第一章 随机变量的抽样方法

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分
- 对某分布抽样，一般先产生 $U(0,1)$ 的随机数，然后再转换为服从目标分布的随机数

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分
- 对某分布抽样，一般先产生 $U(0,1)$ 的随机数，然后再转换为服从目标分布的随机数
- 伪随机数序列的生成：$x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分
- 对某分布抽样，一般先产生 $U(0,1)$ 的随机数，然后再转换为服从目标分布的随机数
- 伪随机数序列的生成：$x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$
  - $U(0,1)$ 随机数发生器：集合 $\{0,1,\ldots,M\}$ 或 $\{1,2,\ldots,M\}$ 上的离散均匀分布，再除以 M 或 $M+1$

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分
- 对某分布抽样，一般先产生 $U(0,1)$ 的随机数，然后再转换为服从目标分布的随机数
- 伪随机数序列的生成：$x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$
  - $U(0,1)$ 随机数发生器：集合 $\{0,1,\ldots,M\}$ 或 $\{1,2,\ldots,M\}$ 上的离散均匀分布，再除以 M 或 $M+1$
  - 取值个数有限，序列一定在某个时间后发生重复

均匀分布随机变量的抽样

- 对随机变量抽样是统计模拟的基本工具
  - 物理方法 (抛硬币、抽签)：真随机，但数量少
  - 计算机抽样：伪随机数，但可以做到与目标分布真正的随机数无法通过统计检验区分
- 对某分布抽样，一般先产生 $U(0,1)$ 的随机数，然后再转换为服从目标分布的随机数
- 伪随机数序列的生成：$x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$
  - $U(0,1)$ 随机数发生器：集合 $\{0,1,\ldots,M\}$ 或 $\{1,2,\ldots,M\}$ 上的离散均匀分布，再除以 M 或 $M+1$
  - 取值个数有限，序列一定在某个时间后发生重复
  - 随机数发生器的周期：序列发生重复的间隔 T

线性同余发生器

定义 (同余)

设 i, j 为整数，M 为正整数，若 $j - i$ 为 M 的倍数，则称 i 与 j 关于 M同余，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

线性同余发生器

定义 (同余)

设 i,j 为整数，M 为正整数，若 $j-i$ 为 M 的倍数，则称 i 与 j 关于 M同余，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例如

$$7 \equiv 2 \pmod{5}, \quad -1 \equiv 4 \pmod{5}.$$

线性同余发生器

定义 (同余)

设 i, j 为整数，M 为正整数，若 $j - i$ 为 M 的倍数，则称 i 与 j 关于 M同余，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例如

$$7 \equiv 2 \pmod{5}, \quad -1 \equiv 4 \pmod{5}.$$

- **线性同余发生器**递推公式

$$x_n = a x_{n-1} + c \pmod{M}, n = 1, 2, \ldots$$

  - a, c: 事先设定的整数
  - $0 \leq x_n < M$, 令 $R_n = x_n / M \in [0, 1)$

线性同余发生器

定义 (同余)

设 i, j 为整数，M 为正整数，若 $j-i$ 为 M 的倍数，则称 i 与 j 关于 M同余，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例如

$$7 \equiv 2 \pmod{5}, \ -1 \equiv 4 \pmod{5}.$$

- **线性同余发生器**递推公式

$$x_n = a x_{n-1} + c \pmod{M}, n = 1, 2, \ldots$$

  - a, c: 事先设定的整数
  - $0 \le x_n < M$, 令 $R_n = x_n / M \in [0, 1)$
- x_n 只有 M 个取值，序列 $x_0, x_1, \ldots$ 一定会重复。使得 $x_n = x_m (n > m)$ 的 $n - m$ 的最小值 T 为随机数发生器在初值 x_0 下的周期 ($T \le M$)

线性同余发生器

定义 (同余)

设 i, j 为整数，M 为正整数，若 $j-i$ 为 M 的倍数，则称 i 与 j 关于 M同余，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例如

$$7 \equiv 2 \pmod{5}, \quad -1 \equiv 4 \pmod{5}.$$

- **线性同余发生器**递推公式

$$x_n = a x_{n-1} + c \pmod{M}, n = 1, 2, \ldots$$

  - a, c: 事先设定的整数
  - $0 \le x_n < M$, 令 $R_n = x_n / M \in [0, 1)$
- x_n 只有 M 个取值，序列 $x_0, x_1, \ldots$ 一定会重复。使得 $x_n = x_m (n > m)$ 的 $n - m$ 的最小值 T 为随机数发生器在初值 x_0 下的周期 $(T \le M)$
  - 练习 *1*：计算线性同余发生器

$$x_n = 3x_{n-1} + 3 \pmod{10}, n = 1, 2, \ldots$$

取初值 $x_0 = 3$ 的周期。

线性同余发生器

- 满周期: 从某个初值 x_0 出发达到最大周期 M

线性同余发生器

- 满周期: 从某个初值 x_0 出发达到最大周期 M

定理

当下列三个条件都满足时，线性同余发生器可以达到满周期：

1. c 与 M 互素
2. 对 M 的任一个素因子 P, $a-1$ 被 P 整除
3. 如果 4 是 M 的因子，则 $a-1$ 被 4 整除

线性同余发生器

- 满周期: 从某个初值 x_0 出发达到最大周期 M

定理

当下列三个条件都满足时，线性同余发生器可以达到满周期：

1. c 与 M 互素
2. 对 M 的任一个素因子 P, $a-1$ 被 P 整除
3. 如果 4 是 M 的因子，则 $a-1$ 被 4 整除

- 取 $a=4m+1$, $c=2n+1$ $(m,n\in\mathbb{N})$ 可达满周期

线性同余发生器

- 满周期: 从某个初值 x_0 出发达到最大周期 M

定理

当下列三个条件都满足时，线性同余发生器可以达到满周期:

1. c 与 M 互素
2. 对 M 的任一个素因子 P, $a-1$ 被 P 整除
3. 如果 4 是 M 的因子，则 $a-1$ 被 4 整除

- 取 $a = 4m+1$, $c = 2n+1$ $(m, n \in \mathbb{N})$ 可达满周期
- Kobayashi 提出了如下满周期 2^{31} 的线性同余发生器

$$x_n = 314159269 x_{n-1} + 453806245 \pmod{2^{31}}.$$

线性同余发生器

好的均匀分布随机数发生器应满足:

- 周期足够长，统计性质符合均匀分布

线性同余发生器

好的均匀分布随机数发生器应满足:

- 周期足够长，统计性质符合均匀分布
- 有很好的随机性，序列元素之间独立性好
  - 满周期的线性同余发生器，序列中前后两项自相关系数的近似公式为

$$\rho(1) \approx \frac{1}{a} - \frac{6c}{aM}\left(1 - \frac{c}{M}\right)$$

所以应将 a 选为较大的值 ($a < M$)。

FSR 发生器

线性同余发生器缺点:

- 产生的多维随机向量相关性大，分布不均匀
- 周期不可能超过 2^L (L 为计算机整数位数)

FSR 发生器

线性同余发生器缺点:

- 产生的多维随机向量相关性大，分布不均匀
- 周期不可能超过 2^L (L 为计算机整数位数)

Tausworthe (1965) 提出反馈位移寄存器法 (FSR):

$$\alpha_k = c_p \alpha_{k-p} + c_{p-1}\alpha_{k-p+1} + \cdots + c_1 \alpha_{k-1} \pmod 2$$

其中 $c_i \in \{0, 1\}$.

FSR 发生器

线性同余发生器缺点：

- 产生的多维随机向量相关性大，分布不均匀
- 周期不可能超过 2^L (L 为计算机整数位数)

Tausworthe (1965) 提出反馈位移寄存器法 (FSR):

$$\alpha_k = c_p \alpha_{k-p} + c_{p-1} \alpha_{k-p+1} + \cdots + c_1 \alpha_{k-1} \pmod 2$$

其中 $c_i \in \{0, 1\}$.

- 递推可以利用逻辑运算快速实现

FSR 发生器

线性同余发生器缺点:

- 产生的多维随机向量相关性大，分布不均匀
- 周期不可能超过 2^L (L 为计算机整数位数)

Tausworthe (1965) 提出反馈位移寄存器法 (FSR):

$$\alpha_k = c_p\alpha_{k-p} + c_{p-1}\alpha_{k-p+1} + \cdots + c_1\alpha_{k-1} \pmod 2$$

其中 $c_i \in \{0, 1\}$.

- 递推可以利用逻辑运算快速实现
- 对二进制序列 $\{\alpha_k : k = 1, 2, \ldots\}$ 后，依次截取长度 M 组合成整数 x_n，再令 $R_n = x_n/2^M \in [0, 1]$
- 作为多维均匀分布随机向量的发生器性质较好

FSR 发生器

线性同余发生器缺点:

- 产生的多维随机向量相关性大，分布不均匀
- 周期不可能超过 2^L (L 为计算机整数位数)

Tausworthe (1965) 提出反馈位移寄存器法 (FSR):

$$\alpha_k = c_p \alpha_{k-p} + c_{p-1} \alpha_{k-p+1} + \cdots + c_1 \alpha_{k-1} \pmod 2$$

其中 $c_i \in \{0, 1\}$.

- 递推可以利用逻辑运算快速实现
- 对二进制序列 $\{\alpha_k : k = 1, 2, \ldots\}$ 后，依次截取长度 M 组合成整数 x_n，再令 $R_n = x_n / 2^M \in [0, 1]$
- 作为多维均匀分布随机向量的发生器性质较好
- 通过选择递推系数和初值可以达到很长的周期，不受计算机字长限制

组合发生器

把若干个发生器组合利用，产生的随机数比单个发生器具有更长的周期和更好的随机性

- MacLaren and Marsaglia (1965) 提出组合同余法，组合两个同余发生器，其中一个用来 “搅乱” 次序

组合发生器

把若干个发生器组合利用，产生的随机数比单个发生器具有更长的周期和更好的随机性

- MacLaren and Marsaglia (1965) 提出组合同余法，组合两个同余发生器，其中一个用来 “搅乱” 次序
- Wichman and Hill (1982) 设计了如下的线性组合发生器。利用三个同余发生器：

$$
\begin{aligned}
U_n &= 171 U_{n-1} \pmod{30269} \\
V_n &= 172 V_{n-1} \pmod{30307} \\
W_n &= 170 W_{n-1} \pmod{30323}
\end{aligned}
$$

做线性组合并求余：

$$
R_n = (U_n/30269 + V_n/30307 + W_n/30323) \pmod{1}
$$

周期约为 7×10^{12}, 超过 $2^{31} \approx 2 \times 10^9$

随机数的检验

对 $U(0,1)$ 随机数发生器产生的序列 $\{R_i : i = 1, 2, \ldots, n\}$, 可以进行各种检验确认其均匀性:

- 把 [0,1] 等分成 K 段，用 Pearson's χ^2 test 检验 $\{R_i : i = 1, 2, \ldots, n\}$ 落在每一段的概率是否近似为 $1/K$

随机数的检验

对 $U(0,1)$ 随机数发生器产生的序列 $\{R_i : i = 1, 2, \ldots, n\}$, 可以进行各种检验确认其均匀性:

- 把 [0,1] 等分成 K 段，用 Pearson's χ^2 test 检验 $\{R_i : i = 1, 2, \ldots, n\}$ 落在每一段的概率是否近似为 $1/K$

- 用 Kolmogorov-Smirnov (K-S) test 检验 $\{R_i : i = 1, 2, \ldots, n\}$ 是否近似服从 $U[0,1]$ 分布

随机数的检验

对 $U(0,1)$ 随机数发生器产生的序列 $\{R_i : i = 1, 2, \ldots, n\}$, 可以进行各种检验确认其均匀性:

- 把 [0,1] 等分成 K 段，用 Pearson's χ^2 test 检验 $\{R_i : i = 1, 2, \ldots, n\}$ 落在每一段的概率是否近似为 $1/K$
- 用 Kolmogorov-Smirnov (K-S) test 检验 $\{R_i : i = 1, 2, \ldots, n\}$ 是否近似服从 $U[0,1]$ 分布
- 把 $\{R_i : i = 1, 2, \ldots, n\}$ 每 d 个组合在一起成为 $\mathbb{R}^d$ 向量，把超立方体 $[0,1]^d$ 每一维均匀分为 K 份，得到 K^d 个子集，用 Pearson's χ^2 test 检验这些 $\mathbb{R}^d$ 向量落在每个子集的概率是否近似为 $1/K^d$
- $\cdots$

非均匀分布随机变量的抽样

- 均匀分布随机数的产生方法是很多非均匀分布抽样方法的基础
- 通用抽样方法:
  - CDF 逆变换
  - Acceptance-Rejection (A-R) 抽样

非均匀分布随机变量的抽样

- 均匀分布随机数的产生方法是很多非均匀分布抽样方法的基础
- 通用抽样方法:
  - CDF 逆变换
  - Acceptance-Rejection (A-R) 抽样
- 其他常用抽样方法:
  - 单调变换，加和
  - Bootstrap
  - 混合分布抽样

CDF 逆变换

适用于对任何 CDF 逆函数已知的分布抽样

定义 (Cumulative distribution function (CDF))

一个随机变量 X 的 CDF $F(x)$ 定义为

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

CDF 逆变换

适用于对任何 CDF 逆函数已知的分布抽样

定义 (Cumulative distribution function (CDF))

一个随机变量 X 的 CDF $F(x)$ 定义为

$$F(x) = P(X \leq x), \ \ x \in \mathbb{R}.$$

- CDF 逆变换的基本想法: 假设随机变量 X 的 PDF $f(x) > 0, \forall x \in \mathbb{R}$, 其 CDF $F(x)$ 的逆函数 F^{-1} 存在，抽取 $U \sim \boldsymbol{U}[0,1]$, 令 $Y = F^{-1}(U)$, 则 $Y \sim F$.

CDF 逆变换

适用于对任何 CDF 逆函数已知的分布抽样

定义 (Cumulative distribution function (CDF))

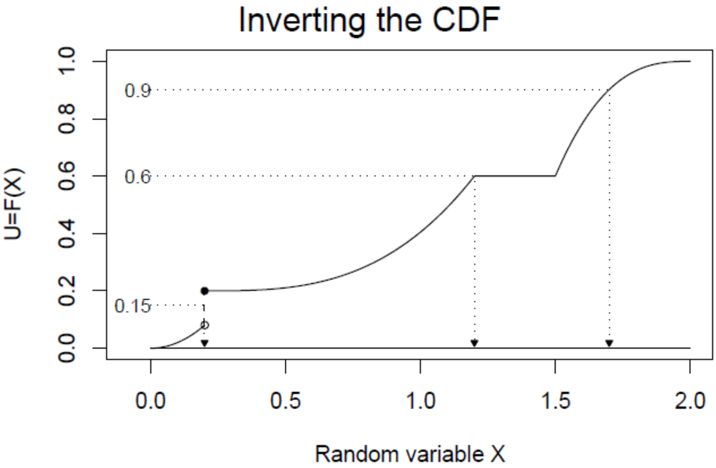
一个随机变量 X 的 CDF $F(x)$ 定义为

$$F(x) = P(X \leq x), \; x \in \mathbb{R}.$$

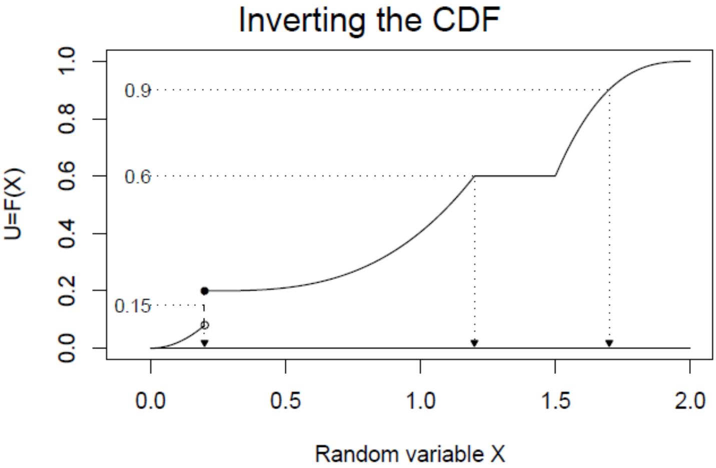
- CDF 逆变换的基本想法: 假设随机变量 X 的 PDF $f(x) > 0, \forall x \in \mathbb{R}$, 其 CDF $F(x)$ 的逆函数 F^{-1} 存在，抽取 $U \sim \boldsymbol{U}[0,1]$, 令 $Y = F^{-1}(U)$, 则 $Y \sim F$.
- 如果 CDF $F(x)$ 不连续 (离散分布) 或不可逆，可以定义如下的广义逆:

$$F^{-1}(u) = \inf\{x \mid F(x) \geq u\}, \quad 0 < u < 1 \tag{1}$$

CDF 逆变换


Inverting the CDF
U=F(X)
0.0
0.2
0.4
0.6
0.8
1.0
0.9
0.6
0.15
0.0
0.5
1.0
1.5
2.0
Random variable X

CDF 逆变换



$$F^{-1}(0.15) = 0.2,\ F^{-1}(0.6) = 1.2,\ F^{-1}(0.9) = 1.7$$

CDF 逆变换

CDF 逆变换: 设 F 是一个 CDF, F^{-1} 是由(1)定义的逆函数，令 $U \sim \boldsymbol{U}[0,1]$, 则 $X = F^{-1}(U) \sim F$

CDF 逆变换

CDF 逆变换: 设 F 是一个 CDF, F^{-1} 是由(1)定义的逆函数，令 $U \sim \boldsymbol{U}[0,1]$, 则 $X = F^{-1}(U) \sim F$

- 如果 $U \sim \boldsymbol{U}[0,1]$, $1 - U \sim \boldsymbol{U}[0,1]$, 因此 $F^{-1}(1-U) \sim F$

CDF 逆变换

CDF 逆变换: 设 F 是一个 CDF, F^{-1} 是由(1)定义的逆函数，令 $U \sim \boldsymbol{U}[0,1]$, 则 $X = F^{-1}(U) \sim F$

- 如果 $U \sim \boldsymbol{U}[0,1]$, $1 - U \sim \boldsymbol{U}[0,1]$, 因此 $F^{-1}(1-U) \sim F$
- 如果 F 是一个连续的 CDF, $X \sim F$, G 是任意分布的 CDF，则随机变量
$$Y = G^{-1}(F(X)) \sim G \tag{2}$$
称函数 $G^{-1}(F(\cdot))$ 为QQ 变换

CDF 逆变换

CDF 逆变换: 设 F 是一个 CDF, F^{-1} 是由(1)定义的逆函数，令 $U \sim \boldsymbol{U}[0,1]$, 则 $X = F^{-1}(U) \sim F$

- 如果 $U \sim \boldsymbol{U}[0,1]$, $1-U \sim \boldsymbol{U}[0,1]$, 因此 $F^{-1}(1-U) \sim F$
- 如果 F 是一个连续的 CDF, $X \sim F$, G 是任意分布的 CDF，则随机变量

$$Y = G^{-1}(F(X)) \sim G \tag{2}$$

  称函数 $G^{-1}(F(\cdot))$ 为QQ 变换

  - QQ 变换将分布 F 的分位数转换为分布 G 下相应的分位数

CDF 逆变换举例

- **指数分布**. 标准指数分布 Exp(1) 的 PDF:

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

  CDF 的逆函数:

  $$F^{-1}(u) = -\log(1-u)$$

  先抽取 $U \sim \boldsymbol{U}(0,1)$, 令 $X = -\log(1-U)$ 或 $X = -\log(U)$ 可得服从 Exp(1) 的样本

CDF 逆变换举例

- **指数分布**. 标准指数分布 Exp(1) 的 PDF:

$$f(x) = e^{-x}, \; x > 0$$

CDF 的逆函数:

$$F^{-1}(u) = -\log(1-u)$$

先抽取 $U \sim \boldsymbol{U}(0,1)$, 令 $X = -\log(1-U)$ 或 $X = -\log(U)$ 可得服从 Exp(1) 的样本

  - $Y \sim \text{Exp}(\lambda)$ 的 PDF:

$$f(y) = \lambda e^{-\lambda y}, \; y > 0$$

如果 $X \sim \text{Exp}(1)$, 则 $X/\lambda \sim \text{Exp}(\lambda)$，因此 $Y = -\log(U)/\lambda \sim \text{Exp}(\lambda)$

CDF 逆变换举例

- **指数分布**. 标准指数分布 Exp(1) 的 PDF:

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

CDF 的逆函数:

$$F^{-1}(u) = -\log(1-u)$$

先抽取 $U \sim \boldsymbol{U}(0,1)$, 令 $X = -\log(1-U)$ 或 $X = -\log(U)$ 可得服从 Exp(1) 的样本

  - $Y \sim \text{Exp}(\lambda)$ 的 PDF:

$$f(y) = \lambda e^{-\lambda y}, \ y > 0$$

    如果 $X \sim \text{Exp}(1)$, 则 $X/\lambda \sim \text{Exp}(\lambda)$，因此 $Y = -\log(U)/\lambda \sim \text{Exp}(\lambda)$
  - 指数分布常用于描述一段时间的分布，它的一个重要特性是无记忆性: 如果 $X \sim \text{Exp}(\lambda)$, 则

$$P(X \geq x + \Delta \mid X \geq x) = \frac{e^{-\lambda(x+\Delta)}}{e^{-\lambda x}} = e^{-\lambda \Delta}$$

CDF 逆变换举例

- **Bernoulli 分布**. 从 Bernoulli 分布 ($P(X=1)=p$, $P(X=0)=1-p$) 抽样可以令 $X=\mathbf{1}(1-U\le p)$ 或 $X=\mathbf{1}(U\le p)$ 实现

CDF 逆变换举例

- **Bernoulli 分布**. 从 Bernoulli 分布 ($P(X=1)=p$, $P(X=0)=1-p$) 抽样可以令 $X=\mathbf{1}(1-U\le p)$ 或 $X=\mathbf{1}(U\le p)$ 实现

- **Cauchy 分布**. Cauchy 分布是 t 分布的一个特例，具有**厚尾**的特性，PDF:

$$f(x)=\frac{1}{\pi\cdot(1+x^2)},\ x\in\mathbb{R}$$

CDF:

$$F(x)=\frac{1}{\pi}\arctan(x)+\frac{1}{2}$$

可利用如下 CDF 逆变换对 Cauchy 分布抽样:

$$X=\tan\left(\pi\cdot(U-1/2)\right),\ U\sim\boldsymbol{U}(0,1)$$

几何解释：Cauchy 变量是一个在 $(-\pi/2,\pi/2)$ 上均匀分布的随机角的正切

CDF 逆变换举例

- **Poisson 分布.** $X \sim \mathrm{Po}(\lambda)$ 的 PMF:

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!},\ \ x=0,1,2,\ldots$$

Devroye (1986) 基于 CDF 逆变换提出以下算法从 Poisson 分布抽样 (考察算法输出 0 和 1 的概率)

Algorithm 1 Sample from Poisson distribution $\mathrm{Po}(\lambda)$

Initialize $X=0,\ p=q=e^{-\lambda}$, generate $U \sim \boldsymbol{U}(0,1)$.
while $U>q$ **do**
  $X=X+1$
  $p=p\lambda/X$
  $q=q+p$
return X

CDF 逆变换举例

- **Poisson 分布**. $X \sim \text{Po}(\lambda)$ 的 PMF:

$$P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!},\ \ x=0,1,2,\ldots$$

Devroye (1986) 基于 CDF 逆变换提出以下算法从 Poisson 分布抽样 (考察算法输出 0 和 1 的概率)

Algorithm 2 Sample from Poisson distribution $\text{Po}(\lambda)$

```
Initialize X = 0, p = q = e^{-λ}, generate U ~ U(0,1).
while U > q do
    X = X + 1
    p = pλ/X
    q = q + p
return X
```

  - 条件 $U > q$ 会被检验 $X+1$ 次，λ 较大时使用该方法抽样会很慢

CDF 逆变换举例

- **Weibull 分布**. Weibull 分布是对指数分布的推广，PDF:

$$f(x) = \frac{k}{\sigma}\left(\frac{x}{\sigma}\right)^{k-1} e^{-(x/\sigma)^k}, \ x > 0$$

参数 $\sigma > 0$, $k > 0$, CDF:

$$F(x) = 1 - \exp\left(-(x/\sigma)^k\right), \ x > 0$$

因此对 $U \sim \boldsymbol{U}(0,1)$ 做变换 $X = \sigma\left(-\log(1-U)\right)^{1/k}$ 可得到 Weibull 分布的样本

CDF 逆变换举例

- **Weibull 分布**. Weibull 分布是对指数分布的推广，PDF:

$$f(x) = \frac{k}{\sigma}\left(\frac{x}{\sigma}\right)^{k-1} e^{-(x/\sigma)^k}, \ x > 0$$

参数 $\sigma > 0$, $k > 0$, CDF:

$$F(x) = 1 - \exp\left(-(x/\sigma)^k\right), \ x > 0$$

因此对 $U \sim \boldsymbol{U}(0,1)$ 做变换 $X = \sigma\left(-\log(1-U)\right)^{1/k}$ 可得到 Weibull 分布的样本

定义 (Hazard function)

对一个随机变量 $X > 0$, 它的 hazard function $h(x)$ 定义为

$$h(x) = \lim_{t \to 0^+} \frac{1}{t} P(X \le x + t \mid X \ge x), \ x > 0.$$

Hazard function 也被称为瞬时失败概率

CDF 逆变换举例

- Weibull 分布的 hazard function:

$$h(x) = k \cdot \frac{x^{k-1}}{\sigma^k}$$

  - $k < 1$ 时, $h(x)$ 随时间 x 递减，比如婴儿的死亡概率
  - $k = 1$ 时，Weibull 退化为指数分布，$h(x)$ 是常数
  - $k > 1$ 时，$h(x)$ 随时间 x 递增，比如非耐用品的生命周期

CDF 逆变换举例

- Weibull 分布的 hazard function:

$$h(x) = k \cdot \frac{x^{k-1}}{\sigma^k}$$

  - $k < 1$ 时，$h(x)$ 随时间 x 递减，比如婴儿的死亡概率
  - $k = 1$ 时，Weibull 退化为指数分布，$h(x)$ 是常数
  - $k > 1$ 时，$h(x)$ 随时间 x 递增，比如非耐用品的生命周期

- **双指数分布**. 标准双指数分布的 PDF:

$$f(x) = \frac{1}{2}\exp(-|x|), \ \ x \in \mathbb{R}$$

练习：求标准双指数分布的 CDF 及其逆函数

CDF 逆变换举例

- Weibull 分布的 hazard function:

$$h(x) = k \cdot \frac{x^{k-1}}{\sigma^k}$$

  - $k < 1$ 时, $h(x)$ 随时间 x 递减，比如婴儿的死亡概率
  - $k = 1$ 时，Weibull 退化为指数分布，$h(x)$ 是常数
  - $k > 1$ 时，$h(x)$ 随时间 x 递增，比如非耐用品的生命周期

- **双指数分布**. 标准双指数分布的 PDF:

$$f(x) = \frac{1}{2}\exp(-|x|), \ \ x \in \mathbb{R}$$

练习：求标准双指数分布的 CDF 及其逆函数

$$F(x) = \begin{cases} \frac{1}{2}e^x, & x < 0 \\ 1 - \frac{1}{2}e^{-x}, & x \geq 0 \end{cases} \qquad F^{-1}(u) = \begin{cases} \log(2u), & 0 < u \leq 1/2 \\ -\log(2(1-u)), & 1/2 < u < 1 \end{cases}$$

CDF 逆变换举例

- **Triangular and power densities**. Triangular 和 power 分布都是 Beta 分布的特例，后面会介绍更高效的抽样方法
  - Triangular density:
    $$f(x) = 2x,\ \ 0 < x < 1$$
    $F^{-1}(U) = \sqrt{U},\ U \sim \boldsymbol{U}(0,1)$
  - Power density ($\alpha > 0$):
    $$f(x) = \alpha x^{\alpha-1},\ \ 0 < x < 1.$$
    $F^{-1}(U) = U^{1/\alpha},\ U \sim \boldsymbol{U}(0,1)$

CDF 逆变换举例

- **Triangular and power densities**. Triangular 和 power 分布都是 Beta 分布的特例，后面会介绍更高效的抽样方法
  - Triangular density:
  $$f(x) = 2x,\ \ 0 < x < 1$$
  $F^{-1}(U) = \sqrt{U},\ U \sim \boldsymbol{U}(0,1)$
  - Power density ($\alpha > 0$):
  $$f(x) = \alpha x^{\alpha-1},\ \ 0 < x < 1.$$
  $F^{-1}(U) = U^{1/\alpha},\ U \sim \boldsymbol{U}(0,1)$
- **截断分布抽样**. 随机变量 $Y \sim F$ (F 是一个连续分布的 CDF), X 服从 F 在 (a,b) 上的截断分布 G, 则 X 的 CDF 为
$$G(x) = \frac{F(x) - F(a)}{F(b) - F(a)}$$

CDF 逆变换举例

- **Triangular and power densities**. Triangular 和 power 分布都是 Beta 分布的特例，后面会介绍更高效的抽样方法
  - Triangular density:
    $$f(x) = 2x,\ 0 < x < 1$$
    $F^{-1}(U) = \sqrt{U}$, $U \sim \boldsymbol{U}(0,1)$
  - Power density ($\alpha > 0$):
    $$f(x) = \alpha x^{\alpha-1},\ 0 < x < 1.$$
    $F^{-1}(U) = U^{1/\alpha}$, $U \sim \boldsymbol{U}(0,1)$
- **截断分布抽样**. 随机变量 $Y \sim F$ (F 是一个连续分布的 CDF), X 服从 F 在 (a,b) 上的截断分布 G, 则 X 的 CDF 为
  $$G(x) = \frac{F(x) - F(a)}{F(b) - F(a)}$$
  因此可通过以下变换从截断分布 G 抽样:
  $$X = F^{-1}\left(F(a) + (F(b) - F(a))U\right),\ U \sim \boldsymbol{U}(0,1)$$

单调变换

- CDF 逆变换通用但不总是很高效，有时利用分布之间的特殊关系可以构造更简洁的抽样变换
- 假设随机变量 X 的抽样方法已知，如果随机变量 $Y = \tau(X)$，$\tau(\cdot)$ 是一个单调递增函数，则 Y 的样本可通过对 X 的样本做单调变换 τ 得到

单调变换

- CDF 逆变换通用但不总是很高效，有时利用分布之间的特殊关系可以构造更简洁的抽样变换

- 假设随机变量 X 的抽样方法已知，如果随机变量 $Y = \tau(X)$，$\tau(\cdot)$ 是一个单调递增函数，则 Y 的样本可通过对 X 的样本做单调变换 τ 得到

  - 对数正态分布：如果 $X \sim N(\mu, \sigma^2)$，$Y = \exp(X)$ 服从对数正态分布

  - 当 $X \sim N(0, 1)$，$Y = \mu + \sigma X \sim N(\mu, \sigma^2)$

单调变换

- CDF 逆变换通用但不总是很高效，有时利用分布之间的特殊关系可以构造更简洁的抽样变换
- 假设随机变量 X 的抽样方法已知，如果随机变量 $Y = \tau(X)$，$\tau(\cdot)$ 是一个单调递增函数，则 Y 的样本可通过对 X 的样本做单调变换 τ 得到
  - 对数正态分布：如果 $X \sim N(\mu, \sigma^2)$，$Y = \exp(X)$ 服从对数正态分布
  - 当 $X \sim N(0,1)$，$Y = \mu + \sigma X \sim N(\mu, \sigma^2)$
  - 任何一个单调变换 $Y = \tau(X)$ 都对应一个 QQ 变换 $F_Y^{-1}(F_X(\cdot))$

Box-Muller 变换

- Box-Muller 变换用两个独立的 $U(0,1)$ 变量产生两个独立的 $N(0,1)$ 变量：

$$\begin{aligned} Z_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2) \\ Z_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2) \end{aligned} \tag{3}$$

其中 $U_1, U_2 \sim \boldsymbol{U}(0,1)$ 且独立

Box-Muller 变换

- Box-Muller 变换用两个独立的 $U(0,1)$ 变量产生两个独立的 $N(0,1)$ 变量：

$$
\begin{aligned}
Z_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2) \\
Z_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2)
\end{aligned} \tag{3}
$$

  其中 $U_1, U_2 \sim \boldsymbol{U}(0,1)$ 且独立

- 原理：将二元随机向量 $(Z_1, Z_2) \sim N(\mathbf{0}, I_2)$ 用极坐标表示

$$
\begin{aligned}
Z_1 &= R\cos(\theta) \\
Z_2 &= R\sin(\theta)
\end{aligned} \tag{4}
$$

  可以证明极角 $\theta \sim \boldsymbol{U}[0, 2\pi)$, 且独立于半径 R; 半径 $R^2 = Z_1^2 + Z_2^2 \sim \chi^2_{(2)}$

Box-Muller 变换

- Box-Muller 变换用两个独立的 $U(0,1)$ 变量产生两个独立的 $N(0,1)$ 变量：

$$
\begin{aligned}
Z_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2) \\
Z_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2)
\end{aligned}
\tag{3}
$$

  其中 $U_1, U_2 \sim \boldsymbol{U}(0,1)$ 且独立

- 原理：将二元随机向量 $(Z_1, Z_2) \sim N(\mathbf{0}, I_2)$ 用极坐标表示

$$
\begin{aligned}
Z_1 &= R\cos(\theta) \\
Z_2 &= R\sin(\theta)
\end{aligned}
\tag{4}
$$

  可以证明极角 $\theta \sim \boldsymbol{U}[0, 2\pi)$, 且独立于半径 R; 半径 $R^2 = Z_1^2 + Z_2^2 \sim \chi^2_{(2)}$

- Box-Muller 不是最快从 $N(0,1)$ 抽样的方法

最大、最小统计量

- 假设 $X_1, \ldots, X_r$ 独立同分布且 CDF 是 F, $Y = \max(X_1, \ldots, X_r)$ 的 CDF:

$$P(Y \leq y) = (F(y))^r$$

最大、最小统计量

- 假设 $X_1, \ldots, X_r$ 独立同分布且 CDF 是 F, $Y = \max(X_1, \ldots, X_r)$ 的 CDF:

$$P(Y \leq y) = (F(y))^r$$

  因此对 CDF 为 $G = F^r$ 的分布抽样，可以先从分布 F 独立抽取 r 个样本，然后只保留最大样本

  - 三角分布的 CDF $G(y) = y^2$, $0 < y < 1$. 用 $\max(U_1, U_2)$ 抽样比逆变换法 $\sqrt{U_1}$ 快

最大、最小统计量

- 假设 $X_1, \ldots, X_r$ 独立同分布且 CDF 是 F, $Y = \max(X_1, \ldots, X_r)$ 的 CDF:

  $$P(Y \leq y) = (F(y))^r$$

  因此对 CDF 为 $G = F^r$ 的分布抽样，可以先从分布 F 独立抽取 r 个样本，然后只保留最大样本

  - 三角分布的 CDF $G(y) = y^2$, $0 < y < 1$. 用 $\max(U_1, U_2)$ 抽样比逆变换法 $\sqrt{U_1}$ 快

- $Y = \min(X_1, \ldots, X_r)$ 的 CDF:

  $$G(y) = 1 - (1 - F(y))^r$$

最大、最小统计量

- 假设 $X_1, \ldots, X_r$ 独立同分布且 CDF 是 F, $Y = \max(X_1, \ldots, X_r)$ 的 CDF:

$$P(Y \leq y) = (F(y))^r$$

  因此对 CDF 为 $G = F^r$ 的分布抽样，可以先从分布 F 独立抽取 r 个样本，然后只保留最大样本

  - 三角分布的 CDF $G(y) = y^2$, $0 < y < 1$. 用 $\max(U_1, U_2)$ 抽样比逆变换法 $\sqrt{U_1}$ 快

- $Y = \min(X_1, \ldots, X_r)$ 的 CDF:

$$G(y) = 1 - (1 - F(y))^r$$

  - 考察 $Y = \min(U_1, U_2)$ 的分布，其中 $U_1, U_2 \sim \boldsymbol{U}(0, 1)$. Y 的 CDF:

$$G(y) = 1 - (1 - y)^2, \ 0 < y < 1$$

    如果用 CDF 逆变换对该分布抽样，需计算 $Y = 1 - \sqrt{1 - U_1}$

顺序统计量

最大、最小统计量都是**顺序统计量**的特例

定义 (顺序统计量)

对于 n 个独立同分布的随机变量，$X_1, \ldots, X_n$. 它们的顺序统计量是将这 n 个变量的取值按从小到大的顺序排列，记为

$$X_{(1)} \leq X_{(2)} \leq \ldots, \leq X_{(n)}.$$

顺序统计量

最大、最小统计量都是**顺序统计量**的特例

定义 (顺序统计量)

对于 n 个独立同分布的随机变量, $X_1, \ldots, X_n$. 它们的顺序统计量是将这 n 个变量的取值按从小到大的顺序排列，记为

$$X_{(1)} \leq X_{(2)} \leq \ldots, \leq X_{(n)}.$$

- 对于 n 个独立的 $\boldsymbol{U}(0,1)$ 随机变量 $U_1, \ldots, U_n$, 证明 $U_{(r)} \sim \text{Beta}(r, n-r+1)$

顺序统计量

最大、最小统计量都是**顺序统计量**的特例

定义 (顺序统计量)

对于 n 个独立同分布的随机变量, $X_1, \ldots, X_n$. 它们的顺序统计量是将这 n 个变量的取值按从小到大的顺序排列，记为

$$X_{(1)} \leq X_{(2)} \leq \ldots, \leq X_{(n)}.$$

- 对于 n 个独立的 $\boldsymbol{U}(0,1)$ 随机变量 $U_1, \ldots, U_n$, 证明 $U_{(r)} \sim \text{Beta}(r, n-r+1)$
- n 个独立同分布的 $Y_i \sim F$, n 很大时，一个快速得到 $Y_{(r)}$ 样本的方法：先抽取 $X \sim \text{Beta}(r, n-r+1)$, 则 $Y_{(r)} = F^{-1}(X)$

顺序统计量

练习：一个系统由 n 个独立的元件组成，每个元件或者工作或者不工作。至少需要 k 个元件工作才能保证系统正常运行。假设在 0 时刻，所有元件都正常工作，用 Y_i 表示元件 i 不工作的时刻，$Y_i > 0$ 且 Y_i 独立服从 Weibull$(\sigma = 1, k = 2)$ 分布（元件会老化），$i = 1, \ldots, n$.
Weibull$(\sigma = 1, k = 2)$ 的 CDF 为

$$F(x) = 1 - \exp(-x^2), \ x > 0.$$

用 S 表示系统停止运行的时刻，如何得到 S 的样本?

加和

如果随机变量 Y 可写为 n 个独立同分布的随机变量之和，$Y = X_1 + \cdots + X_n$，且 X_i 的分布较简单，对 Y 抽样可以先从 X_i 的分布中独立抽取 n 个样本，再加和

- **二项分布**. $Y \sim \text{Bin}(n, p)$ 等于 n 个 Bernoulli 变量的和
- χ^2 **分布**. 如果 $X_i \overset{iid}{\sim} \chi^2_{(\alpha)}$, $i = 1, 2, \ldots, n$. 则

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2_{(n\alpha)}.$$

加和

如果随机变量 Y 可写为 n 个独立同分布的随机变量之和，$Y = X_1 + \cdots + X_n$，且 X_i 的分布较简单，对 Y 抽样可以先从 X_i 的分布中独立抽取 n 个样本，再加和

- **二项分布**. $Y \sim \text{Bin}(n, p)$ 等于 n 个 Bernoulli 变量的和
- χ^2 **分布**. 如果 $X_i \overset{iid}{\sim} \chi^2_{(\alpha)}$, $i = 1, 2, \ldots, n$. 则

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2_{(n\alpha)}.$$

- **Noncentral** χ^2 **分布**. 分布有两个参数, 自由度 n 和参数 $\lambda \geq 0$, 记为 $\chi'^2_{(n)}(\lambda)$. 可如下生成:

$$Y = \sum_{i=1}^{n} X_i^2, \quad X_i \overset{ind}{\sim} N(a_i, 1), \quad \lambda = \sqrt{\sum_{i=1}^{n} a_i^2}$$