Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (18): 3757-3761, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.3757.3761
© 2013 Asian Network for Scientific Information

Calculation and Simulation of Probabilities of Matching Birthdays with Certain Conditions

Linxi He and Zhigang Zhang
School of Mathematics and Physics, USTB, Beijing 100083, China

Abstract: The problem of birthday is an ancient and interesting problem in probability theory. The traditional birthday problem has a conclusion that the probability of the same birthday among people will reach 50% when the number of people reached 23. This study studies the issue of people with a given distribution and discusses the probability that people have the same birthday among a given crowd.

Key words: Birthday problem, simulation, condition distribution, random number

TRADITIONAL BIRTHDAY PROBLEM

Basic formula: The Birthday problem is an important application of probability theory. The traditional model takes into account "the probability at least two people out of a group of n" and this problem has the formula (Clevenson and Watkins, 1991):

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

When people discuss the number of people with the probability of at least two people given, it has the approximate equation (McKinney, 1996):

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

hence:

$$e^{-\frac{(n-1)n}{2N}} \approx 1-p$$

$$-\frac{(n-1)n}{2N} \approx \ln(1-p)$$

$$n^2-n \approx -2N\ln(1-p)$$

thus:

$$n \approx \frac{1}{2}\left[1+\sqrt{1-8N\ln(1-p)}\right]$$

That is, the probability of at least two is P when the number of people is n.

First collision problem: When selecting small balls from boxes randomly, the first collision occurs when the two small balls are selected from the same box. Let X = "Selecting small balls from N boxes randomly, the number of balls selected when the first occurrence of two balls from the same box" and X = 2, 3, ..., N, N+1:

$$p_2 = P\{X=2\} = \frac{1}{N}$$

...

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

...

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

$$p_{N+1} = P\{X=N+1\} = \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{3}{N}\cdot\frac{2}{N}\cdot\frac{1}{N}\right)\cdot\frac{N}{N}$$

among them:

$$\sum_{i=2}^{N+1} p_i = 1$$

Consider the number of ball k most likely to take and then discuss the monotony of p_i first:

Corresponding Author: Linxi He, School of Mathematics and Physics, USTB, Beijing 100083, China

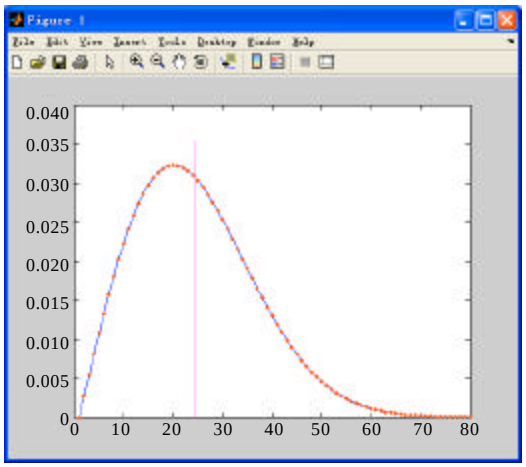

Fig. 1: Relation between the number of extraction and the occurrence of first collision

$$
\begin{aligned}
&p_k = P\{X = k\} = \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\frac{k-1}{N}\\
&= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\left(\frac{N}{N}\cdot\frac{k-1}{N}\right)\\
&p_{k+1} = P\{X = k+1\}\\
&= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\cdot\frac{N-k+1}{N}\right)\cdot\frac{k}{N}\\
&= \left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{N-k+2}{N}\right)\cdot\left(\frac{N-k+1}{N}\cdot\frac{k}{N}\right)\\
&p_{k+1} - p_k = A(N-k+1)k - AN(k-1)\\
&= A\left(-k^2+k+N\right)
\end{aligned}
\tag{3}
$$

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

among them:

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

Hence, $k<k_0$, $p_{k+1}-p_k>0$, the probability monotonically increases. However, if $k>k_0+1$, $p_{k+1}-p_k>0$ the probability would monotonically decrease:

- If k_0 is an integer, then $k = k_0$ or $k= k_0+1$, there is a maximum probability of collision
- If k_0 is not an integer, then, $k = [k_0]+1$, there is a maximum probability of collision

Expectation of the number of balls extracted for the first collision. From the above distribution law of the first collision, average number of first collision can be discussed further. That is the expectation of the number of extracted balls when the first collision occurs:

$$
\begin{aligned}
E(X) &= \sum_{i=2}^{N+1} p_i \cdot i = \frac{1}{N}\cdot 2 + \left(\frac{N-1}{N}\right)\cdot\frac{2}{N}\cdot 3\\
&+\ldots+\left(\frac{N-1}{N}\cdot\frac{N-2}{N}\cdots\frac{3}{N}\cdot\frac{2}{N}\cdot\frac{1}{N}\right)\cdot\frac{N}{N}\cdot(N+1)
\end{aligned}
\tag{4}
$$

Thus, when N = 365, it brings the first collision problem of birthday also when k = 20 the probability reaches the maximum. In addition, EX = 24.61 ≈ 25, DX = 19.61, it shows that an average of 25 people will run into the same birthday.

BIRTHDAY PROBLEM AMONG THE PEOPLE'S BIRTHDAY WITH GIVEN DISTRIBUTION

Under the premise of the given distribution of certain people's birthday, consider the birthday problem of those person selected from such distribution. First, it is called "collision" when the balls drawn from boxes are from the same one. The following two examples illustrate this kind of birthday problem.

Collision of small balls

Example 1: Let six discernible balls occupy the four boxes, whose distribution satisfies:

$$\begin{pmatrix} X & x_1 & x_2 & x_3 & x_4 \\ p & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

Then, select two balls from these six balls arbitrarily and find the probability of these two balls in different boxes.

Solution: The total number of basic events: $C_6^2 = 15$; the number of methods to select boxes: $C_4^2 = 6$; the following shows result of making boxe's number different.

The probability of the two balls in different boxes:

$$q = \frac{13}{15}$$

On the contrary, the probability of the two balls in the same box:

$$p = 1 - q = \frac{2}{15}$$

as there are only two kinds of selection: select two balls from x_1 and x_2, respectively.

If people select three balls from these six balls arbitrarily, the following shows result of making boxe's number different.

The probability of the three balls in different boxes:

$$q = \frac{12}{C_6^3} = \frac{12}{20}$$

Table 1: No. of selection at different combinations (two balls)

Combinations	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	Total
No. of selection	4	2	2	2	2	1	13

Table 2: No. of selection at different combinations (three balls)

Combinations	$x_1x_2x_3$	$x_1x_2x_4$	$x_1x_3x_4$	$x_2x_3x_4$	Total
No. of selection	4	4	2	2	12

Table 3: No. of selection at different combinations (two balls)]

Combinations	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4	Total
No. of selection	3	3	3	1	1	1	12

Table 4: No. of selection at different combinations (three balls)

Combinations	$x_1x_2x_3$	$x_1x_2x_4$	$x_1x_3x_4$	$x_2x_3x_4$	Total
No. of selection	3	3	3	1	10

and in the same box:

$$p = 1 - q = \frac{8}{20}$$

Example 2: Let six discernible balls occupy the four boxes, whose distribution satisfies:

$$\begin{pmatrix} X & x_1 & x_2 & x_3 & x_4 \\ P & \frac{3}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

Then, select two balls from these six balls arbitrarily and find the probability of these two balls in different boxes.

Solution: The total number of basic events: $C_6^2 = 15$; the number of methods to select boxes: $C_4^2 = 6$; the following shows result of making boxe's number different.

The probability of the two balls in the different boxes:

$$q = \frac{12}{15}$$

On the contrary, the probability of the two balls in the same box:

$$p = 1 - q = \frac{3}{15}$$

that is three kinds of selection, select two balls from $x_1C_3^2 = 3$.

If people select three balls from these six balls arbitrarily, the following shows result of making boxes' number different.

The probability of the three balls in different boxes:

$$q = \frac{10}{C_6^3} = \frac{10}{20}$$

and in the same box:

$$p = 1 - q = \frac{10}{20}$$

From the above two examples, the following conclusion can be drawn: The more uniform distribution in the known distribution of small ball collision is, the smaller probability of collision is.

Theorem 1: There are N boxes and n discernible balls occupy m boxes, the distribution law is:

$$\begin{pmatrix} X & x_1 & x_2 & \cdots & x_m \\ P & p_1 & p_2 & \cdots & p_m \end{pmatrix}$$

Select k balls from the n balls ($k \le m \le n$), then the probability of k balls in different boxes is:

$$q = \frac{n^k \sum_{i=1}^{C_m^k} (p_{i_1} p_{i_2} \cdots p_{i_k})}{C_n^k}$$

among them np_i represent the number of balls in i_k-th box.

Proof: Select k different boxes from m boxes and set them $A_{i1}A_{i2}...A_{ik}$, which has C_m^k kinds of selection. Then, choose a ball from each one of the given k boxes and the number of selection is $(np_{i1})(np_{i2})...(np_{ik}) = n^k p_{i1}p_{i2}...p_{ik}$ while the total methods of selecting k balls from n balls is C_n^k. Thus, it is proved that the probability of k selected balls in the different boxes is:

$$q = \frac{n^k \sum_{i=1}^{C_m^k} (p_{i_1} p_{i_2} \cdots p_{i_k})}{C_n^k}$$

Taking people as balls and taking birthday as box, then the distribution of balls is like the known crowd distribution. Similarly, we can discuss the birthday problem of a known crowd.

Same birthday probability when distribute uniformly: The probability of theory 1:

$$q = \frac{n^k \sum_{i=1}^{C_m^k} (p_{i_1} p_{i_2} \cdots p_{i_k})}{C_n^k}$$

it is difficult to compute the probability when n, m,k are large number, since c_m^k items summation is astronomical

figures. However, if the balls in the boxes are evenly distributed:

$$p_{i_1} = p_{i_2} = \cdots = p_{i_k} = \frac{1}{m}$$

there is a simple equation:

$$q_0 = \frac{C_m^k}{C_n^k}\left(\frac{n}{m}\right)^k = \left(\frac{m}{n}\right)\cdot\left(\frac{m-1}{n-1}\right)\cdot\cdots\cdot\left(\frac{m-k+1}{n-k+1}\right)\left(\frac{n}{m}\right)^k \quad (5)$$

to estimate approximately, which avoids calculating directly.

Inference 1: The n known people's birthday satisfy a uniform distribution:

$$\begin{pmatrix} X & x_1 & x_2 & \cdots & x_m \\ P & \frac{1}{m} & \frac{1}{m} & \cdots & \frac{1}{m} \end{pmatrix} (m \le n)$$

then select k person from them, the probability of same birthday of at least two people is:

$$1 - \frac{C_m^k}{C_n^k}\left(\frac{n}{m}\right)^k$$

Theorem 2: The birthday problem of people with given distribution, probability of same birthday of at least two people is smallest when the distribution is uniformly distributed.

Proof: Since, the probability of same birthday is:

$$1 - q = 1 - \frac{n^k \sum_{i=1}^{C_m^k} \left(p_{i_1} p_{i_2} \cdots p_{i_k}\right)}{C_n^k}$$

p_{i1}, p_{i2}, ..., p_{ik} is n non-negative real number and $p_{i1}+p_{i2}+...+p_{ik} = 1$, according to the inequality:

$$\sqrt[n]{p_{i_1} p_{i_2} \cdots p_{i_k}} \le \frac{p_{i_1} + p_{i_2} + \cdots + p_{i_k}}{n}$$

the necessary and sufficient condition for equality is $p_{i1} = p_{i2} = ... = p_{ik}$. Hence, $p_{i1}p_{i2}...p_{ik}$ reach to maximum and 1-q reach to minimum when the distribution is uniformly distributed.

Example 3: At a podium of 117 students, who are known to have the same birthday, assuming that 117 people have

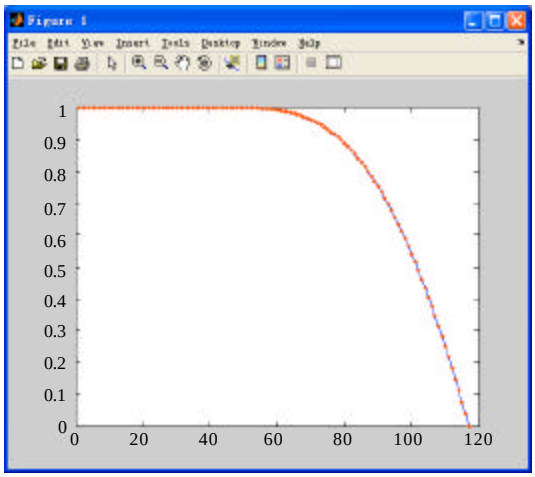

Fig. 2: Relationship between the different birthday day m and probability of at least two people share the same birthday (evenly distributed)

m different birthdays and the distribution of birthday is known; take any 30 people, find the same birthday as the probability of at least two with birthday uniformly distributed.

Solution: When birthday uniformly distributed, take n = 117, k = 30 into Eq. 5:

$$p_0 = 1 - q_0 = 1 - \left(\frac{m}{n}\right)\cdot\left(\frac{m-1}{n-1}\right)\cdot\cdots\cdot\left(\frac{m-k+1}{n-k+1}\right)\left(\frac{n}{m}\right)^k = 1 - \left(\frac{m}{117}\right)\cdot\left(\frac{m-1}{116}\right)\cdot\cdots\cdot\left(\frac{m-29}{89}\right)\left(\frac{117}{m}\right)^k$$

It can be seen that the greater the m is, the lower the probability of the same birthday is.

MONTE CARLO SIMULATION

Take 3 non-uniform distribution of the birthday data as example. The expression can't be calculated directly, so computer simulation (Whitney, 2001) to obtain a given distribution (non-uniform distribution),find the same birthday probability extracted out 30 from 117.

Simulation steps:

- Get the given birthday data's distribution law.
- Test 100 times, resulting 100 sets of random numbers, each group has 30 different random numbers on the type of uniformly distributed
- To judge whether the 30 random numbers corresponding to the birthday are the same and

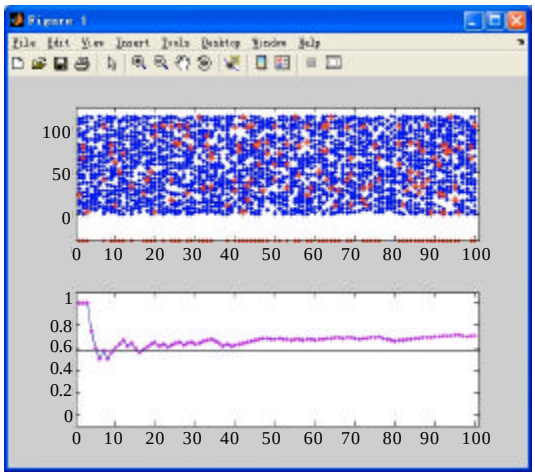

Fig. 3: Probability of same birthday of 30 people extracted from 117 people (non-uniform distribution)

Table 5: Probability and confidence intervals at varies number of tests

No. of tests	5000	10000	15000	20000
Probability	0.7480	0.7417	0.7452	0.7435
Confidence intervals	(0.7375, 0.7585)	(0.7343, 0.7491)	(0.7391, 0.7512)	(0.7382, 0.7487)
No. of tests	25000	30000	35000	40000
Probability	0.7497	0.7468	0.7452	0.7464
Confidence intervals	(0.7451, 0.7544)	(0.7425, 0.7510)	(0.7413, 0.7492)	(0.7428, 0.7501)

record the same down. Finally get the number of "the group what has the same birthday "and divide it by the total number of test. That is the probability of same birthday when the distribution is non-uniform

It can be seen from the above figure that analog value gradually approaches a stable value 0.7450 after 100 times simulation. However, by the formula, at least two of the same birthday probability is 0.5652 when uniformly distributed.

Combine the Monte Carlo simulation with the technology of dual variables (Hammersley and Morton, 1956), the following results obtained by experiment.

By changing the number of the tests, the same birthday probability is about 0.7450 and higher than the same birthday probability when distributed uniformly. Confidence intervals are more and more accurate as the number of tests increases. Moreover, Monte Carlo simulation combined with variance reduction technology can make the same birthday probability more accurate.

CONCLUSION

In this study, expression of the known distribution ball's collision probability was given and the researcher drew the conclusion that uniform distribution has the smallest probability than the same birthday probability of the known people distribution. Besides, they were verified by the theoretical formula and the actual analog.

ACKNOWLEDGMENTS

This article is funded by the Beijing University of Science and Technology Education Research Project (JG2011ZB02).

REFERENCES

Clevenson, M.L. and W. Watkins, 1991. Majorization and the birthday inequality. Math. Mag., 64: 183-188.

Hammersley, J.M. and K.W. Morton, 1956. A new Monte Carlo technique: Antithetic variates. Math. Proc. Cambridge Philos. Soc., 52: 449-475.

McKinney, E.H., 1996. Generalized birthday problem. Am. Math. Monthly, 73: 385-387.

Whitney, M.C., 2001. Exploring the birthday paradox using a Monte Carlo simulation and graphing calculators. Math. Teacher, 94: 258-262.