The Birthday Paradox

An Undergraduate-Level Probability Note

June 2026

1 A Warm-Up: When Is a Match Guaranteed?

Before computing probabilities, start with a simpler question:

How many people are needed to guarantee that at least two people have the same birthday?

Assume again that there are exactly 365 possible birthdays. This question is not about probability yet; it is about certainty.

Theorem 1 (Pigeonhole Principle)**.** *If N objects are placed into m boxes and $N > m$, then at least one box contains at least two objects.*

In the birthday problem, the objects are people and the boxes are the 365 possible birthdays. If there are 366 people, then 366 people are placed into only 365 birthday boxes. By the pigeonhole principle, at least one birthday box must contain two or more people.

Therefore, the smallest number of people needed to make a shared birthday certain is

$$\boxed{366}.$$

For 365 people, a shared birthday is not guaranteed, because it is still possible that each person has a different birthday. But with 366 people, a match is unavoidable.

Important contrast

For a 100% guarantee, we need 366 people. But the birthday paradox asks something subtler: how many people are needed before a shared birthday becomes *more likely than not*? Surprisingly, the answer is only 23 people.

2 The Question

Assume there are 365 possible birthdays and that each person's birthday is equally likely to be any of the 365 days. We ignore leap years and assume birthdays are independent.

The main question is:

For a group of n people, what is the probability that at least two people have the same birthday?

Let

$$A_n = \{\text{at least two people share a birthday}\}.$$

We want to compute

$$P(A_n).$$

The direct event A_n is slightly annoying to count because there can be exactly one pair, two pairs, triples, and many other collision patterns. The key trick is to use the complement.

$$A_n^c = \{\text{all } n \text{ birthdays are different}\}.$$

Then

$$P(A_n) = 1 - P(A_n^c).$$

Key idea

It is easier to count "no shared birthday" than "at least one shared birthday." This is a standard complement method in probability.

3 Exact Formula

For $n \leq 365$, the first person can have any birthday. The second person must avoid the first person's birthday, so there are 364 valid choices. The third person must avoid the first two birthdays, so there are 363 valid choices, and so on.

Therefore,

$$P(A_n^c) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - n + 1}{365}.$$

Equivalently,

$$P(A_n^c) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{365}\right).$$

Thus the exact probability of at least one shared birthday is

$$\boxed{P(A_n) = 1 - \prod_{k=0}^{n-1} \left(1 - \frac{k}{365}\right)}$$

for $0 \leq n \leq 365$.

If $n > 365$, then by the pigeonhole principle at least two people must share a birthday, so

$$P(A_n) = 1.$$

4 Why 23 People Is Enough

For $n = 23$,

$$P(A_{23}) = 1 - \prod_{k=0}^{22} \left(1 - \frac{k}{365}\right).$$

Numerically,

$$P(A_{23}) \approx 0.5073.$$

So with only 23 people, the probability of at least one shared birthday is slightly above 50%.

Table 1: Selected values of the birthday probability.

n	P(at least one match)	Interpretation
10	0.1169	About 11.7%
20	0.4114	About 41.1%
23	0.5073	Just above 50%
30	0.7063	About 70.6%
40	0.8912	About 89.1%
50	0.9704	About 97.0%
57	0.9901	Just above 99%

5 Visualizing the Probability Curve

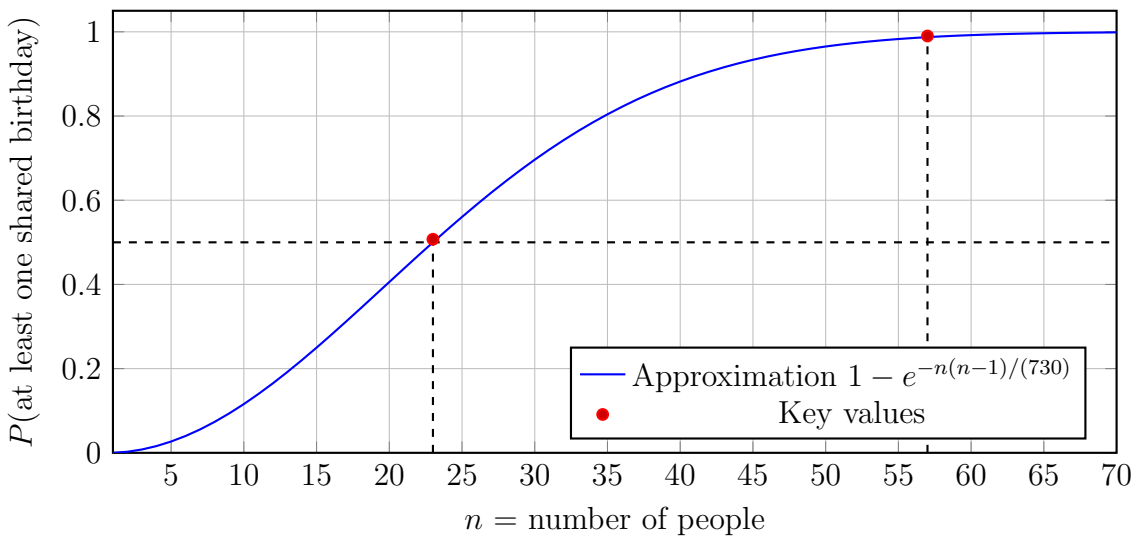


Figure 1: The birthday probability rises quickly because the number of possible pairs grows quadratically in n.

6 The Pair-Counting Intuition

The result feels surprising because people often compare everyone only to themselves. But in a group of n people, birthdays can match across many pairs.

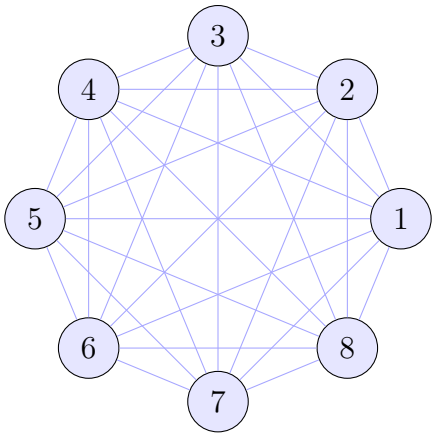
The number of pairs of people is

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

For $n = 23$,

$$\binom{23}{2} = 253.$$

So in a group of 23 people, there are 253 possible pairs that could share a birthday.



For 8 people, there are $\binom{8}{2} = 28$ possible pairs.

Figure 2: Even a small group creates many possible pairs. For n people, the number of pairs is $\binom{n}{2}$.

Each pair has probability 1/365 of matching. The expected number of matching pairs is therefore

$$E[\text{number of matching pairs}] = \binom{n}{2}\frac{1}{365}.$$

For $n = 23$,

$$\binom{23}{2}\frac{1}{365} = \frac{253}{365} \approx 0.693.$$

This does not mean the probability of a match is 0.693. Expected number and probability are not the same. But it explains why a match is already plausible: there are many opportunities for a collision.

Expected Number versus Probability

It is important to distinguish between a probability and an expected value.

A probability measures the chance that an event happens. Therefore, it must lie between 0 and 1:

$$0 \leq P(A_n) \leq 1.$$

For example, when $n = 23$,

$$P(A_{23}) \approx 0.5073.$$

This means that if we repeatedly formed many independent groups of 23 people, then about 50.73% of those groups would contain at least one shared birthday.

An expected value is different. It is a long-run average value of a random variable. In

this section, the random variable is

$$X = \text{number of matching birthday pairs.}$$

Thus $E[X]$ is not the probability of a match. It is the average number of matching pairs we expect to see across many repeated groups.

For $n = 23$,

$$E[X] = \binom{23}{2}\frac{1}{365} \approx 0.693.$$

This means that over many groups of 23 people, the average number of matching pairs per group is about 0.693. Some groups have no matching pairs, some groups have one matching pair, and some groups may have more than one matching pair. The average of those counts is about 0.693.

Probability vs. expected value

A probability is bounded by 1 because it measures the chance of one event. An expected value can be larger than 1 because it can measure the average number of occurrences of something.

For example, if $n = 50$, then

$$E[X] = \binom{50}{2}\frac{1}{365} = \frac{1225}{365} \approx 3.36.$$

This is larger than 1, but that is not a problem. It means that in a typical group of 50 people, the average number of matching birthday pairs is about 3.36. The probability of at least one match is still at most 1; in fact, it is about 0.9704.

So the interpretation is:

$$P(A_n) = \text{chance of at least one matching pair,}$$

while

$$E[X] = \text{average number of matching pairs.}$$

They are related, but they are not the same object.

7 Exponential Approximation

The exact formula is

$$P(A_n^c) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{365}\right).$$

To approximate this, use the fact that for small x,

$$\ln(1-x) \approx -x.$$

Why $\ln(1-x) \approx -x$ for small x

This approximation comes from a Taylor expansion. In general, if a function f is sufficiently differentiable near 0, then its Taylor expansion around 0 is

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f^{(3)}(0)}{3!}x^3 + \cdots.$$

For $f(x) = \ln(1-x)$, this gives the Taylor expansion

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots, \qquad |x| < 1.$$

Here is one quick derivation. For $|t| < 1$, the geometric series gives

$$\frac{1}{1-t} = 1 + t + t^2 + t^3 + \cdots.$$

Integrating both sides from 0 to x gives

$$\int_0^x \frac{1}{1-t}\,dt = \int_0^x (1 + t + t^2 + t^3 + \cdots)\,dt.$$

The left-hand side is

$$-\ln(1-x),$$

and the right-hand side is

$$x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots.$$

Therefore,

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots.$$

When x is small, the terms $x^2, x^3, \ldots$ are much smaller than x, so the first-order approximation is

$$\ln(1-x) \approx -x.$$

More precisely, for $0 \le x < 1$,

$$|\ln(1-x) + x| = \sum_{r=2}^{\infty} \frac{x^r}{r} \le \frac{x^2}{2(1-x)}.$$

Thus the error is of order x^2 when x is small.

In the birthday problem with $n = 23$, the largest value of $k/365$ in the product is $22/365 \approx 0.0603$, so these terms are small enough for the approximation to be useful.

Take logarithms:

$$\ln P(A_n^c) = \sum_{k=0}^{n-1} \ln\left(1 - \frac{k}{365}\right).$$

Using $\ln(1-x) \approx -x$,

$$\ln P(A_n^c) \approx -\sum_{k=0}^{n-1} \frac{k}{365}.$$

Since

$$\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2},$$

we get

$$\ln P(A_n^c) \approx -\frac{n(n-1)}{2 \cdot 365}.$$

Exponentiating both sides gives

$$P(A_n^c) \approx e^{-n(n-1)/(2\cdot 365)}.$$

Therefore,

$$\boxed{P(A_n) \approx 1 - e^{-n(n-1)/(2\cdot 365)}.}$$

For $n = 23$,

$$1 - e^{-23\cdot 22/(2\cdot 365)} \approx 0.5000.$$

The exact value is about 0.5073, so the approximation is already quite good.

8 Finding the 50% Threshold

Using the approximation,

$$P(A_n) \approx 1 - e^{-n(n-1)/(730)}.$$

To find where this is around $1/2$, solve

$$1 - e^{-n(n-1)/730} = \frac{1}{2}.$$

Then

$$e^{-n(n-1)/730} = \frac{1}{2}.$$

Taking logarithms,

$$-\frac{n(n-1)}{730} = \ln\left(\frac{1}{2}\right) = -\ln 2.$$

So

$$n(n-1) = 730 \ln 2.$$

Since

$$730 \ln 2 \approx 505.999,$$

we need

$$n(n-1) \approx 506.$$

Because

$$23 \cdot 22 = 506,$$

we get

$$\boxed{n \approx 23.}$$

This explains why the threshold is 23.

9 A General Version

Suppose there are m equally likely categories instead of 365 days. For example, m could be the number of possible hash values, ID numbers, or bins.

For $n \leq m$,

$$P(\text{at least one collision}) = 1 - \prod_{k=0}^{n-1} \left(1 - \frac{k}{m}\right).$$

Using the same approximation,

$$P(\text{at least one collision}) \approx 1 - e^{-n(n-1)/(2m)}.$$

The 50% threshold approximately satisfies

$$1 - e^{-n(n-1)/(2m)} = \frac{1}{2}.$$

Thus

$$n(n-1) \approx 2m \ln 2.$$

For large m, this gives roughly

$$\boxed{n \approx \sqrt{2m \ln 2}.}$$

This square-root behavior is important: collisions become likely when the number of samples is around the square root of the number of possible categories.

Main takeaway

If there are m possible outcomes, you do not need close to m samples to see a collision. You only need around $\sqrt{m}$ samples. This is the core intuition behind the birthday paradox.

10 Expectation and Random Variables

This section makes the expected-value calculation self-contained.

Formal definition of expectation

Let Y be a discrete random variable with possible values in a set S. The expectation, or expected value, of Y is

$$\boxed{E[Y] = \sum_{y \in S} y \, P(Y = y)}$$

provided the sum is finite or converges absolutely. Intuitively, expectation is the long-run average value of the random variable over many repetitions of the same experiment.

A particularly useful special case is an indicator random variable. For an event B, define

$$\mathbf{1}_B = \begin{cases} 1, & \text{if } B \text{ happens,} \\ 0, & \text{if } B \text{ does not happen.} \end{cases}$$

Then, using the definition of expectation,

$$E[\mathbf{1}_B] = 0 \cdot P(B^c) + 1 \cdot P(B) = P(B).$$

So the expectation of an indicator variable is exactly the probability of the event it indicates.

We also use linearity of expectation. For finitely many random variables $Y_1, \ldots, Y_r$,

$$\boxed{E[Y_1 + \cdots + Y_r] = E[Y_1] + \cdots + E[Y_r].}$$

This formula does not require independence.

Application to the birthday problem

Let X be the number of matching birthday pairs in a group of n people. For each pair (i, j), define

$$I_{ij} = \begin{cases} 1, & \text{if people } i \text{ and } j \text{ share a birthday,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the total number of matching pairs is the sum of all these indicators:

$$X = \sum_{1 \le i < j \le n} I_{ij}.$$

For a fixed pair (i, j), person j matches person i with probability $1/365$, so

$$P(I_{ij} = 1) = \frac{1}{365}.$$

Using the indicator formula,

$$E[I_{ij}] = P(I_{ij} = 1) = \frac{1}{365}.$$

Therefore, by linearity of expectation,

$$E[X] = E\left[\sum_{1 \le i < j \le n} I_{ij}\right] = \sum_{1 \le i < j \le n} E[I_{ij}] = \sum_{1 \le i < j \le n} \frac{1}{365}.$$

There are $\binom{n}{2}$ pairs, so

$$\boxed{E[X] = \binom{n}{2}\frac{1}{365}.}$$

For $n = 23$,

$$E[X] = \binom{23}{2}\frac{1}{365} = \frac{253}{365} \approx 0.693.$$

This means that the long-run average number of matching pairs is about 0.693 per group of 23 people. It does not mean that the probability of at least one match is 0.693.

This is a nice undergraduate-level use of indicator random variables. We did not need the indicators to be independent in order to use linearity of expectation.

11 Key Takeaways

The birthday paradox is surprising, but the mathematics is not mysterious.

1. By the pigeonhole principle, 366 people guarantee a shared birthday when there are 365 possible birthdays.

2. Count the complement: no two people share a birthday.

3. The exact formula is
$$P(A_n) = 1 - \prod_{k=0}^{n-1}\left(1 - \frac{k}{365}\right).$$

4. For $n = 23$, this is approximately 0.5073.

5. The number of possible pairs is $\binom{n}{2}$, which grows like $n^2/2$.

6. The probability $P(A_n)$ means the chance of at least one match, while $E[X]$ means the average number of matching pairs. Formally, for a discrete random variable Y, $E[Y] = \sum_y yP(Y = y)$. The expected value can be larger than 1 because it is not a probability.

7. The approximation
$$P(A_n) \approx 1 - e^{-n(n-1)/(730)}$$
comes from the first-order Taylor approximation $\ln(1 - x) \approx -x$ and explains why the 50% threshold occurs around $n = 23$.

8. In general, collisions become likely when n is around $\sqrt{m}$, not around m.