

粒子物理与核物理实验中的 数据分析

杨振伟
清华大学

第三讲:常用概率密度函数

本讲要点

- 常用的概率密度函数分布的数学形式
- 相应的平均值与方差
- 相关的应用范围

二项分布

N 次独立测量，每次只有成功（概率始终为 p ）或失败（概率为 $1-p$ ）两种可能，得到 n 次成功的概率为

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

平均值：

$$E[n] = \mu = \sum n f = Np$$

适用于仪器探测效率
不确定度的计算

可以证明其满足
归一化条件

$$\sum_n \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} = [(1-p) + p]^N = 1$$

方差：

$$\begin{aligned} V[n] &= \sigma^2 \\ &= E[(n - \mu)^2] \\ &= E[n^2] - E^2[n] \\ &= Np(1-p) \end{aligned}$$

二项分布的适用条件

1. 每次尝试仅有两种可能性;
 2. 每次尝试的成功概率是一样的;
 3. 不同次尝试的结果是独立的。
- 伯努利试验

考虑驾车人被停车检查有否不佩戴安全带的情况是否为一个伯努利试验。

两种结果：佩戴与不佩戴！

如果对所有车都一样，那么驾车人都有同样的概率不佩戴安全带！？（不同年龄人群都是一样的吗？）

检查不同驾车人都佩戴安全带，结果应该是独立的！？
（对于同时同地的前后驾车人都是一样的吗？）

因此，根据数据采样情况，才能分清是否为伯努利试验，才能决定能否应用二项分布。

举例：在效率误差估计中的应用

■ 多层阻性板室(MRPC)的探测效率



闪烁体**1**与**2**同时击中给出



穿过**MRPC**的粒子数 **N**

MRPC记录的击中数目 **N'**

$$p = \frac{N'}{N}$$
$$\Delta p = \frac{\Delta N'}{N} = \sqrt{\frac{p(1-p)}{N}}$$

MRPC探测效率
测量值及其不确定度

二项分布指导决策

我们为大亚湾实验研制生产触发电子学板。按设计在一年内需要修理的电路板为**10%**。如果在实验所需的**20**块板中有**5**块在第一年使用时需要进行维修，那么这种故障率是否可以接受？

解答：首先找出在一年内**20**块板中有**5**块或更多出现问题需要进行维修的概率

$$\begin{aligned}\sum_{n=5}^{20} f(n; 20, 0.1) &= 1 - \sum_{n=0}^4 f(n; 20, 0.1) \\ &= 1 - \sum_{n=0}^4 \frac{20!}{n!(20-n)!} 0.1^n (1-0.1)^{20-n} \\ &= 1 - 0.9568 = 0.0432\end{aligned}$$

5块以上发生故障的概率非常少，所以板的质量不能接受。

从二项到多项分布

类似于二项分布, 但允许结果的可能性 m 大于两种, 概率为

$$\begin{cases} \vec{p} = (p_1, p_2, \dots, p_m) \\ \sum_{i=1}^m p_i = 1 \end{cases} \quad \begin{array}{l} \text{尝试 } N \text{ 次, 结果为} \\ \text{可能性1: } n_1 \\ \text{可能性2: } n_2 \\ \dots \end{array} \quad \longrightarrow \quad \vec{n} = (n_1, n_2, \dots, n_m)$$

得到 (n_1, n_2, \dots, n_m) 概率为

$$f(\vec{n}) = \frac{N!}{n_1! n_2! \dots n_m!} p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}$$

$$\text{平均值: } E[n_i] = Np_i$$

$$\text{方差: } V[n_i] = Np_i(1 - p_i)$$

$$\text{协方差: } V_{ij} = -Np_i p_j \quad (i \neq j)$$

适用于直方图频数
不确定度估计。

泊松分布

泊松分布是二项分布在 $N \rightarrow \infty$, $p \rightarrow 0$ 和 $Np = \text{常数 } \nu$ 的极限形式。

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu}$$

平均值：

$$E[n] = \mu = \sum n f = \nu$$

著名的统计不确定度估计式

$$n \pm \sqrt{n}$$

方差：

$$\begin{aligned} V[n] &= \sigma^2 = E[(n - \mu)^2] \\ &= E[n^2] - E^2[n] \\ &= \nu \end{aligned}$$

泊松分布式是二项分布的近似

概率的第三公理：如果 A_1, A_2, A_3, \dots 是在空间 S 中互斥事例的有限或无限序列，则

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

$$\sum_{n=0}^{\infty} \frac{e^{-\nu} \nu^n}{n!} = e^{-\nu} \sum_{n=0}^{\infty} \frac{\nu^n}{n!} = e^{-\nu} e^{\nu} = 1 \quad (\text{函数为 } e^x \text{ 的麦克劳林公式})$$

$$\begin{aligned} f(n; N, \nu) &= \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{n}{N}\right)^{N-n} \\ &= \frac{N(N-1)(N-2)\dots(N-n+1)}{n! N^n} \nu^n \left(1 - \frac{n}{N}\right)^{N-n} \\ &= \frac{(1 - \frac{1}{N})(1 - \frac{2}{N})\dots(1 - \frac{n-1}{N})}{n!} \nu^n \left(1 - \frac{\nu}{N}\right)^{N-n} \end{aligned}$$

$$\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\dots\left(1 - \frac{n-1}{N}\right) \xrightarrow{N \rightarrow \infty} 1$$

$$\begin{aligned} \left(1 - \frac{\nu}{N}\right)^{N-n} &= \left[\left(1 - \frac{\nu}{N}\right)^{N/\nu}\right]^{\nu} \\ \times \left(1 - \frac{\nu}{N}\right)^{-n} &= \lim_{\substack{x = -\frac{N}{\nu} \rightarrow -\infty \\ \nu}} \left[\left(1 + \frac{1}{x}\right)^{-x}\right]^{\nu} \\ &\rightarrow e^{-\nu} \end{aligned}$$

举例：光电倍增管暗电流影响

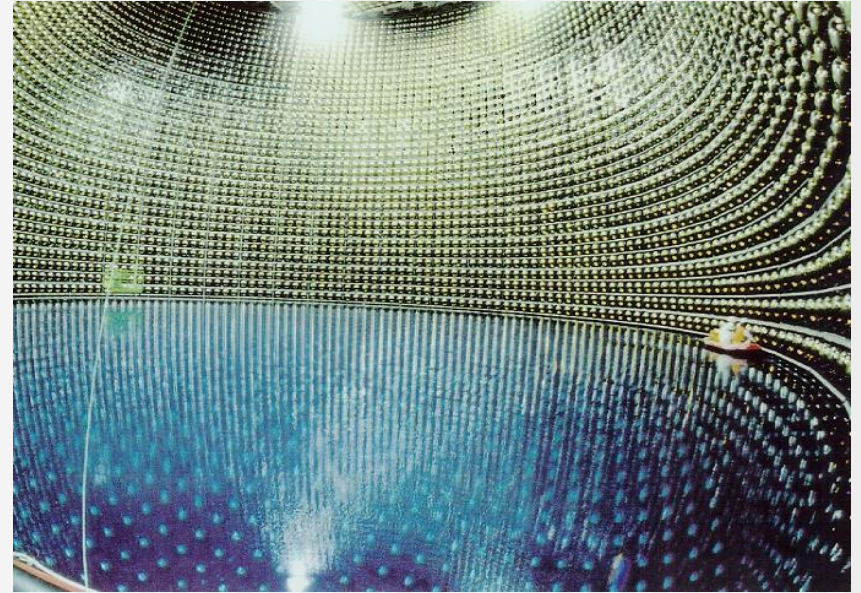
- 在有11146根**PMT**的探测器中，已知每根**PMT**暗电流产生的误响应为3.5kHz。求探测器在任意总长度为500μs时间段观察到每隔10ns **PMT**误击中数目分别为5和6的总次数

在10ns间隔观测到**PMT**误击中的平均数目为

$$10^{-8} \times 11146 \times 3500 = 0.4$$

500μs

平均数



日本超级神冈中微子探测器

$$= 5 \Rightarrow 5 \times 10^4 \times \frac{0.4^5}{5!} \times e^{-0.4} \cong 3(\text{次})$$

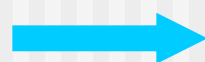
$$= 6 \Rightarrow 5 \times 10^4 \times \frac{0.4^6}{6!} \times e^{-0.4} \cong 0.2(\text{次})$$

这一结论影响到我们在数据分析中应采取的对策。

二项分布与泊松分布

- 假设一学生站在路边想搭便车。过路的汽车平均频率为每分钟一辆，服从泊松分布。而每辆车让搭便车的概率为1%，计算该学生在过了60辆车以后还未能搭上车的可能性

$$N=60, p=0.01, r=0$$



特点：N大 p小

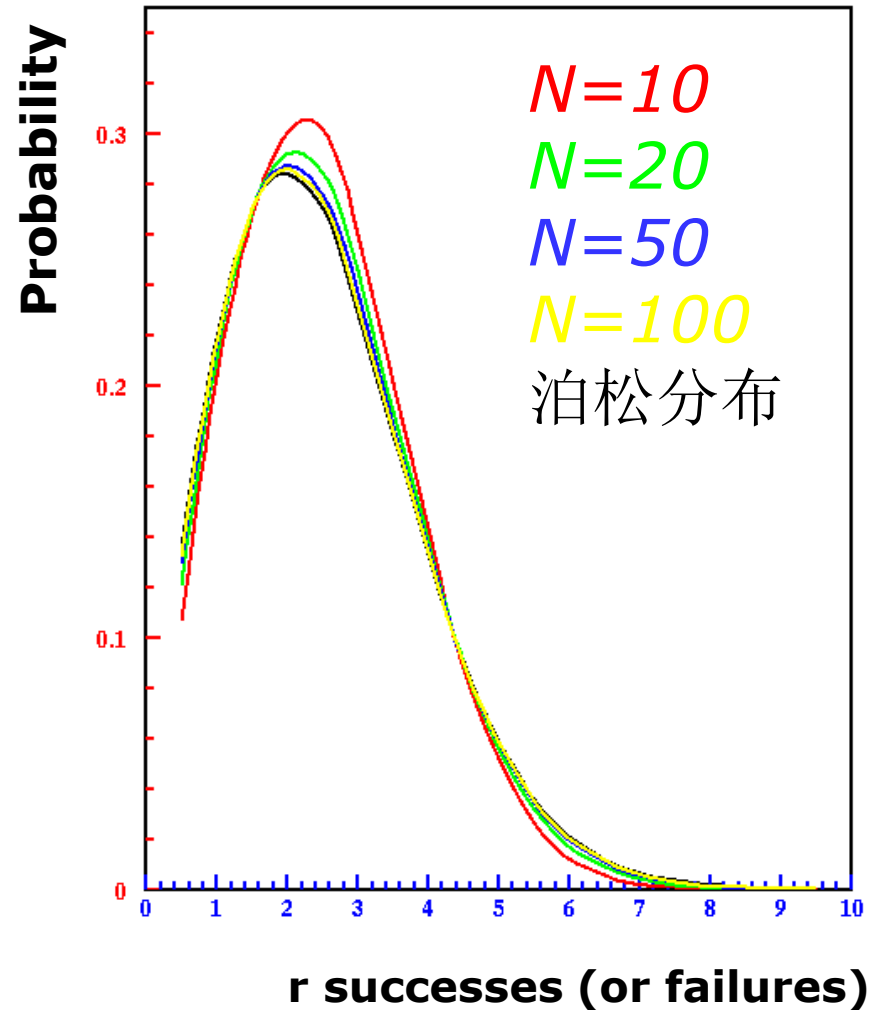
$$\text{根据二项分布: } \frac{60!}{0!(60-0)!} 0.01^0 (1-0.01)^{60-0} = 0.5472$$

$$\text{根据泊松分布: } e^{-60 \times 0.01} \frac{(60 \times 0.01)^0}{0!} = 0.5488$$

泊松分布是二项分布的近似。

泊松分布是二项分布的近似

例如：对于以平均值为**2**的泊松分布而言，相当于二项分布中的 $Np=2$ 。当 N 值增大时，为了保持 Np 不变， p 值相应减小。可以从右图看出，当 N 大于**50**时，两种分布的区别几乎可以忽略。



直方图中的误差处理



一个直方图可看成与

1. 一个事例总数满足泊松分布和在每个区间得到 n_1, n_2, n_3, \dots 事例数为多项分布有关;
2. 或者是直方图中每个区间互相独立的泊松分布有关。

$$\begin{aligned}\Delta N &= \sqrt{N} \\ \text{或} \\ (\Delta N)^2 &= (\Delta n_1)^2 + (\Delta n_2)^2 + (\Delta n_3)^2 + \dots \\ &= n_1 + n_2 + n_3 + \dots \\ &= N\end{aligned}$$

注意：当 $N < 5$ 时误差估计会有很大的偏差。

高斯或正态分布

高斯函数具有连续性与对称性，概率密度为

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{记为 } N(\mu, \sigma)$$

平均值：

$$E[x] = \int xP(x)dx = \mu$$

方差：

$$\begin{aligned} V[x] &= E[(x-\mu)^2] \\ &= E[x^2] - E^2[x] \\ &= \sigma^2 \end{aligned}$$

在所有统计问题扮演核心角色，应用于所有科学研究领域所涉及的分布。测量不确定度，特别是仪器误差通常用高斯函数来描述其概率分布。即使在应用中可能有不恰当的地方，仍然可提供与实际情况相近的很好近似。

中心极限定理

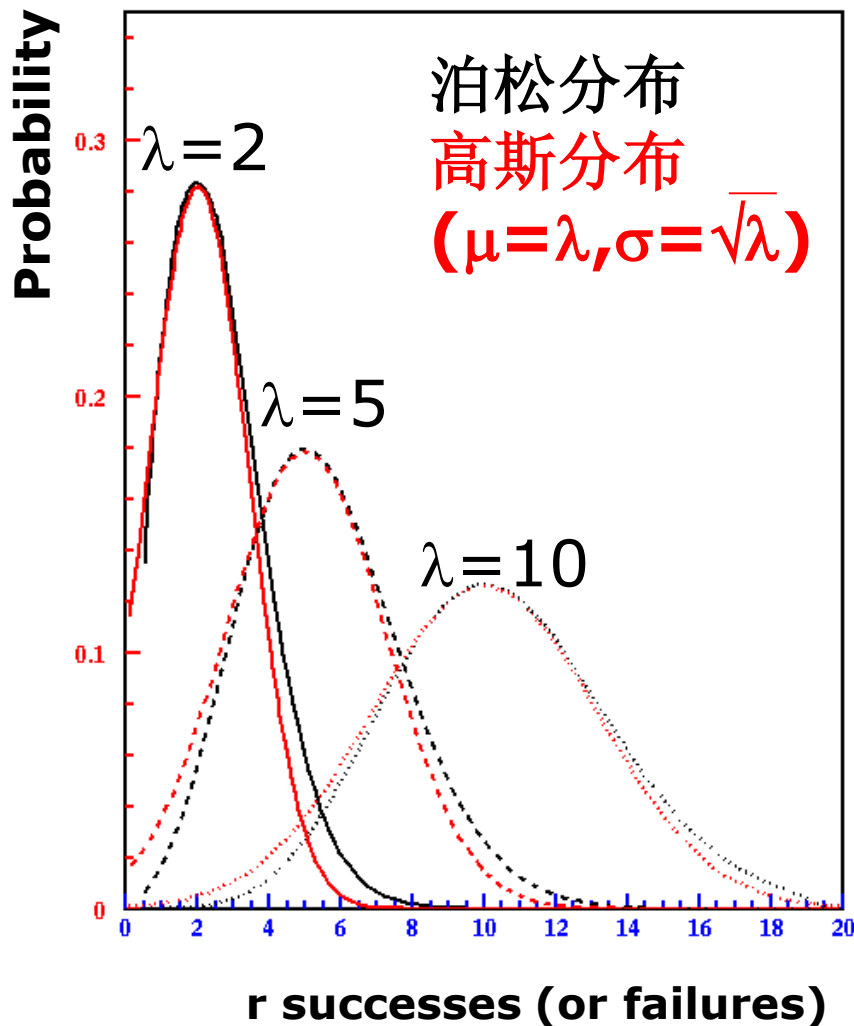
对于 n 个独立的随机变量 x_i , 如果每个 x_i 都服从平均值为 μ_i 和有限的方差 σ_i^2 分布, 那么变量

$$\frac{\left(\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i \right)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \xrightarrow{n \rightarrow \infty} \text{趋于 } N(0,1) \text{ 的正态分布}$$

因此, 如果

$$y = \sum_{i=1}^n x_i \Rightarrow E[y] = \sum_{i=1}^n \mu_i, \quad V[y] = \sum_{i=1}^n \sigma_i^2$$

高斯分布与泊松分布



- 泊松分布只有非负整数定义。
- 高斯分布是连续且可延伸到正负无穷。
- 当泊松分布的平均值越大，与高斯分布的区别就越小。
- 实际应用时，当计数或事例数大于**5**时，可认为误差满足高斯分布。

多维高斯分布

对于随机变量 $\vec{x} = (x_1, \dots, x_n)$ 其多维高斯函数概率密度为

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T V^{-1}(\vec{x} - \vec{\mu})\right]$$

相应的平均值与协方差为 $E[x_i] = \mu_i$, $\text{cov}[x_i, x_j] = V_{ij}$

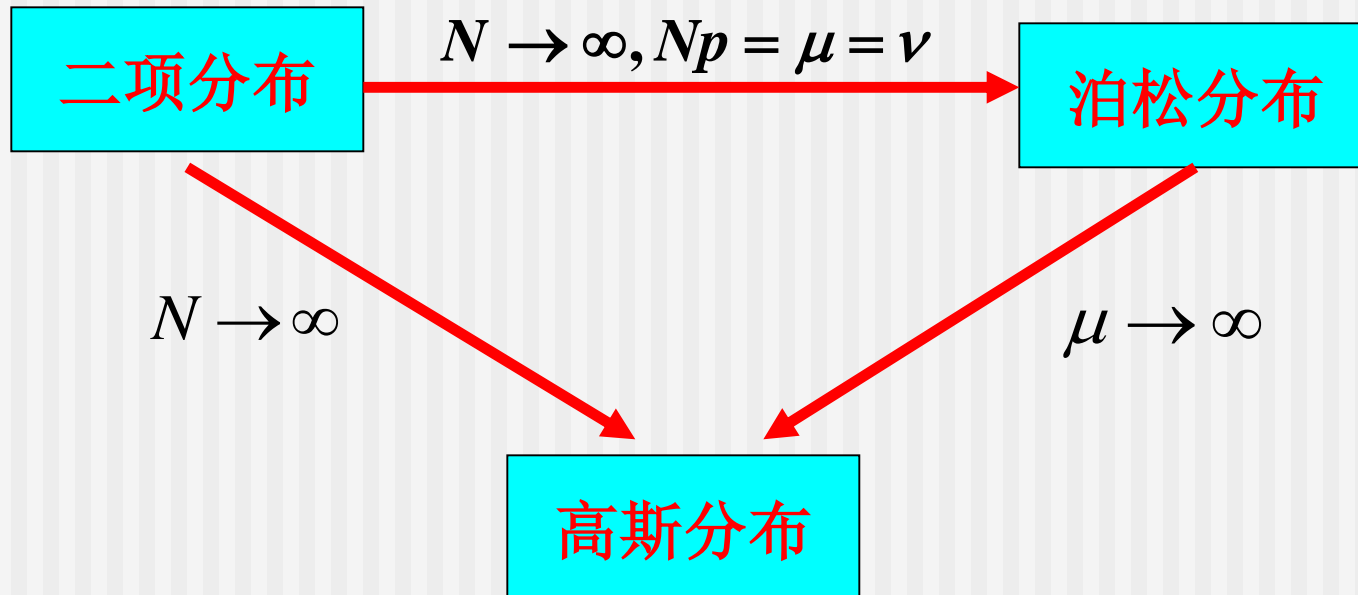
对于二维情形, 其概率密度函数可表示为

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \quad \rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right)\right]\right\}$$

二项、泊松与高斯分布的联系

$$f(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$f(n) = \frac{\nu^n}{n!} e^{-\nu}$$



$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

对数正态分布

如果连续变量 y 是具有均值为 μ 方差为 σ^2 的高斯量，那么 $x = e^y$ 服从对数正态分布。

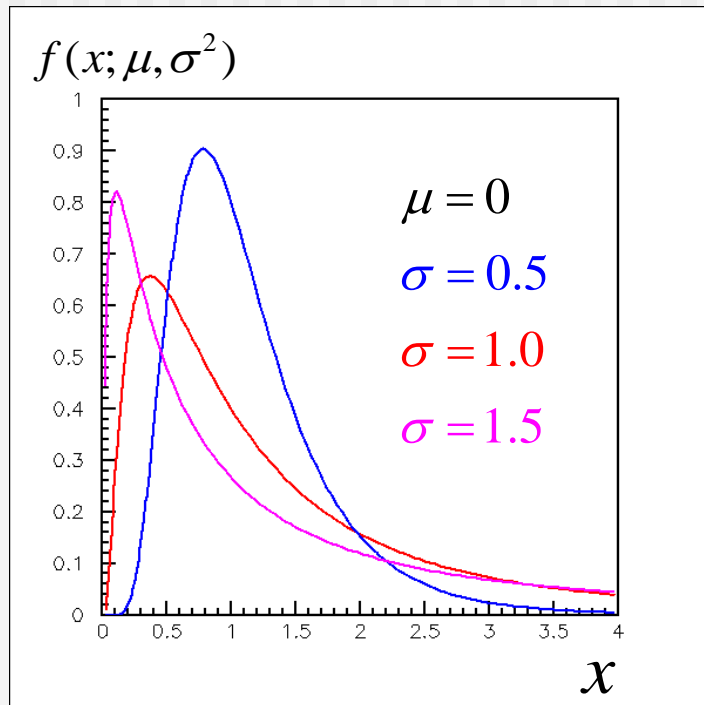
$$f(x; \mu, \sigma^2) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(\frac{-(\log x - \mu)^2}{2\sigma^2}\right) & \text{对于 } x > 0, \sigma > 0 \\ 0 & \text{其它情况} \end{cases}$$

$$\text{平均值: } E[x] = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

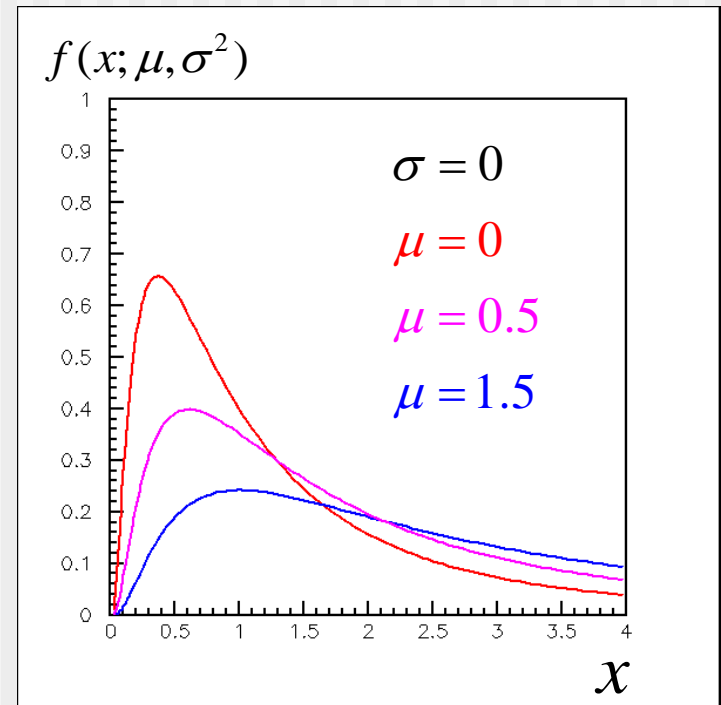
$$\text{方差: } V[x] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$$

不同均值和方差的对数正态分布

同均值不同方差



同方差不同均值



对数正态分布表示一个随机变量其对数服从正态（高斯）分布，提供了一个模型处理类似涉及许多小的倍增误差过程的误差。也适用于观测值是一个正比于过去观测的随机变量。

指定区间的对数正态分布计算

如果需要估计服从对数正态分布的随机变量在区间 $(0 < a < b)$ 的概率值，需要计算

$$\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(\frac{-(\log x - \mu)^2}{2\sigma^2}\right) dx$$

对积分做代换 $y = \log(x)$ ，则可以得到所求的概率值

$$\begin{aligned} & \int_{\log a}^{\log b} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - \mu)^2}{2\sigma^2}\right) dy \quad \text{令 } t = \frac{y - \mu}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log b - \mu}{\sigma}} \exp\left(-\frac{t^2}{2}\right) dt - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log a - \mu}{\sigma}} \exp\left(-\frac{t^2}{2}\right) dt \\ &= F\left(\frac{\log b - \mu}{\sigma}\right) - F\left(\frac{\log a - \mu}{\sigma}\right) \end{aligned}$$

可从正态分布 $N(0,1)$ 表查中出积分值。

对数正态分布用于风险分析

核电站工程师必须采用模型来估计支撑蒸汽发电机的强度，以防止由于地震峰值加速度造成的破坏。专家的意见建议该强度的对数是具有 $\mu=4.0$ 和 $\sigma^2=0.09$ 正态分布。试估计但峰值加速度为 33 时，支撑系统依然可以承受的概率。

解答：

$$1 - F\left(\frac{\log(33) - 4.0}{0.30}\right) = 1 - F(-1.68) = 0.9535$$

或者说系统崩溃瓦解的概率为

$$F\left(\frac{\log(33) - 4.0}{0.30}\right) = F(-1.68) = 0.0465$$

注：大亚湾中微子实验隧道爆破最大允许当量也是采用类似方法估计。

均匀分布

在区间 (a, b) 上均匀分布的连续随机变量 x ，其概率密度函数为

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{当 } \alpha \leq x \leq \beta \\ 0 & \text{其它} \end{cases}$$

$$\text{平均值: } E[x] = \mu = \frac{\alpha + \beta}{2}$$

均匀分布是用蒙特卡罗模拟随机现象的基础。

$$\text{方差: } V[x] = \sigma^2(x) = \frac{(\beta - \alpha)^2}{12}$$

指数分布

对于连续变量 x ($0 \leq x < \infty$) 指数分布,

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

$$\text{平均值: } E[x] = \frac{1}{\xi} \int_0^{\infty} x e^{-x/\xi} dx = \xi$$

$$\text{方差: } V[x] = \frac{1}{\xi} \int_0^{\infty} (x - \xi) e^{-x/\xi} dx = \xi^2$$

常用于描述粒子寿命。

χ^2 - 分布

如果 x_1, \dots, x_n 是相互独立的高斯随机变量，按下列方式求和

$$z = \sum_{i=1}^n (x_i - \mu_i)^2 / \sigma_i^2$$

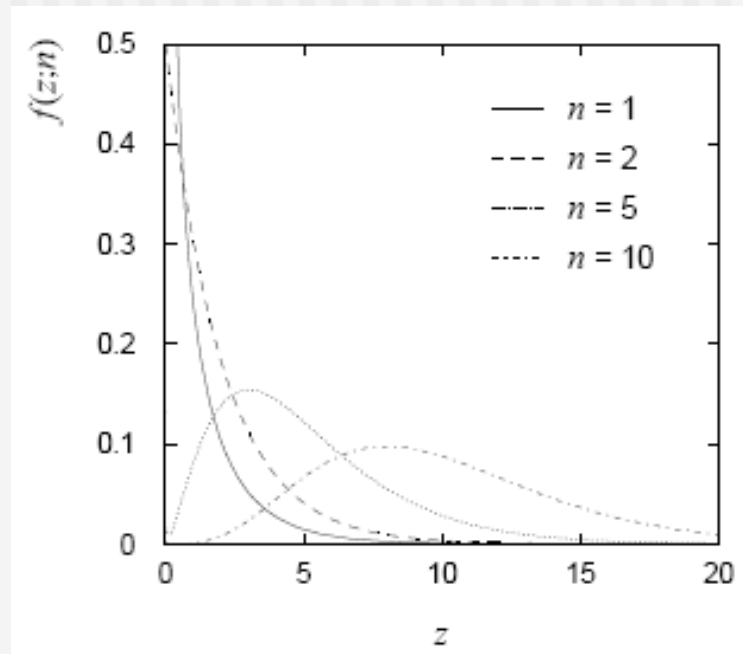
服从自由度为 n 的 χ^2 的 **p.d.f** 为

$$f(z; n) = \frac{z^{n/2-1}}{2^{n/2} \Gamma(n/2)} e^{-z/2}, (z \geq 0)$$

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

平均值: $E[z] = \mu = n$

方差: $V[z] = \sigma^2 = 2n$



χ^2 -分布通常用来检验假设与实际情况的符合程度。

柯西(布莱特-魏格纳)分布

对于连续随机变量 x 的柯西**p.d.f.**为

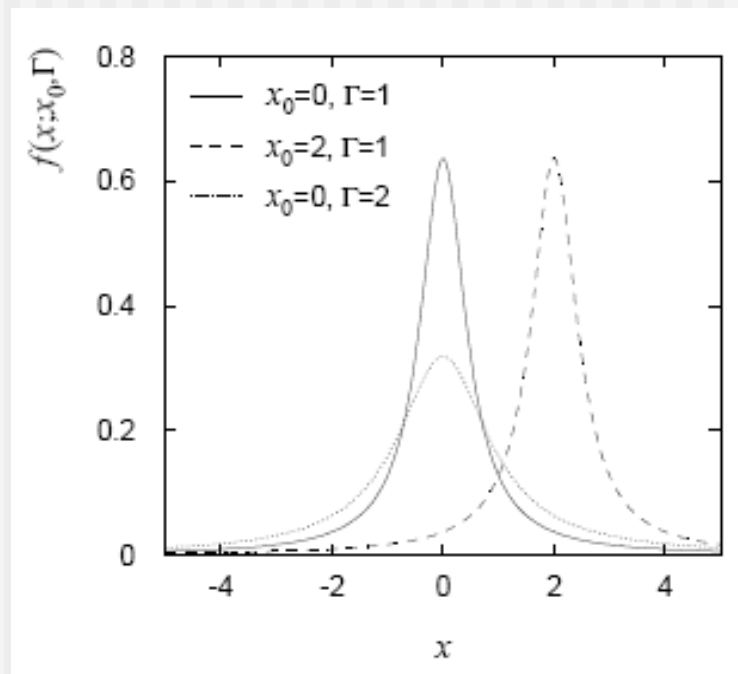
$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

是布莱特-魏格纳**p.d.f.**的一个特例

$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x-x_0)^2}$$

其中, x_0 , Γ =(半高)宽度

在粒子物理中, 常用于描述“共振态”粒子的不变质量分布。



朗道分布

对于具有速度为 $\beta=v/c$ 的带电粒子穿过一层厚度为 d 的物质，其能量损失 Δ 服从朗道p.d.f

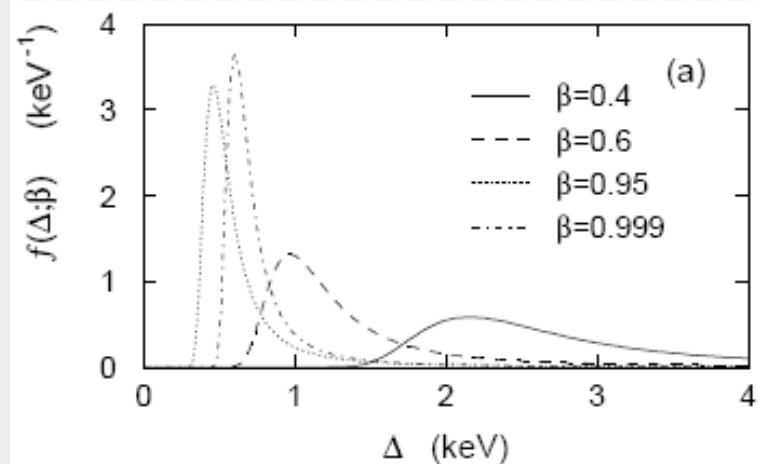
$$f(\Delta; \beta) = \frac{1}{\xi} \phi(\lambda),$$

$$\phi(\lambda) = \frac{1}{\pi} \int_0^\infty \exp(-u \log u - \lambda u) \sin(\pi u) du,$$

$$\lambda = \frac{1}{\xi} \left[\Delta - \xi \left(\log \frac{\xi}{\varepsilon'} + 1 - \frac{1}{\sqrt{1-\beta^2}} \right) \right],$$

$$\xi = \frac{2\pi N_A e^4 z^2 \rho \Sigma Z}{m_e c^2 \Sigma A} \frac{d}{\beta^2},$$

$$\varepsilon' = \frac{I^2 (1-\beta^2) \exp(\beta^2)}{2m_e c^2 \beta^2}, \quad I = \text{平均激发能}$$



厚度 d 增大时，趋于正态分布。

常用于描述粒子的电离能损或能量沉积。

小结

- 二项分布：探测效率，分支比
- 多项分布：直方图的统计误差
- 泊松分布：一定通量下的事例估计
- 均匀分布：常用于蒙特卡罗模拟
- 指数分布：粒子固有衰变时间
- 高斯分布：分辨率
- 多维高斯分布：测量结果的相关性
- 对数正态分布：处理涉及有许多小的倍增误差贡献的误差
- χ^2 分布：拟合结果好坏检验
- 柯西(布莱特-魏格纳)分布：共振态质量与宽度
- 朗道分布：粒子的电离能损

所有分布都可以在
ROOT平台中给出！