Claudio Mattiussi

A tutorial on (Bayesian) probability

Laboratory of Intelligent Systems
Institute of System Engineering
Ecole Polytechnique Fédérale de Lausanne

LIS seminar – 1st July 2005

Probability theory: the science of plausible reasoning

[L]a théorie de la probabilité n'est au fond que le bon sens réduit au calcul: elle fait apprécier avec exactitude, ce que les esprits justes sentent par une sorte d'instinct, sans qu'il puissent souvent s'en rendre compte.

Pierre-Simon Laplace, ***1812***

Probability theory is nothing but common sense reduced to calculation…

The example problem

Estimate and compare the rate of success r of two evolutionary methods M_1 and M_2 applied to a given problem (in short, the rate of success r of two experiments E_1 and E_2), given the observation of the result of two series of runs

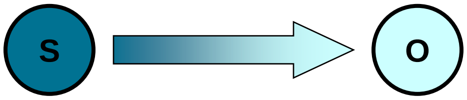
E_1	E_2
13 runs	*12 runs*
4 successes	*7 successes*
9 failures	*5 failures*

Intuitively (“instinctively”), experiment E_2 has a greater rate of success. We would like to express more precisely (“exactly”) the significance of the observed results

What kind of problem is this?

Direct problems (sampling, deductive, …)

We know the system (its state, its parameters…) and we want to tell what we can expect from observing it

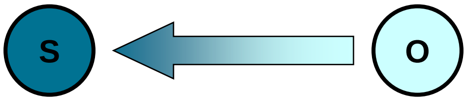


Examples
- Outcome of draws from an urn of know composition
- Outcome of tosses of a fair coin
- Value of optical flow given plane pitch
- Outcome of runs of evolutionary experiment given rate of success
- …

Direct problems are conceptually “easy” (counting, geometry, elementary physics) although they can be technically challenging (combinatorial theory…)

Inverse problems (estimation, inductive, …)

We know the outcome of a series of observations of the system and we want to estimate its properties (state, parameters…)



Examples
- Composition of urn from observation of outcome of draws
- Fairness of coin from observation of outcome of tosses
- Plane pitch from observation of optic flow
- Rate of success of evolutionary experiment from observation of outcome of runs
- …

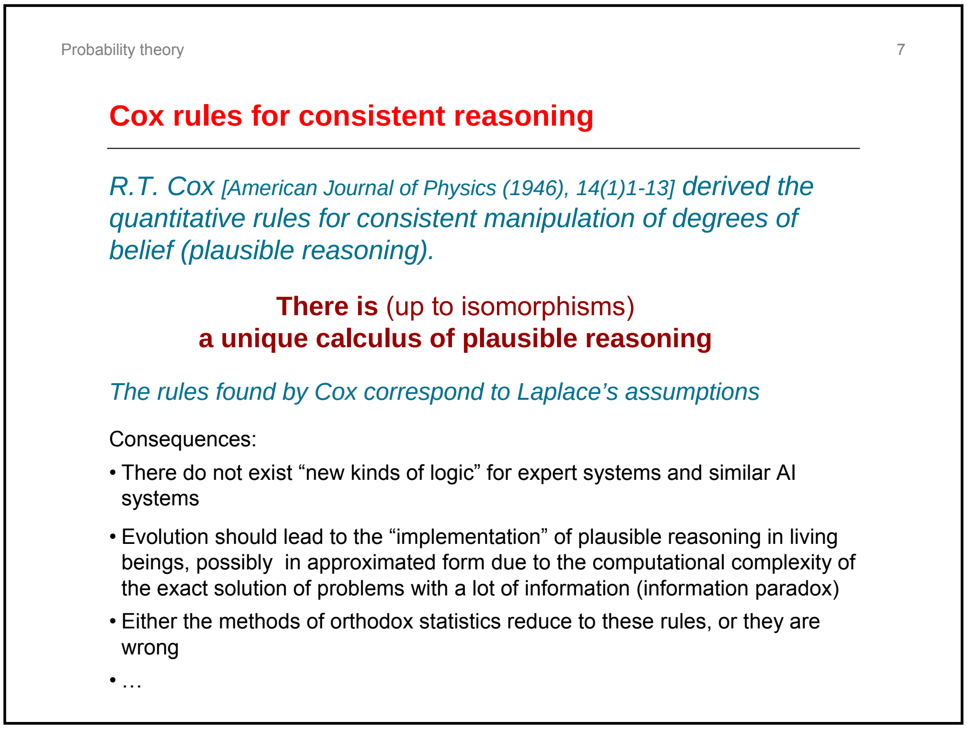
Inverse problems are conceptually “difficult” (we sometimes guess some properties of the system but the complete solution is typically not intuitive) but they are also the most relevant in science and technology

Probability theory

A probability is a numerical value representing our degree of belief (plausibility) in the truth of a proposition

Examples of propositions
- The urn contains four blue balls and six green balls
- The plane pitch is five degrees nose up
- The rate of success of experiment E is 0.7

- This definition is *subjective* (probability depends on our state of information)
- Subjective does not mean *arbitrary*
- The main requisite is *consistency*
  - *Two persons with the same information must obtain to the same numerical value*
  - *Using the same information in different ways (e.g., updating progressively our belief or using all the information at the end…) must give the same result*



Cox rules for consistent reasoning

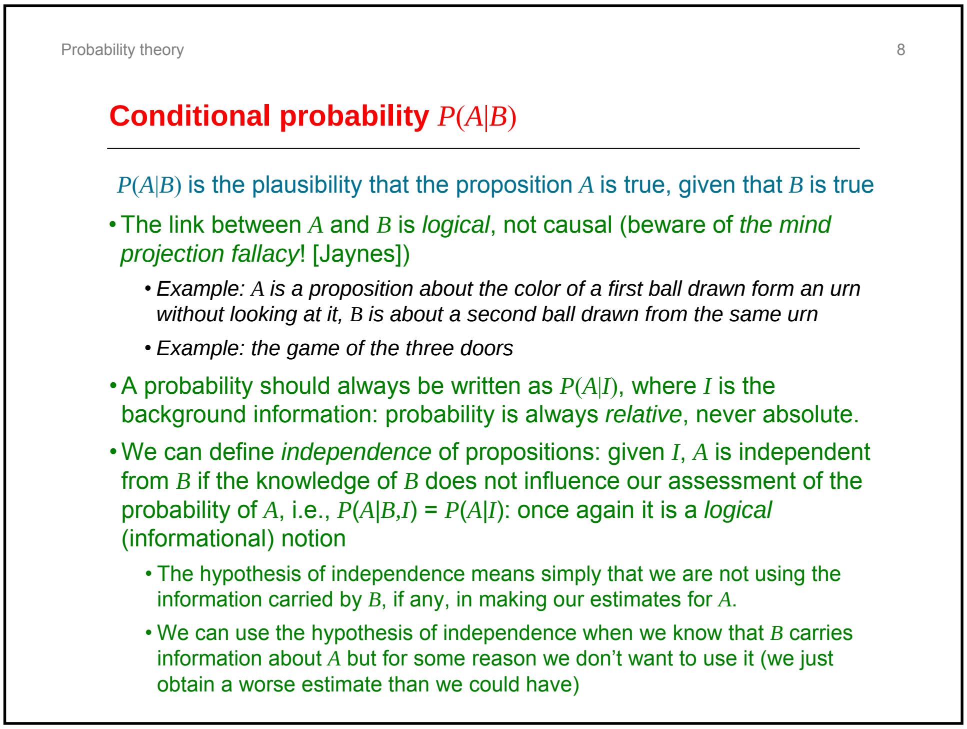
R.T. Cox [American Journal of Physics (1946), 14(1)1-13] derived the quantitative rules for consistent manipulation of degrees of belief (plausible reasoning).

There is (up to isomorphisms)
a unique calculus of plausible reasoning

The rules found by Cox correspond to Laplace's assumptions

Consequences:

- There do not exist "new kinds of logic" for expert systems and similar AI systems
- Evolution should lead to the "implementation" of plausible reasoning in living beings, possibly in approximated form due to the computational complexity of the exact solution of problems with a lot of information (information paradox)
- Either the methods of orthodox statistics reduce to these rules, or they are wrong
- …



Conditional probability $P(A|B)$

$P(A|B)$ is the plausibility that the proposition A is true, given that B is true

- The link between A and B is *logical*, not causal (beware of *the mind projection fallacy*! [Jaynes])
  - *Example: A is a proposition about the color of a first ball drawn form an urn without looking at it, B is about a second ball drawn from the same urn*
  - *Example: the game of the three doors*
- A probability should always be written as $P(A|I)$, where I is the background information: probability is always *relative*, never absolute.
- We can define *independence* of propositions: given I, A is independent from B if the knowledge of B does not influence our assessment of the probability of A, i.e., $P(A|B,I) = P(A|I)$: once again it is a *logical* (informational) notion
  - The hypothesis of independence means simply that we are not using the information carried by B, if any, in making our estimates for A.
  - We can use the hypothesis of independence when we know that B carries information about A but for some reason we don't want to use it (we just obtain a worse estimate than we could have)

Rules of consistent reasoning

- Range of $P(A|I)$, value for certainty of truth and falsity

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

- Sum rule

$$P(A|I) + \mathrm{P}(\overline{A}|I) = 1$$

($P(\overline{A}|I)$ is our degree of belief in the falsity of the proposition A)

- Product rule

$$P(A,B|I) = P(A|B,I) \cdot P(B|I)$$

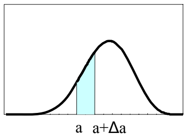
(A,B means "A and B")

Some consequences of the rules

- Set of mutually exclusive (alternative) propositions

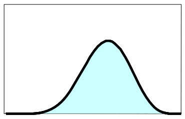
$$\textstyle\sum_i P(A_i|I) \qquad \int_{A}^{A+\Delta A} p(A|I)\,\mathrm{d}A$$

($p(A)$ is a *probability density function*)



- Exhaustive set of alternative propositions

$$\textstyle\sum_i P(A_i|I) = 1 \qquad \int_{-\infty}^{+\infty} p(A|I)\,\mathrm{d}A = 1$$



- Marginalization

$$\textstyle\sum_i P(A,B_i|I) = P(A|I) \qquad \int_{-\infty}^{+\infty} p(A,B|I)\,\mathrm{d}B = p(A|I)$$

- Bayes' theorem

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

Solving our inverse problem via Bayes’ theorem

In general: we know the outcome of a series of observations of the system and we want to estimate its properties (state, parameters…)

likelihood
sampling probability
direct probability

prior probability
“a priori” probability
“the prior”

$$P(S|O,I) = \frac{P(O|S,I)\cdot P(S|I)}{P(O|I)}$$

posterior probability
“a posteriori” probability
inverse probability

evidence

In our case: Estimate the rate of success r of an evolutionary given the observation of the outcomes $\{O_i\}$ of a series of runs

$$p(r|\{O_i\},I) = \frac{p(\{O_i\}|r,I)\cdot p(r|I)}{p(\{O_i\}|I)} \propto p(\{O_i\}|r,I)\cdot p(r|I)$$

the evidence does not depend on *r*

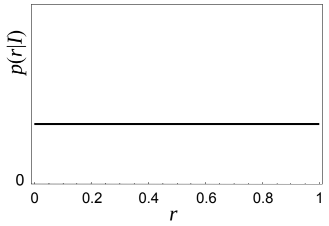
we estimate up to a multiplicative constant: we normalize afterwards

Updating our prior after the observation of a success

- The prior distribution

$$p(r|I) = \text{const}$$

(this kind of information typically follows from symmetry considerations, maximum entropy…)



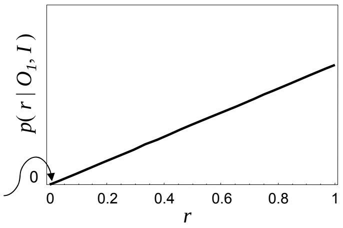
- The likelihood of the observ. $\{O_1 = success\}$

(it’s the probability of success assuming that the success ratio is *r*)

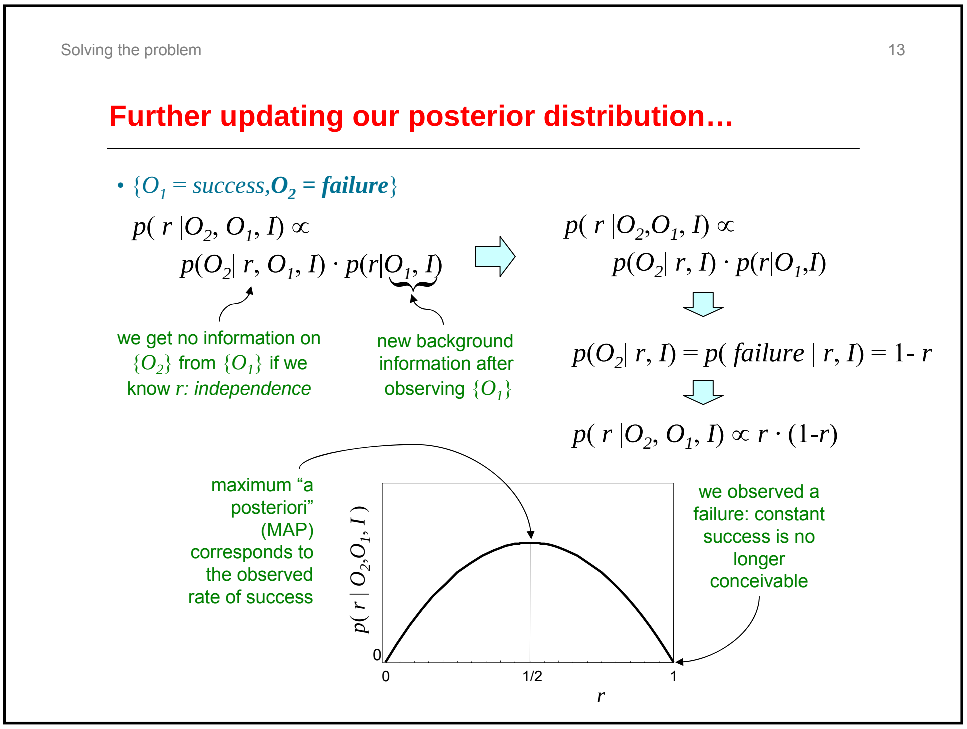
$$p(O_1|r,I) = p(\,success\,|\,r,I) = r$$

- The posterior distribution $P(r|\{O_1\},I)$

$$p(r|O_1,I) \propto p(O_1|r,I)\cdot p(r|I) \propto r$$



we observed a success: constant failure is no longer conceivable



Further updating our posterior distribution…

- $\{O_1 = success, \mathbf{O_2} = \boldsymbol{failure}\}$

$$p(\,r\,|O_2, O_1, I) \propto p(O_2|\,r, O_1, I) \cdot p(r|\underbrace{O_1, I}) \Rightarrow p(\,r\,|O_2,O_1, I) \propto p(O_2|\,r, I) \cdot p(r|O_1,I)$$

we get no information on $\{O_2\}$ from $\{O_1\}$ if we know *r: independence*

new background information after observing $\{O_1\}$

$$\Downarrow$$

$$p(O_2|\,r, I) = p(\,failure\,|\,r, I) = 1\text{-}\,r$$

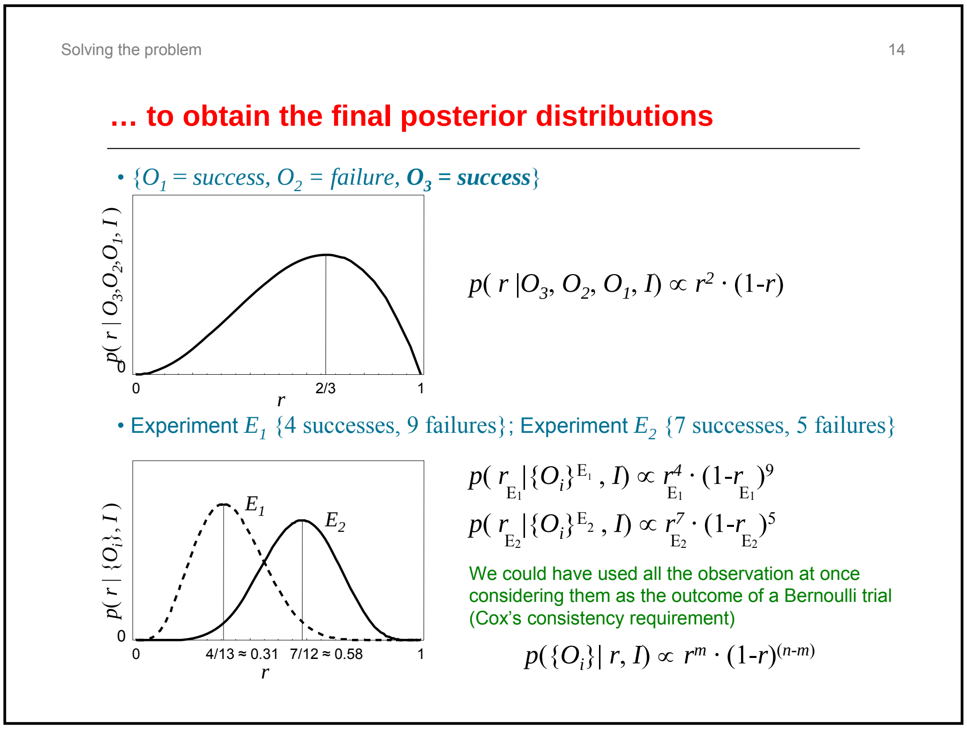
$$\Downarrow$$

$$p(\,r\,|O_2, O_1, I) \propto r \cdot (1\text{-}r)$$

maximum "a posteriori" (MAP) corresponds to the observed rate of success

we observed a failure: constant success is no longer conceivable

$p(\,r\,|\,O_2,O_1,\,I\,)$
0
0
1/2
1
r



… to obtain the final posterior distributions

- $\{O_1 = success,\, O_2 = failure,\, \mathbf{O_3} = \boldsymbol{success}\}$

$p(\,r\,|\,O_3,O_2,O_1,\,I\,)$
0
0
2/3
1
r

$$p(\,r\,|O_3, O_2, O_1, I) \propto r^2 \cdot (1\text{-}r)$$

- Experiment E_1 $\{4$ successes, 9 failures$\}$; Experiment E_2 $\{7$ successes, 5 failures$\}$

$p(\,r\,|\,\{O_i\},\,I\,)$
E_1
E_2
0
0
4/13 ≈ 0.31
7/12 ≈ 0.58
1
r

$$p(\,r_{E_1}|\{O_i\}^{E_1}, I) \propto r_{E_1}^4 \cdot (1\text{-}r_{E_1})^9$$

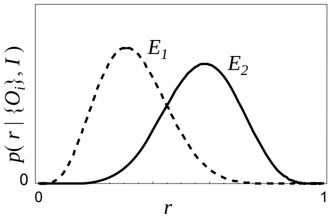
$$p(\,r_{E_2}|\{O_i\}^{E_2}, I) \propto r_{E_2}^7 \cdot (1\text{-}r_{E_2})^5$$

We could have used all the observation at once considering them as the outcome of a Bernoulli trial (Cox's consistency requirement)

$$p(\{O_i\}|\,r, I) \propto r^m \cdot (1\text{-}r)^{(n\text{-}m)}$$

The solution of the inverse problem

Experiment E_1 {4 successes, 9 failures}; Experiment E_2 {7 successes, 5 failures};



By normalizing we obtain:

$$p(\ r_{E_1}|\{O_i\}^{E_1}, I) = \frac{14!}{4!\ 9!}\ r_{E_1}^4 \cdot (1-r_{E_1})^9$$

$$p(\ r_{E_2}|\{O_i\}^{E_2}, I) = \frac{13!}{7!\ 5!}\ r_{E_2}^7 \cdot (1-r_{E_2})^5$$

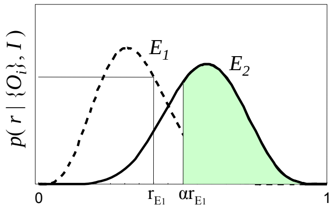
These probability density functions contain all the information conveyed by our observations (combined with our prior information)

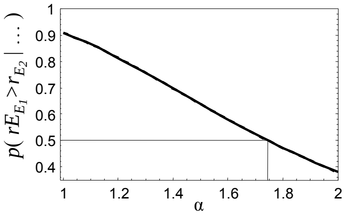
We can still process the pdfs to make the solution more perspicuous

WARNING: No probability *concepts* introduced beyond this point!

Treatment of the solution

Determine $P(r_{E_2} > r_{E_1}|\{O_i\}^{E_1},\{O_j\}^{E_2},I)$ and, more generally $P(r_{E_2} > \alpha\ r_{E_1}|\ldots)$





$$P(r_{E_2} > \alpha\ r_{E_1}|\{O_i\}^{E_1},\{O_j\}^{E_2},I) = \int_0^1 p(r_{E_1}|\{O_i\}^{E_1},I) \int_{\min(\alpha r_{E_1},1)}^1 p(r_{E_2}|\{O_j\}^{E_2},I)\ dr_{E_2}\ dr_{E_1}$$

$$P(r_{E_2} > r_{E_1}|\ldots) \approx 0.91$$
$$P(r_{E_2} > 1.1\ r_{E_1}|\ldots) \approx 0.87$$
$$P(r_{E_2} > 1.2\ r_{E_1}|\ldots) \approx 0.81$$
$$\vdots$$
$$P(r_{E_2} > 1.74\ r_{E_1}|\ldots) \approx 0.5$$
$$\vdots$$

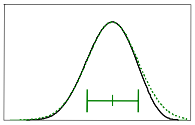
Note that probability theory does not (and cannot) say if the result is “significant”. You need additional criteria for that.

Summarizing the posterior distribution

- Point estimates

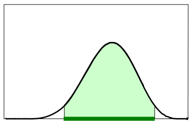
To define a “best” estimate you need to define the cost of being wrong. It’s an issue that pertains to *decision theory*. Different cost functions give different point estimates (median, mean, MAP…)

- Error bars



The posterior distribution is not necessarily a Gaussian function. By approximating it with a Gaussian at the (*unique*!) maximum you obtain an error bar…

- Confidence intervals



Given an percentage of the total probability (95%, 99%, …), you can determine the shortest interval that “contains” this amount of posterior probability…

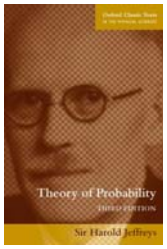
Once again, probability theory does not tell you *how* to summarize

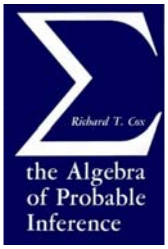
Conclusion

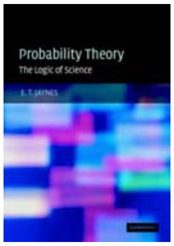
[L]a théorie de la probabilité n’est au fond que le bon sens réduit au calcul: ***elle fait apprécier avec exactitude, ce que les esprits justes sentent par une sorte d’instinct****, sans qu’il puissent souvent s’en rendre compte. […] Par là, elle devient le supplément le plus heureux à l’ignorance et à la faiblesse de l’esprit humain. Si l’on considère […]* ***la vérité des principes qui lui servent de base*** *[…] on verra qu’il n’est point de science plus digne de nos méditations, et qu’il soit plus utile de faire entrer dans le système de l’instruction publique.*

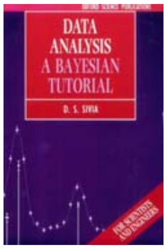
Pierre-Simon Laplace, 1812

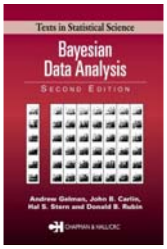
References

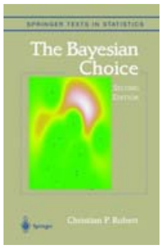









Texts in Statistical Science
Bayesian Data Analysis
SECOND EDITION
Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin
CHAPMAN & HALL/CRC



EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE