Counting on Coincidences

Christopher J. Bishop
SUNY Stony Brook

What is the probability that two people in the room have the same birthday (same month/day)?

What is the probability that two people in the room have the same birthday (same month/day)?

It's easier to compute the probability that everyone has a different birthday.

Let $P(N)$ be the probability that N random people have different birthdays.

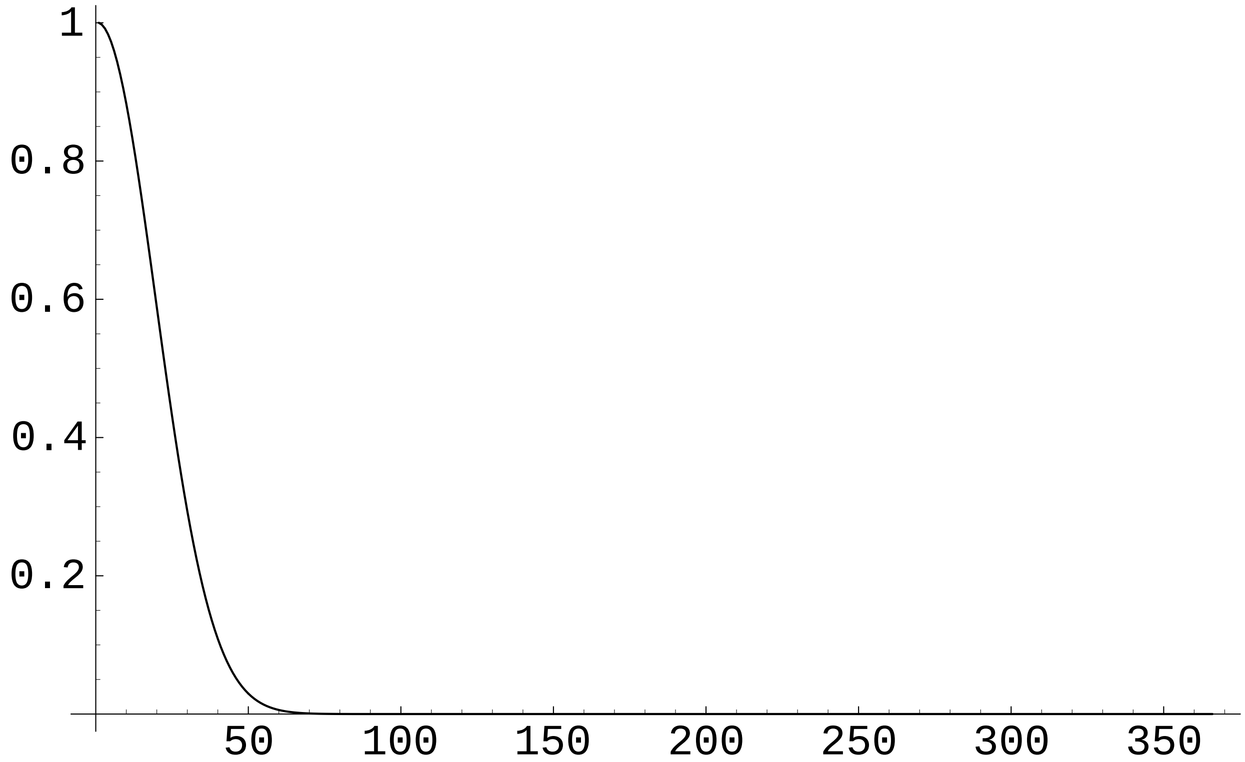
$$P(1) = 1$$

$$P(2) = 1 \cdot \frac{364}{365} \approx .99726$$

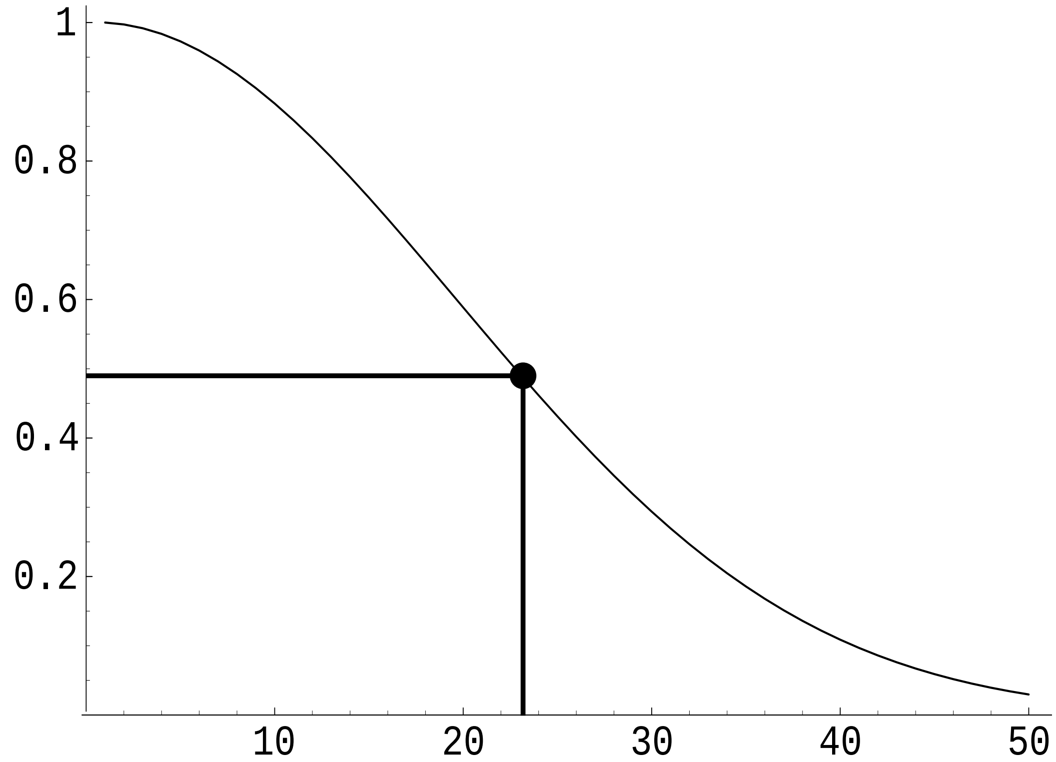
$$P(3) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \approx .991796$$

$$P(4) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \approx .983644$$

$$P(N+1) = P(N) \cdot \frac{365-N}{365}$$



Probability of different birthdays for N people



Probability of different birthdays, $N \leq 50$

$P(22)$	≈ 0.524	$P(50)$	≈ 0.0296
$P(23)$	≈ 0.492	$P(60)$	≈ 0.00587
$P(30)$	≈ 0.293	$P(70)$	≈ 0.000840
$P(40)$	≈ 0.108	$P(100)$	≈ 0.000000307

$$
\begin{aligned}
P(365) &\approx 1.45495521563 \times 10^{-157} \\
&\approx .00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000000000 \\
&\quad\; 00000000000000001455
\end{aligned}
$$

A lottery sells a million tickets each day and chooses one winner every day.

What is the chance that someone wins twice in a year?

(a) about 1 in a trillion
(b) about 1 in a billion
(c) about 1 in a million
(d) about 1 in a thousand
(e) about 1 in a hundred
(f) about 1 in ten
(g) about 50-50

(Assume same million players every day.)

Answer: Probability of 365 different winners is

$$1 \cdot \frac{999,999}{1,000,000} \cdots \frac{999635}{1000000} \approx .9353.$$

6.5% chance of a double winner in one year.

45% chance of double winner in 3 year period.

99.87% chance of a double winner in 10 years.

Suppose we randomly put K balls into N boxes. What is the chance that no box has more than M balls in it?

Call this probability $P(K, N, M)$.

The Birthday Problem is computing $P(K, 365, 1)$.

MIDTERM: What is $P(14400, 9000, 7)$?

(a) .0000000000000000000000000132
(b) .095395
(c) .664954
(d) .999323
(e) .999999999999999999999999845

If we drop 14,400 balls into 9000 boxes, what is the chance no box has more than 7 balls in it?

MIDTERM: What is $P(14400, 9000, 7)$?

(a) .0000000000000000000000000132
(b) .095395
(c) .664954
(d) .999323
(e) .999999999999999999999999845

If we drop 14,400 balls into 9000 boxes, what is the chance no box has more than 7 balls in it?

Why is this an important example?

In 1960 there were 14,400 cases of leukemia in US and 8 cases in Niles, IL, population 20,000.

The average for a town this size would be 1.6 cases.

Is the cluster random?

Population of US in 1960 was

$$180,000,000 = 9,000 \times 20,000.$$

Divide population into 9,000 “boxes” of 20,000 each.

Drop in 14,400 “cases”.

What is the chance that some box has 8 cases?

Answer $= 1 - P(14400, 9000, 7)$.

Calculation gives $P(14400, 9000, 7) = .095395$

The probability of some town of size 20,000 having 8 cases at random is about 90%.

N	Probability biggest cluster $\leq N$	Probability biggest cluster $> N$
6	.000005	.999995
7	.095395	.904605
8	.664954	.335046
9	.937864	.062137
10	.990843	.009157
11	.998788	.001212
12	.999852	.000148

A town of 20,000 with 8 or 9 cases is highly likely.

Probability of 11 cases at random is about 1%.

A counting problem:

A bag contains N numbered balls. You pull one out, look at it and put it back. Continue until you know how big N is. How long does this take?

A counting problem:

A bag contains N numbered balls. You pull one out, look at it and put it back. Continue until you know how big N is. How long does this take?

Answer: Forever. You can never be sure whether or not you missed one.

A counting problem:

A bag contains N numbered balls. You pull one out, look at it and put it back. Continue until you know how big N is. How long does this take?

Answer: Forever. You can never be sure whether or not you missed one.

Better question: How long before you have a "good guess" how many balls are in the bag?

A counting problem:

A bag contains N numbered balls. You pull one out, look at it and put it back. Continue until you know how big N is. How long does this take?

Answer: Forever. You can never be sure whether or not you missed one.

Better question: How long before you have a "good guess" how many balls are in the bag?

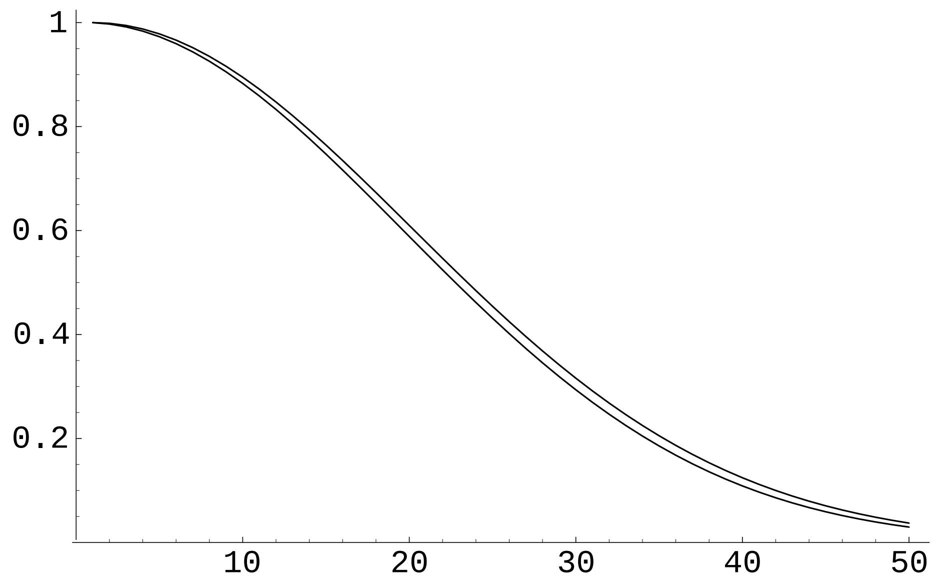
Answer: $\approx \sqrt{N}$

N balls into K boxes. Chance of "no repeats":

Exact Formula:

$$P(N) = 1 \cdot \frac{K-1}{K} \cdots \frac{K-N+1}{K}$$

Approximate Formula: $P(N) \approx e^{-N^2/2 \cdot K}$



Comparing exact and approximate formulas

How long before a 50% chance of a repeat?

Must solve

$$e^{-N^2/2K} = .5$$

$$\frac{-N^2}{2K} = \log .5$$

$$N \quad = \quad \sqrt{2K \log 2} \quad \approx \quad 1.17741\sqrt{K}$$

If we draw $N \approx \sqrt{K}$ random samples (with repetition) from a bag with K items, we have a good chance to get a repeat.

Counting balls in a bag problem

Rough guess:

If first repeat in on the nth draw from the bag, guess $K = (n/1.1774)^2$ as number of balls in bag.

Repeat and take average for better estimate.

Better solution:

Examine and return m samples.

Let t be total number of repeats.

Estimate $K \approx \frac{m(m-1)}{2t}$

Estimate is probably accurate if $m \gg \sqrt{K}$.

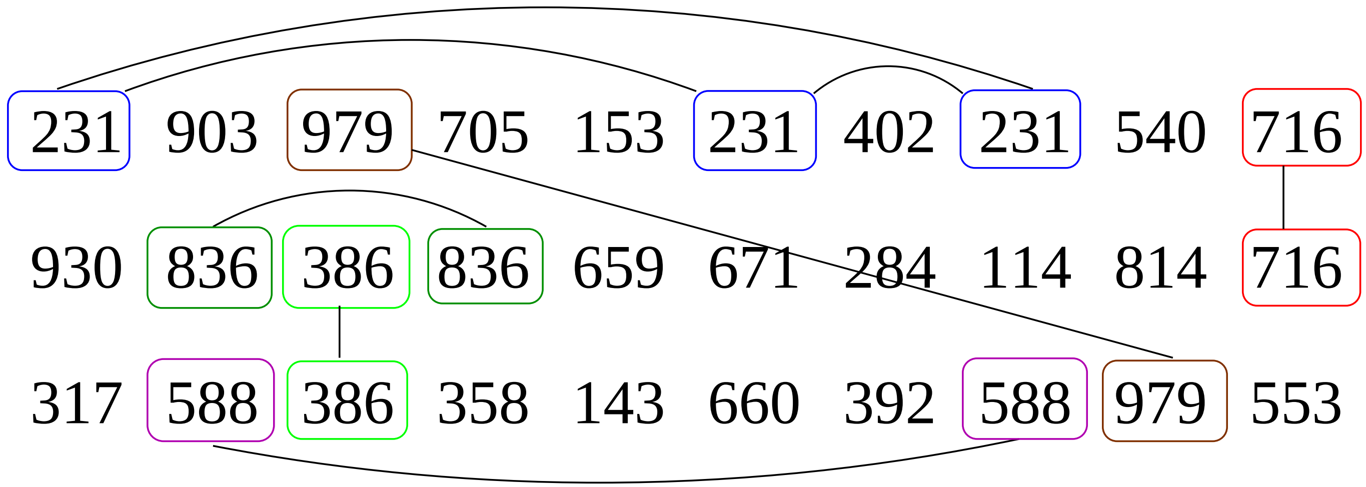
Example: I picked K distinct 3 digit numbers, then drew 30 random samples:

231 903 979 705 153 231 402 231 540 716

930 836 386 836 659 671 284 114 814 716

317 588 386 358 143 660 392 588 979 553

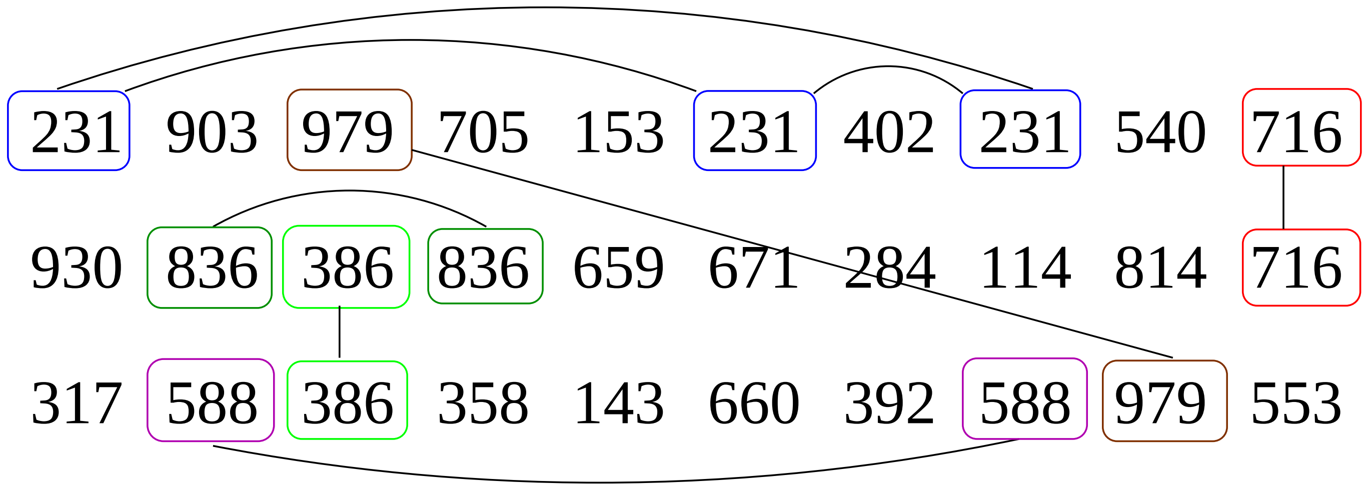
Example: I picked K distinct 3 digit numbers, then drew 30 random samples:



There are $m = 30$ samples and $t = 8$ pairs of repeated numbers, so our guess is

$$\frac{m(m-1)}{2t} = \frac{30 \cdot 29}{16} \approx 54.375.$$

Example: I picked K distinct 3 digit numbers, then drew 30 random samples:



There are $m = 30$ samples and $t = 8$ pairs of repeated numbers, so our guess is

$$\frac{m(m-1)}{2t} = \frac{30 \cdot 29}{16} \approx 54.375.$$

The true K is 57.

Applications:

Counting fish in lake

Counting distinct users on internet

Cryptography (security of digital signatures)

Factoring integers (Pollard rho method)

Many others

FINAL EXAM

Two thieves steal N diamonds with random values between \$1 and \$1,000,000. Can they divide the loot into two piles of equal value?

Impossible if $N = 1$ and unlikely if $N = 2, 3, \ldots ?$.

There is a 50% chance even splitting if $N >?$

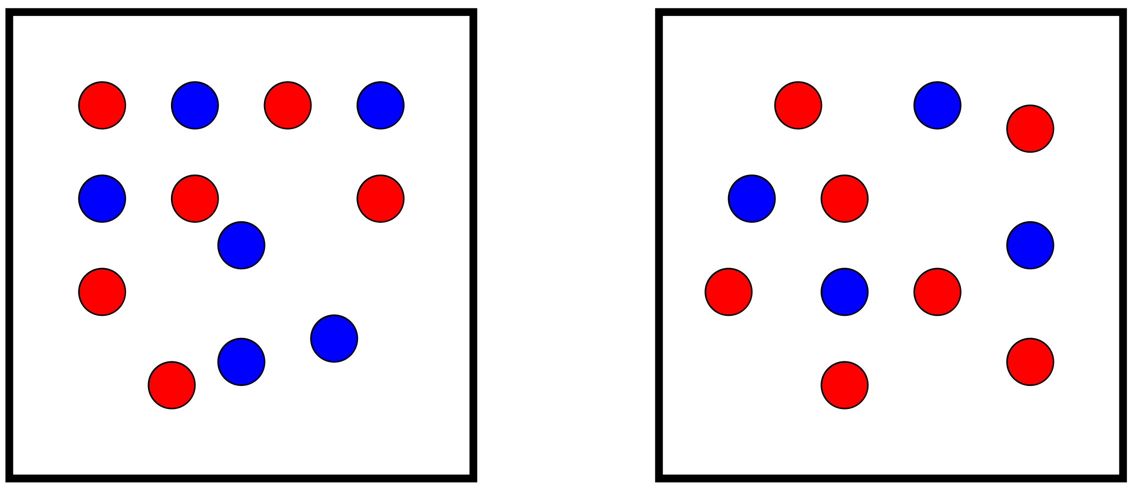
(a) 11
(b) 25
(c) 78
(d) 979
(e) 10,122

Divide diamonds into two groups of size $N/2$.

Randomly divide 1st group into red and blue subsets.

Let R_1, B_1 be the value of each subset.

Let $D_1 = R_1 - B_1$. Similarly $D_2 = R_2 - B_2$.



$D_1 = R_1 - B_1$ $\qquad$ $D_2 = R_2 - B_2$

There are $2^{N/2}$ ways to obtain D_1. Same for D_2.

If $D_1 = D_2$ then

$$R_1 - B_1 = R_2 - B_2,$$

$$R_1 + B_2 = R_2 + B_1,$$

so we get a division into two equal parts.

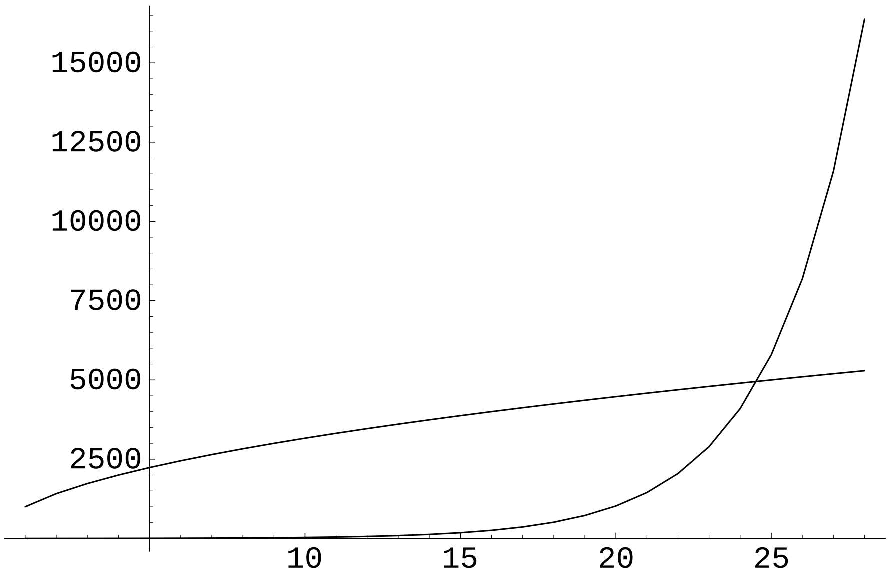
What is the chance $D_1 = D_2$?

What is chance of a repeat among $2^{N/2}$ random numbers of size between $-\frac{N}{2} \cdot 1,000,000$ and $\frac{N}{2} \cdot 1,000,000$?

By Birthday Problem the odds $\approx$ 50-50 if

$$\#\text{ random choices} \approx \sqrt{\#\text{ possible choices}}$$

$$2^{N/2} = \sqrt{N \times 1{,}000{,}000}$$



Answer: An equal division is likely if $N \geq 25$.

Number Partition Problem: given N integers can we divide them into two subsets with same sum?

This is an **NP-hard problem**. Roughly requires checking $\approx 2^N$ possible subsets in worst case.

Randomly choose N numbers in the range $[1, 2^{\kappa N}]$:

- there usually is an equal division if $\kappa < .96$
- there is usually not an equal division if $\kappa > .96$

The Number Partition Problem is only NP-hard problem that has such a precise analysis. Studied by computer scientists, mathematicians and physicists.

Nicknamed the "easiest" NP-hard problem.

References

Application to disease clusters:

W.J. Evans and H.S. Wilf, Computing the distribution of the maximum in balls-in-boxes problems with application to clusters of disease cases, *Proceedings of the National Academies of Science*, 104(2007), pages 11189-11191.

Counting via random sampling:

T. Bajku, S. Dasgupta, R. Kumar and R. Rubinfeld, The complexity of approximating the entropy, *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing*, Montreal, (2002), pages 678–687.

Counting fish by statistics:

Z.E. Schnabel, The estimation of the total fish population of a lake, *The American Mathematics Monthly*, 6(1938), pages 348–352.

Dividing into equal piles:

C. Borgs, J. Chayes and B. Pittel, Phase transition and finite-size scaling for the integer partition problem, *Random Structures Algorithms*, 19 (2001), pages 247–288.

Theory of coincidences:

P. Diaconis and F. Mosteller, Methods for studying coincidences, Journal of the American Statistical Association, 84(1989), pages 853–861.