

Math 181B Worksheet Week 4

Pearson’s Chi-square Test

\[(Y_1, \ldots, Y_K) \sim \text{Multinomial}(n, p_1, \ldots, p_K) \Rightarrow \sum_{k=0}^{K} \frac{(Y_k - np_k)^2}{np_k} \sim \chi^2_{K-1}\]

Examples:

1. **Independent Representation of Multinomial Distribution** Let \(X_1, X_2, \ldots, X_n\) independently drawn from discrete distribution \(\mathbb{P}(X_i = k) = p_k, k = 1, \ldots, m\) and \(p_1 + p_2 + \cdots + p_m = 1\). Define \(Y_k = \sum_{i=1}^{n} I(X_i = k), k = 1, \ldots, m\). Show \((Y_1, Y_2, \ldots, Y_m) \sim \text{Multinomial}(n, p_1, \ldots, p_m)\).

   **Solution:**
   Calculate the joint p.m.f.
   \[
   \mathbb{P}(Y_1 = y_1, \ldots, Y_m = y_m) = \mathbb{P}(X_1 = 1, \ldots, X_{y_1} = 1, X_{y_1+1} = 2, \ldots, X_n = m) \cdot \text{# permutation in X with same Y} \\
   = \mathbb{P}(X_1 = 1) \cdots \mathbb{P}(X_{y_1} = 1) \mathbb{P}(X_{y_1+1} = 2) \cdots \mathbb{P}(X_n = m) \cdot \binom{n}{Y_1, \ldots, Y_m} \\
   = \binom{n}{Y_1, \ldots, Y_m} p_1^{Y_1} \cdots p_m^{Y_m}
   \]

2. **Discrete Goodness of Fit** Test the independence of \(X\) and \(Y\) in the two-way contingency table

<table>
<thead>
<tr>
<th>Variable</th>
<th>(Y=1)</th>
<th>(Y=0)</th>
<th>Marginal of (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X=1)</td>
<td>16</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>(X=0)</td>
<td>398</td>
<td>225</td>
<td>623</td>
</tr>
</tbody>
</table>

Model the problem as \((16, 398, 3, 225) \sim \text{Multinomial}(642, pq, (1-p)q, p(1-q), (1-p)(1-q))\) under null.

(a) Compute the MLE for \(p\) and \(q\).

(b) Do Pearson’s goodness of test using \(\hat{p}\) and \(\hat{q}\).

(c) Derive another test based on the likelihood ratio test.

**Solution:**
(a) Compute the MLE for \(p\) and \(q\).
Likelihood:

\[ L(p, q) = \binom{642}{16, 398, 3, 225} p^{19}(1-p)^{623} q^{414}(1-q)^{228} \]

Similar to the binomial MLE, \( \hat{p} = 19/642 = 0.03 \) and \( \hat{q} = 0.64 \)

(b) Do Pearson’s goodness of test using \( \hat{p} \) and \( \hat{q} \). The estimated expected counts are

\[ E_1 = 642\hat{p}\hat{q} = 12.25, \quad E_2 = 642(1 - \hat{p})\hat{q} = 401.75, \]
\[ E_3 = 642\hat{p}(1 - \hat{q}) = 6.75, \quad E_4 = 642(1 - \hat{p})(1 - \hat{q}) = 221.25, \]
\[ X^2 = 2.4979 < 3.8, \text{ so } H_0 \text{ accepted.} \]

(c) The unconstrained MLE for \((p_1, p_2, p_3, p_4)\) is given by sample proportion

\[ \hat{p}_1 = 12/642, \quad \hat{p}_2 = 398/642, \quad \hat{p}_3 = 3/642, \quad \hat{p}_4 = 225/642 \]

The LR is

\[ \Lambda = \left( \frac{642}{16, 398, 3, 225} \right) \frac{\hat{p}^{19}(1 - \hat{p})^{623}\hat{q}^{414}(1 - \hat{q})^{228}}{\hat{p}_1^{16}\hat{p}_2^{398}\hat{p}_3^3\hat{p}_4^{225}} \]

The null parameter space is of dimension 2. The full parameter space is of dimension 3. Use Wilk’s Theorem, \(-2 \log LR \chi^2_1\). Reject \( H_0 \) if \(-2 \log LR \geq \chi^2_{1,95} = 3.8\).

3. **Continuous Goodness of Fit**

Given the independent samples \( X_1, \ldots, X_{100} \), we are interested in testing the following hypothesis

\[ H_0 : X_1, \ldots, X_{100} \sim Uniform[0, \theta] \quad \text{versus} \quad H_1 : \text{Otherwise} \]

We observe 15 samples in \([0, 1]\), 20 samples in \((1, 2]\), 30 samples in \((2, 3]\), 25 samples in \((3, 10]\) and 10 samples in \((10, +\infty)\).

(a) Suppose \( X_{\text{max}} = 40 \). Test the goodness-of-fit.

(b) Test the goodness-of-fit without \( X_{\text{max}} \).

**Solution:**

Under null hypothesis, the density of \( X_i \’s \) is

\[ f(x) = I(x \in [0, \theta]) / \theta \]
The parameters for the multinomial bins \((O_1, O_2, O_3, O_4, O_5)\) are

\[
p_1(\theta) = \mathbb{P}(X \in [0, 1)) = \int_0^1 f(x)dx = \theta^{-1} \min\{1, \theta\}
\]

\[
p_2(\theta) = \mathbb{P}(X \in [1, 2)) = \int_1^2 f(x)dx = \theta^{-1} I(\theta \geq 1) \min\{1, \theta - 1\}
\]

\[
p_3(\theta) = \mathbb{P}(X \in [2, 3)) = \int_2^3 f(x)dx = \theta^{-1} I(\theta \geq 2) \min\{1, \theta - 2\}
\]

\[
p_4(\theta) = \mathbb{P}(X \in [3, 10)) = \int_3^{10} f(x)dx = \theta^{-1} I(\theta \geq 3) \min\{7, \theta - 3\}
\]

\[
p_5(\theta) = \mathbb{P}(X \in [10, +\infty)) = \int_{10}^{\infty} f(x)dx = \theta^{-1} I(\theta \geq 10)(\theta - 10)
\]

(a) The MLE for Uniform\([0, \theta]\) is \(\hat{\theta} = X_{\text{max}} = 40\). Plug into \(p_i(\theta)\). The expected counts are therefore \(E_i = np_i(\hat{\theta})\). Null parameter space is of dimension 1. Full parameter space is of dimension 4. Use Pearson’s goodness-of-fit test with degree of freedom 3.

(b) If information about \(X_i\)’s is not available, there is another way to find the MLE for \(\theta\) through the likelihood of \(Y_i\)’s. Under null, the likelihood of \(Y_i\)’s is also a function of \(\theta\).

\[
L(\theta) = \binom{n}{Y_1, \ldots, Y_5} p_1(\theta)^{Y_1} \cdots p_5(\theta)^{Y_5}
\]

Use regular MLE process, finding root of derivative with respect to \(\theta\) of log-likelihood, get \(\hat{\theta} = \arg \max_{\theta} L(\theta)\) Then, Plug into \(p_i(\theta)\). The expected counts are therefore \(E_i = np_i(\hat{\theta})\). Null parameter space is of dimension 1. Full parameter space is of dimension 4. Use Pearson’s goodness-of-fit test with degree of freedom 3.