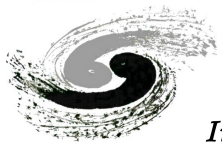
最大似然法和 Fisher 信息

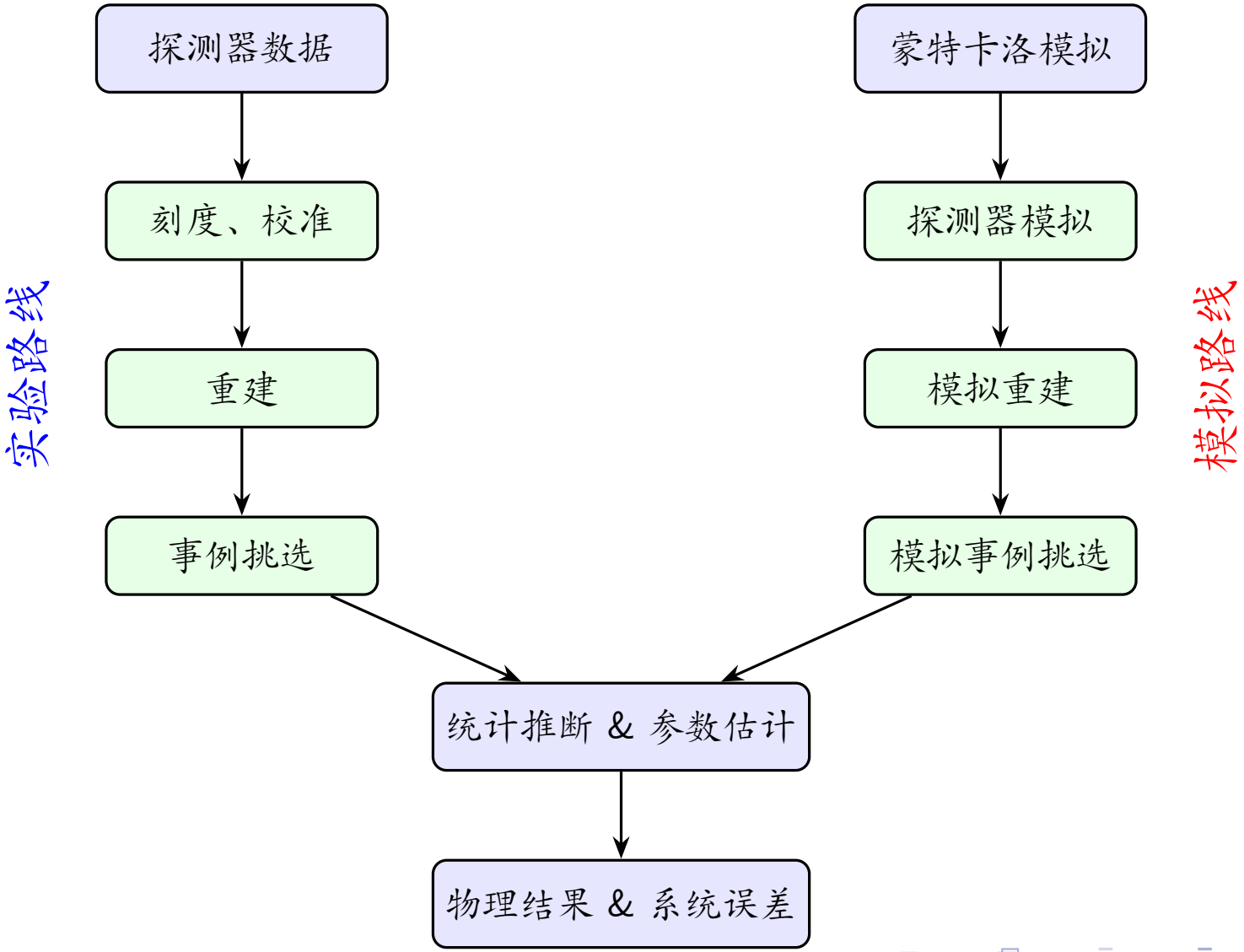
李刚

2026 年 7 月 13 日



中国科学院高能物理研究所
Institute of High Energy Physics, Chinese Academy of Sciences

高能物理数据处理流程


探测器数据
刻度、校准
重建
事例挑选
实验路线
蒙特卡洛模拟
探测器模拟
模拟重建
模拟事例挑选
模拟路线
统计推断 & 参数估计
物理结果 & 系统误差

高能物理数据分析

- 关于 HEP 数据分析，如果只能说两件事

高能物理数据分析

- 关于 HEP 数据分析，如果只能说两件事
  - 统计推断

高能物理数据分析

- 关于 HEP 数据分析，如果只能说两件事
  - 统计推断
  - 假设检验

高能物理数据分析

- 关于 HEP 数据分析，如果只能说两件事
  - 统计推断
  - 假设检验
- 如果可以说第三件事

高能物理数据分析

- 关于 HEP 数据分析，如果只能说两件事
  - 统计推断
  - 假设检验
- 如果可以说第三件事
  - 加上事例挑选，或者用机器学习的说法叫做分类

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；
  - 分布特征：定量，参数空间。

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；
  - 分布特征：定量，参数空间。
- 参数估计：假定总体分布形式已知，但包含未知参数 θ，需要基于样本估计 θ。

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；
  - 分布特征：定量，参数空间。
- 参数估计：假定总体分布形式已知，但包含未知参数 θ，需要基于样本估计 θ。
- 两类估计：

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；
  - 分布特征：定量，参数空间。
- 参数估计：假定总体分布形式已知，但包含未知参数 θ，需要基于样本估计 θ。
- 两类估计：
  - 点估计：构造一个统计量 $\hat{\theta}$ 作为 θ 的估计值。

统计推断与参数估计

- 统计推断：利用（选择后）样本信息对总体分布或其特征作出推测。
  - 总体分布：定性，函数空间（大多数时候知道）；
  - 分布特征：定量，参数空间。
- 参数估计：假定总体分布形式已知，但包含未知参数 θ，需要基于样本估计 θ。
- 两类估计：
  - 点估计：构造一个统计量 $\hat{\theta}$ 作为 θ 的估计值。
  - 区间估计：给出一个随机区间，以一定置信度包含 θ。

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。

“*Don‘t be a regression monkey – understand your model and data.*”

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。
- 懂原理：理解 PDF 和似然函数；注意事例选择、探测效率可能引入偏差：修正。

“*Don‘t be a regression monkey – understand your model and data.*”

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。
- 懂原理：理解 PDF 和似然函数；注意事例选择、探测效率可能引入偏差：修正。
- 重数据：样本量，变量物理意义、分布；统计误差 $\propto 1/\sqrt{N}$；系统误差源于探测器模拟；通过控制样本检验数据与模拟的一致性。

“*Don't be a regression monkey – understand your model and data.*”

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。
- 懂原理：理解 PDF 和似然函数；注意事例选择、探测效率可能引入偏差：修正。
- 重数据：样本量，变量物理意义、分布；统计误差 $\propto 1/\sqrt{N}$；系统误差源于探测器模拟；通过控制样本检验数据与模拟的一致性。
- 讲清楚：拟合结果应包含参数估计值、统计误差、相关性。

“*Don‘t be a regression monkey – understand your model and data.*”

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。
- 懂原理：理解 PDF 和似然函数；注意事例选择、探测效率可能引入偏差：修正。
- 重数据：样本量，变量物理意义、分布；统计误差 $\propto 1/\sqrt{N}$；系统误差源于探测器模拟；通过控制样本检验数据与模拟的一致性。
- 讲清楚：拟合结果应包含参数估计值、统计误差、相关性。
- 做检验：改变拟合范围、bin 宽验证稳定性；分探测器区域拟合检验一致性：关键拟合参数是否稳定。

“*Don‘t be a regression monkey – understand your model and data.*”

避免在统计推断中成为“Reg Monkey”

- 重意义：拟合服务问题，不是为拟合而拟合；避免过度调参。
- 懂原理：理解 PDF 和似然函数；注意事例选择、探测效率可能引入偏差：修正。
- 重数据：样本量，变量物理意义、分布；统计误差 $\propto 1/\sqrt{N}$；系统误差源于探测器模拟；通过控制样本检验数据与模拟的一致性。
- 讲清楚：拟合结果应包含参数估计值、统计误差、相关性。
- 做检验：改变拟合范围、bin 宽验证稳定性；分探测器区域拟合检验一致性：关键拟合参数是否稳定。
- 判断因果：识别真信号，而不是 cut 引起的假信号；用信号区与 Sideband 区类比“处理组”与“对照组”；通过模拟理解产生机制。

“*Don't be a regression monkey – understand your model and data.*”

高能物理数据分析

今天我们只关注从统计角度理解模型，且仅以最大似然法为例

估计量的评价标准

- 无偏性：$\mathbb{E}[\hat{\theta}] = \theta$。

估计量的评价标准

- 无偏性：$\mathbb{E}[\hat{\theta}] = \theta$。
- 有效性：在无偏估计中，方差越小越有效。

估计量的评价标准

- 无偏性：$\mathbb{E}[\hat{\theta}] = \theta$。
- 有效性：在无偏估计中，方差越小越有效。
- 一致性：当样本量 $n \to \infty$ 时，$\hat{\theta} \xrightarrow{P} \theta$（$P$：依概率收敛）。

估计量的评价标准

- 无偏性：$\mathbb{E}[\hat{\theta}] = \theta$。
- 有效性：在无偏估计中，方差越小越有效。
- 一致性：当样本量 $n \to \infty$ 时，$\hat{\theta} \xrightarrow{P} \theta$（$P$：依概率收敛）。
- 这些性质为后续讨论最大似然估计的优良性打下基础。

常用估计方法

- 矩估计法：用样本矩替换总体矩，解方程得到参数估计。

本报告重点介绍最大似然估计及 Fisher 信息。

常用估计方法

- 矩估计法：用样本矩替换总体矩，解方程得到参数估计。
- 最大似然估计法：选择使观测数据出现概率（似然函数）最大的参数值。

本报告重点介绍最大似然估计及 Fisher 信息。

常用估计方法

- 矩估计法：用样本矩替换总体矩，解方程得到参数估计。
- 最大似然估计法：选择使观测数据出现概率（似然函数）最大的参数值。
- 贝叶斯估计：结合先验分布与样本信息得到后验分布，进而获得估计。

本报告重点介绍最大似然估计及 Fisher 信息。

最大似然思想

- 直观想法：给定一组观测数据，我们应选择最可能产生这些数据的参数值。

最大似然思想

- 直观想法：给定一组观测数据，我们应选择最可能产生这些数据的参数值。
- 似然函数 $L(\theta)$ 正是数据出现的概率（密度）作为 θ 的函数。

最大似然思想

- 直观想法：给定一组观测数据，我们应选择最可能产生这些数据的参数值。
- 似然函数 $L(\theta)$ 正是数据出现的概率（密度）作为 θ 的函数。
- 最大似然估计 $\hat{\theta}_{\mathrm{MLE}}$ 即最大化 $L(\theta)$ 的 θ 值。

最大似然思想

- 直观想法：给定一组观测数据，我们应选择最可能产生这些数据的参数值。
- 似然函数 $L(\theta)$ 正是数据出现的概率（密度）作为 θ 的函数。
- 最大似然估计 $\hat{\theta}_{\mathrm{MLE}}$ 即最大化 $L(\theta)$ 的 θ 值。
- 在正则条件下，MLE 具有渐近无偏性、渐近正态性和渐近有效性。

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；
  - 概率密度函数 $f(x;\theta)$ 的支撑集与 θ 无关；

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；
  - 概率密度函数 $f(x;\theta)$ 的支撑集与 θ 无关；
  - $f(x;\theta)$ 关于 θ 三阶可导，且导数与积分可交换；

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；
  - 概率密度函数 $f(x;\theta)$ 的支撑集与 θ 无关；
  - $f(x;\theta)$ 关于 θ 三阶可导，且导数与积分可交换；
  - Fisher 信息 $I(\theta)$ 存在且大于零；

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；
  - 概率密度函数 $f(x;\theta)$ 的支撑集与 θ 无关；
  - $f(x;\theta)$ 关于 θ 三阶可导，且导数与积分可交换；
  - Fisher 信息 $I(\theta)$ 存在且大于零；
  - 对数似然函数 $\ell(\theta)$ 在真实参数附近有唯一最大值。

正则条件 (Regularity Conditions)

- 最大似然估计的优良性质需要满足一些正则条件。
- 常见条件包括：
  - 参数空间 Θ 是开集；
  - 概率密度函数 $f(x;\theta)$ 的支撑集与 θ 无关；
  - $f(x;\theta)$ 关于 θ 三阶可导，且导数与积分可交换；
  - Fisher 信息 $I(\theta)$ 存在且大于零；
  - 对数似然函数 $\ell(\theta)$ 在真实参数附近有唯一最大值。
- 这些条件保证了似然函数的良好行为，使得 MLE 的渐近理论成立。

似然函数

- 考虑来自总体的样本 $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$，其概率密度为 $f(\mathbf{x}; \boldsymbol{\lambda})$。

似然函数

- 考虑来自总体的样本 $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$，其概率密度为 $f(\mathbf{x}; \boldsymbol{\lambda})$。
- 样本的联合概率（密度）为

$$L(\boldsymbol{\lambda}) = \prod_{j=1}^{N} f(\mathbf{x}^{(j)}; \boldsymbol{\lambda})$$

将其视为 $\boldsymbol{\lambda}$ 的函数，称为：似然函数。

似然函数

- 考虑来自总体的样本 $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$，其概率密度为 $f(\mathbf{x}; \boldsymbol{\lambda})$。
- 样本的联合概率（密度）为

$$L(\boldsymbol{\lambda}) = \prod_{j=1}^{N} f(\mathbf{x}^{(j)}; \boldsymbol{\lambda})$$

  将其视为 $\boldsymbol{\lambda}$ 的函数，称为：似然函数。
- 通常使用：（负）对数似然 $NNL = -\ell(\boldsymbol{\lambda}) = -\ln L(\boldsymbol{\lambda})$。

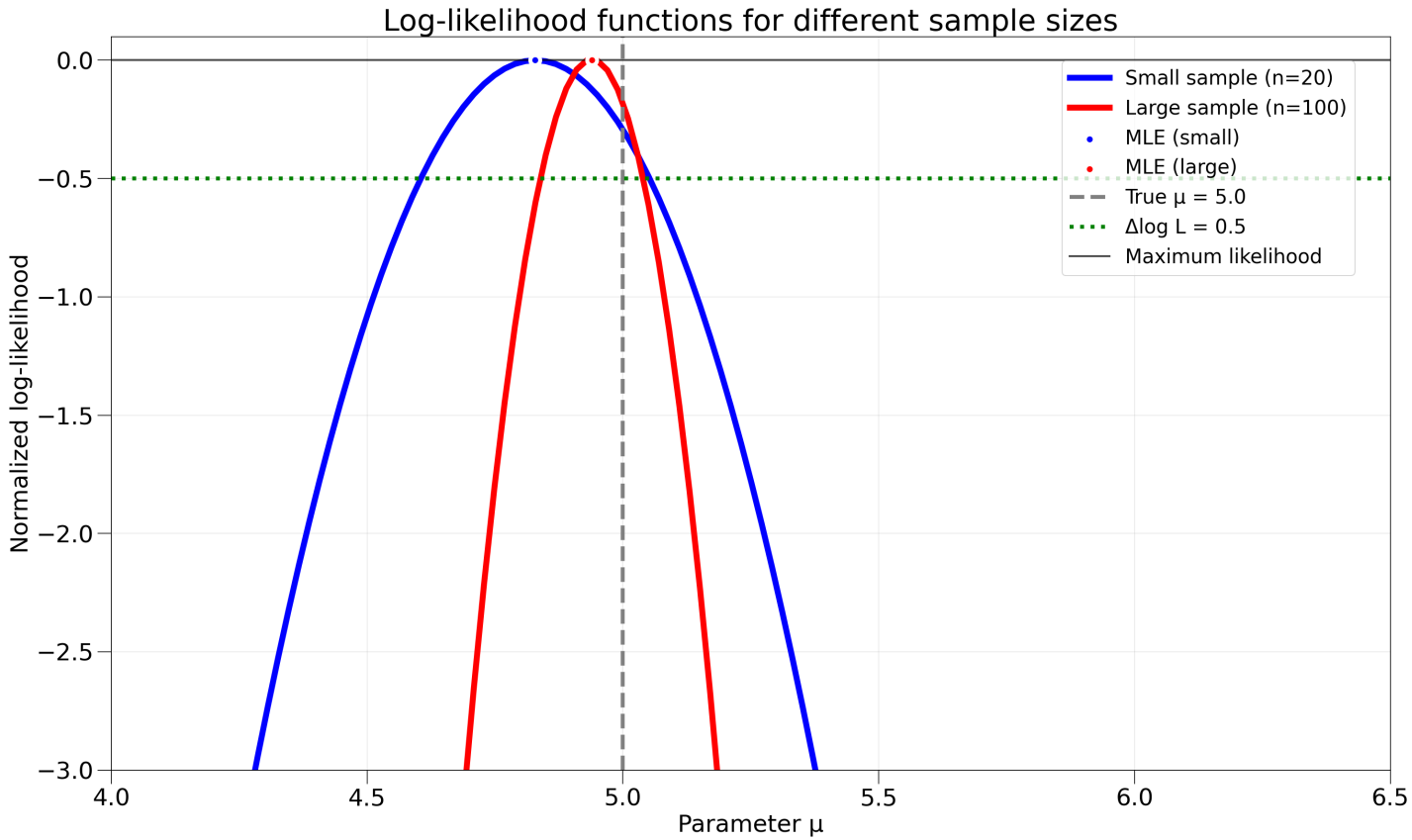


图 1: Likelihood comparison: 20 events and 100 events

最大似然原理

- 最大似然估计量 $\hat{\boldsymbol{\lambda}}$ 是使 $L(\boldsymbol{\lambda})$（或 $\ell(\boldsymbol{\lambda})$）达到最大的 $\boldsymbol{\lambda}$ 值。

最大似然原理

- 最大似然估计量 $\hat{\boldsymbol{\lambda}}$ 是使 $L(\boldsymbol{\lambda})$（或 $\ell(\boldsymbol{\lambda})$）达到最大的 $\boldsymbol{\lambda}$ 值。
- 对于单个参数 λ，极值存在的（必要）条件为

$$\ell'(\lambda) = \frac{d}{d\lambda} \ln L = 0.$$

最大似然原理

- 最大似然估计量 $\hat{\boldsymbol{\lambda}}$ 是使 $L(\boldsymbol{\lambda})$（或 $\ell(\boldsymbol{\lambda})$）达到最大的 $\boldsymbol{\lambda}$ 值。
- 对于单个参数 λ，极值存在的（必要）条件为

$$\ell'(\lambda) = \frac{d}{d\lambda} \ln L = 0.$$

- 对于多个参数，条件为

$$\frac{\partial \ell}{\partial \lambda_i} = 0, \quad i = 1, \ldots, p.$$

最大似然原理

- 最大似然估计量 $\hat{\boldsymbol{\lambda}}$ 是使 $L(\boldsymbol{\lambda})$（或 $\ell(\boldsymbol{\lambda})$）达到最大的 $\boldsymbol{\lambda}$ 值。
- 对于单个参数 λ，极值存在的（必要）条件为

$$\ell'(\lambda) = \frac{d}{d\lambda} \ln L = 0.$$

- 对于多个参数，条件为

$$\frac{\partial \ell}{\partial \lambda_i} = 0, \quad i = 1, \ldots, p.$$

- 该方程称为得分 **(score)** 方程，其解即为 MLE。

例：不同精度的重复测量

- 对同一量 λ 进行 N 次测量 $x^{(j)}$，已知各测量有标准差 σ_j，服从正态分布。

例：不同精度的重复测量

- 对同一量 λ 进行 N 次测量 $x^{(j)}$，已知各测量有标准差 σ_j，服从正态分布。
- 似然函数：$L = \prod \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x^{(j)}-\lambda)^2}{2\sigma_j^2}\right)$。

例：不同精度的重复测量

- 对同一量 λ 进行 N 次测量 $x^{(j)}$，已知各测量有标准差 σ_j，服从正态分布。
- 似然函数：$L = \prod \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x^{(j)}-\lambda)^2}{2\sigma_j^2}\right)$。
- 对数似然：$\ell = -\frac{1}{2}\sum \frac{(x^{(j)}-\lambda)^2}{\sigma_j^2} +$ 常数。

例：不同精度的重复测量

- 对同一量 λ 进行 N 次测量 $x^{(j)}$，已知各测量有标准差 σ_j，服从正态分布。
- 似然函数：$L = \prod \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x^{(j)}-\lambda)^2}{2\sigma_j^2}\right)$。
- 对数似然：$\ell = -\frac{1}{2}\sum \frac{(x^{(j)}-\lambda)^2}{\sigma_j^2} + \text{常数}$。
- 最大似然估计：

$$\hat{\lambda} = \frac{\sum x^{(j)}/\sigma_j^2}{\sum 1/\sigma_j^2},$$

即加权平均。

例：不同精度的重复测量

- 对同一量 λ 进行 N 次测量 $x^{(j)}$，已知各测量有标准差 σ_j，服从正态分布。
- 似然函数：$L = \prod \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x^{(j)}-\lambda)^2}{2\sigma_j^2}\right)$。
- 对数似然：$\ell = -\frac{1}{2}\sum \frac{(x^{(j)}-\lambda)^2}{\sigma_j^2} +$ 常数。
- 最大似然估计：

$$\hat{\lambda} = \frac{\sum x^{(j)}/\sigma_j^2}{\sum 1/\sigma_j^2},$$

即加权平均。

- 启示：*MLE* 自然地处理异方差数据，给出最优加权方案。

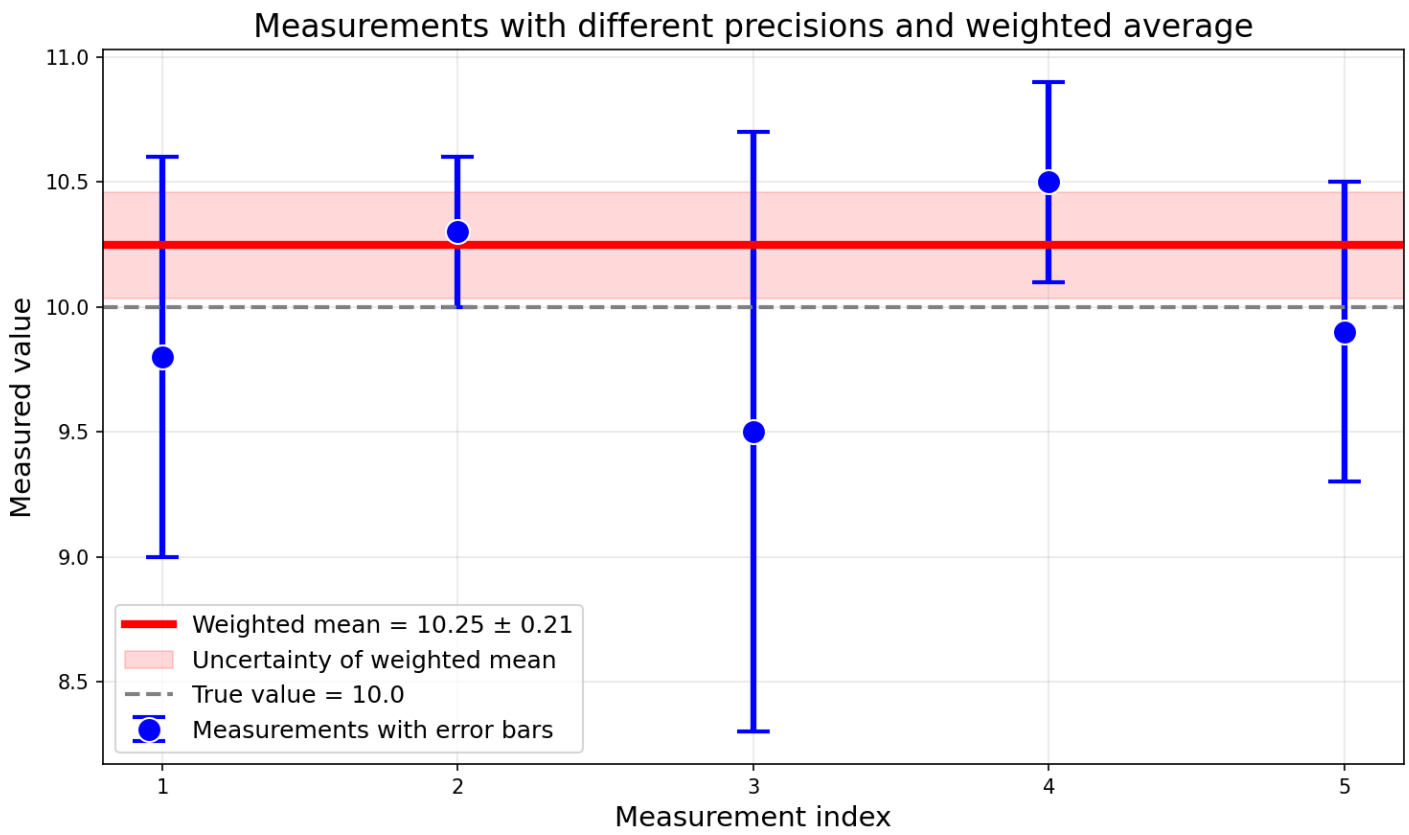


图 2: Weighted average of varying uncertainties

似然比

- 对于两个简单假设 $\boldsymbol{\lambda}_1$ 和 $\boldsymbol{\lambda}_2$，似然比

$$Q = \frac{L(\boldsymbol{\lambda}_1)}{L(\boldsymbol{\lambda}_2)}$$

衡量一组参数相对于另一组参数的可能性有多大。

似然比

- 对于两个简单假设 $\boldsymbol{\lambda}_1$ 和 $\boldsymbol{\lambda}_2$，似然比

$$Q = \frac{L(\boldsymbol{\lambda}_1)}{L(\boldsymbol{\lambda}_2)}$$

衡量一组参数相对于另一组参数的可能性有多大。

- 示例：掷硬币，正面概率为 p。掷 n 次，$L(p) = p^k(1-p)^{n-k}$。若 $p_A = 1/3$，$p_B = 2/3$，观察到 $k = 1$，$n = 5$，则 $Q = \frac{(1/3)(2/3)^4}{(2/3)(1/3)^4} = 8$，支持 p_A。

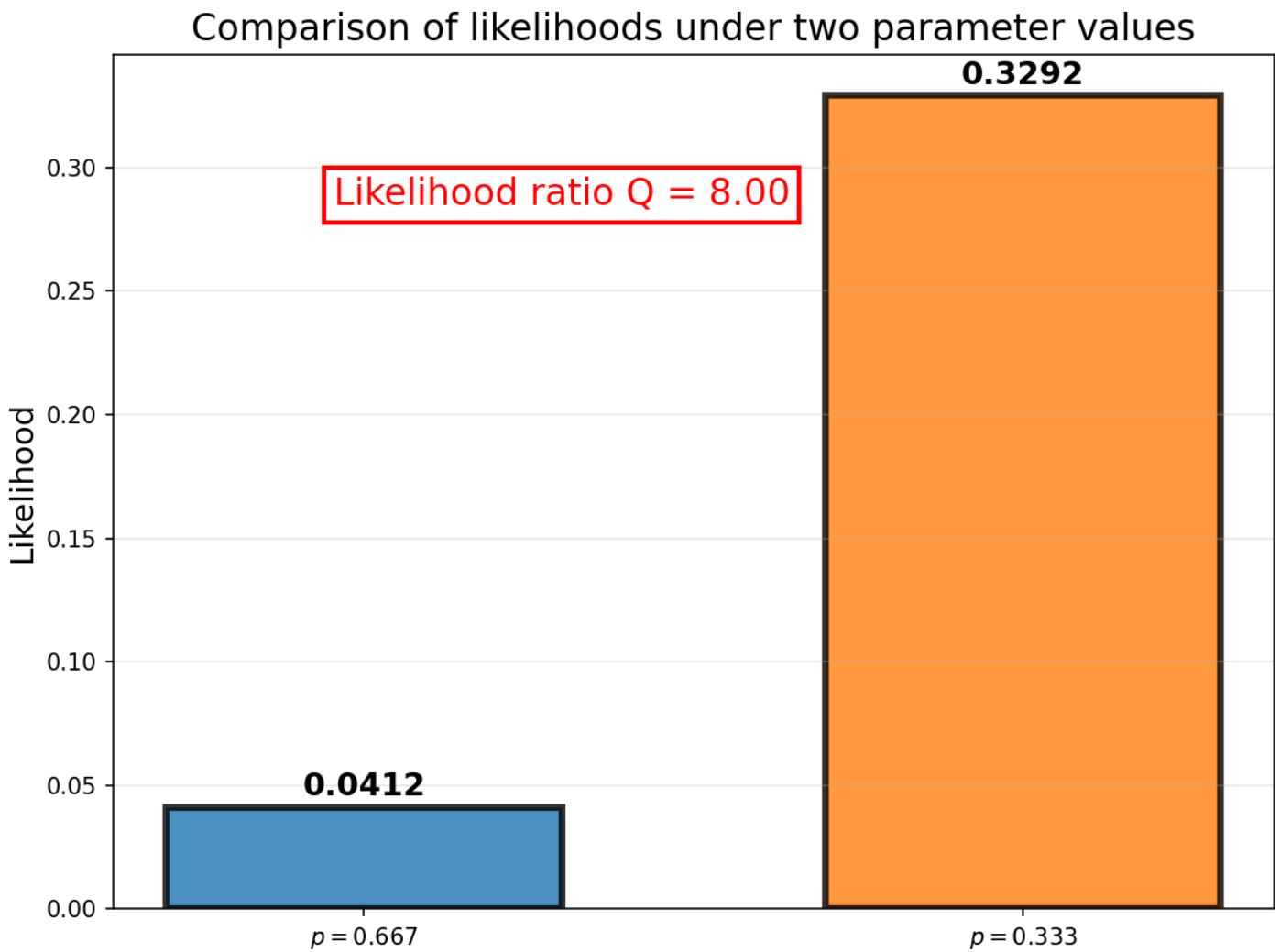


图 3: Likelihood ratio for a binomial distribution with $\hat{p} = 1/3$

动机：数据包含多少参数信息？

- 直观上，若参数微小变化导致数据分布剧烈改变，则数据包含大量信息。

动机：数据包含多少参数信息？

- 直观上，若参数微小变化导致数据分布剧烈改变，则数据包含大量信息。
- 如何度量这种敏感度？考虑对数似然的一阶导数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

称为得分函数。它反映了参数变化时对数似然的“瞬时”变化率。

动机：数据包含多少参数信息？

- 直观上，若参数微小变化导致数据分布剧烈改变，则数据包含大量信息。
- 如何度量这种敏感度？考虑对数似然的一阶导数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

称为得分函数。它反映了参数变化时对数似然的“瞬时”变化率。

- 得分函数的绝对值大意味着数据对参数敏感，但单次观测的得分有正有负，期望为零。

动机：数据包含多少参数信息？

- 直观上，若参数微小变化导致数据分布剧烈改变，则数据包含大量信息。
- 如何度量这种敏感度？考虑对数似然的一阶导数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$$

称为得分函数。它反映了参数变化时对数似然的“瞬时”变化率。

- 得分函数的绝对值大意味着数据对参数敏感，但单次观测的得分有正有负，期望为零。
- 因此用得分函数的方差衡量信息量更为合理：

$$I(\theta) = \mathrm{Var}(s(\theta)) = \mathbb{E}[s(\theta)^2]$$

这就是 **Fisher** 信息的定义。

Fisher 信息的定义

- 设样本 X 服从分布 $f(x;\theta)$，定义得分函数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X;\theta)$$

Fisher 信息的定义

- 设样本 X 服从分布 $f(x;\theta)$，定义得分函数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X;\theta)$$

- **Fisher 信息：**

$$I(\theta) = \mathbb{E}\left[(s(\theta))^2\right]$$

Fisher 信息的定义

- 设样本 X 服从分布 $f(x;\theta)$，定义得分函数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X;\theta)$$

- **Fisher** 信息：

$$I(\theta) = \mathbb{E}\left[(s(\theta))^2\right]$$

- 在正则条件下，还可表示为对数似然二阶矩的负期望：

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta)\right]$$

（证明由 $\mathbb{E}[s] = 0$ 及分部积分可得）

Fisher 信息的定义

- 设样本 X 服从分布 $f(x;\theta)$，定义得分函数：

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(X;\theta)$$

- **Fisher** 信息：

$$I(\theta) = \mathbb{E}\left[(s(\theta))^2\right]$$

- 在正则条件下，还可表示为对数似然二阶矩的负期望：

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta)\right]$$

（证明由 $\mathbb{E}[s] = 0$ 及分部积分可得）

- 对于 N 个**独立同分布 (i.i.d.)** 样本：
总信息量 $I_N(\theta) = NI(\theta)$（可加性）。

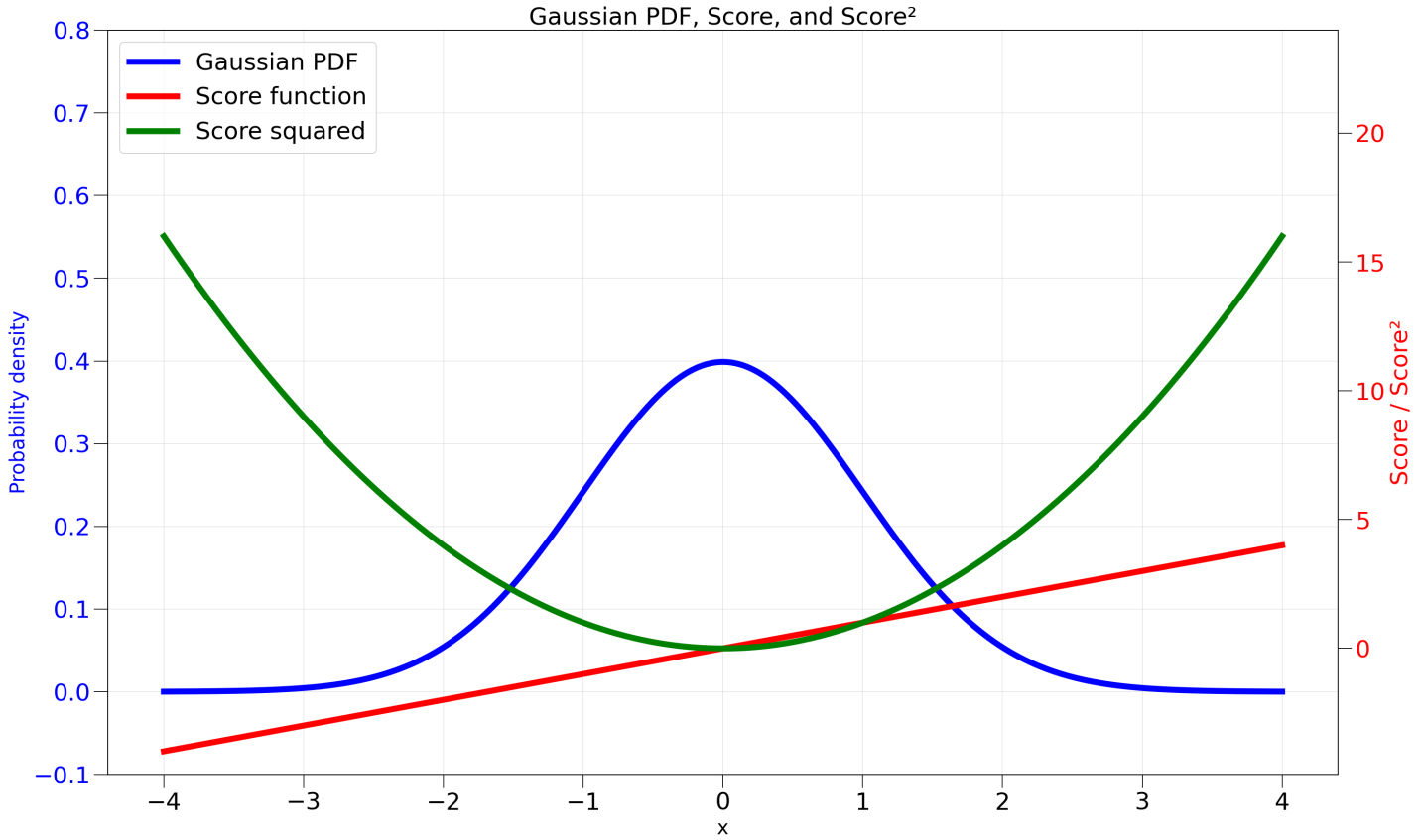


图 4: Gaussian: PDF, score and squared score(不是 Fisher information)

典型分布的 Fisher 信息（单参数）

- 泊松分布 $f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$：

$$I(\lambda) = \frac{1}{\lambda} \quad \text{(单个观测)}$$

- 二项分布 $f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$（$n$ 已知）：

$$I(p) = \frac{n}{p(1-p)}$$

- 指数分布 $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$：

$$I(\tau) = \frac{1}{\tau^2}$$

- 这些表达式将在后续 CRLB 中反复使用。

多参数 Fisher 信息矩阵

- 对于参数向量 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$，定义得分向量：

$$\nabla \ell(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \theta_1}, \ldots, \frac{\partial \ell}{\partial \theta_p} \right)^T$$

多参数 Fisher 信息矩阵

- 对于参数向量 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$，定义得分向量：

$$\nabla \ell(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \theta_1}, \ldots, \frac{\partial \ell}{\partial \theta_p} \right)^T$$

- **Fisher** 信息矩阵 $I(\boldsymbol{\theta})$ 为 $p \times p$ 矩阵，元素为

$$I_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]$$

多参数 Fisher 信息矩阵

- 对于参数向量 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$，定义得分向量：

$$\nabla \ell(\boldsymbol{\theta}) = \left(\frac{\partial \ell}{\partial \theta_1}, \ldots, \frac{\partial \ell}{\partial \theta_p} \right)^T$$

- **Fisher** 信息矩阵 $I(\boldsymbol{\theta})$ 为 $p \times p$ 矩阵，元素为

$$I_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right]$$

- i.i.d. 时总信息矩阵为 $nI(\boldsymbol{\theta})$。

多参数示例：高斯分布

- 正态分布 $N(\mu, \sigma^2)$，参数 $\boldsymbol{\theta} = (\mu, \sigma)$：

$$I(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

- 若参数化为 (μ, σ^2)：

$$I(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

- 柯西分布 $f(x; x_0, \gamma)$：

$$I(x_0, \gamma) = \begin{bmatrix} \frac{1}{2\gamma^2} & 0 \\ 0 & \frac{1}{2\gamma^2} \end{bmatrix}$$

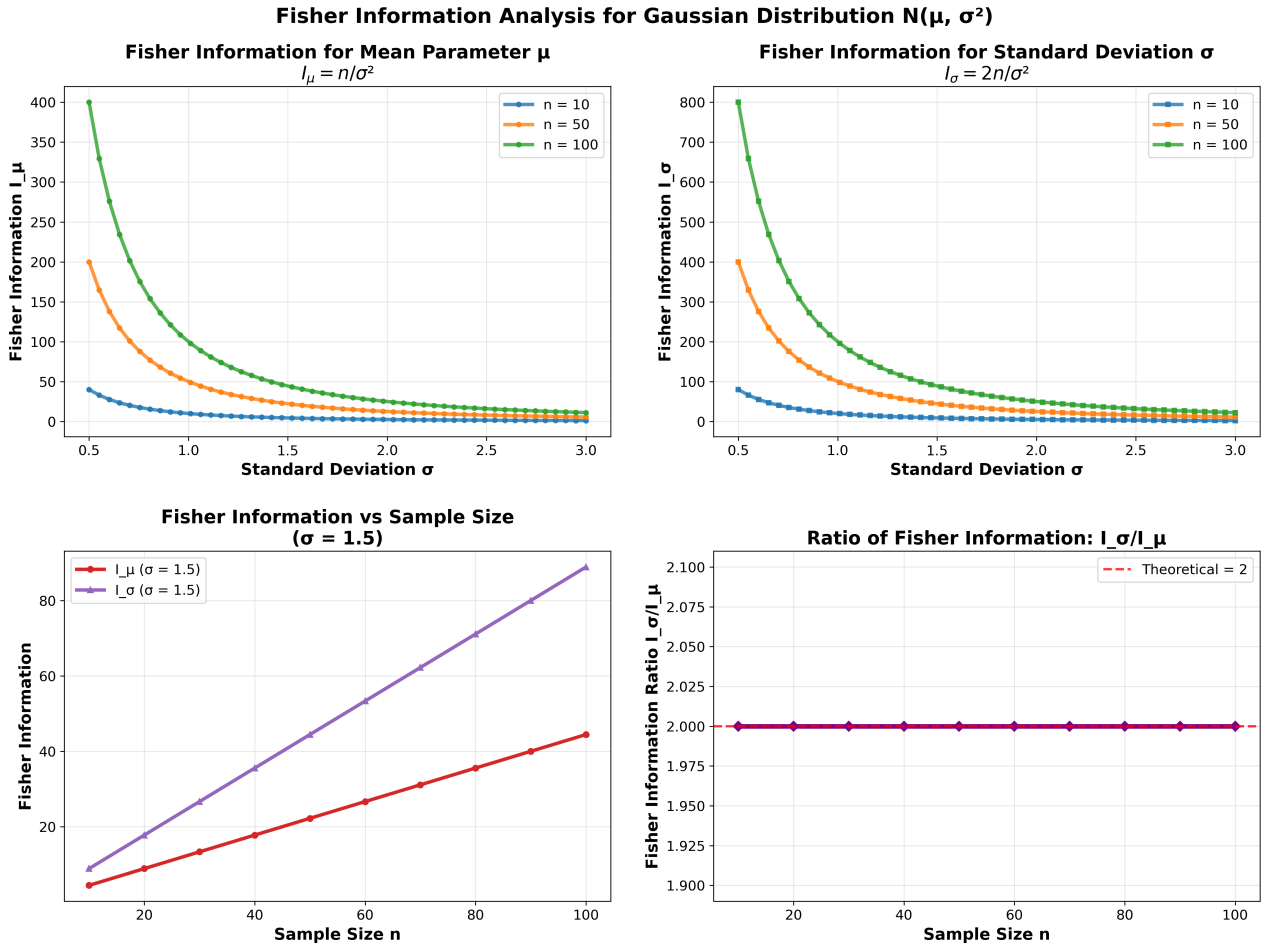


图 5: Gaussian: Fisher information vs. number of event and the parameter. σ 的变化和 n 的可以互相补偿，如果前者变成 2 倍，则后者需要 4 倍事例数来补偿。

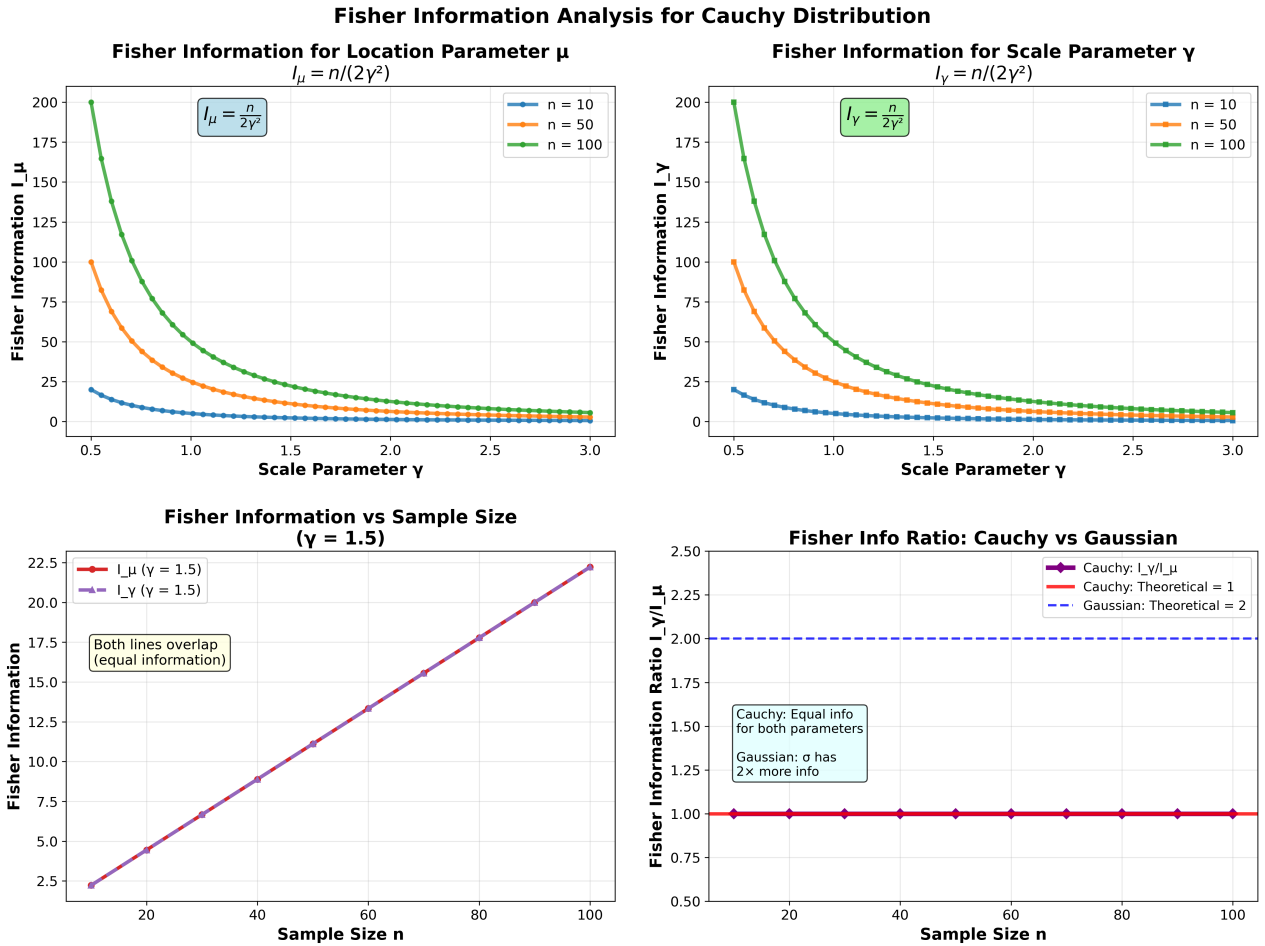


图 6: Cauchy: Fisher information vs. number of event and the parameter.

γ 的变化和 n 的可以互相补偿，如果前者变成 2 倍，则后者需要 4 倍事例数来补偿。

信息不等式与 CRLB

- 设 S 是 θ 的一个估计量，偏差 $B(\theta) = \mathbb{E}[S] - \theta$。

信息不等式与 CRLB

- 设 S 是 θ 的一个估计量，偏差 $B(\theta) = \mathbb{E}[S] - \theta$。
- 信息不等式（Cramér–Rao）：

$$\mathrm{Var}(S) \geq \frac{[1 + B'(\theta)]^2}{I(\theta)}$$

信息不等式与 CRLB

- 设 S 是 θ 的一个估计量，偏差 $B(\theta) = \mathbb{E}[S] - \theta$。
- 信息不等式 (Cramér–Rao)：

$$\mathrm{Var}(S) \geq \frac{[1 + B'(\theta)]^2}{I(\theta)}$$

- 对于无偏估计 ($B(\theta) \equiv 0$)：

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

信息不等式与 CRLB

- 设 S 是 θ 的一个估计量，偏差 $B(\theta) = \mathbb{E}[S] - \theta$。
- 信息不等式 (Cramér–Rao)：

$$\mathrm{Var}(S) \geq \frac{[1 + B'(\theta)]^2}{I(\theta)}$$

- 对于无偏估计 ($B(\theta) \equiv 0$)：

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- 等号成立当且仅当得分函数可表示为

$$\ell'(\theta) = A(\theta)\,(\hat{\theta} - \theta)$$

此时估计量称为有效估计量，达到方差下界。

CRLB 的直观意义

- Fisher 信息越大，方差下界越小，参数可被估计得越精确。

CRLB 的直观意义

- Fisher 信息越大，方差下界越小，参数可被估计得越精确。
- 这类似于“海森堡不确定性原理”：参数估计的精度受限于数据本身所含信息。

CRLB 的直观意义

- Fisher 信息越大，方差下界越小，参数可被估计得越精确。
- 这类似于“海森堡不确定性原理”：参数估计的精度受限于数据本身所含信息。
- 若 $I(\theta) \approx 0$（参数不可识别），则无法获得任何有效估计。

CRLB 的直观意义

- Fisher 信息越大，方差下界越小，参数可被估计得越精确。
- 这类似于“海森堡不确定性原理”：参数估计的精度受限于数据本身所含信息。
- 若 $I(\theta) \approx 0$（参数不可识别），则无法获得任何有效估计。
- CRLB 为实验设计提供指导：要达到所需精度，需要多少样本？

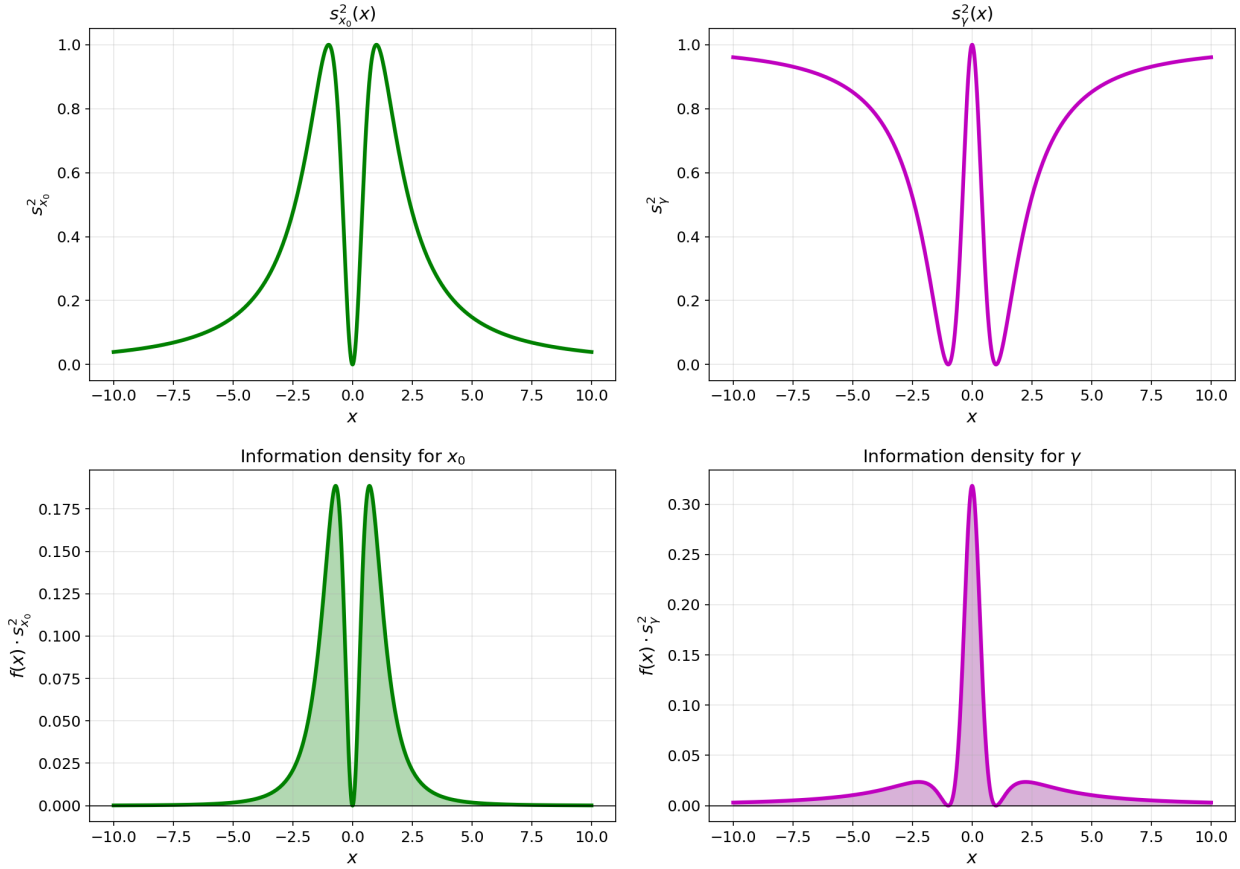


图 7: Cauchy 分布的 score，score^2，以及 Fisher 信息密度分布。据此可以对实验方案做优化，比如测量 Z 粒子的质量和宽度时，选择在哪些能量点取数，取多少数据等。

例：泊松分布达到 CRLB

- $f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$，样本 $k_1, \ldots, k_N$。

例：泊松分布达到 CRLB

- $f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$，样本 $k_1, \ldots, k_N$。
- 对数似然：$\ell(\lambda) = \sum(k_j \ln \lambda - \lambda - \ln k_j!)$

例：泊松分布达到 CRLB

- $f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$，样本 $k_1, \ldots, k_N$。
- 对数似然：$\ell(\lambda) = \sum(k_j \ln\lambda - \lambda - \ln k_j!)$
- 得分函数：$\ell'(\lambda) = \frac{N}{\lambda}(\bar{k} - \lambda)$

例：泊松分布达到 CRLB

- $f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$，样本 $k_1, \ldots, k_N$。
- 对数似然：$\ell(\lambda) = \sum(k_j \ln\lambda - \lambda - \ln k_j!)$
- 得分函数：$\ell'(\lambda) = \frac{N}{\lambda}(\bar{k} - \lambda)$
- 由 CRLB 等号条件 $[\ell'(\theta) = A(\theta)\,(\hat{\theta} - \theta)]$)，$\bar{k}$ 是方差最小的无偏估计，且

$$\mathrm{Var}(\bar{k}) = \frac{\lambda}{N} = \frac{1}{I(\lambda)}$$

例：泊松分布达到 CRLB

- $f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$，样本 $k_1, \ldots, k_N$。
- 对数似然：$\ell(\lambda) = \sum(k_j \ln\lambda - \lambda - \ln k_j!)$
- 得分函数：$\ell'(\lambda) = \frac{N}{\lambda}(\bar{k} - \lambda)$
- 由 CRLB 等号条件 $[\ell'(\theta) = A(\theta)\,(\hat{\theta} - \theta)]$，$\bar{k}$ 是方差最小的无偏估计，且

$$\mathrm{Var}(\bar{k}) = \frac{\lambda}{N} = \frac{1}{I(\lambda)}$$

- 启示：样本均值是泊松分布的最优估计。

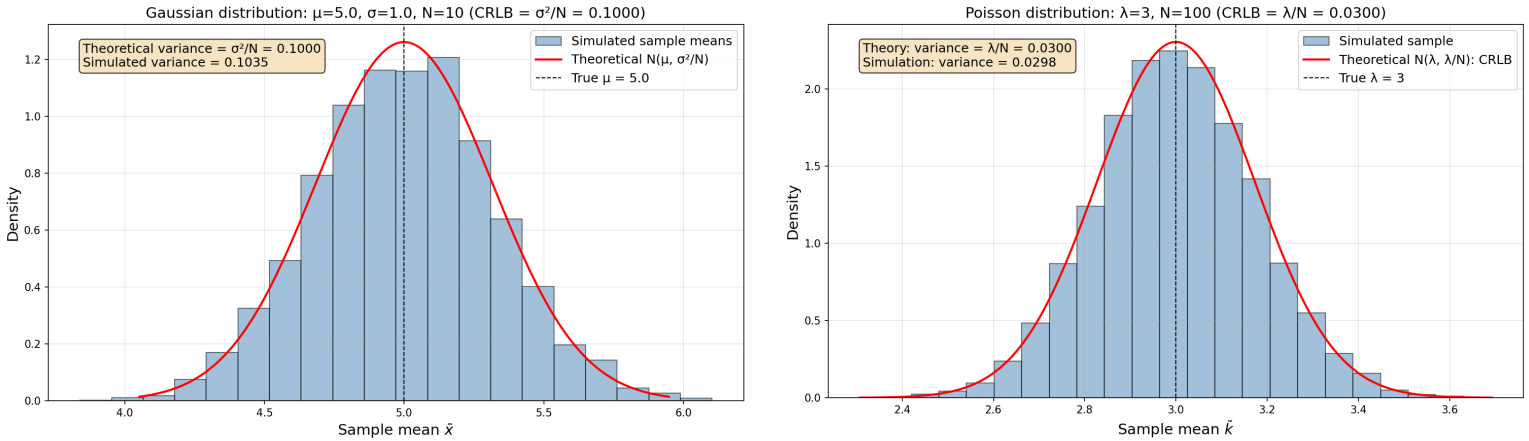


图 8: Gaussian & Poisson distributions: sample mean achieves CRLB

例：加权平均再次达到 CRLB

- 回顾不同精度测量例：$\ell'(\lambda) = \sum \frac{x^{(j)} - \lambda}{\sigma_j^2}$

例：加权平均再次达到 CRLB

- 回顾不同精度测量例：$\ell'(\lambda) = \sum \frac{x^{(j)} - \lambda}{\sigma_j^2}$
- 可改写为 $\left(\sum \frac{1}{\sigma_j^2} \right) (\hat{\lambda} - \lambda)$，其中 $\hat{\lambda}$ 为加权平均。

例：加权平均再次达到 CRLB

- 回顾不同精度测量例：$\ell'(\lambda) = \sum \frac{x^{(j)} - \lambda}{\sigma_j^2}$
- 可改写为 $\left(\sum \frac{1}{\sigma_j^2} \right) (\hat{\lambda} - \lambda)$，其中 $\hat{\lambda}$ 为加权平均。
- 因此 $\hat{\lambda}$ 达到 CRLB，其方差为 $\left(\sum \frac{1}{\sigma_j^2} \right)^{-1}$。

例：加权平均再次达到 CRLB

- 回顾不同精度测量例：$\ell'(\lambda) = \sum \frac{x^{(j)} - \lambda}{\sigma_j^2}$
- 可改写为 $\left(\sum \frac{1}{\sigma_j^2} \right) (\hat{\lambda} - \lambda)$，其中 $\hat{\lambda}$ 为加权平均。
- 因此 $\hat{\lambda}$ 达到 CRLB，其方差为 $\left(\sum \frac{1}{\sigma_j^2} \right)^{-1}$。
- 启示：加权平均是异方差正态分布的最优估计。

Fisher 信息的性质 (I)

- 可加性：独立样本的总信息等于各样本信息之和。对于 i.i.d. 样本，$I_{总}(\theta) = NI(\theta)$。

Fisher 信息的性质 (I)

- 可加性：独立样本的总信息等于各样本信息之和。对于 i.i.d. 样本，$I_{总}(\theta) = NI(\theta)$。
- 参数变换的不变性（链式法则）：若 $\psi = g(\theta)$，g 可微，则

$$I(\psi) = I(\theta)\left(\frac{d\theta}{d\psi}\right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

Fisher 信息的性质 (I)

- 可加性：独立样本的总信息等于各样本信息之和。对于 i.i.d. 样本，$I_{总}(\theta) = NI(\theta)$。
- 参数变换的不变性（链式法则）：若 $\psi = g(\theta)$，g 可微，则

$$I(\psi) = I(\theta)\left(\frac{d\theta}{d\psi}\right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

- 多参数时：$I(\eta) = J^T I(\theta) J$，其中 $J_{ij} = \partial\theta_i/\partial\eta_j$。

Fisher 信息的性质 (I)

- 可加性：独立样本的总信息等于各样本信息之和。对于 i.i.d. 样本，$I_{\text{总}}(\theta) = nI(\theta)$。
- 参数变换的不变性（链式法则）：若 $\psi = g(\theta)$，g 可微，则

$$I(\psi) = I(\theta)\left(\frac{d\theta}{d\psi}\right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

- 多参数时：$I(\eta) = J^T I(\theta) J$，其中 $J_{ij} = \partial\theta_i/\partial\eta_j$。
- 与 **KL** 散度的关系：对于邻近参数，

$$D_{\mathrm{KL}}(p(\cdot|\theta)\|p(\cdot|\theta+\delta)) \approx \frac{1}{2}\delta^T I(\theta)\delta$$

因此 Fisher 信息是统计流形上的黎曼度量（超纲：通向《信息几何学》）。

Fisher 信息的性质 (II)

- 得分方差等价：$I(\theta) = \mathrm{Var}(s(\theta)) = \mathbb{E}[s(\theta)^2]$。

Fisher 信息的性质 (II)

- 得分方差等价：$I(\theta) = \mathrm{Var}(s(\theta)) = \mathbb{E}[s(\theta)^2]$。
- 正交参数：若 $I(\theta)$ 为对角阵，则参数分量称为正交，其 MLE 渐近独立。

Fisher 信息的性质 (II)

- 得分方差等价：$I(\theta) = \mathrm{Var}(s(\theta)) = \mathbb{E}[s(\theta)^2]$。
- 正交参数：若 $I(\theta)$ 为对角阵，则参数分量称为正交，其 MLE 渐近独立。
- 指数族的简洁形式：对于指数族 $f(x;\theta) = h(x)\exp(\eta(\theta)^T T(x) - A(\theta))$，

$$I(\theta) = \mathrm{Var}(T(X)) = \frac{\partial^2 A(\theta)}{\partial\theta\partial\theta^T}$$

地形比喻：似然峰的曲率

- 对数似然函数 $\ell(\theta)$ 在真实参数附近形成一个“山峰”。

地形比喻：似然峰的曲率

- 对数似然函数 $\ell(\theta)$ 在真实参数附近形成一个“山峰”。
- 尖峰（高曲率）：θ 微调 对数似然剧烈下降 数据对参数敏感，信息量大。

地形比喻：似然峰的曲率

- 对数似然函数 $\ell(\theta)$ 在真实参数附近形成一个“山峰”。
- 尖峰（高曲率）：θ 微调 $\Rightarrow$ 对数似然剧烈下降 $\Rightarrow$ 数据对参数敏感，信息量大。
- 平顶（低曲率）：θ 大范围移动 $\Rightarrow$ 对数似然几乎不变 $\Rightarrow$ 信息量小，估计不确定。

地形比喻：似然峰的曲率

- 对数似然函数 $\ell(\theta)$ 在真实参数附近形成一个“山峰”。
- 尖峰（高曲率）：θ 微调 $\Rightarrow$ 对数似然剧烈下降 $\Rightarrow$ 数据对参数敏感，信息量大。
- 平顶（低曲率）：θ 大范围移动 $\Rightarrow$ 对数似然几乎不变 $\Rightarrow$ 信息量小，估计不确定。
- Fisher 信息正是对数似然在真实参数处的期望曲率：

$$I(\theta) = -\mathbb{E}[\ell''(\theta)]$$

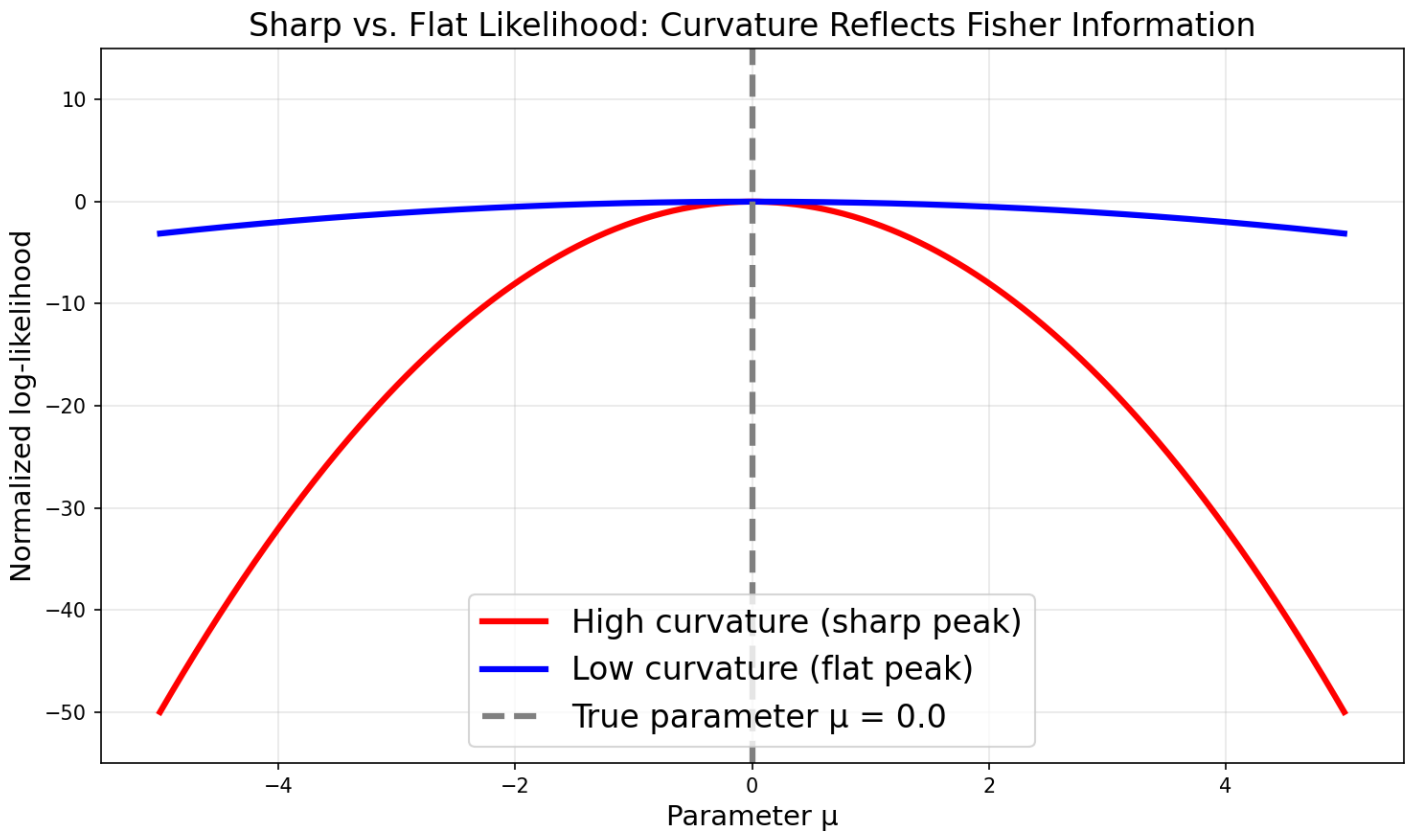


图 9: Sharp vs flat likelihood: high vs low Fisher information(1D)

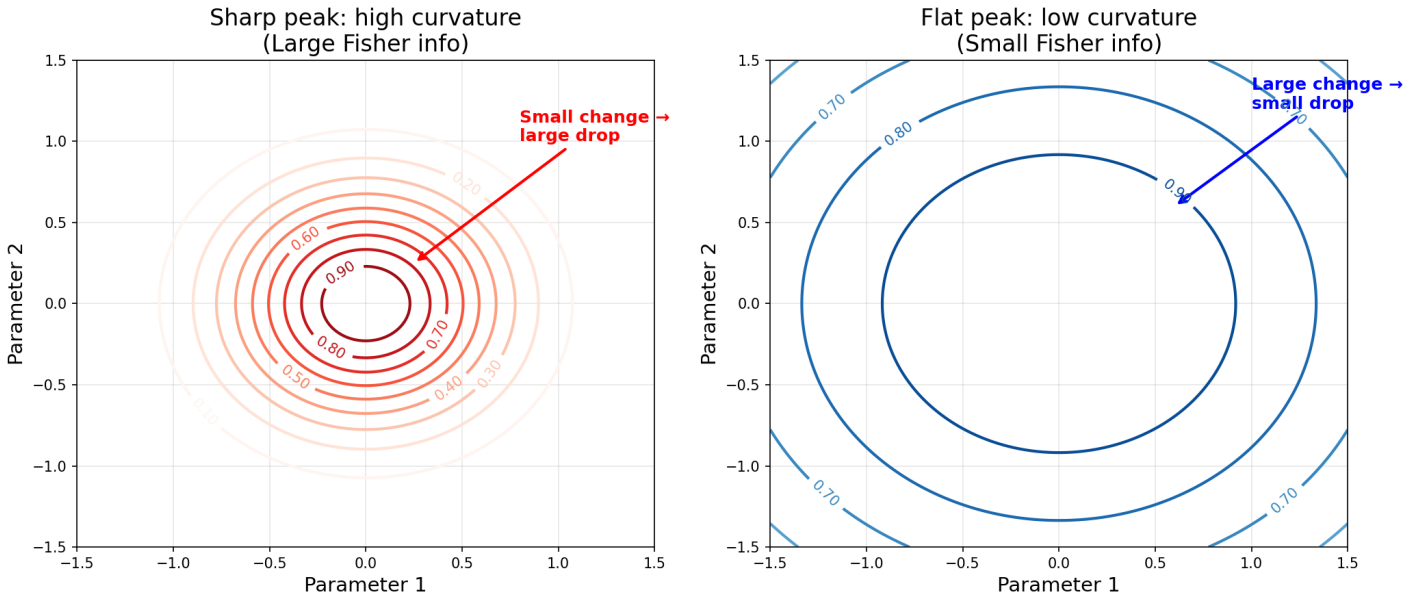


图 10: Sharp vs flat likelihood: high vs low Fisher information (2D)

两个定义的等价性

- 定义一：$I(\theta) = \mathbb{E}[s(\theta)^2]$（得分方差）
- 定义二：$I(\theta) = -\mathbb{E}[\ell''(\theta)]$（期望曲率）
- 为什么相等？关键在于 $\mathbb{E}[s] = 0$，且

$$\frac{\partial}{\partial \theta} \int \frac{\partial f}{\partial \theta} dx = \int \frac{\partial^2 f}{\partial \theta^2} dx = 0$$

通过分部积分可得 $-\mathbb{E}[\ell''] = \mathbb{E}[(\ell')^2]$。

得分分布与曲率

- 若似然很尖：不同样本的得分差异大（方差大）。
- 若似然平坦：得分始终接近零（方差小）。
- 得分方差大 对数似然曲率大 信息量大
- 下图展示高、低曲率下得分的分布：

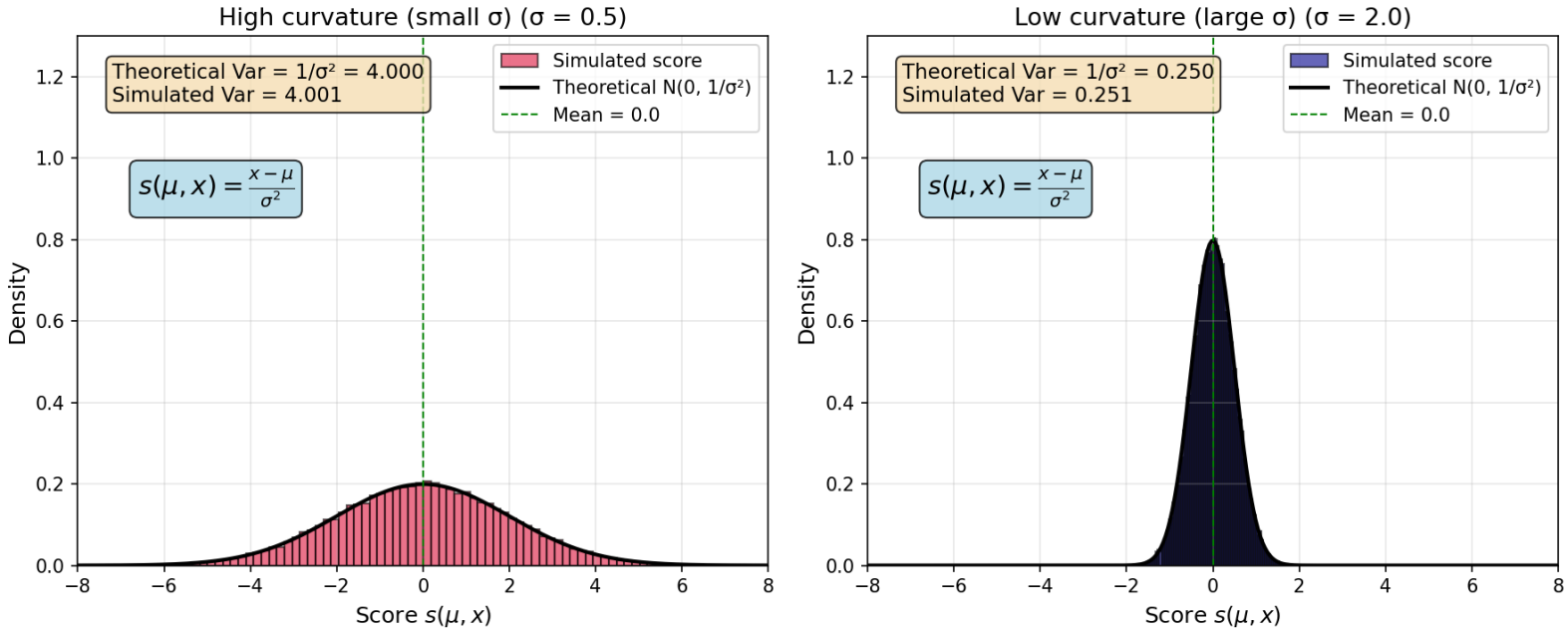


图 11: Score distribution: high vs low curvature

MLE 的渐近性质

- 在正则条件下，MLE $\hat{\theta}_n$ 具有以下性质：

MLE 的渐近性质

- 在正则条件下，MLE $\hat{\theta}_n$ 具有以下性质：
  1. 一致性：$\hat{\theta}_n \xrightarrow{P} \theta$。

MLE 的渐近性质

- 在正则条件下，MLE $\hat{\theta}_n$ 具有以下性质：
  1. 一致性：$\hat{\theta}_n \xrightarrow{P} \theta$。
  2. 渐近正态性：$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, I(\theta)^{-1})$。

MLE 的渐近性质

- 在正则条件下，MLE $\hat{\theta}_n$ 具有以下性质：
  1. 一致性：$\hat{\theta}_n \xrightarrow{P} \theta$。
  2. 渐近正态性：$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, I(\theta)^{-1})$。
  3. 渐近有效性：渐近方差达到 CRLB，即 $I(\theta)^{-1}$。

MLE 的渐近性质

- 在正则条件下，MLE $\hat{\theta}_n$ 具有以下性质：
  1. 一致性：$\hat{\theta}_n \xrightarrow{P} \theta$。
  2. 渐近正态性：$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, I(\theta)^{-1})$。
  3. 渐近有效性：渐近方差达到 CRLB，即 $I(\theta)^{-1}$。
- 这意味着大样本下 MLE 是所有一致估计量中渐近方差最小的。

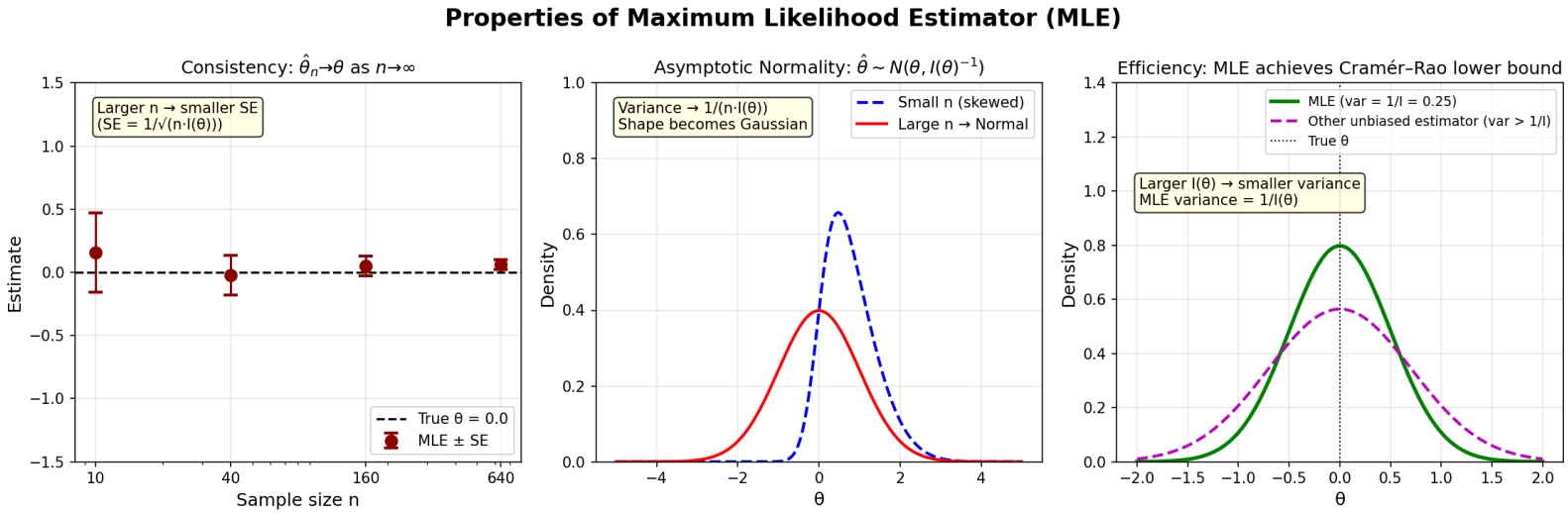


图 12: Illustration of consistency, asymptotic normality and efficiency of MLE

渐近正态性的直观理解

- 对数似然函数在最大值附近可作泰勒展开：

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

渐近正态性的直观理解

- 对数似然函数在最大值附近可作泰勒展开：

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

- 由 WLLN（Weak Law of Large Numbers），$\ell''(\hat{\theta}) \approx -nI(\theta)$。

渐近正态性的直观理解

- 对数似然函数在最大值附近可作泰勒展开：

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

- 由 WLLN（Weak Law of Large Numbers），$\ell''(\hat{\theta}) \approx -nI(\theta)$。
- 因此似然函数近似为 $L(\theta) \approx$ 常数 $\times \exp\left(-\frac{nI(\theta)}{2}(\theta - \hat{\theta})^2\right)$，即关于 θ 的正态核。

渐近正态性的直观理解

- 对数似然函数在最大值附近可作泰勒展开：

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2$$

- 由 WLLN（Weak Law of Large Numbers），$\ell''(\hat{\theta}) \approx -nI(\theta)$。
- 因此似然函数近似为 $L(\theta) \approx$ 常数 $\times \exp\left(-\frac{nI(\theta)}{2}(\theta - \hat{\theta})^2\right)$，即关于 θ 的正态核。
- 误差棒 $\Delta\hat{\theta} = \sqrt{1/(nI(\hat{\theta}))}$ 在大样本下对称。

例：指数分布的平均寿命估计

- $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$，样本 $t_1, \ldots, t_N$。

例：指数分布的平均寿命估计

- $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$，样本 $t_1, \ldots, t_N$。
- MLE：$\hat{\tau} = \bar{t}$。

例：指数分布的平均寿命估计

- $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$，样本 $t_1, \ldots, t_N$。
- MLE：$\hat{\tau} = \bar{t}$。
- 有限 N 时，似然函数不对称，可由 $\ell(\tau) - \ell(\hat{\tau}) = \frac{1}{2}$ 准则构造非对称置信区间。

例：指数分布的平均寿命估计

- $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$，样本 $t_1, \ldots, t_N$。
- MLE：$\hat{\tau} = \bar{t}$。
- 有限 N 时，似然函数不对称，可由 $\ell(\tau) - \ell(\hat{\tau}) = \frac{1}{2}$ 准则构造非对称置信区间。
- 随着 N 增大，似然函数趋于对称，误差棒对称化。

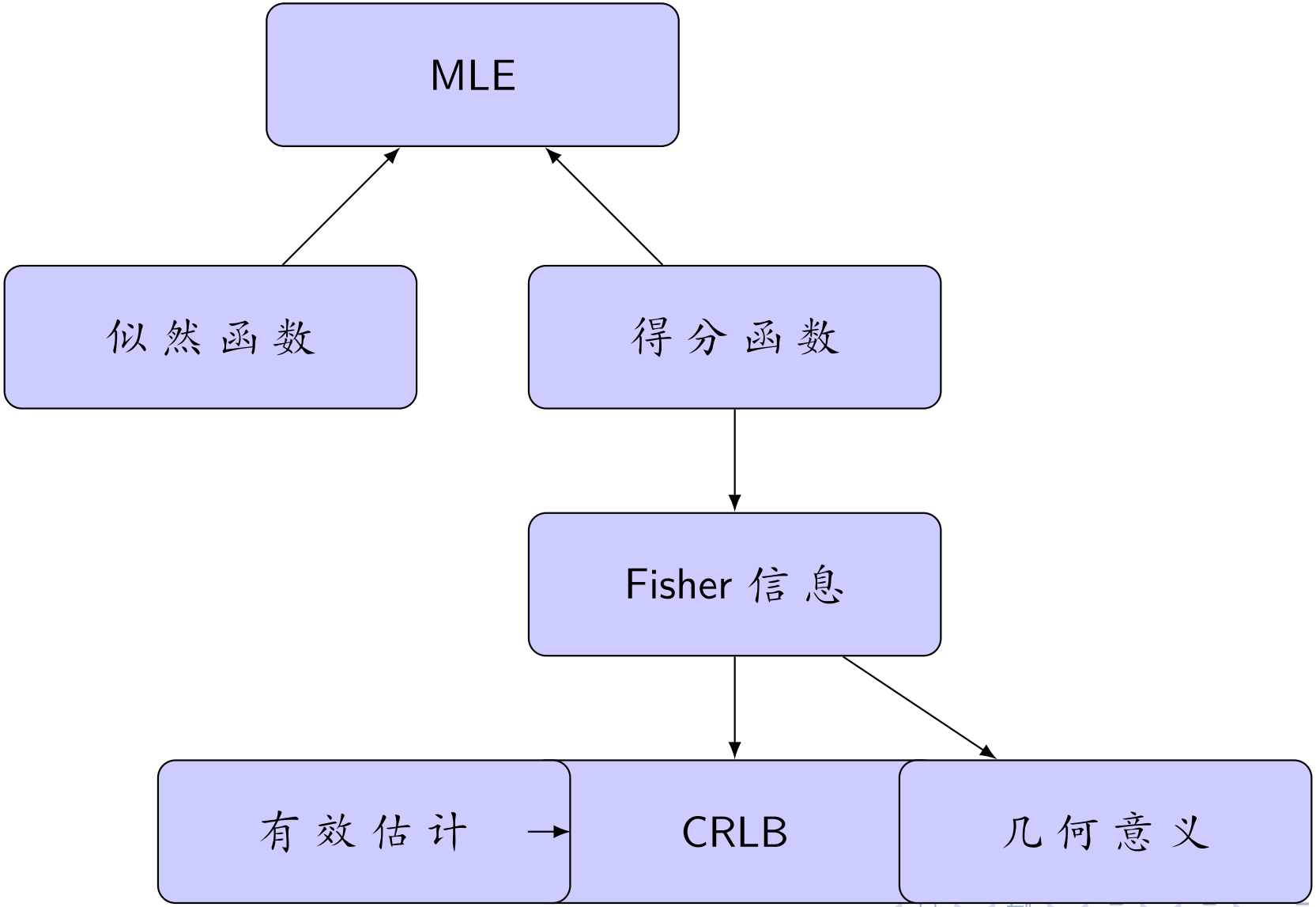
例：指数分布的平均寿命估计

- $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$，样本 $t_1, \ldots, t_N$。
- MLE：$\hat{\tau} = \bar{t}$。
- 有限 N 时，似然函数不对称，可由 $\ell(\tau) - \ell(\hat{\tau}) = \frac{1}{2}$ 准则构造非对称置信区间。
- 随着 N 增大，似然函数趋于对称，误差棒对称化。
- 启示：大样本下 *MLE* 的渐近正态性简化了区间估计。

核心概念回顾

- 最大似然估计：通过最大化似然函数得到参数估计，具有优良渐近性质。
- **Fisher** 信息：度量数据所含参数信息，是得分方差，也是对数似然的期望曲率。
- **Cramér–Rao** 下界：任何无偏估计的方差不低于 $1/I(\theta)$，有效估计达到该下界。
- 几何意义：Fisher 信息是统计流形上的黎曼度量，连接统计与微分几何。
- 应用：Jeffreys 先验、自然梯度法、实验设计等。

概念关系图



应用与拓展

- 实验设计：最大化 Fisher 信息以优化实验方案。

应用与拓展

- 实验设计：最大化 Fisher 信息以优化实验方案。
- 机器学习：自然梯度法加速神经网络训练；Fisher 信息用于确定参数重要性。

应用与拓展

- 实验设计：最大化 Fisher 信息以优化实验方案。
- 机器学习：自然梯度法加速神经网络训练；Fisher 信息用于确定参数重要性。
- 贝叶斯统计：Jeffreys 先验提供参数化不变的无信息先验。

应用与拓展

- 实验设计：最大化 Fisher 信息以优化实验方案。
- 机器学习：自然梯度法加速神经网络训练；Fisher 信息用于确定参数重要性。
- 贝叶斯统计：Jeffreys 先验提供参数化不变的无信息先验。
- 信息几何：将统计模型视为流形，发展出自然梯度、高效优化算法。

应用与拓展

- 实验设计：最大化 Fisher 信息以优化实验方案。
- 机器学习：自然梯度法加速神经网络训练；Fisher 信息用于确定参数重要性。
- 贝叶斯统计：Jeffreys 先验提供参数化不变的无信息先验。
- 信息几何：将统计模型视为流形，发展出自然梯度、高效优化算法。
- 量子估计：量子 Fisher 信息在量子度量学中刻画参数估计精度极限。

总结

- 最大似然法提供了具有优良性质的估计量：一致性、渐近正态性、渐近有效性。

总结

- 最大似然法提供了具有优良性质的估计量：一致性、渐近正态性、渐近有效性。
- Fisher 信息通过 CRLB 量化了估计量的精度极限，其几何解释深化了我们对统计模型的理解。

总结

- 最大似然法提供了具有优良性质的估计量：一致性、渐近正态性、渐近有效性。
- Fisher 信息通过 CRLB 量化了估计量的精度极限，其几何解释深化了我们对统计模型的理解。
- Fisher 信息不仅是理论工具，也在现代统计学习和优化中扮演关键角色。

总结

- 最大似然法提供了具有优良性质的估计量：一致性、渐近正态性、渐近有效性。
- Fisher 信息通过 CRLB 量化了估计量的精度极限，其几何解释深化了我们对统计模型的理解。
- Fisher 信息不仅是理论工具，也在现代统计学习和优化中扮演关键角色。
- 本报告从基础概念出发，逐步深入到几何视角和应用前景，希望帮助听众形成完整的知识图景。