

粒子物理中的统计分析简介

袁 丽

北京航空航天大学

iSTEP 2023 @浙江大学

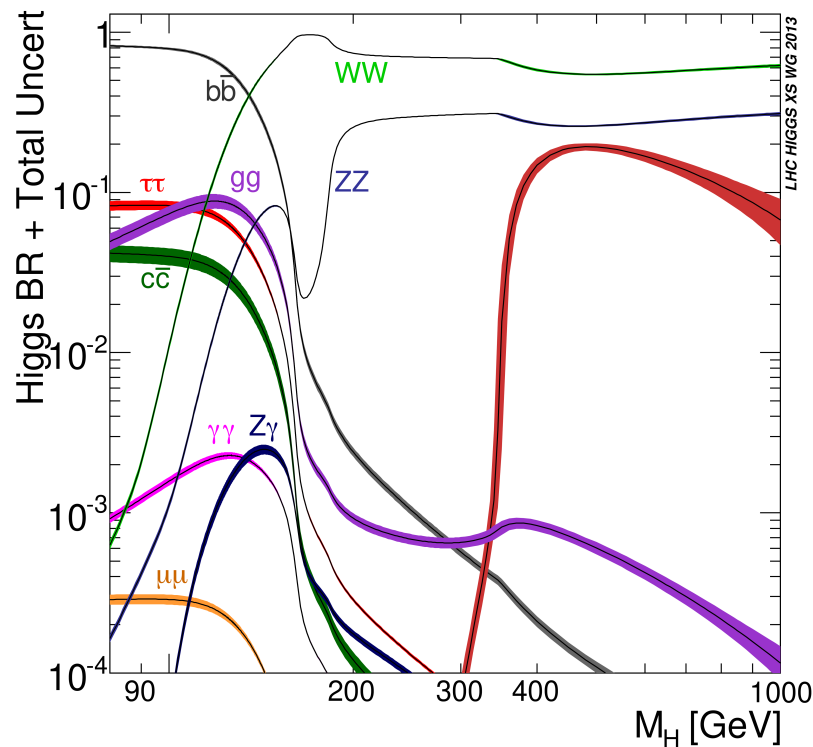
粒子物理中到处是随机现象

- Higgs粒子可衰变到ZZ, WW, bb, $\tau\tau$ 等不同的末态。对于任意一个Higgs粒子, 我们无法准确说出它将衰变为哪一种末态
- Z玻色子的质量测量值: $91.1876 \pm 0.0021 \text{ GeV}$. 由于其具有一定宽度, Z玻色子的质量满足一定的分布
- τ 轻子的寿命测量值: $0.290 \pm 0.001 \text{ ps}$. 但我们无法确切地说出一个特定的 τ 轻子将在何时发生衰变。
- 某一种物理过程 (例如: $J/\psi \rightarrow \mu\mu$), 末态粒子 μ 的飞行方向是任意的。
-

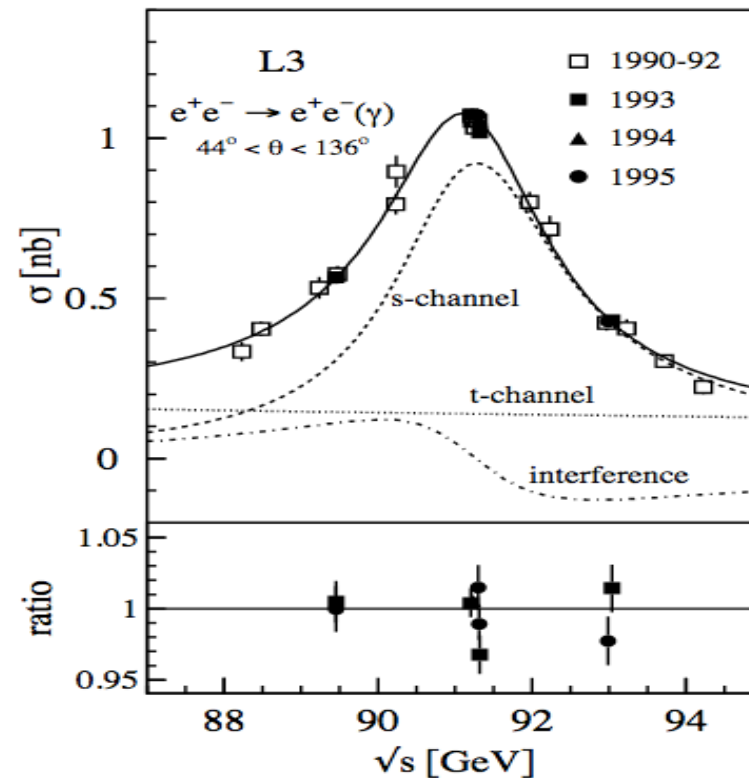
用以研究基本粒子性质的实验结果都是**随机的**。

随机现象满足统计规律

- Higgs粒子的各个衰变末态以一定概率呈现



- Z的质量满足Breit-Wigner分布

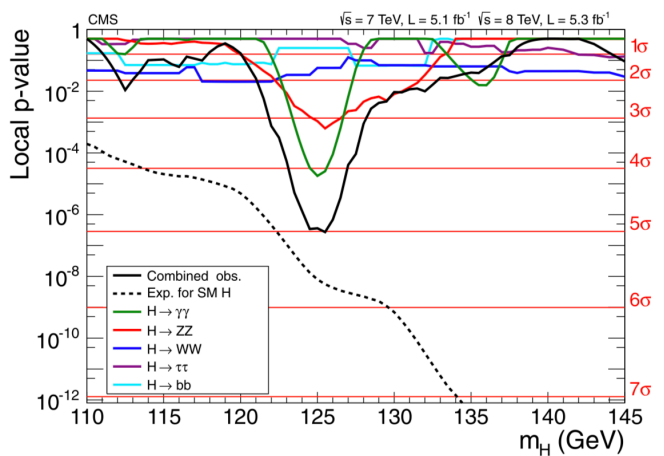
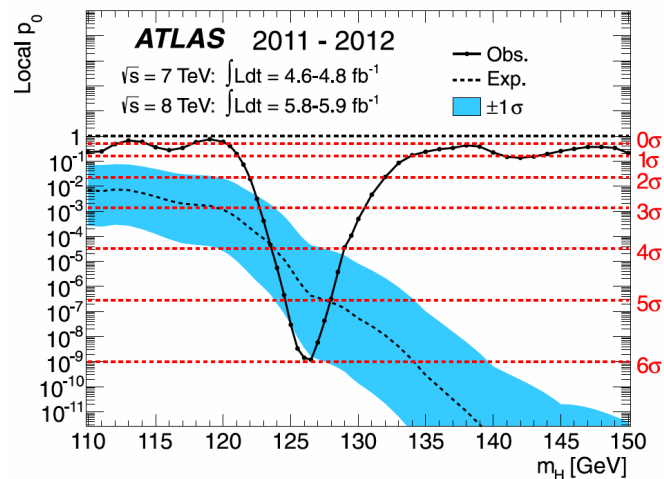


我们应具备关于随机现象的统计学知识，用以理解或描述实验结果。

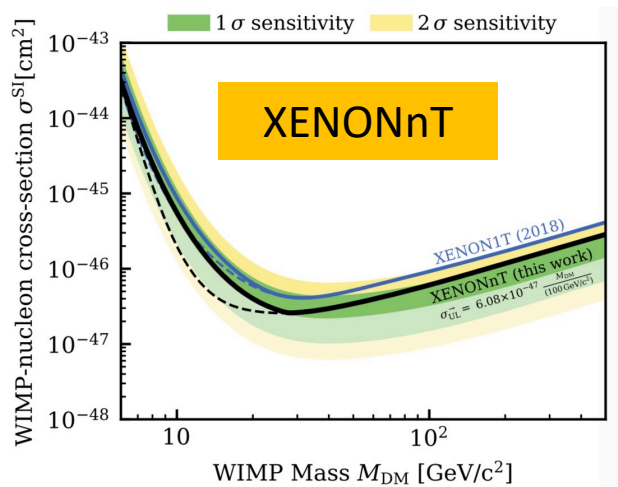
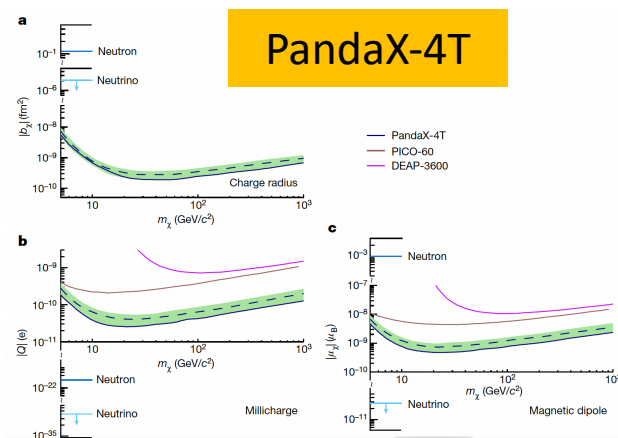
什么是统计 (Statistics) ?

- chatGPT: Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, presenting, and organizing data. It involves methods and techniques for summarizing and describing data, making inferences and predictions based on data, and testing hypotheses. Statistics is widely used in various fields, such as business, economics, social sciences, medicine, and many others, to make informed decisions and draw meaningful conclusions from data.
- 在粒子物理中，统计是一种实验数据分析的共通语言，是用来分析、诠释实验数据，表述物理结果的数学工具。

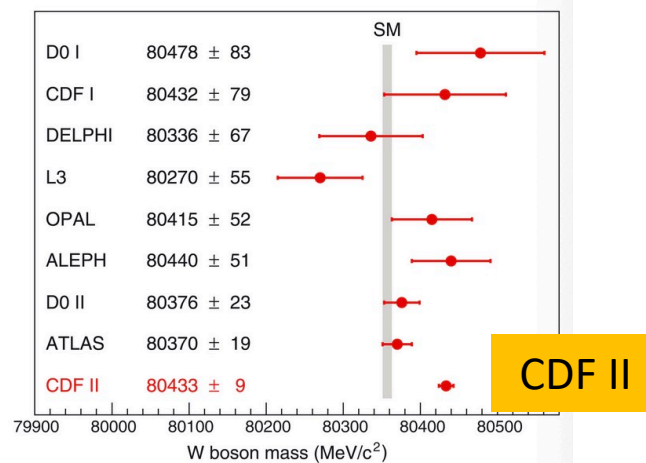
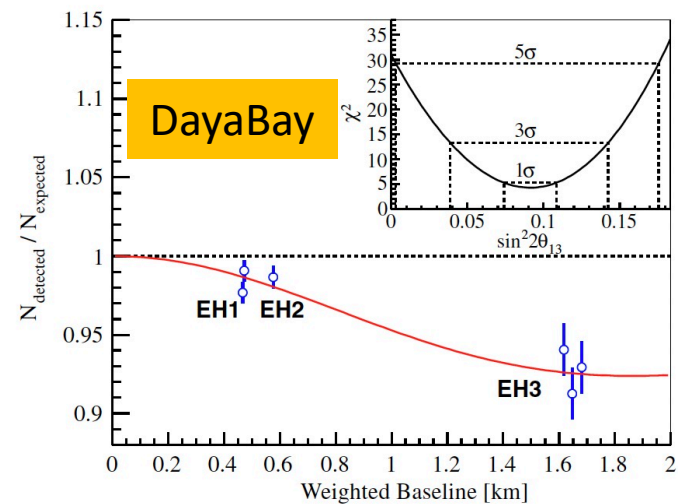
实验结果示例



信号显著性



区间估计



参数测量 (误差)

提 纲

- ❖ 概率论基础
 - ❖ 随机变量、概率
 - ❖ 常见概率分布
 - ❖ 误差传递公式
- ❖ 统计分析
 - ❖ 参数估计
 - ❖ 区间估计
 - ❖ 假设检验

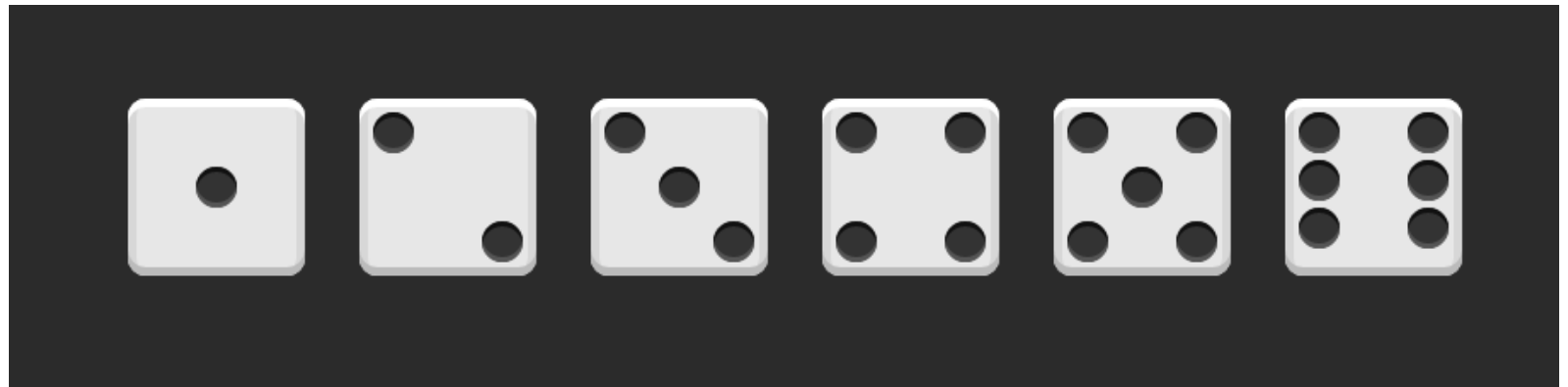
本课件借鉴了陈明水老师和凌家杰老师的课件

参考资料

- 《实验物理中的概率和统计》 朱永生 著
- 《高能物理数据分析：统计方法实用指南》 朱永生/胡红波译
- 《Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods》, O. BEHNKE, K. KRONINGER, G. SCHOTT, T. SCHORNER-SADENIUS
- 《Statistical Data Analysis》, Glen. COWAN
- Particle Data Book: sections on probability, statistics, Monte Carlo, <https://pdg.lbl.gov>

随机事件

- 在相同条件下对同一事物做多次测量或试验，单个测量或试验的结果是不确定的，而大量重复试验的结果则表现出某种规律性，具有这种性质的试验称为**随机试验**。
- 随机试验中可能出现的各种结果称为**随机事件**。
- 随机试验中每一种可能出现的结果是最简单最基本的事件，称为**基本事件**。随机试验 E 的所有基本事件组成的集合称为 E 的**样本空间**



概率的定义

- 重复某种随机试验 N 次，其中事件 A 出现 n 次，事件 A 的概率定义为试验次数 N 趋向于无穷大的极限情形下的频率：

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

事件的概率是随机试验中该事件发生的可能性大小的数量表述

- 设 S 为一随机试验 E 的样本空间，对于 E 的任一事件 A ，满足如下条件的一个非负实数 $P(A)$ 称为事件 A 的概率：

1) $0 \leq P(A) \leq 1$ ，对一切 $A \subset S$ (非负性)

2) $P(S) = 1$ (归一性)

3) 对两两互斥事件 $A_i (i=1, 2, \dots)$ 有 $P(A_1+A_2+\dots) = P(A_1)+P(A_2)+\dots$ (可列可加性)

条件概率

- 假设事件B出现的概率不为零，在给定B的情况下出现A的条件概率定义为：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

如果A与B独立，则

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ P(A|B) &= \frac{P(A) \cdot P(B)}{P(B)} = P(A) \end{aligned}$$

同理可得

$$P(B|A) = P(B)$$

表示事件A、B相互独立时，则事件A的发生与否对事件B的概率没有影响，反之亦然。

贝叶斯定理

由条件概率知： $P(A|B) = \frac{P(A \cap B)}{P(B)}$ ， $P(B|A) = \frac{P(B \cap A)}{P(A)}$

$$\because P(A \cap B) = P(B \cap A)$$

$$\therefore P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

由全概率公式可知，将全集 S 分为相互排斥的子集 A_i ，即 $\sum_i A_i = S$

$$P(B) = \sum_i P(B|A_i) \cdot P(A_i)$$

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)}$$

概率计算例题：

- 有三个相同的钱袋 B_1 , B_2 , B_3 。 B_1 内装2枚金币, B_2 内装金币、银币各一枚, B_3 内装2枚银币。随机地选取一个钱袋并摸出一枚钱币, 设摸到的是一枚金币, 问该钱袋内余下的也是一枚金币的概率为多大?

解：令第一次摸得金币是事件 A , 因为只有选取的钱袋是 B_1 时余下的才是金币, 所以该问题实际上是要求条件概率 $P(B_1|A)$ 。

根据已知条件, 从三个钱袋中摸得一枚金币的条件概率分别是

$$P(A|B_1) = 1, \quad P(A|B_2) = 1/2, \quad P(A|B_3) = 0$$

由于钱袋是随机选定的, 选中 B_1 , B_2 , B_3 的概率相等, 故有:

$$P(B_1) = P(B_2) = P(B_3) = 1/3$$

根据贝叶斯定理可得: $P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^3 P(A|B_i)P(B_i)} = \frac{2}{3}$

随机变量

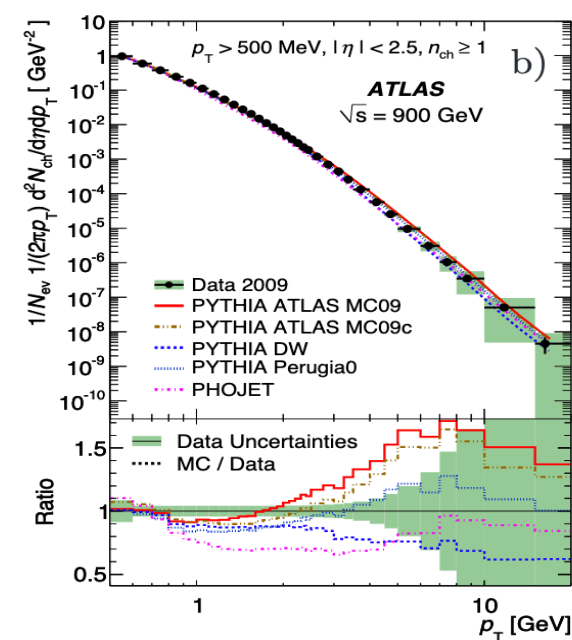
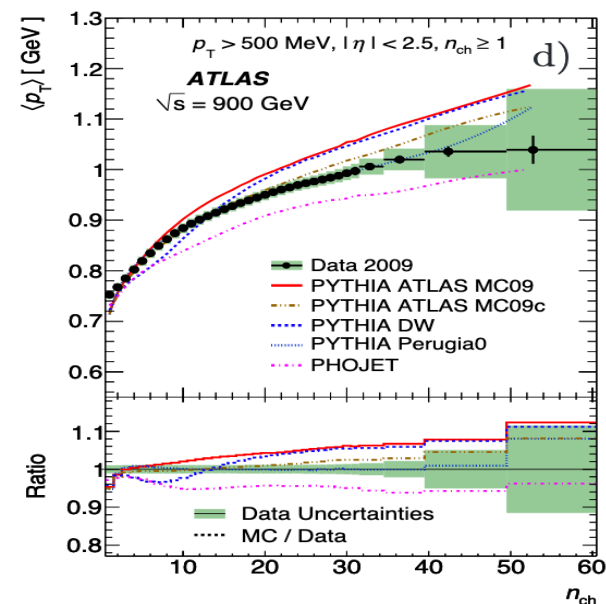
- **随机变量**：随机事件的数量表现
- 根据随机变量在一定区间内取值数量是否有限，分为
 - 离散型 (discrete)
 - 连续型 (continuous)

连续型变量 X :

- **概率密度函数 (p. d. f)** : $f(x) = \frac{P(X \in [x, x+dx])}{dx}$

- **分布函数**: $F(x) = \int_{-\infty}^x f(t)dt = P(X \leq x)$, $f(x) = \frac{dF(x)}{dx}$

离散型变量 X : $f_i = P(x_i)$, $F(x) = \sum_{x \leq x_i} P(x_i)$

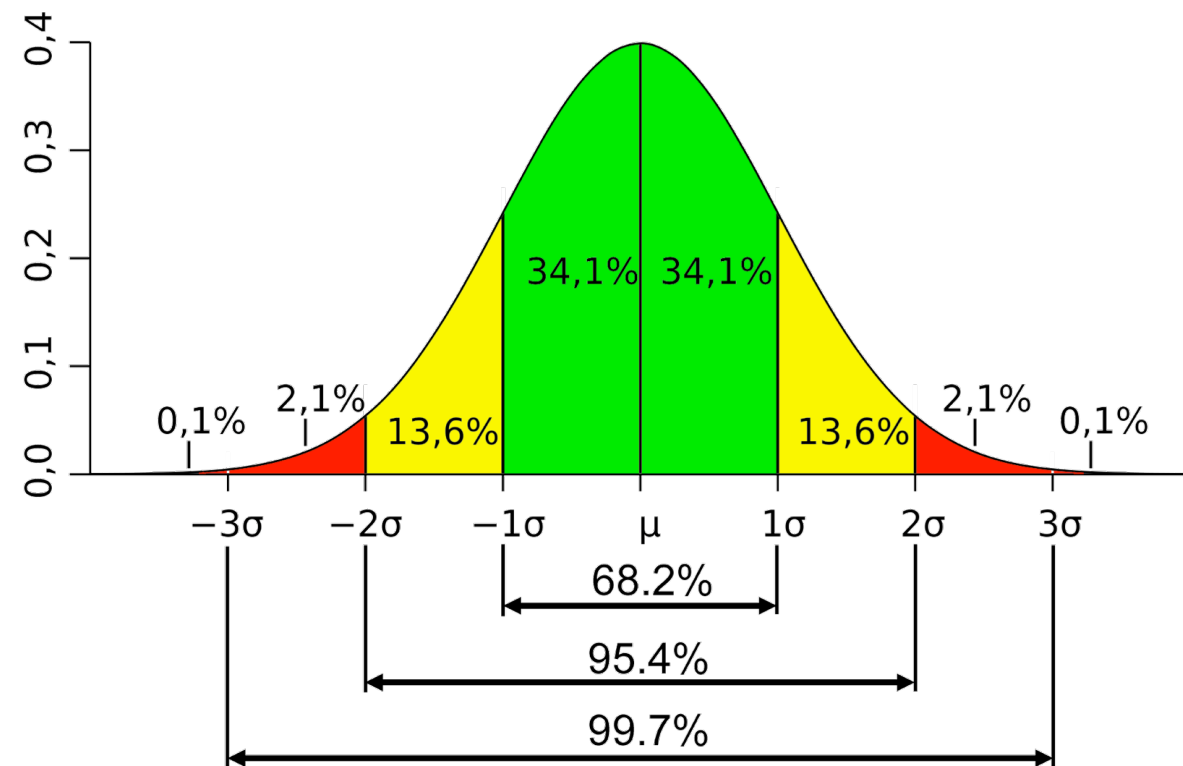


随机变量的数字特征

- **期望值:** $E(x) = \int xf(x)dx = \mu$
- **方差:** $V(x) = E[(x - E[x])^2] = E[x^2] - \mu^2 = \sigma^2$
- **标准方差:** $\sigma = \sqrt{V(x)}$
- **最可几值:** 满足 $f(x_{pro}) = \max_{x \in \Omega} \{f(x)\}$ 的值 x_{pro}
- **协方差:** $Cov[x, y] = E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y$
- **相关系数:** $\rho_{XY} = \frac{Cov[x, y]}{\sigma_X \sigma_Y}, \quad -1 \leq \rho_{XY} \leq 1$

高斯分布

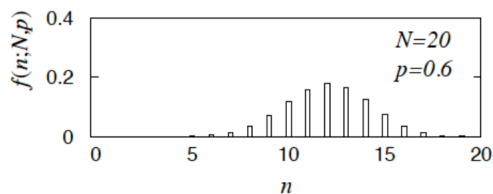
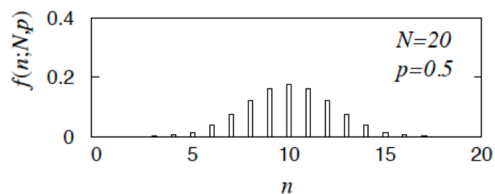
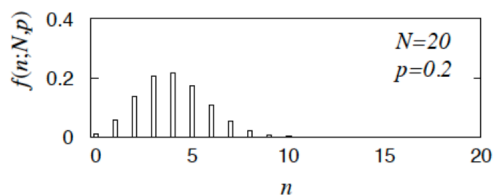
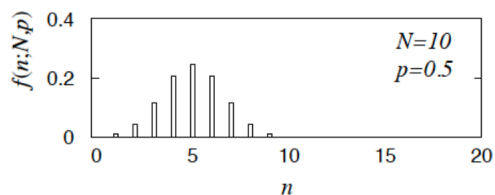
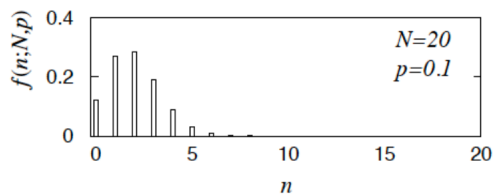
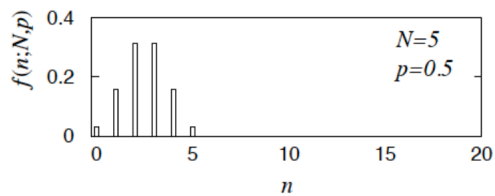
- **概率密度函数**: $f(x; \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- **期望值**: $E(x) = \int_{-\infty}^{\infty} xf(x; \mu, \sigma^2)dx = \mu$
- **方差**: $V(x) = E[x^2] - \mu^2 = \sigma^2$
- 对于许多典型的测量误差，高斯分布是较好的近似。



二项式分布

N次独立测量，每次成功的概率都是 p ，得到 n 次成功的概率是：

- **概率分布函数：** $P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$
- **期望值：** $E(n) = \sum_{n=0}^N n \cdot P(n; N, p) = Np$
- **方差：** $V(n) = E[n^2] - E(n)^2 = Np(1-p)$



二项分布中的参数 p 往往是未知的，须由实验来测定。当试验次数 N 足够大，事件出现次数为 n 时，事件出现的频率（或称效率）即为： $p = \frac{n}{N}$

效率的方差： $V(p) = V\left(\frac{n}{N}\right) = \frac{p(1-p)}{N}$

径迹重建效率例题：

- 实验中一个带电粒子穿过一组6个径迹室，后者用来测量粒子的位置。每一径迹室测量粒子的效率为95%。如果至少有4个径迹室记录到粒子的击中就能确定一条重建的径迹，则径迹的重建效率是多少？

解：根据二项式分布，有4个击中点的概率为：

$$P(4; 6, 0.95) = \frac{6!}{4! 2!} 0.95^4 0.05^2 = 0.031$$

有5个击中点的概率为：

$$P(5; 6, 0.95) = \frac{6!}{5! 1!} 0.95^5 0.05^1 = 0.232$$

有6个击中点的概率为：

$$P(6; 6, 0.95) = \frac{6!}{6! 0!} 0.95^6 0.05^0 = 0.735$$

总效率为： $0.031 + 0.232 + 0.735 = 0.998 = 99.8\%$

多项式分布

与二项式分布类似，但结果有 l 种，概率分别是 $\vec{p} = (p_1, p_2, \dots, p_l)$

每种结果发生的次数是 $\vec{n} = (n_1, n_2, \dots, n_l)$

- 概率分布函数： $P(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! \dots n_l!} p_1^{n_1} \dots p_l^{n_l}$
- 期望值： $E(n_i) = Np_i$
- 方差： $V(n_i) = Np_i(1 - p_i)$
- 协方差： $V_{ij} = -Np_i p_j (i \neq j)$

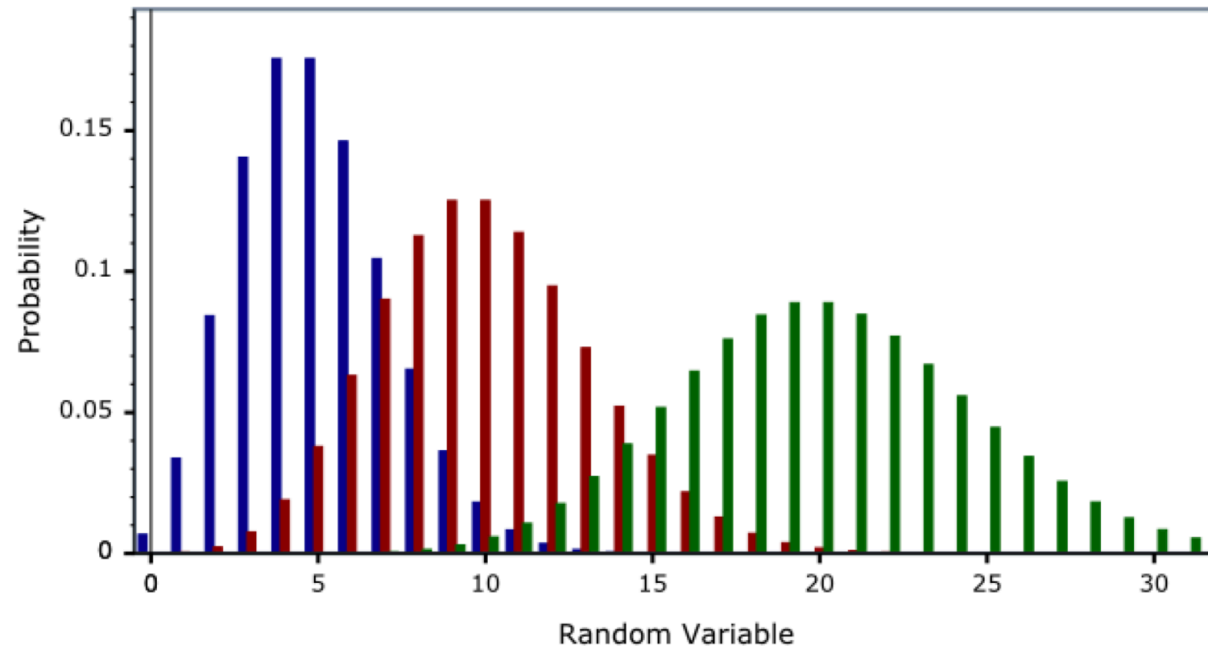
泊松分布

描述随机事件发生的频率的概率分布

- **概率分布函数**: $P(n; \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$
- **期望值**: $E(n) = \sum_{n=0}^{\infty} n \cdot P(n; \lambda) = \lambda$
- **方差**: $V(n) = E[n^2] - \lambda^2 = \lambda$

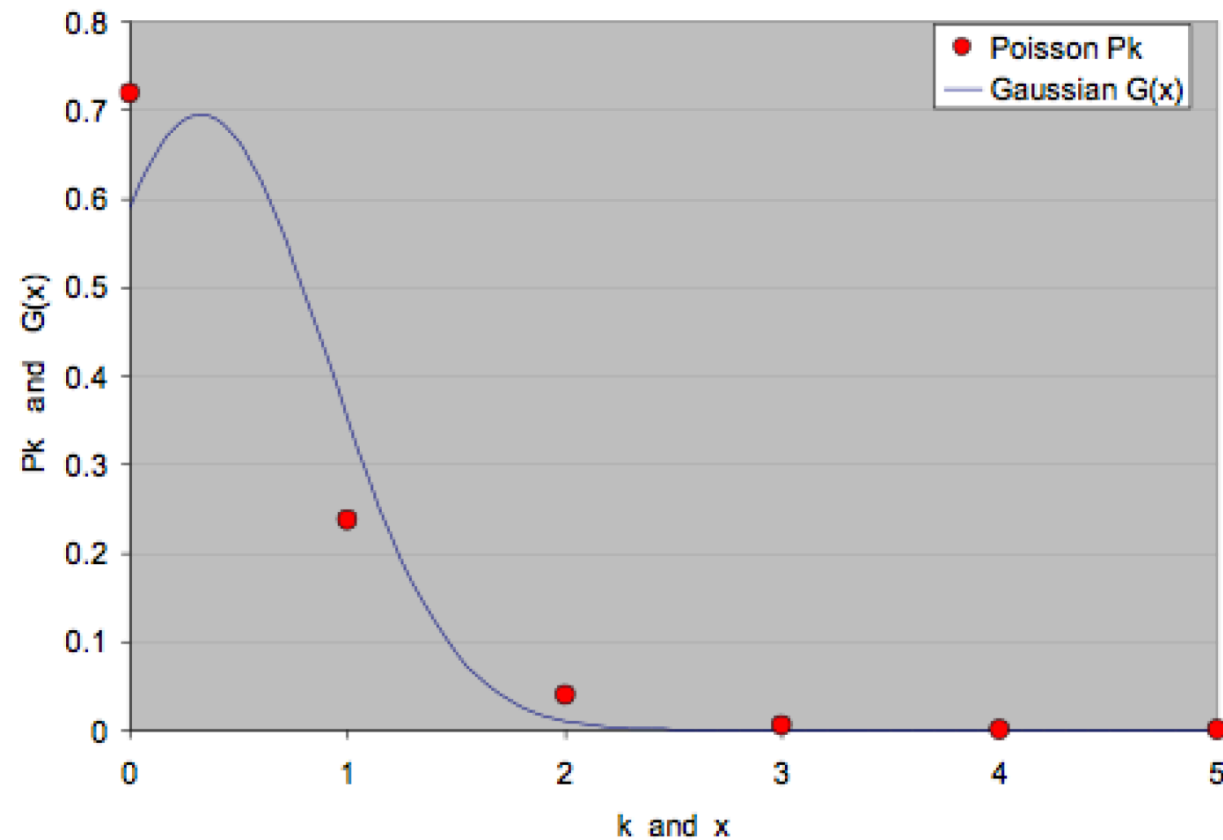
- 直方图中每个区间内的事例数可以看成:
 - 总事例数服从泊松分布, 各自服从多项式分布
 - 服从独立的泊松分布, 误差为 \sqrt{n}

Poisson Distribution PDF

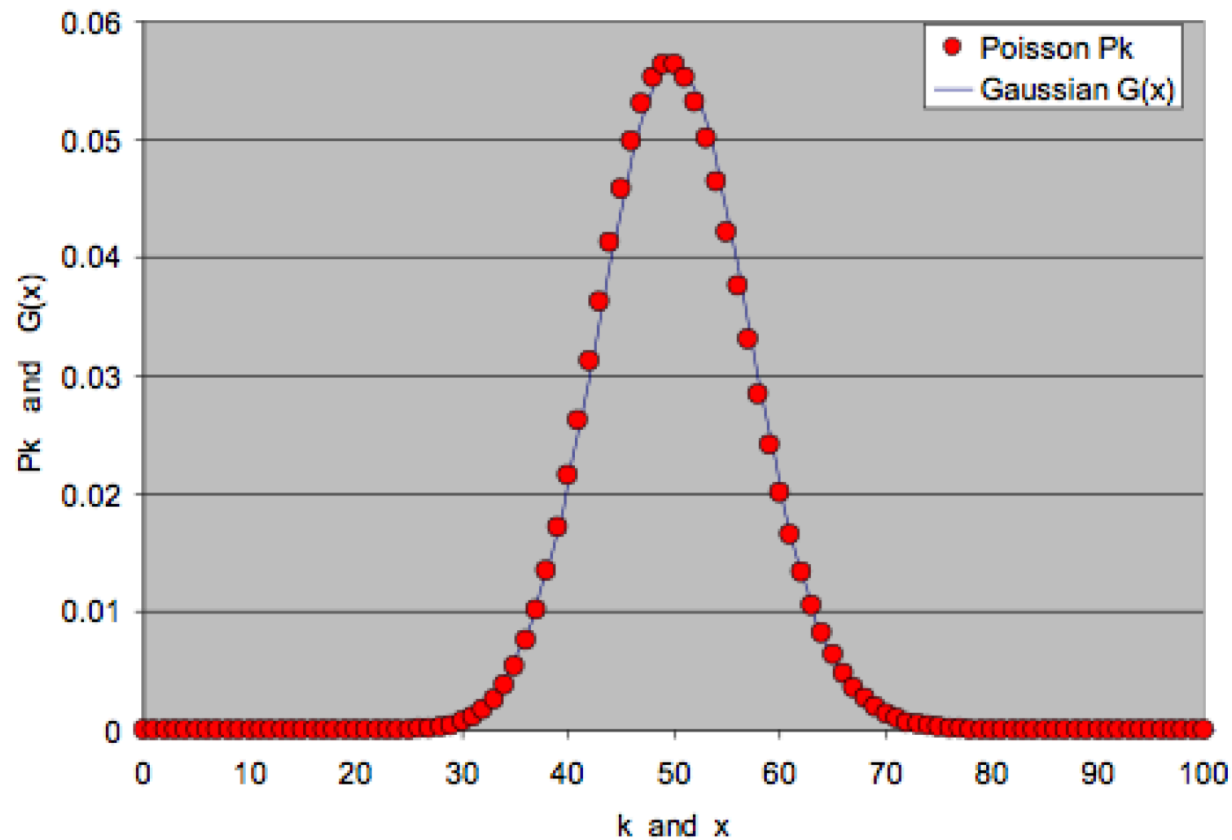


泊松分布 → 高斯分布

Average = 0.33



Average = 50



期望值为 n (远大于1时) 的泊松分布趋近于 $\mu=n, \sigma^2=n$ 的高斯分布

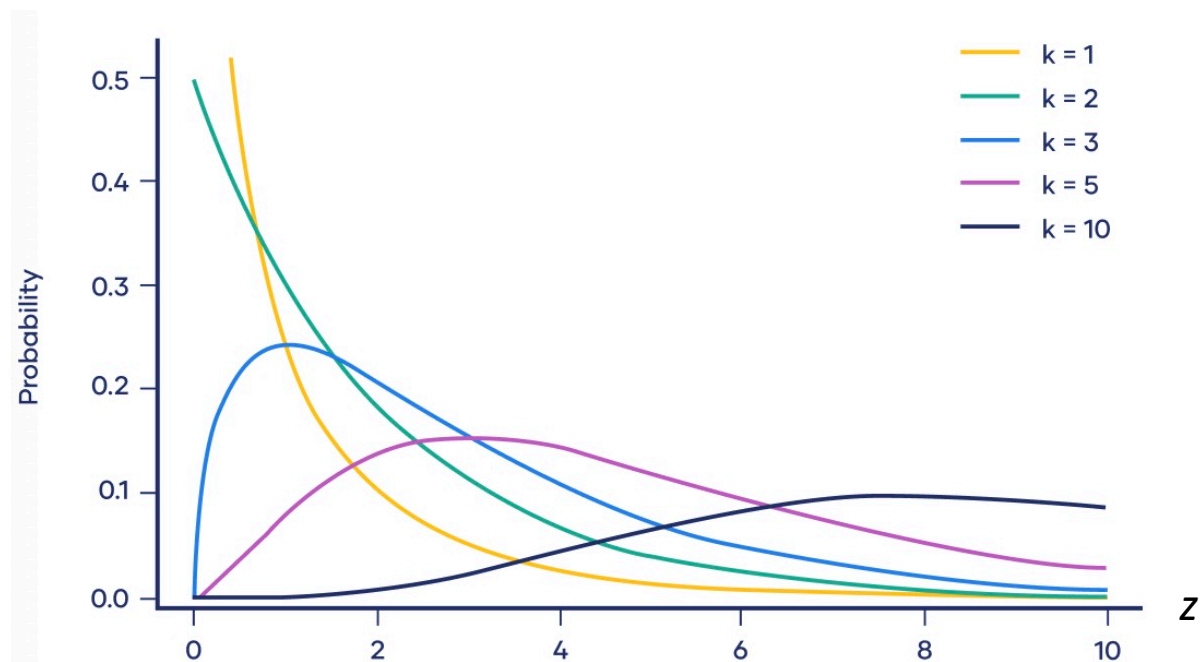
卡方分布

- 如果 x_1, \dots, x_k 服从标准正态分布 $N(0,1)$, 那么 $z = \sum_{i=1}^k x_i^2$ 服从

自由度为 k 的 χ^2 分布 $f(z; k) = \frac{z^{\frac{k}{2}-1}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} e^{-z/2}$ 。

- Γ 函数的定义为: $\Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x} dx$

- **期望值:** $E(z) = \mu = k$
- **方差:** $V(z) = \sigma^2 = 2k$



中心极限定理


- 中心极限定理：对于 n 个互相独立的符合任意的概率分布函数的变量 x_1, \dots, x_n （要求各自的平均值 μ_i 和方差 σ_i^2 是有限的），它们的和 $X = \sum x_i$ 在 $n \rightarrow \infty$ 时符合平均值为 $\sum \mu_i$ ，方差为 $\sum \sigma_i^2$ 的高斯分布。
- 在适当条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量之和近似服从正态分布的条件。


误差传递公式

- 假设 $\vec{x} = (x_1, \dots, x_n)$ 服从联合概率分布 $f(\vec{x})$, 已知协方差 $V_{ij} = \text{Cov}[x_i, x_j]$, 期望 $\vec{\mu} = E[\vec{x}]$, 现有一函数 $y(\vec{x})$, 其期望值 $E[y(\vec{x})] = y(\vec{\mu})$, 如何计算该函数的方差 $V[y(\vec{x})]$?

由方差计算公式可得: $V[y(\vec{x})] = E[y^2(\vec{x})] - (E[y(\vec{x})])^2$,
需要计算 $E[y^2(\vec{x})]$.

将 $y(\vec{x})$ 在 $\vec{\mu}$ 附近做一阶泰勒展开, $y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$

$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \cdot \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} E[(x_i - \mu_i)] +$  第二项 = 0

$E[(\sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)) \cdot (\sum_{j=1}^n \left[\frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j))]$  $\sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$

误差传递公式

$$V[y(\vec{x})] = y^2(\vec{\mu}) + \sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij} - (E[y(\vec{x})])^2$$

因而：
$$V[y(\vec{x})] = \sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

当 x_i 和 x_j 不相关时， $V_{ii} = \sigma_i^2$, $V_{ij} = 0$

例1：
$$y = ax_1 + bx_2$$

$$\sigma_y^2 = a^2\sigma_{x_1}^2 + 2ab\text{Cov}(x_1, x_2) + b^2\sigma_{x_2}^2$$

例2：
$$y = \frac{ax_1}{x_2}$$

$$\sigma_y^2 = \left(\frac{a}{x_2}\right)^2\sigma_{x_1}^2 - 2\frac{a}{x_2}\frac{ax_1}{x_2^2}\text{Cov}(x_1, x_2) + \left(\frac{ax_1}{x_2^2}\right)^2\sigma_{x_2}^2$$

$$\sigma_y^2 = \left(\frac{y}{x_1}\right)^2\sigma_{x_1}^2 - 2\frac{y}{x_1}\frac{y}{x_1x_2}\text{Cov}(x_1, x_2) + \left(\frac{y}{x_2}\right)^2\sigma_{x_2}^2$$

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} - \frac{2\text{Cov}(x_1, x_2)}{x_1x_2} + \frac{\sigma_{x_2}^2}{x_2^2} \xrightarrow{x_1, x_2 \text{不相关}} \frac{\sigma_y^2}{y^2} = \frac{\sigma_{x_1}^2}{x_1^2} + \frac{\sigma_{x_2}^2}{x_2^2}$$

统计分析

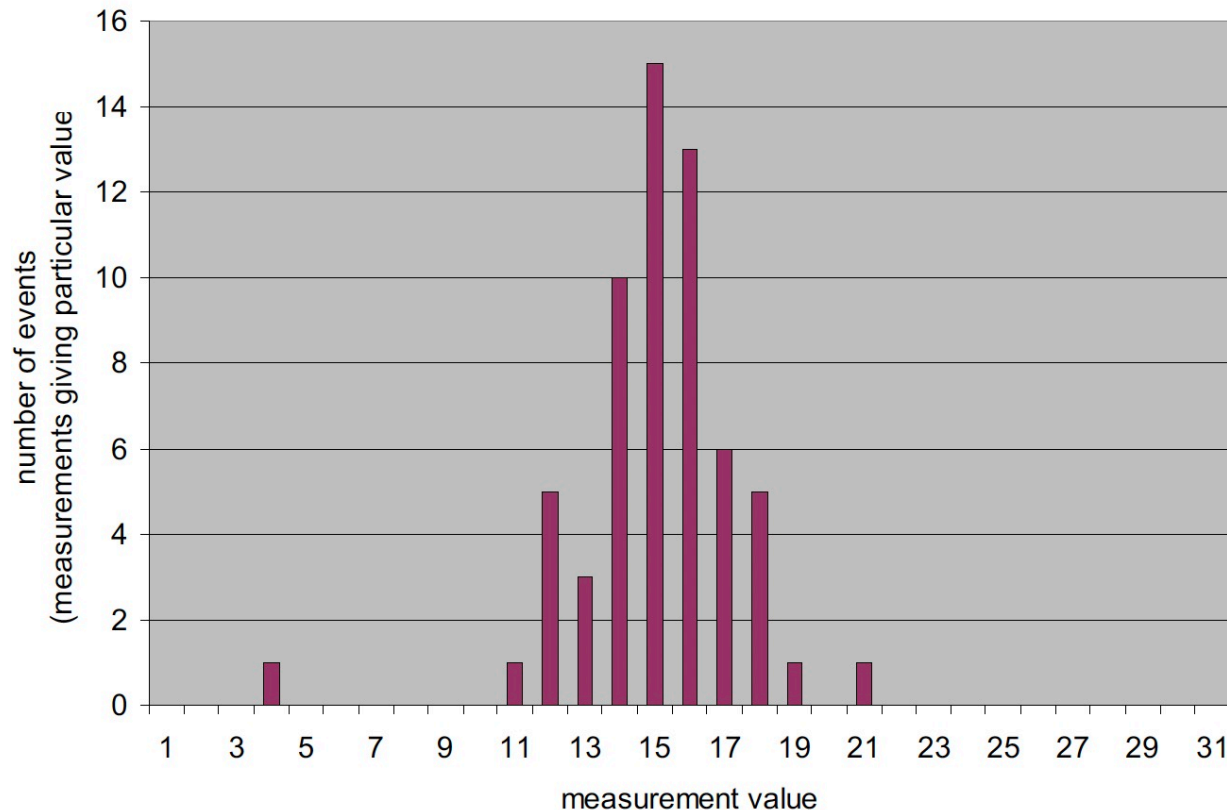
参数估计

区间估计

假设检验

参数估计

- 已知总体分布的概率密度函数形式，但含有未知参数，且已获得一组**有限**的测量值。
- 可通过这组测量值来估计未知参数的数值及评估其误差



- $x_m = \frac{1}{N} \sum_{i=1}^N x_i$
- $\sigma_m^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - x_m)^2$
- $\delta x_m = \frac{\sigma_m}{\sqrt{N}}$
- $\delta \sigma_m = \frac{\sigma_m}{\sqrt{2N}}$

最大似然法

- 假如实验上对具有 $f(x; \vec{\vartheta})$ 分布的变量 x 测量了一组数据： x_1, \dots, x_n ，其中 $\vec{\vartheta}$ 是概率密度函数中的未知参数。测量值为 x_i 的概率为 $f(x_i | \vec{\vartheta})$
- 子样 x_1, \dots, x_n 的联合概率密度由似然函数给出：

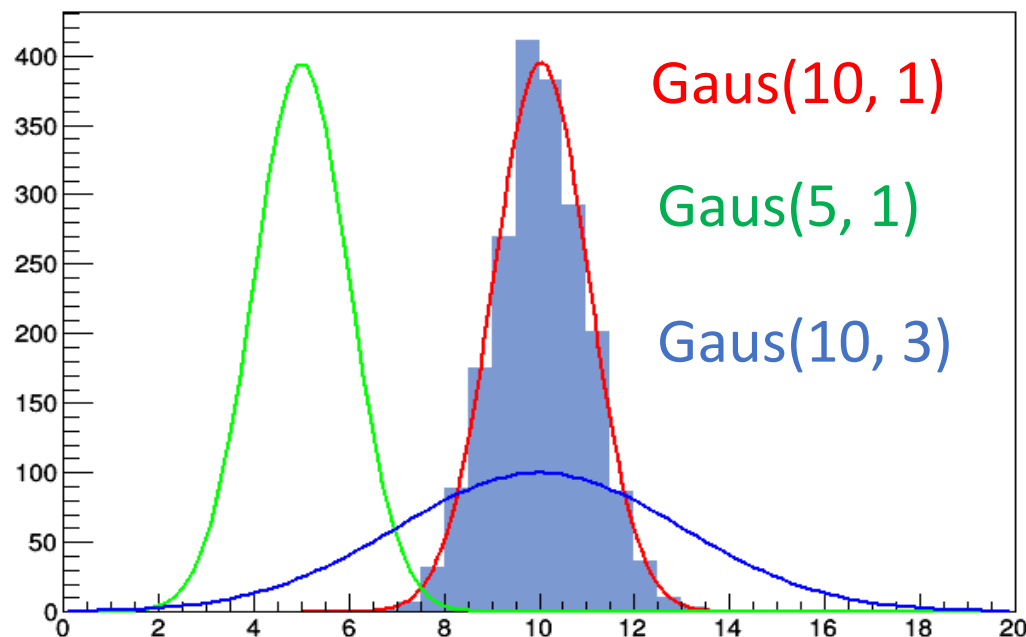
$$L(\vec{x} | \vec{\vartheta}) = L(x_1, \dots, x_n | \vec{\vartheta}) = \prod_{i=1}^n f(x_i | \vec{\vartheta})$$

- 根据极大似然原理，在参数的空间内，应当选择使似然函数 $L(\vec{x} | \vec{\vartheta})$ 达到极大的参数值 $\hat{\vartheta}$ 作为未知参数的估计值：

$$L(\vec{x} | \hat{\vartheta}) \geq L(\vec{x} | \vartheta)$$

最大似然法

Example



- 如果似然函数对 $\vec{\vartheta}$ 的二阶导数存在，极大似然估计 $\hat{\vartheta}$ 可通过求解方程组得到：

$$L(\vec{x}|\hat{\vartheta}) = \prod_{i=1}^n f(x_i|\mu = 10, \sigma = 1) \quad \text{Maximum}$$

$$L(\vec{x}|\vec{\vartheta}) = \prod_{i=1}^n f(x_i|\mu = 5, \sigma = 1)$$

$$L(\vec{x}|\vec{\vartheta}) = \prod_{i=1}^n f(x_i|\mu = 10, \sigma = 3)$$

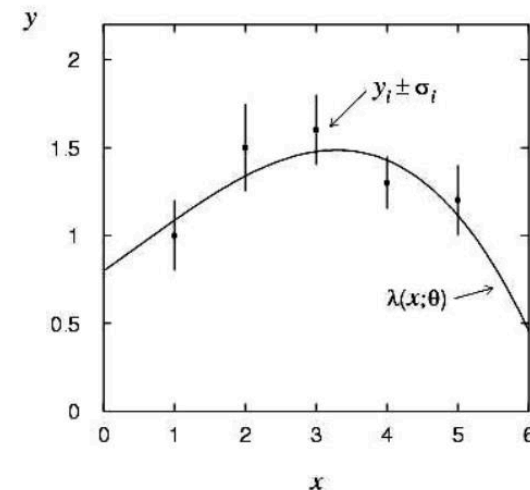
$$\frac{\partial L(\vec{x}|\vec{\vartheta})}{\partial \vec{\vartheta}} = \frac{\partial}{\partial \vec{\vartheta}} \prod_{i=1}^n f(x_i|\vec{\vartheta}) = 0,$$
$$\frac{\partial^2 L(\vec{x}|\vec{\vartheta})}{\partial \vec{\vartheta}^2} \Big|_{\vec{\vartheta}=\hat{\vartheta}} = \frac{\partial^2}{\partial \vec{\vartheta}^2} \prod_{i=1}^n f(x_i|\vec{\vartheta}) \Big|_{\vec{\vartheta}=\hat{\vartheta}} < 0$$

最小二乘法

- 在N个观测点 x_1, \dots, x_N ，通过测量得到**相互独立**的观测值 Y_1, \dots, Y_N ，相应的观测值真值 η_1, \dots, η_N 为未知。假定存在某个理论模型可预测与观测点 x_i 对应的 η_i : $\eta_i = \lambda(\vartheta_1, \dots, \vartheta_L; x_i)$

- 按照最小二乘原理，未知参数 $\vec{\vartheta}$ 的最优估计值是使量：

$$Q^2 = \sum_{i=1}^N (Y_i - \eta_i)^2 w_i$$



达到极小的参数值 $\hat{\vartheta}$ 。式中的权重因子一般取为观测值的方差的倒数。

- 如果将观测量考虑为数学期望等于真值的泊松变量 $w_i = \frac{1}{\sigma_i^2} = \frac{1}{\eta_i} \approx \frac{1}{Y_i}$,

则上式变为 $Q^2 = \sum_{i=1}^N \frac{(Y_i - \eta_i)^2}{Y_i}$ 服从 χ^2 分布

最小二乘法

- 假定观测值 Y_1, \dots, Y_N , 是关于对应真值 η_1, \dots, η_N 的高斯分布: 即 $Y_i \sim N(\eta_i, \sigma_i^2)$, 则N个观测量的似然函数为:

$$L = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \eta_i}{\sigma_i}\right)^2\right] \propto \exp\left[-\frac{1}{2} \sum_{i=1}^N \left(\frac{Y_i - \eta_i}{\sigma_i}\right)^2\right]$$

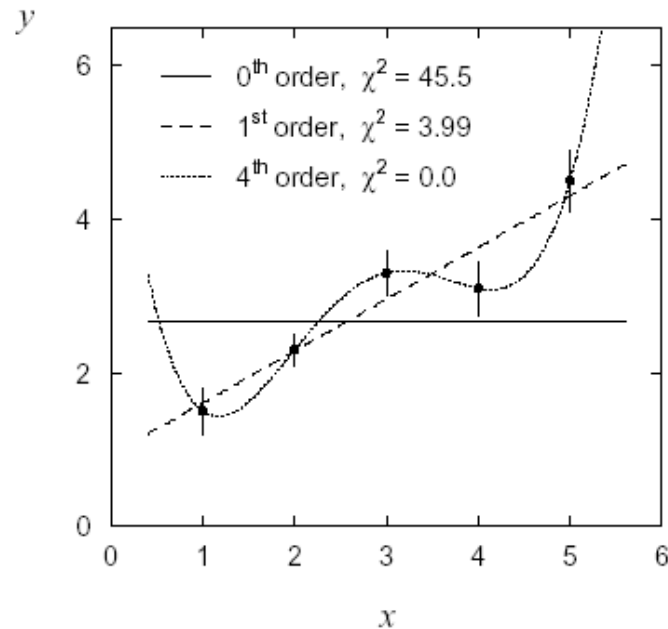
上式的极大似然值对应的即是 $Q^2 = \sum_{i=1}^N \left(\frac{Y_i - \eta_i}{\sigma_i}\right)^2$ 为最小值。

这种情况下, 最大似然法和最小二乘法是等效的。

Q^2 相当于N个标准正态分布平方之和

最小二乘法

- $\overline{Q^2} = N_{measurements} - m_{parameters} = n.d.f$ (自由度数目)
- 服从 χ^2 分布的随机变量的期望值等于自由度数目k(即数据点的数目减去参数的数目)
- 通常也用 χ^2 除以自由度数目来衡量拟合的好坏



用p维多项式拟合5个数据点:

$$\lambda(x; \vartheta_0, \dots, \vartheta_p) = \sum_{i=0}^p \vartheta_i x^i$$

$$\chi^2(\vartheta) = -2\ln L(\vartheta)$$

统计分析

参数估计

区间估计

假设检验

置信区间

- 设某一总体分布含有未知参数 ϑ ， x_1, \dots, x_n 为容量 n 的一个子样。对于给定值 $\gamma (0 < \gamma < 1)$ ，若由子样确定的两个数值 $\vartheta_a(x_1, \dots, x_n)$ 和 $\vartheta_b(x_1, \dots, x_n)$ 满足： $\gamma = P(\vartheta_a \leq \vartheta \leq \vartheta_b)$ ，则称 $[\vartheta_a, \vartheta_b]$ 为参数 ϑ 的概率量 γ 的置信区间（也可以是单边估计： $[-\infty, \vartheta_b]$ 或者 $[\vartheta_a, \infty]$)
- 求置信区间的一般方法：
 - 1) 找到包含未知参数 ϑ 和子样 x_1, \dots, x_n 的函数 $t = t(x_1, \dots, x_n; \vartheta)$ ，它是关于 ϑ 的单调函数，且概率密度 $g(t)$ 为已知，与 ϑ 无关。
 - 2) 求出 $t_a = t(\vartheta_a)$ 和 $t_b = t(\vartheta_b)$ ，使得 $\gamma = P(t_a \leq t(\vartheta) \leq t_b)$ ，
 - 3) 通过 $t \sim \vartheta$ 之间的变换求得参数 ϑ 的置信区间

区间估计：例题1

- 设满足指数分布 $f(x; \lambda) = \lambda e^{-\lambda x}$ 的一组观测值为 x_1, \dots, x_n 。求参数 λ 的置信水平为 γ 的置信区间。

令 $Y = \sum_{j=1}^n 2\lambda x_j$ ，分析其特征函数可知， Y 服从自由度 $2n$ 的 χ^2 分布。

因此，通过查询 $\chi^2(2n)$ 的分位数 $\chi_{\alpha/2}^2$ 和 $\chi_{1-\alpha/2}^2$ 满足

$$P\{\chi_{\alpha/2}^2 \geq Y \geq \chi_{1-\alpha/2}^2\} = 1 - \alpha = \gamma$$

即等同于： $P\left\{\frac{\chi_{\alpha/2}^2}{2 \sum_{j=1}^n x_j} \geq \lambda \geq \frac{\chi_{1-\alpha/2}^2}{2 \sum_{j=1}^n x_j}\right\} = 1 - \alpha = \gamma$

由此得出参数 λ 的置信区间为： $\left[\frac{\chi_{\alpha/2}^2}{2n\bar{x}}, \frac{\chi_{1-\alpha/2}^2}{2n\bar{x}}\right]$

区间估计：例题2

- 设满足正态分布 $N(\mu, \sigma^2)$ 的一组观测值为 x_1, \dots, x_n ， σ^2 为已知。求参数 μ 的置信水平为 γ 的置信区间。

子样平均值 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 服从正态分布 $N(\mu, \sigma^2/n)$ 。

因此，定义 $Y = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ ，则 Y 满足标准正态分布。

可求出给定置信水平 γ 的中心区间的上下限，满足

$$P\{-z_{\alpha/2} \leq Y \leq z_{\alpha/2}\} = 1 - \alpha = \gamma$$

即等同于： $P\left\{\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha = \gamma$

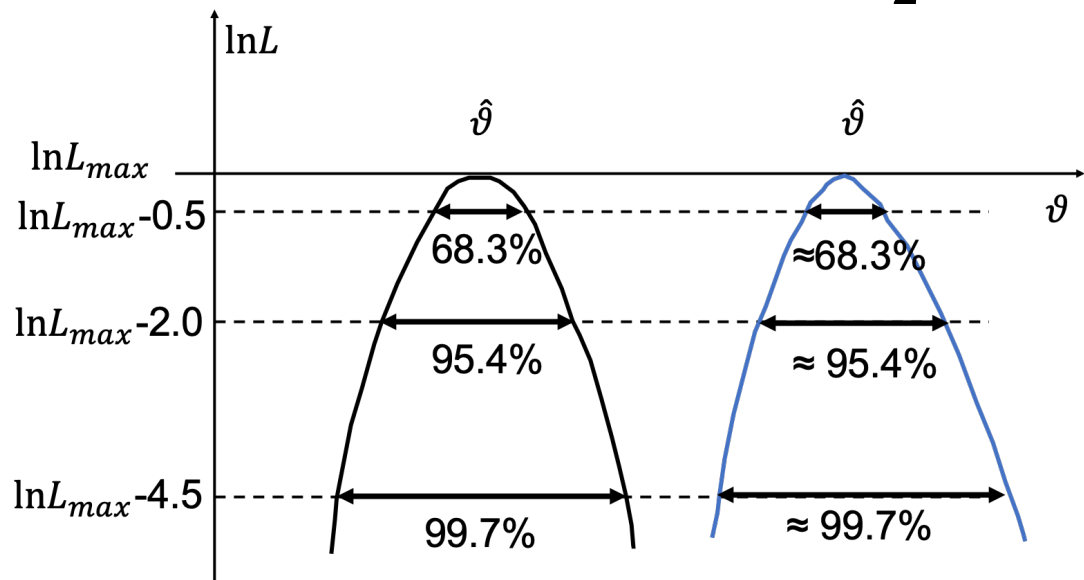
由此得出参数 λ 的置信区间为： $[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

似然函数作区间估计

- 设某一总体分布 $f(x|\vartheta)$ 的一个子样 x_1, \dots, x_n 。当子样容量趋于无穷时，似然函数 $L(\vec{x}|\vartheta) = \prod_{i=1}^n f(x_i|\vartheta)$ 的极大似然估计 $\hat{\vartheta}$ 的分布渐近地服从正态分布，方差 σ^2 达到最小方差界：

$$L(\vec{x}|\vartheta) \rightarrow L(\vartheta) = L_{max} e^{-\frac{1}{2}Q}, \quad Q = \left(\frac{\vartheta - \hat{\vartheta}}{\sigma}\right)^2$$

该式等价于： $\ln L(\vartheta) = \ln L_{max} - \frac{1}{2}Q$ ，即 $\ln L(\vartheta)$ 是 ϑ 的抛物线型函数

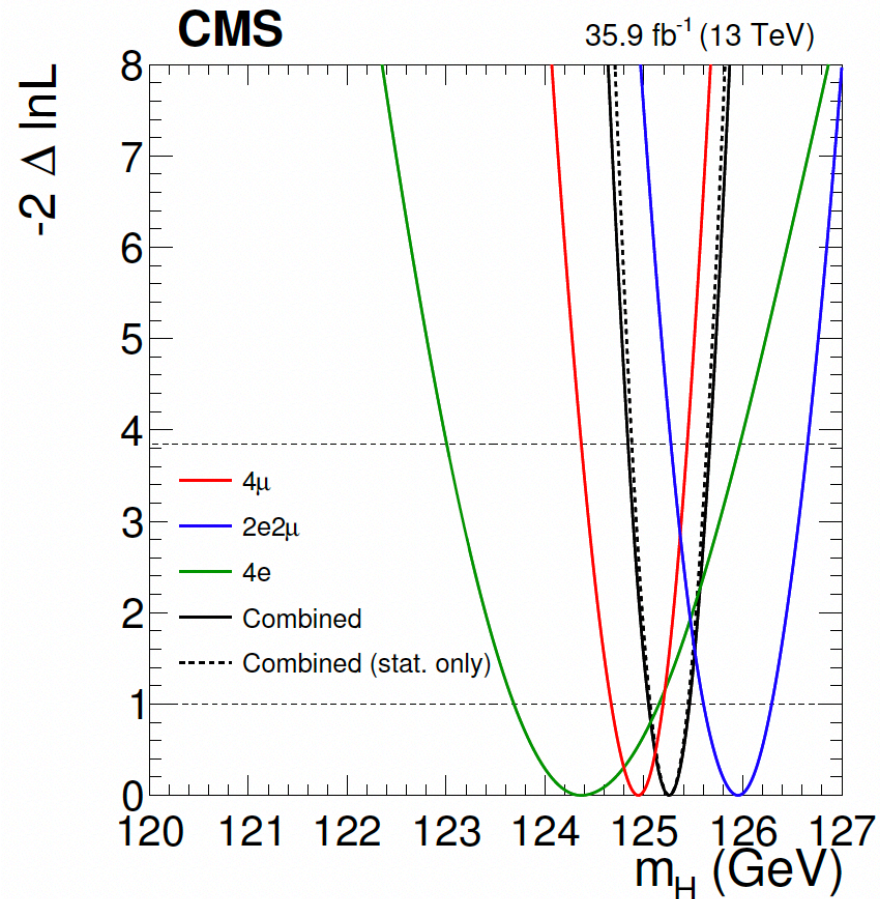
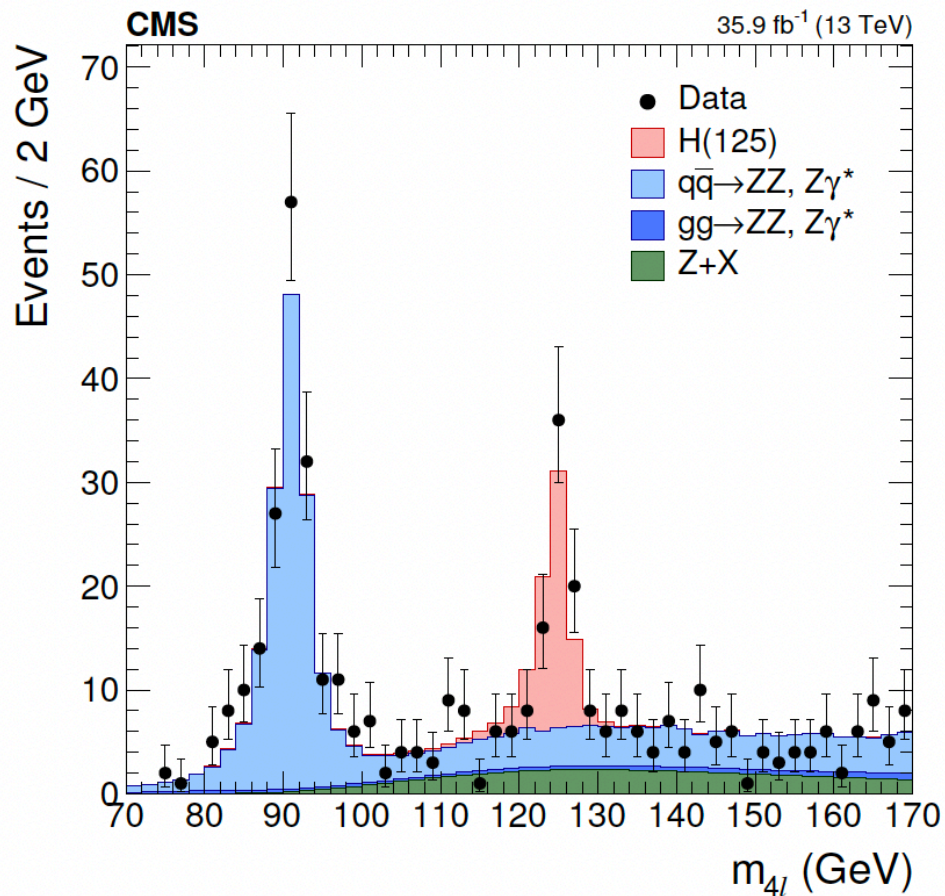


单个未知参数似然区间的确定

左：对称的抛物线型

右：不对称的 $\ln L(\vartheta)$ 函数

似然函数作区间估计



$$\mathcal{P}_{\text{1D stat.}} = \frac{(\mu \cdot S + B)^{n_{\text{obs}}}}{n_{\text{obs}}!} \cdot e^{-(\mu \cdot S + B)} \prod_{i=1}^{n_{\text{obs}}} \frac{\mu \cdot S f_s + B \cdot f_b}{\mu \cdot S + B}$$

最小二乘法作区间估计

- 在最小二乘法中，未知参数 $\vec{\vartheta}$ 的最优估计值是使量：

$$Q^2(\vec{\vartheta}) = \sum_{i=1}^N (Y_i - \eta_i)^2 w_i, \quad w_i = \frac{1}{\sigma_i^2} \quad \text{达到极小的参数值 } \hat{\vartheta}。$$

- 若 $\vec{\vartheta}$ 包含 m 个独立的未知参数，则 $Q^2(\vec{\vartheta})$ 满足 $\chi^2(N - m)$ 分布
- 若参数间存在 k 个独立的约束方程，则 $Q^2(\vec{\vartheta})$ 满足 $\chi^2(N - m + k)$ 分布

- 转换为：

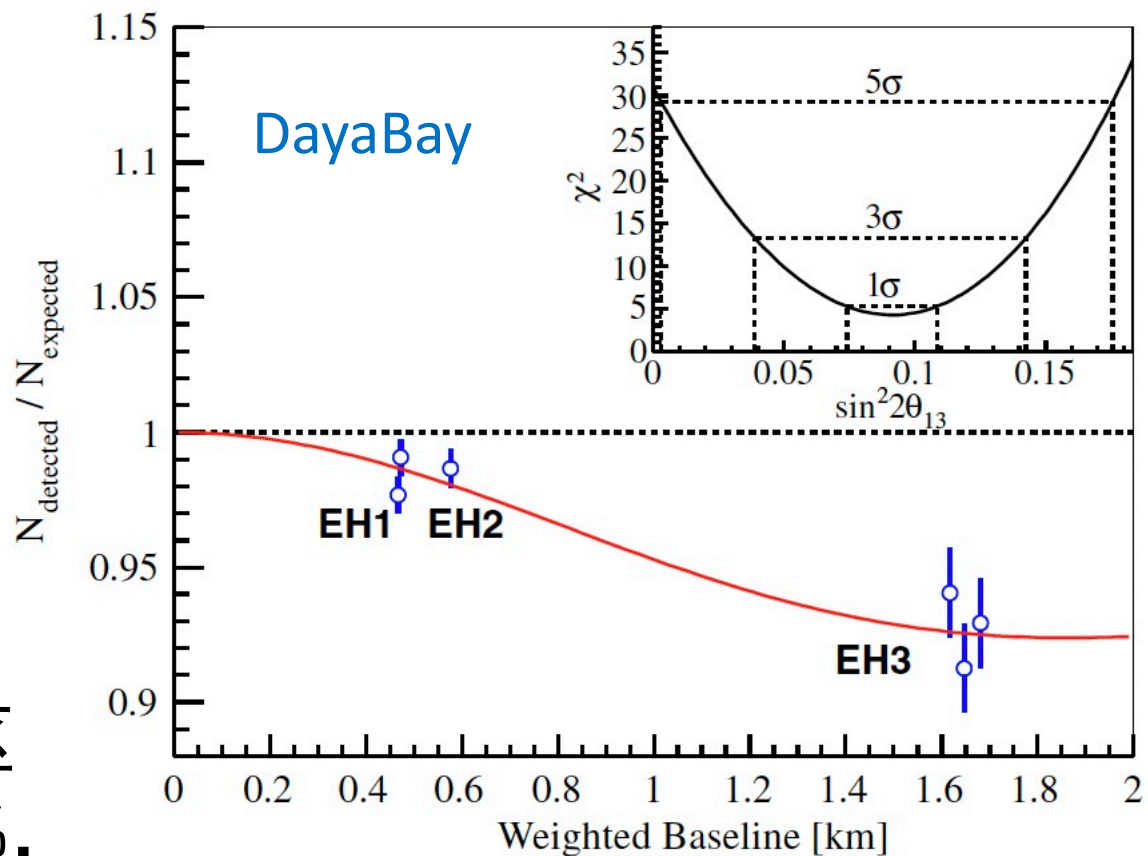
$$Q^2(\vec{\vartheta}) = Q_{\min}^2 + (\vec{\vartheta} - \hat{\vartheta})^T V^{-1}(\hat{\vartheta})(\vec{\vartheta} - \hat{\vartheta}) = Q_{\min}^2 + a$$

如果只包含1个未知参数时， $Q^2(\vartheta) = Q_{\min}^2 + \frac{(\vartheta - \hat{\vartheta})^2}{V(\hat{\vartheta})}$ 是关于 ϑ 的抛物线方程。

最小二乘法作区间估计

$$\chi^2 = \sum_{d=1}^6 \frac{[M_d - T_d (1 + \varepsilon + \sum_r \omega_r^d \alpha_r + \varepsilon_d) + \eta_d]^2}{M_d} + \sum_r \frac{\alpha_r^2}{\sigma_r^2} + \sum_{d=1}^6 \left(\frac{\varepsilon_d^2}{\sigma_d^2} + \frac{\eta_d^2}{\sigma_B^2} \right),$$

$Q^2(\vartheta) = Q_{\min}^2 + a$,
 $a = 1^2, 2^2, 3^2$ 时对应的 ϑ 两个数值区间即是置信水平分别为68.3%, 95.4%, 99.7%的置信区间。



统计分析

参数估计

区间估计

假设检验

假设检验

- 根据检验的对象，可分为参数检验和非参数检验两类
- **参数检验**：分布函数形式已知，检验某个参数是否等于某个规定值
- **非参数检验**：检验随机变量的分布是否具有某个特定的函数形式，或是检验两个总体是否具有相同的分布。

例如：方向相反的高能量正负电子对撞，产生一对 μ 子

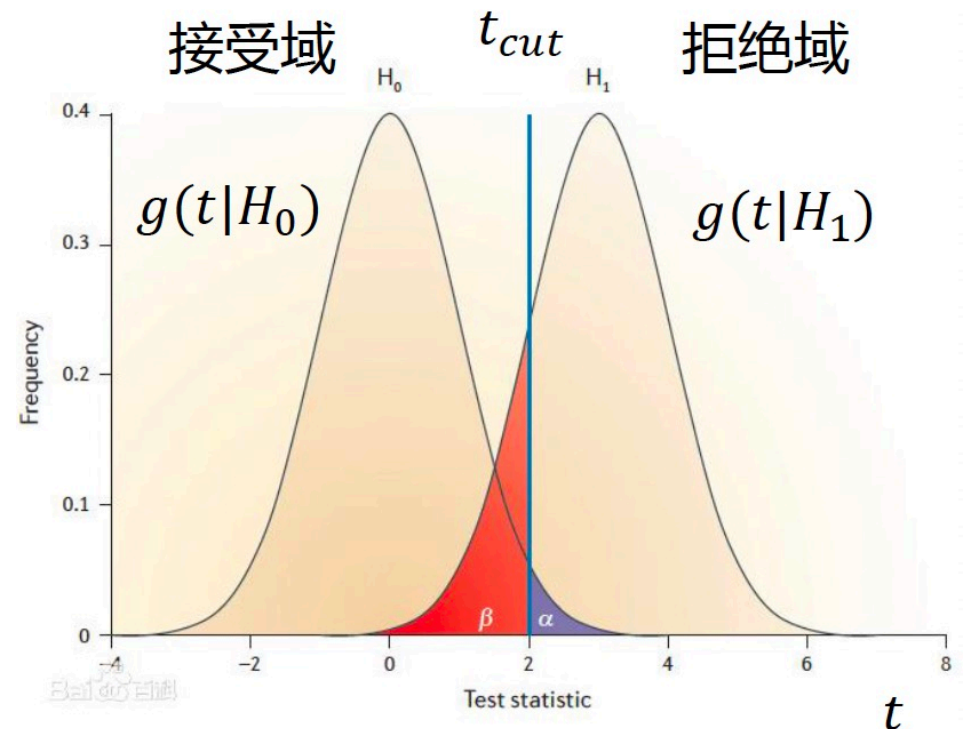


出射的 μ^- 粒子与负电子之间的极角 θ 是一个随机变量，要求确定 θ 的分布是否具有： $C(1 + a\cos^2\theta)$ ， $0 \leq \theta \leq \pi$ 的形式

确定 a 是否等于某个特定值 a_0 ？

假设检验

- 原假设 H_0 : $a = a_0$
- 备择假设 H_1 : $a = a_1$ 或 $a \neq a_0$ 或 $a > a_0$ 或 $a < a_0$
- 一般方法:
 - 1) 构建检验统计量 t , 其概率密度函数形式 $g(t)$ 为 (或间接) 已知
 - 2) 根据给定的概率 α 求出临界值 t_{cut} , 使其满足 $\alpha = \int_{t_{cut}}^{\infty} g(t|H_0)dt$
 - 3) 根据一组观测值计算检验统计量 t_{obs}
 - 4) 如果 $t_{obs} \geq t_{cut}$, 则拒绝 H_0 ; 如果 $t_{obs} < t_{cut}$, 则接受 H_0



似然比检验

- 设某一总体分布 $f(x|\vec{\vartheta})$ ，则 $L(\vec{x}|\vec{\vartheta}) = \prod_{i=1}^n f(x_i|\vec{\vartheta})$ 。假设 H_0 是对某个未知参数加上某个约束条件（如等于某个常数），使得 $\vec{\vartheta}$ 限制在参数空间 Ω 的一个子空间 ω 中。
- $L(\vec{x}|\vec{\vartheta})$ 在空间 Ω 的极大值记为 $L(\hat{\Omega})$ ，在子空间 ω 的极大值记为 $L(\hat{\omega})$
- 定义似然比 $t = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$ 为检验统计量

$$\alpha = \int_0^{t_{cut}} g(t|H_0) dt ; \quad t_{obs} \geq t_{cut}, \text{ 接受 } H_0 ; \quad t_{obs} < t_{cut}, \text{ 拒绝 } H_0$$

假如 H_0 使 r 个参数取固定值，当子样容量很大时， $-2\ln t$ 趋近于 $\chi^2(r)$

信号显著性

- 本底事例期望值是 b ，观测到 n_0 个事例，且 $n_0 > b$
 - 这个实验观测的显著性是多少？
 - 是否发现了一个新效应（导致上述超出）？
 - 或者这仅仅是本底的统计涨落？
- 显著性 S ：假设仅期待本底事例数为 b 的条件下，观测到不少于当前观测事例数的概率：

$$p(n > n_0 | b) = \sum_{k=n_0}^{\infty} p(k|b) = \int_S \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

| significance σ | 1 | 2 | 3 | 4 | 5 |
|---------------------------|-----|------|-------|--------------------|--------------------|
| probability (p -value) | 16% | 2.3% | 0.14% | 3×10^{-5} | 3×10^{-7} |

估计信号显著性的简单方法： S_1

- 对于大统计量情况：

$$S_1 = \frac{\text{signal}}{\sqrt{bkgd}} = \frac{n_{\text{observed}} - b}{\sqrt{b}} = \frac{s}{\sqrt{b}}$$

- 很常用的估计方式，但是对于小统计量的情况，所得结果与正确结果相差较大
- 比如 $b < 100$ 时，这个方法算出来的结果太大（高估了信号显著性）

估计信号显著性的简单方法： S_{cL}

- 更普适的简单方法：

$$S_{cL} = \sqrt{2n_0 \ln(1 + s/b) - 2s}$$

- 源于比较分别基于 $s+b$ 和 b -only两种假设情况下观测到 n_0 事例数的概率，即似然比：

$$S_{cL} = \sqrt{2 \ln Q}, \text{ where } Q = \frac{p(n_0 | s+b)}{p(n_0 | b)}$$

$$Q = \frac{p(n | s+b)}{p(n | b)} = \frac{\frac{(s+b)^n e^{-(s+b)}}{n!}}{\frac{(b)^n e^{-b}}{n!}} = \frac{(s+b)^n e^{-s}}{(b)^n}$$

$$\ln Q = \ln\left[\frac{(s+b)^n e^{-s}}{(b)^n}\right] = n \ln \frac{s+b}{b} - s = n \ln(1 + s/b) - s$$

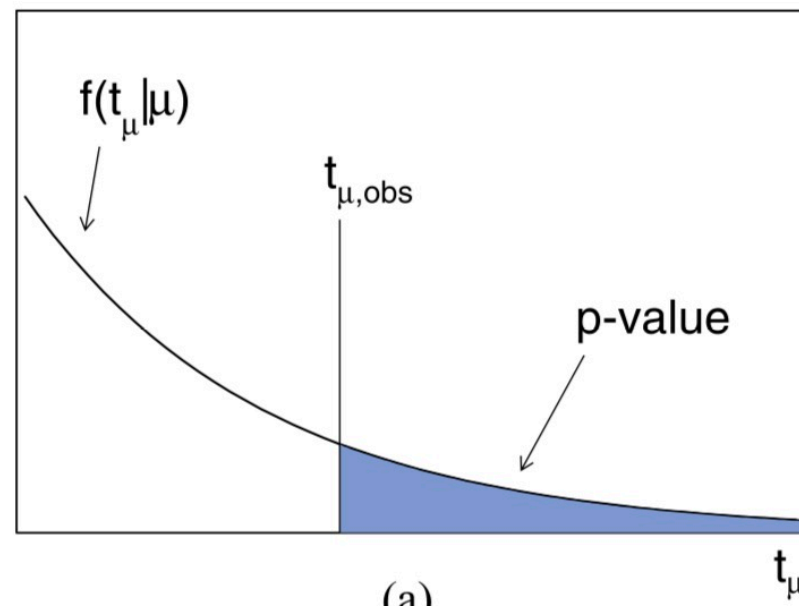
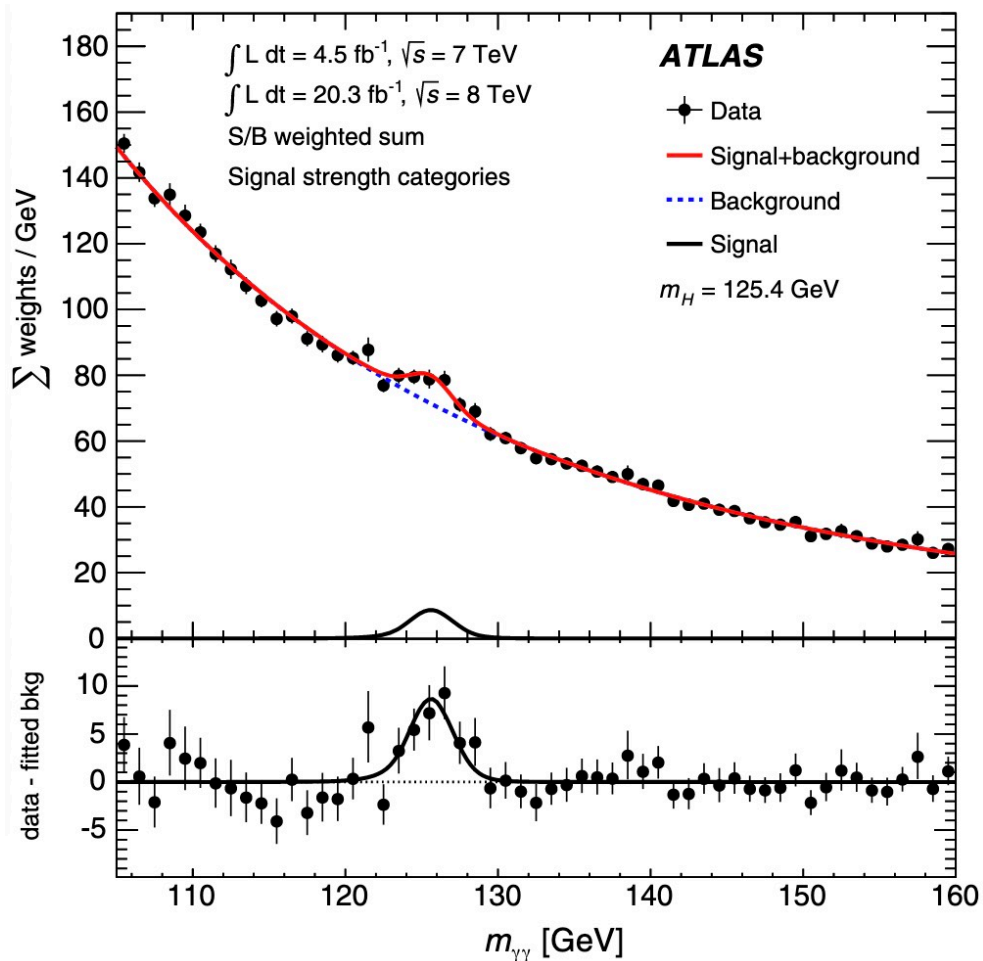
- 所得结果通常与真实显著性非常接近，即使是非常小的统计量，结果也不会偏离超过 0.2σ

估计信号显著性的常规方法：似然比检验

$$\mathcal{L}_c = \text{Pois}(n_c | N_c(\boldsymbol{\theta})) \cdot \prod_{i=1}^{n_c} f_c(m_{\gamma\gamma}^i, \boldsymbol{\theta}) \cdot G(\boldsymbol{\theta})$$

$$N_c = \mu \cdot N_{S,c}(\boldsymbol{\theta}_c^{\text{yield}}, \boldsymbol{\theta}_c^{\text{migr}}, m_H) + N_{\text{bkg},c} + N_{\text{spur},c} \cdot \theta_{\text{spur},c}$$

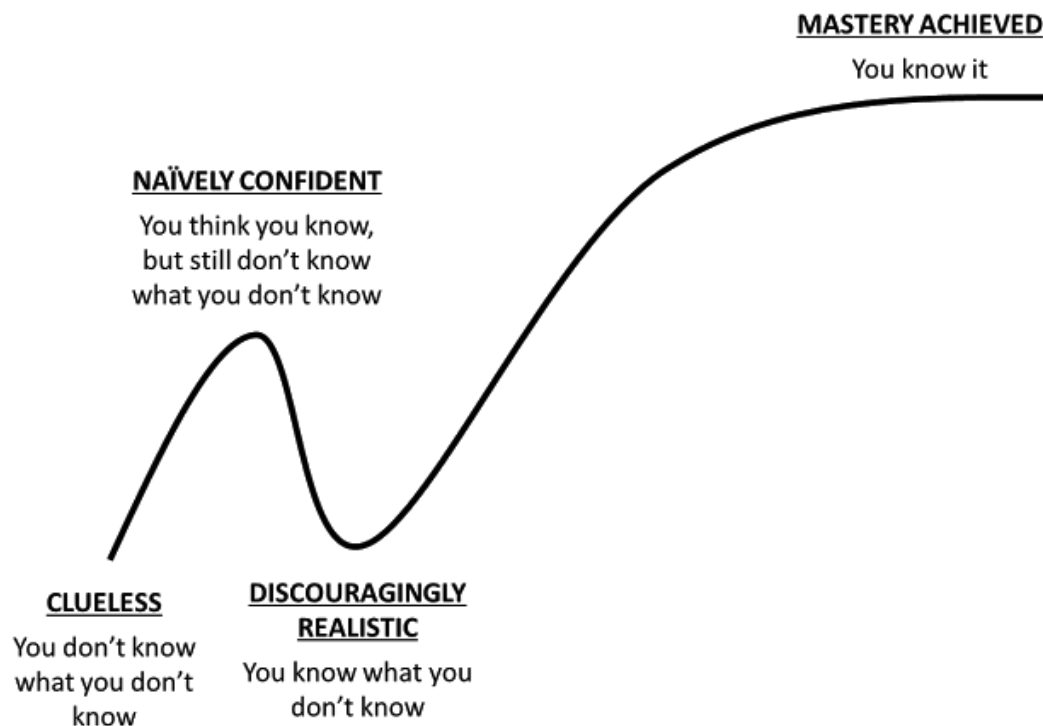
原假设 $H_0: \mu = 0$, 备择假设 $H_1: \mu > 0$



小结

- 介绍了粒子物理实验中概率和统计的一些基本概念和常用方法
- 统计分析是一个物理分析的最后一个环节，也是非常关键的环节
- 任何重要结果的产出都离不开大量的统计检验

- Coverage test
- Local significance,
- Global significance
- Look Elsewhere Effect
- CL_s , CL_{s+b}
-



backup

加权平均值

- 假定自由变量 x 的两个测量值 x_1 和 x_2 的误差分别是 σ_1 和 σ_2 。那么通过**加权平均**，我们可以得到 x 的最佳估计值及其误差：

$$x_m = w_1 x_1 + w_2 x_2, \quad \sigma_m^2 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

$$\text{其中: } w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

- 误差相对很大的测量可以被忽略，因为它权重很小，很难对最终估计值产生影响，也不会改善测量精度（误差）
- 两个一样精度的测量权重一样，加权合并后的误差是单个测量误差的 $1/\sqrt{2}$

最小二乘法

- 如果 $\chi^2/n. d. f$ 远小于1，则在给定的测量误差条件下，拟合好于预期。虽然部件得有问题，但是这时需要仔细检查误差是否被高估，或者是否存在正关联现象（系统误差）
- 如果 $\chi^2/n. d. f$ 远于于1，则有理由怀疑假设的正确性，通常先检查是否低估误差，或者是否存在很大的负关联（系统误差）
- 通过其它cross-checks找出“隐藏”的系统误差

似然函数作区间估计

- 设某一总体分布 $f(x|\vartheta)$ 的一个子样 x_1, \dots, x_n 。参数 ϑ 的不同数值所对应的似然函数值 $L(\vec{x}|\vartheta) = \prod_{i=1}^n f(x_i|\vartheta)$ 可看成是 ϑ 取该数值时可信度的度量，于是参数 ϑ 真值落在区间 $[\vartheta_a, \vartheta_b]$ 内的可信度 γ 可定义为：

$$\gamma = \frac{\int_{\vartheta_a}^{\vartheta_b} L(\vec{x}|\vartheta) d\vartheta}{\int_{-\infty}^{\infty} L(\vec{x}|\vartheta) d\vartheta}$$

由该定义可知，参数 ϑ 真值落在区间 $[-\infty, \infty]$ 内的可信度为1

- 两个统计量 $\vartheta_a(x_1, \dots, x_n)$ 和 $\vartheta_b(x_1, \dots, x_n)$ 满足： $\gamma = P(\vartheta_a \leq \vartheta \leq \vartheta_b)$ ，则称 $[\vartheta_a, \vartheta_b]$ 为参数 ϑ 的概率量 γ 的置信区间。