Laboratory 09
Birthday paradox

Giuseppe Esposito
s302179

I. Preliminary study

A. OVERVIEW

Task 1: Given a probability of conflict equal to 0.5, we want to evaluate the minimum number of students entering in a class such that at least 2 of them have the same birthday (i.e. a conflict has occurred).

Task 2: Given a set of class sizes, we want to compute the probability that at least two students have the same birthday (i.e. a conflict has occurred).

B. Assumptions

Let us assume:

- First case: uniform distribution for birthdays.
- Second case: real distribution of birthdays. **Click here** to access the sources from which I extracted the real distribution.
- To compute the average number of minimum extraction before having a conflict I perform many runs, extracting elements until i have a conflict, then I take the averages of these simulations.
- To compute the probability of conflict in the second task, we used an Taylor based approximated formula (approximation: $e^x \approx 1 - x$)
- Number of students is strictly less then number of days in a year $m < n$.
- I considered also the leap years.

C. Input parameters

- Number of runs (ITERATIONS): number of times we want to repeat a simulation (the higher is this value the more accurate will be the estimates with respect to the theoretical ones, but the higher will be the computational cost);
- List of number of students in a class (M): the possible "sizes" of the classes over which I compute the probability of conflict;
- Distribution (D): it can be either uniform or the one taken from real data.

D. Output metrics

- Average number of minimum extractions such that a conflict occurs;
- Given m number of elements (students), probability that at least one conflict has occurred.

E. Main data structures

For the first metric: I decided to use a dictionary where the keys are all the possible birthdays (int from 0 to 365) and the values are the occurrences of a random extraction.

For the second metric: I decided to use a set to store all the already extracted birthdays.

II. MAIN ALGORITHMS

To compute the real distribution I took as reference the **github repository** csv which stores the data about the births between 2000 and 2014 in U.S. Once I computed the probabilities I used a numpy function which let me assign the corresponding weights to the possible extractions. It firstly random generate an instance of a uniform distribution U(0,1) and then it checks in which interval of the CDF that instance is located so that it returns the index of the upper bound of that interval and then the corresponding possible choice.

These are the 2 main algorithms that I used in order to compute the requested metrics. They could have been easily merged into a single algorithm but I decided to keep them separated for the sake of consistency with the purposes of the laboratory.

Algorithm 1 Task1: Average number of people

```
results ← empty list
for run in range(1; #ITERATIONS) do
    days ← dict with 366 integer keys
    counter ← 0
    while True do
        increment the counter
        sample a random day from the distribution (uniform or real)
        increment the counter of occurrences for that date in the dictionary
        if that value is equal to 2:
            return the counter
    end while
    results[] ← counter
end for
compute the mean over results
```

Algorithm 2 Task 2: Probability to have at least one conflict

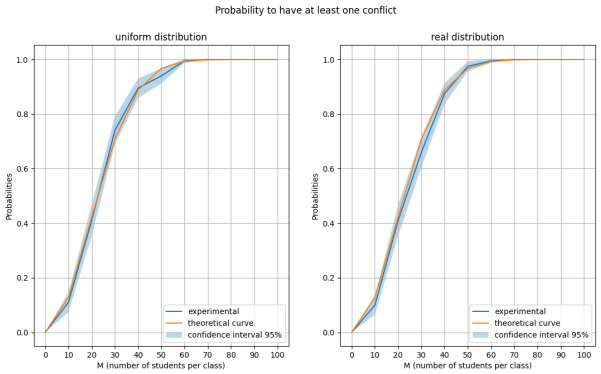
```
counter ← 0
for run in range(1; #ITERATIONS) do
    birthdays_set ← Empty set
    for m in CLASSES_SIZE do
        sample a random day from the distribution (uniform or real)
        if extracted birthday is in birthdays_set
            increment the counter
            break
        else
            add extracted birthday to birthdays_set
    end for
end for
probability ← (counter/#ITERATIONS)
```

III. RESULTS

A. Task1

As result of the experiment I computed the confidence interval over the list of first collisions that occur at each run of the simulation. After 200 runs, we can state that the theoretical value ($1.25 \times \sqrt{n}$ where n is the number of days in a year) is accurate with $confidence\_level = 97\%$.

B. Task2



As we can notice in figure III-B for growing sizes of classes, both curves (theoretical and experimental) has a linear increasing trend until the class size is equal to 60 after which we are always sure to have a conflict, this is due by the fact that if we let more students enter in the class, the probability that all students have distinct birthdays:

$$\hat{p} = \frac{364}{365} + \frac{363}{365} + ... + \frac{365 - m + 1}{365} \tag{1}$$

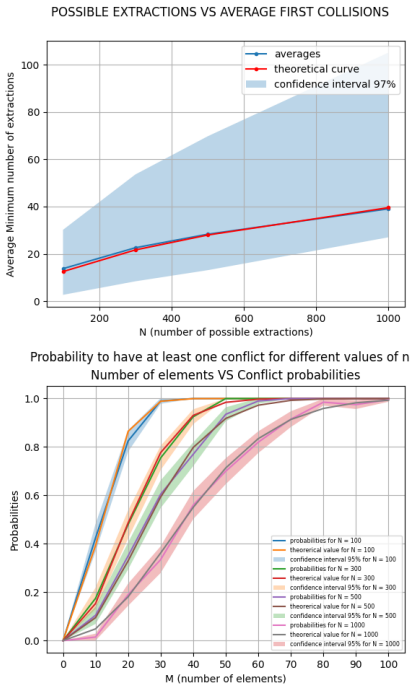
decreases so that his counterpart $1 - \hat{p}$ increases. Moreover we can notice that the theoretical value always falls in the confidence interval of both the uniform and the real distribution generated curves until it reaches the stability where the confidence interval is highly thin, almost negligible, so that the two curves perfectly match.

In the end, running for 200 runs, if we consider the theoretical formula with the following equation:

$$1 - e^{-\frac{m^2}{2 \times 365}} \tag{2}$$

we can assert that it is accurate at $confidence\_level = 95\%$.

IV. EXTENSION



As extension, I decided to look at the generalized case of the birthday paradox: let us consider m elements chosen uniformly at random with repetition from a set of possible choices $n > m$. I will use only the uniform distribution because we do not know what is the real distribution of the possible extractions for different values of n. So the tasks become:

Task1: For each value of n evaluate the minimum number of selections such that a conflict occurs (i.e. from 2 random extractions we have the same selection).

Task2: For each value of n and for each value of m compute the probability that a conflict has occurred.

The IV represents the results of the experiment for the Task1. As we can notice, the confidence interval is very large and this is due by the high variability of the metric under analysis, on the other hand the theoretical curve almost matches the experimental one so the theoretical formula is accurate at $confidence\_level = 95\%$.

In the end for the second experiment, I took as reference 4 different values of possible selections and as we can see from IV it is reasonable that, the more n increases, the slower initial trend increases and so the more elements (m) I need to be sure to encounter at least a conflict. This is confirmed by the increasing trends of the generalized theoretical formula which is:

$$1 - e^{-\frac{m^2}{2n}} \tag{3}$$