中国科学院数学与系统科学研究院
Academy of Mathematics and Systems Science
Chinese Academy of Sciences

第八章 参数估计和评估

洪永淼
中国科学院数学与系统科学研究院
中国科学院大学经济与管理学院

目　录

统计的抽样推断 (Sampling Inference)

- 考察一个来自总体分布 $f_X(x)$ 的随机样本 $\boldsymbol{X}^n = (X_1, \cdots, X_n)$。

- 随机样本 $\boldsymbol{X}^n$ 的一个**实现值** $\boldsymbol{x}^n$ 称为样本容量为 n 的**数据集**。

- **统计推断**的主要目的是使用观测数据 $\boldsymbol{x}^n$ 对总体分布 $f_X(x)$ 进行推断。

参数方法 (Parametric Approach)

- 通常假设一族参数候选概率分布

$$\mathbb{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$$

  - ✓ 其中 $f : \Omega \times \Theta \to \mathbb{R}^+$ 是**已知函数形式**的 PMF/PDF，
  - ✓ Ω 是随机变量 X_i 的支撑，
  - ✓ Θ 是包含 $p \times 1$ 维参数向量 θ 的所有可能取值的参数空间，其中 p 是有限且固定的正整数。
  - ✓ 参数 $\theta \in \Theta$ 的每个值对应分布族 $\mathbb{F}$ 的一个分布，θ 不同取值对应 $\mathbb{F}$ 中不同的概率分布。

正确的模型设定 (Correct Model Specification)

- 假设概率分布族 $\mathbb{F}$ 包含了生成观测数据 $\boldsymbol{x}^n$ 的未知真实总体分布 $f_X(x)$，即**存在某一参数值** $\theta_0 \in \Theta$ 满足

  $f_X(x) = f(x, \theta_0)$, 对几乎所有 $x \in \Omega$ (除一个可数实数集外)

  则称 $\mathbb{F}$ 是对总体分布 $f_X(\cdot)$ 的**正确设定**，且 θ_0 称为参数 θ 的**真实值**。
- 反之，若**不存在任何参数值** $\theta \in \Theta$，使得对几乎所有 $x \in \Omega$ 有 $f_X(x) = f(x, \theta)$，则称 $\mathbb{F}$ 是总体分布 $f_X(\cdot)$ 的**误设**。
- **举例：**若设定一族正态分布模型，但真实总体分布服从伽玛分布，则正态分布模型为误设。此时，不存在任何参数值 θ，可使对几乎所有 $x \in \Omega$，有 $f_X(x) = f(x, \theta)$。

正确模型设定图 vs. 模型误设图

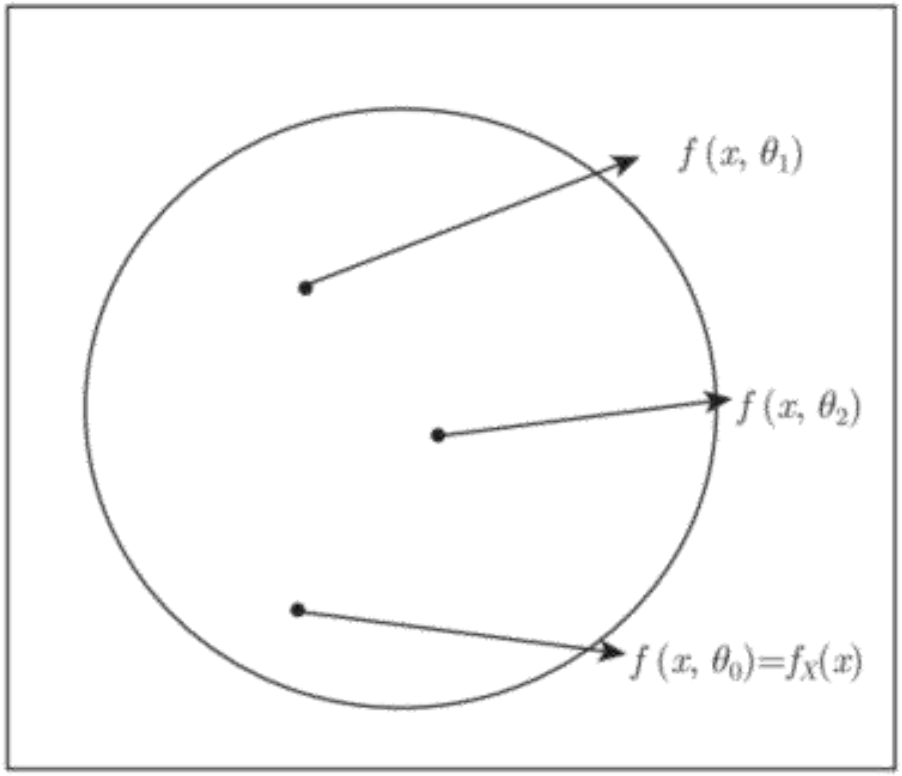
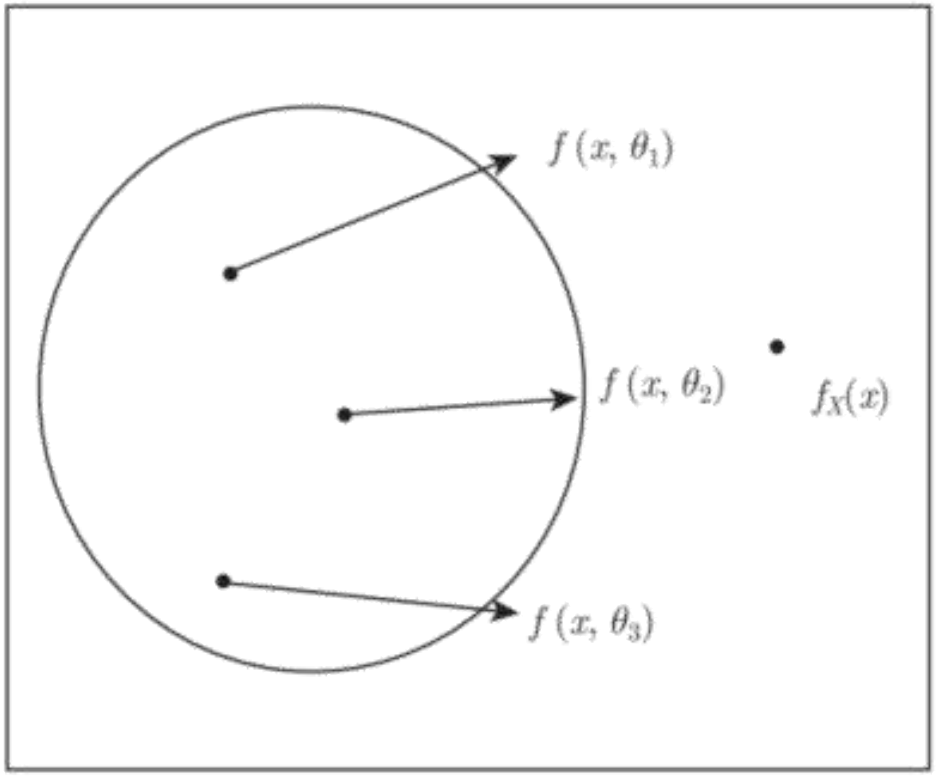


图 8.1: (a) 正确模型设定



(b) 模型误设

例 8.1：[Probit 和 Logit 离散选择模型 (Discrete Choice Models)]

- 当因变量有二元结果，如取 0 和 1 两个可能值，常使用 Probit 和 Logit 模型。
- 例如，职员是否受雇、消费者是否购买某一款汽车，以及金融危机 (如违约风险) 是否发生等。
  - ✓ Probit 模型假设

$$P(Y_i = 1 \mid X_i) = \Phi(\theta_1 + \theta_2 X_i),\ i = 1, \cdots, n$$

其中 $\Phi(\cdot)$ 是 $N(0,1)$ CDF，X_i 是解释变量，$\theta = (\theta_1, \theta_2)$。

  - ✓ Logit 模型则假设

$$P(Y_i = 1 \mid X_i) = \frac{1}{1+e^{-(\theta_1+\theta_2 X_i)}}$$

例 8.2： [经济学和金融学的生存/久期分析 (Survival/Duration Analysis in Economics and Finance)]

- 考虑对如下情况**所需的时间**建模：
  - ✓ 患癌病人还能存活多长时间、
  - ✓ 失业者花多长时间才能重新找到工作、
  - ✓ 两次交易或两个价格变化之间持续多长时间、
  - ✓ 罢工将持续多长时间、
  - ✓ 初创企业将存活多长时间、
  - ✓ 一个家庭多长时间可以脱贫、
  - ✓ 什么时候会爆发金融危机 (如信用违约风险)，等等。
- 对这类问题的研究通常称为**久期分析**或**生存分析**。

例 8.2 (Cont.):

- 假设随机变量 T_i 代表一个已发生了的经济事件的**持续时间**，其概率密度函数为 $f(t)$，概率分布函数为 $F(t)$。
- 则**生存函数** (survival function) 定义为

$$S(t) = P(T_i > t) = 1 - F(t)$$

- **风险率** (hazard rate) 则为

$$\lambda(t) = \lim_{\delta \to 0^+} \frac{P(t < T_i \leq t + \delta | T_i > t)}{\delta} = \frac{f(t)}{S(t)}$$

例 8.2 (Cont.):

- 直觉上，风险率 $\lambda(t)$ 是指事件已持续了 t 时期并将在时间点 t 结束的瞬时概率。上述公式表明，对 $\lambda(t)$ 的模型设定等价于对概率密度函数 $f(t)$ 的模型设定。从经济学视角看，对 $\lambda(t)$ 建模更易于进行经济解释。
- 风险率可能因人而异。为了**控制个体之间的异质性**，可假设个体特质风险率依赖于个体特征变量X_i (如年龄、性别、种族、教育、工作经验)，具体形式如下

$$\lambda_i(t) = e^{X_1'\theta}\lambda_0(t)$$

其中 $\lambda_0(t)$ 为**基准风险函数** (baseline hazard function)。

例 8.2 (Cont.):

- 上述模型由Cox (1972) 首先提出，称为**比例风险模型** (proportional hazard model)。
- 若**模型正确设定**，则**真实参数值**为

$$\theta_0 = \frac{\partial \ln \lambda_i(t)}{\partial X_i} = \frac{1}{\lambda_i(t)} \frac{\partial \lambda_i(t)}{\partial X_i}$$

可解释为个体 i 的特征变量 X_i 对其风险率的相对边际效应。

例 8.2 (Cont.):

- 对 θ_0 的推断可帮助理解个体特征如何对其久期产生影响。
- 例如，假设 T_i 是工人 i 的失业持续时间，则对 θ_0 的推断有助于理解某个人的具体特征，如年龄、教育程度、性别、是否参加在职培训等，如何影响其失业持续时间。这对劳动力市场具有重要的政策启示。
- 可求得给定 X_i 时 T_i 的条件概率密度函数

$$f_i(t) = \lambda_i(t)S_i(t)$$

其中生存函数为

$$S_i(t) = e^{-\int_0^t \lambda_i(s)ds}$$

- 本章即将介绍的极大似然方法可用于估计参数值 θ_0。

- 通常，真实参数值 θ_0 是未知的，需要使用观测数据 $\boldsymbol{x}^n$ 对 θ_0 做出推断。
- 传统上，统计推断问题分为两大部分，
  - ✓ 一是**参数估计** (estimation)，
  - ✓ 二是**假设检验** (hypothesis testing)。
- 本章重点考虑如何估计未知参数值 θ_0。关于 θ_0 的估计量是一个统计量，其实现值可视为对 θ_0 的估计值。
- 以下将介绍两种最常用的估计方法，即**极大似然估计法** (MLE) 和**广义矩估计法** (GMM)。**经典的矩估计方法** (MME) 是广义矩估计法的一个特例。

目　录

◆ 问题

如何根据数据集 $\boldsymbol{x}^n$ 估计未知参数值 θ_0?

- 著名统计学家罗纳德·费希尔 (Ronald A. Fisher) 提出了一种称为极大似然估计 (MLE) 的方法。他证明 MLE 能给出参数 θ 的充分统计量 (只要其存在)，且在某些准则下，MLE 是 θ_0 最有效的估计量，从而展示了该方法的优越性。
- **MLE 的基本思想**是：基于观测数据 $\boldsymbol{x}^n$，选择参数 θ 值使得随机样本 $\boldsymbol{X}^n$ 取值为观测数据 $\boldsymbol{x}^n$ 的概率最大。

定义 8.1

[似然函数 (Likelihood Function)]：给定观测数据集 $\boldsymbol{x}^n$，随机样本 $\boldsymbol{X}^n$ 的联合 PMF/PDF 作为参数 θ 的函数，

$$\hat{L}(\theta \mid \boldsymbol{x}^n) = f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$$

称为随机样本 $\boldsymbol{X}^n$ 在其取值为观测数据 $\boldsymbol{x}^n$ 时的似然函数。此外，$\ln \hat{L}(\theta \mid \boldsymbol{x}^n)$ 称为随机样本 $\boldsymbol{X}^n$ 在其取值为观测数据 $\boldsymbol{x}^n$ 时的**对数似然函数** (log-likelihood function)。

- 似然函数 $\hat{L}(\theta \mid \boldsymbol{x}^n)$ 与随机样本 $\boldsymbol{X}^n$ 取值为 $\boldsymbol{x}^n$ 的联合概率或联合概率密度在数值上是相等的。但二者的概念不同：
  - ✓ 似然函数 $\hat{L}(\theta \mid \boldsymbol{x}^n)$ 是 $\boldsymbol{X}^n$ 取值为 $\boldsymbol{x}^n$ 的概率或概率密度如何随参数 θ 值的变化而变化，
  - ✓ $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 则是给定参数 θ 值时，对 $\boldsymbol{X}^n$ 取不同数据集 $\boldsymbol{x}^n$ 的概率或概率密度的测度。
- 随机样本 $\boldsymbol{X}^n$ 的联合分布 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$不同于总体分布 $f(x, \theta)$。后者是每个随机变量 X_i 的分布。
- 极大似然估计方法就是在参数空间 Θ 上选择最大化似然函数的参数值 θ，该参数值称为**极大似然估计量**。

定义 8.2

[极大似然估计量 (Maximum Likelihood Estimator, MLE)]：令 Θ 为有限维参数空间，假设统计量 $\hat{\theta} \equiv \hat{\theta}_n(\boldsymbol{X}^n)$ 在 $\theta \in \Theta$ 上最大化 $\hat{L}(\theta \mid \boldsymbol{X}^n)$，即

$$\hat{\theta} \equiv \hat{\theta}_n(\boldsymbol{X}^n) = \arg\max_{\theta \in \Theta} \hat{L}(\theta \mid \boldsymbol{X}^n)$$

- 若 $\hat{\theta} \equiv \hat{\theta}_n(\boldsymbol{X}^n)$ 存在，则称之为未知参数值 θ_0 的 MLE。
- 给定随机样本 $\boldsymbol{X}^n$ 的一个样本点 (或数据集) $\boldsymbol{x}^n$，$\hat{\theta}_n(\boldsymbol{x}^n)$ 称为 θ_0 的一个**极大似然估计值**。
- 一般情况下，不同的样本点 $\boldsymbol{x}^n$ 对应不同的极大似然估计值。

- 根据目标函数的性质，MLE 是使观测数据 $\boldsymbol{x}^n$ 发生的概率最大的参数估计值。换言之，通过选择参数估计值 $\hat{\theta}_n(\boldsymbol{x}^n)$，MLE 使 $\boldsymbol{X}^n = \boldsymbol{x}^n$ 的概率最大化，也就是使随机样本 $\boldsymbol{X}^n$ 取观测值 $\boldsymbol{x}^n$ 的概率最大化。

- 在一些情形下，对某些数据集 $\boldsymbol{x}^n$，$\hat{L}(\theta \mid \boldsymbol{x}^n)$ 在 Θ 上的最大值未必存在。此时 MLE 不存在。(问题：可否举出一个例子？)

- 现在提供保证 MLE 存在的充分条件。

定理 8.1

[MLE 的存在性 (Existence of MLE)]：假设 $\hat{L}(\theta \mid \boldsymbol{X}^n)$ 为 $\theta \in \Theta$ 的连续函数的概率为 1，且参数空间 Θ 是**紧集** (compact set)。则存在如下问题的**全局最优解** (global maximizer) $\hat{\theta}$，即

$$\hat{\theta} \equiv \hat{\theta}_n(\boldsymbol{X}^n) = \arg\max_{\theta \in \Theta} \hat{L}(\theta \mid \boldsymbol{X}^n)$$

证明：

- 应用维尔斯特拉斯定理 (Weierstrass theorem)。
  - ✓ 图 8.2 是当参数 θ 为标量时的 MLE 图。

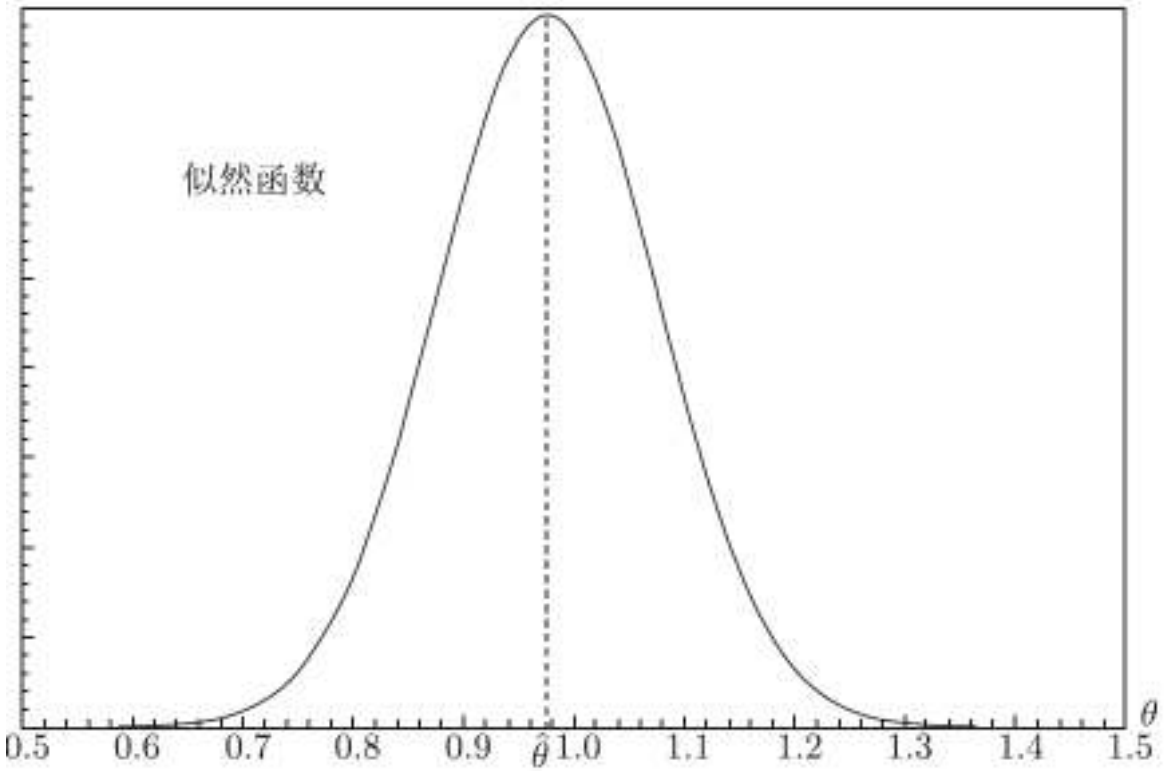


图 8.2：当参数 θ 为标量时的 MLE

- 通常求解 $\max_{\theta \in \Theta} \ln \hat{L}(\theta \mid \boldsymbol{X}^n)$ 比较方便，其中 $\ln \hat{L}(\theta \mid \boldsymbol{X}^n)$ 称为**对数似然函数**，它是 $\hat{L}(\theta \mid \boldsymbol{X}^n)$ 的严格单调增函数。
- MLE 可能并不唯一。给定观测数据集 $\boldsymbol{x}^n$，MLE 可能在参数空间 Θ 上的多个点处取值。因此，可能出现 MLE 有多个解的情形。
- 对数似然函数最大值在参数空间 Θ 上获得，而参数空间 Θ 可能存在一些约束条件。例如，当估计**广义自回归条件异方差** (generalized autoregressive conditional heteroskedasticity, GARCH) 模型时 (Bollerslev, 1986)，需要对参数 θ 施加一些约束条件以保证条件方差总是非负的。

- 当 $\hat{L}(\theta \mid \boldsymbol{X}^n)$ 是 $\theta \in \Theta$ 的平滑函数时，特别地，当 $\ln \hat{L}(\theta \mid \boldsymbol{X}^n)$ 是 $\theta \in \Theta$ 的二次连续可导函数时，$\hat{\theta}$ 容易求解。在此情形下，**MLE 存在的必要条件**是 $\hat{\theta}$ 必须满足**一阶条件**

$$\left.\frac{\partial \ln \hat{L}(\theta \mid \boldsymbol{X}^n)}{\partial \theta}\right|_{\theta=\hat{\theta}} = \mathbf{0}$$

- 若 θ 是 $p \times 1$ 维参数向量，则一阶条件包含 p 个方程。从该一阶条件可求解 $\hat{\theta}$。从图像上看，MLE $\hat{\theta}$ 位于似然函数中斜率为 0 的点，见图 8.2。

- FOC 仅为最大化的必要条件而非充分条件，故其仅为 MLE 提供了可能的候选参数值。一阶导数为零的点可能为**局部最大**(local maxima)、**全局最小**(global minima)、或**拐点**(inflection point)。
- 为了找到全局最大解，需检验**二阶条件**。若 $p \times p$ 维**样本黑塞矩阵**

$$\hat{H}(\theta) = \frac{\partial^2 \ln \hat{L}(\theta \mid \boldsymbol{X}^n)}{\partial \theta \partial \theta'}$$

对所有 $\theta \in \Theta$ 是**负定**，则 $\hat{\theta}$ 是全局最大解。

- 许多情况下可能不容易验证 $\hat{H}(\theta)$ 对所有 $\theta \in \Theta$ 是负定，而验证 $\hat{H}(\hat{\theta})$ 为负定则相对容易，从而可推出 $\hat{\theta}$ 是局部最大解。当 θ 维数较高时，用二阶导数条件检验最大值可能会比较繁冗，可以尝试其他方法。
- 需要强调，一阶导数为零仅可在函数定义域内部定位极值点。若极值出现在边界上，一阶导数可能不为零。因此，需要单独检验边界是否存在极值的情况，这可通过库恩-塔克 (Kuhn-Tucker) 定理完成。

- 当出现无法由一阶条件推出 $\hat{\theta}$ 的**解析解** (closed form solution) 的情况时，需要求 $\hat{\theta}$ 的**数值解** (numerical solution)。
- 绝大多数计算机软件都可计算数值解。

MLE 方法的步骤

1. 求出对数似然函数 $\ln\hat{L}(\theta \mid \boldsymbol{X}^n)$ 的表达式。对具有总体 PMF/PDF $f(x,\theta)$ 的 IID 随机样本，

$$\ln \hat{L}(\theta \mid \boldsymbol{X}^n) = \sum_{i=1}^{n} \ln f(X_i, \theta);$$

2. 解一阶条件 (FOC) 并求得 $\hat{\theta}$；
3. 检验二阶条件 (SOC) 以确保 $\hat{\theta}$ 为全局最大解或至少为局部最大解。

例 8.3：

- 令 $\boldsymbol{X}^n$ 为 IID $N(\mu, 1)$ 随机样本。
- 求 μ 的 MLE 估计量

解：

- 令 $\theta = \mu$。因为 $\boldsymbol{X}^n$ 为 IID $N(\mu, 1)$ 随机样本，随机样本 $\boldsymbol{X}^n$ 的似然函数为

$$\begin{aligned}\hat{L}(\mu \mid \boldsymbol{X}^n) &= \prod_{i=1}^{n} f(X_i, \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i-\mu)^2} \\ &= (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2}\end{aligned}$$

例 8.3 (Cont.)

解 (Cont.)：

- 而因**对数似然函数**为

$$\ln \hat{L}(\mu \mid \boldsymbol{X}^n) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2$$

- 由 **FOC 条件**得

$$\frac{d\ln \hat{L}(\hat{\mu} \mid \boldsymbol{X}^n)}{d\mu} \equiv \left.\frac{d\ln \hat{L}(\mu \mid \boldsymbol{X}^n)}{d\mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n}(X_i - \hat{\mu}) = 0$$

- 解得**样本均值估计量**

$$\hat{\mu} = \bar{X}_n$$

例 8.3 (Cont.)

解 (Cont.)：

- 由 SOC **条件**得

$$\frac{d^2 \ln \hat{L}(\mu \mid \boldsymbol{X}^n)}{d\mu^2} = -n < 0, \qquad \text{对所有 } \mu$$

- 因此 $\hat{\mu} = \bar{X}_n$ 为全局最大解。同时 $\hat{\mu} = \bar{X}_n$ 也是 μ 的充分统计量，参见第六章例 6.11。

例 8.4：

- 假设 $\boldsymbol{X}^n$ 为 IID $N(\mu, \sigma^2)$ 随机样本, 求 (μ, σ^2) 的 MLE。

解：

- 令 $\theta = (\mu, \sigma^2)$，则 $\boldsymbol{X}^n$ 的**对数似然函数**为

$$\ln \hat{L}(\theta \mid \boldsymbol{X}^n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

- **FOC 为**

$$\frac{\partial \ln \hat{L}\left(\hat{\theta} \mid \boldsymbol{X}^n\right)}{\partial \mu} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(X_i - \hat{\mu}) = 0$$

$$\frac{\partial \ln \hat{L}\left(\hat{\theta} \mid \boldsymbol{X}^n\right)}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0$$

例 8.4 (Cont.)：

解 (Cont.)：

- 其中 $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$。注意，此处 σ^2 被视为一个参数，而非 σ。
- 从而有

$$\hat{\mu} = \bar{X}_n$$

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

- σ^2 的 MLE 估计量 $\hat{\sigma}^2$ 与第六章定义的样本方差 $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ 有所不同。

例 8.4 (Cont.):

解 (Cont.):

- 为检验 SOC，计算**样本黑塞矩阵**

$$\hat{H}(\theta)=\begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu) & \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i-\mu)^2 \end{bmatrix}$$

- 当 $\theta=\hat{\theta}$ 时，有 $\hat{H}(\hat{\theta})$ 为负定。

$$\hat{H}(\hat{\theta})=\begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

- 因此 $\hat{\theta}$ 是局部最大解。同时，MLE $\hat{\theta}=(\hat{\mu},\hat{\sigma}^2)$ 也是 $\theta=(\mu,\sigma^2)$ 的充分统计量，参见第六章例 6.12。

◆ **问题 8.2**

假若将 σ 而非 σ^2 视作一个参数，上例中可否获得相同的 MLE 解呢?

- 答案是肯定的。从以下 MLE 不变性可求得

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

定理 8.2

[极大似然估计的不变性 (Invariance of MLE)]：假设 $\hat{\theta}$ 为 $\theta \in \Theta$ 的 MLE 估计量，$g(\cdot)$ 是参数空间 Θ 上的一一映射。则 $g(\hat{\theta})$ 是 $g(\theta)$ 的 MLE 估计量。

证明：

- 因 $g(\theta)$ 是参数空间 Θ 上的一一映射，故存在唯一反函数 $h(\cdot)$ 使得对所有 $\theta \in \Theta$ 有 $h[g(\theta)] = \theta$。
- 定义新参数 $\tau = g(\theta)$。现在求 τ 的 MLE 估计量。由反函数可得 $\theta = h(\tau)$。于是随机样本 $\boldsymbol{X}^n$ 的似然函数为

$$\hat{L}(\theta \mid \boldsymbol{X}^n) = \hat{L}[h(\tau) \mid \boldsymbol{X}^n] = \hat{L}^*(\tau \mid \boldsymbol{X}^n)$$

- 其中 $\hat{L}^*(\tau \mid \boldsymbol{X}^n)$ 是随机样本 $\boldsymbol{X}^n$ 关于新参数 τ 的似然函数。

证明 (Cont.):

- 假设 $\hat{\theta}$ 是 $\theta \in \Theta$ 的全局 MLE 估计量。则有

$$\hat{L}\left(\hat{\theta} \mid \boldsymbol{X}^n\right) \geq \hat{L}(\theta \mid \boldsymbol{X}^n), \qquad \text{对所有 } \theta \in \Theta$$

- 令 $\hat{\tau} = g(\hat{\theta})$，则 $\hat{\theta} = h(\hat{\tau})$。那么对任意 $\theta \in \Theta$，有

$$\begin{aligned}\hat{L}\left(\hat{\theta} \mid \boldsymbol{X}^n\right) &= \hat{L}[h(\hat{\tau}) \mid \boldsymbol{X}^n] \\ &= \hat{L}^*(\hat{\tau} \mid \boldsymbol{X}^n) \\ &\geq \hat{L}(\theta \mid \boldsymbol{X}^n) = \hat{L}[h(\tau) \mid \boldsymbol{X}^n] \\ &= \hat{L}^*(\tau \mid \boldsymbol{X}^n)\end{aligned}$$

- 其中因为 θ 取任意值，故 $\tau = g(\theta)$ 也取任意值。

证明 (Cont.)：

- 因此，对所有 $\tau \in \Gamma$ 有

$$\hat{L}^*(\hat{\tau} \mid \boldsymbol{X}^n) \geq \hat{L}^*(\tau \mid \boldsymbol{X}^n)$$

- 其中 $\Gamma = \{\tau : \tau = g(\theta), \text{对所有 } \theta \in \Theta\}$ 是新参数 τ 的参数空间。故 $\hat{\tau}$ 是 τ 的 MLE 估计量。

证毕。

- 以下定理说明若参数 θ 的充分统计量 $T(\boldsymbol{X}^n)$ 存在，则 MLE 估计量 $\hat{\theta}$ 可通过最大化充分统计量 $T(\boldsymbol{X}^n)$ 的似然函数求得。

定理 8.3

[MLE 的充分性 (Sufficiency of MLE)]：假设给定随机样本取值 $\boldsymbol{X}^n = \boldsymbol{x}^n$，随机样本 $\boldsymbol{X}^n$ 的似然函数为 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$，且 $T(\boldsymbol{X}^n)$ 是 θ 的充分统计量，其中参数 $\theta \in \Theta$。

- 则最大化随机样本 $\boldsymbol{X}^n$ 的似然函数的 MLE 估计量 $\hat{\theta}$ 也是最大化充分统计量 $T(\boldsymbol{X}^n)$ 的似然函数 $f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{X}^n), \theta]$ 的 MLE 估计量。

证明：

- 由定义得 MLE 估计量 $\hat{\theta} = \arg\max_{\theta \in \Theta} \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)$。因为 $T(\boldsymbol{X}^n)$ 是 θ 的充分统计量，故对任意给定样本点 $\boldsymbol{x}^n$，有

$$\begin{aligned} f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta) &= f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{x}^n), \theta] f_{\boldsymbol{X}^n \mid T(\boldsymbol{X}^n)}[\boldsymbol{x}^n \mid T(\boldsymbol{x}^n)] \\ &= f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{x}^n), \theta] h(\boldsymbol{x}^n) \end{aligned}$$

- 其中，给定 $T(\boldsymbol{X}^n) = T(\boldsymbol{x}^n)$ 时，$\boldsymbol{X}^n$ 的条件分布 $f_{\boldsymbol{X}^n \mid T(\boldsymbol{X}^n)}[\boldsymbol{x}^n \mid T(\boldsymbol{x}^n)]$ 不依赖于参数 θ，并记作函数 $h(\boldsymbol{x}^n)$（参见第六章第六节的讨论）。
- 从而有

$$\ln f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta) = \ln f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{x}^n), \theta] + \ln h(\boldsymbol{x}^n)$$

证明 (Cont.)：

- 因此在参数空间 Θ 上最大化 $\ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)$ 等价于在 Θ 上选择 θ 最大化 $\ln f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{X}^n), \theta]$，即

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)$$

$$= \arg\max_{\theta \in \Theta} \ln f_{T(\boldsymbol{X}^n)}[T(\boldsymbol{X}^n), \theta]$$

证毕。

目　录

- 因为 MLE 估计量 $\hat{\theta}$ 通常是随机样本 $\boldsymbol{X}^n$ 的非线性函数，故当随机样本 $\boldsymbol{X}^n$ 并非由正态分布生成时，对推导任意给定样本容量 n 下 MLE $\hat{\theta}$ 的均值、方差以及抽样分布将十分困难。
- 下面，应用第七章介绍的渐近理论来考察 MLE $\hat{\theta}$ 的渐近性质。即当 $n \to \infty$ 时：
  - ✓ $\hat{\theta} \xrightarrow{p} \theta_0$?
  - ✓ $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, ?\ ]$?

正则条件 (regularity conditions)

- 首先提供一组**正则条件**。为简便起见，此处假设参数 θ 为标量。

假设 8.1

[IID]

$\boldsymbol{X}^n = (X_1, \cdots, X_n)$ 为来自某未知总体分布 $f_X(x)$ 的 IID 随机样本。

正则条件 (Cont.)

假设 8.2

[正确的模型设定 (Correct Model Specification)]

(1) 对每个 $\theta \in \Theta$，$f(x,\theta)$ 是未知总体分布 $f_X(x)$ 的一个 PMF/PDF 模型，满足对支撑中的所有 x，$f(x,\theta) > 0$，其中 Θ 是有限维参数空间；

(2) 存在唯一一个参数值 $\theta_0 \in \Theta$ 使得 $f(x,\theta_0)$ 与总体分布 $f_X(x)$ 一致，即对支撑中所有的 x，有 $f(x,\theta_0) = f_X(x)$；

(3) $\ln f(x,\theta)$ 是 (x,θ) 的连续函数，且其绝对值小于非负函数 $b(x)$，满足 $E[b(X_i)] < \infty$，其中期望 $E(\cdot)$ 定义在总体分布 $f_X(x)$ 上。

正则条件 (Cont.)

假设 8.3

[紧参数空间 (Compact Parameter Space)]

参数空间 Θ 为**有界闭集**，或等价地， Θ 为**紧集**。

假设 8.4

[唯一识别 (Unique Identification)]

参数值 θ_0 是 $E[\ln f(X_i, \theta)]$ 的**唯一最优解**。

假设 8.5

[内点解 (Interior Solution)]

θ_0 是参数空间 Θ 的**内点** (interior point)。

正则 (regularity) 条件 (Cont.)

假设 8.6

[平滑和矩条件 (Smoothness and Moment Conditions)]

对每个内点 $\theta \in \Theta$，$f(x,\theta)$ 关于 θ 二阶连续可导，满足

(1) $\frac{\partial}{\partial\theta}\ln f(x,\theta)$ 和 $\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)$ 是 (x,θ) 的连续函数，其绝对值小于非负函数 $b(x)$，且 $E[b(X_i)]<\infty$，$E[b^2(X_i)]<\infty$；

(2) 函数 $H(\theta)=E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_i,\theta)\right]$ 的绝对值在 Θ 不等于零，并且其绝对值为有限值。

- 为便于分析，这一节假设标量参数 θ。以下结论可直接扩展到参数 θ 为向量的情形，但这样做并不能对 MLE 的渐近性质提供新的洞见。

正则条件的意义：

- **假设 8.2** 是关于概率分布模型 $f(x,\theta)$ 的正确设定假设。
- **假设 8.3** 中参数空间 Θ 的紧性 (compactness) 保证了 MLE 的存在 (参见定理 8.1)。
- **假设 8.4** 称为识别条件 (identification condition)，它保证MLE估计量 $\hat{\theta}$ 的概率极限 θ_0 存在并有定义。需要注意，除非 $\theta = \theta_0$，否则一般情形下

$$E[\ln f(X_i,\theta)] = \int_{-\infty}^{\infty} \ln f(x,\theta) f_X(x) dx$$
$$\neq \int_{-\infty}^{\infty} \ln f(x,\theta) f(x,\theta) dx$$

正则条件的意义 (Cont.):

- 在**假设 8.5 和 8.6** 下，可应用泰勒级数展开推导 MLE $\hat{\theta}$ 的渐近分布。
- 统计学中，函数 $\frac{\partial}{\partial\theta}\ln f(x,\theta)$ 称为记分函数 (score function；若 θ 为参数向量，将为向量函数)，而函数 $H(\theta)$ 称为黑塞函数 (Hessian function；若 θ 为参数向量，将称为黑塞矩阵)。

引理 8.1

[极值估计量引理 (Extreme Estimator Lemma)；White (1994，定理 3.4)]：假设：

(1) $Q(\theta)$ 是 $\theta \in \Theta$ 的**非随机连续实值函数**，且 $\theta_0 \in \Theta$ 是 $Q(\theta)$ 在 Θ 上的**唯一最优解**，其中 Θ 为**紧集**；

(2) 随机序列 $\hat{Q}_n(\theta)$ 是 $\theta \in \Theta$ 的连续函数的**概率为 1**；

(3) $\lim\limits_{n\to\infty} \sup\limits_{\theta\in\Theta} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = 0$ 的**概率为 1**。

则当 $n \to \infty$ 时，$\hat{\theta} = \arg\max\limits_{\theta\in\Theta} \hat{Q}_n(\theta)$ 存在，且几乎处处有 $\hat{\theta} \to \theta_0$。

- **证明：**参见 White (1994，定理 3.4 的证明)。

- $Q(\theta)$ 在**假设 (1)** 下的唯一最优解 θ_0 是一个识别条件，可保证**假设 (2)** 下的估计量 $\hat{\theta}$ 有一个明确的概率极限。
- 这里的**假设 (3)** 是一致收敛条件 (uniform convergence condition)，它可由第七章定理 7.3 即一致强大数定律 (USLLN) 推得。
- 这个假设意味着当 $n \to \infty$ 时，$\hat{Q}_n(\theta)$ 和 $Q(\theta)$ 之间在参数空间 Θ 上的最大偏离几乎处处收敛于 0。
- 在以下应用中，令 $\hat{Q}_n(\theta) = n^{-1}\ln\hat{L}(\theta \mid X^n)$ 且 $Q(\theta) = E\left[\ln f\left(X_i, \theta\right)\right]$，可证明 MLE $\hat{\theta}$ 几乎处处收敛于真实值 θ_0。

引理 8.2

[$E[\ln f(X_i,\theta)]$ 的唯一最优解 (Unique Maximizer)]：若假设 8.1 与假设 8.2 (1)-(2) 成立，则模型真实参数值 θ_0 是 $E[\ln f(X_i,\theta)]$ 在 Θ 上的唯一最优解。

证明：

- 定义相对熵

$$I[f_X(\cdot), f(\cdot,\theta)] = -\int_{-\infty}^{\infty} \ln\left[\frac{f(x,\theta)}{f_X(x)}\right] f_X(x)dx$$

证明 (Cont.)：

- 由詹森不等式与对数函数的凹凸性可推出，对所有 $\theta \in \Theta$，有 $I\left[f_X(\cdot), f(\cdot,\theta)\right] \geq 0$。故对所有 $\theta \in \Theta$，

$$\int_{-\infty}^{\infty} \ln[f(x,\theta)] f_X(x)dx \leq \int_{-\infty}^{\infty} \ln[f_X(x)] f_X(x)dx$$

- 显然，若令 $\theta = \theta_0$，在假设模型正确设定下，有 $f(x,\theta_0) = f_X(x)$，因此可在 $\theta = \theta_0$ 处获得 $E[\ln f(X_i,\theta)] = \int_{-\infty}^{\infty} \ln\left[f(x,\theta)\right] f_X(x)dx$ 的最大值。

证毕。

定理 8.4

[MLE 的一致性 (Consistency of MLE)]：若假设 8.1-8.4 成立，且 $\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(X_i, \theta)$。则当 $n \to \infty$ 时，几乎处处有

$$\hat{\theta} \to \theta_0$$

证明：

- 应用上述极值估计量引理。
- 给定假设 8.2，$Q(\theta) = E[\ln f(X_i, \theta)]$ 是 $\theta \in \Theta$ 的连续函数，且由假设 8.3 和 8.4 可知，θ_0 是 $Q(\theta)$ 在紧集 Θ 上的唯一最优解。

证明 (Cont.)：

- 现令 $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^{n} \ln f(X_i, \theta)$。则给定假设 8.1-8.3 以及第七章定理 7.3 的一致强大数定律，可得，当 $n \to \infty$ 时，几乎处处有 $\sup_{\theta \in \Theta} \left| \hat{Q}_n(\theta) - Q(\theta) \right| \to 0$。
- 因此，根据极值估计量引理，当 $n \to \infty$ 时，几乎处处有 MLE $\hat{\theta} \to \theta_0$。

证毕。

- 注意，在确立 MLE $\hat{\theta}$ 的一致性时并未要求 θ_0 是参数空间 Θ 的内点。换言之，一致性定理允许 θ_0 是角点解 (即 θ_0 在 Θ 的边界上)。相应地，亦无需假设对数似然函数 $\ln f(x,\theta)$ 对 θ 可导。
- 事实上，即使 $\ln f(x,\theta)$ 对 θ 可导，当存在角点解时，FOC 条件也可能不成立。

引理 8.3

[记分函数 (Score Function) 的期望为零]：假设 $f(x,\theta)$ 是一个 PDF 模型且 $f(x,\theta)$ 关于 $\theta \in \Theta$ 连续可导，其中 θ 是参数空间 Θ 的内点。则对所有 Θ 内部的 θ，有

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

PMF 模型也有类似结论。

证明：

- 因 $f(x,\theta)$ 是一个 PDF 模型，故对任意给定 $\theta \in \Theta$，$f(x,\theta)$ 是 PDF 。

证明 (Cont.):

- 因此对参数空间 Θ 的任意内点 θ，有

$$\int_{-\infty}^{\infty} f(x,\theta)dx = 1$$

- 求导并交换积分和求导的顺序，有

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \frac{d}{d\theta}(1) = 0$$

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

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

- **证毕。**

- 对数似然函数 $\ln f\ (X_i, \theta)$ 的一阶导数称为随机变量 X_i 的记分函数，即

$$S(X_i, \theta) = \frac{\partial \ln f\ (X_i, \theta)}{\partial \theta}$$

- 直观上，根据引理 8.3，若 X_i 服从 $f(x, \theta)$ 概率分布，关于 θ 的随机变量 X_i 的记分函数 $S(X_i, \theta)$ 的期望值将为 0。
- 也就是说，若在 $f(x, \theta)$ 概率分布下进行大量重复试验，$\ln f(x, \theta)$ 的平均斜率将为 0。

- 需要注意，除非 $\theta = \theta_0$，否则一般情形下

$$E_\theta\left[\frac{\partial \ln f(X_i,\theta)}{\partial\theta}\right] \equiv \int_{-\infty}^{\infty}\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right] f(x,\theta)dx \neq E\left[\frac{\partial \ln f(X_i,\theta)}{\partial\theta}\right]$$

其中 $E_\theta(\cdot)$ 是定义于 $f(x,\theta)$ 的期望，而 $E(\cdot)$ 是定义在未知总体分布 $f_X(x)$ 上的期望。

引理 8.4

[信息等式 (Information Matrix Equality)]：假设 PDF 模型 $f(x,\theta)$ 对关于 $\theta \in \Theta$ 二次连续可导，其中 θ 是参数空间 Θ 的内点。定义

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

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

则对所有 Θ 内点 θ，

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

同样地，PMF 模型也有类似结论。

证明：

- 等式 $\int_{-\infty}^{\infty} f(x,\theta)dx = 1$ 对 θ 求导并交换求导和积分顺序，得

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

- 变型改写为

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

证明 (Cont.)：

- 若该式对 θ 再次求导并交换求导和积分顺序，得

$$\int_{-\infty}^{\infty}\left\{\left[\frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2}\right]f(x,\theta)+\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]\frac{\partial f(x,\theta)}{\partial\theta}\right\}dx=0$$

或等价地

$$\int_{-\infty}^{\infty}\left[\frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2}\right]f(x,\theta)dx+\int_{-\infty}^{\infty}\left[\frac{\partial \ln f(x,\theta)}{\partial\theta}\right]^2 f(x,\theta)dx=0$$

证毕。

- 当随机变量 X_i 服从概率分布 $f(x,\theta)$ 时，$I(\theta)$ 称为 X_i 的**费雪信息 (Fisher information)**。$I(\theta)$ 可用于测度随机变量 X_i 所包含的关于未知参数 θ 的信息量，因为 X_i 的概率分布依赖于 θ。
- $f(x,\theta)$ 刻画了给定参数 θ 值时，我们观测到随机变量 X_i 取 x 值的概率。
  - ✓ 若 $f(x,\theta)$ 关于 θ 的变化呈现尖峰态势，则很容易从随机变量 X_i 中推断出 θ 的“真实”值，或者说，随机变量 X_i 提供了参数 θ 的大量信息。
  - ✓ 相反，若 $f(x,\theta)$ 关于 θ 的变化呈现平峰、延展态势，则很难推断 θ 的“真实”值。

- 因此，$\ln f(X_i,\theta)$ 对数似然函数斜率的绝对值大小可以提供参数值 θ 的信息。鉴于 X_i 是随机的，可以使用记分函数的方差来度量对数似然函数斜率平方的期望值，这就是费雪信息 $I(\theta)$。
- 当 θ 为向量时，$I(\theta)$ 可拓展定义为**费雪信息矩阵(Fisher information matrix)**。

- $H(\theta)$ 是概率分布 $f(x,\theta)$ 的**黑塞函数 (Hessian function)**。当 θ 为向量时， $H(\theta)$ 可拓展定义为黑塞矩阵。
- 若在概率分布 $f(x,\theta)$ 下进行大量重复试验，则 $H(\theta)$ 测度了对数似然函数 $\ln f(X_i,\theta)$ 曲率的大小。
- 引理 8.4 的信息等式表明，记分函数 $S(X_i,\theta)$ 平方的期望值等于对数似然函数 $\ln f(X_i,\theta)$ 的**曲率(即二阶导数)** 绝对值的期望值。$\ln f(X_i,\theta)$ 的曲率越大，对数似然函数从 θ 点峰值处下降越快，对数似然函数的斜率的绝对值变化便越大。
- 现证明经适当标准化处理后，MLE $\hat{\theta}$ 服从渐近正态分布。

定理 8.5

[MLE 估计量的渐近正态性 (Asymptotic Normality of MLE)]:

若假设 8.1 – 8.6 成立，则当 $n \to \infty$ 时，

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N[0, -H(\theta_0)^{-1}]$$

证明:

- 因当 $n \to \infty$ 时，几乎处处有 $\hat{\theta} \to \theta_0$，且 θ_0 是参数空间 Θ 的内点，因此当 n 充分大时，$\hat{\theta}$ 也是 Θ 内点的概率为 1。
- 故**一阶条件**为

$$\left.\frac{d\ln\hat{L}(\theta \mid \boldsymbol{X}^n)}{d\theta}\right|_{\theta=\hat{\theta}} = 0$$

证明 (Cont.):

- 或等价地

$$\frac{d}{d\theta}\sum_{i=1}^{n}\ln f\left(X_i,\hat{\theta}\right)=0$$

- 交换求导和求和顺序，得

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f\left(X_i,\hat{\theta}\right)}{\partial\theta}=0$$

- 由中值定理，有

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f\left(X_i,\theta_0\right)}{\partial\theta}+\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f\left(X_i,\bar{\theta}\right)}{\partial\theta^2}\right]\left(\hat{\theta}-\theta_0\right)=0$$

- 其中，$\bar{\theta}$ 位于 $\hat{\theta}$ 和 θ_0 之间，即存在某个 $\lambda\in(0,1)$，有 $\bar{\theta}=\lambda\hat{\theta}+(1-\lambda)\theta_0$。

证明 (Cont.):

- 当 $n \to \infty$ 时，几乎处处有

$$\left|\bar{\theta} - \theta_0\right| = \left|\lambda\left(\hat{\theta} - \theta_0\right)\right| \leq \left|\hat{\theta} - \theta_0\right| \to 0$$

- 以下，定义样本黑塞函数

$$\hat{H}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln f\left(X_i, \theta\right)}{\partial \theta^2}$$

则有

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \left[-\hat{H}(\bar{\theta})\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial \ln f\left(X_i, \theta_0\right)}{\partial \theta}$$

证明 (Cont.):

- 首先根据中心极限定理 (见第七章定理 7.6)，证明当 $n \to \infty$ 时

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ln f(X_i, \theta_0)}{\partial \theta} \xrightarrow{d} N[0, I(\theta_0)]$$

- 根据定义，记分函数

$$S_i(\theta) = \frac{\partial \ln f(X_i, \theta)}{\partial \theta}, i = 1, \cdots, n$$

- 在随机样本 $\boldsymbol{X}^n$ 为 IID 的假设下，$\{S_i(\theta_0)\}_{i=1}^n$ 也是 IID 序列。

证明 (Cont.):

- 给定假设 8.2 的模型正确设定条件(即总体分布 $f_X(x) = f(x, \theta_0)$)，有

$$
\begin{aligned}
E[S_i(\theta_0)] &= \int_{-\infty}^{\infty} \frac{\partial \ln f(x, \theta_0)}{\partial \theta} f_X(x) dx \\
&= \int_{-\infty}^{\infty} \frac{\partial \ln f(x, \theta_0)}{\partial \theta} f_X(x, \theta_0) dx \\
&= 0 \qquad \text{[引理 8.3]}
\end{aligned}
$$

- $E[S_i(\theta_0)] = 0$ 表明，当 θ_0 是参数空间 Θ 的内点时，$\max_{\theta \in \Theta} E[\ln f(X_i, \theta)]$ 的一阶条件，在 $\theta = \theta_0$ 处成立。

证明 (Cont.):

- 给定 $E[S_i(\theta_0)] = 0$，方差

$$\begin{aligned}\text{var}[S_i(\theta_0)] &= E[S_i(\theta_0)^2] \\ &= E\left[\frac{\partial \ln f(X_i,\theta_0)}{\partial\theta}\right]^2 \\ &= \int_{-\infty}^{\infty}\left[\frac{\partial \ln f(x,\theta_0)}{\partial\theta}\right]^2 f_X(x)dx \\ &= \int_{-\infty}^{\infty}\left[\frac{\partial \ln f(x,\theta_0)}{\partial\theta}\right]^2 f(x,\theta_0)dx \quad \text{[模型正确设定]} \\ &= I(\theta_0) < \infty\end{aligned}$$

- 由 IID 随机序列的中心极限定理（参见定理 7.6）可得，当 $n \to \infty$ 时，

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln f(X_i,\theta_0)}{\partial\theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S(X_i,\theta_0) \xrightarrow{d} N[0, I(\theta_0)]$$

证明 (Cont.)：

- 以下证明当 $n \to \infty$ 时，几乎处处有 $\hat{H}(\bar{\theta}) \to H(\theta_0)$，其中黑塞函数 $H(\theta)$ 由引理 8.4 定义。
- 现在，令

$$\bar{H}(\theta) \equiv E\left[\frac{\partial^2 \ln f(X_i, \theta)}{\partial \theta^2}\right]$$

$$= \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} f_X(x) dx$$

- 注意，除非 $\theta = \theta_0$，否则 $\bar{H}(\theta) \neq H(\theta_0)$。

证明 (Cont.)：

- 进一步有

$$\hat{H}(\bar{\theta}) - H(\theta_0) = [\hat{H}(\bar{\theta}) - \bar{H}(\bar{\theta})] + \left[\bar{H}(\bar{\theta}) - H(\theta_0)\right]$$

- 对第二项 $\left[\bar{H}(\bar{\theta}) - H(\theta_0)\right]$，由引理 7.7 （几乎处处连续性定理）、几乎处处有 $\bar{\theta} \to \theta_0$、概率模型正确设定以及 $\bar{H}(\theta)$ 为 θ 的连续函数（给定假设 8.6），有

$$\bar{H}(\bar{\theta}) - H(\theta_0) = \bar{H}(\bar{\theta}) - \bar{H}(\theta_0) \to 0$$

- 此处为几乎处处收敛于 0。

证明 (Cont.):

- 对第一项 $[\hat{H}(\bar{\theta}) - \bar{H}(\bar{\theta})]$，当 $n \to \infty$ 时，由定理7.3 的一致强大数定律 (USLLN)，有

$$
\begin{aligned}
|\hat{H}(\bar{\theta}) - \bar{H}(\bar{\theta})| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f\left(X_i, \bar{\theta}\right)}{\partial \theta^2} - \left\{ E\left[\frac{\partial^2 \ln f\left(X_i, \theta\right)}{\partial \theta^2} \right] \right\}_{\theta=\bar{\theta}} \right| \\
&\leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f\left(X_i, \theta\right)}{\partial \theta^2} - E\left[\frac{\partial^2 \ln f\left(X_i, \theta\right)}{\partial \theta^2} \right] \right| \\
&= \sup_{\theta \in \Theta} |\hat{H}(\theta) - \bar{H}(\theta)| \\
&\to 0
\end{aligned}
$$

- 此处为几乎处处收敛于 0。

证明 (Cont.):

- 因此，当 $n \to \infty$ 时，几乎处处有 $\hat{H}(\bar{\theta}) - H(\theta_0) \to 0$，且在 $H(\theta_0)$ 非零情况下，几乎处处有

$$\hat{H}(\bar{\theta})^{-1} \to H(\theta_0)^{-1}$$

- 根据斯勒茨基定理（参见定理 7.8），当 $n \to \infty$ 时，

$$\sqrt{n}(\hat{\theta} - \theta_0) = [-\hat{H}(\bar{\theta})]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(X_i, \theta_0)$$

$$\xrightarrow{d} N[0, H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1}]$$

- 根据引理 8.4 的信息等式和概率模型正确设定，可得 $I(\theta_0) = -H(\theta_0)$，因此

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, -H(\theta_0)^{-1}]$$

- 这里 $H(\theta_0)$ 为负，故 $-H(\theta_0)^{-1}$ 为正。**证毕。**

- 渐近正态性：$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N[0,-H(\theta_0)^{-1}]$
- 函数

$$H(\theta) \equiv E_\theta\left[\frac{\partial^2 \ln f\left(X_i, \theta\right)}{\partial \theta^2}\right]=\int_{-\infty}^{\infty} \frac{\partial^2 \ln f\left(x, \theta\right)}{\partial \theta^2} f(x, \theta) dx$$

称为PMF/PDF 模型 $f(x,\theta)$ 的黑塞函数（或黑塞矩阵，当 θ 为向量时），其中期望 $E_\theta(\cdot)$ 在 PDF 模型 $f(x,\theta)$ 下取得。该函数为负，且其绝对值大小测度了似然函数在 θ 点的曲率(degree of curvature)。

- 因此，MLE 估计量 $\hat{\theta}$ 的有效性取决于对数似然函数在真实参数值 θ_0 点处的曲率。

◆ **问题**

为什么 MLE 的渐近正态性在实际应用中有用呢?

- 主要是因为它可用于构建**置信区间估计量** (confidence interval estimators) 和进行**参数假设检验** (hypothesis tests)。比如，关于 θ_0 的一个渐近 $100(1-\alpha)\%$ 置信区间估计量为随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$，其中 $\hat{\theta}_L = \hat{\theta}_L(\boldsymbol{X}^n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(\boldsymbol{X}^n)$，满足

$$\lim_{n\to\infty} P\left(\hat{\theta}_L \le \theta_0 \le \hat{\theta}_U\right) = 1-\alpha$$

- 即当 $n \to \infty$ 时，真实参数值 θ_0 介于 $\hat{\theta}_L$ 和 $\hat{\theta}_{\mathrm{U}}$ 之间的概率趋近于 $1-\alpha$。

MLE 的渐近正态性的应用

- 当 $n \to \infty$ 时，$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, -H(\theta_0)^{-1}]$ 且几乎处处有 $\hat{H}(\hat{\theta}) \to H(\theta_0)$。根据斯勒茨基定理（参见定理 7.8），可得

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

- 因此，当 $n \to \infty$ 时，

$$P\left[-z_{\alpha/2} \leq \sqrt{-n\hat{H}(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z_{\alpha/2}\right] \to 1 - \alpha$$

- 其中 $z_{\alpha/2}$ 是 $\alpha/2$ 水平上 $N(0,1)$ 的**右侧临界值** (upper-tailed critical value)，即

$$P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

其中 $Z \sim N(0,1)$。

MLE 的渐近正态性的应用（Cont.）

- 这等价于当 $n \to \infty$ 时，

$$P\left[\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{1}{-\hat{H}(\hat{\theta})}} \le \theta_0 \le \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{1}{-\hat{H}(\hat{\theta})}}\right] \to 1-\alpha$$

- 从而得到如下渐近 $(1-\alpha)100\%$ 置信区间

$$\hat{\theta} - \sqrt{-\frac{1}{n\hat{H}(\hat{\theta})}}z_{\frac{\alpha}{2}} \le \theta_0 \le \hat{\theta} + \sqrt{-\frac{1}{n\hat{H}(\hat{\theta})}}z_{\frac{\alpha}{2}}$$

- 显然，样本容量 n 越大或对数似然函数在真实参数值 θ_0 处的曲率越大，将获得越狭的 θ_0 置信界 (confidence bounds)，即更精确的区间估计。

MLE $\hat{\theta}$ 的渐近性质

- 当 $n \to \infty$ 时：
  - ✓ $\hat{\theta} \xrightarrow{a.s.} \theta_0$
  - ✓ $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, -H(\theta_0)^{-1}]$

目　录

- **矩估计方法（method of moments estimation，MME）**：是统计学最古老的**参数估计方法**之一。
  - ✓ **基本思想**：通过对总体分布的若干阶矩与其相对应的样本矩进行匹配，获得一定数量的匹配方程以求解总体分布的未知参数值。
- 具体来说，假设 $f(x,\theta)$ 为**未知总体分布** $f_X(x)$ 的 PMF/PDF 模型，其中 $\theta \in \Theta$ 为 $p \times 1$ 维参数向量，且存在一个**未知参数**值 $\theta_0 \in \Theta$ 使得对几乎所有 x，有 $f_X(x) = f(x,\theta_0)$。
- 这意味着参数概率模型 $f(x,\theta)$ 是对总体分布 $f_X(x)$ 的**正确设定**。

- 假设 $\boldsymbol{X}^n$ 来自总体分布 $f_X(x)$ 的 IID 随机样本。$\boldsymbol{X}^n$ 的联合概率分布，$f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta) = \prod_{i=1}^n f(x_i, \theta)$。
- 首先定义一个 $p \times 1$ 维统计向量

$$\hat{m} = \hat{m}_n(\boldsymbol{X}^n)$$

- **[样本矩 (sample moment)]**：$\hat{m}$
- **[总体矩 (population moment)]**：我们可以基于模型分布计算数学期望

$$\begin{aligned} M(\theta) &= E_\theta[\hat{m}_n(\boldsymbol{X}^n)] \\ &= \int_{\mathbb{R}^n} \hat{m}_n(\boldsymbol{x}^n) f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta) d\boldsymbol{x}^n \end{aligned}$$

- 其中，数学期望 $E_\theta(\cdot)$ 定义在随机样本 $\boldsymbol{X}^n$ 的联合 PDF $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)$ 上 (若 $\boldsymbol{X}^n$ 为离散变量随机样本，则 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)$ 为 $\boldsymbol{X}^n$ 的联合 PMF，上述积分改为求和)。当 $\boldsymbol{X}^n$ 为 IID 随机样本时，$f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)=\prod_{i=1}^n f(x_i,\theta)$。数学期望 $M(\theta)$ 可称为**总体矩函数**。
- **[矩匹配 (Moment Matching)]** 其次，求解方程组

$$\hat{m}=M(\hat{\theta})$$

即选择参数值 $\hat{\theta}$ 使样本矩 $\hat{m}$ 等于总体矩 $M(\theta)$。

- 求得的解 $\hat{\theta}=\hat{\theta}_n(\boldsymbol{X}^n)$ 称为真实参数值 θ_0 的**矩估计量 (MME)**。

或等价地，可定义如下**样本矩函数**，对于 $\theta \in \Theta$，

$$\hat{m}(\theta) \equiv \hat{m}_n(\boldsymbol{X}^n) - M(\theta)$$

则参数值 θ_0 的矩估计量 $\hat{\theta}$ 是以下方程组的解

$$\hat{m}(\theta) = 0$$

- 通常，若对于某一参数值 θ_0，$\boldsymbol{X}^n$ 是总体 $f_X(x) = f\ (x, \theta_0)$ 的 IID 随机样本，则可采用以下基本步骤：
  - ✓ (1) 从总体 PMF/PDF 模型 $f(x, \theta)$ 计算若干**总体矩** $E_\theta\left(X_i^k\right)$，$k = 1, 2, \cdots$，即

$$M_k(\theta) = E_\theta\left(X_i^k\right) = \begin{cases} \int_{-\infty}^{\infty} x^k f(x, \theta) dx, & X_i \text{ 是连续随机变量} \\ \sum_{x \in \Omega x} x^k f(x, \theta), & X_i \text{ 是离散随机变量} \end{cases}$$

注意，总体矩 $M_k(\theta)$ 依赖于参数 θ。

✓ (2) 计算随机样本 $\boldsymbol{X}^n$ 的**样本矩**

$$\hat{m}_k = n^{-1}\sum_{i=1}^{n} X_i^k, k = 1,2,\cdots$$

✓ (3) 选择参数值 $\hat{\theta}$，**使样本矩分别等于相应阶数的总体矩。**

一般而言，若 θ 是一个 $p \times 1$ 维参数向量，则需 p 个矩匹配方程

$$\begin{cases} \hat{m}_1 = M_1(\hat{\theta}) \\ \hat{m}_2 = M_2(\hat{\theta}) \\ \cdots \\ \hat{m}_p = M_p(\hat{\theta}) \end{cases}$$

求解这 p 个联立方程，可得 MME 估计量 $\hat{\theta}=\hat{\theta}_n(\boldsymbol{X}^n)$。

- **问题 8.3**

  为什么 MME $\hat{\theta}$ 能够一致估计真实参数值 θ_0？

- 直观上，根据弱大数定律，当 $n \to \infty$ 时，样本矩

$$\hat{m}_k \xrightarrow{p} E\left(X_i^k\right) = \int_{-\infty}^{\infty} x^k f_X(x)dx = \int_{-\infty}^{\infty} x^k f(x,\theta_0)dx = M_k(\theta_0)$$

- 因此，若对任意 n，令 $\hat{m}_k = M_k(\hat{\theta})$，即对样本矩和总体矩进行匹配，其中 $\hat{\theta} = \hat{\theta}_n(X^n)$ 依赖于 n，则当 $n \to \infty$ 时，$M_k(\hat{\theta}) \xrightarrow{p} M_k(\theta_0)$，故当 $n \to \infty$ 时，$\hat{\theta} \xrightarrow{p} \theta_0$。

例 8.5：

- 假设 $\boldsymbol{X}^n$ 为 IID $EXP(\theta)$ 随机样本。
- 分别用矩方法和 MLE 法求参数 θ 的估计量。

例 8.5 (Cont.):

解:

- **(1) 矩估计法:** 因为指数分布的 PDF 为

$$f(x,\theta)=\begin{cases}\frac{1}{\theta}e^{-x/\theta}, & x>0\\ 0, & x\le 0\end{cases}$$

- 可得**总体均值**

$$\begin{aligned}M_1(\theta)&=E_\theta(X_i)\\&=\int_{-\infty}^{\infty}xf(x,\theta)dx\\&=\int_{0}^{\infty}x\frac{1}{\theta}e^{-x/\theta}dx\\&=\theta\end{aligned}$$

例 8.5 (Cont.):

解 (Cont.):

- 另一方面，一阶样本矩是**样本均值**

$$\hat{m}_1 = \bar{X}_n$$

- 将样本均值和总体均值在 $\hat{\theta}$ 处匹配，有

$$\hat{m}_1 = M_1(\hat{\theta}) = \hat{\theta}$$

- 则得**矩估计量**

$$\hat{\theta} = \hat{m}_1 = \bar{X}_n$$

例 8.5 (Cont.):

解 (Cont.):

- **(2) MLE 方法:** 给定 IID $EXP(\theta)$ 假设，随机样本 $\boldsymbol{X}^n$ 的似然函数为

$$\hat{L}(\theta \mid \boldsymbol{X}^n) = \prod_{i=1}^{n} f(X_i, \theta)$$
$$= \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^{n} X_i}$$

- 因此，**对数似然函数**为

$$\ln \hat{L}(\theta \mid \boldsymbol{X}^n) = -n \ln \theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i$$

例 8.5 (Cont.):

解 (Cont.):

- **一阶条件**为

$$\frac{\partial \ln \hat{L}\left(\hat{\theta} \mid \boldsymbol{X}^n\right)}{\partial \theta} = -\frac{n}{\hat{\theta}} + \frac{1}{\hat{\theta}^2}\sum_{i=1}^{n} X_i = 0$$

- 则 **MLE 估计量**为

$$\hat{\theta} = \bar{X}_n$$

- 本例中，MME 和 MLE 两种方法求得的估计量完全相同，因此二者对 θ_0 的估计具有相同的有效性。
  - ✓ **原因**：样本均值或一阶样本矩 $\bar{X}_n$ 是 θ 的充分统计量，均包含了随机样本 $\boldsymbol{X}^n$ 中关于 θ 的所有信息。

例 8.6：

- 假设 X^n 是 IID $N(\mu, \sigma^2)$ 随机样本。
- 求 $\theta = (\mu, \sigma^2)$ 的 MME 估计量。

解：

- 前两阶总体矩和样本矩分别为

$$M_1(\theta) = E_\theta(X_i) = \mu$$
$$M_2(\theta) = E_\theta\left(X_i^2\right) = \sigma^2 + \mu^2$$
$$\hat{m}_1 = \bar{X}_n$$
$$\hat{m}_2 = n^{-1}\sum_{i=1}^{n} X_i^2$$

例 8.6 (Cont.):

解 (Cont.):

- 分别对前两阶样本矩和总体矩进行匹配，得

$$\bar{X}_n = \hat{\mu}$$

$$n^{-1}\sum_{i=1}^{n} X_i^2 = \hat{\sigma}^2 + \hat{\mu}^2$$

- 则有

$$\hat{\mu} = \bar{X}_n$$

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

- MME 估计量和 MLE 估计量相同。因为对正态随机样本 $\boldsymbol{X}^n$，$(\bar{X}_n, S_n^2)$ 是 $\theta = (\mu, \sigma^2)$ 的充分统计量。

MME 的优点

- 当待估参数是总体矩或总体矩的函数时，可在**总体分布**模型 $f(x,\theta)$ **函数形式未知**的情况下使用 MME 进行估计。从这一意义上说，MME 十分**便于应用**。其中一个例子就是稳态分布的参数估计，稳态分布的 PDF 没有解析形式，虽然其特征函数具有解析形式。

MME 的缺点

- MME 仅利用有限数量的样本矩信息，使其可能**无法充分利用**随机样本 $\boldsymbol{X}^n$ 所包含的关于未知参数 θ 的**所有信息**。因此，即使在渐近意义上，**MME 估计量可能不是 θ 的最有效估计量。**
- 另一方面，基于随机样本 $\boldsymbol{X}^n$ 整个联合 PMF/PDF 的 MLE 估计量能够充分利用 $\boldsymbol{X}^n$ 所包含的关于 θ 的所有信息。
- 所以，**MLE 可能比 MME 更有效，除非后者所使用的样本矩是 θ 参数的充分统计量。**

8.4.2 广义矩估计方法

(Generalized Method of Moments Estimation)

- 计量经济学中，总体矩即 $M(\theta) = E_\theta[\hat{m}_n(\boldsymbol{X}^n)]$ 常不可得，因为经济系统的总体分布未知。
- 然而，经济理论通常蕴含了在真实模型参数值 θ_0 处必须满足的若干矩条件。换言之，经济学家经常通过一组矩条件刻画经济理论或经济假说。
- 故可用经济理论所蕴含的这些矩条件来估计真实模型参数值 θ_0。

- 具体而言，假设 θ 为一个 $p \times 1$ 维参数向量，且存在一个 $q \times 1$ 矩函数 $m(X, \theta)$ 使得对某个未知参数值 $\theta_0 \in \Theta$，满足

$$E[m(X, \theta_0)] = \mathbf{0}$$

- 其中，$E(\cdot)$ 是关于随机变量 X 的概率分布（通常未知）的数学期望，这些矩条件可能来自经济理论（如理性期望模型的欧拉方程），且 $q \geq p$。

例 8.7：

- 某投资者最大化其跨期效用函数如下

$$\max_{\{C_t\}} E\left[\sum_{j=0}^{\infty} \beta^j u(C_{t+j}) \mid I_t\right]$$

- 这里，投资者受到跨期预算约束限制，
  - ✓ 参数 β 是时间折现因子，
  - ✓ $u(C_t)$ 是投资者在第 t 期消费 C_t 的效用，
  - ✓ I_t 是投资者在第 t 期所拥有的信息集，
  - ✓ $E_t(\cdot) = E(\cdot \mid I_t)$ 是给定第 t 期的信息集投资者 I_t 下的条件期望。

例 8.7 (Cont.):

- 投资者将选择最佳消费序列 $\{C_t\}$ 满足一阶条件

$$P_t = \beta E\left[\frac{u'(C_{t+1})}{u'(C_t)} Y_{t+1} \middle| I_t\right]$$

- 其中，Y_{t+1} 是投资者在第 $t+1$ 期的资产的随机总收益率，P_t 是第 t 期的资产价格。
- 该一阶条件称为**欧拉方程** (Euler equation)。该方程表示，在均衡状态下，资产的现行价格应等于其风险补偿后的未来资产的预期总收益。
- $\beta \frac{u'(C_{t+1})}{u'(C_t)}$ 称为**随机折现因子** (stochastic discount factor)，其测量了投资者的风险态度。

例 8.7 (Cont.):

- 定义**随机定价误差** (stochastic pricing error) 为

$$\varepsilon_{t+1}(\theta) = \beta \frac{u'(C_{t+1})}{u'(C_t)} Y_{t+1} - P_t$$

- 欧拉方程可等价地由如下条件矩刻画

$$E[\varepsilon_{t+1}(\theta_0) \mid I_t] = 0$$

- 这表明理性投资者每一时期都没有系统性定价误差。

例 8.7 (Cont.):

- 现在，定义矩函数

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

其中 $X_{t+1}=(C_t,C_{t+1},P_t,Y_{t+1},Z_t')'$, $Z_t\in I_t$ 是所谓的工具变量 (instrumental variables)。

- 应用第五章定理 5.24 (重复期望法则)，可得

$$\begin{aligned}E[m(X_{t+1},\theta_0)]&=E\{E[m(X_{t+1},\theta_0)\mid I_t]\}\\&=0\end{aligned}$$

其中 $E(\cdot)$ 是未知总体分布下的无条件期望。

例 8.7 (Cont.):

- 本例中，参数 θ 从哪里来呢?
  - ✓ 除时间折现因子 β 之外，效用函数中的某个 (些) 参数可刻画投资者的风险厌恶程度。
  - ✓ 例如，当投资者具有**不变相对风险厌恶的效用函数** $u(C_t)=\frac{C_t^{\gamma}-1}{\gamma}$ 时，参数 $\gamma=-C_t\frac{u''(C_t)}{u'(C_t)}$ 度量了投资者的**风险厌恶程度**。此处，$\theta=(\beta,\gamma)'$。

例 8.8 [资本资产定价模型 (CAPM)]:

- 定义 Y_t 为第 t 期 k 个资产 (或资产组合) 超额收益率的 $k \times 1$ 维随机向量。这 k 个资产的超额收益率可用市场超额收益率来解释:

$$\begin{aligned} Y_t &= \alpha_0 + \beta_0 R_{mt} + \varepsilon_t \\ &= \theta_0{}' W_t + \varepsilon_t \end{aligned}$$

- 其中
  - ✓ $W_t = (1, R_{mt})'$ 是二元向量,
  - ✓ R_{mt} 是市场组合的超额收益率,
  - ✓ $\theta_0 = (\alpha_0, \beta_0)'$ 是 $2 \times k$ 参数矩阵,
  - ✓ ε_t 是 $k \times 1$ 维的随机扰动项，并有 $E(\varepsilon_t \mid W_t) = 0$。

例 8.8 (Cont.):

- 这是标准资本资产定价模型（CAPM），其表示任何资产的期望超额收益率只取决于不可避免的市场系统风险，而与资产的特质风险无关。
- 令 $X_t = (Y_t', W_t')'$。定义 $q \times 1$ 矩函数

$$m(X_t, \theta) = W_t \otimes (Y_t - \theta' W_t)$$

其中 $q = 2k$，$\otimes$ 表示 Kronecker 积。

- 当 CAPM 成立时，有

$$E[m(X_t, \theta_0)] = \mathbf{0}$$

- 这 q 个矩条件构成了估计和检验标准 CAPM 的基础。

- 一般来说，假设给定 q 个总体矩条件

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

其中

- ✓ $m(X_i, \theta_0)$ 是 $q \times 1$ 维随机向量，
- ✓ θ_0 是 $p \times 1$ 维真实参数向量，
- ✓ $\mathbf{0}$ 是 $q \times 1$ 维零向量。

- 通过选择参数值使样本矩等于总体矩 $E[m(X_i, \theta_0)] = \mathbf{0}$，可得 θ_0 的估计量 $\hat{\theta}$：

$$\hat{m}(\hat{\theta}) \equiv n^{-1} \sum_{i=1}^{n} m\left(X_i, \hat{\theta}\right) = \mathbf{0}$$

- 在实际应用中，可用 q 个矩条件，其中 $q \geq p$，即矩条件的数目不少于未知参数的数目。
- 一般来说，无法求得严格满足方程 $\hat{m}(\theta) = \mathbf{0}$ 的解，因为方程的数目通常大于未知参数的数目。
- 因此，只能选择一个估计量 $\hat{\theta}$ 使 $\hat{m}(\theta)$ 尽量接近零向量。

- 具体而言，GMM 估计量是如下最小化二次型问题的解

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

其中 $\hat{W}$ 为一个 $q \times q$ 随机非奇异对称矩阵，满足当 $n \to \infty$ 时，$\hat{W} \xrightarrow{p} W$，其中 W 是一个 $q \times q$ 非随机非奇异对称矩阵。

- 简单起见，可选择 $\widehat{W} = I$，这里 I 为 $q \times q$ 维单位矩阵。在此情形下，目标函数可写为

$$\widehat{m}(\theta)'\widehat{W}^{-1}\widehat{m}(\theta) = \sum_{k=1}^{q} \widehat{m}_k^2(\theta)$$

即 q 个样本矩的平方和。

- 此处，每个样本矩是等权重的。实际上，第 8.4.1 节介绍的经典 MME 是 GMM 估计的一个特例，其中 $q = p$ 且 $\widehat{W} = I$。

定理 8.6

[GMM 估计量的存在性 (Existence of GMM)]：假设二次型 $\hat{m}(\theta)'\hat{W}^{-1}\hat{m}(\theta)$ 在 $\theta \in \Theta$ 上连续的概率为 1，且参数空间 Θ 是一个紧集。则存在满足如下最小化问题的全局最优解

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

- **证明：**根据维尔斯特拉斯定理 (Weierstrass theorem)。
  - ✓ 类似于 MLE 和 MME，GMM 的矩条件常常是高度非线性的，因此 GMM 估计量 $\hat{\theta}$ 可能不存在解析解，只能通过数值方法求解。

目　录

GMM 估计量 $\hat{\theta}$ 的渐近性质

- 当 $n \to \infty$ 时:
  - ✓ $\hat{\theta} \xrightarrow{a.s.} \theta_0$?
  - ✓ $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N[0, ?]$?
  - ✓ 最优的 $\hat{W}$ 是什么?

正则条件

假设 8.7

[IID]

$\boldsymbol{X}^n = (X_1, \cdots, X_n)$ 为来自未知总体分布 $f_X(x)$ 的 IID 随机样本。

假设 8.8

[矩函数 (Moment Function)]

$q \times 1$ 维矩函数 $m(x, \theta)$ 对 (x, θ) 连续且各元素的绝对值小于非负函数 $b(x)$，满足 $E[b(X_i)] < \infty$，其中期望 $E(\cdot)$ 定义在未知总体分布 $f_X(x)$ 上。

正则条件 (Cont.)

假设 8.9

[唯一识别 (Unique Identification)]

在参数空间 Θ 中，有且仅有一个 $p \times 1$ 维参数值 θ_0，满足 $E[m(X_i, \theta_0)] = \mathbf{0}$，且 $p \leq q$。

假设 8.10

[紧参数空间 (Compact Parameter Space)]

$p \times 1$ 维参数空间 Θ 为有界闭集。

正则条件 (Cont.)

假设 8.11

[权重矩阵 (Weighting Matrix)]

当 $n \to \infty$ 时，几乎处处有 $q \times q$ 随机权重矩阵 $\hat{W} \to W$，其中 W 为对称有界非奇异矩阵。

假设 8.12

[内点解 (Interior Solution)]

$p \times 1$ 维未知参数值 θ_0 是参数空间 Θ 的内点。

正则条件 (Cont.)

假设 8.13

[平滑和矩条件 (Smoothness and Moment Conditions)]

(1) 函数 $\frac{\partial}{\partial\theta}m(x,\theta)$ 和 $\frac{\partial^2}{\partial\theta\partial\theta'}m(x,\theta)$ 对 (x,θ) 连续，且其各元素的绝对值小于非负函数 $b(x)$，满足 $E[b(X_i)]<\infty$；

(2) $q\times q$ 对称矩阵 $V=E[m(X_i,\theta_0)m(X_i,\theta_0)']$ 有界且非奇异；

(3) $q\times p$ 梯度矩阵 (gradient matrix) $G(\theta_0)=E\left[\frac{\partial}{\partial\theta}m(X_i,\theta_0)\right]$ 满秩 (等于 p，给定 $p\leq q$)。

定理 8.7

[GMM 估计量的一致性 (Consistency of GMM)]：若假设 8.7-8.11 成立，则当 $n \to \infty$ 时，几乎处处有

$$\hat{\theta} \to \theta_0$$

- **证明：** 与 MLE 存在性的证明类似。注意此处 θ_0 可为角点解，从而一阶条件未必成立。

定理 8.8

[GMM 估计量的渐近正态性 (Asymptotic Normality of GMM)]:

若假设 8.7-8.13 成立，则

(1) **[渐近正态]** 当 $n \to \infty$ 时

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$$

其中

$$\Omega = \Psi V \Psi'$$

$$V = E[m(X_1, \theta_0) m(X_1, \theta_0)'],$$

且 $\Psi = [G(\theta_0)' W^{-1} G(\theta_0)]^{-1} G(\theta_0)' W^{-1}$。

(2) **[最优权重矩阵]** 若 $W = V$，则当 $n \to \infty$ 时，

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

证明：

- **(1) 定义目标函数**

$$\hat{Q}(\theta) = \hat{m}(\theta)'\hat{W}^{-1}\hat{m}(\theta)$$

- 这里事先设定的权重矩阵 $\hat{W}$ 并非 θ 的函数，故可得如下 $p \times 1$ 维一阶条件

$$\frac{d\hat{Q}(\hat{\theta})}{d\theta} = 2\hat{G}(\hat{\theta})'\hat{W}^{-1}\hat{m}(\hat{\theta}) = 0$$

其中，$q \times p$ 维样本矩阵

$$\hat{G}(\theta) = \frac{d\hat{m}(\theta)}{d\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(X_i,\theta)}{\partial\theta}$$

证明 (Cont.):

- 由中值定理，有

$$\hat{m}(\hat{\theta}) = \hat{m}(\theta_0) + \hat{G}(\bar{\theta})\left(\hat{\theta} - \theta_0\right)$$

其中 $\bar{\theta}$ 位于 $\hat{\theta}$ 和 θ_0 之间，即 $\bar{\theta} = \lambda\hat{\theta} + (1-\lambda)\theta_0$，$\lambda \in [0,1]$。

- 将该式代入上述一阶条件，得

$$\hat{G}(\hat{\theta})'\hat{W}^{-1}\hat{m}(\theta_0) + \hat{G}(\hat{\theta})'\hat{W}^{-1}\hat{G}(\bar{\theta})\left(\hat{\theta} - \theta_0\right) = \mathbf{0}$$

- 类似证明定理 8.5 关于 MLE 渐近正态性时对样本黑塞矩阵 $\hat{H}(\bar{\theta})$ 的推理思路，可证当 $n \to \infty$ 时，几乎处处有

$$\hat{G}(\hat{\theta}) \to G(\theta_0)$$
$$\hat{G}(\bar{\theta}) \to G(\theta_0)$$

- 上述两式的证明需要应用第七章定理 7.3 的一致强大数定律 (USLLN)，梯度函数 (gradient function) $G(\theta) = E\left[\frac{\partial}{\partial\theta} m(X_i, \theta)\right]$ 的连续性，以及当 $n \to \infty$ 趋于无穷时，几乎处处有

$$\left\|\bar{\theta} - \theta_0\right\| \leq \left\|\hat{\theta} - \theta_0\right\| \to 0$$

- 同时，由假设 8.11 可知，当 $n \to \infty$ 时，几乎处处有 $\hat{W} \to W$。因此，当 $n \to \infty$ 时，几乎处处有

$$\hat{G}(\hat{\theta})'\hat{W}^{-1}G(\bar{\theta}) \to G(\theta_0)'W^{-1}G(\theta_0)$$

- 其中，在假设 8.11 和假设 8.13(3) 下，$G(\theta_0)'W^{-1}G(\theta_0)$ 为非奇异矩阵。则对足够大的 n，存在随机逆矩阵 $\left[\hat{G}(\hat{\theta})'\hat{W}^{-1}G(\bar{\theta})\right]^{-1}$，这是因为当 $n \to \infty$ 时，几乎处处有

$$\left[\hat{G}(\hat{\theta})'\hat{W}^{-1}G(\bar{\theta})\right]^{-1} \to \left[G(\theta_0)'W^{-1}G(\theta_0)\right]^{-1}$$

- 由以上的一阶条件，得

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}-\theta_0\right) &= -\left[\hat{G}(\hat{\theta})'\hat{W}^{-1}\hat{G}(\bar{\theta})\right]^{-1}\hat{G}(\hat{\theta})'\hat{W}^{-1}\sqrt{n}\hat{m}(\theta_0)\\ &= -\hat{\Psi}\sqrt{n}\hat{m}(\theta_0)\end{aligned}$$

- 根据 IID 随机序列的中心极限定理 (定理 7.6) 和克拉默-沃尔德 (Cramer-Wold) 方法 (引理 7.13)，有

$$\sqrt{n}\hat{m}(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(X_i, \theta_0) \xrightarrow{d} N(0, V)$$

- 其中 $V = E[m(X_i, \theta_0)m(X_i, \theta_0)']$ 为矩函数 $m(X_i, \theta_0)$ 在 $\theta = \theta_0$ 时的方差-协方差矩阵。此外，当 $n \to \infty$ 时，几乎处处有

$$\begin{aligned}\hat{\Psi} &\equiv \left[\hat{G}(\hat{\theta})'\hat{W}^{-1}\hat{G}(\bar{\theta})\right]^{-1}\hat{G}(\hat{\theta})'\hat{W}^{-1} \\ &\to [G(\theta_0)'W^{-1}G(\theta_0)]^{-1}G(\theta_0)'W^{-1} \equiv \Psi\end{aligned}$$

- 由斯勒茨基定理 (定理 7.8)，当 $n \to \infty$ 时

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Psi V \Psi')$$

- **(2)** 假设 $W = V$，即当权重 W 为矩函数 $m(X_i, \theta_0)$ 的方差-协方差矩阵时，有

$$\begin{aligned}\Psi V \Psi' &= \{[G(\theta_0)'V^{-1}G(\theta_0)]^{-1}G(\theta_0)'V^{-1}\}V\{[G(\theta_0)'V^{-1}G(\theta_0)]^{-1}G(\theta_0)'V^{-1}\}' \\ &= \{[G(\theta_0)'V^{-1}G(\theta_0)]^{-1}G(\theta_0)'V^{-1}\}V\{V^{-1}G(\theta_0)[G(\theta_0)'V^{-1}G(\theta_0)]^{-1}\} \\ &= [G(\theta_0)'V^{-1}G(\theta_0)]^{-1}\end{aligned}$$

- 故当 $W = V$ 时，若 $n \to \infty$，有

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

- **证毕。**

定理 8.9

[GMM 的渐近有效性 (Asymptotic Efficiency of GMM)]：令 $\Omega_0 = [G(\theta_0)'V^{-1}G(\theta_0)]^{-1}$。则对所有有限对称且非奇异矩阵 W，$\Omega - \Omega_0$ 为半正定 (PSD)

其中 Ω 由定理 8.8 给出。

证明：

- 注意到 $\Omega - \Omega_0$ 是半正定矩阵，当且仅当 $\Omega_0^{-1} - \Omega^{-1}$ 是半正定矩阵。
- 为简化表达式，令 $G_0 = G(\theta_0)$，并分解 $q \times q$ 正定对称矩阵 $V = V^{1/2}V^{1/2}$，其中 $V^{1/2}$ 为 $q \times q$ 非奇异对称矩阵，其逆为 $V^{-1/2}$。

证明 (Cont.):

- 对非奇异对称矩阵 A, B, C，有 $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$。
- 因此有

$$\begin{aligned}\Omega_0^{-1} - \Omega^{-1} &= G_0'V^{-1}G_0 - G_0'W^{-1}G_0(G_0'W^{-1}VW^{-1}G_0)^{-1}G_0'W^{-1}G_0 \\ &= G_0'V^{-\frac{1}{2}}\left[I - V^{\frac{1}{2}}W^{-1}G_0(G_0'W^{-1}VW^{-1}G_0)^{-1}G_0'W^{-1}V^{\frac{1}{2}}\right]V^{-\frac{1}{2}}G_0 \\ &= G_0'V^{-\frac{1}{2}}\Pi V^{-\frac{1}{2}}G_0\end{aligned}$$

- 其中，$q \times q$ 维矩阵

$$\Pi \equiv I - V^{\frac{1}{2}}W^{-1}G_0(G_0'W^{-1}VW^{-1}G_0)^{-1}G_0'W^{-1}V^{\frac{1}{2}}$$

是一个幂等矩阵 (idempotent matrix)，即 $\Pi = \Pi'$，$\Pi^2 = \Pi$。

证明 (Cont.)：

- 则

$$\begin{aligned}\Omega_0^{-1}-\Omega^{-1} &= \left(G_0'V^{-\frac{1}{2}}\Pi\right)\left(\Pi V^{-\frac{1}{2}}G_0\right)\\ &= \left(\Pi V^{-\frac{1}{2}}G_0\right)'\left(\Pi V^{-\frac{1}{2}}G_0\right)\end{aligned}$$

  为半正定矩阵 (问题：为什么？)。

证毕。

GMM 估计量 vs. MLE 估计量

- 在实际应用中，若 GMM 估计量 $\hat{\theta}$ 并非充分统计量的函数，则可能通过充分统计量改进 GMM 估计量的有效性。
- 相反，MLE 估计量总为充分统计量的函数，故没有进一步改进的空间。
- 所以，从**有效性角度**考虑，一般**更偏好 MLE**。
- 但是，MLE 要求似然函数正确设定的相关信息，而这在经济理论中通常不容易获得。

目　录

- 一般而言，对同一参数 θ，不同的估计方法将给出不同的估计量。

◆ **问题**

哪个估计量是 θ 的最佳估计量?

- 例如，总体方差 σ^2 至少有两个估计量：
  - ✓ 样本方差：$S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$，
  - ✓ MLE 估计量：$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$。
- 哪个估计量更好呢?

- 直觉上，最佳估计量应该是与未知真实参数 θ 最接近的那一个。为比较不同的估计量，需要定义**适当的准则**（如距离或偏离程度）以测度估计量 $\hat{\theta}$ 和真实参数 θ 之间的接近程度。
- 关于 $\hat{\theta}$ 对 θ 的偏离程度有许多不同的测度。
  - ✓ 一般来说，**绝对距离** $|\hat{\theta} - \theta|$ 的任意增函数都可作为估计量优劣程度的测度。
  - ✓ 然而，以下定义的**均方误准则**的相对优势在于，一方面它非常易于分析，另一方面它可分解为方差和偏差的平方和，且这一分解有很好的解释。

定义 8.3

[均方误 (Mean Squared Error, MSE)]：令 θ 为总体参数，其估计量 $\hat{\theta} = \hat{\theta}_n(\boldsymbol{X}^n)$ 的均方误 (MSE) 定义为

$$\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2$$

其中 $E_\theta(\cdot)$ 表示对随机样本 $\boldsymbol{X}^n$ 的联合分布 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 取期望。

- $\mathrm{MSE}_\theta(\hat{\theta})$ 度量了估计量 $\hat{\theta}$ 和参数 θ 之间的偏离程度。
- $\hat{\theta} - \theta$ 通常称为估计误差 (estimation error)，故 $\mathrm{MSE}_\theta(\hat{\theta})$ 是估计误差大小的测度。
- MSE 并非唯一的估计量优度判断准则。但由于其直观且易于分析，因而是实际中应用最为广泛的准则。

定义 8.4

[偏差 (Bias)]：未知参数 θ 的估计量 $\hat{\theta}$ 的偏差定义为

$$\mathrm{Bias}_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$$

- 若 $\mathrm{Bias}_\theta(\hat{\theta}) = 0$，则称估计量 $\hat{\theta}$ 为 θ 的无偏估计量。
- 偏差 $\mathrm{Bias}_\theta(\hat{\theta})$ 测度了参数 θ 的估计量 $\hat{\theta}$ 的不精确程度。直观上，**估计精度 (accuracy)** 是指很多测量值的平均值与真实参数值之间的接近程度，是对系统误差的一种描述。
- 一个无偏估计量在平均意义上给出了正确估计，即对参数 θ 的估计不存在任何系统性的向上或向下的偏差。

例 8.9：

- 假设 $\boldsymbol{X}^n$ 为来自均值为 μ，方差为 σ^2 的某个总体分布的 IID 随机样本。求 $\mathrm{var}_\theta(\bar{X}_n)$ 的无偏估计量。

解：

- 令 $\theta = (\mu, \sigma^2)$ 且 $\tau = \frac{\sigma^2}{n}$。由于 $\mathrm{var}_\theta(\bar{X}_n) = \frac{\sigma^2}{n}$，$\tau$ 的无偏估计量如下

$$\hat{\tau} = \frac{S_n^2}{n}$$

这里 $E_\theta(\hat{\tau}) = n^{-1}E_\theta\left(S_n^2\right) = \mathrm{var}_\theta(\bar{X}_n) = \tau$，故

$$\mathrm{Bias}_\theta(\hat{\tau}) = E_\theta(\hat{\tau}) - \tau = 0。$$

例 8.10 (Cont.):

解 (Cont.):

- 令 $\theta = (\mu, \sigma^2)$。对参数 $\tau = \mu^2$，其无偏估计量为

$$\hat{\tau} = \bar{X}_n^2 - \frac{S_n^2}{n}$$

- 这是因为

$$\begin{aligned} E(\hat{\tau}) &= E_\theta\left(\bar{X}_n^2\right) - \frac{E_\theta\left(S_n^2\right)}{n} \\ &= \mathrm{var}_\theta(\bar{X}_n) + [E_\theta(\bar{X}_n)]^2 - \frac{E_\theta\left(S_n^2\right)}{n} \\ &= \frac{\sigma^2}{n} + \mu^2 - \frac{\sigma^2}{n} \\ &= \mu^2 \\ &= \tau \end{aligned}$$

例 8.10 (Cont.):

- **解 (Cont.):**
- 直觉上，既然样本均值 $\bar{X}_n$ 是 μ 的一个好的估计量，可预期 $\bar{X}_n^2$ 也是 μ^2 的好的估计量。
- 然而 $\bar{X}_n^2$ 是 $\bar{X}_n$ 的非线性函数，从而产生偏差 $\frac{\sigma^2}{n}$。这个偏差可用无偏估计量 $\frac{S_n^2}{n}$ 代替 $\frac{\sigma^2}{n}$ 加以修正。

定理 8.10

[MSE 分解 (MSE Decomposition)]:

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

证明:

- 利用公式 $(a+b)^2 = a^2 + b^2 + 2ab$，展开

$$\begin{aligned} E_\theta(\hat{\theta}-\theta)^2 &= E_\theta\left[\hat{\theta} - E_\theta(\hat{\theta}) + E_\theta(\hat{\theta}) - \theta\right]^2 \\ &= E_\theta\left[\hat{\theta} - E_\theta(\hat{\theta})\right]^2 + \left[E_\theta(\hat{\theta}) - \theta\right]^2 + 2E_\theta\left\{\left[\hat{\theta} - E_\theta(\hat{\theta})\right]\left[E_\theta(\hat{\theta}) - \theta\right]\right\} \\ &= E_\theta\left[\hat{\theta} - E_\theta(\hat{\theta})\right]^2 + \left[E_\theta(\hat{\theta}) - \theta\right]^2 \end{aligned}$$

- 其中，交叉乘积项

$$\begin{aligned} E_\theta\left\{\left[\hat{\theta} - E_\theta(\hat{\theta})\right]\left[E_\theta(\hat{\theta}) - \theta\right]\right\} &= E_\theta\left\{\left[\hat{\theta} - E_\theta(\hat{\theta})\right]\right\}\left[E_\theta(\hat{\theta}) - \theta\right] \\ &= 0 \cdot \left[E_\theta(\hat{\theta}) - \theta\right] \\ &= 0 \end{aligned}$$

证毕。

- $\mathrm{MSE}_\theta(\hat{\theta})$ 可分解为两部分： $\mathrm{var}_\theta(\hat{\theta})$ 和 $\mathrm{Bias}_\theta(\hat{\theta})^2$ 之和。
  - ✓ $\mathrm{var}_\theta(\hat{\theta})$ 测度了估计量 $\hat{\theta}$ 因抽样变化导致的**变异性 (variability)**，
  - ✓ $\mathrm{Bias}_\theta(\hat{\theta})^2$ 测度了估计方法的估计**精度 (accuracy)**。
- 对任何无偏估计量 $\hat{\theta}$，$\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2 = \mathrm{var}_\theta(\hat{\theta})$。
- 因此，最优无偏估计量是方差最小的无偏估计量。当然，最优无偏估计量可能不如某些有偏估计量，若后者的方差很小以至于可抵消存在的偏差。

- 可用打靶射击的简单例子直观理解 MSE 准则及其分解。
  - ✓ 若绝大多数射击都与目标很接近，那么积分很高，这对应一个很小的 MSE。
  - ✓ 若射击点分布以目标为中心（偏差很小），但在目标周围分散范围大（方差很大），或射击点分布并不分散（方差很小），但却以一个远离目标的点为中心（偏差很大），这两种情况获得的积分都很低。

MSE 的方差和偏差示意图

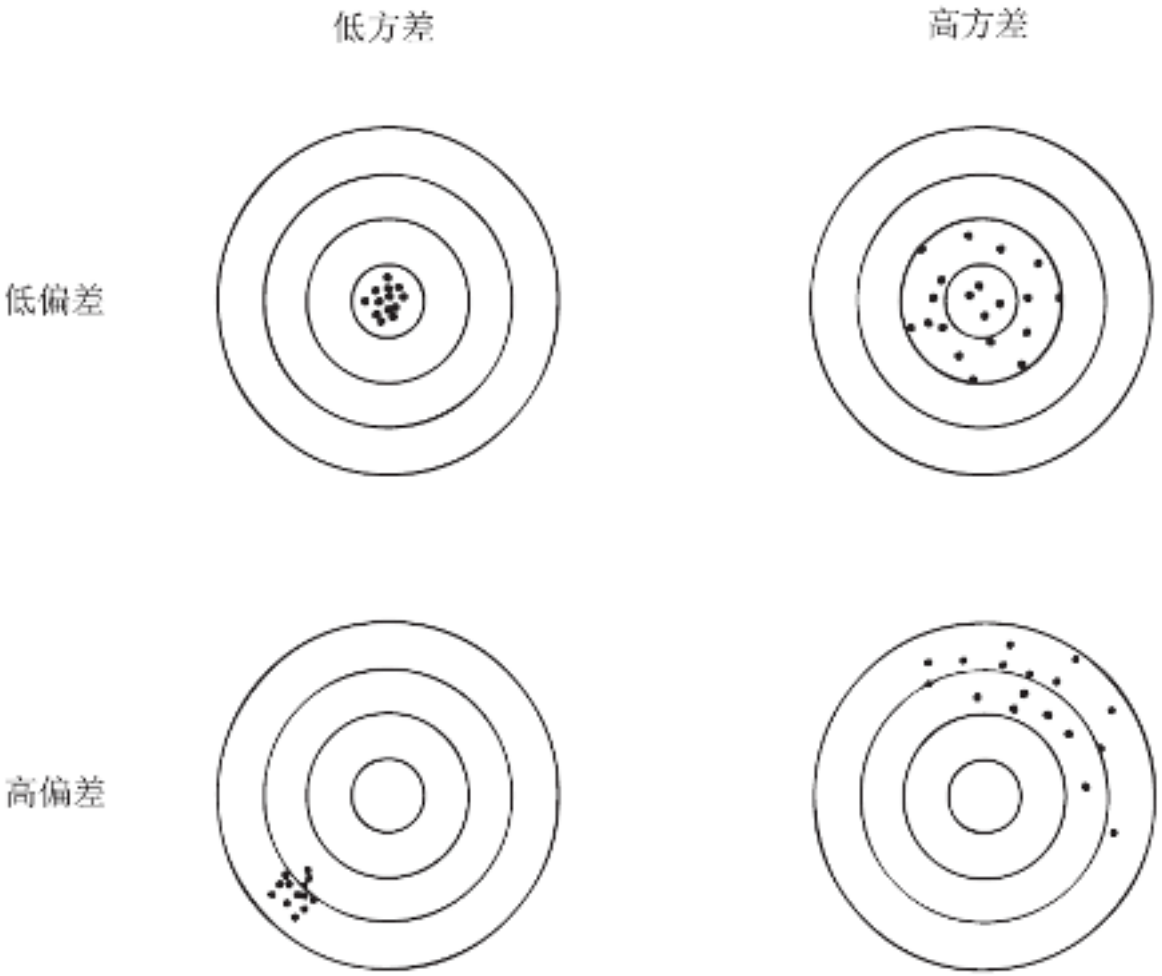


图 8.4：MSE 的方差和偏差示意图

定义 8.5

[估计量的相对有效性 (Relative Efficiency of Estimators)]：若

$$\mathrm{MSE}_\theta(\hat{\theta}) \leq \mathrm{MSE}_\theta(\tilde{\theta})$$

则称在 MSE 准则下，参数 θ 的一个估计量 $\hat{\theta}$ 较之另一估计量 $\tilde{\theta}$ 更有效。

- 评价估计量的相对有效性取决于比较准则。一般而言，不同准则将导致不同的有效性排序。本书采用实际应用最为广泛的 MSE 准则。

例 8.11：

假设 $\{X_i\}_{i=1}^{2n}$ 为来自均值 μ，方差 σ^2 的某个总体分布的 IID 随机样本。为了估计 μ，定义 $\hat{\mu}_1 = n^{-1}\sum_{i=1}^{n} X_i$，$\hat{\mu}_2 = (2n)^{-1}\sum_{i=1}^{2n} X_i$ 哪个估计量在 MSE 准则下更优?

解：

- 应用第六章定理 6.2 和定理 6.3，得
  - ✓ $E_\theta(\hat{\mu}_1) = \mu,\ \mathrm{var}_\theta(\hat{\mu}_1) = \frac{\sigma^2}{n}$,
  - ✓ $E_\theta(\hat{\mu}_2) = \mu, \mathrm{var}_\theta(\hat{\mu}_2) = \frac{\sigma^2}{2n}$。
- 则 $\mathrm{MSE}_\theta(\hat{\mu}_1) = 2\mathrm{MSE}_\theta(\hat{\mu}_2)$。因此 $\hat{\mu}_2$ 更有效。

- 直觉上，在参数估计时总希望能利用更多的样本信息。
- 因此，从统计角度而言，样本分割 (sample splitting) 或样本截断 (sample truncation) 并不是有效方法，因为这些方法未能充分利用包含在 $\{X_i\}_{i=1}^{2n}$ 中的所有样本信息。

例 8.12：

令 (X_1, X_2) 为来自均值为 μ，方差为 σ^2 的某个总体分布的 IID 随机样本。均值 μ 的两个估计量分别为

$$\hat{\mu}_1 = \bar{X}_n = \frac{1}{2}(X_1 + X_2)$$

$$\hat{\mu}_2 = \frac{1}{3}(X_1 + 2X_2)$$

哪个估计量更好?

例 8.12 (Cont.):

解:

- 令 $\theta = (\mu, \sigma^2)$。可验证两个估计量均为 μ 的无偏估计。且

$$\begin{aligned}
\mathrm{var}_\theta(\hat{\mu}_1) &= \frac{1}{2}\sigma^2 \\
\mathrm{var}_\theta(\hat{\mu}_2) &= \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^4 \\
&= \frac{5}{9}\sigma^2 \\
&> \frac{1}{2}\sigma^2
\end{aligned}$$

- 因此，在 MSE 准则下 $\hat{\mu}_1$ 比 $\hat{\mu}_2$ 更有效。

- 直觉上，由于两个随机变量 X_1 和 X_2 服从同分布，没有理由区别对待 X_1 和 X_2 (即对同分布观测值赋不同的权重)。
- 对每个观测值赋予等权重将获得 μ 的最有效估计。

目　录

◆ **问题**

在参数 θ 的一类估计量中哪个是最优估计量?

- 可定义 MSE 最小的估计量为最优。但在实际应用中这种估计量很难求获，因为估计量类型庞大。
- 简便起见，以下只考察一类线性无偏估计量并在其中求最优估计量。
- 首先，定义变换参数 $\tau = \tau(\theta)$ 的广义无偏估计量概念。一个例子是 $\tau = \mu^2$，本章第六节例 8.10 讨论过这个例子。

定义 8.6

[广义无偏估计量 (Generalized Unbiased Estimator)]：

$\hat{\tau} = \hat{\tau}_n(\boldsymbol{X}^n)$ 是参数 $\tau(\theta)$ 的无偏估计量，若

$$E_\theta(\hat{\tau}) = \tau(\theta), \text{对所有 } \theta \in \Theta$$

其中 $E_\theta(\cdot)$ 是对 $\boldsymbol{X}^n$ 的联合概率分布 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 求期望。

当 $\tau(\theta) = \theta$ 时，此定义即回到定义 8.4 (参数 θ 的无偏估计量) 的情形。

定义 8.7

[一致最优无偏估计量 (Uniform Best Unbiased Estimator)]:

令 Γ 为参数 $\tau(\theta)$ 的一类无偏估计量的集合，其中 $\theta \in \Theta$ 且 Θ 是 θ 的参数空间。

- 若估计量 $\hat{\tau}^* \in \Gamma$ 满足:

(1) 对所有 $\theta \in \Theta$，$E_\theta(\hat{\tau}^*) = \tau(\theta)$；

(2) 对 Γ 中 $\tau(\theta)$ 的任意估计量 $\hat{\tau}$ 且对所有 $\theta \in \Theta$，有 $\mathrm{var}_\theta(\hat{\tau}^*) \leq \mathrm{var}_\theta(\hat{\tau})$。

- 则称估计量 $\hat{\tau}^* \in \Gamma$ 为 $\tau(\theta)$ 在参数空间 Θ 上属于 Γ 类所有估计量中的一致最优无偏估计量。

例 8.14：[同方差条件下最优线性无偏估计量 (Best Linear Unbiased Estimator, BLUE)]

- 令 $\boldsymbol{X}^n$ 为 IID (μ, σ^2) 随机样本。定义 μ 的一类线性无偏估计量如下

$$\Gamma = \left\{ \hat{\mu}: \mathbb{R}^n \to \mathbb{R} \mid \hat{\mu} = \sum_{i=1}^{n} c_i X_i, (c_1, \cdots, c_n)' \in \mathbb{R}^n \right\}$$

- (1) 证明：对所有 $\mu \in \mathbb{R}$ 和所有 $n \geq 1$，当且仅当 $\sum_{i=1}^{n} c_i = 1$ 时，$\hat{\mu}$ 为 μ 的无偏估计量；
- (2) 求 μ 在 Γ 类估计量中的一致最有效无偏估计量。

例 8.14 (Cont.)：

解：

- 注意到 $\hat{\tau}=\hat{\mu}$，此处 $\tau(\theta)=\mu$，$\theta=(\mu,\sigma^2)$。
- **(1)** 给定 $\hat{\mu}=\sum_{i=1}^{n} c_i X_i$，取期望得 $E_\theta(\hat{\mu})=\mu\sum_{i=1}^{n} c_i$。因此，若 $\hat{\mu}$ 为 μ 的无偏估计量，即如果

$$E_\theta(\hat{\mu})=\mu, \text{对所有 } \mu \text{ 值}$$

则必有

$$\sum_{i=1}^{n} c_i = 1$$

- 另一方面，若 $\sum_{i=1}^{n} c_i = 1$，则由 $E_\theta(\hat{\mu})=\mu\sum_{i=1}^{n} c_i$，得

$$E_\theta(\hat{\mu})=\mu\cdot 1=\mu, \text{对所有 } \mu$$

- 即 $\hat{\mu}$ 是 μ 的无偏估计。因此，当且仅当 $\sum_{i=1}^{n} c_i = 1$时，$\hat{\mu}$ 为 μ 的无偏估计量。

例 8.14 (Cont.)：

- **解 (Cont.)：**
- **(2)** 为求 μ 的一致最有效无偏估计量，只需在 $\sum_{i=1}^{n} c_i = 1$ 约束条件下，在 Γ 类估计量中找方差最小的那一个估计量。
- 在 IID 的假设下，$\hat{\mu} \in \Gamma$ 的方差为

$$\begin{aligned}\mathrm{var}_\theta(\hat{\mu}) &= \mathrm{var}_\theta\left(\sum_{i=1}^{n} c_i X_i\right) \\ &= \sum_{i=1}^{n} c_i^2 \mathrm{var}_\theta(X_i) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2\end{aligned}$$

例 8.14 (Cont.)：

解 (Cont.)：

- 因 $\hat{\mu}$ 无偏，故 $\sum_{i=1}^{n} c_i = 1$。因此，可求解方差最小化问题

$$\min_{\{c_i\}_{i=1}^{n}} \sigma^2 \sum_{i=1}^{n} c_i^2$$

- 满足约束条件

$$\sum_{i=1}^{n} c_i = 1$$

- 定义拉格朗日函数

$$L(c, \lambda) = \sigma^2 \sum_{i=1}^{n} c_i^2 + \lambda(1 - \sum_{i=1}^{n} c_i)$$

- 其中 $c = (c_1, \cdots, c_n)'$，　λ 为拉格朗日乘子。

例 8.14 (Cont.)：

解 (Cont.)：

- 最小化拉格朗日函数的 $n+1$ 个一阶条件如下:

$$\frac{\partial L(c,\lambda)}{\partial c_i} = 2\sigma^2 c_i - \lambda = 0, i = 1, \cdots, n$$

$$\frac{\partial L(c,\lambda)}{\partial \lambda} = 1 - \sum_{i=1}^{n} c_i = 0$$

- 解这 $n+1$ 个联立方程组，得

$$c_i^* = \frac{1}{n}, i = 1, \cdots, n$$

- 因此，一致最有效无偏估计量为

$$\hat{\mu}^* = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X}_n$$

例 8.14 (Cont.)：

解 (Cont.)：

- 另外，需验证二阶条件以确保 $\hat{\mu}^*$ 是全局最小值。结果的确如此，因为可以证明，对所有 μ，$L(c,\lambda)$ 的黑塞矩阵总为正定(请验证)。

例 8.14 (Cont.)：

- 直觉上，因 n 个随机变量 $\{X_i\}_{i=1}^n$ 具有相同分布，没有理由对其区别对待。
- 那么，最优权重自然是对所有观测值赋予等权重。这与经典线性回归模型的**高斯-马尔可夫定理** (Gauss-Markov theorem) 的思想完全相同。
- 该定理指出，对经典线性回归模型

$$Y_i = X_i'\theta + \varepsilon_i, \qquad i = 1, \cdots, n$$

- 其未知参数 θ 的**普通最小二乘** (ordinary least squares, OLS) **估计量**，在 $\{\varepsilon_i\}_{i=1}^n$ 为 IID $\left(0, \sigma_\varepsilon^2\right)$ 的假设下是最优线性无偏估计量 (BLUE)。

例 8.14 (Cont.)：

- 在线性回归分析中，OLS 估计量 $\hat{\theta}$ 定义为最小化残差平方和的解，即

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - X_i'\theta)^2$$

其对于每个残差 $Y_i - X_i'\theta$ 的权重相同。详细讨论参见第十章。

- 回顾第六章定理 6.1 关于样本均值 $\bar{X}_n$ 是最小化残差平方和的解

$$\bar{X}_n = \arg\min_{a} \sum_{i=1}^{n} (X_i - a)^2$$

- 这其实是线性回归模型的一个特例 (即只包含截距项)。

例 8.15： [异方差条件下最优线性无偏估计量 (Optimal Estimation for μ under Heteroskedasticity])]:

- 假设 $\boldsymbol{X}^n = (X_1, \cdots, X_n)$ 为独立但非同分布随机样本，其中 $E(X_i) = \mu$ 和 $\mathrm{var}(X_i) = \sigma_i^2 < \infty$， $i = 1, \cdots, n$。
- 在如下一类关于 μ 估计量中，求 μ 的一致最优线性无偏估计量

$$\Gamma = \left\{ \hat{\mu} : \mathbb{R}^n \to \mathbb{R} \mid \hat{\mu} = \sum_{i=1}^{n} c_i X_i, (c_1, \cdots, c_n)' \in \mathbb{R}^n \right\}$$

其中 $\sum_{i=1}^{n} c_i = 1$。

例 8.15 (Cont.)：

解：

- 当且仅当 $\sum_{i=1}^{n} c_i = 1$ 时，$\hat{\mu}$ 为 μ 的无偏估计量。
- 应用拉格朗日乘子法，可在 Γ 类估计量中求得最优估计量为

$$\hat{\mu}^* = \sum_{i=1}^{n} c_i^* X_i$$

$$= \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} X_i$$

例 8.15 (Cont.)：

解 (Cont.)：

- 其中，最优权重为

$$c_i^* = \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \propto \frac{1}{\sigma_i^2}, \qquad i = 1, \cdots, n$$

- 该结果表明，在独立但非同分布的随机样本中(所有随机变量的均值 μ 相同，但方差不同) 求 μ 的最有效估计量，需要减小噪声大的观测值的影响（即方差较大的观测值须赋予较小的权重，反之则赋予较大的权重)。
- 最优权重与 σ_i^{-2} 成正比，σ_i^{-2} 是随机变量 X_i 方差的逆。

例 8.15 (Cont.)：

- 这与计量经济学中常用的**广义最小二乘** (generalized least squares, GLS) **估计量**的思想类似。
- 考察线性回归模型

$$Y_i = X_i'\theta + \varepsilon_i, \qquad i = 1, \cdots, n$$

- 其中随机扰动项 $\{\varepsilon_i\}_{i=1}^n$ 为一独立但非同分布序列，其均值 $E(\varepsilon_i) = 0$，方差 $\mathrm{var}(\varepsilon_i^2) = \sigma_i^2$。
- 此处存在无条件异方差 (unconditional heteroskedasticity)，因为 σ_i^2 对不同的 i 可能取不同的值。

例 8.15 (Cont.)：

- 现在考察变换回归模型

$$\frac{Y_i}{\sigma_i} = \left(\frac{X_i}{\sigma_i}\right)' \theta + \frac{\varepsilon_i}{\sigma_i}$$

或等价地

$$Y_i^* = X_i^{*\prime}\theta + \varepsilon_i^*$$

- 其中，新的随机扰动项 $\{\varepsilon_i^*\}_{i=1}^n$ 是具有**零均值**和**单位方差**的 IID 序列。
- 变换线性回归模型的 OLS 估计量 $\hat{\theta}^*$ 称为 **GLS 估计量**。

$$\hat{\theta}^* = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i^* - X_i^{*\prime}\theta)^2$$

例 8.15 (Cont.)：

- GLS 是对原始线性回归模型进行**异方差修正**后所获得的 OLS 估计量。
- 估计量 $\hat{\theta}^*$ 对每个残差 $Y_i - X_i'\theta$ 除以相应的扰动项 ε_i 的标准差 σ_i 以降低噪声大的扰动项影响。
- 可以证明，GLS 估计量为 BLUE。
- 详细讨论参见第十章第九节。

- 拉格朗日乘子法在预算约束下的效用最大化、技术约束下的成本最小化或利润最大化等经济问题中，也都有广泛应用。

- Chow (1975) 提出了一种通过拉格朗日乘子法解决宏观经济学动态最优化问题的方法，建立了一套可应用于随机经济系统的最优控制理论。

◆ **问题**

是否无偏估计量总优于有偏估计量?

- 回答是否定的。以下举一例说明。

例 8.16：

- 令 $\boldsymbol{X}^n$ 为来自正态分布 $N(\mu, \sigma^2)$ 的 IID 随机样本。
- 样本方差 $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ 和 MLE 估计量 $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ 是 σ^2 的两个估计量。
- 依据 MSE，哪个估计量为更有效的估计量?

例 8.16 (Cont.):

解:

- 首先，根据第六章定理 6.6，即 $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$，有 $E_\theta\left(S_n^2\right) = \sigma^2$ 和 $\mathrm{var}_\theta\left(S_n^2\right) = \frac{2\sigma^4}{n-1}$。
- 则有

$$\begin{aligned}\mathrm{MSE}_\theta\left(S_n^2\right) &= E_\theta\left(S_n^2 - \sigma^2\right)^2 \\ &= \mathrm{var}_\theta\left(S_n^2\right) + \left[\mathrm{Bias}_\theta\left(S_n^2\right)\right]^2 \\ &= \frac{2\sigma^4}{n-1}\end{aligned}$$

- 其次，观察到

$$\hat{\sigma}^2 = \frac{n-1}{n} S_n^2$$

例 8.16 (Cont.):

解 (Cont.):

- 有

$$\begin{aligned}\mathrm{Bias}_\theta\left(\hat{\sigma}^2\right) &= E_\theta\left(\hat{\sigma}^2\right) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= -\frac{\sigma^2}{n}\end{aligned}$$

$$\begin{aligned}\mathrm{var}_\theta\left(\hat{\sigma}^2\right) &= \left(\frac{n-1}{n}\right)^2 \mathrm{var}_\theta\left(S_n^2\right) \\ &= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1}\end{aligned}$$

例 8.16 (Cont.):

解 (Cont.):

- 从而对所有 $n > 1$,

$$\begin{aligned}\mathrm{MSE}_\theta\left(\hat{\sigma}^2\right) &= \left(1-\frac{1}{n}\right)^2 \frac{2\sigma^4}{n-1} + \frac{\sigma^4}{n^2} \\ &= \left[\left(1-\frac{1}{n}\right)^2 + \frac{n-1}{2n^2}\right]\frac{2\sigma^4}{n-1} \\ &= \frac{n-1}{n}\frac{2n-1}{2n}\frac{2\sigma^4}{n-1} \\ &< \frac{2\sigma^4}{n-1} = \mathrm{MSE}_\theta\left(S_n^2\right)\end{aligned}$$

- 因此，有偏估计量 $\hat{\sigma}^2$ 优于无偏估计量 S_n^2。

- 例 8.16 表明**无偏估计量未必是更有效的估计量**。正如此例所示，有时在方差和偏差之间存在一种权衡取舍，即用偏差的一个小增量可换取方差较大的减少，从而使 MSE 变小。
- 例 8.16 的结论并非意味着要舍弃 S_n^2 作为 σ^2 的估计量，该结论只是基于 MSE 所得。无法确定 MSE 是否为测度方差估计量优度的最佳方法。此外，当 n 较大时，就 MSE 而言，S_n^2 和 $\hat{\sigma}^2$ 几乎不存在差异。
- 样本方差 $\hat{S}_n^2$ 只是许多无偏估计量中的一个，就 MSE 而言，可能不如有偏估计量。

- 再例如，经典线性回归模型中的 OLS 估计量 $\hat{\beta}$ 是未知参数的无偏估计量。当自变量的数目很大，它们之间可能存在近似多重共线性时，OLS 估计量 $\hat{\beta}$ 将存在巨大方差。
- 一个修正方法就是 Hoerl & Kennard (1970) 提出的所谓**岭回归 (ridge regression)** 估计方法。
- 它通过对平方系数之和施加一个等比例的惩罚项来限制未知参数的大小。这将会带来偏差，但是 OLS 估计量方差会大大减小，从而得到一个更小的 MSE。

- 另外，若许多未知系数为 0 或非常微小，可以通过**套索算法 (least absolute shrinkage and selection operator, LASSO)** 对系数绝对值总和施加一个等比例的惩罚项。
- LASSO 估计量可有效挑选出那些非零系数，从而大大减小 MSE 方差，但其偏差会变大。
- 关于 LASSO 估计的更多讨论，可参见 Tibshirani (1996)。

目　录

- 一般而言，从一类无偏估计量中求最有效估计量是一件困难的事情。
- 当总体分布模型 $f(x,\theta)$ 的函数形式已知时，有另一种评估参数估计量有效性的方法。
- 简单起见，本节假设参数 θ 为标量。

定理 8.11

[克拉默-拉奥下界 (Cramer-Rao Lower Bound)；克拉默-拉奥不等式 (Cramer-Rao Inequality)]：

- 令 $\boldsymbol{X}^n$ 为一个随机样本，其联合 PMF/PDF 为 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)$，并令 $\hat{\tau}=\hat{\tau}_n(\boldsymbol{X}^n)$ 为参数 $\tau(\theta)$ 的任意估计量，且 $E_\theta(\hat{\tau})$ 是 θ 的可导函数，期望 $E_\theta(\cdot)$ 是定义在随机样本 $\boldsymbol{X}^n$ 的联合概率 PMF/PDF $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)$上。对满足 $E_\theta|h(\boldsymbol{X}^n)|<\infty$ 的任意函数 $h:\mathbb{R}^n\to\mathbb{R}$，如果以下条件成立，

$$\frac{d}{d\theta}\int_{\mathbb{R}^n} h(\boldsymbol{x}^n)f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n=\int_{\mathbb{R}^n} h(\boldsymbol{x}^n)\frac{\partial f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)}{\partial\theta}d\boldsymbol{x}^n$$

定理 8.11 (Cont.)

- 则对所有 $n > 0$ 和所有 $\theta \in \Theta$，有

$$\mathrm{var}_\theta(\hat{\tau}) \geq B_n(\theta) \equiv \frac{\left[\frac{dE_\theta(\hat{\tau})}{d\theta}\right]^2}{E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial\theta}\right]^2}$$

- 其中 $B_n(\theta)$ 称为克拉默-拉奥下界 (Cramer-Rao lower bound)。
- 特别地，当 $\hat{\tau}$ 是参数 $\tau(\theta)$ 的无偏估计量时，有

$$B_n(\theta) = \frac{[\tau'(\theta)]^2}{E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial\theta}\right]^2}$$

证明:

- 这里只考察连续分布的情形，离散分布的证明类似。假设有
- (1)

$$E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]^2 = \mathrm{var}_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]$$

- (2)

$$\frac{dE_\theta(\hat{\tau})}{d\theta} = \mathrm{cov}_\theta\left[\hat{\tau},\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]$$

- 根据柯西-施瓦茨 (Cauchy-Schwarz) 不等式，有

$$\left\{\mathrm{cov}_\theta\left[\hat{\tau},\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]\right\}^2 \le \mathrm{var}_\theta(\hat{\tau})\mathrm{var}_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]$$

证明 (Cont.):

- 则从上述结果 (1) 和 (2)，可得

$$\begin{aligned}\mathrm{var}_\theta(\hat{\tau}) &\geq \frac{\left\{\mathrm{cov}_\theta\left[\hat{\tau}, \frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial \theta}\right]\right\}^2}{\mathrm{var}_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial \theta}\right]} \\ &= \frac{\left[\frac{dE_\theta(\hat{\tau})}{d\theta}\right]^2}{E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial \theta}\right]^2} \\ &= B_n(\theta)\end{aligned}$$

- 因此，只需证明结果 (1) 和 (2)。

证明 (Cont.):

- **首先证明结果** (1)，即证明在联合分布 $f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)$ 下，随机样本 $\boldsymbol{X}^n$ 的记分函数 $\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}$ 的均值为零。因为

$$
\begin{aligned}
E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right] &= \int_{\mathbb{R}^n}\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)}{\partial\theta}\right] f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n \\
&= \int_{\mathbb{R}^n}\frac{\partial f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)}{\partial\theta}d\boldsymbol{x}^n \\
&= \frac{d}{d\theta}\int_{\mathbb{R}^n} f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n \quad \text{[交换积分和求导顺序]} \\
&= \frac{d(1)}{d\theta} \\
&= 0
\end{aligned}
$$

证明 (Cont.):

- 由公式 $\mathrm{var}(Y) = E\left(Y^2\right) - \mu_Y^2$，可求得记分函数的方差

$$\mathrm{var}_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right] = E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]^2 - \left\{E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]\right\}^2$$

$$= E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]^2$$

- **结果 (1) 证毕。**

证明 (Cont.):

- **下面证明结果 (2)。** 给定 $E_\theta[\frac{\partial \ln f_{X^n}(X^n,\theta)}{\partial \theta}] = 0$ 和 $\hat{\tau} = \hat{\tau}_n(X^n)$，

  由公式 $\mathrm{cov}(Y,Z) = E(YZ) - \mu_Y \mu_Z$，有

$$\begin{aligned}
\mathrm{cov}_\theta\left[\hat{\tau}, \frac{\partial \ln f_{X^n}(X^n,\theta)}{\partial \theta}\right] &= E_\theta\left[\hat{\tau}\frac{\partial \ln f_{X^n}(X^n,\theta)}{\partial \theta}\right] - E_\theta(\hat{\tau})E_\theta\left[\frac{\partial \ln f_{X^n}(X^n,\theta)}{\partial \theta}\right] \\
&= E_\theta\left[\hat{\tau}\frac{\partial \ln f_{X^n}(X^n,\theta)}{\partial \theta}\right] \\
&= \int_{\mathbb{R}^n} \hat{\tau}_n(\boldsymbol{x}^n)\left[\frac{\partial \ln f_{X^n}(\boldsymbol{x}^n,\theta)}{\partial \theta}\right] f_{X^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n \\
&= \int_{\mathbb{R}^n} \hat{\tau}_n(\boldsymbol{x}^n)\frac{\partial f_{X^n}(\boldsymbol{x}^n,\theta)}{\partial \theta} d\boldsymbol{x}^n \\
&= \frac{d}{d\theta}\int_{\mathbb{R}^n} \hat{\tau}_n(\boldsymbol{x}^n) f_{X^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n \quad \text{[交换积分和求导顺序]} \\
&= \frac{dE_\theta(\hat{\tau})}{d\theta}
\end{aligned}$$

- **结果 (2) 证毕。**

证明 (Cont.)：

- 克拉默-拉奥下界同样可适用于离散分布情形。
- 唯一的变化是以求和替代积分，$f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 代表随机样本 $\boldsymbol{X}^n$ 的联合 PMF 而非联合 PDF。

证毕。

- 需要强调，克拉默-拉奥定理的一个关键假设具有一定限制性，即**可交换积分和微分运算顺序**。该条件称为正则条件(regularity condition)，即在一般情况下成立。
- 这个条件可写为

$$\frac{d}{d\theta}\int_{\mathbb{R}^n} h(\boldsymbol{x}^n) f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n = \int_{\mathbb{R}^n} h(\boldsymbol{x}^n)\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)}{\partial\theta} f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)d\boldsymbol{x}^n$$

或等价地

$$\frac{dE_\theta[h(\boldsymbol{X}^n)]}{d\theta} = E_\theta\left[h(\boldsymbol{X}^n)\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]$$

推论 8.1： [IID 随机样本下的克拉默-拉奥下界 (Cramer-Rao Lower Bound Under an IID Random Sample])]

- 令 $\boldsymbol{X}^n$ 为来自总体 PMF/PDF $f(x,\theta)$ 的 IID 随机样本，并令 $\hat{\tau} = \hat{\tau}_n(\boldsymbol{X}^n)$ 为 $\tau(\theta)$ 的任意估计量，且 $E_\theta[\hat{\tau}_n(\boldsymbol{X}^n)]$ 是 $\theta \in \Theta$ 的可导函数。
- 假设对满足 $E_\theta|h(X_i)| < \infty$ 的任意函数 $h(x)$，有

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

- 则对所有 n，

$$\mathrm{var}_\theta(\hat{\tau}) \geq B_n(\theta) \equiv \frac{\left[\frac{dE_\theta(\hat{\tau})}{d\theta}\right]^2}{nI(\theta)}$$

推论 8.1 (Cont.)：

- 其中

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

  是总体分布 PMF/PDF $f(x,\theta)$ 的**费雪信息**。

- 当 $\hat{\tau}$ 为 $\tau(\theta)$ 的无偏估计量时，则有

$$\mathrm{var}_\theta(\hat{\tau}) \geq B_n(\theta) \equiv \frac{[\tau'(\theta)]^2}{nI(\theta)}$$

证明：

- 给定定理 8.11 的克拉默-拉奥下界，只需证明在随机样本 $\boldsymbol{X}^n$ 为独立同分布的假设下，克拉默-拉奥下界的分母

$$E_\theta\left[\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]^2 = nE_\theta\left[\frac{\partial \ln f(X_i,\theta)}{\partial\theta}\right]^2 = nI(\theta)$$

- 因为 $\boldsymbol{X}^n$ 是来自总体分布 $f(x,\theta)$ 的 IID 样本，有 $f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta)=\prod_{i=1}^n f(X_i,\theta)$，故可得

$$\ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n,\theta) = \sum_{i=1}^n \ln f(X_i,\theta)$$

证明 (Cont.):

- 随机样本 $\boldsymbol{X}^n$ 的记分函数

$$\frac{\partial \ln f_{\boldsymbol{X}^n}(\boldsymbol{X}^n, \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(X_i, \theta)}{\partial \theta}$$

- 由引理 8.3，基于总体分布 $f(x, \theta)$ 的记分函数 $\frac{\partial \ln f(X_i, \theta)}{\partial \theta}$ 的期望为零，即

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

证明 (Cont.):

- 由 IID 假设得

$$E_\theta\left[\frac{\partial \ln f_{X^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right]^2$$

$$= \mathrm{var}_\theta\left[\frac{\partial \ln f_{X^n}(\boldsymbol{X}^n,\theta)}{\partial\theta}\right] = \mathrm{var}_\theta\left[\sum_{i=1}^n \frac{\partial \ln f(X_i,\theta)}{\partial\theta}\right]$$

$$= \sum_{i=1}^n \mathrm{var}_\theta\left[\frac{\partial \ln f(X_i,\theta)}{\partial\theta}\right] = n\mathrm{var}_\theta\left[\frac{\partial \ln f(X_1,\theta)}{\partial\theta}\right]$$

$$= nE_\theta\left[\frac{\partial \ln f(X_1,\theta)}{\partial\theta}\right]^2 = nI(\theta)$$

- 其中，第三和第四个等式由 IID 假设推得，而第五个等式由 $E_\theta\left[\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right] = 0$ 推得。

证毕。

克拉默-拉奥下界

- 在推论 8.1 中，$f(x,\theta)$ 是每个随机变量 X_i 的 PMF/PDF，而定理 8.11 中的 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n,\theta)$ 是随机样本 $\boldsymbol{X}^n$ 的联合 PMF/PDF。可以看出，在 IID 假设下，克拉默-拉奥下界 $B_n(\theta)$ 以速度 n 趋近于零。
- 除尺度因子 n^{-1} 外，克拉默-拉奥下界与费雪信息 $I(\theta)$ 的逆成正比。
  - ✓ 费雪信息测度总体分布 PMF/PDF $f(x,\theta)$ 或等价地每个随机变量 X_i 包含了多少关于 θ 的信息。
  - ✓ $I(\theta)$ 越大，随机变量 X_i 就包含越多关于 θ 的信息，从而克拉默-拉奥下界 $B_n(\theta)$ 就越小。

克拉默-拉奥下界 (Cont.)

- 有些情形下，$I(\theta)$ 的计算比较繁冗。根据引理 8.4 的信息等式$I(\theta)+H(\theta)=0$，可将参数 $\tau(\theta)$ 的无偏估计量 $\hat{\tau}$ 的克拉默-拉奥下界表达为

$$B_n(\theta)=\frac{[\tau'(\theta)]^2}{-nH(\theta)}$$

- 其中

$$H(\theta)\equiv E_\theta\left[\frac{\partial^2 \ln f(X_i,\theta)}{\partial\theta^2}\right]$$
$$=\int_{-\infty}^{\infty}\frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2}f(x,\theta)dx$$

称为总体分布 $f(x,\theta)$ 或随机变量 X_i 的黑塞函数或黑塞矩阵(当 θ 为向量时)。

克拉默-拉奥下界 (Cont.)

- 因此，克拉默-拉奥下界 $B_n(\theta)$ 依赖于对数似然函数 $\ln f(X,\theta)$ 的曲率均值的倒数。对数似然函数的曲率的绝对值越大，克拉默-拉奥下界 $B_n(\theta)$ 越小。

克拉默-拉奥下界 (Cont.)

- 定理 8.5 已证明，当 $n \to \infty$ 时，

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left[0, -H(\theta)^{-1}\right]$$

其中 $\theta \in \Theta$ 为真实参数值。

- 这表明当 n 充分大时，$\hat{\theta} - \theta$ 近似服从正态分布 $N\{0, [-nH(\theta)]^{-1}\}$。
- 此时，MLE 估计量 $\hat{\theta}$ 的方差近似达到克拉默-拉奥下界，这对于任意总体分布 $f(x, \theta)$ 均成立。
- 从定理 8.5 中 $\sqrt{n}(\hat{\theta} - \theta_0)$ 的渐近正态性，可得 MLE 估计量 $\hat{\theta}$ 的渐近偏差等于零，从而 MLE $\hat{\theta}$ 是渐近最有效的估计量。

例 8.17：

- 令 $\boldsymbol{X}^n$ 为 IID Poisson(λ) 随机样本。用克拉默-拉奥下界法证明样本均值 $\bar{X}_n$ 为参数 λ 的最优无偏估计量。

解：

- 在本例中，$\tau(\theta)=\theta=\lambda$，且 $\hat{\tau}=\bar{X}_n$，$E_\lambda(\hat{\tau})=\lambda$。因此，$\hat{\tau}$ 是 λ 的无偏估计量。对泊松分布 Poisson(λ)，$\sigma^2=\lambda$，故有

$$\mathrm{var}_\lambda(\bar{X}_n)=\frac{\sigma^2}{n}=\frac{\lambda}{n}$$

- 因此只需证明克拉默-拉奥下界 $B_n(\lambda)=\frac{\lambda}{n}$。

例 8.17 (Cont.):

解 (Cont.):

- 因为对数似然函数

$$\ln f\left(X_i, \lambda\right) = \ln\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right)$$
$$= -\lambda + X_i \ln\lambda - \ln X_i\,!$$

- 有

$$\frac{\partial^2 \ln f\left(X_i, \lambda\right)}{\partial\lambda^2} = \frac{\partial^2(-\lambda + X_i \ln\lambda - \ln X_i\,!)}{\partial\lambda^2}$$
$$= -\frac{X_i}{\lambda^2}$$

例 8.17 (Cont.):

解 (Cont.):

- 和

$$
\begin{aligned}
H_\lambda(\lambda) &= E_\lambda\left[\frac{\partial^2 \ln f\,(X_i,\lambda)}{\partial\lambda^2}\right] \\
&= E_\lambda\left(-\frac{X_i}{\lambda^2}\right) \\
&= -\frac{E_\lambda(X_i)}{\lambda^2} \\
&= -\frac{1}{\lambda}
\end{aligned}
$$

例 8.17 (Cont.):

解 (Cont.):

- 且因 $\hat{\tau} = \bar{X}_n$，有$E_\lambda(\bar{X}_n) = \lambda$，故

$$\frac{dE_\lambda(\hat{\tau})}{d\lambda} = \frac{d\lambda}{d\lambda} = 1$$

- 因此，

$$\begin{aligned} B_n(\lambda) &= \frac{\left[\frac{d}{d\lambda}E_\lambda(\hat{\tau})\right]^2}{-nH_\lambda(\lambda)} \\ &= \frac{1^2}{-n\left(-\frac{1}{\lambda}\right)} \\ &= \frac{\lambda}{n} \end{aligned}$$

- 因为 $\mathrm{var}_\lambda(\bar{X}_n) = B_n(\lambda)$，即样本均值 $\bar{X}_n$ 达到克拉默-拉奥下界，故 $\bar{X}_n$ 为 λ 的最优无偏估计量。

- 当 $\mathrm{var}_\theta(\hat{\tau})$ 未达到克拉默-拉奥下界时（$\mathrm{var}_\theta(\hat{\tau}) > B_n(\lambda)$），克拉默-拉奥下界可能无法给出确定性结论，即无法判定估计量 $\hat{\tau}$ 是否为 $\tau(\theta)$ 的最有效估计量。
- 以下例子说明克拉默-拉奥下界法的这一缺陷。

例 8.18：

- 令 $\boldsymbol{X}^n$ 为 IID $N(\mu,\sigma^2)$ 随机样本，其中 μ 和 σ^2 未知。
- 证明 S_n^2 未达到克拉默-拉奥下界。

解：

- 本例中 $\theta=(\mu,\sigma^2)$，$\tau(\theta)=\sigma^2$ 以及 $\hat{\tau}=S_n^2$。$N(\mu,\sigma^2)$ 分布的对数似然函数为

$$\ln f(X_i,\theta)=-\ln\sqrt{2\pi}-\frac{1}{2}\ln\sigma^2-\frac{(X_i-\mu)^2}{2\sigma^2}$$

- 因此，有

$$\frac{\partial^2 \ln f(X_i,\theta)}{\partial(\sigma^2)^2}=\frac{1}{2\sigma^4}-\frac{(X_i-\mu)^2}{\sigma^6}$$

例 8.18 (Cont.):

解 (Cont.):

- 而黑塞矩阵为

$$E_\theta\left[\frac{\partial^2 \ln f(X_i,\theta)}{\partial(\sigma^2)^2}\right]=\frac{1}{2\sigma^4}-\frac{E_\theta(X_i-\mu)^2}{\sigma^6}$$
$$=\frac{1}{2\sigma^4}-\frac{1}{\sigma^4}$$
$$=-\frac{1}{2\sigma^4}$$

- 本例中，容易理解为什么信息矩阵 $I(\theta)$ 或黑塞矩阵 $H(\theta)$ 是对随机变量 X_i 所包含 σ^2 的信息的测度。方差 σ^2 越小，随机变量 X_i 的噪声越小，从而可更有效地估计 σ^2。

例 8.18 (Cont.):

解 (Cont.):

- 对 σ^2 的任意无偏估计量 $\hat{\tau} = \hat{\tau}_n(\boldsymbol{X}^n)$，有 $E_\theta[\hat{\tau}_n(\boldsymbol{X}^n)] = \sigma^2$，因此

$$\frac{dE_\theta[\hat{\tau}_n(\boldsymbol{X}^n)]}{d\sigma^2} = 1$$

- 由克拉默-拉奥下界，有

$$\begin{aligned} B_n(\theta) &= \frac{1}{n}\left(1^2 \Big/ \frac{1}{2\sigma^4}\right) \\ &= \frac{2\sigma^4}{n} \\ &< \frac{2\sigma^4}{n-1} = \mathrm{var}_\theta\left(S_n^2\right) \end{aligned}$$

例 8.18 (Cont.):

解 (Cont.):

- 因此，S_n^2 未能达到克拉默-拉奥下界 $2\sigma^4/n$。但是，只能说 S_n^2 未达到克拉默-拉奥下界，而不能判断 S_n^2 并非最优无偏估计量，因为存在两种可能性：
  - ✓ (a) 存在另一个无偏估计量达到 $B_n(\theta)$；
  - ✓ (b) σ^2 的任何无偏估计量都无法达到 $B_n(\theta)$。

克拉默-拉奥下界法的缺陷

- 克拉默-拉奥下界法在求最优无偏估计量时可能有一个缺陷：
  - ✓ 若无偏估计量未能达到克拉默-拉奥下界，则无法判断其是否为最有效估计量。也就是说，克拉默-拉奥下界可能严格小于任意无偏估计量的方差。

定理 8.12

- 假设 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 是随机样本 $\boldsymbol{X}^n$ 的联合 PMF/PDF 且 $\hat{\tau} = \hat{\tau}_n(\boldsymbol{X}^n)$ 是参数 $\tau(\theta)$ 的一个无偏估计量，其中 $f_{\boldsymbol{X}^n}(\boldsymbol{x}^n, \theta)$ 和 $\hat{\tau}$ 满足克拉默-拉奥下界定理 (定理 8.11) 的条件。
- 则估计量 $\hat{\tau}$ 达到克拉默-拉奥下界，当且仅当存在某一函数 $a: \Theta \to R$，有

$$\hat{\tau} - \tau(\theta) = a(\theta) \frac{\partial \ln \hat{L}(\theta \mid \boldsymbol{X}^n)}{\partial \theta}$$

证明：

- 由柯西-施瓦茨 (Cauchy-Schwarz) 不等式，有

$$\left\{\mathrm{cov}_\theta\left[\hat{\tau}, \frac{\partial \ln \hat{L}\left(\theta \mid \boldsymbol{X}^n\right)}{\partial \theta}\right]\right\}^2 \leq \mathrm{var}_\theta(\hat{\tau})\mathrm{var}_\theta\left[\frac{\partial \ln \hat{L}\left(\theta \mid \boldsymbol{X}^n\right)}{\partial \theta}\right]$$

- 当且仅当中心化的参数估计量 $\hat{\tau}-\tau(\theta)$ 与随机样本 $\boldsymbol{X}^n$ 的记分函数 $\frac{\partial \ln \hat{L}(\theta|\boldsymbol{X}^n)}{\partial \theta}$ 成正比时，上述不等式才取等号。这是因为，当 $\hat{\tau}-\tau(\theta)$ 为记分函数 $\frac{\partial}{\partial \theta}\ln \hat{L}(\theta \mid \boldsymbol{X}^n)$ 的线性函数时，二者相关系数的绝对值等于 1。在此情况下，有 $\mathrm{var}_\theta(\hat{\tau}) = B_n(\theta)$。

证毕。

例 8.19 [例 8.18 的延续]:

- 来自 IID $N(\mu, \sigma^2)$ 总体分布的随机样本 $\boldsymbol{X}^n$ 的似然函数为

$$\hat{L}(\theta \mid \boldsymbol{X}^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

其中 $\theta = (\mu, \sigma^2)$。因此，$\boldsymbol{X}^n$ 的记分函数

$$\frac{\partial \ln \hat{L}(\theta \mid \boldsymbol{X}^n)}{\partial \sigma^2} = \frac{n}{2\sigma^4}\left[n^{-1}\sum_{i=1}^{n}(X_i-\mu)^2 - \sigma^2\right]$$

- 当 $a(\theta) = \frac{2\sigma^4}{n}$ 时，可以证明，
  - 若 μ 已知，则 σ^2 的最优无偏估计量为 $\tilde{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i-\mu)^2$。
  - 若 μ 未知，则无法达到克拉默-拉奥下界。

目　录

- 参数估计是统计推断最重要的目的之一。
- 本章首先介绍两种重要的估计方法：
  - ✓ **极大似然估计法 (MLE)**
  - ✓ **矩估计法 (MME)**

  其中矩估计法经计量经济学家扩展为**广义矩估计法 (GMM)**。
- MLE 是基于正确设定的总体分布的参数 PMF/PDF 模型，而 MME 和 GMM 则基于一组关于总体分布的矩条件。
- **MLE的优势：**因为是基于随机样本联合分布的信息，**MLE 相对于 MME 或 GMM 一般来说总是更有效**，除非后者所使用的样本矩是参数的充分统计量。

- **GMM的优势：GMM 不需要数据生成过程的总体分布信息。**
- 为深入分析 MLE 和 MME/GMM 的概率统计性质，本章考察了这两个估计量的渐近性质。
- 为评估同一参数的不同估计量，本章引入均方误准则度量参数估计量和未知参数之间的接近程度，并且介绍两种评估无偏参数估计量的重要方法：
  - ✓ 一是拉格朗日乘子法，该法不需要关于总体分布 PMF/PDF 的信息；
  - ✓ 二是克拉默-拉奥（Cramer-Rao）下界法，此法需要随机样本的似然函数的信息。

中国科学院数学与系统科学研究院

Academy of Mathematics and Systems Science
Chinese Academy of Sciences

Thank You !