

粒子物理与核物理实验中的 数据分析

王喆

清华大学

第五章：参数估计的一般概念

本讲要点

- 估计量
- 样本均值
- 样本方差
- 样本协方差

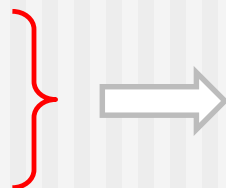
再论统计分析的目标

既包含检验假设，
也包含对假设中可能存在的参数进行估计。

实际问题往往是：

有限样本

参数本身不是直接观测量



如何给出参数的
最佳估计(包括均值、
方差和协方差等)

1. 如何构造参数估计量问题
2. 对参数估计量的评估问题
3. 对参数的有效估计问题

参数估计

概率密度函数的参数是表征其形状的常数，例如

$$f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

随机变量 参数

假设对随机变量 x 的 n 次独立观测得到样本 $\vec{x} = (x_1, \dots, x_n)$

我们希望找到数据样本的适当函数，用来估计参数 θ ：

估计量：

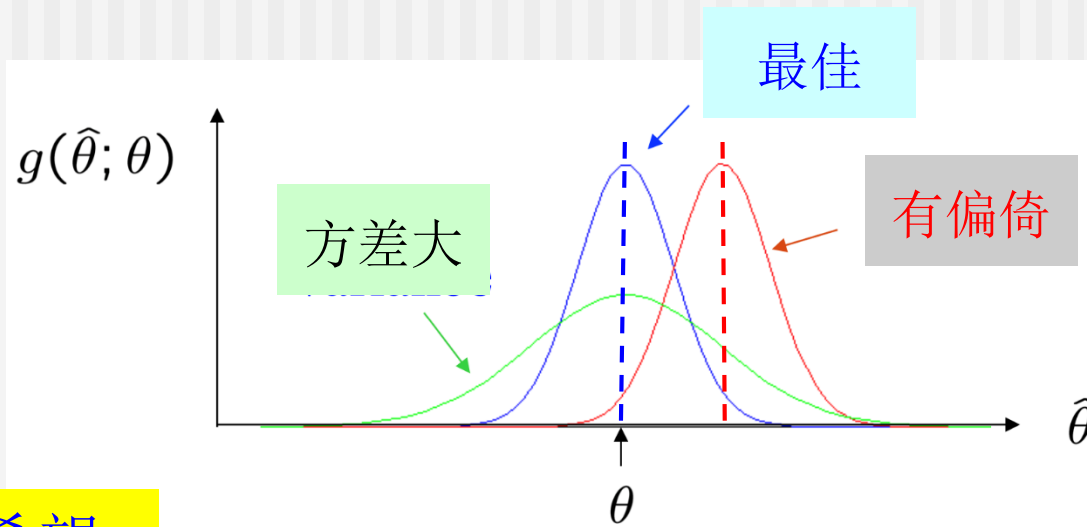
$$\hat{\theta}(\vec{x})$$

估计量用符号“ $\hat{}$ ”表示

估计量是数据样本的函数；对于给定数据样本，估计量的结果称为估计值

估计量的性质

重复整个测量，每次得到的估计值将服从某个分布 $g(\hat{\theta}; \theta)$



我们希望：

偏倚为零或很小（系统不确定度）： $b = E[\hat{\theta}] - \theta$

➔ 多次重复测量的均值应当趋于真值

方差小（统计不确定度小）： $v[\hat{\theta}]$

➔ 偏倚小和方差小通常是相互矛盾的要求

估计量好坏的标准

相合性（一致性）

$$\lim_{n \rightarrow \infty} \hat{\theta} = \theta$$
$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0, \forall \varepsilon > 0 \text{ 成立}$$

偏倚大小（无偏性）

$$b = E[\hat{\theta}] - \theta = 0$$

方差大小（有效性）

对任何估计量 $\hat{\theta}'$ ，都有 $\lim_{n \rightarrow \infty} \frac{V[\hat{\theta}_n]}{V[\hat{\theta}'_n]} \leq 1$ ，则 $\hat{\theta}$ 为有效估计量

参数估计量的构造与收敛性

如何构造参数 θ 的估计量 $\hat{\theta}(\vec{x})$?

没有一个完美的办法

首先是要求相合性 (consistency)

$$\lim_{n \rightarrow \infty} \hat{\theta} = \theta$$

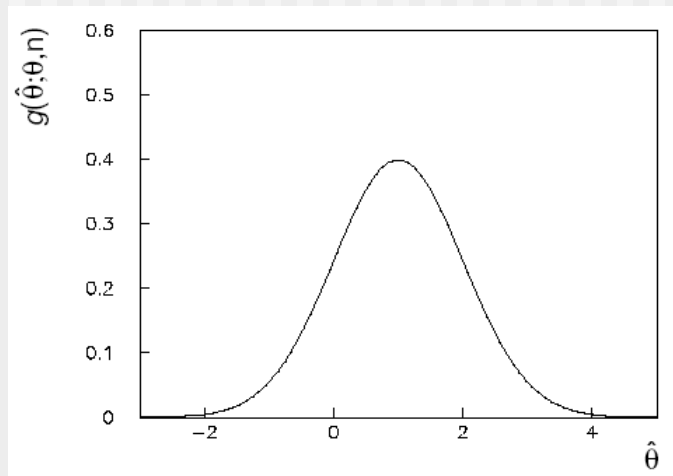
例如，随着样本容量的增大，估计值收敛于真值：

$$\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

注意：在统计意义上的收敛并不保证不会有个别特殊的 $\hat{\theta}_{\text{obs}}$ 与 θ 真值有较大偏离

估计量的偏倚问题

考虑样本容量 n 固定的估计量 $\hat{\theta}$ ，其概率密度函数 $g(\hat{\theta}; \theta, n)$



我们不知道 θ 真值，只能得到 $\hat{\theta}_{\text{obs}}$ 值

$g(\hat{\theta}; \theta, n)$ 的属性包括：

方差 $V[\hat{\theta}] = \sigma_{\hat{\theta}}^2$ ($\sigma_{\hat{\theta}}$: 统计不确定度)

偏离 $b = E[\hat{\theta}] - \theta$ (系统不确定度)

对于大统计量



$$\sigma_{\hat{\theta}} \propto \frac{1}{\sqrt{n}}, b \propto \frac{1}{n}$$

均方差

为了衡量估计量的好坏，会考虑其与真值 θ 的均方误差 (mean square error, MSE)

$$\begin{aligned}\text{MSE} &= E[(\hat{\theta} - \theta)^2] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - 2E[\hat{\theta}E[\hat{\theta}]] + (E[\hat{\theta}])^2 + (E[\hat{\theta}])^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta} - \theta])^2 \\ &= V[\hat{\theta}] + b^2\end{aligned}$$

➡ 通常要求无偏估计量中，对应的方差达到最小。

有效估计量

“有效估计量”：无偏估计量中方差最小的估计量

有时方差与偏离两者可以同时取得

有时这两者无法同时满足，
需要在方差与偏离之间存在平衡点

无参数的样本均值与样本方差估计

考虑对随机变量 x 作 n 次独立测量

➡ 样本大小为 n

等效为对 n 维矢量 $\vec{x} = (x_1, \dots, x_n)$ 作单次独立实验

由于 x_i 相互独立，所以样本的联合概率密度函数可表示为：

$$f_{\text{sample}}(\vec{x}) = f(x_1)f(x_2) \cdots f(x_n)$$

任务：从数据样本推断 $f(x)$ 的属性



构造数据样本的函数，以便估计 $f(x)$ 的各种属性，包括均值，方差，等等

均值的估计量与弱大数定理

考虑物理量 x 的 n 个测量 $x_1, \dots, x_n$ ，我们不知道对应的 pdf，想构造一个 x_i 的函数来估计 x 的均值，一个可能为

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (\text{样本均值})$$

如果 $V[x]$ 有限， $\bar{x}$ 则是一个与 μ 相合的估计量，即

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n x_i - \mu\right| \geq \varepsilon\right) = 0 \quad \text{弱大数定理}$$

计算期待值

$$E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

➡ $\bar{x}$ 是 μ 的无偏估计量。

问题： $\bar{x}$ 的方差是多少？

样本均值的评估量（方差）

对样本均值可靠程度的评估可以用样本均值的方差来估计，

$$\begin{aligned} V[\bar{x}] &= E[(E[\bar{x}] - \bar{x})^2] = E[\bar{x}^2] - (E[\bar{x}])^2 \\ &= E\left[\left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{j=1}^n x_j\right)\right] - \mu^2 \\ &= \frac{1}{n^2} \sum_{i,j=1}^n E[x_i x_j] - \mu^2 \\ &= \frac{1}{n^2} [(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)] - \mu^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

这里 σ^2 是 x 的方差，并利用了 $i \neq j$ 时

$$E[x_i x_j] = E[x_i]E[x_j] = \mu^2 \qquad E[x_i^2] = \mu^2 + \sigma^2$$

例：均值的测量精度

丁肇中在发现 J/ψ 粒子的实验中观测到 25 个 $J/\psi \rightarrow e^+e^-$ 事例，装置的质量测量精度 $\Delta m/m = 1\%$ ，质量分布的平均质量是 3.1 GeV。

装置的质量测量精度

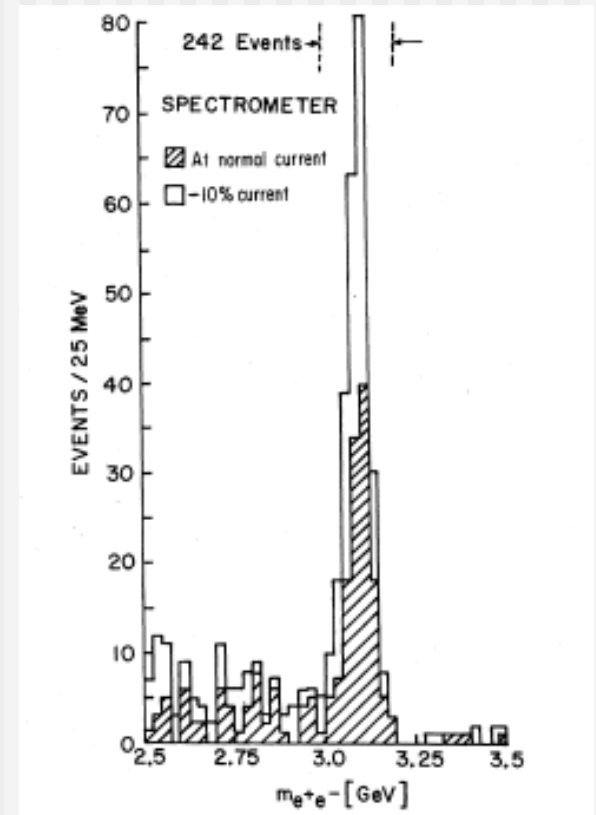
$$\sigma_m = 3100 \times 1\% = 31 \text{ MeV}$$

如果测量的质量分布与装置的测量精度相同，还可以得到

$$\sigma_m' = 3100 \times 1\% \times \frac{1}{\sqrt{25}} = 6.2 \text{ MeV}$$

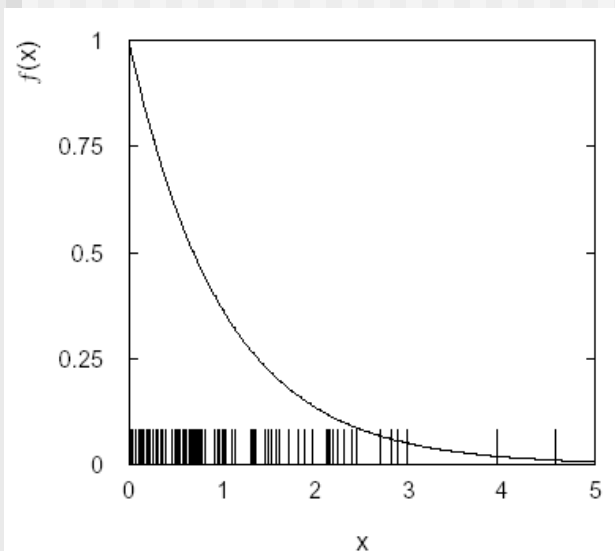
问题：结果应该报告为

3100 $\pm$ 31 MeV 还是 3100 $\pm$ 6 MeV?



Phys. Rev. Lett. 33, 1404 (1974)

例：均值的估计量与评估量

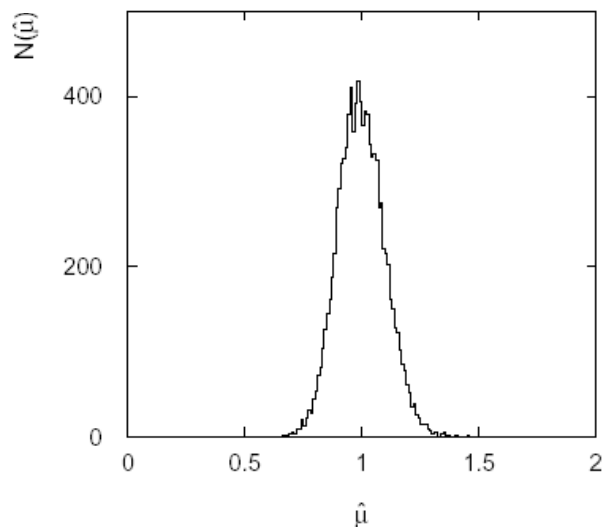


指数分布的蒙特卡罗样本 $n = 100$ ，寿命真实值为 $\mu = 1$ 。

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 1.073$$

重复 10^4 次，每次样本容量都是 $n = 100$ ，把每个样本的样本均值填入直方图

根据中心极限定理， $\hat{\mu}$ 近似服从高斯分布



$\bar{\hat{\mu}} = 0.9981$ ($\hat{\mu}$ 无偏)

$\hat{\mu}$ 的样本标准差为 $0.0995 \approx \sigma/\sqrt{n}$

方差的估计量

假如均值 μ 和方差 $V[x] = \sigma^2$ 都是未知量，样本方差定义为

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} (\overline{x^2} - \bar{x}^2)$$

因子 $\frac{1}{n-1}$ 保证 s^2 无偏，
即 $E[s^2] = \sigma^2$ 。

假如 $\mu = E[x]$ 先验已知（例如某种假设的预期值），则

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = \overline{x^2} - \mu^2$$

s^2 和 S^2 都是 σ^2 的无偏估计量

s^2 的方差

$$V[s^2] = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2^2 \right)$$

可以利用下式可估计 μ_k

$$\mu_k = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^k$$

这里 μ_k 是第 k 阶中心矩，
例如， $\mu_2 = \sigma^2$

协方差与相关系数的估计量

协方差 $V_{xy} = \text{cov}[x, y]$ 的估计量为

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1} (\overline{xy} - \bar{x} \cdot \bar{y}) \quad (\text{无偏})$$

相关系数 $\rho = V_{xy}/(\sigma_x \sigma_y)$ 的估计量为

$$\hat{\rho} = r = \frac{\hat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{k=1}^n (y_k - \bar{y})^2 \right)^{1/2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

r 有偏离。但是当 $n \rightarrow \infty$ 时，该偏倚趋于零。

一般而言，概率密度 $g(r; \rho, n)$ 形式复杂；对于高斯变量 x, y

$$E[r] = \rho - \frac{\rho(1 - \rho^2)}{2n} + \mathcal{O}(n^{-2}); \quad V[r] = \frac{1}{n}(1 - \rho^2)^2 + \mathcal{O}(n^{-2})$$

总结

➤ 估计量

估计量是数据样本的函数 $\hat{\theta}(\vec{x})$

评估估计量的好坏：相合性、偏倚性、有效性

➤ 样本均值

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

➤ 样本方差

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

➤ 样本协方差

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

延续下来的相关系数估计有偏