Summary of Selected Likelihood Results

Mai Zhou

Likelihood functions and statistical inferences based on the likelihood functions is *the major cornerstone* of statistical science. Any statistical curriculum (especially methodology courses at graduate level) should cover this topic.

See Chapter 10 of the Statistics Lecture Notes by Geyer [www.stat.umn.edu/geyer/old/5102/n2.pdf] for a systematic coverage. See the review paper of Reid (2010) on WIREs

[www.utstat.utoronto.ca/reid/research/likelihood-final.pdf] for a quick review aimed at non-statisticians. The textbook of Casella and Berger, have some discussion of MLE in about p.472 etc.

This method has become more important with the wide use of computers, since some steps of the method are automatic and can easily be carried out by computer.

1 Likelihood Function and MLE

• Suppose $X_1, X_2, \cdots, X_n$ are independent and $X_i \sim f(x|\theta)$; where $\theta \in \Theta$ is the parameter. For any $\theta \in \Theta$ $f(x|\theta)$ is a density function. The log likelihood function is

$$l_n(\theta) = \log Lik(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta) \ .$$

If X_i have different densities for each i: $X_i \sim f_i(x_i|\theta)$; then the above definition should be $\sum \log f_i(X_i|\theta)$.

The θ value that maximize the $l_n(\theta)$ is the MLE, we denote it as $\hat\theta_{MLE}$.

The MLE has the so called 'invariance property': i.e. $\widehat{g(\theta)}_{MLE} = g(\hat\theta_{MLE})$.

• (asymptotic distribution of MLE) Under mild regularity conditions (something like two/three derivative of the density f exist and is continuous, or derivative pass under integral sign, and information is positive) we have

$$\hat\theta_{MLE} - \theta_0 \approx N\left(0,\ \sigma^2 = \frac{1}{I(\theta_0)}\right)$$

and

$$I^{1/2}(\hat\theta_{MLE})(\hat\theta_{MLE} - \theta_0) \approx N(0,1)$$

where I can be replaced by J.

Based on this, an approximate 95% confidence interval for θ_0 is

$$\hat{\theta}_{MLE} \pm 1.96 \frac{1}{\sqrt{J(\hat{\theta}_{MLE})}}$$

This type intervals are called Wald confidence intervals. Wald confidence interval DO NOT have 'invariance property'.

2 Likelihood Ratio Test

There are other ways of constructing the confidence intervals. We focus on one that uses the equivalency of confidence interval and testing hypothesis.

- The likelihood ratio test for $H_0 : \theta = \theta^*$ vs. $H_A : \theta \neq \theta^*$ can be based on the

$$-2[l_n(\theta^*) - l_n(\hat{\theta}_{MLE})] = W(\theta^*) \ .$$

The rejection rule is to reject H_0 if $W(\theta^*)$ is too large. (use chi square table to find the critical value. For example if θ is one dimensional, then threshold = 3.84 gives the type I error of 5 percent).

This also leads to confidence intervals if we invert the test. Since the likelihood ratio tests were also called Wilks tests, we call the resulting confidence interval the Wilks interval.

Formally, the confidence interval is

$$\{\theta^* | W(\theta^*) < 3.84\} \ .$$

Please see the document inside the R package emplik (version 1.03 and later) for examples of finding the Wilks confidence interval. [you may find the PDF file inside the doc folder, see the GitHub archive]

The Wilks confidence interval DO have 'invariance property'.

The Wilks confidence interval is always inside the Θ.

3 A Connection

- An interesting relation: for large n and θ^* near θ_0, we have

$$-2[l_n(\theta^*) - l_n(\hat{\theta}_{MLE})] \approx \frac{(\theta^* - \hat{\theta}_{MLE})^2}{J(\hat{\theta}_{MLE})} \ .$$

Notice the left hand side do not explicitly involve $I(\theta)$ or J. *This can be a real advantage*, especially when the Information is hard to calculate or invert.

4 Expected or Observed information?

Based on $X_1, X_2, \cdots, X_n$,

Expected information:

$$I_n(\theta) = -E\left\{\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log Lik(X_i|\theta)\right\}$$

It is non-random but is a function of θ.

Observed information:

$$J_n = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log Lik(X_i|\theta)|_{\theta=\hat{\theta}_{MLE}}$$

It is random but do not involve θ explicitly. It can be computed from the sample. In particular computer programs often do this for us.

Efron and Hinkley (1978) Biometrika and Cao, H. (2013) JHP University PhD Dissertation

try to compare Expected information vs Observed information, with opposite conclusions. However, observed information is easier to calculate with computer.

We also point out, if we use Wilks tests and Wilks confidence intervals, we avoid the information altogether.

5 A picture

A picture is worth a thousand words. [sorry I do not have a picture in PDF]

- Error $\approx N(0, 1/I(\theta_0))$.

- Stepdown $\approx \chi^2/2$

If the likelihood is normal distribution with the parameter of mean, then the two approximates above are exact, and they are also producing exactly the same confidence intervals.

For other distributions, they are approximate and valid only when n is large, and θ near θ_0.