Birthday Paradox

Davide Di Mauro
Politecnico di Torino
Student ID: s306089
s306089@studenti.polito.it

***Abstract*—The objective of this homework is the simulation of the Birthday Paradox, computing the probability of collision and the average number of collisions. An extension is proposed, computing the probability of conflict between two types of categories instead of one like in the standard formulation of the problem.**

I. Assumptions

I made the following assumptions:

1) We don't consider leap years
2) I assume that the distribution of births in America from 1994 to 2003 is representative of the global births distribution

II. Input parameters

- m: the number elements (i.e. the number of people in the set)
- n: the cardinality of the set (i.e. the number of days in a year)
- k: the number of experiments done in order to compute the probability ($k = 1000$) and the average ($k = 500$)

III. Output metrics

- $E[m]$: the average number of people to observe a conflict
- $P(m)$: the probability to observe at least one conflict
- The confidence interval for both the aforementioned statistics

IV. Real distribution

In order to sample data according to a realistic distribution, the following dataset was used: US_births_1994-2003_CDC_NCHS.csv. This file contains the number of people born in the US from 1994 to 2003; in order to use it, the 29th of February is removed and then then average over the years can be computed.

V. Average number of people to observe a conflict

In Figure 1 is shown our estimation of $E[m]$ as the size of the set changes and its confidence interval, together with the exact value of $E[m]$. We can see that as expected the confidence interval gets larger as n increases since it becomes more and more difficult to estimate correctly the average. Nevertheless our estimation it's still reasonable since the exact value of $E[m]$ remains always inside the confidence interval.

Algorithm 1 describe the general process used to compute the average.

Algorithm 1 Average number of people to observe a conflict

Input: m, n, k, $distribution\_type$
Output: $E[m]$
  $m\_list \leftarrow []$
  for $i \leftarrow 0$ to k **do**
    $conflict\_counter \leftarrow 0$
    $birthday\_list \leftarrow []$
    while $True$ **do**
      $birthday\_list \leftarrow sample\_from\_distribution()$
      if $conflict\_in(birthday\_list) == True$ **then**
        $m\_list \leftarrow conflict\_counter$
        break
      else
        $conflict\_counter \leftarrow conflict\_counter + 1$
      end if
    end while
  end for
  return $Avg(m\_list)$

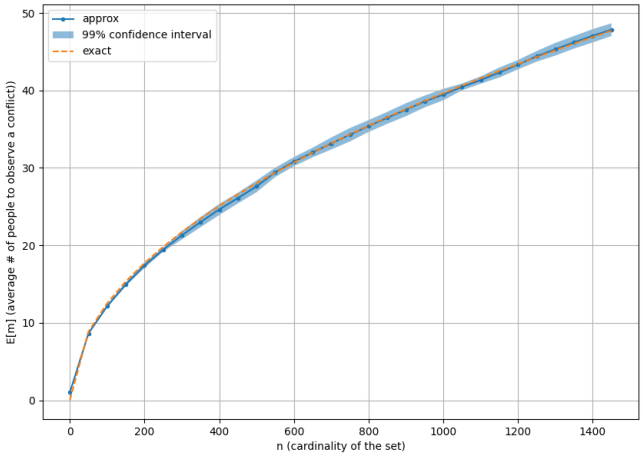

Fig. 1: Average number of people to observe a conflict

VI. Probability of observe a collision

Figure 2 shows the probability of having at least one conflict in both the case of the uniform distribution and the real one. We can conclude that the uniform is a good approximation of the real case, so I will continue in the analysis considering only this case. Moreover, to prove the correctness of the approximation, the exact case is also added to the graph, and the two curves coincide almost perfectly.

Algorithm 2 Probability of observe a collision

Input: m, n, k, $distribution_type$
Output: $P(m)$
  $collision\_counter \leftarrow 0$
  for $i \leftarrow 0$ to k **do**
    $arr \leftarrow sample\_from\_distribution(dim = m)$
    if $conflict\_in(arr) == True$ **then**
      $collision\_counter \leftarrow collision\_counter + 1$
    end if
  end for
  return $collision\_counter/k$

I also wanted to analyze the case in which n changes (Figure 3), and as expected the shape of the curve changes, in fact the probability to observe a collision decreases as n increases. Please notice that also the 99% confidence intervals are plotted in Figure 3, but they are very small and difficult to see, since for estimating the probability in each point $k = 1000$ experiments are computed. The steps used in computing the probability of collision is briefly described in Algorithm 2.

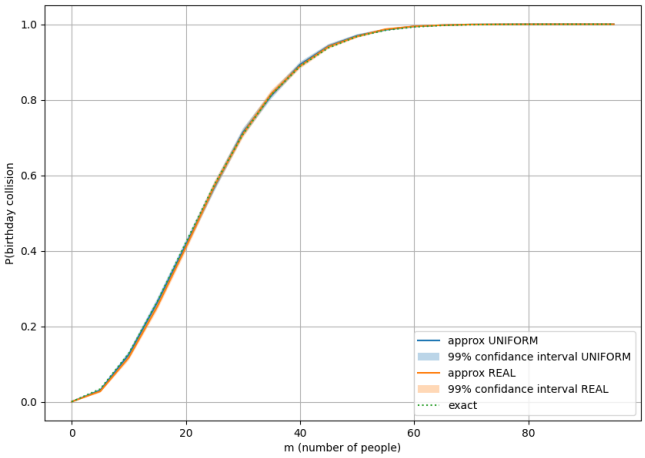

Fig. 2: $P(m)$ - Uniform vs. Real distribution

VII. Extension

The proposed extension is a generalization of the classic problem, considering 2 different types, like men and women. In this case we will have a men and b women, and the problem becomes characterizing the probability of a shared birthday between at least one man and one woman. The probability of **no** shared birthdays is the following equation:

$$p_0 = \frac{1}{d^{a+b}} \sum_{i=1}^{a} \sum_{j=1}^{b} S_2(a,i) S_2(b,i) \prod_{k=0}^{i+j-1} d - k \qquad (1)$$

where $d = 365$ ans S_2 are Stirling numbers of the second kind. Consequently, the desired probability is $1 - p_0$.

Figure 4 show the plot of the simulated probability of at least one shared birthday between at least one men and one women. This generalization is interesting because the results

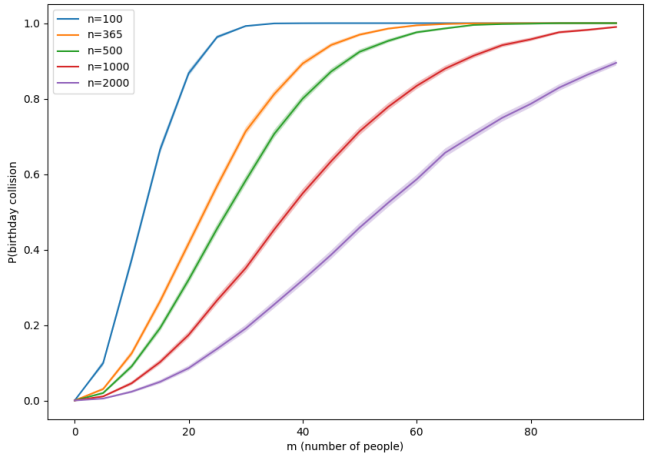

Fig. 3: Effect of n on $P(m)$

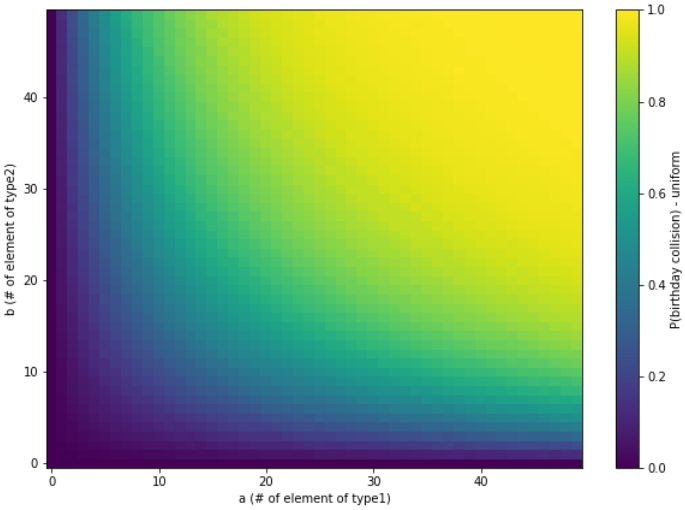

Fig. 4: $1 - p_0$ in the generalized case

show that there is not a unique solution for the total number of people $a + b$ to have a specific value of p_0. An example is that both $a = 16, b = 16$ and $a = 43, b = 6$ are solution to have a 50% probability of collision.

Algorithm 3 Probability of observe a collision **2 types**

Input: a, b, n, k, $distribution_type$
Output: $1 - p_0$
  $collision\_counter \leftarrow 0$
  for $i \leftarrow 0$ to k **do**
    $arr\_type1 \leftarrow sample\_from\_distribution(dim = a)$
    $arr\_type2 \leftarrow sample\_from\_distribution(dim = b)$
    if $conflict\_in(arr\_type1, arr\_type2) == True$ **then**
      $collision\_counter \leftarrow collision\_counter + 1$
    end if
  end for
  return $collision\_counter/k$