

伪随机数随机性的一种新检验

时正华 袁永生

(河海大学理学院, 江苏 南京 210098)

摘要: 伪随机数通过的随机性检验越多, 说明其随机性越好. 利用生产的随机数精度固定这一特点, 根据小数点后面 n 个数字之和, 提出了尾数和的概念, 并给出了求尾数和分布频数的算法. 根据尾数和分布概率提出了一种检验伪随机数的新方法. 当尾数较长时, 尾数和的分布服从渐近正态分布. 求出其期望和方差, 并给出了求尾数和分布概率的近似估计公式.

关键词: 随机数; 尾数和; χ^2 检验

中图分类号: O211.6 **文献标识码:** A **文章编号:** 1000-1980(2005)02-0232-05

Lehmer 把用数学方法产生的随机数定名为“伪随机数”. 数学方法产生的伪随机数是通过迭代过程得到的, 只要它们通过了一系列的统计检验, 那么就可以当作“真正”的随机数来使用^[1]. 伪随机数序列的随机性常常需要采用多种不同的检验方法来检验, 通过的检验愈多, 愈能说明所产生的伪随机数与真正随机数的性质和规律不矛盾^[1~3].

随机性检验常用的组合规律性检验是将所得伪随机数按某种方式组合, 并在虚假设 H_0 (若各伪随机数独立地来自均匀分布, 各类组合的出现有一定的概率)成立的条件下, 比较各类组合的观测频数与理论频数之差是否显著的一种检验方法^[1]. 所以, 如果找到了随机数的一种新的组合方式, 也就找到了一种新的组合规律性检验.

本文在深入分析 $[0, 1]$ 上均匀分布的随机抽样所应具有的各种性质和规律之后, 给出了随机数尾数和的概念, 提出了组合规律性检验的一种新方法——尾数和分布检验法. 尾数和分布检验法利用产生的伪随机数尾数和的分布规律来检验伪随机数的组合规律性和均匀性.

1 尾 数 和

1.1 定义

设 $U(0, 1)$ 上所产生的随机数小数点后面有 n 位, 则小数点后面 n 个数字之和称为这个随机数的尾数和. 这里 n 记为尾数长度.

理论上讲, $U(0, 1)$ 产生的随机数可以有无数个, 但计算机产生的随机数有一定精度, 当精度给定时, 尾数长度是确定的. 当尾数长度给定时, 产生的随机数总数实际上是有限的, 并且这些有限的总数根据尾数和的大小可以进行重新组合分类.

1.2 性质

在取尾数长度为 n 的情况下, 由 $U(0, 1)$ 产生的随机数其尾数和的取值范围为 0 到 $(9 \times n)$ 的整数.

性质的证明显然. 这里需要注意的是: 产生的随机数可能为 1, 因为均匀分布是连续分布, $P(X=1)=0$, 把 $X=1$ 规定为尾数和为 0 的情况, 则在取尾数长度为 n 时, 随机数的总体个数为 10^n 个.

根据上面的性质对产生的随机数进行了一次划分, 只要能求出每个尾数和的理论频数就可以进行随机性检验. 检验方法一般采用 χ^2 检验.

2 χ^2 检验^[1 4]

设将产生的 N 个伪随机数按尾数和的大小分为 r 组, n_i ($i = 1, 2, \dots, r$) 表示第 i 组的观测频数. 虚假设 H_0 是观测结果落在第 i 组的概率 p_i ($\sum_{i=1}^r p_i = 1$), 记 $m_i = Np_i$ 为第 i 组的理论频数. 检验虚假设 H_0 的目的在于比较观测频数 n_i 与理论频数 m_i 的差异是否显著.

a. 统计量的选取:

$$T = \sum_{i=1}^r \frac{(n_i - m_i)^2}{m_i} = \sum_{i=1}^r \frac{(n_i - Np_i)^2}{Np_i}$$

b. 在虚假设 H_0 为真的假定下, T 近似服从 $\chi^2(r - 1)$. 根据显著性水平 α 查 χ^2 表, 当 $T \geq \chi^2_{\alpha}(r - 1)$ 时, 虚假设 H_0 不成立.

3 不同尾数和理论频数的算法

设尾数长度为 n , 当尾数和值 (令为 k) 较小, 如 $k = 0$ 时, 对应的理论频数 $m_0 = 1$, $k = 1$, $m_1 = n$, 理论频数 m_i 比较好求. 但随着尾数和的逐渐增大, $k = \sum_{i=1}^n a_i$ 的组合越来越多 (a_i 属于 0 到 9 的整数, 表示小数点后的第 i 位数), 即使能给出公式 (比如组合论中的反演公式), 公式也极其复杂而没有实际应用价值. 由前面的性质可知, 给定尾数长度 n 时, 随机数可以看作是总体个数为 10^n 的点 (每个点代表 1 个长度为 $1/10^n$ 的区间), 由均匀分布的性质, 可知产生这 10^n 个数是可能的, 由此完全可以使用计算机利用穷举法表示出随机数的所有情况, 并根据尾数和的大小进行分类, 从而统计出每个尾数和的理论频数.

程序简单描述如下: 用 n 个 For 循环来模拟尾数, 每个循环变量 $k[i]$ 代表尾数的第 i 个数字 ($i = 1, 2, \dots, n$), $k[i]$ 取值为 0 到 9 的整数, 求出 $m = \sum_{i=1}^n k[i]$, 根据 m 的大小进行分类 (m 的取值范围为 0 到 $9 \times n$ 的整数), 求出每个分类的个数, 即理论频数. 尾数长度 $n = 5$ 时尾数和理论频数计算结果见表 1.

有了表 1, 对于产生的尾数长度为 5 的伪随机数, 只要利用计算机编程 (与求理论频数类似) 求出尾数和相应的观测频数, 然后利用 χ^2 检验, 就可以检验伪随机数的随机性.

这里需要注意, χ^2 检验通常要求把频数比较小的组并成一个频数较大的组, 如尾数和为 0, 1, 2, 3, 4, 5 的可以合并为一组, 尾数和为 40, 41, 42, 43, 44, 45 的可以合成一个组.

由上可见 “尾数和法” 检验伪随机数的随机性步骤为:

- a. 对产生的伪随机数确定其尾数长度 n ;
- b. 用计算机统计出各个尾数和的观测频数;
- c. 查表 (或通过计算机计算) 找出尾数长度为 n 的各个尾数和的理论频数;
- d. 根据尾数和和理论频数的大小对伪随机数进行分组, 求出统计量 T , 设定一个显著性水平 α , 进行 χ^2 检验^[1 2].

4 尾数和的分布规律和 n 较大时尾数和理论频数的近似估计

求尾数和的理论频数表虽然程序简单, 但随着 n 值的提高, 随机数的个数随 n 成指数级 (10^n) 增加, n 较大时, 计算量是惊人的, 当没有像表 1 那样的数据表时, 对于尾数长度较大的随机数, 利用尾数和的分布规

表 1 尾数长度 $n = 5$ 时尾数和分布情况

Table 1 Distribution of the sum of mantissa for $n = 5$

k	m	k	m	k	m
0	1	16	3 795	32	2 205
1	5	17	4 335	33	1 745
2	15	18	4 840	34	1 340
3	35	19	5 280	35	996
4	70	20	5 631	36	715
5	126	21	5 875	37	495
6	210	22	6 000	38	330
7	330	23	6 000	39	210
8	495	24	5 875	40	126
9	715	25	5 631	41	70
10	996	26	5 280	42	35
11	1 340	27	4 840	43	15
12	1 745	28	4 335	44	5
13	2 205	29	3 795	45	1
14	2 710	30	3 246		
15	3 246	31	2 710		

注: k 代表 “尾数和”, m 代表 “尾数和为 k 时的理论频数”.

律可以得到近似的理论频数(理论频数 = 分布概率 × 10ⁿ,下文为数据处理的方便,用分布概率来代替理论频数).为了研究其规律,作表1中 k 和 $m/10^n$ (即分布概率)的散点图.

经分析与比较,散点图的连线是一个近似的正态分布,可用曲线

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

(1)

来拟合图1, μ 和 σ^2 为参数,参数大小由 n 决定.利用非线性最小二乘法求参数估计,得 $\hat{\mu} = 22.5$, $\hat{\sigma}^2 = 43.4911$,其中 $f(k)$ 的值与对应的 $P(k)$ 值($P(k)$ 表示尾数和为 k 时的理论概率)的最大绝对误差小于 0.001,显然拟合效果非常好.

为了对比,求出表1中 k 的期望值和理论方差.由表1,易得

$$E(k) = 22.5 = 4.5 \times 5$$

$$D(k) = 41.25 = 8.25 \times 5$$

期望值相等,理论方差不等但差不多.为了寻找规律,表2列出了 $n = 2, 3, 4, 5, 6, 7, 8, 9$ 时尾数和分布的理论期望值、理论方差及在模型(1)拟合相应各 n 下尾数和分布概率时参数 μ 和 σ^2 的估计值.

表2 部分 n 值下尾数和分布参数的理论值与拟合值

Table 2 Some theoretical-value and fitting-value of the distribution parameters of the sum of mantissa for different values of n

n	最大尾数和	期望值	理论方差	μ 的估计值 (拟合期望)	σ^2 的估计值 (拟合方差)
2	18	9.0	16.50	9.0	19.40319423
3	27	13.5	24.75	13.5	27.16003323
4	36	18.0	33.00	18.0	35.28534054
5	45	22.5	41.25	22.5	43.49109220
6	54	27.0	49.50	27.0	51.71639065
7	63	31.5	57.75	31.5	59.94958264
8	72	36.0	66.00	36.0	68.18719149
9	81	40.5	74.25	40.5	76.42765219

显然

期望值 = 拟合期望值 = $n \times 4.5$

(2)

作“ n 值与理论方差”和“ n 值与拟合方差(σ^2 的估计)”的散点图,见图2.

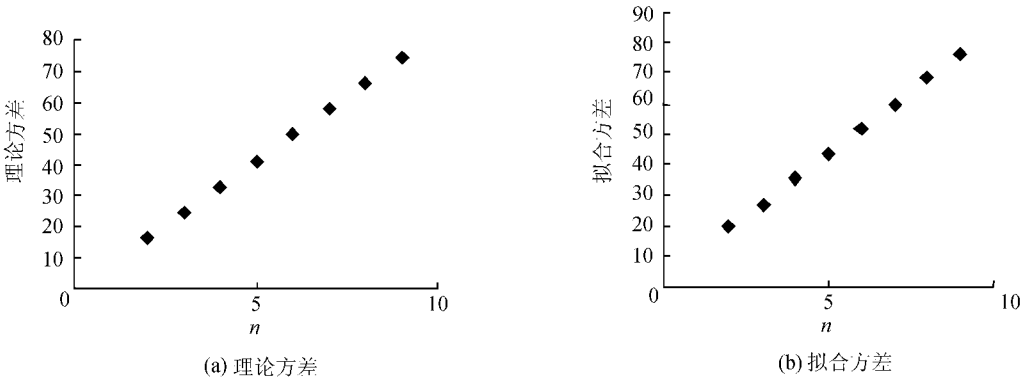


图2 方差和 n 的关系

Fig.2 Relation of n and variance

图2(a)是一个完全的线性关系,其方程为

理论方差 = $8.25 \times n$

(3)

图2(b)虽然看上去像一条直线,却不是一个完全的线性关系,而是近似的线性关系,其回归方程为

拟合方差 = $2.75 + 8.17 \times n + e$

(4)

式中 e 为残差.

注意到拟合方差和理论方差的绝对距离越来越小 ,对表 2 进行处理 ,见表 3.

表 3 理论方差、拟合方差与 n 的关系

Table 3 Relation between n and variances of theoretical-value and fitting-value

n	理论方差	拟合方差 (记为 S 列)	理论方差一阶差分 (记为 A 列)	拟合方差一阶差分 (记为 T 列)	拟合方差 - 理论方差 (记为 Q 列)
2	16.50	19.403 194 23			2.903 194
3	24.75	27.160 033 23	8.25	7.756 839	2.410 033
4	33.00	35.285 340 54	8.25	8.125 307	2.285 341
5	41.25	43.491 092 20	8.25	8.205 752	2.241 092
6	49.50	51.716 390 65	8.25	8.225 298	2.216 391
7	57.75	59.949 582 64	8.25	8.233 192	2.199 583
8	66.00	68.187 191 49	8.25	8.237 609	2.187 191
9	74.25	76.427 652 19	8.25	8.240 461	2.177 652

n 列和 T 列、 n 列与 Q 列的散点图见图 3、图 4.

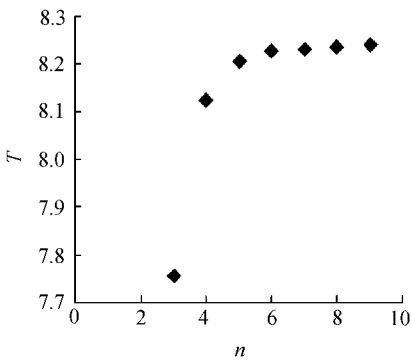


图 3 n 和 T 的关系

Fig.3 Relation of n and T

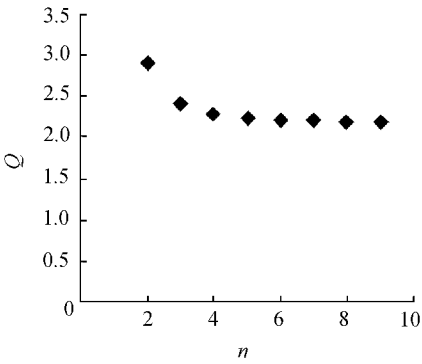


图 4 n 和 Q 的关系

Fig.4 Relation of n and Q

图 3 显示 ,随着 n 值的不断增大 ,拟合方差的变化(即 T 值)趋于稳定 ,在 8.24 左右.图 4 显示 ,随着 n 值的不断增大 ,拟合方差 - 理论方差的值(即 Q 值)趋于稳定 ,在 2.17 左右.

记 $S(i)$ 表示 $n=i$ 时 S 列(拟合方差)的值. $\pi(i)$ 和 $Q(i)$ 意思类似 ,而由于 $S(i)=8.25i+Q(i)$,所以

$$S(i)-S(i-1)=\pi(i)=8.25+Q(i)-Q(i-1)\xrightarrow{Q\text{ 值随 }i\text{ 趋于常数 }i\rightarrow\infty}8.25$$

(5)

由式 (5) 可以得到拟合方差 D 的近似公式

$$S(i)=S(i-1)+\pi(i)=8.25(i-1)+Q(i-1)+\pi(i)$$

(6)

其中 n 较大时(如大于 10) , $\pi(i)$ 可用 8.24 来代替 , $Q(i)$ 可用 2.17 来代替.

由于 $n=9$ 时 , $S(9)>8.24$,在 $S(i-1)$ 已知的情况下 , $S(i)$ 近似估计的误差小于 0.01 ;当 $S(i-1)$ 未知 , $Q(i)$ 虽然极限不能准确地求出 ,但根据 $Q(i)$ 的变化趋势 ,可知极限在 2.16 与 2.14 之间.当 $Q(i)$ 用 2.17 代替时 ,由式 (6) 可以推断出 $S(i)$ 近似估计的误差应该小于 0.03.

由上可见 ,由式 (6) 可以得到 $\hat{\sigma}^2$,利用式 (2) 求得拟合期望 $\hat{\mu}$,再利用模型 (1) 可以求出尾数长度较大时各个尾数和值的近似分布概率 ,然后再把概率转化成频数

$$\text{理论频数}=\text{分布概率}\times 10^n=f(k)\times 10^n$$

就可以进行 χ^2 检验.这里需要注意的是 :用模型 (1) 对概率进行估计误差虽然较小 ,但由于个别尾数和概率(或理论频数)相对很小且尾数和的取值比较多 ,检验时可以根据需要把不同大小的尾数和合并 ,合并后各个组合的概率最好大于 0.05.

5 结 语

本文为检验随机数的随机性提供了一种新的方法 ,并利用该方法对大量随机数^[5 6]进行了检验.检验结果显示 ,尾数和法与均匀性检验、累积频率检验、分拆检验结果一致.而由于尾数和法利用的是一种新的组合

规律性,对于肯定伪随机数的“随机性”又提供了一道“保证”。

参考文献:

- [1] 徐钟济. 蒙特卡罗方法[M]. 上海: 上海科学技术出版社, 1985. 60—97.
- [2] 蔡坤宝. 均匀分布随机数的一种综合评估法[J]. 重庆大学学报, 1996, 20(4): 70—75.
- [3] HAMMERSLEY J M, CHANDSCOMB. Monte carlo method[M]. London: Spottiswoode, Ballantyne & Co Ltd, 1964. 2—10.
- [4] 高惠旋. 统计计算[M]. 北京: 北京大学出版社, 1995. 109—119.
- [5] 杨自强, 魏公毅. 产生伪随机数的若干新方法[J]. 数值计算与计算机应用, 2001, 23(3): 202—216.
- [6] 张传林, 林立东. 伪随机数发生器及应用[J]. 数值计算与计算机应用, 2002, 23(3): 188—208.

A new test on randomness of pseudo-random numbers

SHI Zheng-hua, YUAN Yong-sheng

(College of Sciences, Hohai Univ., Nanjing 210098, China)

Abstract The more randomness tests the pseudo-random numbers pass, the better the randomness of the pseudo-random numbers. In consideration of the fact that the precision of random numbers produced is fixed, the concept of the sum of mantissa is proposed based on the sum of n numbers behind the decimal, and an algorithm for calculation of the distribution frequency of the sum of mantissa is developed. With the distribution probability of the sum of mantissa, a new method for testing the randomness of pseudo-random numbers is put forward. It is found that the distribution of the sum of mantissa is an approximately normal distribution if the mantissa is long. By derivation of the expectation and variance of the sum of mantissa, an approximate estimation formula for the distribution probability of the sum of mantissa is given.

Key words random number; sum of mantissa; χ^2 test

《河海大学学报(自然科学版)》征订启事

《河海大学学报(自然科学版)》是以水资源开发、利用与保护为重点的综合性学术期刊,主要刊登本校在水资源、水文、地质、测量、水利工程、水电工程、水运工程、海洋及海岸工程、水工结构、工程力学、水力学及河流动力学、岩土工程、计算机科学、电力工程、电子技术及自动化工程、工业与民用建筑、管理工程、水利经济、环境工程、机械工程等学科方面的科研成果、学术论文、学术讨论、研究动态等学术性文章,可供上述有关专业的科技工作者及大专院校师生阅读和参考。

本刊创办于1957年,是我国中文核心期刊,在国内工程技术界和学术界有较大影响。刊载的文章中,有不少国家科技攻关(重点)项目和各种科学基金资助项目的研究成果,部分达到了国内领先和国际先进水平,为我国水利、水电、水运工程及其他有关工程建设的规划、设计、施工和管理提供了科学理论、方法和具体建议,发挥了较大的社会效益和经济效益,深受工程界和科技界赞许,并获得全国高校科技期刊优秀编辑出版质量奖,以及中国期刊方阵“双效期刊”、江苏省优秀期刊、全国水利系统优秀期刊称号。

本刊每逢单月出版,国内外公开发行。邮发代号28-63,每本定价3.00元,全年订费48.00元。欢迎广大读者在全国各地邮局订阅或直接与编辑部联系。联系地址:南京市西康路1号《河海大学学报(自然科学版)》编辑部,邮政编码210098。