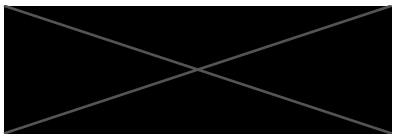
Birthday Paradox



***Abstract*— In probability theory, the birthday problem asks for the probability that, in a set of m randomly chosen people, at least two will share a birthday. The birthday paradox is that, counterintuitively, the probability of a shared birthday exceeds 50% in a group of only 23 people.**

I. Assumptions

The experiments are assumed to be identical independent distribution. The number of experiments is more than 32 to meet the central limit theorem. For both cases (uniform and real distribution of birthdays), we assume that the numbers of a year to be 366 including leap years. To compare to theorical results, we assume that p = 0.5 and n is large enough to use the theorical formulas.

Another assumption is for real distribution of birthday, in a sense that we use numerical results of US_births_1994-2003_CDC_NCHS. The probability distribution and the histogram of the day numbers which are generated by our function are as follows, respectively:

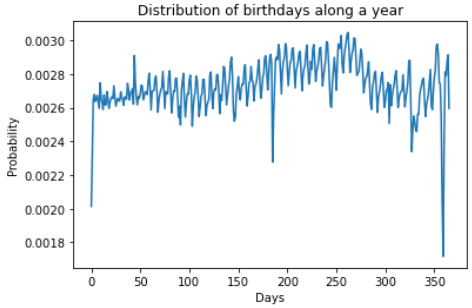


Fig 1. Real distribution for probability

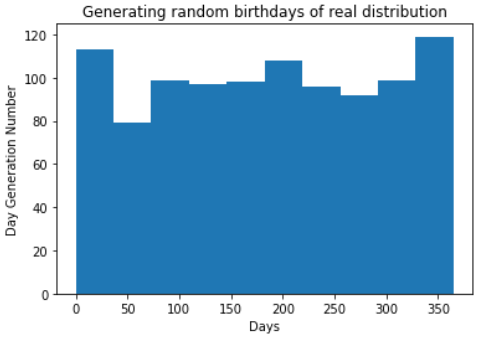


Fig 2. Generated histogram

II. Inputs

A. Average Number of People For a Conflict Case

In this case, simulation iteration number and the number n which is year number could be the inputs.

B. Probability of Birthday Conflict Case

Both the simulation iteration and number of student m to fill the classes are the inputs. However, it should be mentioned that the range for m should be at least more 2 to have a conflict and 100 is enough for upper bound to achieve a desired result. See figure below.

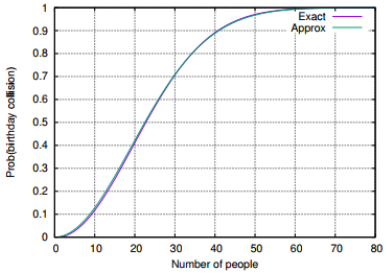


Fig 3. Exact relationship between m and p

III. Outputs

A. Average Number of People For a Conflict Case

There four outputs for this part:

- Average number for uniform case
- Average number for real case
- The confidence interval for uniform
- The confidence interval for real

B. Probability of Birthday Conflict Case

There are two plots for uniform and real cases: probability of birthday conflict in function of m and confidence interval.

IV. Data Structure

A. Average Number of People For a Conflict Case

- Two lists for storing student's birthdays

B. Probability of Birthday Conflict Case

- One list for storing the range number of students being in each class

- Two dictionaries to store probability of conflict for each m

V. MAIN ALGORITHM

There exist two functions which return number of required students in one class in order to have the conflict. The birthdays are randomly generated by uniform and real distributions (mentioned in assumptions). We use random seed to have the same number for multiple times running.

A. *Average Number of People For a Conflict Case*

We run the simulation for 1000 times and append the required number of students (to have a conflict) into a list. After finishing the simulation iteration, an average number of the list is computed and then compared to theorical ones.

Afterwards, in order to compute the confidence interval for mean, we need to calculate average and variance of computed numbers and then determine the interval. Since both average and std are estimated, we are required to use t-distribution with n-1 degree (1000-1) of freedom with alpha 0.5 for the interval.

As it can be seen from figures below, the results are satisfactory.

```
In theory for p=0.5 when n goes to infinity: for n = 365, m ≈ 22.3 and E[m] = 23.9
and our estimated E[m]s are:
***The results of the simulation***
Average number for uniform case: 23.304
Average number for real case: 23.84

The confidence interval for uniform case is: (22.529718683250273, 24.078281316749724)
The confidence interval for real case is: (23.07980502614472, 24.60019497385528)
```

Fig 4. Uniform case results

B. *Probability of Birthday Conflict Case*

In this part, simulation is run based on nested loops. The number of iteration for outer loop is related to how many different numbers are needed to obtain sensible result. We use a list of m students in which m is between 2 and 100. However, inner loop is about the number of classes to put students in, which is 20. So, total number of iterations is 2000.

The idea is that, for each m, we fill the classes with m number of students and then check whether there is a conflict in each class and set its flag to 1 in case of conflict. Finally, the probability for each m is calculated by the summation of the flags divided by the number of the classes.

Furthermore, to compute confidence interval for mean, we have two loops with size 100 and 20 which means the number of samples is 2000. we need to calculate average and variance of computed numbers and then determine the interval. Since both average and std are estimated, we should use t-distribution with n-1 degree (2000-1) of freedom with alpha 0.5 for the interval.

```
Uniform case: mean is 0.7581632653061224, std is 0.3405198514959255 and confidece
interval is(0.7430787660193187, 0.7732477645929262)
Real case: mean is 0.9290816326530612, std is 0.2146334641857016 and confidece
interval is(0.9195737043393071, 0.9385895609668153)
```

Fig 5. Real case result

In comparison with theorical plot, uniform distribution for birthday follows the theory one but the real distribution saturates faster than other two plots. The results show that we need less people to get a conflict in real case since in practice, the real birthdays' distribution is not uniform, we have some season with more birthday so the number 23 is less.

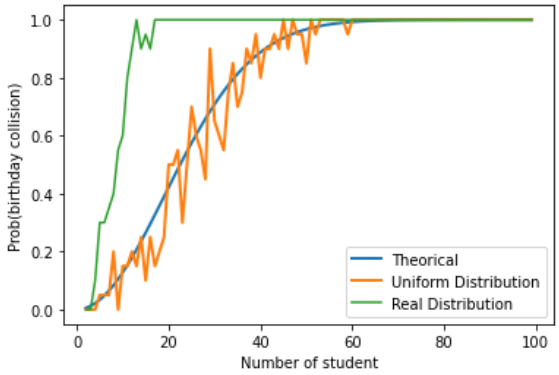


Fig 6. probability of conflict in function of m for all cases

VI. EXTENSION

One possible extension could be to use generalized version of this problem which is choosing arbitrary n instead 365(numbers days in a year) and compare the with theorical results. We run the simulation for uniform case in which n is 1000. The result shows that our simulation works reasonably since theorical m is close to the computed average.

```
In theory for p=0.5 when n goes to infinity: for n = 1000, m ≈ 36.998648623970034 is:
***The results of the simulation***
Average number for uniform case: 38.4
```

Fig 7. Extension results for uniform case