Definitions of Probability

Inductive Probability

An inductive probability of an event the degree of belief which it is rational to place in a hypothesis or proposition on given evidence.

Logical

Principle of indifference

- Principle of Indifference states that two events are equally probable if we have no reason to suppose that one of them will happen rather than the other. (Laplace, 1814)

- Unbiased coin:

probability heads = probability tails = ½

Pierre-Simon, marquis de Laplace
(1749 –1827)
French mathematician, physicist, astronomer

- Symmetric die:

probability of each side = 1/6

Inductive = Naïve probability

- If space S is finite and all outcomes are equally likely, then

  $$\text{Prob(Event E)}=\frac{\text{\# of outcomes in E}}{\text{\# of all outcomes in S}}$$

- Can also work with continuous is # is replaced with Area or Volume
- Unbiased coin: Prob(Heads) = Prob(Tails) = 1/2
- Symmetric die: probability of each side = 1/6
- Lottery outcomes are not symmetric: It is not a 50%-50% chance to win or loose in a lottery

Are you in class?

A. Yes
B. No
C. I am not sure, I am still asleep

Get your i-clickers

Inductive probability can lead into trouble

- Glass contains a mixture of wine and water and proportion of water to wine can be anywhere between 1:1 and 2:1
- (i) We can argue that the proportion of water to wine is equally likely to lie between 1 and 1.5 as between 1.5 and 2.
- (ii) Consider now ratio of wine to water. It is between 0.5 and 1. Based on the same argument it is equally likely in [1/2, 3/4] as it is in [3/4, 1]. But then water to wine ratio is equally likely to lie between 1 and 4/3=1.333… as it is to lie between 1.333.. and 2. This is clearly inconsistent with the previous calculation…
- Paradox solved by clearly defining the experimental design:
  - For (i) use fixed amount of wine (1 liter) and select a uniformly-distributed random number between 1 and 2 for water.
  - For (ii) use 2 liter of water and select uniformly-distributed a random number between 1 and 2 for wine.
  - Different experiments – different answers
- This paradox is old, attributed to (among others) Joseph Bertrand

I have two children. One of them is a boy.
What is the probability I have two boys?

A. 1/2
B. 1/3
C. 2/3
D. 13/27
E. I don’t know

Get your i-clickers

I have two children. One of them is a boy.

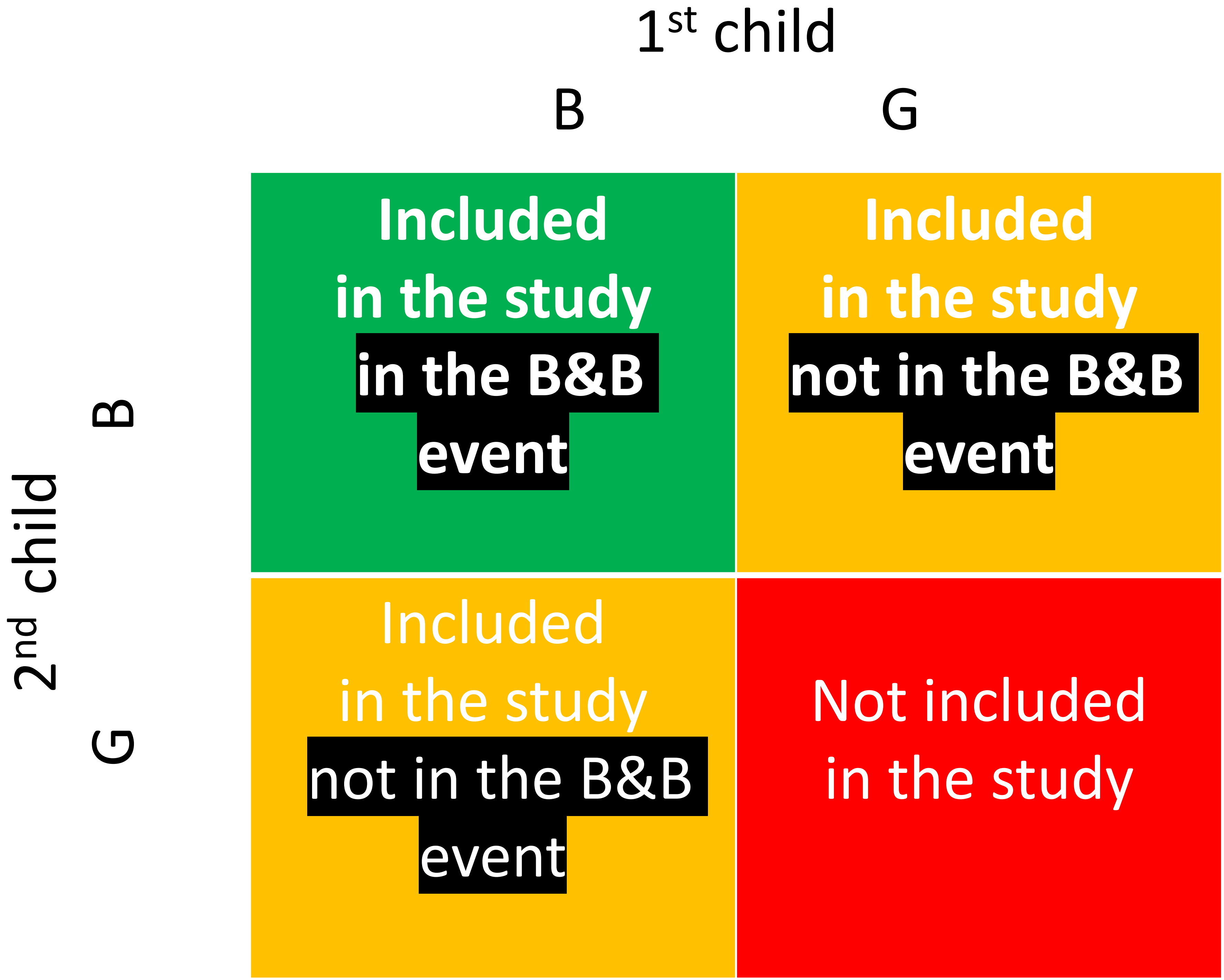
What is the probability I have two boys?

A. 1/2
B. $\boxed{\text{1/3}}$
C. 2/3
D. 13/27
E. I don't know

Get your i-clickers

Solution

- Naïve answer: probability is 1/2
  - It would be correct if I told you that my first child was a boy, and I was asking for a probability that my second child would also be a boy
- Correct answer: probability is 1/3
  - Two children can come in four configurations: 1) boy/girl, 2) girl/boy, 3) boy/boy, 4) girl/girl. Since he has one boy, we are looking at the options 1, 2, or 3. Only the boy/boy combination includes two boys, so the probability is 1/3
- Consider doing an NIH-funded study:
  - recruit 1000 parents with two children
  - send ~250 parents with two girls straight home
  - Out of remaining ~750 parents ~250 (or 1/3 of the total) have two boys. The probability is 1/3


1st child
B
G
2nd child
B
G
Included in the study in the B&B event
Included in the study not in the B&B event
Included in the study not in the B&B event
Not included in the study

I have two children.

One of them is a boy born on Tuesday.

What is the probability I have two boys?

A. 1/2
B. 1/3
C. 2/3
D. 13/27
E. I don't know

Get your i-clickers

I have two children.

One of them is a boy born on Tuesday.

What is the probability I have two boys?

A. 1/2
B. 1/3
C. 2/3
D. $\boxed{\text{13/27}}$
E. I don't know

Get your i-clickers

- 4x7x7=196 outcomes, 13+7+7=27 of which satisfy "Boy born on Tuesday"
- The probability of having two boys =13/(13+7+7)=13/27
- Close but not equal to 14/28=1/2

		Child 1: B						
		M	T	W	R	F	S	S
	M		1					
	T	2	3	4	5	6	7	8
Child 2: B	W		9					
	R		10					
13	F		11					
	S		12					
	S		13					

		Child 1: G						
		M	T	W	R	F	S	S
	M							
	T	1	2	3	4	5	6	7
Child 2: B	W							
	R							
7	F							
	S							
	S							

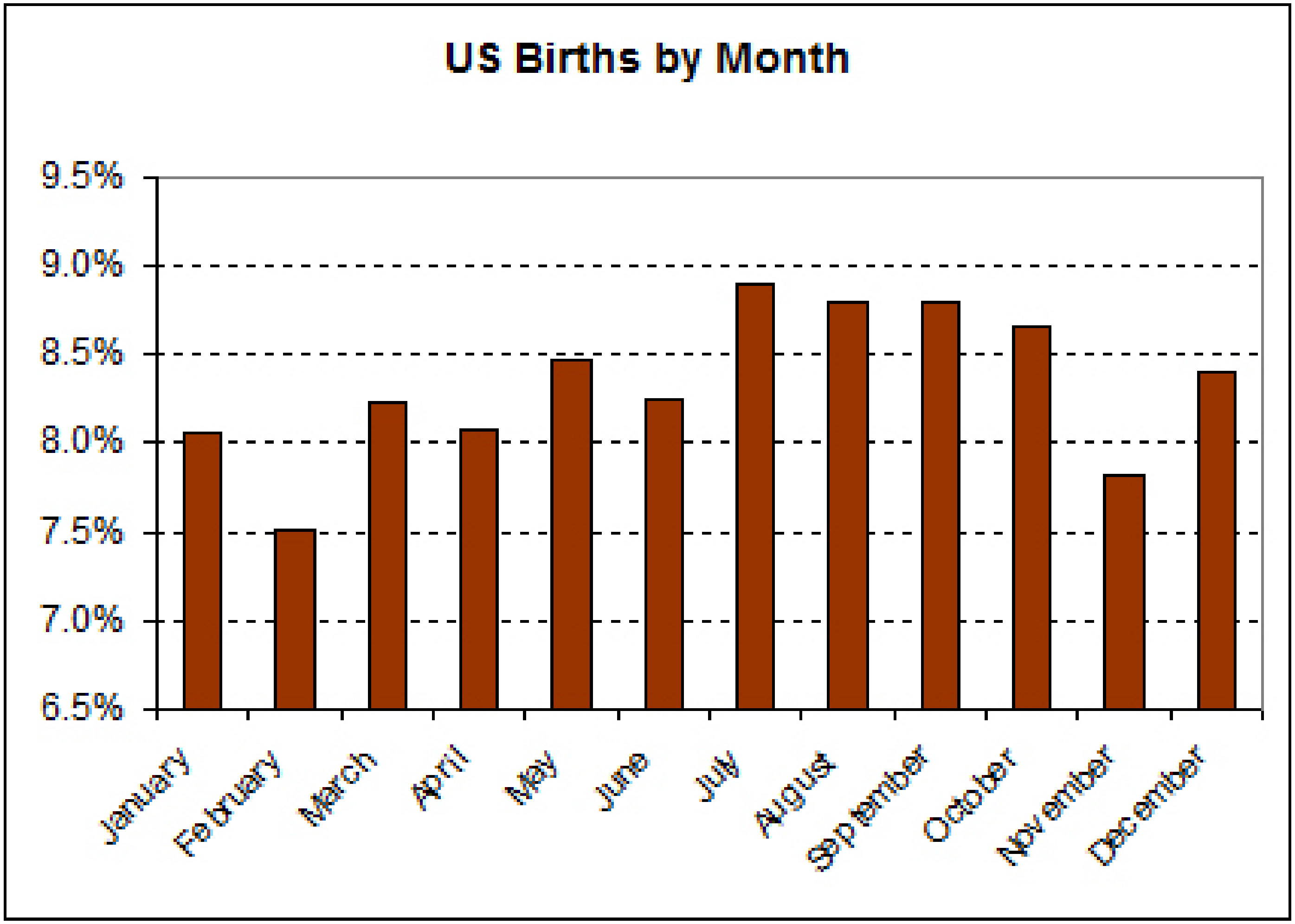
		Child 1: B						
		M	T	W	R	F	S	S
	M		8					
	T		9					
Child 2: G	W		10					
	R		11					
7	F		12					
	S		13					
	S		14					

		Child 1: G						
		M	T	W	R	F	S	S
	M							
	T							
Child 2: G	W							
	R							
	F							
	S							
	S							

How about the probability of having two boys if one boy was born in May or, separately, on May 1)

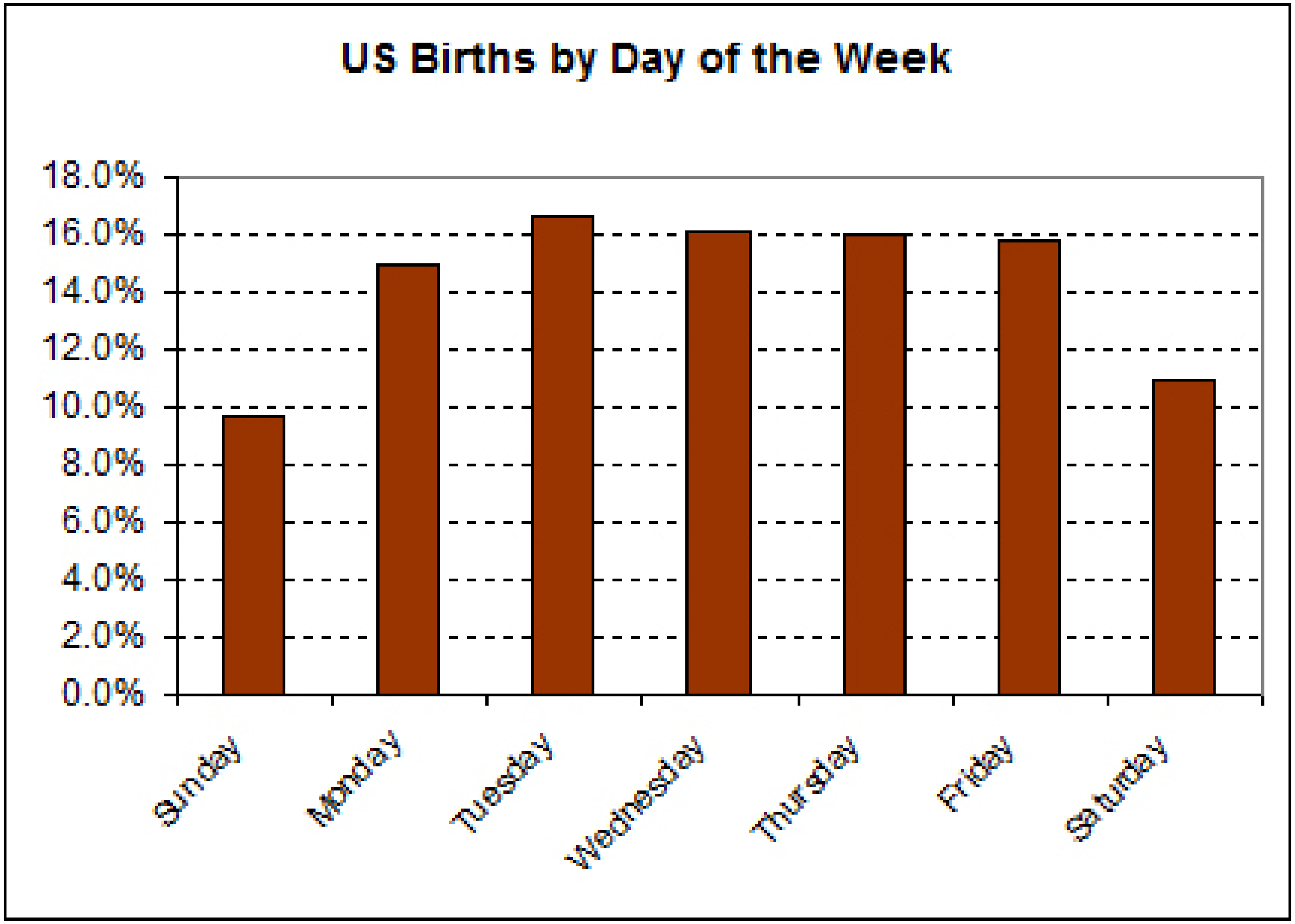
- Assume that giving birth in each day and every month of the year is equally likely
- With days of the week:
Prob(2 boys)=(2*7-1)/(4*7-1)=13/27=0.4815
- With months of the year:
Prob(2 boys)=(2*12-1)/(4*12-1)=23/47= 0.4894
- With days of the year:
Prob(2 boys)=(2*365-1)/(4*365-1)= 729/ 1459 = 0.49966
- The more detailed is the information about the birthday of one of the boys - the closer the probability is to 0.5

Is my assumption that all 12 months are equally likely justified?



Source, birth statistics in 2003:
https://www.cdc.gov/nchs/data/nvsr/nvsr54/nvsr54_02.pdf

Is my assumption that all 7 days of the week are equally likely justified?



Source, birth statistics in 2003:
https://www.cdc.gov/nchs/data/nvsr/nvsr54/nvsr54_02.pdf

Inductive probability
relies on combinatorics
or the art of counting
combinations

Counting – Multiplication Rule

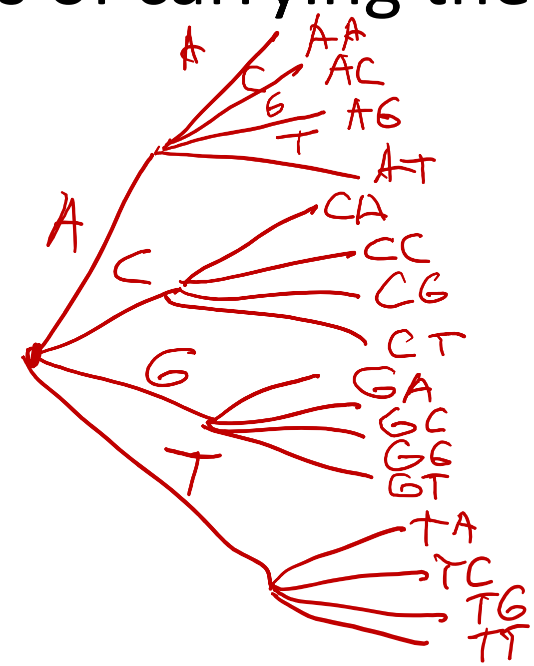
- Multiplication rule:
  - Let an operation consist of k steps and
    - n_1 ways of completing the step 1,
    - n_2 ways of completing the step 2, … and

    …….

    - n_k ways of completing the step k.
  - Then, the total number of ways of carrying the entire operation is:
    - $n_1 * n_2 * \ldots * n_k$

$n_1 = n_2 = 4$

Example: DNA 2-mer



- $S = \{A, C, G, T\}$ the set of 4 DNA bases
  – Number of k-mers is $4^k=4*4*4...*4$ (k –times)
  – There are $4^3=64$ triplets in the genetic code
  – There are only 20 amino acids (AA)+1 stop codon
  – There is redundancy: same AA coded by 1-3 codons
  – Evidence of natural selection: “silent” changes of bases are more common than AA changing ones
- A protein-coding part of the gene is typically 1000 bases long
  – There are $4^{1000}= 2^{2000} \sim 10^{600}$ possible sequences of just one gene
  – Or $(10^{600})^{25,000}=10^{15,000,000}$ of 25,000 human genes.
  – For comparison, the Universe has between 10^{78} and 10^{80} atoms and is $4*10^{17}$ seconds old.

Counting – Permutation Rule

- A permutation is a unique sequence of distinct items.
- If $S = \{a, b, c\}$, then there are 6 permutations
  - Namely: abc, acb, bac, bca, cab, cba (order matters)
- # of permutations for a set of n items is n!
- n! (factorial function) = n*(n-1)*(n-2)*…*2*1
- 7! = 7*6*5*4*3*2*1 = 5,040
- By definition: 0! = 1

Multiplication and permutation rules are two examples of a general problem, in which a set of size k is drawn from a population of n distinct objects

Balls drawn from an urn (or a bag)

1 ball is red

1 ball is blue

1 ball is green

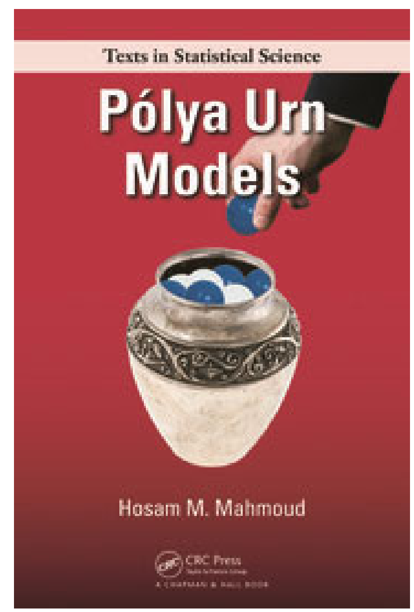
n=3 balls of different colors are in a bag

I draw k=2 balls one at a time

- Do I put each ball back to the bag after drawing it?
  - Yes: problem with replacement
  - No: problem without replacement
- Do I keep track of the order in which balls were drawn?
  - Yes: the order matters
  - No: the order does not matter

George Pólya

- George Pólya (December 13, 1887 – September 7, 1985) was a Hungarian mathematician. He was a professor of mathematics from 1914 to 1940 at ETH Zürich and from 1940 to 1953 at Stanford University. He made fundamental contributions to combinatorics, number theory, numerical analysis and probability theory.

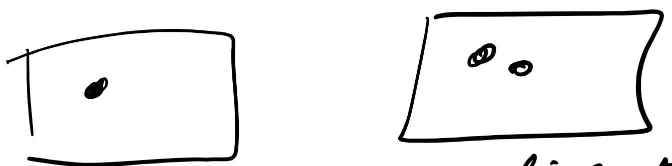


How many ways to choose a sample of K objects out of a population of n objects

	Order matters	Order does not matter
Replace	$n \times n \times n \times \ldots \times n = n^K$	$\cancel{\frac{n^K}{K!}}$ not all objects are different
Do not replace	$n \times (n-1) \times (n-2) \times \ldots \times (n-K+1) = \frac{n!}{(n-K)!}$	All objects are different $\rightarrow$ $\frac{n!}{(n-K)!} \times \frac{1}{K!} = \binom{n}{K}$

How to solve the problem of k out of n with replacement but where order does not matter?

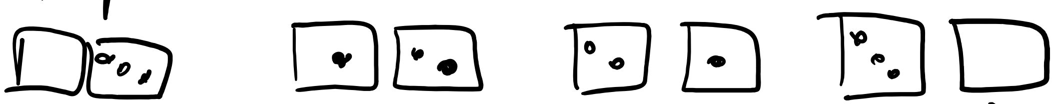
Let's solve $n=2$ problem first:



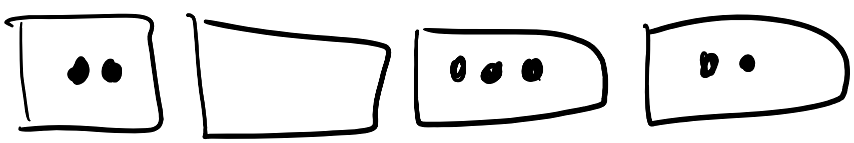
object 1 object 2

$k=3$

4 possibilities



(1) (2) (3) (4)



$n=4,\ k=7$

●●|||●●●|●● k dots, $n-1$ box boundaries

$$\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!\,(n-1)!}$$ ways to distribute

Sampling table

How many ways to choose a sample of k objects out of population of n objects?

	Order matters	Order does not matter
Replacement	$(n)^k$	Difficult: $\binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!k!}$
No replacement	$n(n-1)(n-2)..(n-k+1) = \frac{n!}{(n-k)!}$	$\binom{n}{k} = \frac{n!}{(n-k)!k!}$

Example

- A DNA of 100 bases is characterized by its numbers of each of 4 nucleotides: n_A, n_C, n_G, and n_T ($n_A+n_C+n_G+n_T=100$)

- **I don't care about the sequence** (only about the total numbers of As,Cs,Gs, and Ts

- **How many distinct combinations** of n_A, n_C, n_G, and n_T are out there?

What is n and k in this problem?

A. k=100, n=4
B. k=4, n=100
C. k=100, n=4^{100}
D. k=4^{100}, n=4
E. I don't know

Get your i-clickers

What is n and k in this problem?

A. $k=100$, $n=4$
B. $k=4$, $n=100$
C. $k=100$, $n=4^{100}$
D. $k=4^{100}$, $n=4$
E. I don't know

Is it sampling with or without replacement & does order matter?

A. with replacement, order matters
B. without replacement, any order
C. with replacement, any order
D. without replacement, order matters
E. I don't know

Get your i-clickers

Is it sampling with or without replacement & does order matter?

A. with replacement, order matters
B. without replacement, any order
C. with replacement, any order
D. without replacement, order matters
E. I don't know

$n = 4, \quad k = 100$

$$\overbrace{\circ\;\circ\;\circ}^{A} \,\Big|\, \overbrace{\circ\;\circ\;\circ\;\circ\;\circ}^{C} \,\Big|\, \overbrace{\circ\;\circ}^{G} \,\Big|\, \overbrace{\circ\;\circ\;\circ\;\circ}^{T}$$

100 dots

$$\#\text{ of possibilities} = \frac{(100+4-1)!}{(4-1)!\;100!} = $$

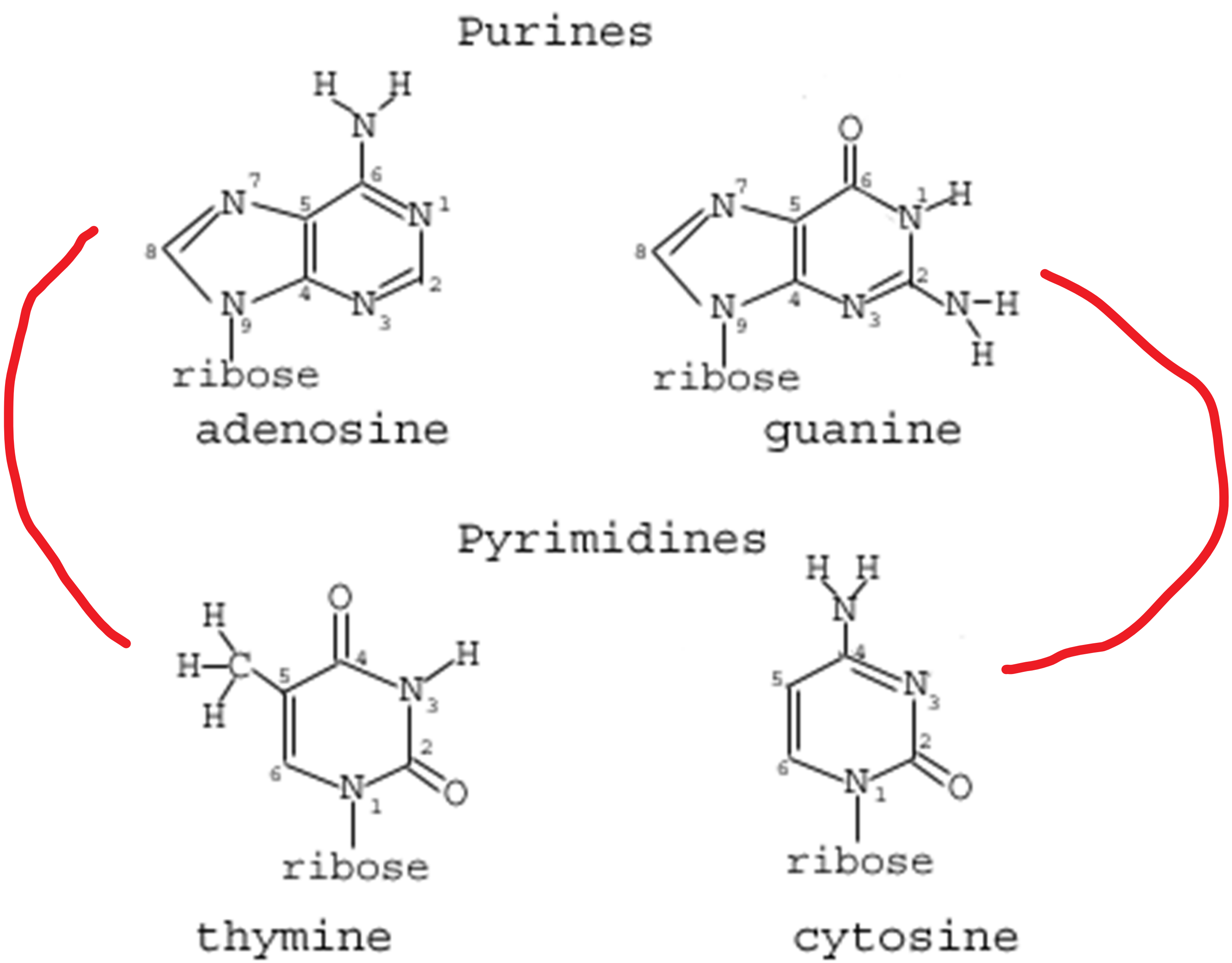
$$= \frac{103 \cdot 102 \cdot 101}{3 \cdot 2 \cdot 1} = 176{,}851$$

If order did matter

$4^{100} = 2^{200} = 10^{60} \approx$ # atoms in a galaxy

Real-life correction: Chagraff’s rule

Chargaff's rule: $n_A = n_T$ and $n_C = n_G$



DNA base frequency & Chargaff rules

Relative proportions (%) of bases in DNA [edit]

The following table is a representative sample of Erwin Chargaff's 1952 data, listing the base composition of DNA from various organisms and support both of Chargaff's rules.[13] An organism such as φX174 with significant variation from A/T and G/C equal to one, is indicative of single stranded DNA.

Organism	%A	%G	%C	%T	A/T	G/C	%GC	%AT
φX174	24.0	23.3	21.5	31.2	0.77	1.08	44.8	55.2
Maize	26.8	22.8	23.2	27.2	0.99	0.98	46.1	54.0
Octopus	33.2	17.6	17.6	31.6	1.05	1.00	35.2	64.8
Chicken	28.0	22.0	21.6	28.4	0.99	1.02	43.7	56.4
Rat	28.6	21.4	20.5	28.4	1.01	1.00	42.9	57.0
Human	29.3	20.7	20.0	30.0	0.98	1.04	40.7	59.3
Grasshopper	29.3	20.5	20.7	29.3	1.00	0.99	41.2	58.6
Sea Urchin	32.8	17.7	17.3	32.1	1.02	1.02	35.0	64.9
Wheat	27.3	22.7	22.8	27.1	1.01	1.00	45.5	54.4
Yeast	31.3	18.7	17.1	32.9	0.95	1.09	35.8	64.4
E. coli	24.7	26.0	25.7	23.6	1.05	1.01	51.7	48.3

exception for a
single stranded virus

DNA base frequency & Chargaff rules

PNAS

Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions

Guenter Albrecht-Buehler*

Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611

Edited by David Botstein, Princeton University, Princeton, NJ, and approved October 3, 2006 (received for review July 5, 2006)

Chargaff's second parity rules for mononucleotides and oligonucleotides (C^{II}_{mono} and C^{II}_{oligo} rules) state that a sufficiently long (>100 kb) strand of genomic DNA that contains *N* copies of a mono- or oligonucleotide, also contains *N* copies of its reverse complementary mono- or oligonucleotide on the same strand. There is

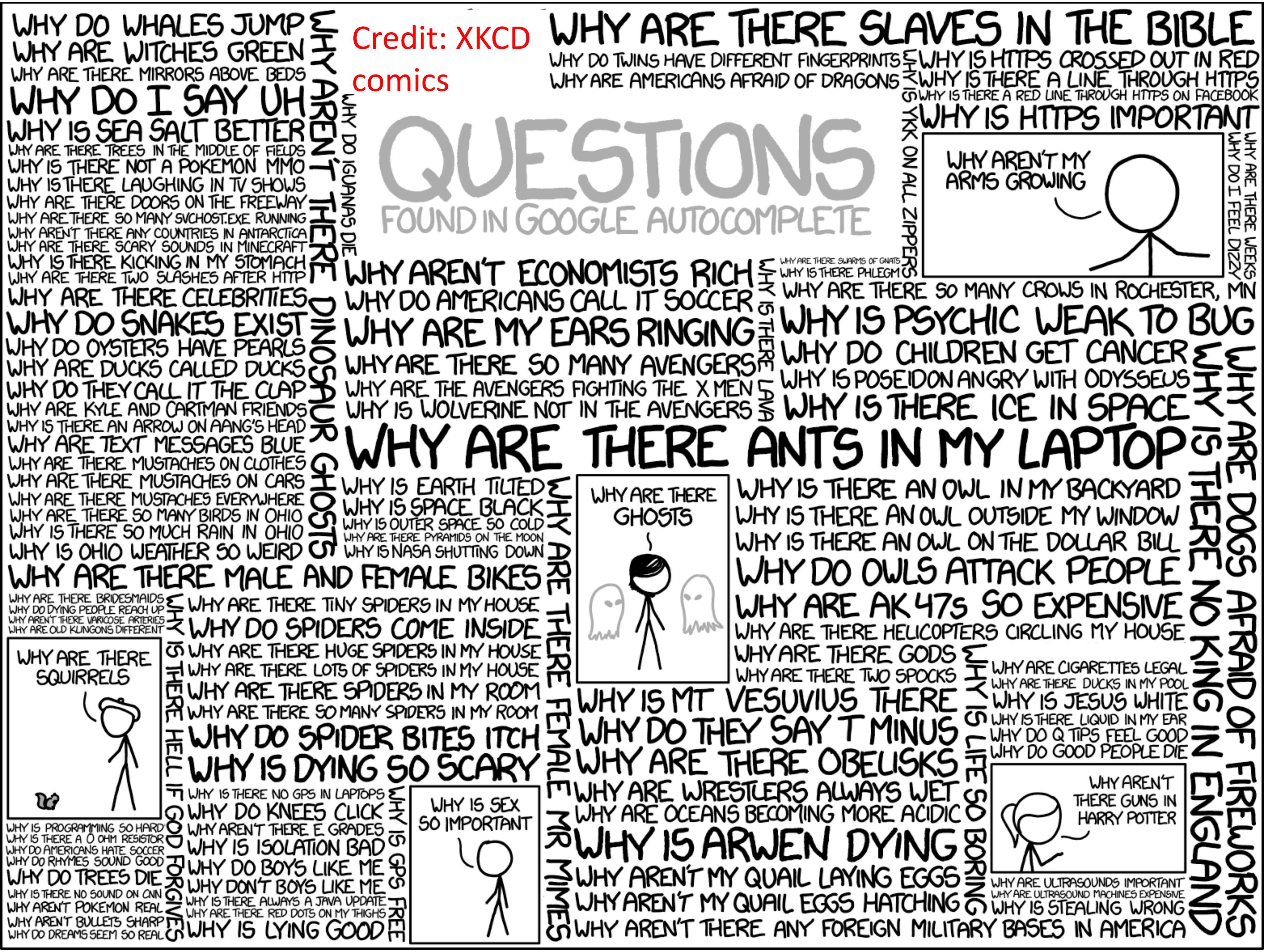
GCA at position $x + 61$! Therefore, Chargaff's second parity rules are not trivial, and have the remarkable implication that some unknown mechanism seems to "count" and adjust the numbers of oligonucleotides and their complements to equal values on each of both strands.

17828–17833 | PNAS | **November 21, 2006** | vol. 103 | no. 47 www.pnas.org/cgi/doi/10.1073/pnas.0605553103

On any segment of a **<u>single strand</u>** of DNA longer than 100kb
One has Chargaff's rules: $n_A = n_T$ and $n_C = n_G$

Reasons are not completely known, but likely to do
with frequent inversions in the course of the genome evolution

Hence nucleotide frequency has only 2 boxes instead of 4



Credit: XKCD comics