Birthday paradox

Introduction and Assumptions

The birthday paradox is a problem that analyzes the probability of having a conflict (probability of having the same birthday of at least a pair of students) by m objects (m in my exercise represent the student's birthday). Few assumptions:

- The number of different days that can be generated is N = 365 (don't consider leap years)
- The number of students is constrained to m < N
- Dataset used for the case-study regarding the realistic distribution case is correct.

Input/Output Metrics

input parameters:

1. MIN_/MAX_M: Left and right limits of students to consider for the computations
2. RUNS_PROB: Numbers of experiments considered for computing the empirical probability of each m
3. RUNS_E: Number of experiments considered to compute the average numbers of people to observe a conflicts
4. CONFIDENCE= Confidence level considered. In this case 0.95 with with respective alpha=0.05

Output metrics:

1. Computing the probability of conflicts and CI for every m considered
2. Compute the average number of people to observe a conflict together with the CI

(Compare the empirical results obtained with the theoretical ones computed from formulas)

Main data structures/Main algorithms

Data structures:

1. Class for realistic distribution: Class to handle the load and the use of the distribution retrieved from file.csv
2. Lists of metrics: used for saving the various outcomes required

Algorithms:

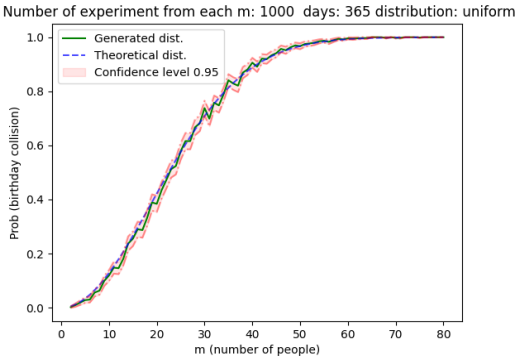
```
# Obtain the probability of conflicts for every m (this
# is run for  experiments=100 and 1000 for
# comparing the results)
initialize metrics_list
for m in range (MIN_M and MAX_M):
        count = 0
        for number of experiments:
                generate m birthday from a
        distribution (normal or realistic)
                if there is a conflict count + 1
        save on metrics list probability of
        conflict(count/runs) and CI of that m.
append on file the results

# Obtain the average number of students to observe
# a conflict (number of experiments= 10 , 50, 100)
initialize conflicts list
for number of experiments:
        initialize metrics_list
        initialize days_list
        for m in range (min_m and max_m):
                generate one "day" from a
distribution (normal or realistic).
                add to days_list  and check if there
is a conflict.
                if there is a conflict break and save
the numbers of days on the conflict list.
save on metrics list average number of days to
generate a conflict and CI of that probability.
append on file the results
```

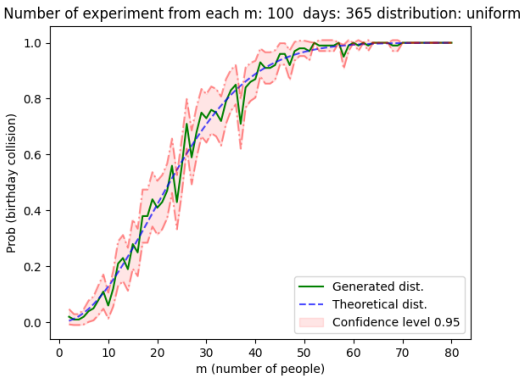
Extensions: As possible extensions that can act on the parameters of the problem I didn't find particularly interesting to change N or MAX_M. In my opinion we would have a behavior foreseeable in both cases. A thing that I found to be a possible extension, is simply to show differences in behavior as the number of experiments (in both the requests) increases.

Analysis and Conclusion

Empirical probability distribution:

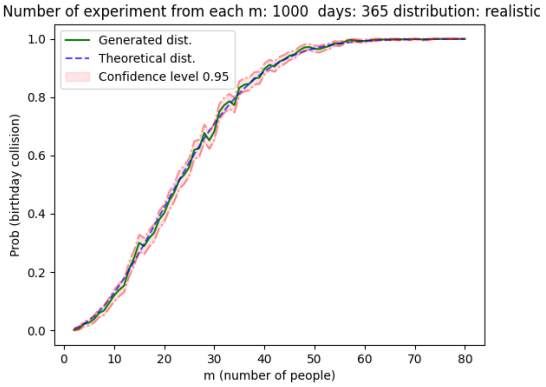
I have computed the probability distribution of “m” from 2 to 80, the curve shows the probability of collision in function of “m”. Below is shown the behavior of the empirical distribution compared with the theoretical formula for two different numbers of experiments (following the constraints that I made regarding m and N in the Assumptions).

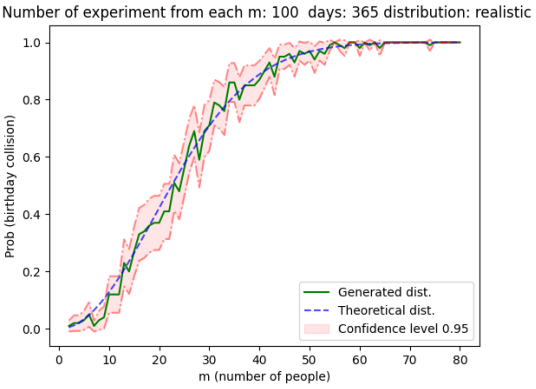




The theoretical distribution is computed using the formula : $p(\mathbf{n}) = 1 - e^{-\frac{m^2}{2\mathbf{n}}}$

As you can notice for $experiements \rightarrow +\infty$, as expected, the empirical curve overlaps with the theoretical. This behavior is also verified by looking at the decrease in variance of the confidence interval between the two graphs proportional to the increase of experiments (Confidence level settings explained before). I want to highlight that even for the minimal experiment (runs=100) the theoretical is roughly included in the C.I. A similar behavior is observed using the probability generated by the realistic distribution (pointed out by the two following graphs).





Average number of people to observe a conflict:

As you can recall , a conflict is when at least one pair of equal elements has been chosen (a collision). Here are the results together with the theoretical comparison:

```
** Average number of people to observe a conflict comparison**
* Runs: 10   Empirical average(uniform): 23.4   C.I.: [ 22.628, 24.172 ]
* Runs: 20   Empirical average(uniform): 24.3   C.I.: [ 23.495, 25.105 ]
* Runs: 40   Empirical average(uniform): 25.275   C.I.: [ 24.662, 25.888 ]
* Runs: 10   Empirical average(realistic): 23.4   C.I.: [ 22.539, 24.261 ]
* Runs: 20   Empirical average(realistic): 23.15   C.I.: [ 22.478, 23.822 ]
* Runs: 40   Empirical average(realistic): 23.5   C.I.: [ 22.724, 24.276 ]
** Theoretical average with N:365 is 23.881 **
```

As you can see, it is interesting to note that the results are more accurate considering the empirical averages and the C.I. of the realistic distribution. Following the hypothesis, despite the less accuracy of the actual averages from the uniform distribution sampling, the C.I. include the theoretical average computed using : $E[m] = 1.25\sqrt{n}$. The formula is used with the assumption of $n \rightarrow +\infty$ and a few approximations from a more general formula.

The entire analysis and conclusion of the results is carried out on Random seeds = [70, 71, 72] .
Here the results/comparisons reported for the analysis are specifically from the random seed = 70