 52</td>
</tr>
<tr>
<td></td>
<td>Small 42</td>
</tr>
</tbody>
</table>

This is a two-way contingency table: each person is classified according to previous car size and new car size. □
Example 9.5  
In Example 9.4, suppose we wish to test $H_0$: the new and previous car sizes are independent.

We obtain:

<table>
<thead>
<tr>
<th>$o_{ij}$</th>
<th>Large</th>
<th>Medium</th>
<th>Small</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>56</td>
<td>52</td>
<td>42</td>
<td>150</td>
</tr>
<tr>
<td>Medium</td>
<td>50</td>
<td>83</td>
<td>67</td>
<td>200</td>
</tr>
<tr>
<td>Small</td>
<td>18</td>
<td>51</td>
<td>81</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>186</td>
<td>190</td>
<td>500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$e_{ij}$</th>
<th>Large</th>
<th>Medium</th>
<th>Small</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>37.2</td>
<td>55.8</td>
<td>57.0</td>
<td>150</td>
</tr>
<tr>
<td>Medium</td>
<td>49.6</td>
<td>74.4</td>
<td>76.0</td>
<td>200</td>
</tr>
<tr>
<td>Small</td>
<td>37.2</td>
<td>55.8</td>
<td>57.0</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>186</td>
<td>190</td>
<td>500</td>
</tr>
</tbody>
</table>

Note the margins are the same.

Then $\sum \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 36.20$, and $df = (3 - 1)(3 - 1) = 4$.

From tables, $\chi^2 (0.05) = 9.488$ and $\chi^2 (0.01) = 13.28$.

So our observed value of 36.20 is significant at the 1% level, i.e., there is strong evidence against $H_0$, so we conclude that the new and present car sizes are not independent.

It may be informative to look at the contributions of each cell to Pearson’s chi-squared:

<table>
<thead>
<tr>
<th></th>
<th>New car</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large</td>
<td>Medium</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Previous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>9.50</td>
<td>0.26</td>
<td>3.95</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>0.00</td>
<td>0.99</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>9.91</td>
<td>0.41</td>
<td>10.11</td>
<td></td>
</tr>
</tbody>
</table>

It seems that more owners of large cars than expected under $H_0$ bought another large car, and more owners of small cars than expected under $H_0$ bought another small car.

Fewer than expected changed from a small to a large car. □