# Large Sample Theory PartII

**Song Xi Chen, Xiaojun Song**

Department of Business Statistics and Econometrics
Center for Statistical Science
Peking University

August 12, 2023

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be iid with distribution $F_\theta$ belonging to a family $\mathcal{F} = \left\{F_\theta : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta\right\}$ and suppose that the distribution $F_\theta$ posses densities $f_\theta(x)$. The likelihood function of the sample $\mathbf{X}$ is defined as

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^n f_{\boldsymbol{\theta}}(X_i).$$

1. The maximum likelihood estimate (MLE) is given by
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \log L(\boldsymbol{\theta}; \mathbf{X}) \ .$$
2. Often, the MLE $\hat{\boldsymbol{\theta}}$ may be obtained by solving a system of likelihood score equations,
$$\left.\frac{\partial \log L(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_j}\right|_{\theta=\hat{\boldsymbol{\theta}}} = 0,\ j = 1, 2, \cdots, k.$$
3. The variance of the score function is crucial for the AN of MLE.

# Fisher Information

## Definition 7.1

Suppose that $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is dominated by a $\sigma$-finite $\mu$. Say $\mathcal{P}$ is Fisher-Information (FI) regular at $\theta \in \Theta$, if there exists an open neighborhood of $\theta$, say $\Theta_\theta$, s.t.

(i) $f_\theta(x) := \frac{dP_\theta(x)}{d\mu} > 0$ for any $x$ and $\theta \in \Theta_\theta$.

(ii) $\forall\, x$, $f_\theta(x)$ is differentiable at $\theta$.

(iii) $\int f_\theta(x)\mu(dx)$ can be differentiable under the integral at $\theta$, i.e.

$$\int \frac{d}{d\theta'} f_{\theta'}(x)\Big|_{\theta'=\theta} \mu(dx) = 0.$$

## Definition 7.2

If a model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is FI regular, then

$$I_1(\theta) = \mathrm{E}_\theta \left[ \frac{d}{d\theta'} \log f_{\theta'}(x)\Big|_{\theta'=\theta} \right]^2$$

is called the FI in $X$ at $\theta$.

# Maximum Likelihood Estimate

## *Remark* 1

(i) *By def. of FI, we have* $\mathrm{E}_\theta\left[\frac{d}{d\theta'}\log f_{\theta'}(X)\big|_{\theta'=\theta}\right]=0$, *so:*

$$I_n(\theta)=\mathrm{var}\left(\frac{d}{d\theta'}\log f_{\theta'}(x)\bigg|_{\theta'=\theta}\right)$$

(ii) *If* $\Theta\subset\mathbb{R}^K$ *for* $K>1$, $\theta=(\theta_1,\cdots,\theta_K)$, *then:*

$$\frac{d}{d\theta'}\log f_{\theta'}(x)=\begin{pmatrix}\frac{d}{d\theta'_1}\log f_{\theta'}(x)\\ \vdots\\ \frac{d}{d\theta'_K}\log f_{\theta'}(x)\end{pmatrix}\in\mathbb{R}^K$$

*and*

$$I_n(\theta)=\mathrm{E}_\theta\left[\frac{d}{d\theta'}\log f_{\theta'}(x)\left(\frac{d}{d\theta'}\log f_{\theta'}(x)\right)^\top\bigg|_{\theta'=\theta}\right]$$

*is the FI matrix.*

# Maximum Likelihood Estimate

If $\mathcal{P}$ is FI regular at $\theta$, and $\forall x$, $f_\theta(x)$ is twice differentiable at $\theta$, and $1 = \int f_\theta(x)\mu(dx)$ can be differentiable w.r.t. $\theta$ under the integral, i.e.

$$\int \frac{d}{d\theta'} f_{\theta'}(x)\bigg|_{\theta'=\theta} \mu(dx) = 0, \quad \int \frac{d^2}{d\theta'^2} f_{\theta'}(x)\bigg|_{\theta'=\theta} \mu(dx) = 0$$

Then,

$$I(\theta) = -\mathrm{E}_\theta \left[ \frac{d^2}{d\theta'^2} \log f_{\theta'}(x)\bigg|_{\theta'=\theta} \right]$$

The proof is evident.

Let $(\mathbb{X}, \mathcal{X}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$ be a p.s. of a r.v. $X$, where $\mathcal{P} \ll$ a $\sigma$-finite $\mu$, $f_\theta(x) = \frac{dP_\theta}{d\mu}$. Suppose that:

(i) $\Theta \subset \mathbb{R}$ is open.

(ii) $A =$ support of $f_\theta$ does not depend on $\Theta$.

(iii) $\forall\, \theta \in \Theta$, $\frac{df_\theta(x)}{d\theta}$ exists.

(iv) $\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta} \log f_\theta(x)\right] = \int \frac{\partial f_\theta(x)}{\partial\theta}\mu(dx) = 0$ for any $\theta \in \Theta$.

(v) $I_n(\theta) > 0$ for any $\theta \in \Theta$.

(vi) $g : \Theta \to \mathbb{R}$ measurable and $\frac{dg(\theta)}{d\theta}$ exists for any $\theta \in \Theta$, and $\hat{g} : \mathbb{X} \to \Theta$ is an unbiased estimator of $g(\theta)$.

(vii) $\frac{d}{d\theta} \int \hat{g}(x) f_\theta(x)\mu(dx) = \int \hat{g}(x)\frac{df_\theta(x)}{d\theta}\mu(dx)$

Then, $\mathrm{var}_\theta(\hat{g}(x)) \geq [g'(\theta)]^2 / I_n(\theta)$ or $\mathrm{var}_\theta(\hat{g}(x)) \geq [g'(\theta)]^\top I_n^{-1}(\theta)[g'(\theta)]$ for multivariate case.

# C-R Lower Bound

## *Remark 2*

(i) $[g'(\theta)]^2/I_n(\theta)$ *is the C-R Lower Bound for unbiased estimator of* $g(\theta)$.

(ii) *Condition (iv) and (vii) are the most restrictive, they can be established under a set of sufficient conditions.*

## Lemma 7.3

*Under the conditions (i)-(iii) in above slides , and if there exists a* $G : \mathbb{X} \times \Theta \to \mathbb{R}$, *s.t.*

(a) $\forall\, \theta \in \Theta$, $G(x, \theta)$ *is* $\mathcal{X}$*-measurable.*

(b) $\mathrm{E}_\theta G^2(x, \theta) < \infty$ *for any* $\theta \in \Theta$.

(c) $\forall\, \theta \in \Theta$, $\exists \epsilon_\theta > 0$, *s.t.*

$$\left| \frac{df_{\theta'(x)}}{d\theta'} \right| \leq G(x, \theta) f_\theta(x), \quad \forall x \in A \text{ and } |\theta - \theta'| < \epsilon_\theta.$$

*then Condition (iv) is satisfied; and for all unbiased estimator of* $g(\theta)$, *say* $\hat{g}(x)$, *if* $\mathrm{E}_\theta(\hat{g}(x))^2 < \infty$, *then Condition (vii) is valid as well.*

# Proof

(Use MVT & DCT):

$\forall\, \theta \in \Theta$, $|\theta - \theta'| < \epsilon_\theta$, $\theta' \in \Theta$, as

$$\int_{\mathcal{X}} f_\theta(x)\mu(dx) = \int_{\mathcal{X}} f_{\theta'}(x)\mu(dx) = 1$$

so,

$$\int_{\mathcal{X}} \frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'}\mu(dx) = 0 \tag{1}$$

From the MVT, Condition (iii), and Condition (c):

$$\left|\frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'}\right| = \left|\frac{df_{\tilde{\theta}}(x)}{d\tilde{\theta}}\right| \le G(x,\theta)f_\theta(x) \tag{2}$$

for some $\tilde{\theta}$ between $\theta$ and $\theta'$.

# Proof

Note that $\int_{\mathcal{X}} G(x,\theta) f_\theta(x)\mu(dx) = \mathrm{E}_\theta G(X,\theta) \le \mathrm{E}_\theta^{1/2} G^2(X,\theta) < \infty$, by DCT,

$$\int_{\mathcal{X}} \frac{df_\theta(x)}{d\theta}\mu(dx) = \int_{\mathcal{X}} \lim_{\theta' \to \theta} \frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'}\mu(dx)$$
$$= \lim_{\theta' \to \theta} \int_{\mathcal{X}} \frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'}\mu(dx) = 0$$

which exactly is the Condition (iv).

On the other hand, suppose $\hat{g}(x)$ is an unbiased estimator of $g(\theta)$ satisfying $\mathrm{E}_\theta \hat{g}^2(x) < \infty$, then:

$$\int_{\mathcal{X}} \hat{g}(x) \frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'}\mu(dx) = \frac{g(\theta) - g(\theta')}{\theta - \theta'} \tag{3}$$

# Proof

From (2), $\forall\, \theta, \theta'$, s.t. $|\theta - \theta'| < \epsilon_\theta$, we have:

$$\left| \hat{g}(x) \frac{f_\theta(x) - f_{\theta'}(x)}{\theta - \theta'} \right| \leq |\hat{g}(x)| G(x, \theta) f_\theta(x)$$

and:

$$\begin{aligned} \int_{\mathcal{X}} |\hat{g}(x)| G(x, \theta) f_\theta(x) \mu(dx) &= \mathrm{E}_\theta |\hat{g}(x)| G(x, \theta) \\ &\leq \left[ \mathrm{E}_\theta \hat{g}^2(x) \mathrm{E}_\theta G^2(x, \theta) \right]^{1/2} < \infty \end{aligned}$$

as $\mathrm{E}_\theta \hat{g}^2(x) < \infty$ and $\mathrm{E}_\theta G^2(x, \theta) < \infty$. Applying DCT on (3) by letting $\theta' \to \theta$, then we get Condition (vii).

# Bhattacharya Inequality: C-R Bound is too low.

## Theorem 7.4

*Suppose the Condition (i) and (ii) in Slide 6. Now, if we give more restrictions on other conditions:*

(iii)* $\frac{\partial^i f_\theta(x)}{\partial \theta^i}$ *exists and* $\int_{\mathcal{X}} \frac{\partial^i f_\theta(x)}{\partial \theta^i} \mu(dx) = 0$, $i = 1, \cdots, K$, $\theta \in \Theta$.

(iv)* $\int_{\mathcal{X}} \frac{1}{f_\theta(x)} \left( \frac{\partial^i f_\theta(x)}{\partial \theta^i} \right)^2 \mu(dx) < \infty$, $i = 1, \cdots, K$, $\theta \in \Theta$.

(v)* $\hat{g}(x)$ *is an unbiased estimator of* $g(\theta)$ *with finite variance, and for any* $i = 1, \cdots, K$, $\theta \in \Theta$,

$$g^{(i)}(\theta) = \frac{\partial^i}{\partial \theta^i} g(\theta) = \int_{\mathcal{X}} \hat{g}(x) \frac{\partial^i f_\theta(x)}{\partial \theta^i} \mu(dx)$$

*Then,* $\mathrm{var}_\theta(\hat{g}(x)) \geq \tilde{g}^\top(\theta) V^{-1}(\theta) \tilde{g}(\theta)$, *where* $V(\theta) = (V_{ij}(\theta))$ *with*

$$V_{ij}(\theta) = \mathrm{E}_\theta \left[ \frac{1}{f_\theta^2(x)} \frac{\partial^i f_\theta(x)}{\partial \theta^i} \frac{\partial^j f_\theta(x)}{\partial \theta^j} \right], \quad \tilde{g}(\theta) = \left( g'(\theta), \cdots, g^{(K)}(\theta) \right)^\top$$

# Proof

Denote $S = S_\theta(x) = (S_\theta^{(1)}(x), \cdots, S_\theta^{(K)}(x))^T$, where:

$$S_\theta^{(i)}(x) = \frac{1}{f_\theta(x)} \frac{\partial^i f_\theta(x)}{\partial \theta^i}$$

From Condition (iii)*, $\mathrm{E}_\theta S = 0$, from Condition (iv)*, $\mathrm{var}_\theta(S) = V(\theta)$, and from Condition (v)*, $\mathrm{cov}_\theta \left( \hat{g}(x), S_\theta^{(i)}(x) \right) = g^{(i)}(\theta)$, hence,

$$A := \mathrm{var}_\theta \begin{pmatrix} \hat{g} \\ S \end{pmatrix} = \begin{pmatrix} \mathrm{var}_\theta(\hat{g}(x)) & \tilde{g}^\top(\theta) \\ \tilde{g}(\theta) & V(\theta) \end{pmatrix}$$

Since $|A| \geq 0$, and

$$|A| = |V(\theta)| \left[ \mathrm{var}_\theta(\hat{g}(x)) - \tilde{g}^\top(\theta) V^{-1}(\theta) \tilde{g}(\theta) \right]$$

which implies $\mathrm{var}_\theta(\hat{g}(x)) - \tilde{g}^\top(\theta) V^{-1}(\theta) \tilde{g}(\theta) \geq 0$.

**Remark 3**

*Bhattacharya Inequality is an extension of C-R Inequality ($K = 1$)!*

# Kullback-Leibler divergence

Kullback-Leibler divergence is a measure on the closeness between two distributions $P_\theta$ and $P_\eta$.

**Definition 7.5 (KL-divergence)**

The Kullback-Leibler (KL) divergence of two probability measure from $P_\theta$ to $P_\eta$

$$D_{KL}(P_\eta \| P_\theta) = -\mathrm{E}_\eta \log \frac{p_\theta}{p_\eta}(\boldsymbol{X}), \quad \boldsymbol{X} \sim P_\eta$$

where $p_\theta, p_\eta$ are the density functions of $P_\theta$ and $P_\eta$ respectively.

- The K-L- divergence is not a true metric, as

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P) \text{ in general.}$$

- By concavity of the log, $D_{KL}(P \parallel Q) \geq 0$ and $= 0$ iff $P = Q$ if the models are identifiable.

# Identifiability

Suppose that we have an i.i.d. samples $X_1, \ldots, X_n \sim X$ where $X$ has probability measure $P_\theta$ dominated by a underlying measure $\mu$ with density $f_\theta(x)$.

**Definition 7.6 (Identifiability)**

A parametric famility (i.e. a class of prob. densities) $\mathbb{P}_\Theta := \{f_\theta(x) : \theta \in \Theta\}$ is identifiable if $\forall \theta_1 \neq \theta_2 (\theta_1, \theta_2 \in \Theta)$, we have

$$\mu\left(x : f_{\theta_1}(x) \neq f_{\theta_2}(x)\right) > 0$$

where $\mu$ is the dominated measure (Lesbegue or counting measure).

- Identifiable parametric famility means no other parameter gives the same probability distribution.
- Identifiability is a sufficient condition in the Consistency of MLE. If the parameter is not identifiable, then consistent estimators cannot exist.

# Cramer's Consistency Condition

## Lemma 7.7 (Minimizing the K-L distance)

*Let $\mathbb{P}_\Theta := \{f_\theta(x) : \theta \in \Theta\}$ be a identifiable parametric famility. If $E_{\theta_0} \log f_{\theta_0}(X) < \infty$, then $M(\theta) := E_{\theta_0} \log[f_\theta / f_{\theta_0}(X)]$, attains its maximum uniquely at its true parameter $\theta_0$, i.e.*

$$E_{\theta_0} \log f_\theta(X) \leq E_{\theta_0} \log f_{\theta_0}(X) < \infty.$$

For $\theta \in \Theta$, since $-\log(t)$ is strictly convex, Jensen¡¯s inequality shows that

$$E_{\theta_0} \log \frac{f_\theta}{f_{\theta_0}}(X) \leq \log E_{\theta_0} \frac{f_\theta}{f_{\theta_0}}(X) = 0.$$

By identifiable condition, the equality holds iff $\theta = \theta_0$. Thus the expected log-likilihood is the largest at the true parameter $\theta_0$.

# Consistency of MLE

## Theorem 7.8

*Let $X_1, \cdots, X_n$ i.i.d. $P_\theta$, $\Theta \subset \mathbb{R}$ and there exists an open neighborhood of $\theta$, say $\Theta_\theta$, s.t.*

(i) $A := \{x | f_\theta(x) > 0\}$ *does not depend on* $\theta$.

(ii) $\forall\, x \in A$, $f_\theta(x)$ *is differentiable at every* $\theta' \in \Theta_\theta$.

(iii) $\mathrm{E}_\theta \log f_{\theta'}(X)$ *exists for all all* $\theta' \in \Theta$, *and is finite.*

(iv) $\mu(x | f_{\theta_1}(x) \neq f_{\theta_2}(x)$ *for* $\theta_1 \neq \theta_2) > 0$, *i.e.* $\mathcal{P} = \{P_\theta\}$ *is identifiable.*

*Then,* $\forall\, \epsilon > 0$, $\delta > 0$, $\exists\, m_{\epsilon,\delta} > 0$, *s.t.* $n > m_{\epsilon,\delta}$ *satisfying:*

$$\mathrm{P}_\theta\{\text{the equation } \frac{d}{d\theta'} \sum_{i=1}^{n} \log f_{\theta'}(X_i) = 0 \text{ has a root within } (\theta - \epsilon, \theta + \epsilon)\} \geq 1 - \delta$$

# Some Remarks

**Remark 4**

(1) *As* $X_i \sim P_\theta$, $\theta$ *is the true parameter. The log likelihood is:*

$$\ell_n(\theta') = \sum_{i=1}^{n} \log f_{\theta'}(X_i)$$

(2) *In (i) and (ii), we can require the properties are still true for any* $x \in \mathcal{X}$ *and* $\theta' \in \Theta$, *which may be more convenience to verify.*

# Proof

WLOG, we assume $\epsilon$ is small enough s.t. $[\theta-\epsilon, \theta+\epsilon] \subset \Theta_\theta$. Note WLLN & (iii):

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\theta\pm\epsilon}(X_i)}{f_\theta(X_i)} \xrightarrow{P_\theta} \mathrm{E}_\theta \log \frac{f_{\theta\pm\epsilon}(X)}{f_\theta(X)} := -\eta_{\theta\pm\epsilon} < 0$$

So $\forall\, \delta > 0$, $\xi > 0$, $\exists\, m = m_{\epsilon,\delta}$, $\forall\, n > m$,

$$\mathrm{P}_\theta \left\{ \left| \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_{\theta\pm\epsilon}(X_i)}{f_\theta(X_i)} + \eta_{\theta\pm\epsilon} \right| < \xi \right\} \geq 1 - \frac{\delta}{2}$$

By choosing $0 < \xi < \min\{\eta_{\theta-\epsilon}, \eta_{\theta+\epsilon}\}$, the above display implies for any $n > m$, we have:

$$\mathrm{P}_\theta(A) := \mathrm{P}_\theta \left( \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i) > \frac{1}{n}\sum_{i=1}^{n} \log f_{\theta+\epsilon}(X_i) \right) \geq 1 - \frac{\delta}{2}$$

$$\mathrm{P}_\theta(B) := \mathrm{P}_\theta \left( \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i) > \frac{1}{n}\sum_{i=1}^{n} \log f_{\theta-\epsilon}(X_i) \right) \geq 1 - \frac{\delta}{2}$$

# Proof

As $\mathrm{P}(AB) = \mathrm{P}(A) - \mathrm{P}(AB^{\mathsf{C}}) \geq \mathrm{P}(A) - \mathrm{P}(B^{\mathsf{C}}) \geq 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta$, we have:

$$\mathrm{P}_\theta\left(\ell_n(\theta - \epsilon) < \ell_n(\theta) \text{ and } \ell_n(\theta + \epsilon) < \ell_n(\theta)\right) \geq 1 - \delta$$

Since $\ell_n(\theta')$ is differentiable,

$$\mathrm{P}_\theta\left(\exists \text{ a local maximum of } \ell_n(\theta') \text{ on } (\theta - \epsilon, \theta + \epsilon)\right) \geq 1 - \delta$$

which actually implies:

$$\mathrm{P}_\theta\left\{\frac{d}{d\theta'}\ell_n(\theta') = 0 \text{ has a root on } (\theta - \epsilon, \theta + \epsilon)\right\} \geq 1 - \delta$$

**Remark 5**

*The root guaranteed by this Theorem is NOT necessary a MLE!*

# Consistency of MLE

**Theorem 7.9**

*Under the conditions of theorem 7.8, define $\hat{\theta}_n$ be the root of the likelihood equation when there is exactly one root (otherwise adopt any definition for $\hat{\theta}_n$). If*

$$\lim_{n\to\infty} \mathrm{P}_\theta\,(\textit{the likelihood equation has a single root}) = 1 \tag{4}$$

*then:*

$$\hat{\theta}_n \xrightarrow{P_\theta} \theta$$

# Proof

For any $\epsilon > 0$ and $\delta > 0$, Theorem 7.8 implies $\exists\, m_{\epsilon,\delta}$, s.t. $\forall\, n > m_{\epsilon,\delta}$,

$$\mathrm{P}_\theta(A) := \mathrm{P}_\theta\,(\text{the LE has a root within } (\theta-\epsilon, \theta+\epsilon)) \geq 1 - \frac{\delta}{2}$$

On the other hand, the extra condition in the Theorem implies $\exists\, m'_\delta$, $\forall\, n > m'_\delta$:

$$\mathrm{P}_\theta(B) := \mathrm{P}_\theta(\text{the LE has a single root}) \geq 1 - \frac{\delta}{2}$$

So as long as $n \geq \max\{m_{\epsilon,\delta}, m'_\delta\}$, we have

$$\mathrm{P}_\theta\left(|\hat{\theta}_n - \theta| < \epsilon\right) = \mathrm{P}_\theta(\hat{\theta}_n \text{ is in } (\theta-\epsilon, \theta+\epsilon)) = \mathrm{P}_\theta(AB) \geq 1 - \delta$$

*Remark 6*

*There is no guarantee that the LE essentially has a single root, i.e. (4), this condition, however, has already an consistent estimator.*

# Asymptotic Normality of MLE

## Theorem 7.10

*Let $X_1, \cdots, X_n$ i.i.d. $P_{\theta_0}$, $\Theta \subset \mathbb{R}$, and there exists an open neighborhood of $\theta_0$, say $\Theta_0$, s.t.*

(i) $f_{\theta'}(x) > 0$ *for all* $x$ *and* $\theta' \in \Theta_{\theta_0}$.

(ii) $\forall x$, $f_{\theta'}(x)$ *is 3-times diffferentiable at* $\forall \theta' \in \Theta_{\theta_0}$.

(iii) $\exists M(x) \geq 0$ *with* $\mathrm{E}_{\theta_0} M(x) < \infty$ *and* $\left| \frac{d^3}{d\theta'^3} \log f_{\theta'}(x) \right| \leq M(x)$, $\forall x$, $\theta' \in \Theta_{\theta_0}$.

(iv) $\int \frac{d^l}{d\theta'^l} f_{\theta'}(x) \Big|_{\theta'=\theta_0} = 0$ *for* $l = 1, 2$. *i.e.* $\int f_{\theta'}(x)\mu(dx) = 1$ *can be differentiable twice w.r.t.* $\theta$ *under the integral at* $\theta$.

(v) $\forall \theta'$, $0 < I_1(\theta') < \infty$ *where* $I_1$ *is the FI based on single observations* $X_1$.

*Let* $\hat{\theta}_n$ *is the MLE of* $\theta$. *Furthermore, we require:*

(vi) $\lim_{n\to\infty} \mathrm{P}_\theta \left( \hat{\theta}_n \text{ is a root of the LE } \right) = 1$ *and* $\mathrm{E}_\theta |\log f_{\theta'}(x)| < \infty$ *for any* $\theta' \in \Theta$.

(vii) $\hat{\theta}_n \xrightarrow{p.} \theta_0$ *and* $\mu\{x | f_\theta(x) = f_{\theta'}(x), \theta \neq \theta'\} = 0$. *Then*

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

# Proof

Let $L_n(\theta') = n^{-1} \sum_{i=1}^{n} \log f_{\theta'}(X_i)$ and:

$$0 = L_n'(\hat{\theta}_n) = L_n'(\theta_0) + L_n''(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2} L_n'''(\theta_1)(\hat{\theta}_n - \theta_0)^2 \tag{5}$$

where $\theta_1$ is between $\hat{\theta}_n$ and $\theta_0$. In writing (5), we note that Condition (vi) and (vii) implies that: $\exists, m, \forall\, n > m$, $\hat{\theta}_n$ is bothe a root of the LE and an element of $\Theta_0$, i.e.

$$\lim_{n \to \infty} \mathrm{P}_\theta \left( \hat{\theta}_n \text{ is the root of the LE } \& \ \hat{\theta}_n \in \Theta_0 \right) = 1 \tag{6}$$

On the other hand, by the WLLN,

$$L_n''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_\theta(X_i) \xrightarrow{P_\theta} -I_1(\theta_0) \overset{\text{(v)}}{\in} (-\infty, 0)$$

So,

$$L_n''(\theta_0) = -I_1(\theta_0) + o_{P_\theta}(1) \tag{7}$$

# Proof

Besides,

$$|L_n'''(\theta_1)| = \frac{1}{n}\left|\sum_{i=1}^{n}\frac{d^3}{d\theta^3}\log f_\theta(X_i)\right|_{\theta=\theta_1} \overset{\text{(iii)}}{\leq} \frac{1}{n}\sum_{i=1}^{n}M(X_i)$$

$$\xrightarrow[WLLN]{P_\theta} \mathrm{E}_{\theta_0}M(X) < \infty.$$

So $\{L_n'''(\theta_1)\}$ is tight, i.e. $L_n'''(\theta_1) = O_{P_\theta}(1)$.

Note that $\hat{\theta}_n \xrightarrow{p.} \theta_0$ as hypothesized, $\hat{\theta}_n - \theta_0 = o_{P_\theta}(1)$, hence,

$$(\hat{\theta}_n - \theta_0)^2 L_n'''(\theta_1) = o_{P_\theta}(\hat{\theta}_n - \theta_0). \tag{8}$$

# Proof

From (5) - (8),

$$0 = L_n'(\theta_0) + \left(-I_1(\theta_0) + o_{P_\theta}(1)\right)(\hat{\theta}_n - \theta_0) + o_{P_\theta}(\hat{\theta}_n - \theta_0).$$

As,

$$\sqrt{n}L_n'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f_\theta(X_i)}{\partial \theta} \xrightarrow{d.} N(0, I_1(\theta_0)),$$

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &= -I_1^{-1}(\theta_0)\sqrt{n}L_n'(\theta_0) + o_{P_\theta}\left(\sqrt{n}(\hat{\theta}_n - \theta_0)\right) \\ &= -I_1^{-1}(\theta_0)\sqrt{n}L_n'(\theta_0) + o_{P_\theta}(1) \xrightarrow{d.} N(0, I_1^{-1}(\theta_0)). \end{aligned}$$

# AN of MLE with compact and convex parameter space

## Theorem 7.11 (Theorem 5.9 in Bijma&Jonker&Van der Vaart(2017))

*Suppose that*

- *The* $\Theta$ *is compact and convex and that* $\theta$ *is identifiable, and let* $\hat{\theta}_n$ *be the maximum likelihood estimator based on a sample of size n from the distribution with (marginal) probability density* $p_\theta$*;*
- *Assume that the map* $\vartheta \mapsto \log p_\vartheta(x)$ *is continuously differentiable for all* $x$*, with derivative* $\ell_\vartheta(x)$ *such that* $|\ell_\vartheta(x)| \leq L(x)$ *for every* $\vartheta \in \Theta$*, where* $L$ *is a function with* $\mathrm{E}_\theta L^2(X_1) < \infty$ *;*
- *If* $\theta$ *is an interior point of* $\Theta$ *and the function* $\vartheta \mapsto I(\vartheta)$ *is continuous and positive.*

*Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, I^{-1}(\theta_0)\right).$$

Bijma, F., Jonker, M., & Van der Vaart, A. (2017). An introduction to mathematical statistics. Amsterdam University Press.

We study the consistency and asymptotic normality of $M$-estimators (proposed by Peter J. Huber) and Z-estimators . MLEs and ME are treated as the special cases of $M$ and $Z$-estimators, respectively.

- Suppose that the parameter $\theta$ (or "functional") of interests attached to the distribution of observations $\boldsymbol{X}_n := (X_1, \ldots, X_n) \sim f_\theta(\boldsymbol{X})$.

**Definition 8.1 ($M$-estimator)**

The $M$-estimator is to find an estimator $\hat{\theta}_n := \hat{\theta}_n (X_1, \ldots, X_n)$ that maximizes a random **criterion function** of the type

$$\theta \mapsto M_n(\theta)$$

For example, $M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta (X_i)$.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 73-101.

# The $Z$-estimators

Often the maximizing value is sought by setting a derivative (or gradient) equal to zero. So, the $Z$-estimators satisfies the **estimating equations**

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta\left(X_i\right) = 0 \tag{9}$$

**Definition 8.2 ($Z$-estimator)**

More generally, the $Z$-estimator is to find an estimator $\hat{\theta}_n$ that solves the the **estimating equations** (9).

- The $M$- and $Z$-Estimators does not require iid or independent structure of the observations.
- The minimization problem for function $-M_n(\theta)$ may be non-convex.
- The $Z$-estimator is often numerically solved by (quasi-) Newton methods, Gradient descent, Stochastic Gradient descent (non-convex).

# Examples: parameter est. from dist. (Location)

## MLE&PMLE

Let $X_1, \ldots, X_n \sim p_\theta$. Then the MLEs maximize the likelihood $\prod_{i=1}^n p_\theta(X_i)$ or equivalently the log-likelihood: $M_n(\theta) := \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$.
**Pseudo-MLE**: $X_i$'s may be dependent, the log-likelihood is still used.

## Two examples of Location estimators

The **sample mean** and **sample median** which are Z-estimators solved by

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n (X_i - \theta) = 0; \quad \text{and} \quad \Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \operatorname{sign}(X_i - \theta) = 0$$

- **Quantile function** for the distribution function $F$:

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

- **Median**: $\mathrm{med}(X) = F^{-1}(0.5)$.
- **Quantiles**: $\theta_0 = \arg\min_\theta \mathrm{E}\rho_\tau(x - \theta) = F^{-1}(\tau)$, (HW).

# Examples: Quantiles

Define the **check function**: $\rho_\tau(y) = y(\tau - \mathrm{I}_{(y<0)})$ as the loss function. Then the $\tau$-**sample quantile** $\hat{\theta}$ can seen as the $M$- and $Z$-estimators.

## Sample quantile

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(X_i - \theta); \text{ and}$$

$$\Psi_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \left((1-\tau)1\left\{X_i < \theta\right\} - \tau 1\{X_i > \theta\}\right) = 0$$

For small sample size $n$, vdv's book gives an alternative def. of the $\tau$-sample quantile: $\hat{\theta}$ solves the inequalities :$-1 < n\Psi_n(\theta) < 1$.

# Examples: Huber estimators

The Huber estimators were motivated by studies in robust statistics concerning the influence of extreme data points on the estimate.

## Huber estimators

Corresponding to the Huber $\Psi$ functions

$$\psi(x) = [x]_{-k}^{k} := \begin{cases} -k & \text{if } x \leq -k \\ x & \text{if } |x| \leq k \\ k & \text{if } x \geq k \end{cases}$$

The **Huber estimators** solves the following estimating equations.

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi\left(X_i - \theta\right) = 0$$

The Huber estimators behave more like the mean (large $k$) or more like the median (small $k$) and thus fill in the gap between the nonrobust mean and very robust median.

# Two Pictures for $Z$-estimator of local parameter

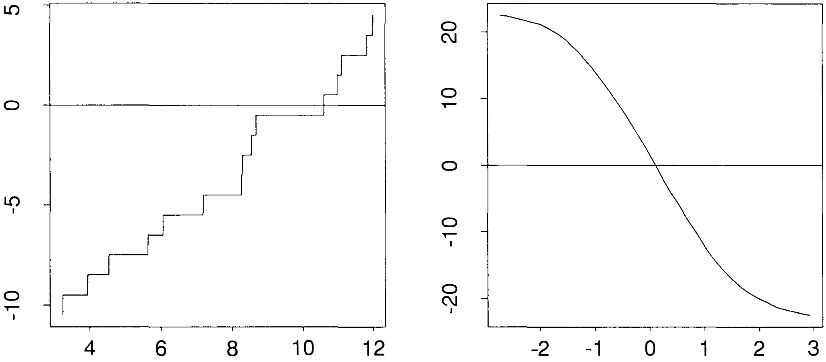


Figure: The functions $\theta \mapsto \Psi_n(\theta)$ for the 80% sample quantile and the Huber estimator from the $\mathrm{gamma}(8, 1)$ and standard normal distribution, respectively. $n = 15$.

# Consistency of $M$-Estimator

The estimator $\hat{\theta}_n$ is used to estimate the parameter $\theta$ aiming at : $\hat{\theta}_n \xrightarrow{P} \theta$, where $\theta \in \Theta$ endowed with metric $d$.

Suppose that the $\hat{\theta}_n$ maximizes the **random** (empirical) criterion function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} M_n(\theta) = \arg\min_{\theta \in \Theta} -M_n(\theta).$$

where $-M_n(\theta) =: L(P_n, P)$ can be seen as the empirical loss function.

## Definition 8.3 (True parameter)

The $\theta_0$ is usually defined as the maximization of the **deterministic** (true) criterion function: $M(\theta) =: \mathrm{E} m_\theta(X)$

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathrm{E} m_\theta(X) = \arg\min_{\theta \in \Theta} \mathrm{E} - m_\theta(X).$$

We wish to prove that $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathrm{P}} 0$ under some regularity conditions

$$M_n(\theta) \xrightarrow{\mathrm{P}} M(\theta), \quad \text{every } \theta.$$

by LLN. The convergence above is not uniformly for $\theta \in \Theta$!

# Consistency of $M$-estimator

Given an arbitrary random function $\theta \mapsto M_n(\theta)$, consider estimators $\{\hat{\theta}_n\}$ satisfies the **nearly maximization condition**:

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - o_P(1) \geq M_n(\theta_0) - o_P(1).$$

Example: $-M_n(\theta)$ is strongly convex. [iff $\ddot{M}_n(\theta) \succeq O(1)\boldsymbol{I}_p > \mathbf{0}\ \forall\theta \in \Theta$.]

## Theorem 8.4 (Consistency of $M$-estimator)

*Let $M_n$ be random functions and let $M$ be a fixed function of $\theta$ such that for every $\varepsilon > 0$, if we have conditions:*

*C1.* **Uniformly convergence**: $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$;

*C2.* **Well-separation(Identifiability)**: $\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0)$;

*C3. The $\{\hat{\theta}_n\}$ satisfies* **nearly maximization condition** . *Then,* $\hat{\theta}_n \xrightarrow{P} \theta$.

# A counterexample for well-separation

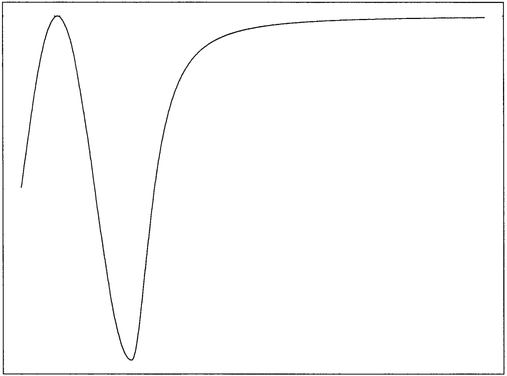

Figure: Example of a M-function whose point of maximum is not well separated.

# Proof of Theorem 8.4

By the well-separation assumption C2, $\forall \varepsilon > 0$, $\exists\ \eta > 0$ s.t.:

$$M(\theta) < M(\theta_0) - \eta \text{ for every } \theta \text{ with } d(\theta, \theta_0) \geq \varepsilon$$

Put $\theta = \hat{\theta}_n$. Thus, $\{d(\hat{\theta}_n, \theta_0) \geq \varepsilon\} \subseteq \{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$. Then

$$P\{d(\hat{\theta}_n, \theta_0) \geq \varepsilon\} \leq P\{M(\theta_0) - M(\hat{\theta}_n) > \eta\} \overset{???}{\to} 0. \quad (10)$$

Next, we show that "$\overset{???}{\to} 0$" is valid by using C3 and C1. By C3, it gives

$$M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1) = M(\theta_0) - o_P(1). \quad (11)$$

where the "$=$" in (11) is by C1: $M_n(\theta) \overset{\mathrm{P}}{\to} M(\theta)$ for $\forall \theta \in \Theta$. Using (11),

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_P(1) \overset{\mathrm{P}}{\to} 0 \quad \text{[by C1]}. \end{aligned}$$

Let $n \to \infty$ in (10), it implies $d(\hat{\theta}_n, \theta_0) \overset{\mathrm{P}}{\to} 0$.

## Corollary 8.5

*Under the (i)* **uniformly convergence C1** *in Theorem 8.4, if we have :*
*(ii).* **Unique maximization***: $M(\theta) =: \mathrm{E}m_\theta(X)$ is uniquely maximized at $\theta_0$;*
*(iii).* **Compactification***: The $\Theta$ is compact;*
*(iv).* **Continuous** *M***-function***: The $M(\theta)$ is continuous. Then, $\hat{\theta}_n \xrightarrow{P} \theta$ for any $\hat{\theta}_n$ satisfying (v) $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$.*

**Proof**: For $\forall \delta > 0$, let $B_\delta(\theta_0) := \{\theta : d(\theta, \theta_0) < \delta\}$. By (ii-iv), we have

$$\sup_{\theta \in \Theta \cap B_\delta^c(\theta_0)} M(\theta) =: M(\theta^*) < M(\theta_0) \text{ for a } \theta^* \in \Theta \cap B_\delta^c(\theta_0).$$

For sufficient large $n$, $\exists\ \varepsilon > 0$ s.t.

$$M(\hat{\theta}_n) \overset{(1)}{>} M_n(\hat{\theta}_n) - \varepsilon/3 \overset{(2)}{>} M_n(\theta_0) - 2\varepsilon/3 \overset{(3)}{>} M(\theta_0) - \varepsilon \qquad (12)$$

In (12): "$\overset{(1)}{>}$" is by (i) and "$\overset{(2)}{>}$" is due to (v) and "$\overset{(3)}{>}$" stems from (i), i.e.

1. From (i) $\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{\mathrm{P}}0$. $E_1$
2. Taking $o_P(1)<\varepsilon/3$, then $M_n(\hat{\theta}_n)\geq M_n(\theta_0)-o_P(1)>M_n(\theta_0)-\varepsilon/3$ with probability approaching 1 (wpa1). $E_2$
3. The same as (1). $E_3$

Let $\varepsilon=M(\theta_0)-M(\theta^*)>0$, plugging this $\varepsilon$ into (12) we get

$$M(\hat{\theta}_n)>M(\theta^*)\ wpa1$$

by $P(E_1\cap E_2\cap E_3)\geq P(E_1)+P(E_2)+P(E_3)-2$.

It should be noted that

$$\{M(\hat{\theta}_n)>M(\theta^*)\}\subseteq\{d(\hat{\theta}_n,\theta_0)<\delta\}$$

since $\theta^*$ maximizes $M(\theta)$ only in $\Theta\cap B_\delta^c(\theta_0)$.

Then by letting $n\to\infty$

$$1\leftarrow P\{M(\hat{\theta}_n)\geq M(\theta^*)\}\leq P\{d(\hat{\theta}_n,\theta_0)<\delta\}\leq 1.$$

# Consistency of $Z$-Estimator

**Theorem 8.6 (Consistency of $Z$-Estimator)**

*Let $\Psi_n$ be random vector-valued functions and let $\Psi$ be a fixed vector-valued function of $\theta$ such that for every $\varepsilon > 0$, if we have :*

*C1*.* **Uniformly convergence***:* $\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \to 0$;

*C2*.* **Well-separation (Identifiability)***:*
$\inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|$;

*C3*. The* $\{\hat{\theta}_n\}$ *satisfies* **nearly zero condition***:* $\Psi_n(\hat{\theta}_n) = o_P(1)$. *Then,*

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

**Proof**: This follows from the Consistency of $M$-estimation by applying $M_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\|\Psi(\theta)\|$.
We can see that nearly maximization turns to nearly zero condition:

$$-\|\Psi_n(\hat{\theta}_n)\| \geq -\|\Psi_n(\theta_0)\| - o_P(1) = -\|\Psi(\theta_0)\| - o_P(1) = -o_P(1).$$

# Consistency without in uniformly convergence

## Lemma 8.7 (p47 of vdv)

*Let $\Theta$ be a subset of the real line and let $\Psi_n$ be random functions and $\Psi$ a fixed function of $\theta$ such that $\Psi_n(\theta) \xrightarrow{P} \Psi(\theta)$ for every $\theta$. Assume that:*
*(a1)Each map $\theta \mapsto \Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$, (a2) or is nondecreasing with $\Psi_n(\hat{\theta}_n) = o_P(1)$ ;*
*(b)Let $\theta_0$ be a point s.t. $\Psi\left(\theta_0 - \varepsilon\right) < 0 < \Psi\left(\theta_0 + \varepsilon\right), \forall \varepsilon > 0$. Then,*

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

## Example 8.8 (Median)

The sample median $\hat{\theta}_n$ is the zero $\theta \mapsto \Psi_n(\theta) = n^{-1} \sum_{i=1}^{n} \operatorname{sign}\left(X_i - \theta\right)$.
By the LLN, for every fixed $\theta$,

$$\Psi_n(\theta) \xrightarrow{\mathrm{p}} \Psi(\theta) = \mathrm{E} \operatorname{sign}(X - \theta) = \mathrm{P}(X > \theta) - \mathrm{P}(X < \theta).$$

# Example 6.8 (con.)

The uniform convergence in **C1$^*$ of Theorem 6.6** is hard to check. It will need the theory of **Empirical Process** (will be studied in the next half-semester) to establish the uniform convergence.

van der Vaart, A. W., Wellner, J. (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. Springer.
Van de Geer, S. A.(2000). Empirical Processes in M-estimation. Cambridge university press.

In this case it is easier to apply Lemma 6.7.

(a) The functions $\theta \mapsto \Psi_n(\theta)$ are non-increasing.

(b) $-\Psi(\theta_0 - \varepsilon) < 0 < -\Psi(\theta_0 + \varepsilon)$. To see (b),

$\Psi(\theta_0 - \varepsilon) = \mathrm{P}(X > \theta_0 - \varepsilon) - \mathrm{P}(X < \theta_0 - \varepsilon) = 1 - 2\mathrm{P}(X < \theta_0 - \varepsilon)$;
$\Psi(\theta_0) = \mathrm{P}(X > \theta_0) - \mathrm{P}(X < \theta_0) = 0 \Rightarrow \mathrm{P}(X > \theta_0) = \mathrm{P}(X < \theta_0) = 0.5$;
$\Psi(\theta_0 + \varepsilon) = \mathrm{P}(X > \theta_0 + \varepsilon) - \mathrm{P}(X < \theta_0 + \varepsilon) = 1 - 2\mathrm{P}(X < \theta_0 + \varepsilon)$.

If $X$ is continuous and the population median is unique, i.e.

$$\mathrm{P}(X < \theta_0 - \varepsilon) < 0.5 < \mathrm{P}(X < \theta_0 + \varepsilon) \quad \forall\, \varepsilon > 0.$$

Applying $\Psi_n(\theta)$ to (a2)+(b) of Lemma 6.7, it implies $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$.

# Wald's Consistency

The semi-continuity is a property that is weaker than continuity. A function $f \in R$ is said to be **upper semi-continuous** (u.s.c.) if

$$\limsup_{x \to x_0} f(x) \leq f(x_0).$$

[to be lower semi-continuous(l.s.c.) if $-f$ is l.s.c.:$\liminf_{x \to x_0} f(x) \geq f(x_0)$.]

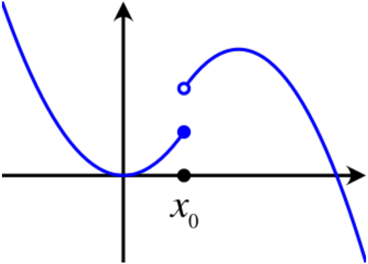


Figure: An l.s.c. function.

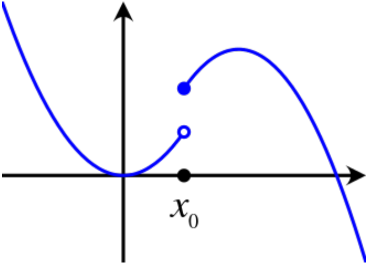


Figure: An u.s.c. function.

The u.s.c. $M$-function is used for Wald's Consistency Condition.

# Wald's Consistency Condition

Let $Pm_\theta := \mathrm{E}m_\theta(X)$. Typically, the map $\theta \mapsto Pm_\theta$ has a unique global maximum at a point $\theta_0$, but here we allow **multiple points of maximum**

$$\Theta_0 := \{\theta_0 \in \Theta : Pm_{\theta_0} = \sup_\theta Pm_\theta\} \neq \emptyset \quad \text{M attains its local maximum.}$$

## Theorem 8.9 (Wald's consistency for $M$-estimator)

*For every compact set $K \subset \Theta$, Wald's consistency Conditions for* $\mathrm{P}(d(\hat{\theta}_n, \Theta_0) \geq \varepsilon,\ \hat{\theta}_n \in K) \to 0$ *is that*

*W1.* **U.S.C. condition***: Let $\theta \mapsto m_\theta(x)$ be u.s.c. for almost all $x$;*

*W2.* **Uniformly bounded on small-balls***: For $\forall$ small-ball $U \subset \Theta$, assume $x \mapsto \sup_{\theta \in U} m_\theta(x)$ is measurable and satisfies the*

$$E\sup_{\theta \in U} m_\theta(X) < \infty. \quad (\textit{a DCT condition}) \tag{13}$$

*W3.* **Nearly maximization on compact set***: The $\{\hat{\theta}_n\}$ satisfies*

$$M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1) \ \textit{for some}\ \theta_0 \in \Theta_0.$$

# Proof of Wald's consistency

- Let $B = \{\theta \in K : d(\theta, \Theta_0) \geq \varepsilon\}$, we are going to show:
$$\mathbb{P}\{\hat{\theta}_n \in B\} \to 0.$$
- If the function $\theta \mapsto Pm_\theta$ is identically $-\infty$, then $\Theta_0 = \Theta$ trivially. We may assume that there exists $\theta_0 \in \Theta_0$ such that $Em_{\theta_0} > -\infty$, thus
$$E\left|m_{\theta_0}(X)\right| < \infty \text{ by (13)}.$$
- Fix some $\theta \in K$ and let $U_l \downarrow \theta$ be a decreasing sequence of open balls around $\theta$ of diameter converging to zero. Let $m_U(x) := \sup_{\theta \in U} m_\theta(x)$. For every $l$, the sequence
$$m_\theta \leq m_{U_l} = \sup_{\theta \in U_l} m_\theta(x) \text{ is decreasing in } l \tag{14}$$
and taking limit in (14) we have
$$m_\theta \leq \lim_{U_l \to \theta} m_{U_l} \leq m_\theta\text{-a.s.}$$
where the last $\leq$ by the upper semi-continuity of $m_\theta(x)$.

- With $m_{U_l} \downarrow m_\theta$ a.e., the monotone convergence theorem for sequence $\{-m_{U_l}\}$ implies that:
$$Em_{U_l}(X) \to Em_\theta(X) \text{ (which may be } -\infty).$$
which shows that $Em_{U_l}(X) \to Em_\theta(X) < Em_{\theta_0}(X)\ \forall\ (\theta \notin \Theta_0)$ since $\theta_0$ maximizes $Em_\theta(X)$.
- Then there exists an open ball $U_k$ who covers $\theta$ such that
$$Em_{U_k}(X) < Em_{\theta_0}(X),\ \forall\ \theta \notin \Theta_0,\ \exists\ k \in \mathbb{N}. \tag{15}$$
- Let $U_\theta$ be the open ball containing $\theta$, then $B$ can be covered by $\{U_\theta : \theta \in B\}$ by the compactness of $K$. Let $U_{\theta_1}, \cdots, U_{\theta_p}$ be the finite subcovers, then we have by **LLN**:
$$\sup_{\theta \in B} \mathbb{P}_n m_\theta \le \sup_{\substack{\theta \in U_{\theta_j} \\ j=1,\cdots,p}} \mathbb{P}_n m_\theta \xrightarrow{a.s.} \sup_{\substack{\theta \in U_{\theta_j} \\ j=1,\cdots,p}} Em_\theta(X) < Em_{\theta_0}(X) \tag{16}$$
where the $=$ is by covering and the last $<$ is from (15).

- On the other hand, by the covering and by the assumption of nearly maximization on compact set, the event $\{\hat{\theta}_n \in B\}$ implies the event

$$\sup_{\theta \in B} \mathbb{P}_n m_\theta \geq \mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - o_P(1) = E m_{\theta_0}(X) - o_P(1)$$

where the $=$ is by LNN applying to $\mathbb{P}_n m_{\theta_0}$.

- Hence, we have $\{\hat{\theta}_n \in B\} \subset \{\sup_{\theta \in B} \mathbb{P}_n m_\theta \geq E m_{\theta_0}(X) - o_P(1)\}$ which leads to

$$\mathbb{P}\{\hat{\theta}_n \in B\} \leq \mathbb{P}\left\{\sup_{\theta \in B} \mathbb{P}_n m_\theta \geq E m_{\theta_0}(X) - o_P(1)\right\} \to 0.$$

the last limit stems from $\sup_{\theta \in B} \mathbb{P}_n m_\theta < E m_{\theta_0}(X)$ a.s. in (16).

# Example: Cauchy MLE

**Example 8.10 (Cauchy likelihood)**

The pdf of Cauchy distribution $\mathrm{Cauchy}(\theta)$ is

$$f_\theta(x) = \frac{1}{\pi\left\{1 + (x-\theta)^2\right\}}.$$

The MLE for location $\theta$ based on the sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Cauchy}(\theta)$ maximizes the log-likelihood function $\theta \mapsto \mathbb{P}_n m_\theta$:

$$m_\theta(x) = -\log\left(1 + (x-\theta)^2\right).$$

The parameter space $\mathbb{R}$ is not compact, but we can enlarge the $\mathbb{R}$ to

$$\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}.$$

i.e. the $\overline{\mathbb{R}}$ is the compactification of $\mathbb{R}$.

## Analysis via Wald's consistency Conditions

1. **U.S.C. condition**: The $m_\theta(x)$ is continuous (also u.s.c.).
2. **Uniformly bounded on small-balls**:

$$\mathrm{E} \sup_{\theta \in U} m_\theta(X) = \int \sup_{\theta \in U} \frac{-\log\{1+(x-\theta)^2\}}{\pi\{1+(x-\theta)^2\}} dx < \infty.$$

$$m_{-\infty}(x) = \limsup_{\theta \mapsto -\infty} m_\theta(x) = -\infty; \qquad m_{\infty}(x) = \limsup_{\theta \mapsto \infty} m_\theta(x) = -\infty$$

3. **Nearly maximization on compact set**:

$$\arg\max_{\theta \in \Theta} M_n(\theta) := M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$$

- Then, we apply Wald's theorem with $\Theta = \overline{\mathbb{R}}$ equipped with the metric

$$d(\theta_1, \theta_2) = |\text{arctg}\, \theta_1 - \text{arctg}\, \theta_2| .$$

- $\Theta_0 = \{\theta_0\}$.
- Thus, taking $K = \overline{\mathbb{R}}$, we obtain that $d(\hat{\theta}_n, \Theta_0) \xrightarrow{\mathrm{P}} 0$.

# Asymptotic Normality of $Z$-estimator

Suppose

- **Measurability**. For each $\theta$ in an open subset of Euclidean space, let $x \mapsto \psi_\theta(x)$ be a measurable vector-valued function.
- **Lipschitz condition**. For every $\theta_1$ and $\theta_2$ in a neighborhood of $\theta_0$ and a measurable function $\dot{\psi}(x)$ with $P[\dot{\psi}(X)]^2 < \infty$, we have

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \leq \dot{\psi}(x) \|\theta_1 - \theta_2\| .$$

- **Moment and differentiability conditions**. Assume that $P \|\psi_{\theta_0}\|^2 < \infty$ and that the map $\theta \mapsto P\psi_\theta$ is differentiable at a zero $\theta_0$, with nonsingular derivative matrix $V_{\theta_0} := \frac{\partial}{\partial\theta} P[\psi_\theta(X)]\big|_{\theta=\theta_0}$.
- **Consistency**. $\mathbb{P}_n \psi_{\hat{\theta}_n} = o_P\left(n^{-1/2}\right)$ and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$.

then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_P(1) \xrightarrow{d} N(0, V_{\theta_0}^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T (V_{\theta_0}^{-1})^T).$$

# Asymptotic Normality of $M$-estimator

Suppose

- **Measurability**. For each $\theta$ in an open subset of Euclidean space, let $x \mapsto m_\theta(x)$ be a measurable function such that $\theta \mapsto m_\theta(x)$ is differentiable at $\theta_0$ for $P$-almost every $x$ with derivative $\dot{m}_{\theta_0}(x)$.
- **Lipschitz condition**. For every $\theta_1$ and $\theta_2$ in a neighborhood of $\theta_0$ and a measurable function $\dot{\psi}(x)$ with $P[\dot{m}(X)]^2 < \infty$, we have
$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot{m}(x) \left\|\theta_1 - \theta_2\right\|.$$
- **Moment and differentiability conditions**. Assume that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion
$$Pm_\theta = Pm_{\theta_0} + (\theta - \theta_0)^T V_{\theta_0} (\theta - \theta_0)/2 + o(\|\theta - \theta_0\|^2).$$
at a point of maximum $\theta_0$ with nonsingular symmetric second derivative matrix $V_{\theta_0} = \frac{\partial^2}{\partial \theta^2} P[m_\theta(X)]\Big|_{\theta=\theta_0}$.
- **Consistency**. $\mathbb{P}_n m_{\hat{\theta}_n} \ge \sup_\theta \mathbb{P}_n m_\theta - o_P\left(n^{-1}\right)$ and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i) + o_P(1) \xrightarrow{d} N(0, V_{\theta_0}^{-1} P \dot{m}_{\theta_0} \dot{m}_{\theta_0}^T V_{\theta_0}^{-1}).$$

# Asymptotic Normality of $M$- and $Z$-estimator

To prove AN of $M$- and $Z$-estimator, the Empirical Process is indispensable. It will be taught detailly in the next half-semester.

## Formal Settings

- Let $X_1, \ldots, X_n$ be a random sample from a $P$ on a measurable space
$$(\mathcal{X}, \mathcal{A}).$$
- We denote the empirical distribution by
$$\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$$
as a discrete uniform measure, where $\delta_x$ is the probability distribution that is degenerate at $x$.
- Given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of $f$ under the empirical measure $\mathbb{P}_n$, and $Pf$ for the expectation under $P$. Thus
$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad Pf = \int f dP.$$

Actually, we treat $\mathbb{P}_n, \mathbb{P}$ as operators rather than the measure.

# Asymptotic Normality of $Z$-estimator

In the heuristic proof for the AN of Z-estimator, we used $\ddot{\psi}_n(\tilde{\theta}_n) = O_p(1)$. A more vigorous proof is created via the so-called Donsker Class.

- Let $G_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(x))$

  A class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} measurable\}$ istb Donsker if $\forall f \in \mathcal{F}$, $G_n f \xrightarrow{\mathrm{d}}$ a tight process in $\ell^\infty(\mathcal{F})$ where $\ell^\infty(\mathcal{F})$ be the set of bounded functions on $\mathcal{F}$.

  Tight here means "$\forall \varepsilon > 0 \; \exists$ a compact set K s.t. $P(x \notin K) < \varepsilon$"

## Example: (parametric class of liptchitz, vdv Ex 19.7)

Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a collection of measurable functions indexed by a bounded subset $\Theta \subset \mathbb{R}^d$. If there exists a measurable function $m(x)$ such that $f_\theta(x)$ is $m(x)$-Lipschitz w.r.t. Euclidean norm

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x) \|\theta_1 - \theta_2\|, \quad \forall \; \theta_1, \theta_2.$$

It can be shown that if $P|m|^r < \infty$ for some $r > 0$, the class of functions $\mathcal{F}$ is P-Donsker.

# A Lemma for Random function

## Lemma 8.11 (Lemma 19.24 in vdv.)

*Suppose that :*

(a) $\mathcal{F}$ *is a P-Donsker class of measurable functions.*

(b) $\{\hat{f}_n\}$ *be a set of random functions that take their values in* $\mathcal{F}$ *such that*

$$\int (\hat{f}_n(x) - f_0(x))^2 dP(x) = P(\hat{f}_n - f_0)^2 \xrightarrow{\mathrm{p}} 0$$

*for some* $f_0 \in L_2(P)$*, i.e.* $Pf_0^2 < \infty$*. Then*

$$\mathbb{G}_n(\hat{f}_n - f_0) \xrightarrow{\mathrm{P}} 0 \text{ and hence } \mathbb{G}_n \hat{f}_n \rightsquigarrow \mathbb{G}_P f_0.$$

# A Lemma for Random function

## *Remark 7*

- *In Lemma 19.24 in vdv, We'll set* $\hat{f}_n := \psi_{\hat{\theta}_n}, f_0 := \psi_{\theta_0}$, *then*

$$P(\psi_{\hat{\theta}_n} - \psi_{\theta_0})^2 \leq P(\dot{\psi})\|\hat{\theta}_n - \theta_0\|^2 \xrightarrow{\mathrm{P}} 0$$

*as* $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, $P(\dot{\psi})^2 < \infty$ *and* $P(\ddot{\psi})^2 < \infty$ *as assumed in the condition of Z-estimator AN.*
*Thus, as* $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ *is P-Donsker, we have*

$$\mathbb{G}_n(\hat{f}_n - f_0) = \sqrt{n}(\mathbb{P}_n \hat{f}_n - P\hat{f}_n) - \sqrt{n}(\mathbb{P}_n f_0 - Pf_0) \xrightarrow{\mathrm{P}} 0$$

$$\Rightarrow \sqrt{n}(\mathbb{P}_n \hat{f}_n - P\hat{f}_n) = \sqrt{n}(\mathbb{P}_n f_0 - Pf_0) + o_p(1)$$

# Proof of AN of $Z$-estimator:

From the lemma and the last of the Remark,

$$\mathbb{G}_n\psi_{\hat{\theta}_n} - \mathbb{G}_n\psi_{\theta_0} = \mathbb{G}_n(\psi_{\hat{\theta}_n} - \psi_{\theta_0}) \xrightarrow{\mathrm{P}} 0. \tag{17}$$

Note $G_n\psi_{\hat{\theta}_n} = \sqrt{n}(\mathbb{P}_n\psi_{\hat{\theta}_n} - P\psi_{\hat{\theta}_n})$ and $\mathcal{F}\psi_{\theta_0} = P\psi_{\theta_0} = 0$, which is natural as $\mathbb{P}_n\psi_{\hat{\theta}_n} = o_p(n^{1/2})$,

$$\mathbb{G}_n\psi_{\hat{\theta}_n} = \sqrt{n}(\mathbb{P}_n\psi_{\theta_0} - P\psi_{\hat{\theta}_n}) + o_p(1). \tag{18}$$

(17) and (18) $\Rightarrow$

$$\mathbb{G}_n\psi_{\theta_0} = \mathbb{G}_n\psi_{\hat{\theta}_n} + o_{\mathrm{P}}(1) = \sqrt{n}(P\psi_{\theta_0} - \mathbb{P}_n\psi_{\hat{\theta}_n}) + o_{\mathrm{P}}(1)$$

By Taylor Exp,

$$\frac{1}{\sqrt{n}}\sum \psi_{\theta_0}(X_i) + o_p(1) = -\sqrt{n}\frac{\partial}{\partial\theta}P\psi_\theta(X)\Big|_{\theta=\theta_0}(\hat{\theta}_n - \theta_0) + \sqrt{n}o_P(\|\hat{\theta}_n - \theta_0\|)$$

# Proof of AN of $Z$-estimator:

Thus,

$$\sqrt{n}V_{\theta_0}(\hat{\theta}_n - \theta_0) = -\mathbb{G}_n\psi_{\theta_0} + \sqrt{n}o_P(\|\hat{\theta}_n - \theta_0\|) \tag{19}$$

and $\sqrt{n}\|V_{\theta_0}(\hat{\theta}_n - \theta_0)\| = O_P(1)$. Note
$\sqrt{n}\|\hat{\theta}_n - \theta_0\| \leq \|V_{\theta_0}^{-1}\|\sqrt{n}\|V_{\theta_0}(\hat{\theta}_n - \theta_0)\| = O_P(1) + o_P(\sqrt{n}\|\hat{\theta}_n - \theta_0\|)$.
So, $\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_p(1)$ and $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$
from (19),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1).$$

## Median: Example 5.24 in vdv's book

- The sample median maximizes the criterion function $\theta \mapsto -\sum_{i=1}^{n}|X_i - \theta|$.
- Assume that the distribution function $F(x)$ of the observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ is differentiable at its median $\theta_0 = F^{-1}(1/2)$ with positive derivative $f(\theta_0)$.
- It follows from Theorem 5.23 applied with centralized M function $m_\theta(x) = |x-\theta| - |x|$. As a consequence of the triangle inequality, this function satisfies the Lipschitz condition with $\dot{m}(x) \equiv 1$:

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)|\theta_1 - \theta_2|$$

  due to $\max\{|a|-|b|, |b|-|a|| \leq |a-b|\}$.
- Furthermore, the map $\theta \mapsto m_\theta(x)$ is differentiable at $\theta_0$ except $x = \theta_0$, with $\dot{m}_{\theta_0}(x) = -\operatorname{sign}(x-\theta_0)$. So

$$\mathrm{E}\left(\dot{m}_{\theta_0}(X)\right)^2 = \mathrm{E}\left(\operatorname{sgn}^2(x-\theta_0)\right) = 1.$$

# Median

- By partial integration,

$$Pm_\theta = \mathrm{E}\left(m_\theta(X)\right) = \int |x-\theta| dF(x) - \int |x| dF(x)$$

$$= \theta F(0) + \int_{(0,\theta]} (\theta - 2x) dF(x) - \theta(1 - F(\theta)) = 2\int_0^\theta F(x)dx - \theta.$$

  If $F(x)$ is sufficiently regular around $\theta_0$, then $Pm_\theta$ is twice differentiable

$$\frac{dPm_\theta}{d\theta} = 2F(\theta) - 1, \frac{d^2 Pm_\theta}{d\theta^2} = 2f(\theta).$$

- More generally, under the minimal condition that $F(x)$ is differentiable at $\theta_0$,

$$Pm_\theta = Pm_{\theta_0} + \frac{1}{2}\left(\theta - \theta_0\right)^2 2f\left(\theta_0\right) + o(|\theta - \theta_0|^2).$$

  Since $V_{\theta_0}^{-1} P[\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T] V_{\theta_0}^{-1} = 1/\left(2f\left(\theta_0\right)\right)^2$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}\left(X_i\right) + o_P(1) \xrightarrow{d} N(0, 1/\left(2f\left(\theta_0\right)\right)^2).$$

## Nonlinear least squares: Example 5.27 in vdv's book

Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ consisting of the "covariates" $X$ and "response variables" $Y$, follows

$$Y = f_{\theta_0}(X) + e, \quad \mathrm{E}(e|X) = 0 \quad \mathrm{Var}\,(e|X) = \sigma^2\,(X) < \infty.$$

- The least squares estimator that minimizes
$$\theta \mapsto \textstyle\sum_{i=1}^n \left(Y_i - f_\theta\left(X_i\right)\right)^2$$
is an M-estimator for $m_\theta(x, y) = -\left(y - f_\theta(x)\right)^2$.
- It should be expected to converge to the minimizer of the limit criterion function

$$\begin{aligned}\theta \mapsto Pm_\theta = \mathrm{E}\left(Y - f_\theta\left(X\right)\right)^2 &= \mathrm{E}[Y - f_{\theta_0}\left(X\right) + \left(f_{\theta_0}\left(X\right) - f_\theta\left(X\right)\right)]^2 \\ &= \mathrm{E}\left(f_{\theta_0} - f_\theta\right)^2 + \mathrm{E}e^2. \qquad (20)\end{aligned}$$

Thus the LS estimator should be consistent if $\theta_0$ is identifiable from the model, in the sense that $\theta \neq \theta_0$ implies that

$$P(f_\theta(X) \neq f_{\theta_0}(X)) > 0.$$

## Nonlinear least squares

- Note that

$$|m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| = \left|(y - f_{\theta_1}(x))^2 - (y - f_{\theta_2}(x))^2\right|$$
$$= |f_{\theta_1}(x) - f_{\theta_2}(x)|\,|2y - f_{\theta_1}(x) - f_{\theta_2}(x)|$$

- We may assume that

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le \dot{f}(x)\,\|\theta_2 - \theta_1\|$$
$$\text{and } \exists\ c(x) \text{ s.t. } f_\theta(x) \le c(x), \quad \forall \theta \in \Theta.$$

Thus

$$|m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| \le |f_{\theta_1}(x) - f_{\theta_2}(x)|\,(2|y| + 2c(x))$$

i.e. $\dot{m}(x,y) := \dot{f}(x)[2|y| + 2c(x)]$.

- Assume that $f_\theta(x)$ is continuous differentiable at $\theta_0$, we check the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion

# Nonlinear least squares

- By (20), we have

$$Pm_\theta = \mathrm{E}\left(Y - f_\theta(X)\right)^2 + \mathrm{E}e^2 = Pm_{\theta_0} + \int [f_\theta(x) - f_{\theta_0}(x)]^2 p(x)dx$$
$$= Pm_{\theta_0} + \int \left[(\theta-\theta_0)^T \dot{f}_{\theta_0}(x) + o(\|\theta-\theta_0\|)\right]^2 p(x)dx$$
$$= Pm_{\theta_0} + \frac{1}{2}(\theta-\theta_0)^T 2\int \dot{f}_{\theta_0}(x)\dot{f}_{\theta_0}^T(x)p(x)dx\,(\theta-\theta_0) + o(\|\theta-\theta_0\|).$$

- So $V_{\theta_0} = 2\int \dot{f}_{\theta_0}(x)\dot{f}_{\theta_0}^T(x)p(x)dx = 2P[\dot{f}_{\theta_0}]\dot{f}_{\theta_0}^T$ and $\dot{m}_{\theta_0}(x,y) = -2\left(y - f_{\theta_0}(x)\right)\dot{f}_{\theta_0}(x) = -2e\dot{f}_{\theta_0}(x)$. If other conditions in Thm 5.23 in vdv are fulfilled, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-V_{\theta_0}^{-1}}{\sqrt{n}}\sum_{i=1}^n \dot{m}_{\theta_0}(X_i, Y_i) + o_P(1) \xrightarrow{d} N(0, V_{\theta_0}^{-1}P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T V_{\theta_0}^{-1}).$$

where (since $e$ and $X$ are independent)

$$V_{\theta_0}^{-1}P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T V_{\theta_0}^{-1} = [2P\dot{f}_{\theta_0}\dot{f}_{\theta_0}^T]^{-1}4Pe^2P[\dot{f}_{\theta_0}\dot{f}_{\theta_0}^T][2P\dot{f}_{\theta_0}\dot{f}_{\theta_0}^T]^{-1} = 2[2P\dot{f}_{\theta_0}\dot{f}_{\theta_0}^T]^{-1}\sigma^2(X).$$

## Examples: Binary regression (GLMs, **vdv's book Ex. 5.11**)

- Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ consisting of $k$-dimensional vectors of "covariates" $X$; , and 0-1 "response variables" $Y$ following
$$P_\theta\left(Y_i = 1 | X_i = x\right) = \Psi\left(\theta^T x\right).$$
- Here $\Psi : \mathbb{R} \mapsto [0,1]$ is a known continuously differentiable, monotone function. The choices $\Psi(t) = 1/\left(1 + e^{-t}\right)$ (the logistic distribution function) and $\Psi = \Phi$ (the normal distribution function) correspond to the **logistic regression** and **probit model**, respectively. The MLE maximizes the (conditional) likelihood function
$$\theta \mapsto \prod_{i=1}^n p_\theta\left(Y_i | X_i\right) := \prod_{i=1}^n \Psi\left(\theta^T X_i\right)^{Y_i}\left(1 - \Psi\left(\theta^T X_i\right)\right)^{1-Y_i}.$$
- For identifiability of $\theta$, we must assume that the distribution of the $X_i$ is not concentrated on a $(k-1)$-dimensional affine subspace of $\mathbb{R}^k$. For simplicity, we assume that the range of $X_i$ is bounded and the non-singularity of the matrix $\mathrm{E}XX^T$.

## Examples: Binary regression (AN)

- **The consistency of $\hat{\theta}_n$ can be proved by combining Theorem 6.6 (Consistency of Z-Estimator).**
- The asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta)$ is now a consequence of Theorem 6.14 (AN of Z-estimator). The score function (Z-function)

$$\psi_\theta(x) := \dot{\ell}_\theta(y|x) = \frac{y - \Psi\left(\theta^T x\right)}{\Psi\left(\theta^T x\right)\left[1 - \Psi\left(\theta^T x\right)\right]} \Psi'(\theta^T x) x$$

  is **uniformly bounded** in $x, y$ and $\theta$ ranging over compacta, and continuous in $\theta$ for every $x, y$.
- The Fisher information matrix is

$$I_\theta = \mathrm{E} \frac{\Psi'\left(\theta^T X\right)^2}{\Psi\left(\theta^T X\right)\left[1 - \Psi\left(\theta^T X\right)\right]} X X^T$$

- Asymptotic distribution for $\hat{\theta}_n$ is given by

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{d}{\rightsquigarrow} N\left(0, I_\theta^{-1}\right).$$

# Estimated Parameter

$$X_1, \cdots, X_n \overset{i.i.d}{\sim} F_{(\theta,\eta)}$$

- $\theta$ is the parameter of interest and $\eta$ is the nuiance paramater.
- Often we plug-in an estimator of $\eta$, say $\hat{\eta}_n$ in the Z-estimating equation,

$$P_n\psi_{(\theta,\eta)} \Rightarrow P_n\psi_{(\theta,\hat{\eta}_n)} = \frac{1}{n}\sum_{i=1}^{n}\psi_{(\theta,\hat{\eta}_n)}(x_i) = 0$$

- This is essentially a 2-step procedure.

# Motivating Example

$$y_i = m_{\theta_0}(x_i) + \epsilon_i, \; E(\epsilon_i | x_i) = 0$$

$y_i$ subject to missingness, and

$$R_i = \begin{cases} 1, & \text{if } y_i \text{ observed,} \\ 0, & \text{if } y_i \text{ missing.} \end{cases}$$

**MAR(Missing at Random assumption)**

$$P(R_i = 1 | x_i, y_i) = P(R_i = 1 | x_i) = \omega_{\eta_0}(x_i)$$

$\omega_{\eta_0}(\cdot)$ istb to missing propensity function. Here is a binary regression model.

- MAR $\Rightarrow$ Given $x_i$, $R_i$ and $y_i$ are independent.
- the so-called ignorable missing at random
- "ignorable": the missing $Y_i$ is ignorable as long as we have $X_i$

# How to estimate $\theta$?

Method 1: Do LSE on data with $R_i = 1$.

$$LS_n(\theta) = \sum_{i=1}^{n} R_i(y_1 - m_\theta(x_i))^2$$

$$\frac{\partial LS_n(\theta)}{\partial \theta} = -2\sum_{i=1}^{n} R_i(y_1 - m_\theta(x_i))\frac{\partial m_\theta(x_i)}{\partial \theta} \tag{21}$$

At $\theta_0$, by assuming $E(\epsilon_i|x_i) = 0$,

$$E\left\{R_i(y_i - m_{\theta_0}(x_i))\frac{\partial m_{\theta_0}(x_i)}{\partial \theta}\right\} = E\left\{\epsilon_i \frac{\partial m_{\theta_0}(x_i)}{\partial \theta}\omega_{\eta_0}(x_i)\right\} \overset{MAR}{=} 0$$

So, the LS estimate that solves (21) is consistent and AN under centain regular conditions.

# How to estimate $\theta$?

Method 2: Inverse Prob Weighted Estimation. (weight (21) by $\omega_{\eta_0}(x_i)$)

$$\sum_{i=1}^{n} \frac{R_i(y_i - m_\theta(x_i))\frac{\partial m_\theta(x_i)}{\partial x_i}}{\omega_{\eta_0}(x_i)} = 0 \tag{22}$$

$$E\left\{\sum_{i=1}^{n} \frac{R_i(y_i - m_\theta(x_i))\frac{\partial m_\theta(x_i)}{\partial x_i}}{\omega_{\eta_0}(x_i)}\right\} = E\left\{\epsilon_i \frac{\partial m_{\theta_0}(x_i)}{\partial \theta}\right\} = 0$$

For the estimator from (21) that ignore missing values and IPW estimator from (22) , which one is more efficient?

# How to estimate $\theta$?

However, $\eta_0$ is unknown, which can be estimated by the binary likelihood,

$$L_n(\eta) = \prod_{i=1}^{n} \omega_\eta^{R_i}(x_i)(1-\omega_\eta(x_i))^{R_i}$$

$$l_n(\eta) = \sum_{i=1}^{n} \{R_i \log \omega_\eta(x_i) + (1-R_i)\log(1-\omega_\eta(x_i))\}$$

$$\frac{\partial l_n(\eta)}{\partial \eta} = \sum_{i=1}^{n} \left\{ \frac{R_i}{\omega_\eta(x_i)} - \frac{1-R_i}{1-\omega_\eta(x_i)} \right\} \frac{\partial \omega_\eta(x_i)}{\partial \eta} \overset{set}{=} 0 \Rightarrow \hat{\eta}$$

The (22) becomes

$$\sum_{i=1}^{n} \frac{R_i(y_i - m_\theta(x_i))\frac{m_\theta}{\partial\theta}}{\omega_{\hat{\eta}}(x_i)} = 0 \tag{23}$$

Chen, Leung and Qin(2008) showed the estimation for $\theta$ based on (23) which estimated $\hat{\eta}$ is more efficient than that using the true $\eta_0$ in (22).

# Motivation Example

The parametric assumption of $P(R_i = 1|x_i) = \omega_{\eta_0}(x_0)$ may be too strong. May consider a nonparametric form

$$P(R_i = 1|x_i, y_i) = P(R_i = 1|x_i) = \omega(x_0)$$

The missing propensity $\omega(\cdot)$ can be estimated via the kernel smoothing method

$$\hat{\omega}_h(x) = \frac{\sum_{i=1}^n K(\frac{x-x_i}{h})R_i}{\sum_{i=1}^n K(\frac{x-x_i}{h})}$$

where K is a kernel, symmetric pdf, h is a smoothing bandwidth, $h \to 0$, $nh \to 0$, as $n \to \infty$.

$$\begin{aligned} E(\hat{\omega}_n(x)) &= E(\sum_{i=1}^n E(\frac{K(\frac{x-x_i}{h})}{\sum K(\frac{x-x_i}{h})} R_i | x_1, \cdots, x_n) \\ &= E(\sum_{i=1}^n \frac{K(\frac{x-x_i}{h})}{\sum K(\frac{x-x_i}{h})} \omega(x_i)) \text{ (a weighted average of } \{\omega(x_i)\}_n) \end{aligned}$$

can show $\sqrt{nh^d}(\hat{\omega}_n(x) - \omega_n(x)) \to K(\mu, v^2)$ $(x \in R^d)$ if $h \approx O(n^{-\frac{1}{4+d}})$.

# Estimated Parameter

## Ex 5.32 (Symmetric Location)

$X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ which is symmetric about $\theta_0$. Let $x \to \psi(x)$ be antisymmetric (odd function). Consider Z-estimator via $\frac{1}{n}\sum_i \psi(\frac{x_i - \theta_0}{\hat{\sigma}})$, $\hat{\sigma}$ is an estimator of $\sigma$.

$$P\psi_{\theta_0,\hat{\sigma}} = \int \psi(\frac{x_0 - \theta_0}{\hat{\sigma}})dF(x) = 0, \ \forall \ \hat{\sigma},$$

since $F(\cdot)$ is symmetric about $\theta_0$ and $\psi(\cdot)$ is an odd function.
Hence, from Th 5.31 ,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0,\eta_0}^{-1}\frac{1}{\sqrt{n}}(P_n\psi_{\theta_0,\eta_0} - P\psi_{\theta_0,\eta_0}) + o_p(1)$$

The estimation is effectively using the true $\eta_0$ as the effect of $\hat{\sigma}$ is not present in the leading order.

Suppose $X_1, \cdots, X_n$ i.i.d. $P \in \mathcal{P}$, and $h : \mathbb{R}^m \to \mathbb{R}$ measurable for a finite positive integer $m < n$, i.e. $h(x_1, \cdots, x_m) = h(x_{i_1}, \cdots, x_{i_m})$ where $(i_1, \cdots, i_m)$ is an arbitrary permutation of $1, \cdots, m$. If not, one can always define and replace by the symmetry:

$$\frac{1}{m!} \sum_{\text{all permutation of } (i_1, \cdots, i_m) \text{ of } (1, \cdots, m)} h(x_{i_1}, \cdots, x_{i_m})$$

Let $\theta = \mathrm{E}h(X_1, \cdots, X_m)$ if $|\mathrm{E}h(X_1, \cdots, X_m)| < \infty$.

**Definition 9.1**

$U_n := \binom{n}{m}^{-1} \sum_{\ell} h(x_{i_1}, \cdots, x_{i_m})$ is called a U-Statistics with kernel $h$ of order $m$, where $\sum_\ell$ denotes the summation over the $\binom{n}{m}$ candidates of $m$-distinct elements $\{i_1, \cdots, i_m\}$ from $\{1, \cdots, m\}$.

# Examples

- $n^{-1}\sum x_i$ is a U-Statistic with kernel $h(x) = x$ of order 1.

- $n^{-1}\sum x_i^k$ is a U-Statistic with kernel $h(x) = x^k$ of order 1.

- $\binom{n}{m}^{-1}\sum_\ell x_{i_1}\cdots x_{i_m}$ is a U-Statistic with kernel:

$$h(x_1, \cdots, x_m) = \prod_{i=1}^{m} x_i$$

  of order $m$.

- $$\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\frac{(x_i - x_j)^2}{2} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

  is a U-statistic of order 2 with $h(x_1, x_2) = \frac{(x_1-x_2)^2}{2}$.

# Variance of U-Statistic

Assume $\mathrm{E}h^2(x_1, \cdots, x_m) < \infty$. For $k \in \{1, \cdots, m\}$, let:

$$\begin{aligned} h_k(x_1, \cdots, x_k) &= \mathrm{E}\left[h(X_1, \cdots, X_m) \,|\, X_1 = x_1, \cdots, X_k = x_k\right] \\ &= \mathrm{E}h(x_1, \cdots, x_k, X_{k+1}, \cdots, X_m) \end{aligned} \tag{24}$$

Clearly, we have $h_m = h$, $h_k(x_1, \cdots, x_k) = \mathrm{E}h_{k+1}(x_1, \cdots, x_k, X_{k+1})$, and:

$$\mathrm{E}h_k(X_1, \cdots, X_k) = \mathrm{E}h(X_1, \cdots, X_m) = \theta$$

Define $\tilde{h}_k(x_1, \cdots, x_k) = h_k(x_1, \cdots, x_k) - \theta$, then:

$$U_n - \mathrm{E}U_n = \binom{n}{m}^{-1} \sum_{\ell} \tilde{h}(X_{i_1}, \cdots, X_{i_m}) \tag{25}$$

# Hoeffding's Theorem

## Theorem 9.2

*Let* $X_1, \cdots, X_n$ *i.i.d.* $P \in \mathcal{P}$ *with* $\mathrm{E}_P h^2(X_1, \cdots, X_m) < \infty$*, then:*

$$\mathrm{var}_P(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \xi_k \tag{26}$$

*where* $\xi_k = \mathrm{var}_P(h_k(X_1, \cdots, X_k))$ *satisfying:*

(i) $\dfrac{m^2}{n}\xi_1 \le \mathrm{var}_P(U_n) \le \dfrac{m}{n}\xi_m$.

(ii) $(n+1)\,\mathrm{var}_P(U_{n+1}) \le n\,\mathrm{var}_P(U_n)$.

(iii) $\mathrm{var}_P(U_n) = \dfrac{k!\binom{m}{k}^2 \xi_k}{n^k} + O\left(n^{-(k+1)}\right)$ *as* $n \to \infty$*, if* $\xi_k \neq 0$ *but* $\xi_j = 0$ *for* $j < k$.

See Shao section 3.2.

# Proof

Let $\{i_1, \cdots, i_m\}$ and $\{j_1, \cdots, j_m\}$ be two sets of m-distinct intergers from $\{1, \cdots, n\}$ s.t. $\{i_1, \cdots, i_m\} \cap \{j_1, \cdots, j_m\} \overset{\text{WLOG}}{=} \{1, \cdots, k\}$, then

$$
\begin{aligned}
&E_p\tilde{h}(x_1, \cdots, x_k, x_{i_k+1}, \cdots, x_{i_m})\tilde{h}(x_1, \cdots, x_k, x_{j_k+1}, \cdots, x_{j_m})\\
=&E_p\{E_p\tilde{h}(x_1, \cdots, x_k, x_{i_k+1}, \cdots, x_{i_m})\tilde{h}(x_1, \cdots, x_k, x_{j_k+1}, \cdots, x_{j_m})|x_1, \cdots, x_k,\\
&\hfill x_{j_k+1}, \cdots, x_{j_m}\}\\
=&E_p\tilde{h}(x_1, \cdots, x_k, x_{j_k+1}\tilde{h}_k(x_1, \cdots, x_k)\\
=&E_p\tilde{h}_k(x_1, \cdots, x_k)E_P(E_p\tilde{h}(x_1, \cdots, x_k, x_{j_k+1}|x_1, \cdots, x_k))\\
=&E_p\tilde{h}_k^2(x_1, \cdots, x_k)\\
=&\text{Var}_p h_k(x_1, \cdots, x_k) = \xi_k \qquad (27)
\end{aligned}
$$

# Proof

$$
\begin{aligned}
\mathrm{Var}_p(U_n) &= E_p(U_n - E(U_n))^2 \\
&= \binom{n}{m}^{-2} \sum_c \sum_c E_p \tilde{h}(x_1, \cdots, x_{i_m}) \tilde{h}(x_1, \cdots, x_{j_m}) \\
&\overset{(27)}{=} \binom{n}{m}^{-2} \sum_{k=1}^{m} \sum_{\#\mathrm{of}\{i_1,\cdots,i_m\}\cap\{j_1,\cdots,j_m\}=k} E_p \tilde{h}(x_1, \cdots, x_{i_m}) \tilde{h}(x_1, \cdots, \\
&= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m} \binom{m}{k} \binom{n-m}{m-k} \xi_k \\
&\to (26)
\end{aligned}
$$

# Proof

- (i) and (ii) can be derived from (26) and the fact that
$$0 = \xi_0 \le \xi_1 \le \xi_2 \le \cdots, \le \xi_m = \mathrm{Var}_p(h)$$
where $\xi_k \le \xi_{k+1}$ for $k = 1, \cdots, m-1$ are implied by Jensen's inequality for conditional expectation.
- To appreciate (iii), note from (26) that
$$\begin{aligned}&\binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \xi_k \\ =&\xi_k \left( \frac{m!}{k!(m-k)!} \right) \frac{k! \left\{(n-m)!\right\}^2}{n!(n-2m+k)!} \\ =&\xi_k \binom{m}{k}^2 k! \frac{(n-m)\cdots(n-2m+k+1)}{n(n-1)\cdots(n-m+1)}\end{aligned}$$
where the last factor is of the order $O(\frac{1}{n^k})$.

# Proof

- For other terms in (26), as

$$\binom{n}{m}^{-1}\binom{m}{j}\binom{n-m}{m-j} = \left\{\binom{m}{j}\right\}^2 j!\frac{\{(n-m)!\}^2}{n!(n-2m+j)!}$$
$$= \left\{\binom{m}{j}\right\}^2 j!\frac{(n-m)\cdots(n-2m+j+1)}{n(n-1)\cdots(n-m+1)}$$
$$\sim \frac{1}{n!} = O(\frac{1}{n^{k+1}}) \quad \text{for} \quad j \geq k+1$$

$$\mathrm{Var}(U_n) = \frac{k!\binom{m}{k}^2 \xi_k}{n^k} + O(\frac{1}{n^{k+1}}) \qquad \square$$

- The leading order of $\mathrm{Var}(U_n)$ is $\frac{1}{n^k}$ where $k$ is the first $\xi_k \neq 0$, which determines the rate of convergence of $U_n - E(U_n)$ to 0, as shown in the next theorem.
- See *Shao §3.2* for examples.

# Asymptotic Normality of U-Statistics

- U-Statistic is NOT a sum of independent r.vs even $X_1, \cdots, X_n$ are independent when $m > 1$, which prevents the use of CLTs for independent r.vs directly.
- The idea now is to find a projection of $U_n$ on $X_1, \cdots, X_n$ respectively, by taking $\mathrm{E}(U_n|X_i), \ i = 1, \cdots, n$. Let $\widetilde{U}_n = \mathrm{E}U_n + \sum_{i=1}^{n} \{\mathrm{E}(U_n|X_i) - \mathrm{E}U_n\}$ which is i.i.d (or independent) which admit CLT. So If we can show $U_n - \widetilde{U}_n$ is negligible, then we can use Slutsky to establish AN of $U_n$.

## Definition 9.3

Let $U_n$ be a U-statistic based on sample $\{X_1, \cdots, X_n\}$. The projection of $U_n$ on $\{x_1, \cdots, x_n\}$ is:

$$\widetilde{U}_n = \mathrm{E}U_n + \sum_{i=1}^{n} \{\mathrm{E}(U_n|X_i) - \mathrm{E}U_n\} := \theta + \sum_{i=1}^{n} \{\varphi_n(X_i) - \theta\} \tag{28}$$

where $\varphi_n(X_i) = \mathrm{E}(U_n|X_i)$.

If $\{X_i\}$ are i.i.d. (or independent), then $\{\varphi_n(X_i)\}$ are i.i.d. (or independent) too. Clearly, $\mathrm{E}\widetilde{U}_n = \mathrm{E}U_n = \theta$ $(= \theta_n$ if $h = h_n)$.

## Lemma 9.4

*Let $U_n$ be a U-Statistic with* $\mathrm{var}(U_n) < \infty$ *for each n. Then:*

$$\mathrm{E}(U_n - \widetilde{U}_n)^2 = \mathrm{var}(U_n - \widetilde{U}_n) = \mathrm{var}(U_n) - \mathrm{var}(\widetilde{U}_n)$$

The proof is based on $\mathrm{cov}(U_n, \widetilde{U}_n) = \mathrm{var}(\widetilde{U}_n)$ which is given in Shao p179.

# Asymptotic Normality of U-Statistics

## Theorem 9.5

*Let $U_n$ be a U-Statistic given in Def 9.3 based on i.i.d. $\{X_i\}_{i=1}^n$ with* $\mathrm{E}h^2(X_1, \cdots, X_m) < \infty$.

(i) *If* $\xi_1 = \mathrm{var}(\tilde{h}_1(X)) > 0$, *then:*

$$\sqrt{n}\left(U_n - \mathrm{E}U_n\right) \xrightarrow{d.} N(0, m^2\xi_1)$$

(ii) *If* $\xi_1 = 0$ *but* $\xi_2 > 0$, *then:*

$$n\left(U_n - \mathrm{E}U_n\right) \xrightarrow{d.} \frac{m(m-1)}{2}\sum_{j=1}^{\infty}\lambda_j\left(\chi_{1,j}^2 - 1\right)$$

*where* $\{\chi_{1,j}^2\}_{j\geq 1}$ *are i.i.d.* $\chi_1^2$ *r.vs and* $\lambda_j$ *are constant satisfying* $\sum_{i=1}^{\infty}\lambda_j^2 = \xi_2$.

# Proof

(i) only. See Serfling (1980) for (ii). Consider:

$$\mathrm{E}(U_n|X_1) = \mathrm{E}\begin{pmatrix} n \\ m \end{pmatrix}^{-1} \sum_{\ell} \mathrm{E}\left(h(X_{i_1}, \cdots, X_{i_m})|X_1\right)$$
$$= \begin{pmatrix} n \\ m \end{pmatrix}^{-1} \left\{ \sum_{\ell_1} \mathbb{E}\left(h(X_{i_1}, \cdots, X_{i_m})|X_1\right) + \sum_{\ell_2} \theta \right\}$$

where $\ell_1$ is all the combinations of $(i_1, \cdots, i_m)$ which contains 1 and $\ell_2$ is other combinations of $(i_1, \cdots, i_m)$ which does not contain 1. It is easy to check:

$$|c_1| = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix}, \quad |c_2| = \begin{pmatrix} n-1 \\ m \end{pmatrix}$$

Hence,

$$\mathrm{E}(U_n|X_1) = \frac{m!(n-m)!}{n!}\left[\frac{(n-1)!}{(m-1)!(n-m)!}h_1(X_1) + \frac{(n-1)!}{m!(n-m-1)!}\theta\right]$$
$$= \frac{m}{n}h_1(X_1) + \frac{n-m}{n}\theta$$

# Proof

Subsequently,

$$\widetilde{U}_n = \theta + \sum_{i=1}^{n} \left\{ \frac{m}{n} h_1(X_i) + \frac{n-m}{n}\theta - \theta \right\}$$
$$= \theta + \frac{m}{n} \sum_{i=1}^{n} \{h_1(X_i) - \theta\} = \theta + \frac{m}{n} \sum_{i=1}^{n} \tilde{h}_1(X_i)$$

From the CLT for i.i.d. r.vs, as $\mathrm{E}\tilde{h}_1^2(X_1) < \infty$, which means:

$$\sqrt{n}\left(\widetilde{U}_n - \theta\right) \xrightarrow{d.} N(0, m^2\xi_1)$$

if $\xi_1 > 0$ since $\mathrm{var}(\widetilde{U}_n) = m^2\xi_1/n$.

# Proof

On the other hand, by Lemma 9.4,

$$\mathrm{E}(U_n - \widetilde{U}_n)^2 = \mathrm{var}(U_n) - \mathrm{var}(\widetilde{U}_n)$$
$$\overset{\text{Thm 9.2}(iii)}{=\!=\!=} \frac{m^2\xi_1}{n} - \frac{m^2\xi_1}{n} + O(n^{-2}) = O(n^{-2})$$

Hence,

$$\mathrm{P}\left(\sqrt{n}|U_n - \widetilde{U}_n| > \epsilon\right) \leq \frac{n\mathrm{E}(U_n - \widetilde{U}_n)^2}{\epsilon^2} = O(n^{-1}) \to 0$$

i.e. $\sqrt{n}(U_n - \widetilde{U}_n) = o_p(1)$. AS a result,

$$\sqrt{n}(U_n - \theta) = \sqrt{n}(\widetilde{U}_n - \theta) + \sqrt{n}(U_n - \widetilde{U}_n)$$
$$= \sqrt{n}(\widetilde{U}_n - \theta) + o_p(1) \xrightarrow{d.} N(0, m^2\xi_1)$$

# Example

Suppose $X_1, \cdots, X_n$ i.i.d. $P$ with $\mathrm{E}_P X_i = \mu$ and $\mathrm{var}_P(X_i) = \sigma^2 > 0$. Let:

$$U_n = \left( \begin{array}{c} n \\ 2 \end{array} \right)^{-1} \sum_{1 \le i_1 < i_2 \le n} X_{i_1} X_{i_2}$$

i.e. $h(x_1, x_2) = x_1 x_2$, $\theta = \mathrm{E} U_n = \mathrm{E} h(X_1, X_2) = \mu^2$. Then,

$$h_1(x) = \mathrm{E}\left\{ h(X_1, X_2) | X_1 = x \right\} = \mathrm{E}\left\{ X_1 X_2 | X_1 = x \right\} = x\mu$$
$$\tilde{h}_1(x) = x\mu - \mu^2 = (x - \mu)\mu$$

and $\xi_1 = \mathrm{var}(\tilde{h}_1(X)) = \mathrm{E}\tilde{h}_1^2(X) = \mu^2\sigma^2 = 0$ iff $\mu = 0$.
Furthermore, since $\tilde{h}_2(x_1, x_2) = \tilde{h}(x_1, x_2) = x_1 x_2 - \mu^2$,

$$\begin{aligned} \xi_2 &= \mathrm{E}\left( (X_1 X_2 - \mu^2) \right)^2 = \mathrm{var}(X_1 X_2) = \mathrm{var}\left( \mathrm{E}(X_1 X_2 | X_1) \right) + \mathrm{E}\,\mathrm{var}(X_1 X_2 | X_1) \\ &= \mathrm{var}(X_1 \mu) + \mathrm{E}(X_1^2 \sigma^2) = \sigma^2 \mu^2 + \sigma^2(\sigma^2 + \mu^2) = \sigma^2(\sigma^2 + 2\mu^2) > 0 \end{aligned}$$

# Example

If $\mu \neq 0$, from the non-degenerated version of CLT,

$$\sqrt{n}(U_n - \mu^2) \xrightarrow{d.} N(0, 4\xi_1) \stackrel{d.}{=\!=} N(0, 4\mu^2\sigma^2)$$

If $\mu = 0$, since $U_n = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} X_{i_1} X_{i_2}$, we have:

$$\bar{X}_n^2 = \frac{1}{n^2} \sum_{i_1, i_2 = 1}^{n} X_{i_1} X_{i_2} = \frac{1}{n^2} \left[ n(n-1) U_n + \sum_{i=1}^{n} X_i^2 \right]$$

Note that $\sqrt{n}\bar{X}_n \xrightarrow{d.} N(0, \sigma^2)$, $n\bar{X}_n^2/\sigma^2 \xrightarrow{d.} \chi_1^2$ and $\frac{1}{n-1} \sum_{i=1} X_i^2 \xrightarrow{p.} \sigma^2$, by Slutsky Theorem,

$$nU_n = \frac{n}{n-1} n\bar{X}_n^2 - \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 \xrightarrow{d.} \sigma^2 \left( \chi_1^2 - 1 \right)$$

# Extensions

Suppose $X_1, \cdots, X_n$ i.i.d. $F$ with *pdf* $f$ and kernel estimation of $f$ with the kernel $K$ and bandwidth $b$:

$$\hat{f}_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b_n}\right) \triangleq \frac{1}{n} \sum_{i=1}^{n} K_{b_n}(x - x_i)$$

where $K_{b_n}(t) = \frac{1}{b_n} K\left(\frac{t}{b_n}\right)$. Assume $b_n \to 0\,, nb_n \to \infty$ as $n \to \infty$.
Consider WT Test:

$$H_0 : f = f_\theta \quad \textit{where } f_\theta \textit{ be a parameter pdf}\,.$$

$\hat{\theta}_n$ be a $\sqrt{n}$-consistent estimation of $\theta$ under $H_0$, i.e. $\hat{\theta}_n - \theta = O_p(n^{-1/2})$, for instance the MLE.

# Extensions

A natural test statistic is

$$\begin{aligned} T_n &= \int \left\{\hat{f}_n(x) - f_{\hat{\theta}_n}(x)\right\}^2 dx \\ &= \int \left\{\hat{f}_n(x) - E\hat{f}_n(x)\right\}^2 dx + \int \left\{E\hat{f}_n(x) - f_{\hat{\theta}_n}(x)\right\}^2 dx \\ &\quad + 2\int \left\{\hat{f}_n(x) - E\hat{f}_n(x)\right\}\left\{E\hat{f}_n(x) - f_{\hat{\theta}_n}(x)\right\} dx \\ &=: T_{n_1} + T_{n_2} + T_{n_3} \end{aligned}$$

The last two terms $T_{n_2}$ and $T_{n_3}$ at most determines the asymptotic mean of $T_n$. Let $\sigma_K^2 := \int u^2 K(u) du$:

$$T_{n_1} = \frac{1}{n^2} \sum_{i,j} \int \{K_{b_n}(x - x_i) - \mu_n(x)\}\{K_{b_n}(x - x_j) - \mu_n(x)\} dx$$

# Extensions

where

$$\mu_n(x) = EK_{b_n}(x - x_i) = f(x) + \frac{1}{2} b_n^2 f''(x) \sigma_K^2 + \cdots$$

Hence,

$$T_{n_1} = \frac{2}{n(n-1)} \sum_{i<j} \int h_n(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} \int \{K_{b_n}(x - x_i) - \mu_n(x)\}^2 \, dx$$
$$=: T_{n_{11}} + T_{n_{12}}$$

$T_{n_{12}}$ contribute to the mean only.

where

$$h_n(x_1, x_2) = \frac{n-1}{n} \int \{K_{b_n}(y - x_1) - \mu_n(y)\} \{K_{b_n}(y - x_2) - \mu_n(y)\} \, dy$$

$h_n$ be symmetric and depend on n.
Hence, we can consider the following question only:

$$U_n := \binom{n}{m}^{-1} \sum_{\ell} h_n(x_{i_1}, \cdots, x_{i_m}) \quad \text{with IID } \{x_i\}_{i=1}^n$$

Let $X_1, \ldots, X_n$ be a random independent sample from a distribution function $F(x)$, $x \in \mathbb{R}$. The empirical distribution function (EDF) is

$$\mathbb{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq t\}.$$

which is a natural estimator for the unknown distribution $F$. Note that $n\mathbb{F}_n(t)$ is binomially distributed with mean $nF(t)$, thus $\mathbb{F}_n(t)$ is unbiased.

### Classical LLN or CLT for EDF

- By the SLLN, $\mathbb{F}_n(t)$ is also consistent: $\mathbb{F}_n(t) \overset{as}{\to} F(t), \quad \forall\, t.$ .
- The centered and scaled version of the empirical measure
  $\mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf(X_i)).$
- Let $\mathcal{F}$ be equal to the collection of all indicator functions of the form $f_t = 1_{(-\infty,t]}$, with $t \in \mathbb{R}$. By the CLT: $\mathbb{G}_n f_t \rightsquigarrow N(0, F(t)(1 - F(t)))$.

## Uniform LLN for EDF

The Glivenko-Cantelli theorem extends the LLN for EDF and gives uniform convergence

$$\|\mathbb{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \stackrel{as}{\to} 0.$$

- **Motivation** 1. Historically, empirical process theory has one of its roots in the study of goodness-of-fit statistics.

[The first goodness-of-fit statistic is Pearson's chi-square statistic. It is performed by discretely binning a continuous distribution into a more tractable multinomial distribution. However, the discretization in chi-square statistic leads to a loss in statistical power. To remedy this problem, Kolmogorov introduced the statistics

$$K_n = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)|$$

to directly measure the maximum functional distance between $\mathbb{F}_n(t)$, $F(t)$.]

# Uniform CLT for EDF

## Kolmogorov distribution

The Kolmogorov distribution is the distribution of the random variable

$$K = \sup_{t \in [0,1]} |B(t)|$$

where $B(t)$ is the Brownian bridge. The cumulative distribution function of $K$ is given by

$$\Pr(K \leq x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}$$

## Uniform CLT for EDF

Under null hypothesis that the sample comes from the distribution $F(x)$

$$\sqrt{n}K_n \overset{n\to\infty}{\longrightarrow} \sup_{t} |B(F(t))|.$$

Theory of Empirical Processes aims to establish the uniform convergence.

# Motivations of Empirical Process

- **Motivation** 2. The uniform convergence condition in Consistency of $M$- and $Z$-estimator is hard to check.

## Theorem 10.1 (Consistency of $M$-estimator)

*Let $M_n$ be random functions and let $M$ be a fixed function of $\theta$ such that for every $\varepsilon > 0$, if we have conditions:*

*C1.* **Uniformly convergence**: $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$;

*C2. Well-separation; C3.The $\{\hat{\theta}_n\}$ satisfies nearly maximization condition. Then $\hat{\theta}_n \xrightarrow{P} \theta$.*

## Theorem 10.2 (Consistency of $Z$-Estimator)

*Let $\Psi_n$ be random vector-valued functions and let $\Psi$ be a fixed vector- valued function of $\theta$ such that for every $\varepsilon > 0$, if we have :*

*C1*.* **Uniformly convergence**: $\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \to 0$;

*C2*. Well-separation; C3*.The $\{\hat{\theta}_n\}$ satisfies nearly zero condition. So, $\hat{\theta}_n \xrightarrow{P} \theta$.*

- **Motivation** 3. When controlling the non-independent summation of a function of the random sample indexed by a common estimator $\hat{\theta}$. It false to use any sort of classical LLN or CLT.

Given an estimator $\hat{\theta}$, we want to study its asymptotic properties for summation some function $f_{\hat{\theta}}(X_i)$,

$$\frac{1}{n}\sum_{i=1}^{n}[f_{\hat{\theta}}(X_i) - \mathrm{E}f_{\theta_0}(X_i)], \text{ is the "true" parameter.}$$

## A Possible Solution

Prove a uniform version (the suprema of empirical processes) for all possible $\hat{\theta}$ on a set $K$, which is usually stronger than what is needed.

$$\frac{1}{n}\sum_{i=1}^{n}[f_{\hat{\theta}}(X_i) - \mathrm{E}f_{\theta_0}(X_i)] \leq \sup_{\theta_0 \in K} |\frac{1}{n}\sum_{i=1}^{n}[f_{\theta_0}(X_i) - \mathrm{E}f_{\theta_0}(X_i)]|.$$

Fortunately, the summation in the sup enjoy independence.