Combinatorics (2.6)
The Birthday Problem (2.7)

Prof. Tesler

Math 186
Winter 2020

Multiplication rule

Combinatorics is a branch of Mathematics that deals with systematic methods of counting things.

Example

- How many outcomes (x, y, z) are possible, where

$$x = \text{roll of a 6-sided die;}$$
$$y = \text{value of a coin flip;}$$
$$z = \text{card drawn from a 52 card deck?}$$

- $(6 \text{ choices of } x) \times (2 \text{ choices of } y) \times (52 \text{ choices of } z) = \boxed{\mathbf{624}}$

Multiplication rule

The number of sequences $(x_1, x_2, \ldots, x_k)$ where there are

$$n_1 \text{ choices of } x_1, \quad n_2 \text{ choices of } x_2, \quad \ldots, \quad n_k \text{ choices of } x_k$$

is $n_1 \cdot n_2 \cdots n_k$.

This assumes the number of choices of x_i is a constant n_i that doesn't depend on the other choices.

Addition rule

Months and days

- How many pairs (m, d) are there where

$$m = \text{month } 1, \ldots, 12;$$
$$d = \text{day of the month?}$$

- Assume it's not a leap year.
- 12 choices of m, but the number of choices of d depends on m (and if it's a leap year), so the total is not "$12 \times$ __"
- Split dates into $A_m = \{\, (m, d) \ : \ d \text{ is a valid day in month } m \,\}$:

$$A = A_1 \cup \cdots \cup A_{12} = \text{whole year}$$
$$|A| = |A_1| + \cdots + |A_{12}|$$
$$= 31 + 28 + \cdots + 31 = 365$$

Addition rule

If $A_1, \ldots, A_n$ are mutually exclusive, then

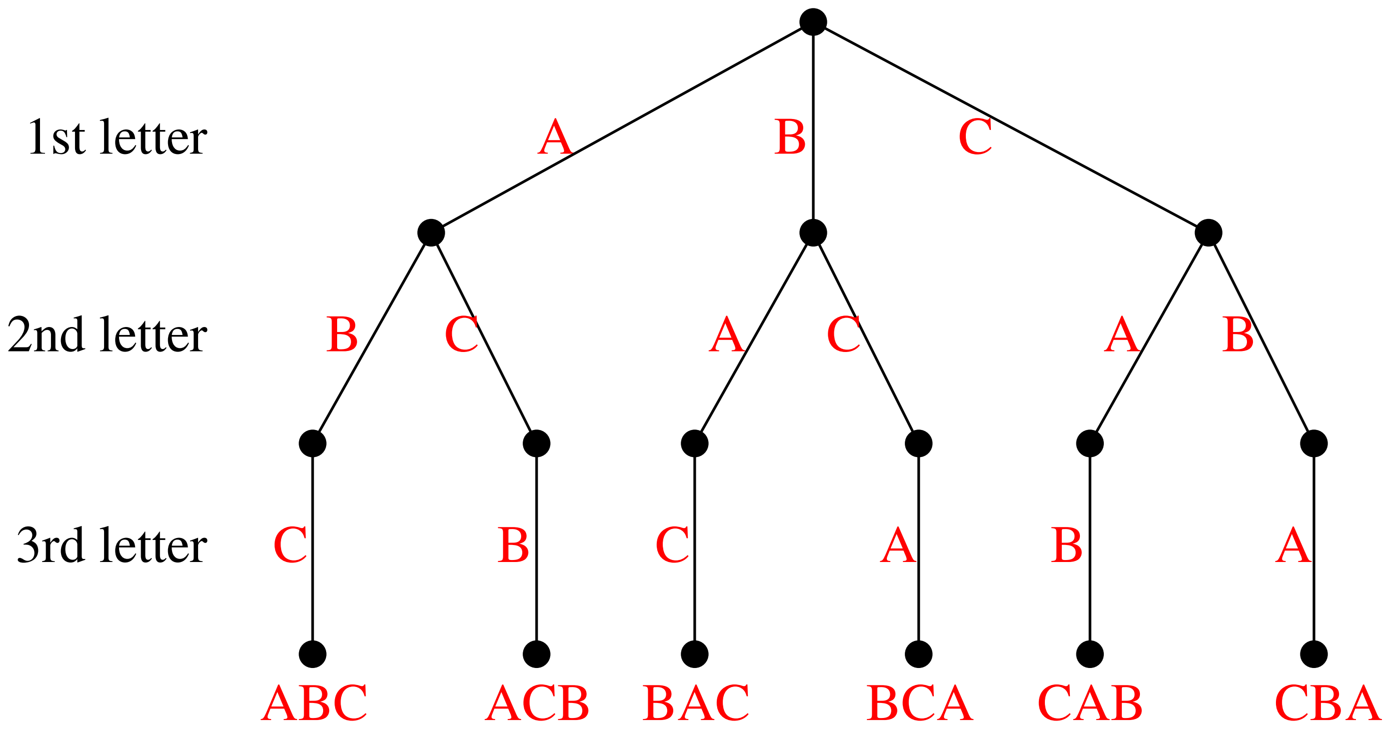
$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{i=1}^{n} |A_i|$$

Permutations of distinct objects

Here are all the permutations of A, B, C:

$$ABC \quad ACB \quad BAC \quad BCA \quad CAB \quad CBA$$

- There are 3 items: A, B, C.
- There are 3 choices for which item to put first.
- There are 2 choices remaining to put second.
- There is 1 choice remaining to put third.
- Thus, the total number of permutations is $3 \cdot 2 \cdot 1 = 6$.



Permutations of distinct objects

- In the example on the previous slide, the specific choices available at each step depend on the previous steps, but the number of choices does not, so the multiplication rule applies.
- The number of permutations of n distinct items is "n-factorial": $n! = n(n-1)(n-2)\cdots 1$ for integers $n = 1, 2, \ldots$

Convention: $0! = 1$

- For integer $n > 1$, $$\begin{aligned} n! &= n \cdot (n-1) \cdot (n-2) \cdots 1 \\ &= n \cdot (n-1)! \end{aligned}$$
  so $(n-1)! = n!/n$.
- E.g., $2! = 3!/3 = 6/3 = 2$.
- Extend it to $0! = 1!/1 = 1/1 = 1$.
- Doesn't extend to negative integers: $(-1)! = \frac{0!}{0} = \frac{1}{0} =$ undefined.

Stirling's Approximation

- In how many orders can a deck of 52 cards be shuffled?
- $52! = 80658175170943878571660636856403766975289505440883277824000000000000$
  (a 68 digit integer when computed exactly)

  $52! \approx 8.0658 \cdot 10^{67}$

- Stirling's Approximation: For large n,
$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n .$$
- Stirling's approximation gives $52! \approx 8.0529 \cdot 10^{67}$

Partial permutations of distinct objects

- How many ways can you deal out 3 cards from a 52 card deck, where the order in which the cards are dealt matters?
  E.g., dealing the cards in order $(A\clubsuit, 9\heartsuit, 2\diamondsuit)$ is counted differently than the order $(2\diamondsuit, A\clubsuit, 9\heartsuit)$.
- $52 \cdot 51 \cdot 50 = 132600$. This is also $52!/49!$.
- This is called an *ordered* 3-card hand, because we keep track of the order in which the cards are dealt.

- How many ordered k-card hands can be dealt from an n-card deck?

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} = {}_nP_k$$

  Above example is ${}_{52}P_3 = 52 \cdot 51 \cdot 50 = 132600$.
- This is also called permutations of length k taken from n objects.

Combinations

- In an *unordered* hand, the order in which the cards are dealt does not matter; only the set of cards matters. E.g., dealing in order $(A\clubsuit, 9\heartsuit, 2\diamondsuit)$ or $(2\diamondsuit, A\clubsuit, 9\heartsuit)$ both give the same hand. This is usually represented by a set: $\{A\clubsuit, 9\heartsuit, 2\diamondsuit\}$.
- How many 3 card hands can be dealt from a 52-card deck if the order in which the cards are dealt does not matter?
- The 3-card hand $\{A\clubsuit, 9\heartsuit, 2\diamondsuit\}$ can be dealt in $3! = 6$ different orders:

$$\begin{array}{ccc} (A\clubsuit, 9\heartsuit, 2\diamondsuit) & (9\heartsuit, A\clubsuit, 2\diamondsuit) & (2\diamondsuit, 9\heartsuit, A\clubsuit) \\ (A\clubsuit, 2\diamondsuit, 9\heartsuit) & (9\heartsuit, 2\diamondsuit, A\clubsuit) & (2\diamondsuit, A\clubsuit, 9\heartsuit) \end{array}$$

- Every unordered 3-card hand arises from 6 different orders.
  So $52 \cdot 51 \cdot 50$ counts each unordered hand $3!$ times; thus there are

$$\frac{52 \cdot 51 \cdot 50}{3 \cdot 2 \cdot 1} = \frac{52!/49!}{3!} = \frac{{}_{52}P_3}{3!}$$

  unordered hands.

Combinations

- The # of unordered k-card hands taken from an n-card deck is

$$\frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{k \cdot (k-1) \cdots 2 \cdot 1} = \frac{(n)_k}{k!} = \frac{n!}{k!\,(n-k)!}$$

- This is denoted $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$ (or ${}_nC_k$, mostly on calculators).
- $\binom{n}{k}$ is the "binomial coefficient" and is pronounced "n choose k."
- The number of unordered 3-card hands is

$$\binom{52}{3} = {}_{52}C_3 = \text{"52 choose 3"} = \frac{52 \cdot 51 \cdot 50}{3 \cdot 2 \cdot 1} = \frac{52!}{3!\,49!} = 22100$$

- **General problem:** Let S be a set with n elements. The number of k-element subsets of S is $\binom{n}{k}$.
- **Special cases:** $\binom{n}{0} = \binom{n}{n} = 1$ $\quad\binom{n}{k} = \binom{n}{n-k}$ $\quad\binom{n}{1} = \binom{n}{n-1} = n$

Binomial Theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

- For $n=4$: $\quad (x+y)^4 = (x+y)(x+y)(x+y)(x+y)$
- On expanding, each factor contributes an x or a y.
  After expanding, we group, simplify, and collect like terms:

$$\begin{aligned}(x+y)^4 &= yyyy \\ &\quad + yyyx + yyxy + yxyy + xyyy \\ &\quad + yyxx + yxyx + yxxy + xyyx + xyxy + xxyy \\ &\quad + yxxx + xyxx + xxyx + xxxy \\ &\quad + xxxx \\ &= y^4 + 4xy^3 + 6x^2y^2 + 4x^3y + x^4\end{aligned}$$

- Exponents of x and y must add up to n (which is 4 here).
- For the coefficient of $x^k y^{n-k}$, there are $\binom{n}{k}$ ways to choose k factors to contribute x's. The other $n-k$ factors contribute y's.
- Thus, $\binom{n}{k}$ unsimplified terms simplify to $x^k y^{n-k}$, giving $\binom{n}{k} x^k y^{n-k}$.

Permutations with repetitions

Here are all the permutations of the letters of ALLELE:

EEALLL	*EELALL*	*EELLAL*	*EELLLA*	*EAELLL*	*EALELL*
EALLEL	*EALLLE*	*ELEALL*	*ELELAL*	*ELELLA*	*ELAELL*
ELALEL	*ELALLE*	*ELLEAL*	*ELLELA*	*ELLAEL*	*ELLALE*
ELLLEA	*ELLLAE*	*AEELLL*	*AELELL*	*AELLEL*	*AELLLE*
ALEELL	*ALELEL*	*ALELLE*	*ALLEEL*	*ALLELE*	*ALLLEE*
LEEALL	*LEELAL*	*LEELLA*	*LEAELL*	*LEALEL*	*LEALLE*
LELEAL	*LELELA*	*LELAEL*	*LELALE*	*LELLEA*	*LELLAE*
LAEELL	*LAELEL*	*LAELLE*	*LALEEL*	*LALELE*	*LALLEE*
LLEEAL	*LLEELA*	*LLEAEL*	*LLEALE*	*LLELEA*	*LLELAE*
LLAEEL	*LLAELE*	*LLALEE*	*LLLEEA*	*LLLEAE*	*LLLAEE*

There are 60 of them, not $6! = 720$, due to repeated letters.

Permutations with repetitions

- There are $6! = 720$ ways to permute the subscripted letters $A_1, L_1, L_2, E_1, L_3, E_2$.
- Here are all the ways to put subscripts on EALLEL:

$$\begin{array}{cccc}
E_1A_1L_1L_2E_2L_3 & E_1A_1L_1L_3E_2L_2 & E_2A_1L_1L_2E_1L_3 & E_2A_1L_1L_3E_1L_2 \\
E_1A_1L_2L_1E_2L_3 & E_1A_1L_2L_3E_2L_1 & E_2A_1L_2L_1E_1L_3 & E_2A_1L_2L_3E_1L_1 \\
E_1A_1L_3L_1E_2L_2 & E_1A_1L_3L_2E_2L_1 & E_2A_1L_3L_1E_1L_2 & E_2A_1L_3L_2E_1L_1
\end{array}$$

- Each rearrangement of ALLELE has
  - $1! = 1$ way to subscript the A's;
  - $2! = 2$ ways to subscript the E's; and
  - $3! = 6$ ways to subscript the L's,

  giving $1! \cdot 2! \cdot 3! = 1 \cdot 2 \cdot 6 = 12$ ways to assign subscripts.
- Since each permutation of ALLELE is represented 12 different ways in permutations of $A_1L_1L_2E_1L_3E_2$, the number of permutations of ALLELE is

$$\frac{6!}{1!\,2!\,3!} = \frac{720}{12} = 60.$$

Multinomial coefficients

- For a word of length n with k_1 of one letter, k_2 of a 2nd letter, ..., the number of permutations is given by the *multinomial coefficient:*
$$\binom{n}{k_1, k_2, \ldots, k_r} = \frac{n!}{k_1!\,k_2!\,\cdots\,k_r!}$$
where $n, k_1, k_2, \ldots, k_r$ are integers $\geqslant 0$ and $n = k_1 + \cdots + k_r$.
- For ALLELE, it's $\binom{6}{1,2,3} = 60$. Read $\binom{6}{1,2,3}$ as "6 choose 1, 2, 3."
- For a multinomial coefficient, the numbers on the bottom must add up to the number on the top ($n = k_1 + \cdots + k_r$), vs. for a binomial coefficient $\binom{n}{k}$, instead it's $0 \leqslant k \leqslant n$.

Multinomial Theorem

- **Binomial theorem:** For integers $n \geqslant 0$,
$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$
$$(x+y)^3 = \tbinom{3}{0}x^0y^3 + \tbinom{3}{1}x^1y^2 + \tbinom{3}{2}x^2y^1 + \tbinom{3}{3}x^3y^0 = y^3 + 3xy^2 + 3x^2y + x^3$$

- **Multinomial theorem:** For integers $n \geqslant 0$,
$$(x+y+z)^n = \underbrace{\sum_{i=0}^{n}\sum_{j=0}^{n}\sum_{k=0}^{n}}_{i+j+k=n} \binom{n}{i,j,k} x^i y^j z^k$$
$$\begin{aligned}(x+y+z)^2 &= \tbinom{2}{2,0,0}x^2y^0z^0 + \tbinom{2}{0,2,0}x^0y^2z^0 + \tbinom{2}{0,0,2}x^0y^0z^2 \\ &\quad + \tbinom{2}{1,1,0}x^1y^1z^0 + \tbinom{2}{1,0,1}x^1y^0z^1 + \tbinom{2}{0,1,1}x^0y^1z^1 \\ &= x^2 + y^2 + z^2 + 2xy + 2xz + 2yz\end{aligned}$$

  $(x_1 + \cdots + x_m)^n$ works similarly with m iterated sums.

- In $(x+y+z)^{10}$, the coefficient of $x^2y^3z^5$ is $\binom{10}{2,3,5} = \frac{10!}{2!\,3!\,5!} = 2520$

Birthday Problem

a.k.a. Hash Collision Problem (in Computer Science)

Fun Party Fact

In a group of 23 or more randomly chosen people, there is over a 50% chance that at least two of them share the same birthday.

General Setup

- n days in a year. Ignore the concept of leap years.
- k people.
- Birthdays are uniform (each person has probability $1/n$ for each possible day) and birthdays of different people are independent:
  - If your club has a party for everyone with a January birthday, the people with January birthdays may be over-represented.
  - In a club for twins, the birthdays also would not be independent.
- What's the probability p that at least two people share a birthday? Equivalently, compute $q = 1 - p$, the probability that all birthdays are different.

Probability all birthdays are different

Example: 3 people

- First person has a unique birthday with probability $\frac{n}{n} = 1$.
- Second person has a birthday different from the first with probability $\frac{n-1}{n}$.
- Given that the first two birthdays were different, the third person has a birthday different from those with probability $\frac{n-2}{n}$.
- $q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n}$

General case

$$q = \prod_{r=1}^{k} P(r\text{th birthday different from first } r-1 \mid \text{first } r-1 \text{ distinct})$$

$$= \prod_{r=1}^{k} \frac{n-r+1}{n} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}$$

Probability all birthdays are different, 2nd derivation

- The sample space is all k-tuples of integers $1, \ldots, n$:

$$S = \{ \, (x_1, x_2, \ldots, x_k) \; : \; 1 \leqslant x_i \leqslant n \, \}$$

  where the ith person has birthday x_i. Note $N(S) = n^k$.
- E.g., number the days of the year $1, 2, \ldots, 365$.
  $(33, 2, 365)$ means the first person is born the 33rd day of the year (Feb. 2), the second is born Jan. 2, the third is born Dec. 31.
- Let A be the event that all birthdays are different.
- $N(A) = {}_nP_k = n(n-1)(n-2)\ldots(n-k+1)$
- $P(A) = N(A)/N(S) = \frac{{}_nP_k}{n^k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}$

Probability all birthdays are different, approximation

We will also give an approximate formula for q:

$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \qquad \approx \exp\left(-\frac{k^2}{2n}\right) \quad \text{for } k \ll n.$$

Question

How large a group of people is needed for at least a 90% chance that at least two share a birthday?

Answer

- $p \geqslant 90\%$ gives $q = 1 - p \leqslant 10\%$.
- We could chug away the exact equation $q = \frac{365}{365}\frac{364}{365} \cdots \frac{366-k}{365}$ on a calculator for $k = 1, 2, 3, \ldots$ until we get $q < 10\%$.
- Or we can solve for k from the approximate formula:

$$q \approx \exp\left(-\frac{k^2}{2n}\right) \qquad \ln(q) \approx -\frac{k^2}{2n} \qquad k \approx +\sqrt{-2n\ln(q)} = +\sqrt{-2n\ln(1-p)}$$

- Note $1 - p < 1$ so $\ln(1-p) < 0$ and $-2n\ln(1-p) > 0$.

Probability all birthdays are different, approximation

$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \qquad \approx \exp\left(-\frac{k^2}{2n}\right) \quad \text{for } k \ll n.$$

- For at least a 90% chance that two people share a birthday, use $k = 41$:

k	q with exact formula	q with approx formula
40	0.1087	0.1117
41	0.0968	0.0999

- How about for $p = 50\%$?

Party problem

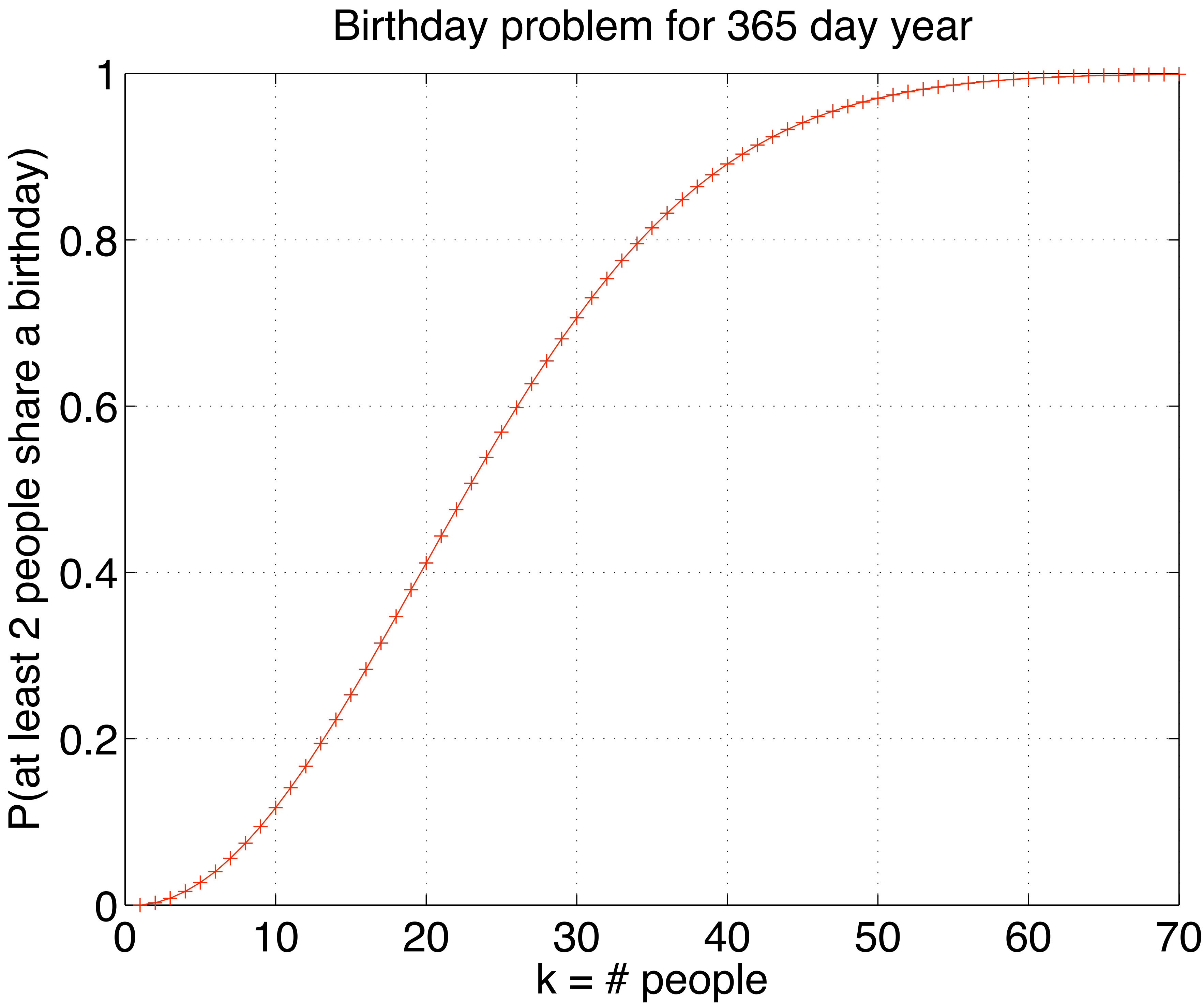
- $q = 1 - p = .50$ and $k \approx \sqrt{-2(365)\ln(.50)} = 22.49$
- In a group of 23 randomly selected people, there's a $p \approx 1 - \exp(-\frac{23^2}{2(365)}) = 51.55\%$ chance that two share a birthday. (The exact formula gives $p = 1 - \frac{365}{365}\frac{364}{365}\cdots\frac{343}{365} \approx 50.73\%$.)
- In a group of 23 or more randomly selected people, there's over a 50% chance that two share a birthday.

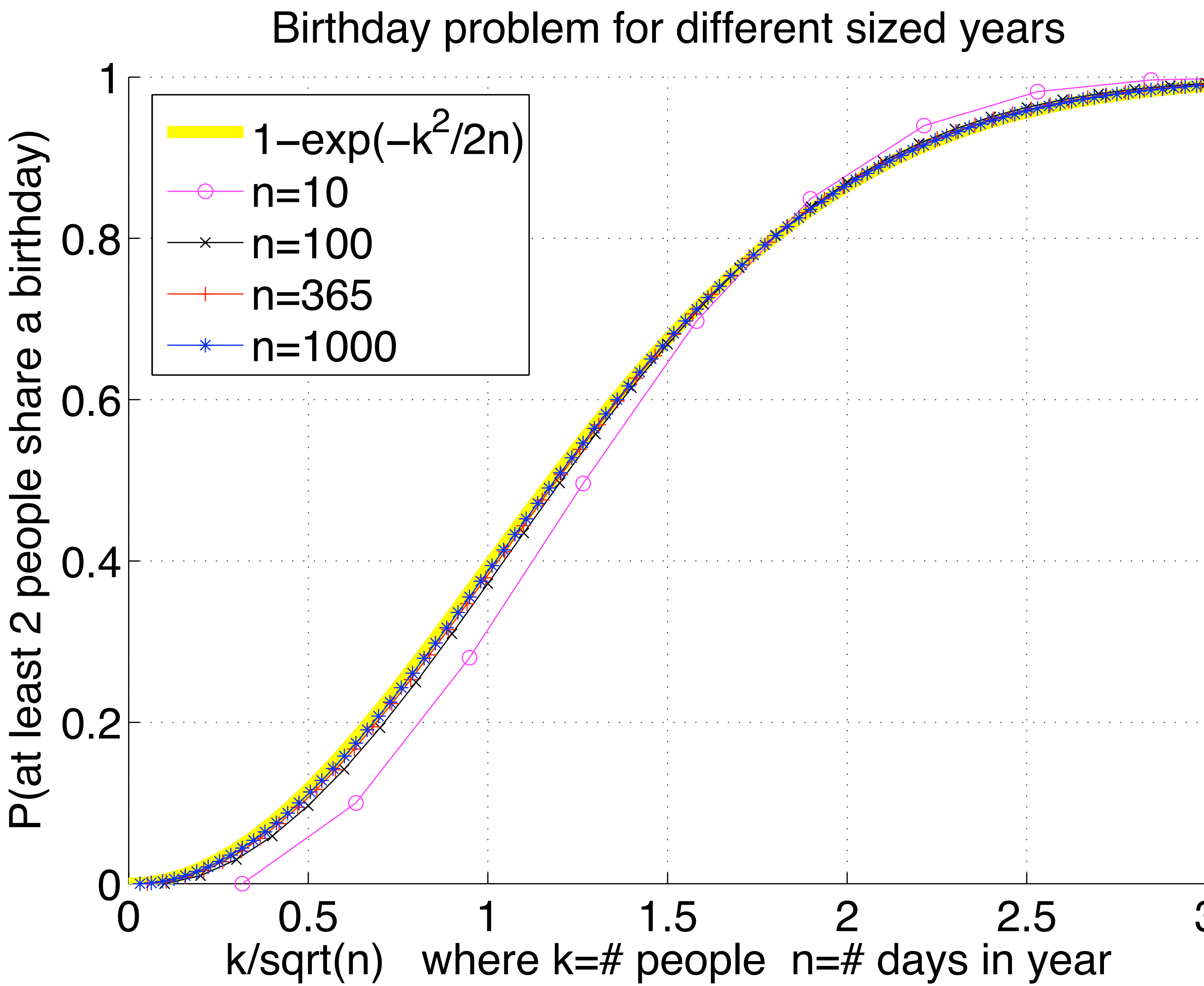
Varying the number of days in a year

- Using $k \approx \sqrt{-2\ln(1-p)}\,\sqrt{n}$ gives

p	**k in n day year**	**k in 365 day year**
.5	$1.18\sqrt{n}$	23
.7	$1.55\sqrt{n}$	30
.9	$2.15\sqrt{n}$	41
.99	$3.03\sqrt{n}$	58

- On the graphs that follow, we plot the exact probability formula.
- First graph: 365 day year.
- Second graph:
  - Multiple year sizes (n) are plotted.
  - We also superimpose the approximate probability formula in yellow.
  - x-axis is $k/\sqrt{n}$, so, for example, in most of the curves,
    probability is $\sim 50\%$ at $k/\sqrt{n} \approx 1.18$
    probability is $\sim 70\%$ at $k/\sqrt{n} \approx 1.55$.


Birthday problem for 365 day year
P(at least 2 people share a birthday)
k = # people
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50
60
70



Birthday problem for different sized years
1−exp(−k^2/2n)
n=10
n=100
n=365
n=1000
P(at least 2 people share a birthday)
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3
k/sqrt(n)   where k=# people n=# days in year

Derivation of approximation formula

- Start from the exact formula
$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}$$
- Take the logarithm to convert the product to a sum:
$$\ln(q) = \ln\left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}\right) = \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right)$$
- **Trick:** Multiply by $1 = n \cdot \frac{1}{n}$ and approximate it as an integral:
$$\ln(q) = n \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right) \frac{1}{n} \approx n \int_{1-k/n}^{1} \ln(x)\, dx$$
Note: bounds are $\frac{n-k}{n} = 1 - \frac{k}{n}$ *and* $\frac{n}{n} = 1$

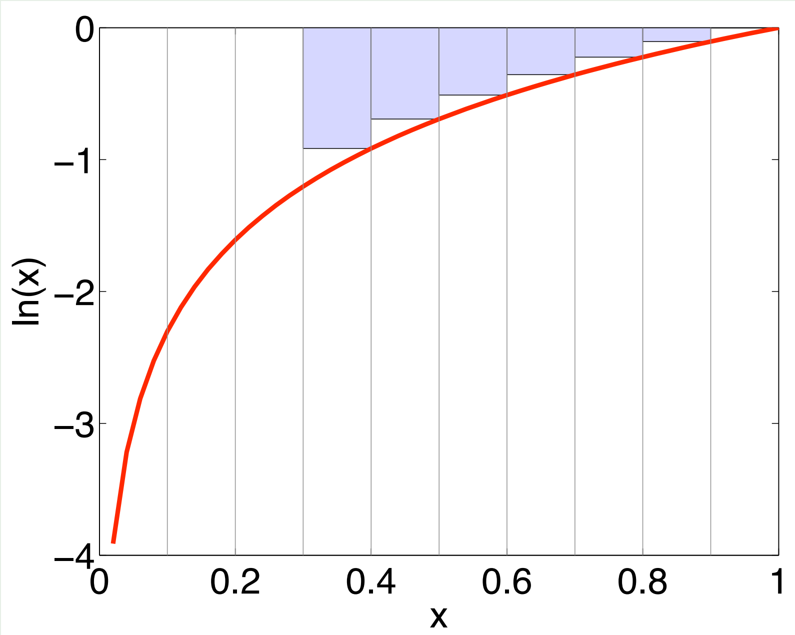
Derivation of approximation formula

$$\ln(q) = n \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right)\frac{1}{n} \approx n \int_{1-k/n}^{1} \ln(x)\,dx$$

Example: $n = 10$, $k = 7$; sum is negative area indicated

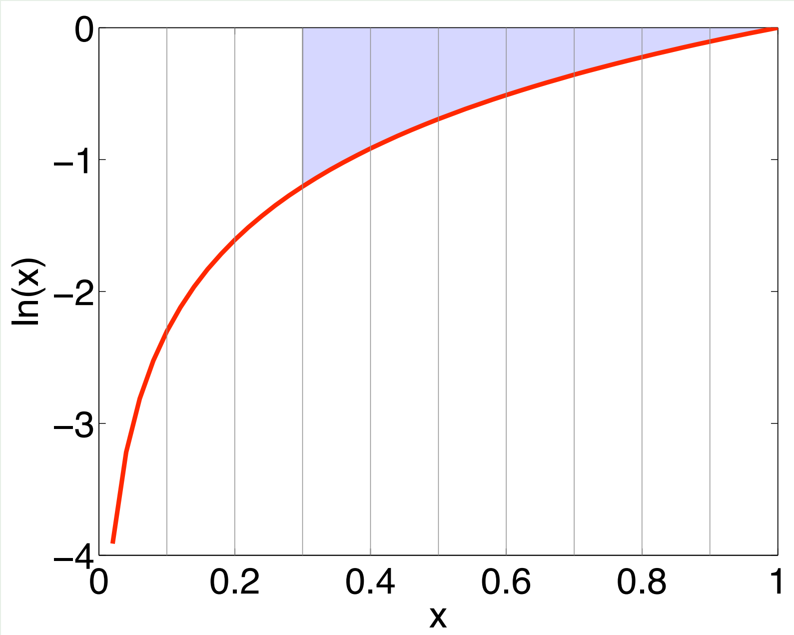
Exact formula for $\ln(q)$

$$\sum_{r=4}^{10} \ln\left(\tfrac{r}{10}\right)\tfrac{1}{10} = -0.280544...$$



Approximate formula for $\ln(q)$

$$\int_{.4}^{1} \ln(x)\,dx = -0.233483...$$



Derivation of approximation formula

$$\ln(q) \approx n \int_{1-k/n}^{1} \ln(x)\,dx = n\Big(x\big(\ln(x)-1\big)\Big)\Big|_{1-k/n}^{1}$$

$$= n\Big(1\big(\ln(1)-1\big) - \big(1-k/n\big)\big(\ln(1-k/n)-1\big)\Big)$$

$$= n\Big(-k/n - \big(1-k/n\big)\big(\ln(1-k/n)\big)\Big)$$

- Using the Taylor series $\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$ gives

$$(1-x)\ln(1-x) = -x + \frac{x^2}{2\cdot 1} + \frac{x^3}{3\cdot 2} + \frac{x^4}{4\cdot 3} + \cdots$$

- Use this (with $x = k/n$) and plug into the approximation for $\ln(q)$. The leading term is

$$\ln(q) \approx n\Big(-\frac{k}{n} + \frac{k}{n} - \frac{k^2}{2\cdot 1\cdot n^2} - \frac{k^3}{3\cdot 2n^3} - \frac{k^4}{4\cdot 3n^4} - \cdots\Big) \approx -\frac{k^2}{2n}\,.$$

  so $p = 1 - q \approx 1 - \exp\left(-\frac{k^2}{2n}\right)$.

- The graphs show this approximation is pretty good except for small n. It's possible to quantify the error analytically also.

Searching for short DNA sequences

Alignment software (such as BLAST); Microarrays

Consider a genome:

Position	1	2	3	4	5	6	7	8	9	10	...
Nucleotide	A	C	A	A	T	G	C	A	T	G	...

- Pick a small value of ℓ; we'll use $\ell = 3$.
- Make a table of coordinates of all *ℓ-mers* (length ℓ substrings):

3-mer	**coordinates**	**3-mer**	**coordinates**
AAT	3	CAA	2
ACA	1	CAT	7
ATG	4, 8	GCA	6
		TGC	5

- In a genome of length m, the coordinates of ℓ-mers are $1, 2, \ldots, m - \ell + 1$.

Birthday Problem	**This example**
$k = \#$ people	$k = \#$ coordinates $= m - \ell + 1$
$n = \#$ days per year	$n = \#$ ℓ-mers $= 4^\ell$

Searching for short DNA sequences

Problem: Search for a short sequence Q ("query") in a long genome T ("text"). We'll do lots of searches against the same T. In the popular alignment software BLAST, T is a database of many genomes.

Strategy:

- In advance: make a table of coordinates of all ℓ-mers in T.
- At search time: See which ℓ-mers are in Q, and use that to find possible locations in T where Q goes.

Given ℓ: At what text length, m, is there $\approx 50\%$ chance of a collision between ℓ-mers in T?

- 4^ℓ ℓ-mers are possible.
- There is $\approx 50\%$ chance of a collision at $\approx 1.18\sqrt{4^\ell}$ ℓ-mers. So $m - \ell + 1 \approx 1.18\sqrt{4^\ell}$, or $m \approx 1.18 \cdot 2^\ell + \ell - 1$.
- Example with $\ell = 6$:
  $m \approx 1.18\sqrt{4^6} + 6 - 1 = 80.52$
  probability is just below 50% at $m = 80$, just above at $m = 81$

Searching for short DNA sequences

Given m: at what ℓ is there $\approx 50\%$ chance of a collision between ℓ-mers in T?

- The human genome is approximately 3 billion nucleotides long. To account for both strands, use text size $m = 6$ billion.
- The # ℓ-mers in T is $m - 2(\ell - 1)$, since we can't start an ℓ-mer at the last $\ell - 1$ positions of either strand. This is $\approx m$ since $\ell \ll m$.
- This is out of 4^ℓ ℓ-mers total.
- There is a 50% chance of collision when $m \approx 1.18\sqrt{4^\ell}$. Solve:

$$\frac{m}{1.18} = \sqrt{4^\ell} = 2^\ell \qquad\qquad \ell = \log_2(m/1.18)$$

  So $\ell = \log_2(6{,}000{,}000{,}000/1.18) = 32.24$.
- The collision probability is above 50% for $\ell \leqslant 32$;
  below 50% for $\ell \geqslant 33$.
- A specific text T might not be so random, however. The human genome has lots of long repeated strings, some much longer than this, as a result of duplication events in evolution.

Hash Collision Problem in Computer Science

Generalizes the birthday problem to other scenarios

A *hash function* maps *keys* to *values* (a.k.a. *buckets* or *codes*):

$$f : \text{Set of keys} \to \text{Set of values (or buckets)}$$

There are n buckets. Assume that keys are independently assigned to buckets with uniform probability $\frac{1}{n}$ per bucket.

Consider a subset of k keys. What is the probability of a *collision* (two keys in the same bucket)?

	Keys	Buckets
Hash collision problem	Keys	Buckets
Birthday problem	People	Days of year
DNA sequence	Coordinates	ℓ-mers

Note: ℓ-mers in overlapping coordinate windows actually are dependent. Assuming independence is an approximation.