

Introduction to Bayesian Methods - 1

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Course webpage:

<http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Bayes.html>

Conditional probabilities and Bayes' Theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

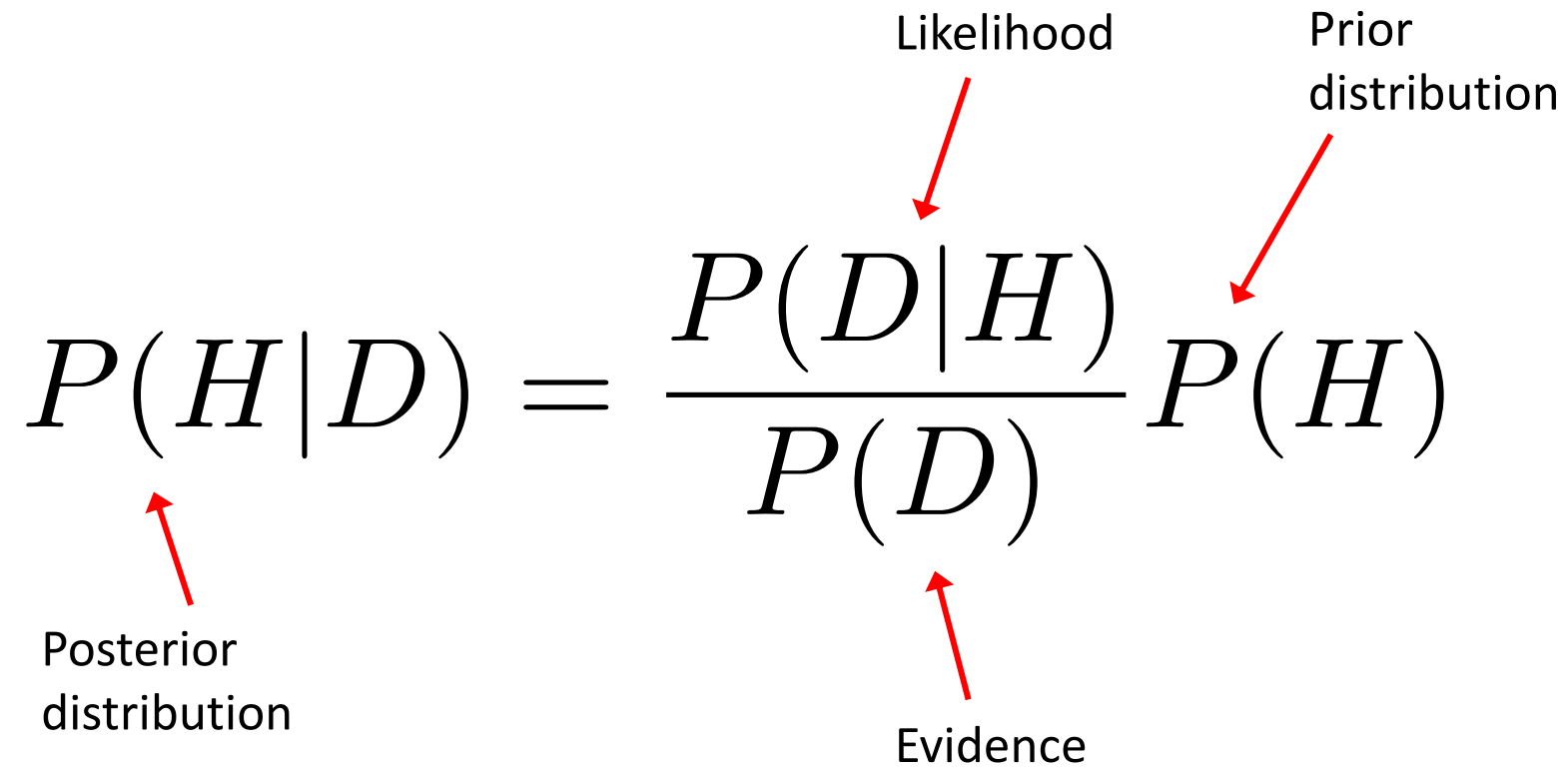
Joint probability and conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem: a purely logical statement

$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

Bayes' theorem again:
now as an inferential
statement



The diagram shows Bayes' theorem with four labels and red arrows pointing to specific parts of the equation:

- Likelihood**: Points to the numerator term $P(D|H)$.
- Prior distribution**: Points to the term $P(H)$.
- Evidence**: Points to the denominator term $P(D)$.
- Posterior distribution**: Points to the left-hand side term $P(H|D)$.

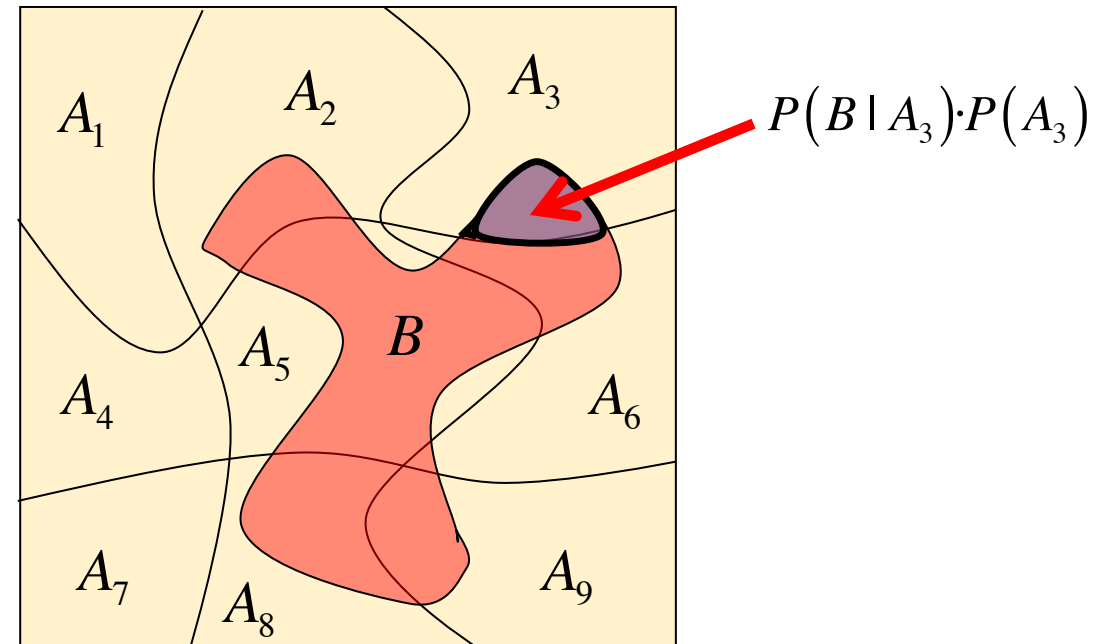
$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(A_k | B) = \frac{P(B | A_k) \cdot P(A_k)}{P(B)} \quad k = 1, \dots, N$$

if the events A_k are mutually exclusive, and they fill the universe

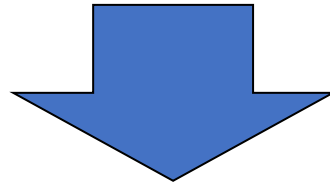
$$P(B) = \sum_{k=1}^N P(B | A_k) \cdot P(A_k)$$



$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

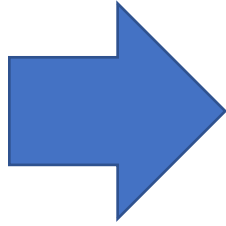


$$P(B) = \sum_{k=1}^N P(B|A_k) \cdot P(A_k)$$



$$P(A_k|B) = \frac{P(B|A_k)}{\sum_{j=1}^N P(B|A_j) P(A_j)} P(A_k)$$

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j)P(H_j)} P(H_k)$$



MAP estimates

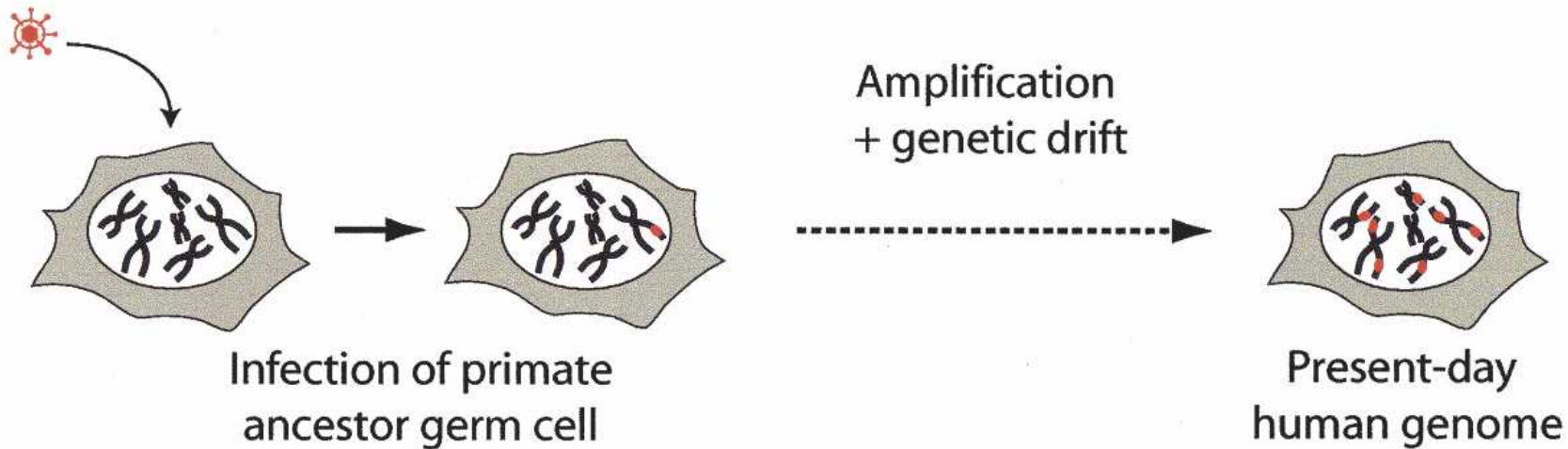
Example of MAP inference: the case of the Phoenix virus

Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements

Marie Dewannieux,^{1,3} Francis Harper,^{2,4} Aurélien Richaud,^{1,4} Claire Letzelter,¹ David Ribet,¹ Gérard Pierron,² and Thierry Heidmann^{1,5}

¹Unité des Rétrovirus Endogènes et Éléments Rétroïdes des Eucaryotes Supérieurs, UMR 8122 CNRS, Institut Gustave Roussy, 94805 Villejuif Cedex, France; ²Laboratoire de Réplication de l'ADN et Ultrastructure du Noyau, UPR1983 Institut André Lwoff, 94801 Villejuif Cedex, France

Human Endogenous Retroviruses are expected to be the remnants of ancestral infections of primates by active retroviruses that have thereafter been transmitted in a Mendelian fashion. Here, we derived in silico the sequence of the putative ancestral “progenitor” element of one of the most recently amplified family—the HERV-K family—and constructed it. This element, *Phoenix*, produces viral particles that disclose all of the structural and functional properties of a bona-fide retrovirus, can infect mammalian, including human, cells, and integrate with the exact signature of the presently found endogenous HERV-K progeny. We also show that this element amplifies via an extracellular pathway involving reinfection, at variance with the non-LTR-retrotransposons (LINEs, SINEs) or LTR-retrotransposons, thus recapitulating ex vivo the molecular events responsible for its dissemination in the host genomes. We also show that in vitro recombinations among present-day human *HERV-K* (also known as *ERV-K*) loci can similarly generate functional HERV-K elements, indicating that human cells still have the potential to produce infectious retroviruses.



Phoenix, the ancestral HERV-K(HML2) retrovirus

To construct a consensus HERV-K(HML2) provirus, we assembled all of the complete copies of the 9.4-kb proviruses that are human specific (excluding those with the 292-nt deletion at the beginning of the *env* gene) and aligned their nucleotide sequence to generate the consensus in silico, taking for each position the most frequent nucleotide. The resulting provirus sequence contains, as expected, ORFs for all of the HERV-K(HML2)-encoded proteins (Gag, Pro, Pol, Env, and the accessory Rec protein), with *gag*, *pro*, and *pol* separated by – 1 frameshifts. Noteworthy, this consensus provirus is distinct from each of the sequences used to generate it, with at least 20 amino acid changes on the overall sequences (Fig. 1).

* provirus = virus genome integrated into DNA of host cell

Notes:

- Nearly 8% of the human genome is composed of sequences of retroviral origin.
- HERV = Human Endogenous RetroVirus
- ORF = Open Reading Frame. The part of the reading frame that has the potential to be translated, a continuous stretch of codons that does not contain the stop codon.
- The three major proteins encoded within the retroviral genome: Gag, Pol, and Env:
 - Gag is a polyprotein and is an acronym for Group Antigens (ag).
 - Pol is the reverse transcriptase.
 - Env is the envelope protein.

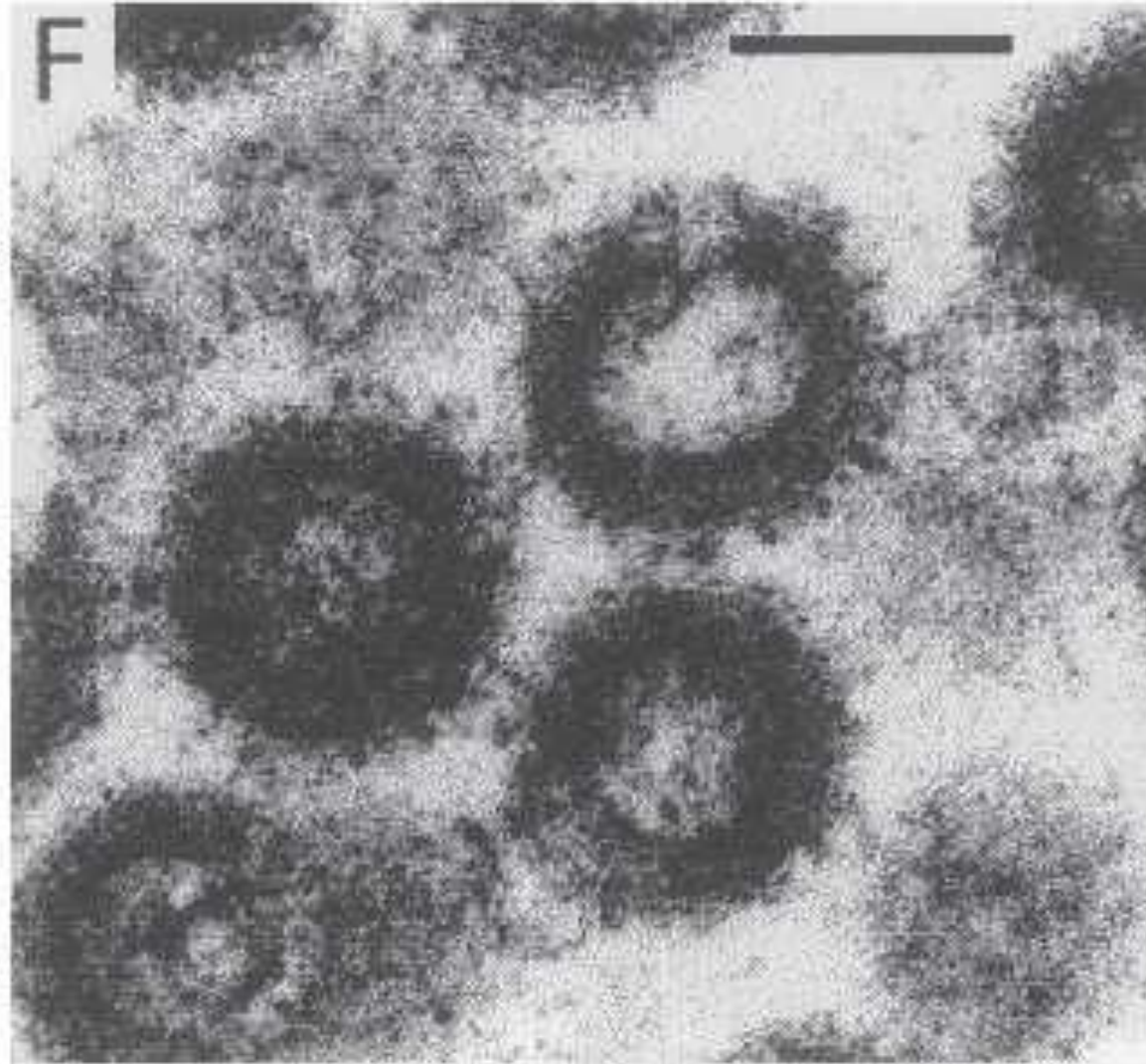


Image of representative particles obtained after transfection with an expression vector for the *Phoenix pro* mutant. Scale bar 100 nm.

A problem of male twins (Efron, 2003)



Pregnant with twins:
fraternal or identical?



Fraternal: $\frac{2}{3}$ of all cases



Identical: $\frac{1}{3}$ of all cases



**What is the probability of
identical twins IF both boys
in sonogram?**



Answer provided by Bayes theorem

$$P(\text{Identical}|\text{Both boys}) = \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})} P(\text{Identical})$$



$$\begin{array}{ll}
 P(\text{Identical}) = 1/3 & \left. \vphantom{\begin{array}{l} P(\text{Identical}) = 1/3 \\ P(\text{Fraternal}) = 2/3 \end{array}} \right\} \text{Prior probabilities} \\
 P(\text{Fraternal}) = 2/3 & \\
 P(\text{Both boys}|\text{Identical}) = 1/2 & \left. \vphantom{\begin{array}{l} P(\text{Both boys}|\text{Identical}) = 1/2 \\ P(\text{Both boys}|\text{Fraternal}) = 1/4 \end{array}} \right\} \text{Conditional probabilities from} \\
 P(\text{Both boys}|\text{Fraternal}) = 1/4 & \text{simple counting argument}
 \end{array}$$

$$\begin{aligned}
 P(\text{Both boys}) &= P(\text{Both boys}|\text{Identical})P(\text{Identical}) \\
 &\quad + P(\text{Both boys}|\text{Fraternal})P(\text{Fraternal}) \\
 &= (1/2)(1/3) + (1/4)(2/3) = 1/3
 \end{aligned}$$

$$\begin{aligned}
 P(\text{Identical}|\text{Both boys}) &= \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})} P(\text{Identical}) \\
 &= \frac{(1/2)}{(1/3)} (1/3) = 1/2
 \end{aligned}$$

A simple application to medical tests (HIV testing)

$$P(\text{positive}|\text{infected}) = 1; \quad P(\text{positive}|\text{not infected}) = 0.015$$

what is the probability $P(\text{infected}|\text{positive})$?

A common answer is 98.5% ... and it is wrong!

Let's use Bayes' theorem ...

$$P(A_k | B) = \frac{P(B | A_k) \cdot P(A_k)}{\sum_{k=1}^N P(B | A_k) \cdot P(A_k)}$$

$$\begin{aligned} P(\text{infected}|\text{positive}) &= \frac{P(\text{positive}|\text{infected}) \times P(\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \\ &= \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected}) \end{aligned}$$

The estimate depends on the size of the infected population

i.e., on the probabilities

P(infected)

P(not infected)

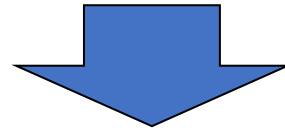
$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected})$$

The posterior estimate depends very strongly on the prior probability

Example: AIDS testing

(data from https://en.wikipedia.org/wiki/List_of_countries_by_HIV/AIDS_adult_prevalence_rate, accessed May 7th 2022)

$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected})$$



$$P_{\text{Italy}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.003 + 0.015 \times 0.997} \times 0.003 \approx 16.7\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.173 + 0.015 \times 0.827} \times 0.173 \approx 93.3\%$$

The large number of false positives and the small probability of finding a sick person mean that the probability of being infected if positive is not actually very high.

Repeating measurements changes the reference population.

We incorporate a new positive result in a repeated measurement by using the previous posterior as the new prior:

$$P_{\text{Italy}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.167 + 0.015 \times 0.833} \times 0.167 \approx 93.0\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.933 + 0.015 \times 0.067} \times 0.933 \approx 99.9\%$$

The first test changes the reference population, and the second test, if positive, gives a significant result.

Prosecutor's fallacy & Defendant's fallacy

Two common mistakes, associated with the wrong reference population

Consider a case where the probability of finding a given DNA subsequence – detected on a crime scene – is 0.00014 in the whole population: what is the probability that an individual who is found to have this rare subsequence in his/her DNA is guilty ???

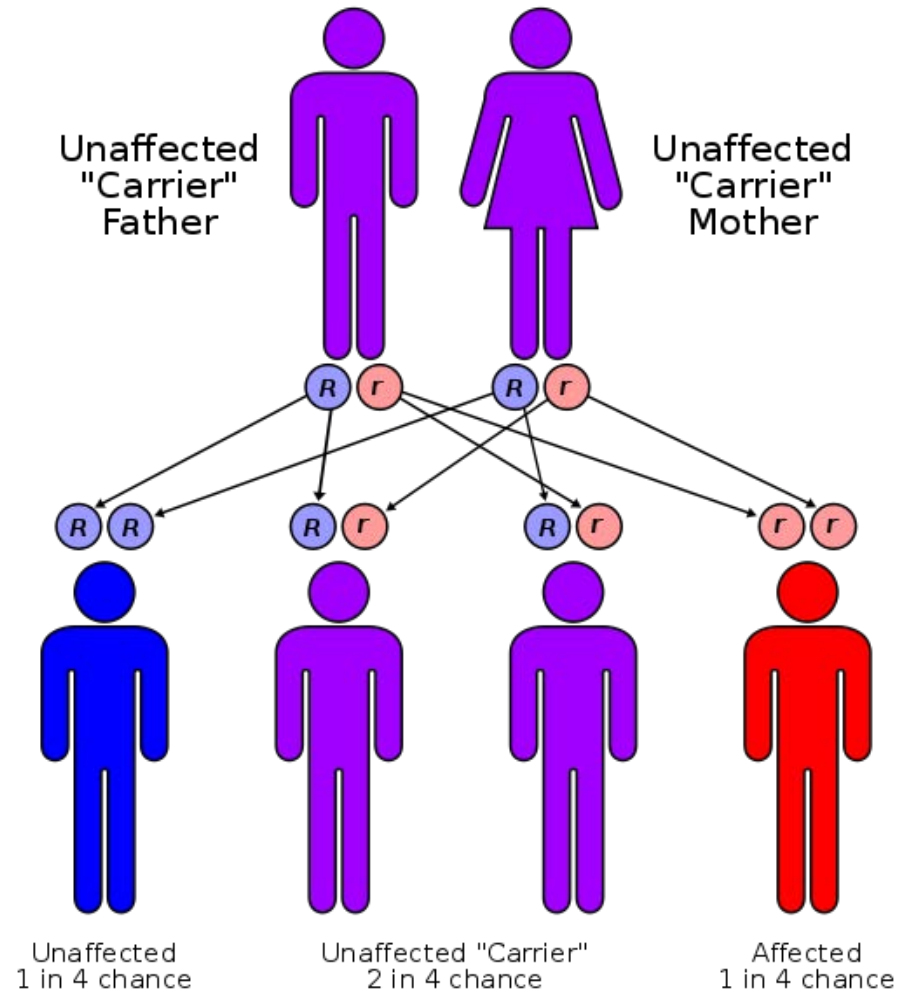
A common answer is $1 - 0.00014 = 0.99986$... but this is WRONG !

$P(\text{innocent} \mid \text{DNA compatible})$

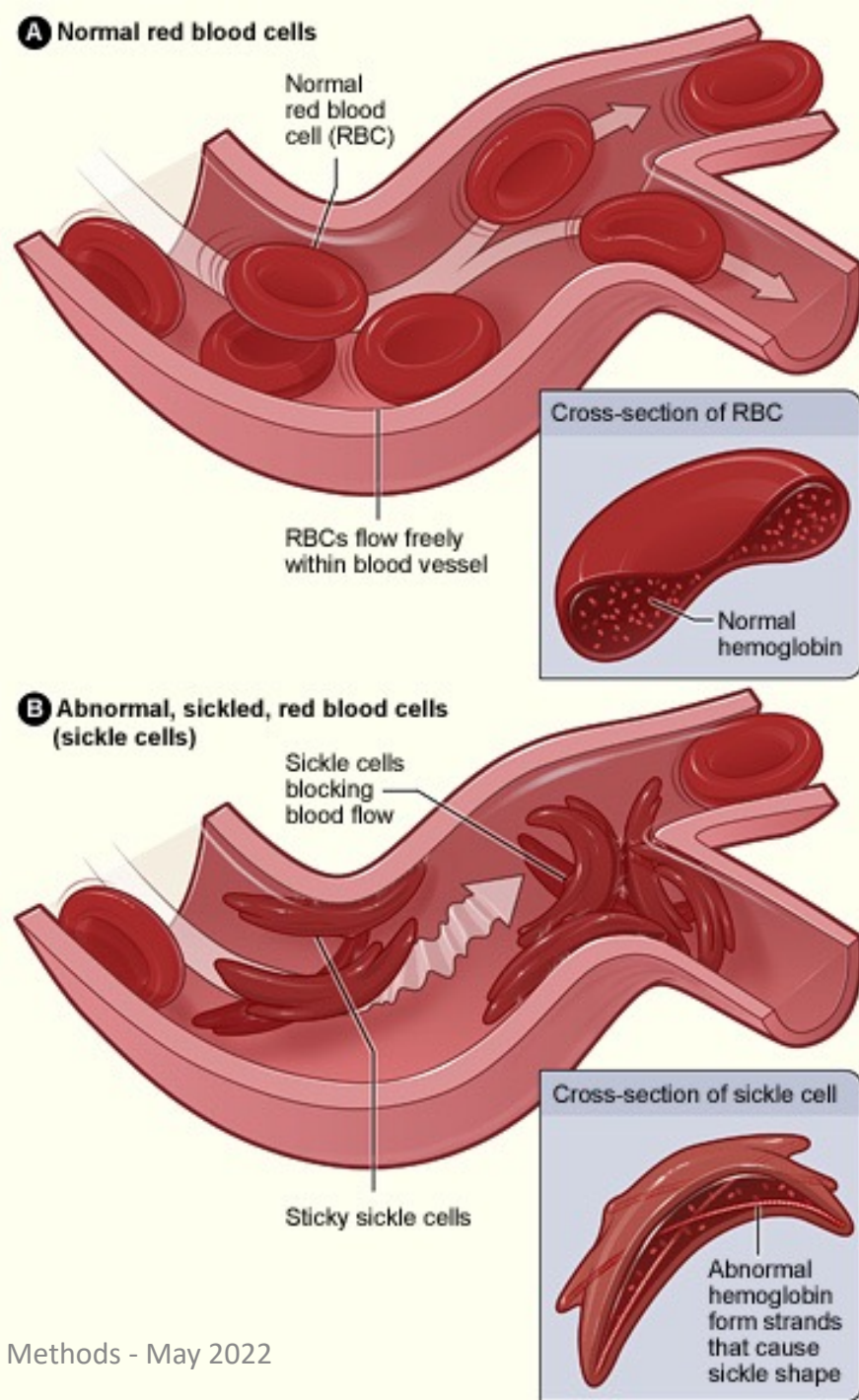
this is what we
actually want!

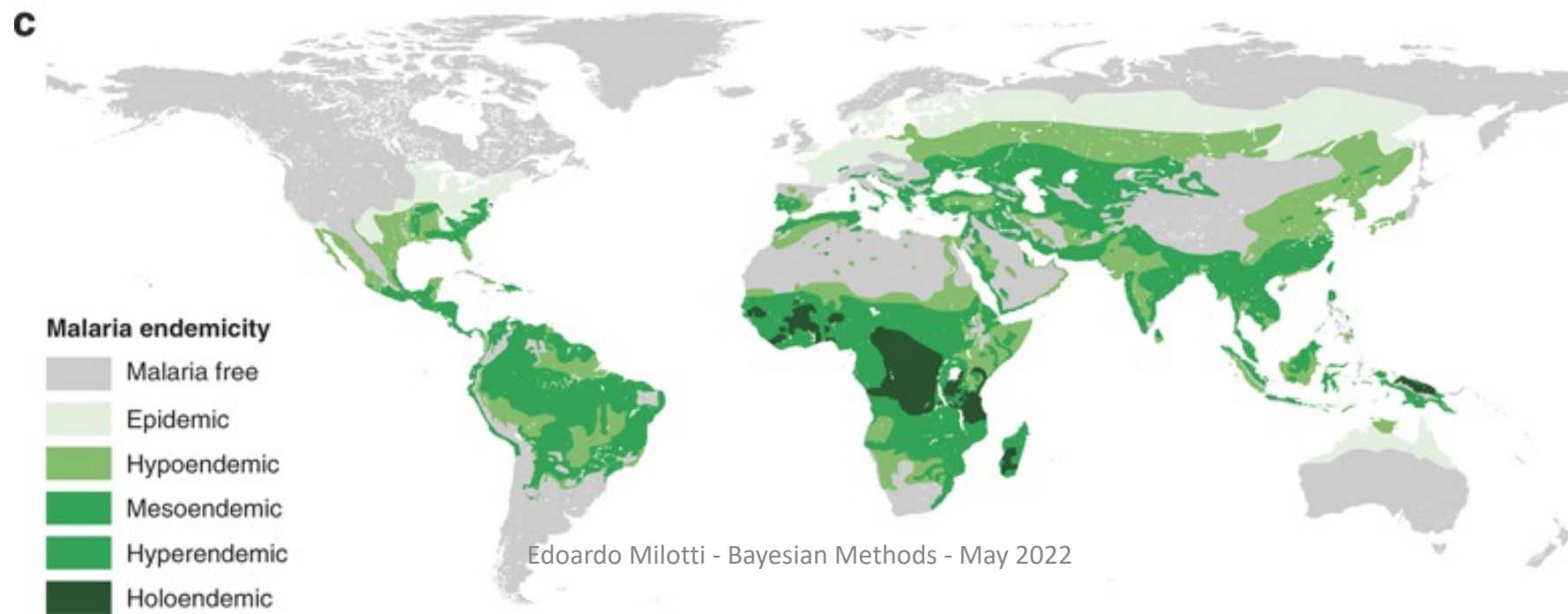
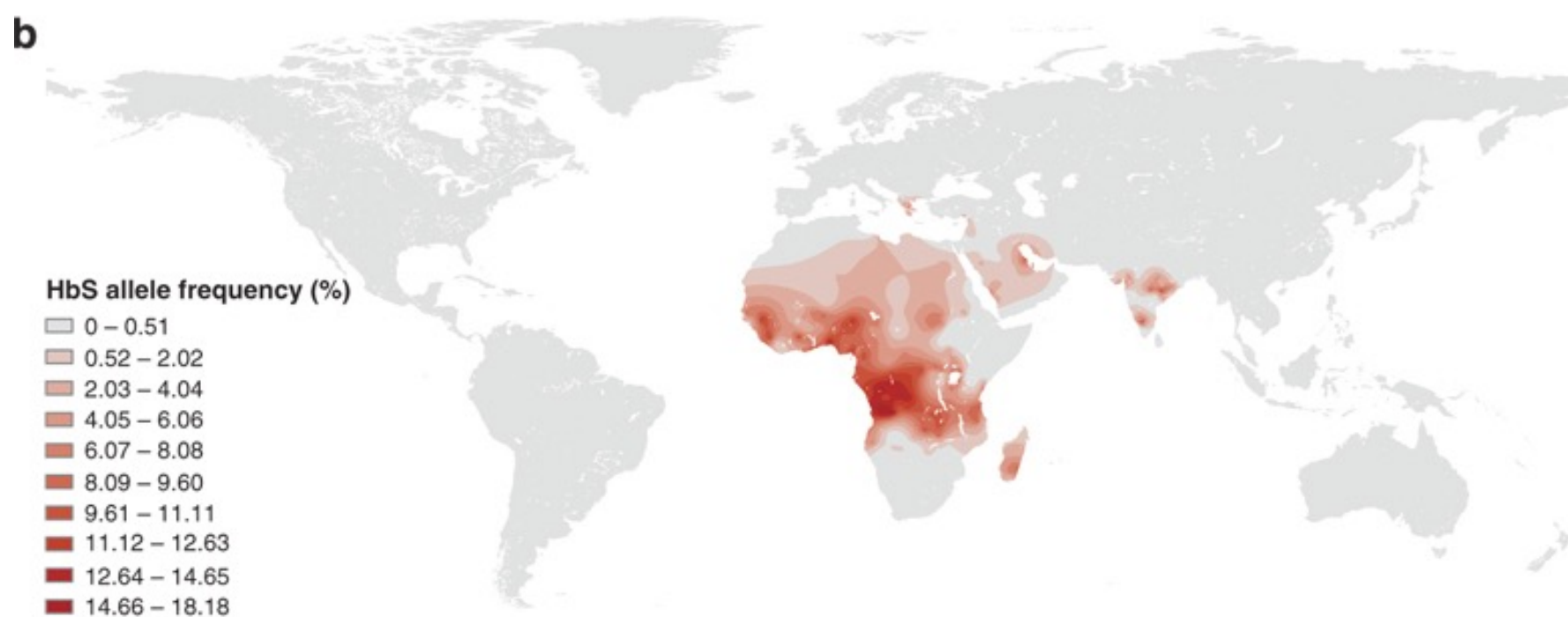
DNA classification - 1: alleles

allele: one of two or more alternative forms of the same gene, at the same position in a chromosome.



example: sickle
cell anemia





from F. B. Piel et al., Nature Comm. 1 (2010) 104

DNA classification - 2: allele frequency

DNA Profile		Allele frequency from database			Genotype frequency for locus		
Locus	Alleles	times allele observed	size of database	Frequency		formula	number
CSF1PO	10	109	432	$p=$	0.25	$2pq$	0.16
	11	134		$q=$	0.31		
TPOX	8	229	432	$p=$	0.53	p^2	0.28
	8						
THO1	6	102	428	$p=$	0.24	$2pq$	0.07
	7	64		$q=$	0.15		
vWA	16	91	428	$p=$	0.21	p^2	0.05
	16						
			profile frequency=				0.00014

taken from <http://www.dna-view.com/profile.htm>

ALFRED: database of human alleles (ALele FREquency Database:
<http://alfred.med.yale.edu/alfred/index.asp>

$\approx 1/7000$, frequency of
 profile in reference
 population

$$P(\text{innocent} \mid \text{DNA compatible}, I) = \frac{P(\text{DNA compatible} \mid \text{innocent}, I)}{P(\text{DNA compatible}, I)} P(\text{innocent} \mid I)$$

$$P(\text{innocent} \mid \text{given allele sequence}, I) = \frac{P(\text{given allele sequence} \mid \text{innocent}, I)}{P(\text{given allele sequence}, I)} P(\text{innocent} \mid I)$$

where

$$\begin{aligned} P(\text{given allele sequence}, I) &= P(\text{given allele sequence} \mid \text{innocent}, I) P(\text{innocent} \mid I) \\ &\quad + P(\text{given allele sequence} \mid \text{guilty}, I) P(\text{guilty} \mid I) \end{aligned}$$

Since the test has a very low error probability, i.e.,

$$P(\text{given allele sequence} \mid \text{guilty}, I) \approx 1$$

we find

$$P(\text{given allele sequence}, I) = 0.00014 \times P(\text{innocent} \mid I) + 1 \times P(\text{guilty} \mid I)$$

Once again, just like in the previous example, we see that it is all-important to determine the prior probabilities $P(\text{innocent}|I)$ and $P(\text{guilty}|I)$. For instance, if we pick a suspect at random in a large population, e.g., in a city with 1 million inhabitants, then

$$P(\text{innocent}|I) = 1 - 10^{-6} = 0.999999; \quad P(\text{guilty}|I) = 10^{-6} = 0.000001$$

$$P(\text{given allele sequence}, I) = 0.00014 \times (1 - 10^{-6}) + 1 \times 10^{-6} \approx 0.000141$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-6}) \approx 0.992907$$

This last result shows that the DNA test is quite inconclusive in this case, because it decreases the probability that the suspect is innocent from 0.999999 to 0.992907, only. How can it be? The reason is that in this case the number of random matches is not small, indeed in this city there are on average $1000000/7000 \approx 143$ people that randomly match the given allele sequence.

The argument can be turned upside down by a cunning lawyer, who might claim that since there are so many random matches, the DNA test is not relevant. However it is not so, and this claim is the “defendant’s fallacy”. Indeed, the problem that we met above was that the starting population was far too large. Other evidence might considerably reduce the number of possible suspects, for instance a surveillance camera might help identify all the people who entered a building and who had a chance to commit the crime, and thus reduce the starting population to, say, 100 people. When we repeat the relevant calculations, we find

$$P(\text{innocent}|I) = 1 - 1/100 = 0.99; \quad P(\text{guilty}|I) = 1/100 = 0.01$$

$$P(\text{given allele sequence}, I) = 0.00014 \times 0.99 + 1 \times 0.01 \approx 0.01014$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.01014}(1 - 10^{-2}) \approx 0.0137$$

We see that the new situation is drastically different, the reason being that on average only $100/7000 \approx 0.0143$ people can randomly match the given allele sequence.

An extremely short history of early Bayesianism

- Rev. Thomas Bayes discovered an early form of Bayes' theorem (second half of 18th century)
- Price discovered the theorem inside Bayes' unpublished notes (end 18th century)
- Laplace reinvented a version of the theorem and later expanded it after studying the Bayes' notes (around 1800)
- Laplace successfully applied the theorem to many experimental data analysis problems (until about 1820)
- Laplace was sometimes ridiculed by people who did not understand some of his approaches
- Laplace discovered the basic version of the Central Limit Theorem and in his later life he abandoned the Bayes theorem in favor of frequency-based methods (until about 1830)
- After the death of Laplace, Bayes' theorem was nearly forgotten and cornered to the darkest parts of statistics (crossing the desert ...)

Bayesian inference

$$\begin{aligned} P(A_k | B) &= \frac{P(B | A_k) \cdot P(A_k)}{\sum_{k=1}^N P(B | A_k) \cdot P(A_k)} \\ &= \frac{P(B | A_k)}{\sum_{k=1}^N P(B | A_k) \cdot P(A_k)} \cdot P(A_k) \end{aligned}$$

$$P(H_k | D, I) = \frac{P(D | H_k, I)}{\sum_{k=1}^N P(D | H_k, I) \cdot P(H_k | I)} \cdot P(H_k | I)$$

$$P(H_k | D, I) = \frac{P(D | H_k, I)}{\sum_{k=1}^N P(D | H_k, I) \cdot P(H_k | I)} \cdot P(H_k | I)$$

(Posterior probability that k -th hypothesis is true, when we observe data D , with prior information I)

=

(Probability of observing data D , given the k -th hypothesis) / Normalization
 ·
 (Prior probability that k -th hypothesis is true)

$$\begin{aligned}
 P(H_k \mid D, I) &= \frac{P(D \mid H_k, I)}{P(D \mid I)} \cdot P(H_k \mid I) \\
 &= \frac{P(D \mid H_k, I)}{\sum_{k=1}^N P(D \mid H_k, I) \cdot P(H_k \mid I)} \cdot P(H_k \mid I)
 \end{aligned}$$

prior distribution

$$P(H_k, I)$$

posterior distribution

$$P(H_k \mid D, I)$$

*likelihood or sampling
distribution*

$$P(D \mid H_k, I)$$

*evidence
(normalizing factor)*

$$P(D \mid I) = \sum_{k=1}^N P(D \mid H_k, I) \cdot P(H_k \mid I)$$

Testing hypotheses

$$P(H_k | D, I) = \frac{P(D | H_k, I)}{P(D | I)} \cdot P(H_k | I)$$

$$\frac{P(H_k | D, I)}{P(H_n | D, I)} = \underbrace{\left(\frac{P(D | H_k, I)}{P(D | H_n, I)} \right)}_{\text{Odds ratio}} \cdot \underbrace{\left(\frac{P(H_k | I)}{P(H_n | I)} \right)}_{\text{Bayes' factor}}$$

When prior probabilities are the same (equally probable hypotheses), the posterior probability ratio depends only on the Bayes' factor:

$$\frac{P(H_k|D, I)}{P(H_n|D, I)} = \underbrace{\frac{P(D|H_k, I)}{P(D|H_n, I)}}_{\text{Bayes' factor}} \times \underbrace{\frac{P(H_k|I)}{P(H_n|I)}}_{\text{Odds ratio}} = \frac{P(D|H_k, I)}{P(D|H_n, I)}$$

Bayes' factor *Odds ratio* *Bayes' factor*

Uniform priors

From discrete sets of hypothesis to the continuum.
The Bayes' theorem in the context of parameter estimation.

$$P(H_k | D, I) = \frac{P(D | H_k, I)}{P(D | I)} \cdot P(H_k | I) = \frac{P(D | H_k, I)}{\sum_{k=1}^N P(D | H_k, I) \cdot P(H_k | I)} \cdot P(H_k | I)$$

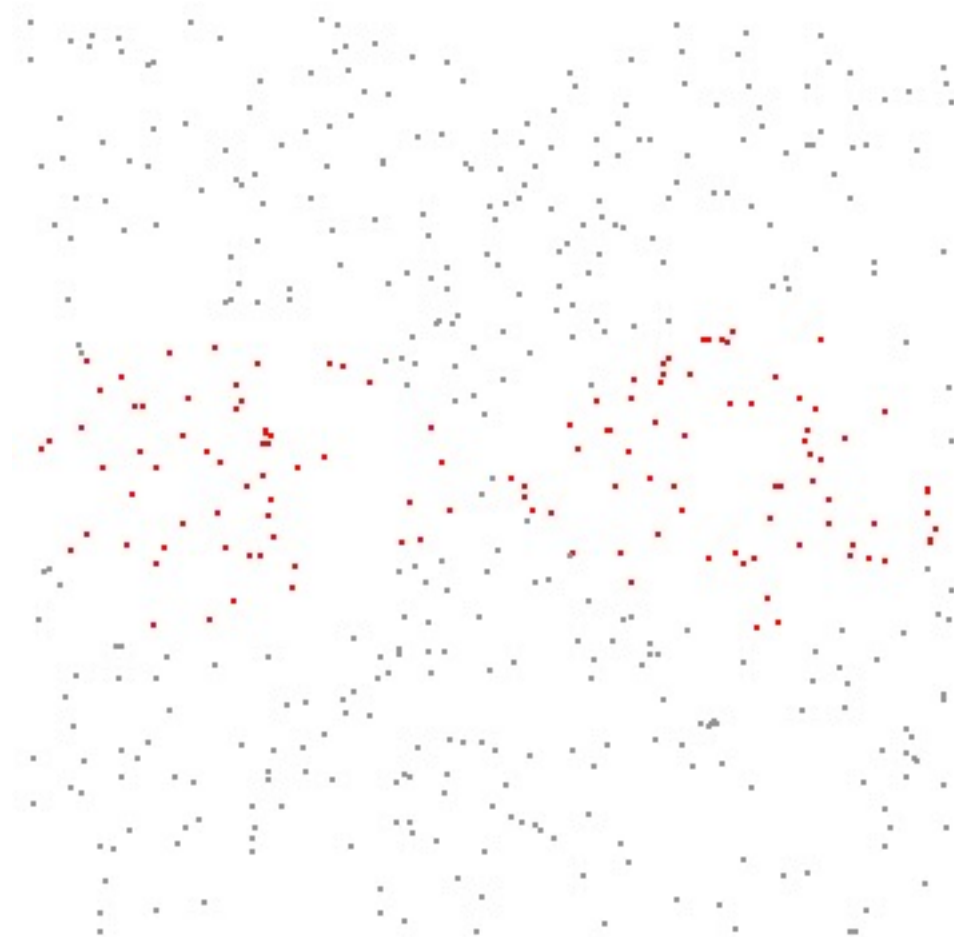


$$p(\boldsymbol{\theta} | D, I) = \frac{P(D | \boldsymbol{\theta}, I)}{\int_{\Theta} P(D | \boldsymbol{\theta}', I) p(\boldsymbol{\theta}' | I) d\boldsymbol{\theta}'} \times p(\boldsymbol{\theta} | I)$$

What if we “measure” a mathematical constant instead of a physical parameter?

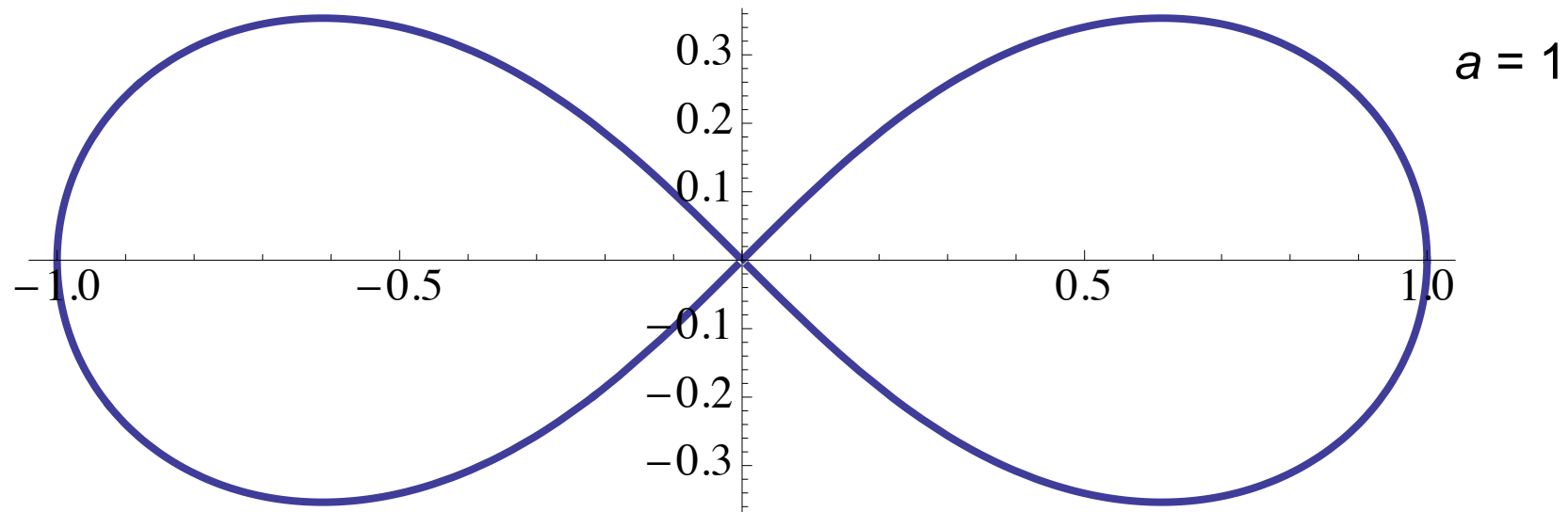
Example:

*area of Bernoulli's lemniscate
obtained with a Monte Carlo
simulation.*



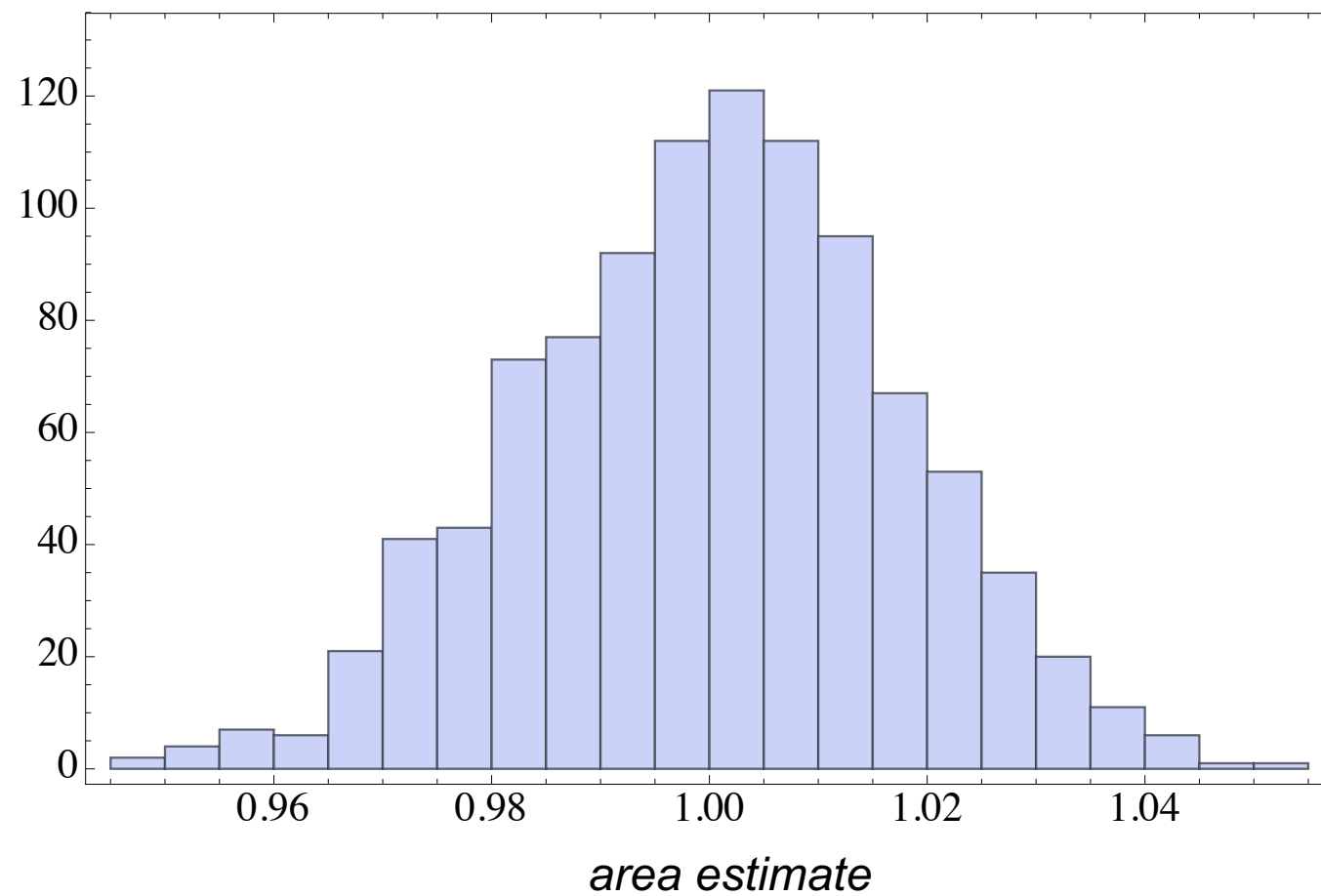
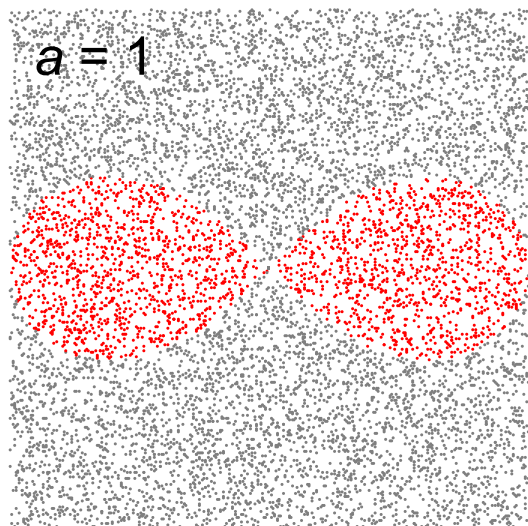
Parametric equation of Bernoulli's lemniscate

$$r = a\sqrt{\cos 2\theta}$$

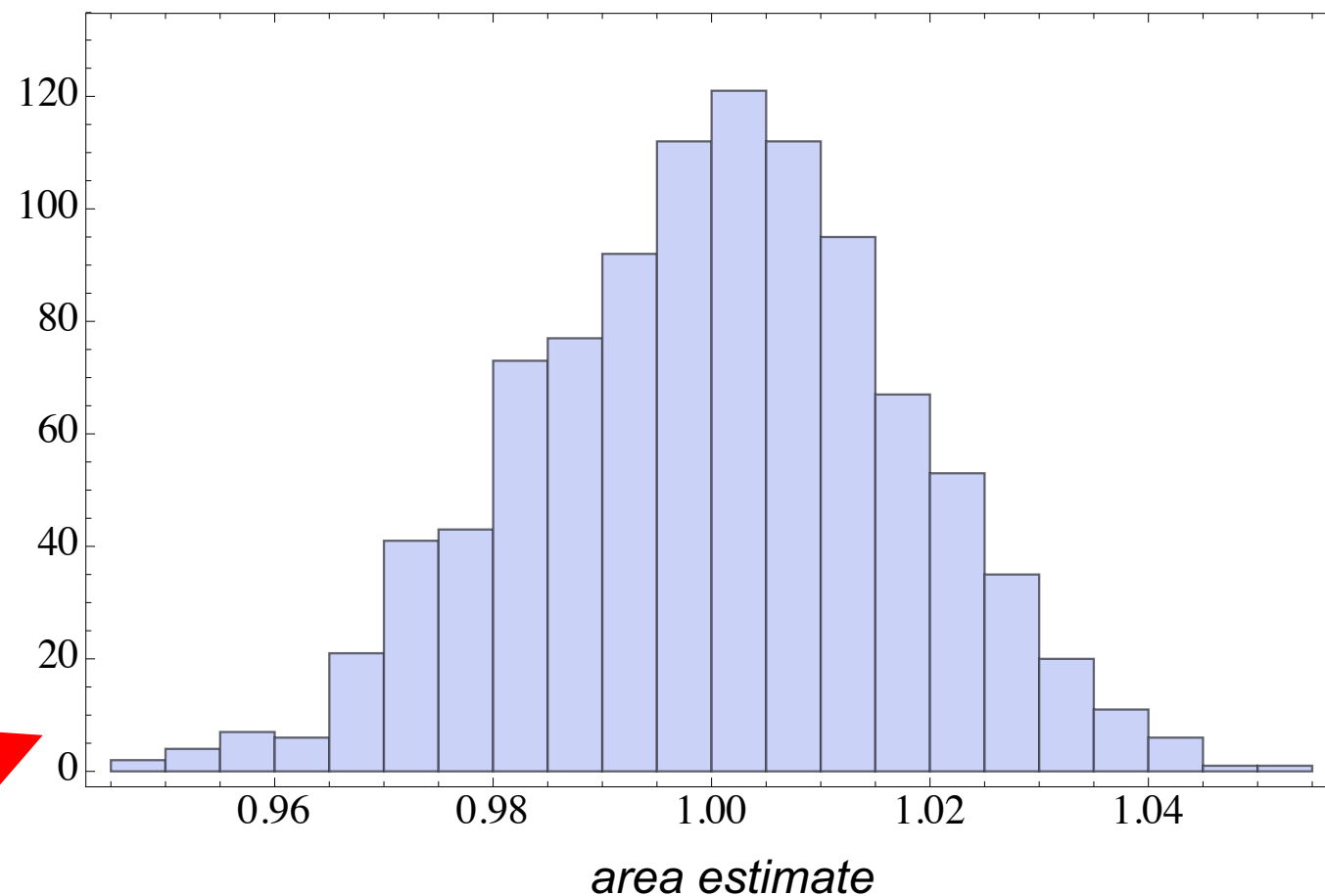
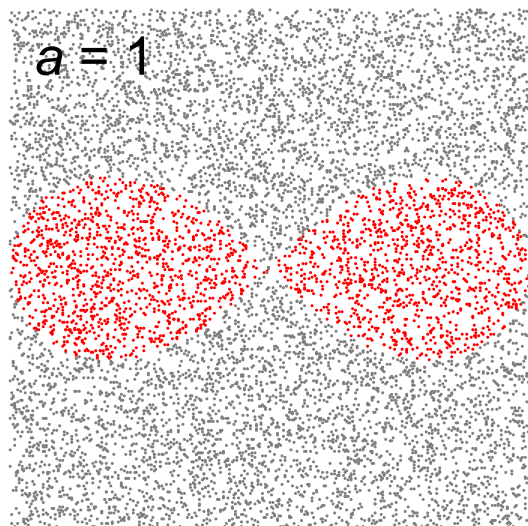


What is its area?

Empirical Monte Carlo distribution of the area estimate



Empirical Monte Carlo distribution of the area estimate



a probability distribution of
a mathematical constant???

Question:

Are we asking a real scientific question?

*If your experiment needs statistics, you
ought to have done a better experiment.*

(Ernest Rutherford, as reported by John Hammersley)

Question:

Why do we use statistics in science?