Probability and Statistics

Fall 2000

Lect. 18: Large Sample Tests CB 8.4

In the previous lecture we discussed in the context of a random sample from a normal distribution with unknown mean μ and known variance σ^2 testing the hypothesis

$$H_0 : \ \mu = \mu_0,$$

against the alternative

$$H_1 : \ \mu = \mu_1 \neq \mu_0.$$

The optimal test at the 5% level corresponded to the critical region

$$C_X = \left\{ x_1, \ldots, x_N \middle| N \cdot (\bar{x} - \mu_0)^2/\sigma^2 > 3.84 \right\},$$

which uses the Chi–squared distribution for the square of a standard normal random variable. Alternatively we could have calculated the test statistic, in this case $N \cdot (\bar{x} - \mu_0)^2/\sigma^2$ which under the null hypothesis has a known distribution, in this case a chi–square one distribution.

We cannot do this so easily for other distributions, In many cases the exact distribution of the sufficient statistic is not easily calculable, and finding exact critical regions is essentially impossible. To get around this we use, just like we did in estimation, large sample approximations. Recall that for the maximum likelihood estimator

$$\sqrt{N} \cdot (\hat{\theta}_{ml} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}).$$

In large samples we can therefore approximate the distribution of $\hat{\theta}_{ml}$ by

$$\hat{\theta}_{ml} \approx \mathcal{N}(\theta_0, \hat{\mathcal{I}}(\hat{\theta}_{ml})^{-1}),$$

for some estimate of the information matrix. Suppose this approximation is exact. In that case we could test the null hypothesis

$$H_0 : \ \theta = \theta_0,$$

against the alternative

$$H_1 : \ \theta \neq \theta_0,$$

by an unbiased most powerful test using the critical region

$$C_X = \left\{ x \middle| N \cdot (\hat{\theta}_{ml} - \theta_0)^2 \cdot \hat{\mathcal{I}}(\hat{\theta}_{ml}) > C_\alpha \right\},$$

where C_α is the critical value of a Chi-square distribution with one degree of freedom, satisfying $Pr(Y > C_\alpha) = \alpha$ if $Y \sim \mathcal{X}^2(1)$, e.g., $C_{0.1} = 2.718$, and $C_{0.05} = 3.84$. This type of test is known as a Wald test. The test statistic is

$$WALD = N \cdot (\hat{\theta}_{ml} - \theta_0)^2 \cdot \hat{\mathcal{I}}(\hat{\theta}_{ml}).$$

Again we reject the null hypothesis if the test statistic is larger than the critical value coming from the chi–squared distribution with one degree of freedom.

There are two other tests for the same null and alternative hypothesis that give very similar results in large samples, and in fact are large sample equivalent. Indirectly all three tests are based on the likelihood function. If the null hypothesis is correct, the log likelihood function should be close to the expected log likelihood function which in that case is maximized at θ_0. Therefore its maximum should be close to the θ_0 (the Wald test), the maximizing value should be close to the value at θ_0 (the likelihood ratio test), and the derivative at θ_0 should be close to zero (the Lagrange multiplier test).

The likelihood ratio test is based on the maximum value of the log likelihood function under the null and under the alternative hypothesis. Define

$$\lambda = \frac{\max_{\theta \in \Theta_0} f_X(x;\theta)}{\max_{\theta \in \Theta_0^c} f_X(x;\theta)}.$$

In our case the only value consistent with the null hypothesis is θ_0, and the value that maximizes the likelihood function under the alternative is the maximum likelihood estimator (except if the maximum likelihood estimator is exactly equal to θ_0, but that is very unlikely). Therefore:

$$\lambda = \frac{f_X(x;\theta_0)}{f_X(x;\hat{\theta}_{mle})}.$$

In addition we have N independent and identically distributed random variables so,

$$\lambda = \frac{\mathcal{L}(x_1,\ldots,x_N;\theta_0)}{\mathcal{L}(x_1,\ldots,x_N;\hat{\theta}_{mle})}.$$

In large samples,

$$LR = 2\cdot\ln\lambda \xrightarrow{d} \mathcal{X}^2(1),$$

a chi–squared distribution with one degree of freedom.

To see the connection with the Wald test, expand $L(\theta_0)$ around $\hat{\theta}_{ml}$:

$$L(\theta_0) = L(\hat{\theta}_{ml}) + \frac{\partial L}{\partial\theta}(\hat{\theta}_{ml})\cdot(\theta_0-\hat{\theta}_{ml}) + \frac{1}{2}\frac{\partial^2 L}{\partial\theta^2}(\tilde{\theta})\cdot(\theta_0-\hat{\theta}_{ml})^2$$

$$= L(\hat{\theta}_{ml}) + \frac{1}{2}\frac{\partial^2 L}{\partial\theta^2}(\tilde{\theta})\cdot(\theta_0-\hat{\theta}_{ml})^2,$$

for some $\tilde{\theta}$ between $\hat{\theta}_{ml}$ and θ_0. Note that $\frac{\partial L}{\partial\theta}(\hat{\theta}_{ml}) = 0$. Substituting this into λ gives

$$2\cdot\ln\lambda = 2\cdot\Big(L(\hat{\theta}_{ml}) - L(\theta_0)\Big) = \frac{\partial^2 L}{\partial\theta^2}(\tilde{\theta})\cdot(\theta_0-\hat{\theta}_{ml})^2.$$

Note that

$$\frac{1}{N}\frac{\partial^2 L}{\partial\theta^2}(\tilde{\theta}) \xrightarrow{p} -\mathcal{I}(\theta_0).$$

Hence

$$2\cdot\ln\lambda \longrightarrow N\cdot(\theta_0-\hat{\theta}_{ml})^2\cdot\mathcal{I}(\theta_0) \approx WALD.$$

The Lagrange multiplier or score test, the third test in the trinity of tests works of the fact that the expectation of the score function is zero: if the null hypothesis is true then

$$E\Big[\frac{\partial \ln f_X}{\partial \theta}(X;\theta_0)\Big] = 0.$$

The variance of the score is

$$\mathcal{I}(\theta_0) = E\Big[\frac{\partial \ln f_X}{\partial \theta}(X;\theta_0) \cdot \frac{\partial \ln f_X}{\partial \theta'}(X;\theta_0)\Big].$$

Hence, using a central limit theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(X_i;\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)).$$

Given that we do not know the exact information matrix, we use an estimate. Typically we use the test statistic

$$LM = \frac{1}{N}\Big(\sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i;\theta_0)\Big)^2 \Big/ \hat{\mathcal{I}}(\theta_0).$$

To see that this is again similar to the Wald test expand the sum of the derivative of the log of the density around $\hat{\theta}_{ml}$:

$$\begin{aligned}
\sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i;\theta_0) &= \sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i;\hat{\theta}_0) + \sum_{i=1}^{N} \frac{\partial^2 \ln f_X}{\partial \theta^2}(x_i;\tilde{\theta}) \cdot (\theta_0 - \hat{\theta}_{ml}) \\
&= \sum_{i=1}^{N} \frac{\partial^2 \ln f_X}{\partial \theta^2}(x_i;\tilde{\theta}) \cdot (\theta_0 - \hat{\theta}_{ml}) \\
&\approx -N \cdot \mathcal{I}(\theta_0) \cdot (\theta_0 - \hat{\theta}_{ml}),
\end{aligned}$$

so that

$$LM = \frac{1}{N}\Big(\sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i;\theta_0)\Big)^2 \Big/ \hat{\mathcal{I}}(\theta_0) \approx \frac{1}{N}\Big(-N \cdot \mathcal{I}(\theta_0) \cdot (\theta_0 - \hat{\theta}_{ml})\Big)^2 \Big/ \hat{\mathcal{I}}(\theta_0)$$

$$\approx N \cdot \mathcal{I}(\theta_0) \cdot (\hat{\theta}_{ml} - \theta_0)^2 \approx WALD.$$

Example

Let us consider an example. Suppose $X_1, \ldots, X_N$ are independent with a Poisson distribution with parameter θ,

$$f_X(x;\theta) = \frac{\theta^x \exp(-\theta)}{x!}.$$

We wish to test the null hypothesis

$$H_0 : \ \theta = 6,$$

against the alternative hypothesis

$$H_1 : \ \theta \neq 6,$$

at the 10% level. It is given that $N = 100$ and $\sum_{i=1}^{N} x_i = 500$.

First consider the maximum likelihood estimator. The log likelihood function is

$$L(\theta) = \sum_{i=1}^{N} x_i \cdot \ln\theta - \theta - \ln x_i!.$$

The score function is

$$\frac{\partial \ln f_X}{\partial \theta}(x|\theta) = \frac{x}{\theta} - 1.$$

The maximum likelihood estimate is

$$\hat{\theta}_{ml} = \bar{x} = 5.$$

The second derivative of the log of the pmf is

$$\frac{\partial^2 \ln f_X}{\partial \theta^2}(x|\theta) = -\frac{x}{\theta^2}.$$

The single-observation information matrix is therefore

$$\mathcal{I}(\theta) = V(X)/\theta^2 = \theta/\theta^2 = 1/\theta,$$

using the square of the first derivatives, or

$$\mathcal{I}(\theta) = E[X]/\theta^2 = \theta/\theta^2 = 1/\theta,$$

using the second derivatives. The information matrix is estimated by evaluating it at the maximum likelihood estimate,

$$\hat{\mathcal{I}}_1 = 1/\hat{\theta}_{ml} = 1/5,$$

or at the value of the parameter under the null:

$$\hat{\mathcal{I}}_2 = 1/\theta_0 = 1/6.$$

Now consider the likelihood ratio test. The value of the log likelihood function under the null is

$$L(\theta_0) = \sum_{i=1}^{N} x_i \cdot \ln \theta_0 - \theta_0 - \ln x_i!$$

$$= 500 \cdot \ln 6 - 100 \cdot 6 - \sum_{i=1}^{N} \ln x_i!.$$

The value of the log likelihood function at the maximum likelihood estimator is

$$L(\hat{\theta}_{ml}) = \sum_{i=1}^{N} x_i \cdot \ln \hat{\theta}_{ml} - \hat{\theta}_{ml} - \ln x_i!$$

$$= 500 \cdot \ln 5 - 100 \cdot 5 - \sum_{i=1}^{N} \ln x_i!.$$

Twice the difference is

$$LR = 2 \cdot (L(\hat{\theta}_{ml}) - L(\theta_0)) = 2 \cdot (500 \ln 5 - 500 - 500 \ln 6 - 600) \approx 17.6$$

Next, consider the Wald test:

$$WALD = N \cdot (\theta_{ml} - \theta_0)^2 \cdot \hat{\mathcal{I}}(\hat{\theta}_{ml}) = 100 \cdot (6-5)^2 \cdot 1/6 = 16.7.$$

Finally, consider the Lagrange multiplier test. The sum of the derivatives is

$$\sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i; \theta_0) = \sum_{i=1}^{N} \frac{x}{\theta_0} - 1 = 500/6 - 100 = -16.67.$$

The Lagrange multiplier test statistic is therefore

$$LM = \frac{1}{N}\Big(\sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i; \theta_0)\Big)^2 \Big/ \hat{\mathcal{I}}(\theta_0) = \frac{1}{100} \cdot (-16.67)^2/(1/5) \approx 13.9.$$

In all cases the test statistic exceeds the critical value for a chi–square one distribution at the 10% level, which is 2.728. Typically the test statistics are close enough that the result (rejection or acceptance) does not depend on the actual test choosen, although this does happen occasionally. □

There is in the end no compelling reason to choose one of the tests rather than another. All have their advantages. The Wald test is easy to calculate once the maximum likelihood estimator and its large sample variance have been calculated. The likelihood ratio test has the advantage of not requiring an arbitrary choice for the information matrix. The Lagrange multiplier has the advantage of not requiring estimation of the model under the alternative hypothesis.

The large sample equivalence of the test can be seen most easily by considering the case where the likelihood function is quadratic with known curvature, that is, in the normal case with known variance. In that case all three tests are exactly identical.