Ch 4.5 连续随机变量函数的分布

回顾前一次课

正太分布的定义，性质

- 一般正太分布与标准正太分布的转化
- $\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2$
- $X \sim N(0,1)$的分布函数 $P(X \geq \epsilon) \leq \frac{1}{2} e^{-\frac{\epsilon^2}{2}} \quad P(|X| \geq \epsilon) \leq \ min\left(1, \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\epsilon^2}{2}}}{\epsilon}\right)$

正态分布的图像

标准正太分布的分布函数、性质

- 对称性、求概率、3σ原则、查表

离散随机变量函数

知道一个随机变量的概率分布后，经常会考虑它的一些函数分布，例如，当知道一个圆的直径X服从均匀分布$U(a,b)$，还可以考虑圆的面积$Y=\pi(X/2)^2$的分布

已知随机变量X的概率分布和函数 $g(x): R \to R$

问题: 随机变量$Y=g(X)$的概率密度$f_Y(y)$？

若X为离散型随机变量, 其分布列为

$$p_k = P(X = x_k) \qquad k = 1,2,\cdots$$

则求$Y=g(X)$的分布列较为简单，将相等的项$g(x_i)=g(x_j)$合并，相应的概率相加即可

例题

若随机变量X的概率分布列为

$$P(X = k) = 1/2^k \qquad\qquad k = 1,2,\cdots$$

求随机变量$Y = \cos(\pi X/2)$ 的分布列

连续型随机变量函数

针对连续型随机变量, 一般采用概率密度函数来刻画其概率分布.

已知：函数$g(x)$和随机变量X的概率密度函数为$f_X(x)$

求解：随机变量$Y = g(X)$的概率密度$f_Y(y)$

数学工具 — 积分求导公式

$$F(y) = \int_{\psi(y)}^{\varphi(y)} f(x)dx$$

$$\boldsymbol{F'(y) = f\big(\varphi(y)\big)\varphi'(y) - f\big(\psi(y)\big)\psi'(y)}$$

求解步骤

- 求解$Y = g(X)$的分布函数

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{g(x)\leq y} f_X(x)dx$$

- 利用分布函数和概率密度关系求解密度函数

$$f_Y(y) = F_Y'(y)$$

例题

设连续型随机变量X的密度函数

$$f(x)=\begin{cases} cx & 0<x<4 \\ 0 & \text{其它} \end{cases}$$

求$Y=2X+8$的密度函数

例题

设随机变量X概率密度为$f_X(x)$, 求$Y = X^2$的概率密度

定理

随机变量X在实数 概率密度为$f_X(x)$, 函数$y = g(x)$处处可导、且严格单调 (即$g'(x) > 0$ 或$g'(x) < 0$), 令其反函数$x = g^{-1}(y) = h(y)$, 则随机变量$Y = g(X)$ 的概率密度为

$$f_Y(y) = \begin{cases} f_X\big(h(y)\big)|h'(y)| & y \in (\alpha, \beta) \\ 0 & \text{其它} \end{cases}$$

$\alpha = \min\{g(-\infty), g(+\infty)\}$, $\beta = \max\{g(-\infty), g(+\infty)\}$

上述定理推广至区间函数$x \in [a, b]$

此时$\alpha = \min\{g(a), g(b)\}$ 和$\beta = \max\{g(a), g(b)\}$

例题

设随机变量$X \sim N\left(\mu, \sigma^2\right)$，则变量$Y = aX + b\ (a > 0)$服从正太分布$N\left(a\mu + b, a^2\sigma^2\right)$

定理

设随机变量X的分布函数$F(x)$是严格单调的连续函数, 则

$$Y = F(X) \sim U(0,1)$$

Ch 4.6 常用分布的随机数生成

定理

随机数在计算机仿真学和密码学等领域具有广泛的应用，通过产生大量的随机数据来实现对真实世界的模拟

可以利用物理随机过程来产生真实的随机数，如反复抛掷硬币、骰子、抽签、摇号等, 这些方法可以得到质量很高的随机数, 但其数量和类型通常较少、难以满足实际的需求

主流方法：使用计算机产生伪随机数，通过确定的算法生成类似真随机数的伪随机数. 由于算法给出的结果总是确定的, 所以伪随机数并不是真正的随机数，但是好的伪随机数序列与真实随机数序列表现几乎相同, 很难进行区分

(0, 1) 上均匀分布的随机数

有很多方法生成(0,1)上均匀分布的随机数，最常用**线性同余法**.
通过迭代方式产生一系列随机数$x_1, x_2, \cdots, x_k$

$$x_i = a\, x_{i-1} + c \quad \mod \ m$$

其中$1 \le i \le k \le m$, 常数x_0为初始给定的种子

为获得较好的随机性, m的取值应足够大, 如$m = 2^k (k = 31,63)$,
常数c与m互质, 常数$a - 1$被m的因子整除, 例如一种可行的选择

$$x_i = 31415926\, x_{i-1} + 453806245 \mod \ 2^{31}$$

在区间(0,1)上均匀分布的随机数$x_1/m, x_2/m \cdots, x_k/m$.

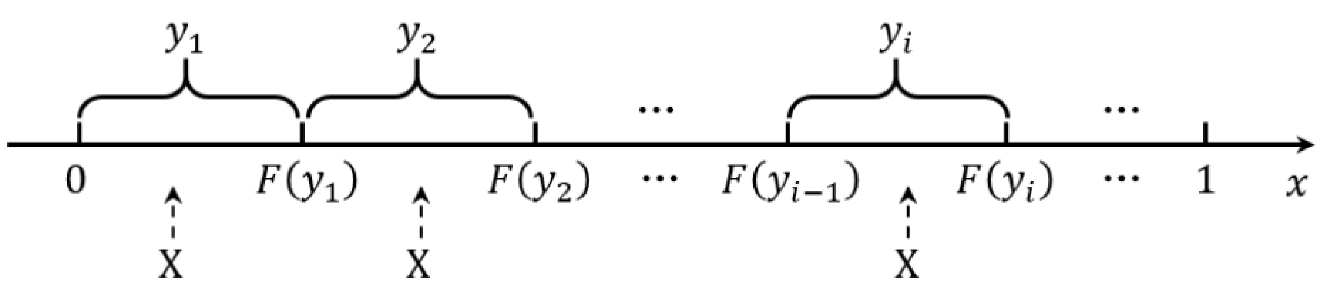
很多时候会根据实际情况选择不同的初始种子x_0

常用离散型分布的随机数

定理： 若随机变量$X \sim U(0,1)$，以及$F(y)$是某个离散型随机变量的分布函数, 其取值分别为$y_1 < y_2 < \cdots$. 设随机变量

$$Y = \begin{cases} y_1 & X \le F(y_1) \\ y_i & X \in (F(y_{i-1}), F(y_i)\ (i \ge 2) \end{cases}$$

则Y的分布函数$F_Y(y) = F(y)$



例题

若已知在区间 $(0,1)$上均匀分布的随机数$x_1, x_2, \cdots, x_k$ $(k \geq 1)$,
如何生成参数为p的伯努利分布的随机数.

常用连续型分布的随机数

定理： 若随机变量$X \sim U(0,1)$, 且$F(y)$ 是某一个连续的分布函数,很显然反函数$F^{-1}(y)$ 存在, 则随机变量

$$Y = F^{-1}(X)$$

的分布函数为$F_Y(y) = F(y)$

例题

若已知在区间 $(0,1)$上均匀分布的随机数$x_1, x_2, \cdots, x_k$ $(k \geq 1)$,
如何生成参数为$\lambda = 4$的指数分布的随机数

标准正太分布的随机数

正太分布的分布函数不存在显示表达式, 难以得到它的反函数, 因此不能利用前面的定理直接生成服从正太分布的随机数

设X和Y是相互独立的标准正太分布随机变量, 则它们的极坐标

$$R = \sqrt{X^2 + Y^2}$$

$$\theta = \arctan(Y/X)$$

相互独立, 且R^2服从参数为$1/2$的指数分布, 而θ服从在 $(0,2\pi)$ 上的均匀分布.

正太分布随机数

由此可得相互独立的标准正太分布随机变量

$$X = R\cos(\theta) = (-2\ln X_1)^{1/2}\cos(2\pi X_2)$$

$$Y = R\sin(\theta) = (-2\ln X_1)^{1/2}\sin(2\pi X_2)$$

这种方法被称为 **Box-Muller方法**