The problem of birthdays distribution by computer simulation

Man Li, ZhiGang Zhang*

School of Mathematics and Physics, USTB, Beijing 100083, China

Received 26 September 2014, www.cmnt.lv

Abstract

Over the years, with the rapid development of computer simulation, the probability distribution of birthdays [1, 2] has been a hot topic. Based on the distinguishable ball-into-box issue, this paper started from a seemingly simple ball placing problem and extended to the question of birthday frequency distribution, i.e. people birthday in different dates probability distribution and the distribution law. In addition, also the most important, the value was obtained by Monte Carlo simulations with different frequency distribution birthday days through computer, finally achieving the theoretical value of the frequency values to estimate the probability.

Keywords: birthday problem, probability distribution, Monte Carlo simulation, Computer simulation

1 Introduction. *Discussion of different conditions when placing balls*

1.1 THE PROBABILITY OF EACH BOX HAVING BALL

First, we have an example. Suppose there are 10 distinguished balls randomly placed into 4 boxes. Calculate the probability of the ball number of each box. To simplify the problem, we put one ball in each box first, which means that there is at least one ball per box. Then the rest of the 6 balls are randomly placed into the boxes. Now we just need to calculate the distribution of the rest of the 6 balls. The corresponding distribution is in Table1.

TABLE 1 Distribution of the number of groups

Distribution using 6 Balls	Number of Groups	Corresponding Distribution of 10 Balls	Number of Ball Placements in Each Group	Number of Ball Placements
6000	$C_4^1=4$	7111	$\frac{10!}{7!}=720$	2880
5100	$C_4^1C_3^1=12$	6211	$\frac{10!}{6!2!}=2520$	30240
4200	$C_4^1C_3^1=12$	5311	$\frac{10!}{5!3!}=5040$	60480
4110	$C_4^1C_3^2=12$	5221	$\frac{10!}{5!2!2!}=7560$	90720
3300	$C_4^2=6$	4411	$\frac{10!}{4!4!}=6300$	37800
3210	$C_4^1C_3^1C_2^1=24$	4321	$\frac{10!}{4!3!2!}=12600$	302400
3111	$C_4^1C_3^3=4$	4222	$\frac{10!}{4!2!2!2!}=18900$	75600
2220	$C_4^3=4$	3331	$\frac{10!}{3!3!3!}=16800$	67200
2211	$C_4^2C_2^2=6$	3322	$\frac{10!}{3!3!2!2!}=25200$	151200
Total Number of Groups	84		Total Number of Placement Methods	818520

As we know, the total number of placement methods is $4^{10}=1048576$.

From the above table, we concluded that the probability of each box having a ball is

$$p=\frac{818520}{4^{10}}=\frac{818520}{1048576}=0.7806.$$

1.2 THE DISTRIBUTION OF THE NUMBER OF BOXES THAT HAVE A BALL

We placed 10 distinguished balls into 4 boxes randomly. The distribution of the number of boxes that has a ball (X) is shown in Table 2.

TABLE 2 Distribution of the number of boxes that has ball with 10 balls and 4 boxes

Number of Boxes *X*	**4**	**3**	**2**	**1**
Number of Placement Methods for one box	818520	55980	1022	1
Number of box choosing method	$C_4^4=1$	$C_4^3=4$	$C_4^2=6$	$C_4^1=4$
Total Number of Placement Methods	818520	223920	6132	4
Probability *p*	$\frac{818520}{1048576}=$ 0.780602	$\frac{223920}{1048576}=$ 0.213547	$\frac{6132}{1048576}=$ 0.005848	$\frac{4}{1048576}=$ 0.000004

* ***Corresponding author*** e-mail: zhn1964@163.com

Theorem 1.1 [3]

If *n* discernible balls are randomly placed in *m* boxes, suppose the number of balls placed in each box is $k_1, k_2, \cdots, k_m$ respectively $(k_1 + k_2 + \cdots + k_m = n)$.

If $K(m,n)$ is the number that satisfies $k_1 + k_2 + \cdots + k_m = n$, the different types of ball placement is described by the set $\{(k_1, k_2, \cdots, k_m)\}$.

$K(m,n)$ is then obtained by the recursion formula:

$$K(m,n) = \begin{cases} 1 & m = 1 \text{ or } n = 0 \\ m & n = 1 \\ K(m-1,n) + K(m,n-1) & n > 1, m > 1 \end{cases} \quad .(1)$$

Next, we use mathematical induction to prove the theorem by considering three cases of the problem.

1) When m=1, n balls are randomly placed into a box, and there is only one case: *n* balls put into the single box.
2) When *n*=1, 1 ball is randomly placed into n boxes, and there are *n* placement methods.

Therefore the collection: $\{(k_1, k_2, \cdots, k_m)\}$ has *m* elements, which are:

$$(1,0,\cdots,0),(0,1,\cdots,0),\cdots,(0,0,\cdots,1) \quad .$$

3) When n>1, m>1, suppose that

$$\begin{cases} K(m-1,n-1) = K(m-2,n-1) + K(m-1,n-2) \\ K(m,n-1) = K(m-1,n-1) + K(m,n-2) \\ K(m-1,n) = K(m-2,n) + K(m-1,n-1) \end{cases} \quad .(2)$$

For *m* boxes, the number of balls in Box 1 is $n, n-1, \cdots, 1, 0$, and there are $n+1$ possibilities in total. Suppose the number of the balls in Box 1 is $i, (i = n, n-1, \cdots 1, 0)$, then the remaining $n-i$ balls are placed into the remaining $m-1$ boxes, Subsequently, the set $\{(k_1, k_2, \cdots, k_m)\}$ becomes $K(m-1, n-i)$

$$\begin{aligned} K(m,n) &= \sum_{i=0}^{n} K(m-1,n-i) = 1 + K(m-1,1) + K(m-1,2) \\ &\quad + \cdots + K(m-1,n-1) + K(m-1,n) \\ &= K(m,1) + K(m-1,2) + \cdots + K(m-1,n-1) + K(m-1,n) \\ &= K(m,2) + \cdots + K(m-1,n-1) + K(m-1,n) \\ &= K(m,n-1) + K(m-1,n) \end{aligned} \quad .(3)$$

Theorem 1.2 [3]

If n distinguished balls,are randomly put into m boxes, then the number of placement methods is:

$$\sum_{i=1}^{K(m,n)} \frac{n!}{k_1!k_2!\cdots k_m!}$$

and the number of balls in each box is: $k_1, k_2, \cdots, k_m$ respectively, $(k_1 + k_2 + \cdots + k_m = n)$

Proof: Corresponding to each $(k_1, k_2, \cdots, k_m)$ the number of placement methods is

$$\begin{aligned} & C_n^{k_1} C_{n-k_1}^{k_2} \cdots C_{n-(k_1+k_2\cdots+k_{m-2})}^{k_{m-1}} C_{n-(k_1+k_2\cdots+k_{m-1})}^{k_m} \\ &= \left(\frac{n!}{k_1!(n-k_1)!}\right)\left(\frac{(n-k_1)!}{k_2!(n-k_1-k_2)!}\right)\cdots\left(\frac{(n-k_1-k_2-\cdots-k_{m-2})!}{k_{m-1}!(n-k_1-k_2-\cdots-k_{m-2}-k_{m-1})!}\right)\left(\frac{(n-k_1-k_2-\cdots-k_{m-1})!}{k_m!(n-k_1-k_2-\cdots-k_{m-1}-k_m)!}\right) \\ &= \frac{n!}{k_1!k_2!\cdots k_m!} \end{aligned}$$

Then, the number of putting is:

$$\sum_{i=1}^{K(m,n)} \frac{n!}{k_1!k_2!\cdots k_m!}$$

From ***Theorem 1.2***, we have the inference.

Inference 1.1

$$\sum_{i=1}^{K(m,n)} \frac{n!}{k_1!k_2!\cdots k_m!} = m^n$$

2 The extension of the distribution of the number of the boxes that has a ball

Theorem 2.1 [4]

Allocate n distinguished balls to m boxes randomly, $(m \le n)$, and the probability of each box having ball is

$$\begin{aligned} P(m,n) &= \frac{1}{m^n} \sum_{i=1}^{K(m,n-m)} \frac{n!}{(k_{i_1}+1)!(k_{i_2}+1)!\cdots(k_{i_m}+1)!} \\ &= \frac{1}{m^n} \sum_{i=1}^{K} \frac{m!}{m_{i_1}!m_{i_2}!\cdots m_{i_t}!} \frac{n!}{(k_{i_1}+1)!(k_{i_2}+1)!\cdots(k_{i_m}+1)!} \end{aligned} \quad . \quad (4)$$

K Indicates the number of balls in descending order of all boxes after allocating *n-m* balls to m boxes. i_t indicates that the amount of the distribution of different numbers of balls in descending order, m_{i_t} indicates the number of the same balls corresponding to this distribution .Such as if the number of balls is "4110", the

distribution is $\begin{pmatrix} 4 & 2 & 0 \\ 1 & 2 & 1 \end{pmatrix}$, k_{i_m} indicates that after allocating n balls to m boxes, corresponding to the i[th] method, the number of each box, then $i_t = 3$, $\left(m_{i_1}, m_{i_2}, m_{i_3}\right) = (1,2,1)$, $\left(k_{i_1}, k_{i_2}, k_{i_3}, k_{i_4}\right) = (4,1,1,0)$.

The number of placement methods that guarantees a ball in each box is $m^n P(m,n)$.

Theorem 2.2 When placing n distinguished balls into N boxes randomly the distribution of the box that has a ball is:

$$P\{X=k\} = \frac{C_N^k k^n P(k,n)}{N^n} = \left(\frac{k}{N}\right)^n C_N^k P(k,n)$$

$$X \in [1 \sim \min(n,N)], \tag{5}$$

where $P(k,n)$ indicates the probability of each box having a ball when placing n balls into k boxes.

Inference 2.1 When placing n balls to k boxes randomlythe probability that any two balls are not in the same box is:

$$P\{X=n\} = \frac{C_N^n n^n P(n,n)}{N^n} = \frac{C_N^n n^n \frac{n!}{n^n}}{N^n} = \frac{C_N^n n!}{N^n} \tag{6}$$

3 The problem of birthday distribution

3.1 THE BASIC QUESTION OF BIRTHDAY DISTRIBUTION

Suppose if there are N people and the probability of at least two people having their birthdays on the same day is p [5,6], then

$$\begin{aligned} p &= 1 - \frac{N\cdot(N-1)\cdots(N-n+1)}{N^n} \\ &= 1 - \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right) \\ &= 1 - \frac{365\cdot 364\cdots(365-n+1)}{365^n} \\ &= 1 - \left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365-n+1}{365}\right) \end{aligned}$$

3.2 THE PROBLEM OF FIRST"COLLISION"

We refer to the phenomenon of people being born on the same day as "collision". When $n = 30$, the probability of collision is high, as shown below:

$$\begin{aligned} q &= \frac{365\cdot 364\cdots(365-n+1)}{365^n} \\ &= \left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365-n+1}{365}\right) \end{aligned}$$

Suppose X is the number of balls when the first collision happens with putting balls in N boxes at random.

Then,

$X = 2,3,\cdots,N,N+1$, $p_2 = P\{X=2\} = \frac{1}{N}$,

$p_3 = P\{X=3\} = \left(\frac{N-1}{N}\right)\cdot\frac{2}{N}$,

$p_4 = P\{X=4\} = \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\right)\cdot\frac{3}{N}$

$\cdots$

$$p_k = P\{X=k\} = \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\frac{k-1}{N}$$

Theorem 3.1 Suppose, that

$$\begin{aligned} p_k &= P\{X=k\} = \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\frac{k-1}{N} \\ &= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\left(\frac{N}{N}\cdot\frac{k-1}{N}\right) \\ p_{k+1} &= P\{X=k+1\} \\ &= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\cdot\frac{N-k+1}{N}\right)\cdot\frac{k}{N} \\ &= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\left(\frac{N-k+1}{N}\cdot\frac{k}{N}\right) \\ p_{k+1} - p_k &= A(N-k+1)k - AN(k-1) \\ &= A\left(-k^2+k+N\right) \end{aligned}$$

Then we have $k_0 = \frac{1}{2}\left(1+\sqrt{1+4N}\right)$.

If $k < k_0$, $p_{k+1} - p_k > 0$, the probability is monotone increasing.

If $k > k_0 + 1$, $p_{k+1} - p_k < 0$, the probability is monotone decreasing.

1) If k_0 is an integer, when $k = k_0$ or $k = k_0 + 1$. There is a maximum probability.

2) If k_0 is not an integer, when $k = [k_0] + 1$. There is a maximum probability.

When $N = 10$, the distribution of ball numbers is shown in Table 3.

What's more, $k_0 = \frac{1}{2}\left(1+\sqrt{1+4N}\right) = 3.7$, when $k = 4$,the probability is highest, and $EX = 4.66 \approx 5$, $DX = 2.94$ (see Figure 1)

TABLE 3 The distribution of ball numbers when n=10

The balls number *X* with first collision	The probability *p*
2	0.1
3	0.18
4	0.216
5	0.2016
6	0.1512
7	0.09072
8	0.042336
9	0.014515
10	0.003266
11	0.000363

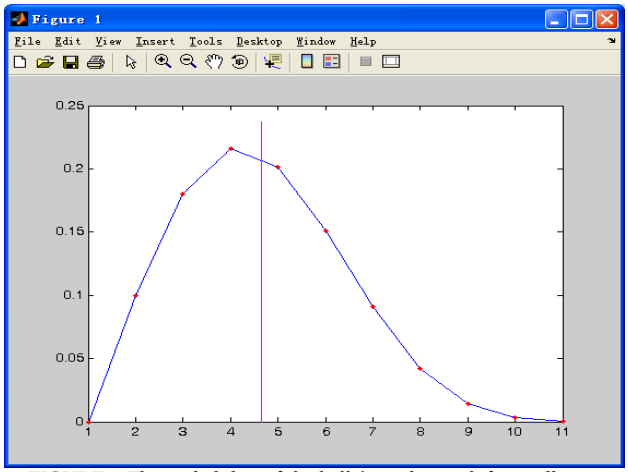

FIGURE 1 The probability of the balls' number with first collision when n=10

When the number of box *N=12*, the Figure 2 we simulated on computer obtained as follows.

From *Theorem 3.1*, we have $k_0 = \frac{1}{2}\left(1+\sqrt{1+4N}\right) = 4$, and from the Figure 2, we can see that ,when $k = 4$ or $k = 5$, the probability is the highest.

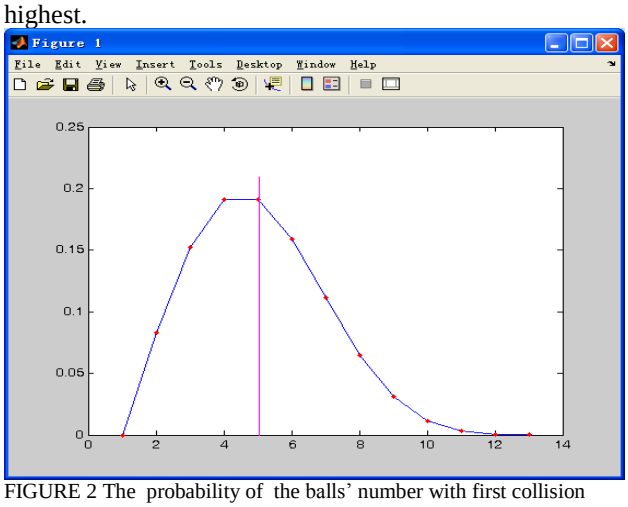

FIGURE 2 The probability of the balls' number with first collision when n=12

3.3 EXTENDING BALL DISTRIBUTION TO BIRTHDAY DISTRIBUTION

To extend ball distribution to birthday distribution [7], 365 boxes will represent the days of the year and each ball will represent a person. Therefore, the problem of placing distinguished balls into boxes becomes a problem of whether or not different people have the same birthdays. Different n, N values correspond to different birthday [8].

However, when the number of people (n) is greater than 50, the amount of computation is enormous. Hence, we can use Monte Carlo simulation [9-12] to generate random numbers within the scope of the day to solve this problem, and then regard the frequency of its appearance as the probability.

1) Under the equally possible birthday conditions i, when the number of birthday (N) equals 365, and *n*=30, we use the formula to calculate the probability of a line after 10000 times of simulation experiments with theoretical values. The birthday number distribution frequency value of the line chart is shown in Figure 3.

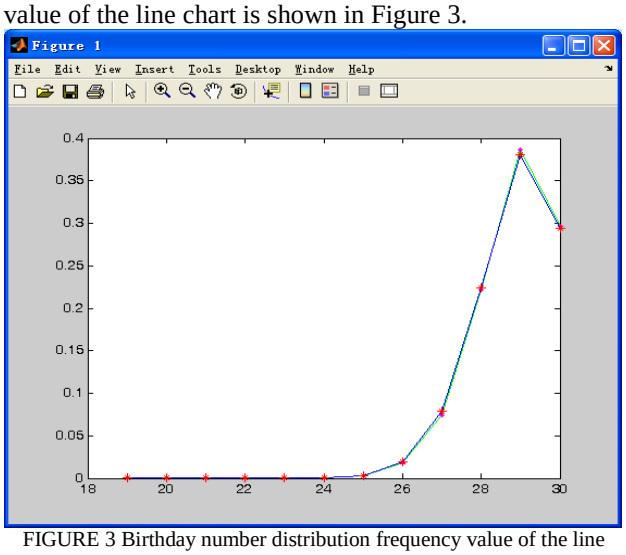

FIGURE 3 Birthday number distribution frequency value of the line chart when n=365,n=30

From Figure 3, it is obvious that when the number is 30, the number of birthdays is mainly distributed in the range of 27 to 30. Its frequency is higher in this scope, and therefore if the number is small, the same birthday probability is smaller and less likely to appear.

The following is the calculation of the probability when the number of people is 30, and the birthday number distribution is 28. We can still assume that if we put 30 balls into 365 boxes, 28 boxes will have a ball.

According to the above example, the problem can be divided into the following two steps:

(1) Choose 28 boxes from 365 boxes, and there are C_{365}^{28} selection methods.

(2) Put 30 balls into these 28 boxes, and the probability of each box having a ball is $28^{30}P(28,30)$. The probability is:

$$P\{X = 28\} = \frac{C_N^k k^n P(k,n)}{N^n} = \frac{C_{365}^{28} 28^{30} P(28,30)}{365^{30}},$$

where $P(28,30)$ indicates that we first put a ball into each of the 28 boxes, which ensures that each box will have a ball, and then put the remaining 2 balls randomly. Thus, we just need to calculate the distribution of these two balls. The distribution and the corresponding placement methods of ten balls are shown in the Table 4.

From the Table 4, we take $P(28,30)$ into the above probability formula, and can directly draw the conclusion: each box has a probability of 0.2238.

2) Figure 4 shows some equally possible conditions. The number of birthdays $N = 365$, the number of people $n = 100$. We get 10000 times birthday number distribution line frequency values through simulation, and use

frequency values to estimate the probability of the theoretical value.

When we increase the number of people greatly, it is obvious that the line becomes relatively stable.

TABLE 4 The distribution of balls with N=30,X =28

Distribution of 2 Balls	Number of Groups	Corresponding Distribution of 30 Balls	Number of Placement Methods of ach group	Total Number of Placement Methods in This Distribution
2000…0	$C_{28}^{1}=28$	3111…1	$\frac{30!}{3!}$	$28\times\frac{30!}{3!}$
1100….0	$C_{28}^{2}=378$	2211…1	$\frac{30!}{2!2!}$	$378\times\frac{30!}{2!2!}$
The total number of group	406		The total number of putting method	2.5066×10^{34}

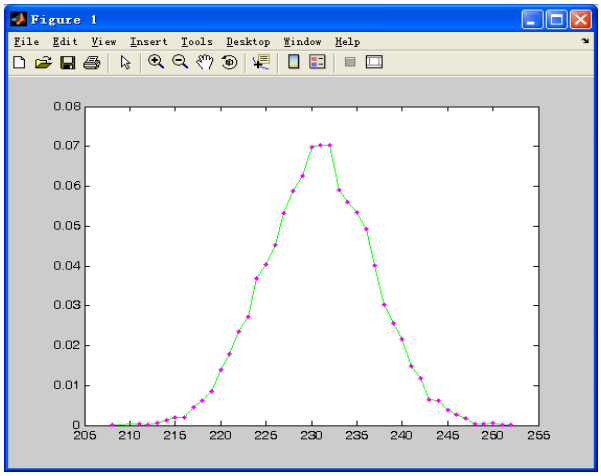

FIGURE 4 365 days, n=100, the 10000 birthday of the distribution of the number of days frequency values line chart

Birthday number distribution increases in probability near 90. Therefore, when the number of people is 100, there is a large probability that the same birthday appears.

Since the theoretical value of the amount of calculation is too large, it is no longer listed.

3) For equally possible conditions, such as birthday, the number of birthday $N=365$, the number of people $n=365$ and $n=1000$, through simulation to get 10000 times birthday number distribution line frequency values, and using frequency values to estimate the probability of the theoretical value, is shown in Figures 5 and 6.

4 Conclusion

Starting from the ball-into-box problem, this article extended the question of distinguished ball distribution and correlated it to the question of birthday frequency distribution. However, the amount of previous discussion about birthday frequency was small, which made the implement relatively easier. In this article, the widely discussed birthday question has been extended, and a different birthday distribution law of the number of days was obtained. This was eventually combined with the Monte Carlo simulation technique, using the principle of random numbers to simulate different people's birthday in order to acquire its frequency. When frequency simulation is large, it becomes closer to the true probability, and the corresponding distribution can be obtained.

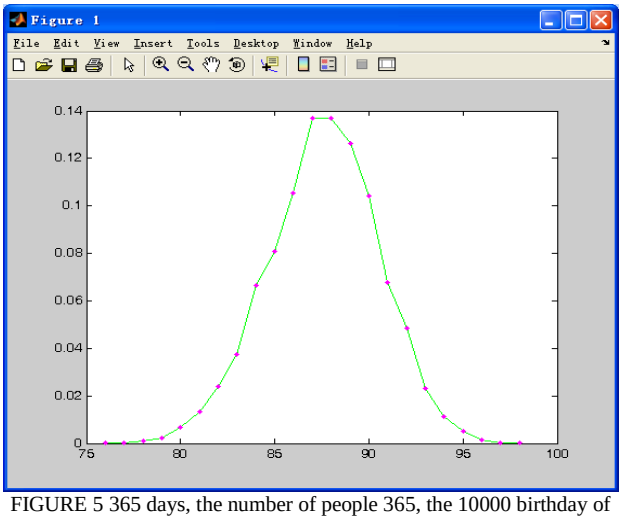

FIGURE 5 365 days, the number of people 365, the 10000 birthday of the distribution of the number of days frequency values line chart

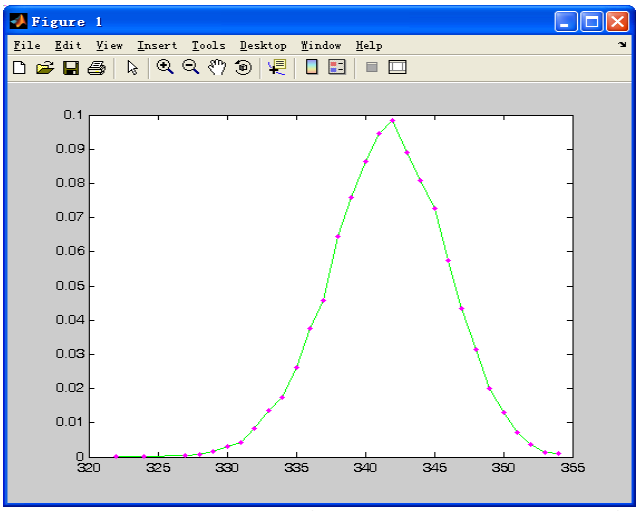

FIGURE 6 365 days, the number of people 1000, the 10000 birthday of the distribution of the number of days frequency values line chart

Acknowledgments

This paper was funded by the University of Science and Technology Beijing education research project (JG2011ZB02).

References

[1] McKinney E 1966 Generalised birthday problem *American Mathematical Monthly* **73**(4) 385-7
[2] Robin A C 1984 The Birthday Distribution: Some More Approximations *The Mathematical Gazette* **68**(445) 204-6
[3] Pesarin F, Salmaso L 2010 Finite-sample consistency of combination-based permutation tests with application to repeated measures designs *Journal of Nonparametric Statistics* **22**(5) 669-84
[4] Chatterjea S K 1961 On Permutations and Combinations *The American mathematical monthly* **68**(3) 279-81
[5] Anichini G 1988 What is the probability of having the same birthday *Archimede* **40**(1) 19-29
[6] Kumar J-D, Proschan F 1992 Birthday Problem with Unlike Probabilities *The American mathematical monthly* **99**(1) 10-2
[7] Clevenson M L, Watkins W 1991 Majorization and the birthday inequality *Mathematics Magazine* **64**(3) 183
[8] Hurley W J 2008 The birthday matching problem when the distribution of birthdays is nonuniform *Chance* **21**(4) 20-4
[9] Aldag S A 2007 *Monte Carlo Simulation of the Birthday Problem*
[10] Kleijnen J, Ridder A, Rubinstein R 2013 Variance Reduction Techniques in Monte Carlo Methods *Encyclopedia of Operations Research and Management Science* 1598-610
[11] Bishwal Jaya P N 2010 Sequential Monte Carlo methods for stochastic volatility models: a review *Journal of Interdisciplinary Mathematics* **13**(6) 619-35
[12] Eftychia C, Marcoulaki G P, Chondrocoukis B P 2013 Optimizing warehouse arrangement using order picking data and Monte Carlo *Journal of Interdisciplinary Mathematics* **8**(2) 253-63

Authors

Man Li, born on November 22, 1991, Shenyang, Liaoning province, China

University studies: Graduated from the Department of Mathematics and Physics, University of Science and Technology Beijing. MSc on Information and Computing Science :
Computer simulation and statistics:
Experience: Got scholarship and honor of “good student” many times during undergraduate**.** Outstanding undergraduate paper, now dedicate to the study of the forecast of mining casualties.

ZhiGang Zhang, born on September 5, 1963, Beijing, China

Current position: Associate Professor, University of Science and Technology Beijing
University studies: Graduated from the Department of Mathematics, Peking University. Masters graduated from Applied Mathematics, University of Science and Technology Beijing
Research interests: Theory and application of database, data structure, C language, computer graphics:
Computer software, applied mathematics, information processing
Publications: A textbook and nearly 30 papers:
Experience: Dedicated to the project "Comprehensive Academic Management System", responsible for eight collaborative research projects, with a total funding of more than one million RMB.During teaching career, got the Outstanding Young Teacher Award and “JianLong“ Outstanding Teacher Award.