Lecture 8: Point estimation: Asymptotics

Ma Xuejun

School of Mathematical Sciences

Soochow University

https://xuejunma.github.io

Outline

1. Asymptotic Theory
2. Consistency of MLE
3. Robustness

Asymptotic Theory

- Until now, what we have introduced are all assumed that the sample size n is given
- Currently, we usually have large sample size. In practice, "$n \geq 30$" usually works
- We write it as $n \to \infty$. Under this condition, what will happen to our estimators?
  - Review of o, O, and convergence
  - Review of distance between probability distributions
  - Consistency
  - Efficiency and Relative Efficiency
  - MLE
  - Robustness

Review

We need some terms to describe what will happen when $n \to \infty$

- $a_n = o(1)$ means that $a_n \to 0$ when $n \to \infty$
- $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$ as $n \to \infty$
- $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{P} 0$ as $n \to \infty$
- $n^b o_p(1) = o_p(n^b)$,so $\sqrt{n} o_p(1/\sqrt{n}) = o_p(1)$
- $a_n = O_p(1)$ if $|a_n|$ is bounded by a constant as $n \to \infty$
- $Y_n = O_p(1)$ if for any $\epsilon > 0$ there exists a constant M such that $\lim_{n\to\infty} P(|Y_n| > M) < \epsilon$ as $n \to \infty$
- $Y_n = o_p(b_n)$ if $Y_n/b_n = O_p(1/\sqrt{n})$
- If $\sqrt{n}(Y_n - c) \xrightarrow{d} Y$,then,$Y_n = O_p(1/\sqrt{n})$
- $O_p(1) \times O_p(1) = O_p(1), o_p(1) \times o_p(1) = o_p(1), o_p(1) \times O_p(1) = o_p(1)$

Distances

Let P and Q be distributions with densities p and q.

- Total Variance distance: $TV(P,Q) = \sup_A |P(A) - Q(A)|$
- L_1 distance: $d_1(P,Q) = \int |p - q|$
- Hellinger distance: $h(P,Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
- Kullback-Leibler distance: $K(P,Q) = \int p \log(p/q)$
- L_2 distance: $d_2(P,Q) = \int (p - q)^2$
- As one type of distance,all of them satisfies that $D(P,Q) > 0$ if $P \neq Q$
- The distances are closely related,say,

$$TV(P,Q) \leq h(P,Q) \leq \sqrt{2TV(P,Q)}$$

Note:all integrals are integrals w.r.t. the probability measure.

Consistency

- Recall: with LLN, the sample average satisfies that $\bar{X}_n \xrightarrow{P} E[X]$
- Generally,for an estimator $\hat{\theta}$, we hope $\hat{\theta} \xrightarrow{P} \theta$

Definition:Consistency

Let $\hat{\theta}_n$ be an estimator for θ. The estimator is said to be consistent if

$$\hat{\theta} \xrightarrow{P} \theta$$

To show consistency,we can:

- Prove that $P(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ for any $\epsilon > 0$
- Recall that L^p convergence indicate convergence in probability,so if $\hat{\theta}_n \xrightarrow{L^2} \theta$,$\hat{\theta}_n$ is consistent.The L^2 distance between $\hat{\theta}_n$ and θ is

$$\int (\hat{\theta}_n - \theta)^2 dF(\hat{\theta}) = E[(\hat{\theta}_n - \theta)^2] = MSE(\hat{\theta}_n)$$

So,if $MSE(\hat{\theta}_n) \to 0$,$\hat{\theta}_n$ is consistent.

Consistency

Example. Let$X_1, \cdots, X_n \sim Bernoulli(p)$.

- Consider the MLE $\hat{p} = \bar{X}_n$.According to LLN,$\bar{X}_n \xrightarrow{P} p$. So, the MLE$\hat{p}$ is consistent.
- Consider the function $\tau(p) = \log(p/(1-p))$.Let the estimator be $W = \log(\bar{X}_n/(1 - barX_n))$.According to the continuous mapping theorem,$W \xrightarrow{P} \tau(p)$.So W is consistent
- Consider the estimator
$$\hat{p} = \frac{\sum X_i + 1}{n+1}.$$
For $\hat{p}$, we have that
$$Bias(\hat{p}) = E[\hat{p}] - p = \frac{1-p}{(n+1)} \to 0 \quad Var(\hat{p}) = \frac{np(1-p)}{(n+1)^2} \to 0.$$
Therefor, $MSE(\hat{p}) = Bias^2 + Var \to 0.\hat{p}$ is consistent.

Consistency

- Consistency $\neq$ Unbiasedness
  - $\hat{p}$ is biased,but consistent
  - X_1 is unbiased for $E[X]$,but not consistent
  - In reality,we prefer consistency(when we have more samples,we can be closer to the truth)
- According to LLN, the MoM estimators are always consistent (the random sample should satisfy the conditions for LNN)
- For Bayes model,it depends on the prior.If we the prior in inappropriate,then the estimator is not consistent.
- How about the the MLE?

Consistency of MLE

Theorem:Consistency of MLE

Let $X_1, X_2, \cdots$ be i.i.d. $f(x;\theta)$,and let $L(\theta)$ be the likelihood function.Let θ be the MLE of θ.If

- θ is identifiable,i.e.,if $\theta_1 \neq \theta_2$,the $f(x|\theta_1) \neq f(x|\theta_2)$;and
- $f(x;\theta)$ have common support w.r.t. different θ,and differentiable in θ;and
- the true parameter θ_0 is an interior point of the parameter space then for any continuous function of θ,$\tau(\theta)$, there is

$$\tau(\hat{\theta}) \xrightarrow{P} \tau(\theta)$$

- If the function is $\tau(\theta) = \theta$,then there is $\hat{\theta} \xrightarrow{P} \theta$,which shows the consistency of MLE
- According to the invariance of MLE,$\tau(\hat{\theta})$ is the MLE $\tau(\theta)$.So the consistency of MLE holds even for a function of θ

Proof

To prove it, we define a new term as the expectation of the one sample log-likelihood function:

$$l(\theta) = E[\log f(X_i; \theta)].$$

The sketch of the proof is

- The MLE $\hat{\theta}$ is the maximiser of the log-likelihood function $l_n(\theta)$,and also $\frac{1}{n} l_n(\theta)$
- The true parameter θ_0 is the maximiser of $l(\theta)$
- For any θ, $\frac{1}{n} l_n(\theta) \xrightarrow{P} l(\theta)$

Combine these conclusions, with the regularity conditions above and the technique in mathematical analysis, we can have the result.

In this class, we show these 3 conclusions only. The last part is easier for compact parameter space, but quite hard when we set such flexible conditions.

Proof

- The MLE $\hat{\theta}$ is the maximiser of the log-likelihood function $l_n(\theta)$.$\Leftarrow$ Definition of MLE.
- The true parameter θ_0 is the maximiser of $l(\theta)$.

$$\begin{aligned} l(\theta) - l(\theta_0) &= E_{\theta_0}[\log f(X;\theta)] - E_{\theta_0}[\log f(X;\theta_0)] \\ &= E_{\theta_0}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) \\ &< \log\left[E_{\theta_0}\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right] = 0^1 \end{aligned}$$

- For any θ, $\frac{1}{n}l_n(\theta) \xrightarrow{p} l(\theta)$. Note that

$$\frac{1}{n}l_n(\theta) = \frac{1}{n}\sum \log f(X_i;\theta) \xrightarrow{P} E[\log f(X;\theta)] = l(\theta)$$

The conergence comes from WLLN.

[1] $Eg(X) \leq g(EX)$: $g(\cdot)$ is a convex function

Inconsistency of an MLE

Example. Let $Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2), Y_{21}, Y_{22} \sim N(\mu_2, \sigma^2), \cdots,$ $Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2)$ Then in this case, the number of parameters increase as n increases, different from the case we discussed before, that the parameter space is fixed.

Solution. The MLE for σ^2 in this problem is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(Y_{ij} - \bar{Y}_i)^2}{2n}, \qquad \bar{Y}_i = (Y_{i1} + Y_{i2})/2$$

which is the average of the MLE of σ^2 in each small group.
Now,for each item in the summation, note that

$$\begin{aligned}\hat{\sigma}_i^2 &= \sum_{j=1}^{2} (Y_{ij} - \bar{Y}_i)^2 = (Y_{i1} - \frac{Y_{i1} + Y_{i2}}{2})^2 + (Y_{i2} - \frac{Y_{i1} + Y_{i2}}{2})^2 \\ &= \frac{(Y_{i1} + Y_{i2})^2}{2} \sim \sigma^2 \chi_1^2.\end{aligned}$$

So,with LLN,$\hat{\sigma}^2 = \frac{1}{2n} \sum \hat{\sigma}_i{}^2 = \frac{1}{n} \sum \hat{\sigma}_i{}^2/2 \xrightarrow{P} \sigma^2/2$.It does not converge to σ^2 in probability! The modified estimator $2\hat{\sigma}^2$ is consistent.

Asymptotic Properties

- Unbiasedness $\longrightarrow$ Asymptotic unbiasedness

$$E[\hat{\theta}_n] - \theta \to 0, \qquad n \to \infty.$$

- Unbiasedness is not that important in the asymptotic theory, since we have the consistency property

- Multiple consistent estimators?
- Cr$\acute{a}$mer-Rao Lower Bound $\longrightarrow$ Asymptotic efficient

$$Var(\hat{\theta}_n) \to Cr\acute{a}mer - Rao\ Lower\ Bound, \qquad n \to \infty$$

- Asymptotic relative efficiency: comparing the variance of these two estimators

Asymptotic Variance

Definition

Let $\hat{\theta}_n$ be an estimator for $\theta = \theta_0$.If for some deterministic sequence (a_n), we have

$$a_n(\hat{\theta}_n - \hat{\theta}_0) \xrightarrow{d} N(0, \sigma^2)$$

The σ^2 is called the asymptotic variance.

- If we study $\hat{\theta}_n - \hat{\theta}_0$ directly, then we have that it converges to 0 in prob. since it is consistent. That does not provide more information to us.
- Usually, for(a_n), we take it as n^c, which increases w.r.t. n, without any constant term that would impact the asymptotic variance.
- Example. In CLT,the average follows the estimation is that

$$\sqrt{n}(\bar{X}_n - E[X]) \xrightarrow{d} N(0, Var(X_1))$$

Here, $a_n = \sqrt{n} = n^{1/2}$ and the asymptotic variance is $Var(X_1)$.If $a_n = n^c$ where $c < 1/2$, then $Var(a_n \bar{X}_n) \to 0$. If $a_n = n^c$ where $c > 1/2$,then $Var(\sqrt{n}\bar{X}_n) \to \infty$

- Note: the asymptotic variance is different from $\lim_{n\to\infty} Var(a_n \hat{\theta}_n)$

Example

Example. Let $Y_n|W_n = w_n \sim N(0, w_n + (1 - w_n)\sigma_n^2)$ with $W_n \sim Bernoulli(p_n)$, where the sequence $(\sigma_n^2), (p_n)$ are known.Now,

$$\begin{aligned} Var(Y_n) &= E[Var(Y_n|W_n)] + Var(E[Y_n|W_n]) \\ &= E(W_n + (1 - W_n)\sigma_n^2) \\ &= p_n + (1 - p_n)\sigma_n^2 \end{aligned}$$

The variance would converge only if $\lim_{n\to\infty}(1 - p_n)\sigma_n^2 < \infty$.
Now we consider the asymptotic variance. First we should figure out the distribution it converges to in distribution. For some fixed a,

$$P(Y_n \leq a) = E[P(Y_n \leq a|W_n)] = (1 - p_n)\Phi(a/\sigma_n) + p_n\Phi(a),$$

where $\Phi(\cdot)$ is the CDF for standard normal distribution.
Therefor, if $p_n \to 1$ then $Y_n \to N(0, 1)$,so that the asymptotic variance is 1. However, if $\lim_{n\to\infty}(1 - p_n)\sigma_n^2 = \infty$ then this is the value of the limiting variance, which is of course different.

Asymptotic Efficiency

Definition

The estimator $\hat{\theta}_n$ is asymptotic efficient for a parameter $\theta = \theta_0$ if

- $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, v(\theta))$ and
- $$v(\theta_0) = \frac{1}{I(\theta_0)} \qquad Crámer - Rao\ Lower\ Bound$$

- One estimator is efficient, as long as it's asymptotic variance exists, and meets the CRLB.
- Obviously, for $\tau(\theta)$, the definition also works,except that the CRLB becomes the CRLB for $\tau(\theta)$
- MLE is always asymptotic efficient

Asymptotic Normality of the MLE

Regularity conditions:

- the dimension of the parameter space does not change with n;
- $f(x, \theta)$ have common support w.r.t. different θ, and is differentiable in θ;
- the differentiation w.r.t. θ is interchangeable with integration over x.

Theorem: Asymptotic Normality of MLE

Let $\hat{\theta}_n$ be the MLE for the parameter θ. Under the regularity condition, there is

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)}).$$

- Hence, $\hat{\theta}_n = \theta + O_p(1/\sqrt{n})$.

Asymptotic Normality of the MLE

Proof. By Taylor's theorem

$$l'(\hat{\theta}) = l'(\theta) + (\hat{\theta} - \theta)l''(\theta) + \cdots$$

Recall that $l'(\hat{\theta}) = 0$ since $\hat{\theta}$ is MLE,so we have

$$0 = l'(\theta) + (\hat{\theta} - \theta)l''(\theta) + \cdots$$

$$(\hat{\theta} - \theta) \approx -\frac{l'(\theta)}{l''(\theta)}$$

$$\sqrt{n}(\hat{\theta} - \theta) \approx -\frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} \equiv \frac{A}{B}$$

Now, $A = \frac{1}{\sqrt{n}}l'(\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i) = \sqrt{n}(\bar{S}_n - 0)$,where $S(\theta, X_i)$ is the score function based on X_i. Recall that $E[S(\theta, X_i)] = 0$ and $Var(S(\theta, X_i)) = I(\theta)$. By CLT,$A \xrightarrow{d} N(0, I(\theta))$.

By WLLN, $B \xrightarrow{P} E[l''(\theta)] = I(\theta)$.Combine them, by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta} - \theta) \approx \frac{A}{B} \xrightarrow{d} \frac{1}{I(\theta)}N(0, I(\theta)) = N(0, 1/I(\theta)).$$

Asymptotic Normality of the MLE

- According to the proof, we can see that

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n}\frac{S(\theta, X_i)}{I(\theta)} + o_p(n^{-1/2}).$$

  The function $\frac{S(\theta,X_i)}{I(\theta)}$ is called the $influencefunction$.
- The asymptotic variance of $\hat{\theta}$ is $1/I(\theta)$
- The estimated asymptotic variance of $\hat{\theta}$ is $1/I(\hat{\theta})$
- If τ is a smooth function of θ, then

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{d} N(0, (\tau'(\theta))^2/I(\theta)),$$

  with asymptotic variance $(\tau'(\theta))^2/I(\theta)$ .The estimated asymptotic variance is $(\tau'(\hat{\theta}))^2/I(\hat{\theta})$

Example

Example. $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Exponential(\theta)$

Now $f(x;\theta) = \theta e^{-\theta x}$ and $L(\theta) = \theta^n e^{-n\theta \bar{X}}$,Hence,the log-likelihood function

$$l(\theta) = -n\theta\bar{X} + n\log\theta$$

and

$$S(\theta) = \frac{n}{\theta} - n\bar{X}, \qquad l''(\theta) = -n/\theta^2 < 0$$

The MLE is $\hat{\theta} = 1/\bar{X}$.The Fisher information is $I_n(\theta) = -E[-n/\theta^2] = n/\theta^2$.Therefore,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta^2)$$

Example

Example. $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Bernoulli(p)$

We can find that the Fisher information for $n = 1$ is $I(p) = 1/(p(1-p))$.So, for the MLE$\hat{p} = \bar{X}$,

$$\sqrt{n}(\bar{X} - p) \xrightarrow{d} N(0, p(1-p)).$$

Now suppose we want to estimate $\tau = p(1-p)$.The MLE is $\hat{\tau} = \bar{X}/(1-\bar{X})$,according to the invariance od MLE.Now

$$\frac{\partial}{\partial p}\frac{p}{1-p} = \frac{1}{(1-p)^2}.$$

The asymptotic distribution for $\hat{\tau}$ is

$$\sqrt{n}\Big(\bar{X}/(1-\bar{X}) - p/(1-p)\Big) \xrightarrow{d} N\Big(0, \frac{1}{(1-p)^4} \times p(1-p)) = N(0, \frac{p}{(1-p)^3}\Big)$$

Asymptotic Efficiency

- Asymptotic variance: different with limiting of variance, since we can ignore some extreme values with small probability
- Asymptotic efficiency: asymptotic variance achieves CRLB
- Asymptotic normality of MLE

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)}).$$

- Comparison with other estimators?

Asymptotic Relative Efficiency

Defination: Asymptotic Relative Efficiency

Suppose that two estimator W_n and V_n statisfy

$$\sqrt{n}(W_n - \tau(\theta_0)) \xrightarrow{d} N(0, \sigma_W^2)$$

$$\sqrt{n}(V_n - \tau(\theta_0)) \xrightarrow{d} N(0, \sigma_V^2)$$

then the asymptotic relative efficiency is defined as

$$ARE(W_n, V_n) = \frac{\sigma_V^2}{\sigma_W^2}$$

- When W_n and V_n have the same convergence rate, the ARE compares the efficiency of them.

Example

Example. Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} Poisson(\lambda)$.The MLE of λ is $\bar{X}$. Let $\tau = P(X_i = 0) = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

$$W_n = \frac{1}{n} \sum Y_i$$

Another estimator is the MLE $V_n = e^{-\bar{X}}$.Compare them.
Since $E[X_1] = \lambda, Var(X_1) = \lambda$, so according to CLT, there is

$$\sqrt{n}(\bar{X} - \lambda) \xrightarrow{d} N(0, \lambda).$$

According to the data method, we have

$$\sqrt{n}(V_n - e^{-\lambda}) \xrightarrow{d} N(0, e^{-2\lambda}\lambda).$$

Note that $Y_i \sim Bernoulli(e^{-\lambda})$, so

$$\sqrt{n}(W_n - e^{-\lambda}) \xrightarrow{d} N(0, e^{-\lambda}(1 - e^{-\lambda})).$$

So we have

$$ARE(W_n, V_n) = \frac{e^{-2\lambda}\lambda}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^{\lambda} - 1} \leq 1.$$

So, for most λ, MLE is more efficient.

Robustness

- MLE is efficient only if the model is right. It can be vary bad if the model is wrong

Example. Suppose $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\theta, \sigma^2)$. The MLE is $\hat{\theta}_n = \bar{X}_n$. Suppose, we have a perturbed model that $X_i \sim N(\theta, \sigma^2)$ with probability $1 - \delta$ and $X_i \sim f(x)$ with probability δ, where

$$f(x) = \frac{1}{\pi(x^2 + 1)}.$$

This is the Cauchy distribution, which is quite famous as an example that $E[X] = +\infty$.Therrfore, $Var(\bar{X}) = \infty$.
If we still apply $\bar{X}_n$, then the Cauchy distribution will destroy its good properties. However, for small δ, the median still keeps the same. On the other hand, if we consider the normal model as correct, then in the next slide we can show that $ARE(M_n, mle) = 0.64 < 1$, which indicates that MLE is better.

- Nonparametric estimation (say, the median) is a solution
- Even when the model is wrong, sometimes the MLE still provides some information

Robustness

Find the asymptotic distribution for the median of X, assuming the model is $N(\theta, 1)$. (Let $\sigma^2 = 1$ for simplicity. The result is the same for any σ^2)
For fixed a, Let $Y_i = I(X_i \leq \theta + a/\sqrt{n})$. Then $Y_i \sim Bernoulli(p_n)$, where

$$p_n = \Phi(\theta + a\sqrt{n}) = \Phi(\theta) + \frac{a}{\sqrt{n}}\phi(\theta) + o(n^{-1/2}) = \frac{1}{2} + \frac{a}{\sqrt{n}}\phi(\theta) + o(n^{-1/2})$$

Also, $\sum_i Y_i$ has mean np_n and standard $\sigma_n = \sqrt{np_n(1-p_n)}$.
Note that, $M_n \leq \theta + a/\sqrt{n} \quad if\ and\ only\ if \quad \sum Y_i \geq \frac{n+1}{2}$.
Then

$$P(\sqrt{n}(M_n - \mu) \leq a) = P(M_n \leq \theta + a/\sqrt{n}) = p(\sum Y_i \geq \frac{n+1}{2})$$
$$= P(\frac{\sum Y_i - np_n}{\sigma_n} \geq \frac{(n+1)/2 - np_n}{\sigma_n}).$$

Now, $\frac{(n+1)/2 - np_n}{\sigma_n} \to -2a\phi(\theta)$, and hence

$$P(\sqrt{n}(M_n - \mu) \leq a) \to P(N(0,1) \geq -2a\phi(\theta)) = P(\frac{N(0,1)}{2\phi(\theta)} \leq a),$$

so that $\sqrt{n}(M_n - \theta) \xrightarrow{d} N(0, \frac{1}{4\phi(\theta)^2})$,and $ARE(M_n, mle) = 0.64$.

Summary: Parameter Estimation

- Unbiased estimator
  - Improve an unbiased estimator: conditional expectation w.r.t. sufficient statistics; U-statistics
  - UMVUE
  - $Crámer - Rao\ Lower\ Bound$ for the UMVUE
  - score function (random with expentation 0); Fisher information(a function of θ);
- Mean squared error
- Multiple loss functions; risk functions
  - Maximum risk $\Rightarrow$ Minimax estimator
  - Bayes risk $\Rightarrow$ Bayes estimator. Bayes estimator is to minimize the posterior risk given the data points
  - Minimax estimator is closely related to Bayes estimator
- Asymptotic properties:
  - Consistency: MLE is always consistent
  - Efficiency: asymptotic variance achieves CRLB; asymptotic relative efficiency
  - Robustness