29

COMMUN. STATIST.-SIMULA. COMPUTA., 11(3), 361-370 (1982)

NUMBER OF MATCHES AND MATCHED PEOPLE IN THE BIRTHDAY PROBLEM

Isaac Meilijson

Tel Aviv University
Tel Aviv, Israel

Marian R. Newborn

Interactive Data Corporation
22 Cortland Street
New York, N.Y. 10006

Aaron Tenenbein

New York University
100 Trinity Place
New York, N.Y. 10006

Uri Yechiali

Tel Aviv University
Tel Aviv, Israel

Key Words and Phrases: probability; expected number of matches; computer storage.

ABSTRACT

The well known birthday problem asks for the probability of at least one match out of a group of n people. Also of interest are the number of matches and the number of matched people. In this paper the means and variances of the number of matches and matched people are obtained. A generalization of the use of these methods to computer storage analysis is discussed.

1. INTRODUCTION

The "Birthday Problem" asks for the probability that at least two people out of a random selection of n people have the same birthday. This problem is well known and is reviewed by Feller (1968). The fact that this probability exceeds .5 when only 23 people are selected is a surprise to most students in probability and statistics classes. Mosteller (1962) discusses this "surprise" effect. Variations on the birthday problem have been discussed by Munford (1977), Bloom (1973), Rust (1976), Glick (1970), and Gihan (1968). McKinney (1966) generalized the birthday problem to compute the probability that at least r people have the same birthday.

Another surprising feature is the apparent large number of matches obtained when large groups of people (50 or more) are involved. Of interest in this situation is the distribution of the number of matches, X, and the number of matched people, Y. In this paper we will derive expressions for $E(X)$, $E(Y)$, $V(X)$, and $V(Y)$ and compare these values for different values of n.

The birthday problem may be generalized and applied to computer storage analysis. Suppose that n records are to be stored in a file. Suppose that the storage is partitioned into $m + 1$ blocks (cells) such that the capacity of cell i $(i = 1,2,\ldots,m)$ is r records and cell $m + 1$ handles the overflow. If any arriving record is equally likely to be hashed to one of the cells $1,2,\ldots,m$ and a cell is overflown as soon as $r + 1$ records are hashed to it, then the number of overflown cells is a generalization of the number of matches and the total number of records that would be put into these overflown cells is equivalent to the number of matched people. Clearly, the birthday problem is a special case of the above when $r = 1$ and $m = 365$.

2. THE NUMBER OF MATCHES

The number of matches, X, is defined as the number of dates which correspond to at least two people from the group of n people selected. We show, under the uniformity assumption that all birthdays are equally likely, that:

$$E(X) = 365\ [1 - (364/365)^n - n(364)^{n-1}/(365)^n] \qquad (2.1)$$

$$V(X) = 365\ (364)\ \theta + E(X) - [E(X)]^2 \qquad (2.2)$$

where

$$\theta = 1 - 2\ (364/365)^n - 2n(364)^{n-1}/(365)^n$$
$$+ (363/365)^n + 2n(363)^{n-1}/(365)^n$$
$$+ n(n-1)\ (363)^{n-2}/(365)^n \qquad (2.3)$$

Equation (2.1) appears on page 446 of Feller (1968) as a solution to problem 18 on page 224.

Proof

Define $X_i = 1$ when at least two people out of n have birthday number i $(i = 1,2,\ldots,365)$. Otherwise $X_i = 0$. The random variables $X_1,X_2,\ldots,X_{365}$ are identically distributed but not independent. Let $p = \Pr[X_i = 1]$ and $\theta = \Pr[X_i = X_j = 1]$; then:

$$p = 1 - \Sigma_{k=o}^{1} \binom{n}{k} \left(\frac{1}{365}\right)^k \left(\frac{364}{365}\right)^{n-k}$$

$$\therefore\ p = 1 - (364/365)^n - n(364)^{n-1}/(365)^n \qquad (2.4)$$

$$\begin{aligned}\theta &= 1 - \Pr[X_i = 1, X_j = 0] - \Pr[X_i = 0, X_j = 1] - \Pr[X_i = X_j = 0]\\ &= 1 - 2\Pr[X_i = 1, X_j = 0] - \Pr[X_i = X_j = 0]\\ &= 1 - 2\ [\Pr[X_j = 0] - \Pr[X_i = X_j = 0]] - \Pr[X_i = X_j = 0]\\ &= 1 - 2\Pr[X_j = 0] + \Pr[X_i = X_j = 0]\\ &= 2\Pr[X_j = 1] + \Pr[X_i = X_j = 0] - 1\end{aligned}$$

$$\therefore\ \theta = 2p - 1 + \Pr[X_i = X_j = 0] \qquad (2.5)$$

Now

$$\Pr[X_i = X_j = 0] = \Pr[N_i \leq 1, N_j \leq 1]$$

where N_k = number of people having birthday k .

$$\Pr[X_i = X_j = 0] = \Pr[N_i = N_j = 0] + 2\Pr[N_i = 0, N_j = 1]$$
$$+ \Pr[N_i = N_j = 1]$$

$$\therefore\ \Pr[X_i = X_j = 0] = (363/365)^n + 2n(363)^{n-1}/(365)^n$$
$$+ n(n-1)\ (363)^{n-2}/365)^n \qquad (2.6)$$

From (2.4), (2.5), and (2.6):

$$\theta = 1 - 2\ (364/365)^n - 2n(364)^{n-1}/(365)^n + (363/365)^n$$
$$+ 2n(363)^{n-1}/(365)^n + n(n-1)\ (363)^{n-2}/(365)^n$$

TABLE I

Mean and Standard Deviation of the
Number of Matches, X , and
Number of Matched People, Y

Number of people in the group	Probability of at least one Match	Number of Matches		Number of Matched People	
n		μ_x	σ_x	μ_y	σ_y
10	.1170	.1215	.2463	.2439	.6842
20	.4114	.5038	.3725	1.0158	1.3632
30	.7063	1.1325	.9483	2.2944	2.0001
40	.8912	1.9942	1.2184	4.0588	2.5971
50	.9704	3.0757	1.4517	6.2893	3.1563
60	.9941	4.3645	1.7609	8.9667	3.6795
70	.9992	5.8487	1.9973	12.0724	4.1685
80	.9999	7.5166	2.2347	15.5886	4.6249
90	1.0000*	9.3572	2.3884	19.4981	5.0503
100	1.0000*	11.3601	2.5682	23.7845	5.3365

* Correct to 4 decimal places.

Now $X = \Sigma_{i=1}^{365} X_i$ = number of matches.

$$E(X) = \Sigma_{i=1}^{365} E(X_i)$$

$$E(X) = \Sigma_{i=1}^{365} \Pr[X_i = 1] = 365\ p$$

from which equation (2.1) follows by substitution of equation (2.4).

$$V(X) = \Sigma_{i=1}^{365} V(X_i) + 2 \mathop{\Sigma\,\Sigma}_{i<j} \operatorname{Cov}(X_i,X_j)$$

$$V(X) = 365\ V(X_i) + 365(364)\ \operatorname{Cov}\ (X_i,X_j) \qquad (2.7)$$

$$V(X_i) = E(X_i^2) - [E(X_i)]^2$$

$$V(X_i) = p(1-p) = E(X)[365-E(X)]/[365]^2$$

$$\mathrm{Cov}(X_i,X_j) = E(X_iX_j) - E(X_i)\ E(X_j)$$

$$\mathrm{Cov}(X_i,X_j) = \theta - p^2 = \theta - (E(X)/365)^2 \ .$$

Equation (2.2) results by substituting for $V(X_i)$ and $\mathrm{Cov}(X_i,X_j)$ into equation (2.7).

Columns 3 and 4 of table 1 show the mean and standard deviation of the number of matched birthdays for various values of n. When $n = 50$, we would expect about 3 matches. As $n \to \infty$, $E(X)$ tends to 365, and $V(X)$ tends to 0, which are expected results.

3. THE NUMBER OF MATCHED PEOPLE

The number of matched people, denoted by Y, is defined as the number of people out of the group of n for whom at least one other person in the group has the same birthday. For example, in a group of 13 people, the following birthdays may result: 2/6, 2/6,2/6,3/19,3/19,6/22,6/22,6/22,6/22,7/11,8/27,12/4,12/9. In this case $Y = 9$ and $X = 3$.

In this section we show that:

$$E(Y) = n[1-(364/365)^{n-1}] \tag{3.1}$$

$$V(Y) = E(Y)[1-E(Y)/n] + n(n-1)\ [(364)(363)^{n-2}/(365)^{n-1} - (364/365)^{2n-2}] \tag{3.2}$$

Proof

Define $Y_i = 1$ if the i^{th} person has the same birthday of at least one other person out of the group $(i = 1,2,\ldots,n)$. Otherwise $Y_i = 0$. $Y_1,Y_2,\ldots,Y_n$ are identically distributed but dependent random variables.

Define

$$r = \Pr[Y_i = 1] = 1 - (364/365)^{n-1} \tag{3.3}$$

Similarly to the derivation of equation (2.5)

$$s = \Pr[Y_i = Y_j = 1] = 2r-1 + \Pr[Y_i = Y_j = 0] \tag{3.4}$$

$$s = 1 - 2\ (364/365)^{n-1} + 364(363)^{n-2}/(365)^{n-1} \tag{3.5}$$

Now

$$Y = \Sigma_{i=1}^{n} Y_i$$

$$E(Y) = \Sigma_{i=1}^{n} E(Y_i) = nr$$

from which equation (3.1) follows.

$$V(Y) = \Sigma_{i=1}^{n} V(Y_i) + 2 \sum_{i<j}\Sigma \operatorname{Cov}(Y_i,Y_j)$$

$$= n[E(Y_i^2) - [E(Y_i)]^2] + n(n-1)[E(Y_iY_j) - E(Y_i)\ E(Y_j)]$$

$$V(Y) = nr(1-r) + n(n-1)[s-r^2]$$

Substitution of (3.3), (3.1) and (3.5) into the above yields equation (3.2).

Columns 5 and 6 of table 1 show the mean and standard deviation of the number of matched people. For 50 people we would expect about 6 matched people. It is interesting to note that $E(Y)$ is roughly twice as large as $E(X)$ indicating that matches are most likely to consist of two matched people for the values of n considered in this table. As $n \to \infty$ it is easy to show that $E(Y)/n$ tends to 1, and $V(Y)$ tends to 0, which are expected results.

4. THE JOINT PROBABILITY DISTRIBUTION FUNCTION OF X AND Y

The birthday problem is also a special case of the following occupancy problem. Consider m cells and n balls, and select x cells and y balls where $x = 1,2,\ldots,m$ and $y = 1,2,\ldots,n$. Let $g_y(x)$ be the number of ways in which these y balls can be arranged within the x cells such that there are at least 2 balls in each cell. A recursive formula to calculate $g_y(x)$ is given by Riordan (1958) as a solution to problem 7 on page 102. The result is

$$g_y(x) = xg_{y-1}(x) + x(y-1)g_{y-1}(x-1) \qquad (4.1)$$

Using (4.1) the joint probability distribution function of X and Y can be calculated.

$$f(x,y) = P[X = x, Y = y] = \binom{m}{x}\binom{n}{y} g_y(x)\gamma/m^n \qquad (4.2)$$

where γ = number of ways in which $(n-y)$ balls can be placed in $(m-x)$ cells such that each cell will contain at most one ball. Clearly,

$$\gamma = (m-x)!/[m-x - (n-y)]!$$

For the birthday problem, setting $m = 365$ we obtain

$$f(x,y) = \frac{\binom{365}{x}\binom{n}{y} g_y(x)(365-x)!/[(365-x) - (n-y)]!}{(365)^n} \qquad (4.3)$$

From this joint distribution the marginal distributions of X and Y can be computed in the usual manner and the means and variances can be evaluated. Clearly this approach requires more effort then the method presented in sections 2 and 3.

5. EXTENSIONS TO COMPUTER STORAGE ANALYSIS

As was pointed in the introduction the birthday problem may be generalized to deal with questions arising in computer storage allocation and overflow. Consider a file with a certain storage capacity. Records are added to the file from time to time. The storage is partitioned into m equal-size blocks, each having a capacity of r records, and an overflow block. A record is hashed to block i $(i = 1,2,\ldots,m)$ with probability $1/m$. If there are already r records in block i, the record is overflown to block $m + 1$ (the overflow block) and stored there. A natural question is what are the "best" values of m and r so that some measure of performance will be optimized. In order to answer such questions it is necessary to find the probability distributions of certain key variables when there are n records in the file.

Let Z_i denote the number of records that "belong" to block i; that is, all records that have been originally hashed to block i .

Clearly,

$$P(Z_i = k) = b(n,1/m;k)$$

$$E(Z_i) = n/m \tag{5.1}$$

where

$$b(n,p;k) = \binom{n}{k}p^k (1-p)^{n-k} \tag{5.2}$$

Let W_i denote the number of records overflown from block i to block $m+1$. $W_i = Z_i - r$ if $Z_i > r, W_i = 0$, otherwise. We have

$$P[W_i = 0] = \Sigma_{k=0}^{r} b(n,1/m;k) \tag{5.3}$$

$$P[W_i = s] = \binom{n}{r+s} (1/m)^{r+s} (1-1/m)^{n-(r+s)} , \quad s = 1,2,\ldots,n-r \tag{5.4}$$

$$E[W_i] = \Sigma_{k=r+1}^{n} (k-r)\, b(n,1/m;k) \tag{5.5}$$

Let W = total number of records in cell $m+1$ (= overflow) .

$$E[W] = \Sigma_{i=1}^{m} E[W_i] = m\, \Sigma_{k=r+1}^{n} k\, b(n,1/m;k) - mr\, \Sigma_{k=r+1}^{n} b(n,1/m;k) \tag{5.6}$$

Substituting (5.2) in (5.6) and simplifying terms we arrive at

$$E[W] = n\, \Sigma_{k=r-1}^{n-1} \binom{n-1}{k} (1/m)^k (1-1/m)^{n-1-k} - mr\, \Sigma_{k=r}^{n} \binom{n}{k} (1/m)^k (1-1/m)^{n-k} \tag{5.7}$$

Define $X_i = 1$ if cell i is overflown, $X_i = 0$ otherwise.

$X = \Sigma_{i=1}^{m} X_i$ is the number of overflown cells.

$$E(X) = m\, E(X_i) = m\, \Sigma_{k=r+1}^{n} b(n,1/m;k) = m[1 - \Sigma_{k=0}^{r} \binom{n}{k} (1/m)^k (1-1/m)^{n-k}] \tag{5.8}$$

The total number of records belonging to all overflown cells is $Y = rX + W$.

Thus,

$$\begin{aligned} E(Y) = E(W) + rE(X) &= n \; \Sigma_{k=r-1}^{n-1} \; b(n-1,1/m;k) \\ &- mr \; \Sigma_{k=r}^{n} \; b(n,1/m;k) \\ &+ rm \; \Sigma_{k=r+1}^{n} \; b(n,1/m;k) \end{aligned}$$

$$E(Y) = n \; \Sigma_{k=r-1}^{n-1} \; b(n-1,1/m;k) - mr\binom{n}{r} \; (1/m)^{r} \; (1-1/m)^{n-r}$$

$$E(Y) = n \; \Sigma_{k=r}^{n-1} \; b(n-1,1/m;k) = n[1- \Sigma_{k=0}^{r-1} \; \binom{n-1}{k} \; (1/m)^{k}(1-1/m)^{n-k-1}] \qquad (5.9)$$

Clearly, setting $m = 365$ and $r = 1$ in equations (5.8) and (5.9) yield equations (2.1) and (3.1), respectively.

BIBLIOGRAPHY

Bloom, D. M. (1973), "A Birthday Problem", *American Mathematical Monthly*, 80, 1141-2.

Feller, W. (1968), *An Introductory to Probability Theory and its Applications*, Volume I, New York: John Wiley and Sons, Inc.

Gihan, E. A. (1968), "Note on the Birthday Problem", *American Statistician*, 22, 2, 28.

Glick, N. (1970), "Hijacking Planes to Cuba: An Up-Dated Version of the Birthday Problem", *American Statistician*, 24, 1, 41-44.

McKinney, E. H. (1966), "Generalized Birthday Problem", *The American Mathematical Monthly*, 73, 385-387.

Mosteller, F. (1962), "Understanding the Birthday Problem", *The Mathematics Teacher*, 55, 322-325.

Munford, A. G. (1977), "A Note on the Uniformity Assumption in the Birthday Problem", *American Statistician*, 31, 3, 119.

Naus, J. I. (1968), "An Extension of the Birthday Problem", *American Statistician*, 22, 1, 27-28.

Riordan, J. (1958), *An Introduction to Combinatorial Analysis*, New York: John Wiley and Sons, Inc.

Rust, P. F. (1976), "The Effect of Leap Years and Seasonal Trends on the Birthday Problem", *American Statistician*, 30, 4, 197-198.

Received by Editorial Board Member August, 1981; Revised December, 1981.

Recommended by Shelemyahu Zacks, State University of New York at Binghamton, Binghamton, NY

Refereed Anonymously.