Chapter 03 Likelihood-Based Tests and Confidence Regions

Part 05

3.2.7 Score Statistic for Multinomial Data

Suppose a r.v. X following $Pr(X=x_i)=p_i$ with $\sum_{i=1}^{k} p_i = 1$.
Test $H_0: p_i = p_{i0}$ v.s. $H_0: p_i \neq p_{i0}$ for at least one i. The goodness-of-fit test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \text{with } E_i = n p_{i0}$$

This statistic is actually a Score statistic.

Proof. Suppose $(N_1, \cdots, N_k) \sim \text{Multinomial}(n; p_1, \cdots, p_k)$.
Let $p = (p_1, \cdots, p_{k-1})^T$, then

$$L(p \mid N_1, \cdots, N_p) = \frac{n!}{N_1! \cdots N_k!} p_1^{N_1} \cdots p_{k-1}^{N_{k-1}} (1 - p_1 - \cdots - p_{k-1})^{N_k}$$

$$l(p) = \log\left(\frac{n!}{N_1! \cdots N_k!}\right) + \sum_{i=1}^{k-1} N_i \log p_i + N_k \log(1 - p_1 - \cdots - p_{k-1})$$

$$\frac{\partial l(p)}{\partial p_j} = \frac{N_j}{p_j} - \frac{N_k}{1 - \sum_{i=1}^{k-1} p_i}, \quad j = 1, \cdots, k-1$$

$$S(p) = \left(\frac{N_1}{p_1} - \frac{N_k}{p_k}, \cdots, \frac{N_{k-1}}{p_{k-1}} - \frac{N_k}{p_k}\right)^T$$

Example 2.13 derived $I_T^{-1}(p_1, \cdots, p_{k-1}) = \frac{1}{n}[\mathrm{diag}(p) - pp^T]$

$$= \frac{1}{n}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} \\ -p_2p_1 & p_2(1-p_2) & & -p_2p_{k-1} \\ \vdots & & & \\ -p_{k-1}p_1 & \cdots & & p_{k-1}(1-p_{k-1}) \end{bmatrix}$$

$$T_S = S^T(\tilde{p})[I_T(\tilde{p})]^{-1}S(\tilde{p})$$

$$= \frac{1}{n}\left\{S^T(\tilde{p})\,\mathrm{diag}(p)\,S(\tilde{p}) - S^T(\tilde{p})pp^TS(\tilde{p})\right\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{k-1}\left(\frac{N_i}{\tilde{p}_i} - \frac{N_k}{\tilde{p}_k}\right)^2\tilde{p}_i - \left[\sum_{i=1}^{k-1}\left(\frac{N_i}{\tilde{p}_i} - \frac{N_k}{\tilde{p}_k}\right)\tilde{p}_i\right]^2\right\}$$

The upper summation limit is set to k rather than $k-1$ because the kth summands are identically 0.

$$= \frac{1}{n}\left\{\sum_{i=1}^{k}\left(\frac{N_i}{\tilde{p}_i} - \frac{N_k}{\tilde{p}_k}\right)^2\tilde{p}_i - \left[\sum_{i=1}^{k}\left(\frac{N_i}{\tilde{p}_i} - \frac{N_k}{\tilde{p}_k}\right)\tilde{p}_i\right]^2\right\}$$

Let $a_i = \frac{N_i}{\tilde{p}_i} - \frac{N_k}{\tilde{p}_k}$ $\qquad a_i\tilde{p}_i = N_i - \frac{N_k \cdot \tilde{p}_i}{\tilde{p}_k}$

$$= \frac{1}{n}\left\{\sum_{i=1}^{k}a_i^2\tilde{p}_i - \left[\sum_{i=1}^{k}a_i\tilde{p}_i\right]^2\right\} = \frac{1}{n}\sum_{i=1}^{k}\left[a_i - \sum_{i=1}^{k}a_i\tilde{p}_i\right]^2\tilde{p}_i$$

$$\sum_{i=1}^{k}\left(N_i - \frac{N_k}{\tilde{p}_k}\tilde{p}_i\right) = n - \frac{N_k}{\tilde{p}_k}\sum_{i=1}^{k}\tilde{p}_i = n - \frac{N_k}{\tilde{p}_k}$$

This allows to invoke the variance equality

$$\sum a_i^2p_i - (\sum a_ip_i)^2 = \sum(a_i - \sum a_jp_j)^2p_j$$

$$EX^2 - (EX)^2 = \mathrm{Var}(X)$$

$$= E[X - E(X)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left[\left(\frac{N_i}{\tilde{p}_i} - \cancel{\frac{N_k}{\tilde{p}_k}}\right) - \left(n - \cancel{\frac{N_k}{\tilde{p}_k}}\right)\right]^2 \tilde{p}_i$$

$$= \frac{1}{n}\sum_{i=1}^{k}\left[\frac{N_i}{\tilde{p}_i} - n\right]^2 \tilde{p}_i$$

$$= \sum_{i=1}^{k}\frac{1}{n}\left(\frac{N_i}{\tilde{p}_i} - \frac{n\tilde{p}_i}{\tilde{p}_i}\right)^2 \tilde{p}_i$$

$$= \sum_{i=1}^{k}\frac{(N_i - n\tilde{p}_i)^2}{n\tilde{p}_i}$$

Here we used the fact that if $P(X=a_i)=p_i$, $i=1,\cdots,k$, then
$E(X^2)-E^2(X)=Var(X)=E[(X-E(X))^2]$.

Note that under H_0, $\tilde{p}=p_0=(p_{10},\cdots,p_{k0})^T$.