



清华大学
Tsinghua University



粒子物理与核物理实验中的 数据分析

第二章：常用的概率分布

杨振伟
清华大学



本章要点

- 常用的概率分布
 - 数学表达
 - 均值与方差
 - 相关的应用范围

更多信息参见:

<http://staff.fysik.su.se/~walck/suf9601.pdf>

<http://pdg.lbl.gov/2019/reviews/rpp2019-rev-probability.pdf>

常用的概率分布

分布	粒子物理核物理中的例子
二项分布 (Binomial)	分支比、效率
多项分布 (Multinomial)	总事例数 N 固定的直方图
泊松分布 (Poisson)	实验发现的事例数
均匀分布 (Uniform)	蒙特卡罗方法
指数分布 (Exponential)	衰变时间
高斯分布 (Gaussian)	测量不确定度、分辨率函数
卡方分布 (Chi-square)	拟合优度
柯西分布 (Cauchy)	共振态的质量
朗道分布 (Landau)	电离能损
贝塔分布 (Beta)	效率的先验概率密度
伽马分布 (Gamma)	指数分布随机变量的和
学生氏分布 (Student's t)	尾部可调的分辨率函数

二项分布

N 次独立测量(伯努利试验), 每次只有成功(概率始终为 p) 或失败(概率为 $1 - p$)两种可能。

定义离散随机变量 n 为成功的次数, $0 \leq n \leq N$ 。

n 服从二项分布

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

可以证明其满足归一化条件

简记 $b(N, p)$

$$\begin{aligned} \sum_{n=1}^N f(n; N, p) &= \sum_{n=1}^N \frac{N!}{n! (N - n)!} p^n (1 - p)^{N - n} \\ &= [p + (1 - p)]^N = 1 \end{aligned}$$

适用于衰变分支比、探测效率不确定度的计算

二项分布：均值与方差

n 的均值（数学期望）为

$$E[n] = \sum_{n=0}^N n \cdot f(n; N, p) = \sum_{n=1}^N \frac{N!}{(n-1)!(N-n)!} p^n (1-p)^{N-n}$$

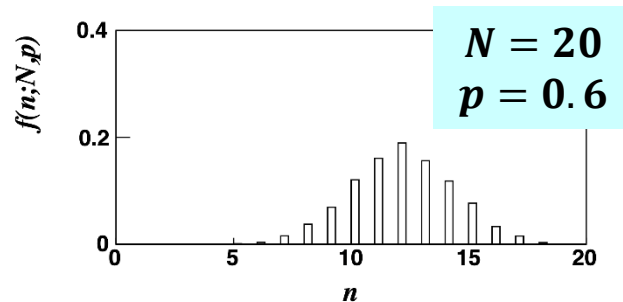
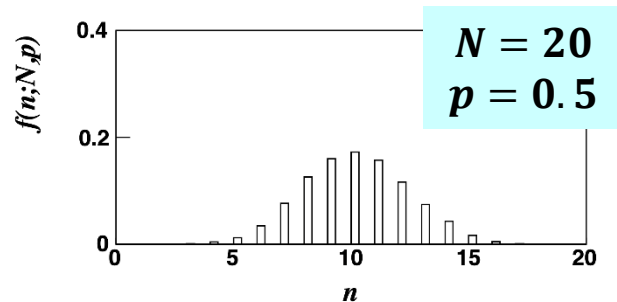
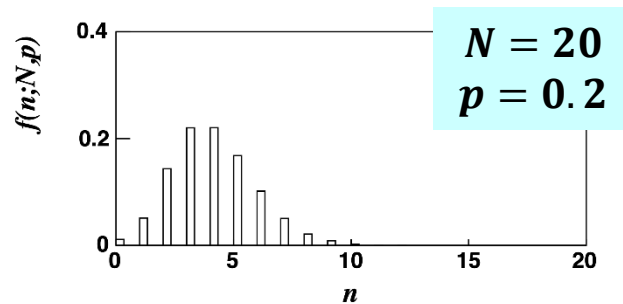
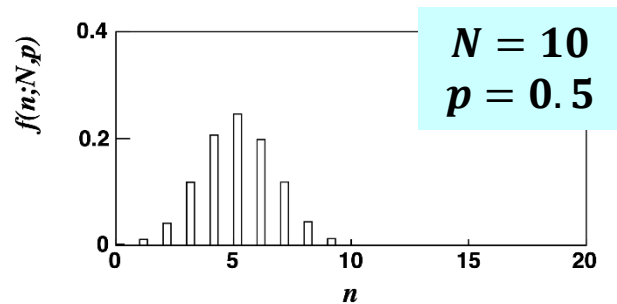
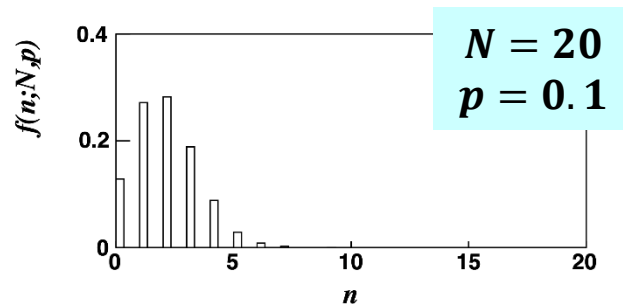
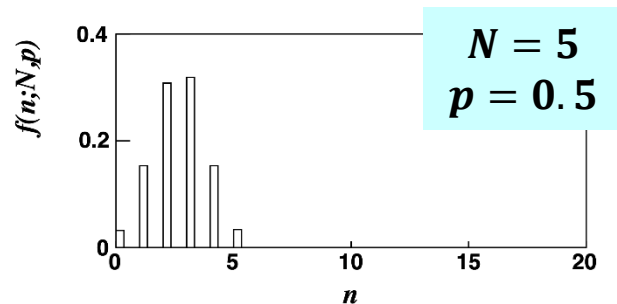
$$= Np \sum_{n=1}^N \frac{(N-1)!}{(n-1)!(N-n)!} p^{n-1} (1-p)^{N-n} = Np$$

n 的方差为

$$\begin{aligned} V[n] &= E[n^2] - E^2[n] = \sum_{n=0}^N n^2 \cdot f(n; N, p) - N^2 p^2 \\ &= Np(1-p) \end{aligned}$$

二项分布：不同参数的对比

不同参数值的二项分布



二项分布：适用条件

伯努利试验

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。

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。

例：效率和效率不确定度的估计

多层阻性板室(MRPC)的探测效率 ε

$$\varepsilon = \frac{N'}{N}$$

N' : MRPC记录的粒子数

N : 穿过MRPC的粒子数

宇宙线



闪烁体1与2
同时击中

穿过MRPC
的粒子数 N

MRPC记录
的击中数 N'

MRPC探测
效率测量值
及不确定度

$$\varepsilon = \frac{N'}{N}$$

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

二项分布指导决策

清华大学为LHCb实验研制闪烁光纤径迹探测器前端电子学板。按设计在一年内需要修理的电路板为0.1%。如果在实验所需的2000块板中有4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: 首先先计算一年内2000块板中 ≥ 4 块需要维修的概率

$$\begin{aligned} P &= \sum_{n=4}^{2000} f(n; 2000, 0.001) = 1 - \sum_{n=0}^3 f(n; 2000, 0.001) \\ &= 1 - \sum_{n=0}^3 \frac{2000!}{n! (2000 - n)!} \times 0.001^n \times (1 - 0.001)^{2000-n} \\ &= 1 - 0.8572 = 0.1428 \end{aligned}$$

一年内 ≥ 4 块发生故障的概率不算小,
所以板的质量可以接受。

多项分布

与二项分布类似，但有 $m > 2$ 种可能结果，每种结果的概率为

$$\vec{p} = (p_1, p_2, \dots, p_m), \text{ 满足 } \sum_{i=1}^m p_i = 1$$

N 次试验，得到的结果可用 m 维矢量表示

$$\vec{n} = (n_1, n_2, \dots, n_m)$$

$$\sum_{i=1}^m n_i = N$$

	次数
可能结果1	n_1
可能结果2	n_2
\vdots	\vdots
可能结果3	n_m

\vec{n} 是服从多项分布的随机变量

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

总频数固定时直方图各个区间的频数服从多项分布。

\vec{n} 实际上就代表一个直方图 (m 个区间，总频数为 N) ！

多项分布

多项分布与二项分布的关系

将第 i 种可能结果视为“成功”，所有其他结果为“失败”：

➡ n_i 服从参数为 (N, p_i) 的二项分布 $f(n_i; N, p_i)$

$$E[n_i] = Np_i$$

$$V[n_i] = Np_i(1 - p_i)$$

可以得到协方差为

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

思考： n_i 和 n_j ($i \neq j$) 之间正相关还是负相关？

泊松分布

考虑二项分布随机变量 n 的极限

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

$$\begin{aligned} N &\rightarrow \infty \\ p &\rightarrow 0 \\ E[n] = Np &\rightarrow \nu \end{aligned}$$

泊松分布

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

$(n \geq 0)$ 简记 $\pi(\nu)$

容易证明

$$\sum_{n=0}^{\infty} \frac{\nu^n}{n!} e^{-\nu} = 1$$

$$E[n] = \nu$$

$$V[n] = \nu$$

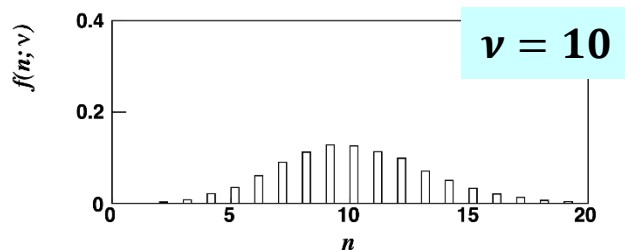
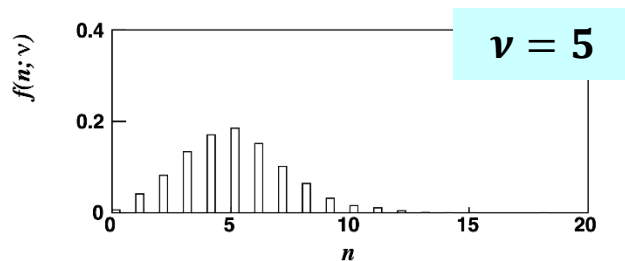
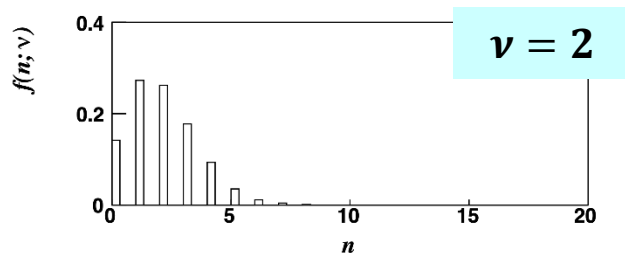
著名的统计不确定度估计式

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

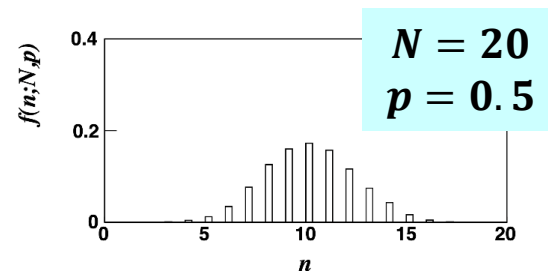
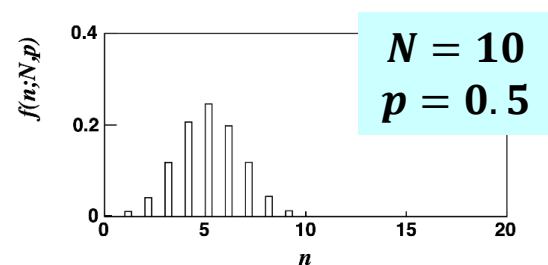
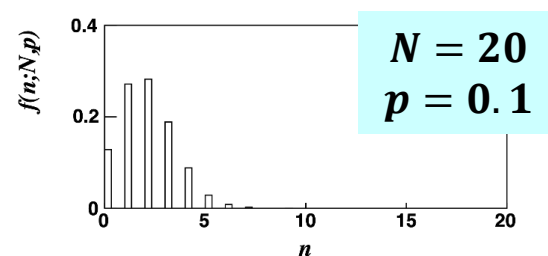
给定积分亮度 $L = \int \mathcal{L} dt$ 截面为 σ 的散射,
散射事例数 n 服从泊松分布

泊松分布

泊松分布



二项分布



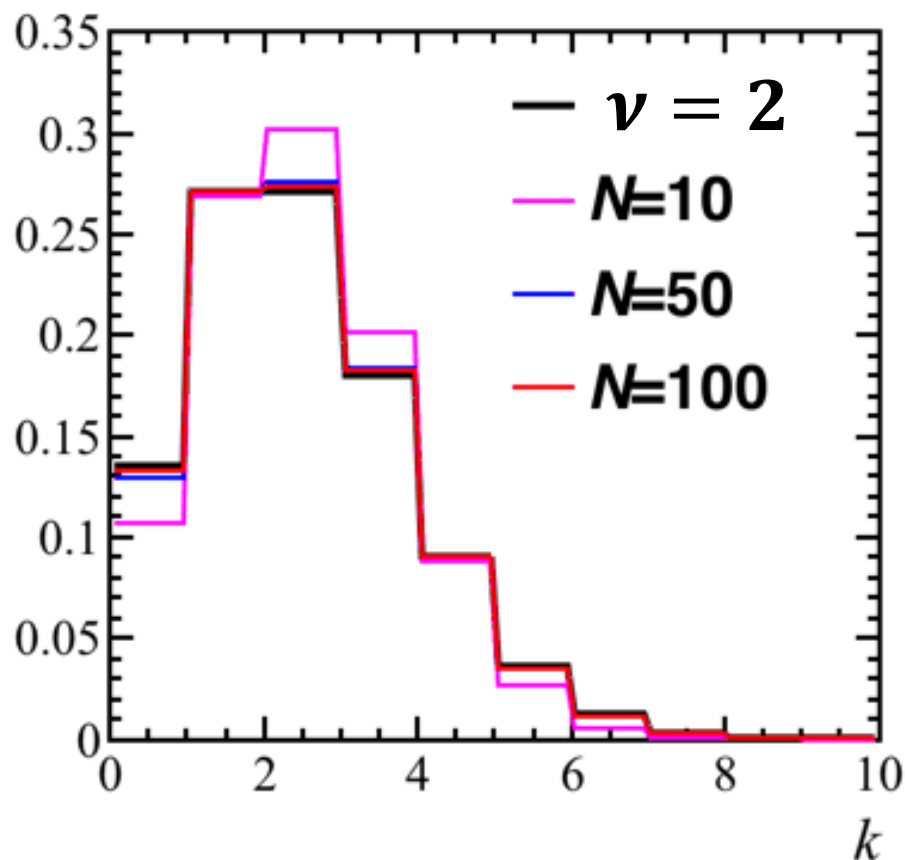
哪两个最相似?

泊松分布是二项分布的极限近似

例如：

对于以 $\nu = 2$ 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。



泊松分布是二项分布的极限近似

假设某人站在路边想搭便车。每分钟过路的汽车数服从泊松分布，平均每分钟过路一辆。假设每辆车让搭便车的概率为1%，并相互独立。

计算过了60辆车后还未能搭上车概率。

假设 $N = 60$ 辆车中同意搭便车的是 n 辆， n 服从二项分布
 $f(n; N = 60, p = 0.01)$

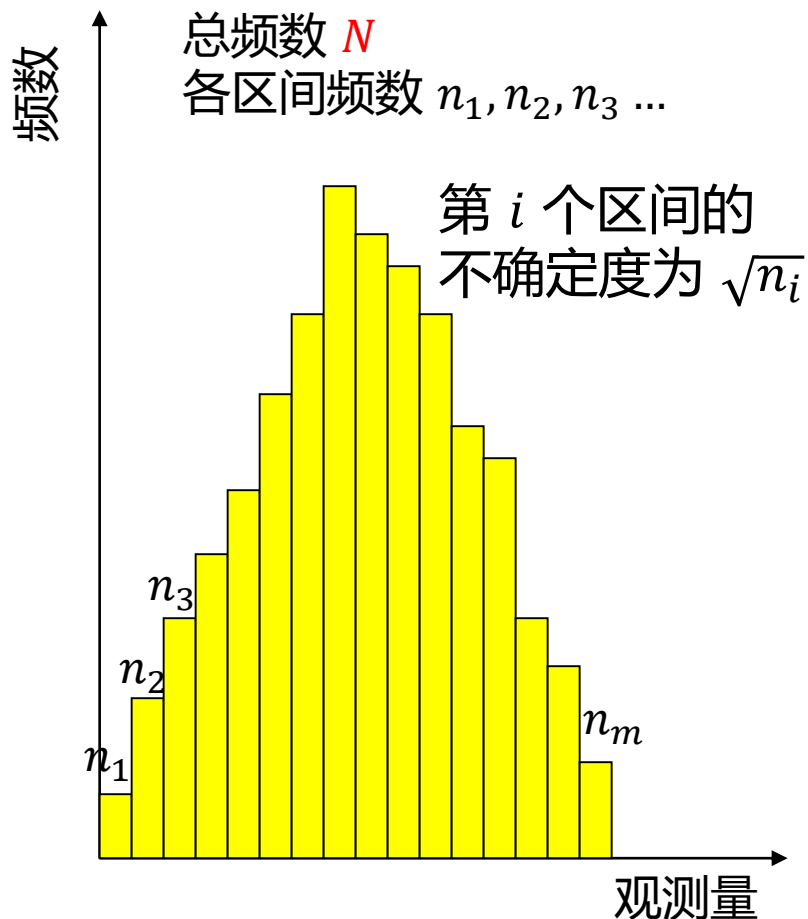
➡
$$f(0; N, p) = \frac{60!}{0!(60-0)!} 0.01^0 (1-0.01)^{60-0} = 0.5472$$

N 大， p 小，可以用 $\nu = Np = 0.6$ 的泊松分布近似

$$f(0; 0.6) = \frac{0.6^{60}}{0!} e^{-0.6} = 0.5488$$

泊松分布是二项分布的近似。

直方图中不确定度的处理



直方图可看成

1. 事例总数 N 服从泊松分布，每个区间频数 $\vec{n} = (n_1, n_2, \dots, n_m)$ 服从多项分布 ($\sum_{i=1}^m n_i = N$)
2. 或者，每个区间相互独立的泊松分布

总频数 N 的不确定度： $\Delta N = \sqrt{N}$

或由独立的 n_i 的不确定度传递得

$$\begin{aligned}(\Delta N)^2 &= (\Delta n_1)^2 + (\Delta n_2)^2 + \dots + (\Delta n_m)^2 \\&= n_1 + n_2 + \dots + n_m \\&= N\end{aligned}$$

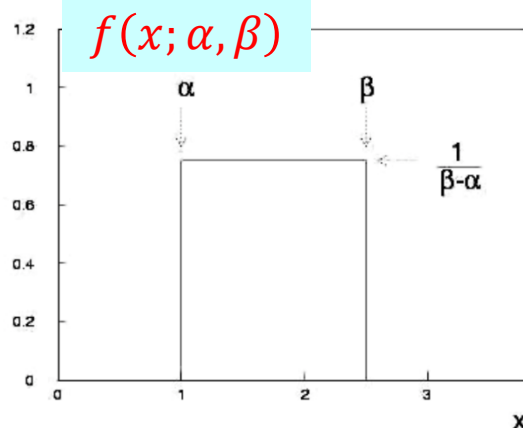
注意：当 $N < 5$ 时不确定度估计会有很大的偏差。

均匀分布

连续随机变量 $-\infty < x < \infty$ 服从均匀分布，概率密度为

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

简记 $U(\alpha, \beta)$



$$E[x] = \frac{1}{2}(\alpha + \beta) \quad V[x] = \frac{1}{12}(\beta - \alpha)^2$$

例：对于 $\pi^0 \rightarrow \gamma\gamma$, E_γ 服从 $[E_-, E_+]$ 间的均匀分布，
其中 $E_\pm = \frac{1}{2}E_\pi(1 \pm \beta)$ 。

注意：若 $F(x)$ 为某连续随机变量 x 的累积分布，则随机变量 $y = F(x)$ 服从 $[0, 1]$ 区间的均匀分布。

均匀分布是用蒙特卡罗模拟随机现象的基础。

指数分布

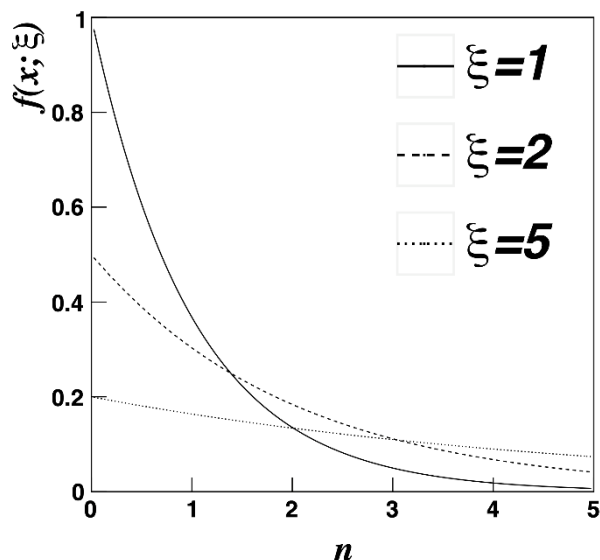
连续随机变量 x 的指数分布定义为

$$f(x; \xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi}, & x \geq 0 \\ 0, & \text{其他} \end{cases}$$

简记 $Exp(\xi)$

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

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



例：不稳定粒子的固有衰变时间

$$f(t; \tau) = \frac{1}{\tau} e^{-t/\tau} \quad (\tau = \text{平均寿命})$$

注意：指数分布没有记忆性： $f(t - t_0 | t \geq t_0) = f(t)$

高斯分布

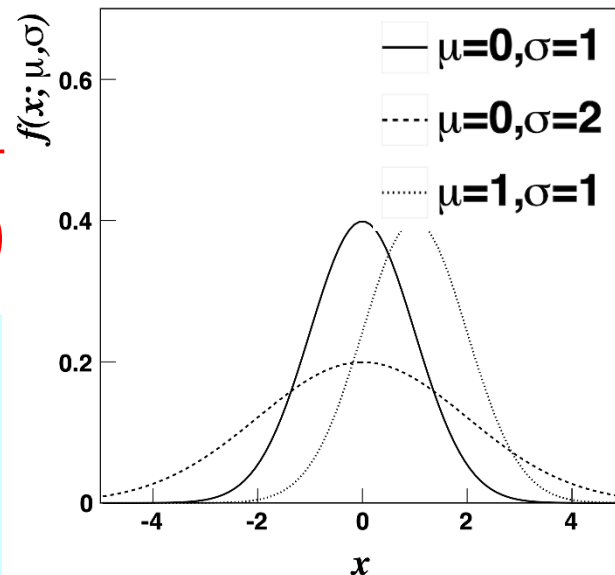
连续随机变量 x 的高斯分布定义为

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

又称正态分布
简记 $N(\mu, \sigma^2)$

$$E[x] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

$$V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$



特例： $\mu = 0, \sigma^2 = 1$ （标准正态分布： $N(0,1)$ ）

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\Phi(x) = \int_{-\infty}^x \varphi(x') dx'$$

如果 $x \sim N(\mu, \sigma^2)$, 则 $y = \frac{x-\mu}{\sigma} \sim N(0,1)$ 。

高斯分布

在所有统计问题扮演核心角色，在所有科学研究领域都会涉及到。

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

高斯分布的重要性在于，如果一个随机变量是由大量小贡献随机变量之和构成的，那么它往往服从高斯分布。

对于 n 个独立的随机变量 x_i ，如果每个 x_i 的方差存在，那么这些变量之和构成的随机变量

$$y = \sum_{i=1}^n x_i$$

在 $n \rightarrow \infty$ 的极限下，服从高斯分布 $N(\mu, \sigma^2)$ ，其中

$$\mu = \sum_{i=1}^n \mu_i, \quad \sigma^2 = \sum_{i=1}^n \sigma_i^2$$

测量不确定度通常来自很多贡献之和，所以重复测量的值可以看作服从高斯分布的随机变量。

中心极限定理（续）

对于 n 有限的情况，如果这 n 个变量之和的涨落不是有一个或少数变量主导，那么中心极限定理近似成立。



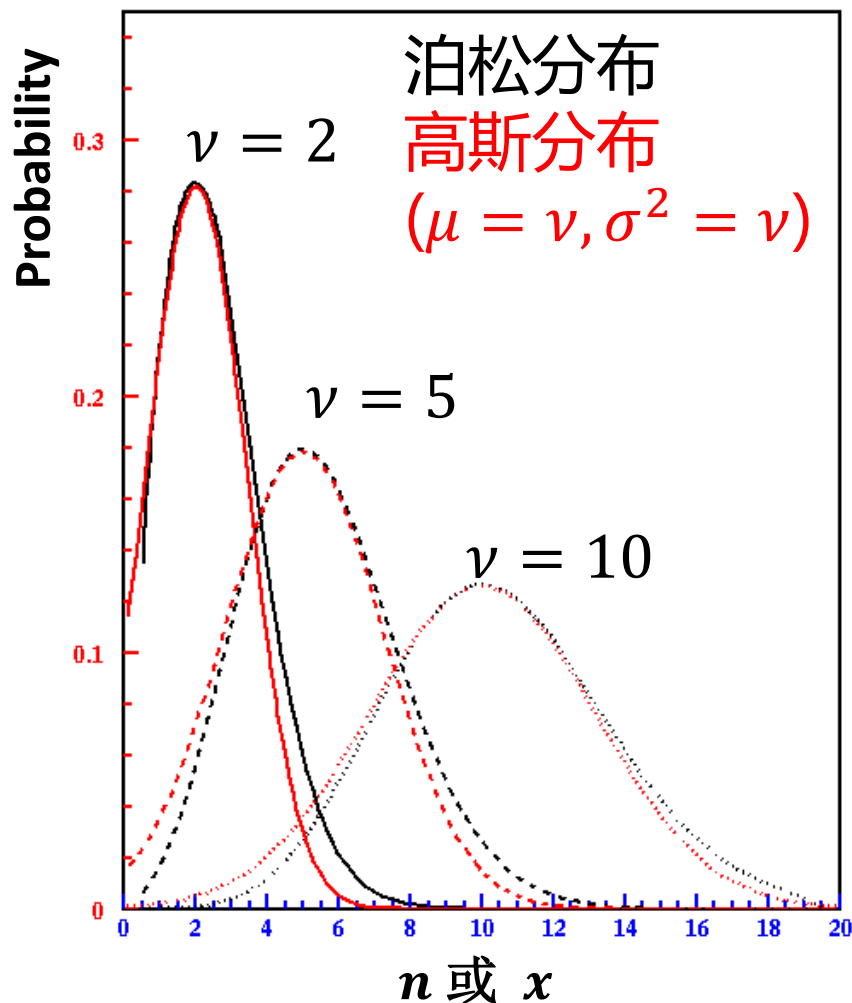
需要特别注意非高斯尾部的测量不确定度。

Good: 空气分子的速度分量 v_x 。

OK: 多重库仑散射引起的粒子偏转。
(稀少的大角度偏转会给出非高斯尾部)

Bad: 带电粒子穿过薄气层引起的能量损失。
(大部分能量损失由稀少的碰撞构成，例如朗道分布)

高斯分布与泊松分布

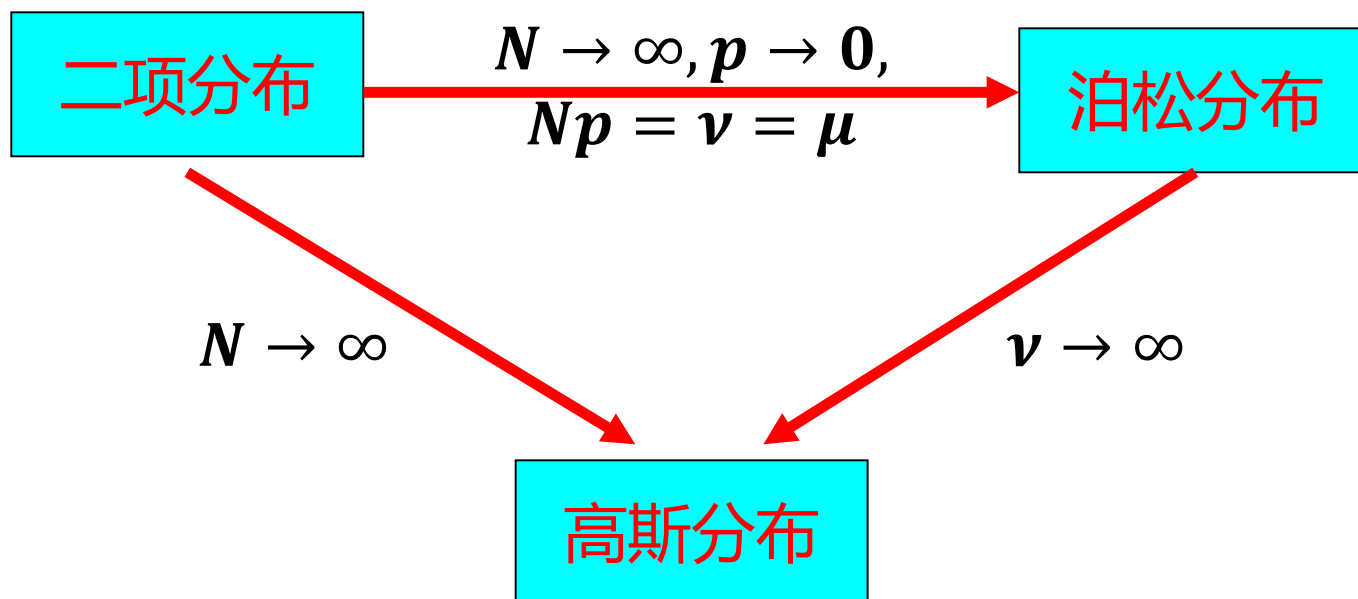


- 泊松分布只有非负整数定义
- 高斯分布是连续且可延伸到正负无穷
- 泊松分布的均值越大，它与高斯分布的区别就越小
- 实际应用时，当计数或频数大于5 时，可认为不确定度满足高斯分布

二项、泊松和高斯分布的联系

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

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



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

多维高斯分布

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

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

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

对于 $n = 2$, 概率密度函数可表示为

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

$\rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ 是相关系数

卡方 (χ^2) 分布

如果 x_1, \dots, x_n 是相互独立的高斯随机变量, 定义

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

$z(\geq 0)$ 服从自由度为 n 的卡方分布:

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

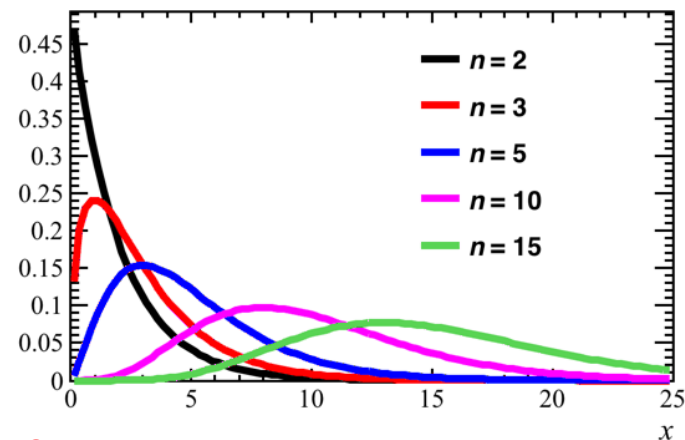
简记 $\chi^2(n)$

Γ 函数定义

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

$$E[z] = n, \quad V[z] = 2n$$

卡方分布通常用来检验假设与实际情况的符合程度, 例如, 最小二乘法拟合的拟合优度检验。



柯西（布莱特-魏格纳）分布

连续随机变量 x 的柯西分布为

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

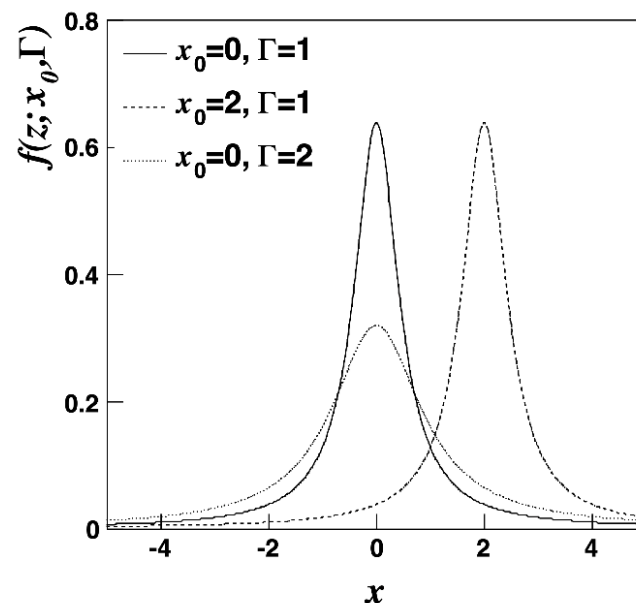
它是布莱特-魏格纳分布的特例

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

x_0 : 模 (most probable value)

Γ : 半高全宽 (full width at half maximum)

$E[x]$ 没有好的定义, $V[x] \rightarrow \infty$



常用于描述“共振态”粒子的不变质量分布, 例如 ρ, K^*, ϕ^0, \dots
 Γ = 衰变率 (平均寿命的倒数)

朗道分布

速度为 $\beta = v/c$ 的带电粒子穿过厚度为 d 的薄物质层，其能量损失 Δ 服从朗道分布：

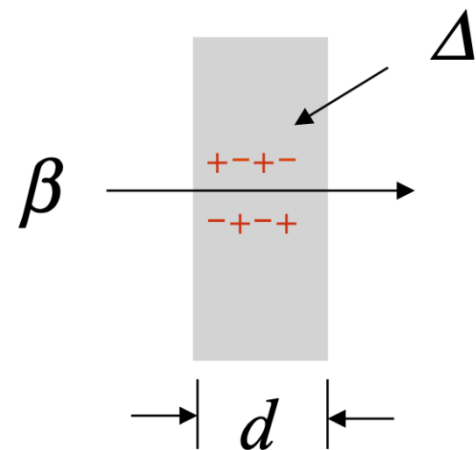
$$f(\Delta; \beta) = \frac{1}{\xi} \phi(\lambda),$$

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

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

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

$$\epsilon' = \frac{I^2 (1 - \beta^2) \exp(\beta^2)}{2m_e c^2 \beta^2}$$



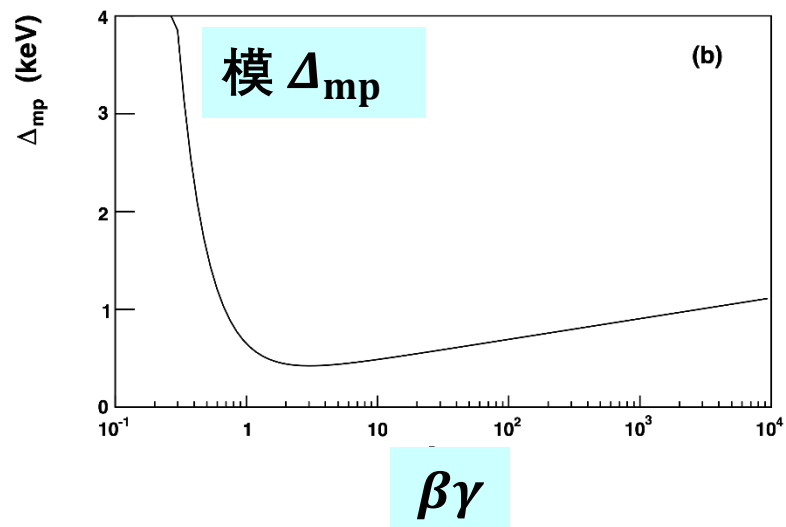
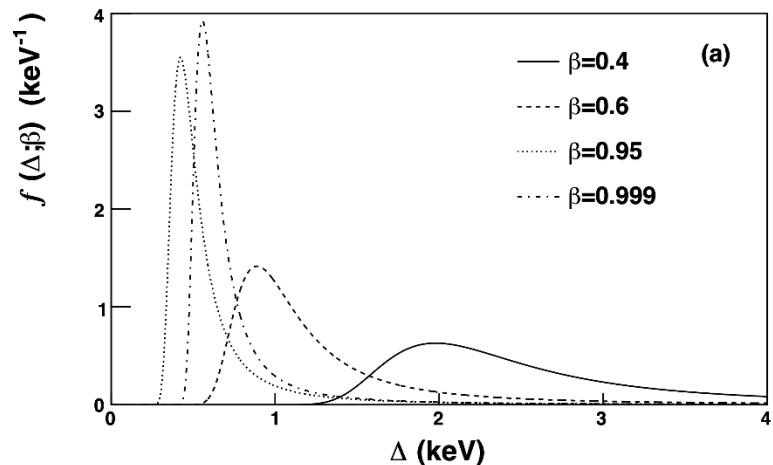
厚度 d 增大时，趋于正态分布。

常用于描述粒子的电离能损或能量沉积。

朗道分布

朗道分布有长的“朗道尾部”
→ 所有的矩都发散

模 (most probable value)
对 β 很敏感
→ 可用于粒子鉴别(PID)



贝塔分布

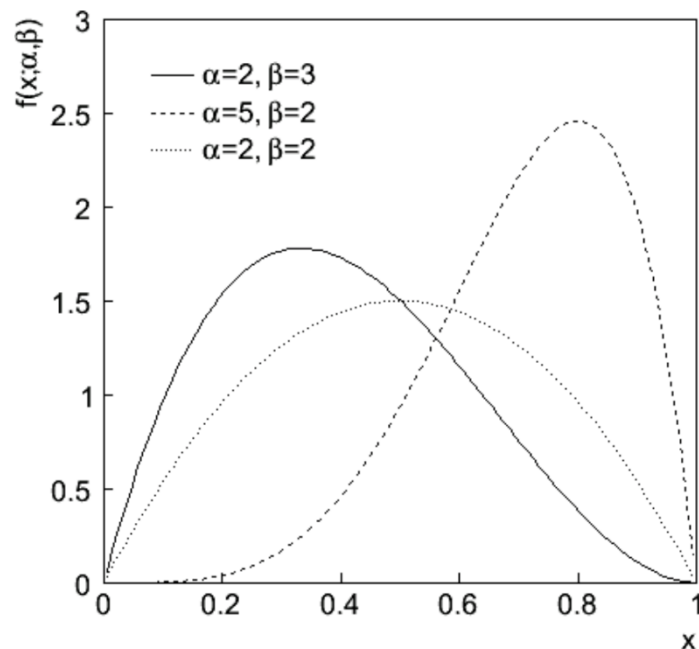
连续随机变量 $0 < x < 1$ 的贝塔分布为：

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha + \beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

简记 $Be(\alpha, \beta)$



$$Be(1,1) = U(0,1)$$

常用来表示只在某个有限区间非零的连续随机变量。

伽马分布

连续随机变量 x 的伽马分布为：

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

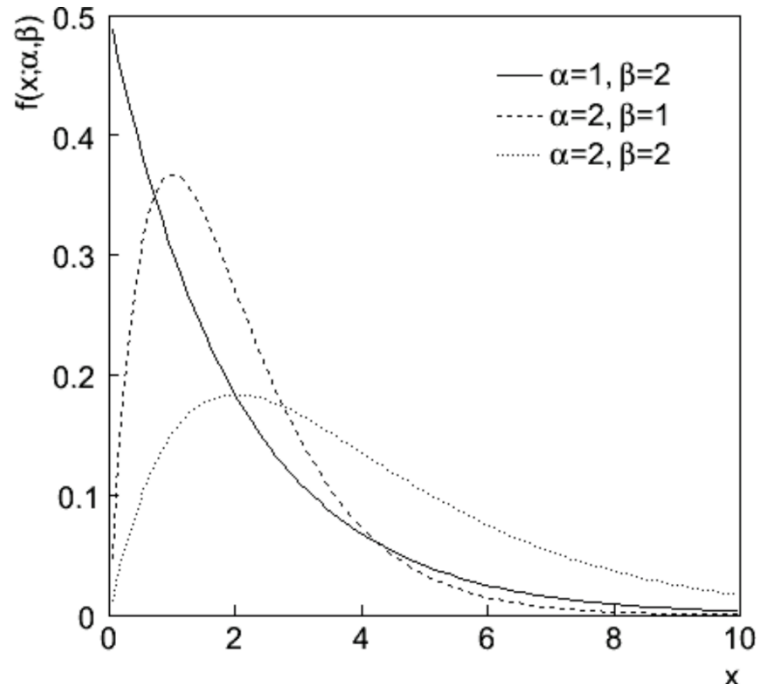
$$E[x] = \alpha\beta$$

简记 $Ga(\alpha, \beta)$

$$V[x] = \alpha\beta^2$$

$$Ga(1, \xi) = Exp(\xi)$$

$$Ga(n/2, 2) = \chi^2(n)$$



常用来表示在 $[0, \infty]$ 内不为零的连续随机变量。

例： n 个指数分布随机变量的和，泊松过程第 n 个事件发生的时间。

学生氏分布

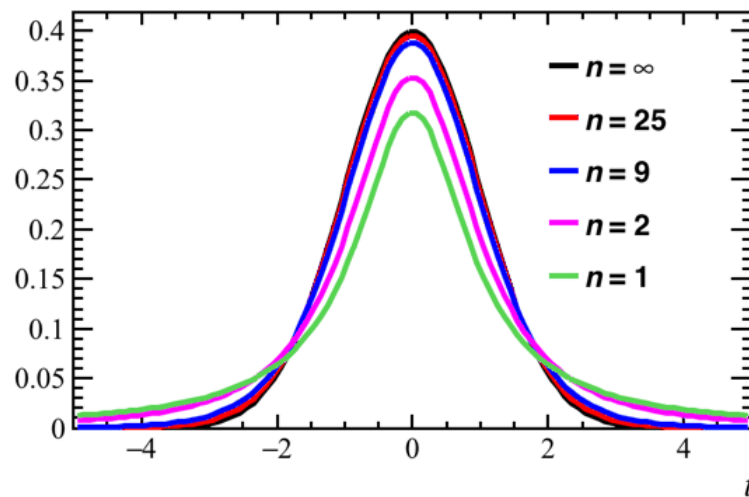
连续随机变量 x 的自由度为 ν 的学生氏分布为：

$$f(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$E[x] = 0 \quad (\nu > 0)$$

$$V[x] = \frac{\nu}{\nu - 2}$$

简记 $t(\nu)$



ν ：自由度的数目（可以不是整数）

$t(1)$ = 柯西分布

$t(\infty) = N(0,1)$

小结

分布	粒子物理核物理中的例子
二项分布 (Binomial)	分支比、效率
多项分布 (Multinomial)	总事例数 N 固定的直方图
泊松分布 (Poisson)	实验发现的事例数
均匀分布 (Uniform)	蒙特卡罗方法
指数分布 (Exponential)	衰变时间
高斯分布 (Gaussian)	测量不确定度、分辨率函数
卡方分布 (Chi-square)	拟合优度
柯西分布 (Cauchy)	共振态的质量
朗道分布 (Landau)	电离能损
贝塔分布 (Beta)	效率的先验概率密度
伽马分布 (Gamma)	指数分布随机变量的和
学生氏分布 (Student's t)	尾部可调的分辨率函数

ROOT平台给出所有这些分布!