Lab report
Birthday problem

Francesco Pagano - 299266

1 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?

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.

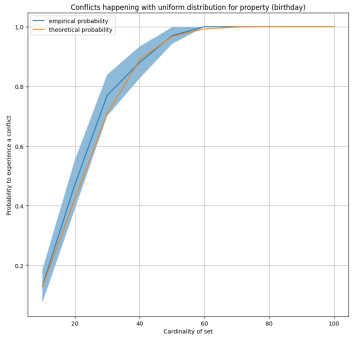
2 Structure of the simulator

The way the simulator is built takes into account the opportunity for such (easy) generalization by defining two simple classes:

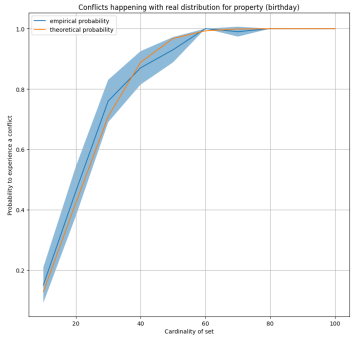
- Obj: stores the information relative to the instance's specific property, be it an instance of a uniformely distributed random variable or an instance of a random variable distributed according to a real distribution.
- Property: is a general way of encapsulating the specific property the software simulates, in this case a birthday date.

Since the simulator is dealing with birthdays, each Property's instance is upper-bounded by 365 (days of a non-leap year) in its value. By iterating among two distribution types (Uniform or Real[1]) and a fixed set of cardinalities, the simulator generates a set with the given cardinality and distribution and checks for the presence of any conflict. If so happens, results are then processed and logged.

[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

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

3 Results

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:

$$E[m] = \sqrt{\frac{\pi}{2} \times n} \approx 1.25m \tag{1}$$

And, 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)
Theoretial avg. cardinality	23.88	23.88

Table 1: Simulator's output metrics