随机变量的产生方法

王璐

生成随机变量是统计模拟的一个基本工具。我们可以用物理方法得到一组真实的随机数，比如反复抛掷硬币、骰子、抽签、摇号等，这些方法得到的随机数质量好，但是数量不能满足随机模拟的需要。主流的方法是使用计算机产生**伪随机数**。伪随机数是由计算机算法生成的序列 $\{x_i, i = 1, 2, \dots\}$, 因为计算机算法的结果是固定的，所以伪随机数不是真正的随机数，但是好的伪随机数序列可以做到与理论上真正的分布 F 无法通过统计检验区分开，所以我们也把计算机生成的伪随机数视为随机数。

需要生成某种分布的随机数时，一般先产生服从均匀分布的随机数，然后再将其转换为服从其它分布的随机数。

1 均匀分布随机变量的产生

计算机中伪随机数序列是迭代生成的，即 $x_n = g(x_{n-1}, x_{n-2}, \dots, x_{n-p})$, g 是确定的函数。均匀分布随机数发生器首先生成的是在集合 $\{0, 1, \dots, M\}$ 或 $\{1, 2, \dots, M\}$ 上离散取值的服从离散均匀分布的随机数，然后除以 M 或 $M + 1$ 变成 $[0, 1]$ 内的值当作服从连续均匀分布的随机数。这种方法实际上只取了有限个值，因为取值个数有限，根据算法 $x_n = g(x_{n-1}, x_{n-2}, \dots, x_{n-p})$ 可知序列一定在某个时间后发生重复，使得序列发生重复的间隔 T 叫做随机数发生器的周期。好的随机数发生器可以保证 M 很大且周期很长。现在常用的均匀分布随机数发生器由线性同余法、反馈位寄存器法以及随机数发生器的组合。这部分内容主要参考李东风 (2016) 第二章 2.1 节。

1.1 线性同余发生器

Definition 1.1 (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** (congruential)，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例如

$$11 \equiv 1 \pmod{10},\ -9 \equiv 1 \pmod{10}.$$

对于整数 A, 用 $A \pmod M$ 表示 A 除以 M 的余数，显然 A 和 $A \pmod M$ 同余，且 $0 \le A \pmod M < M$。

线性同余发生器利用求余运算生成随机数，其递推公式为

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

其中 a 和 c 是事先设定的整数。取某个整数初值 x_0 后可以往下递推得到序列 $\{x_n\}$。注意到 $0 \le x_n < M$, 令 $R_n = x_n/M$, 则 $R_n \in [0, 1)$, 最后把序列 $\{R_n\}$ 作为均匀分布的随机数序列输出。

因为线性同余法的递推算法仅依赖于前一项，序列元素取值只有 M 个可能值，所以产生的序列 $x_0, x_1, \ldots$ 一定会重复。若存在正整数 n 和 m 使得 $x_n = x_m (n > m)$, 则必有 $x_{n+k} = x_{m+k}$, $k = 1, 2, \ldots$, 即 $x_n, x_{n+1}, x_{n+2}, \ldots$ 重复了 $x_m, x_{m+1}, x_{m+2}, \ldots$, 称这样的 $n - m$ 的最小值 T 为此随机数发生器在初值 x_0 下的周期。由序列取值的有限性可知 $T \le M$。

练习 *1*: 计算线性同余发生器

$$x_n = 7x_{n-1} + 7 \pmod{10}, n = 1, 2, \ldots$$

取初值 $x_0 = 7$ 的周期。(数列为 $7, 6, 9, 0, 7, 6, 9, 0, 7, \ldots$, 周期为 $T = 4$)

练习 *2*: 计算线性同余发生器

$$x_n = 5x_{n-1} + 1 \pmod 8, n = 1, 2, \ldots$$

取初值 $x_0 = 1$ 的周期。(数列为 $1, 6, 7, 4, 5, 2, 3, 0, 1, 6, 7, \ldots$, 周期为 $T = 8 = M$, 达最大周期)

当线性同余发生器从某个初值 x_0 出发达到最大周期 M, 也称**满周期**，则初值 x_0 取任意整数产生的序列都会达到满周期，序列总是从 x_M 开始重复。如果发生器从 x_0 出发不是满周期的，那么它从任何整数出发都不是满周期的。适当选取 M, a, c 可以使产生的随机数序列和真正的 $U[0, 1]$ 随机数表现接近。

Theorem 1. 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

常取 $M = 2^L$, L 为计算机中整数的位数。根据定理1, 可取 $a = 4m + 1$, $c = 2n + 1$ (m 和 n 是任意正整数)，这样的线性同余发生器是满周期的。例如 Kobayashi 提出了如下的满周期 2^{31} 的线性同余发生器

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

其周期较长，统计性质比较好。

- 好的均匀分布随机数发生器应该周期足够长，统计性质符合均匀分布。把同余法生成的数列看成随机变量序列 $\{X_n\}$，在满周期时，可认为 X_n 是从 $\{0,1,\ldots,M-1\}$ 中随机等可能选取的，即

$$P(X_n = i) = 1/M,\ i = 0,1,\ldots,M-1$$

此时

$$E(X_n) = \sum_{i=0}^{M-1} i\frac{1}{M} = \frac{M-1}{2}$$

$$Var(X_n) = E(X_n^2) - [E(X_n)]^2 = \sum_{i=0}^{M-1} i^2\frac{1}{M} - \frac{(M-1)^2}{4} = \frac{1}{12}(M^2-1)$$

于是当 M 很大时

$$E(R_n) = E(X_n/M) = \frac{1}{2} - \frac{1}{2M} \approx \frac{1}{2}$$

$$Var(R_n) = Var(X_n/M) = \frac{1}{12} - \frac{1}{12M^2} \approx \frac{1}{12}$$

可见生成数列的期望和方差很接近均匀分布。

- 好的随机数发生器还应该有很好的随机性，产生的序列不应该有规律，序列之间独立性好。但是随机数发生器产生的序列是由确定的公式生成，不可能做到真正独立，至少我们要求序列的自相关性较弱。对于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

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

所以应该将 a 选为较大的值 ($a < M$)。

1.2 FSR 发生器

线性同余发生器产生一维均匀分布随机数效果很好，但产生的多维随机向量相关性大，分布不均匀。而且线性同余法的周期不可能超过 2^L。Tausworthe (1965) 提出一种新的做法——反馈位移寄存器法 (FSR), 对这些方面有改进。

FSR 按照如下递推法则生成一系列取值为 0 或 1 的数 $\alpha_1, \alpha_2, \ldots$, 每个 α_k 由前面若干个 $\{\alpha_i\}$ 的线性组合除以 2 的余数产生：

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

其中每个系数 c_i 只取 0 或 1，这样的递推可以利用程序语言中的逻辑运算快速实现。比如，如果 FSR 算法中的系数 $(c_1, c_2, \ldots, c_p)$ 仅有两个为 1，e.g. $c_p = c_{p-q} = 1 (1 < q < p)$, 递推法则可写为：

$$\begin{aligned}\alpha_k &= \alpha_{k-p} + \alpha_{k-p+q} \pmod 2 \\ &= \begin{cases} 0 & \text{if } \alpha_{k-p} = \alpha_{k-p+q} \\ 1 & \text{if } \alpha_{k-p} \neq \alpha_{k-p+q}. \end{cases}\end{aligned}$$

这可以用计算机的异或运算 $\oplus$ 进行快速计算：

$$\alpha_k = \alpha_{k-p} \oplus \alpha_{k-p+q},\ k = 1, 2, \ldots$$

给定初值 $(\alpha_{-p+1}, \alpha_{-p+2}, \ldots, \alpha_0)$ 递推得到序列 $\{\alpha_k : k = 1, 2, \ldots\}$ 后，依次截取长度为 M 的二进制序列组合成整数 x_n, 再令 $R_n = x_n / 2^M$。巧妙选择递推系数和初值（种子）可以得到很长的周期，且作为多维均匀分布随机向量的发生器性质较好。在上述 $c_p = c_{p-q} = 1 (1 < q < p)$ 的例子中，递推算法只需要异或运算，不受计算机字长限制，适当选取 p, q 后周期可以达到 $2^p - 1$ (如取 $p = 98$)。

1.3 组合发生器法

随机数设计中比较困难的是独立性和多维的分布。可以考虑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) 提出了组合同余法，组合两个同余发生器，一个用来 “搅乱” 次序。将两个同余发生器记为 A 和 B。用 A 产生 m 个随机数 (e.g. m=128)，存放在数组 $T = (t_1, t_2, \ldots, t_m)$。需要产生 x_n 时，从 B 中生成一个随机下标 $j \in \{1, 2, \ldots, m\}$，取 $x_n = t_j$, 然后从 A 再生成一个新随机数替代 T 中的 t_j, 如此重复。这样组合可以增强随机性，加大周期 (可超过 2^L)。也可以只使用一个发生器，用 x_{n-1} 来选择下标。

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：

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

这个组合发生器的周期约有 7×10^{12}, 超过 $2^{31} \approx 2 \times 10^9$.

在 R 软件中，用 `runif(n)` 产生 n 个 $U(0,1)$ 均匀分布的随机数。R 提供了若干种随机数发生器，可以用 `RNGkind()` 函数切换。在使用随机数进行模拟时，如果希望模拟的结果可重复，就需要在模拟开始时设置固定的随机数种子。在 R 中，可以用函数 `set.seed(m)` 来设置种子，其中 m 是任意整数。

1.4 随机数的检验

对均匀分布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。一些检验的想法有：

- 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 $1/K$.
  - Pearson's χ^2 test 可以检验样本落入若干互斥分类的概率分布是否等于某个特定的离散分布 (reference distribution)。其原理是通过每个分类的实际观察次数与理论期望次数之差构造统计量。
- 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。
  - K-S test 可以检验样本是否服从某个特定的一维连续分布 (reference distribution)，其原理是通过定义样本的 empirical CDF 与 reference CDF 的距离构造统计量。
- 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 检验这些组合得到的 $\mathbb{R}^d$ 向量落在每个子集的概率是否近似为 $1/K^d$.
- $\cdots$

2 非均匀分布随机变量的产生

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石。常用的科学计算软件, 如 R、Matlab，都提供了很多常见的非均匀分布的抽样函数，如 normal, Poisson, binomial, exponential, gamma, beta, etc. 但是有时候我们可能需要从某个特殊分布中抽样，而这些软件没有提供现成的抽样函数。因此我们需要理解这些非均匀分布的随机数是如何生成的，以便在必要的时候 make a custom solution. 在这部分我们会学习一些通用型的方法，如 CDF 逆变换, acceptance-rejection sampling 等。这部分内容主要参考Owen (2013) Chapter 4.

2.1 CDF 逆变换

将均匀分布的随机变量转化为非均匀分布的随机变量最直接的方法是 CDF 逆变换 (inverse CDF transform). 理论上这个方法适用于任何 CDF 的逆函数已知的分布。

Definition 2.1 (Cumulative distribution function (CDF))**.** 一个随机变量 X 的 CDF $F(x)$ 定义为

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

任一分布可由它的 CDF 完全刻画。CDF 有如下性质:

- $F(+\infty) = 1$
- $F(-\infty) = 0$
- 右连续: $\lim\limits_{x \to y^+} F(x) = F(y)$

对于一个连续分布，CDF 和它的 PDF (probability density function) $f(x) \geq 0$ 的关系是

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a).$$

假设随机变量 X 的 PDF $f(x) > 0, \forall x \in \mathbb{R}$. 那么它的 CDF $F(x)$ 单调且连续，因此存在逆函数 F^{-1}. 由于 $0 \leq F(x) \leq 1$, 如果我们抽取 $U \sim \boldsymbol{U}[0,1]$, 然后考察 $Y = F^{-1}(U)$ 的分布会发现

$$\begin{aligned} P(Y \leq y) &= P\left(F^{-1}(U) \leq y\right) \\ &= P\left(F\left(F^{-1}(U)\right) \leq F(y)\right) \\ &= P(U \leq F(y)) \\ &= F(y) \end{aligned}$$

因此 Y 服从 CDF 为 $F(\cdot)$ 的分布，记为 $Y \sim F$. 这就是 **CDF 逆变换的基本想法**。然而在实践中，它还面临很多问题。比如对离散分布

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

它的 CDF $F(x)$ 不连续且不可逆; 有时我们需要抽样的分布既有离散又有连续的部分，设 F_d 是一个离散分布的 CDF, F_c 是一个连续分布的 CDF, $0 < \lambda < 1$, 则 $\lambda F_d + (1-\lambda)F_c$ 也是一个 CDF，如图1所示，它在一些点的逆函数也无法定义。

上述问题可以通过为 CDF 定义如下的**广义逆**得到解决。

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

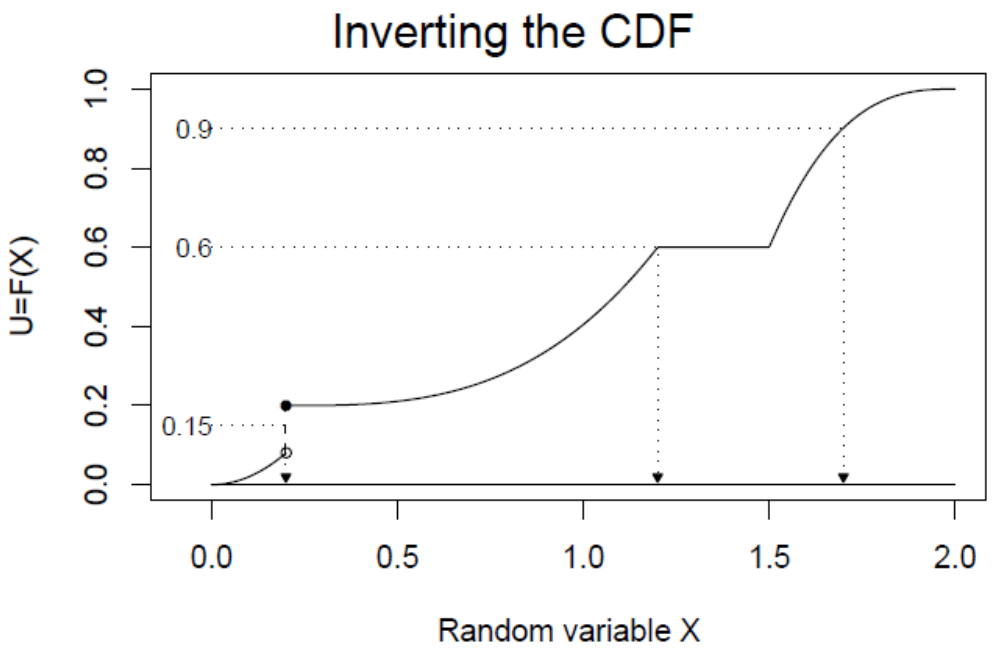

Figure 1: 随机变量 X 在 [0,2] 上取值，但它在区间 [1.2,1.5] 取值的概率为 0，图中的实线是 X 的 CDF, 它在 $x = 0.2$ 处有一个跳跃。

由于 $F(+\infty) = 1$，(1)中的集合总是非空的，因此总可以找到下确界。图1展示了 F 的广义逆在几个点的值：$F^{-1}(0.15) = 0.2$, $F^{-1}(0.6) = 1.2$, $F^{-1}(0.9) = 1.7$.

CDF 逆变换: 设 F 是一个 CDF, F^{-1} 是由(1)定义的逆函数。如果随机变量 $U \sim \boldsymbol{U}[0,1]$, 令 $X = F^{-1}(U)$, 则变换得到的随机变量 $X \sim F$.

Remarks

- 如果 $U \sim \boldsymbol{U}[0,1]$, 那么 $1 - U \sim \boldsymbol{U}[0,1]$, 因此 $F^{-1}(1-U) \sim F$. 有时候 $F^{-1}(1-U)$ 具有更简单的形式。
- 在 CDF 逆变换中，我们是对服从 $\boldsymbol{U}[0,1]$ 的随机变量做逆变换，上述想法可以进一步推广。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

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

  服从分布 G, 因为 $F(X) \sim \boldsymbol{U}[0,1]$. 函数 $G^{-1}(F(\cdot))$ 也被称为 **QQ 变换**，因为它将分布 F 的分位数 (quantile) 转换为分布 G 下相应的分位数。即如果 x 是 F 的 0.1-quantile ($P(X \le x) = F(x) = 0.1$), 则 $y = G^{-1}(F(x))$ 是 G 的 0.1-quantile ($P(Y \le y) = G(y) = 0.1$).
- 有时 $G^{-1}(F(\cdot))$ 的形式比 G^{-1} 或 F 都简单，我们可以直接将 $X \sim F$ 转化为 $Y = G^{-1}(F(X)) \sim G$，而不需要知道 F 或 G^{-1} 的具体形式。比如知道如何从 $N(0,1)$ 抽样

后，如果想得到 $N(\mu,\sigma^2)$ 的样本，可以做变换 $Y=\mu+\sigma Z$, $Z\sim N(0,1)$，这其实是一个 QQ 变换(2).

2.1.1 CDF 逆变换举例

很多重要的一元分布可以用逆变换的方法进行抽样，这里列举一些有用的例子。

- **指数分布 (Exponential distribution)**. 标准的指数分布 Exp(1) 的 PDF 是

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

它的 CDF 是

$$F(x)=P(X\le x)=\int_0^x f(t)dt=1-e^{-x},\ x>0.$$

CDF 的逆是

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

因此先抽取 $U\sim \boldsymbol{U}(0,1)$, 再令 $X=-\log(1-U)$ 可以产生服从 Exp(1) 的样本。或者可直接令 $X=-\log(U)$ ~Exp(1), 因为 $1-U\sim \boldsymbol{U}(0,1)$, $F^{-1}(1-U)=-\log(U)$.

一般的指数分布有一个 rate parameter $\lambda>0$, 记为 Exp(λ), 有时人们会用 scale parameter $\theta=1/\lambda$ 取代 λ 描述指数分布。$Y\sim$ Exp(λ) 的期望是 $E(Y)=1/\lambda$, PDF 是

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

如果 $X\sim$ Exp(1), 则 $X/\lambda\sim$ Exp(λ). 因此可以通过变换 $Y=-\log(U)/\lambda$ 获得 Exp(λ) 的样本。

  - 指数分布常用于描述一段时间的分布，但它的一个重要特性是**无记忆性** (memoryless). 如果 $X\sim$ Exp(λ), 则

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

它与 x 无关。因此指数分布不适合描述一个非耐用品的生命周期，比如灯泡的寿命等。后面介绍的 Weibull 分布更适合描述这种情况。

- **Bernoulli 分布**。如果 $X\sim$ Bern(p), 则 $P(X=1)=p$, $P(X=0)=1-p$. 从 Bernoulli 分布抽样可以利用变换 $X=\mathbf{1}(1-U\le p)$ 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** (heavy tails) 的特性。Cauchy 分布的 PDF 是

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

密度函数 $f(x)$ 在 $x \to \pm\infty$ 时下降地很慢以至于 $E(|X|) = \infty$. Cauchy 的 CDF 如下

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

对 CDF 求逆，我们可以利用如下变换从 Cauchy 分布抽样，

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

从几何角度看，Cauchy 变量是一个在 $(-\pi/2, \pi/2)$ 上均匀分布的随机角的正切 (tangent). Cauchy 分布可用于描述看似像正态分布，又存在一些极端大或小的观察值的情形。

- **离散均匀分布**。对于在 $\{0, 1, \ldots, k-1\}$ 上均匀分布的离散随机变量，可以令

$$X = \lfloor kU \rfloor,\ U \sim \boldsymbol{U}(0, 1)$$

其中 $\lfloor kU \rfloor$ 表示 $\leq kU$ 的最大整数。如果我们想从 $\boldsymbol{U}\{1, 2, \ldots, k\}$ 上抽样，可令 $X = \lceil kU \rceil$, 其中 $\lceil kU \rceil$ 表示 $\geq kU$ 的最小整数。

- **Poisson** 分布。如果 $X \sim \text{Po}(\lambda)$, 则

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

且 $E(X) = \lambda$. Algorithm 1 是 Devroye (1986) 基于 CDF 逆变换提出的从 Poisson 分布抽样的方法。算法中条件 $U > q$ 会被检验 $X + 1$ 次，因此算法平均需要的迭代步数是 $E(X + 1) = \lambda + 1$. 显然对 λ 较大的情形，使用该方法抽样会很慢。

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

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

- **正态分布**。用 Φ 表示 $N(0,1)$ 的 CDF，如果 $U \sim \boldsymbol{U}(0,1)$, 那么 $Z = \Phi^{-1}(U) \sim N(0,1)$. 但是 Φ 和 Φ^{-1} 都没有解析形式，这使得用 CDF 逆变换从正态分布抽样变得十分困难。不过通过对 Φ^{-1} 做精确的数值近似，使用逆变换抽样仍然是可行的，比如 Wichura (1988) 提出的算法 AS241, 其中使用的近似函数的精度非常高

$$\frac{|\hat{\Phi}_W^{-1}(u) - \Phi^{-1}(u)|}{|\Phi^{-1}(u)|} < 10^{-15}, \quad \min(u, 1-u) > 10^{-316}.$$

后面会介绍另一种更简便的方法——Box-Muller 变换。

- **Weibull 分布**。Weibull 分布是对指数分布的推广，它的 PDF

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

有两个参数 $\sigma > 0$, $k > 0$. $k = 1$ 时, Weibull 分布退化为指数分布 Exp$(1/\sigma)$. Weibull 分布的 CDF 是

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

因此对 $U \sim \boldsymbol{U}(0,1)$ 做变换 $X = \sigma\left(-\log(1-U)\right)^{1/k}$ 可实现从 Weibull 分布抽样。

之前我们讨论过指数分布的无记忆性，为了探讨 Weibull 分布的基本特性，引入一个新的概念——**hazard function**.

Definition 2.2 (Hazard function)**.** 对一个随机变量 $X > 0$, 它的 hazard function $h(x)$ 定义为

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

如果用 X 表示一个灯泡的生命周期，hazard function $h(x)$ 表示给定灯泡在 x 时刻正常工作，它瞬间出现故障的概率（密度），也称瞬时失败概率（instantaneous probability of failure）。

练习: 求指数分布 Exp(λ) 和 Weibull 分布 Weibull(k, σ) 的 hazard function. (一个有用的近似: $e^x \approx 1 + x$, $x \to 0$)

- 指数分布 Exp(λ) 的 hazard function

$$h(x) = \lim_{t \to 0^+} \frac{1}{t}(1 - e^{-\lambda t}) = \lim_{t \to 0^+} \frac{\lambda t}{t} = \lambda$$

可以看到指数分布的瞬时失败概率是常数，与当前时刻 x 无关。

– Weibull 分布的 hazard function

$$\begin{aligned}
h(x) &= \lim_{t\to 0^+}\frac{1}{t}\cdot\frac{P(x\le X\le x+t)}{P(X\ge x)}\\
&= \lim_{t\to 0^+}\frac{1}{t}\cdot\frac{F(x+t)-F(x)}{1-F(x)}\\
&= \lim_{t\to 0^+}\frac{1}{t}\cdot\frac{-\exp\left[-\left(\frac{x+t}{\sigma}\right)^k\right]+\exp\left[-\left(\frac{x}{\sigma}\right)^k\right]}{\exp\left[-\left(\frac{x}{\sigma}\right)^k\right]}\\
&= \lim_{t\to 0^+}\frac{1}{t}\cdot\left\{-\exp\left[-\left(\frac{x+t}{\sigma}\right)^k+\left(\frac{x}{\sigma}\right)^k\right]+1\right\}\\
&= \lim_{t\to 0^+}\frac{1}{t}\left[\left(\frac{x+t}{\sigma}\right)^k-\left(\frac{x}{\sigma}\right)^k\right]\\
&= \frac{d}{dx}\left(\frac{x}{\sigma}\right)^k\\
&= k\cdot\frac{x^{k-1}}{\sigma^k}
\end{aligned}$$

可以看到

* $k<1$ 时, 瞬时失败概率 $h(x)$ 随时间 x 递减。这种事件有可能发生，比如婴儿在刚出生时死亡概率很高，但随时间推移死亡概率在下降。
* $k=1$ 时，Weibull 退化为指数分布，瞬时失败概率是常数。
* $k>1$ 时，瞬时失败概率 $h(x)$ 随时间 x 递增。这种事件很常见，比如任何会老化的产品的生命周期。

- **双指数分布 (Double exponential distribution)**。标准的双指数分布的 PDF 是

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

练习：求标准双指数分布的 CDF 及其逆函数。

$$F(x)=\begin{cases}\frac{1}{2}e^x, & x<0\\ 1-\frac{1}{2}e^{-x}, & x\ge 0\end{cases}$$

$$F^{-1}(u)=\begin{cases}\log(2u), & 0<u\le 1/2\\ -\log(2(1-u)), & 1/2<u<1\end{cases}$$

- **Gumbel 分布**。Gumbel 分布常用来描述一种极值的分布。如果 $Y_1, Y_2, \ldots, Y_n$ 独立同分布，当 n 很大时，$X = \max(Y_1, Y_2, \ldots, Y_n)$ 近似服从 Gumbel 分布。标准 Gumbel 分布的 CDF 为

$$F(x) = \exp(-e^{-x}),\ x \in \mathbb{R}$$

利用变换 $X = -\log(-\log(U))$, $U \sim \boldsymbol{U}(0,1)$ 可以得到服从标准 Gumbel 分布的样本。一般的 Gumbel 分布有两个参数 $\mu \in \mathbb{R}$ 和 $\sigma > 0$, 它的 CDF 的形式为

$$F(x) = \exp(-e^{-(x-\mu)/\sigma}),\ x \in \mathbb{R}.$$

虽然看起来参数 μ, σ 是对一个标准 Gumbel 分布做位移和放缩，它们并不代表 Gumbel 分布的期望和标准差。一般的 Gumbel 分布对应的逆变换形式为 $X = \mu - \sigma\log(-\log(U))$.

- **Triangular and power densities**.

练习：用逆变换法从 triangular 分布抽样。Triangular density

$$f(x) = 2x,\ 0 < x < 1$$

CDF $F(x) = x^2$, 因此 $F^{-1}(U) = \sqrt{U}$, $U \sim \boldsymbol{U}(0,1)$.

练习：用逆变换法从 power 分布抽样。Power density ($\alpha > 0$)

$$f(x) = \alpha x^{\alpha-1},\ 0 < x < 1.$$

CDF $F(x) = x^\alpha$, 因此 $X = U^{1/\alpha}$, $U \sim \boldsymbol{U}(0,1)$.

Triangular 和 power 分布都是 Beta 分布的特例，后面还会介绍其它从 Beta 分布抽样的方法。

- **截断分布抽样**。有时我们会面临从一个分布的特定区间上抽样，比如只想得到 $N(0,1)$ 大于 5 的样本。我们当然可以先从 $N(0,1)$ 抽取大量样本，然后只保留在 $(5, +\infty)$ 上的样本，但这种抽样方法往往很低效。下面介绍较高效的利用逆变换从截断分布抽样的方法。假设我们想从一个连续分布的 (a,b) 区间上抽样，F 是这个分布的 CDF, f 是分布的 PDF 且 $f(x) > 0$, $\forall x \in (a,b)$. 如果随机变量 $Y \sim F$, X 服从 F 在 (a,b) 上的截断分布，，则 X 的 CDF 为

$$\begin{aligned} G(x) &= P(X \le x), \quad a < x < b \\ &= P(Y \le x \mid a < Y < b) \\ &= \frac{P(a < Y \le x)}{P(a < Y < b)} \\ &= \frac{F(x) - F(a)}{F(b) - F(a)} \end{aligned}$$

为使 $G(x)$ 是一个有效的 CDF，规定 $G(x)=0,\ x\le a$; $G(x)=1,\ x\ge b$. 因此可通过以下变换从 F 在 (a,b) 上的截断分布抽样

$$X=F^{-1}\left(F(a)+(F(b)-F(a))U\right),\ U\sim\boldsymbol{U}(0,1).$$

2.2 离散分布的逆变换法

由于离散分布的 CDF 不连续，使用逆变换方法时，可以利用一些特殊技巧提高抽样的效率。我们在 Section 2.1.1介绍了如何从离散均匀分布中抽样，对于一般的定义在有限个点上的离散分布

$$P(X=k)=p_k>0,\ k=1,\dots,N$$

定义累积概率

$$P_k=\sum_{i=1}^{k}p_i,\ k=1,\dots,N$$

令

$$F^{-1}(u)=k,\quad P_{k-1}<u\le P_k$$

其中 $P_0=0$. 对于 $U\sim\boldsymbol{U}(0,1)$, 如果我们按上式从 $k=1$ 逐个搜索，抽取 X 平均需要比较 $E(X)$ 次，计算量是 $O(N)$. 使用二分法搜索的计算量是 $O(\log(N))$, Algorithm 2展示了在逆变换中使用二分法搜索的算法。

Algorithm 2 Bisection-based inversion of a CDF on $\{1,\dots,N\}$

```
Input u, N, P_{0:N}.
Initialize L = 0, R = N.
while L < R − 1 do
    k = ⌊(L + R)/2⌋
    if u > P_k then
        L = k
    else
        R = k
return R
```

R 中用 `sample(x, size=n, prob=p, replace=TRUE)` 可以得到有限个点的离散分布的 n 个独立样本，其中向量 `x` 是离散分布可取值的集合，向量 `p` 是每个值对应的概率。

当离散分布的 support 有无穷个点，上述二分法就失效了。这种情况下，如果离散分布的 CDF 有较简单的解析形式，我们可以先从一个连续分布中抽样，再对样本做一些截断处理变成整数。以**几何分布**的抽样为例介绍具体做法。

Definition 2.3 (几何分布 (Geometric distribution))**.** 如果每次试验成功的概率是 θ, 不断独立地做试验直到取得一次成功, 在第一次成功之前所经历的失败次数 X 服从几何分布:

$$P(X=k)=\theta(1-\theta)^k,\ k=0,1,\ldots.$$

考察几何分布的 CDF

$$\begin{aligned}G(n)=P(X\le n)=\sum_{k=0}^{n}P(X=k)&=\theta\sum_{k=0}^{n}(1-\theta)^k=\theta\frac{1-(1-\theta)^{n+1}}{1-(1-\theta)}\\&=1-(1-\theta)^{n+1}=1-\exp\left[(n+1)\log(1-\theta)\right]\end{aligned}$$

它的形式与指数分布 $Y\sim\mathrm{Exp}(1)$ 的 CDF

$$F(y)=1-e^{-y}$$

有些相似。因此如果 $Y\sim\mathrm{Exp}(1)$, 利用 QQ 变换 $X=G^{-1}\left(F(Y)\right)=\lceil -1-Y/\log(1-\theta)\rceil$ 可以得到几何分布的样本。当 $U\sim\boldsymbol{U}(0,1)$, $-\log(U)\sim\mathrm{Exp}(1)$, 因此 $X=\lceil -1+\log(U)/\log(1-\theta)\rceil$ 服从几何分布。

2.3 其它变换

逆变换法的思路很简单, 但有时会面临数值计算方面的问题。有时利用分布之间的特殊关系可以构造更简洁的变换进行抽样。下面我们列举一些其它重要的变换方法。

2.3.1 单调变换

假设随机变量 X 在 $\mathbb{R}$ 上的 PDF 为 f_X. $\tau(\cdot)$ 是一个可逆的增函数。令随机变量 $Y=\tau(X)$, 则 Y 的 CDF 为

$$F_Y(y)=P(Y\le y)=P(\tau(X)\le y)=P(X\le\tau^{-1}(y))=F_X\left(\tau^{-1}(y)\right).\tag{3}$$

Y 的 PDF 为

$$f_Y(y)=\frac{d}{dy}F_Y(y)=\frac{d}{dy}F_X\left(\tau^{-1}(y)\right)=f_X\left(\tau^{-1}(y)\right)\frac{d}{dy}\tau^{-1}(y).\tag{4}$$

比如对数正态 (log-normal) 分布就是通过单调变换定义的。如果 $X\sim N(\mu,\sigma^2)$, 则 $Y=\exp(X)$ 服从对数正态分布, 它的 PDF 为

$$f_Y(y)=\frac{1}{y\sqrt{2\pi}\sigma}\exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right),\ y>0$$

假设随机变量 X 的期望是 0，标准差为 1，我们可以平移 X 并对它进行缩放，$Y = \mu + \sigma X$，使得 Y 的期望是 μ, 标准差为 σ. 此时 Y 的 PDF 为

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{y-\mu}{\sigma}\right). \tag{5}$$

由此可以证明，当 $X \sim N(0,1)$, $Y = \mu + \sigma X \sim N(\mu, \sigma^2)$. 根据(3), 使用单调变换 $Y = \tau(X) \sim F_Y$ 对 Y 抽样等价于使用 QQ 变换 $F_Y^{-1}(F_X(\cdot))$.

2.3.2 Box-Muller 变换

著名的 Box-Muller 变换方法可以用两个独立的 $U(0,1)$ 变量产生两个独立的 $N(0,1)$ 变量:

$$\begin{aligned} Z_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2) \\ Z_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2) \end{aligned} \tag{6}$$

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

Box-Muller 的原理简单解释如下。如果将服从二元标准正态分布 $N(\mathbf{0}, I_2)$ 的随机向量 (Z_1, Z_2) 用极坐标表示，它对应的角度 $\theta \sim \boldsymbol{U}[0, 2\pi)$, 且独立于半径 R, 因此可以用 $\theta = 2\pi U_2$ 产生极坐标下的角度。而它的半径 $R^2 = Z_1^2 + Z_2^2 \sim \chi^2_{(2)}$, 等价于 Exp(1/2) 或 2×Exp(1)，所以可以用 $R = \sqrt{-2\log(U_1)}$ 产生极坐标下的半径，这里使用了逆变换法从指数分布抽样。最后再通过极坐标变换

$$\begin{aligned} Z_1 &= R\cos(\theta) \\ Z_2 &= R\sin(\theta) \end{aligned} \tag{7}$$

映射到正常的坐标系得到两个独立的 $N(0,1)$ 变量。

有关正态分布上述性质的证明如下。对于两个独立的 $N(0,1)$ 随机变量 Z_1, Z_2, 它们的 joint PDF 为

$$f_{Z_1,Z_2}(z_1, z_2) = f_{Z_1}(z_1) \cdot f_{Z_2}(z_2) = \frac{e^{-z_1^2/2}}{\sqrt{2\pi}} \frac{e^{-z_2^2/2}}{\sqrt{2\pi}} = \frac{1}{2\pi} e^{-(z_1^2+z_2^2)/2}.$$

根据(4)在多元变量的推广，极坐标变换(7)对应的 (R, θ) 的联合 PDF 为

$$f_{R,\theta}(r,\theta) = f_{z_1,z_2}\left(r\cos(\theta), r\sin(\theta)\right) |\det(J)|$$

其中 J 是 Jacobian matrix

$$J = \frac{\partial(z_1, z_2)}{\partial(r,\theta)} = \begin{pmatrix} \dfrac{\partial z_1}{\partial r} & \dfrac{\partial z_1}{\partial \theta} \\ \dfrac{\partial z_2}{\partial r} & \dfrac{\partial z_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix} \tag{8}$$

因此 $|\det(J)| = r$, 注意到 $z_1^2 + z_2^2 = r^2$, (R, θ) 的联合 PDF 的具体形式为

$$f_{R,\theta}(r,\theta) = \frac{1}{2\pi} \cdot re^{-r^2/2},\ 0 \le \theta < 2\pi,\ r > 0 \tag{9}$$

$$= f_\theta(\theta) \cdot f_R(r) \tag{10}$$

由于联合 PDF $f_{R,\theta}(r,\theta)$ 可以分解为 θ 的函数 $f_\theta(\theta) = 1/2\pi$ 与 r 的函数 $f_R(r) = r\exp(-r^2/2)$ 的乘积，因此 R 和 θ 是独立的，且 $\theta \sim \boldsymbol{U}[0, 2\pi)$.

下面我们来分析 R 的分布。由于 $R^2 = Z_1^2 + Z_2^2$, 猜测 $R^2 \sim \chi^2_{(2)}$. 严格证明如下。令 $S = \tau(R) = R^2$, $R > 0$. 因此 τ 的逆变换 $\tau^{-1}(S) = \sqrt{S}$ 且

$$\frac{d}{dS}\tau^{-1}(S) = \frac{1}{2}S^{-1/2}.$$

根据(4), S 的 PDF 为

$$f_S(s) = f_R(\sqrt{s})\frac{1}{2}s^{-1/2} = s^{1/2}e^{-s/2}\frac{1}{2}s^{-1/2} = \frac{1}{2}e^{-s/2},$$

因此 $S = R^2 \sim \chi^2_{(2)}$ 或 Exp(1/2) 或 2×Exp(1).

幸运的是 θ 既没有出现在联合 PDF (10) 也没有出现在 Jacobian matrix (8)中。否则寻找 θ 的分布需要将 θ 表示成关于 Z_1 和 Z_2 的 arctangent 函数，还要对落在 arctangent 函数的不同区间进行分类讨论。

Box-Muller 方法因为操作简单所以很流行。实践中我们可能不需要用到 Z_2，可以只输出 Z_1. 但是 Box-Muller 方法不是最快从 $N(0,1)$ 抽样的方法，因为计算 cos, sin, log 和 $\sqrt{}$ 提高了计算成本。

2.3.3 Maxima, minima and order statistics

如果 $Y = \max(X_1, \ldots, X_r)$, 其中 X_i's 独立同分布且 CDF 是 F. 则

$$P(Y \le y) = P\left(\max_{1\le i\le r} X_i \le y\right) = \prod_{i=1}^{r} P(X_i \le y) = (F(y))^r$$

因此如果我们想从 CDF 是 $G = F^r$ 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即可。例如

- 令 F 表示 $\boldsymbol{U}(0,1)$ 的 CDF, $r = 2$, 即 $Y = \max(U_1, U_2)$, 则 $G(y) = F(y)^2 = y^2$, $0 < y < 1$. Y 的 PDF 为

  $$g(y) = 2y,\ 0 < y < 1.$$

  是我们之前介绍过的 triangular density. 显然用 $\max(U_1, U_2)$ 从三角分布抽样比逆变换法 $\sqrt{U_1}$ 快。

另一方面，如果 $Y = \min(X_1, \ldots, X_r)$, 则 Y 的 CDF 为

$$G(y) = P(Y \le y) = P\left(\min_{1\le i\le r} X_i \le y\right) = 1 - P\left(\min_{1\le i\le r} X_i > y\right)$$
$$= 1 - \prod_{i=1}^{r} P(X_i > y) = 1 - (1 - F(y))^r$$

我们来看用 minima 抽样的一个例子。

- 考虑 $Y = \min(U_1, U_2)$ 的分布，其中 $U_k \sim \boldsymbol{U}(0,1)$. 则 Y 的 CDF 和 PDF 为

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

  这也是一种 triangular density. 如果用逆变换法从该分布中抽样，需计算 $Y = 1 - \sqrt{1-U}$, $U \sim \boldsymbol{U}(0,1)$, 比使用 minima 抽样效率低。

最大最小统计量都是 **order statistics** 的特例。

Definition 2.4 (Order statistics)**.** 对于 n 个独立同分布的随机变量, $X_1, \ldots, X_n$. 它们的 order statistics 是将这 n 个变量的取值按从小到大的顺序排列，记为

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

我们来研究一下 $\boldsymbol{U}(0,1)$ 的 order statistics. 对于 n 个独立的 $\boldsymbol{U}(0,1)$ 随机变量 $U_1, \ldots, U_n$，令 $U_{(r)}$ 表示 r-th order statistic. 如果 $x \le U_{(r)} < x + \Delta$, 且 Δ 非常小，使得有超过一个 U_i 落在区间 $[x, x+\Delta)$ 的概率可以忽略不记，则区间 $(0, x)$,$[x, x+\Delta)$, $(x+\Delta, 1)$ 分别包含 $r-1$, 1, 以及 $n-r$ 个 $\{U_1, \ldots, U_n\}$ 中的变量。因此

$$P(x \le U_{(r)} < x + \Delta) = P\left\{(r-1)\text{个}U_i's \in (0,x),\ \text{且}1\text{个}U_i \in [x, x+\Delta),\ \text{且剩下的}(n-r)\text{个}U_i's \in [x+\Delta, 1)\right\}$$
$$= \binom{n}{r-1} x^{r-1} \cdot (n-r+1) \cdot \Delta \cdot (1-x-\Delta)^{n-r}$$
$$= \frac{n!}{(n-r)!(r-1)!} x^{r-1}(1-x-\Delta)^{n-r}\Delta$$

则 $U_{(r)}$ 的 PDF 为

$$f_{U_{(r)}}(x) = \lim_{\Delta\to 0} \frac{P(x \le U_{(r)} < x + \Delta)}{\Delta} = \frac{n! x^{r-1}(1-x)^{n-r}}{(n-r)!(r-1)!},\ 0 < x < 1.$$

这其实是 Beta$(r, n-r+1)$ 分布。

Remarks

1. 如果想从 Beta(α,β) 分布抽样，且参数 α, β 都是正整数，可以先产生 $n=(\alpha+\beta-1)$ 个 $\boldsymbol{U}(0,1)$ 变量，然后只保留 $U_{(\alpha)}$. 当然这种方法在 $\alpha+\beta$ 很大的情况下可能比较慢。

2. R 中用 `sort(x)` 可以将 `x` 中的元素从小到大排列；`order(x)` 可以获得把 `x` 的元素从小到大排列的下标，比如 `order(c(3,1,7,4))` 结果为 `(2,1,4,3)`.

3. 对于 n 个独立同分布的 $Y_i\sim F$, F 是 CDF 且 F^{-1} 有解析形式。当 n 很大的时候，如果想得到 $Y_{(r)}$ 的样本，一个快速的方法是：先抽取 $X\sim\text{Beta}(r,n-r+1)$, 则 $Y_{(r)}=F^{-1}(X)$. 因为 $F(Y_i)\overset{ind}{\sim}\boldsymbol{U}(0,1)$, $i=1,\ldots,n$, CDF F 是增函数，单调变换不改变统计量的大小关系，因此 $F(Y_{(r)})\sim U_{(r)}\sim\text{Beta}(r,n-r+1)$.

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?

2.3.4 Sums

有时我们要抽样的 Y 的分布可以写成 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。此时我们可以先从 X_i 的分布中独立抽取 n 个样本，再把它们加在一起就得到了一个 Y 的样本。例如

- **二项分布**。如果 $Y\sim\text{Bin}(n,p)$, 则

$$P(Y=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,\ldots,n.$$

  二项分布描述的是 n 次独立随机试验中成功的次数 Y，每次试验成功的概率都是 p. 因此可以将 Y 写成 n 个 Bernoulli 变量的和

$$Y=\sum_{i=1}^{n}X_i,\quad X_i\overset{iid}{\sim}\text{Bern}(p),i=1,2,\ldots,n.$$

- χ^2 **分布**。如果 $X_i\overset{iid}{\sim}\chi^2_{(\alpha)}$, $i=1,2,\ldots,n$. 则

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

当 $\alpha = 1$, 可以令 $X_i = Z_i^2$, $Z_i \sim N(0,1)$. 当 $\alpha = 2$, 可以从期望为 2 的指数分布中抽取 $X_i \sim \text{Exp}(1/2)$. χ^2 分布是后面将要介绍的 Gamma 分布的一个特例。

- **Noncentral χ^2 分布**。Noncentral χ^2 分布有两个参数, 自由度 n 和参数 $\lambda \geq 0$, 记为 $\chi'^2_{(n)}(\lambda)$. 它可以按如下方式生成:

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

Noncentral χ^2 分布可以用来构造 noncentral F 分布

$$F = \frac{Y_1/n}{Y_2/d}, \quad Y_1 \sim \chi'^2_n(\lambda), Y_2 \sim \chi^2_{(d)}.$$

其中 Y_1 与 Y_2 独立。上述 noncentral F 分布 $F'_{n,d}(\lambda)$ 经常用于计算假设检验的 power.

2.4 Bootstrap

在统计推断中我们经常面临的问题是，假设观察值独立同分布 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} F$，需要计算某个统计量 $\hat{T} = T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ 的分布。当 $\hat{T}$ 的方差很难计算，或者不知道 $\hat{T}$ 具体服从什么分布时，可以用 Bootstrap 方法得到 $\hat{T}$ 的近似分布。

Bootstrap 是一种基于观察值的**经验分布** (empirical distribution) 进行重抽样的方法。

Definition 2.5 (经验分布)**.** 观察值 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ 的经验分布的 CDF 定义为

$$\hat{F}_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\boldsymbol{x}_i \leq \boldsymbol{x}).$$

即 $\hat{F}_n$ 相当于一个均匀的离散分布。根据观察值的经验分布进行重抽样时，每个样本 $\boldsymbol{x}_i$ 被抽到的概率都是 $1/n$。比如可以用如下方式得到一个重抽样的样本:

$$\text{抽取 } k \sim \boldsymbol{U}\{1, \ldots, n\}, \text{ 然后令 } \boldsymbol{X}^\star = \boldsymbol{x}_k.$$

使用 Bootstrap 时，我们用上述重抽样的方法生成 B 组数据，每组数据 b 仍然包含 n 个样本，即

$$\boldsymbol{X}_i^{\star b} \overset{iid}{\sim} \hat{F}_n, \quad i = 1, \ldots, n; \; b = 1, \ldots, B.$$

注意独立地从 $\hat{F}_n$ 中抽样是一个有放回的抽样过程，因此在新产生的数据 b 中，样本的值可能有重复。对于每个重抽样产生的数据 b, 我们可以计算 $\hat{T}$ 的一个值

$$\hat{T}^{\star b} = T(\boldsymbol{X}_1^{\star b}, \ldots, \boldsymbol{X}_n^{\star b}), \quad b = 1, \ldots, B$$

通常 $\{\hat{T}^{\star b} : b = 1, \ldots, B\}$ 可以很好地近似 $\hat{T}$ 服从的分布。因此 $\mathrm{Var}(\hat{T})$ 可以如下估计:

$$\mathrm{Var}(\hat{T}) \approx \frac{1}{B-1} \sum_{b=1}^{B} (\hat{T}^{\star b} - \bar{T})^2$$

其中 $\bar{T} = 1/B \sum_{b=1}^{B} \hat{T}^{\star b}$. $\hat{T}$ 的 95% 置信区间近似为 $\{\hat{T}^{\star b} : b = 1, \ldots, B\}$ 的 2.5% 分位数与 97.5% 分位数所夹的区间。

2.5 Acceptance-Rejection

当逆变换法失效时，即无法找到一个变换将一个 $\boldsymbol{U}(0,1)$ 变量转化服从目标分布 F 的随机变量，此时可以尝试另一种抽样方法：先从另一个容易抽样的分布 G 中抽样，然后按照某种规则去掉一些样本，最后保留下来的的样本即为服从目标分布 F 的样本。人们把这种抽样方法称为 rejection sampling 或 acceptance-rejection (A-R) sampling.

假设 F 和 G 是两个连续分布的 CDF, 它们的 PDF 分别是 f 和 g. 使用 A-R 方法需满足以下 3 个条件:

1. 可以从分布 g 中抽样
2. 函数 f/g 可计算
3. 存在常数 $c > 0$ 使得

$$f(x) \le c \cdot g(x), \ \forall x \in \mathbb{R}$$

A-R 的具体做法如 Algorithm 3 所示。首先从分布 g 中抽样 $Y \sim g$. 假设 $Y = y$ (显然 $g(y) > 0$), 样本 y 被接受的概率是

$$A(y) = \frac{f(y)}{c \cdot g(y)}.$$

如果 y 没有被接受，继续从 g 中抽样直到有一个样本被接受，保留该样本作为 f 的一个样本。

Algorithm 3 Acceptance-rejection sampling

given c with $f(x) \le c \cdot g(x), \forall x \in \mathbb{R}$

repeat

$Y \sim g$

$U \sim \boldsymbol{U}(0,1)$

until $U \le f(Y)/(c \cdot g(Y))$

$X = Y$

return X

Remarks

1. 条件 $f(x) \le c \cdot g(x)$ 保证了每个候选样本 y 被接受的概率 $A(y) \le 1$. 如果函数 $f(x)/g(x)$ 在某些点无界，则不能在 A-R 中使用分布 g.

2. A-R 方法的基本想法是：如果候选样本 y 来自分布 g 且以概率 $A(y)$ 被接受，则它出现的概率（密度）$\propto g(y)A(y)$. 因此如果令 $A(y) \propto f(y)/g(y)$ 则有 $g(y)A(y) \propto f(y)$. 严格证明见定理2. Markov chain Monte Carlo 方法 Metropolis-Hastings 的接受规则（acceptance rule）也使用了该想法。

Theorem 2. *如果两个 PDF $f(x)$ 和 $g(x)$ 满足*

$$f(x) \le c \cdot g(x), \ \forall x \in \mathbb{R}$$

则 Algorithm 3 产生的 $X \sim f$.

Proof. Y 被接受的概率是

$$P(Y \text{ is accepted}) = \int_{-\infty}^{+\infty} g(y)A(y)dy = \frac{1}{c}\int_{-\infty}^{+\infty} f(y)dy = \frac{1}{c}.$$

Algorithm 3 产生的 X 的 CDF 为

$$\begin{aligned} P(X \le x) = P\left(Y \le x \mid Y \text{ is accepted}\right) &= \frac{P(Y \le x \text{ and } Y \text{ is accepted})}{P(Y \text{ is accepted})} \\ &= \frac{\int_{-\infty}^{x} g(y)A(y)dy}{1/c} = \frac{1/c\int_{-\infty}^{x} f(y)dy}{1/c} \\ &= \int_{-\infty}^{x} f(y)dy \end{aligned}$$

因此 $X \sim f$. □

Remarks

1. 从上述证明我们看到，在 A-R 方法中，g 分布产生的样本被接受的概率是 $1/c$. 因此为了得到分布 f 的一个样本，需要从分布 g 中抽取的样本数 N 服从几何分布

   $$P(N = k) = \frac{1}{c}\left(1 - \frac{1}{c}\right)^{k-1},$$

   则 $E(N) = c$，即如果每个来自 g 的候选样本被接受的概率是 $1/c$, 则平均要产生 c 个候选样本才能有一个被接受。当然总有 $c \ge 1$，因为

   $$1 = \int_{-\infty}^{+\infty} f(x)dx \le \int_{-\infty}^{+\infty} cg(x)dx = c.$$

2. 为了减少样本损失，提高算法效率，我们希望 c 越小越好。一般可将 c 选为

$$c = \sup_x \frac{f(x)}{g(x)}$$

这里的 supremum 只需考察 f 的 support, 因为条件 $f(x) \le c \cdot g(x), \forall x$ 要求 g 的 support 必须覆盖 f 的 support.

3. 定理2 可以推广到多维分布，只需在证明中将区间 $(-\infty, x]$ 替换为高维空间的长方体形式 $\prod_{j=1}^{d}(-\infty, x_j] \subset \mathbb{R}^d$ 即可。

- **从 Cauch 分布生成正态分布**。$N(0,1)$ 的 PDF 为 $f(x) = 1/\sqrt{2\pi} \cdot \exp(-x^2/2)$. Cauchy 分布的 PDF 为 $g(x) = \pi^{-1}(1+x^2)^{-1}$. 寻找

$$c = \sup_x \sqrt{\frac{\pi}{2}}(1+x^2)\exp(-x^2/2)$$

易证等式右边的方程在 $x = \pm 1$ 取到最大值 (作业), 因此选取 $c \approx 1.52$. 我们可以先从 Cauchy 分布抽样 $Y \sim g$ (比如使用 CDF 逆变换方法) , 然后以概率 $f(y)/(cg(y))$ 接受样本 $Y = y$. 此时 Cauchy 分布的样本被接受的概率为 $1/c \approx 0.658$.

2.5.1 A-R 的几何解释

从几何角度理解 A-R 方法需要下面的定理。

Theorem 3. *如果 $Y \sim g$ (g 是 PDF) 且 $Z|Y \sim \boldsymbol{U}(0, cg(Y))$ (常数 $c > 0$)，则 (Y, Z) 的联合分布是区域*

$$S_c(g) = \{(y, z) \mid 0 \le z \le cg(y), y \in \mathbb{R}\} \subset \mathbb{R}^2$$

上的均匀分布。

Proof. 这里假设 g 是 $\mathbb{R}$ 上的连续函数, 更严格的证明见 A. B. Owen (2013, Theorem 4.4). 则 (Y,Z) 的联合 PDF 为

$$f_{Y,Z}(y,z) = f_Y(y) f_{Z|Y}(z \mid y) = g(y)\frac{1}{cg(y)} = \frac{1}{c}$$

是一个常数。可验证

$$\int_{S_c(g)} f_{Y,Z}(y,z)dydz = \int_{S_c(g)} \frac{1}{c}dydz = \frac{1}{c} \cdot \text{Size}\left(S_c(g)\right) = \frac{1}{c}\int_{-\infty}^{+\infty} cg(y)dy = 1$$

因此 (Y,Z) 服从 $S_c(g)$ 上的均匀分布。 □

Theorem 4. *如果 $(X, Z) \sim \boldsymbol{U}(S_M(f))$, 其中*

$$S_M(f) = \{(x, z) \mid 0 \le z \le Mf(x), x \in \mathbb{R}\}$$

$M > 0$, f 是 $\mathbb{R}$ 上的一个 PDF, 则 $X \sim f$.

Proof. 因为 $(X, Z) \sim \boldsymbol{U}(S_M(f))$, 所以

$$P(X \le x) = \frac{\text{Size}\left(S_M(f) \cap (-\infty, x] \times [0, +\infty)\right)}{\text{Size}\left(S_M(f)\right)} = \frac{\int_{-\infty}^{x} Mf(y)dy}{\int_{-\infty}^{+\infty} Mf(y)dy} = \frac{M\int_{-\infty}^{x} f(y)dy}{M} = \int_{-\infty}^{x} f(y)dy$$

因此 $X \sim f$. □

对 A-R 方法的几何解释如图2所示, 为了得到 $X \sim f$ 的样本, 我们可以先获取在曲线 $cg(x)$ 下均匀分布的点 (Y, Z)(根据 Theorem 3, 方法是先抽 $Y \sim g$, 再抽 $Z \sim \boldsymbol{U}(0, cg(Y))$), 然后只保留在曲线 $f(x)$ 下的点 (X, Z). 此时点 (X, Z) 在曲线 $f(x)$ 下也是均匀分布的, 根据 Theorem 4, 边际分布 $X \sim f$.

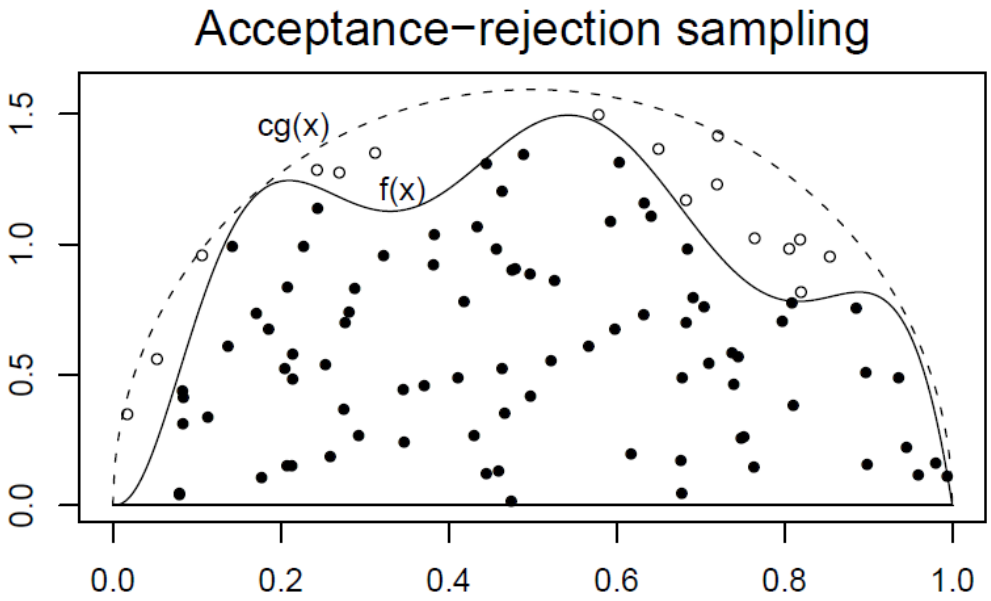

Figure 2: A-R 方法的几何解释。图中实线代表目标分布的 PDF $f(x)$, 虚线代表 $cg(x)$; 图中的点在曲线 $cg(x)$ 下均匀分布, 在曲线 $f(x)$ 下的实心黑点代表被接受的点, 空心点代表被拒绝的点。实心黑点对应的横坐标服从 PDF 为 f 的分布。

Remark

1. 在设计 Algorithm 3 时, 我们使用的是精确的 PDF f 和 g. 有时我们可能只知道 f 或 g 正比于某个函数但不知道确切的归一化常数 (normalizing constant). 比如 f 可能是一个截断分布

— $N(0,1)$ 限制在某个集合 A 上的分布，而 $P(A)$ 未知，此时我们只知道

$$f(x) = \frac{1/\sqrt{2\pi} \cdot \exp(-x^2/2)\mathbf{1}(x \in A)}{\int_A 1/\sqrt{2\pi} \cdot \exp(-x^2/2)dx} \propto \mathbf{1}(x \in A) \cdot \exp(-x^{2/}/2)$$

A-R 方法的几何解释告诉我们，如果 $\tilde{f} \propto f$, $\tilde{g} \propto g$ 是 unnormalized PDF，也可以使用 Algorithm 3 抽样。此时 Algorithm 3 相当于在正比于 $g(x)$ 的曲线下均匀撒点，然后只保留那些正比于 $f(x)$ 曲线下的点，而这些被接受的点的横坐标仍服从 PDF 为 f 的分布。

下面来看一些 A-R 方法的应用：

- **从 $N(0,1)$ 尾部抽样**。我们想得到 $N(0,1)$ 在区间 $[5,+\infty)$ 上的样本，当然可以使用截断分布抽样法，这里再介绍一种用 A-R 方法抽样的思路。目标分布的 PDF 为

$$f(x) \propto \tilde{f}(x) = \exp\left(-x^2/2\right),\ x \geq 5$$

将 g 选为一个经过平移和缩放的指数分布

$$Y \sim 5 + \text{Exp}(1)/5$$

显然可以使用逆变换法对 Y 抽样：$Y = 5 - \log(U)/5$, $U \sim \boldsymbol{U}(0,1)$. 根据随机变量平移缩放的 PDF 关系(5), Y 的 PDF 为

$$g(y) \propto \tilde{g}(y) = \exp\left(-5(y-5)\right),\ y \geq 5$$

然后选取常数 $\tilde{c}$ 使得 $\tilde{f}(x) \leq \tilde{c}\tilde{g}(x)$, $\forall x \geq 5$. 注意到函数

$$\frac{\tilde{f}(x)}{\tilde{g}(x)} = \exp\left(5(x-5) - \frac{x^2}{2}\right)$$

在 $[5,+\infty)$ 上递减，因此选取 $\tilde{c}$ 为

$$\tilde{c} = \sup_{x \geq 5} \frac{\tilde{f}(x)}{\tilde{g}(x)} = \frac{\tilde{f}(5)}{\tilde{g}(5)} = \exp(-5^2/2)$$

即可使用 Algorithm 3 得到目标分布的样本。

练习：计算此时来自 g 的样本总体被接受的概率。

References

李东风 (2016). 统计计算. 高等教育出版社.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.*

