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

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

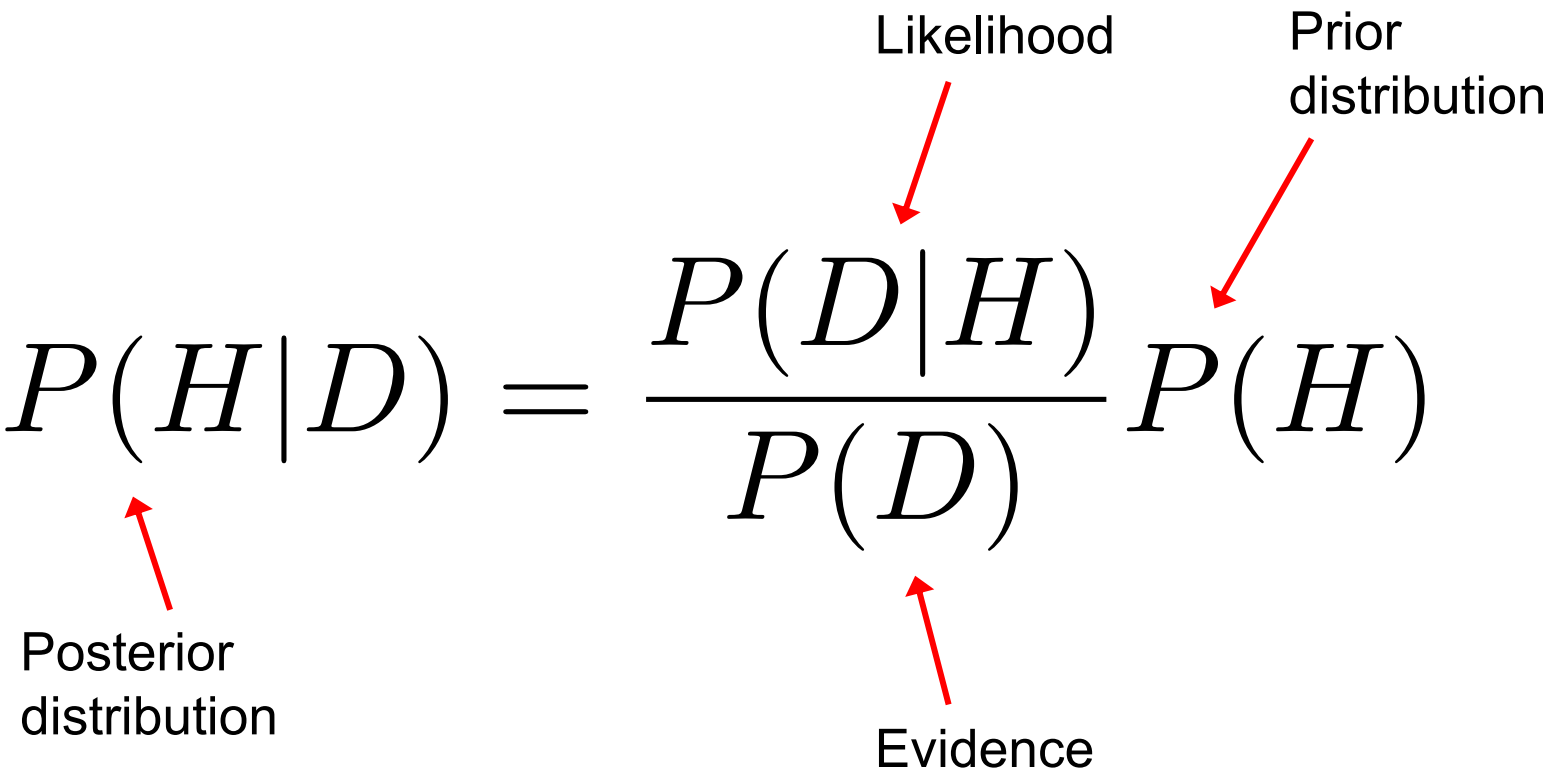
Joint probability and conditional probabilities

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

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

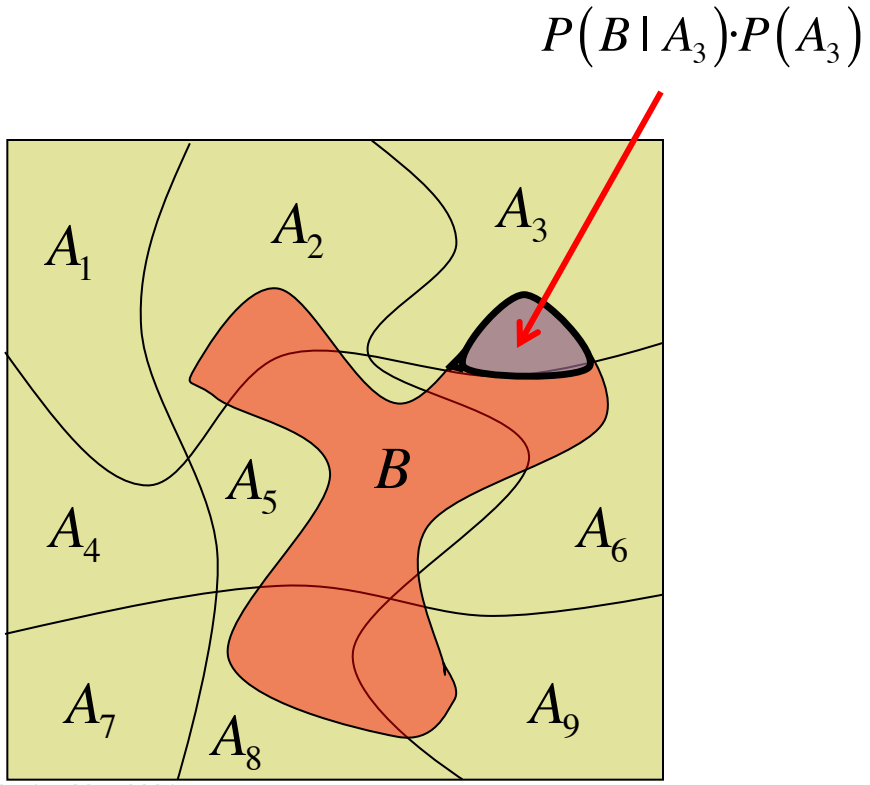

Likelihood
Prior distribution
$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$
Posterior distribution
Evidence

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

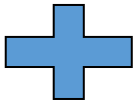
$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{P(B)} \qquad k = 1, \ldots, N$$

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

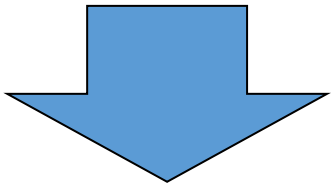
$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$



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

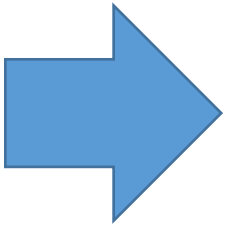


$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$



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

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



MAP estimates

Probabilities and 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]

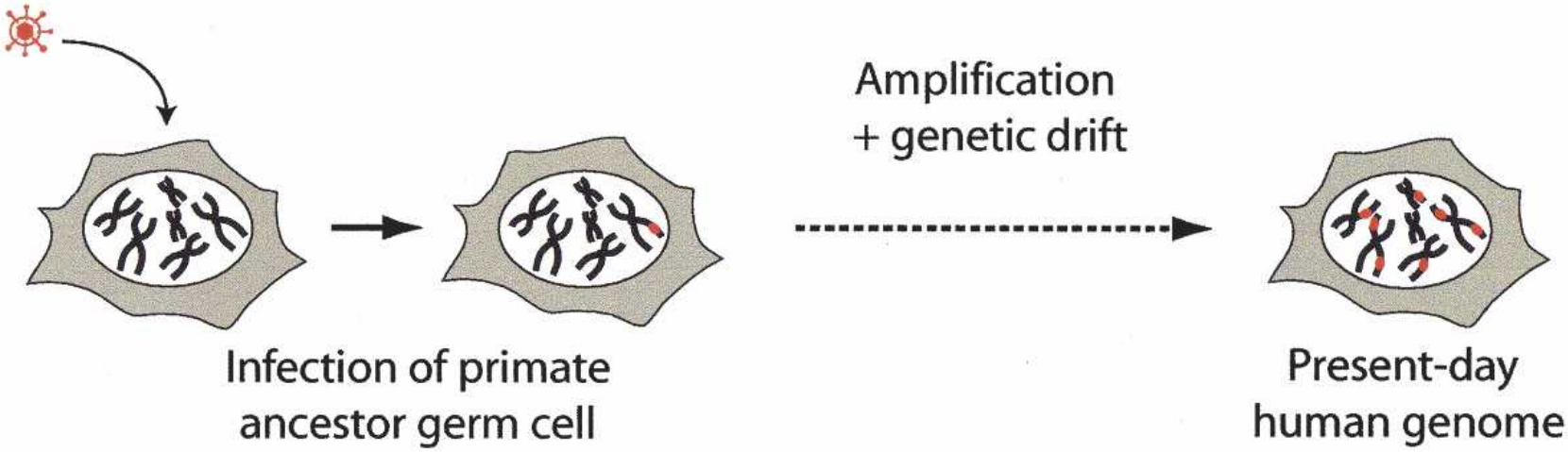
[1]*Unité des Rétrovirus Endogènes et Eléments Rétroïdes des Eucaryotes Supérieurs, UMR 8122 CNRS, Institut Gustave Roussy, 94805 Villejuif Cedex, France;* [2]*Laboratoire de Réplication de l'ADN et Ultrastructure du Noyau, UPR1983 Institut André Lwoff, 94801 Villejuif Cedex, France*

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 *ERVK*) loci can similarly generate functional HERV-K elements, indicating that human cells still have the potential to produce infectious retroviruses.

1548 **Genome Research** www.genome.org

16:1548–1556 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

•DOI: 10.1101/gr.5565706



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. Noteworthily, 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

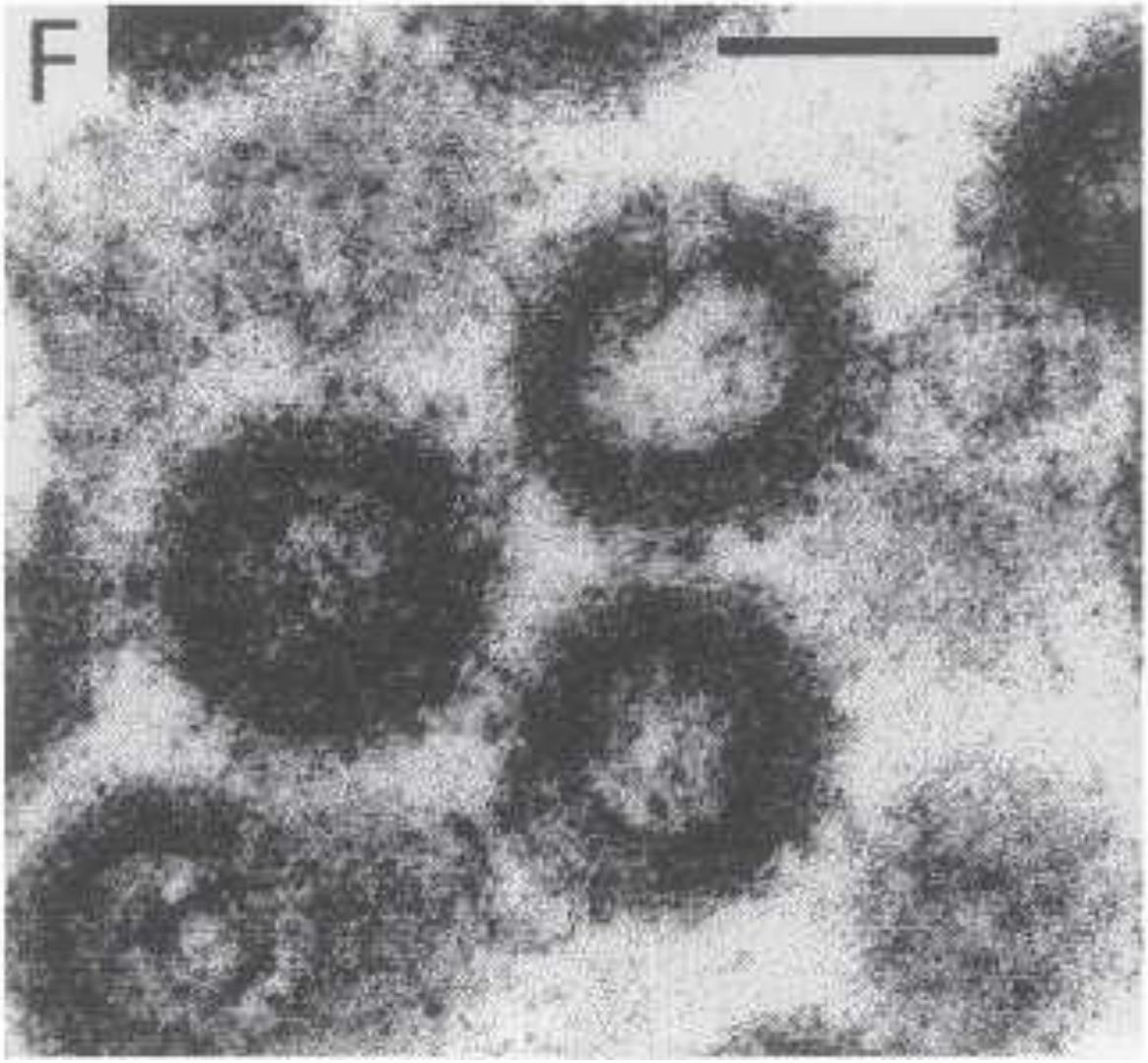


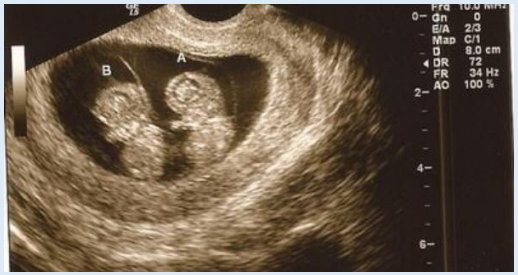
Image of representative particles obtained after transfection with an expression vector for the *Phoenix pro* mutant. Scale bar 100 nm.

A problem of male twins (Efron, 2003)

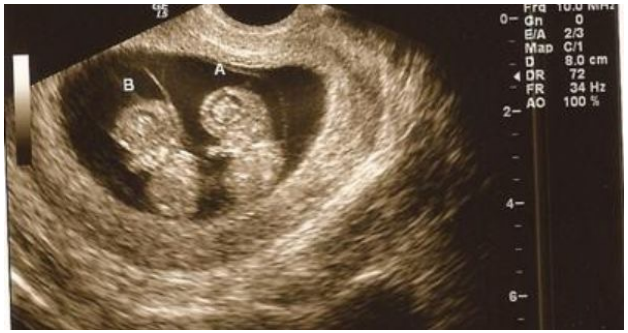
Pregnant with twins:
fraternal or identical?

Fraternal: 2/3 of all cases

Identical: 1/3 of all cases



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



Answer provided by Bayes theorem

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

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

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

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

A simple application to medical tests (example of HIV test)

$$P(positive \mid infect) = 1 \qquad P(positive \mid not\ infect) = 1.5\%$$

what is the probability *P(infect|positive)* ?

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

Let's use Bayes' theorem ...

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

$$P(infect \mid positive) = \frac{P(positive \mid infect) \cdot P(infect)}{P(positive \mid infect) \cdot P(infect) + P(positive \mid not\ infect) \cdot P(non\ infect)}$$

$$= \frac{P(positive \mid infect)}{P(positive \mid infect) \cdot P(infect) + P(positive \mid not\ infect) \cdot P(non\ infect)} \cdot P(infect)$$

The estimate depends on the size of the infect population

i.e., on the probabilities

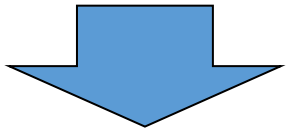
P(infect) P(not infect)

$$P(infect \mid positive) = \frac{P(positive \mid infect)}{P(positive \mid infect)\cdot P(infect) + P(positive \mid not\ infect)\cdot P(non\ infect)} \cdot P(infect)$$

The posterior estimate strongly depends on the prior probability

Example: AIDS frequency in Italy 0.4 %

AIDS frequency in South Africa 18.1%



$$P(infect \mid positive) = \frac{1}{1 \cdot 0.004 + 0.015 \cdot 0.996} \cdot 0.004 \approx 21.1\%$$ Italy

$$P(infect \mid positive) = \frac{1}{1 \cdot 0.181 + 0.015 \cdot 0.819} \cdot 0.181 \approx 93.6\%$$ South Africa

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.

If we find a positive result in a repeated measurement:

$$P\left(infect \middle| \{positive, positive\}\right) = 94.7\%$$ Italy

$$P\left(infect \middle| \{positive, positive\}\right) = 99.9\%$$ South Africa

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 to the wrong reference population

$$P(DNA\,compatible \mid innocent)$$

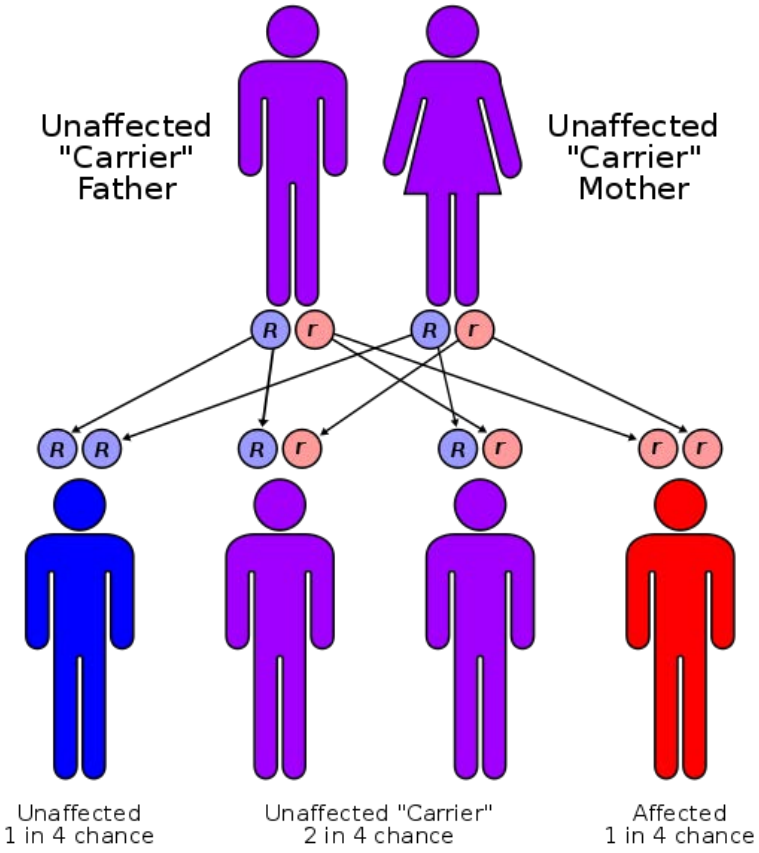
$$P(innocent \mid DNA\,compatible)$$

this is what we want!

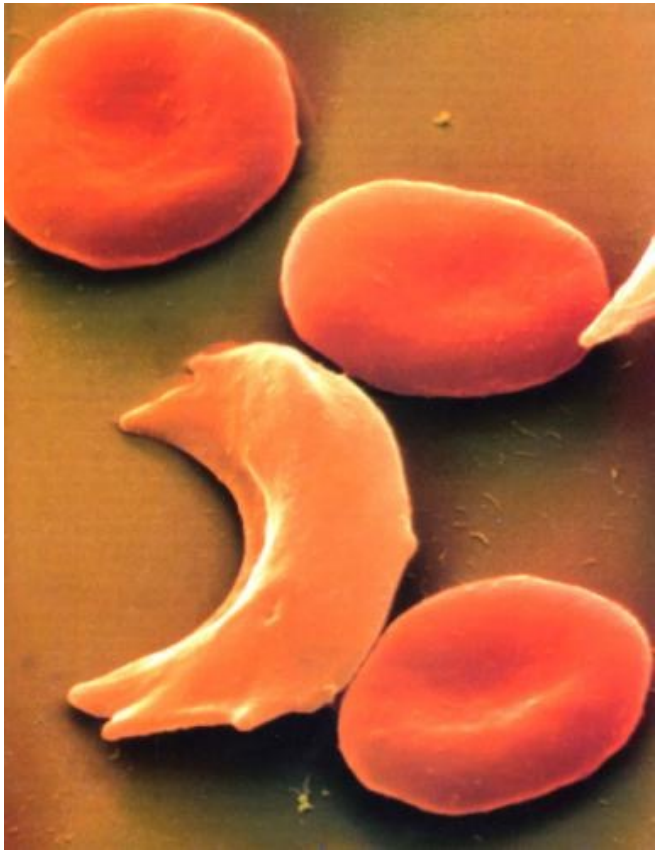
$$P(innocent \mid DNA\,compatible, I) = \frac{P(DNA\,compatible \mid innocent, I)}{P(DNA\,compatible, I)} P(innocent \mid I)$$

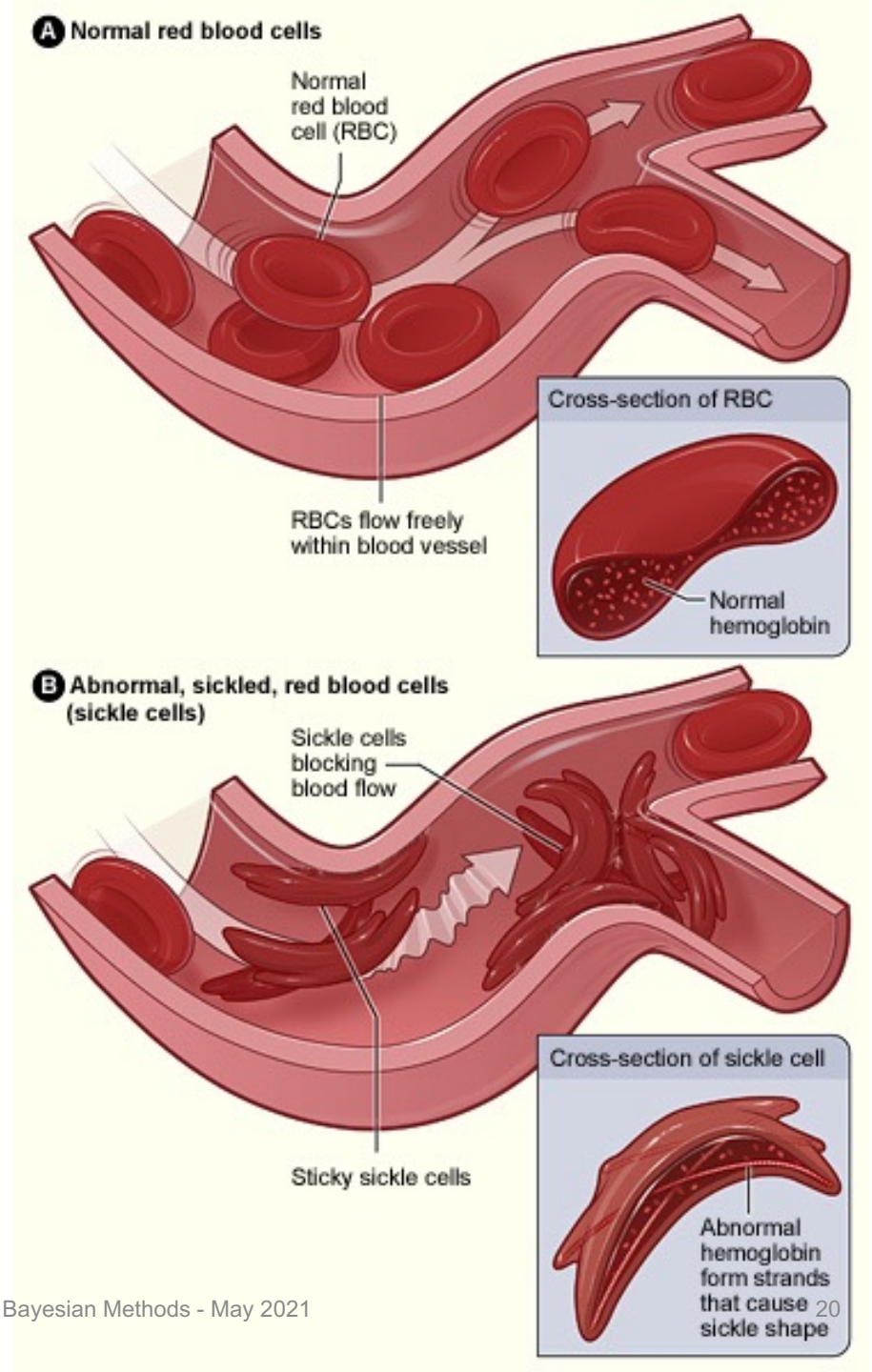
DNA classification - 1: alleles

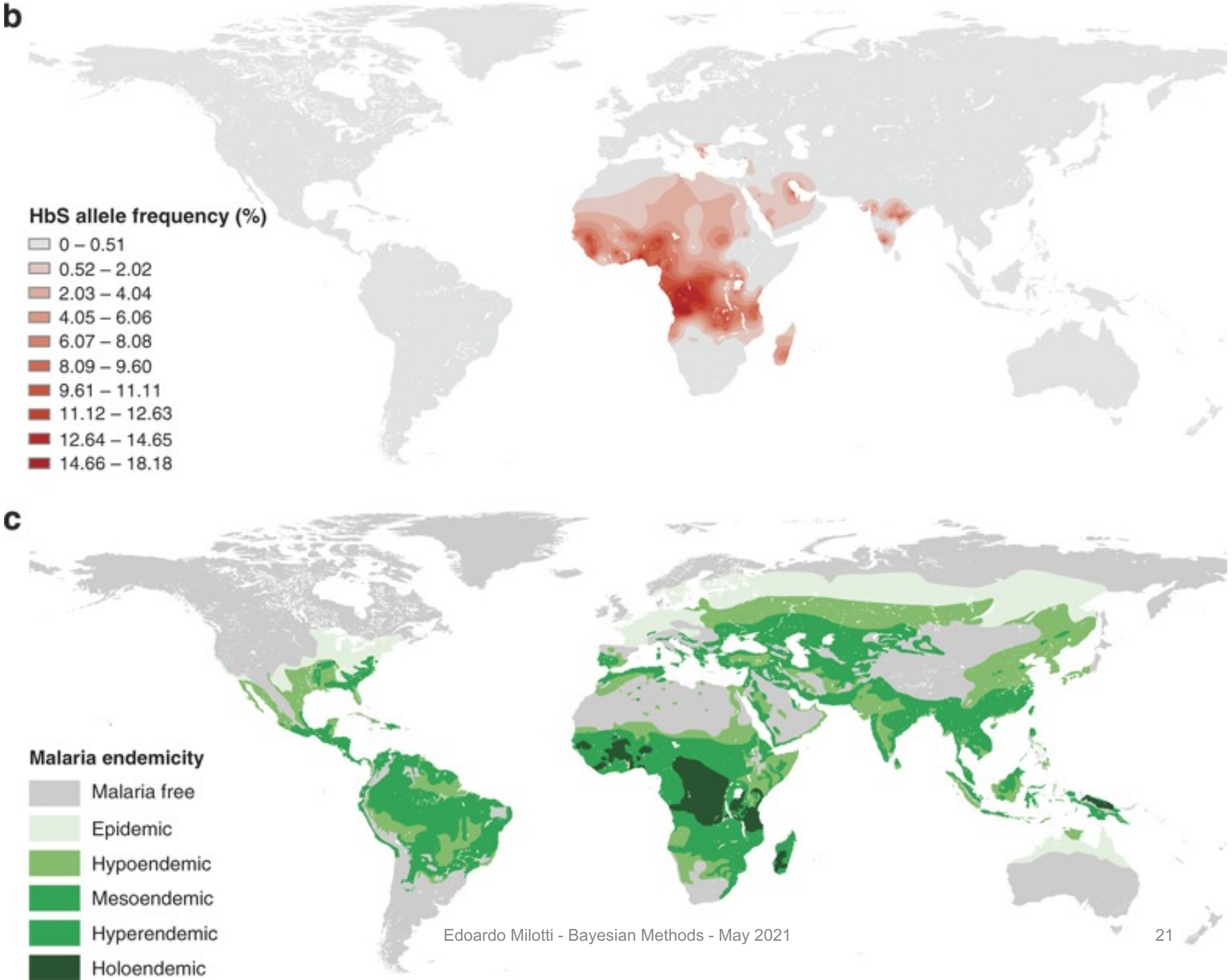
***allele**: one of two or more alternative forms of the same gene, at the same position in a chromosome.*



example: sickle cell anemia




A Normal red blood cells
Normal red blood cell (RBC)
RBCs flow freely within blood vessel
Cross-section of RBC
Normal hemoglobin
B Abnormal, sickled, red blood cells (sickle cells)
Sickle cells blocking blood flow
Sticky sickle cells
Cross-section of sickle cell
Abnormal hemoglobin form strands that cause sickle shape



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

DNA classification - 2: allele frequency

A copy

B copy

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	$p2$	0.05
	16						
			profile frequency=				**0.00014**

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

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

≈ 1/7000, frequency of profile in reference population

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

where

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

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

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

we find

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

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 favour 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

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

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

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

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

(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)

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

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

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

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

$$\frac{P(H_k \mid D, I)}{P(H_n \mid D, I)} = \left(\frac{P(D \mid H_k, I)}{P(D \mid H_n, I)} \right) \cdot \left(\frac{P(H_k \mid I)}{P(H_n \mid I)} \right)$$

Bayes' factor

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

$$\frac{P\left(H_k \mid D, I\right)}{P\left(H_n \mid D, I\right)} = \left(\frac{P\left(D \mid H_k, I\right)}{P\left(D \mid H_n, I\right)} \right)$$

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

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

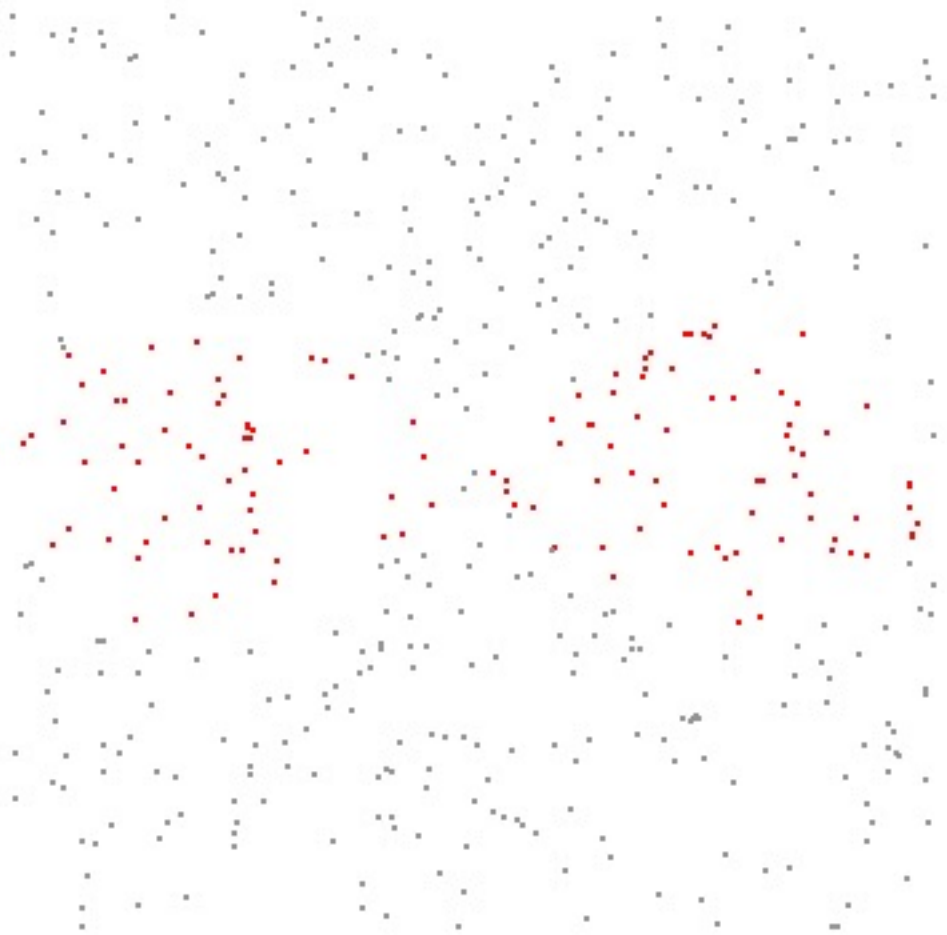
$$dP(\theta \mid D,I) = \frac{P(D \mid \theta,I)}{\int_{\Theta} P(D \mid \theta,I) \cdot dP(\theta \mid I)} \cdot dP(\theta \mid I)$$

$$\frac{dP(\theta \mid D,I)}{d\theta} = \frac{P(D \mid \theta,I)}{\int_{\Theta} P(D \mid \theta,I) \cdot \frac{dP(\theta \mid I)}{d\theta} d\theta} \cdot \frac{dP(\theta \mid I)}{d\theta}$$

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

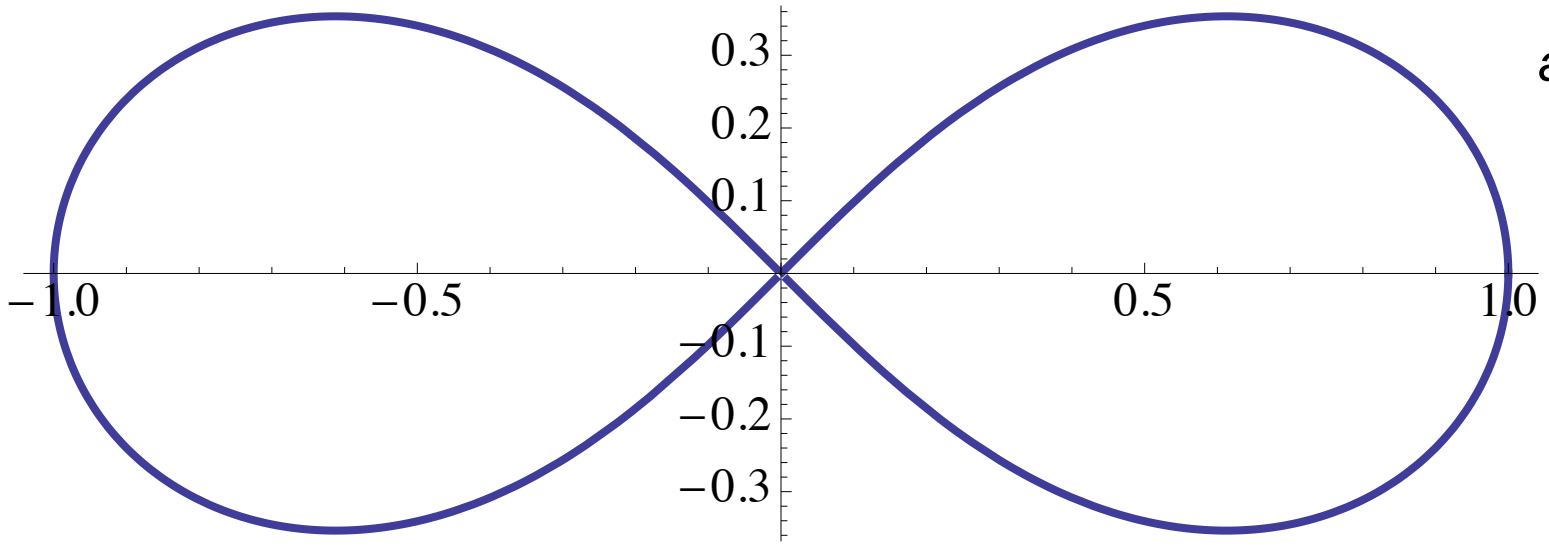
Example:

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



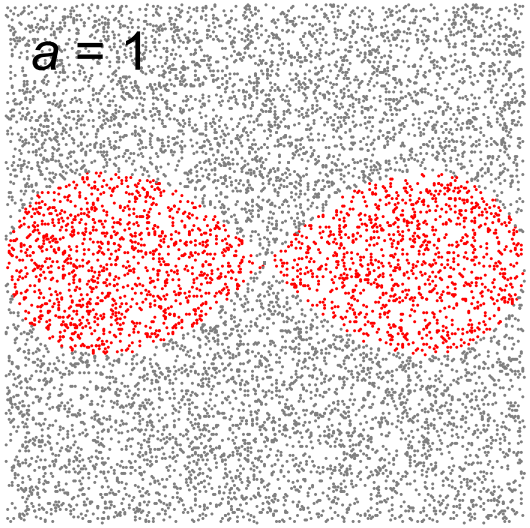
Parametric equation of Bernoulli's lemniscate

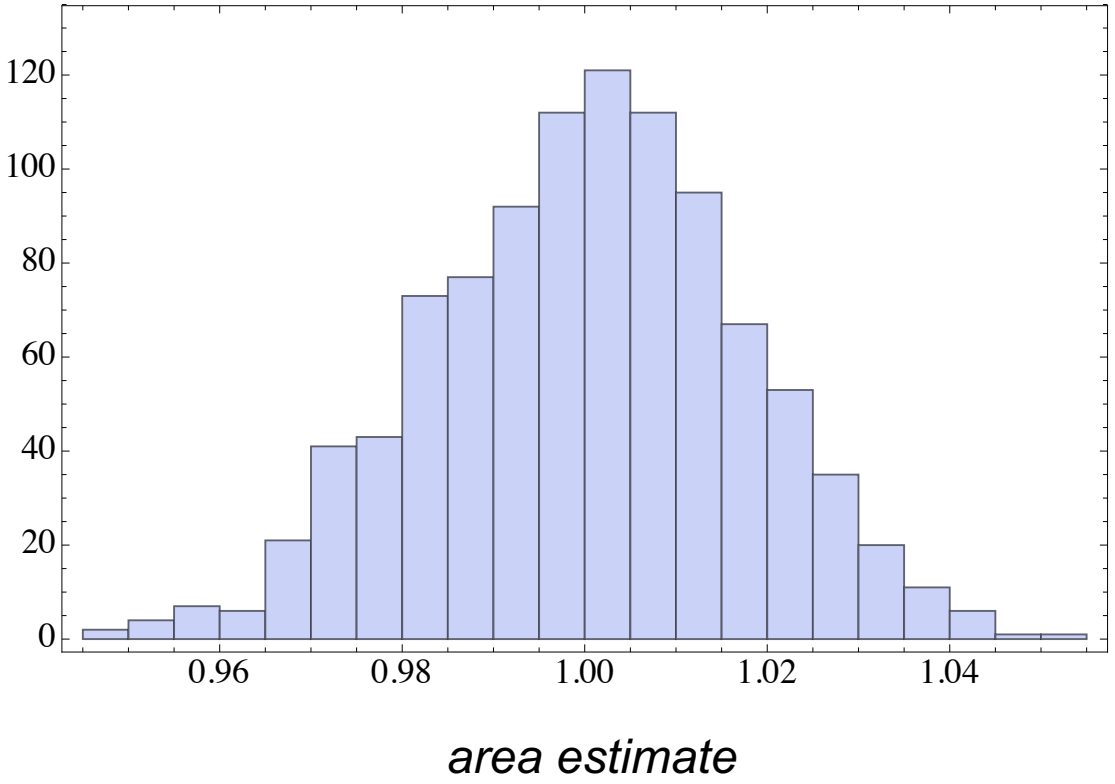
$$r = a\sqrt{\cos 2\theta}$$



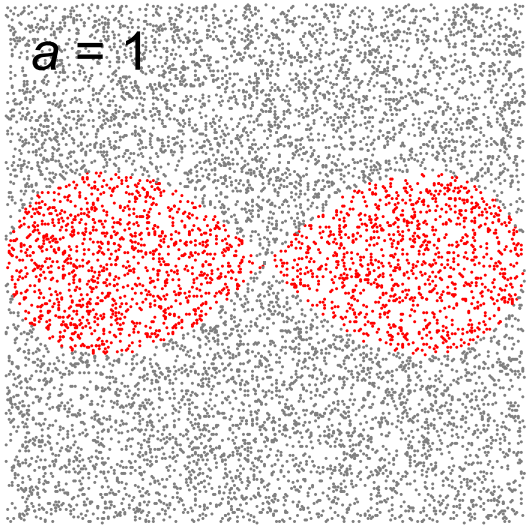
What is its area?

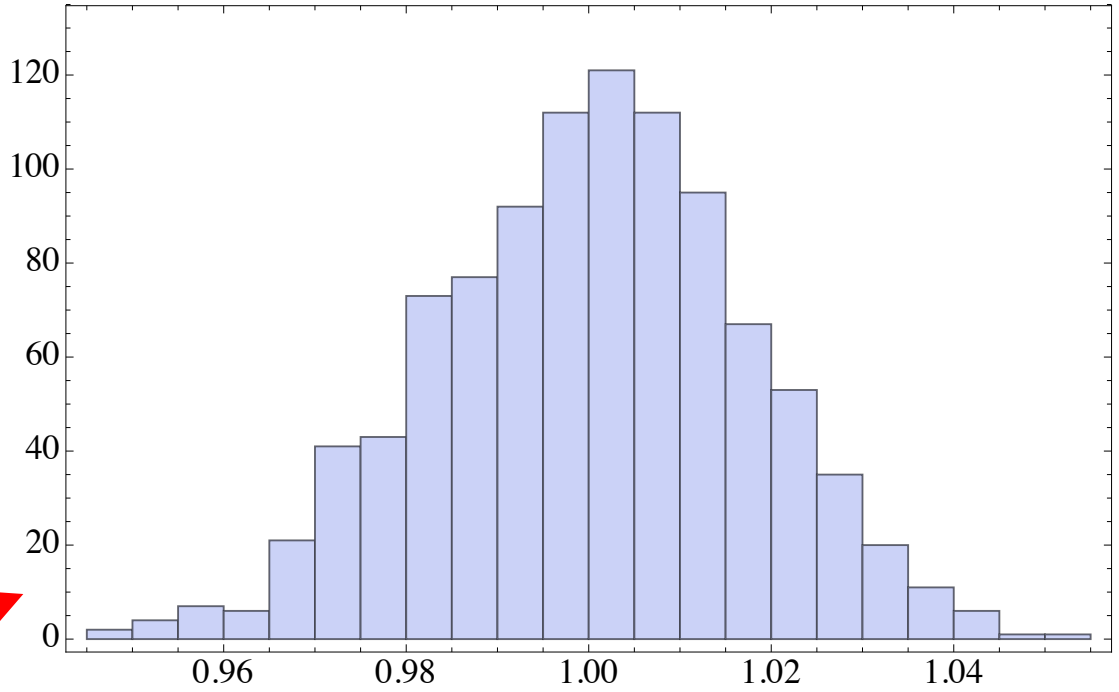
Empirical Monte Carlo distribution of the area estimate





Empirical Monte Carlo distribution of the area estimate





area estimate

a probability distribution of
a mathematical constant???

Frequentist view: this is the distribution of an estimate, it does not make sense to talk of the distribution of a constant.

Bayesian view: while in this case the value to be estimated is unmistakably “true”, this is not a real experiment where the model itself is not certain, and probability applies to it as well.

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?