Birthday paradox probability*

- N birthdays
- k persons

What is the probability $\mathcal{P}$ that *at least* 2 *people* share the same birthday?

Theorem 1.

$$\mathcal{P} = 1 - \prod_{i=1}^{k-1}\left(1 - \frac{i}{N}\right)$$

and

$$\mathcal{P} \geq \frac{1}{2} \text{ if } k \geq \frac{1}{2} + \frac{1}{2}\sqrt{1 + 8N\ln 2}$$

For instance with $N = 365$, we have that $23 > \frac{1}{2}+\frac{1}{2}\sqrt{1+8*365\ln 2}$ (indeed: $2025 = 45^2 > 1 + 2920\ln 2$ with $\ln 2 < 0.69315$, and $(1+45)/2$ is fine), so 23 persons suffice to have more than one chance over two that two of them have the same birthday.

Proof.

1. At least 2 have the same birthday is the converse of all have a different birthday. Now for the latter, the first person can choose whatever she wants, the second one has $N-1$ chances over N to have a different birthday, the third one has $N-2$ chances over N to be different from the first two, etc. Overall

$$\mathcal{P} = 1 - \prod_{i=1}^{k-1}\left(\frac{N-i}{N}\right) = 1 - \prod_{i=1}^{k-1}\left(1 - \frac{i}{N}\right).$$

2. Consider $f(x) = e^x - 1 - x$, then its derivative is $f'(x) = e^x - 1$, negative before 0 and positive afterwards. Thus $f(x)$ is decreasing before 0 and increasing afterwards, where $f(0) = 1 - 1 - 0 = 0$. Thus $f(x) \geq 0$ for all x. In particular $f(-x) \geq 0$ and

$$\forall x,\ e^{-x} \geq 1 - x$$

3. The latter gives that $\prod_{i=1}^{k-1}\left(1 - \frac{i}{N}\right) \leq e^{\sum_{i=1}^{k-1} -\frac{i}{N}} = e^{-\frac{k(k-1)}{2N}}$ and thus that:

$$\mathcal{P} = 1 - \prod_{i=1}^{k-1}\left(1 - \frac{i}{N}\right) \geq 1 - e^{-\frac{k(k-1)}{2N}}.$$

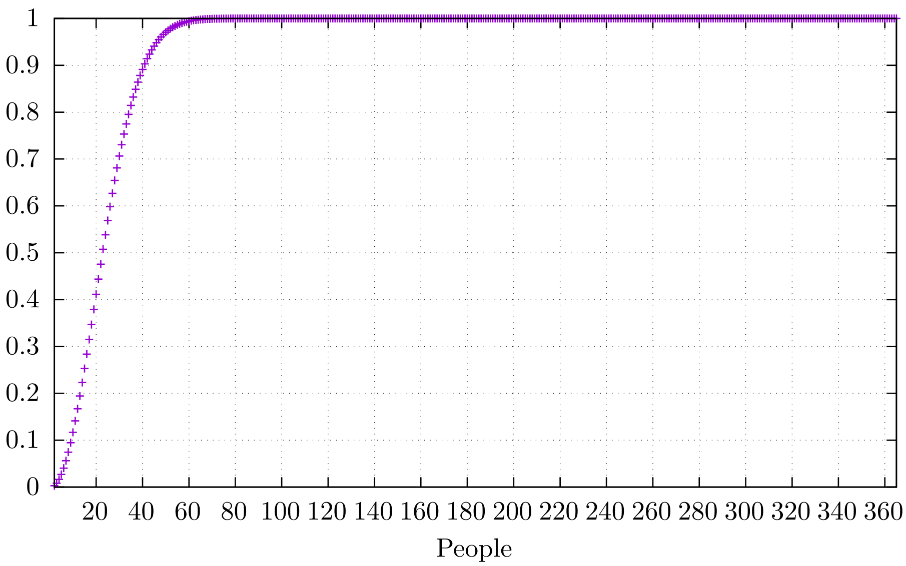
*Jean-Guillaume Dumas April 6, 2020

4. Now $e^{-\frac{k_0(k_0-1)}{2N}} = \frac{1}{2}$ if $k_0^2 - k_0 - 2N\ln 2 = 0$, that is if $k_0 = \frac{1}{2} + \frac{1}{2}\sqrt{1+8N\ln 2}$.

5. Finally, $1 - e^{-x}$ is increasing so $\mathcal{P} \geq \frac{1}{2}$ if $k \geq k_0$.

□

Figure 1: Probability of at least one common birthday



More upper bounds can be derived, for instance:

- $\frac{1}{2} + \frac{1}{2}\sqrt{1+8N\ln 2} < \frac{1}{2} + \frac{1}{2}\sqrt{1} + \frac{1}{2}\sqrt{8N\ln 2} \approx 1 + 1.17741\sqrt{N}$;
- $\frac{1}{2} + \sqrt{N}\sqrt{\frac{1}{4N} + 2\ln 2} < \frac{1}{2} + \sqrt{N}\sqrt{\frac{1}{4} + 2\ln 2} \approx 0.5 + 1.27918\sqrt{N}$;

and if N is large one also has the approximation:

$$k_0 \approx 0.5 + 1.17741\sqrt{N}.$$

Other useful approximations:

- $1 - e^{-ln(2)} = 1 - 1/2 = 1/2$, so if k is such that $\frac{k(k-1)}{2N} > ln(2) \approx 0.693$ then $\mathcal{P} \geq 1/2$;
- $1 - e^{-1/2} \approx 0.39$, so if N is large enough and k is close to $\sqrt{N}$, then $\mathcal{P}$ is close to $2/5$.