Lab Report
Birthday problem

Francesco Pagano - 299266

I. Introduction

This lab's objective is simulating the birthday problem. Such phenomenon arises when comparing the birthdays of a set of people with given cardinality, resulting in two or more having the same birthday. The question we are trying to answer is:

What is the minimum cardinality of a set of people such that at least two of them happen to have the same birthday with given probability p?

Assuming **m** is the cardinality of the set under study, with $m < 366$, then

$$\hat{p} = \prod_{i=1}^{m-1} \frac{365 - i}{365} \tag{1}$$

$$\bar{p} = 1 - \hat{p} \tag{2}$$

are the probabilities of every member of the set having a different birthday and at least two members having the same birthday, respectively. Equation (2) can also be expressed in a more computationally-handy form using a trivial Taylor expansion:

$$\bar{p} = 1 - e^{-\frac{m^2}{2 \times 365}} \tag{3}$$

The analysis of this problem can be further extended and generalized to study any problem where a sampling of a certain number of elements from a set with given cardinality is required to compare one of their specific properties, which can be upper bounded by any value.

II. Simulator's architecture

The way the simulator is built takes into account the opportunity of analyzing the generalized version of the birthday problem. The general workflow starts from computing the average cardinality to experience a conflict alongside its confidence interval, then the simulation for computing the probability in function of a given set of cardinalities starts. This process is repeated one time for each distribution employed for generating the birthdays of the set.

A. Assumptions

The process for generating the initial birthday is carried out either by means of a uniformely distributed random variable, or a random variable distributed according to a real dataset[1].

B. Input parameters

The simulator accepts as input the following parameters:

- Property's upper bound: in this case the maximum index of the year's day, 365.
- Maximum considered cardinality: the largest cardinality possible for the test set.
- Real dataset path: the path pointing to the dataset used for generating the discrete (real) distribution from which instantiating birthdays.

C. Data structues and algorithms

With the goal of generalization in mind, the birthday property is represented by the Property class, which is just a general way of encapsulating informations circa the property generation together with methods for generating each test set of given cardinality, while another class, Obj, generalizes each member of every test set with each instance being a placeholder for the property itself, in this case, the birthday date. The pseudocode for the simulator is:

```
for each dist_type:
    compute_avg_card()
    compute_avg_ci()
    for each test_cardinality:
        generate_test_set()
        compute_statistics()
    log_results()
```

Where each Confidence Interval is computed according to the standard laws:

$$I_m = [\hat{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \hat{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}] \tag{4}$$

$$I_p = [\hat{p} - z_{\frac{\alpha}{2}} \hat{s}, \hat{p} + z_{\frac{\alpha}{2}} \hat{s}] \tag{5}$$

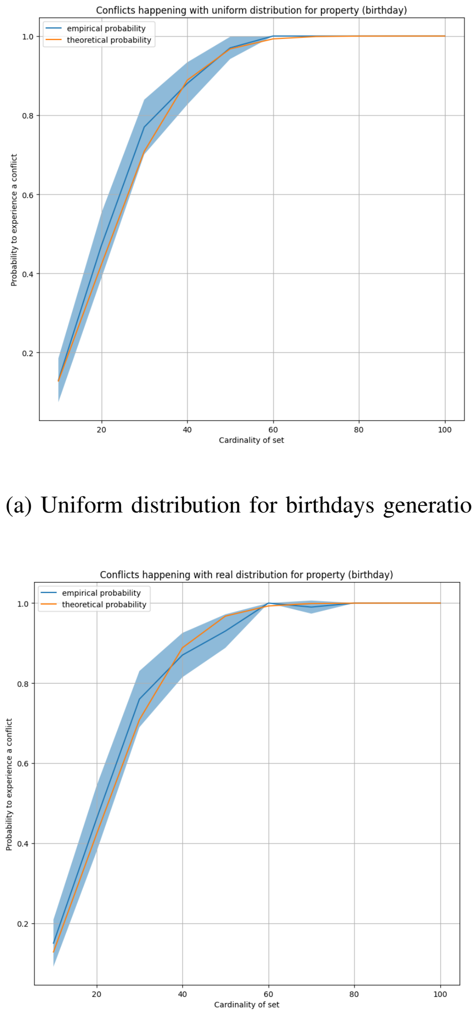
Here we have:

- $\hat{x}$, $\hat{p}$ are the variable and probability estimator, respectively.
- s and $\hat{s}$ are the variable's standard deviation and the probability's standard deviation's estimator.
- n is the number of experiments/trials corresponding to the appropriate value.
- $t_{n-1,\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ are Student's t and Gaussian value at confidence level $\frac{\alpha}{2}$.

D. Output metrics

The simulator outputs its computations on the average cardinality to experience a conflict alongside its confidence

[1] To get a real distribution for the birthdays the dataset that has been used is available at https://github.com/fivethirtyeight/data/blob/master/births/US_births_2000-2014_SSA.csv, which contains U.S. births data for the years 2000 to 2014, as provided by the Social Security Administration.



(a) Uniform distribution for birthdays generation

(b) Real distribution for birthdays generation

Fig. 1: Probabilities vs Cardinalities for different birthdays' distributions.

interval and the theoretical value. These results are shown in the following table and graphs for both uniform and real distributions.

III. Results

A. Birthday simulation

Fig.1 shows the probabilities of experiencing a conflict when the birthdays are sampled from a uniform distribution (Fig.1a) or real (Fig.1b) alongside their respective confidence interval and the curve predicted by theory. According to the mathematics of the problem, the greater the cardinality is, the greater the probability of experiencing a conflict (at least two members having the same birthday) in the set, with cardinalities greater than 60 featuring a stable probability $p \approx 1$. Theoretical results show that the average cardinality to experience a conflict is sharply concentrated around its average:

$$E[m] = \sqrt{\frac{\pi}{2} \times n} \approx 1.25\sqrt{n} \tag{6}$$

where n is the property upper bound. According to the following table, the obtained output metrics are coherent with the theory.

	Uniform dist.	Real dist.
Average cardinality	21.74	25.09
Confidence interval	(13.50, 29.98)	(12.68, 37.49)
Theoretical avg. cardinality	23.88	23.88

TABLE I: Simulator's output metrics

B. Generalization

Since the generalization of the birthday problem is very versatile, simulating different scenarios in which, using the previous fixed set of cardinalities, the key property to study has a different upper bound than the simple number of days in a yeah results useful. Fig.2 shows that the such upper bound doesn't really affect the general trend governing such phenomenon.

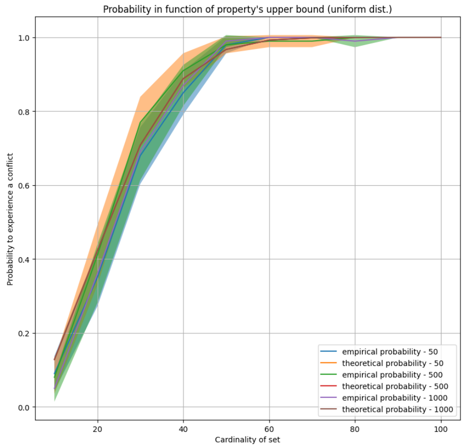


Fig. 2: Probabilities vs Cardinalities for different property's upper bounds: 50, 500, 1000. Property's generation is carried out only by means of the uniform distribution since we are dealing with an hypothetical phenomenon useful just for visualization purposes.