Computer aided simulations and performance evaluation

Lab 7 - The Birthday Problem

Gabriele Cuni 277957

7.1 Assumptions

- case1: The birthday distribution is uniform and $n = 365$
- case2: The generalization when $n \neq 365$ and $m < n$
- Number of runs must be selected in order to have a smooth curve, given that we output a probability, the smoothness will be heavily reliant on the number of runs.
- $P(collision) = \frac{number\ of\ runs\ in\ which\ a\ collision\ occurs}{number\ of\ runs}$
- $pTheo = 1 - e^{\frac{-m^2}{2n}}$

7.2 Input Parameters

- **seed**: The initial seed
- **noRuns**: Number of runs for each experiment
- **noProbs**: Number of probabilities $P(collision)$ must be computed for any number of people.
- **m**: Number of people.
- **n**: Number of days, useful for future generalization without the need of changing the code.
- **ciCl**: Confidence level of the confidence interval

7.3 Output parameters

- **x_hat**: The simulated probability that at least one pair of equal elements has been chosen
- **pTheo**: It is the theoretical probability that at least one pair of equal elements has been chosen and it is an approximate formula.
- **ciUb**: Confidence interval upper bound
- **ciLb**: Confidence interval lower bound
- **ciRe**: Confidence interval relative error (I always output it, because I use it to assess if the number of run is good both with ciUb and ciLb, it doesn't mean I use it for plotting)

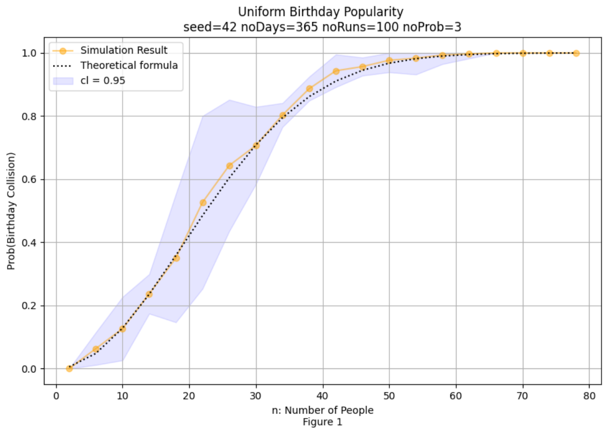
7.4 Data Structure

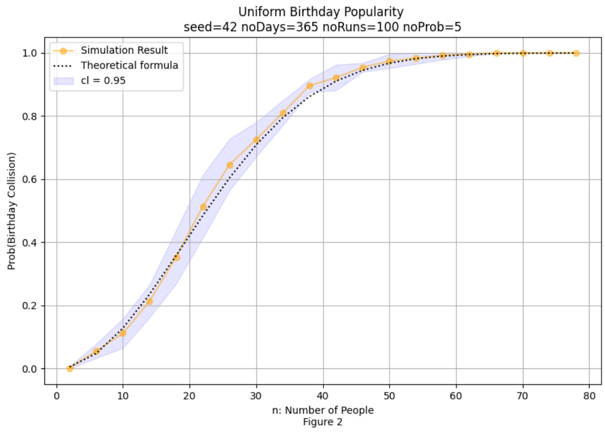
- **simulation_results** is a dictionary that stores all the simulation output parameters and input parameters in order to dump on file.
- **pArray** is a one dimensional floating point array which stores the $P(collision)$'s computed with the same input parameters in order to compute the confidence interval for that experiment.
- **runResult** is a one dimensional integer array that acts like a boolean one. It stores the result of any runs.
- **days** is a one dimensional integer array which supports the birthday problem simulation.

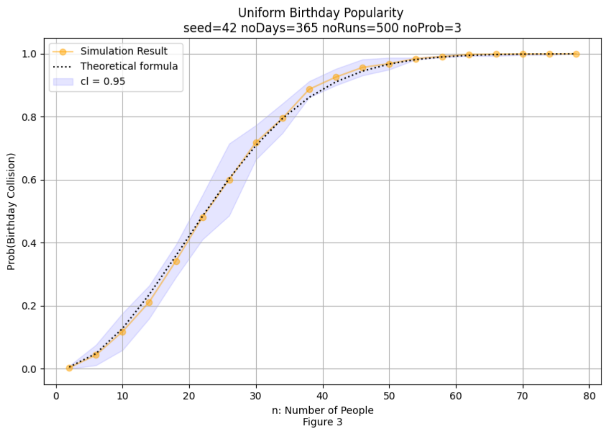
7.5 Main Algorithm

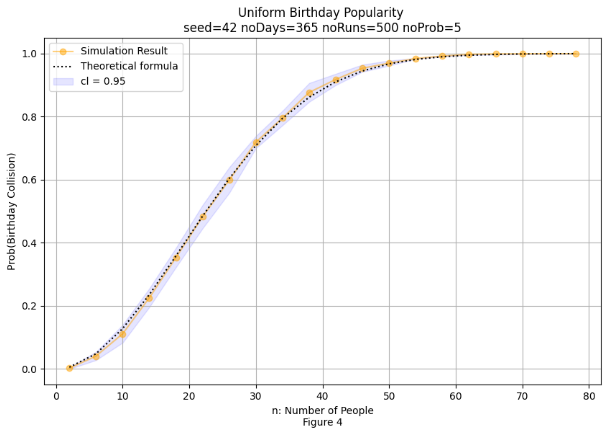
The main algorithm is all in the function **runSimulatorGeneralization.** Any time this function is called with a given input parameter set, it performs **noProb** experiments. For each experiment it performs **noRuns** run which are needed to computed the collision probability $P(collision)$. For each run the birthday conflict under uniform birthday popularity problem is simulated and $P(collision)$ is computed. **pArray** is sent as input parameter to the function **confidenceInterval** which computes all the output parameters but **pTheo** that is computed in the main algorithm. Finally the confidence interval upper and lower bound are clipped to be between zero and one.

7.6 Birthday conflict under uniform birthday popularity accuracy discussion









The four figures show that the simulation follows very well the theoretical approximated formula.
The comparison between figure one and two shows how increasing **noProb** makes the confidence interval shrinking as expected because **noProb** directly conditions the delta calculation.
The comparison between figure one and three shows how increasing **noRuns** makes the confidence interval shrinking, because more runs means the standard deviation of the experiment result **pArray** is lower and also this standard deviation conditions the delta calculation.
Figure four shows how increasing both **noRuns** and **noProb** makes the confidence interval shrinking much more.

7.7 Generalized version of the birthday problem accuracy discussion

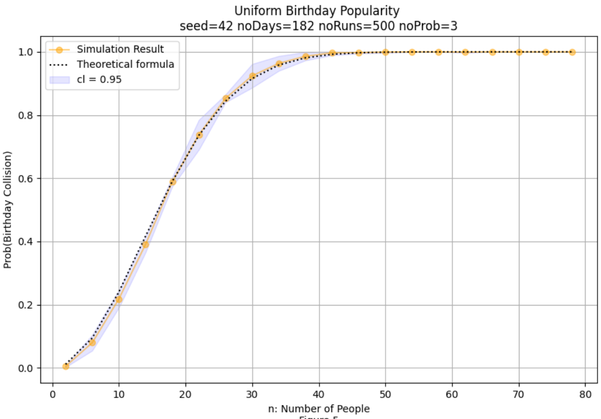


Figure 5

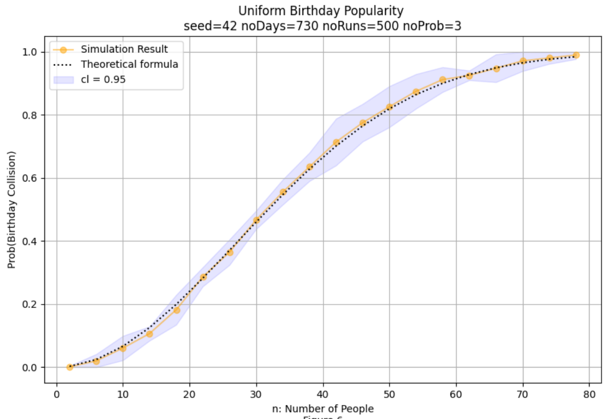


Figure 6

Figure five and six are the result of the generalization when the number of days is not 365. As before the simulation follows very well the approximated theoretical formula **pTheo**.

They also show how increasing the number of days makes the confidence interval expand. This behavior is, one again, because of the standard deviation of the experiment result **pArray**, which increases with the number of days.

Finally, as expected, fewer days make the $P(collision)$ converges to one faster than more days.