Statistics 3858 : Likelihood Ratio for Multinomial Models

Suppose $\underline{X}$ is multinomial on M categories, that is $X \sim \text{Multinomial}(n, \underline{p})$, where $\underline{p} = (p_1, p_2, \ldots, p_M) \in \mathcal{A}$, and the parameter space is

$$\mathcal{A} = \{\underline{p} : p_j \geq 0, \quad \sum_{j=1}^{M} p_j = 1 \quad \}$$

The dimension of this parameter space is $M - 1$. It is a simplex of dimension $M - 1$.

The likelihood function is

$$L(\underline{p}) = c(n, X_1, \ldots, X_M) \prod_{j=1}^{M} p_j^{X_j}$$

where the data is $\underline{X} = (X_1, X_2, \ldots, X_M)$. Notice that $X_j \geq 0$ and $X_1 + X_2 + \ldots + X_M = n$ and

$$c(n, x_1, \ldots, x_M) = \binom{n}{x_1 \quad \ldots x_M} = \frac{n!}{x_1! x_2! \ldots x_M!}$$

is the multinomial coefficient. The MLE is easily found using the log-likelihood and Lagrange multipliers and is

$$\hat{\underline{p}} = \left(\frac{X_1}{n}, \ldots, \frac{X_M}{n} \right)$$

A special multinomial model in certain models is of the form

$$\underline{p} = \underline{p}(\theta) = (p_1(\theta), \ldots, p_M(\theta))$$

where the components are of a functional form of some other parameter. For example in the Hardy-Weinberg model with $M = 3$

$$\underline{p} = ((1-\theta)^2, 2\theta(1-\theta), \theta^2)$$

where $\theta \in \Theta = (0, 1)$. We can view this as a particular 1 dimensional subset or sub-manifold, say $\mathcal{A}_0$ of the $M - 1$ dimensional simplex $\mathcal{A}$ that is the general parameter space for multinomials on M categoriess.

This section constructs the generalized likelihood ratio (GLR) statistic for

$$H_0 : \underline{p} \in \mathcal{A}_0$$

versus

$$H_A : \underline{p} \in \mathcal{A}_A = \mathcal{A} \setminus \mathcal{A}_0$$

We often write the alternative as $H_A : \underline{p}$ is not in $\mathcal{A}_0$ or simply refer to is as the general alternative (in this context).

Let $\hat{\theta}$ be the MLE of θ. Then the MLE of $\underline{p}(\theta)$ is given by $\underline{p}(\hat{\theta})$. Since $\mathcal{A}_A \cup \mathcal{A}_0 = \mathcal{A}$, the denominator for the GLR is the likelihood evaluated at the general or unrestricted MLE of $\underline{p}$. Thus the GLR is

$$\Lambda(X) = \frac{\prod_{j=1}^{M} p_j(\hat{\theta})^{X_j}}{\prod_{j=1}^{M} \hat{p}_j^{X_j}}$$

The rejection region is of the form

$$\Lambda(x) < c$$

where c is determined by

$$\alpha = P_0(\Lambda(X) < c)$$

By Theorem 9.4A c is obtained as $c_1 = -2\log(c)$ where

$$\alpha = P_0\left(-2\log(\Lambda(X)) > c_1 = -2\log(c) \right)$$

The constant c_1 is approximately the upper $1-\alpha$ quantile of a $\chi^2_{(d)}$ distribution where the degrees of freedom is $d = M - 1 - \dim(\mathcal{A}_0)$.

For the Hardy-Weinberg model this is

$$M - 1 - \dim(\mathcal{A}_0) = 3 - 1 - 1 = 1$$

A size $\alpha = .05$ test will have $c_1 = (1.96)^2 = 3.84$ and $c = e^{-3.84/2} = e^{-1.92} = 0.146$.

Consider the function $g : \mathcal{R}^+ \mapsto \mathcal{R}$ given by

$$g(y) = y\log(y/y_0)$$

where y_0 is a given number. The first two derivatives are

$$\begin{aligned} g'(y) &= \log(y/y_0) + y\frac{1}{y} \\ &= \log(y/y_0) + 1 \\ g''(y) &= \frac{1}{y} \end{aligned}$$

and

$$\begin{aligned} g(y_0) &= y_0\log(y_0/y_0) = 0 \\ g'(y_0) &= \log(y_0/y_0) + 1 = 1 \\ g''(y_0) &= \frac{1}{y_0} \end{aligned}$$

When we take the negative 1 times the log of the GLR $\Lambda(X)$ we see, after gathering up some common terms, that it contains

$$-\log(p_j(\hat{\theta})) + \log(\hat{p}_j) = \log\left(\frac{\hat{p}_j}{p_j(\hat{\theta})}\right) .$$

Aside : We are interested in a negative number times the log GLR, since the GLR ≤ 1, and this will result in the negative log being positive. If one were to define the GLR with the ratio reversed this would not be the case, but by convention GLR is defined as this ratio. Some text books unfortunately do not follow this convention.

Thus for a given j, taking $y_0 = p_j(\hat{\theta})$ and $y = \hat{p}_j$

$$\begin{aligned} g(\hat{p}_j) &\approx g(p_j(\hat{\theta})) + g'(p_j(\hat{\theta}))\left(\hat{p}_j - p_j(\hat{\theta})\right) + \frac{1}{2}g''(p_j(\hat{\theta}))\left(\hat{p}_j - p_j(\hat{\theta})\right)^2 \\ &= \left(\hat{p}_j - p_j(\hat{\theta})\right) + \frac{\left(\hat{p}_j - p_j(\hat{\theta})\right)^2}{2p_j(\hat{\theta})} \end{aligned}$$

Below consider g_j to be the function g above with $y_0 = p_j(\hat{\theta})$.

It then follows that

$$\begin{aligned} -2\log\left(\Lambda(X)\right) &= -2\sum_{j=1}^{M} X_j \log(p_j(\hat{\theta})/\hat{p}_j) \\ &= 2n\sum_{j=1}^{M} \frac{X_j}{n} \log(\hat{p}_j/p_j(\hat{\theta})) \\ &= 2n\sum_{j=1}^{M} \hat{p}_j \log(\hat{p}_j/p_j(\hat{\theta})) \\ &= 2n\sum_{j=1}^{M} g_j(\hat{p}_j) \\ &\approx n\sum_{j=1}^{M} \left\{ 2\left(\hat{p}_j - p_j(\hat{\theta})\right) + \frac{\left(\hat{p}_j - p_j(\hat{\theta})\right)^2}{p_j(\hat{\theta})} \right\} \\ &= 2n\sum_{j=1}^{M} \left\{\hat{p}_j - p_j(\hat{\theta})\right\} + \sum_{j=1}^{M} \frac{n^2\left(\hat{p}_j - p_j(\hat{\theta})\right)^2}{np_j(\hat{\theta})} \\ &= 2n\left(1-1\right) + \sum_{j=1}^{M} \frac{\left(n\hat{p}_j - np_j(\hat{\theta})\right)^2}{np_j(\hat{\theta})} \end{aligned}$$

$$= \sum_{j=1}^{M} \frac{\left(n\hat{p}_j - np_j(\hat{\theta})\right)^2}{np_j(\hat{\theta})}$$

This last expression is often written as $n\hat{p}_j = X_j = O_j$ where O_j is the observed counts in the j-th category, and $np_j(\hat{\theta}) = \hat{E}_j$ (or sometimes E_j) as the expected counts for the best fit for the statistical model with parameter θ, that is the restricted multinomial model that corresponds to the null hypothesis. When doing this we obtain

$$\begin{aligned} \chi^2 &= \sum_{j=1}^{M} \frac{\left(n\hat{p}_j - np_j(\hat{\theta})\right)^2}{np_j(\hat{\theta})} \\ &= \sum_{j=1}^{M} \frac{\left(O_j - \hat{E}_j\right)^2}{\hat{E}_j} \end{aligned}$$

This last formula is called the Pearson's chi-squared statistic.

Thus in this multinomial setting the Pearson's chi-squared statistic is equivalent to the generalized likelihood ratio test. It also has a very natural property of comparing the observed and fitted model. We reject if the GLR Λ is very small, or equivalently when $-2\log(\Lambda) = \chi^2$ is very large. This of course is a measure which is large if O_j is far from the expected counts for the best fitted model in the null hypothesis.

In order to assess when the observed value of χ^2 is large, we need to compute for a given α the critical value so that

$$\alpha = P_0(\chi^2 > c)$$

By Theorem 9.4A (Rice) when the statistical model is one according to the null hypothesis, the sampling distribution of χ^2 converges to $\chi^2_{(d)}$ where the degrees of freedom is $d = M - 1 - \dim(\mathcal{A}_0)$.

In the Hardy-Weinberg example, $M = 3$ and the null hypothesis is that $\underline{p} \in \mathcal{A}_0$ in the notation at the beginning this handout, thus the degrees of freedom is $3 - 1 - 1 = 1$. The size $\alpha = .05$ critical value to determine the rejection region is thus $c = 3.84$. The decision rule is thus to reject if

$$\Lambda(X) \leq e^{-3.84/2} = 0.146$$

or equivalently if

$$\chi^2 = \sum_{j=1}^{3} \frac{\left(O_j - \hat{E}_j\right)^2}{\hat{E}_j} > 3.84$$

Alternatively we could observe the corresponding statistic and calculate the p-value. If we observe χ^2_{Obs} then the p-value is

$$\text{p-value} = P(Y > \chi^2_{\text{Obs}}) \ .$$

Remark : This is of course the value of the critical constant c so that χ^2_{Obs} falls on the boundary of the rejection region.