# 数据科学基础

# Foundations of Data Science

## 3.4 连续概率分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

# 均匀分布

**【均匀分布】** 如果随机变量$X$的概率密度为:

$$f(x)=\begin{cases}\dfrac{1}{b-a}, a<x<b\\ 0, x\le a, \text{或}, x\ge b\end{cases}$$

则称$X$在区间$(a,b)$内服从均匀分布,记为$X\sim U(a,b)$, 其分布函数为

$$F(x)=\begin{cases}0, x<a\\ \dfrac{x-a}{b-a}, a\le x<b\\ 1, x\ge b\end{cases}$$

# 均匀分布的矩

$$E(X)=\int_{-\infty}^{+\infty}xf(x)dx=\int_{-\infty}^{+\infty}\frac{x}{b-a}dx=\frac{a+b}{2}$$

$$E\left(X^2\right)=\int_{-\infty}^{+\infty}x^2f(x)dx==\int_{-\infty}^{+\infty}\frac{x^2}{b-a}dx\ =\frac{b^3-a^3}{3(b-a)}=\frac{a^2+b^2+ab}{3}$$

$$Var(X)=E\left(X^2\right)-\ (EX)^2=\frac{a^2+b^2+ab}{3}-\frac{a^2+b^2+2ab}{4}=\frac{(b-a)^2}{12}$$

更多详情，请咨询课程助教。

# 指数分布

**【指数分布】** 如果随机变量$X$的概率密度为:

$$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}}, x>0\\ 0, x\leq 0\end{cases}$$

则称$X$服从参数为$\lambda(\lambda>0)$的指数分布, 记为$X\sim E(\lambda)$, 其分布函数为:

$$F(x)=\begin{cases}1-e^{-\frac{x}{\theta}}, x>0\\ 0, x\leq 0\end{cases}$$

# 指数分布的无记忆性

$$P\{X > s + t|X > s\} = P\{X > t\}$$

这一性质称为指数分布的无记忆性.

事实上可以证明指数分布是唯一具有上述性质的连续型分布.

证明:

$$P\{X > s + t|X > s\} = \frac{P\{(X > s + t) \cap (X > s)\}}{P\{X > s\}} = \frac{P\{X > s + t\}}{P\{X > s\}}$$

$$= \frac{P\{X > s + t\}}{P\{X > s\}} = \frac{1 - E(s + t)}{1 - E(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda(s)}} = e^{-\lambda t} = P\{X > t\}$$

# 指数分布的矩

【**定理**】设随机变量$X$指数分布, 即概率密度为

$$f(x)=\frac{1}{\theta}e^{-\frac{x}{\theta}},x>0$$

则$E(X)=\theta,Var(X)=\theta^2$

证:

$$E(X)\int_{-\infty}^{\infty}xf(x)dx=\int_0^{\infty}x\frac{1}{\theta}e^{-\frac{x}{\theta}}dx$$

$$=-xe^{-\frac{x}{\theta}}\Big|_0^{\infty}+\int_0^{\infty}e^{-\frac{x}{\theta}}=\theta$$

$$E\left(X^2\right)\int_{-\infty}^{\infty}x^2f(x)dx=\int_0^{\infty}x^2\frac{1}{\theta}e^{-\frac{x}{\theta}}dx$$

$$=-x^2e^{-\frac{x}{\theta}}\Big|_0^{\infty}+\int_0^{\infty}2xe^{-\frac{x}{\theta}}=2\theta^2$$

$$Var(X)=E\left(X^2\right)-E(X)^2=\theta^2$$

# 正态分布

- 正态分布最早是棣莫弗在1718年著作的书籍的及1734年发表的一篇关于二项分布文章中提出的，当二项随机变量的位置参数$n$很大及形状参数为$\frac{1}{2}$时，则所推导出二项分布的近似分布函数就是正态分布。
- 拉普拉斯在1812年发表的《分析概率论》中对棣莫弗的结论作了扩展到二项分布的位置参数为$n$及形状参数为$p$时。现在这一结论通常被称为棣莫佛 - 拉普拉斯定理。拉普拉斯在误差分析试验中使用了正态分布。
- 勒让德于1805年引入最小二乘法这一重要方法；而高斯则宣称他早在1794年就使用了该方法，并通过假设误差服从正态分布给出了严格的证明。

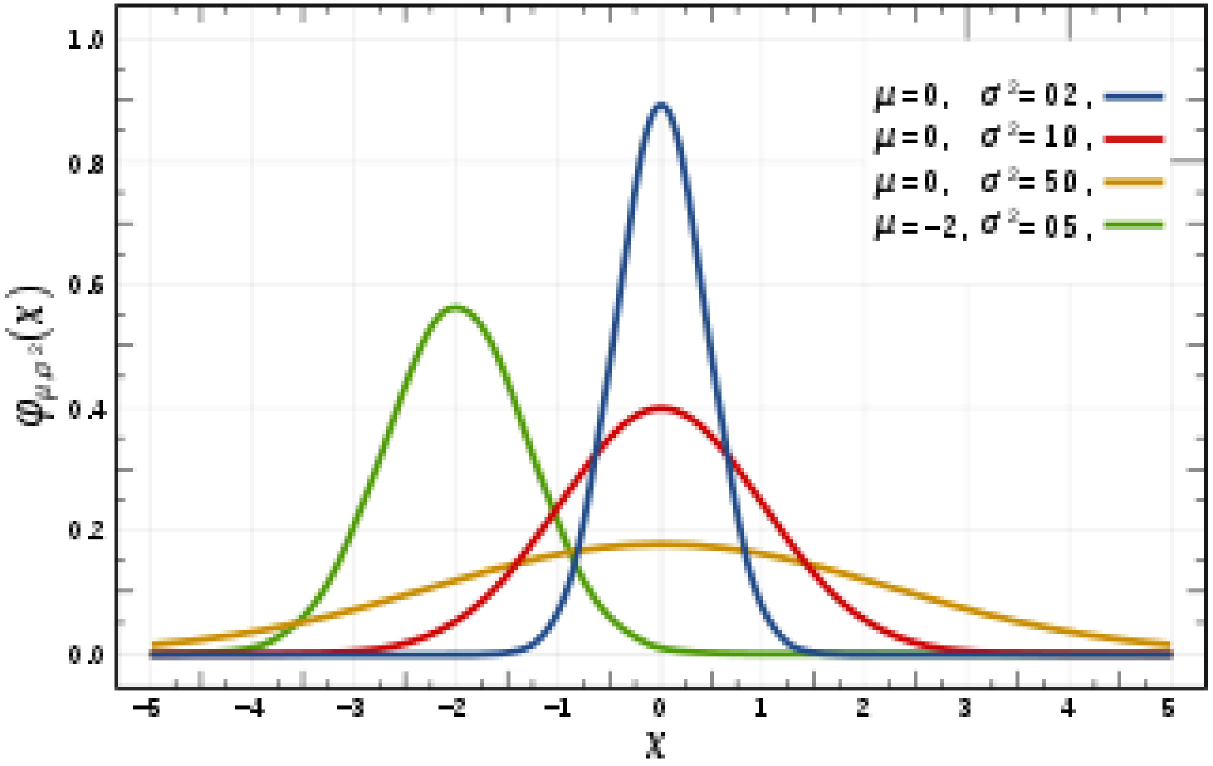
# 正态分布

【**正态分布**】如果随机变量$X$ 的概率密度为：

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

则称$X$ 服从参数为$\mu,\sigma^2$的正态分布,记为$X\sim N\left(\mu,\sigma^2\right)$, 其中参数$\mu\in R,\sigma>0.$

# 正态分布




正态分布性质：

- 曲线关于$x = \mu$对称.
- 当$x = \mu$时取到最大值$\frac{1}{\sqrt{2\pi}\sigma}$
- 固定$\sigma$, 改变$\mu$, 曲线沿$O_x$轴平移;
- 固定$\mu$, 改变$\sigma$, 由于最大值为$\frac{1}{\sqrt{2\pi}\sigma}$. 曲线变得越尖, 因而$X$落在$\mu$附近的概率越大.

# 标准正态分布

【标准正态分布】当$\mu=0,\sigma=1$,称$X$服从标准正态分布,

记为$X\sim N(0,1)$, 分布函数记为$\Phi(x)$. 有

$$\Phi(-x)=1-\Phi(x)$$

- 标准正态分布查表方法

# 正态分布标准化

**【定理】**若$X \sim N\left(\mu,\sigma^2\right)$, 则$Z=\frac{X-\mu}{\sigma} \sim N(0,1)$.

证明:

$$P\{Z \leq x\}=P\{\frac{(X-\mu)}{\sigma} \leq x\}=P\{X \leq \mu+\sigma x\}$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\mu+\sigma x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

令$y=\frac{t-u}{\sigma}$, 则

$$P\{Z \leq x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy=\Phi(x)$$

所以 $Z=\frac{X-\mu}{\sigma} \sim N(0,1)$.

# 标准正态分布的矩

**【定理】**$Z\sim N(0,1)$, 证明$E(X)=0, Var(X)=1.$

证明:

$$E(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} te^{-\frac{t^2}{2}}dt=0$$

$$Var(Z)=E(Z^2)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t^2e^{-\frac{t^2}{2}}dt$$

$$=-\frac{1}{\sqrt{2\pi}}t\Big|_{-\infty}^{+\infty}+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}dt=1$$

# 正态分布的矩

【**定理**】对于任意$X \sim N(\mu, \sigma^2)$, $E(X) = \mu$, $Var(Z) = \sigma^2$

证明:

因为$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. 所以

$$E(X) = E(\mu + \sigma Z) = \mu$$

$$Var(X) = Var(\mu + \sigma Z) = \sigma^2$$

# 正态分布计算

**【例】**一种电子元件的使用寿命$X\sim\mathbb{N}\left(100,15^2\right)$,某仪器上装有3个这种元件，三个元件损坏与否是相互独立的.求：使用的最初90小时内无一元件损坏的概率.

# 正态分布计算

**【例】**一批钢材(线材)长度$X \sim N\left(\mu, \sigma^2\right)$

1. 若$\mu = 100, \sigma = 2$, 求这批钢材长度小于$97.8cm$的概率;
2. 若$\mu = 100$, 要使这批钢材的长度至少有$90\%$落在区间$(97,103)$内, 问$\sigma$至多取何值?

# 正态分布计算

解: (1) $P\{X < 97.8\} = \Phi\left(\frac{97.8-100}{2}\right) = 1 - \Phi(1.1) = 1 - 0.8643 = 0.1357$

(2) $0.90 \le P\{97 < X < 103\} = \Phi\left(\frac{103-100}{\sigma}\right) - \Phi\left(\frac{97-100}{\sigma}\right)$

$$= 2\Phi\left(\frac{3}{\sigma}\right) - 1 \Rightarrow \Phi\left(\frac{3}{\sigma}\right) \ge 0.95 \Rightarrow \frac{3}{\sigma} \ge 1.645$$

所以 $\sigma \le 1.8237.$

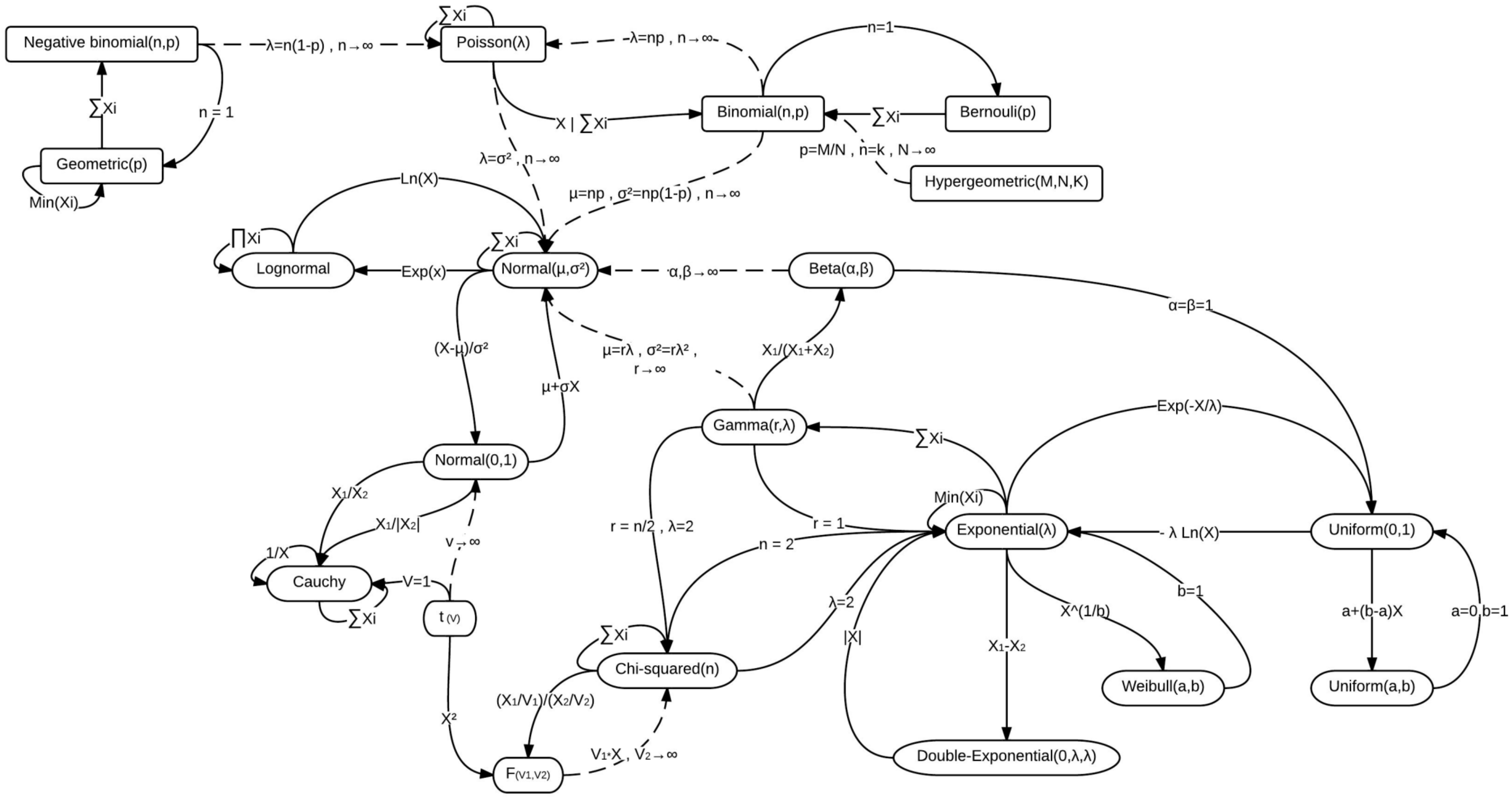
# 正态分布计算

【例】将一温度调节器放置在存储着某种液体的容器内, 调节器定在$d$, 液体的温度$X$是一个随机变量, 且$X \sim N\left(d, 0.5^2\right)$.

(1)若$d = 90$, 求$X < 89$的概率;

(2)若要求保持液体的温度至少为80的概率不低于0.99, 问$d$至少为多少?

# 总结


Negative binomial(n,p)
λ=n(1-p) , n→∞
∑Xi
Poisson(λ)
λ=np , n→∞
n=1
∑Xi
n = 1
Geometric(p)
Min(Xi)
X | ∑Xi
Binomial(n,p)
∑Xi
Bernouli(p)
λ=σ² , n→∞
p=M/N , n=k , N→∞
Hypergeometric(M,N,K)
Ln(X)
μ=np , σ²=np(1-p) , n→∞
∏Xi
Lognormal
Exp(x)
∑Xi
Normal(μ,σ²)
α,β→∞
Beta(α,β)
α=β=1
(X-μ)/σ²
μ=rλ , σ²=rλ² ,
r→∞
$X_1/(X_1+X_2)$
μ+σX
Exp(-X/λ)
Gamma(r,λ)
∑Xi
Normal(0,1)
$X_1/X_2$
Min(Xi)
$X_1/|X_2|$
r = n/2 , λ=2
r = 1
1/X
v→∞
n = 2
Exponential(λ)
- λ Ln(X)
Uniform(0,1)
Cauchy
V=1
λ=2
b=1
∑Xi
t (V)
∑Xi
X^(1/b)
a+(b-a)X
a=0,b=1
|X|
Chi-squared(n)
$X_1$-$X_2$
Weibull(a,b)
Uniform(a,b)
$(X_1/V_1)/(X_2/V_2)$
X²
$V_1$*X , $V_2$→∞
Double-Exponential(0,λ,λ)
F(V1,V2)

更多详情，请咨询课程助教。