# 1 POINT ESTIMATION: DEFINITION OF ESTIMATORS

**Point estimator:** any function $W(X_1, \ldots, X_n)$ of a data sample.

The exercise of point estimation is to use particular functions of the data in order to estimate certain unknown population parameters.

**Examples:** Assume that $X_1, \ldots, X_n$ are drawn i.i.d. from some distribution with unknown mean $\mu$ and unknown variance $\sigma^2$.

Potential point estimators for $\mu$ include: sample mean $\bar{X}_n = \frac{1}{n}\sum_i X_i$; sample median med$(X_1, \ldots, X_n)$.

Potential point estimators for $\sigma^2$ include: the sample variance $\frac{1}{n}\sum_i (X_i - \bar{X}_n)^2$.

*Any point estimator is a random variable*, whose distribution is that induced by the distribution of $X_1, \ldots, X_n$.

**Example:** $X_1, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$.
Then sample mean $\bar{X}_n \sim N(\mu_n, \sigma_n^2)$ where $\mu_n = \mu,\ \forall n$ and $\sigma_n^2 = \sigma^2/n$.

For a particular realization of the random variables $x_1, \ldots, x_n$, the corresponding point estimator evaluated at $x_1, \ldots, x_n$, i.e., $W(x_1, \ldots, x_n)$, is called the *point estimate*.

In these lecture notes, we will consider three types of estimators:

1. Method of moments
2. Maximum likelihood
3. Bayesian estimation

## Method of moments:

Assume: $X_1, \ldots, X_n \sim$ i.i.d. $f(x|\theta_1, \ldots, \theta_K)$

Here the unknown parameters are $\theta_1, \ldots, \theta_K$ ($K \leq N$).

Idea is to find values of the parameters such that the population moments are as close as possible to their "sample analogs". This involves finding values of the parameters to solve the following $K$-system of equations:

$$
\begin{aligned}
m_1 &\equiv \frac{1}{n}\sum_i X_i = EX = \int x f(x|\theta_1,\ldots,\theta_K)dx \\
m_2 &\equiv \frac{1}{n}\sum_i X_i^2 = EX^2 = \int x^2 f(x|\theta_1,\ldots,\theta_K)dx \\
&\vdots \\
m_K &\equiv \frac{1}{n}\sum_i X_i^K = EX^K = \int x^K f(x|\theta_1,\ldots,\theta_K)dx.
\end{aligned}
$$

**Example:** $X_1,\ldots,X_n \sim$ i.i.d. $N(\theta,\sigma^2)$. Parameters are $\theta$, $\sigma^2$.

Moment equations are:

$$
\begin{aligned}
&\frac{1}{n}\sum_i X_i = EX = \theta \\
&\frac{1}{n}\sum_i X_i^2 = EX^2 = VX + (EX)^2 = \sigma^2 + \theta^2.
\end{aligned}
$$

Hence, the MOM estimators are $\theta^{MOM} = \bar{X}_n$ and ${\sigma^2}^{MOM} = \frac{1}{n}\sum_i X_i^2 - (\bar{X}_n)^2$.

**Example:** $X_1,\ldots,X_n \sim$ i.i.d. $U[0,\theta]$. Parameter is $\theta$.

MOM: $\bar{X}_n = \frac{\theta}{2} \Longrightarrow \theta^{MOM} = 2\cdot\bar{X}_n$.

■■■

The model assumption that $X_1,\ldots,X_n \sim$ i.i.d. $f(\cdot|\vec{\theta})$ implies a number of moment equations equal to the number of moments, which can be $>> K$. This leaves room for evaluating the model specification.

For example, in the uniform distribution example above, another moment condition which should be satisfied is that

$$
\frac{1}{n}\sum_i X_i^2 = EX^2 = VX + (EX)^2 = \frac{\theta^2}{3} + \frac{\theta}{2}. \tag{1}
$$

At the MOM estimator $\theta^{MOM}$, one can see whether

$$
\frac{1}{n}\sum_i X_i^2 = \frac{{\theta^{MOM}}^2}{3} + \frac{\theta^{MOM}}{2}.
$$

(Later, you will learn how this can be tested more formally.) If this does not hold, then that might be cause for you to conclude that the original specification that $X_1, \ldots, X_n \sim$ i.i.d. $U[0, \theta]$ is inadequate. Eq. (1) is an example is an **overidentifying restriction**.

## Maximum Likelihood Estimation

Let $X_1, \ldots, X_n \sim$ i.i.d. with density $f(\cdot|\theta_1, \ldots, \theta_K)$.

**Define:** the likelihood function, for a continuous random variable, is the joint density of the sample observations:

$$L(\vec{\theta}|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\vec{\theta}).$$

View $L(\vec{\theta}|\vec{x})$ as a function of the parameters $\vec{\theta}$, for the data observations $\vec{x}$.

The **maximum likelihood estimator (MLE)** are the parameter values $\vec{\theta}^{ML}$ which maximize the likelihood function:

$$\vec{\theta}^{ML} = \text{argmax}_{\vec{\theta}} L(\vec{\theta}|\vec{x}).$$

Usually, in practice, to avoid numerical overflow problems, maximize the log of the likelihood function:

$$\vec{\theta}^{ML} = \text{argmax}_{\vec{\theta}} \log\ L(\vec{\theta}|\vec{x}) = \sum_i \log f(x_i|\vec{\theta}).$$

Analogously, for discrete random variables, the likelihood function is the joint probability mass function:

$$L(\vec{\theta}|\vec{x}) = \prod_{i=1}^{n} P(X = x_i|\vec{\theta}).$$

■■■

**Example:** $X_1, \ldots, X_n \sim$ i.i.d. $N(\theta, 1)$.

- $\log L(\theta|\vec{x}) = \log(n/\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$
- $\max_\theta \log L(\theta|\vec{x}) = \min_\theta \frac{1}{2}\sum_i (x_i - \theta)^2$

- FOC: $\frac{\partial \log L}{\partial \theta} = \sum_i (x_i - \theta) = 0 \Rightarrow \theta^{ML} = \frac{1}{n}\sum_i x_i$ (sample mean)

  Also should check second order condition: $\frac{\partial^2 \log L}{\partial \theta^2} = -n < 0$ : so satisfied.

**Example:** $X_1, \ldots, X_n \sim$ i.i.d. Bernoulli with prob. $p$. Unknown parameter is $p$.

- $L(p|\vec{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$
- $$\begin{aligned}\log L(p|\vec{x}) &= \sum_{i=1}^n [x_i \cdot \log p + (1-x_i)\cdot \log(1-p)] \\ &= y \log p + (n-y)\log(1-p) : \quad y \text{ is number of 1's}\end{aligned}$$
- FOC: $\frac{\partial \log L}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} \Longrightarrow p^{ML} = \frac{y}{n}$

  For $y = 0$ or $y = n$, the $p^{ML}$ is (respectively) 0 and 1: corner solutions.
- SOC: $\left.\frac{\partial \log L}{\partial p^2}\right|_{p=p^{ML}} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} < 0$ for $0 < y < n$.

  When parameter is multidimensional: check that the *Hessian matrix* $\frac{\partial^2 \log L}{\partial \theta \partial \theta'}$ is negative definite.

**Example:** $X_1, \ldots, X_n \sim\ U[0, \theta]$.

Likelihood function

$$L(\vec{X}|\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \max(X_1, \ldots, X_n) \leq \theta \\ 0 & \text{if } \max(X_1, \ldots, X_n) > \theta \end{cases}$$

which is maximized at $\theta_n^{MLE} = \max(X_1, \ldots, X_n)$.

## Bayesian inference

Philosophically different view of the world. Model the unknown parameters $\vec{\theta}$ as random variables, and assume that researcher's beliefs about $\theta$ are summarized in a *prior distribution* $f(\theta)$.

In this sense, Bayesian approach is "subjective", because researcher's beliefs about $\theta$ are accommodated in inferential approach.

$X_1, \ldots, X_n \sim$ i.i.d. $f(x|\theta)$: the Bayesian views the density of each data observation as a *conditional density*, conditional on a realization of the random variable $\theta$.

Given data $X_1, \ldots, X_n$, we can update our beliefs about the parameter $\theta$ by computing the posterior density (using Bayes Rule):

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta) \cdot f(\theta)}{f(\vec{x})} = \frac{f(\vec{x}|\theta) \cdot f(\theta)}{\int f(\vec{x}|\theta) f(\theta) d\theta}.$$

A Bayesian point estimate of $\theta$ is some feature of this posterior density. Common point estimators are:

- Posterior mean:
$$E\left[\theta|\vec{x}\right] = \int \theta f(\theta|\vec{x}) d\theta.$$
- Posterior median: $F^{-1}_{\theta|\vec{x}}(0.5)$, where $F_{\theta|\vec{x}}$ is CDF corresponding to the posterior density: i.e., $F_{\theta|\vec{x}}(\tilde{\theta}) = \int_{-\infty}^{\tilde{\theta}} f(\theta|\vec{x}) d\theta$.
- Posterior mode: $\max_\theta f(\theta|\vec{x})$. This is the point at which the density is highest.

Note that $f(\vec{x}|\theta)$ is just the likelihood function, so that the posterior density $f(\theta|\vec{x})$ can be written as:

$$f(\theta|\vec{x}) = \frac{L(\theta|\vec{x}) \cdot f(\theta)}{\int L(\theta|\vec{x}) f(\theta) d\theta}.$$

*But there is a difference in interpretation:* in Bayesian world, the likelihood function is random due to both $\vec{x}$ and $\theta$, whereas in classical world, only $\vec{x}$ is random.

**Example:** $X_1, \ldots, X_n \sim$ i.i.d. $N(\theta, 1)$, with prior density $f(\theta)$.

Posterior density $f(\theta|\vec{x}) = \frac{\exp\left(-\frac{1}{2}\sum_i (x_i - \theta)^2\right) f(\theta)}{\int \exp\left(-\frac{1}{2}\sum_i (x_i - \theta)^2\right) f(\theta) d\theta.}$

Integral in denominator can be difficult to calculate: computational difficulties can hamper computation of posterior densities.

However, note that the denominator is not a function of $\theta$. Thus

$$f(\theta|\vec{x}) \propto L(\theta|\vec{x}).$$

Hence, if we assume that $f(\theta)$ is constant (ie. uniform), for all possible values of $\theta$, then the posterior mode argmax$_\theta f(\theta|\vec{x})$ = argmax$_\theta L(\theta|\vec{x}) = \theta^{ML}$.

**Example:** Bayesian updating for normal distribution, with normal priors

$X \sim N(\theta, \sigma^2)$, assume $\sigma^2$ is known.

Prior: $\theta \sim N(\mu, \tau^2)$, assume $\tau$ is known.

Then posterior distribution

$$\theta|X \sim N(E(\theta|X), V(\theta|X)),$$

where

$$\begin{aligned} E(\theta|X) &= \frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu \\ V(\theta|X) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \end{aligned}$$

This is an example of a *conjugate* prior and conjugate distribution, where the posterior distribution comes from the same family as the prior distribution.

Posterior mean $E(\theta|X)$ is weighted average of $X$ and prior mean $\mu$.

In this case, as $\tau \to \infty$ (so that prior information gets worse and worse): then $E(\theta|X) \to X$ (a.s.). These are just the MLE (for just one data observation).

When you observe an i.i.d. sample $\vec{X}_n \equiv (X_1, \ldots, X_n)$, with sample mean $\bar{X}_n$:

$$\begin{aligned} E(\theta|\vec{X}_n) &= \frac{n\tau^2}{n\tau^2+\sigma^2}\bar{X}_n + \frac{\sigma^2}{\sigma^2+n\tau^2}\mu \\ V(\theta|\vec{X}_n) &= \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}. \end{aligned}$$

In this case, as the number of observations $n \to \infty$, the posterior mean $E(\theta|\vec{X}_n) \to \bar{X}_n$. So as $n \to \infty$, the posterior mean converges to the MLE: when your sample becomes arbitrarily large, you place no weight on your prior information.

■■■

Properties of Bayesian updating typify "rational" behavior (learning) in economics.

Note that, by the law of iterated expectations:

$$E\left[E(\theta|X_1, X_2)|X_1\right] = E(\theta|X_1)$$

that is: the sequence of posterior means is a *martingale*. Your best guess at time $t = 1$ of the posterior mean at $t = 2$ is just the posterior mean at $t = 1$.

Equivalently, the "forecast error" has mean zero:

$$E\left[E(\theta|X_1, X_2) - E(\theta|X_1)|X_1\right] = 0.$$

This property is being "correct on average" characterizes *rational expectations*.

## 2 FINITE-SAMPLE PROPERTIES OF ESTIMATORS (CB 7.3)

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

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

Consider:
$\hat{\mu}_n \equiv \frac{1}{n}\sum_i X_i$, estimator of $\mu$
$\hat{\sigma}_n^2 \equiv \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2$, estimator of $\sigma^2$.

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

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

$$\begin{aligned}
E\hat{\sigma}^2 &= E(\frac{1}{n}\sum_i (X_i - \bar{X}_n)^2) \\
&= \frac{1}{n} \cdot \sum_i (EX_i^2 - 2EX_i\bar{X}_n + E\bar{X}_n^2) \\
&= \frac{1}{n} \cdot n\left[(\mu^2 + \sigma^2) - 2(\mu^2 + \frac{\sigma^2}{n}) + \frac{\sigma^2}{n} + \mu^2\right] \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Hence it is biased.

To fix this bias, consider the estimator $s_n^2 \equiv \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2$, and $Es_n^2 = \sigma^2$ (unbiased).

■■■

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

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

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

**Example:** $X_1, \ldots, X_n \sim\ U[0,\theta]$. $f(X) = 1/\theta,\ x \in [0,\theta]$.

- 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
  $MSE(\hat{\theta}_N) = V\hat{\theta}_n = \frac{4}{n^2}\sum_i VX_i = \frac{\theta^2}{3n}$
- Consider estimator $\tilde{\theta}_n \equiv \max(X_1, \ldots, X_n)$.
  In order to derive moments, start by deriving CDF:

$$P(\tilde{\theta}_n \le z) = P(X_1 \le z, X_2 \le z, \ldots, X_n \le z) = \prod_{i=1}^{n} P(X_i \le z)$$
$$= \begin{cases} \left(\frac{z}{\theta}\right)^n & \text{if } z \le \theta \\ 1 & \text{otherwise} \end{cases}$$

  Therefore $f_{\tilde{\theta}_n}(z) = n \cdot \left(\frac{z}{\theta}\right)^{n-1}\frac{1}{\theta}$, for $0 \le x \le \theta$.

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

  Bias$(\tilde{\theta}_n) = -\theta/(n+1)$
  $E(\tilde{\theta}_n^2) = \frac{n}{\theta^n}\int_0^\theta z^{n+1}dz = \frac{n}{n+2}\theta^2$.
  Hence $V\tilde{\theta}_n = \theta^2\left(\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right) = \theta^2\frac{n}{(n+2)(n+1)^2}$.
  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? $V\hat{\theta}_n = \frac{\theta^2}{3n} = O(1/n)$.
$V\tilde{\theta}_n = \left(\frac{n+1}{n}\right)^2 V\left(\max(X_1, \ldots, X_n)\right) = \theta^2\left(\frac{1}{n(n+2)}\right) = O(1/n^2)$.
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 $\theta$.

■■■

By itself, the UMVUE concept is not constructive, as one cannot use it to evaluate whether an estimator satisfies the concept. Fortunately, 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 $f(\vec{X}|\theta)$, and let $W(\vec{X})$ be any estimator satisfying

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

$$(ii) \quad V_\theta W(\vec{X}) < \infty.$$

  Then

$$V_\theta W(\vec{X}) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\vec{X})\right)^2}{E_\theta\left(\frac{\partial}{\partial\theta}\log f(\vec{X}|\theta)\right)^2}.$$

  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 $E_\theta W(X) = \int W(X) f(X|\theta) dX$:

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

Applying the Cauchy-Schwarz inequality, we have

$$[\frac{d}{d\theta} E_\theta W(X)]^2 \leq Var_\theta W(X) \cdot E_\theta [\frac{\partial}{\partial\theta} \log f(X|\theta)]^2$$

or

$$Var_\theta W(X) \geq \frac{[\frac{d}{d\theta} E_\theta W(X)]^2}{E_\theta [\frac{\partial}{\partial\theta} \log f(X|\theta)]^2.}$$

■

- 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 $\theta$.

  A crucial step in the derivation of the CR-bound is the interchange of differentiation and integration. This implies

$$\begin{aligned}
E_\theta \frac{\partial}{\partial\theta} \log f(\vec{X}|\theta) &= \int \frac{1}{f(\vec{X}|\theta)} \frac{\partial f(\vec{X}|\theta)}{\partial\theta} f(\vec{X}|\theta) dx \\
&= \frac{\partial}{\partial\theta} \int f(\vec{X}|\theta) dx = \frac{\partial}{\partial\theta} \cdot 1 = 0
\end{aligned} \tag{2}$$

- (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 Generalized MOM (GMM) estimation based on the moment condition

$$E_{\theta_0}\frac{\partial}{\partial\theta}\log f(x_i|\theta) = 0. \tag{3}$$

From the previous derivation, we see that this condition holds at the true value $\theta = \theta_0$.

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

  **Corollary 7.3.10**: if $X_1, \ldots, X_n \sim$ *i.i.d.* $f(X|\theta)$, then

$$V_\theta W(\vec{X}) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\vec{X})\right)^2}{n \cdot E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2}. \tag{4}$$

- Up to this point, Cramer-Rao results not operational, because the estimator $W(\vec{X})$ is on both sides of the inequality. However, for an unbiased estimator, $E_\theta W(\vec{X}) = \theta$, so that $\frac{d}{d\theta}E_\theta W(\vec{X}) = 1$. Then

$$V_\theta W(\vec{X}) \geq \frac{1}{E_\theta\left(\frac{\partial}{\partial\theta}\log f(\vec{X}|\theta)\right)^2}.$$

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

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

  What is CRLB for an unbiased estimator of $\mu$?

  Unbiased $\rightarrow$ numerator =1.

$$\log f(x|\theta) = \log\sqrt{2\pi} - \log\sigma - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$\frac{\partial}{\partial\mu}\log f(x|\theta) = -\left(\frac{x-\mu}{\sigma}\right)\cdot\left(\frac{-1}{\sigma}\right) = \frac{x-\mu}{\sigma^2}$$

$$E\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2 = E\left((X-\mu)^2\sigma^4\right) = \frac{1}{\sigma^4}VX = \frac{1}{\sigma^2}. \tag{5}$$

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. (2) above, we get that LHS of (*) =0.

RHS of (*):

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

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

- The LHS of the above condition (*) is just $\frac{d}{d\theta} \int \left( \frac{\partial}{\partial \theta} \log f(X|\theta) \right) f(X|\theta) dX$. 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.

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

  Go back to derivation of CRLB (final line of top display), by the Cauchy-Schwarz inequality and assuming unbiasedness ($E_\theta W(X) = \theta$) we have:

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

  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:

$$\frac{\partial \log L}{\partial \theta} = a(\theta) * (W(X) - \theta)$$

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:** $X_1, \ldots, X_n \sim$ *i.i.d.* $N(\mu, \sigma^2)$

Consider estimating $\mu$:

$$\begin{aligned}
\frac{\partial}{\partial\mu} \log L(\theta|\vec{X}) &= \frac{\partial}{\partial\mu}\left(\sum_{i=1}^n \log f(X|\theta)\right) \\
&= \frac{\partial}{\partial\mu}\left(\sum_{i=1}^n -\log\sqrt{2\pi} - \log\sigma - \frac{1}{2}\left(\frac{(X-\mu)^2}{\sigma^2}\right)\right) \\
&= \sum_{i=1}^n \left(\frac{X-\mu}{\sigma^2}\right) \quad = \frac{n}{\sigma^2}\cdot(\bar{X}_n - \mu).
\end{aligned}$$

Hence, CRLB can be attained (in fact, we showed earlier that CRLB attained by $\bar{X}_n$)

Additional examples: Poisson, Binomial.

# 3 LARGE SAMPLE PROPERTIES OF ESTIMATORS (CB 10.1)

Large-sample properties: exploit LLN, CLT

Consider data $\{X_1, X_2, \ldots\}$ by which we construct a sequence of estimators $W_n \equiv \left\{W(\vec{X}_1), W(\vec{X}_2), \ldots\right\}$. $W_n$ 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 $W_n \xrightarrow{as} \theta$. *Weak consistency* obtains when $W_n \xrightarrow{p} \theta$.

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 $f_\theta \equiv f(x_1, \ldots, x_n, \ldots; \theta)$ 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 $\theta_0$, $P(\theta_n \in \mathcal{N}) \to 1$.

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

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

The event "$\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2$" implies it holds at $\theta_n$

$$Q_n(\theta_n) - Q_0(\theta_n) < \epsilon/2 \Leftrightarrow Q_0(\theta_n) > Q_n(\theta_n) - \epsilon/2 \tag{6}$$

Similarly, it holds at $\theta_0$

$$Q_n(\theta_0) - Q_0(\theta_0) > -\epsilon/2 \Rightarrow Q_n(\theta_0) > Q_0(\theta_0) - \epsilon/2. \tag{7}$$

Since $\theta_n =$ argmax$_\theta Q_n(\theta)$, Eq. (6) implies

$$Q_0(\theta_n) > Q_n(\theta_0) - \epsilon/2. \tag{8}$$

Hence, adding Eqs. (7) and (8), we have

$$Q_0(\theta_n) > Q_0(\theta_0) - \epsilon. \tag{9}$$

So we have shown that

$$\begin{aligned}
&\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2 \Longrightarrow Q_0(\theta_n) > Q_0(\theta_0) - \epsilon \\
\Leftrightarrow &P\left(Q_0(\theta_n) > Q_0(\theta_0) - \epsilon\right) \geq P\left(\sup_{\theta\in\Theta} |Q_n(\theta) - Q_0(\theta)| < \epsilon/2\right) > 1 - \delta.
\end{aligned}$$

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

$$\begin{aligned}
Q_0(\theta_n) > Q_0(\theta_0) - \epsilon &\Rightarrow Q_0(\theta_n) > \max_{\theta\in\Theta\cap\bar{\mathcal{N}}} Q_0(\theta) \\
&\Rightarrow \theta_n \in \mathcal{N} \\
&\Leftrightarrow P\left(\theta_n \in \mathcal{N}\right) \geq P\left(Q_0(\theta_n) > Q_0(\theta_0) - \epsilon\right) > 1 - \delta.
\end{aligned}$$

Since the argument above holds for any arbitrarily small neighborhood of $\theta_0$, and for any $(\epsilon, \delta)$ we are done. ■

[**] 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 $\theta$ with

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

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

To understand this more intuitively, go back to the structure of the proof.The key is equations (6), (7) which ensure that $Q_n$ is close to $Q_0$ at the points $\theta_n$ and $\theta_0$ *with the same* $\epsilon$, an "equal closeness" condition. Given this, we can derive that $Q_0(\theta_n)$ and $Q_0(\theta_0)$ are close (Eq. (9)) which by continuity of $Q_0$ (A3) ensures that $\theta_n$ and $\theta_0$ are close.

However, pointwise convergence is not enough to ensure this "equal closeness" condition. At any given $n$, $Q_n(\theta_0)$ being close to $Q_0(\theta_0)$ does not imply this at other points. This requires an "equicontinuity" type of condition which ensures that $Q_n$ lies with high probability within a "sleeve" of the limiting function $Q_0(\cdot)$ (see picture below). Since $Q_n()$ is a stochastic function, stochastic equicontinuity is the right notion of continuity here.

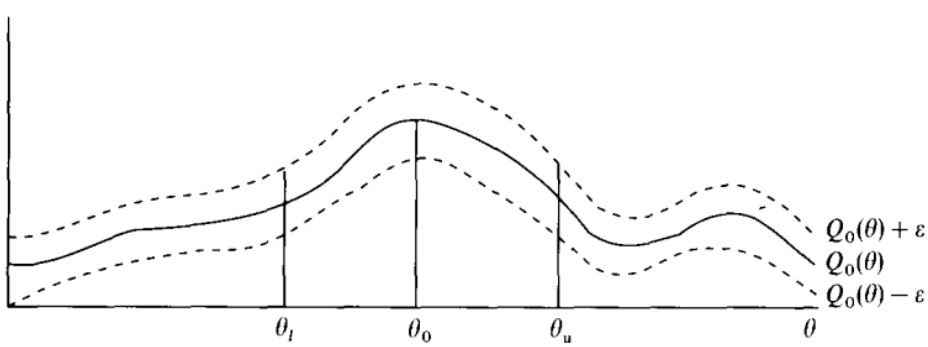


Figure 2.

**Asymptotic normality for M-estimators** Define the "score vector"

$$\bigtriangledown_{\tilde{\theta}} Q_n(\theta) = \left[ \frac{\partial Q_n(\theta)}{\partial \theta_1}\bigg|_{\theta=\tilde{\theta}}, \ldots, \frac{\partial Q_n(\theta)}{\partial \theta_K}\bigg|_{\theta=\tilde{\theta}} \right]'.$$

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

$$\left[ \bigtriangledown_{\tilde{\theta}\tilde{\theta}} Q_n(\theta) \right]_{i,j} = \frac{\partial^2 Q_n(\theta)}{\partial \theta_i \partial \theta_j}\bigg|_{\theta=\tilde{\theta}}, \quad 1 \leq i, j \leq K.$$

Make the following assumptions:

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

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

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow{d} N(0, H_0^{-1} \Sigma H_0^{-1})$$

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

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

$$\begin{aligned}
&0 = \bigtriangledown_{\theta_n} Q_n(\theta) = \bigtriangledown_{\theta_0} Q_n(\theta) + \bigtriangledown_{\bar{\theta}_n \bar{\theta}_n} Q_n(\theta)(\theta_n - \theta_0) \\
\Rightarrow &\underbrace{\bigtriangledown_{\bar{\theta}_n \bar{\theta}_n} Q_n(\theta)}_{\xrightarrow{p} H_0 \text{ (using A5)}} \sqrt{n}(\theta_n - \theta_0) = - \underbrace{\sqrt{n} \bigtriangledown_{\theta_0} Q_n(\theta)}_{\xrightarrow{d} N(0,\Sigma) \text{ (using A4)}} \\
\Leftrightarrow &\sqrt{n}(\theta_n - \theta_0) \xrightarrow{d} -H(\theta_0)^{-1} N(0, \Sigma) = N(0, H_0^{-1} \Sigma H_0^{-1}). \qquad \blacksquare
\end{aligned}$$

Note: A5 is a uniform convergence assumption on the sample Hessian. Given previous discussion, it ensures that the sample Hessian $\bigtriangledown_{\theta\theta} Q_n(\theta)$ evaluated at $\bar{\theta}_n$ (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 $\theta_0$.

**Efficiency**

Efficiency 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 "regular" 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 $v(\theta) = \frac{1}{E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2}$

■■■

By comparison with Eq. (4), note that the asymptotic variance $\frac{1}{E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2}$ is equivalent to the CRLB for *one observation* ($n = 1$). $I(\theta) \equiv E_{\theta_0}\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2$ 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. (5)) So asyptotically efficient estimators are asymptotically equivalent to such an estimation problem. This is a deep topic and a full discussion is beyond this class. But essentially a normal distribution with zero mean and variance equal to the reciprocal of the Fisher information is the most "concentrated" distribution for an estimator of the mean parameter of a normal distribution involving one single observation. Thus asymptotically efficient estimators are asymptotically equivalent to such an estimation problem.

■■■

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:

$$0 = \frac{\partial \log L(\theta|\vec{X}_n)}{\partial \theta}\Big|_{\theta=\hat{\theta}_n^{MLE}}.$$

Using the mean value theorem:

$$0 = \frac{\partial \log L(\theta|\vec{X}_n)}{\partial \theta}\Big|_{\theta=\theta_0} + \frac{\partial^2 \log L(\theta|\vec{X}_n)}{\partial \theta^2}\Big|_{\theta=\theta_n^*}\left(\hat{\theta}_n^{MLE} - \theta_0\right)$$

$$\Longrightarrow \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \sqrt{n}\frac{-\frac{\partial \log L(\theta|\vec{X}_n)}{\partial \theta}\Big|_{\theta=\theta_0}}{\frac{\partial^2 \log L(\theta|\vec{X}_n)}{\partial \theta^2}\Big|_{\theta=\theta_n^*}} = \sqrt{n}\frac{-\frac{1}{n}\sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}}{\frac{1}{n}\sum_i \frac{\partial^2 \log f(x_i|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_n^*}} \quad (**)$$

By the LLN,

$$\frac{1}{n}\sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \xrightarrow{p} E_{\theta_0}\frac{\partial \log f(X|\theta)}{\partial \theta}\Big|_{\theta=\theta_0} = \int \frac{\partial f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0} dx.$$

Using the familiar argument, the last term evaluates to zero: $\int \frac{\partial f}{\partial \theta} dx = \frac{\partial}{\partial \theta}\int f dx = 0$. Hence, the CLT can be applied to the numerator of (**):

$$\text{numerator of (**)} \xrightarrow{d} N\left(0, E_{\theta_0}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2\right).$$

By LLN, and uniform convergence of Hessian term:

$$\text{denominator of (**)} \xrightarrow{p} E_{\theta_0}\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}.$$

Hence, by Slutsky theorem:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{E_{\theta_0}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2}{\left[E_{\theta_0}\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}\right]^2}\right).$$

By the information inequality:

$$E_{\theta_0}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2 = -E_{\theta_0}\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}$$

so that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{E_{\theta_0}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2}\right)$$

so that the asymptotic variance is the CRLB.

## Additional topics:

### 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 $MSE(W(\vec{X})) = E_\theta \left(W(\vec{X}) - \theta\right)^2$, the MSE is actually the risk function associated with the *quadratic loss function* $\mathcal{L}(\theta, W(\vec{X})) = \left(W(\vec{X}) - \theta\right)^2$.

The exercise of minimizing risk takes a given value of $\theta$ as given. But typically, we want an estimator which does well regardless of which value of $\theta$ you are considering. (Analogous to the focus on the *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 $\theta$.

**Consistency of MLE: Wald's argument**

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 $\theta_n$:

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

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.

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

- Define $\hat{\theta}_n^{MLE} \equiv \text{argmax}_\theta \frac{1}{n}\sum_i \log f(x_i|\theta)$. 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 $\theta$.[1]
- Hence:

$$\begin{aligned}
&E_{\theta_0}\log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right) < 0,\ \forall\theta \\
\Longrightarrow &E_{\theta_0}\log f(x|\theta) < E_{\theta_0}\log f(x|\theta_0),\ \forall\theta \\
\Longrightarrow &E_{\theta_0}\log f(x|\theta) \text{ is maximized at the true } \theta_0.
\end{aligned}$$

  This is the "identification" assumption from the M-estimator consistency theorem.

[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 $\theta$.