Locally most powerful and other large sample tests

BS2 Statistical Inference, Lecture 12
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 19, 2004

Locally Most Powerful Tests

Recall that for testing a simple hypothesis with $\Theta_0 = \{\theta_0\}$ vs. a simple alternative $\Theta_A = \{\theta_1\}$, the LRT with critical region

$$C = \left\{ x \mid \frac{L(\theta_1; x)}{L(\theta_0; x)} > K \right\}$$

is optimal of its size, i.e. for fixed size $\alpha = P_{\theta_0}(C)$, it has maximal power $\phi(\theta_1) = P_{\theta_1}(C)$ under the alternative.

It is difficult to find a test which is optimal against all alternatives, i.e. *uniformly most powerful.*

As almost any test will detect large deviations from the hypothesis we could attempt to find a test which is most powerful for *small deviations* from the hypothesis.

Suppose now that $\theta_1 = \theta_0 + \delta$ is close to θ_0, i.e. δ is small. Taylor's theorem now yields

$$\log \frac{L(\theta_1; x)}{L(\theta_0; x)} = S(x; \theta_0)\delta + R(x, \delta) \approx S(x; \theta_0)\delta,$$

where S is the score statistic. Thus the test with critical region

$$C = \{x \mid S(x; \theta_0) > K\}$$

is optimal for alternatives close to θ_0. We also say that this test is *Locally Most Powerful* in the direction δ.

Note that if $\theta_1 < \theta_0$, the sign changes and the LMP test against local alternatives in the opposite direction takes the form

$$C = \{x \mid S(x; \theta_0) < K\}.$$

Example

Consider $X = (X_1, \ldots, X_n)$ independent and identically distributed as $\mathcal{N}(\mu, \mu^2)$ with $\mu > 0$ unknown.

If we let $S = \sum X_i$ and $SS = \sum X_i^2$ the score statistic is

$$S(x; \mu) = -\frac{n}{\mu} + \frac{SS}{\mu^3} - \frac{S}{\mu^2}.$$

Thus the critical region for the LMP test of the hypothesis $\mu = 1$ against local alternatives with $\mu > 1$ has the form

$$S(x; 1) = -n + SS - S > K$$

or, equivalently

$$SS - S > K'.$$

The score test

To determine the critical value K for the LMP we need in principle to know the distribution of the score statistic.

However, this is generally intractable so we rely on the fact that the score statistic has an approximate normal distribution with the Fisher information as its variance.

Thus, we can construct an asymptotic test by calculating the critical region as

$$S(x;\theta_0) > \sqrt{ni(\theta_0)}z_{1-\alpha}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile in the standard normal distribution $\mathcal{N}(0,1)$ and $i(\theta_0)$ is the Fisher information in a single observation.

This large sample test is known as a *score test*.

For the example just considered we have $i(\mu) = 3n/\mu^2$ so the critical region of the score test for $\mu = 1$ against $\mu > 1$ becomes

$$S(x; 1) = -n + SS - S > \sqrt{3n} z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile in the standard normal distribution $\mathcal{N}(0, 1)$.

This test is particularly simple to calculate in this and many other cases where the Fisher information has a simple form.

The score test can easily be modified to a test for a two-sided alternative by alluding to the fact that since

$$S(X; \theta) \overset{\text{a}}{\sim} \mathcal{N}\{0, ni(\theta)\},$$

Slutsky's theorem implies that

$$\frac{\{S(X;\theta)\}^2}{ni(\theta_0)} \xrightarrow{D} \chi^2(1),$$

where $\chi^2(1)$ is the χ^2-distribution with 1 degree of freedom. So the *two-sided score test* has critical region

$$\{S(x;\theta_0)\}^2 > ni(\theta_0)\chi^2(1)_{1-\alpha}.$$

In the example considered, this takes the form

$$(SS - S - n)^2 > 3n\chi^2(1)_{1-\alpha}$$

where $\chi^2(1)_{1-\alpha}$ is the $(1-\alpha)$-quantile in the χ^2-distribution.

The χ^2 test

Another alternative to construct a test statistic for the hypothesis is to directly use the asymptotic distribution of the maximum likelihood estimate. Since it holds generally that

$$\hat{\theta} \overset{\mathrm{a}}{\sim} \mathcal{N}\left\{\theta, \frac{1}{ni(\theta)}\right\}.$$

Again alluding to Slutsky's theorem, a test statistic for the hypothesis $\theta = \theta_0$ against the two-sided alternative $\hat{\theta} \neq \theta_0$ can be constructed by using the critical region

$$X^2 = ni(\theta_0)(\hat{\theta} - \theta_0)^2 > \chi^2(1)_{1-\alpha}.$$

This test is known as the χ^2-*test*.

Returning to the example

In the example considered, the MLE $\hat{\mu}$ was given as

$$\hat{\mu} = \frac{-S + \sqrt{S^2 + 4nSS}}{2n}$$

so the χ^2 test for $\mu = 1$ has critical region

$$X^2 = 3n \left(\frac{-S + \sqrt{S^2 + 4nSS}}{2n} - 1 \right)^2 > \chi^2(1)_{1-\alpha}.$$

Note that this appears to be much more complex than the score test. However, as one would often have calculated $\hat{\mu}$ anyway, this complication may not play a practical role.

Wald's test

There are many alternatives for these large sample tests.

Wald's test uses the estimated Fisher information $i(\hat{\theta})$ instead of $i(\theta_0)$ in the χ^2-test yielding the critical region

$$W = ni(\hat{\theta})(\hat{\theta} - \theta_0)^2 > \chi^2(1)_{1-\alpha}.$$

This is convenient when $\hat{\theta}$ has been calculated by the method of scoring, as then $i(\hat{\theta})$ has already been calculated during the iterative step, e.g. in the `glm` routine.

In the example, the Wald test statistic becomes

$$W = \frac{3n}{\hat{\mu}^2}\left(\frac{-S + \sqrt{S^2 + 4nSS}}{2n} - 1\right)^2 > \chi^2(1)_{1-\alpha}.$$

The danger in using Wald's test is that it may have small power when θ very different from θ_0, as then $i(\hat{\theta})$ might become small, as is the case in the example above.

Comparing the test statistics in the example, we find

$$W = X^2/\hat{\mu}^2$$

so for large μ, X^2 is better at detecting deviations from H_0:

$$\lim_{\hat{\mu}\to\infty} W(\hat{\mu}) = 3n \text{ whereas } \lim_{\hat{\mu}\to\infty} X^2(\hat{\mu}) = \infty.$$

On the other hand

$$\lim_{\hat{\mu}\to 0} W(\hat{\mu}) = \infty \text{ whereas } \lim_{\hat{\mu}\to 0} X^2(\hat{\mu}) = 3n,$$

so the Wald test is more powerful for small alternatives in this particular example.

Observed information tests

A finalvariant of the Wald test, we may of course use the observed rather than the expected information to yield the critical region

$$\tilde{W} = j(\hat{\theta})(\hat{\theta} - \theta_0)^2 > \chi^2(1)_{1-\alpha},$$

Where

$$j(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} \log f(X; \hat{\theta})$$

which is positive when $\hat{\theta}$ is the MLE.

As the Wald statistic is easy to calculate when the method of scoring is used is easy to calculate, this test statistic is convenient when the Newton–Raphson method is used.

In the example we get

$$\tilde{W} = \left(\frac{n}{\hat{\mu}^2} + \frac{SS}{\hat{\mu}^4}\right)\left(\frac{-S + \sqrt{S^2 + 4nSS}}{2n} - 1\right)^2.$$

But this variant of the Wald statistic has the same problem with large deviations from the hypothesis as the version based on expected information.

The maximized likelihood ratio

Instead of trying to obtain high power close to the hypothesis, as in a locally most powerful test, it seems sensible to *maximize power at the most likely value of the parameter* θ i.e. at the MLE $\hat{\theta}$.

This leads to the *(maximized) likelihood ratio test* MLRT, commonly just known as the likelihood ratio test LRT, with critical region of the form

$$\Lambda = -2 \log \frac{L(X; \theta_0)}{L(X; \hat{\theta})} > K.$$

We have taken logarithms and multiplied by 2.

As we shall see in the next lecture, the Wald and χ^2-tests can just be seen as Taylor-approximations to this test.

In our example we get

$$\Lambda = 2n \log \hat{\mu} - \frac{SS}{\hat{\mu}^2} + \frac{2S}{\hat{\mu}} + SS - 2S$$

which again yields an important alternative to the other large sample tests.