7 Large Sample Theory

7.1 Convergence

Definition 1 (Convergence in distribution)**.** A sequence of random vectors X_n converges in distribution to X if $\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x)$ as $n \to \infty$ for all points x at which $x \mapsto \mathrm{P}(X \le x)$ is continuous. It is denoted by $X_n \stackrel{d}{\to} X$ or $X_n \stackrel{\mathcal{L}}{\to} X$.

Example 1. Example: Let X_n be uniform random variable on the set $\{1/n, 2/n, ..., n/n\}$. Then, for any fixed $x \in [0,1]$,

$$\mathrm{P}(X_n \le x) = \sum_{i=1}^{n} \frac{1}{n} \times 1\left(x \le \frac{i}{n}\right) \to x.$$

Convergence holds because

$$x = \int_0^x 1dx = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{n} \times 1\left(x \le \frac{i}{n}\right).$$

Hence, $X_n \stackrel{d}{\to} X = U(0,1)$ uniform distribution on $[0,1]$. So $\mathrm{P}(X_n \in A)$ converges to $\mathrm{P}(X \in A)$ for all A of the form $\{x : x \le a\}$.

Remark 1. If the densities converge as $f_n \to f$, then $\mathrm{P}(X_n \in A)$ converges to $\mathrm{P}(X \in A)$ for all Borel sets A. But $\mathrm{P}(X_n \in A)$ does not converge to $\mathrm{P}(X \in A)$ for all A. If $A = \{x : x \text{ is rational}\}$, then $\mathrm{P}(X_n \in A) = 1$ but $\mathrm{P}(X \in A) = 0$!

Remark 2. Convergence in distribution does not mean that $\mathrm{E}[X_n] \to \mathrm{E}[X]$. Suppose that

$$X_n \sim \left(1 - \frac{1}{n}\right) N(0,1) + \frac{1}{n} N\left(n^2, 1\right).$$

Then,

$$\mathrm{P}(X_n \le x) = \left(1 - \frac{1}{n}\right) \mathrm{P}\{N(0,1) \le x\} + \frac{1}{n} \mathrm{P}\left\{N\left(n^2,1\right) \le x\right\} \to \mathrm{P}\{N(0,1) \le x\}.$$

But $\mathrm{E}[X_n] = n$ and $\mathrm{E}[X] = 0$.

Lemma 1 (Portmanteau Lemma)**.** $X_n \stackrel{d}{\to} X$ *if and only if* $E[f(X_n)] \to E[f(X)]$ *for all bounded and continuous functions* f.

Let $d(x,y)$ be a distance function on $\mathbb{R}^k$. For example, we can consider the Euclidean distance

$$d(x,y) = \|x - y\|_2 = \sqrt{(x-y)^T (x-y)}.$$

Definition 2 (Convergence in Probability)**.** A sequence of random vectors X_n converges in probability to X if for every $\epsilon > 0$, $\mathrm{P}(d(X_n, X) > \epsilon) \to 0$ (or equivalently $\mathrm{P}(d(X_n, X) \le \epsilon) \to 1$) as $n \to \infty$. It is denoted by $X_n \stackrel{P}{\to} X$.

Example 2. Suppose that $X_1, ..., X_n$ form an iid sample from some distribution with finite variance. Then,

$$\mathrm{P}\left(\left|\bar{X} - \mathrm{E}[X]\right| \ge \epsilon\right) \le \frac{\mathrm{Var}\left[\bar{X}\right]}{\epsilon^2} = \frac{\mathrm{Var}[X]}{n\epsilon^2} \to 0.$$

Definition 3 (Convergence Almost Surely)**.** A sequence of random vectors X_n converges almost surely to X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) \quad = \quad 1.$$

It is denoted by $X_n \overset{a.s.}{\to} X$. An equivalent definition of convergence almost surely is that $X_n \overset{a.s.}{\to} X$ if and only if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(d\left(X_k, X\right) < \epsilon, \text{ for all } k \geq n\right) \quad = \quad 1.$$

Example 3. Let $Z \sim \text{Uniform}\,(0,1)$ and $X_n = 1\left(Z < n^{-1}\right)$. Then,

$$\left\{\lim_{n\to\infty} X_n = 0\right\} \quad = \quad \{Z > 0\}\,.$$

Hence,

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = 0\right) \quad = \quad \mathrm{P}\left(Z > 0\right) = 1,$$

that is $X_n \overset{a.s.}{\to} 0$.

Relations among the above types of convergence are included in the following theorem.

Theorem 1. *Basic relationships are as follows.*

1. $X_n \overset{a.s.}{\to} X$ *implies* $X_n \overset{P}{\to} X$.
2. $X_n \overset{P}{\to} X$ *implies* $X_n \overset{d}{\to} X$.
3. *If* $c \in \mathbb{R}^d$ *is a constant vector, then* $X_n \overset{d}{\to} c$ *implies* $X_n \overset{P}{\to} c$.

Example 4. Let $X \sim N\,(0,1)$. Let $X_1 = X$, $X_2 = -X$, $X_3 = X$, $X_4 = -X$, etc. Then X_i has the same distribution as X for any i. Hence, $X_n \overset{d}{\to} X$. But

$$\mathrm{P}\left(|X_n - X| > 1\right) \quad = \quad \begin{cases} \mathrm{P}\left(|X - X| > 1\right) = 0 & \text{n is odd} \\ \mathrm{P}\left(|-X - X| > 1\right) = \mathrm{P}\left(|X| > \frac{1}{2}\right) \approx 0.62. & \text{n is even} \end{cases}$$

Hence, X_n does not converge in probability to X.

Example 5. Let $Z \sim \text{Uniform}\,(0,1)$. Let $X_1 = 1$, $X_2 = 1\left(Z < \frac{1}{2}\right)$, $X_3 = 1\left(\frac{1}{2} \leq Z < 1\right)$, $X_4 = 1\left(Z < \frac{1}{4}\right)$, $X_5 = 1\left(\frac{1}{4} \leq Z < \frac{1}{2}\right)$, ... In general, if $n = 2^k + m$ for $k \geq 0$ and $0 \leq m < 2^k$, then

$$X_n \quad = \quad 1\left(\frac{m\,(n)}{2^{k(n)}} \leq Z < \frac{m\,(n) + 1}{2^{k(n)}}\right),$$

and $\mathrm{P}\left(X_n = 1\right) = \frac{1}{2^{k(n)}}$. Hence,

$$\mathrm{P}\left(|X_n - 0| \geq \epsilon\right) \quad = \quad \mathrm{P}\left(X_n = 1\right) = \frac{1}{2^{k(n)}} \to 0.$$

But, for any $Z < 1$, X_n does not converge to any value since we just move the interval $\frac{m}{2^k} \leq Z < \frac{m+1}{2^k}$ from 0 to 1. Hence, we don't have $X_n \overset{a.s.}{\to} 0$.

Theorem 2 (Continuous Mapping Theorem)**.** *Let* $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ *be continuous at every point of a set* C *such that* $P\,(X \in C) = 1$. *Then,*

1. *If* $X_n \overset{d}{\to} X$, *then* $g\,(X_n) \overset{d}{\to} g\,(X)$.
2. *If* $X_n \overset{P}{\to} X$, *then* $g\,(X_n) \overset{P}{\to} g\,(X)$.
3. *If* $X_n \overset{a.s.}{\to} X$, *then* $g\,(X_n) \overset{a.s.}{\to} g\,(X)$.

7.2 Consistency of MLE

Let X_1, ..., X_n be iid with density $p(x \mid \theta)$. The log-likelihood is

$$\ell(\theta) = \log L(\theta) \quad = \quad \sum_{i=1}^{n} \log p(x_i \mid \theta).$$

Lemma 2. *Let $\{p(\cdot \mid \theta) : \theta \in \Theta\}$ be a collection of densities such that the corresponding probability measures satisfy*

$$P_\theta \neq P_{\theta_0} \text{ for every } \theta \neq \theta_0 \quad \text{(identification).}$$

Then

$$-KL\left(p(X \mid \theta_0),\ p(X \mid \theta)\right) \quad = \quad E\left[\log\left(\frac{p(X \mid \theta)}{p(X \mid \theta_0)}\right) \mid \theta_0\right]$$

attains its maximum uniquely at θ_0.

Proof. It is a direct consequence of Kullback-Leibler inequality in Estimation chapter. □

Theorem 3 (Consistency of MLE: Unidimensional)**.** *Suppose that*

1. *the distributions P_θ of the observations are distinct,*
2. *the observations are i.i.d. with probability density $p(\cdot \mid \theta)$ with respect to some measure μ,*
3. *the distributions P_θ have common support so that $\{x : p(x \mid \theta) > 0\}$ is independent of θ,*
4. *the parameter space Θ contains an open set ω of which the true parameter value θ_0 is an interior point,*
5. *for almost all x, $p(x \mid \theta)$ is differentiable with respect to θ in ω.*

Then, with probability tending to 1 as $n \to \infty$, the likelihood equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log p(X_i \mid \theta)}{\partial \theta} \quad = \quad 0$$

has a root $\hat{\theta}_n$ such that $\hat{\theta}_n$ tends to θ_0 in probability.

Proof. Let $\mathcal{N}(\theta_0, \epsilon)$ be a open interval with the center θ_0 and ϵ as radius. By Assumption 4, we can find a small enough $\epsilon > 0$ such that $\mathcal{N}(\theta_0, \epsilon) \subset \Theta$. Define $x = (x_1, ..., x_n)$, and

$$S_n \quad = \quad \left\{x : \sum_{i=1}^{n} \log p(x_i \mid \theta_0) > \sum_{i=1}^{n} \log p(x_i \mid \theta_0 - \epsilon) \text{ and } \sum_{i=1}^{n} \log p(x_i \mid \theta_0) > \sum_{i=1}^{n} \log p(x_i \mid \theta_0 + \epsilon)\right\}.$$

By Jensen's inequality,

$$\mathrm{E}\left[\log\left(\frac{p(X \mid \theta)}{p(X \mid \theta_0)}\right) \mid \theta_0\right] \quad \leq \quad \log\left(\mathrm{E}\left[\frac{p(X \mid \theta)}{p(X \mid \theta_0)} \mid \theta_0\right]\right) = \log\left(\int p(x \mid \theta)\, dx\right) = 0,$$

where the equality holds if and only if $\theta = \theta_0$ [Assumption 1]. By LLN,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(x_i \mid \theta)}{p(x_i \mid \theta_0)}\right) \quad \overset{P}{\to} \quad \mathrm{E}\left[\log\left(\frac{p(X \mid \theta)}{p(X \mid \theta_0)}\right) \mid \theta_0\right] < 0, \quad \forall \theta \neq \theta_0,$$

that is, for any $\delta > 0$,

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(x_i \mid \theta)}{p(x_i \mid \theta_0)}\right) - \mathrm{E}\left[\log\left(\frac{p(X \mid \theta)}{p(X \mid \theta_0)}\right) \mid \theta_0\right]\right| < \delta\right) \quad \to \quad 1.$$

Since the limit is negative, then

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p(x_i \mid \theta)}{p(x_i \mid \theta_0)}\right) < 0\right) = \mathrm{P}\left(\sum_{i=1}^{n}\log p(x_i \mid \theta_0) > \sum_{i=1}^{n}\log p(x_i \mid \theta)\right) \quad \to \quad 1.$$

Consequently, $\mathrm{P}(S_n) \to 1$. Because $p(x \mid \theta)$ is differentiable with respect to θ in $\mathcal{N}(\theta_0, \epsilon)$ by Assumption 5, then there must exist a local maximum so that $n^{-1}\sum_{i=1}^{n}\frac{\partial \log p(X_i|\theta)}{\partial\theta} = 0$. Hence, for any small $\epsilon > 0$, there exists a sequence $\hat{\theta}_n(\epsilon)$ such that $\mathrm{P}\left(\left|\hat{\theta}_n(\epsilon) - \theta_0\right| < \epsilon\right) \to 1$. To eliminate the dependence of the estimator on ϵ, we simply choose $\hat{\theta}_n$ that is closest to θ_0 for each n. The resulting sequence $\hat{\theta}_n^*$ satisfies $\mathrm{P}\left(\left|\hat{\theta}_n^* - \theta_0\right| < \epsilon\right) \to 1$. $\square$

Example 6. Consider a random sample X_1, ..., X_n from $N(\theta, 1)$ with density

$$p(x \mid \theta) \quad = \quad \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta)^2}{2}\right\}, \quad x > 0, \theta \in \mathbb{R}.$$

The MLE is $\hat{\theta} = \bar{X}$.

1. We can show directly from the definition of consistency that

$$\mathrm{P}\left(\left|\bar{X} - \theta\right| < \epsilon\right) \quad = \quad \mathrm{P}\left(-\sqrt{n}\epsilon < \frac{\bar{X} - \theta}{\sqrt{1/n}} < \sqrt{n}\epsilon\right) \to 1.$$

2. We can also verify the conditions of the theorem. For example, for the identification assumption, if we have $\theta_1 \neq \theta_2$ such that $p(x \mid \theta_1) = p(x \mid \theta_2)$ for almost all x. Then, we must have

$$-\frac{(x-\theta_1)^2}{2} \quad = \quad -\frac{(x-\theta_2)^2}{2},$$

which means that θ_1 must be the solution of

$$\theta_1^2 - 2x\theta_1 - \theta_2^2 + 2x\theta_2 \quad = \quad 0.$$

If $x = 0$, we must have $\theta_1^2 = \theta_2^2$. But when $\theta_1 = -\theta_2$, we have

$$\theta_2^2 - 2x(-\theta_2) - \theta_2^2 + 2x\theta_2 \quad = \quad 4x\theta_2,$$

which cannot be zero unless $x = 0$. Thus, we must have $\theta_1 = \theta_2$.

Example 7 (Inconsistency of MLE: Ferguson's Example)**.** Let X_1, ..., X_n be with probability θ i.i.d. Uniform $(-1, 1)$, and be with probability $1 - \theta$ i.i.d. with a triangular distribution with pdf

$$\frac{1}{c(\theta)}\left(1 - \frac{|x-\theta|}{c(\theta)}\right), \text{ for } |x - \theta| \leq c(\theta),$$

where $c(\theta)$ is a continuous and decreasing function in θ with $c(0) = 1$ and $0 < c(\theta) \leq 1 - \theta$ for $0 < \theta < 1$. The parameter space is a compact set $\Theta = [0, 1]$. If $c(\theta) \to 0$ sufficiently fast as $\theta \to 1$, then $\hat{\theta}_n \overset{a.s.}{\to} 1$ whatever be the true value of $\theta \in \Theta$.

- In this exmaple, there is no common support, since the triangular distribution part depends on the parameter.

- The inconsistency arises because the likelihood explodes if $c(\theta)$ is too small. Only one observation is enough to make it explode.

Theorem 4 (Strong Consistency of MLE)**.** *Suppose that X_1, ..., X_n are i.i.d., and satisfy*

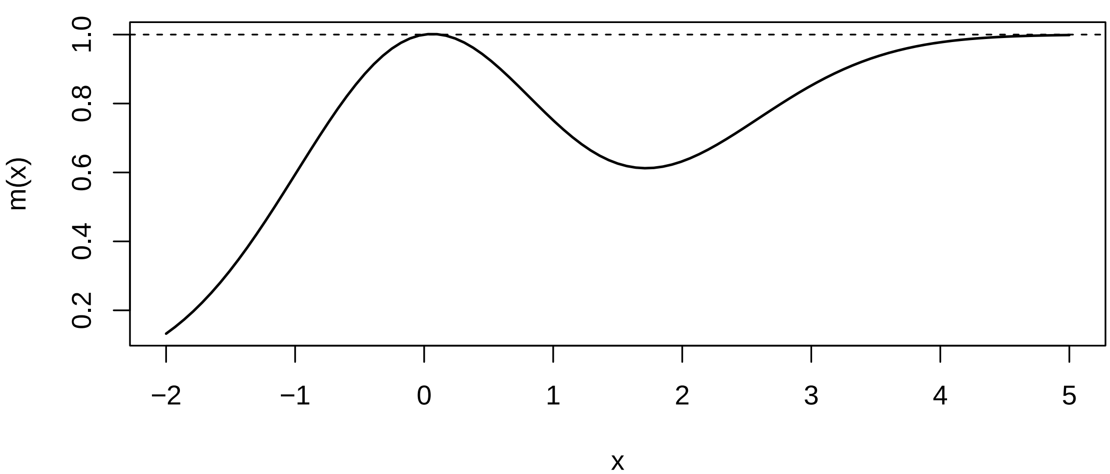
1. *The parameter space Θ is a compact set,*
2. *The density $p(x \mid \theta)$ with respect to μ is continuous in θ for all x,*
3. *There exists a function $K(x)$ such that $E[|K(X)| \mid \theta_0] < \infty$ and*
$$U(x,\theta) = \log p(x \mid \theta) - \log p(x \mid \theta_0) \le K(x),$$
for all x and θ,
4. *for all $\theta \in \Theta$ and sufficiently small $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} p(x \mid \theta')$ is measurable in x,*
5. *$p(x \mid \theta) = p(x \mid \theta_0)$ almost everywhere with respect to μ implies $\theta = \theta_0$.*

Then, any sequence of maximum likelihood estimator $\hat{\theta}_n$ of θ satisfies
$$\hat{\theta}_n \overset{a.s.}{\to} \theta.$$

In a general case, let $\hat{\theta}_n$ be an M-estimator that maximizes $M_n(\theta)$. We want our estimator to be consistent: $d\left(\hat{\theta}_n, \theta_0\right) \overset{P}{\to} 0$, where $d(\cdot,\cdot)$ is a distance function. Heuristically,

1. Pointwise convergence: for every θ, $M_n(\theta) \overset{P}{\to} M(\theta)$, where $M_n(\theta)$ is a random function and $M(\theta)$ is a deterministic function,
2. $\hat{\theta}_n = \arg\sup_{\theta} M_n(\theta)$,
3. $\theta_0 = \arg\sup_{\theta} M(\theta)$.

Then it is reasonable to expect $\hat{\theta}_n \overset{P}{\to} \theta_0$. However, pointwise convergence is too weak. An illustration is as follows.



Theorem 5 (Consistency of General M-Estimator)**.** *Let M_n be random functions and let M be a fixed function of θ such that for every $\epsilon > 0$,*

$$\begin{aligned}\sup_{\theta\in\Theta}\left|M_n(\theta)-M(\theta)\right| &\xrightarrow{P} 0, \quad \textit{(uniform convergence)}\\ \sup_{\theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) &< M(\theta_0). \quad \textit{(well-separated function)}\end{aligned}$$

Then, any sequence of estimators $\hat{\theta}_n$ with

$$M_n\left(\hat{\theta}_n\right) \geq M_n(\theta_0) - o_P(1) \quad \textit{(nearly maximizer)}$$

converges in probability to θ_0.

Proof. The uniform convergence assumption implies $M_n(\theta_0) \xrightarrow{P} M(\theta_0)$, or equivalently

$$M(\theta_0) - o_{\mathrm{P}}(1) \leq M_n(\theta_0) \leq M(\theta_0) + o_{\mathrm{P}}(1).$$

Hence, by the nearly maximizer assumption,

$$M_n\left(\hat{\theta}_n\right) \geq M_n(\theta_0) - o_{\mathrm{P}}(1) \geq M(\theta_0) - o_{\mathrm{P}}(1).$$

This implies that

$$\begin{aligned}M(\theta_0) - M\left(\hat{\theta}_n\right) &\leq M_n\left(\hat{\theta}_n\right) + o_{\mathrm{P}}(1) - M\left(\hat{\theta}_n\right)\\ &\leq \sup_{\theta\in\Theta}\left|M_n(\theta)-M(\theta)\right| + o_{\mathrm{P}}(1) \xrightarrow{P} 0,\end{aligned}$$

where the convergence holds by the uniform convergence assumption.

The well-separated function assumption implies that, there exists $\eta(\epsilon) > 0$ such that for any $\theta \in \{\theta : d(\theta,\theta_0) \geq \epsilon\}$, we have $M(\theta) < M(\theta_0) - \eta$. If $\hat{\theta}_n \in \{\theta : d(\theta,\theta_0) \geq \epsilon\}$, then we have $M\left(\hat{\theta}_n\right) < M(\theta_0) - \eta$ and

$$\mathrm{P}\left(d\left(\hat{\theta}_n,\theta_0\right) \geq \epsilon \mid \theta_0\right) \leq \mathrm{P}\left(M\left(\hat{\theta}_n\right) < M(\theta_0) - \eta \mid \theta_0\right) = \mathrm{P}\left(M(\theta_0) - M\left(\hat{\theta}_n\right) > \eta \mid \theta_0\right) \to 0,$$

where the convergence holds since we have shown above that $M(\theta_0) - M\left(\hat{\theta}_n\right) \xrightarrow{P} 0$. □

Remark 3. If θ_0 is a unique maximizer of $M(\theta)$, then the well-separated function assumption means that the supremum is only attained at θ_0. If $M_n\left(\hat{\theta}_n\right) \geq \sup_\theta M_n(\theta) - o_{\mathrm{P}}(1)$, then $\hat{\theta}_n$ nearly maximizes M_n. But $M_n\left(\hat{\theta}_n\right) \geq M_n(\theta_0) - o_{\mathrm{P}}(1)$ is enough for consistency.

Theorem 6 (Consistency of General Z-Estimator)**.** *Let Ψ_n be random functions and let Ψ be a fixed function of θ such that for every $\epsilon > 0$,*

$$\sup_{\theta\in\Theta}\left\|\Psi_n(\theta)-\Psi(\theta)\right\| \xrightarrow{P} 0, \quad \textit{(uniform convergence)}$$

$$\inf_{\theta:d(\theta,\theta_0)\geq\epsilon}\left\|\Psi(\theta)\right\| > 0 = \left\|\Psi(\theta_0)\right\|. \quad \textit{(well separated function)}$$

Then, any sequence of estimators $\hat{\theta}_n$ with

$$\Psi_n\left(\hat{\theta}_n\right) = o_P(1) \quad \textit{(nearly a zero)}$$

converges in probability to θ_0.

We often do not need so strong assumptions as in the general case.

Lemma 3 (Consistency of Z-Estimator: Weaker Assumption)**.** *Let Θ be a subset of the real line and let Ψ_n be random functions and let Ψ be a fixed function of θ such that*

$$\Psi_n(\theta) \quad \xrightarrow{P} \quad \Psi(\theta) \text{ for every } \theta. \quad \text{(pointwise convergence)}$$

Assume that each map $\theta \mapsto \Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$, or is nondecreasing with $\Psi_n\left(\hat{\theta}_n\right) = o_P(1)$. Let θ_0 be a point such that $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$ for every $\epsilon > 0$. Then, $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. There are two sets of assumptions. We prove them separately.

1. Because $\Psi_n(\theta)$ is a continuous function with a unique zero at $\hat{\theta}_n$, then if we know $\Psi_n(\theta_0 - \epsilon) < 0 < \Psi_n(\theta_0 + \epsilon)$, the solution must be between $\theta_0 - \epsilon$ and $\theta_0 + \epsilon$. This meant that

$$\mathrm{P}\left(\Psi_n(\theta_0 - \epsilon) < 0 < \Psi_n(\theta_0 + \epsilon)\right) \quad = \quad \mathrm{P}\left(\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\right).$$

By the assumption of pointwise convergence, we have $\Psi_n(\theta_0 - \epsilon) \xrightarrow{P} \Psi(\theta_0 - \epsilon)$ and $\Psi_n(\theta_0 + \epsilon) \xrightarrow{P} \Psi(\theta_0 + \epsilon)$. Thus, together with the assumption $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$, we further get

$$\mathrm{P}\left(\Psi_n(\theta_0 - \epsilon) < 0 < \Psi_n(\theta_0 + \epsilon)\right) \quad \to \quad 1.$$

Hence,

$$\mathrm{P}\left(\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\right) \quad \to \quad 1.$$

2. If we indeed have $\Psi_n\left(\hat{\theta}_n\right) = 0$, then a nondecreasing $\Psi_n(\theta)$ means that $\Psi_n(\theta_0 - \epsilon) \le 0 \le \Psi_n(\theta_0 + \epsilon)$ implies $\theta_0 - \epsilon \le \hat{\theta}_n \le \theta_0 + \epsilon$. We can use the same reasoning as in the first assumption set. Next we consider $0 \ne \Psi_n\left(\hat{\theta}_n\right) = o_{\mathrm{P}}(1)$. For a $\eta > 0$, the nondecreasing $\Psi_n(\theta)$ means that

$$\begin{matrix} \Psi_n(\theta_0 - \epsilon) < -\eta \\ \hat{\theta}_n \le \theta_0 - \epsilon \end{matrix} \qquad \text{implies} \qquad \Psi_n\left(\hat{\theta}_n\right) \le \Psi_n(\theta_0 - \epsilon) < -\eta.$$

Since $\Psi_n\left(\hat{\theta}_n\right) = o_{\mathrm{P}}(1)$, then $\mathrm{P}\left(\Psi_n\left(\hat{\theta}_n\right) < -\eta\right) \to 0$. If $\Psi_n(\theta_0 - \epsilon) < -\eta$ and $\Psi_n(\theta_0 + \epsilon) > \eta$, then $\hat{\theta}_n$ must satisfy $\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon$ since $\Psi_n(\theta)$ is nondecreasing. Thus,

$$\mathrm{P}\left(\Psi_n(\theta_0 - \epsilon) < -\eta \text{ and } \Psi_n(\theta_0 + \epsilon) > \eta\right) \quad \le \quad \mathrm{P}\left(\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\right).$$

By the assumption $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$ and pointwise convergence, we have

$$\mathrm{P}\left(\Psi_n(\theta_0 - \epsilon) < -\eta \text{ and } \Psi_n(\theta_0 + \epsilon) > \eta\right) \quad \to \quad 1$$

for sufficiently small η.

$\square$

Example 8 (Consistency of Median)**.** The sample median of continuous random variable is a zero of the map

$$\theta \quad \mapsto \quad \Psi_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{sign}(X_i - \theta).$$

- By LLN, we have pointwise convergence:

$$-\frac{1}{n}\sum_{i=1}^{n}\text{sign}\left(X_i-\theta\right) \quad \overset{P}{\to} \quad \Psi\left(\theta\right)=-\text{E}\left[\text{sign}\left(X-\theta\right)\right]=\text{P}\left(X<\theta\right)-\text{P}\left(X>\theta\right).$$

- $\Psi_n\left(\theta\right)$ is a nondecreasing function and $\Psi_n\left(\hat{\theta}_n\right)=0$.

- The last assumption becomes

$$\begin{aligned} & \text{P}\left(X<\theta_0-\epsilon\right)-\text{P}\left(X>\theta_0-\epsilon\right)<0<\text{P}\left(X<\theta_0+\epsilon\right)-\text{P}\left(X>\theta_0+\epsilon\right), \\ \text{or} \quad & 2\text{P}\left(X<\theta_0-\epsilon\right)+\text{P}\left(X=\theta_0-\epsilon\right)<1<2\text{P}\left(X<\theta_0+\epsilon\right)+\text{P}\left(X=\theta_0+\epsilon\right). \end{aligned}$$

  It holds for example if we have a continuous random variable and the population median is unique, i.e., $\text{P}\left(X<\theta_0-\epsilon\right)<0.5<\text{P}\left(X<\theta_0+\epsilon\right)$.

7.3 Asymptotic Normality

Theorem 7 (Cramré-Rao Conditions for MLE: Univariate)**.** *Suppose that X_1, ..., X_n are i.i.d., and satisfy*

1. *The parameter space Θ is an open set such that the true parameter value θ_0 is an interior point,*

2. *The distributions P_θ have common support $A=\{x : p\left(x \mid \theta\right)>0\}$,*

3. *For every $x\in A$, the density $p\left(x\mid\theta\right)$ is three times differentiable with respect to θ, and the third derivative is continuous in θ,*

4. *The integral $\int p\left(x\mid\theta\right)d\mu\left(x\right)$ can be twice differentiable under the integral sign,*

5. *The Fisher information $\mathcal{I}\left(\theta\right)$ satisfies $0<\mathcal{I}\left(\theta\right)<\infty$,*

6. *There exists a function $M\left(x\right)$ such that*

$$\left|\frac{\partial^3\log p\left(x\mid\theta\right)}{\partial\theta^3}\right| \quad \leq \quad M\left(x\right)$$

   for all $x\in A$ and θ in a neighborhood of θ_0, and that $E\left[M\left(X\right)\mid\theta_0\right]<\infty$.

Then, any consistent sequence $\hat{\theta}_n$ of roots of the likelihood equation satisfies

$$\sqrt{n}\left(\hat{\theta}_n-\theta\right) \quad \overset{d}{\to} \quad N\left(0,\ \mathcal{I}^{-1}\left(\theta\right)\right).$$

Proof. Let $\ell\left(\theta\right)=\sum_{i=1}^{n}\log p\left(x_i\mid\theta\right)$ be the log-likelihood. By the third assumption, we get the Taylor's theorem:

$$0=\frac{d\ell\left(\hat{\theta}_n\right)}{d\theta} \quad = \quad \frac{d\ell\left(\theta_0\right)}{d\theta}+\left(\hat{\theta}_n-\theta_0\right)\frac{d^2\ell\left(\theta_0\right)}{d\theta^2}+\frac{1}{2}\left(\hat{\theta}_n-\theta_0\right)^2\frac{d^3\ell\left(\theta_n^*\right)}{d\theta^3},$$

where θ_n^* lies between θ_0 and $\hat{\theta}_n$. Thus,

$$\sqrt{n}\left(\hat{\theta}_n-\theta_0\right) \quad = \quad \frac{n^{-1/2}\frac{d\ell(\theta_0)}{d\theta}}{-\frac{1}{n}\frac{d^2\ell(\theta_0)}{d\theta^2}-\frac{1}{2n}\left(\hat{\theta}_n-\theta_0\right)\frac{d^3\ell(\theta_n^*)}{d\theta^3}},$$

provided that the denominator is nonzero.

1. The numerator satisfies

$$n^{-1/2}\frac{d\ell(\theta_0)}{d\theta} \quad = \quad \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{d\log p(x_i \mid \theta_0)}{d\theta} - 0\right] \xrightarrow{d} N(0, \mathcal{I}(\theta_0)),$$

by CLT since

$$\mathrm{E}\left[\frac{d\log p(x_i \mid \theta_0)}{d\theta} \mid \theta_0\right] = 0, \qquad \mathrm{Var}\left[\frac{d\log p(x_i \mid \theta_0)}{d\theta} \mid \theta_0\right] = \mathcal{I}(\theta_0)$$

by Assumptions 1, 2, 4, 5.

2. Regarding the first term in denominator,

$$\frac{1}{n}\frac{d^2\ell(\theta_0)}{d\theta^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{d^2\log p(x_i \mid \theta_0)}{d\theta^2} \quad \xrightarrow{P} \quad \mathrm{E}\left[\frac{d^2\log p(x_i \mid \theta_0)}{d\theta^2} \mid \theta_0\right] = -\mathcal{I}(\theta_0),$$

by LLN and Assumptions 1, 2, 4, 5.

3. Regarding the second term in denominator,

$$\left|\frac{1}{n}\frac{d^3\ell(\theta_n^*)}{d\theta^3}\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{d^3\log p(x_i \mid \theta_n^*)}{d\theta^3}\right| \quad \leq \quad \frac{1}{n}\sum_{i=1}^{n}\left|\frac{d^3\log p(x_i \mid \theta_n^*)}{d\theta^3}\right| \leq \frac{1}{n}\sum_{i=1}^{n}M(x_i),$$

by Assumption 6, if θ_n^* is in a small neighborhood of θ_0. Our assumptions 1, 2, 3 implies that $\hat{\theta}_n \xrightarrow{P} \theta_0$ by Theorem 3. Hence, the probability that θ_n^* is in a small neighborhood of θ_0 approaches to 1. Note that by assumption $\mathrm{E}[M(X) \mid \theta_0] < \infty$, then LLN implies that

$$\frac{1}{n}\sum_{i=1}^{n}M(x_i) \quad \xrightarrow{P} \quad \mathrm{E}[M(X) \mid \theta_0],$$

that is,

$$\mathrm{P}\left(\mathrm{E}[M(X) \mid \theta_0] - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n}M(x_i) \leq \mathrm{E}[M(X) \mid \theta_0] + \epsilon\right) \quad \to \quad 1.$$

Hence,

$$\mathrm{P}\left(\left|\frac{1}{n}\frac{d^3\ell(\theta_n^*)}{d\theta^3}\right| \leq \mathrm{E}[M(X) \mid \theta_0] + \epsilon\right) \quad \geq \quad \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}M(x_i) \leq \mathrm{E}[M(X) \mid \theta_0] + \epsilon\right) \to 1,$$

in other words, $\frac{1}{n}\frac{d^3\ell(\theta_n^*)}{d\theta^3}$ is bounded in probability. This means that

$$\frac{1}{n}\left(\hat{\theta}_n - \theta_0\right)\frac{d^3\ell(\theta_n^*)}{d\theta^3} \quad = \quad o_{\mathrm{P}}(1).$$

Thus, we have reached

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \quad = \quad \frac{n^{-1/2}\frac{d\ell(\theta_0)}{d\theta}}{\mathcal{I}(\theta_0) + o_{\mathrm{P}}(1)} = \frac{n^{-1/2}\frac{d\ell(\theta_0)}{d\theta}}{\mathcal{I}(\theta_0)} + o_{\mathrm{P}}(1).$$

□

Example 9. Consider $p(x \mid \theta) = \theta^{-1} \exp\left\{-\theta^{-1} x\right\}$, $\theta > 0$. The second-order derivative is

$$\frac{\partial^2 p(x \mid \theta)}{\partial \theta^2} = \left(\frac{x^2}{\theta^5} - \frac{4x}{\theta^4} + \frac{2}{\theta^3}\right) \exp\left\{-\frac{x}{\theta}\right\},$$

and $\int \frac{\partial^2 p(x|\theta)}{\partial \theta^2} dx = 0$. In fact, this distribution belongs to the exponential family. Hence, we can change the order of integration and differentiation. Note that

$$\frac{\partial \log p(x \mid \theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{x}{\theta^2}, \qquad \frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2} = \frac{1}{\theta^2} - \frac{2x}{\theta^3}.$$

Then, the Fisher information is

$$\operatorname{var}\left(\frac{\partial \log p(x \mid \theta)}{\partial \theta}\right) = \frac{1}{\theta^2} \in (0, \infty), \quad \text{or} \quad -\mathrm{E}\left(\frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^2}\right) = \frac{1}{\theta^2}.$$

The theorem above obtains the MLE by solving

$$\sum_{i=1}^{n} \frac{d \log p(x_i \mid \theta)}{d\theta} = 0.$$

Hence, it can be viewed as a Z-estimator. The following theorem considers the general Z-estimator under classic conditions.

Theorem 8 (Normality of Z-Estimator: Classic Condition)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d., and consider $\Psi_n(\theta) = n^{-1} \sum_{i=1}^{n} \psi_\theta(X_i)$. Suppose that*

1. *For each θ in an open subset of Euclidean space, let $\theta \mapsto \psi_\theta(x)$ be twice continuously differentiable for every x.*
2. *Suppose that $E[\psi_{\theta_0}(X_1) \mid \theta_0] = 0$, $E\left[\|\psi_{\theta_0}(X_1)\|^2 \mid \theta_0\right] < \infty$ and that the matrix $E\left[\frac{d\psi_{\theta_0}(X_1)}{d\theta} \mid \theta_0\right]$ exists and is nonsingular.*
3. *Assume that the second-order partial derivatives are dominated by a fixed integrable function $\phi(x)$ for every θ in a neighborhood of θ_0.*

Then every consistent estimator sequence $\hat{\theta}_n$ such that $\Psi_n\left(\hat{\theta}_n\right) = 0$, the sequence $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ is asymptotically normal with mean zero and covariance matrix

$$\left(E\left[\frac{d\psi_{\theta_0}(X_1)}{d\theta^T} \mid \theta_0\right]\right)^{-1} E\left[\psi_{\theta_0}(X_1) \psi_{\theta_0}^\top(X_1) \mid \theta_0\right] \left\{\left(E\left[\frac{d\psi_{\theta_0}(X_1)}{d\theta \mathring{A}T} \mid \theta_0\right]\right)^{-1}\right\}^\top.$$