CASE STUDY 045
[Python]
Proving the Birthday Paradox with 2016 Olympics Athletes

Difficulty Level: 3 of 3

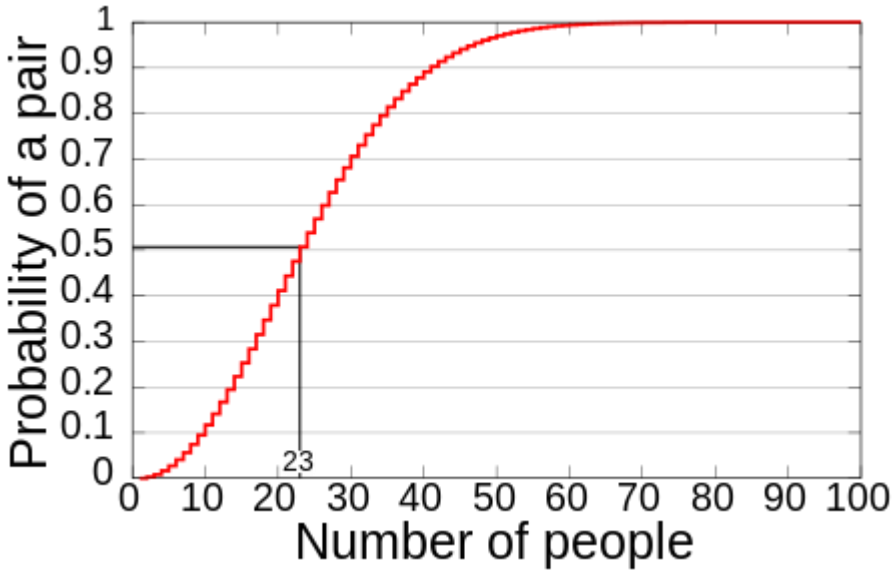
How would you answer the following questions?

1) In a classroom of 23 students, what is the probability of 2 of them share the same birthday (day and month)?

2) What about a classroom of 40 students?

The answers are **50%** and roughly **90%**, respectively.

How is that possible? It is strange, counter-intuitive, and completely true, but the birthday problem (birthday paradox) is a thing. It's only a "paradox" because our brains can't handle the compounding power of exponents. We expect probabilities to be linear and only consider the scenarios we're involved in (both faulty assumptions). But the comparison is made pair-wise. And there are a lot of comparisons in samples like that.



Let's use the birthdates of Rio 2016 Olympics' athletes to randomly select our sample and prove the converging probabilities with different sample sizes.

1) Load the dataset.

a) Convert the column dob (date of birthday) to the correct date data type.

2) Build a function called sample() that receives a number representing the sample size and returns the sample dataset.

3) Build a function called isSameMonthDay() that receives two birth dates and compare of them have the same day and month or not.

4) Build a function called oneRound() that receives a sample dataset, combines every possible pairs and returns True if at least one of the pair of athletes share the same day and month of birth.

5) Build a function called trial() that receives the number of trials and sample size, iterate over the number of trials sampling (sample), comparing(oneRound and isSameMonthDay) and print the percentage of times that we find shared birthdates.

6) Run the trial function with 100 trials and:

a) sample size of 23. You should expect to find shared birthdates 50% of the times.

b) sample size of 30. You should expect to find shared birthdates 70% of the times.

c) sample size of 40. You should expect to find shared birthdates 89% of the times.

d) sample size of 50. You should expect to find shared birthdates 97% of the times.

e) sample size of 60. You should expect to find shared birthdates 99% of the times.

I challenge you to reproduce this study over a group of people of your choice, like your own classroom or football team.

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file.