

粒子物理与核物理实验中的 数据分析

王喆

清华大学

第六章：
极大似然法

本讲要点

- ❑ 似然函数，最大似然估计量
- ❑ 最大似然估计量的方差估计
- ❑ 分区间数据拟合
- ❑ 最大似然法的拟合优度检验
- ❑ 不等精度观测结果的并合
- ❑ 最大似然法与贝叶斯估计

参数估计与最大似然法

前面给出了在不知道概率密度函数 pdf 的情况下，如果没有未知参数，如何从有限的数据样本中，估计出随机变量的期待值、方差与相关系数。

如果已经知道概率密度函数 pdf 的具体形式，但是函数中包含有未知参数，如何从有限的样本中估计未知参数的期待值、方差与相关系数...

最大似然法

含参数情况下如何估计参数

通常先给出 $f(x)$ 的假设形式，其中包含未知参数 θ

→ 利用给定的 $f(x, \theta)$ 形式和数据样本估计参数 θ

统计量(statistic)：数据样本的函数（不含未知参数）

估计量(estimator)：用来估计pdf某些属性的统计量

记号： θ 的估计量为 $\hat{\theta}$

估计值(estimate)：估计量的观测值，通常记为 $\hat{\theta}_{\text{obs}}$

参数拟合：利用 x 的数据样本估计参数 θ 的过程

每次实验给出的估计量 $\hat{\theta}$ 是满足概率密度 $g(\hat{\theta}; \theta)$ 的随机变量。
估计量 $\hat{\theta}$ 的均值为

$$E[\hat{\theta}(\vec{x})] = \int \hat{\theta} g(\hat{\theta}; \theta) d\hat{\theta} = \int \cdots \int \hat{\theta}(\vec{x}) f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n$$

参数估计量好坏的标准

符合程度(一致性)

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

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

偏置大小(无偏性)

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

方差大小(有效性)

$$\text{对任何估计量 } \hat{\theta}', \text{ 都有 } \lim_{n \rightarrow \infty} \frac{V[\hat{\theta}_n]}{V[\hat{\theta}'_n]} \leq 1, \text{ 则 } \hat{\theta} \text{ 渐进效果好。}$$

参数估计与概率大小的关系

考虑有数据样本 $\vec{x} = (x_1, \dots, x_n)$, 这里 x 服从pdf分布 $f(x; \theta)$ 。

目标: 估计 θ 。或者更为一般地, 估计 $\vec{\theta} = (\theta_1, \dots, \theta_m)$

如果 $f(x; \theta)$ 为真, 则有

$$P(\text{对所有在 } [x_i, x_i + dx_i] \text{ 观察到的 } x_i) = \prod_{i=1}^n f(x_i, \theta) dx_i$$

如果假设(包括 θ 的取值)为真

➡ 可以预料会使观测结果具有高的概率。

如果假设的 θ 取值远离真值或假设不对

➡ 会使观测结果具有低的概率。

似然函数

根据参数好坏与概率大小的关系，可以认为真实的 θ 应使得下式定义的似然函数

$$L(\theta) = \prod_{i=1}^n f(x_i, \theta)$$

注意：虽然 $L(\theta) = f_{\text{sample}}(\bar{x}; \theta)$ ，但是 $L(\theta)$ 只是 θ 的函数。这是因为在实验完成以后， $\bar{x}$ 就可以被当做常数。

有大的数值。

在经典统计理论里， $L(\theta)$ 并不是 θ 的概率密度函数。

θ 不是一个随机变量，但 $\hat{\theta}$ 却是。

在贝叶斯统计理论里，把 $L(\theta) = L(\vec{x}|\theta)$ 看作给定 θ 情况下， $\vec{x}$ 的概率密度函数，然后利用贝叶斯定理得到验后概率密度函数 $p(\theta|\vec{x})$ 。

最大似然估计量

定义最大似然估计量 $\hat{\theta}$ 为使得 $L(\theta)$ 最大的 θ 值。通过解下列方程

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, m$$

通常可以找到对于 m 个参数的解 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 。

有时候 $L(\theta)$ 可以有好几个极大值



取最大值

注意，**1)** 该方法利用了所有信息，与如何划分数据分布区间无关；
2) 定义的最大似然估计量并不保证它们总是最优的。

 需要对诸如无偏性，有效性等问题进行研究

多数情况下对于足够大样本，最大似然法的确能给出了期待的好结果。

即使是小样本的情况，虽然并不总是达到最优，但它通常仍然能给出最好的实用解。

最大似然估计量的参数选取

考虑 θ 的最大似然估计值是下列方程的解

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

对数形式(对同样的参数值, 该定义不会改变最大值的位置)。

如果选用另一个等价参数 $h(\theta)$, 则 h 的最大似然估计值是下列方程的解

$$\frac{\partial \log L(\theta)}{\partial h(\theta)} = 0$$

因为

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \log L(\theta)}{\partial h} \frac{\partial h}{\partial \theta}$$

因此, 只要 $\frac{\partial h}{\partial \theta} \neq 0$, 就有

$$\left. \frac{\partial \log L(\theta)}{\partial h(\theta)} \right|_{\theta=\hat{\theta}} = \left. \frac{\partial \log L(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

因此, L 的最大似然估计值与参数选取无关, 具有唯一性。

$$\longrightarrow \hat{h} = h(\hat{\theta})$$

例子:指数概率密度函数参数

考虑指数概率密度函数

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

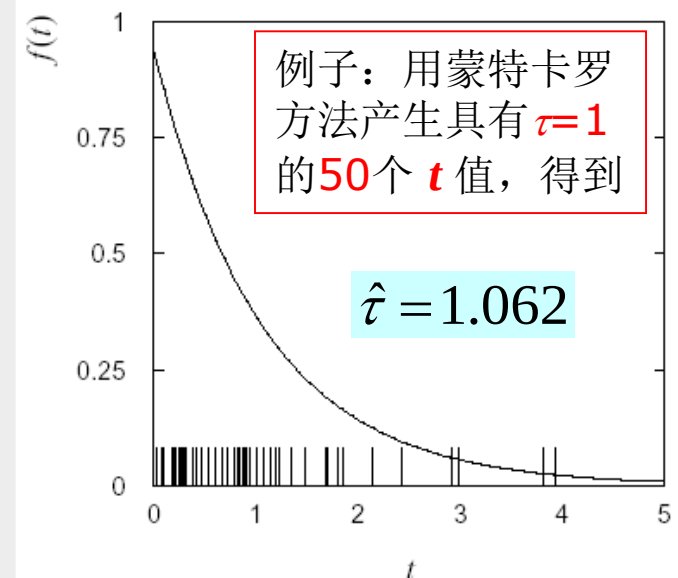
并假设有一数据样本 $t_1, \dots, t_n$ 。通常为了方便起见, 可采用

$$\log L(\tau) = \sum_{i=1}^n \log f(t_i; \tau) = \sum_{i=1}^n \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

$$\text{令 } \frac{\partial \log L}{\partial \tau} = 0, \text{ 并求解 } \tau,$$

$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

是平均寿命的
最大似然估计



指数型分布平均值的偏向性问题

对样本求平均

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

所得到的平均值是 τ 的一个无偏估计量吗？

原则上可以通过找出概率密度函数(例如采用解析或蒙特卡罗方法)

$$g(\hat{\tau}; \tau)$$

并计算出偏置的大小

$$b = E[\hat{\tau}] - \tau$$



来检查估计量是否是无偏的

指数型最大似然估计是无偏的

一种方法是解析计算 $E[\hat{\tau}]$

$$\begin{aligned} E[\hat{\tau}(t_1, \dots, t_n)] &= \int \dots \int \hat{\tau}(\vec{t}) f_{\text{joint}}(\vec{t}; \tau) dt_1 \dots dt_n \\ &= \int \dots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \dots dt_n \\ &= \frac{1}{n} \sum_{i=1}^n \left(\int t_i \frac{1}{\tau} e^{-t_i/\tau} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-t_j/\tau} dt_j \right) \\ &= \frac{1}{n} \sum_{i=1}^n \tau \\ &= \tau \end{aligned}$$

因此， $\hat{\tau}$ 是 τ 的一个无偏估计量。

指数型分布量的倒数情况

假设指数概率密度函数可写为

$$f(t; \lambda) = \lambda e^{-\lambda t}$$

这里 $\lambda = 1/\tau$ ，是衰变常数或寿命的倒数。

在测量值是时间 t 的情况下，如何找出 λ 的最大似然估计量？

根据最大似然估计量的唯一性，当 λ 是 τ 的等价参数时，

使 $L_\lambda(\lambda)$ 达到最大值的是 $\lambda(\hat{\tau})$ ，变量 $\hat{\tau}$ 也同时使 $L_\tau(\tau)$ 达到最大。



函数 $\lambda(\theta)$ 的最大似然估计量是 $\hat{\lambda} = \lambda(\hat{\tau})$

指数型分布量倒数的有偏估计

所以，对于衰变常数，有

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^{-1}$$

那么 $\hat{\lambda}$ 是 λ 的一个无偏估计量吗？

一般而言，一个无偏估计量的非线性函数对参数的函数是有偏向性的。

对于 $\hat{\lambda}$ ，利用特殊函数处理方法可以证明

$$E[\hat{\lambda}] = \lambda \frac{n}{n-1}$$

→ $\hat{\lambda}$ 有偏置。但当 $n \rightarrow \infty$ 时，该偏置将趋于零。

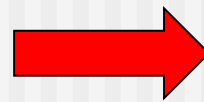
高斯概率密度函数中的参数

考虑一个样本服从高斯概率密度函数，其参数 μ , σ^2 未知。
其对数似然函数为

$$\begin{aligned}\log L(\mu, \sigma^2) &= \sum_{i=1}^n \log f(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^n \left(\log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \log \sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right)\end{aligned}$$

对 μ , σ^2 偏微分后的函数取零，解方程得到

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0$$



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

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^4} = 0$$



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

取“帽”号表示方程的解是参数的估计值。

高斯函数参数估计值的偏向性

前面已证明 $\hat{\mu}$ 是 μ 的无偏估计量。计算 $\widehat{\sigma^2}$ 的平均值得到

$$E[\widehat{\sigma^2}] = \frac{n-1}{n} \sigma^2$$

利用

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

$$E[x_i x_j] = \mu^2 \quad i \neq j$$

因此， σ^2 的最大似然估计量有偏向性。这种偏向性随 n 趋于无穷大时而消失。但是方差的统计估计量

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

对任何概率密度函数的方差估计都是无偏的。
但是最大似然法给不出这样的估计量。

估计量的方差：解析法

指数分布平均值的估计量为：

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

对应的概率密度函数 $g(\hat{\tau}; \tau, n)$ 分布的宽度可以估计

$$\begin{aligned} V[\hat{\tau}] &= E[\hat{\tau}^2] - (E[\hat{\tau}])^2 \\ &= \int \dots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^2 \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \dots dt_n \\ &\quad - \left(\int \dots \int \left(\frac{1}{n} \sum_{i=1}^n t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \dots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \dots dt_n \right)^2 \\ &= \frac{\tau^2}{n} \rightarrow \hat{\tau} \text{ 的方差比 } t \text{ 的方差小 } n \text{ 倍。} \end{aligned}$$

估计量的方差：解析法(续)

注意：对于未知真值 τ 的 $V[\hat{\tau}], \sigma_{\hat{\tau}}$ 函数，其估计可以采用

$$\hat{\sigma}_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

通常情况下，“统计误差”由上式给出。例如，

$$\hat{\tau} \pm \hat{\sigma}_{\hat{\tau}} = 1.062 \pm 0.150$$

这就意味着

最大似然法对 τ 的估计为1.062；

最大似然法对 $g(\hat{\tau}; \tau, n)$ 的 σ_{τ} 估计为0.150。

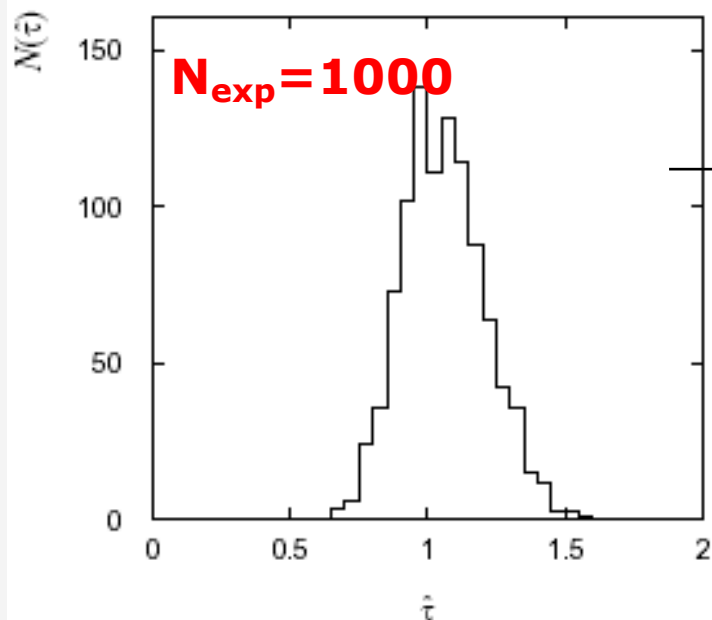
如果 $g(\hat{\tau}; \tau, n)$ 是高斯函数，则 $[\hat{\tau} - \hat{\sigma}_{\hat{\tau}}, \hat{\tau} + \hat{\sigma}_{\hat{\tau}}]$ 为所谓的
“68%的置信区间”

估计量的方差：蒙特卡罗方法

通常情况下， $\hat{\theta}$ 的具体形式 $g(\hat{\theta}; \theta, n)$ 并不知道。对于此类情况，

→ 可采用蒙特卡罗方法得到 $g(\hat{\theta}; \theta, n)$

例如，对指数的pdf，我们有 $\hat{\tau}=1.062$ 。在蒙特卡罗中，将其作为 τ 的真值。产生 $n=50$ 的样本，并重复1000次实验。计算每次实验的 $\hat{\tau}$ ，并填入直方图。



蒙特卡罗实验可给出标准偏差

$$s = 0.151$$

类似于前面估计的 $\hat{\tau} / \sqrt{n} = 0.150$ 。

注意： $g(\hat{\tau}; \tau, n)$ 近似服从高斯分布。根据中心极限定理，可以断定在大样本极限下它确实是高斯分布。

RCF边界问题(信息不等式)

任何估计量(不仅仅是最大似然法)的方差下界为

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right] \quad (b \text{ 为偏置})$$

也称为 Rao-Cramér-Frechet 不等式(信息不等式)。

如果等式满足, 就可以说 $\hat{\theta}$ 是有效的。



最大似然估计量对大的样本统计量 n 几乎总是有效的。

通常假设上述结论为真, 利用RCF边界估计 $V[\hat{\theta}]$

朱永生 P197, P244,
Frederick James P148

$$E(T) = \int \cdots \int T(X_1, \cdots, X_n) L(X_1, \cdots, X_n | \vartheta) dX_1 \cdots dX_n$$

$$= \tau(\vartheta) + b(\vartheta).$$

T是 θ 的一个函数，我们可以不直接估计 θ ，b是偏离，E是期待值

上式两侧对 θ 求导

$$\int \cdots \int T \frac{\partial L}{\partial \vartheta} dX = \int \cdots \int T \frac{\partial \ln L}{\partial \vartheta} \cdot L dX \quad (\text{A})$$

$$= \frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta},$$

再有概率归一化条件

$$\int \cdots \int L dX = 1$$

上式两侧对 θ 求导

$$\int \cdots \int \frac{\partial L}{\partial \vartheta} dX = \int \cdots \int \frac{\partial \ln L}{\partial \vartheta} L dX = 0 \quad (\text{B})$$

**(A) - (B)
 $\times (\tau(\theta) + b(\theta))$**

$$\int \cdots \int [T - \tau(\vartheta) - b(\vartheta)] \frac{\partial \ln L}{\partial \vartheta} L dX = \frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta}$$

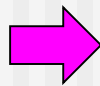
利用施瓦兹不等式: $V(X) \cdot V(Y) \geq [\text{cov}(X, Y)]^2$

$$\int \cdots \int [T - \tau(\vartheta) - b(\vartheta)] \frac{\partial \ln L}{\partial \vartheta} L dX = \frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta}$$

将这里的 $[T - \tau(\theta) - b(\theta)]$ 和 $\partial \ln L / \partial \theta$ 看做
两个随机变量

$$\int \cdots \int [T - \tau(\vartheta) - b(\vartheta)]^2 L dX \cdot \int \cdots \int \left(\frac{\partial \ln L}{\partial \vartheta} \right)^2 L dX \geq \left(\frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta} \right)^2$$

这里的第一项方差的
定义式—— $V(T)$ 和第
二项是 $\partial \ln L / \partial \theta$ 的期
待值



$$V(T) \geq \frac{\left(\frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta} \right)^2}{E \left[\left(\frac{\partial \ln L}{\partial \vartheta} \right)^2 \right]}$$

再将前面的
(B) 式对 θ
求导

$$\begin{aligned}
 0 &= \frac{\partial}{\partial \vartheta} \int \cdots \int \frac{\partial \ln L}{\partial \vartheta} L d\mathbf{X} = \int \cdots \int \frac{\partial}{\partial \vartheta} \left(\frac{\partial \ln L}{\partial \vartheta} \cdot L \right) d\mathbf{X} \\
 &= \int \cdots \int \left[\frac{\partial^2 \ln L}{\partial \vartheta^2} \cdot L + \frac{\partial L}{\partial \vartheta} \frac{\partial \ln L}{\partial \vartheta} \right] d\mathbf{X} \\
 &= \int \cdots \int \left[\frac{\partial^2 \ln L}{\partial \vartheta^2} + \left(\frac{\partial \ln L}{\partial \vartheta} \right)^2 \right] L d\mathbf{X} \\
 &= E \left(\frac{\partial^2 \ln L}{\partial \vartheta^2} \right) + E \left[\left(\frac{\partial \ln L}{\partial \vartheta} \right)^2 \right], \quad \text{注意这两项的关系, 带回}
 \end{aligned}$$

重写上页的结果, 得到更方便的形式, 即 RCF 不等式

$$V(T) \geq \frac{\left(\frac{\partial \tau}{\partial \vartheta} + \frac{\partial b}{\partial \vartheta} \right)^2}{E \left[-\frac{\partial^2 \ln L}{\partial \vartheta^2} \right]}$$

如果我们正在研究的问题中T就是 θ 自身

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / E \left[-\frac{\partial^2 \log L}{\partial \theta^2} \right]$$

如果b为零

$$V[\hat{\theta}] \geq 1/E \left[-\frac{\partial^2 \log L}{\partial \theta^2} \right]$$

这个式子是我们未来一些重要应用的“基础”

RCF边界问题(续一)

例如，对于前面的指数概率密度函数的例子，我们可以得到

$$\frac{\partial^2 \log L}{\partial \tau^2} = \frac{n}{\tau^2} \left(1 - \frac{2}{\tau} \frac{1}{n} \sum_{i=1}^n t_i \right) = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right)$$

已知 $b = 0$ ，所以

$$V[\hat{\tau}] \geq \frac{1}{E \left[-\frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \right]} = \frac{1}{-\frac{n}{\tau^2} \left(1 - \frac{2E[\hat{\tau}]}{\tau} \right)} = \frac{\tau^2}{n}$$

=真的方差



最大似然法的估计量 $\hat{\tau}$ 对任何
样本统计量大小 n 都是有效的。

对于 $\vec{\theta} = (\theta_1, \dots, \theta_m)$ 具有有效估计量以及零偏置的情况，

$$(V^{-1})_{ij} = E \left[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right] = -n \int f(x; \vec{\theta}) \frac{\partial^2 \log f(x; \vec{\theta})}{\partial \theta_i \partial \theta_j} dx$$

有效估计量的
方差正比于 $1/n$

RCF边界问题(续二)

在RCF边界里, $\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}$ 的期待值是参数真值的函数。可用下式估计, 即用最佳拟合值处的结果来代表

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta} = \hat{\theta}}$$

对于只有一个参数的情况, 可以得到

$$\widehat{\sigma^2}_{\hat{\theta}} = \left(\frac{-1}{\frac{\partial^2 \log L}{\partial \theta^2}} \right) \bigg|_{\theta = \hat{\theta}}$$

通常求 $\log L$ 的最大值是通过数值计算来完成, 二阶导数的矩阵(也就是Hessian矩阵)是通过有限差值来估计。

➡ 调用 CERN 的 MINUIT 软件包中的 HESSE 程序

例子：估计实验所需的统计量

假设我们想通过质子与反质子弹性散射来研究反质子的极化，观测量为散射角 $x = \cos\phi$ ，已知对应的概率密度函数为 $f(x; \alpha) = \frac{1}{2}(1 + \alpha x)$ ，其中 α 是反映反质子极化的参数。如果过去的实验已经测量出 α 在 0.10 ± 0.02 左右(即 α 估计值的相对误差为 20%)，那么要想在统计上将相对误差减少到 5%，总共需要多少个事例？

由信息不等式，任何估计量的方差下界为

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right]$$

对于本题， $b=0$ ，

$$L(\alpha) = \prod_{i=1}^n f(x_i, \alpha) = \left(\frac{1}{2}\right)^n \prod_{i=1}^n (1 + \alpha x_i)$$

例子：实验所需的统计量（续）

$$\frac{\partial^2 \log L}{\partial \alpha^2} = \frac{\partial^2}{\partial \alpha^2} \left[n \log \frac{1}{2} + \sum_{i=1}^n \log(1 + \alpha x_i) \right] = \frac{\partial}{\partial \alpha} \left[\sum_{i=1}^n \frac{x_i}{1 + \alpha x_i} \right] = - \sum_{i=1}^n \frac{x_i^2}{(1 + \alpha x_i)^2}$$

$$\begin{aligned} E \left[- \frac{\partial^2 \log L}{\partial \alpha^2} \right] &= n \int_{-1}^1 \frac{x^2}{(1 + \alpha x)^2} \frac{1}{2} (1 + \alpha x) dx = \frac{n}{2} \int_{-1}^1 \frac{x^2}{1 + \alpha x} dx \\ &= \frac{n}{2} \frac{1}{2\alpha^3} \left[(1 + \alpha x)^2 - 4(1 + \alpha x) + 2 \ln |1 + \alpha x| \right] \Big|_{-1}^1 \\ &= \frac{n}{2\alpha^3} \left[\ln \left| \frac{1 + \alpha}{1 - \alpha} \right| - 2\alpha \right] \end{aligned}$$

由信息不等式，

$$V[\hat{\alpha}] \geq \frac{1}{\frac{n}{2\alpha^3} \left[\ln \left| \frac{1 + \alpha}{1 - \alpha} \right| - 2\alpha \right]} \quad \leftarrow \quad \begin{aligned} \hat{\alpha} &= 0.10 \\ V[\hat{\alpha}] &= 0.05^2 = 2.5 \times 10^{-5} \end{aligned}$$

带入不等式可以求得 $n \geq 1.2 \times 10^5$ 。

估计量的方差：图解法

考虑单参数 θ 情况下，将 $\log L(\theta)$ 在 $\hat{\theta}$ 附近展开，

$$\log L(\theta) = \log L(\hat{\theta}) + \left[\frac{\partial \log L}{\partial \theta} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!} \left[\frac{\partial^2 \log L}{\partial \theta^2} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots$$

令 $\log L(\hat{\theta}) = \log L_{\max}$ ，并且上式第二项为零，因此有

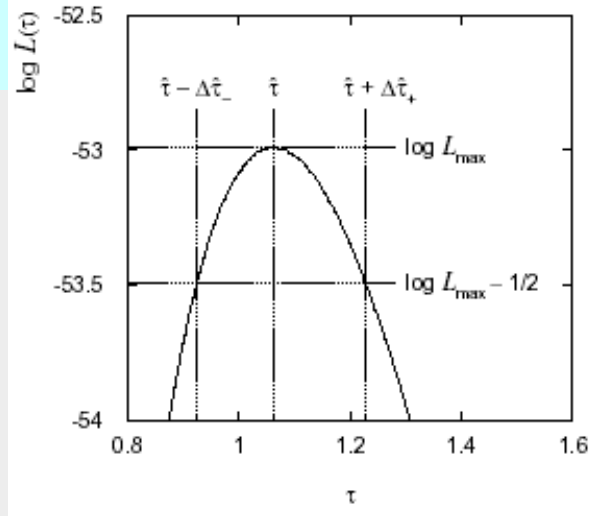
$$\log L(\theta) = \log L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}_{\hat{\theta}}^2}$$

也就是

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{\max} - \frac{1}{2}$$

为了得到 $\hat{\sigma}_{\hat{\theta}}$ ，可以让 θ

→ 偏离 $\hat{\theta}$ ，使得 $\log L$ 值减掉一个 $1/2$ 数值。



指数函数例子

$$\hat{\tau} = 1.062$$

$$\Delta \hat{\tau}_- = 0.137$$

$$\Delta \hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta \hat{\tau}_-$$

$$\approx \Delta \hat{\tau}_+ = 0.15$$

最大似然估计量的渐进正态性

如果 $\vec{x} = (x_1, \dots, x_n)$ 是分布 $f(x; \theta)$ 的随机样本， $\hat{\theta}$ 是参数 θ 的最大似然估计。则当样本容量 $n \rightarrow \infty$ 时， $\hat{\theta}$ 的分布趋近于一个正态分布，即

$$f(\hat{\theta}; \theta) = N(\hat{\theta}; \theta, V[\theta])$$

根据似然函数定义，极值问题处理和误差传递，方差可写为

$$V[\theta] = - \left(\frac{\partial^2 \log L(\theta)}{\partial^2 \theta} \right)^{-1} = - \frac{1}{n} \left(\frac{\partial^2 \log f(x; \theta)}{\partial^2 \theta} \right)^{-1}$$



在推断大样本的最大似然估计误差时，可以利用测量误差理论中最常见的正态分布进行推断。

注意：样本容量多大，才能近似利用极限正态分布，才可以看作最有效的估计，这将依赖于观测量的概率密度函数的具体形式。但对于指数型分布会有一些最优性质。

双参数拟合

由联合概率密度得到样本的似然函数

$$L(\vec{x}; \alpha, \beta) = \prod_{i=1}^{2000} f(x_i; \alpha, \beta)$$

用数值解找到 $\log L$ 的最大值

$$\hat{\alpha} = 0.50 \pm 0.05$$

$$\hat{\beta} = 0.51 \pm 0.11$$

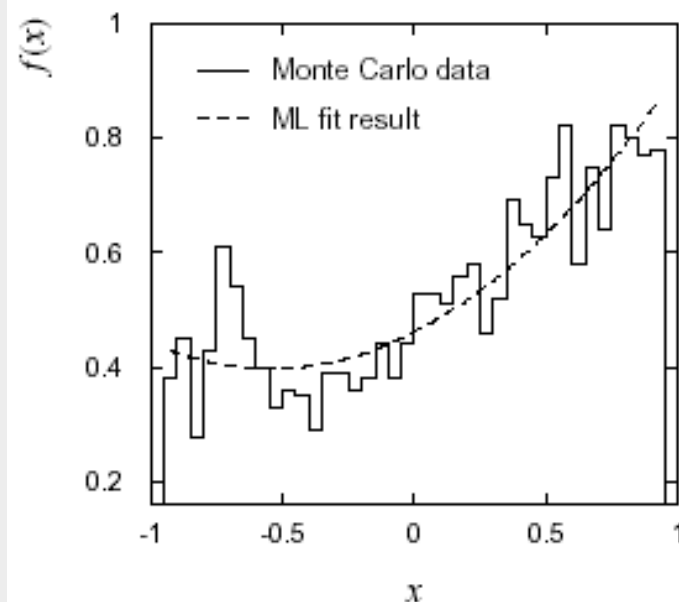
$$\widehat{\text{cov}}[\hat{\alpha}, \hat{\beta}]$$

$$= 0.0024$$

$$\hat{\rho} = r = 0.42$$

“误差”：

$$\hat{\sigma}_{\hat{\alpha}}, \hat{\sigma}_{\hat{\beta}}$$



注意：本例子中采用的是点拟合，与直方图的区间大小划分无关。

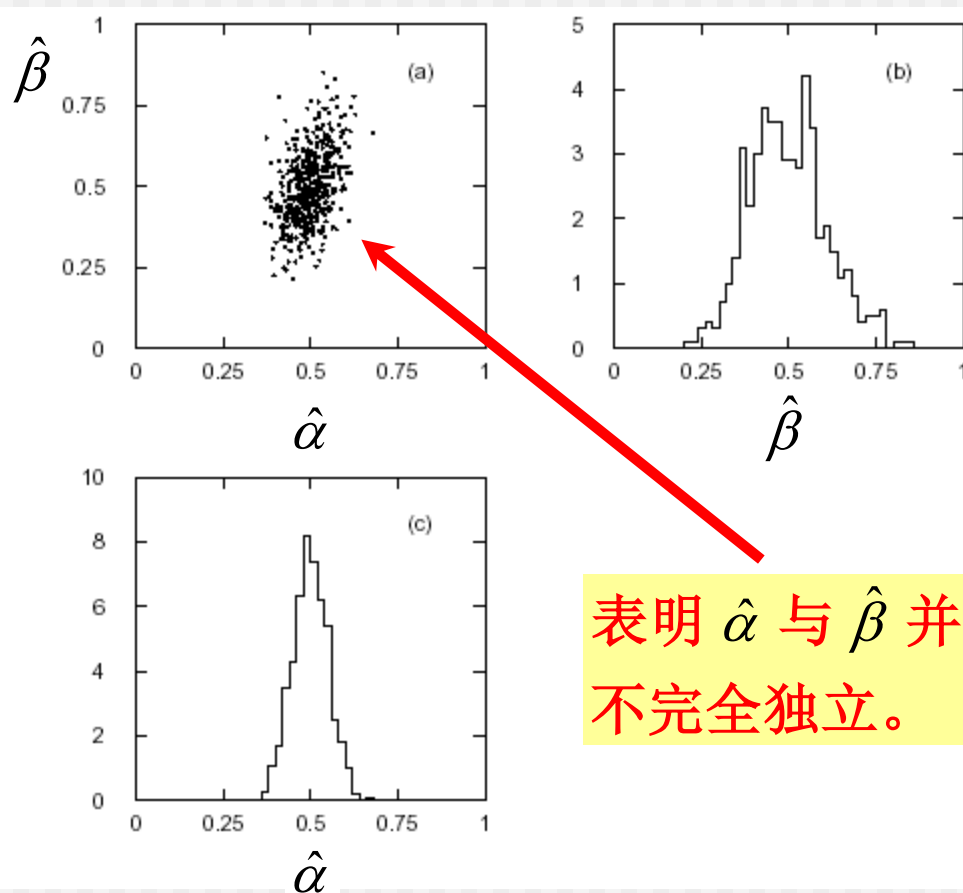
协方差

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta} = \vec{\hat{\theta}}}$$

(MINUIT HESSE)

双参数拟合结果的蒙特卡罗检验

做500次模拟实验，重复最大似然拟合，每次都是模拟 $n=2000$ 个事例，得到500组 $\hat{\alpha}$ 和 $\hat{\beta}$ 。



表明 $\hat{\alpha}$ 与 $\hat{\beta}$ 并不完全独立。

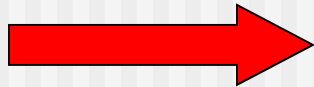
$$\begin{aligned}\overline{\hat{\alpha}} &= 0.499 & \overline{\hat{\beta}} &= 0.498 \\ s_{\hat{\alpha}} &= 0.051 & s_{\hat{\beta}} &= 0.111 \\ \overline{\text{cov}[\hat{\alpha}, \hat{\beta}]} &= 0.0024 \\ \hat{\rho} = r &= 0.42\end{aligned}$$

$\hat{\alpha}, \hat{\beta}$ 正相关

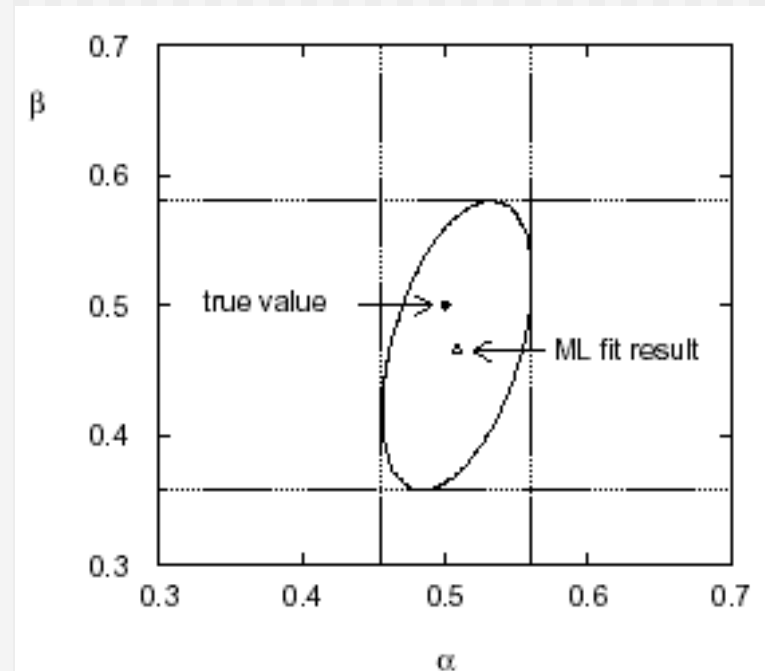
边缘概率密度函数近似为高斯分布。

双参数拟合的 $\log L$ 等高线形式

在前面蒙特卡罗样本中，其中的一次拟合结果在 α, β 平面表示为



对于大的样本容量 n ,
 $\log L$ 具有下列形式:



$$\log L(\alpha, \beta) = \log L_{\max}$$

$$-\frac{1}{2(1-\rho^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right]$$

$\log L$ 等高线对应的方差

等高线 $\log L(\alpha, \beta) = \log L_{\max} - 1/2$ 是由下式定义得到的

$$\frac{1}{1-\rho^2} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2 - 2\rho \left(\frac{\alpha - \hat{\alpha}}{\sigma_{\hat{\alpha}}} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_{\hat{\beta}}} \right) \right] = 1$$

其中，等高线的切线给出了标准偏差



$$\sigma_{\hat{\alpha}}, \sigma_{\hat{\beta}}$$

椭圆的倾角与相关系数有关



$$\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$$

注意：参数之间的相关性体现在对估计量的误差或方差方面的影响(或偏大或偏小)。

推广的最大似然法(名字不好)

到目前为止, 只考虑了**固定样本大小** n 的情形, 有时候, n 被看作泊松分布的**随机变量**, 平均值为 ν 。

→ 实验结果定义: $n, x_1, \dots, x_n$

则推广的似然函数为
$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \vec{\theta})$$

假设理论给出 $\nu = \nu(\vec{\theta})$, 去掉与 θ 无关的项, 有

$$\begin{aligned} \log L(\vec{\theta}) &= n \log \nu(\vec{\theta}) - \nu(\vec{\theta}) + \sum_{i=1}^n \log f(x_i; \vec{\theta}) \\ &= -\nu(\vec{\theta}) + \sum_{i=1}^n \log(\nu(\vec{\theta}) \cdot f(x_i; \vec{\theta})) \end{aligned}$$

推广的最大似然法： $\nu, \vec{\theta}$ 独立

假设 $\nu, \vec{\theta}$ 是在函数上相互独立的，

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \vec{\theta})$$

$$\frac{\partial L}{\partial \nu} = \left(\frac{n}{\nu} - 1 \right) \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^n f(x_i; \vec{\theta}) = 0 \quad \longrightarrow \hat{\nu} = n$$

$$\frac{\partial L}{\partial \theta_i} = 0 \quad \longrightarrow \text{与普通最大似然法确定 } \hat{\theta}_i \text{ 的要求一样}$$

适用于对处理 $f(x; \vec{\theta})$ 是已知分量叠加的情况。

例如：是信号与本底分量的叠加。

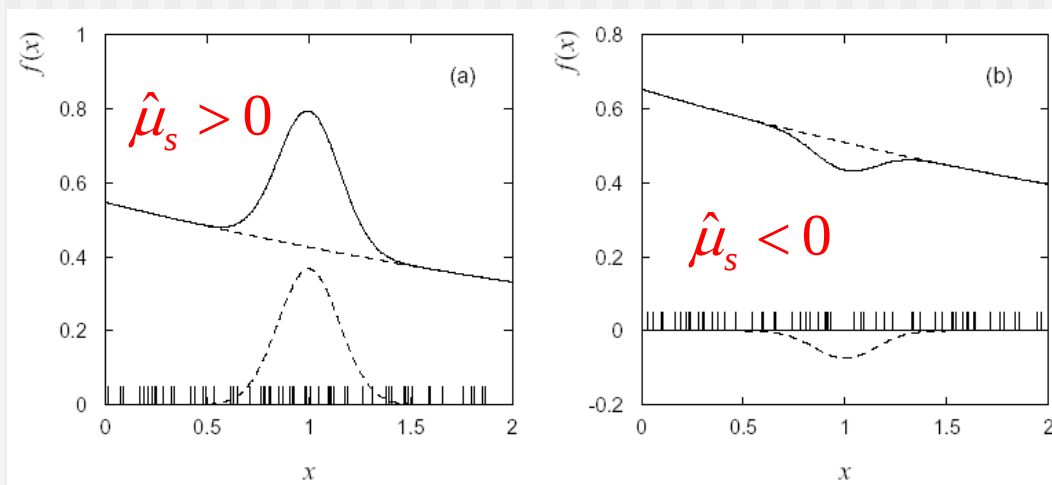
可以将其分解成单独求 ν 与 θ 估计值问题。

非物理结果问题

假设有两类事例：信号 (s) 与本底 (b)

$$f(x) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

假设 $f_s(x)$ 与 $f_b(x)$ 已知，需要估计 μ_s 与 μ_b ，



问题：如果出现负值，应该如何报告结果？



1. 报道负的 $\hat{\mu}_s$ (非物理结果！)？
2. 报道含偏置的真值估计量为： $\hat{\mu}_s^{physical} = \max(0, \hat{\mu}_s)$ ？

最大似然法处理分区数据

通常数据 $\vec{x}$ 在划分为 N 个区间直方图中的频数为

$$\vec{n} = (n_1, \dots, n_N), \quad n_{tot} = \sum_{i=1}^N n_i$$

通常称为“对直
方图拟合”。

在某种假设下，有期待值

$$\vec{v} = (v_1, \dots, v_N), \quad v_{tot} = \sum_{i=1}^N v_i(\vec{\theta})$$



$$v_i(\vec{\theta}) = v_{tot} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx$$

如果样本的联合概率密度函数(n_{tot} 为常数)为

$$f_{sample}(\vec{n}; \vec{v}) = \frac{n_{tot}!}{n_1! \dots n_N!} \left(\frac{v_1}{n_{tot}} \right)^{n_1} \dots \left(\frac{v_N}{n_{tot}} \right)^{n_N} \rightarrow \log L(\vec{\theta}) = \sum_{i=1}^N n_i \log v_i(\vec{\theta})$$

扩展的分区间数据拟合(名字不好)

n_{tot} 不为常数
样本的联合概率密度函数为

$$f_{sample}(\vec{n}; \vec{v}) = \frac{v_{tot}^{n_{tot}} e^{-v_{tot}}}{n_{tot}!} \frac{n_{tot}!}{n_1! \dots n_N!} \left(\frac{v_1}{n_{tot}} \right)^{n_1} \dots \left(\frac{v_N}{n_{tot}} \right)^{n_N}$$

如果 n_i 是泊松随机变量, 有推广的对数似然函数

$$\log L(v_{tot}, \vec{\theta}) = -v_{tot} + \sum_{i=1}^N n_i \log v_i(v_{tot}, \vec{\theta})$$

例子：指数分布

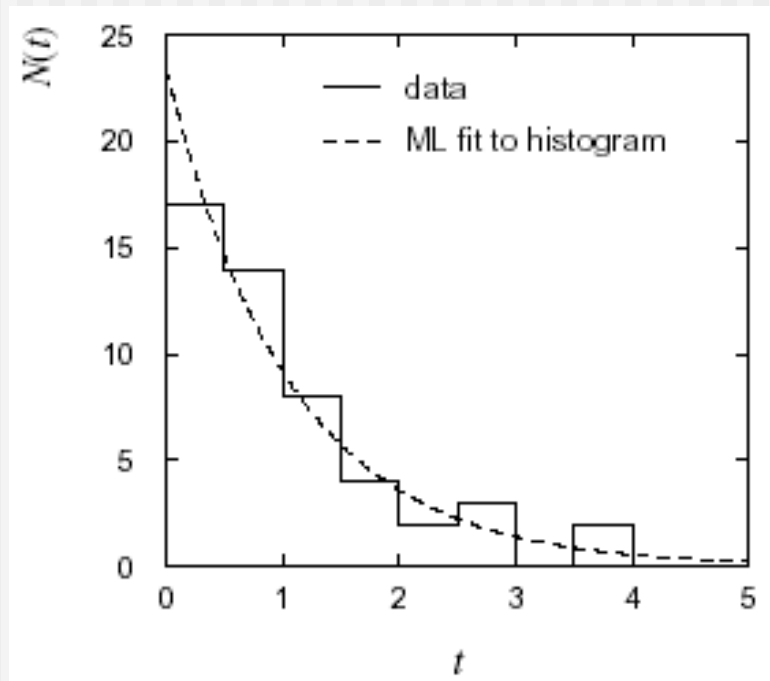
指数分布例子

直方图拟合: $\hat{\tau} = 1.07 \pm 0.17$

点估计: $\hat{\tau} = 1.06 \pm 0.15$



两者结果吻合。点估计结果误差较小。



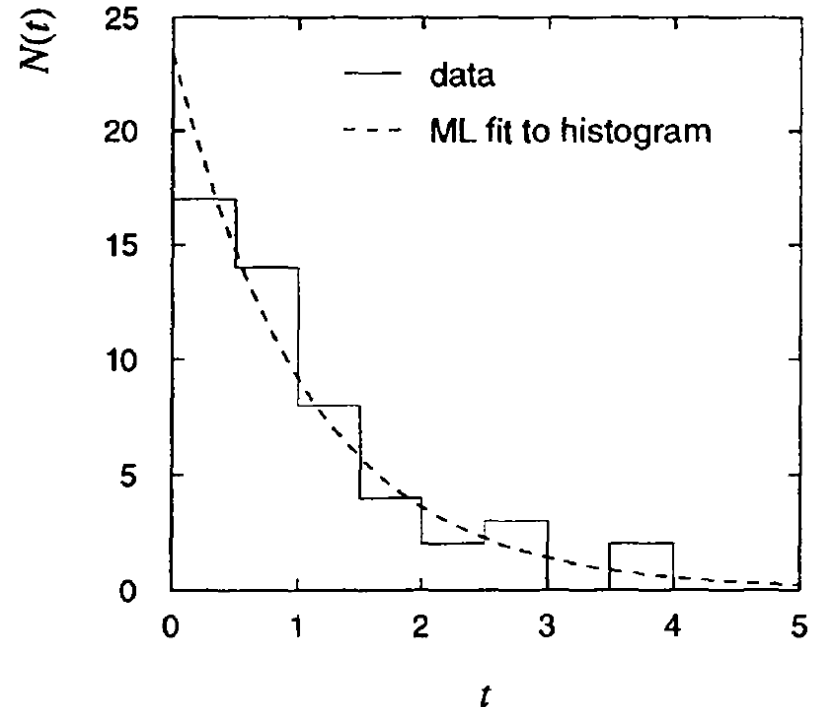
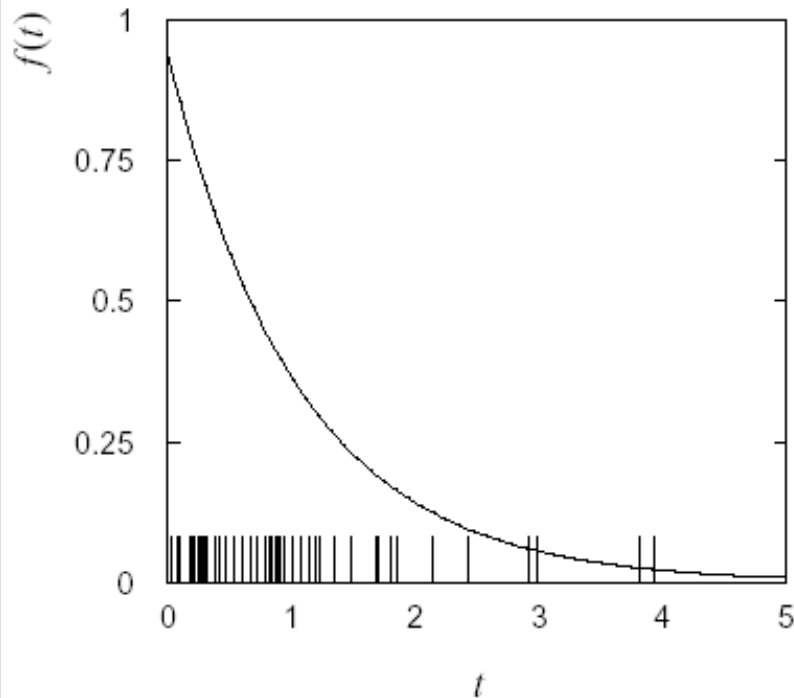
当区间宽度为零时,

直方图拟合



点估计

不分bin的数据拟合，与分bin的直方图拟合



由于一些情况，例如本底谱输入，只能拟合直方图，但由于分bin的原因，每个bin内的细节信息实际上是丢失了

分区处理数据的问题

分区处理数据有时会因区间宽度过大而造成部分信息丢失影响到参数的估计。例如

$$f(x; \alpha, \beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta / 3}$$

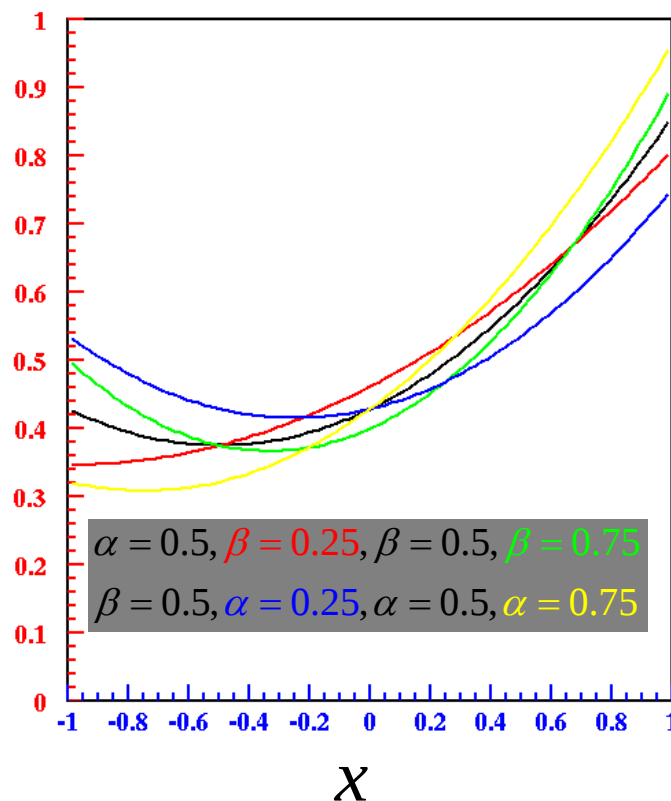
由于 β 对函数变化影响较大，过宽的分区，对 β 的估计影响较大。

例如，在 $x \in [-1, 1]$ 范围内分50个区间

$$\hat{\alpha} = 0.47 \pm 0.05, \hat{\beta} = 0.39 \pm 0.11 \quad \text{☹}$$

在 $x \in [-1, 1]$ 范围内分200个区间

$$\hat{\alpha} = 0.50 \pm 0.05, \hat{\beta} = 0.50 \pm 0.11 \quad \text{☺}$$

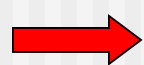


因此，对直方图拟合，一定要确认区间的大小对结果无明显影响。注意：区间无穷小时，与点估计结果一致。

检验最大似然法的拟合优度

对非分区(点估计)情况, 检验很困难。可以尝试检验统计量 $t = \log L_{\max}$

原则上可行, 但检验统计量的分布 $g(t)$ 通常未知



需要蒙特卡罗计算P-值

对分区或对拟合非分区数据后所做的直方图情况,

$$\hat{v}_i = n_{tot} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\hat{\theta}}) dx \quad (\text{假设有 } m \text{ 个拟合参数})$$

可以考虑研究比值的分布情况

$$\lambda = \frac{L(\vec{n} | \vec{v})}{L(\vec{n} | \vec{n})} = \frac{f_{\text{joint}}(\vec{n}; \vec{v})}{f_{\text{joint}}(\vec{n}; \vec{n})}$$

检验最大似然法的拟合优度(续)

例如对于满足多项式或泊松分布的随机变量 n_i , 可用检验统计量

$$\chi_M^2 = -2 \log \lambda_M = 2 \sum_{i=1}^N n_i \log \frac{n_i}{\hat{v}_i}$$

$$\chi_P^2 = -2 \log \lambda_P = 2 \sum_{i=1}^N \left(n_i \log \frac{n_i}{\hat{v}_i} + \hat{v}_i - n_i \right)$$

对于大统计量样本, 两者都服从最小二乘的pdf, 自由度 = $N - m$

或利用皮尔逊的 χ^2 检验, 

用 $\hat{v}_i = v_i(\hat{\theta})$ 代替 v_i


$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \hat{v}_i)^2}{\hat{v}_i}$$

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - \hat{p}_i n_{tot})^2}{\hat{p}_i n_{tot}}$$

泊松分布: 自由度 = $N - m$

多项式分布: 自由度 = $N - m - 1$

不等精度观测结果的并合

如果 $(x_1, \dots, x_n)$ 是对同一固定量 μ 的 n 个不等精度测量值，对应的标准误差是 $(\sigma_1, \dots, \sigma_n)$ ，其中任一测量值 x_i 的分布是方差为 σ_i^2 的正态分布 $N(x_i, \mu, \sigma_i^2)$ ，而且各观测值相互独立，方差已知。则不等精度观测样本的似然函数为

$$L(\vec{x}; \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma_i} \right)^2 \right]$$

因此，取对数并解似然方程

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^n \frac{x_i - \mu}{\sigma_i^2} = 0$$



$$\hat{\mu} = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} = \frac{\sum_{i=1}^n \omega_i x_i}{\sum_{i=1}^n \omega_i}$$

权重因子 $\omega_i = \frac{1}{\sigma_i^2}$

$$\begin{aligned} \widehat{\sigma_{\hat{\mu}}^2} &= \left(\frac{-1}{\frac{\partial^2 \text{Log} L}{\partial^2 \mu}} \right) \bigg|_{\mu=\hat{\mu}} \\ &= \frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} \end{aligned}$$

例子：求 μ 寿命的世界平均值

世界上一共有五个实验精确测量了 μ 的平均寿命(10^{-9} s)

τ_i	权重因子	$\tau_i - \hat{\tau}$	$\chi_i^2 = \omega_i (\tau_i - \hat{\tau})^2$
2197.078 ± 0.073	187.65	0.048	0.432
2197.025 ± 0.155	41.62	-0.005	0.001
2196.95 ± 0.06	277.78	-0.08	1.778
2197.11 ± 0.08	156.25	0.08	1.000
2197.3 ± 0.3	11.11	0.27	0.810

$$\hat{\tau} = \frac{\sum_{i=1}^5 \omega_i \tau_i}{\sum_{i=1}^5 \omega_i} = 2197.03 \times 10^{-9} \text{秒}, \quad \hat{\sigma}_{\hat{\tau}} = \left(\sum_{i=1}^5 \omega_i \right)^{-\frac{1}{2}} = 0.04 \times 10^{-9} \text{秒}$$

$$\sum_{i=1}^5 \chi_i^2 / (5-1) = 1.005$$



与期待值相符。

似然函数与贝叶斯定理

在贝叶斯统计中, θ 与 $\vec{x}$ 都是随机变量

$$L(\theta) = L(\vec{x} | \theta) = f_{\text{joint}}(\vec{x} | \theta)$$

是 $\vec{x}$ 在给定 θ 的条件概率密度函数。

贝叶斯方法

- 对假设 (θ) 采用主观概率来描述
- 实验前, $\pi(\theta)$ 反应了对过程的了解(先验概率密度函数)
- 利用贝叶斯定理, 根据数据来改进先验概率密度函数

→ 验后概率:
$$p(\theta | \vec{x}) = \frac{L(\vec{x} | \theta)\pi(\theta)}{\int L(\vec{x} | \theta')\pi(\theta')d\theta'}$$

似然法估计量与贝叶斯估计量

$p(\theta | \vec{x})$ 是一个复杂的函数。



利用估计量 $\hat{\theta}_{\text{Bayes}}$ 来归纳出对 θ 的认识。

实际应用中，会碰到如何得到 $\pi(\theta)$ 的问题？

无金科玉律(主观的!)，通常假定它在全部取值区域上均匀分布。



$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$

注意：均匀假设会遇到诸如 $\lambda = 1/\theta$ 时， $\pi(\lambda)$ 不是均匀分布之类的问题。

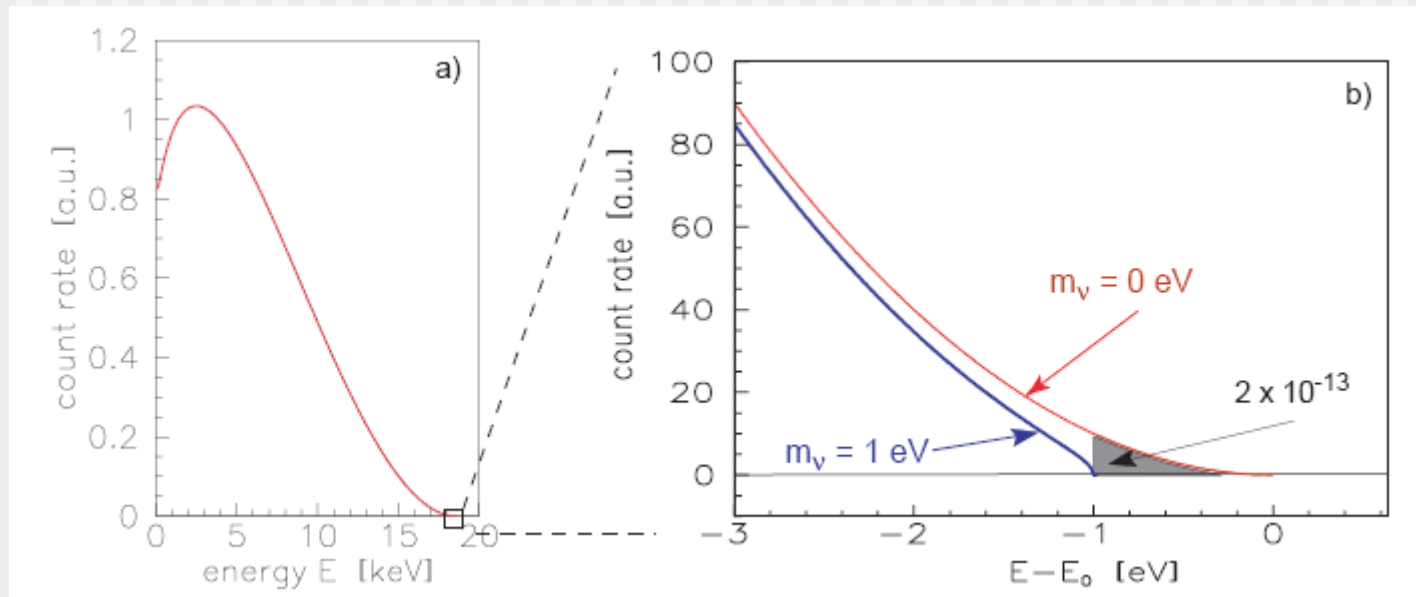
非贝叶斯方法目前在统计学中占主要地位。但是，如果被测定参数是随机变量，而且它的验前分布已知，就应该用贝叶斯方法。

例子：电子中微子质量测量

氚的贝塔衰变 ${}^3\text{H} \rightarrow {}^3\text{He}^+ + e^- + \bar{\nu}_e$ ($n \rightarrow p + e^- + \bar{\nu}_e$)

理论上，电子的能谱满足

$$\frac{dN}{dE} = C \times F(Z, E) p E (E_0 - E) [(E_0 - E)^2 - m_\nu^2]^{1/2}$$



显然，不能假设先验概率密度函数 $\pi(m_\nu)$ 在全空间中均匀分布，而应采取阶跃函数形式， m_ν 大于某值时 $\pi(m_\nu) = 0$ 。

本章要点

- ❑ 似然函数，最大似然估计量
- ❑ 最大似然估计量的方差估计
- ❑ 分区间数据拟合
- ❑ 最大似然法的拟合优度检验
- ❑ 不等精度观测结果的并合
- ❑ 最大似然法与贝叶斯估计

$$-2 \ln \lambda_p = 2 \sum_i (n_i - \nu_i + n_i \log \frac{n_i}{\nu_i})$$

$$n_i = \sigma_i + \nu_i$$

$$= 2 \sum_i \left[-\sigma_i + (\nu_i + \sigma_i) \log \left(1 + \frac{\sigma_i}{\nu_i} \right) \right]$$

$$= 2 \sum_i \left[-\sigma_i + (\nu_i + \sigma_i) \left(\frac{\sigma_i}{\nu_i} - \frac{1}{2} \left(\frac{\sigma_i}{\nu_i} \right)^2 \right) \right]$$

$$= 2 \sum_i \left[-\sigma_i + \sigma_i + \frac{\sigma_i^2}{\nu_i} - \frac{1}{2} \frac{\sigma_i^2}{\nu_i} - \frac{1}{2} \frac{\sigma_i^3}{\nu_i^2} \right]$$

$$= \sum_i \frac{\sigma_i^2}{\nu_i} = \sum_i \frac{(n_i - \nu_i)^2}{\nu_i}$$