第二章 随机数的产生与检验

一. 概述

二. 均匀随机数的生成

三. 均匀随机数的检验

用系统仿真方法解决实际问题时，首先要解决的是随机数的产生方法(或称随机变量的抽样方法)。

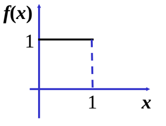
一. 概述

随机数：从某个分布为$F(x)$的总体中随机地抽取的样本观察值。

1. 均匀分布U(0,1)随机数的特殊作用

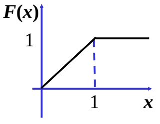
U(0,1)的概率密度函数：

$$f(x)=\begin{cases}1, & 0\le x\le 1\\ 0, & \text{其它}\end{cases}$$



均匀分布U(0,1)的分布函数：

$$F(x)=\begin{cases}0, & x\le 0\\ x, & 0<x\le 1\\ 1, & x>1\end{cases}$$



均值和方差分别为：

$$E(U)=\int_0^1 x\,dx=\frac{1}{2}$$

和

$$Var(U)=\int_0^1 (x-\frac{1}{2})^2 dx=\frac{1}{12}$$

U(0, 1)随机变量与其它形式分布$F(x)$的随机变量的互相转换：

若随机变量U～U(0, 1)，$F(x)$为任一严格单调递增分布函数，F^{-1}为其反函数，令$X=F^{-1}(U)$，则X的分布函数为$F(x)$，即$X\sim F(x)$。

$$\begin{aligned}P\{X\le x\}&=P\{F^{-1}(U)\le x\}\\&=P\{U\le F(x)\}\\&=F(x)\end{aligned}$$

反之，设随机变量X有严格单调递增连续的分布函数$F(x)$，F^{-1}为其反函数，令$U=F(X)$，有 $U\sim$ U(0, 1) 。显然U是随机变量，且由$0\le F(X)\le 1$知$0\le U\le 1$,于是对任一u ($0\le u\le 1$)，

$$\begin{aligned}P\{U\le u\}&=P\{F(X)\le u\}\\&=P\{X\le F^{-1}(u)\}\\&=F[F^{-1}(u)]\\&=u\end{aligned}$$

利用U(0, 1)随机数生成$F(x)$随机数的方法：

(1) 生成相互独立的U(0, 1)随机数 u_i ($i=1, 2,\dots,n$)，

(2) 令$x_i=F^{-1}(u_i)$ ($i=1, 2,\dots,n$)，则 x_i ($i=1, 2,\dots$) 是分布为$F(x)$的随机数，且也相互独立。

类似地，也可以给出由其它分布$F(x)$的随机数生成U(0, 1)随机数的方法。

随机数的生成

随机变量的生成

2. 生成随机数的一般方法

- 手工方法
  - 抽签
  - 掷骰子
  - 摸球
- 物理方法
  - 噪声发生器
  - 放射源激励计数器
- 数学方法

按照一定的算法(递推公式)来生成“随机数”序列(也称随机数流)。

3. 伪随机数

伪随机数：用数学方法生成的随机数。

随机数发生器：用数学方法生成随机数所依赖的算法和程序。

从本质上说伪随机数并不具有真正的随机性，但如果精心设计算法，可以生成具有真正随机数的一些统计性质的伪随机数。

通常，只要所生成的伪随机数能通过一系列统计检验(如独立性、均匀性等)，就可以把它们作为真正的随机数使用。

伪随机数发生器应具备的特征：

① 生成的随机数流要具有均匀随机数的统计性质，如所服从分布的均匀性，抽样的随机性、序列间的独立性等；

② 生成的随机数流要有足够长的周期，以满足仿真计算的需要；

③ 生成随机数流的速度快，占用计算机的内存少，具有完全可重复性。

二. 均匀随机数的生成

1. 早期的随机数发生器

(1) 平方取中 (Mid-Square Approach)

算法：

① 任取一个2k位整数 (k为任意正整数)作为种子值(初值)；

② 将种子值平方得4k位整数(不足4k位时高位补0)；

③ 取此4k位的中间2k位整数作为下一个种子值；

④ 规范化种子值，可得一均匀随机数。

重复上述过程，即可得一系列随机数。

例3.1 取 k =1，2k=2位整数76为第一个种子值x_0，则由平方取中法可得

n	② 平方$x_{n-1}\cdot x_{n-1}$	③ 取中x_n	④ 规范化u_n
1	$(76)^2$= 5776	x_1=77	u_1=0.77
2	$(77)^2$= 5929	x_2=92	u_2=0.92
3	$(92)^2$= 8464	x_3=46	u_3=0.46
4	$(46)^2$= 2116	x_4=11	u_4=0.11
5	$(11)^2$= 0121	x_5=12	u_5=0.12
	………………………………		
11	$(84)^2$= 7056	x_{11}=05	u_{11}=0.05
12	$(05)^2$= 0025	x_{12}=02	u_{12}=0.02
13	$(02)^2$= 0004	x_{13}=00	u_{13}=0.00
14	$(00)^2$= 0000	x_{14}=00	u_{14}=0.00

退化!!!

平方取中法的递推公式为：

$$\begin{cases} x_n = [\dfrac{x_{n-1}^2}{10^k}] \bmod 10^{2k} \\ u_n = x_n / 10^{2k} \qquad (n = 1,2,\cdots) \\ \text{初值 } x_0 \end{cases}$$

其中x_0为2k位的非负整数，$[x]$表示取x的整数部分，N mod M 表示对 N 进行模为M的求余运算，即

$$N \bmod M = N - [\frac{N}{M}] \times M$$

(2) 乘积取中法 (Mid-Multiplication Approach)

乘积取中法需要取两个2k位的(k为任意正整数)种子值 x_0 , x_1 ，其递推公式：

$$\begin{cases} x_{n+1}=[\dfrac{x_{n-1}x_n}{10^k}]\bmod 10^{2k} \\ u_{n+1}=x_{n+1}/10^{2k} \qquad (n=1,2,\cdots) \\ \text{初值 } x_0, x_1 \end{cases}$$

例3.2 取 $k=2$, $x_0=5167$, $x_1=3729$，则由乘积取中法可得

n	乘积$x_{n-1}\cdot x_n$	取中x_{n+1}	规范化u_n
1	5167 · 3729 = 19267743	x_2=2677	u_2=0.2677
2	3729 · 2677 = 09982533	x_3=9825	u_3=0.9825
3	2677 · 9825 = 26301525	x_4=3015	u_4=0.3015
4	9825 · 3015 = 29622375	x_5=6223	u_5=0.6223
5	3015 · 6223 = 18762345	x_6=7623	u_6=0.7623
6	6223 · 7623 = 47437929	x_7=4379	u_7=0.4379
7	7623 · 4379 = 33381117	x_8=3811	u_8=0. 3811

………………………………

2. 线性同余法 (Linear Congruence Generator)

线性同余法简称为LCG方法或线性同余发生器，其递推公式为

$$\begin{cases} x_n=(ax_{n-1}+c)\bmod m \\ u_n=x_n/m \qquad (n=1,2,\cdots) \\ \text{初值} x_0 \end{cases}$$

其中m为模数，a为乘子(乘数)，c为增量(加数)，且 x_0, m, a, c 均为非负整数。

显然由上述递推得到的x_n满足：$0\le x_n<m$。从而x_n至多能取m个不同的整数。

周期：对初值x_0，同余法$x_n=(a x_{n-1}+c)\bmod m$产生的数列$\{x_n\}$($n=1,2,\ldots$)，其重复数之间的最短长度(循环长度)称为此初值下LCG的周期，记为T。若 T = m ，则称之为满周期。

注：用线性同余法产生随机数时，参数a, c, x_0, m的选取十分关键！

例3.3 取 $m=8$, $a=3$, $c=1$, $x_0=1$，则由线性同余法可得 x_n , u_n 如下：

$$\begin{cases} x_n=(3x_{n-1}+1)\bmod 8 \\ u_n=x_n/8 \qquad (n=1,2,\cdots) \\ x_0=1 \end{cases}$$

n	1	2	3	4	5	6	7	8	9	…
$3x_{n-1}+1$	4	13	16	1	4	13	16	1	4	…
x_n	4	5	0	1	4	5	0	1	4	…
u_n	0.5	0.625	0	0.125	0.5	0.625	0	0.125	0.5	…

易见 $x_1=x_5=4$，且从 $n=5$开始 x_n (u_n)循环取x_1 (u_1)到x_4 (u_4)的值，周期T = 4 < m =8，非满周期 。

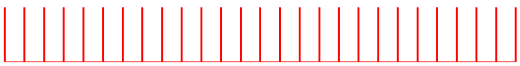
例3.4 取 $m=8$, $a=5$, $c=5$, $x_0=5$，则由线性同余法可得 x_n , u_n 如下：

$$\begin{cases} x_n=(5x_{n-1}+5)\bmod 8 \\ u_n=x_n/8 \qquad (n=1,2,\cdots) \\ x_0=5 \end{cases}$$

n	1	2	3	4	5	6	7	8	9	…
$5x_{n-1}+5$	30	35	20	25	10	15	40	5	30	…
x_n	6	3	4	1	2	7	0	5	6	…
u_n	0.75	0.375	0.5	0.125	0.25	0.875	0	0.625	0.75	…

此例中 $x_1=x_9=6$，且从 $n=9$开始 x_n (u_n)循环取x_1 (u_1)到x_9 (u_9)的值，周期T = 8= m ，满周期 。

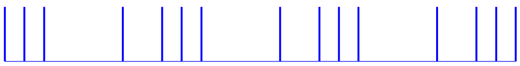
满周期



6 3 4 1 2 7 0 5 6 3 4 1 2 7 0 5........ .... m=8 T=8

均匀

非满周期



4 5 0 2 6 7 1 3 4 5 0 2 6 7 1 3........ .... m=8 T=4

不均匀

线性同余法为满周期的参数的选取准则。

定理 若参数a, c, m满足下列三个条件，则线性同余法LCG可达到满周期：

① c 与 m 互素，即可以同时整除 c 和 m 的正整数只有1；

② 对任一素数 q，若 q 能整除 m，则 q 也能整除 $a-1$；

③ 若 4 能整除 m，则 4 也能整除 $a-1$。

条件 ①，②，③ 成立 $\Longrightarrow$ LCG满周期

(1) 混合同余法

在线性同余法的递推公式中当参数 c > 0 时的LCG方法称为混合同余法。

为延长随机数发生器的周期，通常取 $m=2^{b-1}$, b 为所用计算机字长。例. b=32, $m=2^{31}$=2147483648.

优点：

① 可使随机数的周期尽可能地大，适当选取 a, c, 使周期 T = m 取到最大值；

② 算法上利用计算机的“整数溢出”原理可简化计算，提高产生随机数的效率。

(2) 乘同余法

在线性同余法的递推公式中当参数 c= 0 时的LCG方法称为乘同余法，其递推公式为：

$$\begin{cases} x_n = (a x_{n-1}) \bmod m \\ u_n = x_n / m \qquad\qquad (n=1,2,\cdots) \\ \text{初值 } x_0 \end{cases}$$

乘同余法不可能达到满周期。

关注问题：乘同余法的最大周期T=？如何选取参数可使乘同余法达到最大周期？统计特征？

三. 均匀随机数的检验

1. 均匀性检验(频率检验)

该方法检验经验频率与理论频率的差异是否显著。

设 u_i(i=1, …, n)是待检验的随机数，作统计假设：

$$H_0:\ u_i \sim U(0,1)$$

(1) χ^2 检验

χ^2检验的具体步骤如下：

① 将[0, 1]区间分成 m 个不相交的小区间:

$$[\frac{i-1}{m}, \frac{i}{m})(i=1,2,\cdots,m)$$

② 由假设$\{u_j\}$落入第 i 个小区间的概率为$p_i = 1/m$，计算理论频数

$$np_i = n\left(\frac{i}{m}-\frac{i-1}{m}\right) = \frac{n}{m} \quad (i=1,2,\cdots,m)$$

③ 计算$\{u_j\}$落在第 i 个区间中的个数n_i (i=1, 2,…, m)，称之为经验频数。

④ 由于统计量

$$\chi^2 = \sum_{i=1}^{m}\frac{(n_i - np_i)^2}{np_i} = \frac{m}{n}\sum_{i=1}^{m}(n_i - \frac{n}{m})^2$$

渐近服从$\chi^2(m\text{-}1)$分布，对给定水平α，查χ^2分布表得临界值 $\chi^2_\alpha(m-1)$: $P\{\chi^2 > \chi^2_\alpha(m-1)\} = \alpha$ 。

⑤ 利用④计算出 χ^2 的值，若

$$\chi^2 \le \chi^2_\alpha(m-1)$$

则可认为经验频数与理论频数没有显著差异；否则差异显著，从而认为假设不成立。

例3.5 给定水平α＝0.05，试检验下面随机数序列(n=100))的均匀性：

0.34 0.90 0.25 0.89 0.87 0.44 0.12 0.21 0.46 0.67 0.83 0.76 0.79
0.64 0.70 0.81 0.94 0.74 0.22 0.74 0.96 0.99 0.77 0.67 0.56 0.41
0.52 0.73 0.99 0.02 0.47 0.30 0.17 0.82 0.56 0.05 0.45 0.31 0.78
0.05 0.79 0.71 0.23 0.19 0.82 0.93 0.65 0.37 0.39 0.42 0.99 0.17
0.99 0.46 0.05 0.66 0.10 0.42 0.18 0.49 0.37 0.51 0.54 0.01 0.81
0.28 0.69 0.34 0.75 0.49 0.72 0.43 0.56 0.97 0.30 0.94 0.96 0.58
0.73 0.05 0.06 0.39 0.84 0.24 0.40 0.64 0.40 0.19 0.79 0.62 0.18
0.26 0.97 0.88 0.64 0.47 0.60 0.11 0.29 0.78

将[0,1)区间等分成10个子区间(即 m ＝10)，并统计出随机数序列落在各子区间中的个数(经验频数)n_i 分别为：7, 9, 8, 9, 14, 7, 10, 15, 9, 12，易见理论频数为np_i＝n / m＝10。由此计算得

$$\chi^2 = \frac{m}{n}\sum_{i=1}^{m}(n_i - \frac{n}{m})^2 = 0.1\sum_{i=1}^{10}(n_i - 10)^2 = 7$$

查表得 $\chi^2_{0.05}(9) = 16.92, \chi^2 < 16.92$，故可接受假设$H_0$, 即认为$\{u_j\}$ 的分布函数与U(0,1)分布没有显著差异。

(2) 柯尔莫哥洛夫－斯米尔诺夫检验
(Kolmogorov－Smirnov)

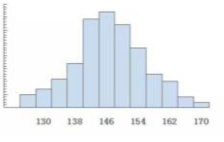
K－S检验的具体步骤如下：

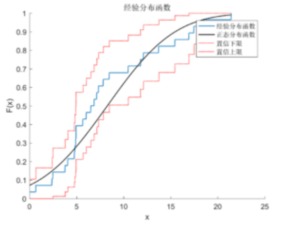
① 将随机数$\{u_i\}$ 从小到大排序后记为 $\{u_{(i)}\}$, 其经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < u_{(1)} \\ \dfrac{i}{n}, u_{(i)} \le x < u_{(i+1)} & i = 1,2,\cdots,n-1 \\ 1, & x \ge u_{(n)} \end{cases}$$

② 将$F_n(x)$与$U(0, 1)$的分布函数$F(x)=x$ ($0 \le x \le 1$)比较，计算最大偏差为

$$D_n = \max\{D_n^+, D_n^-\}$$





其中

$$D_n^+ = \max_{1\le i\le n}\left\{\left|F_n(u_{(i)}) - F(u_{(i)})\right|\right\} = \max_{1\le i\le n}\left\{\left|\frac{i}{n} - u_{(i)}\right|\right\}$$

$$D_n^- = \max_{1\le i\le n}\left\{\left|F(u_{(i)}) - F_n(u_{(i-1)})\right|\right\} = \max_{1\le i\le n}\left\{\left|u_{(i)} - \frac{i-1}{n}\right|\right\}$$

③ 注意D_n 渐近服从柯尔莫哥洛夫－斯米尔诺夫分布。对给定水平α，查K－S分布表得临界值

$$D_\alpha(n): P\{D_n > D_\alpha(n)\} = \alpha$$

④ 若 $D_n \le D_\alpha(n)$ ，则可接受假设，即认为经验分布函数与均匀分布函数之间没有显著差异；否则，有显著差异。

例3.6 给定水平α＝0.05，对例3.5中前10个随机数用K－S法检验其均匀性。

把随机数由小到大排列，并将相关数据列表如下：

i	1	2	3	4	5	6	7	8	9	10
$u(i)$	0.12	0.21	0.25	0.34	0.44	0.46	0.67	0.87	0.89	0.90
i/n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\|i/n-u(i)\|$	0.02	0.01	0.05	0.06	0.06	0.14	0.03	0.07	0.01	0.1
$\|u(i)-(i-1)/n\|$	0.12	0.11	0.05	0.04	0.04	0.04	0.07	0.17	0.09	0

由此表可得D^+＝0.14，D^-＝0.17，故 Dn＝max{0.14, 0.17}＝0.17。查K－S分布表得$D_{0.05}(10)$＝0.41 > 0.17，因此可以接受总体分布与均匀分布之间无显著差异的假设。

2. 独立性检验

(1) 相关系数检验

(2) 游程检验

(3) 扑克检验

(4) 连贯性检验

相关系数检验/自相关检验

自相关检验利用相关系数进行随机数的独立性检验。相关系数反映了随机变量之间的线性相关程度。

设给定n个随机数$x_1, x_2, \dots, x_n$，假设j阶自相关系数为$\rho_j = 0\ (j = 1, 2, \dots, m)$。样本的j阶自相关系数为

$$\rho_j = C_j / C_0$$

$$C_j = Cov(X_i, X_{i+j}) = E(X_i X_{i+j}) - E(X_i)E(X_{i+j})$$

$$C_0 = Var(X_i)$$

对于U(0, 1)分布，$\rho_j = 12E(X_i X_{i+j}) - 3$，考虑滞后m的自相关系数$\rho_{im}$（即$u_i, u_{i+m}, u_{i+2m}, \dots, u_{i+(h+1)m}$）

$$\hat{\rho}_{im} = \frac{12}{k+1}\sum_{k=0}^{h} u_{i+km}u_{i+(k+1)m} - 3$$

$h = \lfloor (n-i)/m \rfloor - 1$是使得$i + (h+1)m \le n$成立的最大整数，$\lfloor (n-i)/m \rfloor$代表下确界。

当h值很大，如果$u_i, u_{i+m}, u_{i+2m}, \dots, u_{i+(h+1)m}$之间不相关，则$\hat{\rho}_{im}$近似于正态分布$N(0, \hat{\sigma}_{im}^2)$，则利用统计量Z可以对数列相关性进行检验。对于给定的显著性水平α，若$|Z| < Z_{\alpha/2}$，则认为u_n具有统计上的独立性；反之则不具有独立性。

$$\hat{\sigma}_{im}^2 = \frac{13h+7}{(h+1)^2}$$

游程检验

二分变量

随机排列

1.男\男，女\女\女，男，女\女，男\男\男\男

2.男\男\男\男\男\男\男，女\女\女\女\女

3.男，女，男，女，男，女，男，女，男，女，男\男

连续出现男或女的区段为游程

扔硬币

正面是1，反面是0。

001101110001001

游程总数检验 或者 最大游程检验