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 $\{x_i, i = 1, 2, \ldots\}$, 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 $x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$, 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 $\{0, 1, \ldots, M\}$ 或 $\{1, 2, \ldots, M\}$ 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 $x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$ 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

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

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 $x_0, x_1, \dots$ 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 $x_{n+k} = x_{m+k}$, $k = 1, 2, \dots$, 即 $x_n, x_{n+1}, x_{n+2}, \dots$ 重复了 $x_m, x_{m+1}, x_{m+2}, \dots$, 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

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

取初值 $x_0 = 7$ 的周期。

练习 2：计算线性同余发生器

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

取初值 $x_0 = 1$ 的周期。

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，序列之间独立性好。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

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

于是当 M 很大时

$$\begin{aligned}
E(R_n)=E(X_n/M)&=\frac{1}{2}-\frac{1}{2M}\approx\frac{1}{2}\\
Var(R_n)=Var(X_n/M)&=\frac{1}{12}-\frac{1}{12M^2}\approx\frac{1}{12}
\end{aligned}$$

可见生成数列的期望和方差很接近均匀分布。

随机数序列还需要有很好的随机性。数列的排列不应该有规律，序列中的两项不应该有相关性。因为序列由确定性公式生成，所以不可能真正独立。至少我们要求序列自相关性弱。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

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

所以应该选 a 值较大 $(a<M)$。

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), 对这些方面有改善。

FSR 按照某种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:

$$\alpha_k=c_p\alpha_{k-p}+c_{p-1}\alpha_{k-p+1}+\cdots+c_1\alpha_{k-1}\ (\mathrm{mod}\ 2)$$

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. $c_p = c_{p-q} = 1 (1 < q < p)$, 则算法变成

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

可以用计算机的异或运算 $\oplus$ 进行快速递推

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

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 $p = 98$)。

References

李东风 (2016). 统计计算. 高等教育出版社.