粒子物理与核物理实验中的数据分析

陈少敏
清华大学

第五讲：统计检验(续)
参数估计中的基本概念

本讲要点

- 检验拟合优度，*P*-值定义与应用
- 信号观测的显著程度
- 皮尔逊的χ^2检验

- 样本、估计量与偏置
- 估计量的平均值、方差与协方差

一个古老的随机性问题

任意投掷一枚硬币，得到结果为正面与反面的概率都是0.5。

如果有人声称他(她)对此进行了检验。投了20次，得到了17次正面的结果。那么他(她)能否断定得到正面的概率应该是

$$p_h = 0.85 \pm 0.08$$

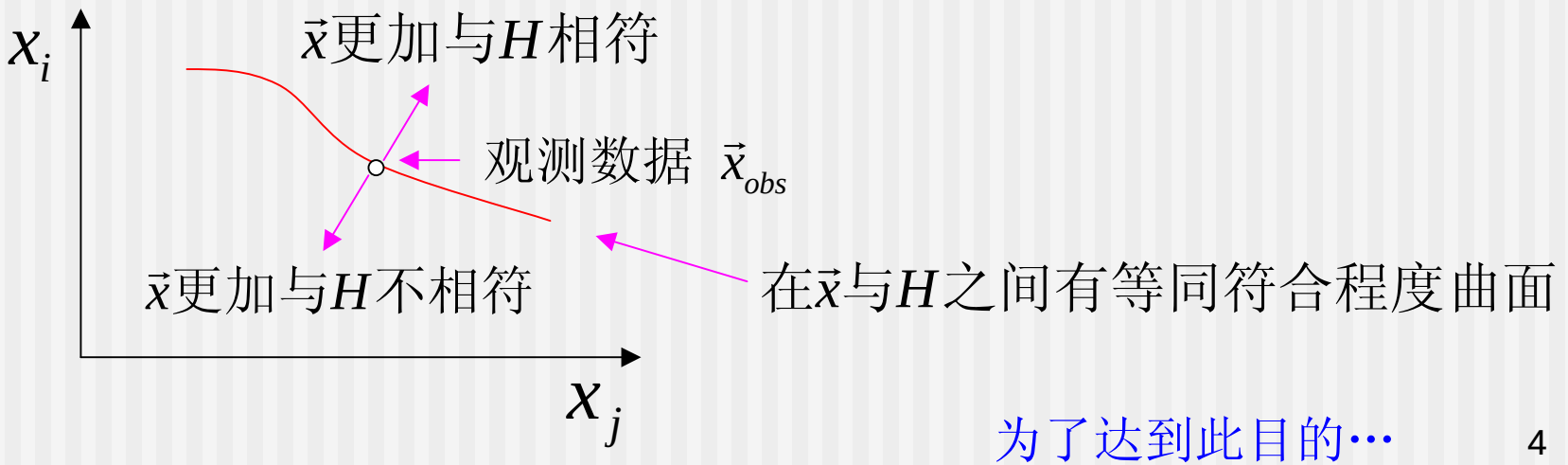
也就是说与预期值0.5有4个标准偏差呢？

问题：理论上允许这样的极端情况出现吗？或者说与这样一种极端情况相等或更高的概率有多大？

检验拟合优度

如果假设H对数据中的一部分矢量$\vec{x}=(x_1,\ x_2,...,x_n)$给出了预言 $f(\vec{x}\,|\,H)$。我们在 $\vec{x}$-空间观察到一个点:$\vec{x}_{obs}$。从数据来看，对假设H的正确与否会得出什么样的结论呢?

需要决定$\vec{x}$–空间中哪一部分比观测点$\vec{x}_{obs}$更能代表与假设H的不相符。



为了达到此目的…

检验统计量与拟合优度

通常需要构造统计检验量$t(\vec{x})$，它的大小可以反映出在$\vec{x}$与H之间符合的程度。例如

小的 t ⟹ 数据与 H 更符合

大的 t ⟹ 数据与 H 更不符合

由于概率密度函数 $f(\vec{x}\,|\,H)$已知，因此在H假设条件下检验统计量t 的概率密度函数 $g(t\,|\,H)$是完全可以确定。

P-值定义

将拟合优度用*P*-值表示 (也称为观察的显著水平或置信水平)

P=观察到实验数据 $\vec{x}$ 或 $t(\vec{x})$像 $\vec{x}_{obs}$ 或 $t(\vec{x}_{obs})$一样，
与假设H具有相同或较小符合程度的概率。

注意: 这不是 H 为真的概率。

在经典统计学上，我们从不涉及 $P(H)$ 。

而在贝叶斯统计理论中，则把 H 当成了随机变量，并利用贝叶斯定理得到

$$P(H \mid t) = \frac{P(t \mid H)\pi(H)}{\int P(t \mid H)\pi(H)dH}$$

$\pi(H)$:H 的先验概率

对所有可能性进行归一化积分

P-值与假设检验

根据*P*-值的定义，对 *H* 假设拟合优度的检验可以通过计算*P*-值的大小来完成。

但是，应注意以下两点:

- 在*P*-值定义中不涉及别的假设。
- *P*-值是一个随机变量。前面采用的显著水平在检验时已经被指定为常数。

→ 如果*H* 为真，则对于连续的$\vec{X}$，*P* 在[0,1]范围内均匀分布.

→ 如果*H* 非真，则*P*的概率密度函数通常很接近零。

例子:拟合优度检验

投 N 次硬币，观察到 n_h 次头朝上的概率服从二项式分布：

$$f(n_h; p_h, N) = \frac{N!}{n_h!(N-n_h)!} p_h^{n_h} (1-p_h)^{N-n_h}$$

假设H：硬币是公平的(朝上的 p_h = 朝下的 $p_t = 0.5$)

取拟合优度检验统计量 $t = |n_h - N/2|$

投 N=20 次硬币，观察到17次头朝上，则 $t_{obs} = |17 - 20/2| = 7$

在 t-空间中，具有相同或较少符合的区域为

$$t = (n_h - N/2) \geq 7$$

$$P-\text{值} = P(n_h = 0,1,2,3,17,18,19,20) = \sum_{i \leq 8} f_i \approx 0.0026$$

拟合优度检验中的问题

问题:当 *P*-值等于0.0026,是否意味着 *H* 假设是错的?

P-值并不回答此问题。它只是给出与观察到的结果一样,与 *H* 假设不符或者高于 *H* 假设($p_h=p_t=0.5$)的概率。

P-值=“偶然”得到如此奇怪结果的概率

一种实用的检验方法是在同样的假设下,产生同样数目的事例足够多次。检查如此奇怪的结果发生的概率是否与***P*-值**相当。

观测到一个信号的显著程度

假设观测 n 个事例，包含了

n_b = 已知过程(或本底)的事例数　　n_s = 新过程(或信号)的事例数

如果n_b，n_s服从泊松分布，均值为ν_b，ν_s，它们之和$n = n_b + n_s$也是服从泊松分布，均值为$\nu = \nu_b + \nu_s$：

$$P(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}e^{-(\nu_s+\nu_b)}$$

如果 ν_s=0.5而且观测到 n_{obs}=5 → 可否就此声称该迹象为新的发现？

假设H：ν_s=0，即只有本底过程出现。

→ 也就是所谓的“无效假设”

观测到一个信号的显著程度(续)

对应的P–值

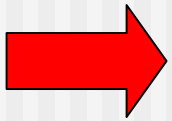
$$P-\text{值} = P(n \geq n_{obs})$$

$$= \sum_{n=n_{obs}}^{\infty} P(n;\nu_s = 0, \nu_b)$$

$$= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}$$

$$\approx 1.7 \times 10^{-4}$$

$$(\neq P(\nu_s = 0)!)$$



给出了得到这种极端结果的概率：虽然很小但不为零！

潜在的问题之一

一个误导读者但又常常被使用的结果表示...

对 ν_s 估计时得到：$n_{obs}=5$

估计 n 的标准偏差为：$\sqrt{n}=2.2$

信号

ν_s的估计值：$n_{obs}-\nu_b=4.5\pm2.2$

即与零有两倍的标准偏差

实际想要的是：均值 ν_b=0.5的泊松变量给出观测量大于 5 概率是多少？

概率为1.7×10^{-4}

但上面的结果表示隐含了均值为4.5，σ=2.2的高斯变量给出零或更少的概率：

$$\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}\times2.2}\exp\left(\frac{-(x-4.5)^2}{2\times2.2^2}\right)dx=0.021$$

如果$\nu_s \gg 1$，没有问题，
即 n 服从高斯分布。

潜在的问题之二

实际问题中会涉及系统误差，例如 $\nu_b=0.8$，则概率变为

$$P\text{-值} = P(n \geq 5; \nu_b = 0.8, \nu_s = 0)$$

$$= \sum_{n=n_{obs}}^{\infty} P(n;\ \nu_b = 0.8, \nu_s = 0)$$

$$= 1 - \sum_{n=0}^{n_{obs}-1} \frac{\nu_b^n}{n!} e^{-\nu_b}$$

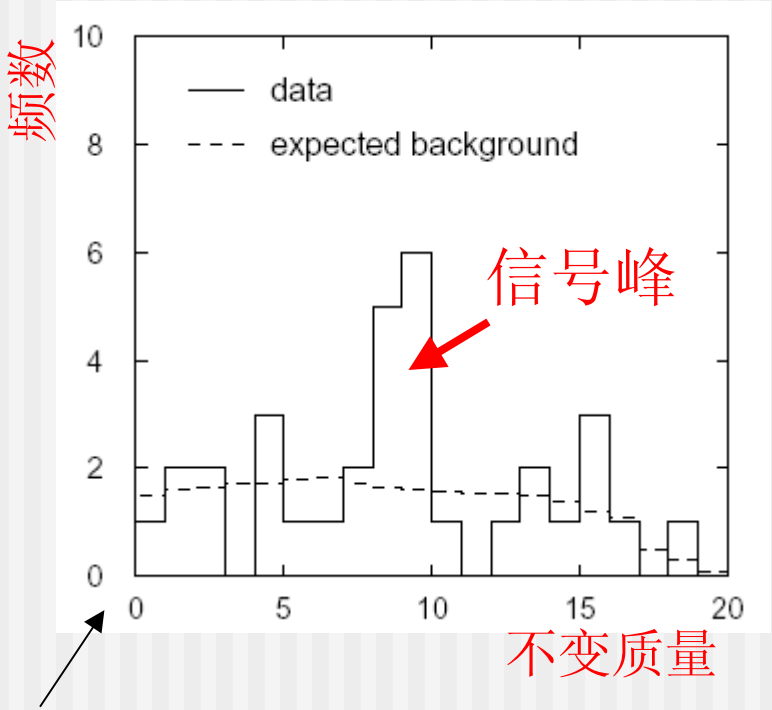
$$= 1.4 \times 10^{-3}$$

比 $\nu_b=0.5$ 时小了一个量级。

建议给出与 ν_b 合理变化相对应的P-值范围

信号峰的显著性

假设我们不但测量了总的事例数，还测量了每个事例对应的不变质量。



观察到的实验数据与期待本底大小的直方图，每个区间是泊松分布的一个变量。

在显示信号峰的两个区间，有11个事例，本底估计为$\nu_b = 3.2$

$$P(n \geq 11; \nu_b = 3.2; \nu_s = 0) = 5.0 \times 10^{-4}$$

Q1：在哪寻找信号峰？

➡ 计算任何两相连区间的P-值

Q2：信号宽度与分辨率相符吗？

➡ 将区间增大至分辨率的几倍

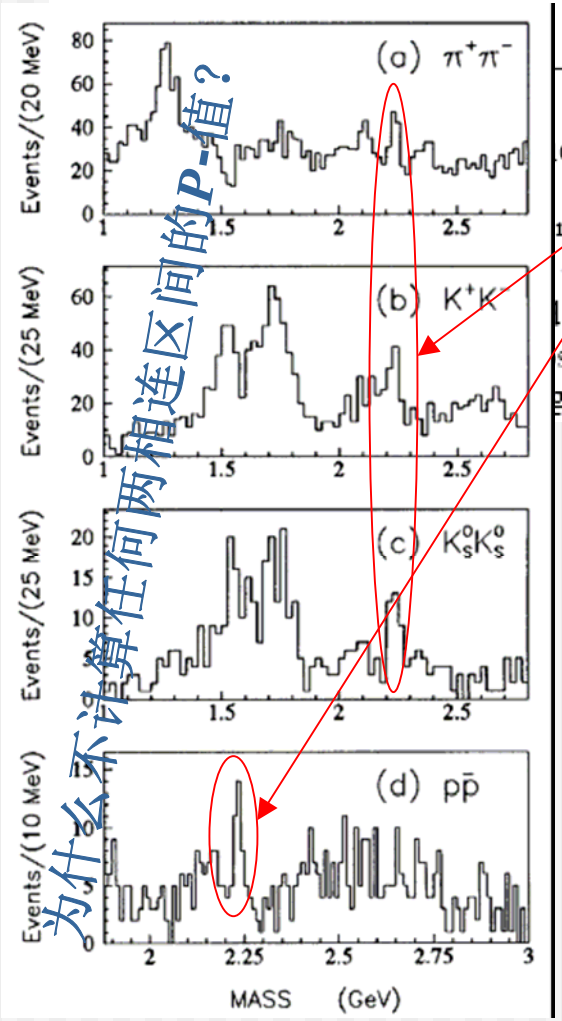
Q3：信号峰是人为制造出来的吗？

➡ 调整选择条件，分析新数据

…

Qn：能发表信号峰结果吗？

例子：ξ(2230)的观测



为什么不计算任何两相连区间的P-值？

PHYSICAL REVIEW LETTERS 6 MAY 1996

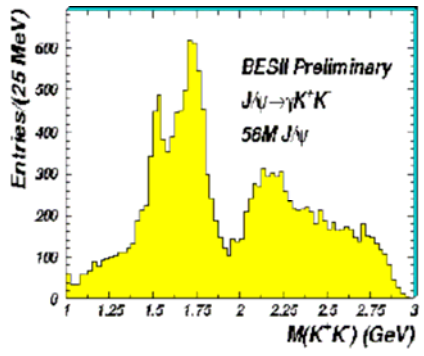
dies of $\xi(2230)$ in J/ψ Radiative Decays

n,[2] S. J. Chen,[1] S. M. Chen,[1] Y. Chen,[1] Y. B. Chen,[1] Y. Q. Chen,[1] B. S. Cheng,[1]

the Beijing electron-positron collider has observed the $\xi(2230)$ signal in $p\bar{p}$ final states with 4.6σ, 4.1σ, 4.0σ, and 3.8σ statistical significances. servations of two nonstrange decay modes of $\xi \to \pi^+\pi^-$ and $p\bar{p}$ are glueball interpretation of the $\xi(2230)$. [S0031-9007(96)00037-3]

陈老师也是作者之一

Status of ξ(2230) at BES II



- So far, no clear signal of ξ(2230) has been observed.
- All possible problems are still being checked.

重复实验得不到先前的结果！

皮尔逊的 χ^2 检验

在观测的数据 $\vec{n}=(n_1,\ldots,n_N)$ 与预言的期待值 $\vec{\nu}=(\nu_1,\ldots,\nu_N)$ 之间进行比较的检验统计量

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

如果 n_i 是相互独立而且服从均值为 ν_i 泊松分布，所有 ν_i 并不太小(≥5)，那么 χ^2 将服从N个自由度的最小二乘概率密度函数分布。所观察的χ^2可给出P-值

$$P-\text{值}=\int_{\chi^2}^{\infty} f(z;N)dz$$

这里，$f(z;N)$ 自由度为N 的最小二乘概率密度函数。

皮尔逊的 χ^2 检验(续)

自由度为N 的最小二乘概率密度函数的期待值为 $E(z)=N$

➡ 通常以χ^2/N 来体现符合的程度

最好分别给出 χ^2，N，例如

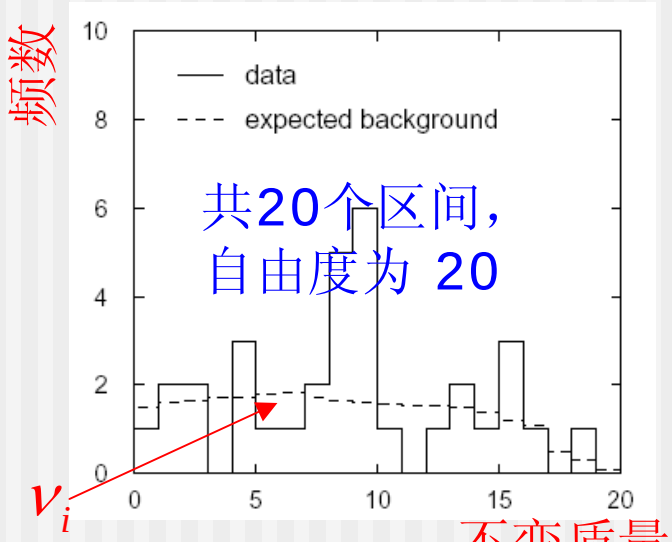
$$\chi^2 = 15,\ N = 10 \rightarrow P - 值 = 0.13$$

$$\chi^2 = 150, N = 100 \rightarrow P - 值 = 9.0 \times 10^{-4}$$

如果$n_{tot} = \sum_{i=1}^{N} n_i$固定，$n_i$服从二项式分布，$p_i = \nu_i / n_{tot}$,则

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i n_{tot})^2}{p_i n_{tot}}$$ ⬅ 服从N-1自由度的χ^2分布($p_i\ n_{tot} \gg 1$)

例子：χ^2 检验



$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

1）由于许多区间只有很少或根本没有计数，它将不服从 χ^2 的概率密度函数分布。

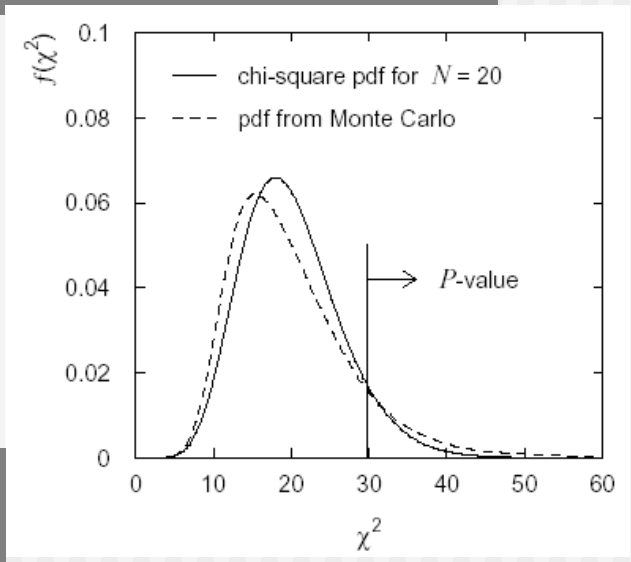
2）皮尔森 χ^2 仍可以作为一个检验统计量。

为计算*P*-值，先用蒙特卡罗方法得到 $f(\chi^2)$

产生 n_i 均值为 ν_i 的泊松分布，$i=1,\ldots,N$

计算 χ^2，填入直方图

重复足够多次



MC pdf：*P*-值=0.11

χ^2 pdf：*P*-值=0.073

对于统计检验的评论

在实际问题中，我们常常遇到对低统计量的情况下，需要判断所观察到的现象是否为真正的物理信号。利用*P*-值的大小可以表示结果是否为已知过程的极端情形。由于每个人的信心不同，会造成同一个*P*-值，结论却完全不一样的现象。

在统计误差范围内无新迹象。

结果虽然在统计误差范围，但有可能是新物理的信号。

发现了新物理的信号，误差为...

历史上类似故事的发生很多：*J*/Ψ粒子的发现，*W*粒子的发现，顶夸克的发现...

参数估计:基本概念

考虑对一随机变量 x 进行 n 次独立测量

样本大小为 n

等效为一个 n 维矢量的单次观测

$$\vec{x} = (x_1, \ldots, x_n)$$

由于 x_i 相互独立，因此样本的联合概率密度函数可以表示为：

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

任务：从一数据样本中推断 $f(x)$ 属性

构造数据的函数，以便估计 $f(x)$ 各种属性，包括平均值，方差，...等等。

参数估计:基本概念(续)

通常情况下，首先给出 $f(x)$ 的假设形式，其中包含未知参数 θ

利用 $f(x,\theta)$ 的给定形式和数据样本估计参数 θ

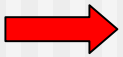
统计=所研究数据的函数

估计量=用来估计pdf某些属性的统计

记号：θ 的估计值为 $\hat{\theta}$

估计=估计量的观测值(通常记为 $\hat{\theta}_{obs}$)

注意：$\hat{\theta}(\vec{x})$是随机变量的函数。



其自身也是随机变量，由含期待值(均值)与方差等参数的概率密度函数来刻划。

估计量

如何构造估计量 $\hat{\theta}(\vec{x})$?

没有一个完美无缺的办法去构造估计量

构造估计量是为了满足某些与假设不相符而设立的判别条件。
首先是要求符合的程度

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

例如，随着样本的增大，估计值收敛于真值：

$$\text{对任何的}\varepsilon > 0, \lim_{n\to\infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

注意：在统计意义上的收敛并不能保证不会有个别特殊的$\hat{\theta}_{obs}$与θ值有较大的偏离。

估计量的各种属性

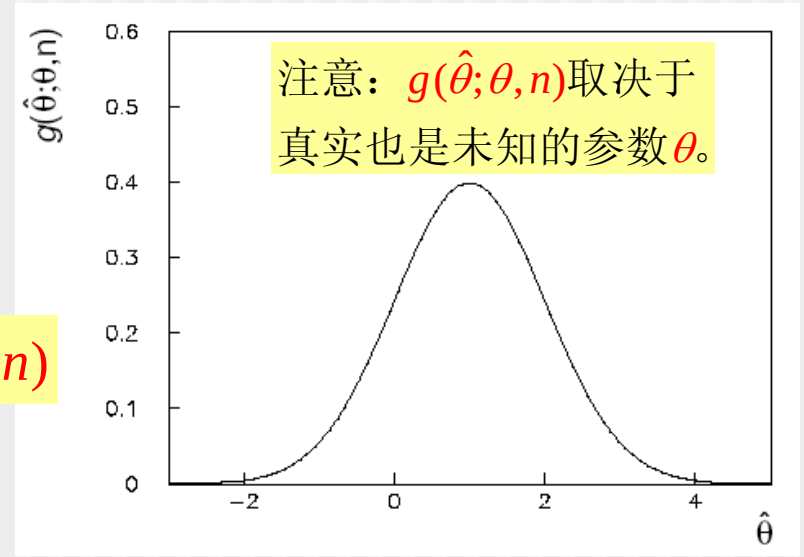
考虑对于一个固定样本大小 n 的估计量 $\hat{\theta}$ 有概率密度函数

我们不知道θ，只知道一个$\hat{\theta}_{obs}$值。

$g(\hat{\theta};\theta,n)$的属性包括：

方差$V[\hat{\theta}]=\sigma_{\hat{\theta}}^2$ ($\sigma_{\hat{\theta}}$="统计误差")

偏置$b=E[\hat{\theta}]-\theta$ ("系统误差",取决于n)



对于大量的估计量，我们将有

$\Rightarrow \sigma_{\hat{\theta}} \propto \frac{1}{\sqrt{n}}, b \propto \frac{1}{n}$ 有时会考虑均方误差 $\Rightarrow \mathrm{MSE}=V[\hat{\theta}]+b^2$

一般来说，在方差与偏置之间存在平衡点

$\Rightarrow$ 通常要求在零偏置的估计量中，对应的方差达到最小。

平均值(期待值)的估计量

考虑对物理量 x 的 n 个测量 $x_1,\ldots,x_n$，我们想要估计量

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ (样本的平均值)

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

$$\text{对任何}\ \varepsilon > 0, \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \geq \varepsilon\right) = 0$$

这就是大数弱定理。计算期待值

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

→ $\bar{x}$是一个对于μ 的无偏估计量。计算方差...

平均值(期待值)的评估量(续)

方差可以计算为

$$V[\bar{x}] = E[\bar{x}^2] - (E[\bar{x}])^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2$$

$$= \frac{1}{n^2}[(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)] - \mu^2$$

$$= \frac{\sigma^2}{n}$$

这里σ^2是 x 的方差，并利用了

当 $i \neq j$ 时, $E[x_i x_j] = E[x_i]E[x_j] = \mu^2$,

以及$E[x_i^2] = \mu^2 + \sigma^2$

方差的估计量

假设平均值 μ 和方差 $V[x]=\sigma^2$ 都是未知量。根据样本的方差估计σ^2

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

这里因子 1/(*n*-1) 使得 $E(S^2) = \sigma^2$，即无偏的。

假如 $\mu=E[x]$ 作为先验已知，则

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

是 σ^2 的一个无偏估计量。可以计算 s^2 的方差

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

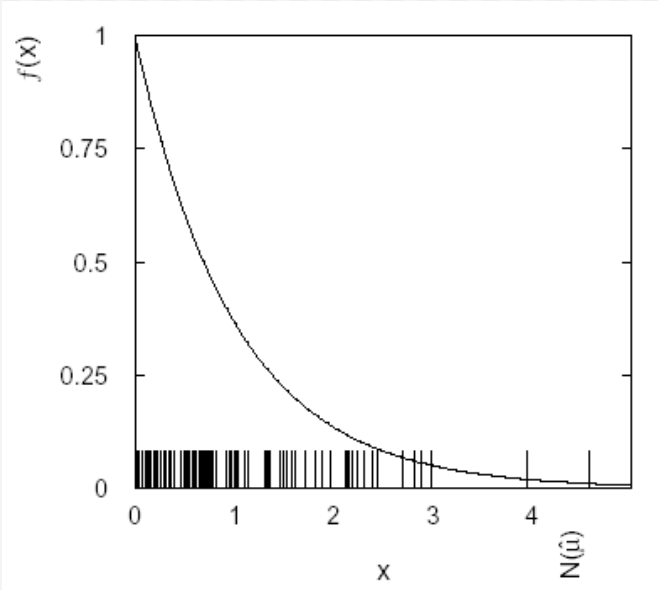
这里 μ_k 是第*k*阶中心矩

(例如：$\mu_2=\sigma^2$)

可以利用下式可估计 μ_k

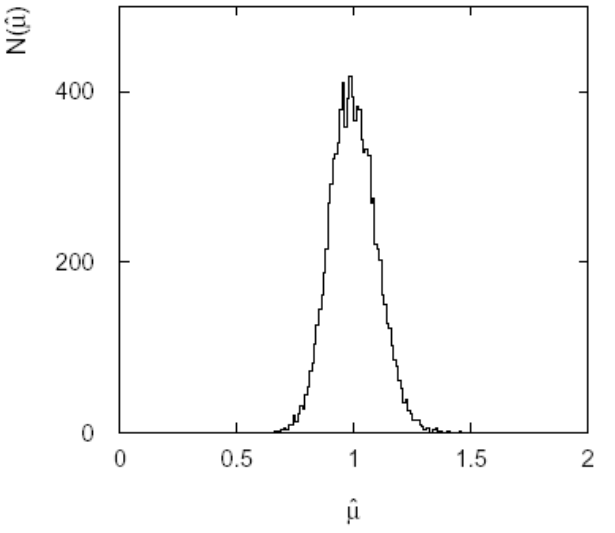
$$m_k = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

例子:平均值的估计量



数据样本 $n=100$
数值来自蒙特卡罗
$\mu=1$，$\sigma^2=1$

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



$\overline{\hat{\mu}}=0.9981$($\hat{\mu}$是无偏的)
$\hat{\mu}$值的样本标准偏差
$=0.0995\approx\sigma/\sqrt{n}$

重复10^4次，每次都是$n=100$，把每一次样本的平均值填入直方图

根据中心极限定理, $\hat{\mu}$ 的概率密度函数近似于高斯。

协方差与相关系数的评价量

为了估计协方差 $V_{xy}=\mathrm{cov}[x,y]$，利用

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

(是无偏的)

对于相关系数 $\rho=\dfrac{V_{xy}}{\sigma_x\sigma_y}$ 的估计，可以利用下式进行估计

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

此时的 r 有系统偏置。但是当 $n\to\infty$ 时，该偏置将趋于零。

一般而言，概率密度函数 $g(r,\rho,n)$ 形式较复杂；对于高斯分布量 x,y

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

小结

1. **检验拟合优度，*P*-值定义与应用**

   P-值为得到数据像已观测的结果一样与假设不符或更不符合的概率。

2. **信号观测的显著程度**

   很复杂，许多具有10^{-4}效应的结果最终证明是统计涨落的受害者。

3. **皮尔逊 χ^2 检验**

   广泛用于检验统计量。对于小样本数据，它将不服从 χ^2 的概率密度函数分布。但仍可用蒙特卡罗得到概率密度函数分布。

- **估计量**

  无最佳方法构造估计量，可以按符合程度，偏置，方差来构造。

- **均值，方差与协方差的估计量**

  虽然推导中无太深奥的理论，但它们非常接近理想情况。

习题

习题5.1:在正负电子对撞实验中，观测具有特殊运动学性质的事例数服从泊松分布。对于一定的积分亮度(即在给定的束流密度下收集数据的时间), 预期3.9个事例的过程实际观测到16个事例。计算假如无新的物理过程贡献到所观测的事例数的*P*-值。为了对泊松分布进行求和,可以利用下式

$$\sum_{n=0}^{m} P(n;\nu) = 1 - F_{\chi^2}(2\nu; n_{dof})$$

这里$P(n{:}\nu)$是平均值为ν观测到n个事例的泊松概率, F_{χ^2}是 χ^2 对于自由度为 $n_{\mathrm{dof}}=2(m+1)$ 分布的累积分布。

习题(续一)

习题5.2: 在放射性实验中,卢瑟福与盖革对一固定时间间隔发生的α衰变的数目进行了统计(数据见下表)。假设源中包含大量的放射性原子,它们中的任何一个在短时间内发射出一个α粒子的概率很小,可以认为在时间间隔 Δt 衰变的数目 m 服从泊松分布。

m	0	1	2	3	4	5	6	7
n_m	57	203	383	525	532	408	273	139
m	8	9	10	11	12	13	14	>14
n_m	45	27	10	4	0	1	1	0

与该假设的任何偏离将意味着衰变不是独立的。可以想象例如发射的α粒子可能会引起邻近的原子发生衰变,导致短时间内发生级联衰变。

习题(续二)

(a) 利用表中的数据,找出样本的平均值

$$\overline{m} = \frac{1}{n_{tot}} \sum_m n_m m$$

和对应的样本方差

$$s^2 = \frac{1}{n_{tot} - 1} \sum_m n_m (m - \overline{m})^2$$

这里, n_m是有m次衰变的数目。$n_{tot} = \sum_m n_m = 2608$ 是总的时间间隔。求和是从 m=0 到在一个间隔中所能观测到的最大衰变数目(在该实验为m=14)。从 $\overline{m}$ 与 s^2找出祢散的指标

$$t = \frac{s^2}{\overline{m}}$$

习题(续三)

由于$\overline{m}$ 与 s^2 是 m 平均值与方差的估计量,而且如果 m 为泊松分布变量时,它们是相等的,可以预计 t 为 1 左右。可以证明对于泊松分布的量 m 和大的 n_{tot}, $(n_{tot}-1)\,t$ 服从 n_{tot}-1自由度下的χ^2分布。而且,对于大的n_{tot}，将服从平均值为 n_{tot}-1 与方差等于 $2(n_{tot}-1)$ 的高斯分布。

(b)假设m 服从泊松分布，对应的P-值是多少?什么样的 t 值可选为代表比已观测的 t 完全或较少符合泊松假设?

(c) 用蒙特卡罗方法产生大量的数据样本，每个样本均包含了n_{tot}=2608个 m 次α衰变，这里m服从泊松分布，均值取由表中的数据得到的$\overline{m}$。对每一个数据样本，定出 t 值并填入直方图。从直方图和从卢瑟福数据中得到的t 值，定出对于泊松分布假设的P-值。将结果与(a)得到的结果相比较。如果填入的是$(n_{tot}-1)\,t$，请检查分布是否符合高斯分布?