

Introduction to Bayesian Statistics - 1

PhD Physics course (XXVIII ciclo)

Università di Trieste

Edoardo Milotti

Webpage:

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

The operational definition of probability of an event A

$$P(A) = \frac{N(A)}{N}$$

The relative frequency

$$f_n(A) = \frac{n(A)}{n}$$

The law of large numbers

$$\lim_{n \rightarrow \infty} f_n(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n} = P(A)$$

The algebra of probabilities

Let A and B be statements that can be either true or false, and such that we can assign probabilities. Then the following rules apply:

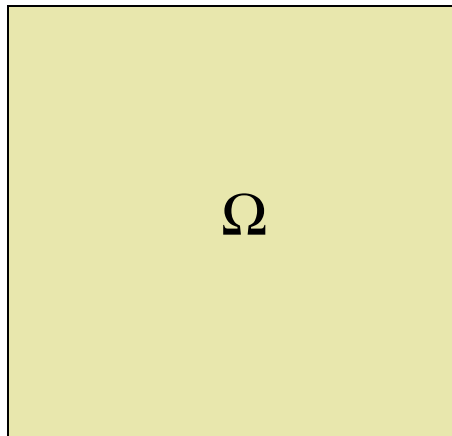
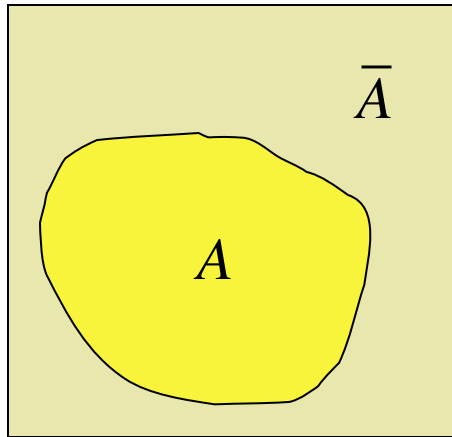
$$0 \leq P(A) \leq 1$$

$$P(\Omega) = 1$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Probability space and measure theory

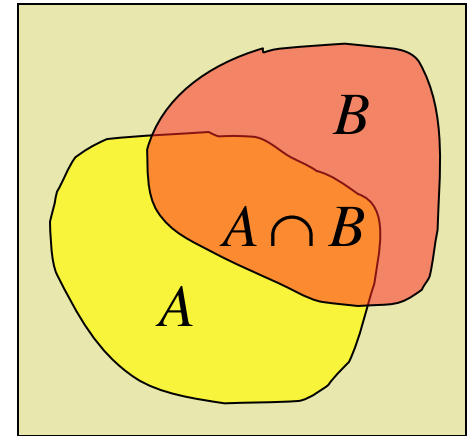


$$0 \leq P(A) \leq 1$$

$$P(\Omega) = 1; \quad P(A) + P(\bar{A}) = 1$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$



Bayes' Theorem

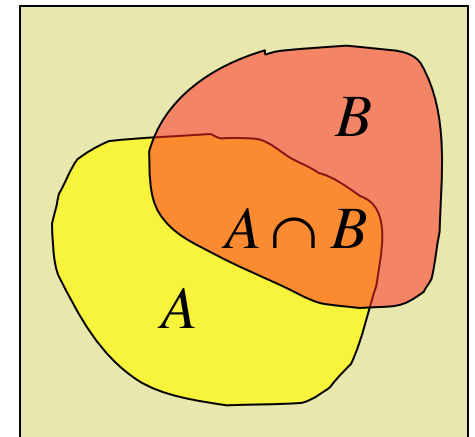
$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Independent events:

$$P(A \text{ e } B) = P(A) \cdot P(B)$$

Dependent events:

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



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

Bayes' theorem

rev. Thomas Bayes (1702-1761)



Thomas Bayes was the son of a London Presbyterian minister, Joshua Bayes born perhaps in Hertfordshire. In 1719 he enrolled at the University of Edinburgh to study logic and theology.

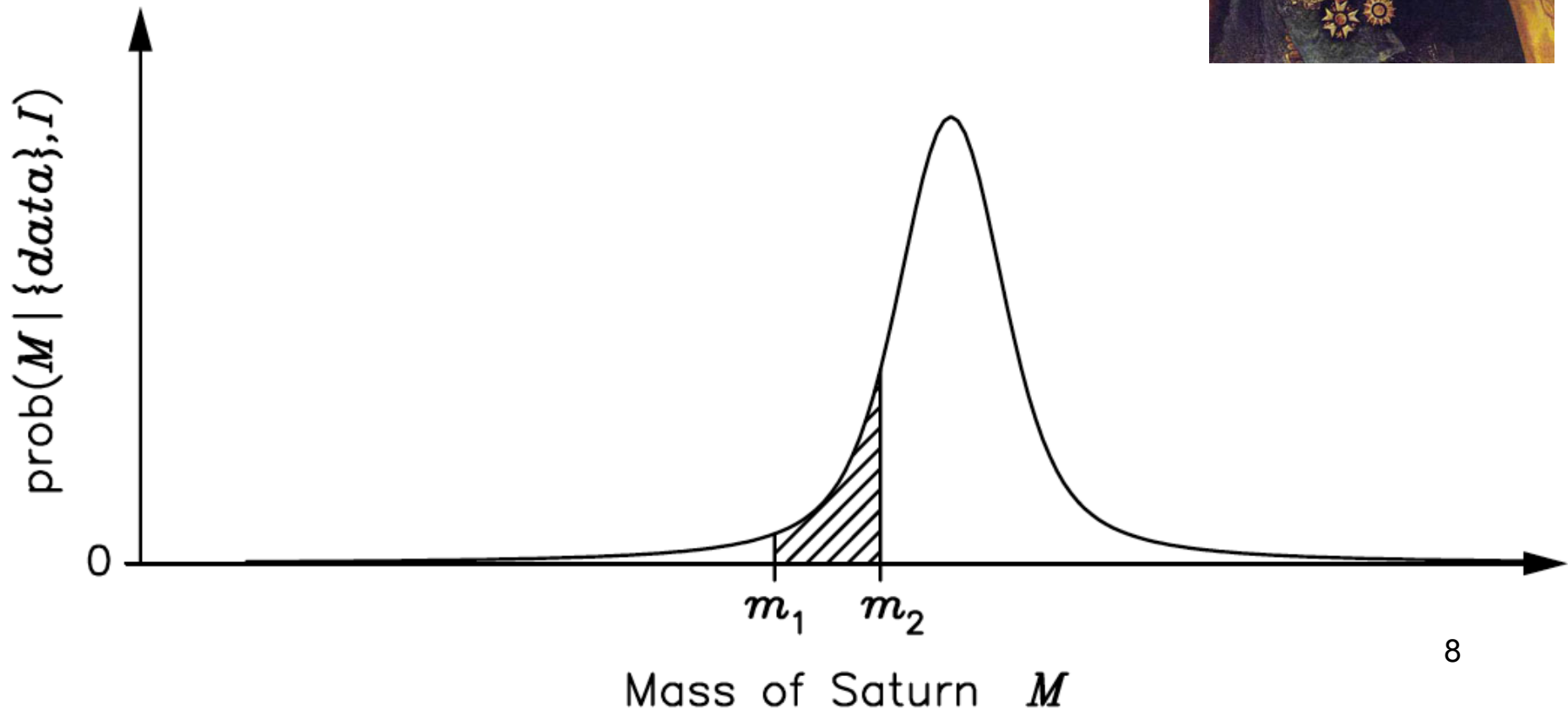
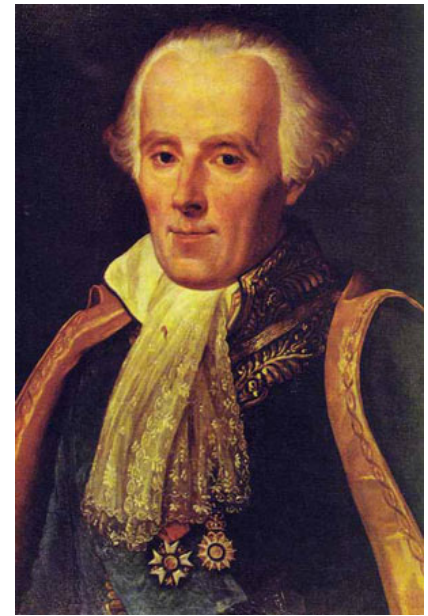
He is known to have published two works in his lifetime: *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* (1731), and *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst* (published anonymously in 1736), in which he defended the logical foundation of Isaac Newton's calculus against the criticism of George Berkeley, author of *The Analyst*.

It is speculated that Bayes was elected as a Fellow of the Royal Society in 1742 on the strength of *the Introduction to the Doctrine of Fluxions*, as he is not known to have published any other mathematical works during his lifetime. Some feel that he became interested in probability while reviewing a work written in 1755 by Thomas Simpson, but others think he learned mathematics and probability from a book by de Moivre.

Bayes died in Tunbridge Wells, Kent. He is buried in Bunhill Fields Cemetery in London where many Nonconformists are buried.

The ideas of Bayes were clarified, extended and put to good use by Pierre Simon, Marquis de Laplace

“In order to give some interesting applications of it I have profited by the immense work which M. Bouvard has just finished on the movements of Jupiter and Saturn ... His calculations give him the mass of Saturn equal to 3512th part of that of the sun. Applying to them my formulae of probability, I find that it is a bet of 11,000 against one that the error of this result is not 1/100th of its value ...”



Bayesians stress the subjective aspects. Examples can be found in (D'Agostini 2003). D'Agostini cites Schrödinger, who took a stand very much like to De Finetti:

Definition of probability:

. . . a quantitative measure of the strength of our conjecture or anticipation, founded on the said knowledge, that the event comes true.

Subjective nature of probability:

Since the knowledge may be different with different persons or with the same person at different times, they may anticipate the same event with more or less confidence, and thus different numerical probabilities may be attached to the same event.

Conditional probability:

Thus whenever we speak loosely of 'the probability of an event', it is always to be understood: probability with regard to a certain given state of knowledge.

logical foundations of probability theory

AMERICAN JOURNAL *of* PHYSICS

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1

JANUARY-FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

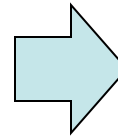
R. T. COX

The Johns Hopkins University, Baltimore 18, Maryland

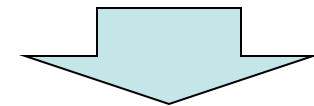
Cox's paper has three main parts:

- general considerations on probability
- axiomatic derivation of the rules of probability from standard logic rules

$$p(c, b|a) = F(p(c|a, b), p(b|a))$$
$$p(\tilde{b}|a) = S(p(b|a))$$

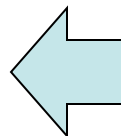


functional
equations for F, S



•

standard rules for
probability of combined
events



$$F[F(x, y), z] = F[x, F(y, z)]$$
$$xS[S(y)/x] = yS[S(x)/y]$$

- frequentist probabilities as a reasonable special case

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

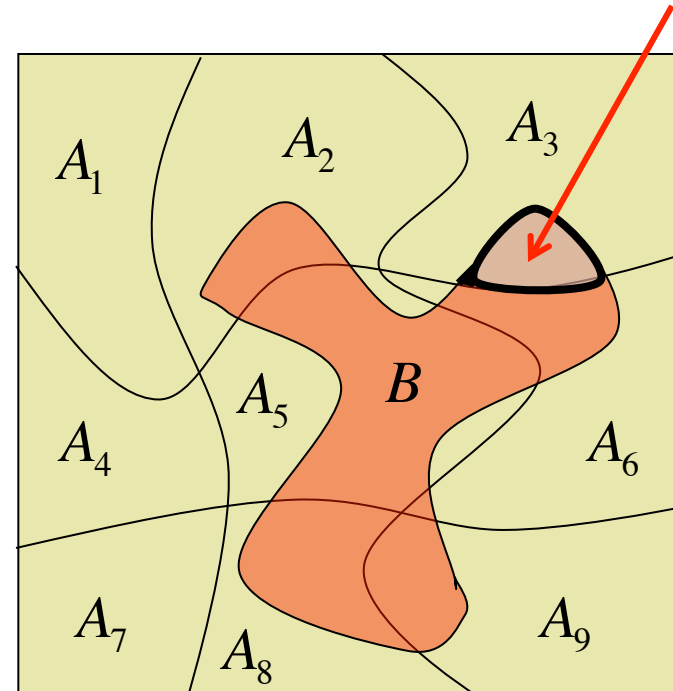
$$P(A_k | B) = \frac{P(B | A_k) \cdot P(A_k)}{P(B)}$$

$$k = 1, \dots, N$$

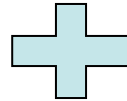
$$P(B | A_3) \cdot P(A_3)$$

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

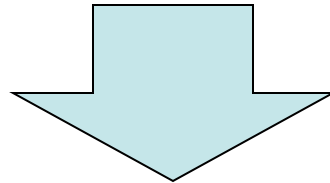
$$P(B) = \sum_{k=1}^N P(B | A_k) \cdot P(A_k)$$



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

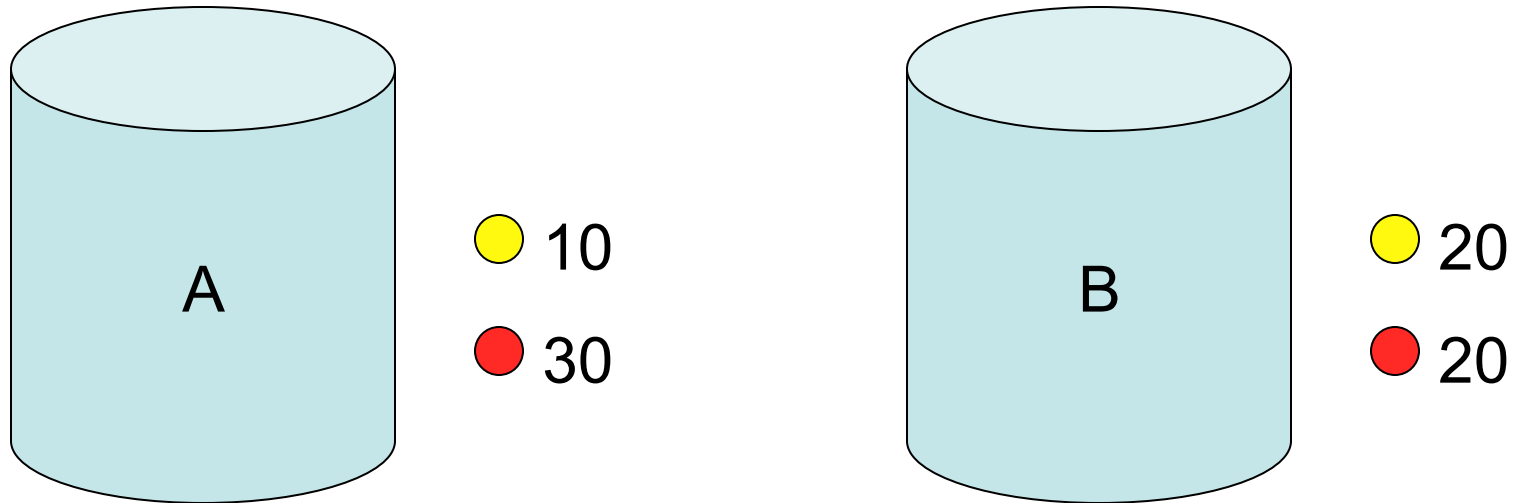


$$P(B) = \sum_{k=1}^N P(B|A_k) \cdot P(A_k)$$



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

A simple example



Here we choose a ball as follows:

1. We choose the urn first
2. We draw a ball from that urn

What is the probability of drawing one red ball?

$P(A) = P(B) = 1/2$ (probability of choosing either A or B)

$P(G|A) = 1/4$ (probability of drawing a yellow ball from A)

$P(R|A) = 3/4$ (probability of drawing a red ball from A)

$P(G|B) = 1/2$ (probability of drawing a yellow ball from B)

$P(R|B) = 1/2$ (probability of drawing a red ball from A)

and therefore

$$\begin{aligned} P(R) &= P(R|A) \cdot P(A) + P(R|B) \cdot P(B) \\ &= (3/4) \cdot (1/2) + (1/2) \cdot (1/2) = 5/8 = 0.625 \end{aligned}$$

Inverse problem: if we drew a red ball, what is the probability that we drew it from urn A?

(NB: here we assume that the “physical model” is known, i.e., we assume we know how many red and yellow balls are in each urn)

“a priori” probability: $P(A) = 1/2$

Now we apply Bayes’ theorem

$$P(A | R) = \frac{P(R | A) \cdot P(A)}{P(R)} = \frac{(3/4) \cdot (1/2)}{(5/8)} = \frac{3}{5} = 0.6$$

“a posteriori” probability

This is a simple example of Bayesian inference

We draw another red ball, still from the same urn, (however we do not know whether this is A or B). Since now

$$P(R) = P(R|A) \cdot P(A) + P(R|B) \cdot P(B) = 0.65$$

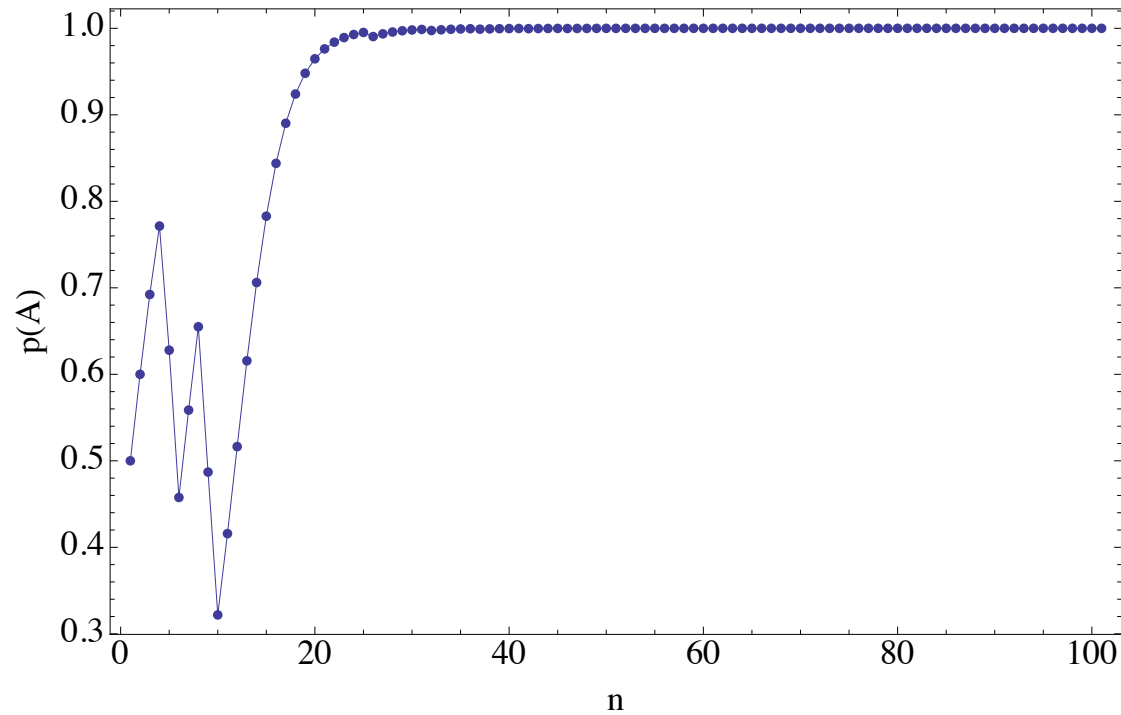
we find

$$P(A|\{R,R\},I) = \frac{P(R|A,I) \cdot P(A|R,I)}{P(R,I)} \approx 0.692308$$

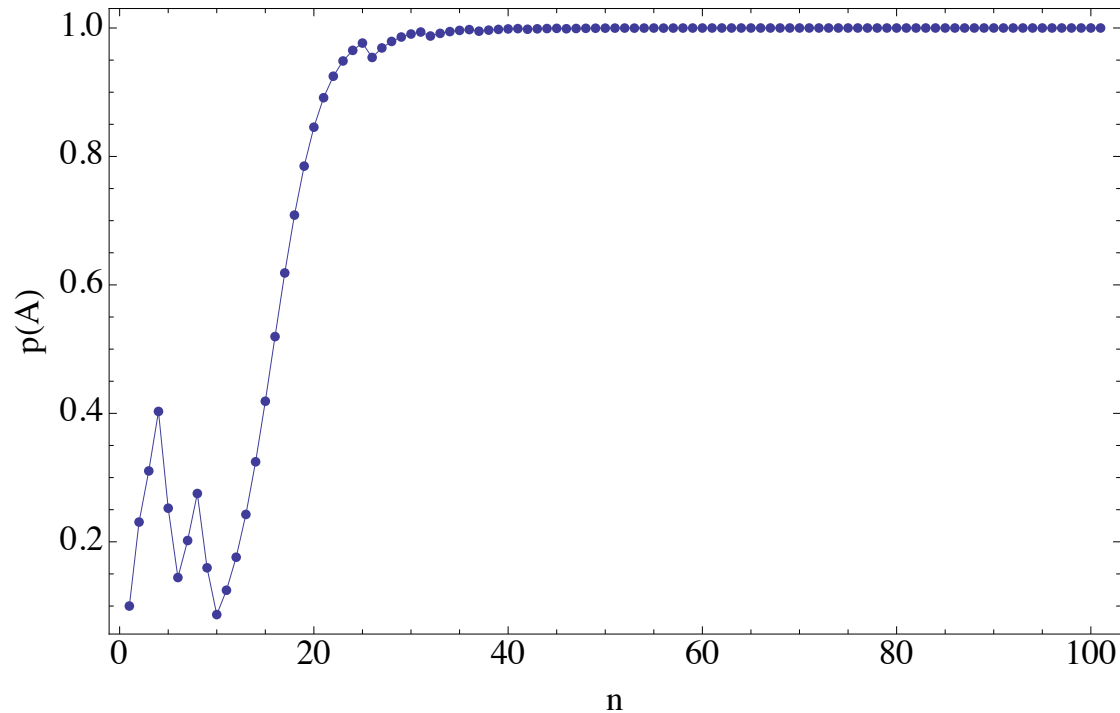
Notice that data can be inserted one by one!

100 successive draws ...

R, R, R, Y, Y, R, R, Y, Y, R, R, R, R, R, R, R, R, R, R, R, R, R, R, Y,
R, R, R, R, R, Y, R, R, R, R, Y, R, R, R, R, Y, R, R, R, Y, R, R, R, R, R,
R, Y, R, R, Y, R, R, R, R, R, R, Y, R, R, R, R, Y, R, R, Y, R, Y, R, R, Y,
Y, R, R, Y, R, R, R, Y, R, R, Y, R, R, R, R, R, R, R, R, R, R, Y, Y, R, R



... a different starting point: here the initial prior probability is 0.05 instead of 0.5.



Frequentist and Bayesian statistics

Where's the difference?

Why are there sharply different opinions?

The opinion of a Bayesian physicists (M. Goldstein tries to express contrasting views, in *Advanced Statistical Techniques in Particle Physics*, Grey College, Durham, 18 - 22 March 2002)

PRO's (Bayesian view)

BAYES IS CORRECT

[C1] Other approaches are wrong, as argued through the well-rehearsed counter-examples about the failure of meaning of the core concepts of more traditional inference, such as significance and coverage properties. Thus, a valid confidence interval may be empty, a statistically significant result obtained with high power may be almost certainly false, and so forth.

[C2] The Bayes approach is right, as argued on the grounds that the method evaluates the relevant kinds of uncertainty judgements, namely the uncertainties over the quantities that we want to learn about, given the quantities that we observe, based on careful foundational arguments using ideas such as coherence and exchangeability to show why this is the unavoidable way to analyse our uncertainties *

BAYES IS USEFUL

[U1] The methodology gives good solutions for standard problems, as argued through individual cases. The solutions appear paradox-free, and correspond well with intuition.

[U2] The methodology offers the only way to tackle many non-standard problems, as there is a unified approach for all problems in uncertainty. It offers a method which can always be followed, unlike most other approaches which rely on ad hoc tricks for each individual case.

CON's (Frequentist view)

BAYES IS INAPPROPRIATE

[I1] Bayesian methodology answers problems wrongly. Usually, this is attributed to unnecessary and unhelpful appeal to arbitrary prior assumptions, which should not belong in scientific analyses.

[I2] Bayesian methodology answers the wrong problems. This argument replaces the blanket criticism of the Bayes approach by recognition that the Bayes solution may indeed tell us something meaningful about what an individual might conclude from the data, but still argues that such individual subjective reasoning is inappropriate as a way of reaching sound and objective scientific conclusions, which are related to consensus within the scientific community.

BAYES IS HARD

[H1] Every problem is hard for Bayesian analysis. This is a reflection of the difficulty, even in the simplest problem, of finding an objectively justifiable prior distribution for the quantities of interest. In general how do we find prior distributions and what should we do if experts disagree?

[H2] Hard problems are hard for Bayesian analysis. Even if we could solve the prior specification issue for simple problems, the difficulty involved in constructing a full Bayes specification for more complicated problems renders the approach infeasible.

The above arguments have been simplified down to their essential form to suggest that there are (at least!) two levels at which we may debate the correct use of statistical methodology:

(i) the **current practice** debate: [C1],[U1], versus [I1],[H1]

(ii) the **underlying issues** debate: [C2],[U2], versus [I2],[H2]

Of course, the two debates are intimately linked, and starting in one debate we may easily find ourselves dipping into the other. However, unless we are clear as to which debate we are in, it is easy to become confused, especially as the structure of the two debates appears so similar.

* it has been argued that quantum probabilities are to be interpreted in a Bayesian way, *thereby leading to a different meaning of the interpretation of outer reality*

(see Caves et al., PRA 65 (2002) 022305)

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

$$P(\text{positive} \mid \text{infect}) = 1$$

$$P(\text{positive} \mid \text{not infect}) = 1.5\%$$

what is the probability $P(\text{infect} \mid \text{positive})$?

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

Let's use Bayes' theorem ...

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

$$\begin{aligned} P(\text{infect} \mid \text{positive}) &= \frac{P(\text{positive} \mid \text{infect}) \cdot P(\text{infect})}{P(\text{positive} \mid \text{infect}) \cdot P(\text{infect}) + P(\text{positive} \mid \text{not infect}) \cdot P(\text{non infect})} \\ &= \frac{P(\text{positive} \mid \text{infect})}{P(\text{positive} \mid \text{infect}) \cdot P(\text{infect}) + P(\text{positive} \mid \text{not infect}) \cdot P(\text{non infect})} \cdot P(\text{infect}) \end{aligned}$$

The estimate depends on the size of the infect population
i.e., on the probabilities

P(infect) P(not infect)

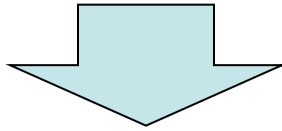
$P(\text{infect} \mid \text{positive})$

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

The posterior estimate strongly depends on the prior probability

Example: AIDS frequency in Italy 0.4 %

AIDS frequency in South Africa 18.1%



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

Italy

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

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:

$$P(\textit{infect} | \{\textit{positive}, \textit{positive}\}) = 94.7\% \quad \text{Italy}$$

$$P(\textit{infect} | \{\textit{positive}, \textit{positive}\}) = 99.9\% \quad \text{South Africa}$$

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

$P(DNA\ compatible \mid innocent)$

$P(innocent \mid DNA\ compatible)$

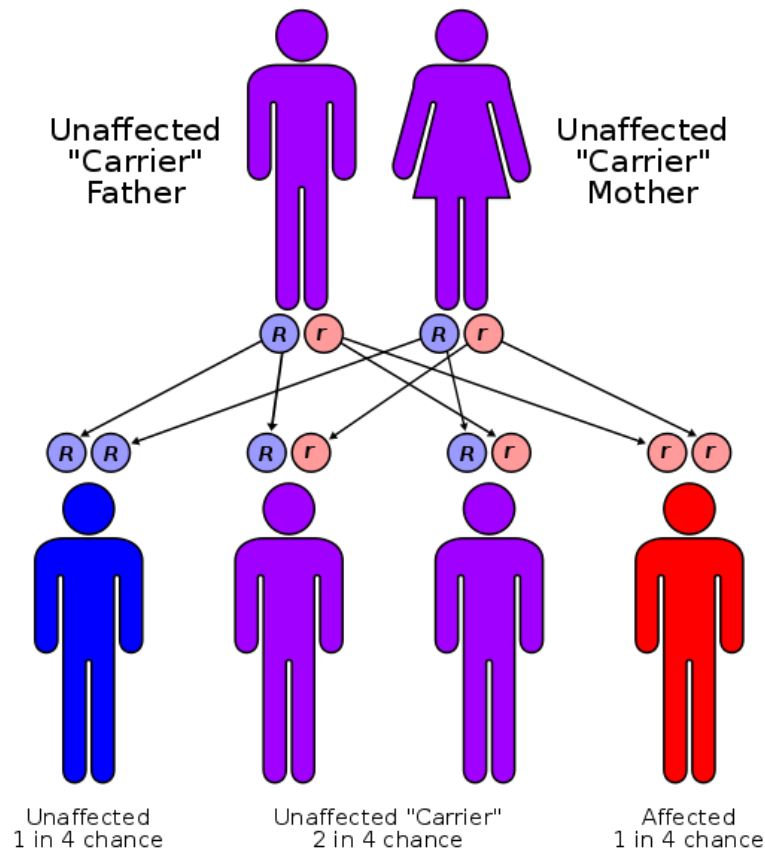
this is
what we
want!



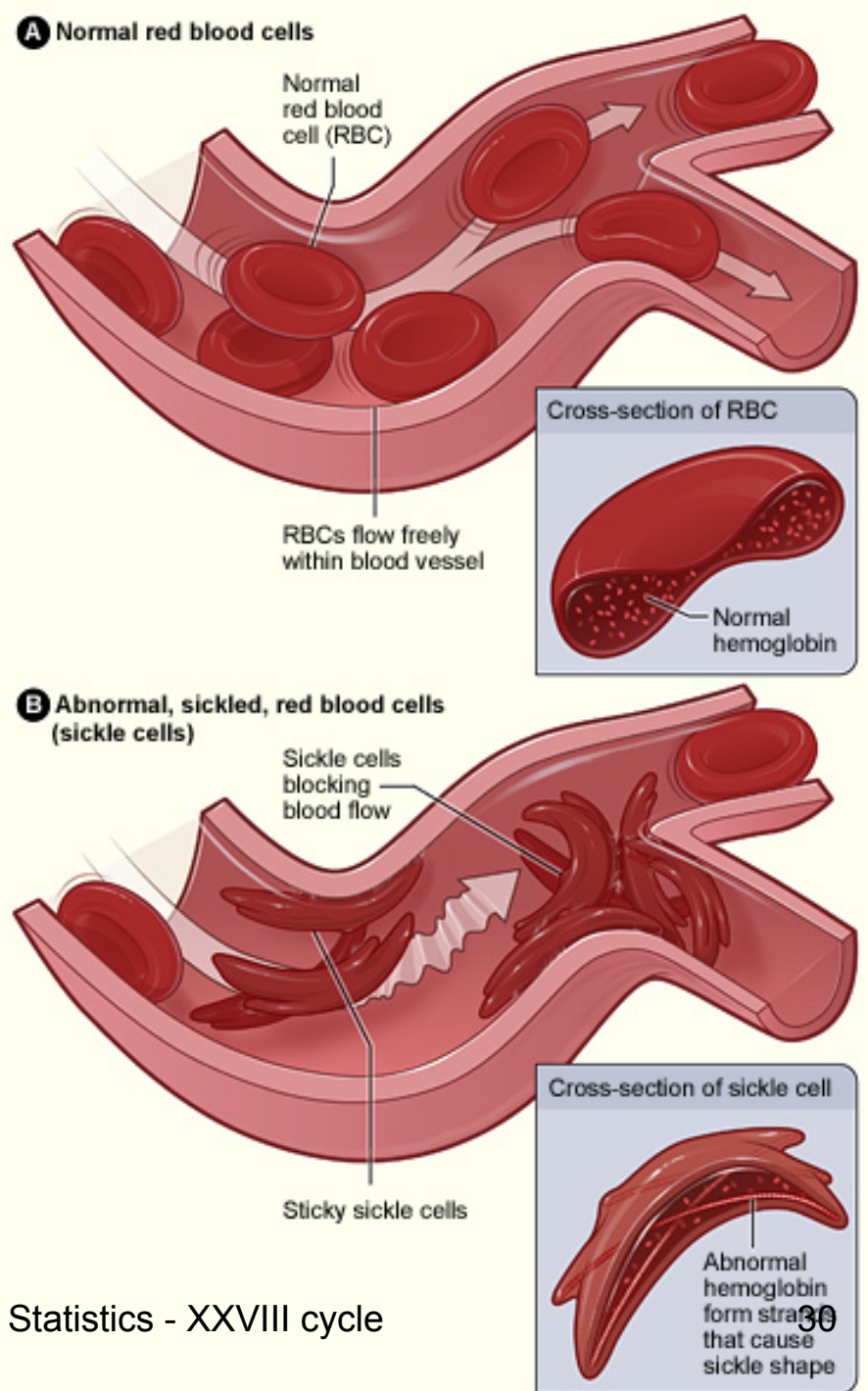
$$P(innocent \mid DNA\ compatible, I) = \frac{P(DNA\ compatible \mid innocent, I)}{P(DNA\ compatible, I)} P(innocent \mid I)$$

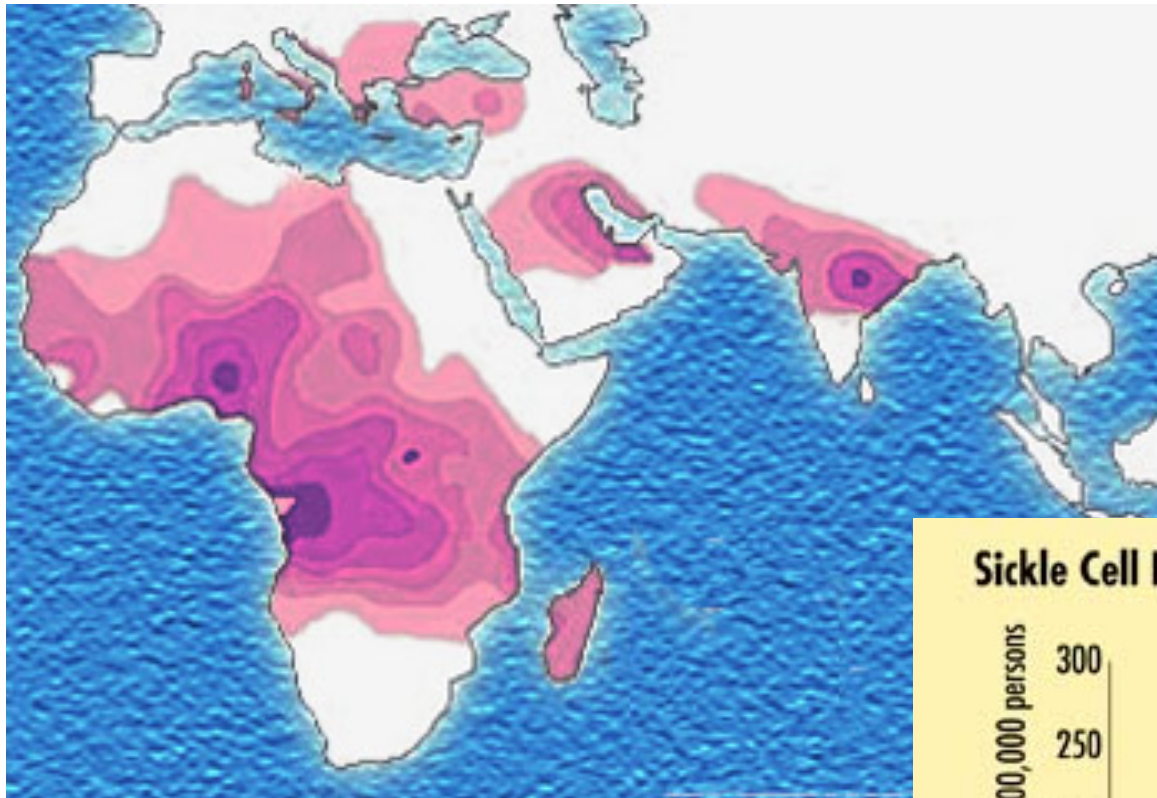
Classificazione del DNA-1: alleli

allele: una di due o più forme alternative di un gene affetto da mutazioni, e che si trovano nello stesso posto in un cromosoma



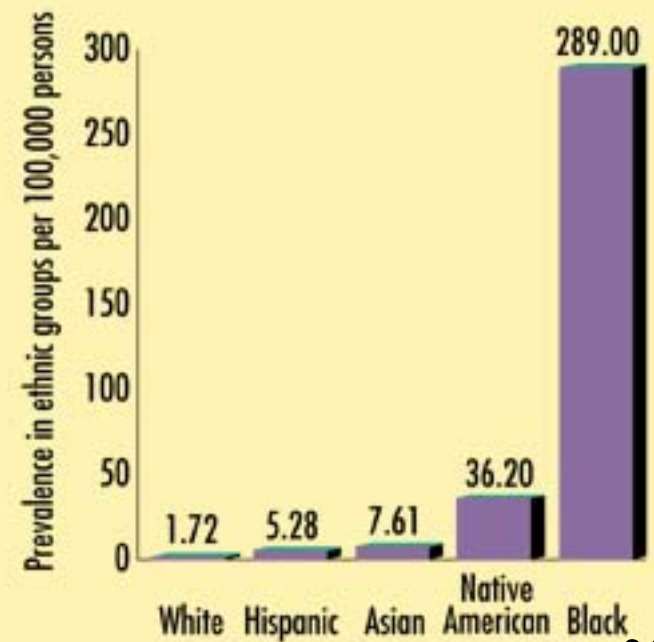
esempio: malattie
genetiche come
l'anemia falciforme





sickle-cell disease frequency

Sickle Cell Disease in the United States



Classificazione del DNA-2: frequenza degli alleli

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

tratto da <http://www.dna-view.com/profile.htm>

Database di alleli umani (ALeLe FREquency Database:
<http://alfred.med.yale.edu/alfred/index.asp>

≈ 1/7000, frequenza del
 profilo nella popolazione
 di riferimento

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

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

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-6}) \approx 0.992908$$

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 0.992908, 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

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-2}) \approx 0.0137$$

References:

- G. D' Agostini, Rep. Prog. Phys. **66** (2003) 1383
- V. Dose, Rep. Prog. Phys. **66** (2003) 1421
- W. C. Thompson and E. L. Schulman, Law and Human Behavior **11** (1987) 167
- M. Botje, lecture notes: <http://www.nikhef.nl/~h24/bayes/>