Point Estimation: properties of estimators

- finite-sample properties (CB 7.3)
- large-sample properties (CB 10.1)

1 FINITE-SAMPLE PROPERTIES

How an estimator performs for finite number of observations $n$.

Estimator: $W$
Parameter: $\theta$

Criteria for evaluating estimators:

- **Bias**: does $EW = \theta$?
- Variance of $W$ (you would like an estimator with a smaller variance)

**Example**: $X_1, \ldots, X_n \sim \text{i.i.d.} (\mu, \sigma^2)$

Unknown parameters are $\mu$ and $\sigma^2$.

Consider:

- $\hat{\mu}_n = \frac{1}{n} \sum_{i} X_i$, estimator of $\mu$
- $\hat{\sigma}^2_n = \frac{1}{n} \sum_{i} (X_i - X_n)^2$, estimator of $\sigma^2$.

**Bias**: $E\hat{\mu}_n = \frac{1}{n} \cdot n\mu = \mu$. So **unbiased**.

**Var** $\hat{\mu}_n = \frac{1}{n^2} n\sigma^2 = \frac{1}{n} \sigma^2$.

$$E\hat{\sigma}^2 = E\left(\frac{1}{n} \sum_{i} (X_i - \bar{X}_n)^2\right)$$

$$= \frac{1}{n} \cdot \sum_{i} \left(EX_i^2 - 2EX_i\bar{X}_n + E\bar{X}_n^2\right)$$

$$= \frac{1}{n} \cdot n \left[ (\mu^2 + \sigma^2) - 2(\mu^2 + \frac{\sigma^2}{n}) + \frac{\sigma^2}{n} + \mu^2 \right]$$

$$= n - \frac{1}{n} \sigma^2.$$
Hence it is biased.

To fix this bias, consider the estimator 
\[ s_n^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \]
and 
\[ E s_n^2 = \sigma^2 \] (unbiased).

Mean-squared error (MSE) of \( W \) is \( E(W - \theta)^2 \). Common criterion for comparing estimators.

Decompose: \( MSE(W) = VW + (EW - \theta)^2 = \text{Variance} + (\text{Bias})^2 \).

Hence, for an unbiased estimator: \( MSE(W) = VW \).

Example: \( X_1, \ldots, X_n \sim U[0, \theta] \). \( f(X) = \frac{1}{\theta}, x \in [0, \theta] \).

- Consider estimator \( \hat{\theta}_n \equiv 2\bar{X}_n \).
  \[ E\hat{\theta}_n = 2 \frac{1}{n} \cdot E \sum_i X_i = 2 \cdot \frac{1}{n} \cdot \frac{1}{2} n \cdot \theta = \theta. \] So unbiased
  \[ MSE(\hat{\theta}_N) = \frac{4}{n^2} \sum_i VX_i = \frac{\theta^2}{3n}. \]

- Consider estimator \( \tilde{\theta}_n \equiv \max(X_1, \ldots, X_n) \).
  In order to derive moments, start by deriving CDF:
  \[ P(\tilde{\theta}_n \leq z) = P(X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z) \]
  \[ = \prod_{i=1}^{n} P(X_i \leq z) \]
  \[ = \left\{ \begin{array}{ll} \left( \frac{z}{\theta} \right)^n & \text{if } z \leq \theta \\ 1 & \text{otherwise} \end{array} \right. \]
  Therefore \( f_{\tilde{\theta}_n}(z) = n \cdot \left( \frac{z}{\theta} \right)^{n-1} \frac{1}{\theta}, \) for \( 0 \leq x \leq \theta \).

  \[ E(\tilde{\theta}_n) = \int_0^{\theta} z \cdot n \cdot \left( \frac{z}{\theta} \right)^{n-1} \frac{1}{\theta} dz \\ = \frac{n}{\theta^n} \int_0^{\theta} z^n dz = \frac{n}{n+1} \theta. \]

  Bias(\( \tilde{\theta}_n \)) = \(-\theta/(n + 1)\)
  \[ E(\tilde{\theta}_n^2) = \frac{n}{\theta^2} \int_0^{\theta} z^{n+1} dz = \frac{n}{n+2} \theta^2. \]

  Hence \( V \tilde{\theta}_n = \theta^2 \left( \frac{n}{n+2} - \left( \frac{n}{n+1} \right)^2 \right) = \theta^2 \frac{n}{(n+2)(n+1)}. \)

  Accordingly, \( MSE = \frac{2\theta^2}{(n+2)(n+1)} \).
Continue the previous example. Redefine $\tilde{\theta}_n = \frac{n+1}{n} \max(X_1, \ldots, X_n)$. Now both estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ are unbiased.

Which is better? $V\hat{\theta}_n = \frac{\theta^2}{3n} = O(1/n)$.

$V\tilde{\theta}_n = (\frac{n+1}{n})^2 V(\max(X_1, \ldots, X_n)) = \theta^2 \left( \frac{1}{n(n+2)} \right) = O(1/n^2)$.

Hence, for $n$ large enough, $\tilde{\theta}_n$ has a smaller variance, and in this sense it is “better”.

Best unbiased estimator: if you choose the best (in terms of MSE) estimator, and restrict yourself to unbiased estimators, then the best estimator is the one with the lowest variance.

A best unbiased estimator is also called the “Uniform minimum variance unbiased estimator” (UMVUE).

Formally: an estimator $W$ is a UMVUE of $\theta$ satisfies:

(i) $E_\theta W = \theta$, for all $\theta$ (unbiasedness)

(ii) $V_\theta W \leq V_\theta \tilde{W}$, for all $\theta$, and all other unbiased estimators $\tilde{W}$.

The “uniform” condition is crucial, because it is always possible to find estimators which have zero variance for a specific value of $\theta$.

It is difficult in general to verify that an estimator $W$ is UMVUE, since you have to verify condition (ii) of the definition, that $VW$ is smaller than all other unbiased estimators.

Luckily, we have an important result for the lowest attainable variance of an estimator.

• Theorem 7.3.9 (Cramer-Rao Inequality): Let $X_1, \ldots, X_n$ be a sample with joint pdf $f(X|\theta)$, and let $W(X)$ be any estimator satisfying

(i) $\frac{d}{d\theta} E_{\theta} W(X) = \int \frac{\partial}{\partial \theta} \left[ W(X) \cdot f(X|\theta) \right] dX$;

(ii) $V_\theta W(X) < \infty$.

Then

$$V_\theta W(X) \geq \frac{\left( \frac{d}{d\theta} E_{\theta} W(X) \right)^2}{E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2}.$$
The RHS above is called the Cramer-Rao Lower Bound.

Proof: (Cramer, Mathematical Methods of Statistics, p. 475ff) Start with the following manipulation of the equality $E_\theta W(X) = \int W(X)f(X|\theta)dX$:

$$
\frac{d}{d\theta}E_\theta W(X) = \frac{d}{d\theta} \int W(X)f(X|\theta)dX
= \int W(X)\frac{\partial}{\partial\theta}f(X|\theta)dX
= \int (W(X) - E_\theta W(X))\frac{\partial}{\partial\theta}f(X|\theta)dX \quad \text{(note } \int E_\theta W(X)\frac{\partial}{\partial\theta}f(X|\theta)dX = 0) \\
= \int W(X) - E_\theta W(X) \left( \frac{\partial}{\partial\theta} \log f(X|\theta) \right) f(X|\theta)dX
$$

Applying the Cauchy-Schwarz inequality, we have

$$
\left[ \frac{d}{d\theta}E_\theta W(X) \right]^2 \leq \text{Var}_\theta W(X) \cdot E_\theta \left[ \frac{\partial}{\partial\theta} \log f(X|\theta) \right]^2
$$
or

$$
\text{Var}_\theta W(X) \geq \frac{\left[ \frac{d}{d\theta}E_\theta W(X) \right]^2}{E_\theta \left[ \frac{\partial}{\partial\theta} \log f(X|\theta) \right]^2}.
$$

- The LHS of condition (i) above is $\frac{d}{d\theta} \int W(\vec{X})f(X|\theta)dX$, so by Leibniz’ rule, this condition rules out cases where the support of $X$ is dependent on $\theta$.

The crucial step in the derivation of the CR-bound is the interchange of differentiation and integration which implies

$$
E_\theta \frac{\partial}{\partial\theta} \log f(\vec{X}|\theta) = \int \frac{1}{f(\vec{X}|\theta)} \frac{\partial f(\vec{X}|\theta)}{\partial\theta} f(\vec{X}|\theta)dx
= \frac{\partial}{\partial\theta} \int f(\vec{X}|\theta)dx
= \frac{\partial}{\partial\theta} \cdot 1 = 0
$$

- (skip) The above derivation is noteworthy, because

$$
\frac{\partial}{\partial\theta} \log f(\vec{X}|\theta) = 0
$$
is the FOC of maximum likelihood estimation problem.
In the i.i.d. case, this becomes the sample average
\[ \frac{1}{n} \sum_i \frac{\partial}{\partial \theta} \log f(x_i|\theta) = 0. \]

And by the LLN:
\[ \frac{1}{n} \sum_i \frac{\partial}{\partial \theta} \log f(x_i|\theta) \xrightarrow{p} E_{\theta_0} \frac{\partial}{\partial \theta} \log f(x_i|\theta), \]

where \( \theta_0 \) is the true value of \( \theta_0 \). This shows that maximum likelihood estimation of \( \theta \) is equivalent to GMM estimation based on the moment condition
\[ E_{\theta_0} \frac{\partial}{\partial \theta} \log f(x_i|\theta) = 0. \] (2)

From the previous derivation, we see that this condition holds at the true value \( \theta = \theta_0 \). (What happens when we evaluate LHS of (2) at values \( \theta \neq \theta_0 \)?)

- In the iid case, the CR lower bound can be simplified

**Corollary 7.3.10**: if \( X_1, \ldots, X_n \sim \text{i.i.d. } f(X|\theta) \), then
\[ V_\theta W(\bar{X}) \geq \frac{\left( \frac{\partial}{\partial \theta} E_{\theta} W(\bar{X}) \right)^2}{n \cdot E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2}. \] (3)

- Up to this point, Cramer-Rao results not operational, because the estimator \( W(\bar{X}) \) is on both sides of the inequality. However, for an unbiased estimator, \( E_{\theta} W(\bar{X}) = \theta \), so that \( \frac{\partial}{\partial \theta} E_{\theta} W(\bar{X}) = 1 \). Then
\[ V_\theta W(\bar{X}) \geq \frac{1}{E_{\theta} \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2}. \]

Hence, if we find an unbiased estimator with variance corresponding to CRLB, then we’ve identified a UMVUE.

- **Example**: \( X_1, \ldots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2) \).

What is CRLB for an unbiased estimator of \( \mu \)?

Unbiased \( \rightarrow \) numerator =1.
\[
\log f(x|\theta) = \log \sqrt{2\pi} - \log \sigma - \frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2
\]
\[
\frac{\partial}{\partial \mu} \log f(x|\theta) = - \left( \frac{x - \mu}{\sigma} \right) \cdot \left( -\frac{1}{\sigma} \right) = \frac{x - \mu}{\sigma^2}
\]
\[
E \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 = E \left( (X - \mu)^2 \right) = \frac{1}{\sigma^4} V X = \frac{1}{\sigma^2}.
\]

Hence the CRLB = \( \frac{1}{n} \cdot \frac{1}{\sigma^2} = \frac{\sigma^2}{n} \).

This is the variance of the sample mean, so that the sample mean is a UMVUE for \( \mu \).

**Lemma 7.3.11 (Information inequality):** if \( f(X|\theta) \) satisfies

\[
(*) \quad \frac{d}{d\theta} E_\theta \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right) = \int \frac{\partial}{\partial \theta} \left[ \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right) f(X|\theta) \right] dx,
\]

then
\[
E_\theta \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right).
\]

**Proof:**

LHS of (*): Using Eq. (1) above, we get that LHS of (*) =0.

RHS of (*):

\[
\int \frac{\partial}{\partial \theta} \left[ \left( \frac{\partial}{\partial \theta} \log f \right) f \right] dx
\]
\[
= \int \frac{\partial^2 \log f}{\partial \theta^2} f dx + \int \frac{1}{f} \left( \frac{\partial f}{\partial \theta} \right)^2 dx
\]
\[
= E \frac{\partial^2 \log f}{\partial \theta^2} + E \left( \frac{\partial \log f}{\partial \theta} \right)^2.
\]

Putting the LHS and RHS together yields the desired result. ■

- The LHS of the above condition (*) is just \( \frac{d}{d\theta} \int (\frac{\partial}{\partial \theta} \log f(X|\theta)) f(X|\theta) dX \). As before, the crucial step is the interchange of differentiation and integration.

- The CRLB is a feature of a particular model (or sampling distribution \( f(X|\theta) \)), and not a feature of the estimator.
• **Example:** for the previous example, consider CRLB for unbiased estimator of \( \sigma^2 \).

We can use the information inequality, because condition (*) is satisfied for the normal distribution. Hence:

\[
\frac{\partial}{\partial \sigma^2} \log f(x|\theta) = -\frac{1}{2 \sigma^2} + \frac{1}{2} \frac{(x - \mu)^2}{\sigma^4}
\]

\[
\frac{\partial}{\partial \sigma^2} \left( \frac{\partial}{\partial \sigma^2} \log f(x|\theta) \right) = \frac{1}{2 \sigma^4} - \frac{(x - \mu)^2}{\sigma^6}
\]

\[-E \left( \frac{\partial}{\partial \sigma^2} \left( \frac{\partial}{\partial \sigma^2} \log f(x|\theta) \right) \right) = - \left( \frac{1}{2 \sigma^4} - \frac{1}{\sigma^4} \right) = \frac{1}{2 \sigma^4}.
\]

Hence the CRLB is \( \frac{2 \sigma^4}{n} \).

• **Example:** \( X_1, \ldots, X_n \sim U[0, \theta] \). Check conditions for CRLB for an unbiased estimator \( W(\vec{X}) \) of \( \theta \).

\[
\frac{d}{d\theta} EW(\vec{X}) = 1 \text{ (because it is unbiased)}
\]

\[
\int \frac{\partial}{\partial \theta} \left[ W(\vec{X}) f(\vec{X}|\theta) \right] d\vec{X} = \int W(\vec{X}) \cdot \left( -\frac{1}{\theta^2} \right) d\vec{X} \neq \frac{d}{d\theta} EW(\vec{X}) = 1
\]

Hence, condition (i) of theorem not satisfied.

• **But when can CRLB (if it exists) be attained?**

Go back to derivation of CRLB, by the Cauchy-Schwarz inequality we have:

\[
\left[ \int (W(X) - \theta) \left( \frac{\partial \log L}{\partial \theta} \right) dF(X|\theta) \right]^2 \leq \int (W(X) - \theta)^2 dF(X|\theta) \int \left( \frac{\partial \log L}{\partial \theta} \right)^2 dF(X|\theta).
\]

Note that the equality binds when (turns out this is iff) \( \frac{\partial \log L}{\partial \theta} \) is proportional to \( W(X) - \theta \), that is:

\[
\frac{\partial \log L}{\partial \theta} = a(\theta) \ast (W(X) - \theta)
\]

where the proportionality constant \( a(\theta) \) can depend on \( \theta \) but does not depend on \( X \).

Under this condition then, the \( Var(W(X)) \) is exactly the CRLB; that is, the CRLB is attainable.

**Corollary 7.3.15:** \( X_1, \ldots, X_n \sim i.i.d. f(X|\theta) \), satisfying the conditions of CR theorem.
– Let $L(\theta|\vec{X}) = \prod_{i=1}^{n} f(X_\theta)$ denote the likelihood function.
– Estimator $W(\vec{X})$ unbiased for $\theta$
– $W(\vec{X})$ attains CRLB iff you can write
\[
\frac{\partial}{\partial \theta} \log L(\theta|\vec{X}) = a(\theta) \left[ W(\vec{X}) - \theta \right]
\]
for some function $a(\theta)$.

**Example:** $X_1, \ldots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2)$
Consider estimating $\mu$:
\[
\frac{\partial}{\partial \mu} \log L(\theta|\vec{X}) = \frac{\partial}{\partial \mu} \left( \sum_{i=1}^{n} \log f(X_\theta) \right)
= \frac{\partial}{\partial \mu} \left( \sum_{i=1}^{n} - \log \sqrt{2\pi} - \log \sigma - \frac{1}{2} \left( \frac{(X - \mu)^2}{\sigma^2} \right) \right)
= \sum_{i=1}^{n} \left( \frac{X - \mu}{\sigma^2} \right)
= \frac{n}{\sigma^2} \cdot (\bar{X}_n - \mu).
\]
Hence, CRLB can be attained (in fact, we showed earlier that CRLB attained by $\bar{X}_n$).
Additional examples: Poisson, Binomial.

■■■

**Loss function optimality**

Let $\vec{X} \sim f(\vec{X}|\theta)$.

Consider a *loss function* $\mathcal{L}(\theta, W(\vec{X}))$, taking values in $[0, +\infty)$, which penalizes you when your $W(\vec{X})$ estimator is “far” from the true parameter $\theta$. Note that $\mathcal{L}(\theta, W(\vec{X}))$ is a random variable, since $\vec{X}$ (and $W(\vec{X})$) are random.

Consider estimators which *minimize expected loss*: that is
\[
\min_{W(\cdots)} E_{\theta} \mathcal{L}(\theta, W(\vec{X})) \equiv \min_{W(\cdots)} R(\theta, W(\cdots))
\]
where $R(\theta, W(\cdots))$ is the *risk function*. (Note: the risk function is not a random variable, because $\vec{X}$ has been integrated out.)
Loss function optimality is a more general criterion than minimum MSE. In fact, because $MSE(W(\bar{X})) = E_\theta \left( W(\bar{X}) - \theta \right)^2$, the MSE is actually the risk function associated with the \textit{quadratic loss function} $L(\theta, W(\bar{X})) = (W(\bar{X}) - \theta)^2$.

Other examples of loss functions:

- **Absolute error loss**: $|W(\bar{X}) - \theta|
- **Relative quadratic error loss**: $\frac{(W(\bar{X}) - \theta)^2}{|\theta| + 1}$

The exercise of minimizing risk takes a given value of $\theta$ as given, so that the minimized risk of an estimator depends on whichever value of $\theta$ you are considering. You are typically interested in an estimator which does well regardless of which value of $\theta$ you are considering. (Analogous to the focus on the \textit{uniform} minimal variance.)

For this different problem, you want to consider a notion of risk which does not depend on $\theta$. Two possible criteria are:

- **“Average” risk**:
  \[
  \min_{W(\cdots)} \int R(\theta, W(\cdots)) h(\theta) d\theta.
  \]
  where $h(\theta)$ is some weighting function across $\theta$. (In a Bayesian interpretation, $h(\theta)$ is a prior density over $\theta$.)

- **Minmax criterion**:
  \[
  \min_{W(\cdots)} \max_{\theta} R(\theta, W(\cdots)).
  \]
  Here you choose the estimator $W(\cdots)$ to minimize the maximum risk $= \max_\theta R(\theta, W(\cdots))$, where $\theta$ is set to the “worse” value. So minmax optimizer is the best that can be achieved in a “worst-case” scenario. This can represent a decision criterion for “ambiguous” settings where you have little information about the possible values of $\theta$. 


2 LARGE SAMPLE PROPERTIES OF ESTIMATORS

Large-sample properties: exploit LLN, CLT

Consider data \( \{X_1, X_2, \ldots \} \) by which we construct a sequence of estimators \( W_n \equiv \{W(X_1), W(X_2), \ldots \} \). \( W_n \) is a random sequence.

Define: we say that \( W_n \) is consistent for a parameter \( \theta \) iff the random sequence \( W_n \) converges stochastically to \( \theta \). Strong consistency obtains when \( W_n \xrightarrow{\text{st}} \theta \). Weak consistency obtains when \( W_n \xrightarrow{\text{p}} \theta \).

For estimators like sample-means, consistency can be proved using a LLN. Now we want to consider estimators which are not sample means.

Define: an M-estimator is an estimator of \( \theta \) which a maximizer of an objective function \( Q_n(\theta) \).

Examples:

- MLE: \( Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta) \)
- Least squares: \( Q_n(\theta) = \sum_{i=1}^{n} [y_i - g(x_i; \theta)]^2 \). OLS is special case when \( g(x_i; \theta) = \alpha + X_i'\beta \).
- GMM: \( Q_n(\theta) = G_n(\theta)'W_n(\theta)G_n(\theta) \) where

\[
G_n(\theta) = \left[ \frac{1}{n} \sum_{i=1}^{n} m_1(x_i; \theta), \frac{1}{n} \sum_{i=1}^{n} m_2(x_i; \theta), \ldots, \frac{1}{n} \sum_{i=1}^{n} m_M(x_i; \theta) \right]',
\]

an \( M \times 1 \) vector of sample moment conditions, and \( W_n \) is an \( M \times M \) weighting matrix.

Notation: For each \( \theta \in \Theta \), let \( f_\theta \equiv f(x_1, \ldots, x_n; \theta) \) denote the joint density of the data for the given value of \( \theta \). For \( \theta_0 \in \Theta \), we denote the limit objective function \( Q_0(\theta) = \text{plim}_{n \to \infty, f_{\theta_0}} Q_n(\theta) \) (pointwise for each \( \theta \)).

Consistency of M-estimators  Make the following assumptions:

1. For each \( \theta_0 \in \Theta \), the limiting objective function \( Q_0(\theta) \) is uniquely maximized at \( \theta_0 \) (“identification”)
2. Parameter space $\Theta$ is a compact subset of $\mathbb{R}^K$. 

3. $Q_0(\theta)$ is continuous in $\theta$ 

4. $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$; that is:

$$
\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| \xrightarrow{P} 0.
$$

**Theorem: (Consistency of M-Estimator)** Under assumption 1,2,3,4, $\theta_n \xrightarrow{P} \theta_0$. 

**Proof:** We need to show: for any arbitrarily small neighborhood $\mathcal{N}$ containing $\theta_0$, $P(\theta_n \in \mathcal{N}) \to 1$.

For $n$ large enough, the uniform convergence conditions that, for all $\epsilon, \delta > 0$,

$$
P \left( \sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2 \right) > 1 - \delta.
$$

The event “$\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2$” implies

$$
Q_n(\theta_n) - Q_0(\theta_n) < \epsilon/2 \iff Q_0(\theta_n) > Q_n(\theta_n) - \epsilon/2 
$$

Similarly,

$$
Q_0(\theta_0) - Q_n(\theta_0) > -\epsilon/2 \Rightarrow Q_n(\theta_0) > Q_0(\theta_0) - \epsilon/2.
$$

Since $\theta_n = \arg\max_{\theta} Q_n(\theta)$, Eq. (5) implies

$$
Q_0(\theta_n) > Q_n(\theta_0) - \epsilon/2.
$$

Hence, adding Eqs. (6) and (7), we have

$$
Q_0(\theta_n) > Q_0(\theta_0) - \epsilon.
$$

So we have shown that

$$
\sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2 \Rightarrow Q_0(\theta_n) > Q_0(\theta_0) - \epsilon
$$

$$
\iff P \left( Q_0(\theta_n) > Q_0(\theta_0) - \epsilon \right) \geq P \left( \sup_{\theta \in \Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2 \right) \to 1.
$$

Now define $\mathcal{N}$ as any open neighborhood of $\mathbb{R}^K$, which contains $\theta_0$, and $\bar{\mathcal{N}}$ is the complement of $\mathcal{N}$ in $\mathbb{R}^K$. Then $\Theta \cap \bar{\mathcal{N}}$ is compact, so that $\max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta)$ exists.

Set $\epsilon = Q_0(\theta_0) - \max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta)$. Then

$$
Q_0(\theta_n) > Q_0(\theta_0) - \epsilon \Rightarrow Q_0(\theta_n) > \max_{\theta \in \Theta \cap \bar{\mathcal{N}}} Q_0(\theta)
$$

$$
\Rightarrow \theta_n \in \mathcal{N}
$$

$$
\iff P(\theta_n \in \mathcal{N}) \geq P(\theta_n \in \mathcal{N}) \Rightarrow P(Q_0(\theta_n) > Q_0(\theta_0) - \epsilon) \to 1.
$$
Since the argument above holds for any arbitrarily small neighborhood of \( \theta_0 \), we are done.

- In general, the limit objective function \( Q_0(\theta) = \lim_{n \to \infty} Q_n(\theta) \) may not be that straightforward to determine. But in many cases, \( Q_n(\theta) \) is a sample average of some sort:

\[
Q_n(\theta) = \frac{1}{n} \sum_i q(x_i|\theta)
\]

(eg. least squares, MLE). Then by a law of large numbers, we conclude that (for all \( \theta \))

\[
Q_0(\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_i q(x_i|\theta) = E_{x_i} q(x_i|\theta)
\]

where \( E_{x_i} \) denote expectation with respect to the true (but unobserved) distribution of \( x_i \).

** Let’s unpack the uniform convergence condition. Sufficient conditions for this conditions are:

1. Pointwise convergence: For each \( \theta \in \Theta \), \( Q_n(\theta) - Q_0(\theta) = o_p(1) \).
2. \( Q_n(\theta) \) is stochastically equicontinuous: for every \( \epsilon > 0, \eta > 0 \) there exists a sequence of random variable \( \Delta_n(\epsilon, \eta) \) and \( n^*(\epsilon, \eta) \) such that for all \( n > n^* \), \( P(|\Delta_n| > \epsilon) < \eta \) and for each \( \theta \) there is an open set \( \mathcal{N} \) containing \( \tilde{\theta} \) with

\[
\sup_{\theta \in \mathcal{N}} |Q_n(\tilde{\theta}) - Q_n(\theta)| \leq \Delta_n, \ \forall n > n^*.
\]

Note that both \( \Delta_n \) and \( n^* \) do not depend on \( \theta \): it is uniform result.

To understand this more intuitively, consider what we need for consistency. By continuity of \( Q_0 \), we know that \( Q_0(\theta) \) is close to \( Q_0(\theta_0) \) for \( \theta \in \mathcal{N}(\theta_0) \). By pointwise convergence, we have \( Q_n(\theta) \) converging to \( Q_0(\theta) \) for all \( \theta \). However, what we need is that even if \( Q_n(\theta) \) is not optimized by \( \theta_0 \), the optimizer \( \theta_n = \arg\max_\theta Q_n(\theta) \) should not be far from \( \theta_0 \). Pointwise convergence does not guarantee this.

For the last part, we need \( Q_n(\theta) \) to be “equally close” to \( Q_0(\theta) \) for all \( \theta \), because then the optimizers of \( Q_n \) and \( Q_0 \) cannot be too far apart. However, pointwise convergence is not enough to ensure this “equally closeness”. At any given \( n \), \( Q_n(\theta_0) \) being close to \( Q_0(\theta_0) \) does not imply this at other points. Uniform convergence ensures that at any given \( n \), \( Q_n \) and \( Q_0 \) are “equally close” at all points \( \theta \). This was exploited in the proof (eqs. (5), (6)).
allows us to argue that with high probability, \( Q_0(\theta_n) \) is within a small neighborhood of \( Q_0(\theta_0) \). By continuity of \( Q_0 \), this implies that \( \theta_n \) lies within a small neighborhood of \( \theta_0 \) with high probability.

**Asymptotic normality for M-estimators** Define the “score vector”

\[
\nabla_\theta Q_n(\theta) = \left[ \frac{\partial Q_n(\theta)}{\partial \theta_1} \Big|_{\theta = \hat{\theta}}, \ldots, \frac{\partial Q_n(\theta)}{\partial \theta_K} \Big|_{\theta = \hat{\theta}} \right]'.
\]

Similarly, define the \( K \times K \) Hessian matrix

\[
\begin{bmatrix}
\nabla_{\hat{\theta} \hat{\theta}} Q_n(\theta)
\end{bmatrix}_{i,j} = \frac{\partial^2 Q_n(\theta)}{\partial \theta_i \partial \theta_j} \Big|_{\theta = \hat{\theta}}, \quad 1 \leq i,j \leq K.
\]

Note that the Hessian is symmetric.

Make the following assumptions:

1. \( \theta_n = \arg\max_{\theta} Q_n(\theta) \overset{p}{\to} \theta_0 \)
2. \( \theta_0 \in \text{interior}(\Theta) \)
3. \( Q_n(\theta) \) is twice continuously differentiable in a neighborhood \( \mathcal{N} \) of \( \theta_0 \).
4. \( \sqrt{n} \nabla_{\theta_0} Q_n(\theta) \overset{d}{\to} N(0, \Sigma) \)
5. Uniform convergence of Hessian: there exists the matrix \( H(\theta) \) which is continuous at \( \theta_0 \) and \( \sup_{\theta \in \mathcal{N}} ||\nabla_{\theta \theta} Q_n(\theta) - H(\theta)|| \overset{p}{\to} 0. \)
6. \( H(\theta_0) \) is nonsingular

**Theorem (Asymptotic normality for M-estimator):** Under assumptions 1,2,3,4,5,

\[
\sqrt{n}(\theta_n - \theta_0) \overset{d}{\to} N(0, H_0^{-1}\Sigma H_0^{-1}).
\]

where \( H_0 \equiv H(\theta_0) \).

**Proof:** (sketch) By Assumptions 1,2,3, \( \nabla_{\theta_n} Q_n(\theta) = 0 \) (this is FOC of maximization problem). Then using mean-value theorem (with \( \hat{\theta}_n \) denoting mean value):

\[
0 = \nabla_{\theta_n} Q_n(\theta) = \nabla_{\theta_0} Q_n(\theta) + \nabla_{\theta_n \hat{\theta}_n} Q_n(\theta)(\theta_n - \theta_0)
\]

\[
\Rightarrow \nabla_{\hat{\theta}_n \hat{\theta}_n} Q_n(\theta) \sqrt{n}(\theta_n - \theta_0) = -\sqrt{n} \nabla_{\theta_0} Q_n(\theta)
\]

\[
\overset{p}{\to} H_0 \text{ (using A5)} \quad \overset{d}{\to} N(0, \Sigma) \text{ (using A4)}
\]

\[
\Leftrightarrow \sqrt{n}(\theta_n - \theta_0) \overset{d}{\to} -H(\theta_0)^{-1}N(0, \Sigma) = N(0, H_0^{-1}\Sigma H_0^{-1}). \quad \blacksquare
\]
Note: A5 is a uniform convergence assumption on the sample Hessian. Given previous discussion, it ensures that the sample Hessian $\nabla_{\theta} Q_n(\theta)$ evaluated at $\hat{\theta}_n$ (which is close to $\theta_0$) does not vary far from the limit Hessian $H(\theta)$ at $\theta_0$, which is implied by a type of “continuity” of the sample Hessian close to $\theta_0$.

2.1 Maximum likelihood estimation

The consistency of MLE can follow by application of the theorem above for consistency of M-estimators.

Essentially, as we noted above, what the consistency theorem showed above was that, for any M-estimator sequence $\theta_n$:

$$\text{plim}_{n \to \infty} \theta_n = \arg\max_{\theta} Q_0(\theta).$$

For MLE, there is a distinct and earlier argument due to Wald (1949), who shows that, in the i.i.d. case, the “limiting likelihood function” (corresponding to $Q_0(\theta)$) is indeed globally maximized at $\theta_0$, the “true value”. Thus, we can directly confirm the identification assumption of the M-estimator consistency theorem. This argument is of interest by itself.

**Argument:** (summary of Amemiya, pp. 141–142)

- Define $\hat{\theta}_n^{MLE} \equiv \arg\max_{\theta} \frac{1}{n} \sum_i \log f(x_i|\theta)$. Let $\theta_0$ denote the true value.
- By LLN: $\frac{1}{n} \sum_i \log f(x_i|\theta) \xrightarrow{P} E_{\theta_0} \log f(x_i|\theta)$, for all $\theta$ (not necessarily the true $\theta_0$).
- By Jensen’s inequality: $E_{\theta_0} \log \left( \frac{f(x|\theta)}{f(x|\theta_0)} \right) < \log E_{\theta_0} \left( \frac{f(x|\theta)}{f(x|\theta_0)} \right)$
- But $E_{\theta_0} \left( \frac{f(x|\theta)}{f(x|\theta_0)} \right) = \int \left( \frac{f(x|\theta)}{f(x|\theta_0)} \right) f(x|\theta_0) = 1$, since $f(x|\theta)$ is a density function, for all $\theta$.
- Hence:

$$E_{\theta_0} \log \left( \frac{f(x|\theta)}{f(x|\theta_0)} \right) < 0, \forall \theta$$

$$\implies E_{\theta_0} \log f(x|\theta) < E_{\theta_0} \log f(x|\theta_0), \forall \theta$$

$$\implies E_{\theta_0} \log f(x|\theta) \text{ is maximized at the true } \theta_0.$$

---

In this step, note the importance of assumption A3 in CB, pg. 516. If $x$ has support depending on $\theta$, then it will not integrate to 1 for all $\theta$. 

---

\[\text{1}\]In this step, note the importance of assumption A3 in CB, pg. 516. If $x$ has support depending on $\theta$, then it will not integrate to 1 for all $\theta$. 

---

14
This is the “identification” assumption from the M-estimator consistency theorem.

Now we introduce another idea, efficiency, which is a large-sample analogue of the “minimum variance” concept.

For the sequence of estimators $W_n$, suppose that

$$k(n)(W_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

where $k(n)$ is a polynomial in $n$. Then $\sigma^2$ is denoted the asymptotic variance of $W_n$.

In “usual” cases, $k(n) = \sqrt{n}$. For example, by the CLT, we know that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. Hence, $\sigma^2$ is the asymptotic variance of the sample mean $\bar{X}_n$.

**Definition 10.1.11:** An estimator sequence $W_n$ is asymptotically efficient for $\theta$ if

- $\sqrt{n}(W_n - \theta) \xrightarrow{d} N(0, v(\theta))$, where
- the asymptotic variance $v(\theta) = \frac{1}{E_{\theta_0}(\frac{\partial}{\partial \theta} \log f(X|\theta))^2}$

By comparison with Eq. (3), note that the asymptotic variance $\frac{1}{E_{\theta_0}(\frac{\partial}{\partial \theta} \log f(X|\theta))^2}$ is equivalent to the CRLB for one observation ($n = 1$). $I(\theta) \equiv E_{\theta_0}(\frac{\partial}{\partial \theta} \log f(X|\theta))^2$ is called the Fisher information.

Some intuition: Recall that $N(0, 1/I(\theta))$ is the distribution for the sample mean estimator for the mean parameter of a normal distribution using only one observation. (cf. Eq. (4)) So essentially asymptotically efficient estimators are asymptotically equivalent to such an estimation problem. A fuller discussion of efficiency is deep and beyond this course.

By asymptotic normality result for M-estimator, we know what the asymptotic distribution for the MLE should be. However, it turns out given the information inequality, the MLE’s asymptotic distribution can be further simplified.
Theorem 10.1.12: Asymptotic efficiency of MLE

Proof: (following Amemiya, pp. 143–144)

- \( \hat{\theta}_n^{MLE} \) satisfies the FOC of the MLE problem:
  \[
  0 = \frac{\partial \log L(\theta|\bar{X}_n)}{\partial \theta} |_{\theta = \hat{\theta}_n^{MLE}}.
  \]

- Using the mean value theorem:
  \[
  0 = \frac{\partial \log L(\theta|\bar{X}_n)}{\partial \theta} |_{\theta = \theta_0} + \frac{\partial^2 \log L(\theta|\bar{X}_n)}{\partial \theta^2} |_{\theta = \theta_n} (\hat{\theta}_n^{MLE} - \theta_0)
  \]

  \[
  \implies \sqrt{n} (\hat{\theta}_n - \theta_0) = \sqrt{n} \frac{\partial \log L(\theta|\bar{X}_n)}{\partial \theta} |_{\theta = \theta_0} = \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta} |_{\theta = \theta_0} = \frac{1}{n} \sum_i \frac{\partial^2 \log f(x_i|\theta)}{\partial \theta^2} |_{\theta = \theta_0} \quad (**)
  \]

- Note that, by the LLN,
  \[
  \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta} |_{\theta = \theta_0} \xrightarrow{p} E_{\theta_0} \frac{\partial \log f(X|\theta)}{\partial \theta} |_{\theta = \theta_0} = \int \frac{\partial f(x_i|\theta)}{\partial \theta} |_{\theta = \theta_0} dx.
  \]

  Using same argument as in the information inequality result above, the last term is:
  \[
  \int \frac{\partial f}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int f dx = 0.
  \]

- Hence, the CLT can be applied to the numerator of (**):
  \[
  \text{numerator of (**)} \xrightarrow{d} N \left( 0, E_{\theta_0} \left( \frac{\partial \log f(x_i|\theta)}{\partial \theta} |_{\theta = \theta_0} \right)^2 \right).
  \]

- By LLN, and uniform convergence of Hessian term:
  \[
  \text{denominator of (**)} \xrightarrow{p} E_{\theta_0} \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} |_{\theta = \theta_0}.
  \]

- Hence, by Slutsky theorem:
  \[
  \sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{d} N \left( 0, \left[ E_{\theta_0} \left( \frac{\partial \log f(x_i|\theta)}{\partial \theta} |_{\theta = \theta_0} \right)^2 \right] \left[ E_{\theta_0} \left( \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} |_{\theta = \theta_0} \right)^2 \right] \right).
  \]
• By the information inequality:
\[
E_{\theta_0} \left( \frac{\partial \log f(x_i|\theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^2 = -E_{\theta_0} \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \bigg|_{\theta=\theta_0}
\]
so that
\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N \left( 0, \frac{1}{E_{\theta_0} \left( \frac{\partial \log f(x_i|\theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^2} \right)
\]
so that the asymptotic variance is the CRLB.
Hence, the asymptotic approximation for the finite-sample distribution is
\[
\hat{\theta}_n^{MLE} \sim N \left( \theta_0, \frac{1}{n E_{\theta_0} \left( \frac{\partial \log f(x_i|\theta)}{\partial \theta} \bigg|_{\theta=\theta_0} \right)^2} \right).
\]