# Chapter 8 Parameter Estimation and Evaluation

## 8.1 Population and Distribution Model

- **Problem:** Consider a random sample $X_1, \cdots, X_n$ from some unknown population distribution $f_X(x)$. A realization $x_1, \cdots, x_n$ of random sample $X_1, \cdots, X_n$ is observed. We want to make inference of the population distribution $f_X(x)$ using the observed data.

- **Distribution Model:** To make inference of $f_X(x)$, we often consider a class of parametric candidate distributions

$$\mathbb{F} = \{f(x;\theta) : \theta \in \Theta\}.$$

Each value of $\theta \in \Theta$ gives a *distribution model* for $f_X(x)$. If the true population distribution

$$f_X(x) = f(x;\theta_0) \in \mathbb{F},$$

we call that $\mathbb{F}$ is *correctly specified* and $\theta_0$ is called the true parameter value. In contrast, $\mathbb{F}$ is said to be *misspecified* for the population distribution $f_X(x)$ if there exists no value for $\theta \in \Theta$ such that $f_X(x) = f(x;\theta)$.

## 8.1 Population and Distribution Model

- **Definition: Point Estimator.** Let $X_1, \cdots, X_n$ be a random sample of size $n$ from the population $f(x;\theta)$ with parameter $\theta$. A *point estimator* of $\theta$ is any function $T(X_1, \cdots, X_n)$ of the random sample, that is, any statistic is a point estimator.

- **Definition: Interval Estimator.** An *interval estimator* of $\theta$ is a pair of statistics $L(X_1, \cdots, X_n)$ and $U(X_1, \cdots, X_n)$ that satisfy $L(x_1, \cdots, x_n) \leq U(x_1, \cdots, x_n)$ for any $x_1, \cdots, x_n$.

- **Remark:** When a realization $x_1, \cdots, x_n$ is observed, we make the inference that $\theta = T(x_1, \cdots, x_n)$ by point estimation and $L(x_1, \cdots, x_n) \leq \theta \leq U(x_1, \cdots, x_n)$ by interval estimation.

- **Remarks:**
  - Compared to interval estimation, the point estimator provides a single value as the estimation of the parameter.
  - Sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a point estimator of the true mean $E(X_i)$.
  - Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ is a point estimator of the true variance $\mathrm{Var}(X_i)$.
  - There can be many different point estimators of the same parameter.

- **Example:** Suppose that we take a random sample $X_1, \cdots, X_n$ from the $U[0, \theta]$ distribution so that

$$f(x;\theta) = \begin{cases} 1/\theta & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to estimate $\theta$.

- **ANS:** Let $X_{(1)}, \cdots, X_{(n)}$ be the order statistics of $X_1, \cdots, X_n$, we have $E(X_{(k)}) = \frac{k}{n+1}\theta$. Some different estimators of $\theta$ are as follows.
  - $T_1 = X_{(n)}$.
  - $T_2 = \frac{n+1}{n} X_{(n)}$.
  - $T_3 = X_{(1)} + X_{(n)}$.
  - $T_4 = (n+1)X_{(1)}$
  - $T_5 = 2\overline{X}_n$.

- **Example:** Let $X_1, \cdots, X_n$ be a random sample from $U(0,1)$. Then the pdf of $X_{(k)}$ is

$$f_{X_{(k)}}(u) = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{n-k}, \quad 0 < u < 1,$$

  and $X_{(k)}$ follows a $\text{Beta}(k, n-k+1)$ distribution.

  – $E(X_{(k)}) = \frac{k}{n+1}$.

  – $\text{Var}(X_{(k)}) = \frac{k(n-k+1)}{(n+1)^2(n+2)}$.

  – $\rho_{X_{(1)}X_{(n)}} = \frac{\text{Cov}(X_{(1)}, X_{(n)})}{\sqrt{\text{Var}(X_{(1)})\text{Var}(X_{(n)})}} = 1/n$.

- **Order Statistics**

## 8.2 Maximum Likelihood Estimation

- **Definition: Maximum Likelihood Estimators (MLE).** Let $X_1, \cdots, X_n$ be a random sample of size $n$ from the population $f(x;\theta)$ with parameter $\theta$. Given $(X_1, \cdots, X_n) = (x_1, \cdots, x_n)$, the value $\widehat{\theta}$ that maximizes the likelihood function

$$L(\theta) = L(\theta; x_1, \cdots, x_n) \triangleq \prod_{i=1}^{n} f(x_i; \theta),$$

  over $\theta \in \Theta$ is called the *maximum likelihood estimate* of $\theta$.

- **Remarks:**
  - The MLE is the value of $\theta$ which makes the observed data $x_1, \cdots, x_n$ most likely to occur.
  - It is often convenient to maximize the log likelihood function

$$\log[L(\theta)] = \sum_{i=1}^{n} \log[f(x_i; \theta)].$$

## 8.2 Maximum Likelihood Estimation

- **Theorem: Existence of MLE.** Suppose that with probability one, $\hat{L}(\theta|\mathbf{X}^n)$ is a continuous function of $\theta \in \Theta$, and $\Theta$ is a compact set. Then there exists a global maximizer $\hat{\theta}_n$ that solves the problem,

$$\hat{\theta}_n \equiv \hat{\theta}(\mathbf{X}^n) = \arg\max_{\theta \in \Theta} \hat{L}(\theta|\mathbf{X}^n).$$

- **Remarks:** Assume $L(\theta; x_1, \cdots, x_n)$ is twice continuously differentiable about $\theta$.
  - The MLE $\widehat{\theta}$ must satisfy the *first order condition* (FOC) $\frac{\partial L(\theta; x_1, \cdots, x_n)}{\partial \theta}\Big|_{\theta=\widehat{\theta}} = 0$ if $\widehat{\theta}$ is in the interior of $\Theta$. The boundary need to be checked separately.
  - Suppose $\theta$ is a $p \times 1$ vector. If $\widehat{\theta}$ satisfies the FOC, and the $p \times p$ Hessian matrix

$$H(\widehat{\theta}) = \frac{\partial^2 L(\theta; x_1, \cdots, x_n)}{\partial \theta \partial \theta'}\Big|_{\theta=\widehat{\theta}}$$

is negative definite, then $\widehat{\theta}$ is a local maximum estimator of $L(\theta; x_1, \cdots, x_n)$.

- Remark
  - If the Hessian matrix $H(\theta)$ is negative definite for all $\theta \in \Theta$, then $\widehat{\theta}$ is the global maximum estimator.
  - Given different initial values, computer softwares may find different local maximum solutions of the likelihood function.

- **Example:** In five independent Bernoulli trials with probability of success $\theta$, three successes and two failures were observed. Find the MLE of the probability of success $\theta$.

- **Solution:**
  - The likelihood function is
  $$L(\theta) = \prod_{i=1}^{5} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^3(1-\theta)^2.$$
  - The log-likelihood function is $\log[L(\theta)] = 3\log(\theta) + 2\log(1-\theta)$.
  - The MLE is $\theta = \sum_{i=1}^{n} x_i/n = 3/5$. The MLE is a function of the sufficient statistic $\sum_{i=1}^{n} X_i$ of $\theta$.

- **Remark:** If a **unique** MLE of $\theta$ exists, then it is a function of any sufficient statistic of $\theta$.

- **Example:** Suppose that we observe values $x_1, \cdots, x_n$ from a $N(\mu, \sigma^2)$ distribution. Find the MLE of $\mu$ and $\sigma^2$.

- **Solution:**

  – The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

  – The log-likelihood function is

$$\log[L(\theta)] = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

  – The MLE is

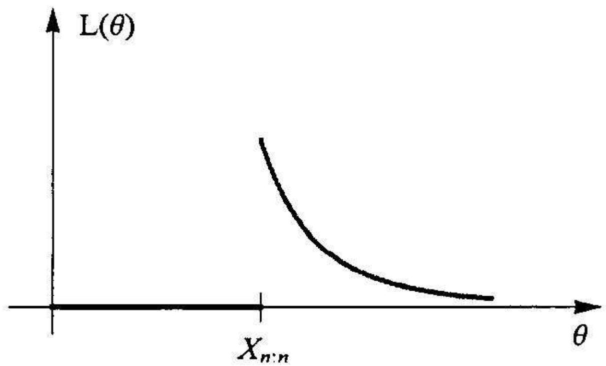
$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

# 8.2 Maximum Likelihood Estimation

- **Example:** Given the sample $x_1, \cdots, x_n$ from $U[0, \theta]$, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } \theta \geq x_i \text{ for all } i, \\ 0 & \text{otherwise.} \end{cases}$$

The MLE is $\theta = X_{(n)}$.



- **Example:** Let $X_1, \cdots, X_n$ be a random sample from a $U[\theta - 1/2, \theta + 1/2]$ distribution. Find the MLE of $\theta$.

- **Solution.**

  – The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \begin{cases} 1 & \text{if } x_{(n)} - 1/2 \le \theta \le x_{(1)} + 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

  – All values between $X_{(n)} - 1/2$ and $X_{(1)} + 1/2$ are the MLE's of $\theta$.

- **Example:** To estimate parameter $\lambda$ in the EXP($\lambda$) distribution with pdf $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x > 0$, a typical experiment consists of putting $n$ pieces of the equipment to test and observing the lifetimes $X_1, X_2, \cdots, X_n$.
  - Suppose that the experiment is interrupted after some time $T$.
  - We can only observe the lifetimes $x_1, \cdots, x_m$ of $m$ equipments.
  - About the remaining equipments we only know that $X_{m+1}, \cdots, X_n > T$.
  - We observe $X_i^* = \min\{X_i, T\}$, $i = 1, \cdots, n$.

- **Solution:**
  - The likelihood function is

$$L(\lambda) = \prod_{i=1}^{m} f(x_i;\lambda) \prod_{i=m+1}^{n} P(X_i > T) = \lambda^m e^{-\lambda(x_1+\cdots+x_m)} e^{-(n-m)\lambda T}.$$

  - The MLE is $\widehat{\lambda} = \frac{m}{x_1+\cdots x_m+(n-m)T}$.

## 8.2 Maximum Likelihood Estimation

- **Example: Linear Regression.** Let $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$ be a random sample from a population, where $\boldsymbol{X}_i = (X_{i,1}, \cdots, X_{i,p})'$ is a random vector. Suppose

$$\boldsymbol{X}_i \sim f_X(\boldsymbol{x}), \quad Y_i = \boldsymbol{X}_i'\beta + \varepsilon_i,$$

where $\beta = (\beta_1, \cdots, \beta_p)'$ is a constant vector, $\varepsilon_i \sim N(0, \sigma^2)$ is independent with $\boldsymbol{X}_i$. Find the MLE of $\beta$ and $\sigma^2$.

- **Solution:**
  - Because $Y_i = \boldsymbol{X}_i'\beta + \varepsilon_i$, $f_{Y|X}(y_i \mid \boldsymbol{x}_i) \sim N(\boldsymbol{x}_i'\beta, \sigma^2)$.
  - $f_{XY}(\boldsymbol{x}_i, y_i) = f_X(\boldsymbol{x}_i) f_{Y|X}(y_i \mid \boldsymbol{x}_i) = f_X(\boldsymbol{x}_i) \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\beta)^2\}$.
  - The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_X(\boldsymbol{x}_i) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\beta)^2\}.$$

## 8.2 Maximum Likelihood Estimation

- – The log-likelihood function is

$$\log[L(\theta)] = \sum_{i=1}^{n} \log[f_X(\boldsymbol{x}_i)] - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\beta)^2.$$

  – The MLE of $\beta$ satisfies

$$\partial \log[L(\theta)]/\partial\beta_1 = -\tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^{n} x_{i,1}(y_i - \boldsymbol{x}_i'\beta) = 0,$$
$$\vdots$$
$$\partial \log[L(\theta)]/\partial\beta_p = -\tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^{n} x_{i,p}(y_i - \boldsymbol{x}_i'\beta) = 0.$$

  Let $\boldsymbol{X} = (\boldsymbol{X}_1', \cdots, \boldsymbol{X}_n')'$, $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$, then

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\beta) = 0.$$

  – MLE of $\beta$ is $\widehat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$.

  – MLE of $\sigma^2$ is $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\widehat{\beta})^2$.

- **Theorem: Invariance property of the MLE.** Suppose $\widehat{\theta}$ is the MLE of $\theta$; and $g(\theta)$ is a one-to-one function over $\Theta$. Then $g(\widehat{\theta})$ is also the MLE of $g(\theta)$.

- **Example:** Suppose we observe $x_1, \cdots, x_n$ from $N(0, \sigma^2)$. Find the MLE of $\sigma^2$ and $\sigma$ $(\sigma > 0)$.

- **Solution:**

  – The log-likelihood function is

$$\log[L(\theta)] = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}.$$

  – The MLE is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ and } \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}.$$

## 8.2 Maximum Likelihood Estimation

- **Theorem: Sufficiency of the MLE.** Suppose $\mathbf{X}^n$ is a random sample with the likelihood function $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$, and $T(\mathbf{X}^n)$ is a sufficient statistic for $\theta$, where $\theta \in \Theta$ is a parameter. Then the MLE $\widehat{\theta}$ that maximizes the likelihood function $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ of the random sample $\mathbf{X}^n$ is also the MLE that maximizes the likelihood function $f_{T(\mathbf{X}^n)}(T(\mathbf{x}^n), \theta)$ of the sufficient statistic $T(\mathbf{X}^n)$.

- **Solution:**
  - $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = f_{T(\mathbf{X}^n)}(T(\mathbf{x}^n), \theta) f_{T(\mathbf{X}^n|\mathbf{X}^n)}(\mathbf{x}^n|T(\mathbf{x}^n)) = f_{T(\mathbf{X}^n)}(T(\mathbf{x}^n), \theta) h(\mathbf{x}^n)$,
  - $\ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \ln f_{T(\mathbf{X}^n)}(T(\mathbf{x}^n), \theta) + \ln h(\mathbf{x}^n)$,
  - $\widehat{\theta} = \arg\max_{\theta \in \Theta} \ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) = \arg\max_{\theta \in \Theta} \ln f_{T(\mathbf{X}^n)}(T(\mathbf{x}^n), \theta)$.

## 8.3 Asymptotic Properties of MLE

- **Assumptions:**

  – **A.1:** $X_1, X_2, \cdots$ are i.i.d. from some population distribution.

  – **A.2:**

  (1). There exists a parameter value $\theta_0$ in the **interior** of $\Theta$ such that $f(x;\theta_0)$ coincides with the population distribution.

  (2). For each $\theta \in \Theta$, $f(x;\theta)$ is a probability pdf/pmf with $f(x;\theta) > 0$ for all $x \in \mathcal{X} = \{x : f(x;\theta_0) > 0\}$.

  (3). $\theta_0$ is the **unique** maximizer of $\max_{\theta\in\Theta} E\{\log[f(X_1,\theta)]\}$. Here

$$E\{\log[f(X_1,\theta)]\} = \int_{\mathcal{X}} \log(f(x,\theta))f(x,\theta_0)\,dx.$$

  (4). The function $\log[f(x;\theta)]$ is continuous on $\mathcal{X} \times \Theta$, and its absolute value is bounded by a nonnegative function $b(x)$ with $E[b(X_1)] < \infty$.

  – **A.3:** $\Theta$ is closed and bounded ($\Theta$ is a compact set).

- **A.4:** $\theta_0$ is the unique maximizer of $E\{\log[f(X_1, \theta)]\}$.
- **A.5:** $\theta_0$ is in the interior of parameter space $\Theta$.
- **A.6:** For each interior point $\theta \in \Theta$, $f(x, \theta)$ is twice continuously differentiable with respect to $\theta$ such that

(1). The functions $\frac{\partial}{\partial\theta}\log[f(x;\theta)]$, $\frac{\partial^2}{\partial\theta^2}\log[f(x;\theta)]$ are continuous in $(x, \theta)$, and their absolute values are bounded by a nonnegative function $b(x)$ with $\int_{-\infty}^{\infty} b(x)f(x;\theta_0)dx < \infty$.

(2). The absolute value of the function $H(\theta) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta^2}\left\{\log[f(x;\theta)]\right\} f(x, \theta)dx$ is bounded by some constant and is nonzero.

## 8.3 Asymptotic Properties of MLE

- **Lemma: Extrema Estimator Lemma.** Suppose

(1) $Q(\theta)$ is a nonstochastic function continuous in $\theta \in \Theta$, and $\theta_0 \in \Theta$ is the unique maximizer of $Q(\theta)$ over $\Theta$, where $\Theta$ is a compact set;

(2) with probability one, $\hat{Q}(\theta)$ is a sequence of random functions continuous in $\theta \in \Theta$;

(3) $\lim_{n\to\infty} \sup_{\theta\in\Theta} |\hat{Q}(\theta) - Q(\theta)| = 0$ almost surely.

Then $\hat{\theta}_\theta = \arg\max_{\theta\in\Theta} \hat{Q}(\theta)$ exists and $\hat{\theta}_n \to \theta_0$ almost surely as $n \to \infty$.

- **Theorem: Consistency of MLE.** Suppose Assumptions **A.1-A.4** hold, $\widehat{\theta}_n$ is the MLE of $\theta$. Then as $n \to \infty$,

$$\widehat{\theta}_n \xrightarrow{a.s.} \theta_0.$$

## 8.3 Asymptotic Properties of MLE

- **Lemma.** Suppose $f(x, \theta)$ is a PDF model and $f(x, \theta)$ is continuously differentiable with respect to $\theta \in \Theta$, where $\theta$ is an interior point in parameter space $\Theta$. Then for all $\theta$ in the interior of $\Theta$,

$$\int_{-\infty}^{\infty} [\frac{\partial \ln f(x, \theta)}{\partial \theta}] f(x, \theta) dx = 0.$$

- **Lemma: Information Matrix Equality.** Suppose a PDF model $f(x, \theta)$ is twice continuously differentiable respect to $\theta \in \Theta$, where $\theta$ is an interior point in parameter space $\Theta$. Define

$$\begin{aligned} I(\theta) &= \int_{-\infty} \infty [\frac{\partial \ln f(x, \theta)}{\partial \theta}]^2 f(x, \theta) dx, \\ H(\theta) &= \int_{-\infty}^{\infty} [\frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2}] f(x, \theta) dx. \end{aligned}$$

Then for all $\theta$ in the interior of $\Theta$,

$$I(\theta) + H(\theta) = 0.$$

A similar result holds for a PMF model.

- **Proof.**

$$
\begin{aligned}
H(\theta) &= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta^2}\left\{\log[f(x;\theta)]\right\} f(x,\theta)dx \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left\{\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}\right\} f(x,\theta)dx \\
&= -\int_{-\infty}^{\infty} \frac{1}{f^2(x;\theta)}\left[\frac{\partial f(x;\theta)}{\partial\theta}\right]^2 f(x,\theta)dx + \int_{-\infty}^{\infty} \frac{1}{f(x;\theta)}\frac{\partial^2 f(x;\theta)}{\partial\theta^2} f(x,\theta)dx \\
&= -\int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta}\log[f(x,\theta)]\right]^2 f(x,\theta)dx + \int_{-\infty}^{\infty} \frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx \\
&= -I(\theta) + 0.
\end{aligned}
$$

# 8.3 Asymptotic Properties of MLE

- **Theorem: Asymptotic Normality of MLE.** Suppose Assumptions **A.1**-**A.6** hold, $\widehat{\theta}_n$ is the MLE of $\theta$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N[0, -H^{-1}(\theta_0)] \stackrel{d}{=} N[0, I^{-1}(\theta_0)],$$

where

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta} \log[f(x,\theta)]\right]^2 f(x,\theta)dx$$

is called the *Fisher information matrix* for $f(x,\theta)$.

- **Remarks:**
  - The Fisher information matrix $I(\theta_0)$ measures the degree of the curvature of the log-likelihood function at $\theta_0$. If $I(\theta_0)$ is large, it is easy to estimate $\theta$, if $I(\theta_0)$ is small, it is difficult to estimate $\theta$.
  - Asymptotic normality of MLE can be used to construct *confidence interval* or *test* of parameter $\theta$.
  - When $X_1, X_2, \cdots$ are not independent, *e.g.*, when $\{X_t\}$ is a time series, the consistency of MLE also holds under certain conditions, but its *asymptotic efficiency* can be different.

# 8.4 Method of Moments and Generalized Method of Moments

## 8.4.1 Method of Moments Estimation

- **Method of Moments Estimator(MME):** Suppose $X_1, \cdots, X_n$ is a random sample from the population distribution $f_X(x;\theta)$, where $\theta \in \Theta$ is a $p \times 1$ vector.
  - First find a $p \times 1$ vector $W(X_i)$, and compute its expectation $M(\theta) = E_\theta[W(X_i)]$.
  - Then solve equations
  $$M(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} W(X_i).$$
  The solution $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$ is called the *method of moments estimator*.

- **Remark:** Usually, we will choose $W(X_i) = (X_i, \cdots, X_i^p)'$, then $M(\theta) = (E(X_i), \cdots, E(X_i^p))'$.

### 8.4.1 Method of Moments Estimation

- **Example:** Let $X_1, \cdots, X_n$ be a random sample from the EXP($\lambda$) distribution with the pdf $f(x) = \lambda e^{-\lambda x}$ for $x > 0$.
  - The first moment is $E_\lambda(X_1) = 1/\lambda$. Solving equation
  $$\frac{1}{n}\sum_{i=1}^{n} X_i = 1/\widehat{\lambda}_1,$$
  we obtain $\widehat{\lambda}_1 = 1/\overline{X}_n$.
  - The second moment is $E_\lambda(X_1^2) = 2/\lambda^2$. Solving equation
  $$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = 2/\widehat{\lambda}_2^2,$$
  we obtain $\widehat{\lambda}_2 = \sqrt{2n/\sum_{i=1}^{n} X_i^2}$.

- **Remark:** In this example, the estimator obtained by using the first moment is the same as the MLE. But in many cases, the MME is not as *efficient* as the MLE (*i.e.*, MME has a larger MSE).

### 8.4.1 Method of Moments Estimation

- **Example:** Suppose that we want to estimate both $\mu$ and $\sigma^2$ based on a random sample $X_1, \cdots, X_n$ from some distribution with mean $\mu$ and variance $\sigma^2$.

- **Solution:**
  - Solve equations

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \widehat{\mu},$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \widehat{\mu}^2 + \widehat{\sigma}^2.$$

  - We obtain $\widehat{\mu} = \overline{X}_n$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

- **Remark:** In this example, we do not require any knowledge on the functional form of the population distribution $f(x;\theta)$.

## 8.4.1 Method of Moments Estimation

- **Example:** Suppose $X_1, \cdots, X_n$ is a random sample from uniform distribution $U[-\theta, \theta]$. Find the estimator for $\theta$.

- **Solution:**

  – Because $E(X) = 0$, we can not obtain an estimator from the first moment.

  – The second moment is

$$E(X_i^2) = \int_{-\theta}^{\theta} \frac{1}{2\theta} x^2 \, dx = \frac{1}{3}\theta^2,$$

which gives the estimator

$$\widehat{\theta} = \sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}.$$

## 8.4.2 Generalized Method of Moments Estimation: GMM

- **Generalized Method of Moments Estimator(GMM):** In econometrics, $E_\theta[W(X_i)]$ can not be computed in some cases, due to the fact that the population distribution of $X_i$ is unknown. However, some constraints

$$E[m(X_i, \theta_0)] = 0$$

must hold for the true parameter $\theta_0$. This may follow from some economic and financial theory. We can obtain an estimator $\widehat{\theta}$ for $\theta_0$ by solving equation

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \widehat{\theta}) = 0$$

without assuming the population distribution of $X_i$.

## 8.4.2 Generalized Method of Moments Estimation: GMM

- **Example:** An investor who maximizes an intertemporal utility function

$$\max_{\{c_t\}} U(c_t, c_{t+1}) = \max_{\{c_t\}} \{u(c_t) + \beta E_t[u(c_{t+1})]\}$$

subject to an intertemporal budget constraint will choose a sequence of consumptions $\{c_t\}$ that satisfies the first order condition

$$P_t = \beta E\left[\frac{u'(c_{t+1})}{u'(c_t)} Y_{t+1} \mid I_t\right],$$

where $Y_{t+1}$ is the payoff of an asset at time $t+1$, $P_t$ is the price of the asset at time $t$, and $E_t(\cdot) = E(\cdot \mid I_t)$ is the conditional expectation given the information set $I_t$ available at time $t$. Then

$$E[m(X_{t+1}, \theta)] = E\left\{\left[\beta \frac{u'(c_{t+1})}{u'(c_t)} Y_{t+1} - P_t\right] Z_t\right\} = 0$$

where $X_{t+1} = (c_t, c_{t+1}, P_t, Y_{t+1}, Z_t)$, $Z_t$ is the *instrumental variable*, which is a function of $I_t$. $\theta$ denotes the parameter in $u(\cdot)$ and the discount factor $\beta$.

## 8.4.2 Generalized Method of Moments Estimation: GMM

- **GMM:** Suppose the parameter $\theta \in \Theta$ is a $p \times 1$ vector. We can use $q > p$ constraints and obtain the estimator from solving the optimization problem

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \{\widehat{m}_n(\theta)' \widehat{W}_n^{-1} \widehat{m}_n(\theta)\},$$

where $\widehat{m}_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(X_i, \theta)$, $\widehat{W}_n \xrightarrow{p} W$, a positive definite matrix.

- **Remarks:**
  - We can choose $\widehat{W}_n \equiv I_q$, the identity matrix.
  - There exists some optimal choice of $\Sigma$, which can give an asymptotically most efficient estimator.
  - MME is a special case of GMM estimator with $q = p$ constraints.
  - Under certain conditions, MLE can also be considered as a special case of GMM estimator with $q = p$ constraints (FOC).

## 8.4.2 Generalized Method of Moments Estimation: GMM

- **Remarks:**
  - In GMM, it does not require any knowledge on the functional form of the population distribution $f(x;\theta)$.
  - GMM estimator may be less efficient than MLE if MLE assumes the correct functional form of $f(x;\theta)$ .

- **Theorem: Existence of GMM Estimator.** Suppose that with probability one, $\widehat{m}_n(\theta)'\widehat{W}_n\widehat{m}_n(\theta)$ is continuous over $\Theta$ and $\Theta$ is a compact set. Then there exists a global minimizer $\widehat{\theta}$ that solves the problem

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\{\widehat{m}_n(\theta)'\widehat{W}_n^{-1}\widehat{m}_n(\theta)\}.$$

## 8.5 Asymptotic Properties of GMM

**Assumptions:**

- **A.1:** $X_1, X_2, \cdots$ are i.i.d. from some population distribution.
- **A.2:** The $q \times 1$ vector function $m(x, \theta)$ is continuous in $(x, \theta)$ and the absolute value of each dimension is bounded by a nonnegative function $b(x)$ with $E[b(X)] < \infty$, where the expectation $E(\cdot)$ is taken under the unknown population distribution.
- **A.3:** There exists a **unique** $p \times 1$ parameter value $\theta_0$ in the interior of $\Theta$ such that $E[m(X_i, \theta_0)] = 0$, where the expectation $E(\cdot)$ is taken under the population distribution.
- **A.4:** $\Theta$ is closed and bounded ($\Theta$ is a compact set).

## 8.5 Asymptotic Properties of GMM

- **A.5:** $\widehat{W}_n \xrightarrow{a.s.} W$, where $\widehat{W}_n^{-1}$ is bounded, $\widehat{W}_n$ and $\Sigma$ are positive definite and nonsingular.

- **A.6:** The parameter value $\theta_0$ is an interior point of the parameter space $\Theta$.

- **A.7:**

  (1) The functions $\frac{\partial}{\partial\theta}m(x;\theta)$, $\frac{\partial^2}{\partial\theta^2}m(x;\theta)$ are continuous in $(x,\theta)$, and the absolute values of their component functions are bounded by a nonnegative function $b(x)$ with $\int_{-\infty}^{\infty} b(x)f(x;\theta_0)dx < \infty$.

  (2) The $q \times q$ matrix $V \triangleq E[m(X_1,\theta_0)'m(X_1,\theta_0)]$ is bounded and nonsingular.

  (3) The $q \times p$ matrix $G(\theta_0) \triangleq E\left[\frac{\partial}{\partial\theta}m(x;\theta_0)\right]$ is of full rank.

## 8.5 Asymptotic Properties of GMM

- **Theorem: Consistency of GMM.** Suppose Assumptions **A.1**-**A.5** hold, $\widehat{\theta}_n$ is the GMM estimation of $\theta$. Then $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$.

- **Theorem: Asymptotic Normality.** Suppose Assumptions **A.1**-**A.7** hold, $\widehat{\theta}_n$ is the GMM estimation of $\theta$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Psi V \Psi'),$$

where $V = E[m(X_1,\theta_0)m(X_1,\theta_0)']$, $\Phi = \left[G(\theta_0)'W^{-1}G(\theta_0)\right]^{-1} G(\theta_0)'W^{-1}$. Moreover, if $W = V$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, [G(\theta_0)V^{-1}G(\theta_0)']^{-1}).$$

## 8.5 Asymptotic Properties of GMM

- **Theorem: Asymptotic Efficiency.** Put $\Omega_0 = [G(\theta_0)V^{-1}G(\theta_0)']^{-1}$. Then

$$\Omega - \Omega_0 \text{ is positive semi-definite (PSD)}$$

for all finite and nonsingular matrix W.

## 8.6 Mean Squared Error Criterion

- **Definition: Mean Squared Error(MSE).** Let $\theta$ be a population parameter. The MSE of an estimator $\hat{\theta}_n = \hat{\theta}(\mathbf{X}^n)$ of the parameter $\theta$ is defined as

$$MSE(\hat{\theta}_n) = E_\theta(\hat{\theta}_n - \theta)^2,$$

where $E_\theta(\cdot)$ denotes the expectation which is taken under the joint distribution $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ of the random sample $\mathbf{X}^n$, or equivalently under the sampling distribution of $\mathbf{X}^n$.

- **Definition: Bias.** The bias of a point estimator $\hat{\theta}_n$ of parameter $\theta$ is defined as

$$Bias_\theta(\hat{\theta}_n) = E_\theta(\hat{\theta}_n) - \theta.$$

An estimator $\hat{\theta}_n$ for $\theta$ is called an *unbiased* estimator for $\theta$ if the bias $E_\theta(\hat{\theta}_n) - \theta = 0$.

## 8.6 Mean Squared Error Criterion

- **Example.** Suppose $\mathbf{X}^n$ is an IID random sample from some population with mean $\mu$ and variance $\sigma^2$. Find an unbiased estimator for $var_\theta(\bar{X}_n)$.

- **Theorem: MSE Decomposition.**

$$E_\theta(\hat{\theta}_n - \theta)^2 = \mathrm{var}_\theta(\hat{\theta}_n) + [\mathrm{Bias}_\theta(\hat{\theta})]^2.$$

- **Definition: Relative Efficiency.** An estimator $\hat{\theta}_n$ for parameter $\theta$ is said to be more efficient than another estimator $\tilde{\theta}_n$ for the same parameter in terms of MSE if

$$MSE(\hat{\theta}_n) \leq MSE(\tilde{\theta}_n).$$

- **Example.** Let $(X_1, X_2)$ be an IID random sample. Two estimators for $\mu$ are $\hat{\mu}_1 = \bar{X}_n = \frac{1}{2}(X_1 + X_2)$, $\hat{\mu}_2 = \frac{1}{3}(X_1 + 2X_2)$. Which estimator is better?

## 8.7 Best Unbiased Estimators

- **Problem:** We want to find the best estimator that has the smallest MSE.
  - Unfortunately, such a best estimator is very difficult to obtain, because the class of estimators we have to compare is very huge.
  - For simplicity, we focus on the class of all *unbiased estimators* and find the best estimator within this class.

- **Definition: Generalized Unbiased Estimator.** $\hat{\gamma}_n = \gamma(\mathbf{X}^n)$ is an unbiased estimator for the parameter $\tau(\theta)$ if

$$E_\theta(\hat{\gamma}_n) = \tau(\theta), \quad \text{for all } \theta \in \Theta.$$

  When $\tau(\theta) = \theta$, we return to the previous definition of the unbiased estimator for parameter $\theta$.

- **Remark:** If $\hat{\gamma}_n$ is an unbiased estimator of $\theta$, $\tau(\hat{\gamma}_n)$ is **not** necessary to be an unbiased estimator of $\tau(\theta)$.

## 8.7 Best Unbiased Estimators

- **Definition: Uniform Best Unbiased Estimator.** Let $\Gamma$ be a class of unbiased estimators of parameter $\tau(\theta)$, where $\theta \in \Theta$, $\Theta$ is a known parameter space. An estimator $\hat{\gamma}_n \in \Gamma$ is a *uniformly best unbiased estimator* for $\tau(\theta)$ within the class $\Gamma$ if

(1) $E_\theta(\hat{\gamma}_n^*) = \tau(\theta)$ for all $\theta \in \Theta$,

(2) For any estimator $\hat{\gamma}_n \in \Gamma$ of $\hat{\gamma}_n$, $\text{var}_\theta(\hat{\gamma}_n^*) \le \text{var}_\theta(\hat{\gamma}_n)$ for all $\theta \in \Theta$.

- **Remarks:**
  - $\hat{\gamma}_n^*$ is the "best" in terms of variance (or MSE), it is also called the *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.
  - In some cases, we can use some other criteria instead of MSE to measure the performance of estimators.

- **Example.** Let $\mathbf{X}^n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The sample variance $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and the MLE estimator $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are two estimators for $\sigma^2$. Which is more efficient in terms of MSE?

- **Solution.**

$$\mathrm{MSE}(S_n^2) = E_\theta(S_n^2 - \sigma^2)^2 = \mathrm{var}_\theta(S_n^2) + [\mathrm{Bias}_\theta(S_n^2)]^2 = \frac{2\sigma^4}{n-1}.$$

$$\begin{aligned}\mathrm{MSE}(\hat{\sigma}_n^2) &= (1-\frac{1}{n})^2\frac{2\sigma^4}{n-1} + \frac{\sigma^4}{n^2} = \frac{n-1}{n}\frac{2n-1}{2n}\frac{2\sigma^4}{n-1} \\ &< \frac{2\sigma^4}{n-1} = \mathrm{MSE}(S_n^2).\end{aligned}$$

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Theorem: Cramer-Rao Lower Bound; Cramer-Rao Inequality; Information Inequality.** Let $\mathbf{X}^n$ be a random sample with joint PMF/PDF $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$, and let $\hat{\gamma}_n = \gamma(\mathbf{X}^n)$ be any estimator of parameter $\tau(\theta)$ where $E_\theta(\hat{\gamma}_n)$ is a differentiable function of $\theta$. Suppose the joint PMF/PDF $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$ of the random sample $\mathbf{X}^n$ satisfies the condition that
$$\frac{d}{d\theta}\int_{\mathcal{R}^n} h(\mathbf{x}^n) f_{\mathbf{X}^n}(\mathbf{x}^n, \theta) d\mathbf{x}^n = \int_{\mathcal{R}^n} h(\mathbf{x}^n) \frac{\partial f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{\partial \theta} d\mathbf{x}^n,$$
for any function $h : \mathcal{R}^n \to \mathcal{R}$ with $E_\theta|h(\mathbf{X}^n)| < \infty$, where $E_\theta(\cdot)$ is taken over $f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)$. Then for all $n > 0$ and all $\theta \in \Theta$,
$$\mathrm{var}_\theta(\hat{\gamma}_n) \geq B_n(\theta) \equiv \frac{\left[\frac{dE_\theta(\hat{\gamma}_n)}{d\theta}\right]^2}{E_\theta\left[\frac{\partial \ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{\partial \theta}\right]^2}.$$
In particular, when $E_\theta(\hat{\gamma}_n)$ is unbiased for parameter $\tau(\theta)$, we have
$$B_n(\theta) = \frac{[\tau'(\theta)]^2}{E_\theta\left[\frac{\partial \ln f_{\mathbf{X}^n}(\mathbf{x}^n, \theta)}{\partial \theta}\right]^2}.$$

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Theorem: Cramer-Rao Lower Bound Under IID Random Samples.** Let $\mathbf{X}^n$ be an IID random sample from population PMF/PDF $f(x,\theta)$, and let $\hat{\gamma}_n = \gamma(\boldsymbol{X}^n)$ be any estimator of $\tau(\theta)$, where $E_\theta(\gamma(\boldsymbol{X}^n))$ is a differentiable function of $\theta \in \Theta$. Suppose

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} h(x)f(x,\theta)dx = \int_{-\infty}^{\infty} h(x)\frac{\partial f(x,\theta)}{\partial\theta}dx$$

for all $h(x)$ with $E_\theta|h(X)| < \infty$. Then for all $n$,

$$\mathrm{var}_\theta(\hat{\gamma}_n) \geq B_n(\theta) \triangleq \frac{\left[\frac{d}{d\theta}E_\theta(\hat{\gamma}_n)\right]^2}{nI(\theta)},$$

where $I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right]$ is called the *Fisher information matrix.*

When $E_\theta(\hat{\gamma}_n)$ is unbiased for $\tau(\theta)$, then

$$\geq B_n(\theta) = \frac{[\tau'(\theta)]^2}{nE_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right]^2} = \frac{[\tau'(\theta)]^2}{nI(\theta)},$$

- **Proof.**

  – $\text{var}_\theta(\hat{\gamma}_n) \geq \frac{\text{cov}^2_\theta\left[\hat{\gamma}_n, \frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta)\right]}{\text{var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta)\right]}$ (why?).

  – Because

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right] &= \int\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] f(x;\theta)\,dx \\ &= \int\left[\frac{\partial}{\partial\theta}f(x;\theta)\right] dx = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = 0, \end{aligned}$$

  so

$$\text{var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta)\right] = n\text{var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right] = nE_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right]^2.$$

  – $\text{cov}_\theta\left[\hat{\gamma}_n, \frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta)\right] = E_\theta\left[\hat{\gamma}_n\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}_n;\theta)\right] = \frac{d}{d\theta}E_\theta(\hat{\gamma}_n)$.

  – Hence, $\text{var}_\theta(\hat{\gamma}_n) \geq B_n(\theta)$.

- **Remarks:**
  - Usually, $\text{var}_\theta(\hat{\gamma}_n) = B_n(\theta)$ does not imply that $\hat{\gamma}_n$ is the estimator with the smallest variance, because the value of $B_n(\theta)$ depends on $\hat{\gamma}_n$.
  - In the class of unbiased estimators, $B_n(\theta)$ does not depend on $\hat{\gamma}_n$. If $\hat{\gamma}_n$ is an unbiased estimator of $\tau(\theta)$ and $\text{MSE}_{\hat{\gamma}_n}(\theta) = \text{var}_\theta(\hat{\gamma}_n) = B_n(\theta)$, then $\hat{\gamma}_n$ is UMVUE.
  - For biased estimator of $\tau(\theta)$, it is possible that the MSE is less than $B_n(\theta) = \frac{[\tau'(\theta)]^2}{nI(\theta)}$.
  - MLE approximately achieves the Cramer-Rao lower bound of unbiased estimators when $n$ is large. (MLE may be biased.)

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Example:** Let $X_1, \cdots, X_n$ be an i.i.d random sample from Poisson($\lambda$) distribution with pmf $f(x;\lambda) = e^{-\lambda}\lambda^x/x!$, $x = 0, 1, \cdots$. (For Poisson($\lambda$) distribution, $E_\lambda(X_1) = \lambda$, $E_\lambda(X_1^2) = \lambda^2 + \lambda$.)
  - Consider the unbiased estimator $\overline{X}_n$ of $\lambda$,we have $\mathrm{Var}_\lambda(\overline{X}_n) = \lambda/n$.
  - The Cramer-Rao lower bound of unbiased estimators is

$$\begin{aligned}\frac{1}{nI(\lambda)} &= \frac{1}{nE_\lambda\left[\frac{\partial}{\partial\lambda}\log f(X_1;\lambda)\right]^2}\\ &= \frac{1}{nE_\lambda\left[\frac{X_1}{\lambda} - 1\right]^2}\\ &= \lambda/n.\end{aligned}$$

  - $\overline{X}_n$ is UMVUE of $\lambda$.

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Example:** Let $X_1, \cdots, X_n$ be an i.i.d. $N(\mu, \sigma^2)$ random sample. Show that $S_n^2$ does not attain the Carmer-Rao lower bound.

- **Solution:**

  – $\mathrm{var}(S_n^2) = 2\sigma^4/(n-1)$.

  – $\log f(x, \theta) = -\log\sqrt{2\pi} - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$.

  – $\frac{\partial}{\partial(\sigma^2)}\log f(x, \theta) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}$.

  – The Cramer-Rao lower bound of unbiased estimators of $\sigma^2$ is

$$\frac{1}{nE_\theta\left[\frac{\partial}{\partial(\sigma^2)}\log f(X_1;\theta)\right]^2} = \frac{2\sigma^4}{n}.$$

  – $\mathrm{Var}(S_n^2) = \frac{2\sigma^4}{n-1} > B_n(\theta)$. Because $S_n^2$ is UMVUE of $\sigma^2$ (it is a function of complete statistic), the Carmer-Rao lower bound is not attainable in this case.

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Theorem.** Suppose $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ is the joint PMF/PDF of the random sample $\boldsymbol{X}^n$ and $\hat{\gamma}_n = \gamma(\boldsymbol{X}^n)$ is an unbiased estimator for parameter $\tau(\theta)$, where $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ and $\hat{\gamma}_n$ satisfy the conditions in the Cramer-Rao lower bound theorem. Then the estimator $\hat{\gamma}_n$ attains the Cramer-Rao lower bound if and only if

$$\hat{\gamma} - \tau(\theta) = a(\theta)\frac{\partial \ln L_n(\theta|\boldsymbol{X}^n)}{\partial \theta}$$

for some function $a : \Theta \to \mathbb{R}$.

- **Theorem: Rao-Blackwell.** Let $\hat{\gamma}$ be any unbiased estimator of $\tau(\theta)$, and let $T_n = T(\boldsymbol{X}^n)$ be a sufficient statistic for $\theta$. Define $\phi(T_n) = E_\theta(\hat{\gamma} \mid T_n)$, then $\phi(T_n)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

- **Proof.**
  - Because $T_n$ is a sufficient statistic for $\theta$, $P(X_1, \cdots, X_n \mid T_n)$ does not depend on $\theta$. Hence $\phi(T_n) = E_\theta(\hat{\gamma} \mid T_n)$ also does not depend on $\theta$, it is an estimator.
  - Because $E_\theta[\phi(T_n)] = E_\theta[E(\hat{\gamma} \mid T_n)] = E_\theta(\hat{\gamma}) = \tau(\theta)$, $\phi(T_n)$ is unbiased.
  - Because
$$\mathrm{var}_\theta(\hat{\gamma}) = \mathrm{var}_\theta[E(\hat{\gamma} \mid T_n)] + E_\theta[\mathrm{var}(\hat{\gamma} \mid T_n)],$$
    so $\mathrm{var}_\theta[\phi(T_n)] \leq \mathrm{var}_\theta(\hat{\gamma})$.
- **Remarks:** If the best unbiased estimator exists, it must be a function of any sufficient statistic.

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Definition: Complete Statistic.** Statistic $T_n = T(\boldsymbol{X}^n)$ is called a *complete statistic* for distribution family $f(x, \theta)$ if $E_\theta[g(T_n)] = 0$ for all $\theta \in \Theta$ implies $P_\theta[g(T_n) = 0] = 1$ for all $\theta \in \Theta$.

- **Theorem: Complete Statistics in the Exponential Family.** Let $\boldsymbol{X}^n$ be IID random variables from an exponential family with PMF/PDF of the form

$$f(x, \theta) = h(x)c(\theta) \exp\big\{ \sum_{j=1}^{k} w_j(\theta) t_j(x) \big\}, \text{ for } -\infty < x < \infty,$$

where $\theta = (\theta_1, \cdots, \theta_k)'$. Then the statistic

$$T(X^n) = \big( \sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i) \big)$$

is complete if $\{(w_1(\theta), \cdots, w_k(\theta)) : \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$.

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Theorem:** Let $T_n$ be a **complete sufficient** statistic for a parameter $\theta$, and let $\phi(T_n)$ be a function of $T_n$. Then $\phi(T_n)$ is the uniform best unbiased estimator (or UMVUE) of $\tau(\theta) = E_\theta[\phi(T)]$.

- **Remark:** If $\hat{\gamma}$ is an unbiased estimator of $\tau(\theta)$ and $T_n$ is a complete sufficient statistic for a parameter $\theta$, then $E(\hat{\gamma} \mid T_n)$ is the uniform best unbiased estimator of $\tau(\theta)$.

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Example:** Suppose $X_1, \cdots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$.
  - $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a complete sufficient statistic for $(\mu, \sigma^2)$.
  - $\overline{X}_n$ is the UMVUE of $\mu$.
  - $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\overline{X}_n^2\right)$ is the UMVUE of $\sigma^2$. However, $S_n^2$ is not the best among all estimators.
  - $\text{MSE}_{S_n^2}(\sigma^2) = E\left(S_n^2 - \sigma^2\right)^2 = \frac{2\sigma^4}{n-1}$. ($\frac{n-1}{\sigma^2}S_n^2$ follows a $\chi_{n-1}^2$ distribution.)
  - Let $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. Then

$$
\begin{aligned}
\text{MSE}_{\widehat{\sigma^2}}(\sigma^2) &= \text{Bias}^2_{\widehat{\sigma^2}}(\sigma^2) + \text{Var}(\widehat{\sigma^2}) \\
&= \frac{1}{n^2}\sigma^4 + \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} \\
&= \frac{2n-1}{n^2}\sigma^4 < \text{MSE}_{S_n^2}(\sigma^2).
\end{aligned}
$$

## 8.8 Cramer-Rao Lower Bound - An Alternative Method

- **Example:** Let $X_1, \cdots, X_n$ be i.i.d. random variables from $N(\theta, 1)$. Then $\overline{X}_n$ is an unbiased estimator of $\theta$.
  - $\mathrm{Var}_\theta(\overline{X}_n) = 1/n$.
  - Let $\phi = E_\theta(\overline{X}_n \mid X_1)$, then $\phi = \frac{1}{n}X_1 + \frac{n-1}{n}\theta$.
  - $E(\phi) = \theta$, $\mathrm{Var}_\theta(\phi) = 1/n^2$.
  - However, $\phi$ is not an estimator because it depends on $\theta$.

- **Conlusion**

## 8.8' Loss Function and Risk Function

- **Problem:** How to evaluate the performance of different estimators?

- **Definition: Loss Function.** A *loss function* $\mathcal{L}(\widehat{\theta}, \theta)$ is a nonnegative function that measures the difference between the estimated value $\widehat{\theta}$ and the true value $\theta$. Usually, $\mathcal{L}(\widehat{\theta}, \theta)$ is a non-decreasing function of $|\widehat{\theta} - \theta|$.

- **Example:** Suppose $\widehat{\theta} = T(X_1, \cdots, X_n)$. The loss function can be
  - Absolute error loss:

$$\mathcal{L}[T(X_1, \cdots, X_n), \theta] = |T(X_1, \cdots, X_n) - \theta|.$$

  - Squared error loss:

$$\mathcal{L}[T(X_1, \cdots, X_n), \theta] = [T(X_1, \cdots, X_n) - \theta]^2.$$

# 8.8' Loss Function and Risk Function

- **Definition: Risk Function.** The performance of an estimator $T(X_1, \cdots, X_n)$ can be evaluated as the average loss

$$
\begin{aligned}
R_T(\theta) &= E_\theta\{\mathcal{L}[T(X_1, \cdots, X_n), \theta]\} \\
&= \int \mathcal{L}[T(x_1, \cdots, x_n), \theta]\, dF(x_1, \cdots, x_n; \theta) \\
&= \begin{cases} \sum_{x_1, \cdots, x_n} \mathcal{L}[T(x_1, \cdots, x_n), \theta] f(x_1, \cdots, x_n; \theta) & \text{for d.r.v.,} \\ \int \cdots \int \mathcal{L}[T(x_1, \cdots, x_n), \theta] f(x_1, \cdots, x_n; \theta)\, dx_1 \cdots dx_n, & \text{for c.r.v..} \end{cases}
\end{aligned}
$$

$R_T(\theta)$ is called a *risk function.*

## 8.8' Loss Function and Risk Function

- **Definition:** The estimator $T_1$ is *R-dominating* estimator $T_2$, or is *R-better* than $T_2$, if we have

$$R_{T_1}(\theta) \leq R_{T_2}(\theta) \quad \text{for all } \theta \in \Theta \ .$$

- **Remark:** There are cases that two estimators are *not comparable*: the two risk functions cross each other, *i.e.*, one is below the other for some $\theta$, and above it for some other $\theta$.

## 8.8' Loss Function and Risk Function

- **Definition:** The *mean squared error* (MSE) of an estimator $T$ is

$$\text{MSE}_T(\theta) = E_\theta[T(X_1, \cdots, X_n) - \theta]^2.$$

- **Definition:** The difference

$$B_T(\theta) = E_\theta(T) - \theta$$

is called the *bias* of estimator $T$. An estimator $T$ such that $B_T(\theta) = 0$ for every $\theta$ will be called *unbiased*.

- **Theorem: MSE Decomposition.** The mean squared error of an estimator is the sum of its variance and square of the bias, that is

$$\text{MSE}_T(\theta) = \text{Var}_\theta(T) + B_T^2(\theta).$$

- **Example:** Suppose that we take a random sample $X_1, \cdots, X_n$ from the $U[0,\theta]$ distribution. Find the MSE of different estimators of $\theta$.

- **ANS:**

  – $X_{(k)}/\theta \sim \text{Beta}(k, n-k+1)$.

  – $T_1 = X_{(n)}$, then $E_\theta(T_1) = \frac{n\theta}{n+1}$, and

$$\begin{aligned}\text{Var}_\theta(T_1) &= \frac{n\theta^2}{(n+1)^2(n+2)},\\ \text{MSE}_{T_1}(\theta) &= \frac{2\theta^2}{(n+1)(n+2)}.\end{aligned}$$

  – $T_2 = \frac{n+1}{n}X_{(n)}$, then $E_\theta(T_2) = \theta$, and

$$\text{MSE}_{T_2}(\theta) = \text{Var}_\theta(T_2) = \frac{\theta^2}{n(n+2)}.$$

  $T_2$ is MSE-better than $T_1$.

- – $T_3 = X_{(1)} + X_{(n)}$, then $E_\theta(T_3) = \theta$, and

$$\begin{aligned}
\mathrm{MSE}_{T_3}(\theta) &= \mathrm{Var}_\theta(T_3) \\
&= \mathrm{Var}(X_{(1)}) + \mathrm{Var}(X_{(n)}) + 2\mathrm{Cov}(X_{(1)}, X_{(n)}) \\
&= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{n\theta^2}{(n+1)^2(n+2)} + 2\,\frac{1}{n}\,\frac{n\theta^2}{(n+1)^2(n+2)} \\
&= \frac{2\theta^2}{(n+1)(n+2)}.
\end{aligned}$$

  – $T_4 = (n+1)X_{(1)}$, then $E_\theta(T_4) = \theta$, and

$$\mathrm{MSE}_{T_4}(\theta) = \mathrm{Var}_\theta(T_4) = \frac{n\theta^2}{n+2}.$$

- – $T_5 = 2\overline{X}_n$, then $E_\theta(T_5) = \theta$, and

$$\mathrm{MSE}_{T_5}(\theta) = \mathrm{Var}_\theta(T_5) = \frac{\theta^2}{3n}.$$

  – $T_6 = \frac{n+2}{n+1} X_{(n)}$, then $E_\theta(T_6) = \frac{n(n+2)\theta}{(n+1)^2}$, and

$$\begin{aligned} \mathrm{Var}_\theta(T_6) &= \frac{n(n+2)\theta^2}{(n+1)^4}, \\ \mathrm{MSE}_{T_6}(\theta) &= \frac{\theta^2}{(n+1)^2}. \end{aligned}$$

  – $T_6$ is MSE-better than $T_1, T_2, \cdots, T_5$.

- **Remark:** Unbiased estimators are not necessary MSE-better than biased estimators.

## Additional 8.9 Bayes Estimators

- **Bayes Estimators:** In classical statistical methods (for example, in MME and MLE), $\theta$ is considered as a fixed value in the parameter space $\Theta$. In the Bayesian method, the parameter $\theta$ is considered as a random variable $\theta : S \to \Theta$, following some distribution $\pi(\theta)$. Here $S$ is the sample space.

- **Remark:** Suppose $\pi(\theta)$ is the *prior* distribution of the parameter $\theta$. Given the i.i.d. observations $x_1, \cdots, x_n$ from the population $f(x \mid \theta)$, the conditional pdf/pmf of $\theta$ given $x_1, \cdots, x_n$, referred to as the *posterior* distribution, is

$$f(\theta \mid x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n, \theta)}{f(x_1, \cdots, x_n)} = \frac{\pi(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)}{\int_\Theta \pi(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)\, d\theta}.$$

## Additional 8.9 Bayes Estimators

- **Example:** Suppose random variables $X_1, \cdots, X_n$ are from independent Bernoulli trials with the same probability of success $\theta$, and suppose $\theta$ follows a beta distribution with parameters $\alpha$ and $\beta$, that is,

$$\pi(\theta) = C\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1,$$

where $C$ is the normalizing constant. If a realization $x_1, \cdots, x_n$ is observed, find the posterior distribution of $\theta$.

- **ANS:**
  - We have

$$f(x_1, \cdots, x_n, \theta) = C\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

  - Conditional on $x_1, \cdots, x_n$, the posterior distribution of $\theta$ follows a beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$.

## Additional 8.9 Bayes Estimators

- **Average Risk:** In Bayesian methods, the performance of an estimator $\widehat{\theta} = T(\boldsymbol{X}_n) = T(X_1, \cdots, X_n)$ can be measured by the *average risk*

$$E\left[R_T(\theta)\right] = \int_{\Theta} R_T(\theta)\pi(\theta)d\theta.$$

- **Remarks:**
  - Estimators are always comparable under Bayesian setup because the average risk is a real value.
  - We can rewrite the average risk as

$$\begin{aligned}\int_{\Theta} R_T(\theta)\pi(\theta)d\theta &= \int_{\Theta}\left[\int_{\mathcal{X}} L(T,\theta)f(\boldsymbol{x}_n \mid \theta)d\boldsymbol{x}_n\right]\pi(\theta)d\theta \\ &= \int_{\mathcal{X}}\left[\int_{\Theta} L(T,\theta)f(\theta \mid \boldsymbol{x}_n)d\theta\right]f(\boldsymbol{x}_n)d\boldsymbol{x}_n,\end{aligned}$$

    where $L(T,\theta)$ is the loss function. $\int_{\Theta} L(T,\theta)f(\theta \mid \boldsymbol{x}_n)d\theta$ is called the *posterior expected loss.*

## Additional 8.9 Bayes Estimators

- **Theorem:** Consider point estimator $T$ for a real value $\theta$.
  - For squared error loss, the posterior expected loss is
    $$\int_{\Theta} (T-\theta)^2 f(\theta \mid \boldsymbol{x}_n) d\theta = E[(T-\theta)^2 \mid \boldsymbol{X}_n = \boldsymbol{x}_n],$$
    it is minimized for any given $\boldsymbol{x}$ if $T = E(\theta \mid \boldsymbol{X}_n)$. So $T = E(\theta \mid \boldsymbol{X}_n)$ is the estimator with the smallest average risk.
  - For absolute error loss, the posterior expected loss is
    $$\int_{\Theta} |T-\theta| f(\theta \mid \boldsymbol{x}_n) d\theta = E[|T-\theta| \mid \boldsymbol{X}_n = \boldsymbol{x}_n].$$
    The posterior expected loss and the average risk are minimized if $T$ is the median of $f(\theta \mid \boldsymbol{X}_n)$.

## Additional 8.9 Interval Estimation: Bayesian Intervals

- **Bayesian Intervals:** Suppose given the observation $x_1, \cdots, x_n$, the posterior density of $\theta$ is $\pi(\theta \mid x_1, \cdots, x_n)$. We can compute the probability that $\theta$ lies between two values $a$ and $b$ as

$$P(a \leq \theta \leq b \mid x_1, \cdots, x_n) = \int_a^b \pi(\theta \mid x_1, \cdots, x_n)\, d\theta.$$

## Additional 8.9 Interval Estimation: Bayesian Intervals

- **Example:** Suppose that we observe $n = 3$ Bernoulli trials with unknown probability of success $\theta$, where $\theta$ has the prior distribution beta(2,2). Assume that we record 2 successes. Then the posterior density is

$$C\theta^3(1-\theta)^2 = 60\theta^3 - 120\theta^4 + 60\theta^5, \quad 0 < \theta < 1.$$

The probability that the true value of $\theta$ lies in interval $(0, 0.2)$ equals

$$\int_0^{0.2} 60\theta^3 - 120\theta^4 + 60\theta^5 \, d\theta = 0.017.$$

- **Remark:** For a given probability $p$, there are many different choices of Bayesian interval $(a, b)$ such that $P(a \leq \theta \leq b) = p$. Usually, we want the Bayesian interval with the shortest length.

## Additional 8.9 Interval Estimation: Confidence Intervals

- **Definition: Confidence Intervals (CI).** A pair of statistics $L = L(\boldsymbol{X}_n), U = U(\boldsymbol{X}_n)$ is an level $1-\alpha$ *confidence interval* for the parameter $\theta$ if for all $\theta \in \Theta$,

$$P_\theta[L(\boldsymbol{X}_n) \leq \theta \leq U(\boldsymbol{X}_n)] = 1 - \alpha.$$

- **Remarks:**
  - $\theta$ is a fixed (nonstochastic) parameter, the end points of the interval $L(\boldsymbol{X}_n)$ and $U(\boldsymbol{X}_n)$ are random.
  - When a realization $\boldsymbol{x}_n = (x_1, \cdots, x_n)$ is observed, $\theta$ is either in $[L(\boldsymbol{x}_n), U(\boldsymbol{x}_n)]$ or not. There is no uncertainty.
  - On the contrary, Bayesian setup allows us to say that $\theta$ is inside $[L(\boldsymbol{x}_n), U(\boldsymbol{x}_n)]$ with a certain probability for a given realization $\boldsymbol{x}_n$ .
  - For a given confidence level $1-\alpha$, there are many different level $1-\alpha$ confidence intervals. Usually, we want to find the confidence interval with the shortest length.

## Additional 8.9 Interval Estimation

- **Definition:** A random variable W is called *pivotal* for $\theta$ if it depends on the sample $\boldsymbol{X}_n = (X_1, \cdots, X_n)$ and on the unknown parameter $\theta$, while its distribution does not depend on $\theta$.

- **General steps of finding the confidence interval:**
  - Find a pivotal random variable $W(\boldsymbol{X}_n; \theta)$.
  - Determine the values $q_\alpha^*$ and $q_\alpha^{**}$ (not depend on $\theta$) such that

$$P[q_\alpha^* \leq W(\boldsymbol{X}_n; \theta) \leq q_\alpha^{**}] = 1 - \alpha.$$

  - Convert the inequality $q_\alpha^* \leq W(\boldsymbol{X}_n; \theta) \leq q_\alpha^{**}$ into the form

$$L(\boldsymbol{X}_n; q_\alpha^*, q_\alpha^{**}) \leq \theta \leq U(\boldsymbol{X}_n; q_\alpha^*, q_\alpha^{**}).$$

- **Example:** Consider the random sample $X_1, \cdots, X_n$ from the distribution $N(\mu, \sigma^2)$, where $\mu$ is the unknown parameter while $\sigma^2$ is known. Find a $(1-\alpha)$-level confidence interval for $\mu$.

- **ANS:**
  - Because $\overline{X}_n - \mu$ follows a $N(0, \sigma^2/n)$ distribution, it is a pivotal random variable.
  - We have
$$P\left[-z_{\alpha/2} \leq \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \leq z_{\alpha/2}\right] = 1 - \alpha,$$
    where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$.
  - Therefore,
$$P\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right] = 1 - \alpha.$$

- **Example:** Let $X_1, \cdots, X_n$ be a random sample from the distribution $N(\mu, \sigma^2)$. Find a $(1-\alpha)$-level confidence interval for $\sigma^2$ under the cases (1) $\mu$ is known; (2) $\mu$ is unknown.

- **ANS:**

  – (1) When $\mu$ is known, the pivotal random variable is

  $$U = \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2,$$

  which follows a chi squared distribution with degrees of freedom $n$. Thus

  $$P\left[\chi^2_{1-\alpha/2,n} \leq \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2 \leq \chi^2_{\alpha/2,n}\right] = 1 - \alpha,$$

  where $\chi^2_{\alpha,n}$ is the upper $\alpha$-quantile of $\chi^2_{\alpha,n}$.

- – Therefore the $(1-\alpha)$-level confidence interval is

$$\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\alpha/2,n}} \le \sigma^2 \le \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\alpha/2,n}}.$$

  – (2) When $\mu$ is unknown, the pivotal random variable is

$$V = \sum_{i=1}^n (X_i - \overline{X}_n)^2/\sigma^2,$$

  which follows a chi squared distribution with degrees of freedom $n-1$. The $(1-\alpha)$-level confidence interval is

$$\frac{\sum_{i=1}^n (X_i-\overline{X}_n)^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{\sum_{i=1}^n (X_i-\overline{X}_n)^2}{\chi^2_{1-\alpha/2,n-1}}.$$

## Additional 8.9 Interval Estimation

- **Example:** Construct asymptotic confidence interval (or hypothesis testing) using the asymptotic normality of MLE.
  - Let $\widehat{H}_n(\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log[f(X_i, \widehat{\theta}_n)]$, we can show that $\widehat{H}_n(\widehat{\theta}_n) \xrightarrow{p} H(\theta)$.
  - According to the asymptotic normality of MLE,
$$\sqrt{-n\widehat{H}_n(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1).$$
  - Let $z_{\alpha/2}$ be the upper $\alpha/2$-quantile of $N(0,1)$. As $n \to \infty$,
$$P\left(-z_{\alpha/2} < \sqrt{-n\widehat{H}_n(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta_0) < z_{\alpha/2}\right) \to 1-\alpha,$$
which can be rewritten as
$$\lim_{n\to\infty} P\left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{-n\widehat{H}_n(\widehat{\theta}_n)}} < \theta_0 < \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{-n\widehat{H}_n(\widehat{\theta}_n)}}\right) \to 1-\alpha.$$