Likelihood Ratio tests

For general composite hypotheses optimality theory is not usually successful in producing an optimal test. instead we look for heuristics to guide our choices. The simplest approach is to consider the likelihood ratio

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

and choose values of $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$ which are reasonable estimates of θ assuming respectively the alternative or null hypothesis is true. The simplest method is to make each θ_i a maximum likelihood estimate, but maximized only over Θ_i.

Example 1: Consider a sample of size n from the $N(\mu, 1)$ model and test $\mu \le 0$ against $\mu > 0$. (Remember the uniformly most powerful test.) The log-likelihood is

$$-n(\bar{X} - \mu)^2/2$$

If $\bar{X} > 0$ then the global maximum in Θ_1 at $\bar{X}$. If $\bar{X} \le 0$ the global maximum in Θ_1 is at 0. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$ and 0 if $\bar{X} \le 0$. Similarly, $\hat{\mu}_0$ is $\bar{X}$ if $\bar{X} \le 0$ and 0 if $\bar{X} > 0$. Hence

$$\frac{f_{\hat{\theta}_1}(X)}{f_{\hat{\theta}_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

which simplifies to

$$\exp\{n\bar{X}|\bar{X}|/2\}$$

This is a monotone increasing function of $\bar{X}$ so the rejection region will be of the form $\bar{X} > K$. To get level α we must reject if $n^{1/2}\bar{X} > z_\alpha$. Notice that a simpler statistic with the same rejection region is the *log-likelihood ratio*

$$\lambda \equiv 2\log\left(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)}\right) = n\bar{X}|\bar{X}|$$

Example 2: In the $N(\mu, 1)$ problem suppose we make the null $\mu = 0$. Then the value of $\hat{\mu}_0$ is simply 0 while the maximum of the log-likelihood over the alternative $\mu \ne 0$ occurs at $\bar{X}$. This gives

$$\lambda = n\bar{X}^2$$

which has a χ^2_1 distribution. This test leads to the rejection region $\lambda > (z_{\alpha/2})^2$ which is the usual two sided t-test.

Example 3: For the $N(\mu, \sigma^2)$ problem testing $\mu = 0$ against $\mu \neq 0$ we must find two estimates of μ, σ^2. The maximum of the likelihood over the alternative occurs at the global mle $\bar{X}, \hat{\sigma}^2$. We find

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -n/2 - n\log(\hat{\sigma})$$

First we maximize ℓ over the null hypothesis. Recall that

$$\ell(\mu, \sigma) = -\frac{1}{2\sigma^2}\sum (X_i - \mu)^2 - n\log(\sigma)$$

On the null $\mu = 0$ so find we $\hat{\sigma}_0$ by maximizing

$$\ell(0, \sigma) = -\frac{1}{2\sigma^2}\sum X_i^2 - n\log(\sigma)$$

This leads to

$$\hat{\sigma}_0^2 = \sum X_i^2/n$$

and

$$\ell(0, \hat{\sigma}_0) = -n/2 - n\log(\hat{\sigma}_0)$$

This gives

$$\lambda = -n\log(\hat{\sigma}^2/\hat{\sigma}_0^2)$$

Since

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2 + n\bar{X}^2}$$

we can write

$$\lambda = n\log(1 + t^2/(n-1))$$

where

$$t = \frac{n^{1/2}\bar{X}}{s}$$

is the usual t statistic. Thus the likelihood ratio test rejects for large values of $|t|$ — the usual test. Notice that if n is large we have

$$\lambda \approx n[1 + t^2/(n-1) + O(n^{-2})] \approx t^2 \,.$$

Since the t statistic is approximately standard normal if n is large we see that

$$\lambda = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

has nearly a χ_1^2 distribution.

This is a general phenomenon when the null hypothesis being tested is of the form $\phi = 0$. Here is the general theory. Suppose that the vector of $p+q$ parameters θ can be partitioned into $\theta = (\phi, \gamma)$ with ϕ a vector of p parameters and γ a vector of q parameters. To test $\phi = \phi_0$ we find two mles of θ. First the global mle $\hat\theta = (\hat\phi, \hat\gamma)$ maximizes the likelihood over $\Theta_1 = \{\theta : \phi \neq \phi_0\}$ (because typically the probability that $\hat\phi$ is exactly ϕ_0 is 0).

Now we maximize the likelihood over the null hypothesis, that is we find $\hat\theta_0 = (\phi_0, \hat\gamma_0)$ to maximize

$$\ell(\phi_0, \gamma)$$

The log-likelihood ratio statistic is

$$2[\ell(\hat\theta) - \ell(\hat\theta_0)]$$

Now suppose that the true value of θ is ϕ_0, γ_0 (so that the null hypothesis is true). The score function is a vector of length $p+q$ and can be partitioned as $U = (U_\phi, U_\gamma)$. The Fisher information matrix can be partitioned as

$$\begin{bmatrix} \mathcal{I}_{\phi\phi} & \mathcal{I}_{\phi\gamma} \\ \mathcal{I}_{\gamma\phi} & \mathcal{I}_{\gamma\gamma} \end{bmatrix} .$$

According to our large sample theory for the mle we have

$$\hat\theta \approx \theta + \mathcal{I}^{-1} U$$

and

$$\hat\gamma_0 \approx \gamma_0 + \mathcal{I}_{\gamma\gamma}^{-1} U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat\theta)$ and $\ell(\hat\theta_0)$ around θ_0 you get

$$\ell(\hat\theta) \approx \ell(\theta_0) + U^t \mathcal{I}^{-1} U + \frac{1}{2} U^t \mathcal{I}^{-1} V(\theta) \mathcal{I}^{-1} U$$

where V is the second derivative matrix of ℓ. Remember that $V \approx -\mathcal{I}$ and you get

$$2[\ell(\hat\theta) - \ell(\theta_0)] \approx U^t \mathcal{I}^{-1} U .$$

A similar expansion for $\hat\theta_0$ gives

$$2[\ell(\hat\theta_0) - \ell(\theta_0)] \approx U_\gamma^t \mathcal{I}_{\gamma\gamma}^{-1} U_\gamma .$$

If you subtract these you find that

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

can be written in the approximate form

$$U^t M U$$

for a suitable matrix M. It is now possible to use the general theory of the distribution of $X^t M X$ where X is $MVN(0, \Sigma)$ to demonstrate that

Theorem 1 *The log-likelihood ratio statistic*

$$\lambda = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

has, under the null hypothesis, approximately a χ^2_p *distribution.*

Aside:

Theorem 2 *Suppose* $X \sim MVN(0, \Sigma)$ *with* Σ *non-singular and* M *is a symmetric matrix. If* $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ *then* $X^t M X$ *has a* χ^2_ν *distribution with df* $\nu = trace(M\Sigma)$.

Proof: We have $X = AZ$ where $AA^t = \Sigma$ and Z is standard multivariate normal. So $X^t M X = Z^t A^t M A Z$. Let $Q = A^t M A$. Since $AA^t = \Sigma$ condition in the theorem is

$$AQQA^t = AQA^t$$

Since Σ is non-singular so is A. Multiply by A^{-1} on the left and by $(A^t)^{-1}$ on the right to get the identity $QQ = Q$.

The matrix Q is symmetric so $Q = P\Lambda P^t$ where Λ is a diagonal matrix containing the eigenvalues of Q and P is orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. So rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ) .$$

Notice that $W = P^t Z$ is $MVN(0, P^t P = I)$; i.e. W is standard multivariate normal. Now

$$W^t \Lambda W = \sum \lambda_i W_i^2$$

We have established that the general distribution of any quadratic form $X^t M X$ is a linear combination of χ^2 variables. Now go back to the condition

$QQ = Q$. If λ is an eigenvalue of Q and $v \neq 0$ is a corresponding eigenvector then $QQv = Q(\lambda v) = \lambda Qv = \lambda^2 v$ but also $QQv = Qv = \lambda v$. Thus $\lambda(1 - \lambda)v = 0$. It follows that either $\lambda = 0$ or $\lambda = 1$. This means that the weights in the linear combination are all 1 or 0 and that $X^t M X$ has a χ^2 distribution with degrees of freedom, ν, equal to the number of λ_i which are equal to 1. This is the same as the sum of the λ_i so

$$\nu = trace(\Lambda)$$

But

$$\begin{aligned} trace(M\Sigma) &= trace(MAA^t) \\ &= trace(A^t MA) \\ &= trace(Q) \\ &= trace(P\Lambda P^t) \\ &= trace(\Lambda P^t P) \\ &= trace(\Lambda) \end{aligned}$$

In the application Σ is $\mathcal{I}$ the Fisher information and $M = \mathcal{I}^{-1} - J$ where

$$J = \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{I}_{\gamma\gamma}^{-1} \end{bmatrix}$$

It is easy to check that $M\Sigma$ becomes

$$\begin{bmatrix} I & 0 \\ -\mathcal{I}_{\gamma\phi}\mathcal{I}_{\phi\phi} & 0 \end{bmatrix}$$

where I is a $p \times p$ identity matrix. It follows that $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ and $trace(M\Sigma) = p$.