文章编号:1001-9081(2015)S2-0038-03

非均匀分布下生日问题的模拟研究

刘天昕[1*],耿凤杰[2],赵俊芳[2]
(1.北京科技大学 计算机与通信工程学院,北京 100083; 2.中国地质大学(北京) 数理学院,北京 100083)
(*通信作者电子邮箱 ltx_fly@163.com)

摘 要:生日问题的解是在均匀分布的假设下给出的,其原理用于密码学中的生日攻击。实际生日分布是非均匀的,有必要研究与均匀分布下生日概率解的定量差别。利用蒙特卡罗模拟方法研究了非均匀生日分布下的生日概率问题。在生日分布为余弦变化的模型下模拟计算了生日概率,并将结果与均匀分布的解析解作了对比分析。模拟结果表明,即使在20%和10%的非均匀度下,二者与均匀分布的结果仍然非常接近,最大差别分别小于0.008和0.003;但在100%极端非均匀度下,最大差别可达0.15。该模拟算法可推广应用于复杂概率问题的数值计算,利用计算机模拟可以给出较为精确的结果。

关键词:生日问题;生日攻击;模拟计算;非均匀分布;概率计算
中图分类号:TP301.6 **文献标志码**:A

Simulation study on birthday problem in non-uniform distribution

LIU Tianxin[1*], GENG Fengjie[2], ZHAO Junfang[2]
(1. *School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*;
2. *School of Science, China University of Geosciences(Beijing), Beijing* 100083, *China*)

Abstract: The birthday problem is solved under the assumption of uniform distribution, and its principle is used as the birthday attack in cryptography. It is necessary to quantitatively compare the solutions when real birthday distribution is not uniform. The birthday problem in non-uniform distribution was studied using the Monte Carlo simulation. In a postulated cosine-distributed model, the birthday probabilities were computed and compared to the analytical solution in uniform distribution. The simulation results show that even at the levels of 20% and 10% non-uniformity, the probabilities are still very close to the uniform situation and the maximum deviations are 0.008 and 0.003, respectively; At the special case of 100% non-uniformity, the maximum deviation reaches 0.15. The simulation algorithm can be used in the numerical calculations of complicated probability problems, and results with high precision can be obtained using computer simulation.

Key words: birthday problem; birthday attack; simulation computing; non-uniform distribution; probability calculation

0 引言

生日问题起源于一个数学问题:随机找 n 个人($2 \leqslant n \leqslant 365$),至少有两人生日相同的概率 p 有多大?这个问题之所以出名,就在于答案出乎大多数人的意料。当 $n = 50$ 时,$p = 97\%$,这意味着在任何一个有50名学生的班级中,几乎就有2名同学的生日相同! 类似生日问题由Richard von Mises于1939年最早提出[1],之后有许多相关的推广与应用[2-3],如计算机领域中密码学里的生日攻击[4-5],用以减小哈希函数(Hash function)的冲突率[6-9]。

生日问题属于古典概率问题,有严格解析解,其求解过程不需要高深的数学知识。先考虑 n 个人生日都不相同的概率:假设一年365天(有闰年情况的讨论见文献[10]),所有人的生日都是独立的(不考虑双胞胎等特殊情况)并且是完全随机的,那么第2个人与第1个人生日不同的概率是364/365,第3个人和前面2个人生日都不同的概率是363/365,依此类推,最后1个人和前面所有人生日都不同的概率是$(365-n+1)/365$,将这些数字相乘就得到 n 个人生日都不相同的概率。令 $N = 365$,由此可得出 n 个人中至少有2人生日相同的概率公式:

$$p(n) = 1 - \frac{N!}{N^n(N-n)!} = 1 - \exp\left(\sum_{i=1}^{n}\ln(N-i+1) - n\ln N\right) \quad (1)$$

当 $n = 23$ 时,$p = 50.7\%$;当 $n \to N$ 时,$p \to 1$;当 $n > N$ 时,由抽屉原理可知,$p = 1$。式(1)的计算结果见图1,由于 $n > 72$ 时 p 的值已非常接近1,后面的部分不再画出。

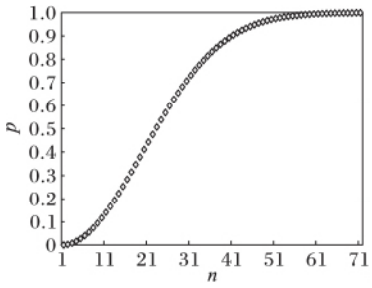


图1 不同人数 n 下至少2人生日相同的概率 p

特别需要指出的是,式(1)是在生日分布为均匀分布的条件下给出的。事实上,一个地区,一个国家,乃至整个世界,出生日

收稿日期:2015-04-27;修回日期:2015-06-10。
基金项目:国家自然科学基金资助项目(11202192);中央高校基本业务经费资助项目(2652012141)。
作者简介:刘天昕(1994-),女,北京人,主要研究方向:计算机仿真、数据库; 耿凤杰(1979-),女,河北保定人,副教授,博士,主要研究方向:数学与应用、概率论; 赵俊芳(1982-),女,河北邯郸人,副教授,博士,主要研究方向:数学与应用、概率论。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

期并非是完全随机的,由于风俗和文化的原因,在某些天或月份偏离均匀分布的峰值可能高达 15%(参见文献[11]的统计数据)。Bloom 于 1973 年一般地证明了,在各种生日分布中,均匀分布的生日概率最小[12]。也就是说,实际的生日概率曲线要高于图 1 的曲线,至于高出多少,取决于实际的生日分布。

当出生日期不是均匀分布时,生日概率问题如何计算?它们与均匀分布的定量差别有多大?利用传统方法求解上述问题将变得非常困难甚至不可能。随着计算机技术的发展,利用模拟方法求解类似问题变得相对容易。计算机模拟,又称为统计实验法,直接实验法,或蒙特卡罗(Monte Carlo)方法,旨在利用计算机的快速计算能力解决包含随机过程的问题,目前在各个研究领域获得了广泛的应用。

几乎所有的计算机系统都配有随机数产生器,可根据用户需要产生任意分布的随机变量。假设根据人口普查得到了实际的出生日期分布,一次实验就可以产生 n 个人的随机生日,假如进行了 M 次实验并统计出了有两人生日相同的次数 m,那么概率 p 就近似等于 m/M。根据统计学的原理,当 M 足够大时,p 的精度可以达到 $O(1/\sqrt{M})$ 的量级。

本文主要研究非均匀分布下的生日概率计算问题,旨在给出非均匀分布与均匀分布下相应概率的定量差别,并给出具体实现和一些模拟结果。其中的算法可推广应用于复杂概率的数值计算。

1　算法与检验

1.1　算法描述

步骤 1　在[1,N]区间内,产生 n 个满足某种分布(均匀或非均匀)的随机自然数,代表 n 个人的生日,存入整型数组 $b(n)$。

步骤 2　比较数组 $b(n)$ 中是否有相同的元素,若有则计数 m 加 1。

步骤 3　重复步骤 1 和步骤 2 共 j 次,统计最后的 m,计算概率 $p = m/j$。

步骤 4　重复步骤 1,2,3 共 k 次,得到 k 个概率值 $p(k)$。

步骤 5　由 k 个 $p(k)$ 值,利用式(2)计算生日概率的平均值 p' 和方差 σ,输出结果。

$$\begin{cases} p' = \sum_{i=1}^{k} p(i)/k \\ \sigma = \sqrt{\dfrac{\sum_{i=1}^{k}(p(i) - p')^2}{k(k-1)}} \end{cases} \tag{2}$$

上述模拟算法实际上包含了一般概率问题数值计算的全部过程。步骤 1 产生样本空间,步骤 2 统计满足某种条件的事例出现次数,步骤 3 增大样本统计量,步骤 4、5 计算大统计量下的概率与统计误差,因此本文的算法可应用于复杂概率计算。如随机找 n 个人,计算至少 3 人生日相同的概率[13],只需简单修改步骤 2 的条件即可。无论随机变量的分布形式如何,由(0,1]区间产生的均匀随机数原则上可以变换产生任意分布的随机变量。

1.2　算法检验

为检验模拟算法的正确性,在均匀分布的基础上,模拟计算了 $n = 10,20,30,40,50$ 下的 $p'(n)$ 值,并与式(1)的计算结果作了比较,见表 1。两种模拟结果分别是在 $M = j \times k = 10^3 \times 10 = 10^4$ 和 $M = j \times k = 10^4 \times 10^2 = 10^6$ 抽样实验中得到的,精度分别在百分之一和千分之一的量级。可以看出,当抽样次数越多时,模拟结果越接近理论值。

表 1　不同抽样次数下模拟结果与理论结果的比较

n	$p' \pm \sigma$ ($M = 10^4$)	$p' \pm \sigma$ ($M = 10^6$)	p
10	0.1151 ± 0.0023	0.11679 ± 0.00030	0.116948
20	0.4210 ± 0.0074	0.41155 ± 0.00055	0.411438
30	0.7033 ± 0.0057	0.70697 ± 0.00044	0.706316
40	0.8870 ± 0.0034	0.89151 ± 0.00033	0.891232
50	0.9697 ± 0.0013	0.97032 ± 0.00017	0.970374

2　非均匀分布模型与模拟计算结果

2.1　非均匀生日分布模型

对于出生日期的非均匀分布,假定分布是近似连续变化的,本文构建了一种简单模型,其概率密度函数(Probablistic Density Function, PDF)为:

$$f(i) = A(1 - \varepsilon \cos(2\pi i l/N)) \tag{3}$$

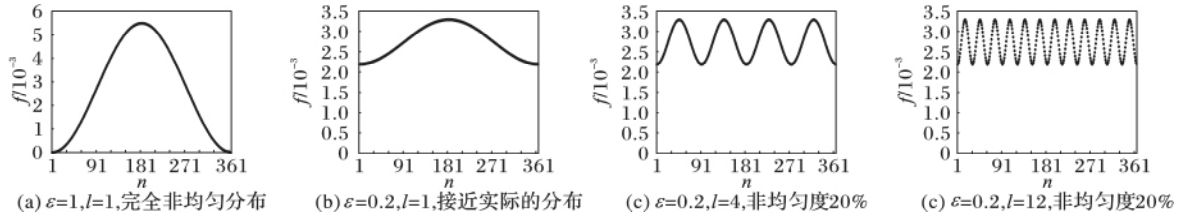
这是一个离散分布($i = 1,2,\cdots,N$),相当于在均匀分布的基础上叠加一个余弦变化的非均匀分布。其中,A 为归一化因子,满足 $\sum_{i=1}^{N} f(i) = 1$;$0 \leqslant \varepsilon \leqslant 1$,代表非均匀度,$\varepsilon = 0$ 表示均匀分布,$\varepsilon = 1$ 表示非均匀度为 100%;l 为整数,$0 < l < N$,表示一年中有几个出生高峰。式(3)用于 1.1 节算法步骤 1 中非均匀生日分布的抽样,图 2 给出了以下 4 种分布:

(a) $\varepsilon = 1$, $l = 1$,完全非均匀分布,这是一种极端假想情况,对应于一年中有一个夏季出生高峰,年初和年末出生人数极少。利用该分布可近似检验计算结果的正确性。

(b) $\varepsilon = 0.2$, $l = 1$,这是一种接近实际的分布,对应于一年夏季一个出生高峰。保守估计实际生日分布的非均匀度应当小于 15%,即不同生日的人数与平均数的差别小于 15%。这里用 20% 的不均度来检验与均匀分布的差别,本文也给出了 $\varepsilon = 0.1$ 下各种分布的模拟计算结果,见 2.2 节。

(c) $\varepsilon = 0.2$, $l = 4$,非均匀度 20%,相当于一年四个出生高峰。

(d) $\varepsilon = 0.2$, $l = 12$,非均匀度 20%,相当于每个月都有一个出生高峰。



(a) ε=1,l=1,完全非均匀分布　(b) ε=0.2,l=1,接近实际的分布　(c) ε=0.2,l=4,非均匀度20%　(c) ε=0.2,l=12,非均匀度20%

图 2　不同生日分布情况下的概率密度函数

2.2　模拟计算结果

为了定量确定非均匀分布与均匀分布生日概率的差异,以下的计算均是在样本统计量 $M = j \times k = 10^4 \times 10^2 = 10^6$ 下进行的,先讨论假想分布(a)的情况。图 3 是模拟计算结果 p' 与均匀分布公式计算结果 p 的比较,可以看出,$p' > p$,这与文献[12]的结论一致,即均匀分布的生日概率最小。图 4(a)

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表示 $p'-p$ 随 n 的变化，最大差异(峰位处)小于0.15，对应于 $n=23$ 附近，偏离23越远，差别越小。从图2(a)可以看出，由于一年中出生日期集中在夏季，年初和年末出生人数都很少，因此抽取相同人数，至少有两人生日相同的机会自然要比均匀分布的情况要大。在这种情况下，$p'(40)\approx p(50)$。在(a)分布情况下，随机找40个人，几乎就有两人生日相同!

对于(b)、(c)、(d)3种分布，它们的非均匀度均为20%，由于模拟计算的结果与均匀分布相差很小，因此这里不再画出 p' 与 p 的比较结果，只给出 $p'-p$ 随 n 的变化，见图4(b)~(d)。

虽然这3种分布的形状非常不同，考虑到模拟计算的统计误差，它们的模拟结果却非常相似，随周期个数变化不大，它们与均匀分布的最大差异都小于0.008(0.8%)！远远小于(a)分布下给出的0.15的结果。这个结果表明，只要整体生日分布均匀，即便在此基础上有20%的非均匀度，生日概率并不发生明显的变化。

为进一步与较为实际的生日分布对比，本文还考察了非均匀度为10%的以下3种情况:

(e) $\varepsilon=0.1, l=1$;

(f) $\varepsilon=0.1, l=4$;

(g) $\varepsilon=0.1, l=12$。

这里不再画出它们的分布图，只给出它们的模拟计算结果 $p'-p$ 随 n 的变化，见图4(e)~(g)。

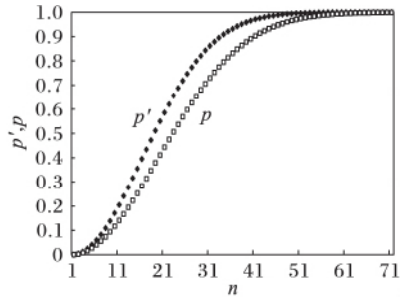


图3 (a)分布下模拟计算结果 p' 与均匀分布理论结果 p 比较

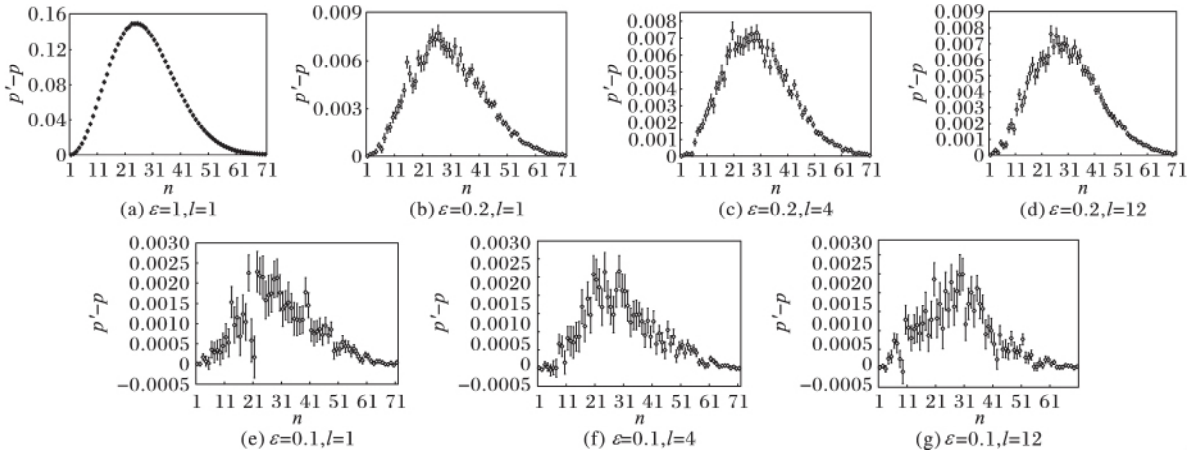


图4 非均匀分布下 $p'-p$ 随 n 的变化

可以看出，对绝大多数计算点 $p'-p>0$，但对 $n<10$ 的个别点 $p'-p<0$，这是由于统计误差引起的，因为此时 $p'-p$ 的绝对值与它们的误差相当。随着统计量的增大，预期它们的值将变为正值。从图4(e)~(g)还可以看出，这三种分布的计算结果也很相似，表明它们与周期次数关联不大，且它们与均匀分布的最大差异小于0.003(0.3%)，即非常接近均匀分布的结果。

需要指出的是，图4(e)~(g)的误差看起来好像比图4(b)~(d)要大，事实上它们的差别不大，只是因为 $p'-p$ 的值较小，对比之下显得它们的误差较大而已，这可从它们的纵轴标度不同看出。其实，图3中 p' 和图4中的 $p'-p$ 也是有误差的，由于它们本身的值较大，误差才显得较小。如果要计算非均匀度 ε 等于5%或更小的情况，由于这时 $p'-p$ 的值与它们的误差已经接近，要得出可靠的结论必需增加统计量，即提高抽样次数(M)，以减小统计误差，当然，这时模拟计算的速度也会急剧下降。

3 结语

本文研究了非均匀生日分布下的生日概率问题。原则上，在已知生日分布(均匀或非均匀)的情况下，可以通过模拟计算很好地求解生日概率。本文的模拟算法可推广应用于一般更为复杂的概率问题的数值计算。为了定量地确定非均匀分布下生日概率与均匀分布的差别，本文提出了接近实际生日分布的余弦变化模型，在20%和10%的非均匀度下，二者的最大差异分别小于0.008和0.003，与均匀分布的结果相差无几。这个结论像生日概率问题一样多少有点出乎意料!

参考文献:

[1] Wikipedia. Birthday problem [EB/OL]. [2015-03-21]. http://en.wikipedia.org/wiki/Birthday_problem.

[2] KLAMKIN M S, NEWMAN D J. Extensions of the birthday surprises [J]. Journal of Combinatorial Theory, 1967, 3(3): 279-282.

[3] ABRAMSON M, MOSER W O J. More birthday surprises [J]. American Mathematical Monthly, 1970, 77: 856-858.

[4] WIKIPEDIA. Birthday attack [EB/OL]. [2015-03-21]. http://en.wikipedia.org/wiki/Birthday_attack.

[5] 冯登国. 密码分析学 [M]. 北京: 清华大学出版社, 2000: 113-114.

[6] 姬东耀，冯登国. 对两个双方密码协议运行模式的攻击及改进 [J]. 计算机科学, 2003, 30(6): 72-73.

[7] 程宽，韩文根. MD5选择前缀碰撞算法的改进及复杂度分析 [J]. 计算机应用, 2014, 34(9): 2650-2655.

[8] 李婧，房鼎益，何路. 混沌文本零水印的词法主动攻击 [J]. 计算机应用, 2012, 32(9): 2603-2605.

[9] 冯伟，冯登国. 基于串空间的可信计算协议分析 [J]. 计算机学报, 2015, 38(4): 701-716.

[10] ERDELSKY P J. The birthday paradox [EB/OL]. [2015-03-25]. http://www.efgh.com/math/birthday.html.

[11] MURPHY R. An analysis of the distribution of birthdays in a calendar year [EB/OL]. [2015-04-21]. http://www.panix.com/~murphy/bday.html.

[12] BLOOM D M. A birthday problem [J]. American Mathematical Monthly, 1973, 80: 1141-1142.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net