粒子物理与核物理实验中的数据分析

陈少敏
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

二项式分布

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

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

可以证明其满足归一化条件

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

平均值：

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

方差：

$$V[n]=\sigma^2$$
$$=E[(n-\mu)^2]$$
$$=E[n^2]-E^2[n]$$
$$=Np(1-p)$$

适用于仪器探测效率误差的计算

二项式分布均值证明

$$E[n]=\sum nf=\sum_{r=0}^{N} n\cdot\frac{N!}{n!(N-n)!}p^{n}(1-p)^{N-n}$$

$$\frac{n}{n!}=\frac{1}{(n-1)!}$$

$$N!=N(N-1)!$$

$$\mu=Np\sum_{n=1}^{N}\frac{(N-1)!}{(n-1)!(N-n)!}p^{n-1}(1-p)^{N-n}$$

$$n'=n-1,\qquad N'=N-1$$

关键点：均值计算
要求 N 趋于无穷大。

$$\mu=Np\sum_{n'=0}^{N'}\frac{(N'-1)!}{(n'-1)!(N'-n')!}p^{n'}(1-p)^{N'-n'}=Np\sum f=Np$$

二项式分布的适用条件

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式分布。

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
测量值及其误差

二项式分布指导决策

我们为大亚湾实验研制生产触发电子学版。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

$$\sum_{n=5}^{20} f(n;20,0.1) = 1 - \sum_{n=0}^{4} f(n;20,0.1)$$

$$= 1 - \sum_{n=0}^{4} \frac{20!}{n!(20-n)!} 0.1^n (1-0.1)^{20-n}$$

$$= 1 - 0.9568 = 0.0432$$

允许有**5**块以上有故障的概率非常小，故板的质量不能接受。

从二项式到多项式分布

类似于二项式分布，但允许结果的可能性 m大于两种，概率为

$$\begin{cases} \vec{p} = (p_1, p_2 \ldots, p_m) \\ \sum_{i=1}^{m} p_i = 1 \end{cases}$$

尝试N次，结果为

可能性1：n_1

可能性2：n_2

…

$$\Rightarrow \quad \vec{n} = (n_1, n_2, \ldots, n_m)$$

得到$(n_1, n_2, \ldots, n_m)$概率为

$$f(\vec{n}) = \frac{N!}{n_1! n_2! \ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_3}$$

$$\text{平均值}: E[n_i] = Np_i$$

$$\text{方差}: V[n_i] = Np_i(1 - p_i)$$

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

适用于直方图频数误差估计。

泊松分布

泊松分布是二项式分布在 $N\to\infty$，$p\to 0$ 和 Np=常数 ν 的极限形式。

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

平均值：

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

著名的统计误差估计式

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

方差：

$$V[n]=\sigma^2=E[(n-\mu)^2]$$
$$=E[n^2]-E^2[n]$$
$$=\nu$$

泊松分布式是二项式分布的近似

概率的第三公理：如果 A_1，A_2，A_3，… 是在空间 S 中互斥事例一个有限或无限的序列，则

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

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

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

$$= \frac{N(N-1)(N-2)\ldots(N-n+1)}{n!N^n}\nu^n\left(1-\frac{n}{N}\right)^{N-n}$$

$$= \frac{(1-\frac{1}{N})(1-\frac{2}{N})\ldots(1-\frac{n-1}{N})}{n!}\nu^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

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

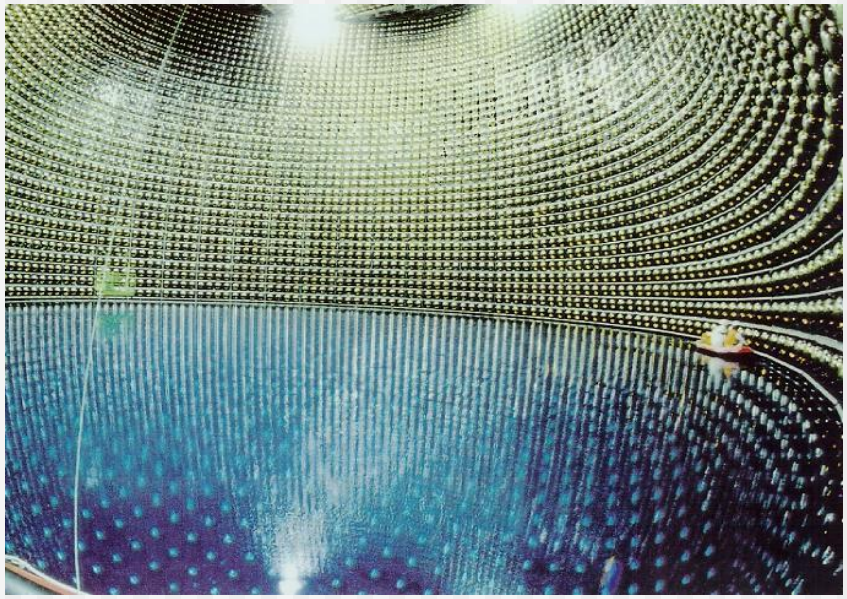
$$\left(1-\frac{\nu}{N}\right)^{N-n} = \left[\left(1-\frac{\nu}{N}\right)^{N/\nu}\right]^{\nu}$$

$$\times\left(1-\frac{\nu}{N}\right)^{-n} = \lim_{x=-\frac{N}{\nu}\to-\infty}[(1+\frac{1}{x})^{-x}]^{\nu}$$

$$\to e^{-\nu}$$

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



日本超级神冈中微子探测器

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

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

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

这一结论影响到我们在数据分析中应采取的对策。

二项式分布与泊松分布

- 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

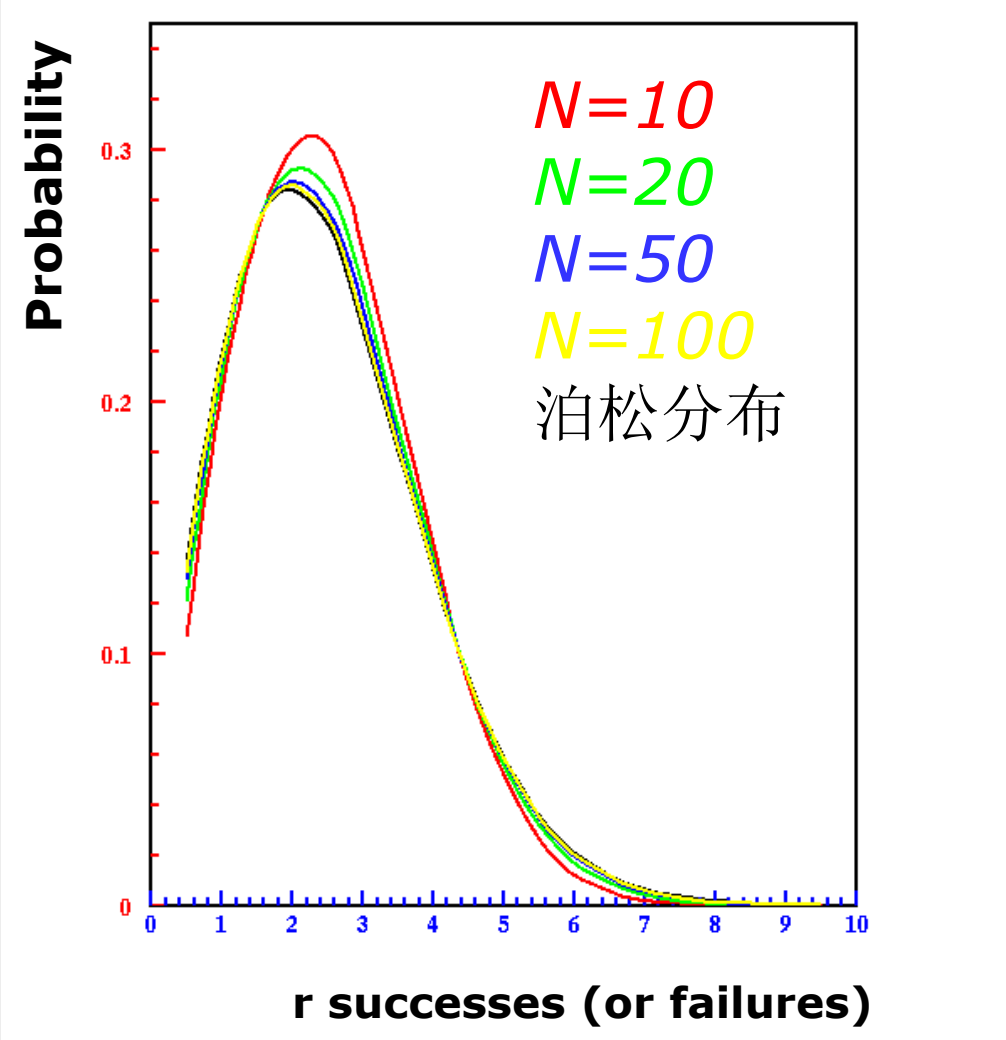
根据二项式分布：$\dfrac{60!}{0!(60-0)!}0.01^0(1-0.01)^{60-0}=0.5472$

根据泊松分布：$e^{-60\times0.01}\dfrac{(60\times0.01)^0}{0!}=0.5488$

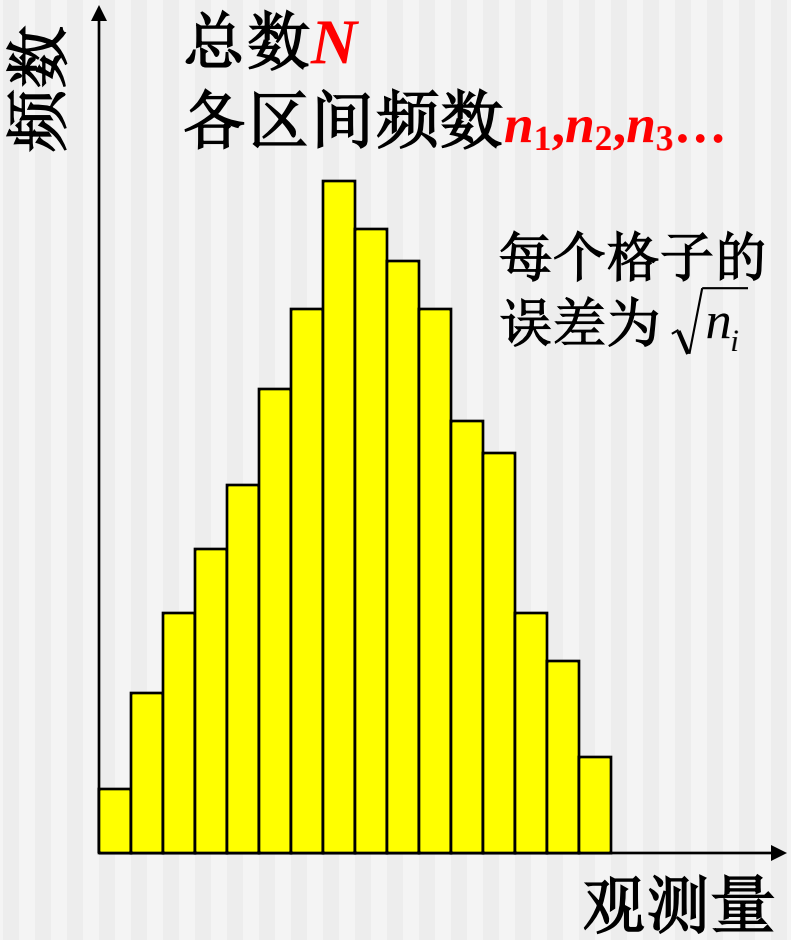
泊松分布是二项式分布的近似。

泊松分布是二项式分布的近似

例如：对于以平均值为2的泊松分布而言，相当于二项式分布中的$Np=2$。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$n_1,n_2,n_3\ldots$事例数为多项式分布有关；
2. 或者是直方图中每个区间互相独立的泊松分布有关。

$$\Delta N = \sqrt{N}$$

或

$$(\Delta N)^2 = (\Delta n_1)^2 + (\Delta n_2)^2 + (\Delta n_3)^2 + \ldots$$
$$= n_1 + n_2 + n_3 + \ldots$$
$$= N$$

注意：当 $N < 5$ 时误差估计会有很大的偏差。

分数电荷夸克的寻找

实验原理：当高速带电粒子穿过云室时，所产生的电离会引起液滴的形成，它们在云室中形成可以观测的径迹。假设每单位径迹形成一粒液滴的概率是常数，并且正比于粒子电荷的开方。由于几乎所有粒子穿过云室时都带一个电荷单位，可以通过寻找一个径迹引起的单位液滴数目，明显比所有径迹的平均值要低的粒子的存在证据，来从实验上确立“夸克”这种理论上认为只带2/3单个（质子）电荷的存在证据。

VOLUME 23, NUMBER 12 PHYSICAL REVIEW LETTERS 22 SEPTEMBER 1969

EVIDENCE OF QUARKS IN AIR-SHOWER CORES*

C. B. A. McCusker and I. Cairns
Cornell-Sydney University Astronomy Center, Physics Department, The University of Sydney, Sydney, Australia
(Received 3 September 1969)

In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

泊松分布诠释

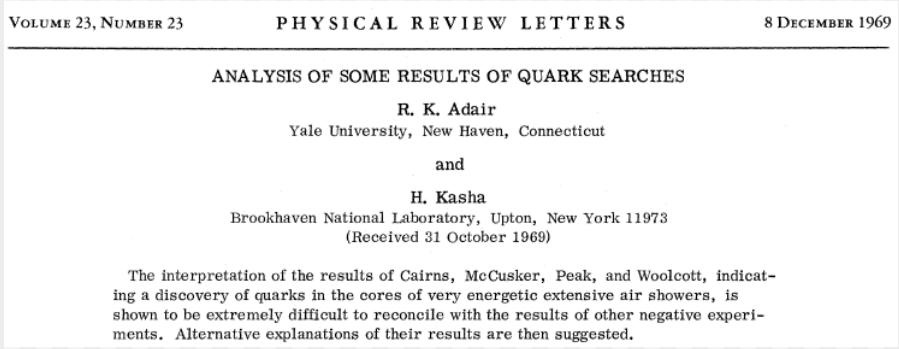
从假设中得知，径迹在一个给定长度中的液滴数目应该服从泊松分布。通过对正常径迹在固定长度引起的液滴数目进行测量得到均值为**229**。在同样长度下，对**55,000**根径迹进行测量，发现有一根径迹只产生**110**个液滴。出现该情况的正常概率为：

泊松统计（主观概率）　　频率概率

$$f(n \le 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \approx 1.6 \times 10^{-18} \quad \Leftrightarrow \quad \frac{1}{55,000} \approx 2 \times 10^{-5}$$

结果比频率概率相差很大，出现显著的反常（即：不可能出现只含**110**液滴的径迹）。

重新研究发现泊松分布的现象与粒子的散射有关。



VOLUME 23, NUMBER 23　　PHYSICAL REVIEW LETTERS　　8 DECEMBER 1969

ANALYSIS OF SOME RESULTS OF QUARK SEARCHES

R. K. Adair
Yale University, New Haven, Connecticut

and

H. Kasha
Brookhaven National Laboratory, Upton, New York 11973
(Received 31 October 1969)

The interpretation of the results of Cairns, McCusker, Peak, and Woolcott, indicating a discovery of quarks in the cores of very energetic extensive air showers, is shown to be extremely difficult to reconcile with the results of other negative experiments. Alternative explanations of their results are then suggested.

新物理+泊松分布的诠释

假设每个“泊松”事例精确产生 4 个液滴，原来实验观察到的平均液滴数应是 4 的倍数，也就是

$110 \xrightarrow{4\text{的倍数}} 112$　　$229 \xrightarrow{4\text{的倍数}} 228$

重新计算概率

泊松统计（主观概率）　　频率概率

$$f(n' \le 112/4 = 28)$$

$$= \sum_{i=0}^{28} \frac{(228/4 = 57)^i e^{-(228/4=57)}}{i!}$$

$$\approx 6.7 \times 10^{-6}$$

$\Longleftrightarrow$

$$\frac{1}{55,000} \approx 2 \times 10^{-5}$$

结果与频率概率接近

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

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

记为 N(μ,σ)

平均值：

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

方差：

$$V[x]=E[(x-\mu)^2]$$
$$=E[x^2]-E^2[x]$$
$$=\sigma^2$$

在所有统计问题扮演中心角色，应用于所有科学研究领域所涉及的分布。测量误差，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

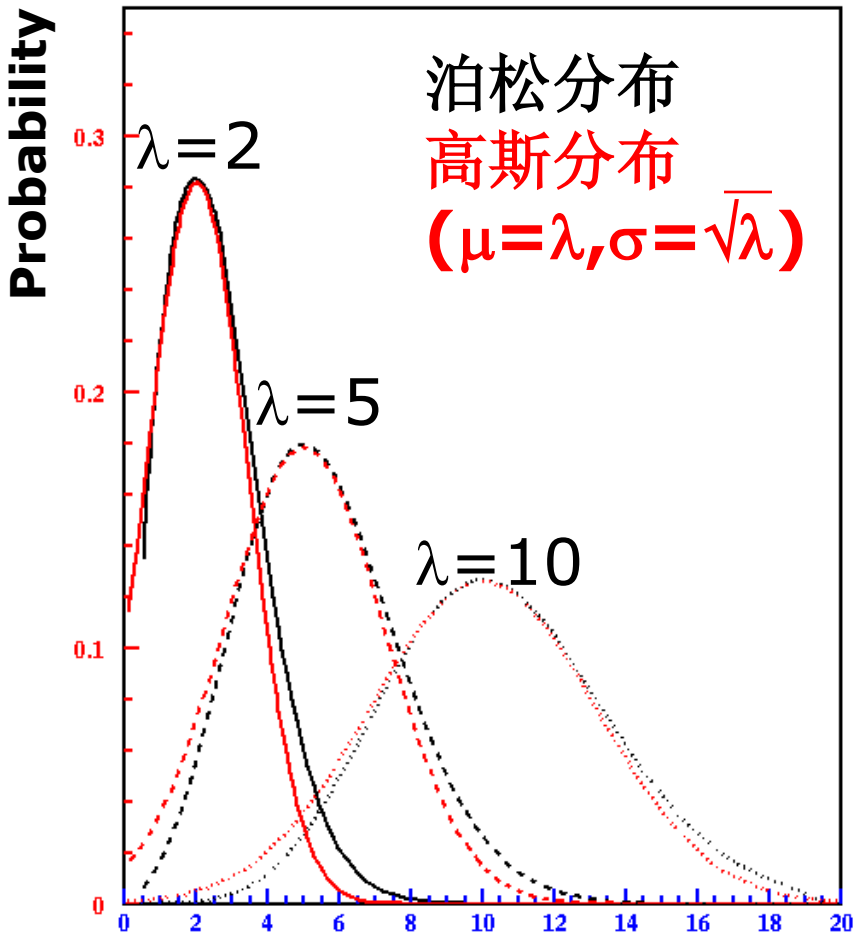
对于 n 个独立的随机变量 x_i，如果每个 x_i 都服从平均值为 μ_i 和有限的方差 σ_i^2 分布，那么变量

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

因此，如果

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

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$\vec{x}=(x_1,\cdots,x_n)$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 $E[x_i]=\mu_i,\quad \mathrm{cov}[x_i,x_j]=V_{ij}$

对于二维情形，其概率密度函数可表示为

$$f(x_1,x_2;\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

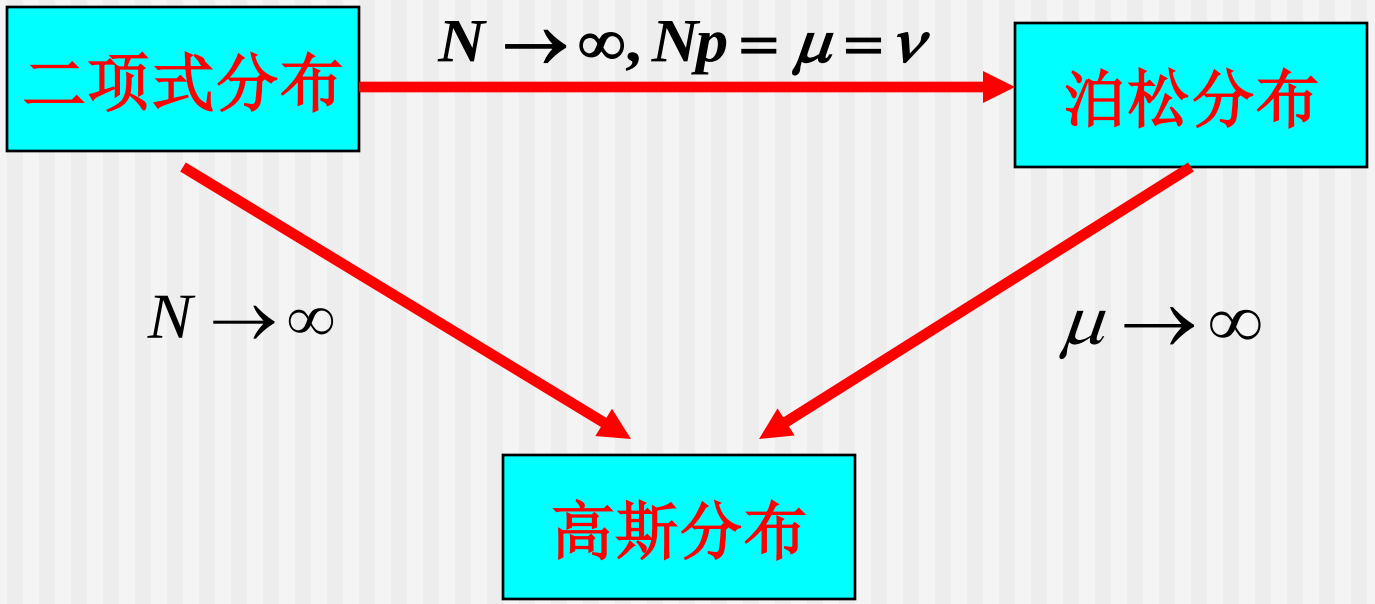
$$\rho=\mathrm{cov}[x_1,x_2]/(\sigma_1\sigma_2)$$

$$\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

二项式，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。

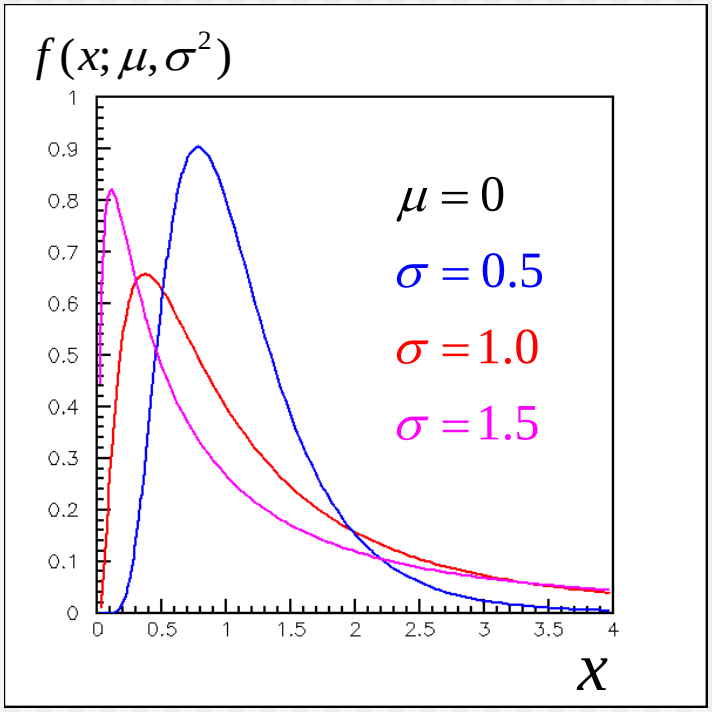
$$f(x;\mu,\sigma^2)=\begin{cases}\dfrac{1}{\sqrt{2\pi\sigma^2}}\dfrac{1}{x}\exp\left(\dfrac{-(\log x-\mu)^2}{2\sigma^2}\right) & \text{对于} x>0,\sigma>0\\ 0 & \text{其它情况}\end{cases}$$

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

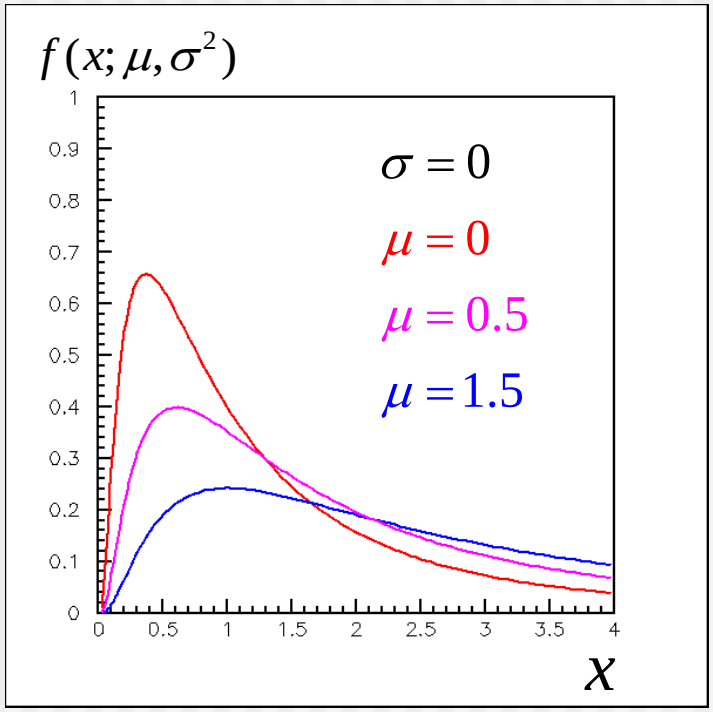
$$\text{方差：}\ V[x]=\exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]$$

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

$$\int_{\log a}^{\log b}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy \qquad 令 t=\frac{y-\mu}{\sigma}$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\log b-\mu}{\sigma}}\exp(-\frac{t^2}{2})dt-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\log a-\mu}{\sigma}}\exp(-\frac{t^2}{2})dt$$

$$=F\left(\frac{\log b-\mu}{\sigma}\right)-F\left(\frac{\log a-\mu}{\sigma}\right)$$

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 μ=4.0 和 σ^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

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

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

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 $x\ (0 \le x < \infty)$ 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

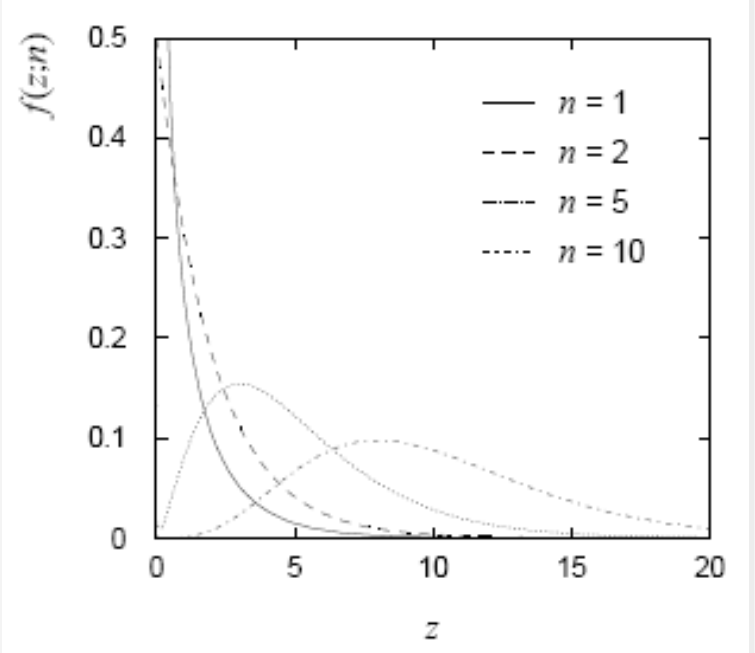
如果 $x_1,\ldots,x_n$ 是相互独立的高斯随机变量，按下列方式求和

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

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 $\Gamma(r)\equiv\int_0^\infty x^{r-1}e^{-x}dx$



$$\text{平均值}:E[z]=\mu=n$$

$$\text{方差}:V[z]=\sigma^2=2n$$

χ^2-分布通常用来检验假设与实际情况的符合程度。

科西(布莱格-魏格纳)分布

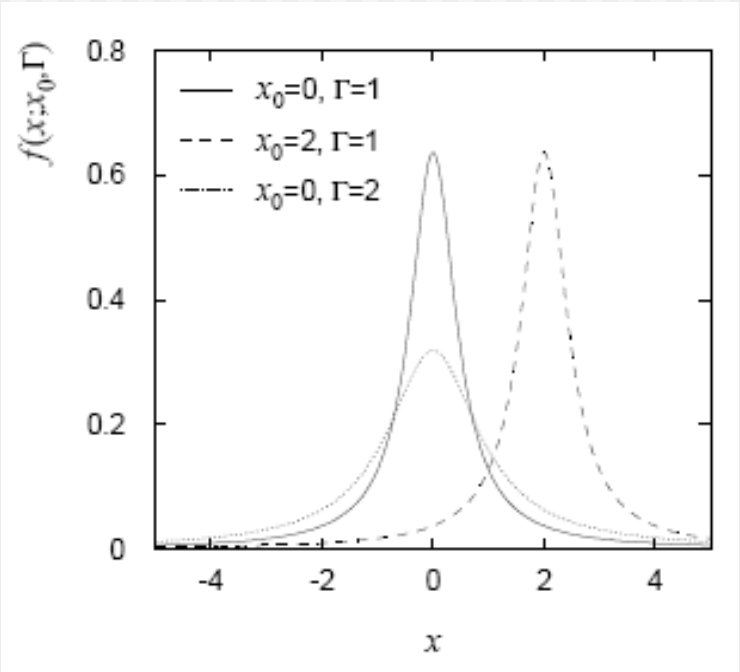
对于连续随机变量 x 的科西p.d.f.为

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

是布莱格-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**

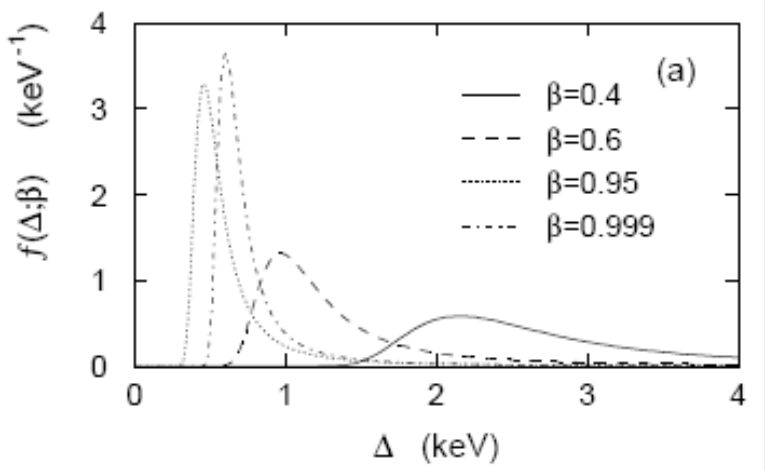
$$f(\Delta;\beta)=\frac{1}{\xi}\phi(\lambda),$$

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

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

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

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



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

- 二项式分布：探测效率，分支比
- 多项式分布：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
- 科西(布莱格-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！