Statistique bayésienne, data sciences, nouveaux risques (et apprentissage: une courte introduction)

Arthur Charpentier[1]

[1] Université du Québec à Montréal

Septembre 2022

Agenda

Préliminaire

Exposé en 2014, à la Cass Business School (aujourd'hui Bayes Business School)...

Getting into Bayesian Wizardry...

(with the eyes of a muggle actuary)

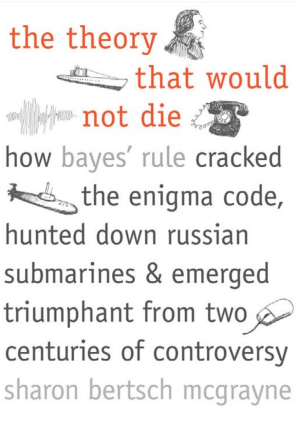
Arthur Charpentier

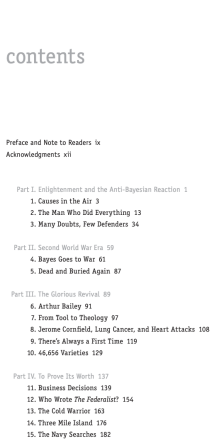
charpentier.arthur@uqam.ca

http://freakonometrics.hypotheses.org/

R in Insurance, London, July 2014

Un tout petit peu d'histoire




contents

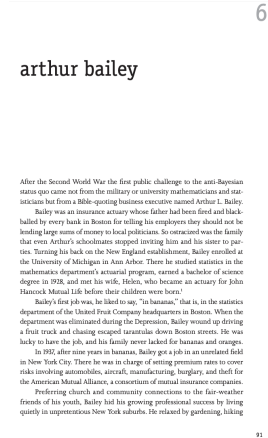
Preface and Note to Readers ix
Acknowledgments xii

Part I. Enlightenment and the Anti-Bayesian Reaction 1
1. Causes in the Air 3
2. The Man Who Did Everything 13
3. Many Doubts, Few Defenders 34

Part II. Second World War Era 59
4. Bayes Goes to War 61
5. Dead and Buried Again 87

Part III. The Glorious Revival 89
6. Arthur Bailey 91
7. From Tool to Theology 97
8. Jerome Cornfield, Lung Cancer, and Heart Attacks 108
9. There's Always a First Time 119
10. 46,656 Varieties 129

Part IV. To Prove Its Worth 137
11. Business Decisions 139
12. Who Wrote *The Federalist*? 154
13. The Cold Warrior 163
14. Three Mile Island 176
15. The Navy Searches 182


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 1932, 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.[1] 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.[2] 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."[3] 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."[4] 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.[5] 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), qui mentionne Bailey (1950) (mais pas Whitney (1918))

Incertitude, assurance et économie (un peu de provocation) III

- pour le souscripteur, $\pi \preceq X$ (prix de réservation $\geq \pi$)
  formellement, $\preceq$ est caractérisé par une utilité u et des croyances $\mathbb{Q}_S$
- pour l'assureur, $X + \sum_{i=1}^{n} X_i \leq \pi + \sum_{i=1}^{n} \pi_i$ (les primes payent les sinistres)
  formellement, cette inégalité se traduit en espérance, ou en probabilité
  ou plutôt en croyance, $\mathbb{Q}_A$, e.g. $\mathbb{Q}_A\left(X + \sum_{i=1}^{n} X_i \leq \pi + \sum_{i=1}^{n} \pi_i\right) = 90\%$

Probabilités et variables aléatoires I

"*Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means* ", Russell (1929), cité par Bell (1945) et Stevens (1951).

La base des probabilités et de la statistique est l'espace probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$,

- Ω est un espace abstrait des "états de la nature"
- $\mathcal{F}$ est l'ensemble des parties de Ω, $A \in \mathcal{F}$ étant un évènement
- $\mathbb{P}$ est une fonction $\mathcal{F} \to [0, 1]$ vérifiant quelques propriétés

e.g. $\mathbb{P}(\Omega) = 1$; pour des évènements disjoints, une propriété d'additivité: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$; une propriété d'inclusion, si $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$, que l'on retrouve chez Cardano (1564) ou Bernoulli (1713), ou, pour des évènements disjoints $A_1, \cdots, A_n, \cdots$, $\mathbb{P}(A_1 \cup \cdots \cup A_n \cup \cdots) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \cdots$, chez Kolmogorov (1933), inspiré par Lebesgue (1918), etc. Et dans ce cadre, on peut définir des variables aléatoires

- X est une fonction $\Omega \to \mathbb{R}$ ou plus généralement $\Omega \to \mathcal{X}$.

Probabilités et variables aléatoires II

On a des objets formels, mathématiquement bien définis, mais dans un contexte de modélisation a-t-on un sens univoque d'interprétation du résultat du calcul ? cf "*la probabilité est-elle inhérente à l'évènement, ou à notre jugement ?* " Martin (2009)

Il existe de nombreux paradoxes philosophiques quand on parle de probabilité (et de hasard), e.g. je lance une pièce de monnaie, qui retombe, hors de ma vue

- $\mathbb{P}(X = \text{face}) = \mathbb{P}(X = \text{pile}) = 1/2$?
- $\mathbb{P}(X = \text{face}) = 1$ ou $\mathbb{P}(X = \text{pile}) = 1$?

Ou dans un contexte juridique "*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* ", cité dans Fenton and Neil (2018).

Voir aussi Hájek (2002) sur le sens philosophique de "probabilité".

Probabilités et variables aléatoires III

Comme le dit Martin (2009),

- "*attribuer une signification objective à la probabilité qu'un évènement se réalise, c'est admettre que cet évènement n'est pas nécessaire, autrement dit, qu'il n'est pas intégralement déterminé,*"
- "*si on suppose un déterminisme intégral et universel, la probabilité ne saurait recevoir qu'une signification subjective, et la probabilité dépend de notre connaissance et de notre ignorance*"

On attribue trop d'importance à cette probabilité $\mathbb{P}$ supposée objective.

La probabilité (mathématique) n'est pas née comme un concept bien défini entrant dans le cadre d'un formalisme mathématique, mais comme un outil pour quantifier et maîtriser des situations d'incertitude, appliqué à la mesure de la probabilité de durée de vie les tables de mortalité (en vue du calcul des rentes viagères), le calcul des risques d'erreur (dans les opérations de mesure), l'étude de la probabilité des témoignages et des jugements, etc.

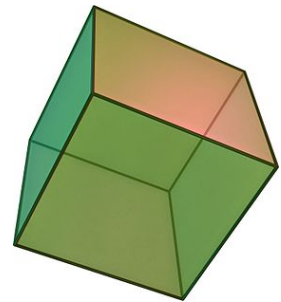
Probabilités et variables aléatoires IV

"*la théorie des probabilités 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'ils puissent souvent s'en rendre compte*", Laplace (1774)

Cournot (1843) distinguait ainsi une signification objective de la probabilité (comme mesure de la possibilité physique de réalisation d'un événement aléatoire) et une signification subjective (la probabilité étant un jugement porté sur un événement, ce jugement étant lié à l'ignorance des conditions de la réalisation de l'événement).

Remarque une probabilité non définie en termes de fréquence peut recevoir un sens objectif :

Il n'est nullement besoin répéter des lancés de dés pour affirmer que (avec un dé parfaitement équilibrée) la probabilité d'obtenir 6 lors d'un lancer est égale à $1/6$ (par symmétrie du cube)



Probabilités et variables aléatoires V

Mais bien souvent, les probabilités "physiques" ne reçoivent une valeur objective qu'a posteriori sur la base de la loi des grands nombres, la fréquence empirique convergent vers la probabilité (théorie fréquentiste des probabilités)

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \in A)}_{\text{fréquence (empirique)}} \overset{\text{p.s.}}{\to} \underbrace{\mathbb{P}(X \in A)}_{\text{probabilité}} \text{ quand } n \to \infty$$

(cf confusion dans de nombreux livres entre "probabilité" et "fréquence")

$$\text{Loi des Grands Nombres : } \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{p.s.}}{\to} \mathbb{E}(X) \text{ quand } n \to \infty \text{ ou } \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}(X)$$

Probabilités et variables aléatoires VI

Mais cette approche est incapable de donner du sens à la probabilité d'un évènement singulier unique, comme le notait 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.*"

Probabilités et variables aléatoires VII

Pour Popper (1959), les probabilités correspondent à des dispositions physiques ("propensions") inhérentes au système. Cette propension a une existence physique, mais elle n'est pas directement observable.

Les fréquences d'occurrence sont des manifestations de ces propensions. Dans le cas contraire, il est malgré tout possible d'estimer la probabilité de réalisation de l'événement singulier, en considérant celle-ci comme mesurée non par une fréquence "réelle", mais par une fréquence "potentielle" (ou "virtuelle").

Enfin, lorsqu'un individu énonce un jugement, le degré de crédibilité ou de croyance qu'il lui accorde dépend des connaissances dont cet individu dispose (Pettigrew (2016)). Ce degré de croyance sera associé à une probabilité, qui n'aura alors qu'une signification subjective. "*la probabilité d'un diagnostic, d'un témoignage, etc., ne mesure pas la conformité de ce jugement à la réalité, mais le degré avec lequel on peut faire l'hypothèse de cette conformité*", Martin (2009).

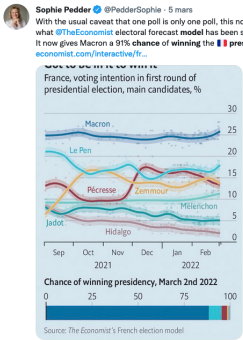
Probabilités et variables aléatoires VIII

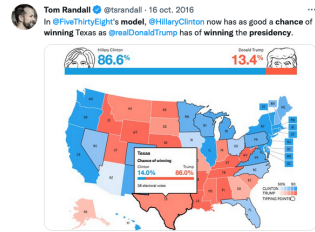
Cette subjectivité pose des soucis lors de leur utilisation, en particulier en matière crimininelle, "*Sometimes the 'balance of probability' standard is expressed mathematically as '50+% probability', but this can carry with it a danger of pseudo-mathematics, as the argument in this case demonstrated. 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* cité dans Hunt and Mostyn (2020).

On pourra aussi lire Jonakait (1983), Saini (2011) ou Fenton et al. (2016).

Probabilité ? Probabilité de gagner une élection ?

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





Comme interpréter cette “probabilité de gagner" ?

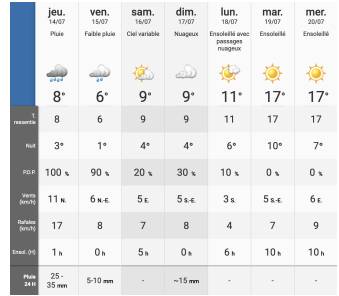
Comment interpréter un “intervalle de confiance" sur cette probabilité ? (@AdamSinger)

Probabilité ? Probabilité de pluie ? I

Comment interpréter la 'P.D.P.' ("probabilité de pluie") des sites de météo ?

	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	-	-	-

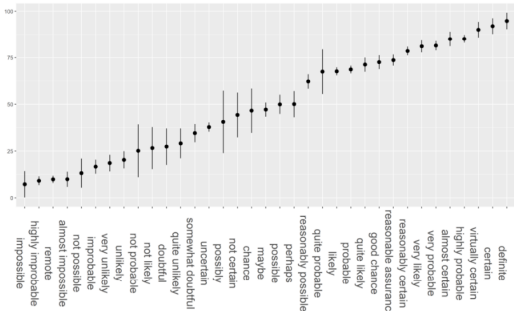
"*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated*, Silver (2012)

Murphy and Epstein (1967), Roberts (1968)

Gneiting and Raftery (2005) sur les méthodes ensemblistes pour de la prévision météo.

Probabilité ? Probabilité de pluie ? II

Plus généralement, on peut penser aux probabilités mentionnées par le GIEC/IPCC, Mastrandrea et al. (2010) discuté par Stoerk et al. (2020) ou Kause et al. (2022)



(source Vogel et al. (2022))

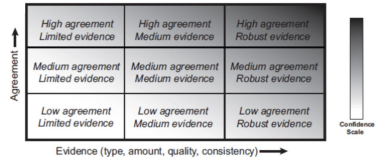
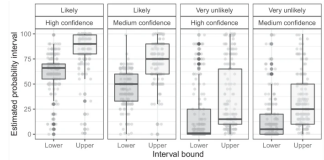


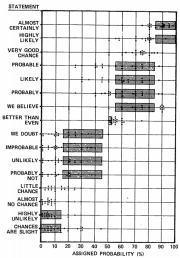
Table 1. Likelihood Scale

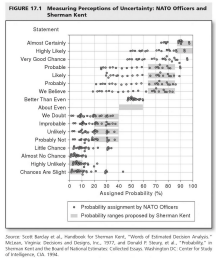
Term*	Likelihood of the Outcome
Virtually certain	99-100% probability
Very likely	90-100% probability
Likely	66-100% probability
About as likely as not	33 to 66% probability
Unlikely	0-33% probability
Very unlikely	0-10% probability
Exceptionally unlikely	0-1% probability

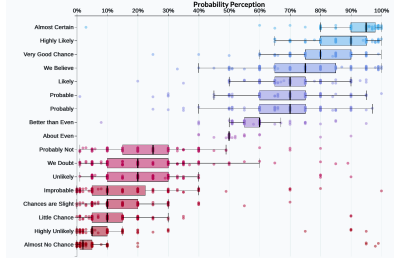


Probabilité ? Probabilité de pluie ? III

Note : “Règle de Cromwell”: il ne faut pas donner une probabilité de 1 à un événement dont la logique ne permet pas de démontrer qu’il est vrai, et il ne faut jamais donner une probabilité de 0 à un événement, sauf s’il peut être démontré logiquement qu’il est faux, Lindley (2013), Barclay et al. (1977) et Pherson and Pherson (2012).

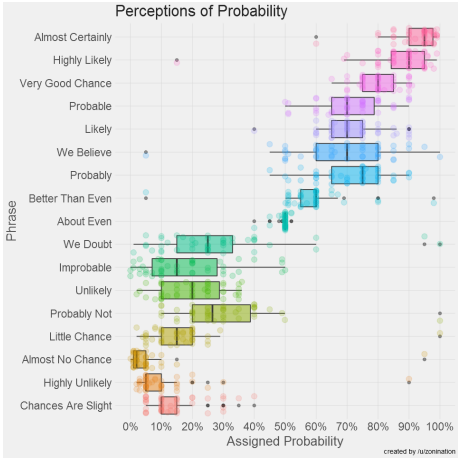


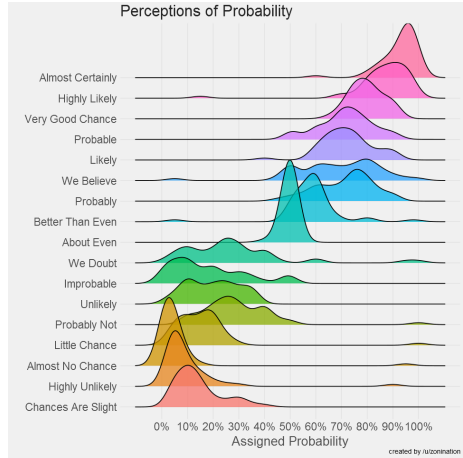




Probabilité ? Probabilité de pluie ? IV

Voir aussi @zonination sur les “perceptions des probabilités"





Statistique bayésienne ?

- Formule de Bayes (ou “probabilités inversées"), Bayes (1763), Laplace (1774)

Soient deux évènements A et B tels que $\mathbb{P}(B) \neq 0$,

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

“*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)

Statistique bayésienne ?

- Formule de Bayes,

  Bayes (1763), Laplace (1774)

Soient deux évènements A et B tels que $\mathbb{P}(B) \neq 0$,

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

- Probabilités subjectives,

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

- Approche non-fréquentiste des probabilités,

  Neyman (1977), Bayarri and Berger (2004)

- Crédibilité et "*experience rating*"

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

Statistique bayésienne ?

- Formule de Bayes,

  Bayes (1763), Laplace (1774)

Soient deux évènements A et B tels que $\mathbb{P}(B) \neq 0$,

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

- Un problème inverse (on tente de déterminer les causes d'un phénomène à partir de l'observation expérimentale de ses effets)
- Une mise à jour des croyances (passant de la loi *a priori* $\mathbb{P}(A)$ à la loi *a posteriori* $\mathbb{P}(A|B)$)

Statistique bayésienne ?

Une personne tousse (évènement B). Quelle est lhypothèse la plus crédible ?
(tiré de Dehaene (2012))

$$\begin{cases} A_1 : \text{elle a un cancer du poumon} \\ A_2 : \text{elle a une gastro-entérite} \\ A_3 : \text{elle a une grippe} \end{cases}$$

La règle de Bayes est $\mathbb{P}[\text{maladie}|\text{symptôme}] \propto \mathbb{P}[\text{symptôme}|\text{maladie}] \cdot \mathbb{P}[\text{maladie}]$

$$\begin{cases} A_1 : \mathbb{P}[\text{maladie}] \approx 0 \text{ (même si } \mathbb{P}[\text{symptôme}|\text{maladie}] \approx 1) \\ A_2 : \mathbb{P}[\text{symptôme}|\text{maladie}] \approx 0 \text{ (même si } \mathbb{P}[\text{symptôme}|\text{maladie}] \text{ grande}) \\ A_3 : \text{les deux probabilités sont raisonnables} \end{cases}$$

Probabilités conditionnelles ? En pratique...

Problème dit “de Monty Hall”
(tiré de *Let's make a deal*)

$$\mathbb{P}(\text{trésor derrière la porte}) = \frac{1}{3}$$

Probabilités conditionnelles ? En pratique...

Problème dit “de Monty Hall”
(tiré de *Let's make a deal*)

- stratégie 1 : on change toujours de porte
- stratégie 2 : on ne change jamais de porte

stratégie 2 : on garde la porte

$$\begin{aligned}
&\mathbb{P}(\text{stratégie 2 gagnante}) \\
&= \mathbb{P}(\text{trésor derrière la porte choisie initialement}) \\
&= \frac{1}{3}
\end{aligned}$$

(faire apparaître la chèvre derrière la troisième porte n'apporte aucune information sur ce qu'il y a derrière la première porte)

Probabilités conditionnelles ? En pratique...

Problème dit “de Monty Hall”
(tiré de *Let's make a deal*)

- stratégie 1 : on change toujours de porte
- stratégie 2 : on ne change jamais de porte

stratégie 1 : on change de porte

$$
\begin{aligned}
&\mathbb{P}(\text{stratégie 1 gagnante}) \\
&= \mathbb{P}(\text{trésor derrière l'autre porte}) \\
&= \mathbb{P}(\text{trésor derrière l'autre porte}|\ \text{correct}) \cdot \mathbb{P}(\ \text{correct}) \\
&+ \mathbb{P}(\text{trésor derrière l'autre porte}|\ \text{faux}) \cdot \mathbb{P}(\ \text{faux}) \\
&= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}
\end{aligned}
$$

Statistique bayésienne ? En pratique...

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

1% des adultes sont atteints d'un cancer. La grande majorité de ces cancers (90%) peuvent être détectés par un test. Il y a 9% de chances que le test s' avère positif chez une personne qui n'a pas le cancer. Si le test est positif, quelle est la probabilité que la personne ait effectivement un cancer?

A) 9 sur 10 (réponse choisie par 50% des gynécologistes)

B) 8 sur 10

C) 1 sur 2

D) 1 sur 10

E) 1 sur 100

Statistique bayésienne ? En pratique...

1% des adultes sont atteints d'un cancer. La grande majorité de ces cancers (90%) peuvent être détectés par un test. Il y a 9% de chances que le test s'avère positif chez une personne qui n'a pas le cancer. Si le test est positif, quelle est la probabilité que la personne ait effectivement un cancer?

Réponse: En formalisant

$$\begin{cases} \mathbb{P}[\text{cancer}] = 1\% \\ \mathbb{P}[\text{test positif}|\text{cancer}] = 90\% \\ \mathbb{P}[\text{test positif}|\text{pas de cancer}] = 9\% \end{cases}$$

puis en appliquant la règle de Bayes

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

la bonne réponse est D, "1 sur 10".

Statistique bayésienne ? En pratique...

Pour Gigerenzer and Hoffrage (1995), la formulation bayésienne est (trop) complexe.

Autre présentation du problème:

Sur 10,000 personnes, 100 ont un cancer. Sur ces 100, 90%, soit 90, auront un test positif. Sur les 9,900 autres, 9% soit 899 auront un test positif. Parmi un échantillon de gens qui ont un test positif, quelle fraction ont vraiment un cancer?

Réponse: 90 parmi (90+899), soit environ “1 sur 10”.

L'axiomatique des croyances I

Les axiomes de l'approche Bayésienne, Titelbaum (2022a), (2022b), sont

- étape 1 : les degrés de croyance

Les croyances sont quantifiées sur une échelle allant de 0 à 1

La "rationalité des croyances" signifie que les croyances sont des mesures de probabilités (et vérifient les axiomes associés), Buehler (1976).

Note: une version plus faible de cohérence peut être définie à l'aide des capacités (au sens de Choquet (1954)), constuite sur l'axiome : si $A \subset B$, alors $\mathbb{Q}[A] \leq \mathbb{Q}[B]$ (et non plus l'additivité d'évènements disjoints).

L'axiomatique des croyances II

- étape 2 : la mise à jour des croyances

Pour Popper (1955), un agent qui croit A au degré $\mathbb{Q}[A]$, s'il apprend B, il alors croire A au degré $\mathbb{Q}[A|B]$

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

Jeffrey (1965) a proposé une généralisation si B est associé à une croyance $\mathbb{Q}'[B]$,

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

Autrement dit, "*raisonner consiste à graduer ses croyances et à réviser ses degrés de croyance par conditionalisation bayésienne à mesure que de nouvelles informations deviennent disponibles*", Drouet (2016).

L'axiomatique des croyances 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)

"Je ne cherche pas à savoir pourquoi le fait que je prévois se réalisera, mais pourquoi je prévois que le fait va se réaliser. Ce ne sont plus les faits qui ont besoin d'une cause pour se produire : c'est notre esprit qui trouve commode d'imaginer des rapports causaux afin de les expliquer, de les coordonner et d'en rendre la prédiction possible"

Le pari hollandais I

Ramsey (1926) et De Finetti (1937) ont suggéré de comprendre la rationalité des croyances à l'aide de paris (formalisé par Lehman (1955) Kemeny (1955), Teller (1973), Lindley et al. (1979) et Skyrms (1987)) et d'arbitrages (on parle de Bayésianisme subjectif).

On attribue la croyance q à un pari (loterie) associé à A, rapportant a si A survient et 0 sinon si et seulement si la valeur de la loterie est qa, Hájek (2009)

L'argument du "pari hollandais" (dutch book) est que si un individu a des croyances qui violent les probabilités et que s'il parie sur la base de ces croyances, alors il est prêt à accepter un ensemble de paris dont il est certain de repartir perdant, Pettigrew (2020).

Note: Lehman (1955) a utilisé le terme "dutch book", mais il correspond à la notion d'arbitrage en mathématiques financières.

Le pari hollandais 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*"

Cette caractérisation est aussi appelé théorème de Cox-Jaynes, Cox (1946) repris par Jaynes (1988) et Jaynes (2003) : les probabilités (caractérisées par les axiomes de Kolmogorov) sont le seul mécanisme normatif de l'induction de la plausibilité.

Mais aussi chez Good (1966)

Voir aussi Eisenberg and Gale (1959) et Baron and Lange (2006) sur les notions de parimutuel, et les nouveaux risques (voir aussi Chen and Pennock (2010), Charpentier (2017), Charpentier (2019) sur les marchés prédictifs).

Supposons que I joueurs parient sur J chevaux. Chaque joueur possède une somme b_i, et on normalisera ($b_1 + \cdots + b_I = 1$).

Le pari hollandais III

Le joueur i mise $\beta_{i,j}$ sur le cheval j ($b_i = \beta_{i,1} + \cdots + b_{i,J}$).
On note π_j le montant parié sur le cheval j ($\pi_j = \beta_{1,j} + \cdots + b_{I,j}$).
Comme $\pi_j \in (0, 1)$ et $\pi_1 + \cdots + \pi_J = 1$ est interprété comme une probabilité, décrivant une "croyance collective".

On peut aussi ajouter des contraintes empiriques, et associer les croyances à des fréquences connues) (on parle de Bayésianisme empirique)

Williamson (2004) a introduit un Bayésianisme objectif, inspiré par Jaynes (1957), basé sur la maximisation de l'entropie (approche maxmin), associé à un principe de précaution.

Logique non booléenne I

Note On peut aussi trouver des liens avec la logique.
Classiquement, si on a la proposition "Si A est vrai, alors B est vrai"

$$\begin{cases} \text{Si j'observe que } A \text{ est vrai, j'en conclus que } B \text{ est vrai} \\ \text{Si j'observe que } B \text{ est faux, j'en conclus que } A \text{ est faux.} \end{cases}$$

Dans le cadre de la logique booléenne, ce sont les seules assertions équivalentes
($A \Longrightarrow B$ et $\neg B \Longrightarrow \neg A$).
Mais il peut y avoir des raisonnements plausibles, Pólya (1958)

$$\begin{cases} \text{Si j'observe que } A \text{ est faux, il me semble que } B \text{ devient moins plausible} \\ \text{Si j'observe que } B \text{ est vrai, il me semble que } A \text{ devient plus plausible.} \end{cases}$$

Que signifie ici "plausible" ? (On reviendra sur ces croyances en neuroscience plus tard)

Bayésianisme, statistique et calcul I

$$\text{a posteriori } = \pi(\theta|\boldsymbol{y}) = \frac{\pi(\theta)\cdot\mathbb{P}(\boldsymbol{y}|\theta)}{\mathbb{P}(\boldsymbol{y})} = \frac{\text{a priori}\cdot\text{vraisemblance}}{\text{évidence}}$$

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

- Lois conjuguées : **Binomiale - Beta**

La vraisemblance pour un modèle de variables binomiales est

$$\begin{cases} \boldsymbol{x} \mapsto f(\boldsymbol{x};p) = p^s(1-p)^{n-s} \text{ où } s = \boldsymbol{x}^\top\mathbf{1} = x_1 + \cdots + x_n \\ p \mapsto p^s(1-p)^{n-s} \text{ sur } [0,1] \text{ est une loi Beta} \end{cases}$$

$$\text{Si } \begin{cases} x_i|\theta \sim \mathcal{B}(\theta) \\ \theta \sim \mathcal{B}eta(a,b) \text{ a priori} \end{cases} \qquad \text{alors } \theta|\boldsymbol{x} \sim \mathcal{B}eta(a+s, b+n-s) \text{ a posteriori}$$

(qui se généralise en **Multinomiale - Dirichlet**)

Bayésianisme, statistique et calcul II

- Lois conjuguées : **Poisson - Gamma**

La vraisemblance pour un modèle de variables Poisson est

$$\begin{cases} \boldsymbol{x} \mapsto f(\boldsymbol{x};\lambda) = \dfrac{e^{n\lambda}\lambda^s}{x_1!\cdots x_n!} \text{ où } s = \boldsymbol{x}^\top \mathbf{1} = x_1 + \cdots + x_n \\ \lambda \mapsto e^{n\lambda}\lambda^s \text{ sur } \mathbb{R}_+ \text{ est une loi Gamma} \end{cases}$$

Si

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

On retrouve en particulier

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

utilisé en théorie de la crébilité, Bühlmann (1967).

- Lois conjuguées : **Normale - Normale**

Si la variance $\boldsymbol{\Sigma}$ est connue

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

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

Utilisé classiquement en économétrie Bayésienne,

Bayésianisme, statistique et calcul IV

- Lois conjuguées : **Normale - Inverse Wishart**

Si la moyenne $\boldsymbol{\mu}$ est connue

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

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

Utilisé classiquement en économétrie Bayésienne, pour des modèles VAR, Adjemian and Pelgrin (2008), ou en gestion de portefeuilles, Black and Litterman (1990, 1992) (voir aussi Satchell and Scowcroft (2000) pour une mise en perspective).

Bayésianisme, statistique et calcul V

Les méthodes bayésiennes peuvent être très puissantes pour estimer des modèles de panels, hiérarchiques ou multi-niveaux, Gelman and Hill (2006).

- **Modèle hiérarchique**

Quand l'individu i appartient au groupe j,

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

avec des constantes et des pentes qui dépendent des groupes.

(généralement dans un modèle GLM).

Sinon, soit on utilise des simulations (voir MCMC), soit on passe par des hypothèses simplificatrices.

Considèrons des symptomes $s_1, \cdots, s_k$ et des maladies $m_1, \cdots, m_j$ (dans $\{0,1\}$)

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

"Naïve Bayes" repose sur des hypothèses (Spiegelhalter et al. (1993))

- les maladies sont mutuellement exclusives $\mathbb{P}\big[\boldsymbol{M}=\boldsymbol{m}\big|\boldsymbol{S}=\boldsymbol{s}\big]=0$ si $\boldsymbol{m}^\top\mathbf{1}>1$,
- les symptomes sont conditionnellement indépendants

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

Et dans ce cas,

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

On peut améliorer le modèle à l'aide d'un réseau bayésien (on en reparlera plus loin).

Bayésianisme, statistique et calcul VIII

Pour déterminer $\mathbb{P}\big[M_i = m_i \big| \boldsymbol{S} = \boldsymbol{s}\big]$, on a besoin de connaître

- la prévalence des maladies $\mathbb{P}\big[M_i = 1\big]$
- les sensibilités $\mathbb{P}\big[S_j = 1 | M_i = 1\big]$
- les spécificités $\mathbb{P}\big[S_j = 0 | M_i = 0\big]$

pour tous les symptomes S_j et toutes les maladies M_i.

Notons que $\mathbb{P}\big[S_j = s_j | M_i = m_i\big]$ ont une interprétation causale: ce sont les maladies qui causent les symptomes.

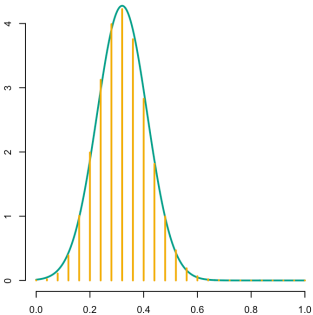
Voir Sadegh-Zadeh (1980) sur les diagnostiques bayésiens, ou Donnat et al. (2020).

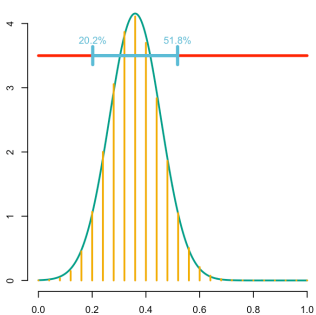
Bayésianisme, statistique et calcul I

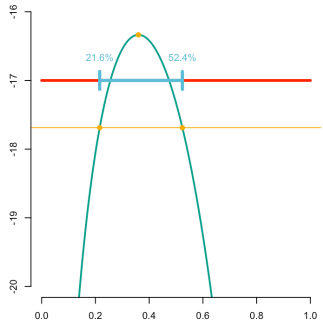
- Loi a posteriori

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

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





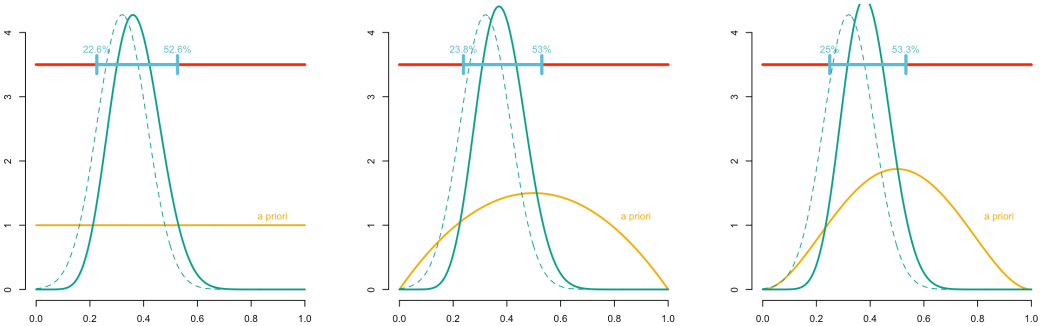


Bayésianisme, statistique et calcul II

- Loi a posteriori

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

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

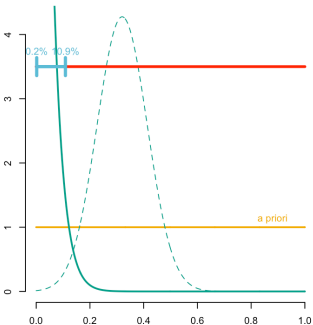


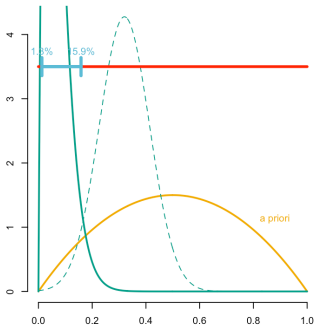
Bayésianisme, statistique et calcul XIX

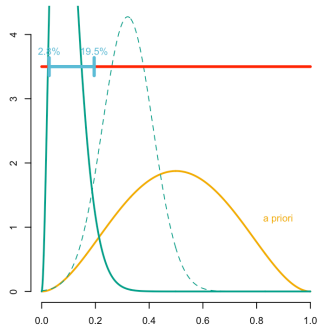
- Loi a posteriori

Et si $\boldsymbol{x} = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$, $\mathcal{B}(\theta)$?

Approche Bayésienne, $\widehat{\theta}|\boldsymbol{x} \sim \mathcal{B}eta(\alpha_0, \beta_0 + n)$, comme $\sum_{i=1}^{n} x_i = 0$



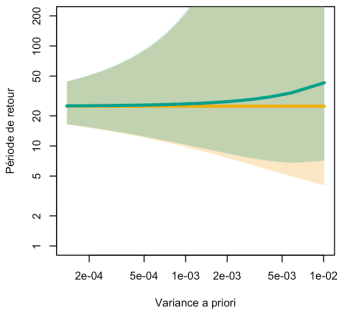


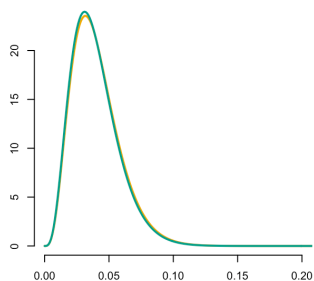


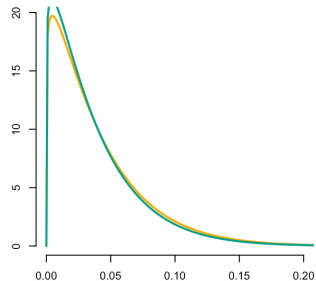
Bayésianisme, statistique et calcul XX

- Loi a posteriori

Ministère de l'intérieur (2019) "*seuil unique pour qualifier une sécheresse géotechnique d'anormale: une durée de retour supérieure ou égale à 25 ans* " (probabilité $1/25$) On n'a observé aucune sécheresse sur 2 ans ($\{0, 0\}$), que devient notre croyance sur la période de retour ?

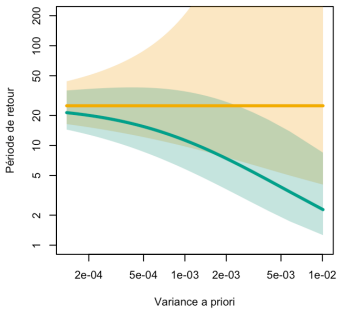


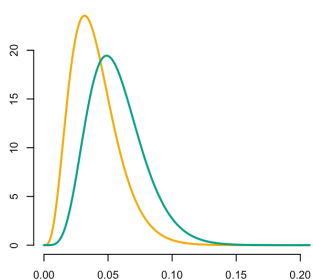


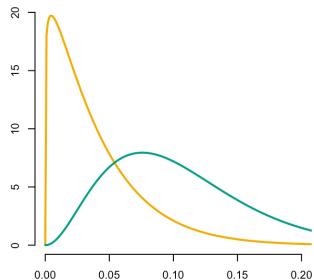


- Loi a posteriori

A titre de comparaison, si on a observé deux sécheresses majeures ($\{1, 1\}$), nos croyances a posteropri sont très influencées par ces évènements inatendus







Bayésianisme, statistique et calcul XXII

- De la loi à l'estimateur

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

La moyenne a posteriori est aussi la solution du problème

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

- "Intervalle de confiance" ou plutôt "intervalle de crédibilité"

Pour l'intervalle de confiance, on cherche $[\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]$ tel que $\mathbb{P}[\theta \in [\widehat{a}_{\mathcal{D}}, \widehat{b}_{\mathcal{D}}]] \geq 95\%$.

Pour l'intervalle de crédibilité, on cherche $[a, b]$ tel que $\mathbb{P}[\theta \in [a, b]|\mathcal{D}] \geq 95\%$.

Bayésianisme, statistique et calcul XXV

- "Intervalle de confiance"

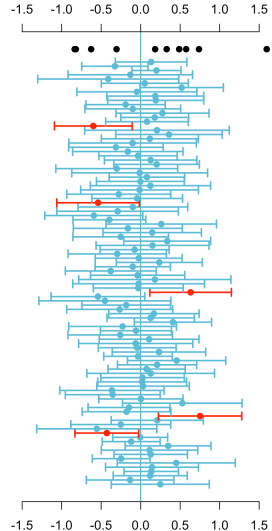
Supposons que $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
(ici $\theta = 0$)

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

Si on génère $\mathcal{D}' = \{x'_1, \cdots, x'_n\}$ suivant $\mathcal{N}(\theta, \sigma^2)$ on veut

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

interprété comme une fréquence, en répétant l'expérience.
Ici, $\alpha = 5\%$: dans 5% des simulations, 0 n'est pas dans $[a, b]$.

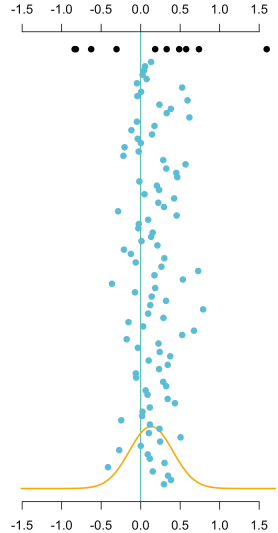


Bayésianisme, statistique et calcul XXVIII

- "Intervalle de crédibilité"

Supposons que $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
On se donne une loi a priori $\pi(\cdot)$ pour θ
et $\pi(\cdot|\mathcal{D})$ est la loi a posteriori (potentiellement compliquée)

On suppose qu'on peut générer $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ suivant $\pi(\cdot|\mathcal{D})$.



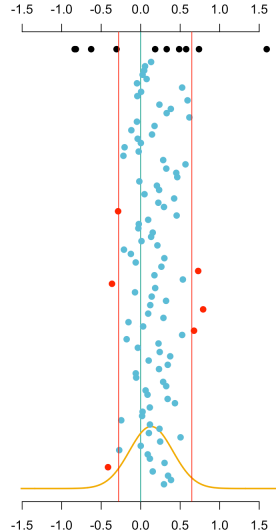
Bayésianisme, statistique et calcul XXIX

- "Intervalle de crédibilité"

Supposons que $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
On se donne une loi a priori $\pi(\cdot)$ pour θ
et $\pi(\cdot|\mathcal{D})$ est la loi a posteriori (potentiellement compliquée)

On suppose qu'on peut générer $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ suivant $\pi(\cdot|\mathcal{D})$.
On considère

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



Bayésianisme, statistique et calcul XXX

- "Intervalle de crédibilité"

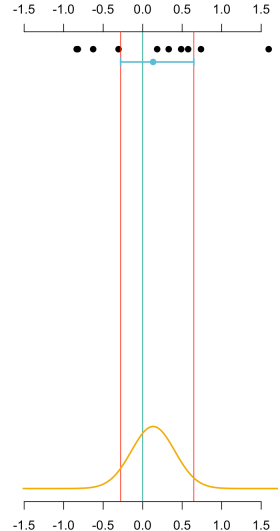
Supposons que $\mathcal{D} = \{x_1, \cdots, x_n\}$, $X_i \sim \mathcal{N}(\theta, \sigma^2)$
On se donne une loi a priori $\pi(\cdot)$ pour θ
et $\pi(\cdot|\mathcal{D})$ est la loi a posteriori (potentiellement compliquée)

On suppose qu'on peut générer $\tilde{\theta}_1, \cdots, \tilde{\theta}_k$ suivant $\pi(\cdot|\mathcal{D})$.
On considère

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

alors

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



Bayésianisme, statistique et calcul XXXI

On peut aussi évoquer de la modélisation Bayésienne nonparamétrique, Ferguson (1973). Au lieu de supposer $X_i \sim f \in \mathcal{F}_\Theta$ où $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$, on considère une famille plus générale,

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

On peut toujours calculer une loi a posteriori,

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

où π est une loi a priori sur $\mathcal{F}$. Très proche des problèmes d'urnes de Pólya (infinies), du "Chinese restaurant process" et des processus de Dirichlet, Blackwell and MacQueen (1973), Ghosh and Ramamoorthi (2003), Orbanz and Teh (2010).

Bayésianisme, statistique et calcul XXXII

Par exemple, si $X_1, \cdots, X_n$ i.i.d. de loi F. La loi a priori π est un processus de Dirichlet, $D(\alpha, F_0)$, où $F_0 \in \mathcal{F}$ est une loi a priori pour X, alors que α indique la dispersion autour de F_0.
Pour tirer suivant $D(\alpha, F_0)$,

- on tire $z_1, z_2, \cdots$ suivant F_0,
- on tire $v_1, v_2, \cdots$ suivant une loi Beta $\mathcal{B}(1, \alpha)$,
- on définit itérativement des poids, $\omega_1 = v_1$ et $\omega_j = v_j(1 - v_{j-1}) \cdots (1 - v_1)$
- $F(x) = \sum_{j \geq 1} \omega_j \mathbf{1}(x \leq z_j)$

Si a priori $\pi \sim D(\alpha, F_0)$, alors a posteriori, $\pi | \mathcal{D} \sim D(\alpha + n, F_n)$ où

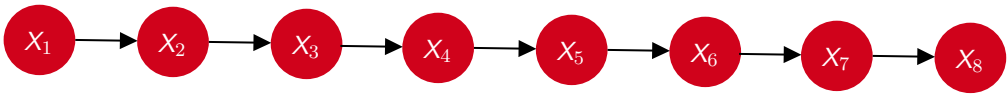
$$F_n = \frac{n}{n+\alpha} \widehat{F}_n + \frac{\alpha}{n+\alpha} F_0, \text{ où } \widehat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(x \leq x_j)$$

Bayes et propriété de Markov I

- Propriété de Markov

Cette propriété permet de simplifier l'écriture (et le calcul) de la loi a posteriori

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



Pour rappels, moyennant quelques hypothèses techniques, le noyau de transition $p(x_{t+1}|x_t)$ converge ($t \to \infty$) vers une mesure stationnaire $p^*(x)$.

Si $x_t \in \mathcal{X}$ de cardinal fini, $p(\cdot|\cdot)$ se lit dans une matrice (stochastique) P.

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

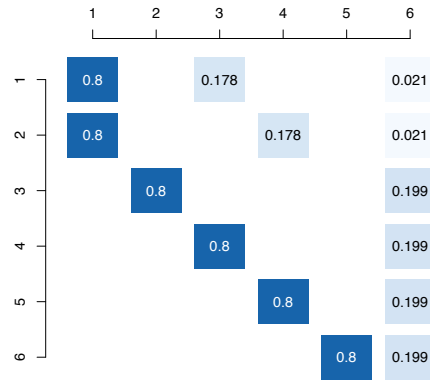
Bayes et propriété de Markov II

Exemple système bonus-malus, 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.



Si la fréquence de sinistre est $N \sim \mathcal{P}(0.225)$, $\mathbb{P}(N = 0) = 20\%$.

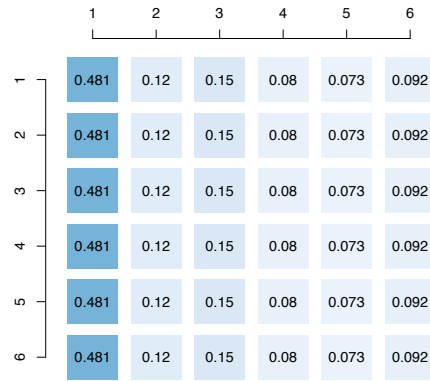
Bayes et propriété de Markov XI

Exemple système bonus-malus, Lemaire (1995),

HONG KONG
Table B-9. Hong Kong System

Class	Premium	Class After 1 Claims: 0	1	≥2
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.



Si la fréquence de sinistre est $N \sim \mathcal{P}(0.225)$, $\mathbb{P}(N = 0) = 20\%$.

Bayes et propriété de Markov XII

- Calcul d'espérance et MCMC

Loi des grands nombres

$$\text{si } X_1, \cdots, X_n, \cdots \text{ i.i.d. de loi } p^*, \frac{1}{n}\sum_{i=1}^{n} X_i \overset{p.s.}{\to} \mathbb{E}_{p^*}(X) = \int x dp^*(x)$$

Théorème ergodique (si $p(\cdot|\cdot)$ a pour loi invariante p^*)

$$\text{si } X_1, \cdots, X_t, X_{t+1}, \cdots \text{ est généré suivant } p(\cdot|\cdot), \frac{1}{n}\sum_{t=t_0+1}^{t_0+n} X_t \overset{p.s.}{\to} \mathbb{E}_{p^*}(X) = \int x dp^*(x)$$

et générer (X_t) suivant $p(\cdot|\cdot)$ peut se faire par l'algorithme d'Hasting-Metropolis ou de Gibbs, Andrieu et al. (2003) ou Kruschke (2014).

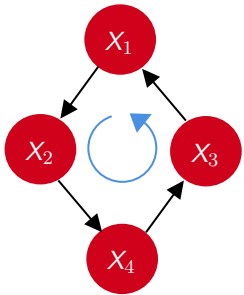
Bayes et propriété de Markov XIII

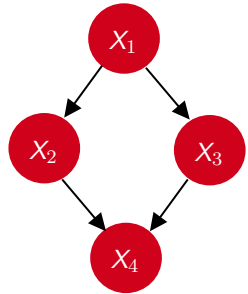
La propriété de Markov permet d'écrire

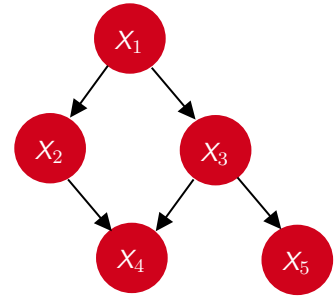
$$\mathbb{P}(\boldsymbol{x}) = \prod_{i=2}^{p} \mathbb{P}(x_i | x_{i-1}) \cdot \mathbb{P}(x_1)$$

On peut étendre cette propriété quand il existe un DAG pour les p variables.

- Graph dirigé acyclique (DAG)



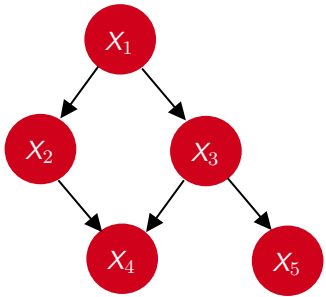




Bayes et propriété de Markov XIV

- Réseau Bayésien

Un couple $\{G, \mathbb{P}\}$ est un Réseau Bayésien, si $G = \{V, E\}$ est un DAG et s'il vérifie la condition de Markov suivante : chaque variable X dans V est indépendante de ses non-descendants, dans G, conditionnellement à ses parents,



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

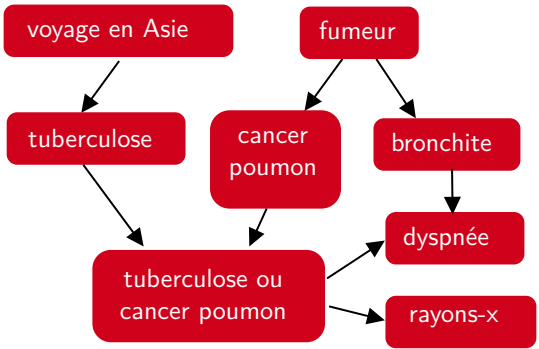
$$\begin{cases} X_2 \perp\!\!\!\perp \{X_3, X_4\} \mid X_1 \\ X_3 \perp\!\!\!\perp X_2 \mid X_1 \\ X_4 \perp\!\!\!\perp \{X_1, X_5\} \mid \{X_2, X_3\} \\ X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\} \mid X_3 \end{cases}$$

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

Bayes et propriété de Markov XV

- Réseau Bayésien et Diagnostique Médical

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



On a le graph (orienté acyclique, DAG) et les probabilités conditionnelle

Bayésianisme et apprentissage statistique I

L'économétrie est fondée sur un modèle probabiliste, contrairement à la plupart des approches de machine learning, cf Charpentier et al. (2018)

- dans les SVM, la distance à la droite de séparation est utilisé comme un score qui peut être ensuite interprété comme une probabilité - Platt scaling, Platt et al. (1999) ou isotonic regression Zadrozny and Elkan (2001, 2002) (voir aussi Niculescu-Mizil and Caruana (2005) "good probabilities")
- les modèles GLM (sous certaines conditions) vérifient la propriété d'autocalibration, Denuit et al. (2021), pas les modèles d'apprentissage machine, i.e.

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

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

Bayésianisme et apprentissage statistique II

Comme l'explique la page méthodologique de Scikit-learn, "*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.*"

Très proche de ce qui existe pour quantifier l'incertitude dans les modèles météorologiques,

"*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)

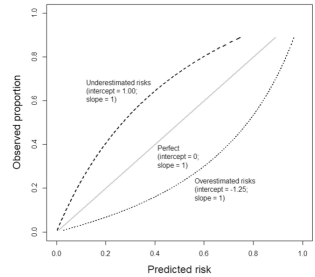
Bayésianisme et apprentissage statistique III

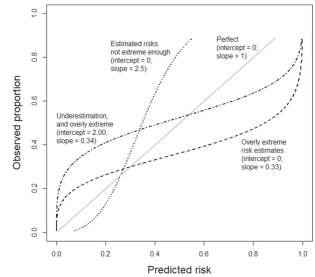
Comme l'explique Van Calster et al. (2019), "*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*",

- si on observe que dans ce groupe, 40 sur 100 ont la maladie, on a sous-estimé le risque
- si on observe que dans ce groupe, 10 sur 100 ont la maladie, on a sur-estimé le risque

On peut sur-sous estimer différement pour les petits ou les grands risques.

Test de Hosmer-Lemeshow test (Hosmer Jr et al. (2013)) dans le cas logistique (à généraliser).





Bayésianisme et apprentissage statistique IV

- Estimateur Ridge, Hoerl and Kennard (1970) (régression linéaire)

On cherche $\widehat{\beta}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2 \right\}$, "équivalent" au problème d'optimisation sous contrainte $\underset{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_2\leq c}{\text{argmin}} \left\{ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \right\}$.

On va considérer le modèle suivant

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

L'estimateur Maximum a Posteriori (MAP) vérifie

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

Bayésianisme et apprentissage statistique V

- Estimateur LASSO, Tibshirani (1996) (régression linéaire)

On cherche $\widehat{\beta}_\lambda = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \right\}$, "équivalent" (Gill et al. (2019)) au problème d'optimisation contraint $\underset{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_1\leq c}{\text{argmin}} \left\{ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \right\}$.

On va considérer le modèle suivant (Tibshirani (1996) et Park and Casella (2008))

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

L'estimateur Maximum a Posteriori (MAP) vérifie

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

Bayésianisme et apprentissage statistique 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

- Réseaux de neurones

Rumelhart et al. (1985), Rumelhart et al. (1986) Hertz et al. (1991) et Buntine and Weigend (1991) ont proposé de formaliser la rétro-propagation dans un contexte bayésien, repris par MacKay (1992) et Neal (1992).

Synthèse de Neal (2012), il y a plus de 25 ans (ou plus récemment Neal (2012) Theodoridis (2015), Gal and Ghahramani (2016) et Goulet et al. (2021))

Le Bayésianisme comme processus d'apprentissage I

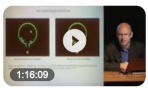
Vieux sujet,
Shepard (1987) ou Tenenbaum (1998).

"*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)

Repris par Dehaene (2012) dans son cours au Collège de France.

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

Le Bayésianisme comme processus d'apprentissage II

Les simplifications oppérées par le cerveau sont connu depuis longtemps, Goodman (1955).

On a une urne contenant 100 boules, une personne tire une boule bleue, que peut-on dire ?
A priori pas grande chose sauf si par le passé, on a observé que toutes les urnes contenaient toujours des boules de la même couleur. Une observation unique peut alors être très informative
Permet dapprendre à apprendre, Kemp and Tenenbaum (2008), Kemp et al. (2010), Tenenbaum et al. (2011)

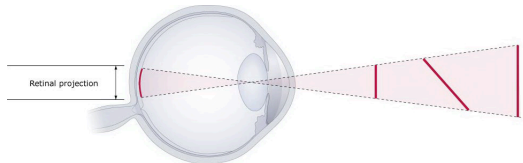
Apprentissage du langage, Stolcke (1994), Watanabe and Chien (2015), Duh (2018) ou Murawaki (2019).

Depuis Shepard (1992), de nombreuses expériences sur la vision.

Le Bayésianisme comme processus d'apprentissage III

Von Helmholtz (1867) parlait de "unbewusste Schluss", ou inférence inconsciente.

La vue est construite (plus ou moins) comme une projection, or (cf cours algèbre linéaire) les projections ne sont pas inversibles: plusieurs images pourraient avoir la même projection. Notre cerveau cherche l'image la plus vraisemblable

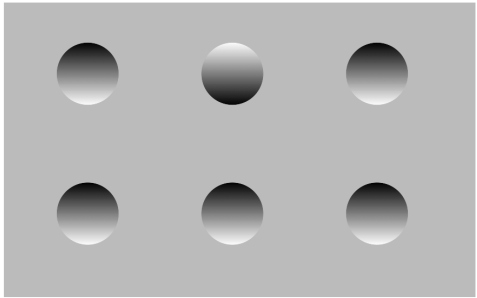


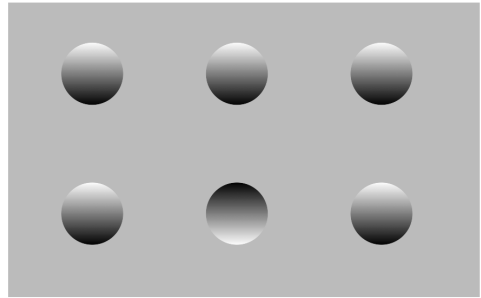
Les entrées sensorielles sont toujours ambigües, notre système perceptif doit donc sélectionner, parmi une infinité de solutions possibles, celle qui est la plus plausible, Ernst and Banks (2002).

Sur la vision comme processus Bayésien d'apprentissage Yuille and Kersten (2006), Clark (2013) Moreno-Bote et al. (2011)

Le Bayésianisme comme processus d'apprentissage VI

Exemple classique sur des "biais" de perception des images, par exemple les formes.





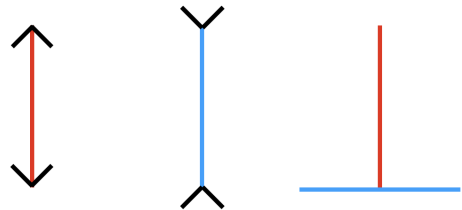
Il s'agir pourtant de la même figure (ayant subie une rotation de $180°$. (rectangle gric avec 6 disques avec un gradient noir/blanc). Problème ambigu, Ramachandran (1988).

Note: notre oeil fait une inférence sur la source de lumière (vient d'en haut, sans aucune autre information - hypothèse a priori) pour inférer la forme.

Le Bayésianisme comme processus d'apprentissage VIII

Exemple classique sur des "biais" de perception, par exemple les longueurs.

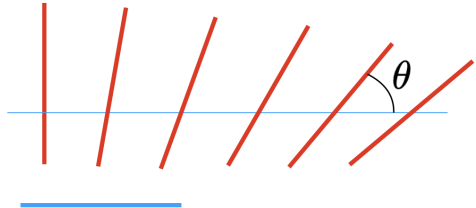
Parmi les traits rouge et bleu,
lequel est le plus grand ?



Comme le souligne Dehaene (2012), "*linférence Bayésienne rend bien compte des processus de perception: étant donné des entrées ambigües, notre cerveau en reconstruit linterprétation la plus probable*.

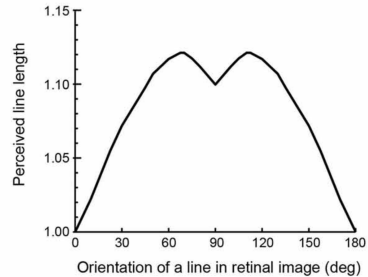
Le Bayésianisme comme processus d'apprentissage X

Parmi les traits rouge et bleu, lequel est le plus grand ?



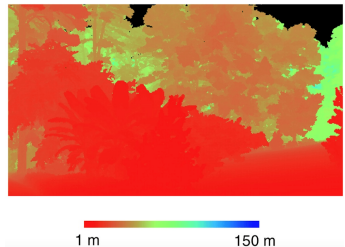
Plusieurs études sur la perception de la taille dun objet, en fonction de son orientation (angle θ)

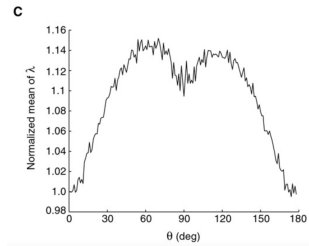
Shipley et al. (1949), Pollock and Chapanis (1952), Cormack and Cormack (1974) et Purves et al. (2008) a noté que le trait vertical semble 10% plus grand que le trait horizontal.



Le Bayésianisme comme processus d'apprentissage XI

La déformation faite par le cerveau correspond à des distributions a priori que lon peut observer sur des images dans la nature, Howe and Purves (2002), Purves (2009), Girshick et al. (2011) ou Purves et al. (2011) (sur la base de distances (réelles) mesurés, par télémétrie laser et comparées au mesure sur la rétine)





Autrement dit, notre rétine a appris à corriger les distances perçues en fonction de l'angle d'inclinaison, dans un environnement quotidien (3d), mais continue à le reproduire pour un dessin sur une feuille (2d).

Le Bayésianisme comme processus d'apprentissage XII

On peut aussi apprendre des Méthodes d'ensemble et par aggrégation d'avis. Par exemple, deviner le poids d'une vache, Cornwall, England, 1906, Galton (1907).

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

Prédiction unique x_j v.s moyenne $\overline{x}$,

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

où t est la vérité ("ambiguity decomposition").

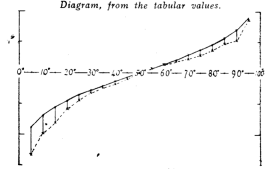
"*Bayesian methods are sometimes proposed as mathematical aggregations of expert judgements*", Hanea et al. (2021)

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	− 40	+ 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	+ 40	− 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.

Le Bayésianisme comme processus d'apprentissage 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) sur de l'agrégation de modèles $m_1, \cdots, m_k$,

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

avec des poids $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$ dans le simplexe $\mathcal{S}_k$. On suppose a priori une distribution de Dirichlet.

Voir aussi Mongin (1995, 2001), inspiré de Karni et al. (1983).

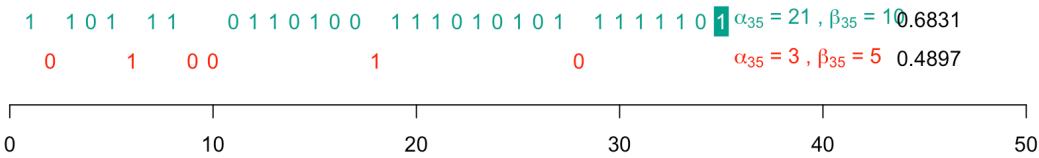
Le Bayésianisme comme processus d'apprentissage

Thompson sampling (ou posterior sampling et probability matching), par Thompson (1933, 1935), et les bandits Beta-Bernoulli.

On a le choix entre K alternatives, rapportant $\boldsymbol{X} = (X_1, \cdots, X_K)$, avec $X_k \sim \mathcal{B}(\theta_k)$. A priori, on suppose que $\theta_k \sim \mathcal{B}eta(\alpha_k, \beta_k)$. A la date t, on tire K variables Beta (indépendantes) $B_k \sim \mathcal{B}eta(\alpha_k, \beta_k)$, et on retient $k^\star = \underset{k=1,\cdots,K}{\text{argmin}} \{B_k\}$.

On fait alors la mise à jour $(\alpha_{k^*}, \beta_{k^*}) \leftarrow (\alpha_{k^*} + x_{k^*}, \beta_{k^*} + (1 - x_{k^*}))$,

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

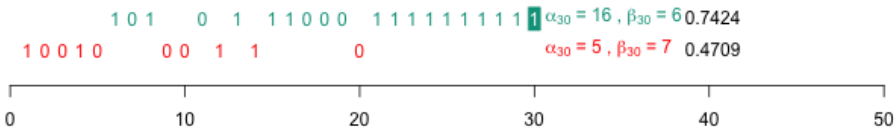


Le Bayésianisme comme processus d'apprentissage

On peut utiliser cette approche dans le contexte du problème de Monty Hall

- stratégie 1 : on change toujours de porte
- stratégie 2 : on ne change jamais de porte

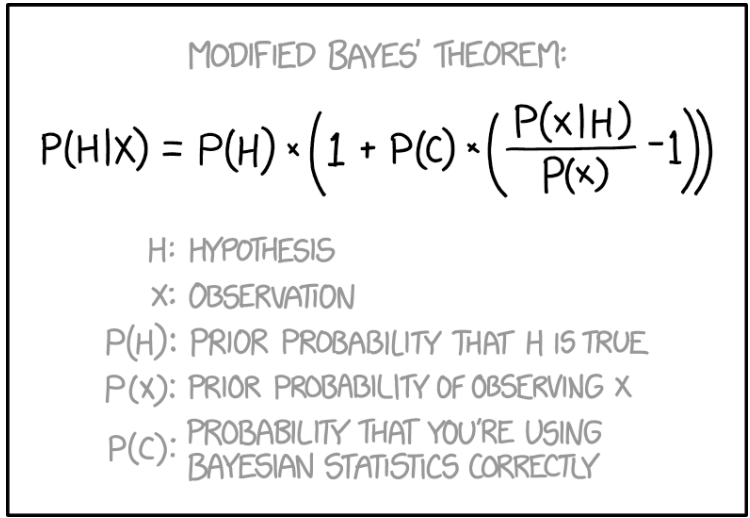
stratégie 1 : on change de porte



“Conclusion” ou résumé (succint)

- l’approche bayésienne est intéressante pour décrire des croyances face à des évènements incertains, en particulier si les évènements ne se produiront qu’une fois
- le calcul bayésien peut s’interpréter comme une mise à jour de croyance ou comme un problème inverse
- très fortement lié aux graphs causaux
- permet de tenir compte d’avis d’expert, et propose une méthdode d’ensemble
- la modélisation bayésienne décrit l’apprentissage tant humain que par des machines

“Conclusion” ou résumé (succint)



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

Références I

, ,

Adjemian, S. and Pelgrin, F. (2008). Un regard bayésien sur les modèles dynamiques de la macroéconomie. *Economie prevision*, (2):127–152.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning*, 50(1):5–43.

Anscombe, F. J., Aumann, R. J., et al. (1963). A definition of subjective probability. *Annals of mathematical statistics*, 34(1):199–205.

Bailey, A. L. (1950). *Credibility Procedures: Laplace's generalization of Bayes' Rule and the combination of collateral knowledge with observed data*. New York State Insurance Department,.

Barclay, S. et al. (1977). Handbook for decisions analysis.

Baron, K. and Lange, J. (2006). *Parimutuel applications in finance: new markets for new risks*. Springer.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Références II

Bell, E. T. (1945). *The development of mathematics*. Courier Corporation.

Berliner, L. M., Levine, R. A., and Shea, D. J. (2000). Bayesian climate change assessment. *Journal of Climate*, 13(21):3805–3820.

Bernoulli, J. (1713). *Ars conjectandi: opus posthumum: accedit Tractatus de seriebus infinitis; et Epistola gallice scripta de ludo pilae reticularis*. Impensis Thurnisiorum.

Black, F. and Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 115.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial analysts journal*, 48(5):28–43.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Buehler, R. J. (1976). Coherent preferences. *The Annals of Statistics*, 4(6):1051–1064.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin: The Journal of the IAA*, 4(3):199–207.

Buntine, W. L. and Weigend, A. S. (1991). Bayesian back-propagation. *Complex Systems*, 5.

Cardano, G. (1564). *Liber de ludo aleae*. Franco Angeli.

Références III

Charpentier, A. (2017). Les marchés prédictifs comme technique de prévision. *Risques*, (111).

Charpentier, A. (2019). Du pari au marché prédictif. *Variance.eu*.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, 505(1):147–169.

Chen, Y. and Pennock, D. M. (2010). Designing markets for prediction. *AI Magazine*, 31(4):42–52.

Chipman, H., George, E., and McCulloch, R. (2006). Bayesian ensemble learning. *Advances in neural information processing systems*, 19.

Choquet, G. (1954). Theory of capacities. In *Annales de l'Institut Fourier*, volume 5, pages 131–295.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Cormack, E. O. and Cormack, R. H. (1974). Stimulus configuration and line orientation in the horizontal-vertical illusion. *Perception & Psychophysics*, 16(2):208–212.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. Hachette.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13.

Références IV

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

De Finetti, B. (1931). *Probabilismo: saggio critico sulla teoria delle probabilità e sul valore della scienza*. Francesco Perrella.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

Dehaene, S. (2012). *Le cerveau statisticien : la révolution Bayésienne en sciences cognitives*. Collège de France.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Donnat, C., Miolane, N., Bunbury, F., and Kreindler, J. (2020). A bayesian hierarchical network for combining heterogeneous data sources in medical diagnoses. In *Machine Learning for Health*, pages 53–84. PMLR.

Drouet, I. (2016). *Le bayésianisme aujourdhui. Fondements et pratiques*. Éditions Matériologiques.

Duh, K. (2018). Bayesian Analysis in Natural Language Processing. *Computational Linguistics*, 44(1):187–189.

Références V

Eisenberg, E. and Gale, D. (1959). Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics*, 30(1):165–168.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Fenton, N. and Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Fenton, N., Neil, M., and Berger, D. (2016). Bayes and the law. *Annual Review of Statistics and Its Application*, 3:51.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Feynman, R. P. (2005). *The pleasure of finding things out: The best short works of Richard P. Feynman*. Basic Books.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

Références VI

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A., et al. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Verlag.

Gigerenzer, G. and Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *Bmj*, 327(7417):741–744.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684.

Gill, P. E., Murray, W., and Wright, M. H. (2019). *Practical optimization*. SIAM.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, 310(5746):248–249.

Références VII

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier.

Goodman, N. (1955). Axiomatic measurement of simplicity. *The Journal of Philosophy*, 52(24):709–722.

Goulet, J.-A., Nguyen, L. H., and Amiri, S. (2021). Tractable approximate gaussian inference for bayesian neural networks. *J. Mach. Learn. Res.*, 22:251–1.

Hájek, A. (2002). Interpretations of probability. *Stanford Encyclopedia of Philosophy*.

Hájek, A. (2009). Dutch book arguments. In *The Handbook of Rational and Social Choice*. Oxford University Press.

Hanea, A., Wilkinson, D. P., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., et al. (2021). Mathematically aggregating experts predictions of possible futures. *PloS one*, 16(9):e0256919.

Hasselmann, K. (1998). Conventional and bayesian approach to climate-change detection and attribution. *Quarterly Journal of the Royal Meteorological Society*, 124(552):2541–2565.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the theory of neural computation*. CRC Press.

Références VIII

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). *Graphical models with R*. Springer Verlag.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Howe, C. Q. and Purves, D. (2002). Range image statistics can explain the anomalous perception of length. *Proceedings of the National Academy of Sciences*, 99(20):13184–13188.

Hunt, I. and Mostyn, J. (2020). Probability reasoning in judicial fact-finding. *The international Journal of evidence & Proof*, 24(1):75–94.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jaynes, E. T. (1988). How does the brain do plausible reasoning? In *Maximum-entropy and Bayesian methods in science and engineering*, pages 1–24. Springer.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffrey, R. (1965). *The logic of decision*. University of Chicago press.

Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge University Press.

Références IX

Jonakait, R. N. (1983). When blood is their argument: probabilities in criminal cases, genetic markers, and, once again, bayes' theorem. *University of Illinois Law Review*, page 369.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

Karni, E., Schmeidler, D., and Vind, K. (1983). On state dependent preferences and subjective probabilities. *Econometrica:*, pages 1021–1031.

Karvetski, C. W., Olson, K. C., Mandel, D. R., and Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, 10(4):305–326.

Kause, A., Bruine de Bruin, W., Persson, J., Thorén, H., Olsson, L., Wallin, A., Dessai, S., and Vareman, N. (2022). Confidence levels and likelihood terms in ipcc reports: a survey of experts from different scientific disciplines. *Climatic Change*, 173(1):1–18.

Kemeny, J. G. (1955). Fair bets and inductive probabilities1. *The Journal of Symbolic Logic*, 20(3):263–273.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692.

Kemp, C., Tenenbaum, J. B., Niyogi, S., and Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, 114(2):165–196.

Références X

Klugman, S. A. (1991). *Bayesian statistics in actuarial science: with emphasis on credibility*, volume 15. Springer Science & Business Media.

Kolmogorov, A. (1933). Grundbegriffe der wahrscheinlichkeitsrechnung.

Kremer, W. (2014). Do doctors understand test results. *BBC World Service*.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Laplace, P. S. (1774). Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

Lebesgue, H. (1918). Remarques sur les théories de la mesure et de l'intégration. *Annales scientifiques de l'École Normale Supérieure*, 35:191–250.

Lehman, R. S. (1955). On confirmation and rational betting. *The Journal of Symbolic Logic*, 20(3):251–262.

Références XI

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance*, volume 19. Springer science & business media.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, pages 275–324.

Lindley, D. V. (2013). *Understanding uncertainty*. John Wiley & Sons.

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society: Series A (General)*, 142(2):146–162.

Longley-Cook, L. H. (1962). An introduction to credibility theory. Casualty Actuarial Society.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Martin, T. (2009). la probabilité, un concept pluriel. *Pour la science)*, (385):46–50.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., et al. (2010). Guidance note for lead authors of the ipcc fifth assessment report on consistent treatment of uncertainties.

McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C*. Yale University Press.

Références XII

Merrick, J. R. (2008). Getting the right mix of experts. *Decision Analysis*, 5(1):43–52.

Ministère de l'intérieur (2019). Procédure de reconnaissance de l'état de catastrophe naturelle - révision des critères permettant de caractériser l'intensité des épisodes de sécheresses-réhydrations des sols a l'origine des mouvement de terrains différentiels. Technical report.

Mongin, P. (1995). Consistent bayesian aggregation. *Journal of Economic Theory*, 66(2):313–351.

Mongin, P. (2001). The paradox of the bayesian experts. In *Foundations of Bayesianism*, pages 309–338. Springer.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496.

Murawaki, Y. (2019). Bayesian Learning of Latent Representations of Language Structures. *Computational Linguistics*, 45(2):199–228.

Murphy, A. H. and Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755.

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid monte carlo method. Technical report, Citeseer.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Références XIII

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, pages 97–131.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Oakes, D. (1985). Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339.

Orbanz, P. and Teh, Y. W. (2010). Bayesian nonparametric models. *Encyclopedia of machine learning*, 1.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Pettigrew, R. (2020). *Dutch book arguments*. Cambridge University Press.

Pherson, K. H. and Pherson, R. H. (2012). *Critical thinking for strategic intelligence*. CQ Press.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Pollock, W. T. and Chapanis, A. (1952). The apparent length of a line as a function of its inclination. *Quarterly Journal of Experimental Psychology*, 4(4):170–178.

Références XIV

Pólya, G. (1958). *Les mathématiques et le raisonnement plausible*. Paris, Gauthier-Villars.

Popper, K. R. (1955). Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science*, 6(21):51–57.

Popper, K. R. (1959). The propensity interpretation of probability. *The British journal for the philosophy of science*, 10(37):25–42.

Purves, D. (2009). Vision. *Handbook of neuroscience for the behavioral sciences*.

Purves, D., Wojtach, W. T., and Howe, C. (2008). Visual illusions: an empirical explanation. *Scholarpedia*, 3(6):3706.

Purves, D., Wojtach, W. T., and Lotto, R. B. (2011). Understanding vision in wholly empirical terms. *Proceedings of the National Academy of Sciences*, 108(supplement_3):15588–15595.

Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American*, 259(2):76–83.

Ramsey, F. P. (1926). *Truth and probability*. Cambridge University Press.

Roberts, H. V. (1968). On the meaning of the probability of rain. In *first national conference on statistical meteorology*.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, 81(3):247–264.

Références XV

Rougier, J. and Crucifix, M. (2018). Uncertainty in climate science and climate policy. In *Climate Modelling*, pages 361–380. Springer.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Sadegh-Zadeh, K. (1980). Bayesian diagnostics: A bibliography part 1. *Metamedicine*, 1(1):107–124.

Saini, A. (2011). A formula for justice. *The Guardian*, October 2nd.

Satchell, S. and Scowcroft, A. (2000). A demystification of the black–litterman model: Managing quantitative and traditional portfolio construction. *Journal of Asset Management*, 1(2):138–150.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Shepard, R. N. (1992). *L'Oeil qui pense: visions, illusions, perceptions*. Editions du Seuil.

Shipley, W. C., Nann, B. M., and Penfield, M. J. (1949). The apparent length of tilted lines. *Journal of experimental psychology*, 39(4):548.

Références XVI

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Skyrms, B. (1987). Dynamic coherence and probability kinematics. *Philosophy of science*, 54(1):1–20.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical science*, pages 219–247.

Stevens, S. S. (1951). *Mathematics, measurement, and psychophysics*. Wiley.

Stoerk, T., Wagner, G., and Ward, R. E. (2020). Policy brief- – recommendations for improving the treatment of risk and uncertainty in economic estimates of climate impacts in the sixth intergovernmental panel on climate change assessment report. *Review of Environmental Economics and Policy*.

Stolcke, A. (1994). *Bayesian learning of probabilistic language models*. University of California, Berkeley.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540.

Teller, P. (1973). Conditionalization and observation. *Synthese*, 26(2):218–258.

Références XVII

Tenenbaum, J. (1998). Bayesian modeling of human concept learning. *Advances in neural information processing systems*, 11.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Theodoridis, S. (2015). *Machine learning: a Bayesian and optimization perspective*. Academic press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Thompson, W. R. (1935). On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Titelbaum, M. G. (2022a). *Fundamentals of Bayesian Epistemology 1: Introducing Credences*. Oxford University Press.

Titelbaum, M. G. (2022b). *Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives*. Oxford University Press.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

Références XVIII

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.

Vogel, H., Appelbaum, S., Haller, H., and Ostermann, T. (2022). The interpretation of verbal probabilities: A systematic literature review and meta-analysis. *German Medical Data Sciences 2022–Future Medicine: More Precise, More Integrative, More Sustainable!*, pages 9–16.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, volume 9. Voss.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

Watanabe, S. and Chien, J.-T. (2015). *Bayesian speech and language processing*. Cambridge University Press.

Whitney, A. W. (1918). Theory of experience rating. *Proceedings of the Casualty Actuarial Society*, 4.

Williamson, J. (2004). *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press.

Références XIX

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616. Citeseer.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.