AND YOU THOUGHT THE BIRTHDAY PROBLEM WAS ONLY A CURIOSITY?

Steven J. Wilson
Johnson County Community College
AMATYC 2014, Nashville, Tennessee

THE BIRTHDAY PROBLEM

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

JOHNNY CARSON

Johnny Carson's stab at it …

http://www.cornell.edu/video/the-tonight-show-with-johnny-carson-feb-6-1980-excerpt

Also Feb 7, Feb 8

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

SOLUTION

- P(at least 2 share) = 1 – P(no one shares)
- To find P(no one shares), do probabilities of choosing a non-matching birthday:

$$= \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \times \left(1 - \frac{n-1}{365}\right)$$

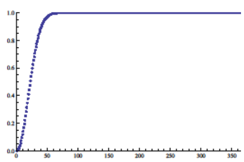
$$= \frac{365 \times 364 \times 363 \times \cdots \times (365-(n-1))}{365^n}$$

$$= \frac{365!}{(365-n)! \times 365^n} = \frac{{}_{365}P_n}{365^n}$$

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

STANDARD RESULT

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



n	P(sharing)
5	2.71%
10	11.69%
15	25.29%
20	41.14%
25	56.87%
30	70.63%
35	81.44%
40	89.12%
45	94.10%
50	97.04%

- For n = 23,
  P(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

THE LEGENDARY SOURCE

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.

BUT EVEN EARLIER …

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.

COMPUTING BEFORE CALCULATORS?

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)

$$= \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \times \frac{365-(n-1)}{365}$$

$$= \left(1-\frac{0}{365}\right) \times \left(1-\frac{1}{365}\right) \times \left(1-\frac{2}{365}\right) \times \cdots \times \left(1-\frac{n-1}{365}\right)$$

- Recall the Taylor Series:

$$e^x = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!} + \cdots$$

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

$$= e^{-0/365} \times e^{-1/365} \times e^{-2/365} \times \cdots \times e^{-(n-1)/365}$$

$$= 1 \times e^{-(1+2+\cdots+(n-1))/365}$$

$$= e^{-(n(n-1)/2)/365}$$

AN APPROXIMATE SOLUTION

Solving P(at least 2 share) > 0.50

$$1 - e^{-(n(n-1)/2)/365} > 0.5$$

$$e^{-(n(n-1)/2)/365} < 0.5$$

$$\frac{-n(n-1)}{2(365)} < \ln(0.5)$$

$$\boxed{n(n-1) > (365)\ln 4} \approx 505.997$$

$$n^2 - n - 506 > 0$$

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

OR MORE GENERALLY…

$$n(n-1) > d \ln 4$$

$$n^2 - n - d \ln 4 > 0$$

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

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

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

… OR WITH MENTAL MATH

$$n > 0.5 + \sqrt{0.25 + d \ln 4}$$
$$n \approx 0.5 + 1.177\sqrt{d}$$
$$n \approx 1.2\sqrt{d}$$

So *n* is roughly proportional to the square root of *d*, with proportionality constant about 1.2.

$$n \approx 1.2\sqrt{365} \approx 22.93$$
$$n \approx 0.5 + 1.177\sqrt{365} \approx 22.99$$

(Pat'sBlog 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 = ${}_nC_2 = \frac{n(n-1)}{2}$
  - For 23 people, there are 253 pairs!
  - That's 253 chances of a birthday match!
- P(2 of n people sharing) ≈ $1 - \left(\frac{364}{365}\right)^{\frac{n(n-1)}{2}}$
  - Alarm Bells! Independence assumed!

IGNORING THE ALARM

- P(2 of n people sharing) ≈ $1 - \left(\frac{364}{365}\right)^{\frac{n(n-1)}{2}}$
  - 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?

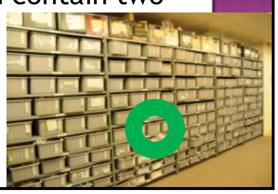
- Discrete or Continuous?

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

- These answer the wrong question

BALLS IN BINS

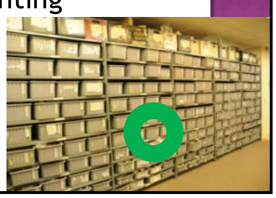
- Place *n* indistinguishable balls into *d* distinguishable bins.
- Balls → People, Bins → Birthdays
- When does one bin contain two (or more) balls?



- 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:

$$P(X=x)=\frac{(x-1)}{d^x}\left({}_dP_{x-1}\right)\quad \text{for}\quad x\in\{2,3,4,\ldots,d+1\}$$

PROBABILITY DISTRIBUTIONS

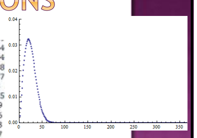
MEAN

E(X) ≈ 24.6166

MODE

The probability of obtaining the first match is highest when n = 20

	PDF	CDF
10	0.022324343821910956	0.11694817771107764
11	0.024193200610655406	0.14114137832173304
12	0.02588341051633133	0.16702478883806438
13	0.027385486394365006	0.19441027523242937
14	0.02869223677254361	0.223102512004973
15	0.029798807758713363	0.25290131976368635
16	0.03070268548916357	0.2836040052528499
17	0.03140366004371069	0.3150076652965606
18	0.031903752575228685	0.3469114178717893
19	0.03220710815974737	0.3791185260315367
20	0.0323198575490433	0.41143838358058
21	0.03224995158462575	0.4436883351652057
22	0.03200697249734433	0.4756953076625501
23	0.03160192666143534	0.5072972343239854
24	0.031047023590543385	0.5383442579145288
25	0.030355446054935092	0.5686997039694639
26	0.029541116166475076	0.5982408201359389
27	0.028618462127302978	0.626859282263242
28	0.027602190079157443	0.6544614723423994
29	0.02650706513537758	0.680968537477777
30	0.025347705241491693	0.7063162427192686
31	0.024138391009375176	0.7304546337286438
32	0.022892894121676825	0.7533475278503207
33	0.02162432632545134	0.774971854175772
34	0.02034501044438226	0.7953168646201543
35	0.01906637425456097	0.8143832388747152
36	0.017798867505164292	0.8321821063798794
37	0.016551901836505033	0.8487340082163845
38	0.01533381286573636	0.8640678210821208
39	0.01415184328460111	0.8782196643667218
40	0.013012145451226961	0.8912318098179488



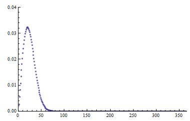
MEDIAN

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) ?

$$\frac{x-1}{d^x}\left({}_dP_{x-1}\right) > \frac{x}{d^{x+1}}\left({}_dP_x\right)$$
$$\frac{(x-1)}{d^x}\frac{d!}{(d-x+1)!} > \frac{x}{d^{x+1}}\frac{d!}{(d-x)!}$$
$$d(x-1) > x(d-x+1)$$
$$x^2 - x - d > 0$$
$$x > \frac{1+\sqrt{1+4d}}{2}$$



EXACT FORMULA!

- For d=365, we get x>19.61, so 19 or 20
- Approximating: $x > \sqrt{d}$ for large *d*

THE MEAN (EXPECTED VALUE)

- Since $P(X=x) = \frac{(x-1)}{d^x}\left({}_dP_{x-1}\right)$ for $x \in \{2,3,4,\ldots,d+1\}$
  then $E(X) = \sum_{x=2}^{d+1} x\frac{x-1}{d^x}\left({}_dP_{x-1}\right)$
- which is related to Ramanujan's Q-function, specifically $E(X) = 1 + Q(d)$
- which is asymptotic:
  $$E(X) \sim 1 + \sqrt{\frac{\pi d}{2}} - \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2d}} - \frac{4}{135d} + \cdots$$

- and for d=365, we get:
  $$E(X) \approx \sqrt{\frac{365\pi}{2}} + \frac{2}{3} + \frac{1}{12}\sqrt{\frac{\pi}{730}} - \frac{4}{49275} + \cdots \approx 24.6166$$

BIRTHDAY PROBLEM ON JUPITER?

- A Jovian day = 9.9259 Earth-hours
- A Jovian year = 11.86231 Earth-years

$$1\ Jyr = 11.86231\ Eyrs\left(\frac{365.24\ Edays}{1\ Eyr}\right)\left(\frac{24\ Ehrs}{1\ Eday}\right)\left(\frac{1\ Jday}{9.9259\ Ehrs}\right)$$
$$\approx 10476\ Jdays$$
$$P(X=x) = \frac{(x-1)}{10476^x}\left({}_{10476}P_{x-1}\right) \quad \text{for} \quad x \in \{2,3,4,\ldots,10477\}$$

- Both 10476^{25} and ${}_{10476}P_{25}$ cause a TI-84 overflow!

JOVIAN MEDIAN ESTIMATES

- Proportionality?

$$n \approx 1.2\sqrt{10476} \approx 122.823$$

- Heuristically? $1-\left(\frac{10475}{10476}\right)^{\frac{n(n-1)}{2}} > 0.5$

$$n(n-1) > \frac{\ln 4}{\ln\left(\frac{10476}{10475}\right)}$$

$$n \approx 121.009$$

MORE JOVIAN MEDIAN ESTIMATES

- Series Approximation:

$$\frac{10476}{10476} \times \frac{10475}{10476} \times \cdots \times \frac{10476-(n-1)}{10476} < 0.5$$

$$n(n-1) > (10476)\ln 4$$

$$n \approx 121.012$$

- Mathematica: for n = 121, P(match) = 50.13%

JOVIAN BIRTHDAY AVERAGES

- Mean:

$$E(X) \approx \sqrt{\frac{\pi d}{2}} + \frac{2}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2d}} - \frac{4}{135d} + \cdots \approx 128.947$$

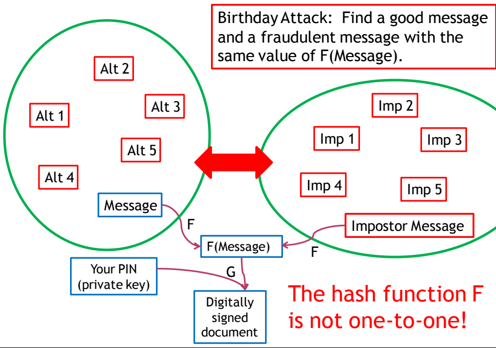
- Median:

$$n \approx \frac{1+\sqrt{1+4d\ln 4}}{2} \approx 121.012$$

- Mode:

$$x \approx \frac{1+\sqrt{1+4d}}{2} \approx 102.854$$

DIGITAL SIGNATURES



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(message 1 in set 1 doesn't match set 2) = $1-\frac{n}{d}$

MATCHING BETWEEN 2 SETS

- P(no message in set 1 matches anything in set 2) = $\left(1-\frac{n}{d}\right)^n \approx \left(e^{-n/d}\right)^n = e^{-n^2/d}$
- 50% probability of at least one match between the sets: $1-e^{-n^2/d} > 0.50$
  $$n^2 > d \ln 2$$
  $$n \approx 0.83\sqrt{d}$$
- Probability *p* of at least one match between the sets: $n \approx \sqrt{d \ln\left(\frac{1}{1-p}\right)}$

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