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

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

(Ernest Rutherford, as reported by John Hammerseley)

Question:

What is the purpose of statistics in science?

Conditional probabilities and Bayes' Theorem

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

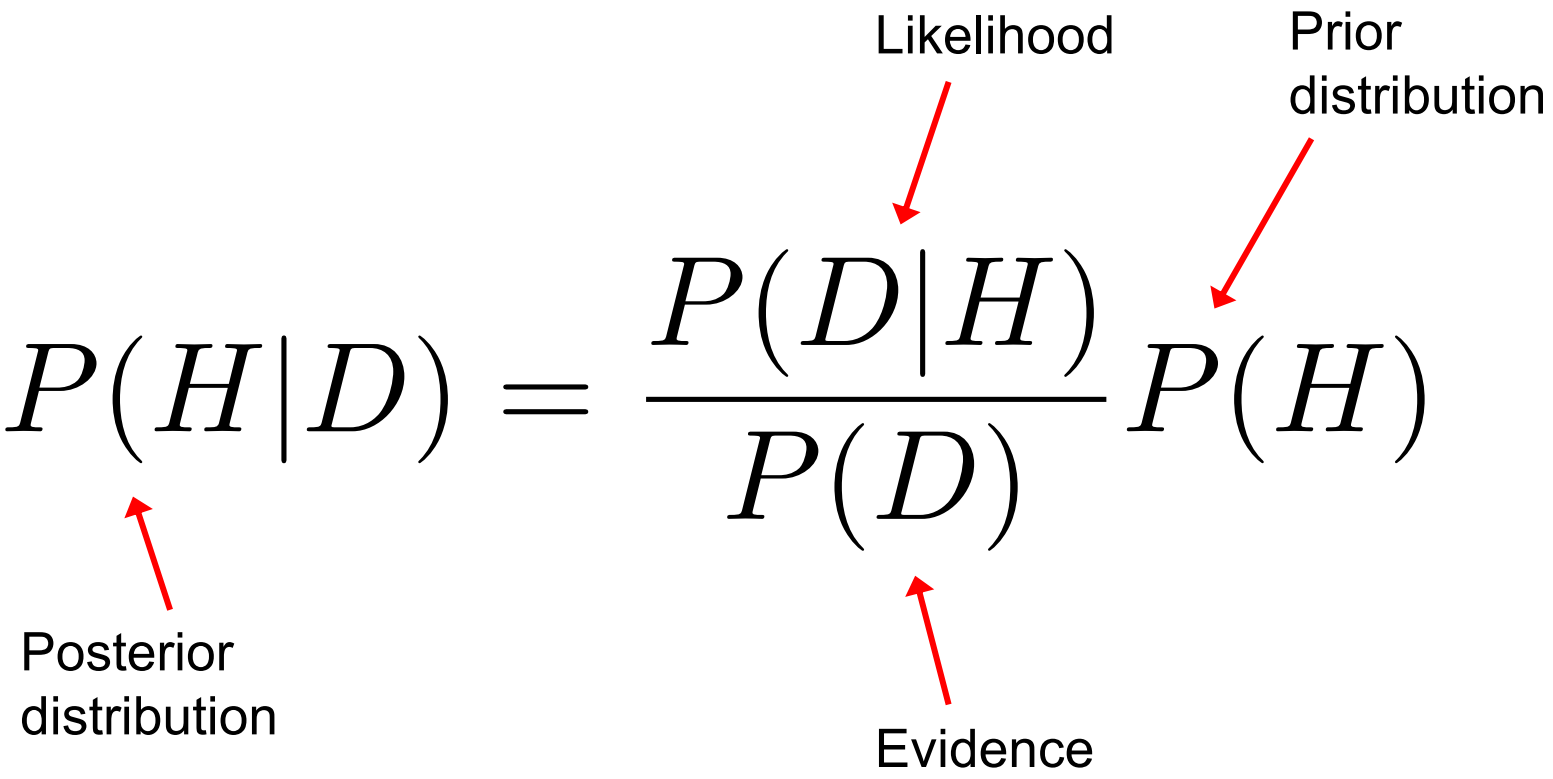
Joint probability and conditional probabilities

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

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

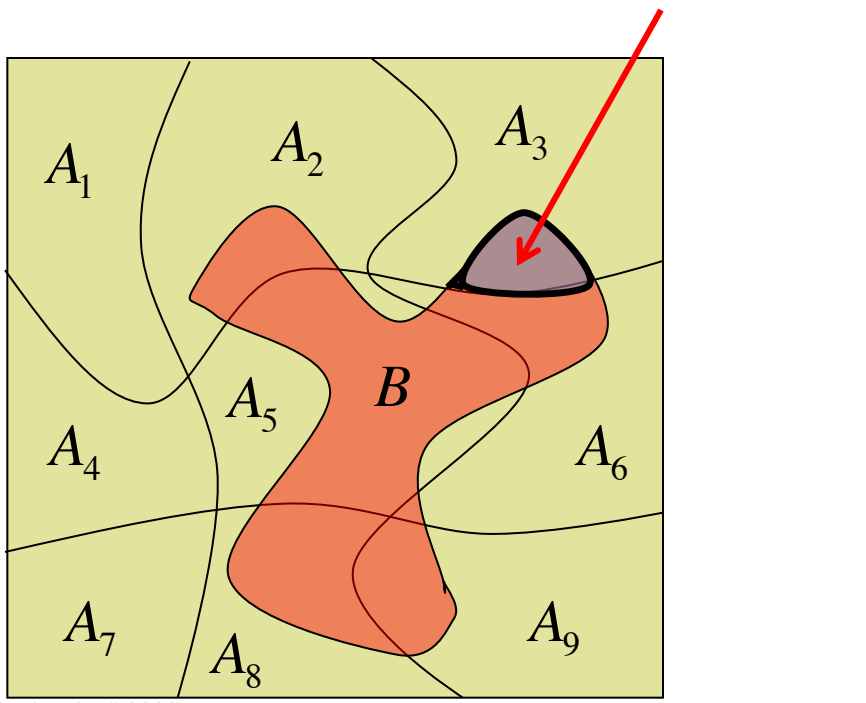

Likelihood
Prior distribution
$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$
Posterior distribution
Evidence

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

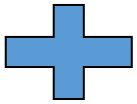
$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{P(B)} \qquad k = 1, \ldots, N$$

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

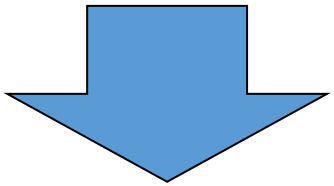
$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$



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

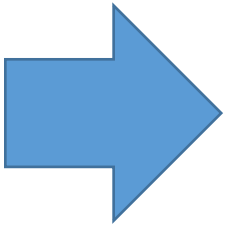


$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$



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

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



MAP estimates

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

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

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

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

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

Probabilities and inference: the case of the Phoenix virus

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

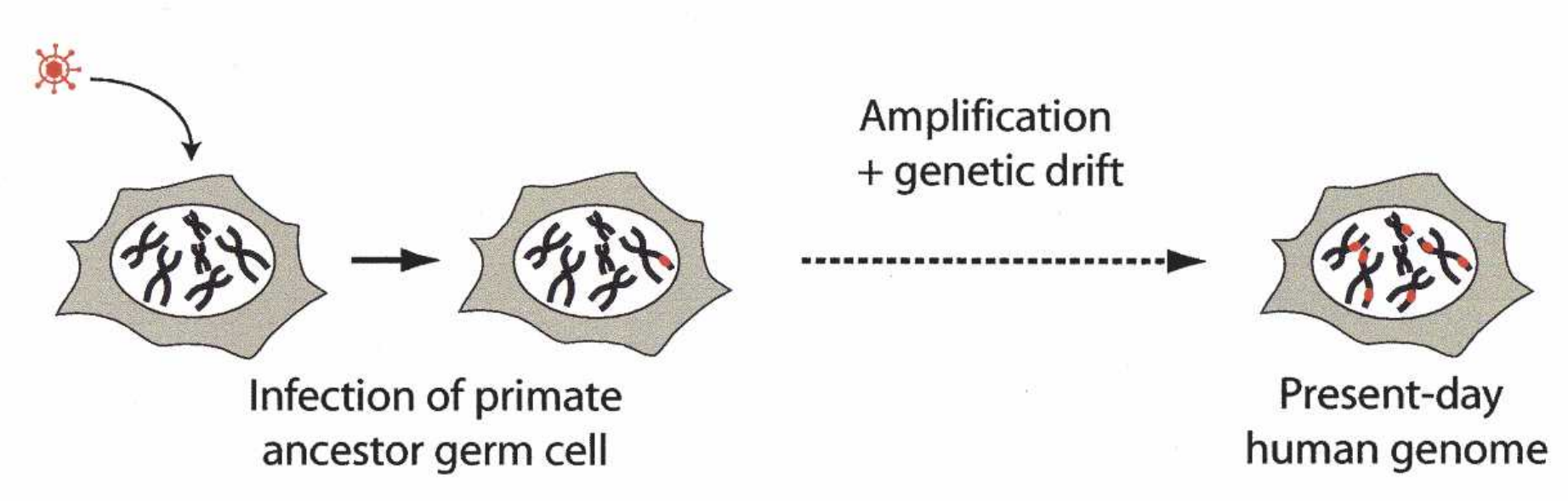
Marie Dewannieux,[1,3] Francis Harper,[2,4] Aurélien Richaud,[1,4] Claire Letzelter,[1] David Ribet,[1] Gérard Pierron,[2] and Thierry Heidmann[1,5]

[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



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.

* provirus = virus genome integrated into DNA of host cell

Name	GenBank Accession Number	FirstPosition	Last position	Orientation
K104	AC116309	123,567	114,122	—
K108-1	AC072054	47,417	37,947	—
K108-2	AC072054	38,914	29,443	—
K109	AC055116	139,321	148,740	+
K113	AY037928	1	9,472	+
K115	AY037929	1	9,463	+
	Y178333	1	8,629	+
	AP000776	101,084	110,549	+
	AC025420	37,159	46,615	+

This table provides the GenBank coordinates of the human endogenous HERV-K copies used to generate the *Phoenix* provirus.

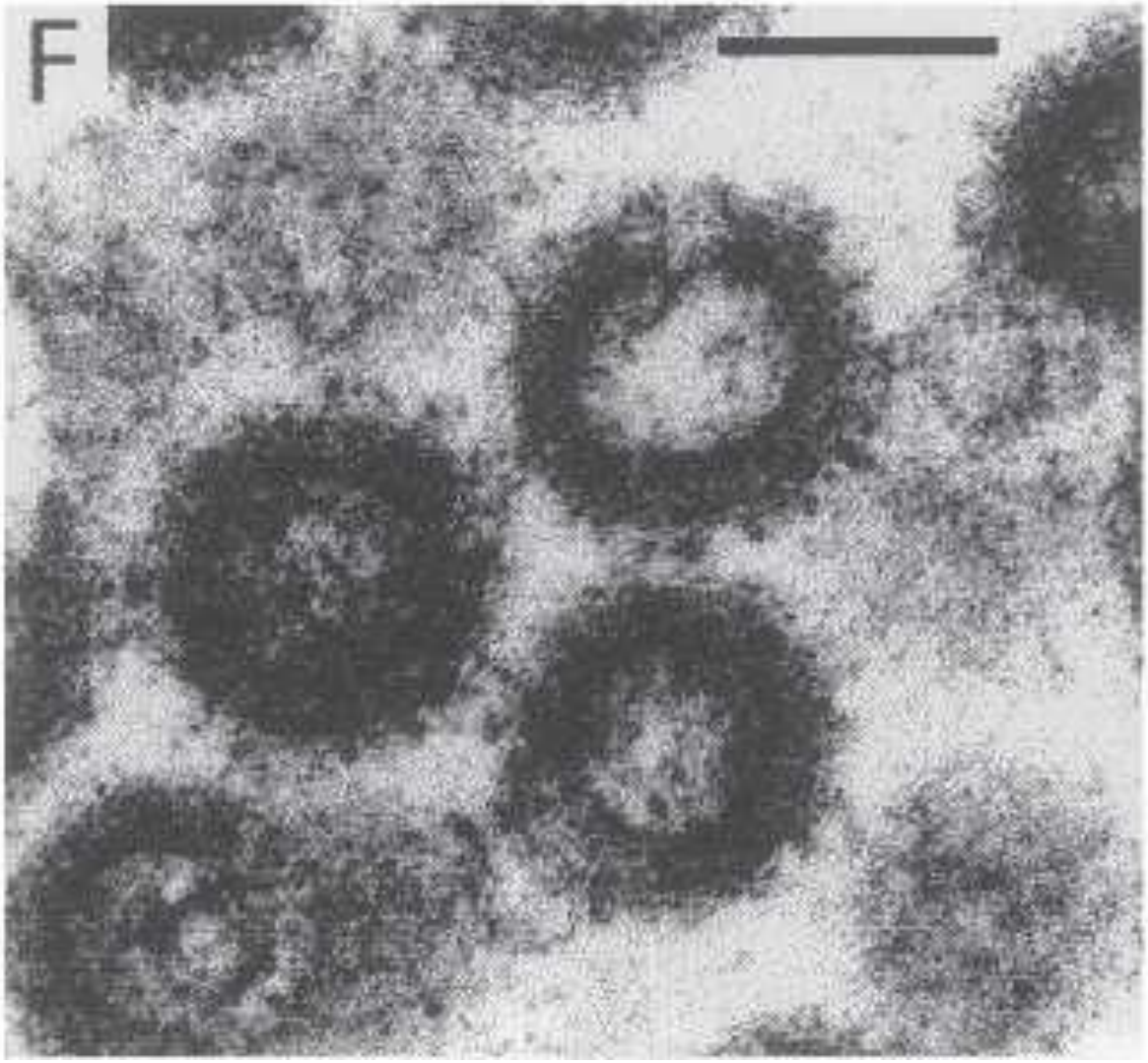


Image of representative particles obtained after transfection with an expression vector for the *Phoenix pro* mutant. Scale bar 100 nm.

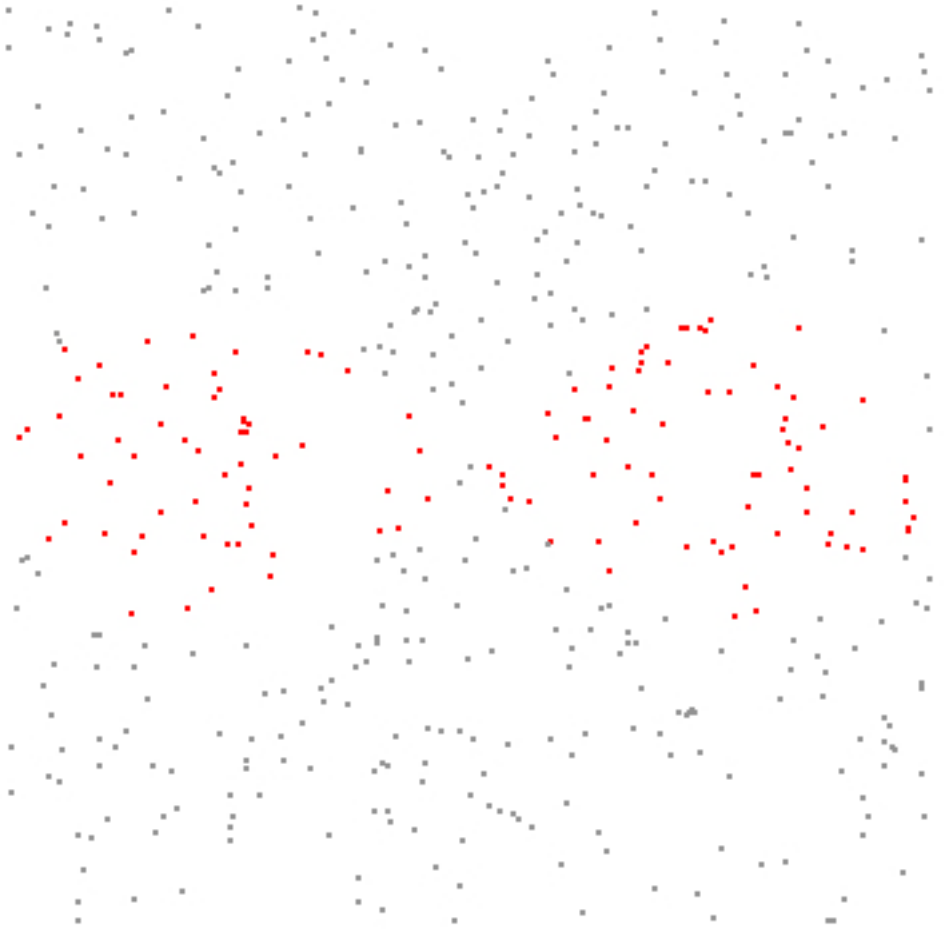
An extremely short history of early Bayesianism

- 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 ...)

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

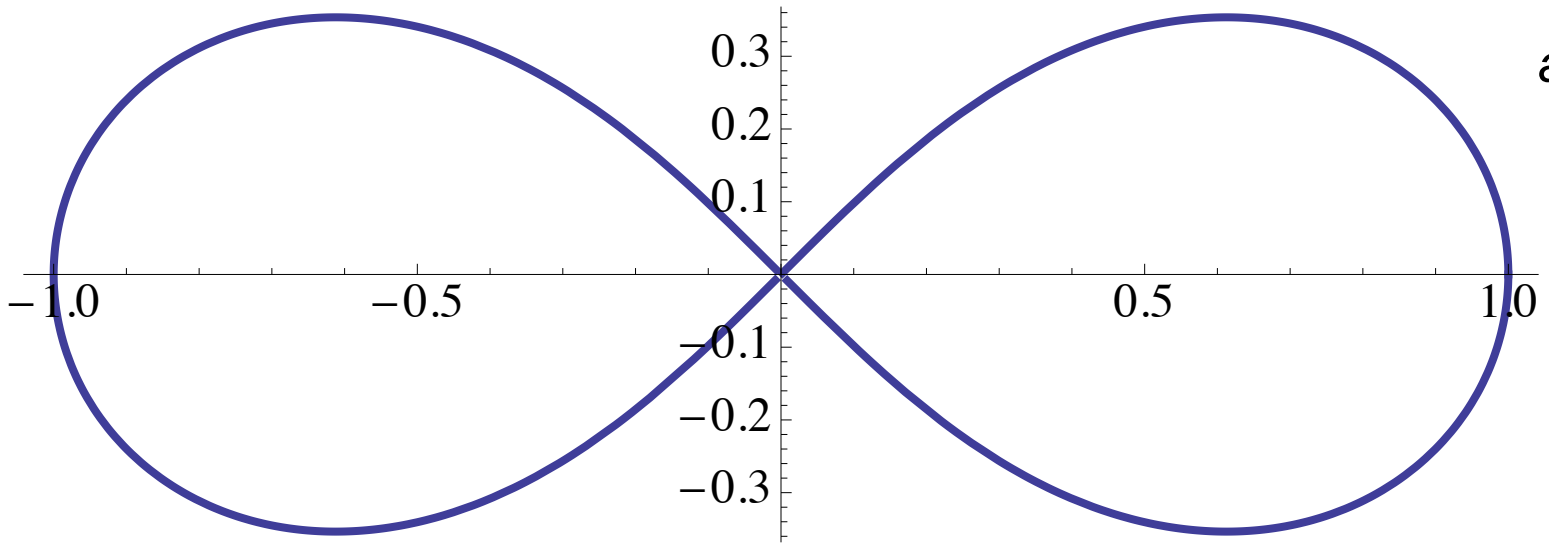
Example:

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



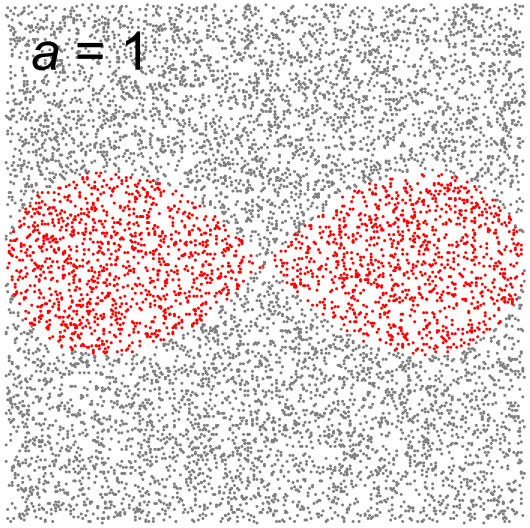
Parametric equation of Bernoulli's lemniscate

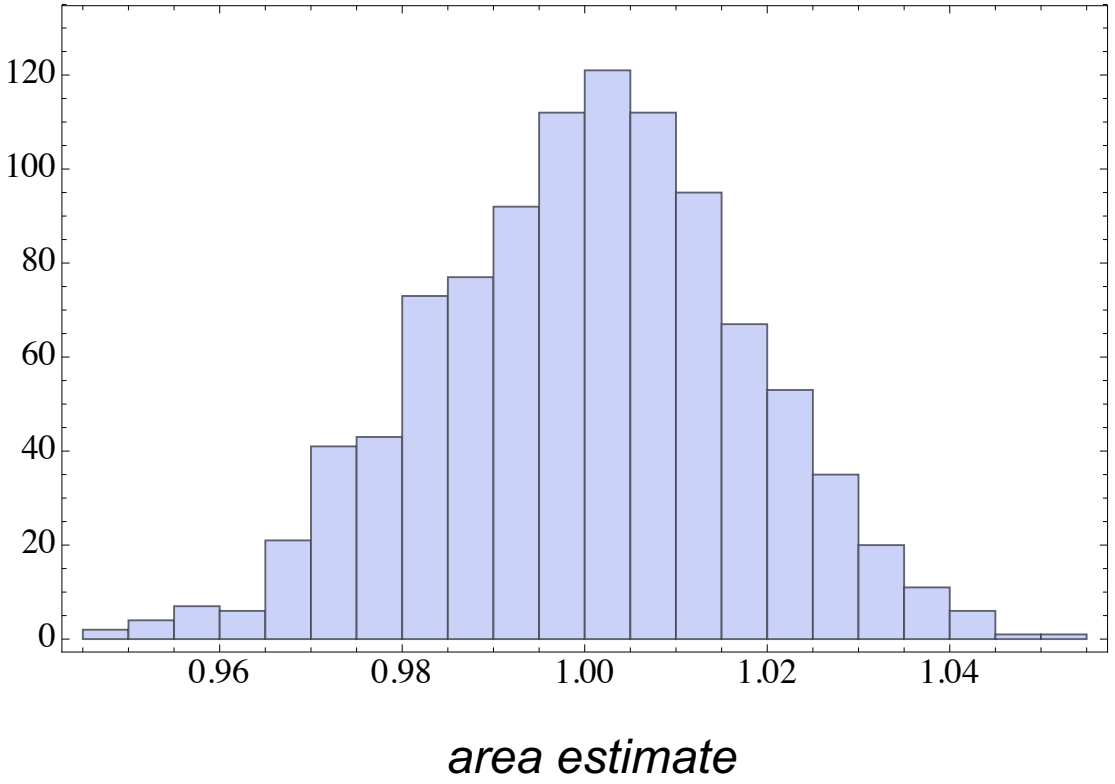
$$r = a\sqrt{\cos 2\theta}$$



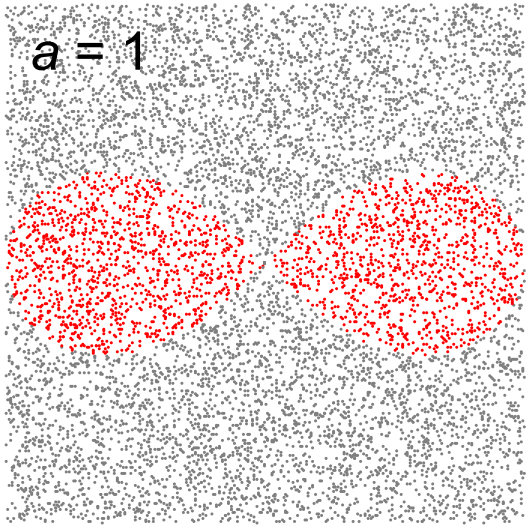
What is its area?

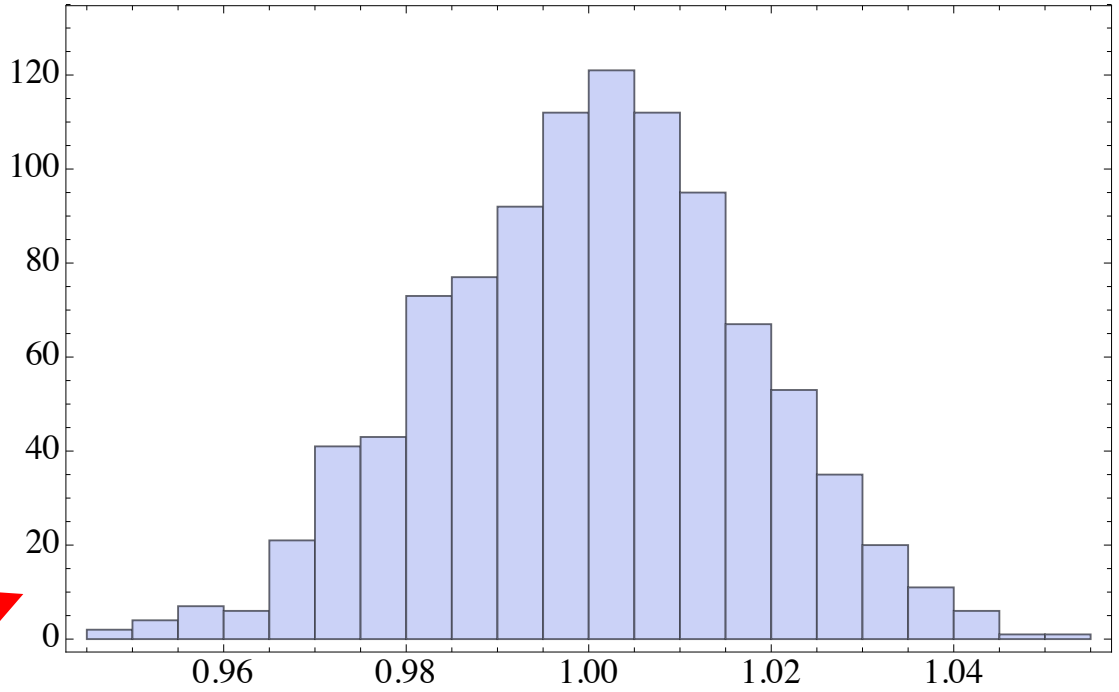
Empirical Monte Carlo distribution of the area estimate





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.

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

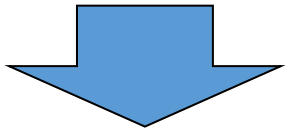
P(infect) P(not infect)

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

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

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

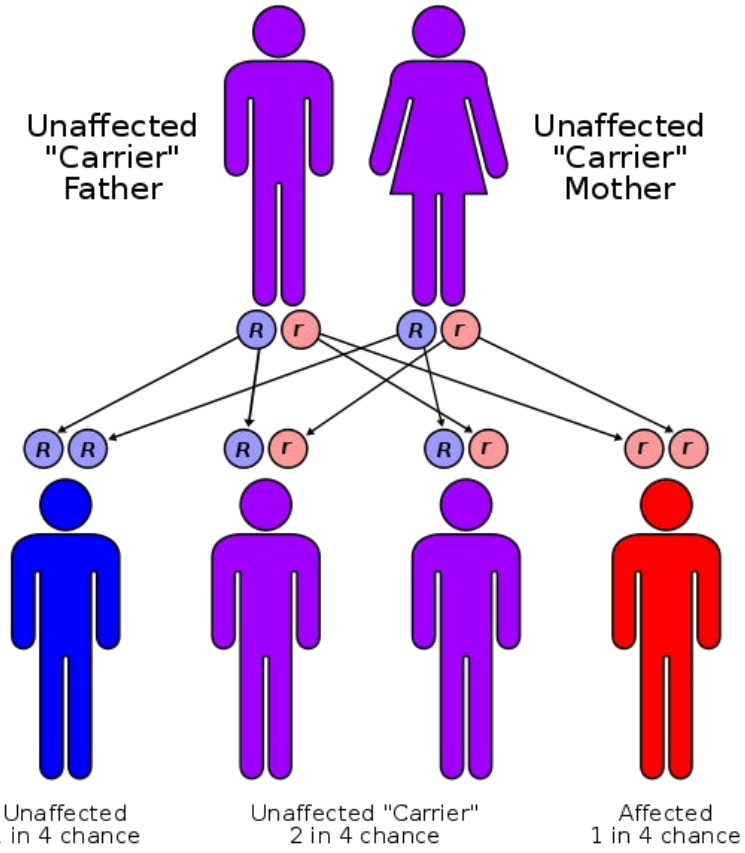
$$P(innocent \mid DNA\,compatible)$$

this is what we want!

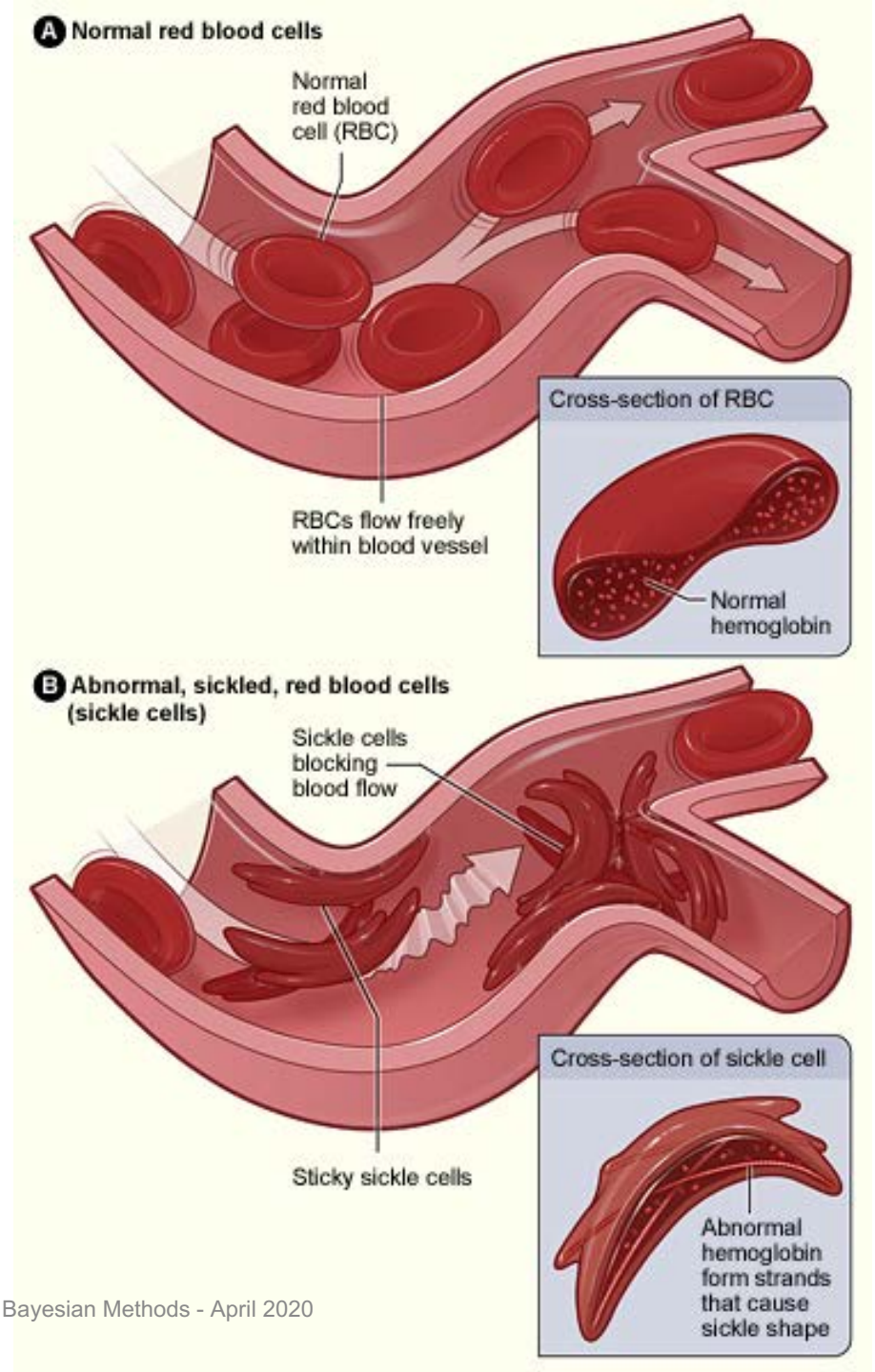
$$P(innocent \mid DNA\,compatible, I) = \frac{P(DNA\,compatible \mid innocent, I)}{P(DNA\,compatible, I)} P(innocent \mid I)$$

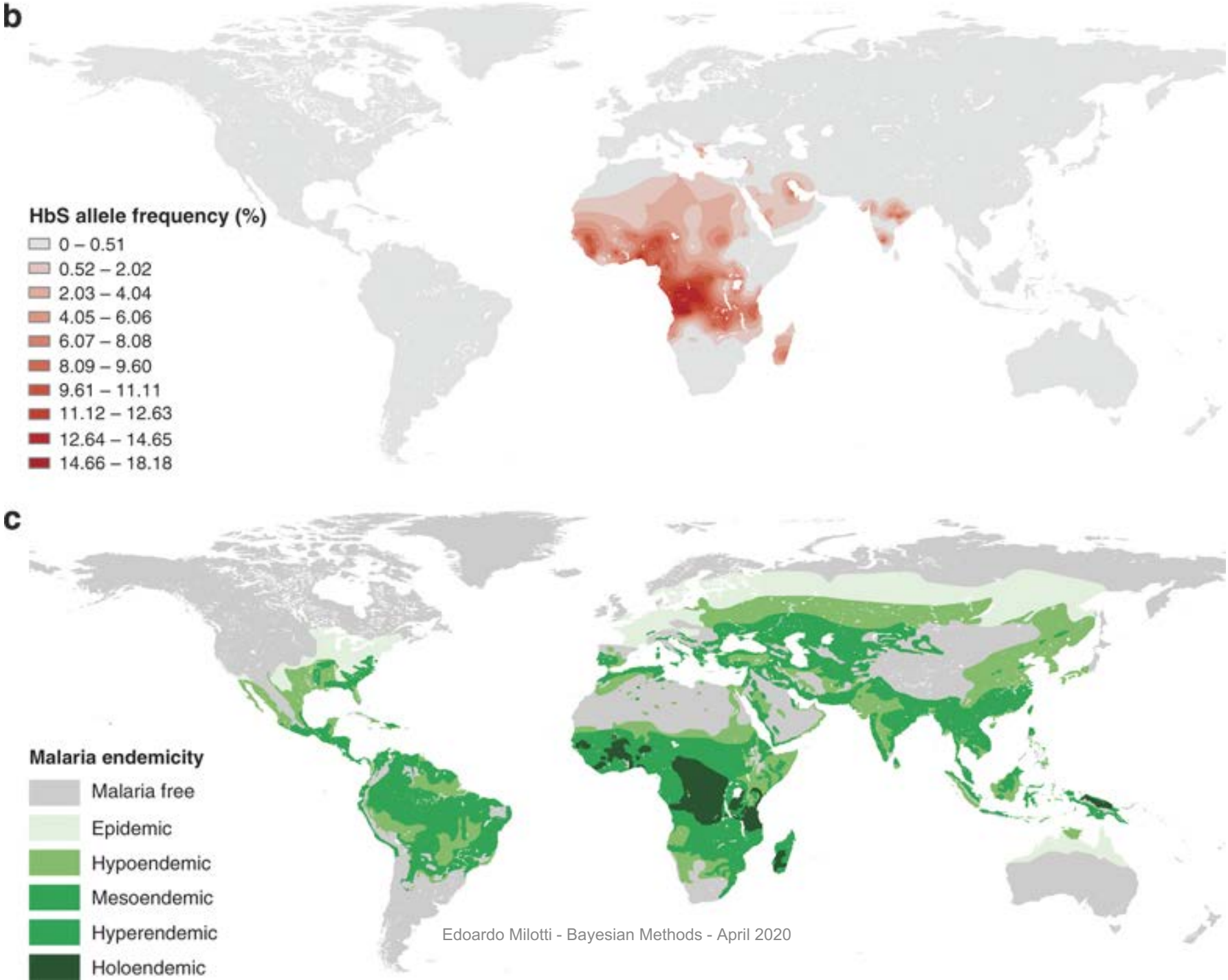
DNA classification - 1: alleles

***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.

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.

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

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

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

1. *Example of Bayesian inference*: estimate of the (probability) parameter of the binomial distribution

$$P(n \mid \theta, N) = \binom{N}{n} (1-\theta)^{N-n} \theta^{n}$$

this is the parameter that we want to infer from data

$$p(\theta \mid n, N) = \frac{P(n \mid \theta, N)}{\int_0^1 P(n \mid \theta, N) \cdot p(\theta) d\theta} \cdot p(\theta) =$$

uniform distribution: the least informative prior

$$= \frac{\binom{N}{n} (1-\theta)^{N-n} \theta^{n}}{\int_0^1 \binom{N}{n} (1-\theta)^{N-n} \theta^{n} \cdot p(\theta) d\theta} \cdot p(\theta) = \frac{(1-\theta)^{N-n} \theta^{n}}{\int_0^1 (1-\theta)^{N-n} \theta^{n} \, d\theta}$$

final result is a beta distribution

$$p(\theta \mid n,N) = \frac{(1-\theta)^{N-n}\,\theta^{n}}{\int_0^1 \theta^{n}(1-\theta)^{N-n}\,d\theta} = \frac{(1-\theta)^{N-n}\,\theta^{n}}{B(n+1,N-n+1)}$$

$$B(m,n) = \int_0^1 t^{m-1}(1-t)^{n-1}\,dt$$

beta function

$$= \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

$$p(\theta \mid n,N) = \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)}(1-\theta)^{N-n}\,\theta^{n}$$

$$= \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\,\theta^{n}$$

Mathematical digression: relationship between gamma and beta function

$$\Gamma(m)\Gamma(n) = \int_0^\infty s^{m-1}e^{-s}\,ds \int_0^\infty t^{n-1}e^{-t}\,dt$$

$$s = x^2; \qquad t = y^2; \qquad \Rightarrow$$

$$\Gamma(m)\Gamma(n) = 4\int_0^\infty x^{2m-1}e^{-x^2}\,dx \int_0^\infty y^{2n-1}e^{-y^2}\,dy$$

$$x = r\cos\theta; \qquad y = r\sin\theta; \qquad \Rightarrow$$

$$\Gamma(m)\Gamma(n) = 4\int_0^\infty r^{2m+2n-1}e^{-r^2}\,dr \int_0^{\pi/2} \cos^{2m-1}\theta \sin^{2n-1}\theta\, d\theta$$

$$= \Gamma(m+n)\left(2\int_0^{\pi/2} \cos^{2m-1}\theta \sin^{2n-1}\theta\, d\theta \right) \qquad \left(t = \cos^2\theta; \quad dt = -2\cos\theta\sin\theta d\theta\right)$$

$$= \Gamma(m+n)\int_0^1 t^{m-1}(1-t)^{n-1}\,dt$$

$$= \Gamma(m+n)B(m,n)$$

$$\Rightarrow \qquad B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \qquad \Rightarrow \qquad B(m+1,n+1) = \frac{m!n!}{(m+n+1)!}$$

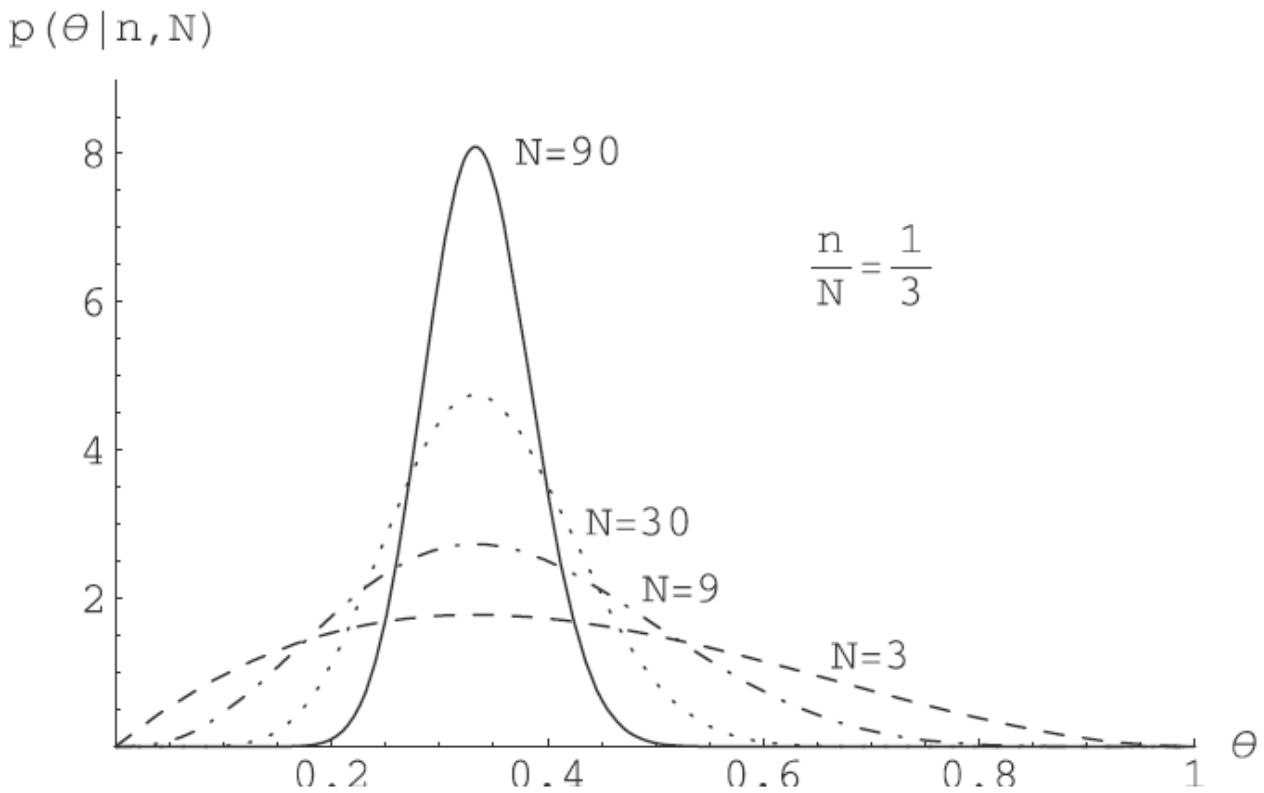


Figure 1. Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

From the knowledge of the posterior pdf we obtain all the momenta of the distribution

$$p(\theta \mid n, N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\,\theta^{n}$$

$$\langle\theta\rangle = \int_0^1 p(\theta \mid n, N)\,\theta d\theta = \frac{(N+1)!}{n!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+1}\,d\theta$$

$$= \frac{(N+1)!}{n!(N-n)!}B(n+2, N-n+1)$$

$$= \frac{(N+1)!}{n!(N-n)!}\cdot\frac{(n+1)!(N-n)!}{(N+2)!}$$

$$= \frac{n+1}{N+2} \to \frac{n}{N}$$

biased, asymptotically unbiased, estimator

$$\left\langle \theta^2 \right\rangle = \int_0^1 p\left(\theta \mid n,N\right) \theta^2 d\theta = \frac{(N+1)!}{n!(N-n)!} \int_0^1 \left(1-\theta\right)^{N-n} \theta^{n+2}\, d\theta$$

$$= \frac{(N+1)!}{n!(N-n)!} B\left(n+3, N-n+1\right)$$

$$= \frac{(N+1)!}{n!(N-n)!} \cdot \frac{(n+2)!(N-n)!}{(N+3)!}$$

$$= \frac{\left(n+2\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)}$$

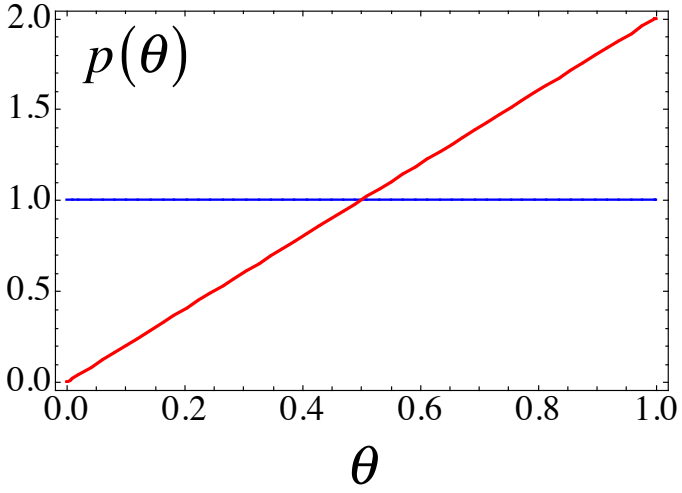
$$\operatorname{var}\theta = \left\langle \theta^2 \right\rangle - \left\langle \theta \right\rangle^2 = \frac{\left(n+2\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)} - \left(\frac{n+1}{N+2} \right)^2 =$$

$$= \frac{\left(N-n+1\right)\left(n+1\right)}{\left(N+3\right)\left(N+2\right)^3}$$

What happens if we try a different prior?

Let's try with a linear prior

$$p(\theta) = 2\theta$$



$$p(\theta \mid n,N) = \frac{P(n \mid \theta,N)}{\int_0^1 P(n \mid \theta,N)\cdot p(\theta)\,d\theta}\cdot p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)^{N-n}\theta^n \cdot 2\theta\,d\theta}\cdot 2\theta = \frac{(1-\theta)^{N-n}\theta^{n+1}}{\int_0^1 (1-\theta)^{N-n}\theta^{n+1}\,d\theta}$$

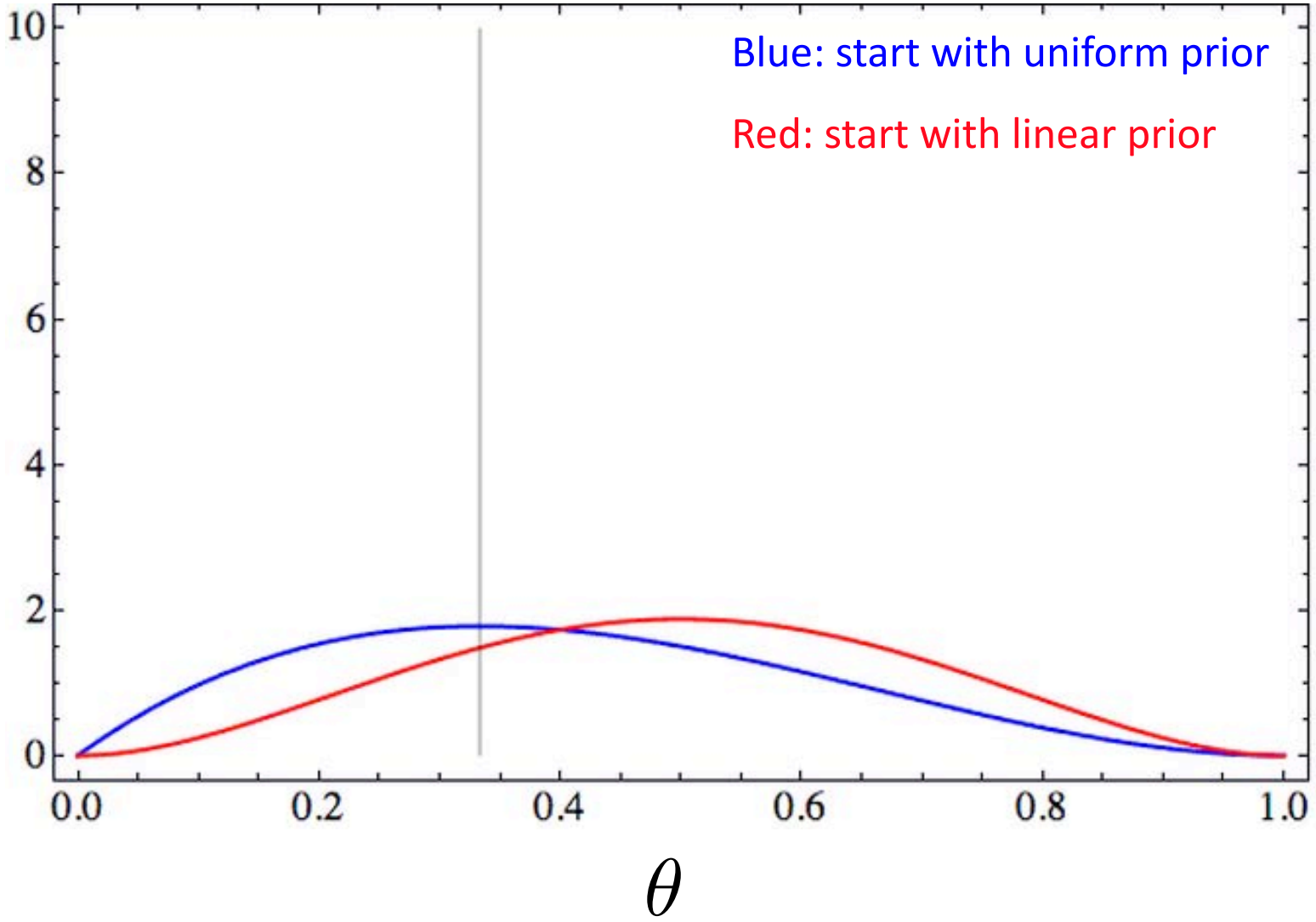
$$p(\theta \mid n,N) = \frac{(N+2)!}{(n+1)!(N-n)!}(1-\theta)^{N-n}\,\theta^{n+1}$$

$$\langle\theta\rangle = \int_0^1 p(\theta \mid n,N)\,\theta d\theta = \frac{(N+2)!}{(n+1)!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+2}\,d\theta$$

$$= \frac{(N+2)!}{(n+1)!(N-n)!}B(n+3,N-n+1)$$

$$= \frac{(N+2)!}{(n+1)!(N-n)!}\cdot\frac{(n+2)!(N-n)!}{(N+3)!}$$

$$= \frac{n+2}{N+3} \to \frac{n}{N}$$

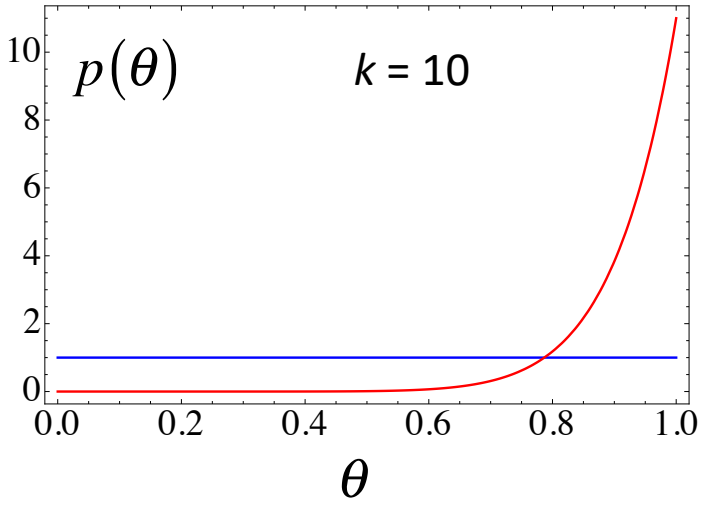


Taking few coin throws, the posterior from the linear prior is considerably biased. The bias disappears when the number of coin throws is large.

Now we try with a very non-uniform prior

We take

$$p(\theta) = (k+1)\theta^k; \qquad k \gg 1$$



$$p(\theta \mid n,N) = \frac{p(n \mid \theta,N)}{\int_0^1 P(n \mid \theta,N)\cdot p(\theta)\,d\theta} \cdot p(\theta)$$

$$= \frac{\binom{N}{n}(1-\theta)^{N-n}\theta^n}{\int_0^1 \binom{N}{n}(1-\theta)^{N-n}\theta^n \cdot (k+1)\theta^k\,d\theta} \cdot (k+1)\theta^k = \frac{(1-\theta)^{N-n}\theta^{n+k}}{\int_0^1 (1-\theta)^{N-n}\theta^{n+k}\,d\theta}$$

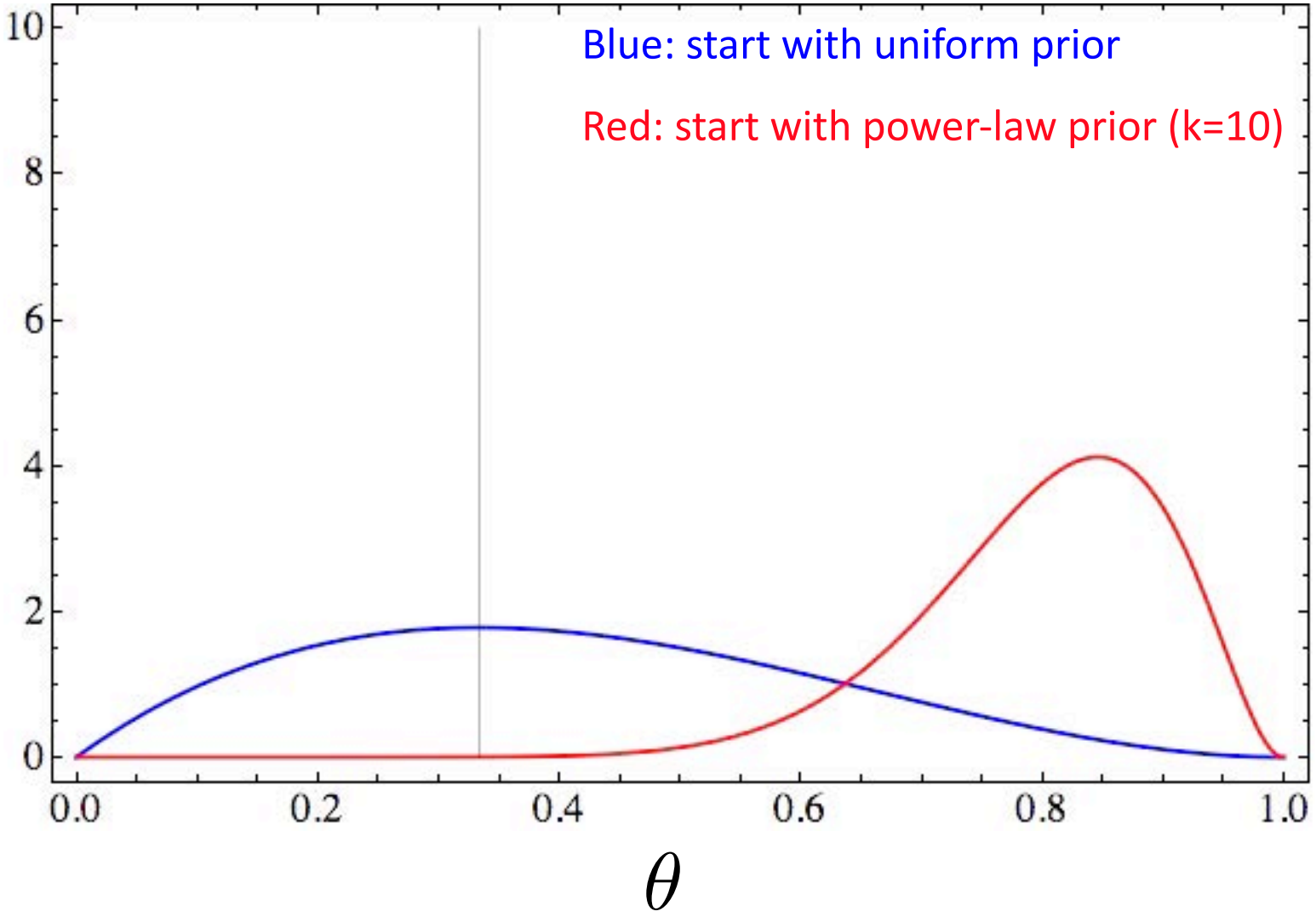
$$p(\theta \mid n,N) = \frac{(N+k+1)!}{(n+k)!(N-n)!}(1-\theta)^{N-n}\,\theta^{n+k}$$

$$\langle\theta\rangle = \int_0^1 p(\theta \mid n,N)\,\theta d\theta = \frac{(N+k+1)!}{(n+k)!(N-n)!}\int_0^1 (1-\theta)^{N-n}\,\theta^{n+k+1}\,d\theta$$

$$= \frac{(N+k+1)!}{(n+k)!(N-n)!}B(n+k+2,N-n+1)$$

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

$$= \frac{n+k+1}{N+k+2} \to \frac{n}{N}$$



In this case, initial bias due to the prior is very large.

Note on posterior distributions:

the relationship between binomial distribution and beta function is quite important and common, and leads to the formal definition of the Beta distribution:

$$\mathrm{B}\left(\theta \middle| a,b\right) = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}\theta^{a-1}\left(1-\theta\right)^{b-1}$$

There are other important dualities between distributions. This topic is discussed in depth in

J. M. Bernardo: “Reference Posterior Distributions for Bayesian Inference”, J. R. Statist. Soc. B **41** (1979), 113

Lessons learned:

1. The prior information is not neutral, a careful choice of the prior distribution is a necessity.

   Question: how do we choose a prior?

2. If we want to keep all possibilities alive, we must heed the Cromwell's rule: "Prior probabilities 0 and 1 should be avoided" (Lindley, 1991)

   The reference is to Oliver Cromwell's phrase:
   I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

3. Convergence as the dataset size grows seems to be granted, however it may be very slow with a bad choice of prior distribution

   Question: is convergence really granted???

The Bernstein-Von Mises theorem

- Convergence can only be defined with respect to a frequentist approach.
- The theorem that grants convergence under very weak hypotheses is the Bernstein-Von Mises theorem.
- It is interesting to note that even here we can find inconsistencies.

Maximum a posteriori (MAP) estimate – MAP is not mean value!

Consider the case with a uniform prior: from the posterior distribution

$$p(\theta \mid n, N) = \frac{(N+1)!}{n!(N-n)!}(1-\theta)^{N-n}\,\theta^{n}$$

we easily find that the posterior pdf is maximized by the parameter value

$$\theta = n/N$$

which is the unbiased estimate of the parameter (unlike the mean value!)

Credible intervals (case of initial uniform prior), Bayesian analog of confidence intervals.