STA701 - ADVANCED STATISTICAL INFERENCE I

Notes on large sample theory for Maximum Likelihood Estimator and other unbiased estimating functions

Author: Prof. Mai Zhou
First version of Notes taken by: Yuhua Su (Fall 2000)

Department of Statistics

The University of Kentucky

Revised, Spring, 2003; Spring 2005

1 Definitions

1.1 **Definition** Suppose we have observation $x_1, \cdots, x_n$ which has density $f_\theta(x_1, \cdots, x_n)$. The *Likelihood function* is a function of parameter θ: $f_\theta(x_1, \cdots, x_n)$, for $\theta \in \Theta$ (the Parameter Space).

It is the probability density function evaluated at the observed sample. Or the infinitesimal probability of the observed value, if we include the dx_i's in the likelihood function. The interpretation of "probability of observe the sample" is nice, unite the discrete case and continuous case and also to the empirical likelihood case. Since dx_i are considered as constants, they are often dropped. But it often clerify things when several possible candidate of likelihood function are competing with each other.

In the i.i.d. case $L(\theta) = f_\theta(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$.

1.2 **Definition** The MLE ($\hat\theta_{MLE}$) is the value of θ where the log likelihood function attains its maximum. It depends on $x_1, \cdots, x_n$, i.e. $\hat\theta_{MLE} = \hat\theta_{MLE}(x_1, \cdots, x_n)$ and we assume it is a measurable function.

Sometime, $\hat\theta_{MLE}$ is also defined as the solution of the equation (score equation)

$$\frac{\partial}{\partial\theta} \log f_\theta(x_1, \cdots, x_n) = 0 \ .$$

Obviously, these two definitions are not always equivalent, but we shall use them interchangeably with comments. Our first goal is to show that $\hat\theta_{MLE}$ is consistent.

In many other cases, an estimator (not necessarily MLE) can also be defined by minimizing of a function, or the solution to an equation. Example include the least squares estimator, LASSO etc. When this function is convex but may not have continuous derivative, please see Pollard notes.

2 Consistency of MLE

Lemma *For $t > 0$ we have $t - 1 \geq \log(t)$, and the inequality is strict except when $t = 1$.*

Now replace t in the above Lemma by $\frac{g(X)}{f(X)}$, and take expectation E_f. ($X \sim f(x)$).

Recall that we showed, for any two densities $f(x)$ and $g(x)$

$$\int \log \frac{g(x)}{f(x)} f(x) dx \leq 0 \ .$$

The equality holds only when $f(x) = g(x)$, a.s. (dF). Another way of writing the above inequality is that

$$E \log \frac{g(X)}{f(X)} \leq 0 \text{ where } X \sim f(\cdot)$$

or

$$E_f \log g(X) \le E_f \log f(X) \ .$$

Now consider a family of distributions $f_\theta(x), \theta \in \Theta$ and we require the following condition (1) Whenever $\theta_0 \neq \theta'$, $f_{\theta_0}(x) \neq f_{\theta'}(x)$ a.s. f_{θ_0}, where θ_0 is such that we have observations $X_0, X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta_0}(x)$.

Thus, we have, for any fixed $\theta' \neq \theta_0$,

$$E_{\theta_0} \log \frac{f_{\theta'}(X)}{f_{\theta_0}(X)} < 0 \ .$$

Note that the above strict inequality is obtained by the condition (1). Let us denote

$$E_{\theta_0} \log \frac{f_{\theta'}(X)}{f_{\theta_0}(X)} = c \in \mathcal{R}^-,$$

(notice $c < 0$. The existence of the expectation is an assumption) which, in turn, implies

$$\log \prod \frac{f_{\theta'}(X_i)}{f_{\theta_0}(X_i)} = \sum_{i=1}^n \log \frac{f_{\theta'}(X_i)}{f_{\theta_0}(X_i)} \xrightarrow[\text{as } n \to \infty]{\text{in } \mathcal{P} \text{ or a.s.}} -\infty \ (\approx \lim n \cdot c),$$

or

$$E_{\theta_0} \log f_{\theta_0}(X) = E_{\theta_0} \log f_{\theta'}(X) - c$$

By SLLN, we have

$$\frac{\sum_{i=1}^n \log f_{\theta_0}(x_i)}{n} \xrightarrow[\text{as } n \to \infty]{\text{a.s.}} E_{\theta_0} \log f_{\theta_0}(X)$$

and

$$\frac{\sum_{i=1}^n \log f_{\theta'}(x_i)}{n} \xrightarrow[\text{as } n \to \infty]{\text{a.s.}} E_{\theta_0} \log f_{\theta'}(X),$$

therefore,

$$P(\sum_{i=1}^n \log f_{\theta_0}(x_i) > \sum_{i=1}^n \log f_{\theta'}(x_i) \text{ as } n \to \infty) = 1$$

Now, let us take $\theta' = \theta_0 + \delta$ and $\theta'' = \theta_0 - \delta$. Then, on a set with probability greater than $1 - 2\epsilon$, we have, for $n > N$,

$$\sum_{i=1}^n \log f_{\theta_0}(x_i) > \sum_{i=1}^n \log f_{\theta_0+\delta}(x_i)$$

and

$$\sum_{i=1}^{n} \log f_{\theta_0}(x_i) > \sum_{i=1}^{n} \log f_{\theta_0-\delta}(x_i)$$

this implies that as a function of θ, $\sum_{i=1}^{n} \log f_\theta(x_i)$ has a (at least local) maximum inside $(\theta_0-\delta, \theta_0+\delta)$. Or, taking the maximum to be $\hat{\theta}_{MLE}$, we have shown

$$P_{\theta_0}(|\hat{\theta}_{MLE}-\theta_0| < \delta) > 1-2\epsilon \text{ for } n > N.$$

Notice that δ, ϵ are arbitrary small positive numbers. This is weak consistency of MLE in i.i.d. case. We actually only proved a weaker result, i.e. there is a solution of $(\frac{\partial}{\partial\theta}\log f = 0)$ that is weakly consistent. Notice there may be multiple solutions.

Remark $\hat{\theta}_{MLE}$ is also strongly consistent, i.e. $\hat{\theta}_{MLE} \xrightarrow[\text{as } n\to\infty]{\text{a.s.}} \theta_0$, but we shall not prove that conclusion.

Now here is a second consistent proof that can be used for high dimensional θ. And it can also be used for other estimators that are defined by maximum or minimum of a random function $S_n(\theta)$.

The condition required in this theorem, is stronger: Uniform convergence of $S_n(\theta)$ to $S(\theta)$. (How we usually proof such convergence?).

Theorem Suppose $S(\theta)$ is such a non-random function that, for any $\delta > 0$, $\exists\ \eta_\delta > 0$ and for a δ-NBHD of θ_0, $N_\delta(\theta_0) \triangleq \{\theta : \|\theta-\theta_0\| \leq \delta\}$, we have

$$\inf_{\theta\notin N_\delta(\theta_0)} S(\theta) - S(\theta_0) \geq \eta_\delta > 0. \qquad (1)$$

This implies that θ_0 is the unique (globle) minimum of $S(\cdot)$. Further, suppose $S_n(\theta)$ is a random function based on a sample of n observations (no i.i.d. assumption) and it attains its global minimum (not necessary unique) value at $\hat{\theta}_n$. If $S_n(\theta) \to S(\theta)$ in probability uniformly for $\theta \in \Theta$ as $n \to \infty$, i.e.

$$\sup_{\theta\in\Theta} |S_n(\theta) - S(\theta)| \overset{\mathcal{P}}{\to} 0, \text{ as } n \to \infty, \qquad (2)$$

then $\hat{\theta}_n \overset{\mathcal{P}}{\to} \theta_0$.

Proof: WLOG, suppose $\hat{\theta}_n$ is a global minimizer of $S_n(\theta)$. Let $A_n \triangleq \{\hat{\theta}_n \in \Theta : S_n(\hat{\theta}_n) - S_n(\theta_0) \leq 0\}$. Then $P(A_n) \equiv 1$. Now let $B_n \triangleq \{\hat{\theta}_n \in \Theta : \hat{\theta}_n \in N_\delta(\theta_0)\}$ for an arbitrary but fixed $\delta > 0$. Notice we always have

$$\begin{aligned} 1 &= P(A_n) \\ &= P(A_n \cap B_n) + P(A_n \cap B_n^c) \\ &\leq P(B_n) + P(A_n \cap B_n^c) \end{aligned}$$

Thus, if we can show $P(A_n \cap B_n^c) \to 0$ as $n \to \infty$, it implies that $P(B_n) \to 1$ and that $\hat{\theta}_n \overset{\mathcal{P}}{\to} \theta_0$. Here comes the proof of the above claim ($P(A_n \cap B_n^c) \to 0$).

Because of (2), $\forall \epsilon > 0$, $\exists N_\epsilon$, for $\forall n \geq N_\epsilon$, $P(\sup_\theta |S_n(\theta) - S(\theta)| > \epsilon) < \epsilon$. This allows us to substitute $S_n(\cdot)$ with $S(\cdot)$, at least with probability $> 1 - \epsilon$. We compute

$$\begin{aligned} 0 &\leq P(A_n \cap B_n^c) \\ &\leq \epsilon + P(A_n \cap B_n^c \cap \{\sup_\theta |S_n(\theta) - S(\theta)| \leq \epsilon\}) \qquad (3) \\ &\leq \epsilon + P(\{|S(\hat{\theta}_n) - S(\theta_0)| \leq 2\epsilon\} \cap B_n^c) \qquad (4) \end{aligned}$$

where inequality (3) is because of

$$P(A_n \cap B_n^c \cap \{\sup_\theta |S_n(\theta) - S(\theta)| > \epsilon\}) \leq P(\{\sup_\theta |S_n(\theta) - S(\theta)| > \epsilon\}) < \epsilon$$

and inequality (4) is due to that

$$A_n \cap \{\sup_\theta |S_n(\theta) - S(\theta)| \leq \epsilon\} \subset \{S(\hat{\theta}_n) - S(\theta_0) \leq 2\epsilon\}. \qquad (5)$$

When $2\epsilon < \eta_\delta$, the last probability in expression (4) is zero since that is an impossible event. Thus, for $\epsilon < \eta_\delta/2$, $n > N_\epsilon$, we have $0 \leq P(A_n \cap B_n^c) \leq \epsilon$. Since ϵ is arbitrary, it follows that $P(A_n \cap B_n^c) \to 0$, as $n \to \infty$. QED.

Notes on (5). Since we have uniform closeness, therefore

$$|S_n(\hat{\theta}_n) - S(\hat{\theta}_n)| \leq \epsilon$$

and

$$|S(\theta_0) - S_n(\theta_0)| \leq \epsilon \ .$$

Add them together, we have

$$|S_n(\hat{\theta}_n) - S_n(\theta_0) + S(\theta_0) - S(\hat{\theta}_n)| \leq 2\epsilon$$

Notice $S_n(\hat{\theta}_n) - S_n(\theta_0) \leq 0$ (globle min). This imply

$$S(\theta_0) - S(\hat{\theta}_n) \geq -2\epsilon$$

which is

$$S(\hat{\theta}_n) - S(\theta_0) \leq 2\epsilon$$

Since θ_0 is the globle min of $S(\cdot)$, the above is also ≥ 0.

3 The efficiency of MLE

In the discussion that follows, the MLE is always thought of as the solution of the likelihood equation.

$$\frac{\partial}{\partial \theta} \log f_\theta(x_1, \cdots, x_n) = 0 \ .$$

In i.i.d. case, the equation can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(x_i) = 0 \ .$$

Notice $E_{\theta_0} \frac{\partial}{\partial \theta} \log f_\theta(x_1, \cdots, x_n) = 0$ when $\theta = \theta_0$.

In general, when a function of both θ and $x, g(\theta, x)$ satisfy

$$E_{\theta_0} g(\theta, X) = 0 \text{ when } \theta = \theta_0$$

then $g(\cdot)$ is called an unbiased estimating function. An estimate of θ based on a sample $(\{X_i\}_{i=1}^n)$ could be obtain by solving, for θ

$$\frac{1}{n} \sum_{i=1}^{n} g(\theta, x_i) = 0$$

For example, an unbiased estimate of θ, $T(X)$, could be thought of as the solution of the following unbiased estimating equation

$$g(x, \theta) = T(x) - \theta = 0$$

This function satisfies $E_{\theta_0}(T(X) - \theta_0) = 0$. However, the solution of unbiased estimating equation may be biased, though asymptotically the bias goes to zero.

Definition The *efficiency* (or *information*) of an estimator obtained by solving the unbiased estimating equation $g(x, \theta) = 0$ is

$$I_g(\theta) = \frac{[E_\theta(\frac{\partial}{\partial \theta} g(X, \theta))]^2}{E_\theta g^2(X, \theta)} \ .$$

Remark If $g(x,\theta) = \frac{\partial}{\partial\theta}\log f_\theta(x)$, the above information becomes the usual Fisher information.

Theorem For any unbiased estimating function $g(x,\theta)$ that satisfies the regularity conditions ((i) the information $I_g$is well defined, (ii) $\frac{\partial}{\partial\theta}\int g = \int \frac{\partial}{\partial\theta} g$), we have

$$I_g(\theta) \leq E_\theta[\frac{\partial}{\partial\theta}\log f_\theta(X)]^2 = \text{ Fisher Information}$$

Proof: Under the integral sign, differentiating the equation $E_\theta g(X,\theta) = 0$, we have

$$\frac{\partial}{\partial\theta}[\int g(x,\theta) f_\theta(x)dx] = \frac{\partial}{\partial\theta}[0],$$

that is

$$Eg' + Ef'/fg = 0$$

move the term and then square the equation we have,

$$(Eg')^2 = (Ef'/fg)^2$$

Now use Cauchy-Schwartz inequality for the right hand side term.

$$(Eg')^2 \leq E(f'/f)^2 \times Eg^2$$

Move the term involve g to the left and notice $E(f'/f)^2 = I_{fisher}$. QED.

Remark This theorem includes the usual Cramer-Rao inequality as a special case.

Remark The equality sign of the theorem holds iff

$$g(x,\theta) = \lambda(\theta)\cdot\frac{\partial}{\partial\theta}\log f_\theta(x).$$

Now let us study the estimator "$\hat\theta_{MLE}$" defined by equation $\sum_{i=1}^n g(\hat\theta_{MLE}, x_i) = 0$.

By Taylor expansion,

$$\sum_{i=1}^n [g(x_i,\theta_0) + (\hat\theta_{MLE}-\theta_0)g'(x_i,\theta_0) + \frac{(\hat\theta_{MLE}-\theta_0)^2}{2}g''(x_i,\bar\theta)] = 0 \tag{6}$$

where $\bar\theta$ is between $\hat\theta_{MLE}$ and θ_0 and the derivatives are w.r.t. θ (assume exist).

Rearranging equation (6) yields

$$\sqrt{n}(\hat\theta_{MLE}-\theta_0) \quad = \quad -\frac{\sum_{i=1}^n g(x_i,\theta_0)}{\sqrt{n}} \times \frac{1}{\frac{1}{n}\{\sum_{i=1}^n g'(x_i,\theta_0) + \frac{(\hat\theta_{MLE}-\theta_0)}{2}\sum_{i=1}^n g''(x_i,\bar\theta)\}}$$

Now we want to show two things. First, Since $g(x_i, \theta_0)$ are i.i.d. r.v.s with $Eg(X, \theta_0) = 0$ and $Eg^2(X, \theta_0) < \infty$ (assumption), by the CLT, we have

$$-\frac{\sum_{i=1}^{n} g(x_i, \theta_0)}{\sqrt{n}} \overset{\mathcal{D}}{\to} N(0, Eg^2(X_1, \theta_0) = k)$$

Second, By the WLLN, $[\frac{1}{n}\sum_{i=1}^{n} g'(x_i, \theta_0) + \frac{\hat{\theta}_{MLE} - \theta_0}{2}\frac{1}{n}\sum_{i=1}^{n} g''(x_i, \bar{\theta})]^{-1} \overset{\mathcal{P}}{\to} \frac{1}{c}$.

(i) We know $\frac{1}{n}\sum_{i=1}^{n} g'(x_i, \theta_0) \overset{\mathcal{P}}{\to} Eg'(x_i, \theta_0)$ again, by i.i.d. of the r.v.s $g'(X_i, \theta_0)$

(ii) Need to show $|\frac{1}{n}\sum_{i=1}^{n} g''(x_i, \bar{\theta})| < \frac{1}{n}\sum_{i=1}^{n} M(x_i) \overset{\mathcal{P}}{\to} E_{\theta_0}M(X_1) < \infty$ for θ_0, i.e. $|g''(x_i, \theta)| < M(x_i)$.

(iii) We know $\frac{\hat{\theta}_{MLE} - \theta_0}{2} \overset{\mathcal{P}}{\longrightarrow} 0$. (by section 2)

Finally, by Slutsky Theorem, $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = X_n \cdot a_n \overset{\mathcal{D}}{\to} X \cdot a$

We have proved a theorem for the asymptotic distribution of MLE above. Please formulate the theorem yourself.

Remark the whole proof also works for other estimating functions.

Identification of the limiting distribution: The constant $c = Eg'(x_1, \theta_0)$.

In the case of MLE, $g(x_i, \theta_0) = \frac{\partial}{\partial\theta} \log f_\theta(x_1)$, and thus

$$\begin{aligned} c &= E\frac{\partial^2}{\partial\theta^2} \log f_\theta(x_1)|_{\theta_0} \\ &= \int \frac{\partial^2}{\partial\theta^2} \log f_\theta(x)|_{\theta_0} \cdot f_{\theta_0}(x)dx \\ &= -I_{\text{Fisher}}(\theta_0). \end{aligned}$$

On the other hand, the variance of $g(x_1, \theta_0)$ is

$$\begin{aligned} E[g(X_1, \theta_0)]^2 &= E[\frac{\partial}{\partial\theta} \log f_\theta(X_1)]^2 \\ &= I_{\text{Fisher}}(\theta_0). \end{aligned}$$

Therefore, the limiting distribution is $N(0, \dfrac{1}{I_{Fisher}(\theta)})$.

Remark In the case of a general $g(x, \theta)$ function, the limiting distribution is $N(0, \dfrac{1}{I_g})$.

Remark For (purely) discrete random variables, the above proof still works, with density replaced by PMF.

Even for mixed distributions, the proof is OK. The Key is to have a fixed, σ-finite dominated measure, so that the density can be defined.

Remark Generalization for multidimesional θ.

Remark Generalization for independent but not identically distributed observations.

Example (Neyman-Scott). Two observations each from $N(\theta_i, \sigma^2)$ population. Both θ_i and σ^2 are unknown. Let the number of populations go to infinite. The MLE of σ^2 is not even consistent.

Therefore for infinite dimensinal parameters we need to be careful. But also see the empirical likelihood ratio result.

Remark Expected information and observed information.

4 Nuisance Parameters and parameter of interest

How does the Fisher information for the parameter of interest change when there are nuisance parameter(s)?

Suppose we have parameters θ and η.

Example Fisher information matrix for θ, η. Cremer-Rao inequality in the matrix form. Specialize to parameter θ alone.

The information for θ alone is defined as follows:

the second derivative wrt θ can be decomposed as the orthoganal sum of a component in the direction of $\partial\eta$ and a component perpendicular to $\partial\eta$.

The length of the perpendicular part is the information for θ.

For infinite dimensional nuisance parameter, 'derivative' needs careful work.

Stein (1956) "a nonparametric problem is at least as difficult as any of the parametric problems obtained by assume we have enough knowledge of the unknown state of nature to restric it to a finite dimensional set"

Therefore he define the information for the parameter as the infemum of all those finite dimentional parametric sub-models.

5 Likelihood Ratio Statistic

Theorem (Wilks theorem in $\mathcal{R}^1$) Suppose $X_1, X_2, \cdots$ are i.i.d with density $f(x, \theta)$, $\theta \in \Theta$, where Θ is an open set in $\mathcal{R}^1$. Let $\hat{\theta}_n = \hat{\theta}$ denote the MLE of θ based on n observations. If the null hypothesis $H_0 : \theta = \theta_0$ is true, then

$$
\begin{aligned}
W &= -2 \log \frac{\sup\limits_{\theta \in \Theta_0} \prod\limits_{i=1}^{n} f(x_i, \theta)}{\sup\limits_{\theta \in \Theta} \prod\limits_{i=1}^{n} f(x_i, \theta)} \\
&= -2 \log \frac{\prod_{i=1}^{n} f(x_i, \theta_0)}{\prod_{i=1}^{n} f(x_i, \hat{\theta})}
\end{aligned}
$$

has an asymptotic χ^2 distribution with $df = 1$.

PROOF:

$$W = -2 \left[\sum \log f(x_i, \theta_0) - \sum \log f(x_i, \hat{\theta}) \right] .$$

Use Taylor expansion on the first summation above, around $\hat{\theta}$, (since θ_0 and $\hat{\theta}_n$ are close to each other.)

$$\sum \log f(x_i, \theta_0) = \sum \log f(x_i, \hat{\theta}_n) + (\theta_0 - \hat{\theta}_n) \sum \frac{\partial}{\partial \theta} + 1/2 (\theta_0 - \hat{\theta}_n)^2 \sum \frac{\partial^2}{\partial \theta^2}$$

Notice the first derivative in the Taylor expansion is zero, (derivative at $\hat{\theta}$) since $\hat{\theta}$ is MLE.

We get

$$W = n(\theta_0 - \hat{\theta})^2 \frac{1}{n} \sum - \frac{d^2}{d\theta^2} \log f(x_i, \bar{\theta}).$$

By the CLT for MLE we have

$$\sqrt{n}(\theta_0 - \hat{\theta}) \to N(0, I^{-1})$$

which imply

$$I \times n(\theta_0 - \hat{\theta})^2 \to \chi^2_{df=1} .$$

In view of Slutsky theorem, we only need to show

$$-\frac{1}{n} \sum \frac{d^2}{d\theta^2} \log f(x_i, \bar{\theta}) \to I$$

in probability. This can be shown easily, assuming for example

$$\frac{d^2}{d\theta^2} \log f(x, \theta)$$

is continuous at θ_0 uniformly for x.

QED.

6 Three types of tests related to the likelihood functions

6.1 Likelihood ratio tests (Wilks)

To test $H_0 : \theta = \theta_0$ use

$$-2 \log \frac{\sup\limits_{\theta \in \Theta_0} \Pi f(x_i, \theta)}{\sup\limits_{\theta \in \Theta} \Pi f(x_i, \theta)} \ .$$

Under H_0 this has an approximate χ^2 distribution.

6.2 Score test (Rao)

Let $g = g(\theta, x)$ be the score functions

$$g_1(\theta, x) = \partial \log Lik \partial \theta$$

To test $H_0 : \theta = \theta_0$ use

$$\sum g(\theta_0, x_i) \sqrt{var} g(\theta_0, x) \ .$$

Under H_0 this has an approximate $N(0,1)$ distribution. The proof can be easily obtained from results of previous sections. The advantage of this method is that we do not need to find the MLE $\hat{\theta}$.

6.3 Wald type tests

To test $H_0 : \theta = \theta_0$ use

$$\frac{\hat{\theta}_{MLE} - \theta_0}{\sqrt{Var(\hat{\theta}_{MLE})}} \ .$$

Under H_0 this has an approximate $N(0,1)$ distribution, equivalently the square of it has chi square distribution.

Confidence intervals can be obtained by inverting the tests.

7 Transformation of parameter(s)

Consider a 1 to 1, differentiable transformation of the parameter:

$$\theta_{new} = T(\theta_{old}) \ .$$

After a 1 to 1 trandformation of the parameter, we can do the three tests same as before the trandformation. However, the Wald test/confidence interval will change after the transformation. In other words, the P-value of the test can be different, the confidence interval is different.

This is why we see arc-sin, log, log-log, square, logit, square-root, Z-transformations in the statistical literature. They all aim at improving the inference after applying the said transformation. (when coupled with Wald method.)

However, the likelihood ratio test/confidence interval is *invariant* under transformation.

Example Suppose we have a sample from binomial population and we would like to estimate p the success probability. The "plain" Wald confidence interval is

$$\hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we use a logit transformation on the parameter, $\theta = \log p/(1-p)$ then we get

$$\log\frac{\hat{p}}{1-\hat{p}} \pm 1.645\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}$$

A good reference is May 2001 Statistical Science.

Nuisance parameters. When there are more than one parameter in the likelihood and we are only testing/estimating for one of the parameter (parameter of interest) then the rest of the parameters are called nuisance parameter.

How are the three testing method handle the nuisance parameter?

For the Wilks likelihood ratio method, we simply "profileing them out". i.e. we form the ratio where the numerator only fix the parameter of interest at the null value.

The resulting chi square statistics will have degree of freedom equal to the difference of the number of free parameters (i.e. parameters that are not fixed) in the numerator and the denominator. For example the likelihood have 5 parameters and we are interested to test if two of them are zero. Then the numerator likelihood will fix the two parameters of interest at zero, and maximize over the other three nuisance parameters. The denominator will be maximized over 5 parameters; and the degree of freedom of the likelihood ratio statistics is 5-3=2.

8 Large Sample Property of Bayes Estimation

Two types of results:

(1) The posterior distribution is asymptotically normal with mean $\hat{\theta}_{mle}$ and variance equal to the inverse of the observed information matrix. Ref: Walker, JRSSB 1969, p. 80-88.

To make it precise: as $n \to \infty$, the posterior probability that $\hat{\theta} + b\sigma_n < \theta < \hat{\theta} + a\sigma_n$, namely

$$\int_{\hat{\theta}+b\sigma_n}^{\hat{\theta}+a\sigma_n} \text{posterior } d\theta \longrightarrow \int_b^a \phi(\theta) d\theta = \Phi(a) - \Phi(b)$$

(converge in probability) where $\Phi(\cdot)$ is the CDF of standard normal distribution. In the above $\hat{\theta}$ is the MLE; σ_n^2 is the inverse of the observed Fisher information number: $\{-\log Lik^{''}(\hat{\theta})\}^{-1}$; and a and b are any two finite constants. There are many regularity conditions, but one of them is that the Fisher information must grow to infinity as n grows.

One of the insight from the proof is that the likelihood function, $\prod_{i=1}^n f(x_i, \theta)$, as a function of θ is similar to a normal density function with mean $\hat{\theta}$, variance σ_n^2.

(2) The difference of $\hat{\theta}_{mle}$ and a Bayes estimator is asymptotically smaller than $\frac{1}{\sqrt{n}}$. (sample size n). Ref: Zhao, Ann. Statist. 1970, p. 601-608.

9 Re-sampling Estimation Equation

Sometimes the estimation equation is easily defined but the variance of the estimator (?) is difficult. For example, the estimation equation for the median. (the variance for the sample median is difficult).

Suppose $X_1, \ldots, X_n$ is iid from a distribution with median θ_0. Consider

$$0 = \sum \phi(X_i - \theta)$$

where $\phi(t) = 2I_{[t>0]} - 1$. Let us denote the solution by $\hat{\theta}$ (the sample median).

Consider the estimation equation

$$Z = \sum \phi(X_i - \theta)$$

where Z is a random variable generated from the distribution $2 \times bin(n, 0.5) - n$ (and independent of everything).

Denote the solution of the above estimation equation by $\hat{\theta}^*$. Theory show that

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \sqrt{n}(\hat{\theta}^* - \hat{\theta}) \qquad \textit{in distribution.}$$

Significance: the distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ can be obtained by Monte Carlo method. We can repeatedly generate Z from $2 \times bin(n, 0.5) - n$ distribution while keep X_i fixed, solving the estimation equation for many Z give us many $\hat{\theta}^*$. And the sample distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ can be obtained. With fast and cheap computing power, this is easy.

Q: does this approximation correct to the second order?

For general estimating function $g()$, we may generate Z from the distribution $N(0, \sigma^2)$ with $\sigma^2 = 1/n \sum g^2(X_i, \hat{\theta})$.

Some problems:

0. Sketch the picture for the function $S(\theta) = E \log f_\theta(X)$ where the densities come from (i) a normal location family with $\theta = mean$ (ii) exponential family of distributions ($\theta = \lambda$).

1. Let $f(x, \theta)$ be a family of densities, with $\theta \in \Theta$ where Θ is an open subset of the plain R^2.

Suppose $X_1, X_2, \cdots, X_n$ are iid random variables from the density $f(x, \theta_0)$ for a fixed θ_0 in the interior of Θ.

Show that (under regularity conditions) for any given $\theta \in \Theta$ and $\theta \neq \theta_0$, we have

$$P(L_n(\theta) < L_n(\theta_0)) \to 1 \quad as\ n \to \infty$$

where $L_n(\theta)$ is the likelihood function $\prod_{i=1}^n f(X_i, \theta)$.

Regularity conditions:

2. Suppose we have two binomial populations with success probabilities p_1 and p_2 and number of trials n_1 and n_2.

Also assume we observed n_{11} successes and n_{12} failures from the first population ($n_{11} + n_{12} = n_1$). We observed n_{21} successes and n_{22} failures from the second population ($n_{21} + n_{22} = n_2$.

Assume $p_1 = p_2 = p$ and $0 < p < 1$. Show that as $n_1 \to \infty$ and $n_2 \to \infty$ we have

$$\sum_{jk} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}$$

converge in distribution to a chi-square distribution (what is the df of the limiting chi-square distribution?)

where the summation is over the four terms as follows

$$E_{11} = \frac{n_1}{n_1 + n_2}(n_{11} + n_{21}); E_{12} =; E_{21} =; E_{22} = \ .$$

3. Suppose $X_1, X_2, \cdots X_n$ are iid r.v.s from exponential distribution with parameter $\lambda > 0$. (i.e.

$$f(x, \lambda) = \lambda \exp(-\lambda t) \quad \text{for} \ \ t > 0.$$

Show that

$$2 \log L_n(\hat{\lambda}) - 2 \log L_n(\lambda)$$

converge in distribution to chi-square distribution. where $\hat{\theta}$ is the MLE of λ.

4. Suppose $X_1, \cdots, X_n$ are iid uniform $[\mu - 1, \mu + 1]$ r.v.s. Define a function $g(t)$ as follows:

$$g(t) = \begin{cases} 0 & \text{for } |t| \leq 0.1 \\ (t - 0.1)^2 & \text{for } t > 0.1 \\ (t + 0.1)^2 & \text{for } t < -0.1 \end{cases}$$

Now define an estimator, $\hat{\theta}$ of μ as the (any) minimizer of the following function

$$\min_{\theta} \sum_{i=1}^{n} g(X_i - \theta) \ .$$

(a) show that the estimator $\hat{\theta}$ is (weakly) consistant. (i.e. $\hat{\theta} \rightarrow_P \mu$ as $n \rightarrow \infty$.)

(b) show that the estimator $\hat{\theta}$ is asymptotically normally distributed.

(c) Give the expression of the asymptotic variance in (b).

(5) Given two sequences of r.v.s X_n and Y_n. Suppose $X_n = O_p(1/\sqrt{n})$ and $Y_n = O_p(1/\sqrt{n})$. Further suppose $cor(X_n, Y_n) = 1 - O_p(1/\sqrt{n})$. Show that $X_n - Y_n = o_p(1/\sqrt{n})$.

(6) Suppose we get 3 successes in 20 flip of a coin, please obtain 90% confidence interval for the unknown p. By "plain" Wald, Wald with logistic transform, and by likelihood ratio.

10 Infinite dimensional nuisance parameter

Proof of ELT:

First use Lagrange multiplier calculation to the empirical likelihood to get an expression of $-2\log LR$:

$$-2\log LR = -2\log\prod_{i=1}^{n} nw_i = 2\sum_{i=1}^{n}\log[1-\lambda/n(X_i-\mu)],$$

with λ being the solution of the equation

$$0 = \frac{1}{n}\sum\frac{X_i-\mu}{1-\lambda/n(X_i-\mu)}\ . \tag{7}$$

Notice the identity

$$\frac{1}{1+\epsilon} = 1-\frac{\epsilon}{1+\epsilon} = 1-\epsilon+\frac{\epsilon^2}{1+\epsilon} = 1-\epsilon+\epsilon^2-\frac{\epsilon^3}{1+\epsilon} = \cdots\ .$$

For $\epsilon \to 0$ the last term on the right hand side has same order as its numerator.

Apply this identity to

$$\frac{1}{1-\lambda/n(X_i-\mu)}$$

(ASSUME we showed $\lambda = O_P(1/\sqrt{n})$) we get

$$\sum_{i=1}^{n}\frac{X_i-\mu}{1-\lambda/n(X_i-\mu)} = \sum(X_i-\mu)+\lambda/n\sum(X_i-\mu)^2+\lambda^2/n^2\sum\frac{(X_i-\mu)^3}{1-\lambda/n(X_i-\mu)}\ .$$

From here we can get the approximate solution to the equation (?)

$$\lambda^*/n = \frac{-\sum(X_i-\mu)}{\sum(X_i-\mu)^2}+o_P(1/\sqrt{n})\ .$$

Next consider the empirical likelihood ratio

$$-2\log LR = 2\sum\log[1-\lambda^*/n(X_i-\mu)]\ .$$

Use Taylor expansion to the function $\log(1+\epsilon)$, we have

$$-2\log LR = 2[-\lambda^*/n\sum(X_i-\mu)+1/2(\lambda^*)^2/n^2\sum(X_i-\mu)^2-O_P((\lambda^*)^3)\sum(X_i-\mu)^3]\ .$$

Finally plug in the λ^* and carefully re-arrange the terms to get

$$-2\log LR = \frac{[\sum(X_i-\mu)]^2}{\sum(X_i-\mu)^2}+o_P(1)\ .$$

By CLT we know

$$\frac{1}{\sqrt{n}} \sum (X_i - \mu) \longrightarrow N(0, \sigma^2)$$

in distribution and by WLLN we know

$$\frac{1}{n} \sum (X_i - \mu)^2 \longrightarrow \sigma^2 = Var(X)$$

in probability.

Combine the above two convergence results and use Slutsky theorem, we see that $-2 \log LR \to \chi^2$ in distribution.

Remark : If the observations X_i were only independent but not identically distributed with CDF's $F_i(t)$, we can still prove the following.

Suppose for each $i : 1 \le i \le n$

$$\mu = \int g_i(t) dF_i(t) = E g_i(X_i) \qquad \sigma_i^2 = Var(g_i(X_i)) \ .$$

(Same μ but different σ_i^2.)

Suppose the Lindeberg or the Liapounoff condition hold for the sequence of the r.v. $g_i(X_i)$:

$$\sum_{i=1}^{n} E \left[\frac{g_i(X_i) - \mu}{\sum \sigma_i^2} \right]^3 \to 0$$

as $n \to \infty$.

This guarentees we have

$$\frac{\sum (g_i(X_i) - \mu)}{\sqrt{\sum \sigma_i^2}} \longrightarrow N(0, 1)$$

in distribution.

Assume also $1/M < \sigma_i^2 < M$ for some $M > 0$. We still need to show also

$$\max(g_i(X_i) - \mu) = o_p(\sqrt{n})$$

and show

$$\frac{\sum [g_i(X_i) - \mu]^2}{\sum \sigma_i^2} \to 1$$

But they can be done.

Remark: The mean of a distribution $\int g(t) dF(t)$ do not seems to be that hard to do by other inference method. Why use EL then? Try the case where the data has a few censored observations.

For censored observations, we assume the lifetimes are independent, non-identically distributed; whereas the censoring times are iid.