统计计算期末论文

0-1均匀分布随机数发生器实验研究

作者：	专业：
姚智元（1310100）	应用数学
杨帅（1310096）	统计学

南开大学数学科学学院

2016 年 5 月 30 日

摘要

随着信息技术的飞速发展，随机数发生器在诸多方面都有着广泛的应用，如信息安全、科学实验模拟、雷达中的测距信号、遥控遥测中的测控信号等等。

虽然通过提取自然物理噪声来制作的真随机数发生器在随机性上有着来自软件和固定算法制作的伪随机数发生器不可逾越的优势，但是真随机数发生器存在生成过程不能大量重复，生成速度慢等缺点，而且在一些领域中，我们并不需要成本高昂的真随机数，伪随机数即可满足要求。故能生成随机性质好的伪随机数序列尤为重要。而所有的分布的随机数获取都与0到1上均匀分布随机数的获取分不开。

本文主要介绍了一些0到1上均匀分布的随机数的生成算法以及统计检验方法，并对现有的一些伪随机数发生器进行了统计检验和Monte-Carlo模拟。

关键字：0-1均匀分布　　伪随机数生成器　　统计检验　　Monte-Carlo模拟

目录

1 前言

1.1 均匀分布随机序列的意义

随机序列在不同的领域有许多不同类型的应用。如雷达中的测距信号，遥控遥测中的测控信号，数字通信中的群同步和加扰解扰信号，无线通信码分多址系统中的扩频信号等都要用到随机序列。在通信系统仿真中，系统决策需要进行多次重复的仿真对结果进行对比研究，因此生成可重复使用的伪随机数序列以及对该生成的随机数序列的统计特性进行检验显得尤为重要。伪随机数序列的生成是计算机仿真由其是离散事件系统仿真的基础，其品质是直接影响计算机仿真结果的有效的最重要的因素之一。模拟计算表明，通过改进的伪随机数的品质，常常可使计算机仿真结果的有效性得到明显地提高。在有些情况下甚至可使仿真结果的数值相对误差缩小到原来地功或更小。反之，品质低劣的伪随机数的使用，则有可能导致不可信的模拟结果。因此进行伪随机数发生器的研究与检验具有重大的理论意义和实际应用价值。

随机序列的生成的重点是生成服从均匀分布的随机序列。从数学的角度出发，只要能得到来自均匀分布的序列，就可以由如反函数法、筛选法等生成来自确定分布的随机序列，故本文重点讨论来自均匀分布的随机序列生成方法。

1.2 随机数的概念和分类

要定义随机数的概念，必须先理解“随机”这两个字的意思。通常，随机意味着不确定性，不可预测性，可能出现也可能不出现等概念。简单举一个例子, 假如一个随机序列由“0”序列和“1”序列组成,那么我们不能确定当前状态是“0”序列还是“1”序列，也不能根据当前状态或以前的状态来推测下一个状态，它的结果是完全不可预测的。

衡量一个随机序列质量的主要指标有周期性、相关性、分布特性等。周期性要求随机序列具有无限长的周期，序列的任何两个子序列都不可能重复。相关性是指一个随机序列中各个随机数互不相关，而且各个子序列间也互不相关。对于真随机数序列当然满足上述两个条件，但对于伪随机数序列而言，要做到这点就很难，因为伪随机数都是用特定的数学公式生成，其生成的序列都有确定的周期。因此我们应慎重地选择数学公式和算法，使序列的周期尽可能的长，相关性尽可能接近于

零。分布特性是指把随机数的区间均匀地分成 n 等份，然后统计每个区间上的随机数个数，然后画出其分布曲线，是服从均匀分布、正态分布还是指数分布。

不同的文献对随机数的分类都不一样，这里我们采用最多的一种分类方法。一般而言，随机数可以分为三类真随机数、准随机数、伪随机数。

真随机数只能由某些随机物理过程来产生，如电子设备的热噪声、宇宙噪声、放射性衰变等。真随机数是完全不可预测的，在任何情况下都不可能重复产生两个完全相同的随机数序列。由于真随机数是完全不可预测的，所以真随机数广泛地应用于加密领域来保证信息安全。但真随机数的生成速率较低，而且在物理实现上也有一定的难度。

准随机数并不具有随机性质，仅仅用来处理问题的时候，能够得到正确的结果。准随机数属于伪随机数的一种，所以要实现严格数学意义上的随机性，在理论上是不可能的，在实际应用中也完全没有这个必要。关键是保证“随机”数序列能产生所需要的结果，符合某个特定的应用。

伪随机数通常由数学公式产生，如果生成随机数的算法确定了，那么这个随机数序列也就确定了。所以从数学意义上讲，伪随机数是完全不随机的，只要知道随机数的种子以及相应的参数，我们就可以预测下一状态的随机数。但是只要伪随机数通过一系列测试检验，而且其有足够长的周期，那么在一定范围内，还是可以把它当作真随机数来使用，而且伪随机数产生器在物理实现上比真随机数简单得多，它的生成速度比真随机数快得多。因此伪随机数广泛地应用于各个领域。对伪随机数而言,其实现算法显得尤为重要。

本文主要讨论利用计算机及数学算法生成的服从[0,1]均匀分布的伪随机数序列。

1.3 好的伪随机数应具备的条件

在计算机上用数学方法产生均匀随机数是按照一定的计算方法而产生的，它们具有类似于均匀随机变量的独立抽样序列的性质，这些数既然是依照确定算法产生的，便不可能是真正的随机数，因此常把用数学方法产生的随机数称为伪随机数。虽然如此，如果计算方法经过细心的设计，可以产生看起来是相互独立的，区间上均匀分布的随机数，并且可以通过一系列的统计检验如均匀性、独立性等等。也就是说，只要具有真正均匀随机数的一些统计性质，我们就可以把伪随机数作为真正

的随机数使用。

按照某一递推公式 $\eta_n = f(\eta_{n-1}, \eta_{n-2}, \cdots, \eta_{n-k})$ 产生数列 $\eta_1, \eta_2, \cdots, \eta_n, \cdots$ 使得当 n 充分大时，此数列具有均匀分布随机变量的独立抽样序列的性质，此数列就称为伪随机数列。

显然按照递推公式计算产生的伪随机数，到一定长度之后或退化为零或周而复始地出现周期现象同时按照递推公式进行递推计算，如果我们确定了初始值，整个序列就可以精确地给出。可见，这些与随机数应具有的基本统计性质是矛盾的。一个用递推公式产生具有均匀分布随机变量的独立抽样序列性质的数学方法，即一个好的随机数发生器应当具备以下几点条件:

1. 产生的随机数序列要具有均匀总体随机样本的统计性质，如分布的均匀性，抽样的随机性，数列间的独立性等等。

2. 产生的随机数序列要具有较好的理论支持或具有好的格结构，如分布的谱检验和高维分布检验一。

3. 产生的随机数序列要有足够长的周期，以满足模拟计算的需要。

4. 产生的随机数序列的速度快，占用计算机的内存少，具有完全可重复性。

随着计算机运算速度的不断提高，存储容量和字长的不断扩大，3和4两点一般都能满足至于条件2属于理论方面的问题，由于涉及到很多数论等基础数学方面的内容，本文不做过多的研究。因此，关键是如何构造满足1的随机数发生器，即做大量的统计检验工作。最后利用Monte-Carlo方法进行数学实验，测试每个随机数的效果。

2 伪随机数发生器

2.1 取中法

2.1.1 平方取中法

最早产生随机数的方法是平方取中法。以生成4位随机数为例，将一个4位数平方后得到一个8位数（不足8 位在前面补0），然后截取中间的4位为新的随机数，重

复上述过程即可得到一位随机数序列。

平方取中法递推公式：

$$x_{n+1} = (x_n^2/10^2)\ mod\ 10^4$$

产生 $[0,1]$ 之间随机数序列

$$r_n = x_n/10^4$$

优点：

在计算机上易于实现，内存占用少。

缺点：

均匀性不好，对小数目存在偏倚的现象，均匀性不好，数列的长度和周期难以确定，对初始数据依赖很大。

例：

利用平方取中法产生两位数的随机数序列，种子数取为 $x_0 = 44$ 通过计算得到：

$$44^2 = 1936;\ x_1 = 93,$$
$$93^2 = 8649;\ x_2 = 64,$$
$$64^2 = 4096;\ x_3 = 09,$$
$$09^2 = 0081;\ x_4 = 08,$$
$$08^2 = 0064;\ x_5 = 06,$$
$$06^2 = 0036;\ x_6 = 03,$$
$$03^2 = 0009;\ x_7 = 00,$$
$$00^2 = 0000;\ x_8 = 00,$$

这样，随机数就无法继续产生了。

利用平方取中法的另一个问题是这个方法可能产生退化，即总是得到相同的 x_i 值。

例：

设在产生四位数的随机数过程中，得到了一个 $x_i = 4500$ 从而

$$4500^2 = 20250000;\ x_{i+1} = 2500,$$
$$2500^2 = 06250000;\ x_{i+2} = 2500,$$

相继产生的数值总为2500。由于这些原因，平方取中法已经被许多新的能提供更可靠的随机数序列的算法所取代。

2.1.2 乘积取中法

迭代公式为：

$$x_{n+2} = (x_{n+1}x_n/10^2)\ mod\ 10^4$$

产生 $[0,1]$ 之间随机数序列

$$r_n = x_n/10^4$$

乘积取中法较之于平方取中法，产生的随机数序列长度和均匀性都有改善，但是数列长度还是不够，而且对初始值的依赖很大。

2.2 移位法

电子计算机善于进行位移等逻辑运算，利用计算机这一特点而产生伪随机数列的方法称为移位法。

一个字长为32位的计算机，取一初始值 x_0 化为二进制数后向左移7位得到 x_{01}，右移7位得到 x_{02}，将 x_{01} 和 x_{02} 相加得到 x_1，再对 x_1 进行上述运算过程得到 x_2，这样重复多次可得正整数序列 $\{x_n\}$，取 $r_n = x_n 2^{-32}$ 为 $[0,1]$ 上均匀分布的伪随机数

列通项。这一算法的递推公式为：

$$x_{n+1} = x_n 2^7 + x_n 2^{-7} \ mod \ 2^{32}$$

产生伪随机数列：

$$r_n = x_n 2^{-32}$$

2.3 同余法

同余法是目前产生伪随机数比较好的方法，其中包括混合同余法、加同余法和乘同余法，还有一些改进的同余法如二次平方同余法等。

2.3.1 线性同余法

线性递推公式：

$$x_{n+1} = (ax_n + c) \ mod \ m$$

产生伪随机数列：

$$r_n = x_n / m$$

其中 a , c , m 以及初始值 x_0 都是正整数，容易看出 r_n 满足

$$0 \leqslant r_n < 1$$

当 $c \neq 0$ 且 $a \neq 1$ ，则称这种方法为混合同余法。

当 $c = 0$ 时，这种方法称为乘同余法。

当 $c \neq 0$ 且 $a = 1$ 时，这种方法称为加同余法。

虽然加同余法产生的序列周期长，电子计算机实现也很方便，只要进行加法及移位运算即可完成，但从理论上看，所得到的随机数列的性质一般不如乘同余法和混合同余法。

若 $a = c = 1$ ，则有 $(x_{n+1} = x_n + 1) \ mod \ m$ 。这样构成的序列虽然周期可以达到 m ，但不是随机的，因此乘子 a ,增量 c 的选取是非常重要的，下面给出 a 和 c 的所

有使得周期最大的选取方法，以便从中挑选符合统计性质的伪随机数列。

1. m 的选择：
   由于重复周期的长度总是小于 m，因此需要将 m 取大的数值，更进一步，所选用的 m 的值应能简化同余关系的解，在计算机中数字都是用二进制表达的，因此已经证明 m 取值为 $(2^k - 1)$ 是很好的，k 为计算机的字长。

2. a 和 c 的选取：
   当且仅当下列条件满足时，一个有线性同余法产生的随机数序列的重复周期为 m。

   - c 与 m 互质。
   - $(a - 1)$ 是每个能整除 m 的质数的倍数。即如果 q 是整除 m 的质数，则 q 能整除 $(a - 1)$。
   - 如果 m 能被4整除，则 $(a - 1)$ 也能被4整除。这些限制使得乘子值满足形式为 $a = z^p + 1$, 其中 z 是计算机中用于表示数字的基数；模数的形式为 $m = z^k$，k 为计算机的字长，且有 $z \leqslant p < k$。特别的，当选择 $a = 2^{16} + 5 = 65541$ 或 $a = 2^{16} + 3 = 65539$ 时可以得到满意的结果，至于 c 的选择，只要满足 c 与 m 互为质数的条件即可。

3. x_0 的选取：
   如果随机数序列的周期为 m，因为能产生完全的序列，即在一个周期内可以取到 0 至 $(m - 1)$ 的所有值，因此 $x_0 = 0$ 时会产生退化的序列。

用线性同余法产生随机数，有一些缺陷，主要指该方法产生的均匀随机数作为 $m(m > 1)$ 维均匀随机向量时相关性大。在整数的尾数字长为 L 位的计算机上，不可能得到 $T > 2^L$ 均匀随机数列。

2.3.2 改进的加同余法

改进的加同余法需要 n 个数的序列作为它的种子，这 n 个数的序列可以应用其他的方法产生，应用加同余法可以使这个序列不断的扩大，改进的加同余法的算法

是：

$$x_j = (x_{j-1} + x_{j-n}) \ mod \ m$$

这种方法的好处是速度快，因为它不需要作乘法运算。它可以得到大于 m 的周期，但这种方法产生随机数的过程不象混合同余法那样清楚，因此由此方法产生的随机数序列需要经过仔细的确认。

2.3.3 二次平方同余法

这种方法适用于 m 为2的幂次情况，它的递推关系式为：

$$x_{n+1} = [x_n(x_n + 1)] \ mod \ m, \ n > 0$$

种子数 x_0 必须满足条件 $x_0 \ mod \ 4 = 2.$

2.4 移位寄存器序列

移位寄存器序列也称 *Tausworthe* 序列。这本是随机产生0,1二进制位的方法。我们称这种方法为反馈位移寄存器法(Feedback ShiftRegister Methods)，简称FSR方法或FSR发生器。

Tausworthe（1965）提出FSR方法，用线性反馈递推公式：

$$a_k = (c_p a_{k-p} + c_{p-1} a_{k-p+1} + \cdots + c_1 a_{k-1}) \ mod \ 2$$

对寄存器中的二进制数码 a_k 作递推运算,其中 p 是给定正整数，$c_p = 1, c_i = 0$ 或1 $(i = 1, 2, \cdots, p-1)$ 为给定常数。给定初值 $(a_p, a_{p+1}, \cdots, a_1)$，由公式产生的0或1值组成二进制数列 $\{a_n\}$。截取数列 $\{a_n\}$ 中连续的L位构成一个位二进制整数

接着截取L位又形成一个整数,以此类推,即得:

$$
\begin{aligned}
x_1 &= (a_1, a_2, \cdots, a_L)_2 \\
x_1 &= (a_{L+1}, a_{L+2}, \cdots, a_{2L})_2 \\
&\cdots\cdots\cdots\cdots \\
x_n &= (a_{(n-1)L+1}, a_{(n-1)L+2}, \cdots, a_{nL})_2 \ (n = 1, 2, 3, \cdots)
\end{aligned}
$$

令 $r_n = \frac{x_n}{2^L} \quad (n = 1, 2, 3, \cdots)$

则 $\{r_n\}$ 即为FSR方法产生的均匀随机序列。通过改变 p 及 $c_1, c_2, \cdots, c_p$ 的值将得到不同的FSR发生器。

关于 *Fibonacci* 序列：历史上曾使用此序列产生随机数，其递推公式为：

$$x_{i+1} = (x_i + x_{i-1}) \ mod \ m, \ i = 1, 2, \cdots$$

例如 $m = 100, x_0 = 1, x_1 = 1$ 则有序列：

$$1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 44, 33, 77, 10, 87, \cdots$$

此发生器没有乘法运算，产生速度快，但是它存在着令人不能容忍的不居中现象，即由前两个数得到的第三个数要不是同时大于就是同时小于前二者而永不居中。此序列的另一个缺点是显著的序列相关，即取小值的数后面出现也取小值的趋势。所有这些都说明它不是一个好的随机数发生器。

3　伪随机数发生器的效果检验

3.1　检验类型

在一般情况下,统计检验分为以下两类不同的检验方法:

1. 经验检验:
   是以发生器产生的均匀伪随机数序列 $\{r_i\}$ 为基础,根据 [0, 1] 区间上均匀总体简

单随机样本 $\{u_i\}$ 的性质,如特征向量、均匀性、随机性等,研究产生的随机数序列 $\{r_i\}$ 的相应性质,进行比较、鉴别,视其差异是否显著,决定取舍。

2. 理论检验:
   理论检验从统计意义上讲并不是一种检验。它用一种综合的方法来评估发生器的数值参数,而根本不必产生任何随机数序列 $\{r_i\}$,即它是一种理论上的研究。由于理论检验从数学上推导困难,故不予讨论。

3.2 检验步骤

统计检验的一般步骤:

1. 提出原假设 H_0 :总体分布为均匀分布 $u(0,1)$;
2. 构建适当的统计量 $T = T(x_1, x_2, \cdots, x_n)$ 为样本,其中 $(x_1, x_2, \cdots, x_n)$ 为样本,并求出统计量 T 在原假设 H_0 成立时的分布;
3. 给出显著水平量 α,给出拒绝域 W : W 使 $P\{T(x_1, x_2, \cdots, x_n) \in W\} = \alpha$.本文 α 取0.95;
4. 由观测值(样本值)计算 T 值;
5. 作统计推断,当 $T \in W$ 时,否定 H_0 ;否则接受 H_0。

3.3 检验统计量

在均匀随机数的统计检验中,用到如下两类统计量:

1. 正态分布统计量 η
   设 $\eta_1, \eta_2, \cdots, \eta_n, \cdots$ 是相互独立同 $F(x)$ 分布,且 $E(\eta_i) = \mu, Var(\eta_i) = \sigma^2$ 记 $\bar{\eta} = \frac{1}{n}\sum_1^n \eta_i$ ， 则 $U = \frac{\bar{\eta}-\mu}{\eta}\sqrt{n}$ 以 $N(0,1)$ 为极限分布。
2. 卡方 (χ^2) 统计量
   将总体 η 的简单子样 $\eta_1, \eta_2, \cdots, \eta_n, \cdots$ 按一定规则分为互不相交的 m 个组,记

第 i 组的观测频数为 $n_i(i = 1, 2, \cdots, m)$。若随机变量 η 属于第 i 组的概率为 P_i ,记理论频数 $\mu_i = np_i$, 由 n_i, μ_i 构造统计量:

$$V = \sum_{i=1}^{m} \frac{(n_i - \mu_i)^2}{\mu_i}$$

渐进服从 $\chi^2(k), k = m - l - 1$， l 是附加在概率分布 $\{P_i\}$ 上独立约束条件的个数(即确定概率 P_i 时利用样本估计总体参数的个数)。

3.4 统计检验方法

统计检验的一般方法:

1. 参数检验:
   均匀随机数的参数检验是检验由某个发生器产生的随机数序列 $\{r_i\}$ 的均值、方差和各阶矩等与均匀分布的理论值是否有显著的差异。

2. 均匀性检验:
   随机数的均匀性检验又称频率检验,它是来检验由某个发生器产生的随机数序列 $\{r_i\}$ 是否均匀地分布在 $[0, 1]$ 区间上。也就是检验经验频率与理论频率的差异是否显著。

3. 独立性检验:
   独立性检验主要是检验随机数序 $r_1, r_2, r_3, \cdots, r_n$ 之间的统计相关性是否显著。

3.4.1 参数检验

均匀随机数的参数检验是检验由某个发生器（或者算法）产生的随机数序列 r_i 的均值、方差、各阶矩等与均匀分布的理论值是否有显著的差异。若随机变量 $R \sim U(0, 1)$,则 $E(R) = \frac{1}{2}, Var(R) = \frac{1}{12}, E(R^2) = \frac{1}{3}$。 $R_1, R_2, R_3, \cdots, R_n$ 是U(0,1) 总体的

简单的随机样本,即 $R_1, R_2, R_3, \cdots, R_n$ 相互独立同 $U(0,1)$ 分布。记:

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$$
$$\bar{R}^2 = \frac{1}{n}\sum_{i=1}^{n} R_i^2$$
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(R_i - \frac{1}{2})^2$$

则有:

$$E(\bar{R}) = \frac{1}{2} \qquad Var(\bar{R}) = \frac{1}{12n}$$
$$E(\bar{R}^2) = \frac{1}{3} \qquad Var(\bar{R}^2) = \frac{4}{45n}$$
$$E(S^2) = \frac{1}{12} \qquad Var(S^2) = \frac{1}{180n}$$

设 $r_1, r_2, r_3, \cdots, r_n$ 是某个发生器产生的随机数。首先对特征量作统计检验,在 $\{r_i\}$ 是均匀总体的简单随机样本的假设下,而以下统计量:

$$u_1 = \frac{\bar{r} - E(\bar{R})}{\sqrt{Var(\bar{R})}} = \sqrt{12n}(\bar{r} - \frac{1}{2})$$
$$u_2 = \frac{\bar{r}^2 - E(\bar{R}^2)}{\sqrt{Var(\bar{R}^2)}} = \frac{\sqrt{45n}}{2}(\bar{r}^2 - \frac{1}{3})$$
$$u_2 = \frac{S^2 - E(S^2)}{\sqrt{Var(S^2)}} = \sqrt{180n}(S^2 - \frac{1}{12})$$

由中心极限定理渐进服从 $N(0,1)$。给定显著水平 α 后,查标准正态数值表得 $\lambda : P\{|u| > \lambda\} = \alpha, u_i \sim N(0,1)$,否定域 $W_i = \{|u_i| > \lambda\}$。由随机数序列 $\{r_i\}$ 计算 u_i 的值,若 $|u_i| \leqslant \lambda$,则认为产生的随机数序列的特征量与均匀总体的特征量没有显著的差异;否则,由于 $\{r_i\}$ 的特征量与均匀总体的特征量有显著的差异,故不能认为 $\{r_i\}$ 是均匀总体的简单样本。

3.4.2 均匀性检验

随机数的均匀性检验又称频率检验,它是来检验由某个发生器(或者算法)产生的随机数序列 $\{r_i\}$ 是否均匀地分布在 $[0,1]$ 区间上,即检验经验频率与理论频率的差异是否显著。均匀性检验种常用的方法有:卡方检验（ $Pearson\,\chi^2$ 拟合优度检验）、柯氏检验（ $Kolmogorov$ 检验）与序列检验。

1. 卡方检验

   设 $r_1, r_2, r_3, \cdots, r_n$ 是待检验的一组随机数,假设 $H_0 : r_1, r_2, r_3, \cdots, r_n$ 为均匀总体的简单样本。将 $[0,1]$ 区间分为 m 个小区间,以 $[\frac{(i-1)}{m}, \frac{i}{m}]\ (i = 1, 2, \cdots, m)$ 表示第 i 个小区间,设 $\{r_j\}\ (j = 1, 2, \cdots, m)$ 落入第 i 个区间的数目为 $n_i\ (i = 1, 2, \cdots, m)$。根据均匀性假设,落入每个小区间的概率为 $\frac{1}{m}$, 第 i 个小区间的理论频数为 $u_i = \frac{n}{m}\ (i = 1, 2, \cdots, m)$, 则由 $Pearson\,\chi^2$ 拟合优度检验可知，统计量

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

   渐进服从 $\chi^2(m-1)$。给定显著水平 α,查 χ^2 分布表得临界值后,即可对经验频率与理论频率得差异作显著性检验。

2. K-S检验(柯氏检验)

   K-S($Kolmogorov-Smirnov$)检验是连续分布的拟合性检验。它检验样本的经验分布函数与总体的分布函数间的差异是否显著。设随机数为 $r_1, r_2, r_3, \cdots, r_n$,从小到大排列后得 $\hat{r}_1, \hat{r}_2, \hat{r}_3, \cdots, \hat{r}_n$。记经验分布函数为 $F_n(x)$,将 $F_n(x)$ 与均匀分布的分布函数 $F(x) = x$ 比较,其最大偏差即 $K - S$ 检验统计量为:

   $$V_2 = sup|F_n(x) - F(x)|$$

   利用 V_2 极限零分布进行显著检验（在R语言中可以用ks.test()函数进行检验）。

3. 序列检验

   序列检验实际上是用于多维分布地均匀性检验,它也间接地检验序列地独立

性。已知随机数序 $\{r_i\}$ $(i = 1, 2, \cdots, 2p)$,将容量为 $2p$ 的随机数依次配对为:

$$v_1 = (r_1, r_2), v_2 = (r_3, r_4), \cdots, v_p = (r_{2p-1}, r_{2p})$$

如果 $\{r_i\}$ $(i = 1, 2, \&, 2p)$ 是均匀随机数序列,那么它们应该构成平面上正方形内的二维均匀随机向量的样本。将单位正方形分成 k^2 个等面积的小正方形。 $\{n_{ij}\}$ 表示点 $\{v_i\}$ $(i = 1, 2, \cdots, 2p)$ 落入第 (i, j) 个小正方形的频数;理论频数 $u_{ij} = \frac{2p}{k^2}$ 。则检验统计量:

$$V_3 = \frac{k^2}{2p} \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} - \frac{2p}{k^2})^2$$

在 $\{r_i\}$ $(i = 1, 2, \cdots, 2p)$ 为均匀分布的独立抽样序列成立时渐进地服从 $\chi^2(k^2-1)$ (此检验也可以作为列联表检验来检验序列的独立性)。应该注意到当生成的随机数的数量较少时，所选的 k 值用该用较小值，否则会因为正方形分割得太细而使得一些区间没有随机数出现从而不能通过检验。而且这种检验方法不一定能通过，但在未能通过检验时比较统计量 V_3 的值也从一定程度上看出随机数生成器（或者算法）的好坏。

3.4.3 独立性检验

独立性检验主要是检验随机数序列 $r_1, r_2, r_3, \cdots, r_n$ 之间的统计相关性是否显著。下面介绍几种常见的检验方法。独立性检验中常用的方法有:相关系数检验I、相关系数检验II、游程检验。

1. 相关系数检验I

   两个随机变量的相关系数反映它们之间的线性相关程度, 故可以利用相关系数来检验随机数的独立性。设 $\{r_i\}$ $(i = 1, 2, \cdots, n)$ 为待检测的随机数组。原假设 H_0 : 相关系数 $\rho = 0$. 考虑样本的 j 阶自相关系数：

$$W_1(j) = \rho(j) = \frac{\frac{1}{n-j} \sum_{i=1}^{n-j} (r_i - \bar{r})(r_{i+j} - \bar{r})}{\frac{1}{n} \sum_{i=1}^{n-j} (r_i - \bar{r})^2}$$

当 $n - j$ 充分大,且 $\rho = 0$ 成立时 $W_1(j)$ 渐进服从 $N(0, 1)$ 分布, $(j = 1, 2, \cdots, m;$ 在

实际检验中,常取 $m = 10 \sim 20$)。

2. 相关系数检验II

$\rho(j)$ 还可以定义为：

$$\rho(j) = \frac{1}{n}\sum_{i=1}^{n}\frac{(r_i - \bar{r})(r_k - \bar{r})}{s^2} \ (j = 1, 2, \cdots, m)$$

其中， $k = (i + j)\ mod\ n, s^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2$。

记 $C_j = \frac{1}{n}\sum_{i=1}^{n} r_i r_k$， 则：

$$\rho(j) = \frac{(C_j - \bar{r}^2)}{s^2} \approx \frac{(C_j - \frac{1}{2})}{12}$$

易得 $E(C_j) = \frac{1}{4}, Var(C_j) = \frac{13}{144n}$。

这时 H_0 可以用检验假设 $\hat{H}_0 : E(C_j) = \frac{1}{4}$ 来代替。检验统计量为：

$$W_2(j) = \frac{C_j - \frac{1}{4}}{\sqrt{\frac{13}{144n}}} \sim N(0, 1)$$

3. 游程检验

游程检验是直接检验随机数序列 r_i 的独立性。对随机数序列 r_i ,把它分为许多子序列,使得其中每一个子序列内的值都是上升的,则称每个子序列为一个上升游程。例如以下13 个随机数:

$$0.12, 0.132, 0.67, 0.122, 0.113, 0.23, 0.14, 0.56, 0.78, 0.97, 0.98, 0.53, 0.31$$

可以分成5个上升游程,第一个游程长度为3,其余的上升游程长度分别为1,2,5,2。

首先统计游程长度为1,2,3,4,5和 $\geqslant 6$ 的游程数目,分别记为: $q_1, q_2, q_3, q_4, q_5, q_6$。

则统计量:

$$W_3 = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij}(q_i - nb_i)(q_j - nb_i)$$

渐进服从 $\chi^2(6)$ 分布，上式中 n 为样本容量， a_{ij} 为下列矩阵的各元素:

$$\begin{pmatrix} 4529.4 & 9044.9 & 13568 & 18091 & 22615 & 27892 \\ 9044.9 & 18097 & 27139 & 36187 & 45234 & 55789 \\ 13568 & 27139 & 40721 & 54281 & 67852 & 83685 \\ 18091 & 36187 & 54281 & 72414 & 90470 & 111580 \\ 22615 & 45234 & 67852 & 90470 & 113262 & 139476 \\ 27892 & 55789 & 83685 & 111580 & 139476 & 172860 \end{pmatrix}$$

b_i 是下列向量的各元素:

$$(\frac{1}{6}, \frac{5}{24}, \frac{11}{120}, \frac{19}{720}, \frac{29}{5040}, \frac{1}{840}) = (b_1, b_2, b_3, b_4, b_5, b_6)$$

行游程检验时,样本容量应大于400。当显著水平 α 为0.05时临界值为12.59。

上述游程检验是按上升游程进行检验的,也可按下降游程进行游程检验,即各个游程是按单调递减划分的,这时检验统计量的形式不变游程检验用来检验随机数的独立性。

序列检验是一项比较难通过的检验，即使每次对于对很好的随机数的生成进行检验，绝大部分都不能通过。（实验检验所得到的结果，原因未明）

3.5 Monte-Carlo方法实验检验

3.5.1 Monte-Carlo方法设计思想

Monte-Carlo方法的要点是：对要解决的数值计算问题，构造适当的概率模型，使要得到的解正好重合于概率模型中随机变量的概率分布或数字特征，其后在计算机上用伪随机数列对随机变量进行模拟得到一个大子样的观测数据,进行统计整理以后，给出问题的一个近似估计。因此，Monte-Carlo 方法是双重近似，一是将数值计算问题用概率模型作近似，二是在计算机上用伪随机数作近似抽样值进行统计整理作出一些估计。

3.5.2 Monte-Carlo方法实验原理

对生成伪随机数的方法进行比较,设计如下的数学实验: 设 $g(x)$ 是 $[0,1]$ 的连续函数，且 $0 \leqslant g(x) \leqslant 1$ 。考虑定积分在 $[0,1]$ 上的值。在单位正方形内，曲线 $y = g(x)$ 下面的阴影的面积 A 就是积分值 s。如果在边长为1的正方形内均匀地投点 (x, y)，则该随机点落入曲线 $y = g(x)$ 下阴影的概率为由此想法构造计算定积分的投点模型。

向正方形 $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$ 内均匀投点 (ξ_i, η_i), $\{\xi_i\}$, $\{\eta_i\}$ 是相互独立的均匀随机数列，第 i 次试验成功，即 (ξ_i, η_i) 落入 A 中，也就是满足 $\eta_i \leqslant g(\xi_i)$，若每次成功的概率为 p，进行 n 次试验成功了 k 次，则由大数定理知 $\lim\limits_{n\to\infty} \frac{k}{n} \overset{a.s}{=} p$.

3.6 实验设计

3.6.1 实验工具

利用R Version 3.2.5, RStudio Version 0.99.896 进行随机序列生成、经验检验、Monte-Carlo模拟。

3.6.2 实验过程

1. 生成随机数序列

   - 平方取中法
     取种子为测试时间： $x_1 = 1704$，生成来自平方取中法的随机序列，长度为4000.
   - 乘积取中法
     取种子为测试时间： $x_1 = 0528, x_2 = 1704$，生成来自乘积取中法的随机序列，长度为4000.
   - 移位法
     移位法 $\{x_i\}$的范围较大，故取种子为测试时间加上日期后一位：
     $x_1 = 81704$，生成来自移位法的随机序列，长度为4000.

- 混合同余法
  根据文献，找到一种比较好的参数设置：

$$x_1 = 1,\ a = 2^8 - 1,\ c = 28399,\ m = 454383$$

  取种子 $x_1 = 1$. 生成来自混合同余法的随机序列，长度为4000.

- 移位寄存器
  取给定的常数向量：

$$c = (1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1)$$

  给定初值：

$$(a_1, a_2, \cdots, a_{16}) = (0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1)$$

  生成来自移位寄存器的随机序列，长度为64000. 按照算法获得[0,1]的随机序列 $\{r_i\}$，长度为4000.

2. 经验检验
   依照上述方法对生成序列进行参数检验、均匀性检验、独立性检验，返回检验统计量及结果。

3. Monte-Carlo方法实验

   - 将生成序列中的奇数点序列命名 $\{\eta_i\}$，偶数点序列命名 $\{\xi_i\}$。
   - 进行Monte-Carlo方法实验。令曲线

$$g_1(x) = x^2,\ g_2(x) = x^3,\ g_3(x) = x^4$$

     故由定积分容易算出三条曲线与 $x = 0,\ x = 1,\ y = 0$ 围成的阴影面积分别 $\frac{1}{3},\ \frac{1}{4},\ \frac{1}{5}$.

- 计算落入阴影面积的 (η_i, ξ_i) 的个数 k_1, k_2, k_3，计算相对偏差：

$$\sigma_1 = \frac{|\frac{k_1}{n} - \frac{1}{3}|}{\frac{1}{3}}$$
$$\sigma_2 = \frac{|\frac{k_2}{n} - \frac{1}{4}|}{\frac{1}{4}}$$
$$\sigma_3 = \frac{|\frac{k_3}{n} - \frac{1}{5}|}{\frac{1}{5}}$$

3.7 实验结果及分析

3.7.1 平方取中法

Monte-Carlo模拟实验结果图中只画出了 $g(x) = x^2$ 的曲线，下同。

```
> fcs(x)
$开始检验
[1] "!"

$参数1检验
$参数1检验[[1]]
[1] "参数检验未通过"

$参数2检验
$参数2检验[[1]]
[1] "参数检验未通过"

$参数3检验
$参数3检验[[1]]
[1] "参数检验未通过"
```

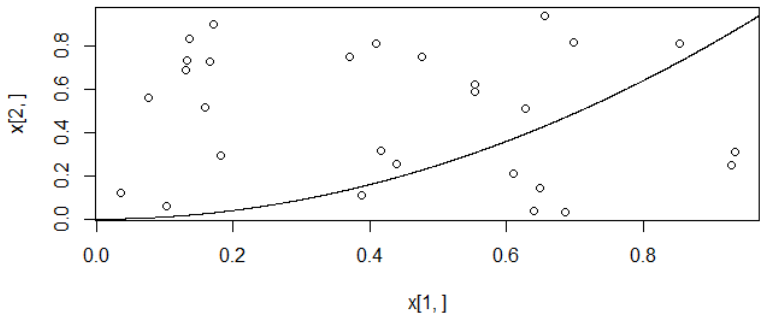
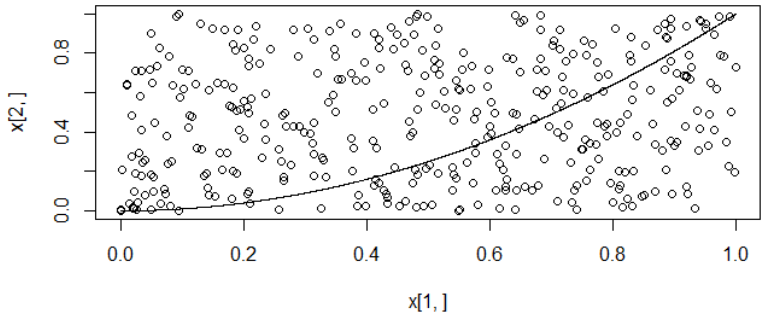


图 1: 平方取中法的检验结果及Monte-Carlo模拟实验结果

由Monte-Carlo检验方法画出的图可以看出，平方取中法生成随机数周期太短，因为图中散点只有很少的一些，而实际上在同一个地方画出了很多点。值得注意的是这种情形在种子数的选取较好时可以稍微变好，可是这种算法仍然表现很差。另

外其参数检验没通过，可以看出其性质很差。故不再继续讨论。

3.7.2 乘积取中法




```
> fcs(x2)
$开始检验
[1] "!"

$参数1检验
$参数1检验[[1]]
[1] "参数检验未通过"


$参数2检验
$参数2检验[[1]]
[1] "参数检验未通过"


$参数3检验
$参数3检验[[1]]
[1] "参数检验未通过"
```

图 2: 乘积取中法的检验结果及Monte-Carlo模拟实验结果

由Monte-Carlo检验方法画出图形容易看出乘积取中法生成的随机数的均匀性有所提高，且周期变长了，但参数检验仍然不能通过，故性质仍然不太好。值得注意的是其周期之于平方取中法之所以增长了是因为每一迭代新增了元素造成。

3.7.3 移位法

参数检验：

```
> fcs(x3)
$开始检验
[1] "!"

$参数1检验
$参数1检验[[1]]
[1] "参数检验通过"


$参数2检验
$参数2检验[[1]]
[1] "参数检验通过"


$参数3检验
$参数3检验[[1]]
[1] "参数检验通过"
```

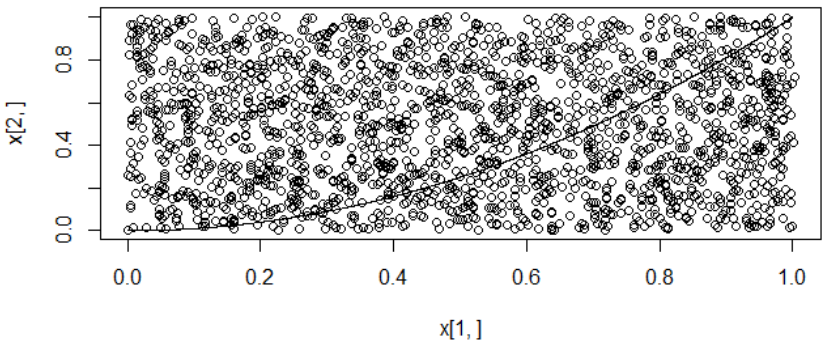
独立性检验：

```
> tr(x3,15)
$结果
[1] "通过检验"

$样本值
[1] 0.04565504

$标准正态的0.05上分位数
[1] 1.644854

>
```

均匀性检验：

```
> fjy(x3,10,4)
$均匀性的卡方检验结果
$均匀性的卡方检验结果$样本卡方参数估计
[1] 4.735

$均匀性的卡方检验结果$理论参数
[1] 16.91898

$均匀性的卡方检验结果$理论p值
[1] 0.8569264

$均匀性的卡方检验结果[[4]]
[1] "通过检验"


$柯式检验结果
$柯式检验结果$p值
[1] 0.7917205

$柯式检验结果[[2]]
[1] "通过K-S检验"


$序列检验的结果
$序列检验的结果$样本参数估计
[1] 12.98

$序列检验的结果$理论参数
[1] 24.99579

$序列检验的结果[[3]]
[1] "通过检验"
```

Monte-Carlo模拟：



```
> fmonte(x3)
$比较结果为
$比较结果为$第一个积分的积分相对误差
[1] 0.062

$比较结果为$第二个积分的积分相对误差
[1] 0.058

$比较结果为$第三个积分的积分相对误差差
[1] 0.055
```

图 3: 移位法Monte-Carlo模拟实验结果

首先由Monte-Carlo检验方法画出的图可以看出其周期比较长，且分布均匀，独立性在所测试的四种方法中是最好，其独立性检验有最小值0.04566（四舍五入）。而且Monte-Carlo模拟效果也比较好，相对误差在5%左右,我们用Monte-Carlo模拟估计积分的方法来综合比较随机数算法的好坏，故从移位法最后一项实验输出可看出其效果还是不错的。这儿我们只用了4000个数来耦合了2000个点，当用更多的数来耦合时，相对误差将会减小（我们已经经过计算实验验证，此篇论文没有列出当随机数增多时的效果）。

3.7.4 混合同余法

参数检验：

```
> fcs(x4)
$开始检验
[1] "!"

$参数1检验
$参数1检验[[1]]
[1] "参数检验通过"


$参数2检验
$参数2检验[[1]]
[1] "参数检验通过"


$参数3检验
$参数3检验[[1]]
[1] "参数检验通过"
```

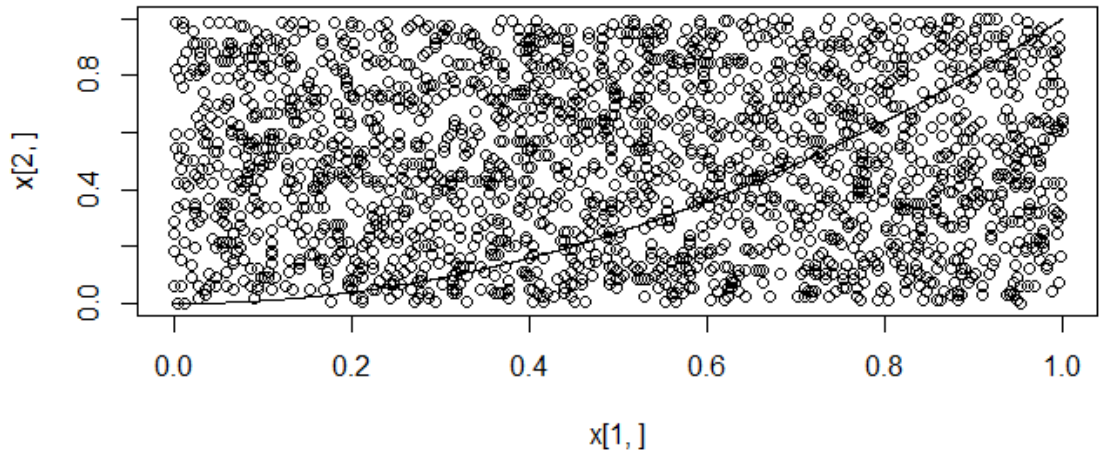
独立性检验：

```
> fr(x4,15)
$结果
[1] "通过检验"

$样本值
[1] 0.3932017

$标准正态的0.05上分位数
[1] 1.644854
```

均匀性检验：

```
> fjy(x4,10,6)
$均匀性的卡方检验结果
$均匀性的卡方检验结果$样本卡方参数估计
[1] 7.565

$均匀性的卡方检验结果$理论参数
[1] 16.91898

$均匀性的卡方检验结果$理论p值
[1] 0.578506

$均匀性的卡方检验结果[[4]]
[1] "通过检验"


$柯式检验结果
$柯式检验结果$p值
[1] 0.8294535

$柯式检验结果[[2]]
[1] "通过K-S检验"


$序列检验的结果
[1] "未通过均匀性的序列检验"
```

Monte-Carlo模拟：




```
> fmonte(x4)
$比较结果为
$比较结果为$第一个积分的积分相对误差
[1] 0.064

$比较结果为$第二个积分的积分相对误差
[1] 0.074

$比较结果为$第三个积分的积分相对误差差
[1] 0.0875
```

图 4: 混合同余法的Monte-Carlo模拟实验结果

混合同余法生成的随机数（4000个数）各方面性质都比较好:同样由Monte-Carlo检验方法画出的图可以看出周期比较长，且比较均匀，其各个检验统计量虽然没有最突出的表现，但在Monte-Carlo模拟方面也会随随机数个数的增多表现出越来越小的相对误差,，故用来估计积分时，线性同余法也是很好的，但还是没有反馈移位寄存器表现好。其各方面统计检验值如图所示。

3.7.5 反馈移位寄存器

参数检验：

```
> fcs(x5)
$开始检验
[1] "!"

$参数1检验
$参数1检验[[1]]
[1] "参数检验通过"


$参数2检验
$参数2检验[[1]]
[1] "参数检验通过"


$参数3检验
$参数3检验[[1]]
[1] "参数检验通过"
```

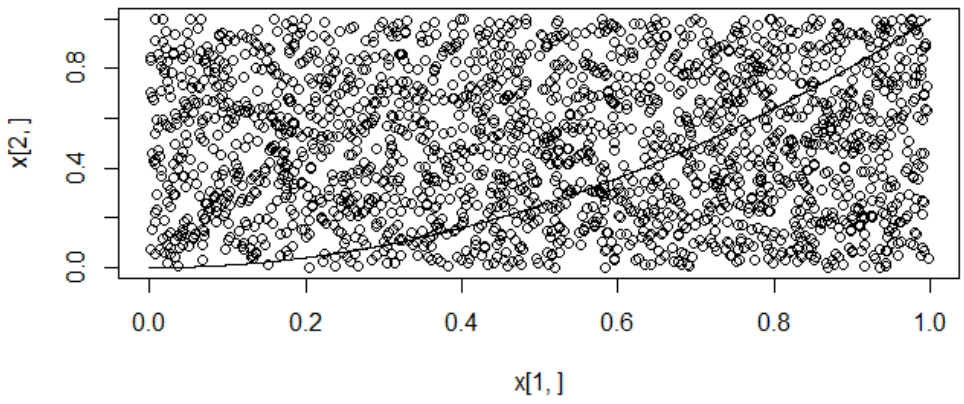
独立性检验：

```
> fr(x5,15)
$结果
[1] "通过检验"

$样本值
[1] 0.4973203

$标准正态的0.05上分位数
[1] 1.644854
```

均匀性检验：

```
> fjy(x5,10,10)
$均匀性的卡方检验结果
$均匀性的卡方检验结果$样本卡方参数估计
[1] 2.3

$均匀性的卡方检验结果$理论参数
[1] 16.91898

$均匀性的卡方检验结果$理论p值
[1] 0.9857876

$均匀性的卡方检验结果[[4]]
[1] "通过检验"


$柯式检验结果
$柯式检验结果$p值
[1] 0.9186192

$柯式检验结果[[2]]
[1] "通过K-S检验"


$序列检验的结果
[1] "未通过均匀性的序列检验"
```

Monte-Carlo模拟：



```
> fmonte(x5)
$比较结果为
$比较结果为$第一个积分的积分相对误差
[1] 0.0065

$比较结果为$第二个积分的积分相对误差
[1] 0

$比较结果为$第三个积分的积分相对误差差
[1] 0.0225
```

图 5: FSR的Monte-Carlo模拟实验结果

首先由Monte-Carlo检验方法画出的图可以看出产生随机数的周期比较长，由均匀性检验知道其均匀性是最好的，虽然只是一次生成效果的比较，但效果比较明显，样本的P值极高，独立性也比较好。在以上所述的所有方法中，移位寄存器的方法是Monte-Carlo模拟结果最好的,它在每一个积分的相对误差都是很小的，误差第一个比其他两个小了一个数量级，第二个为零（这很大程度上是偶然因素，但无法否认效果是在更好的趋势上的），第三个误差比其他的要小很多（相对而言），我们认为Monte-Carlo模拟效果好的原因也许是其分布的均匀性，因为此次随机数组的检验的均匀性统计检验量的效果最好，意味着其分布很均匀，换句话说就是在相等面积上的点个数差不多，这也许就是为什么其能更精确估计面积积分，这儿之所以用“也许”描述均匀性对积分的估计的作用是因为我们认为独立性、还有随机数列周期性也会影响积分，因为若周期小，相关性强，则我们耦合的点容易聚集于第一象限的对角线上，这也会影响对积分的估计。

4 总结

4.1 实验结果总结

实验对5种较为常见的伪随机数发生器进行了检验。

从实验结果可以看出，历史上最先提出的平方取中法是效果相对较差的一种方法，该算法中包含乘法，运算速度不如移位法等，而且有周期非常明显，容易产生退化等缺点。而作为改进的乘积取中法确实在周期上有所改善，但仍然不尽人意，从实验结果可以看出，与后来的方法比较起来，生成的随机序列是无法满足我们的需求的。

对于移位法，这是一种计算效率较高的方法，而且生成的随机序列效果也令人满意，相对于其他方法，移位法的独立性检验结果是非常优秀的，是一个值得使用的伪随机数发生器。

混合同余法是目前较为流行的一种伪随机数发生器，如我们较为熟知的C++中得 *rand*() 函数就是用的混合同余法，也说明了该算法的效果被广大使用者所接受。但是在我们的测试中发现，混合同余法对参数的选择有较高的要求，选定一个让周期尽量大的参数设置显然是保证混合同余法效果的关键。

反馈移位寄存器(FSR)是本次实验中表现最好的方法。主要体现在其算法运行效率高、生成随机序列的独立性、均匀性都是非常好的。最后的Monte-Carlo模拟的效果也再次证实了我们的结论，是最值得关注的一种伪随机数发生器。

4.2 需要改进之处

1. 在选择进行检验及实验的算法时并未将之前陈述的所有算法进行分析。在时间充足的情况下，应对各个伪随机数发生器进行检验及实验模拟。

2. 在对各个伪随机数发生器进行实验时，我们对一些发生器取测试时间为种子，而各个发生器对种子的要求并不都是相同的，所以对某些个别的发生器，我

们为了使其效果尽量好，调整了参数和种子。但这样可能会影响结论的严谨性，应继续分析各个随机数发生器对参数及种子的灵敏度。

3. 序列检验的失效原因未知，在时间充足的情况下应多查阅文献，了解原因，改进调整该检验方法。

参考文献

[1] 杨振海&张国志. 随机数生成[J].数理统计与管理,2006,25(2):244-252

[2] 金畅.蒙特卡洛方法中随机数发生器和随机抽样方法的研究[D].大连:大连理工大学,2015.

[3] 张广强.均匀随机数发生器的研究和统计检验。[D].大连:大连理工大学，2005.

[4] 李存华.伪随机数及其生成.[J].淮海工学院学报(自然科学版),1997,6(2):10-17

[5] 郑列&宋正义.伪随机数生成算法及比较[J].湖北工业大学学报,2008,23(5):65-88

[6] 王水花&张煜东&吴乐南.伪随机数与准随机数的比较[J].计算机与信息技术:32-36

[7] 张志军&杨育捷&段新涛.一中紧凑高效LFSR结构的均匀分布随机数产生器[J].新乡学院学报,2014,31(6):16-18.

[8] 赵雪峰.一种伪随机数生成算法的研究与实现[N].电脑学习,2005-12(6).

[9] 张咏.随机数发生器和随机数性能检验方法研究[D].电子科技大学,2006.

[10] 谭阳.关于随机数生成算法的研究[D].湖南,湖南师范大学,2008.