An introduction to Bayesian (thinking and) modeling

Arthur Charpentier[1]

[1] Université du Québec à Montréal

November 2022

Agenda

Preliminaries

Keynote in 2014 at the Cass Business School (now Bayes Business School)...

Getting into Bayesian Wizardry...

(with the eyes of a muggle actuary)

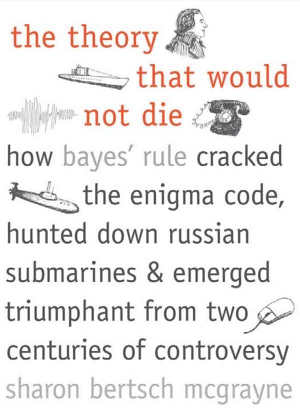
Arthur Charpentier

charpentier.arthur@uqam.ca

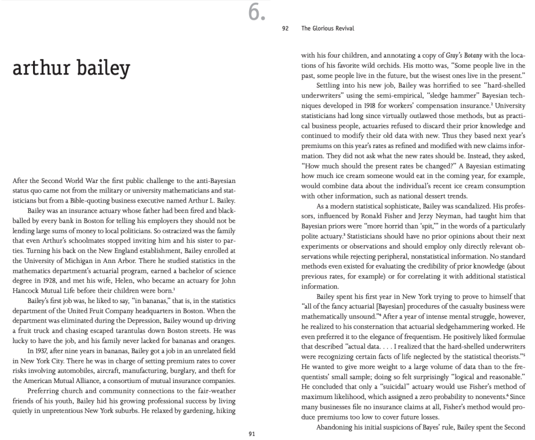
http://freakonometrics.hypotheses.org/

R in Insurance, London, July 2014

A little bit of history



contents



6.

arthur bailey

After the Second World War the first public challenge to the anti-Bayesian status quo came not from the military or university mathematicians and statisticians but from a Bible-quoting business executive named Arthur L. Bailey.

Bailey was an insurance actuary whose father had been fired and blackballed by every bank in Boston for telling his employers they should not be lending large sums of money to local politicians. So ostracized was the family that even Arthur's schoolmates stopped inviting him and his sister to parties. Turning his back on the New England establishment, Bailey enrolled at the University of Michigan in Ann Arbor. There he studied statistics in the mathematics department's actuarial program, earned a bachelor of science degree in 1928, and met his wife, Helen, who became an actuary for John Hancock Mutual Life before their children were born.[1]

Bailey's first job was, he liked to say, "in bananas," that is, in the statistics department of the United Fruit Company headquarters in Boston. When the department was eliminated during the Depression, Bailey wound up driving a fruit truck and chasing escaped tarantulas down Boston streets. He was lucky to have the job, and his family never lacked for bananas and oranges.

In 1937, after nine years in bananas, Bailey got a job in an unrelated field in New York City. There he was in charge of setting premium rates to cover risks involving automobiles, aircraft, manufacturing, burglary, and theft for the American Mutual Alliance, a consortium of mutual insurance companies.

Preferring church and community connections to the fair-weather friends of his youth, Bailey hid his growing professional success by living quietly in unpretentious New York suburbs. He relaxed by gardening, hiking

91

92 The Glorious Revival

with his four children, and annotating a copy of *Gray's Botany* with the locations of his favorite wild orchids. His motto was, "Some people live in the past, some people live in the future, but the wisest ones live in the present."

Settling into his new job, Bailey was horrified to see "hard-shelled underwriters" using the semi-empirical, "sledge hammer" Bayesian techniques developed in 1918 for workers' compensation insurance.[2] University statisticians had long since virtually outlawed those methods, but as practical business people, actuaries refused to discard their prior knowledge and continued to modify their old data with new. Thus they based next year's premiums on this year's rates as refined and modified with new claims information. They did not ask what the new rates should be. Instead, they asked, "How much should the present rates be changed?" A Bayesian estimating how much ice cream someone would eat in the coming year, for example, would combine data about the individual's recent ice cream consumption with other information, such as national dessert trends.

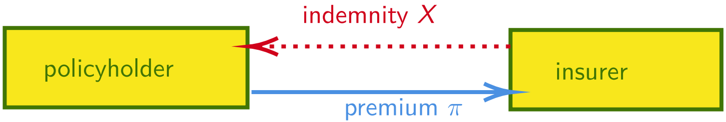
As a modern statistical sophisticate, Bailey was scandalized. His professors, influenced by Ronald Fisher and Jerzy Neyman, had taught him that Bayesian priors were "more horrid than 'spit,'" in the words of a particularly polite actuary.[3] Statisticians should have no prior opinions about their next experiments or observations and should employ only directly relevant observations while rejecting peripheral, nonstatistical information. No standard methods even existed for evaluating the credibility of prior knowledge (about previous rates, for example) or for correlating it with additional statistical information.

Bailey spent his first year in New York trying to prove to himself that "all of the fancy actuarial [Bayesian] procedures of the casualty business were mathematically unsound."[4] After a year of intense mental struggle, however, he realized to his consternation that actuarial sledgehammering worked. He even preferred it to the elegance of frequentism. He positively liked formulae that described "actual data. . . . I realized that the hard-shelled underwriters were recognizing certain facts of life neglected by the statistical theorists."[5] He wanted to give more weight to a large volume of data than to the frequentists' small sample; doing so felt surprisingly "logical and reasonable." He concluded that only a "suicidal" actuary would use Fisher's method of maximum likelihood, which assigned a zero probability to nonevents.[6] Since many businesses file no insurance claims at all, Fisher's method would produce premiums too low to cover future losses.

Abandoning his initial suspicions of Bayes' rule, Bailey spent the Second

McGrayne (2011), that mentioned Bailey (1950) (but not Whitney (1918))

Uncertainty, insurance and economics III



for the policyholder, $\pi \preceq X$ (reservation price$\geq \pi$)

formally, $\preceq$ is characterized by some utility function u and belifs $\mathbb{Q}_p$

for the insurer, $X + \sum_{i=1}^{n} X_i \leq \pi + \sum_{i=1}^{n} \pi_i$

formally, that inequality holds on average, or on probability

based on some beliefs $\mathbb{Q}_i$, e.g. $\mathbb{Q}_i\left(X + \sum_{i=1}^{n} X_i \leq \pi + \sum_{i=1}^{n} \pi_i\right) = 90\%$

Probabilities and random variables I

"*Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means* ", Russell (1929), quoted in Bell (1945)

Probabily and statistics rely on the concept of probability spaces, $(\Omega, \mathcal{F}, \mathbb{P})$,

- Ω (or S in some textbooks) is the sample space, the set of all possible outcomes
- $\mathcal{F}$ a set of events on Ω, $A \in \mathcal{F}$ is an "event"
- $\mathbb{P}$ is a function $\mathcal{F} \to [0,1]$ satisfying some properties

e.g. $\mathbb{P}(\Omega) = 1$; for disjoint events, an additiviy property: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$; aa subset property, if $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$, as inCardano (1564) or Bernoulli (1713), or for multiple disjoint events as in Kolmogorov (1933), $A_1, \cdots, A_n, \cdots$,

$$\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \cdots) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \cdots$$

inspired by Lebesgue (1918), etc. In this (mathematical) framework, we can finally define random variables

- X is a function $\Omega \to \mathbb{R}$ or more generally $\Omega \to \mathcal{X}$.

Probabilities and random variables II

We have formal objects, mathematically well defined, but in a context of modeling does one have a univocal sense of interpretation of the result of the calculation? cf "*Is the probability inherent to the event, or to our judgment?* " Martin (2009)

There are many philosophical paradoxes when we talk about probability (and chance), e.g. I throw a coin, which falls back, out of my sight

- $\mathbb{P}(X = \text{heads}) = \mathbb{P}(X = \text{tails}) = 1/2$?
- $\mathbb{P}(X = \text{heads}) = 1$ or $\mathbb{P}(X = \text{tails}) = 1$?

Or in a legal context, *Look, the guy either did it or he didn't do it. If he did then he is 100% guilty and if he didn't then he is 0% guilty; so giving the chances of guilt as a probability somewhere in between makes no sense and has no place in the law* , quoted in Fenton and Neil (2018).

See also Hájek (2002) on the philosophical approach of "probability".

Probabilities and random variables III

As said by Martin (2009),

- "*To attribute an objective meaning to the probability that an event will occur is to admit that this event is not necessary, in other words, that it is not completely determined,*"
- "*If we suppose an integral and universal determinism, the probability can only receive a subjective meaning, and the probability depends on our knowledge and our ignorance*"

Too much importance is attributed to this supposedly objective probability $\mathbb{P}$.

The (mathematical) probability was not born as a well defined concept within the framework of a mathematical formalism mathematical formalism, but as a tool to quantify and control situations of uncertainty, applied to the measurement of the probability of life mortality tables (for the calculation of life annuities), the calculation of the risks of error (in of error (in measurement operations), the study of the probability of testimonies and judgments, etc.

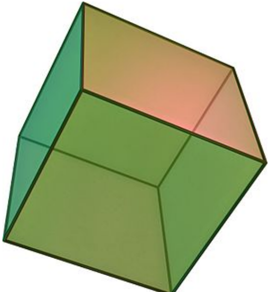
Probabilities and random variables IV

"*The theory of probabilities is basically only common sense reduced to calculation: it makes appreciate with exactitude, what the just minds feel by a kind of instinct, without them often being able to realize it*", Laplace (1774)

Cournot (1843) thus distinguished a objective meaning of the probability (as measure of the physical possibility of realization of a random event) and a subjective meaning (the probability being a judgement made on an event, this judgement being linked to the ignorance of judgment being linked to the ignorance of the conditions of the realization of the event).

Note: a probability not defined in terms of frequency can receive an objective meaning: :

There is no need to repeat throws of dice to affirm that (with a perfectly balanced die) the probability of obtaining 6 at the time of a throw is equal to $1/6$ (by symmetry of the cube)



Probabilities and random variables V

But very often, the "physical" probabilities receive an objective value only posterior on the basis of the law of large numbers, the empirical frequency converge towards the probability (frequentist theory of probabilities)

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i\in A)}_{\text{(empirical) frequency}} \overset{\text{a.s.}}{\rightarrow} \underbrace{\mathbb{P}(X\in A)}_{\text{probability}} \text{ as } n\rightarrow\infty$$

(in some textbooks, there is a confusion between "probability" and "frequency")

$$\text{Law of large numbers : }\frac{1}{n}\sum_{i=1}^{n}X_i \overset{\text{a.s.}}{\rightarrow} \mathbb{E}(X) \text{ as } n\rightarrow\infty \text{ or } \frac{1}{n}\sum_{i=1}^{n}X_i\approx\mathbb{E}(X)$$

Probabilities and random variables VI

But this approach is unable to make sense of the probability of a "(single singular event", as noted by von Mises (1928, 1939).

"*When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all.*"

Probabilities and random variables VII

For Popper (1959), probabilities correspond to physical dispositions ("propensions") inherent to the system. This propensity has a physical existence, but it is not directly observable.

The frequencies of occurrence are manifestations of these propensities. In the contrary case, it is nevertheless possible to estimate the probability of realization of the singular event, by considering this one as measured not by an "actual" frequency, but by a "potential" (or "virtual") frequency.

Finally, when an individual makes a judgment, the degree of credibility or belief that he or she gives it depends on the knowledge that the individual has (Pettigrew (2016)). depends on the knowledge that this individual has (Pettigrew (2016)). This degree of belief will be associated with a probability, which will then only have a subjective meaning. "*The probability of a diagnosis, a testimony, etc., does not measure the conformity of this judgment to reality, but the degree to which one can hypothesize this conformity. This conformity can be hypothesized*", Martin (2009).

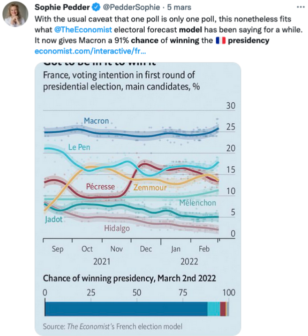
Probabilities and random variables VIII

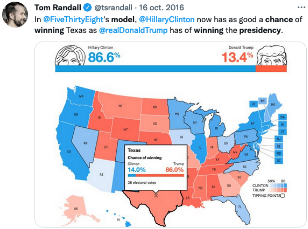
This subjectivity raises concerns about their use, especially in criminal matters, "*Sometimes the 'balance of probability' standard is expressed mathematically as '50+% probability', but this can carry with it a danger of pseudo-mathematics, as the argument in this case demonstrated. When judging whether a case for believing that an event was caused in a particular way is stronger than the case for not so believing, the process is not scientific (although it may obviously include evaluation of scientific evidence) and to express the probability of some event having happened in percentage terms is illusory*, *Nulty & Ors v Milton Keynes Borough Council* cited in Hunt and Mostyn (2020).

See also Jonakait (1983), Saini (2011) or Fenton et al. (2016).

Probability ? Probability to win an election ?

@PedderSophie (The Economist), vs @HuffPost or @tsrandall (Bloomberg)





How to interpret this "probability of winning" ?

How to interpret a "confidence interval" on that probability ? (@AdamSinger)

Probability ? Probability of precipitation ? I

How to interpret the 'P.o.P.' ("Probability of Precipitation") on weather websites ?

	jeu. 14/07	ven. 15/07	sam. 16/07	dim. 17/07	lun. 18/07	mar. 19/07	mer. 20/07
	Ensoleillé avec passages nuageux	Ensoleillé	Ensoleillé	Ensoleillé avec passages nuageux	Ensoleillé	Ensoleillé avec passages nuageux	Risque d'averses
	31°	27°	28°	30°	35°	38°	28°
T. ressentie	30	26	27	28	32	35	28
Nuit	16°	15°	15°	18°	22°	21°	19°
P.D.P.	20 %	0 %	0 %	20 %	0 %	10 %	40 %
Vents (km/h)	19 N.-O.	13 N.-E.	15 N.-E.	17 N.-E.	15 E.	20 E.	19 S.-O.
Rafales (km/h)	28	19	22	25	23	30	29
Ensol. (H)	11 h	15 h	14 h	12 h	15 h	12 h	12 h
Pluie 24 H	-	-	-	-	-	~1 mm	~1 mm

	jeu. 14/07	ven. 15/07	sam. 16/07	dim. 17/07	lun. 18/07	mar. 19/07	mer. 20/07
	Nuageux avec orages dispersés	Généralement ensoleillé	Ciel variable	Possibilité d'orages	Risque d'averses	Risque d'averses	Nuageux avec éclaircies
	24°	26°	27°	28°	28°	28°	29°
T. ressentie	27	29	31	33	35	34	35
Nuit	15°	16°	19°	20°	20°	22°	21°
P.D.P.	40 %	10 %	20 %	40 %	40 %	40 %	30 %
Vents (km/h)	15 N.	15 O.	19 S.-O.	20 S.-O.	6 O.	28 S.-O.	26 S.-O.
Rafales (km/h)	23	23	29	30	9	42	39
Ensol. (H)	3 h	13 h	9 h	4 h	4 h	6 h	3 h
Pluie 24 H	<1 mm	-	-	~1 mm	<1 mm	~5 mm	~5 mm

	jeu. 14/07	ven. 15/07	sam. 16/07	dim. 17/07	lun. 18/07	mar. 19/07	mer. 20/07
	Pluie	Faible pluie	Ciel variable	Nuageux	Ensoleillé avec passages nuageux	Ensoleillé	Ensoleillé
	8°	6°	9°	9°	11°	17°	17°
T. ressentie	8	6	9	9	11	17	17
Nuit	3°	1°	4°	4°	6°	10°	7°
P.D.P.	100 %	90 %	20 %	30 %	10 %	0 %	0 %
Vents (km/h)	11 N.	6 N.-E.	5 E.	5 S.-E.	3 S.	5 S.-E.	6 E.
Rafales (km/h)	17	8	7	8	4	7	9
Ensol. (H)	1 h	0 h	5 h	0 h	6 h	10 h	10 h
Pluie 24 H	25 - 35 mm	5-10 mm	-	~15 mm	-	-	-

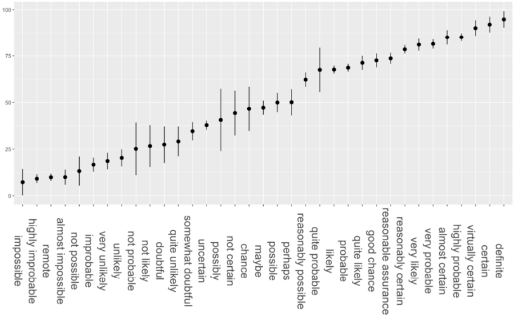
"*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated*, Silver (2012)

Murphy and Epstein (1967), Roberts (1968)

Gneiting and Raftery (2005) on ensemble methods for weather forecasting.

Probability ? Probability of precipitation ? II

More generally, we can think of the "probabilities" mentioned by the IPCC, Mastrandrea et al. (2010) discussed in Stoerk et al. (2020) or Kause et al. (2022)



(source Vogel et al. (2022))

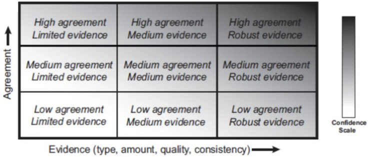
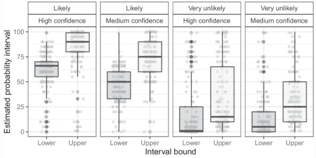


Table 1. Likelihood Scale

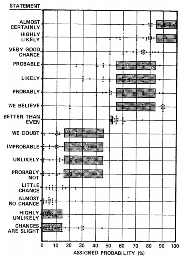
Term*	Likelihood of the Outcome
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

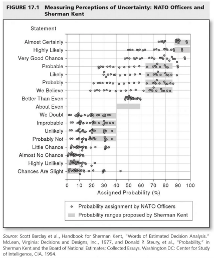


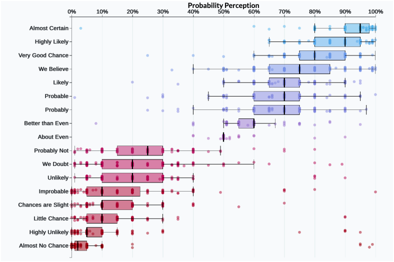
Probability ? Probability of precipitation ? III

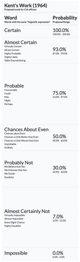
Note : "Cromwell's rule": one should not give a probability of 1 to an event that cannot logically be shown to be true, and one should never give a probability of 0 to an event unless it can logically be shown to be false,

Lindley (2013), Barclay et al. (1977) et Pherson and Pherson (2012).



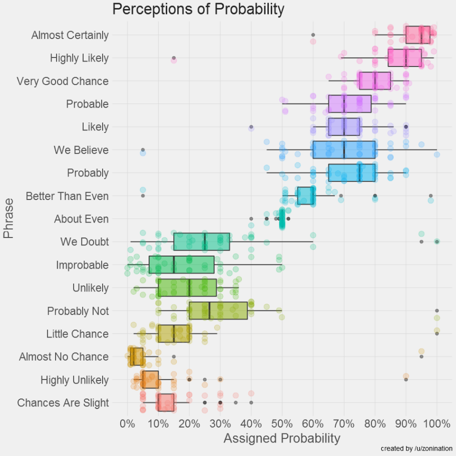


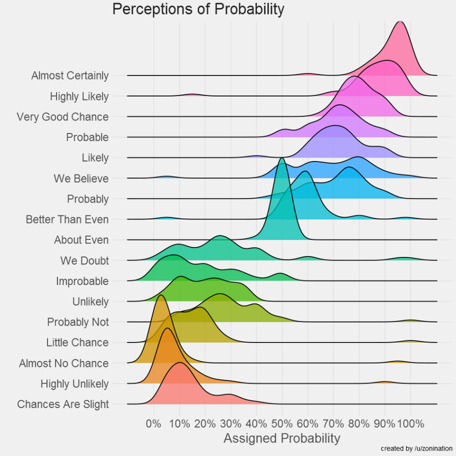




Probability ? Probability of precipitation ? IV

See also @zonination on "probability perceptions"





Bayesian statistics ?

- Bayes formula (the “inverse problem"),
  Bayes (1763), Laplace (1774)

Given two events A and B such that $\mathbb{P}(B) \neq 0$,

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

“*If a person has an expectation depending on the happening of an event, the probability of the event is [in the ratio] to the probability of its failure as his loss if it fails [is in the ratio] to his gain if it happens* ", Proposition 2, Bayes (1763)

“*The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the chance of the thing expected upon its happening* ", Bayes (1763)

Bayesian statistics ?

- Bayes formula (the “inverse problem"),

  Bayes (1763), Laplace (1774)

Given two events A and B such that $\mathbb{P}(B) \neq 0$,

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

- subjective probabilities,

  De Finetti (1937), Anscombe et al. (1963), Kahneman and Tversky (1972) Savage (1972), Jeffrey (2004)

- Non-frequentist approach of probablities,

  Neyman (1977), Bayarri and Berger (2004)

- Credibility and “*experience rating*”

  Whitney (1918), Longley-Cook (1962), Bühlmann (1967), Klugman (1991)

Bayesian statistics ?

- Bayes formula (the "inverse problem"),
  Bayes (1763), Laplace (1774)

Given two events A and B such that $\mathbb{P}(B) \neq 0$,

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

- An inverse problem (we try to determine the causes of a phenomenon of a phenomenon from the experimental observation of its effects)
- An update of beliefs (from a *prior* distribution $\mathbb{P}(A)$ to a *posterior* distribution $\mathbb{P}(A|B)$)

Bayesian statistics ?

A person coughs (event B). Which hypothesis is the most credible?
(from Dehaene (2012))

$$\begin{cases} A_1 : \text{she has lung cancer} \\ A_2 : \text{she has gastroenteritis} \\ A_3 : \text{she has the flu} \end{cases}$$

With Bayes' rule $\mathbb{P}[\text{disease}|\text{symptom}] \propto \mathbb{P}[\text{symptom}|\text{disease}] \cdot \mathbb{P}[\text{disease}]$

$$\begin{cases} A_1 : \mathbb{P}[\text{disease}] \approx 0 \text{ (even if } \mathbb{P}[\text{symptom}|\text{disease}] \approx 1) \\ A_2 : \mathbb{P}[\text{symptom}|\text{disease}] \approx 0 \text{ (even if } \mathbb{P}[\text{symptom}|\text{disease}] \text{ high)} \\ A_3 : \text{two reasonable probabilities} \end{cases}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

$$\mathbb{P}(\text{treasure behind the door}) = \frac{1}{3}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

$$\mathbb{P}(\text{treasure behind the door}) = \frac{1}{3}$$

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

- strategy 1 : always switch the door
- strategy 2 : never switch the door

$$\mathbb{P}(\text{strategy 2 winning})$$
$$= \mathbb{P}(\text{treasure behind the door choisie initialement})$$
$$= \frac{1}{3}$$

(making the goat appear behind the third door does not bring no information on what's behind the first door)

The practice of conditional probabilities

"Monty Hall" problem
(from *Let's make a deal*)

- strategy 1 : always switch the door
- strategy 2 : never switch the door

$$
\begin{aligned}
&\mathbb{P}(\text{strategy 1 winning})\\
&= \mathbb{P}(\text{treasure behind the other door})\\
&= \mathbb{P}(\text{treasure behind the other door} \mid \text{ correct }) \cdot \mathbb{P}(\text{ correct })\\
&+ \mathbb{P}(\text{treasure behind the other door} \mid \text{ false }) \cdot \mathbb{P}(\text{ false })\\
&= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}
\end{aligned}
$$

Practice of Bayesian Statistics

"*Do doctors understand test results?* ", Kremer (2014):

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

A) 9 out of 10

B) 8 out of 10

C) 1 out of 2

D) 1 out of 10

E) 1 out of 100

Practice of Bayesian Statistics

"*Do doctors understand test results?* ", Kremer (2014):

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

A) 9 out of 10 (chosen by 50% gynecologists)

B) out of 10

C) 1 out of 2

D) 1 out of 10

E) 1 out of 100

Practice of Bayesian Statistics

1 percent of adults have cancer. The vast majority of these cancers (90 percent) can be detected by a test. There is a 9 percent chance that the test will be positive in a person who does not have cancer. If the test is positive, what is the likelihood that the person actually has cancer?

Answer: when formalizing

$$\begin{cases} \mathbb{P}[\text{cancer}] = 1\% \\ \mathbb{P}[\text{test positive}|\text{cancer}] = 90\% \\ \mathbb{P}[\text{test positive}|\text{no cancer}] = 9\% \end{cases}$$

then, using Bayes' rule

$$\mathbb{P}[\text{cancer}|\text{test positive}] = \frac{\mathbb{P}[\text{test positive}|\text{cancer}] \cdot \mathbb{P}[\text{cancer}]}{\mathbb{P}[\text{test positive}]} = \frac{90\% \times 1\%}{9\% \times 99\% +} = \frac{9}{9+89} \simeq \frac{1}{10}$$

valid answer is D, "1 out of 10".

Practice of Bayesian Statistics

For Gigerenzer and Hoffrage (1995), the Bayesian formulation is (too) complex.

Another presentation of the problem:

Out of 10,000 people, 100 have cancer. Of these 100, 90%, or 90, will test positive. Of the remaining 9,900, 9 percent, or 899, will test positive. Of a sample of people who test positive, what fraction actually have cancer?

Answer: 90 among (90+899), i.e. about “1 out of 10”.

Axiomatic of beliefs I

Axioms of Bayesian approach, Titelbaum (2022a), (2022b), are

- step 1 : beliefs

Beliefs are quantified on a scale from 0 to 1

The "rationality of beliefs" means that beliefs are measures of probabilities (and verify the associated axioms), Buehler (1976).

Note: a weaker version of coherence can be defined using capacities (in the sense of Choquet (1954)), based on the axiom : if $A \subset B$, then $\mathbb{Q}[A] \leq \mathbb{Q}[B]$ (and no longer the additivity of disjoint events)

Axiomatic of beliefs II

- step 2 : updating beliefs

For Popper (1955), an agent who believes A to the degree $Q[A]$, if he learns B, he then believes A to the degree $Q[A|B]$

$$\mathbb{Q}[A] \mapsto \mathbb{Q}[A|B]\cdot\underbrace{\mathbb{Q}[B]}_{=1}+\mathbb{Q}[A|\neg B]\cdot\underbrace{\mathbb{Q}[\neg B]}_{=0}=\mathbb{Q}[A|B]=\mathbb{Q}_B[A]$$

Jeffrey (1965) proposed a generalization if B is associated with a belief $Q'[B]$,

$$\mathbb{Q}[A] \mapsto \mathbb{Q}'[A]=\mathbb{Q}[A|B]\cdot\mathbb{Q}'[B]+\mathbb{Q}[A|\neg B]\cdot\mathbb{Q}'[\neg B]$$

In other words, "reasoning consists of graduating one's beliefs and revising one's degrees of belief by Bayesian conditionalization as new information becomes available", Drouet (2016).

Axiomatic of beliefs III

"*La differenza essenziale da rilevare è nell'attribuzione del 'perchè': non cerco perchè IL FATTO che io prevedo accadrà, ma perchè IO prevedo che il fatto accadrà. Non sono più i fatti che hanno bisogno di una causa per prodursi : è il nostro pensiero che trova comodo di immaginare dei rapporti di causalità per spiegarli, coordinarli, e renderne possibile la previsione*", De Finetti (1931)

"I do not seek to know why the fact that I foresee will come true, but why I foresee that the fact will come true. It is no longer the facts that need a cause to happen: it is our mind that finds it convenient to imagine causal relationships in order to explain them, to coordinate them and to make the prediction possible"

The Dutch book I

Ramsey (1926) and De Finetti (1937) suggested to understand the rationality of beliefs with the help of bets (formalized by Lehman (1955) Kemeny (1955), Teller (1973), Lindley et al. (1979) and Skyrms (1987)) and "arbitrage" (we speak of Subjective Bayesianism).

We assign the belief q to a bet (lottery) associated to A, yielding a if A occurs and 0 otherwise if and only if the value of the lottery is qa, Hájek (2009)

The dutch book argument is that if an individual has beliefs that violate the probabilities and if he bets based on those beliefs, then he is willing to accept a set of bets that he is certain to lose, Pettigrew (2020).

Note: Lehman (1955) used the term "dutch book", but it corresponds to the notion of "arbitrage" in financial mathematics.

The Dutch book II

Lehman (1955) "*if a set of betting prices violate the probability calculus, then there is a Dutch Book consisting of bets at those prices.*"

Kemeny (1955), "*if a set of betting prices obey the probability calculus, then there does not exist a Dutch Book consisting of bets at those prices*"

This characterization is also called Cox-Jaynes theorem, Cox (1946) taken up by Jaynes (1988) and Jaynes (2003) : probabilities (characterized by Kolmogorov axioms) are the only normative mechanism for plausibility induction

See also Good (1966)

or Eisenberg and Gale (1959) and Baron and Lange (2006), Chen and Pennock (2010) on parimutuel, and predictive markets

Suppose that I payers bet on J horses. Each player bets b_i, and normalize $(b_1 + \cdots + b_I = 1)$.
Player i bets $\beta_{i,j}$ on horse j ($b_i = \beta_{i,1} + \cdots + \beta_{i,J}$).

The Dutch book III

We note π_j the amount bet on the horse j ($\pi_j = \beta_{1,j} + \cdots + \beta_{I,j}$).
Since $\pi_j \in (0,1)$ and $\pi_1 + \cdots + \pi_J = 1$ is interpreted as a probability, describing a "collective belief".

We can also add empirical constraints, and associate the beliefs to known frequencies) (this is called Empirical Bayesianism)

Williamson (2004) introduced an objective Bayesianism, inspired by Jaynes (1957), based on entropy maximization (maxmin approach), associated with a precautionary principle.

Non-boolean logic I

Note We can also find links with logic.
Classically, if we have the proposition "If A is true, then B is true"

$$\begin{cases} \text{If I observe that } A \text{ is true, I conclude that } B \text{ is true} \\ \text{If I observe that } B \text{ is false, I conclude that } A \text{ is false.} \end{cases}$$

With boolean logic, these are the only equivalent assertions

$$(A \Longrightarrow B \text{ and } \neg B \Longrightarrow \neg A)$$

But there may be some plausible reasoning, Pólya (1958)

$$\begin{cases} \text{If I observe that } A \text{ is false, it seems to me that } B \text{ becomes less plausible} \\ \text{If I observe that } B \text{ is true, it seems to me that } A \text{ becomes more plausible.} \end{cases}$$

What means "plausible" here ?

Bayesianism, statistics and calculus I

$$\text{posterior} = \pi(\theta|\boldsymbol{y}) = \frac{\pi(\theta)\cdot\mathbb{P}(\boldsymbol{y}|\theta)}{\mathbb{P}(\boldsymbol{y})} = \frac{\text{prior}\cdot\text{likelihood}}{\text{evidence}}$$

$$\text{posterior} = \pi(\theta|\boldsymbol{y}) \propto \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}\cdot\binom{s}{n}\theta^{s}(1-\theta)^{n-s}$$

- Conjugate distributions: **Binomial - Beta**

The likelihood for binomial (Bernoulli) variables

$$\begin{cases}\boldsymbol{x}\mapsto f(\boldsymbol{x};p) = p^{s}(1-p)^{n-s} \text{ where } s = \boldsymbol{x}^{\top}\mathbf{1} = x_1+\cdots+x_n\\ p\mapsto p^{s}(1-p)^{n-s} \text{ on } [0,1] \text{ is a Beta distribution}\end{cases}$$

$$\text{If } \begin{cases} x_i|\theta\sim\mathcal{B}(\theta)\\ \theta\sim\mathcal{B}eta(a,b) \text{ prior}\end{cases} \quad \text{then } \theta|\boldsymbol{x}\sim\mathcal{B}eta(a+s,b+n-s) \text{ posterior}$$

(that can be extended to **Multinomial - Dirichlet**)

Bayesianism, statistics and calculus II

- Conjugate distributions : **Poisson - Gamma**

The likelihood for Poisson variables is

$$\begin{cases} \boldsymbol{x} \mapsto f(\boldsymbol{x};\lambda) = \dfrac{e^{n\lambda}\lambda^s}{x_1!\cdots x_n!} \text{ where } s = \boldsymbol{x}^\top\mathbf{1} = x_1 + \cdots + x_n \\ \lambda \mapsto e^{n\lambda}\lambda^s \text{ on } \mathbb{R}_+ \text{ is a Gamma distribution} \end{cases}$$

If

$$\begin{cases} x_i|\lambda \sim \mathcal{P}(\lambda) \\ \theta \sim \mathcal{G}amma(a,b) \text{ a priori} \end{cases} \quad \text{then } \lambda|\boldsymbol{x} \sim \mathcal{G}amma(a+s, b+n) \text{ a posteriori}$$

Hence

$$\text{a priori } \mathbb{E}(\lambda) = \frac{a}{b} \text{ and a posteriori } \mathbb{E}(\lambda|\boldsymbol{x}) = \frac{a+s}{b+n}$$

intensively used in credibility theory Bühlmann (1967).

Bayesianism, statistics and calculus III

- Conjugate distributions : **Normal - Normal**

If variance $\boldsymbol{\Sigma}$ is known

$$\begin{cases} \boldsymbol{x}_i|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{cases} \quad \text{then } \boldsymbol{\mu}|\boldsymbol{x} \sim \mathcal{N}\big(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big)$$

$$\text{where } \begin{cases} \boldsymbol{\mu}_x = \big(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\big)^{-1}\big(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}\big) \\ \boldsymbol{\Sigma}_x = \big(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\big)^{-1} \end{cases}$$

used classically in Bayesian econometrics.

Bayesianism, statistics and calculus IV

- Conjugate distributions : **Normal - Inverse Wishart**

If mean $\boldsymbol{\mu}$ is known

$$\begin{cases} \boldsymbol{x}_i|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} \sim IW(\nu_0, \boldsymbol{\Psi}_0) \end{cases} \quad \text{then } \boldsymbol{\Sigma}|\boldsymbol{x} \sim IW(\nu_x, \boldsymbol{\Psi}_x)$$

$$\text{where } \begin{cases} \nu_x = n + \nu \\ \boldsymbol{\Psi}_x = \boldsymbol{\Psi} + \sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^\top \end{cases}$$

Classically used in Bayesian econometrics, for VAR models, Adjemian and Pelgrin (2008), or in portfolio management, Black and Litterman (1990, 1992) (see also Satchell and Scowcroft (2000) for a perspective).

Bayesianism, statistics and calculus V

Bayesian methods can be very powerful for estimating panel, hierarchical, or multilevel models, Gelman and Hill (2006).

- **Hierarchical model**

When the individual i belongs to the group j,

$$y_{i,j} = \alpha_j + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \varepsilon_{i,j}, \text{ where } \begin{cases} \alpha_j = a_0 + \boldsymbol{z}_j^\top \boldsymbol{\beta}_1 + u_j \\ \boldsymbol{\beta}_j = \boldsymbol{b}_0 + \boldsymbol{Z}_j^\top \boldsymbol{B}_1 + \boldsymbol{u}_j \end{cases}$$

with constants and slopes depending on the groups.

(usually in a GLM model).

Bayesianism, statistics and calculus VI

Otherwise, either simulations are used (see MCMC) or simplifying assumptions are made.

Consider symptoms $s_1, \cdots, s_k$ and diseases $m_1, \cdots, m_j$ (in $\{0, 1\}$)

$$\mathbb{P}\big[\boldsymbol{M} = \boldsymbol{m}\big|\boldsymbol{S} = \boldsymbol{s}\big] = \frac{\mathbb{P}\big[\boldsymbol{M} = \boldsymbol{m}\big] \cdot \mathbb{P}\big[\boldsymbol{S} = \boldsymbol{s}\big|\boldsymbol{M} = \boldsymbol{m}\big]}{\sum_{\boldsymbol{x}} \mathbb{P}\big[\boldsymbol{M} = \boldsymbol{x}\big] \cdot \mathbb{P}\big[\boldsymbol{S} = \boldsymbol{s}\big|\boldsymbol{M} = \boldsymbol{x}\big]}$$

"Naïve Bayes" relies on assumptions (Spiegelhalter et al. (1993))

- diseases are mutually exclusive $\mathbb{P}\big[\boldsymbol{M} = \boldsymbol{m}\big|\boldsymbol{S} = \boldsymbol{s}\big] = 0$ si $\boldsymbol{m}^\top \mathbf{1} > 1$,
- the symptoms are conditionally independent

$$\mathbb{P}\big[\boldsymbol{S} = \boldsymbol{s}\big|M_i = m_i\big] = \prod_{j=1}^{k} \mathbb{P}\big[S_j = s_j\big|M_i = m_i\big]$$

Bayesianism, statistics and calculus VII

In that case

$$\mathbb{P}\big[M_i = m_i \big| \boldsymbol{S} = \boldsymbol{s}\big] = \frac{\mathbb{P}\big[M_i = m_i\big] \cdot \displaystyle\prod_{j=1}^{k} \mathbb{P}\big[S_j = s_j \big| M_i = m_i\big]}{\mathbb{P}\big[M_i = 0\big] \cdot \displaystyle\prod_{j=1}^{k} \mathbb{P}\big[S_j = s_j \big| M_i = 0\big] + \mathbb{P}\big[M_i = 1\big] \cdot \displaystyle\prod_{j=1}^{k} \mathbb{P}\big[S_j = s_j \big| M_i = 1\big]}$$

We can improve the model by using a Bayesian network (we will talk about it later).

Bayesianism, statistics and calculus VIII

To determine $\mathbb{P}\big[M_i = m_i \big| \boldsymbol{S} = \boldsymbol{s}\big]$, we need to know

- prevalence of disease $\mathbb{P}\big[M_i = 1\big]$
- sensitivity $\mathbb{P}\big[S_j = 1 | M_i = 1\big]$
- specificity $\mathbb{P}\big[S_j = 0 | M_i = 0\big]$

for all symptoms S_j and all disease M_i.

Note that $\mathbb{P}\big[S_j = s_j | M_i = m_i\big]$ have a causal interpretation: it is the diseases that cause the symptoms.

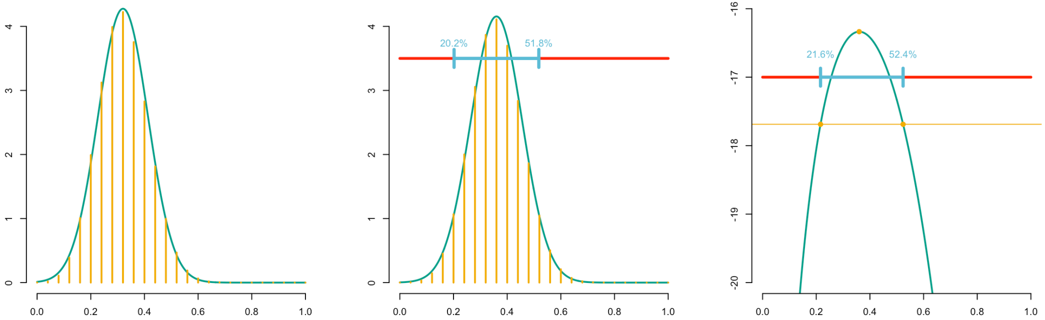
See Sadegh-Zadeh (1980) on Bayesian diagnostics, or Donnat et al. (2020).

Bayesianism, statistics and calculus I

- Posterior distribution

Suppose $\boldsymbol{x} = \{0,0,0,1,0,1,1,0,0,0,0,0,1,0,1,0,1,1,0,1,1,0,0,0,0\}$, $\mathcal{B}(\theta)$

Frequentist approach, $\widehat{\theta} \approx \mathcal{N}\left(\theta, \dfrac{\theta(1-\theta)}{n}\right)$, $\mathbb{P}\left(\theta \in \left[\overline{x} \pm 1.64\sqrt{\dfrac{\overline{x}(1-\overline{x})}{n}}\right]\right) \approx 90\%$

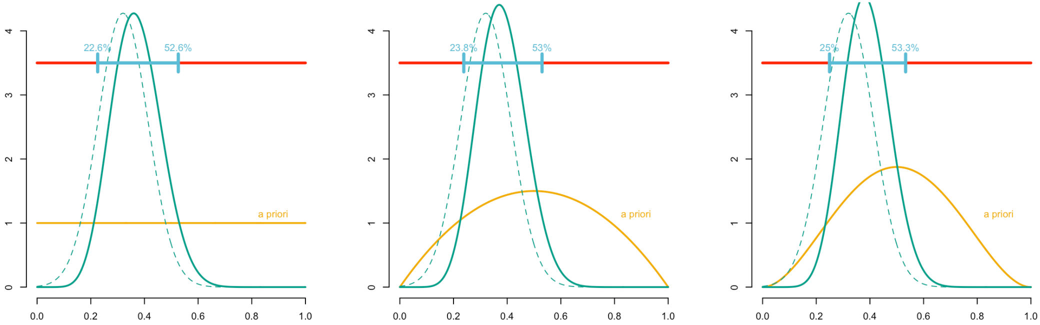


Bayesianism, statistics and calculus XVIII

- Posterior distribution

and finally $\boldsymbol{x} = \{0,0,0,1,0,1,1,0,0,0,0,0,1,0,1,0,1,1,0,1,1,0,0,0,0\}$, $\mathcal{B}(\theta)$

Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0 + s, \beta_0 + n - s)$, $s = \displaystyle\sum_{i=1}^{n} x_i$

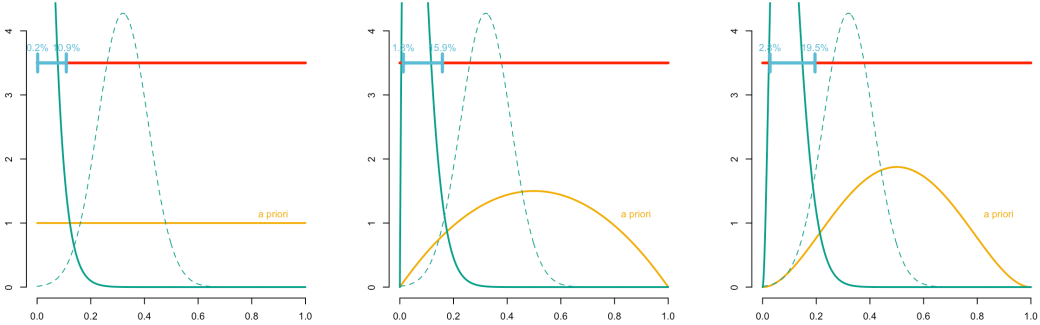


Bayesianism, statistics and calculus XIX

- Posterior distribution

What if $\boldsymbol{x} = \{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0\}$, $\mathcal{B}(\theta)$?

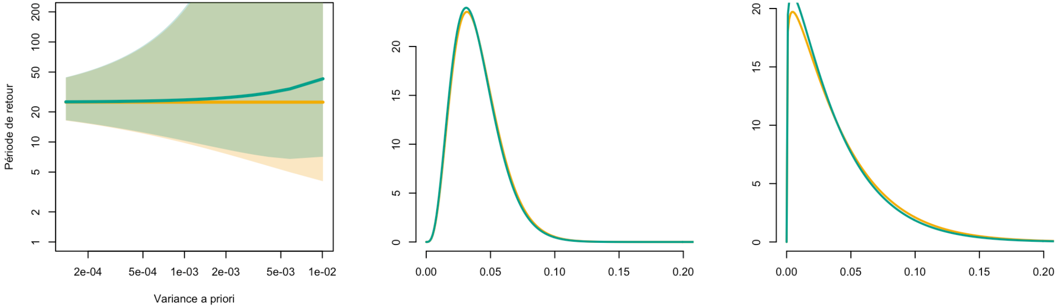
Bayesian approach, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0, \beta_0 + n)$, since $\displaystyle\sum_{i=1}^{n} x_i = 0$



Bayesianism, statistics and calculus XX

- Posterior distribution

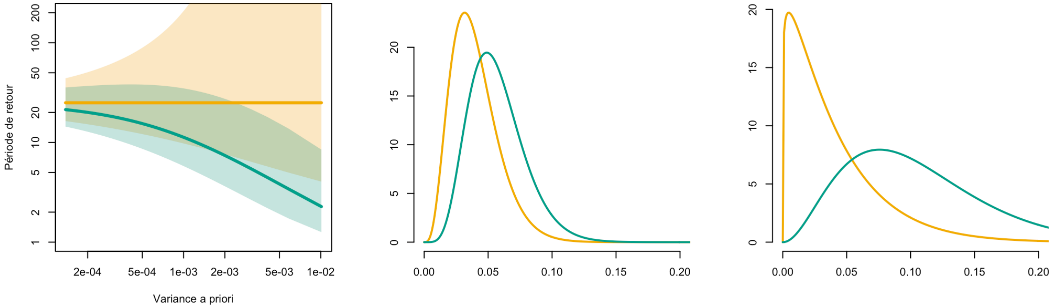
Ministère de l'intérieur (2019) "*A single threshold for qualifying a geotechnical drought as abnormal: a return period greater than or equal to 25 years* " (probabilité $1/25$) (probability $1/25$) No drought has been observed over 2 years ($\{0,0\}$), what happens to our belief about the return period?



Bayesianism, statistics and calculus XXI

- Posterior distribution

As a comparison, if we have observed two major droughts ($\{1, 1\}$), our beliefs a posteriori are very influenced by these unexpected events



Bayesianism, statistics and calculus XXII

- From the distribution to the estimator

$$\begin{cases} \text{posterior average} & \widehat{\theta} = \mathbb{E}\big[\theta|\mathcal{D}\big] \\ \text{maximum a posteriori (MAP)} & \widehat{\theta} = \max\big\{\pi(\theta|\mathcal{D})\big\} \text{ i.e. the mode} \end{cases}$$

The average posterior is also the solution of the problem

$$\widehat{\theta} = \underset{\tau}{\text{argmin}}\left\{\mathbb{E}\big[(\theta-\tau)^2|\mathcal{D}\big]\right\} = \underset{\tau}{\text{argmin}}\left\{\int(\theta-\tau)^2\pi(\theta|\mathcal{D})d\theta\right\}$$

- "confidence interval" or "credibility interval"

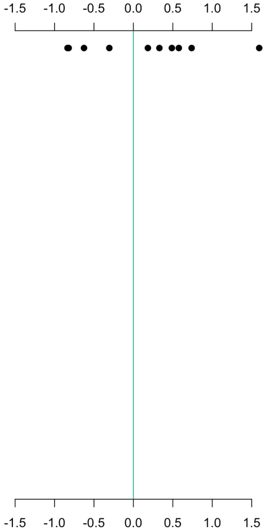
For the confidence interval, we look for $[\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]$ such that $P[\theta \in [\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]] \geq 95\%$.

For the credibility interval, we look for $[a, b]$ such that $\mathbb{P}[\theta \in [a, b]|\mathcal{D}] \geq 95\%$.

Bayesianism, statistics and calculus XXIII

- "confidence interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$ (here $\theta = 0$)

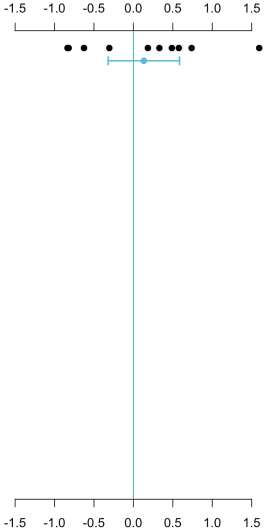


Bayesianism, statistics and calculus XXIV

- "confidence interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$ (here $\theta = 0$)

Consider $[a, b] = \left[\overline{x} \pm q_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}\right]$



Bayesianism, statistics and calculus XXV

- "confidence interval"

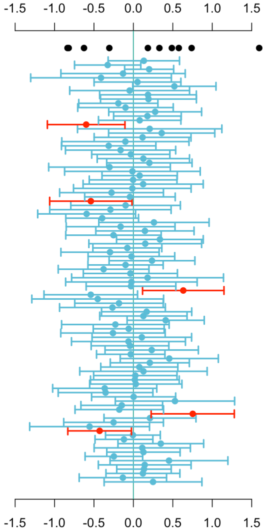
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$ (here $\theta = 0$)

Consider $[a, b] = \left[\overline{x} \pm q_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}\right]$

Generate $\mathcal{D}' = \{x'_1, \cdots, x'_n\}$ from $\mathcal{N}(\theta, \sigma^2)$, we want

$$\mathbb{P}\left[\theta \notin \left[\overline{x}' \pm q_\alpha \frac{\widehat{\sigma}'}{\sqrt{n}}\right]\right] \approx \alpha$$

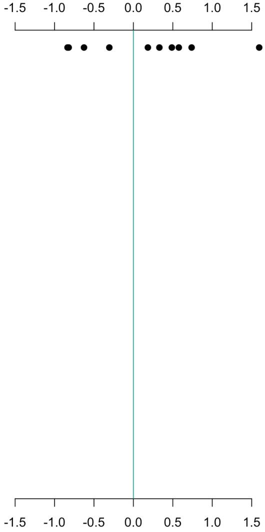
interpreted as a frequency, and repeating the experience.
Here, $\alpha = 5\%$: in 5% of the simulations, 0 is not in $[a, b]$.



Bayesianism, statistics and calculus XXVI

- "credibility interval"

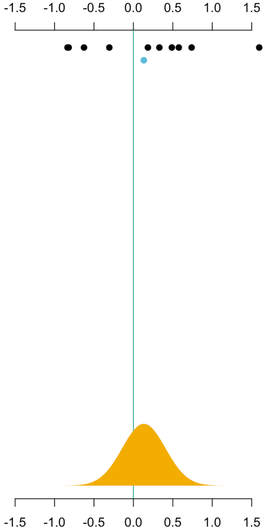
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ



Bayesianism, statistics and calculus XXVII

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

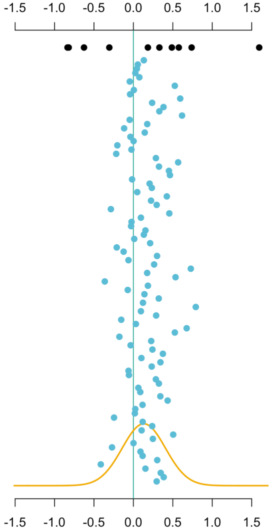


Bayesianism, statistics and calculus XXVIII

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.



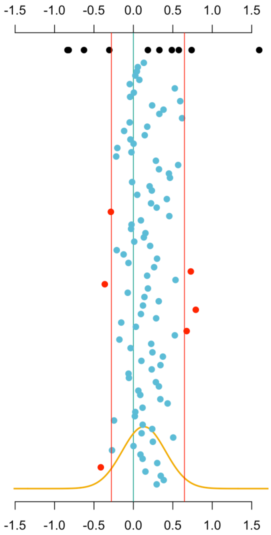
Bayesianism, statistics and calculus XXIX

- "credibility interval"

Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.
Consider

$$\begin{cases} a = \widehat{\Pi}^{-1}(\alpha/2|\mathcal{D}) \text{ quantile with level } \alpha/2 \\ b = \widehat{\Pi}^{-1}(1-\alpha/2|\mathcal{D}) \text{ quantile with level } 1-\alpha/2 \end{cases}$$



Bayesianism, statistics and calculus XXX

- "credibility interval"

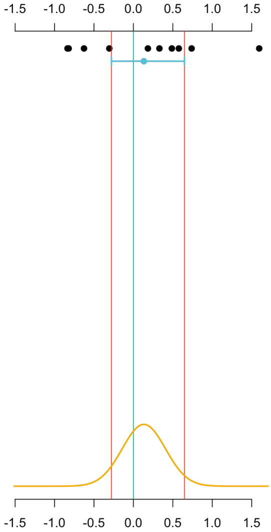
Suppose $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
Consider some prior distribution $\pi(\cdot)$ for θ
and $\pi(\cdot|\mathcal{D})$ is the posterior distribution
(potentially complicated)

Suppose we generate $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ given $\pi(\cdot|\mathcal{D})$.
Consider

$$\begin{cases} a = \widehat{\Pi}^{-1}(\alpha/2|\mathcal{D}) \text{ quantile with level } \alpha/2 \\ b = \widehat{\Pi}^{-1}(1-\alpha/2|\mathcal{D}) \text{ quantile with level } 1-\alpha/2 \end{cases}$$

then

$$\mathbb{P}\left[\theta \notin \left[\widehat{\Pi}^{-1}(\alpha/2|\mathcal{D}); \widehat{\Pi}^{-1}(1-\alpha/2|\mathcal{D})\right]\right] \approx \alpha$$



Bayesianism, statistics and calculus XXXI

We can also evoke the nonparametric Bayesian modeling, Ferguson (1973). Instead of assuming $X_i \sim f \in \mathcal{F}_\Theta$ where $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$, we consider a more general family,

$$X_i \sim f \in \mathcal{F} = \left\{ f : \int_{\mathbb{R}} [f''(y)]^2 dy < \infty \right\}$$

We can always compute a posterior law,

$$\pi(f \in A|\mathcal{D}) = \mathbb{P}(X \in A|\mathcal{D}) = \frac{\int_A \mathcal{L}_n(f) d\pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f) d\pi(f)}, \text{ where } \mathcal{L}_n(f) = \prod_{i=1}^n f(x_i)$$

where π is an a prior distribution on $\mathcal{F}$. Very close to the Pólya urn problems (infinite), to the Chinese restaurant process and to the Dirichlet processes, Blackwell and MacQueen (1973), Ghosh and Ramamoorthi (2003), Orbanz and Teh (2010).

Bayesianism, statistics and calculus XXXII

For example, if $X_1,\cdots,X_n$ i.i.d. of distribution F. The a priori law π is a Dirichlet process, $D(\alpha,F_0)$, where $F_0\in\mathcal{F}$ is a prior distribution for X, while α indicates the dispersion around F_0.
To draw according to $D(\alpha,F_0)$,

- we draw $z_1,z_2,\cdots$ according to F_0,
- we draw $v_1,v_2,\cdots$ according to a Beta law $\mathcal{B}(1,\alpha)$,
- we define iteratively weights, $\omega_1=v_1$ and $\omega_j=v_j(1-v_{j-1})\cdots(1-v_1)$
- $F(x)=\displaystyle\sum_{j\geq 1}\omega_j\mathbf{1}(x\leq z_j)$

If prior $\pi\sim D(\alpha,F_0)$, then the posterior is, $\pi|\mathcal{D}\sim D(\alpha+n,F_n)$ where

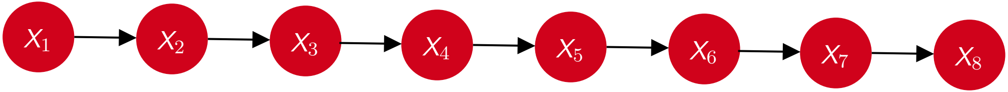
$$F_n=\frac{n}{n+\alpha}\widehat{F}_n+\frac{\alpha}{n+\alpha}F_0,\text{ where }\widehat{F}_n(x)=\frac{1}{n}\sum_{j=1}^n\mathbf{1}(x\leq x_j)$$

Bayes and Markov property I

- Markov property

This property allows to simplify the writing (and the calculation) of the posterior distribution

$$\mathbb{P}\big[X_{t+1}=x_{t+1}\big|X_t=x_t,X_{t-1}=x_{t-1},\cdots\big]=\mathbb{P}\big[X_{t+1}=x_{t+1}\big|X_t=x_t\big]$$



As a reminder, under some technical assumptions, the transition kernel $p(x_{t+1}|x_t)$ converges ($t\rightarrow\infty$) to a stationary measure $p^*(x)$.

If $x_t\in\mathcal{X}$ of finite cardinal, $p(\cdot|\cdot)$ reads in a (stochastic) matrix P.

$$\mathbb{P}\big[X_{t+k}=j\big|X_t=i\big]=[P^k]_{ij}\ \text{(Chapman Kolmogorov)}$$

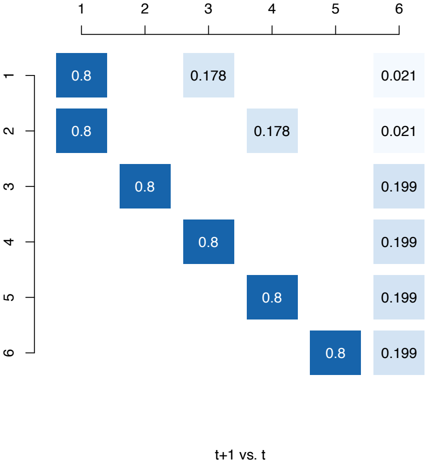
Bayes and Markov property II

Example bonus-malus schemes Lemaire (1995),

HONG KONG
Table B-9. Hong Kong System

Class	Premium	Class After 0 Claims	Class After 1 Claims	Class After ≥2 Claims
6	100	5	6	6
5	80	4	6	6
4	70	3	6	6
3	60	2	6	6
2	50	1	4	6
1	40	1	3	6

Starting class: 6.



If claims frequency is $N \sim \mathcal{P}(0.225)$,
$\mathbb{P}(N = 0) = 20\%$.

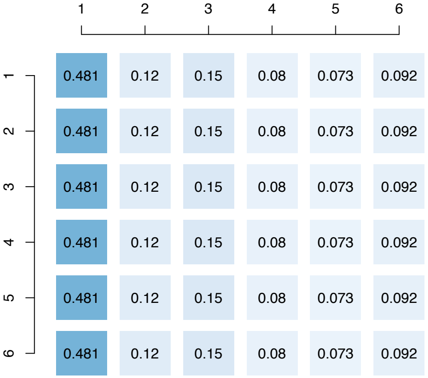
Bayes and Markov property XI

Example bonus malus schemes Lemaire (1995),

HONG KONG
Table B-9. Hong Kong System

Class	Premium	Class After 0 Claims	Class After 1 Claims	Class After ≥2 Claims
6	100	5	6	6
5	80	4	6	6
4	70	3	6	6
3	60	2	6	6
2	50	1	4	6
1	40	1	3	6

Starting class: 6.



t+100 vs. t

If claims frequency is $N \sim \mathcal{P}(0.225)$,
$\mathbb{P}(N = 0) = 20\%$.

Bayes and Markov property XII

- Expected values and MCMC

Law of large numbers

$$\text{if } X_1, \cdots, X_n, \cdots \text{ i.i.d. with law } p^*, \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}_{p^*}(X) = \int x dp^*(x)$$

Ergodic theorem (if $p(\cdot|\cdot)$ has invariant distribution p^*)

$$\text{if } X_1, \cdots, X_t, X_{t+1}, \cdots \text{ is generated from } p(\cdot|\cdot), \frac{1}{n}\sum_{t=t_0+1}^{t_0+n} X_t \xrightarrow{a.s.} \mathbb{E}_{p^*}(X) = \int x dp^*(x)$$

where (X_t) is generated from $p(\cdot|\cdot)$ using either d'Hasting-Metropolis or Gibbs sampler, Andrieu et al. (2003) or Kruschke (2014).

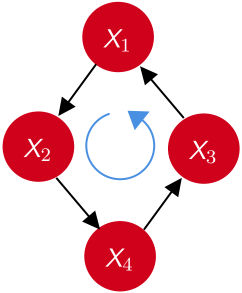
Bayes and Markov property XIII

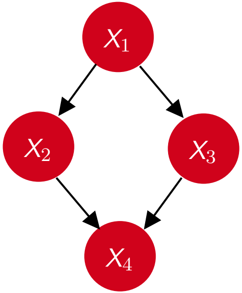
Using Markov property

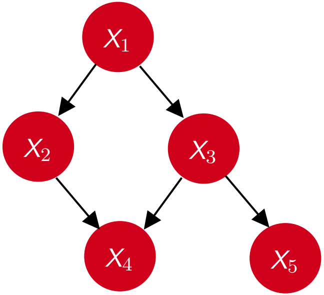
$$\mathbb{P}(\boldsymbol{x}) = \prod_{i=2}^{p} \mathbb{P}(x_i | x_{i-1}) \cdot \mathbb{P}(x_1)$$

That can be extended on a DAG for the p variables.

- Directed acyclic graph (DAG)



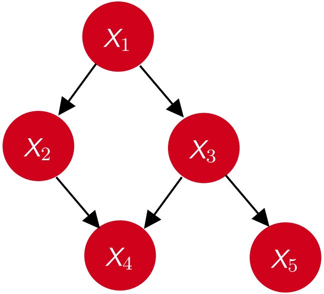




Bayes and Markov property XIV

- Bayesian Network

A couple $\{G, \mathbb{P}\}$ is a Bayesian network, if $G = \{V, E\}$ is a DAG and if it satisfies the Markov property : each variable X in V is independent from its non-descendants, in G, conditional on its parents,



$$\mathbb{P}(\boldsymbol{x}) = \prod_{i=1}^{p} \mathbb{P}(x_i | \boldsymbol{x}_{\text{parents}_i})$$

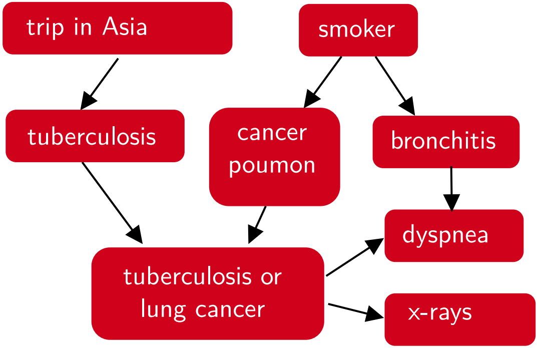
$$\begin{cases} X_2 \perp\!\!\!\perp \{X_3, X_4\} \mid X_1 \\ X_3 \perp\!\!\!\perp X_2 \mid X_1 \\ X_4 \perp\!\!\!\perp \{X_1, X_5\} \mid \{X_2, X_3\} \\ X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\} \mid X_3 \end{cases}$$

$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(x_5|x_3)\mathbb{P}(x_4|x_2, x_3)\mathbb{P}(x_3|x_1)\mathbb{P}(x_2|x_1)\mathbb{P}(x_1)$$

Bayes and Markov property XV

- Bayesian Network and Medical Diagnostics

via Lauritzen and Spiegelhalter (1988) and Højsgaard et al. (2012)



We have network (DAG)
and conditional probabilities

Bayesianism and statistical learning I

Econometrics is based on a probabilistic model, unlike most machine learning approaches, see Charpentier et al. (2018)

- in SVMs, the distance to the separation line is used as a score which can then be interpreted as a probability - Platt scaling, Platt et al. (1999) or isotonic regression Zadrozny and Elkan (2001, 2002) (see also Niculescu-Mizil and Caruana (2005) "good probabilities")
- GLM models (under additional conditions) satisfy the autocalibration property, Denuit et al. (2021), not machine learning models, i.e.

$$\mathbb{E}[Y|\widehat{Y}=y]=y,\ \forall y$$

Lichtenstein et al. (1977), Dawid (1982) or Oakes (1985), Gneiting et al. (2007)

Bayesianism and statistical learning II

As mentioned on Scikit-learn's methodological page, "*Well calibrated classifiers are probabilistic classifiers for which the output can be directly interpreted as a confidence level. For instance, a well calibrated (binary) classifier should classify the samples such that among the samples to which it gave a [predicted probability] value close to 0.8, approximately 80% actually belong to the positive class.*"

Very close to what exists to quantify uncertainty in weather models,

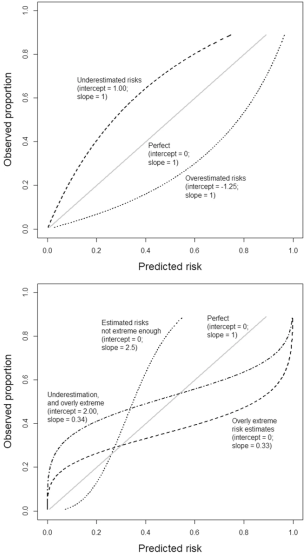
"*Suppose that a forecaster sequentially assigns probabilities to events. He is well calibrated if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent*", Dawid (1982) ou "*we desire that the estimated class probabilities are reflective of the true underlying probability of the sample*, Kuhn et al. (2013)

Bayesianism and statistical learning III

As explained in Van Calster et al. (2019), "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*",

- if 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

Hosmer-Lemeshow test (Hosmer Jr et al. (2013)) for the logistic model.



Bayesianism and statistical learning IV

- Ridge estimate, Hoerl and Kennard (1970) (linear model)

We look for $\widehat{\beta}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda\|\boldsymbol{\beta}\|_2^2\right\}$, "equivalent" to the constrained optimization problem $\underset{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_2\leq c}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})\right\}$.

Consider

$$\begin{cases}\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \text{ or } \boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{X\beta},\sigma^2\mathbb{I})\\ \boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},\tau^2\mathbb{I}) \text{ posterior}\end{cases}$$

Maximum a posteriori (MAP) satisfies

$$\widehat{\beta}_{MAP} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})+\frac{\sigma^2}{\tau^2}\|\boldsymbol{\beta}\|_2^2\right\}$$

Bayesianism and statistical learning V

- LASSO estimate, Tibshirani (1996) (linear regression)

We look for $\widehat{\beta}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda\|\boldsymbol{\beta}\|_1\right\}$, "equivalent" (Gill et al. (2019)) to the constrained optimization problem $\underset{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_1\leq c}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})\right\}$.

Consider (Tibshirani (1996) and Park and Casella (2008))

$$\begin{cases}\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \text{ ou } \boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta}\sim\mathcal{N}(\boldsymbol{X\beta},\sigma^2\mathbb{I})\\ \boldsymbol{\beta}\sim\mathcal{L}(\tau) \text{ posterior, i.e. } \pi(\boldsymbol{\beta})=(\tau/2)^p\exp\left[-\tau\|\boldsymbol{\beta}\|_1\right]\end{cases}$$

Maximum a posteriori (MAP) satisfies

$$\widehat{\beta}_{MAP}=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})+\sigma^2\tau\|\boldsymbol{\beta}\|_1\right\}$$

Bayesianism and statistical learning VI

Tibshirani (1996) suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors

- Neural nets

Rumelhart et al. (1985), Rumelhart et al. (1986) Hertz et al. (1991) and Buntine and Weigend (1991) proposed to formalize back-propagation in a Bayesian context, taken up by MacKay (1992) and Neal (1992).

State of the art in Neal (2012), more than 25 years ago (or more recently Neal (2012) Theodoridis (2015), Gal and Ghahramani (2016) and Goulet et al. (2021))

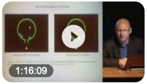
Bayesianism as a learning process I

Old topic, see
Shepard (1987) or Tenenbaum (1998).

"*How does abstract knowledge guide learning and reasoning from sparse data? How does the mind get so much from so little?*", Tenenbaum et al. (2011)

Discussed in Dehaene (2012),

www.youtube.com › watch
la révolution Bayésienne... (1) - Stanislas Dehaene (2011-2012)



Enseignement **2011-2012 : Le cerveau statisticien : la révolution Bayésienne en sciences cognitives** Cours du ma...
YouTube · Sciences de la vie - Collège de France · Il y a 1 semaine

Le cerveau statisticien : la révolution Bayésienne en sciences cognitives
Présentation

10 janvier 2012 ~ 09:30 ~
Cours
Introduction au raisonnement Bayésien et à ses applications
Stanislas Dehaene

17 janvier 2012 ~ 09:30 ~
Cours
Les mécanismes Bayésiens de l'induction chez l'enfant
Stanislas Dehaene

24 janvier 2012 ~ 09:30 ~
Cours
Les illusions visuelles : des inférences optimales ?
Stanislas Dehaene

31 janvier 2012 ~ 09:30 ~
Cours
Combinaison de contraintes et sélection d'un percept unique
Stanislas Dehaene

07 février 2012 ~ 09:30 ~
Cours
La prise de décision Bayésienne
Stanislas Dehaene

14 février 2012 ~ 09:30 ~
Cours
L'implémentation neuronale des mécanismes Bayésiens
Stanislas Dehaene

21 février 2012 ~ 09:30 ~
Cours
Le cerveau vu comme un système prédictif
Stanislas Dehaene

Bayesianism as a learning process II

The simplifications managed by the brain are known since a long time, Goodman (1955).

We have an urn containing 100 balls, a person draws a blue ball, what can we say ?
A priori not much... except if in the past, we observed that all the urns always contained balls of the same color. A single observation can then be very informative
Allows to learn how to learn, Kemp and Tenenbaum (2008), Kemp et al. (2010), Tenenbaum et al. (2011)

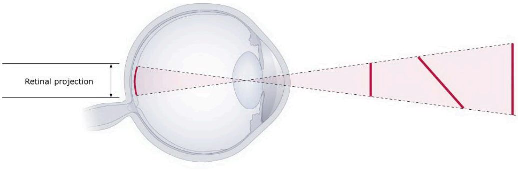
Language learning, Stolcke (1994), Watanabe and Chien (2015), Duh (2018) or Murawaki (2019).

Since Shepard (1992), many experiences on vision

Bayesianism as a learning process III

Von Helmholtz (1867) defined “unbewusste Schluss”, or unconscious inference.

The view is constructed (more or less) as a projection, but (see linear algebra course) projections are not invertible: several images could have the same projection. Our brain looks for the most likely image

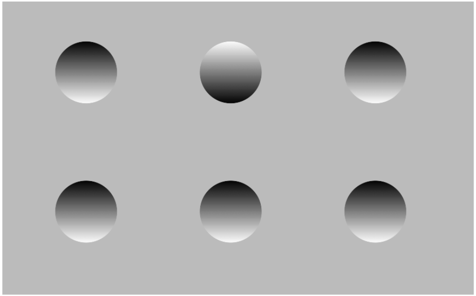


Sensory inputs are always ambiguous, so our perceptual system must select, among an infinite number of possible solutions, the one that is most plausible, Ernst and Banks (2002).

On vision as a Bayesian learning process Yuille and Kersten (2006), Clark (2013) Moreno-Bote et al. (2011)

Bayesianism as a learning process IV

Classic example on "biases" of image perception, for example the forms.

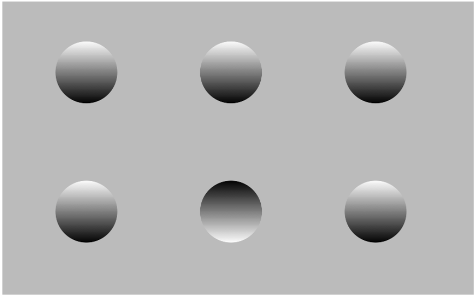


Consider the image above, what do we see?

Classically, we see 5 "holes" and 1 "bump"

Bayesianism as a learning process V

Classic example on "biases" of image perception, for example the forms.

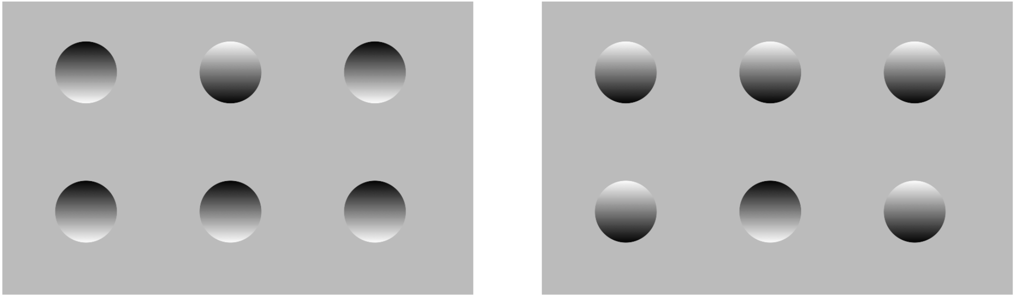


Consider the picture above, what do you see ?

Classically, 5 "bumps" et 1 "hole"

Bayesianism as a learning process VI

Classic example on "biases" of image perception, for example the forms.



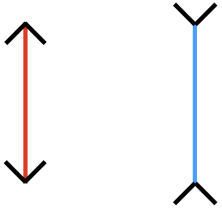
It is however the same figure (having undergone a rotation of $180°$. (grey rectangle with 6 disks with a black/white gradient). Ambiguous problem, Ramachandran (1988).

Note: our eye makes an inference about the light source (comes from above, without any other information - a priori assumption) to infer the shape.

Bayesianism as a learning process VII

Classic example on "biases" of image perception, for example the lenghts

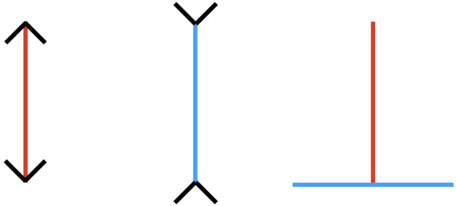
Among red and blue lines,
which one is the longuest?



Bayesianism as a learning process VIII

Classic example on "biases" of image perception, for example the lenghts

Among red and blue lines, which one is the longuest?



As mentioned by Dehaene (2012), "*Bayesian inference gives a good account of perception processes: given ambiguous inputs, our brain reconstructs the most likely interpretation*.

Bayesianism as a learning process IX

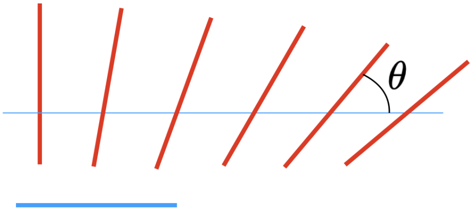
Classic example on "biases" of image perception, for example the lenghts

Among red and blue lines, which one is the longuest?

Generally, all strokes red are seen as larger than the stroke blue.

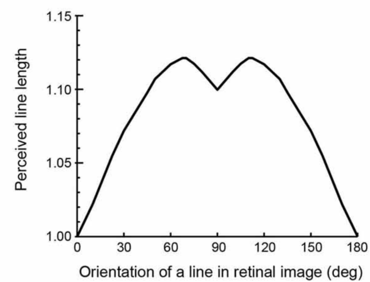
Bayesianism as a learning process X

Which of the lines red and blue is larger?



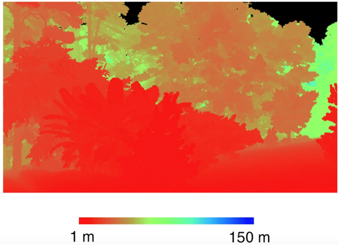
Several studies on the perception of the size of an object, according to its orientation (angle θ)

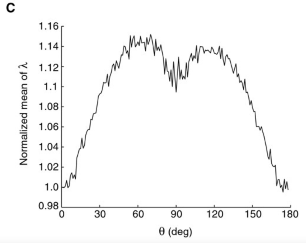
Shipley et al. (1949), Pollock and Chapanis (1952), Cormack and Cormack (1974) and Purves et al. (2008) noted that the vertical line appears 10% larger than the horizontal line.



Bayesianism as a learning process XI

The deformation made by the brain corresponds to a priori distributions that can be observed on images in nature, Howe and Purves (2002), Purves (2009), Girshick et al. (2011) or Purves et al. (2011) (based on (real) distances measured, by laser telemetry and compared to the measurement on the retina)





In other words, our retina has learned to correct the perceived distances according to the angle of inclination, in an everyday environment (3d), but continues to reproduce it for a drawing on a sheet (2d).

Bayesianism as a learning process XII

One can also learn from Ensemble methods and by aggregation of opinions. For example, guess the weight of a cow, Cornwall, England, 1906, Galton (1907).

787 participants, $x_1, \cdots, x_n$.

Unique prediction x_j v.s average $\overline{x}$,

$$\mathbb{E}\big[(x_j - t)^2\big]\big] = (\overline{x} - t)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

where t is the truth ("ambiguity decomposition").

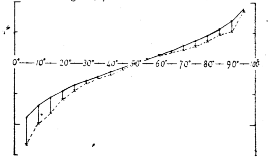
"*Bayesian methods are sometimes proposed as mathematical aggregations of expert judgements*", Hanea et al. (2021)

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Diagram, from the tabular values.



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Bayesianism as a learning process XIII

"*I have approximate answers and possible beliefs and different degrees of certainty about different things*", Feynman (2005)

"*Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise*", Surowiecki (2005)

Merrick (2008), Karvetski et al. (2013) on model aggregation $m_1, \cdots, m_k$,

$$m(\boldsymbol{x}) = \sum_{i=1}^{k} \theta_i m_i(\boldsymbol{x}, \alpha_i)$$

with weights $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$ in the simplex $\mathcal{S}_k$. We assume a prior Dirichlet distribution.

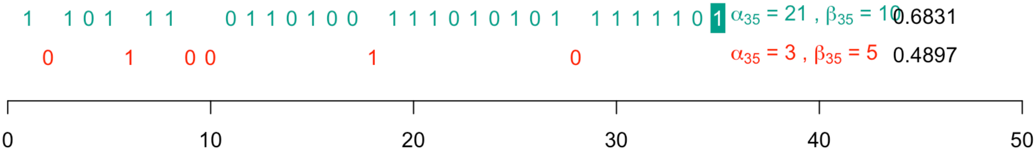
See also Mongin (1995, 2001), inspired by Karni et al. (1983).

Bayesianism as a learning process

Thompson sampling (or posterior sampling and probability matching), by Thompson (1933, 1935), and Beta-Bernoulli bandits.

We have to choose among K alternatives, that yield $\boldsymbol{X} = (X_1, \cdots, X_K)$, with $X_k \sim \mathcal{B}(\theta_k)$.
Assume (prior) $\theta_k \sim \mathcal{B}eta(\alpha_k, \beta_k)$. At time t, draw K Beta variables (independents) $B_k \sim \mathcal{B}eta(\alpha_k, \beta_k)$, and select $k^\star = \underset{k=1,\cdots,K}{\text{argmin}}\{B_k\}$.
Consider updating $(\alpha_{k^*}, \beta_{k^*}) \leftarrow (\alpha_{k^*} + x_{k^*}, \beta_{k^*} + (1 - x_{k^*}))$,

- simulated data, i.i.d., $X_1 \sim \mathcal{B}(72\%)$
- simulated data, i.i.d., $X_2 \sim \mathcal{B}(24\%)$



Bayesianism as a learning process

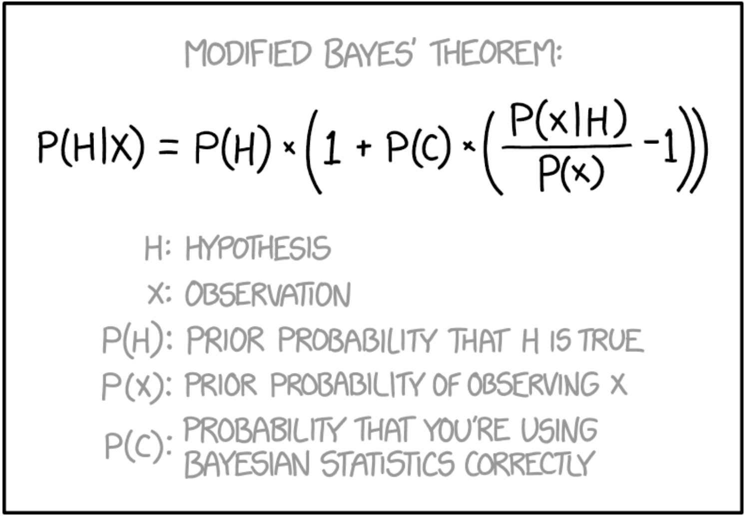
We can use that approach in the context of Monty Hall

- strategy 1 : always switch the door
- strategy 2 : never switch the door

"Conclusion" or wrap-up

- the Bayesian approach is interesting to describe beliefs in front of uncertain events, in particular if the events will occur only once
- Bayesian computation can be interpreted as a belief update or as an inverse problem
- is very strongly linked to causal graphs
- allows to take into account expert opinions, and proposes an ensemble method modeling describes both human and machine learning

"Conclusion" or wrap-up



(via https://xkcd.com/2059/)