Chapter 5: Asymptotics of Likelihood Inference *

5 Asymptotics of Likelihood Inference

5.1 Notation & Basic Assumptions

- Suppose that $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ is identifiable (i.e., $\theta_1\neq\theta_2\Rightarrow P_{\theta_1}\neq P_{\theta_2}$) and dominated by a σ -finite measure μ. Let $f_\theta(x)=\frac{dP_\theta}{d\mu}(x)$. We'll consider asymptotics for inference based on the likelihood function $f_\theta(x)$ (a random function of θ). We focus on the iid (one sample) case.

  The basic one-observation model is $(\mathcal{X},\mathcal{B},P_\theta)$, where $\theta\in\Theta\subset\mathbb{R}^k$, and $f_\theta(x)=\frac{dP_\theta}{d\mu}(x)$ is the density for one observation.

- Notations:
  $X^n=(X_1,\ldots,X_n)$: iid observables through stage n
  $\mathcal{X}^n$: observation space (n -fold product space) for X^n
  $\mathcal{B}^n$: n -fold product σ -algebra on X^n
  $P_\theta\equiv P_\theta^n$: distribution of X^n (i.e., n -fold probability product)
  μ^n : dominating measure (n -fold product measure) on $\mathcal{X}^n$
  $f_\theta^n=\frac{dP_\theta^n}{d\mu^n}$: $f_\theta^n(x^n)=\prod_{i=1}^n f_\theta(x_i)$ for $x^n=(x_1,\ldots,x_n)$
  $L_n(\theta)=\log f_\theta^n(X^n)$: log-likelihood function $L_n(\theta)=\sum_{i=1}^n\log f_\theta(X_i)$

  Note: We mostly focus on convergence in probability (i.e., consistency) and convergence in distribution (i.e., asymptotic normality) in results to follow.

- Definition 90: A statistic $T:\mathcal{X}^n\to\Theta$ is called the maximum likelihood estimator (MLE) of θ if $T(x^n)$ maximizes $f_\theta^n(x^n)$ over $\theta\in\Theta$, or

$$f_{T(x^n)}^n(x^n)=\sup_{\theta\in\Theta}f_\theta^n(x^n)$$

  for all $x^n\in\mathcal{X}^n$ Note: For a fixed $x^n\in\mathcal{X}^n$, a value $\theta\in\Theta$ maximizing $f_\theta^n(x^n)$ is called a maximum likelihood estimate (MLE), even if there are other possible outcomes $y^n\in\mathcal{X}^n$ for which $f_\theta(y^n)$ cannot be maximized over θ. That is, a MLE (maximum likeli- hood estimator or estimate) need not always exist.

  - maximizer out of parameter space
  - density blow up to infinity

    A simple way to fix the unbounded likelihood problem is to replace continuous densities $f_\theta^n(x^n)$ with discrete probability masses so that the likelihood is bounded by 1.

    For example, divide the outcome region $\mathcal{X}$ (for a single observation X_i) into disjoint parts, say $\mathcal{D}_1,\ldots,\mathcal{D}_k$ for some $k\geq 1$, and determine $p_\theta(\mathcal{D}_j)=\int_{\mathcal{D}_j}f_\theta(x)d\mu(x)=P_\theta(X_1\in\mathcal{D}_j),j=1,\ldots,k$. Then, define the likelihood at $x^n=(x_1,\ldots,x_n)$ as

$$\prod_{i=1}^n\prod_{j=1}^k P_\theta(X_i\in\mathcal{D}_j)^{I[x_i\in\mathcal{D}_j]}=\prod_{j=1}^k[p_\theta(\mathcal{D}_j)]^{\sum_{i=1}^n I[x_i\in\mathcal{D}_j]}$$

- Definition 91: Suppose $\{T_n\}$ is a sequence of statistics with $T_n:\mathcal{X}^n\to\Theta\subset\mathbb{R}^k$

  1. $\{T_n\}$ is (weakly) consistent at $\theta_0\in\Theta$ if, for every $\epsilon>0$

$$P_{\theta_0}[\|T_n-\theta_0\|>\epsilon]\to 0\quad\text{as } n\to\infty$$

  2. $\{T_n\}$ is strongly consistent at $\theta_0\in\Theta$ if $T_n\to\theta_0$ as $n\to\infty$, a.s. (P_{θ_0}).

  (Weak) consistency of T_n at θ_0 means convergence of T_n to θ_0 in θ_0 -probability: $T_n\xrightarrow{p_{\theta_0}}\theta_0$ as $n\to\infty$

*STA643: Advanced Theory of Statistical Inference. Instructed by Dr. Daniel Nordman. Arranged by Zhiling Gu

It turns out that it is quite hard to show that an MLE exists and is consistent; see Theorem 7.49 and Lemma 7.54 of Schervish. We'll take an easier and much more common approach due to Cramér. Rather than consider maximizers of the likelihood, the basic idea is to instead focus on statements about how roots of the likelihood equation can exist and provide consistent estimation.

- Definition 92 : In the case where $\Theta \subset \mathbb{R}^k$ and f_θ is differentiable in $\theta=(\theta_1,\ldots,\theta_k)$, the likelihood equations are

$$\frac{\partial}{\partial\theta_i}L_n(\theta) \equiv \frac{\partial}{\partial\theta_i}\log f_\theta^n\left(x^n\right)=0, \quad i=1,2,\ldots,k$$

where $L_n(\theta)=\log f_\theta^n\left(x^n\right)$

Note: If f_θ is differentiable in θ and a MLE exists in the interior of Θ, then the MLE will satisfy the likelihood equations.

- Theorem 93: Suppose that $k=1$ (i.e., a real-valued parameter) and there exists an open neighborhood of θ_0, say O, such that

1. $f_\theta(x)>0$ for all $x \in \mathcal{X}$ and $\theta \in O$
2. for any $x \in \mathcal{X}$, $f_\theta(x)$ is differentiable at every $\theta \in O$, and
3. $\mathrm{E}_{\theta_0}\log f_\theta\left(X_1\right)$ exists for all $\theta \in O$ and $\mathrm{E}_{\theta_0}\log f_{\theta_0}\left(X_1\right)$ is finite.

Then, for any $\epsilon>0$ and $\delta>0$, there exists an $N \equiv N(\epsilon,\delta)$ such that for any $n \geq N$ $P_{\theta_0}\Big($ likelihood equation $\frac{d}{d\theta}L_n(\tilde{\theta})=0$ has a root $\tilde{\theta} \in\left[\theta_0-\epsilon,\theta_0+\epsilon\right]\Big)>1-\delta$

Proved based on WLLN combined with KL information $I(\theta_0,\theta)$. Define the average log-likelihood as $\bar{L}_n(\theta)=n^{-1}\sum_{i=1}^n\log f_\theta\left(X_i\right)$ and the difference

$$\Delta_n(\theta)=\bar{L}_n\left(\theta_0\right)-\bar{L}_n(\theta), \quad \theta \in \Theta$$

Note that

$$\mathrm{E}_{\theta_0}\Delta_n(\theta)=\mathrm{E}_{\theta_0}\log\left(\frac{f_{\theta_0}\left(X_1\right)}{f_\theta\left(X_1\right)}\right)=I\left(\theta_0,\theta\right)>0 \quad \theta \neq \theta_0$$

Pick/fix $\epsilon>0$ so that $\left[\theta_0-\epsilon,\theta_0+\epsilon\right] \subset O$(open set around θ_0). Then, by WLLN

$$\Delta_n(\theta+\epsilon) \xrightarrow{p_{\theta_0}} I\left(\theta_0,\theta+\epsilon\right)>0 \quad \& \quad \Delta_n(\theta-\epsilon) \xrightarrow{p_{\theta_0}} I\left(\theta_0,\theta-\epsilon\right)>0 \quad \text{as } n \rightarrow \infty$$

- Corollary 94: Under the assumptions of Theorem 93, suppose in addition that Θ is open and $f_\theta(x)$ is differentiable at every point $\theta \in \Theta$ so that

$$\frac{d}{d\theta}L_n(\theta) \equiv \frac{d}{d\theta}\log f_\theta^n\left(x^n\right)$$

makes sense at all $\theta \in \Theta$. Define ρ_n to be the root of the likelihood equation when there is exactly one; otherwise, adopt any definition for ρ_n.

If, with θ_0 -probability approaching 1, the likelihood equation has a single root (i.e, $\lim_{n\rightarrow\infty}P_{\theta_0}($ the likelihood equation has exactly one solution $)=1$), then

$$\rho_n \xrightarrow{p_{\theta_0}} \theta_0 \quad \text{as } n \rightarrow \infty$$

i.e., ρ_n converges to θ_0 in θ_0 -probability.

Proved by $A_n \equiv$ "likelihood equation has a root $\tilde{\theta}_n \in\left[\theta_0-\epsilon,\theta_0+\epsilon\right]$" $\& B_n \equiv$ "likelihood equation has one root". By Theorem 93, there exists $N_1 \equiv N_1(\epsilon,\delta)$ such that $n \geq N_1$ implies $P_{\theta_0}\left(A_n\right)>1-\delta/2$; also by assumption, there exists $N_2 \equiv N_2(\delta)$ such that $n \geq N_2$ implies $P_{\theta_0}\left(B_n\right)>1-\delta/2$ Hence, for $n \geq \max\left\{N_1,N_2\right\}$, $P_{\theta_0}\left(A_n \cap B_n\right)>1-\delta$. On the event $A_n \cap B_n$, we have $\rho_n=\tilde{\theta}_n \in\left[\theta_0-\epsilon,\theta_0+\epsilon\right]$; that is, $P_{\theta_0}\left(\left|\rho_n-\theta_0\right| \leq \epsilon\right)>1-\delta$ for $n \geq \max\left\{N_1,N_2\right\}$

- Note: Theorem 93 says that we're guaranteed (for large n) that some root/solution exists to the likelihood equations which is close to the true parameter θ_0. Corollary 94 says that, if the likelihood equation has just one solution for large n, then that solution must be consistent: that's because (for large n) the solution from Corollary 94 must correspond to the one given in Theorem 93 .

- Corollary 95: Under the assumptions of Theorem 93, if $\{T_n\}$ is a sequence of estimators consistent at θ_0 and $\hat{\theta}_n = \begin{cases} T_n & \text{if the likelihood equation has no roots,} \\ \text{the root of the likelihood equation closest to } T_n & \text{otherwise} \end{cases}$

  then $\hat{\theta}_n \xrightarrow{p_{\theta_0}} \theta_0$ as $n \to \infty$.

  Note: Corollary 95 says that, when faced with multiple roots, one strategy is just to select the root closest to another consistent estimator T_n. In practice, expect for the simplest cases, finding any root of the likelihood equation is a numerical problem. It is often difficult to know if any algorithm for solving the likelihood equation has converged or if a root we find is really the one closest to T_n.

- Another way for possibly improving a consistent estimator T_n is to use a one-step Newton improvement on T_n as follows. For $k=1$, write $L_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$ and the likelihood equation as

$$L'(\theta) = 0$$

  Treating a as an initial approximation to a root and assuming enough differentiability, we have

$$L_n'(\theta) \approx L_n'(a) + L_n''(a)(\theta - a)$$

  Setting $0 = L_n'(a) + L_n''(a)(\theta - a)$ and solving for θ gives

$$\theta = a - \frac{L_n'(a)}{L_n''(a)}$$

  provided that $L_n''(a) \neq 0$ This suggests a one-step Newton improvement on a consistent estimator T_n of $\theta \in \mathbb{R}$ as

$$\tilde{\theta}_n = T_n - \frac{L_n'(T_n)}{L_n''(T_n)}$$

  provided that $L_n''(T_n) \neq 0$

  For $k > 1$, the one-step Newton improvement on an estimator T_n of $\theta \in \mathbb{R}^k$ becomes

$$\tilde{\theta}_n = T_n - \left[L_n''(T_n)\right]^{-1} L_n'(T_n)$$

  for $L_n'(T_n)$ as the $k \times 1$ vector of first-order partial derivatives of $L_n(\theta)$ and $L_n''(T_n)$ as the $k \times k$ matrix of second-order partial derivatives of $L_n(\theta)$

  Note: It is often true that estimators of this type have asymptotic behaviors (including consistency and asymptotic normality) similar to those of real roots of the likelihood equations. See, for example, Schervish's development around his Theorem 7.75 for a more general (M-estimation) version of this.

- Theorem 96: Suppose that $k=1$ (i.e., a real-valued parameter) and there exists an open neighborhood of θ_0, say O, such that

  1. $f_\theta(x) > 0$ for all $x \in \mathcal{X}$ and $\theta \in O$
  2. for any $x \in \mathcal{X}$, $f_\theta(x)$ is three-times differentiable at every $\theta \in O$
  3. there exists $M(x) \geq 0$ with $\mathrm{E}_{\theta_0} M(X_1) < \infty$ and

$$\left| \frac{d^3}{d\theta^3} \log f_\theta(x) \right| \leq M(x)$$

     for all $x \in \mathcal{X}$ and $\theta \in O$
  4. $1 = \int f_\theta(x) d\mu(x)$ can be differentiated twice with respect to θ under the integral at θ_0
  5. $I_1(\theta) \in (0, \infty)$ for all $\theta \in O$.

  Then, if $\hat{\theta}_n$ is a root of the likelihood equation with θ_0 -probability approaching 1 and $\hat{\theta}_n \xrightarrow{p_{\theta_0}} \theta_0$, then under θ_0

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

  as $n \to \infty$, i.e., $P_{\theta_0}\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \leq x\right) \to \Phi\left(x\sqrt{I_1(\theta_0)}\right)$ as $n \to \infty$ for each $x \in \mathbb{R}$

  Prove by Taylor Explansion+CLT+WLLN+Slutsky's theorem.

- The following corollaries provide practical large-sample, Wald-type confidence limits for θ_0 :

  $\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{n I_1(\hat{\theta}_n)}}$ $\left(I_1\left(\hat{\theta}_n\right)\right.$ as (estimated) "expected Fisher information"

  $\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{-L_n''(\hat{\theta}_n)}}$ $\left(-n^{-1} L_n''\left(\hat{\theta}_n\right)\right.$ as "observed Fisher information"

- Corollary 97: Under the assumptions of Theorem 96, if $I_1(\theta)$ is continuous at θ_0 then under θ_0

$$\sqrt{nI_1\left(\hat{\theta}_n\right)}\left(\hat{\theta}_n-\theta_0\right) \xrightarrow{d} N(0,1) \quad \text{as } n\to\infty$$

- Corollary 98 : Under the assumptions of Theorem 96, then under θ_0

$$\sqrt{-L_n''\left(\hat{\theta}_n\right)}\left(\hat{\theta}_n-\theta_0\right) \xrightarrow{d} N(0,1) \quad \text{as } n\to\infty$$

  Note: For a sequence of estimators $\left\{\hat{\theta}_n^*\right\}$ with $\sqrt{n}\left(\hat{\theta}_n^*-\theta_0\right) \xrightarrow{d} N\left(0,V\left(\theta_0\right)\right)$ under θ_0, then

$$\frac{\frac{1}{I_1(\theta_0)}}{V\left(\theta_0\right)}$$

  is called the asymptotic efficiency of $\left\{\hat{\theta}_n^*\right\}$.

  Typically, this ratio is bounded by 1, implying that the asymptotic variance $1/I_1\left(\theta_0\right)$ of a likelihood root $\hat{\theta}_n$ (or MLE) is generally optimal (smallest).

  However, it is possible (even in regular problems) at some θ_0 to have an asymptotic efficiency which is larger than 1 (i.e., where $V\left(\theta_0\right)$ is strictly smaller than $1/I_1\left(\theta_0\right)$). Such points θ_0 are called a "super-efficiency point" of $\left\{\hat{\theta}_n^*\right\}$. There are theorems, though, that say we cannot have too many super-efficiency points of $\left\{\hat{\theta}_n^*\right\}$.

- Theorem 99: Under the assumptions 1-5 of Theorem 96 , suppose that, under θ_0, the sequence of estimators T_n is $\sqrt{n}$ -consistent for θ (i.e., meaning that $\sqrt{n}\left(T_n-\theta_0\right)$ is tight or bounded in θ_0 -probability). Define

$$\tilde{\theta}_n=T_n-\frac{L_n'\left(T_n\right)}{L_n''\left(T_n\right)}$$

  Then, under θ_0

$$\sqrt{n}\left(\tilde{\theta}_n-\theta_0\right) \xrightarrow{d} N\left(0,\frac{1}{I_1\left(\theta_0\right)}\right) \quad \text{as } n\to\infty$$

  Note: Versions of Corollaries $97-98$ also hold when $\hat{\theta}_n$ is replaced by $\tilde{\theta}_n$.

- For $\mathcal{P}=\left\{P_\theta:\theta\in\Theta\right\},\Theta_0\subset\Theta$ and testing $H_0:\theta\in\Theta_0$ vs. $H_1:\theta\in\Theta_1=\Theta\backslash\Theta_0$ consider statistics

$$LR(x)=\frac{\sup_{\theta\in\Theta_1} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)} \quad \text{or} \quad \lambda(x)=\max\{1,LR(x)\}=\frac{\sup_{\theta\in\Theta} f_\theta(x)}{\sup_{\theta\in\Theta_0} f_\theta(x)}$$

  and we reject H_0 for large values of $\lambda(x)$. Suppose that there is a MLE of θ, say $\hat{\theta}(x)$ where

$$\sup_{\theta\in\Theta} f_\theta(x)=f_{\hat{\theta}(x)}(x)$$

  in the numerator of $\lambda(x)$. This suggests the possibility of using "MLE-type" asymptotics to establish limiting distributions for likelihood ratio-type test statis- tics.

- Theorem 100: Under the assumptions of Theorem 96 (where $\hat{\theta}_n$ assumed to be a likelihood root that is consistent for θ_0) and letting

$$\Lambda_n\equiv 2\log\frac{f_{\hat{\theta}_n}\left(X^n\right)}{f_{\theta_0}\left(X^n\right)}=2\left(L_n\left(\hat{\theta}_n\right)-L_n\left(\theta_0\right)\right)$$

  then under θ_0

$$\Lambda_n \xrightarrow{d} \chi_1^2 \quad \text{as } n\to\infty$$

  Note: This again is the $k=1$ version. Similar results hold for $k>1$ with χ_k^2 limits. Note: For $H_0:\theta=\theta_0$, if $\hat{\theta}_n$ above is not only a consistent root of the likelihood equations but also a MLE, then

$$\Lambda_n=2\log\lambda\left(X^n\right)$$