

The Birthday Problem

What is the smallest positive integer k so that the probability is > 0.5 that at least two people in a group of k people have the same birthday?

Ignore February 29.

Assume each birthday is equally likely.

The probability that k people all have different birthdays is

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - k + 1}{365}$$

which is

$$\frac{365!}{(365 - k)! \times (365)^k}.$$

Thus the probability that at least two of k people have the same birthday is

$$P(k) = 1 - \frac{365!}{(365 - k)! \times (365)^k}.$$

More generally, suppose we are given an integer-valued random variable with uniform distribution between 1 and n . Choose k instances of this random variable. What is the probability $P(n, k)$ that at least two of the k instances are the same value?

As for birthdays, we find

$$P(n, k) = 1 - \frac{n!}{(n-k)!n^k}.$$

Write this as

$$P(n, k) = 1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \times \dots \times \left(1 - \frac{k-1}{n}\right).$$

To estimate this, note that $1 - x \leq e^{-x}$ for all $x \geq 0$ and $1 - x \approx e^{-x}$ when x is small.

This gives

$$P(n, k) > 1 - e^{-1/n}e^{-2/n}e^{-3/n} \times \dots \times e^{-(k-1)/n}$$

$$P(n, k) > 1 - e^{-(1/n+2/n+3/n+\dots+(k-1)/n)}$$

$$P(n, k) > 1 - e^{-k(k-1)/(2n)}.$$

We will have $P(n, k) = 0.5$ when

$$0.5 = 1 - e^{-k(k-1)/(2n)}$$

or $2 = e^{k(k-1)/(2n)}$, that is, when

$$\ln 2 = k(k-1)/(2n).$$

When k is large, the percentage difference between k and $k-1$ is small, and we may approximate $k-1 \approx k$. This gives $k^2 \approx 2n \ln 2$ or

$$k \approx \sqrt{2(\ln 2)n} \approx 1.18\sqrt{n}.$$

For $n = 365$, we find

$$k \approx 1.18\sqrt{365} \approx 22.54,$$

or $k \approx 23$.

Suppose $H(M)$ is a hash function with m -bit output. There are $n = 2^m$ possible hash values.

If H is applied to k random inputs, the probability of finding a duplicate ($H(M) = H(M')$) is $P(2^m, k)$. The minimum number of k needed for a duplicate to occur with probability > 0.5 is about

$$k = 1.18\sqrt{2^m} = 1.18 \times 2^{m/2}.$$

The overlap between two sets

Given an integer random variable with uniform distribution between 1 and n , and two sets of k ($k \leq n$) instances of the random variable, what is the probability $R(n, k)$ that the two sets overlap, that is, at least one of the n values appears in both sets?

We assume k is small enough ($k < \sqrt{n}$) so that the k instances of the random variable in each set are all different. (A few duplicates won't hurt this analysis.)

The probability that one given element of the first set does not match any element of the second set is $(1 - \frac{1}{n})^k$.

The probability that the two sets are disjoint is

$$((1 - \frac{1}{n})^k)^k = (1 - \frac{1}{n})^{k^2}$$

so $R(n, k) = 1 - (1 - \frac{1}{n})^{k^2}$.

Using $1 - x \leq e^{-x}$, we get

$$R(n, k) > 1 - (e^{-1/n})^{k^2} = 1 - e^{-k^2/n}.$$

We will have $R(n, k) = 0.5$ when $\frac{1}{2} = 1 - e^{-k^2/n}$

or $2 = e^{k^2/n}$ or $\ln 2 = k^2/n$ or

$$k = \sqrt{(\ln 2)n} \approx 0.83\sqrt{n}$$

Suppose a hash function H with $n = 2^m$ possible values is applied to k random inputs to produce a set X and again to k additional random inputs to produce a set Y . What is the minimum value of k so that the probability is at least 0.5 of finding at least one match between the two sets, that is, $H(x) = H(y)$, where $x \in X$ and $y \in Y$? Using the approximation above, the minimum k is about

$$k = 0.83\sqrt{2^m} = 0.83 \times 2^{m/2}.$$