TESTS OF NON-PARAMETRIC HYPOTHESES: GENERAL IDEAS

Good-bye to hypotheses about parameters! We turn now to tests of hypotheses about the distributions of random variables. Let $F(x)$ denote the CDF of a random variable $X$.

SIMPLE NULL HYPOTHESES. We will consider first hypotheses of the following form:

$$H_0 : F(x) = F_0(x)$$
$$H_1 : F(x) \neq F_0(x).$$

In other words, we want to test the hypotheses that the CDF of $X$ is a specified function $F_0(x)$ which is suspected to be the true CDF of $X$ (of course, if $X$ is of discrete type we may consider the PMF $f_0(x)$, and if $X$ is of continuous type its PDF $f_0(x)$ may be considered). There are many tests of such non-parametric hypotheses.

As always, we start with a random sample

$$X_1, X_2, \ldots, X_n$$

from $X$ of size $n$. We have to construct a test statistic $D = D(X_1, X_2, \ldots, X_n)$ such that

a) its distribution under $H_0$ or some approximation to this distribution is known;

b) $d = D(x_1, \ldots, x_n)$ has to be small if the true CDF is $F_0(x)$ and large if it is not $F_0(x)$. Or, in other words, $d$ has to show how good the observed sample fits the null hypothesis.

These properties suggest that it’s reasonable to choose the critical region of the form:

$$C = \{d : d \geq d_{tr}\}$$

where the threshold $d_{tr}$ is chosen in such a way that, for the chosen significance level of the test $\alpha$, we have $P(D \in C; H_0) = \alpha$ (or $P(D \in C) \approx \alpha$ if the distribution of $D$ is known only approximately).

COMPOSITE NULL HYPOTHESES. We will consider now hypotheses of the following form:

$$H_0 : F(x) = F_0(x; \theta_1, \theta_2, \ldots, \theta_d)$$
$$H_1 : F(x) \neq F_0(x; \theta_1, \theta_2, \ldots, \theta_d).$$

where $\theta_1, \theta_2, \ldots, \theta_d$ are unknown parameters.

Now $H_0$ is a composite hypothesis because it does not specify the distribution (unknown parameters!). A method to test such hypotheses will be considered below.
**PEARSON’S CHI-SQUARE GOODNESS-OF-FIT TEST**

**Simple null hypothesis.** The test statistic is constructed in the following way. We divide the range of the r.v. $X$ into $k$ semi-intervals:

$$A_1 = (a_0, a_1], A_2 = (a_1, a_3], \ldots, A_k = (a_{k-1}, a_k].$$

Let $Y_i$ be the number of those random variables $X_1, \ldots, X_n$ that are located in the semi-interval $A_i$. Under $H_0$ we can calculate the probabilities $p_{i0} = F_0(a_i) - F_0(a_{i-1})$ that $X \in A_i$. It is clear that under $H_0$ the r.v. $Y_i \sim b(n, p_{i0})$. The expected number of observations in $A_i$ (under $H_0$) is

$$e_{i0} \overset{\text{def}}{=} E(Y_i) = np_{i0}.$$ 

The test statistic is

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - e_{i0})^2}{e_{i0}}.$$ 

We can expect that, under $H_0$, $Y_i$ are not far from $e_{i0}$, so $Q_{k-1}$ possesses property a). It enjoys also property b) as the following theorem shows.

**Theorem.** If $n$ is large enough then the distribution of $Q_{k-1}$ is close to $\chi^2(k-1)$, the $\chi^2$-distribution with $k-1$ degrees of freedom.

The approximation is fair if $e_{i0} \geq 5$ for all $i = 1, \ldots, k$.

**Proof** for the case $k = 2$ (two semi-intervals). In this case

$$Y_1 + Y_2 = n, p_{10} + p_{20} = 1, e_{10} + e_{20} = np_{10} + np_{20} = n(p_{10} + p_{20}) = n,$$

$$(Y_1 - e_{10}) + (Y_2 - e_{20}) = (Y_1 + Y_2) - (e_{10} + e_{20}) = 0$$

and

$$Y_1 - e_{10} = -(Y_2 - e_{20}).$$

Hence

$$Q_1 = \frac{(Y_1 - e_{10})^2}{e_{10}} + \frac{(Y_2 - e_{20})^2}{e_{20}} = \frac{(Y_1 - e_{10})^2}{e_{10}} + \frac{(Y_1 - e_{10})^2}{e_{20}} =$$

$$\frac{(Y_1 - np_{10})^2}{np_{10}} + \frac{(Y_1 - np_{10})^2}{np_{10}(1-p_{10})} = \frac{(Y_1 - np_{10})^2}{np_{10}(1-p_{10})} \overset{\text{def}}{=} Z^2(n)$$

where $Z(n) = \frac{Y_1 - np_{10}}{\sqrt{np_{10}(1-p_{10})}}$. If $n$ is large then, by the CLT, $Z(n)$ is distributed approximately as $Z \sim N(0, 1)$ so $Z^2(n)$ is distributed approximately as $\chi^2$. It remains to recall that for i.i.d. random variables $Z_1, \ldots, Z_r \sim N(0, 1)$ the r.v. $Z_1^2 + \ldots + Z_r^2$ has the distribution $\chi^2(r)$: in our case $r = 1$ and $Z^2 \sim \chi^2(1)$!
We see that we "loose" one degree of freedom because of the (unique!) linear relation between the random variables $Y_i - e_{i0}, i = 1, ..., k$: since $\sum_{i=1}^{k} Y_i = n$ and $\sum_{i=1}^{k} e_{i0} = n$

$$\sum_{i=1}^{k} (Y_i - e_{i0}) = 0.$$ 

From the $\chi^2$ table we can find $\chi^2_{\alpha}(k - 1)$ such that for the r.v. $W \sim \chi^2(k - 1)$ we have $P(W \geq \chi^2_{\alpha}(k - 1)) = \alpha$. The Theorem shows that for large $n$ we have:

$$P\{Q_{k-1} \geq \chi^2_{\alpha}(k - 1)\} \approx \alpha.$$ 

Let $x_1, x_2, ..., x_n$ be the observed sample, $y_1, ..., y_k$ the observed frequencies. The value of the statistic $Q_{k-1}$ is

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - e_{i0})^2}{e_{i0}}.$$ 

The critical region with the approximate significance level $\alpha$ is

$$C_\alpha = \{q_{k-1} : q_{k-1} \geq \chi^2_{\alpha}(k - 1)\}$$

is the CR with approximate significance level $\alpha$. 


Composite null hypothesis. The hypothetic CDF is

\[ F_0(x; \theta_1, \theta_2, \ldots \theta_d) \]

(depends on \( r \) unknown parameters). Now we have to estimate these unknown parameters. We may use the Maximum Likelihood Estimators \( \hat{\Theta}_1, \hat{\Theta}_2, \ldots \hat{\Theta}_d \) and consider the estimated CDF

\[ \hat{F}_0(x) = F_0(x; \hat{\Theta}_1, \hat{\Theta}_2, \ldots \hat{\Theta}_d). \]

Instead of the probabilities \( p_{i0} = F_0(a_i) - F_0(a_{i-1}) \) we have to consider the estimated probabilities

\[ \hat{p}_{i0} = \hat{F}_0(a_i) - \hat{F}_0(a_{i-1}) \]

and instead of the expected frequencies \( e_{i0} = np_{i0} \) we have the estimated expected frequencies

\[ \hat{e}_{i0} = n\hat{p}_{i0}. \]

The test statistic is

\[ \hat{Q}_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - \hat{e}_{i0})^2}{\hat{e}_{i0}}. \]

**Theorem 2.** When \( n \) is large the distribution of \( \hat{Q}_{k-1} \) is approximately \( \chi^2(k - 1 - d) \). The critical region with the approximate significance level \( \alpha \) is

\[ C_\alpha = \{ \hat{q}_{k-1} : \hat{q}_{k-1} \geq \chi^2_\alpha(k - 1 - d) \}. \]
Let $X$ denote the repair time in days required for a certain component in an airplane. We wish to test whether a Poisson model with a mean of three days appears to be a reasonable model for this variable. The repair times for 40 components were recorded, with the results shown in Table 13.11. In some cases the component could be repaired immediately on-site, which is interpreted as zero days.

Under $H_0: X \sim \text{POI}(3)$, we have $f(x) = e^{-3}3^x/x!$, and the cell probabilities are given by $p_{10} = P[X = 0] = f(0) = e^{-3} = 0.050$, $p_{10} = f(1) = e^{-3}3 = 0.149$, $p_{10} = f(2) = 0.224$, and so on. The expected numbers are then $e_j = np_{10}$. The

<table>
<thead>
<tr>
<th>Repair Time (Days)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>&gt; 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed $(a_j)$</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Probabilities $(p_j)$</td>
<td>0.050</td>
<td>0.149</td>
<td>0.224</td>
<td>0.224</td>
<td>0.168</td>
<td>0.101</td>
<td>0.050</td>
<td>0.034</td>
</tr>
<tr>
<td>Expected $(e_j)$</td>
<td>2.00</td>
<td>5.96</td>
<td>8.96</td>
<td>8.96</td>
<td>8.72</td>
<td>4.04</td>
<td>2.00</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>7.96</td>
<td></td>
<td>8.96</td>
<td>8.96</td>
<td>8.72</td>
<td>4.04</td>
<td>2.00</td>
<td>1.36</td>
</tr>
</tbody>
</table>

right-hand tail cells are pooled to achieve an $e_j \geq 5$, and the first two cells also are pooled. This leaves $e = 5$ cells and

$$
\chi^2 = \frac{(4 - 7.96)^2}{7.96} + \frac{(7 - 8.96)^2}{8.96} + \cdots + \frac{(13 - 7.40)^2}{7.40} = 9.22 > 7.78
$$

$$
\chi^2 \sim \chi^2_{(4)}
$$

so we can reject $H_0$ at the $\alpha = 0.10$ significance level.
Example 2

Consider again the data given in Example 1 and suppose now that we wish to test \( H_0 : X \sim \text{POI}(\mu) \). The usual MLE of \( \mu \) is the average of the 40 repair times, which in this case is computed as

\[
\hat{\mu} = \frac{[0(1) + 1(3) + 2(7) + \cdots + 6(6)]}{40} = 3.65
\]

Under \( H_0 \) the estimated probabilities are now \( \hat{p}_j = e^{-3.65}(3.65)^j/j! \) and the estimated expected values, \( \hat{e}_j = np\hat{p}_j \), are computed using a Poisson distribution with \( \mu = 3.65 \). Retaining the same five cells used before, the results are as shown in Table 2.

<table>
<thead>
<tr>
<th>Repair Times (Days)</th>
<th>(0, 1)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>( \geqslant 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed (( o_j ))</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Probabilities (( \hat{p}_j ))</td>
<td>0.121</td>
<td>0.173</td>
<td>0.211</td>
<td>0.192</td>
<td>0.303</td>
</tr>
<tr>
<td>Expected (( \hat{e}_j ))</td>
<td>4.84</td>
<td>6.92</td>
<td>8.44</td>
<td>7.68</td>
<td>12.12</td>
</tr>
</tbody>
</table>

We have

\[
\chi^2 = \frac{(4 - 4.84)^2}{4.84} + \cdots + \frac{(13 - 12.12)^2}{12.12} = 1.62 < 6.25 = \chi^2_{0.05}(3)
\]

so a Poisson model appears to be quite reasonable for these data, although the Poisson model with \( \mu = 3 \) was found not to fit well. The number of degrees of freedom here is 3, because one parameter is estimated.