COLLISIONS OF BIRTH DATES

Jiří Slavík
University of Žilina, Faculty of Management Science and Informatics

Abstract: *The article deals with the possible colliding situations in the case, when a person is identified only by its name and surname and its date of birth. Real data about people born in individual years, statistics concerning names and surnames and other relevant information were gained via Internet pages of MVČR. By collision we understand the occurrence of a situation, when there is a pair of people, having the same name and surname and being born in the same year, month and day.*

Keywords: *Probability of a collision, distribution law of the number of collisions, mean value of the number of collisions, Monte Carlo simulation*

1. Introduction

Identification of people by their personal data is a sensitive problem, the highest degree of privacy should be preserved and as few personal data as possible should be revealed. In this connection a question appeared, whether the name, surname and the birth date are sufficient information to identify uniquely a person. We can find on the internet pages of MVČR statistics concerning frequency of surnames in ČR and also statistics of names given to children born in the individual years. Using some simplifications, the number of people born in the same year with equal name and surname can be estimated and the probability of a collision can be computed. By collision we understand the occurrence of two different people with equal name, surname and birthdate. Approximation of the probability distribution of the collision rate is derived and the results are compared with the Monte Carlo simulation experiment results.

2. Solution of the problem

Simplifications

The following conditions are assumed to be true:

- Year has 365 days
- Name and surname are independent
- Frequency of surnames occurrence does not change in time
- Every day in a year is equaly probable to be a birthday of a person
- Every person has only one name and one surname

These simplifications of a real situation are rather realistic, are not excesively limiting and reflect the characteristics of a majority.

Probability of a collision

Let'have a random sample of *n* people. We will find the probability that at least two of them celebrate the birthday at the same day and month. (So far without respect to the year). This is also the probability that at least one collision occurs. Let's denote as X the random variable defined as the number of collisions among *n* people. It is easy to see that the probability that no collision occurs is

$$P(X=0)=\frac{365.364.363\ ....(365-n+1)}{365^{n}} \qquad (1)$$

and therefore the probability of at least one collision is

$P(X>0)=1-P(X=0)$

But we are more ambicious, we would like to know the distribution law of the collision number, i.e.the distribution of the random variable X. It means that we should find how many collisions occur and with what probabilities. This task is more difficult. We ned to know all cases where at least two people are born the same day and month and their probabilities. Then, if *k* people are borne the same day and month, the number of collisions is *k(k*-1)/2, i.e. the number of all pairs among *k* people. We shall find an approximate solution to this problem and compare its results with the simulation results.

Approximate solution

The approximate solution is based on the folowing idea. Let's divide *n* people into pairs. The number of pairs is N = *n(n*-1*)* /2. A collision in each pair occurs with the probability $p=\frac{1}{365}$ (In a fixed day the probability of collision is $\frac{1}{365^2}$ and there are 365 days, so $p=\frac{1}{365^2}.365=\frac{1}{365}$). Collision does not occure with the probability $q=\frac{364}{365}$. We can repeat it for each pair. The series of these trials resembles Bernoulli independent trials and therefore we use the formula

$$P(X=k)=\frac{N!}{k!(N-k)!}p^k.q^{n-k} \quad \text{for } k = 0,1,...,\text{N} \qquad (2)$$

for the probability of *k* collisions.

However, there is a trap. The trials are not independent. We can see it clearly for small values of *n*. If *n* = 3 for example, the number of collisions can't be 2, so P(X=2) =0, which, of course, does not comply with the formula (2).

For sufficiently large values of *n*, however, the dependence is rather week and the formula (2) gives good results. Moreover, because parameter *p* is sufficiently small and N is large enough, we can use the approximation by Poisson distribution. Parameter l is equal to $N.p$.

$$P(X=k)=\frac{l^k}{k!}e^{-l} \qquad (3)$$

The approximation of P(X=*k*) can be improved. We know, using formula (1), the exact probability $P_0=P(X=0)$. So we can correct the value of parameter *p* in (2) the following way: (Corrected values are denoted as p_1, q_1, l_1)

$$P(X=0)=q^N \Rightarrow q_1=\sqrt[N]{P_0} \quad p_1=1-q_1 \qquad (4)$$

similarly, the value of l in (3) can be corrected:

$$P(X=0)=e^{-l} \Rightarrow l_1=-\ln(P_0) \qquad (5)$$

The three series of Monte Carlo experiments each with 10^6 repetitions were performed and average value of its results were used to verify the viability of these approximations. The results of these experiments are presented in the Table 1. The following values were chosen:

n = 23 , N = 253

Initial approximation parameters:

p = 1/365 = 0.0027397

q = 1-p = 0.9972603

l = 0.6931501

E(X)=0.6931501

P(X=0) = 0.492703 exact value from (1)

Corrected parameters values:

p_1= 0.002794

q_1 = 0.997206

l_1 =0.70785

E(X)=0.70785

Table 1: Approximations of collisions distribution

No of collisions	**P (X = *k*)**				
	Monte Carlo	**Bin(*p*,*q*)**	**Poiss(l)**	**Bin(p_1,q_1)**	**Poiss(l_1)**
0	0.49265	0.49952	0.49999	0.49270	0.49270
1	0.36333	0.34720	0.34657	0.34923	0.34876
2	0.11095	0.12018	0.12011	0.12329	0.12343
3	0.02579	0.02765	0.02775	0.02890	0.02912
4	0.00593	0.00474	0.00481	0.00506	0.00515
5	0.00101	0.00065	0.00067	0.00071	0.00073
6 and more	0.00125	0.00006	0.00010	0.00011	0.00011

Estimates for mean and dispersion obtained from Monte Carlo simulation are:

E(x)=0.69347

D(x)=0.69187

Their near equivalence also supports the hypothesis of Poisson distribution of the number of collisions.

We may observe a satisfactory fit of experimental results with approximations. The correction of parameters p and l does not seem to be significant.

Similar accuracy of results was obtained for different choices of parameter ***n***.

3. Example

We want to investigate the collisions of persons called Tomáš Novák born in the year 1989. There are 4914852 different surnames registered in the Czech Republic. Let’s see the excerpt from MVCR statistics of names and surnames:

Table 2: Frequency of surnames in ČR

Surname	**Absolute frequency**	**Relative frequency**
Novák	34476	0,007015
Svoboda	25311	0,005150
Novotný	24388	0,004962
Dvořák	22342	0,004546
Černý	17936	0,003649
Procházka	16177	0,003291

Table 3: Frequency of names in 1989

Name	**How many born in 1989**
Jiří	***3162***
Jan	5363
Josef	1166
Petr	3763
Jaroslav	1158
Pavel	2344
Miroslav	1150
František	518
Martin	4741
Zdeněk	947
Václav	1069
Tomáš	5056

Taking into account the simplifications given in 1.1, we may estimate the number of persons born in the year 1989 and bearing the name and surname Tomáš Novák. We obtain: $n = 5056 \text{ x } 0.007015 = 35$ (rounded to integer). We estimate the distribution of random variable X – the number of collisions - by Poisson distribution with parameter l , which may me estimated as follows:

$N = n\ (n - 1\ )/2 = 595$

$p = 1 / 365 = 0.0027397$

$l = N\ p = 1.63014 \ \Rightarrow\ E(X) = 1.63$ ….average number of collisions

The distribution of the number of collisions using the approximation by Poisson law is in the next table:

Table 4: Distribution of collisions for 35 people

Number of collisions *k*	**Probability P(X=*k*)**
0	***0.19590***
1	0.31935
2	0.26029
3	0.14144
4	0.05764
5	0.01879
6	0.00511
7 and more	0.00148

4. Conclusion

The exact formula for the distribution law of the number of collisions seems to be difficult to derive, so far I don’t know how to find it. However, there is a simple way how to substitute the exact solution by its approximation. This approximation is described and some variants of it are investigated in this article. The result is demonstrated on the interesting case of collisions in the identification of persons by their name, surname and birthdate in the Czech Republic. The example, at the same time, gives a hint how to do it generally for any name, surname and year of birth.

References:

[1] PARZEN, E.:*Modern Theory of Probability and Its Applications.* John Wiley&Sons, N.Y., 1960.

[2] Internet Page of MVČR http://www.mvcr.cz/statistiky

Contact address:

Jiří Slavík
University of Žilina
Faculty of Management Science and Informatics
Slovak Republic
Email: slavik@frdsa.fri.utc.sk