How many people must be in a room before the probability of at least two sharing a birthday is greater than 50%?

Johnny Carson’s stab at it ...


Also Feb 7, Feb 8

Johnny Carson, 1925-2005
Ed McMahon, 1923-2009
**SOLUTION**

- \( P(\text{at least 2 share}) = 1 - P(\text{no one shares}) \)

- To find \( P(\text{no one shares}) \), do probabilities of choosing a non-matching birthday:
  
  \[
  P(\text{no one shares}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{n-1}{365}
  \]

- So \( P(\text{at least 2 share}) = 1 - \frac{365!}{(365-n)!365^n} \cdot \frac{n!}{365^n} \)

**STANDARD RESULT**

- \( P(\text{at least 2 share}) = 1 - \frac{365^P_n}{365^n} \)

<table>
<thead>
<tr>
<th>( n )</th>
<th>( P(\text{sharing}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.71%</td>
</tr>
<tr>
<td>10</td>
<td>11.69%</td>
</tr>
<tr>
<td>15</td>
<td>25.29%</td>
</tr>
<tr>
<td>20</td>
<td>41.14%</td>
</tr>
<tr>
<td>25</td>
<td>56.87%</td>
</tr>
<tr>
<td>30</td>
<td>70.63%</td>
</tr>
<tr>
<td>35</td>
<td>81.44%</td>
</tr>
<tr>
<td>40</td>
<td>89.12%</td>
</tr>
<tr>
<td>45</td>
<td>94.10%</td>
</tr>
<tr>
<td>50</td>
<td>97.04%</td>
</tr>
</tbody>
</table>

- For \( n = 23 \), \( P(\text{at least 2 share}) = 50.73\% \)

**REACTIONS?**

The birthday problem used to be a splendid illustration of the advantages of pure thought over mechanical manipulation ...

... what calculators do not yield is understanding, or mathematical facility, or a solid basis for more advanced, generalized theories.

--- Paul Halmos (1916-2006), *I Want to Be a Mathematician*, 1985
MY QUESTIONS? (MY OUTLINE)
- Is this only a curiosity?
- How was this computed historically?
- What is the underlying distribution?
- Generalizations?
- Applications?

NOT THEORETICAL?
[After solving the Birthday Problem]

“The next example in this section not only possesses the virtue of giving rise to a somewhat surprising answer, but it is also of theoretical interest.”

- Sheldon Ross, *A First Course in Probability*, 1976

NOVEL? SO-CALLED?
[After developing a formula for the probability that no point appears twice when sampling with replacement]

“A novel and rather surprising application of [the formula] is the so-called birthday problem.”

- Hoel, Port, & Stone, *Introduction to Probability Theory*, 1971
Richard von Mises (1883-1953) often gets credit for posing it in 1939, but...

...he sought the expected number of repetitions as a function of the number of people.

Ball & Coxeter included it in their 11th edition of *Mathematical Recreations and Essays*, published in 1939.

They gave credit to Harold Davenport (1907-1969), who shared it about 1927.

But he did not think he was the originator either.

So how did they do the computations back then?

Paul Halmos: “The birthday problem used to be a splendid illustration of the advantages of pure thought over mechanical manipulation...”
A SERIES APPROXIMATION

- P(no one shares)

\[
P = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-(n-1)}{365}
\]

- Recall the Taylor Series:

\[
e^x = 1 + x + \frac{x^2}{2!} + \cdots
\]

- First-order approximation: P(no one shares)

\[
e^{-\frac{n(n-1)}{2 \times 365}} \approx e^{-\frac{n(n-1)}{2 \times 365}} 
\]

\[
e^{-\frac{n(n-1)}{2 \times 365}} = 1 \times e^{-\frac{3}{506} - \frac{(n-1)}{365}}
\]

\[
e^{-\frac{n(n-1)}{2 \times 365}} 
\]

AN APPROXIMATE SOLUTION

Solving P(at least 2 share) > 0.5

\[1 - e^{-\frac{n(n-1)}{2 \times 365}} > 0.5 \]

\[e^{-\frac{n(n-1)}{2 \times 365}} < 0.5 \]

\[-\frac{n(n-1)}{2 \times 365} < \ln(0.5) \]

\[n(n-1) > (365) \ln(0.5) \approx 505.997 \]

\[n^2 - n - 506 > 0 \]

\[n > \frac{1 + \sqrt{1 + 4(506)}}{2} = \frac{1 + \sqrt{2025}}{2} = 23 \]

OR MORE GENERALLY...

\[n(n-1) > d \ln 4 \]

\[n^2 - n - d \ln 4 > 0 \]

\[n > \frac{1 + \sqrt{1 + 4d \ln 4}}{2} \]

\[n > 0.5 + \sqrt{0.25 + d \ln 4} \]

\[n \approx 0.5 + 1.177\sqrt{d} \]
So $n$ is roughly proportional to the square root of $d$, with proportionality constant about 1.2.

$$n = 1.2\sqrt{365} \approx 22.93$$

$$n \approx 0.5 + 1.177\sqrt{d}$$

(Pat’s Blog attributes the factor 1.2 to Persi Diaconis)

---

**A HEURISTIC EXPLANATION**

- $P(2$ people not sharing) $= \frac{364}{365}$
- Among $n$ people, number of pairs $= \binom{n}{2} = n(n-1)/2$
  - For 23 people, there are 253 pairs!
  - That’s 253 chances of a birthday match!

- $P(2$ of $n$ people sharing) $= 1 - \left(\frac{364}{365}\right)^{\binom{n}{2}}$
  - Alarm Bells! Independence assumed!

---

**IGNORING THE ALARM**

- $P(2$ of $n$ people sharing) $= 1 - \left(\frac{364}{365}\right)^{\binom{n}{2}}$
  - Alarm Bells temporarily suppressed!

- Solving: $1 - \left(\frac{364}{365}\right)^{\frac{n(n-1)}{2}} > 0.5$

$$n^2 - n - 505.3 > 0$$

- gives $n > 22.98$
Underlying Distribution?

- Discrete or Continuous?
  - Binomial
    - Distribution: Discrete
    - Random Variable: No. of Successes
    - Assumptions: Fixed n, constant p, independent
  - Hypergeometric
    - Distribution: Discrete
    - Random Variable: No. of Successes
    - Assumptions: Fixed n, p varies, dependent
  - Poisson
    - Distribution: Discrete
    - Random Variable: No. of Successes
    - Assumptions: Fixed time, constant λ, independent
  - Geometric
    - Distribution: Discrete
    - Random Variable: First Success
    - Assumptions: n varies, constant p, independent
  - Negative Binomial
    - Distribution: Discrete
    - Random Variable: k-th Success
    - Assumptions: n varies, constant p, independent
  - Multinomial
    - Distribution: Discrete
    - Random Variable: No. of Each Type
    - Assumptions: Fixed n, constant p, independent

These answer the wrong question.

Balls in Bins

- Place n indistinguishable balls into d distinguishable bins.
- Balls → People, Bins → Birthdays
- When does one bin contain two (or more) balls?
  - Bernoulli ???
  - Multinomial ???

First Match

- Want “first match”!
  - AKA: occupancy problem, collision counting
THE FIRST MATCH PATTERN
- How many trials to get the first match?
- Variables:
  - $x =$ number of trials to get the first match
  - $d =$ number of equally likely options (birthdays)
- Pattern:
  - Not Not Not ... Not Match
  - $\frac{d}{d} \times \frac{d-1}{d} \times \frac{d-2}{d} \times \cdots \times \frac{d-(x-2)}{d} \times \frac{x-1}{d}$

THE FIRST MATCH PROBABILITY
- Pattern:
  - Not Not Not ... Not Match
  - $\frac{d}{d} \times \frac{d-1}{d} \times \frac{d-2}{d} \times \cdots \times \frac{d-(x-2)}{d} \times \frac{x-1}{d}$
- Probability:
  - $P(X = x) = \frac{(x-1)}{d}$ for $x \in \{2, 3, 4, \ldots, d + 1\}$

PROBABILITY DISTRIBUTIONS
- **MEAN**
  - $E(X) = 24.6166$
- **MODE**
  - The probability of obtaining the first match is highest when $n = 20$
- **MEDIAN**
  - When $n=23$, the probability of at least one match first exceeds 50%
THE MODE ANALYTICALLY

- The PDF is discrete, so avoid calculus.
- For which \( x \) is \( P(x) > P(x+1) \)?
- For \( d=365 \), we get \( x > 19.61 \), so 19 or 20
- Approximating: \( x > \sqrt{d} \) for large \( d \)

EXACT FORMULA!

THE MEAN (EXPECTED VALUE)

- Since \( P(X = x) = \frac{(x-1)!}{d!} \) for \( x \in \{2, 3, 4, ..., d+1\} \)
- \( E(X) = \sum_{x=2}^{d+1} \frac{x-1}{d} \)

which is related to Ramanujan's \( Q \)-function, specifically \( E(X) = 1 + Q(d) \)
- which is asymptotic:
  \[
  E(X) \approx 1 + \frac{\pi d}{\sqrt{24}} - \frac{1}{3} + \frac{1}{12} \sqrt{\frac{\pi}{2d}} - \frac{4}{135d} + ...
  \]
- and for \( d=365 \), we get:
  \[
  E(X) = \frac{365\pi}{2\sqrt{24}} + \frac{1}{3} + \frac{1}{12} \sqrt{\frac{\pi}{730}} - \frac{4}{49275} + ... \approx 24.6166
  \]

BIRTHDAY PROBLEM ON JUPITER?

- A Jovian day = 9.9259 Earth-hours
- A Jovian year = 11.86231 Earth-years
- \( 1 \text{ Jyr} = 11.86231 \text{ Eys} \) \( \left( \frac{365.24 \text{ Edays}}{1 \text{ Eyr}} \right) \left( \frac{24 \text{ Ehrs}}{1 \text{ Eday}} \right) \left( \frac{1 \text{ Jday}}{1 \text{ Jyr}} \right) \) = 10476 Jdays
- \( P(X = x) = \frac{(x-1)!}{10476^{x-1} \text{ P}_{10476}} \) for \( x \in \{2, 3, 4, ..., 10477\} \)
- Both \( 10476^{25} \) and \( 10476^25 \) cause a TI-84 overflow!
JOVIAN MEDIAN ESTIMATES

- Proportionality?
  \[ n \approx 1.2\sqrt{10476} \approx 122.823 \]

- Heuristically?
  \[ 1 - \left( \frac{10475}{10476} \right)^{n(n-1)} > 0.5 \]
  \[ n(n-1) > \frac{\ln 4}{\ln \left( \frac{10476}{10475} \right)} \]
  \[ n \approx 121.009 \]

MORE JOVIAN MEDIAN ESTIMATES

- Series Approximation:
  \[ \frac{10476}{10476} \times \frac{10475}{10476} \times \cdots \times \frac{10476-n}{10476} < 0.5 \]
  \[ n(n-1) > (10476)\ln 4 \]
  \[ n \approx 121.012 \]

- Mathematica: for \( n = 121 \),
  \[ P(\text{match}) = 50.13\% \]

JOVIAN BIRTHDAY AVERAGES

- Mean:
  \[ E(X) = \sqrt{\frac{\pi d}{2}} + \frac{2}{3} + \frac{1}{12} \sqrt{2\pi} - \frac{4}{135d} + \cdots = 128.947 \]

- Median:
  \[ n = \frac{1 + \sqrt{1+4d \ln 3}}{2} = 121.012 \]

- Mode:
  \[ x = \frac{1 + \sqrt{1+4d}}{2} = 102.854 \]
**DIGITAL SIGNATURES**

Birthday Attack: Find a good message and a fraudulent message with the same value of \( F(\text{Message}) \).

**The hash function \( F \) is not one-to-one!**

---

**MATCHING BETWEEN 2 SETS**

Given \( n \) good messages, \( n \) fraudulent messages, and a hash function with \( d \) possible outputs, what is the probability of a match between 2 sets?

- Assume \( n \ll d \), so \( n \) outputs are probably all different.

- \( P(\text{message 1 in set 1 doesn’t match set 2}) = 1 - \frac{n}{d} \)

---

**MATCHING BETWEEN 2 SETS**

- \( P(\text{no message in set 1 matches anything in set 2}) = \left( 1 - \frac{n}{d} \right) = \left( e^{-\alpha d} \right) = e^{-\alpha/d} \)

- 50% probability of at least one match between the sets: 
  \[ 1 - e^{-\alpha/d} > 0.50 \]
  \[ n^2 > d \ln 2 \]
  \[ n = 0.83 \sqrt{d} \]

- Probability \( p \) of at least one match between the sets: 
  \[ n = \sqrt{d \ln \left( \frac{1}{1-p} \right)} \]
ATTACKING A 16-BIT HASH

- With a 16-bit hash function, there are $2^{16} = 65,536$ possible outputs
- 50% probability of a match with messages: $n \approx 0.83 \sqrt{65536} \approx 213$
- 1% probability of a match with messages $n \approx \sqrt{65536 \ln \left( \frac{1}{0.99} \right)} \approx 26$

IMPROVING THE DEFENSE

Increase the size of the hash function:
- With 64-bits, $2^{64} = 18,446,744,073,709,551,616$ outputs are possible
- 50% probability of a match with 3.5 billion messages (11 messages per US citizen)
- 1% probability of a match with 431 million messages

ANSWERS TO MY QUESTIONS

- Is this only a curiosity?
  - No, it’s part of a class of problems
- How was this computed historically?
  - Using approximations (including Taylor Series)
- What is the underlying distribution?
  - “First Match” Distribution
- Generalizations?
  - Vary days in a year (other planets)
- Applications?
  - Digital signatures
THANK YOU!

- Steven J. Wilson
  Johnson County Community College
  Overland Park, KS 66210
  swilson@jccc.edu

- A PDF of the presentation is available at:
  http://www.milefoot.com/about/presentations/birthday.pdf