COMPOSITE HYPOTHESES FOR MULTINOMIAL DISTRIBUTIONS

1. Definitions and an example

Let $X_1, ..., X_k$ be observed with a multinomial $(n, \pi_1, ..., \pi_k)$ distribution where $\pi_1, ..., \pi_k$ are unknown with $\pi_j \geq 0$ and $\sum_{j=1}^k \pi_j = 1$. The dimension of this full multinomial model H_1 with k categories is $d = k-1$. Suppose we have an m-dimensional composite hypothesis H_0 under which $\pi_j = p_j(\theta)$ for a parameter θ in an m-dimensional set M_0 with $m < k-1$. An example for $k = 3$ and $m = 1$ is the Hardy–Weinberg equilibrium model with $\theta = p$, $0 < p < 1$, $p_1(p) = p^2$, $p_2(p) = 2p(1-p)$, and $p_3(p) = (1-p)^2$. Here M_0 is the interval $[0, 1]$.

It would be inconvenient if any $p_j(\theta)$ could be 0, especially if θ is the true value of the parameter. If we then estimated θ by some θ' so that $p_j(\theta')$ is close to 0, then $np_j(\theta')$ might be less than 5 and one could not apply chi-squared tests under the usual rule. So, it will be assumed that $p_j(\theta) > 0$ for all $j = 1, ..., k$ and all $\theta \in M_0$.

2. The Wilks test

Let X be the data vector $(X_1, ..., X_k)$. It is given as grouped data. If we can test H_0 by the Wilks likelihood ratio test, then we must be able to find the maximum likelihood estimate $\widehat{\theta}$ of $\theta \in M_0$ based on X. In many cases $\widehat{\theta}$ is hard to compute, but we're assuming it can be computed in this case. Then we can also evaluate the χ^2 statistic

$$\widehat{X}_n^2 = \sum_{j=1}^k \frac{(X_j - np_j(\widehat{\theta}))^2}{np_j(\widehat{\theta})}. \tag{1}$$

If H_0 is true, then the distribution of $\widehat{X}_n^2$ converges as $n \to \infty$ to that of $\chi^2(k-1-m)$, as shown in "χ^2 tests for composite hypotheses, asymptotic distributions," chisqcmp.pdf. It follows that $\widehat{X}_n^2$ remains bounded in probability as $n \to \infty$, i.e. for any $\varepsilon > 0$ there is an $M < +\infty$ such that $\Pr(\widehat{X}_n^2 > M) < \varepsilon$ for all n. This implies that for any

Date: 18.650, Oct. 19, 2015.

sequence $a_n \to +\infty$, $\Pr(\widehat{X}_n^2 > a_n) \to 0$. Thus for each j

$$(2)\quad \Pr(|X_j - np_j(\widehat{\theta})| > \sqrt{n}a_n = \Pr\left(\left|\frac{X_j}{n} - p_j(\widehat{\theta})\right| > n^{-1/2}a_n\right) \to 0.$$

For example, let $a_n = n^\delta$ for $0 < \delta < 1/2$. Also, by Wilks's theorem, the Wilks statistic $W = -2\log(\Lambda)$, where $\Lambda = \Lambda(\widehat{\theta})$ from its definition, has the same limit distribution as $n \to \infty$.

3. Asymptotic equality of two statistics if H_0 holds

Under H_0, not only do $\widehat{X}^2$ and W have the same limiting distribution, but their difference approaches 0:

Theorem 1. *If H_0 holds then $\widehat{X}^2 - W \to 0$ in probability as $n \to \infty$.*

Proof. For H_0 to hold means there exists a true $\theta_0 \in M_0$, so that $\{X_j\}_{j=1}^k$ have a multinomial $(n, \{p_j(\theta_0)\}_{j=1}^k)$ distribution. From the assumptions, $p_j(\theta_0) > 0$ for all j. Since $k \geq 2$ it also follows that $p_j(\theta_0) < 1$. Each X_j has a binomial $(n, p_j(\theta_0))$ distribution, with mean $np_j(\theta_0)$ and variance $np_j(\theta_0)(1 - p_j(\theta_0))$. It follows from Chebyshev's inequality that

$$(3)\qquad \Pr(X_j \leq np_j(\theta_0)/2) \leq \frac{np_j(\theta_0)}{(np_j(\theta_0)/2)^2} = \frac{4}{np_j(\theta_0)} \to 0$$

as $n \to \infty$. It also follows from Chebyshev's inequality that

$$(4)\quad \Pr(|X_j - np_j(\theta_0)| > \sqrt{n}a_n) = \Pr\left(\left|\frac{X_j}{n} - p_j(\theta_0)\right| > n^{-1/2}a_n\right) \to 0.$$

Combining with (2) gives that for each $j = 1, ..., k$, $p_j(\widehat{\theta}) \to p_j(\theta_0)$ in probability, i.e. for every $\varepsilon > 0$, $\Pr(|p_j(\widehat{\theta}) - p_j(\theta_0)| > \varepsilon) \to 0$ as $n \to \infty$.

On H_1 we have the likelihood function

$$(5)\qquad L_1(X, \pi) = \binom{n}{X_1, ..., X_k} \prod_{j=1}^k \pi_j^{X_j}.$$

On H_0 the likelihood is

$$(6)\qquad L(X, \theta) = \binom{n}{X_1, ..., X_k} \prod_{j=1}^k p_j(\theta)^{X_j}.$$

For the Wilks test, the MLEs (maximum likelihood estimates) of π_j under the full multinomial model H_1 are simply $\hat{\pi}_j = X_j/n$ for $j =$

$1, ..., k$, since under H_1, X_j has a binomial (n, π_j) distribution for each j. The likelihood (5) maximized over H_1 is

$$\binom{n}{X_1, ..., X_k} \prod_{j=1}^{k} \left(\frac{X_j}{n}\right)^{X_j}. \tag{7}$$

The ratio $\Lambda(\theta)$ of the likelihood at $\theta \in M_0$ given in (6) to the likelihood (7), which is maximized over H_1, is

$$\Lambda(\theta) = \prod_{j=1}^{k} (np_j(\theta)/X_j)^{X_j}. \tag{8}$$

In forming the Wilks statistic, $W = -2 \ln \Lambda$, the likelihood ratio Λ used is the maximum of $\Lambda(\theta)$ over $\theta \in M_0$, which is $\Lambda(\widehat{\theta})$.

Maximizing $\Lambda(\theta)$ is equivalent to maximizing its (natural) logarithm, which at $\widehat{\theta}$ equals

$$\begin{aligned} \log(\Lambda(\widehat{\theta})) &= \sum_{j=1}^{k} X_j \log(np_j(\widehat{\theta})/X_j) \\ &= \sum_{j=1}^{k} X_j \log\left(\frac{X_j - (X_j - np_j(\widehat{\theta}))}{X_j}\right) \\ &= \sum_{j=1}^{k} X_j \log\left(1 - \frac{X_j - np_j(\widehat{\theta})}{X_j}\right). \end{aligned} \tag{9}$$

To relate this to X^2 statistics, an idea is to use the Taylor series $\log(1-u) = -u - u^2/2 - u^3/3 - \cdots$, valid for $|u| < 1$, and moreover, to use the series when $|u|$ is small enough so that the first two terms $-u - u^2/2$ give a sufficient approximation. In our case, for

$$u_j = \frac{X_j - np_j(\widehat{\theta})}{X_j} = 1 - \frac{np_j(\widehat{\theta})}{X_j}, \tag{10}$$

we'd like each $X_j|u_j|^3$ to be small, for which it suffices that $n|u_j|^3$ are small for all j. By (2) with $a_n = n^{0.1}$, the probability that $|X_j - np_j(\widehat{\theta})| \le n^{3/5}$ converges to 1. Then by (3),

$$u_j = |X_j - np_j(\widehat{\theta})|/X_j \le n^{3/5}/(np_j(\theta_0))$$

and $n|u_j^3| \le n^{-1/5}/p_j(\theta_0) \to 0$ as $n \to \infty$, as desired. Further, we get that

$$np_j(\widehat{\theta})/X_j \to 1 \tag{11}$$

in probability as $n \to \infty$. Then

$$\log(\Lambda(\widehat{\theta})) \doteq \sum_{j=1}^{k} X_j \left(\frac{-X_j + np_j(\widehat{\theta})}{X_j} \right) - \frac{X_j}{2} \left(\frac{X_j - np_j(\widehat{\theta})}{X_j} \right)^2$$

$$= \sum_{j=1}^{k} -X_j + np_j(\widehat{\theta}) - \frac{1}{2} \frac{(X_j - np_j(\widehat{\theta}))^2}{X_j}. \tag{12}$$

For the first order terms, we have

$$\sum_{j=1}^{k} -X_j + np_j(\widehat{\theta}) = -n + n = 0, \tag{13}$$

because $\sum_{j=1}^{k} p_j(\widehat{\theta}) = 1$. [Note however that if p_j are not exact but only computed to some fixed number of decimal places, say four, then their sum might equal for example 1.0001 or .9999 and then the expression in (13 would have absolute value $0.0001n \to \infty$ as $n \to \infty$.] Returning to the situation where (13) holds exactly (if necessary by some adjustment to rounded numbers), we get for $W(\widehat{\theta}) = -2 \log \Lambda(\widehat{\theta})$ by (11) and (12)

$$W(\widehat{\theta}) - \widehat{X}^2 \to 0 \tag{14}$$

in probability, proving the theorem. $\square$

It also follows from (14) that choosing $\theta' \in M_0$ to maximize the likelihood is approximately the same as the "minimum χ^2 estimate," choosing $\theta' \in M_0$ to minimize the right side of (1).