Cours 2012:

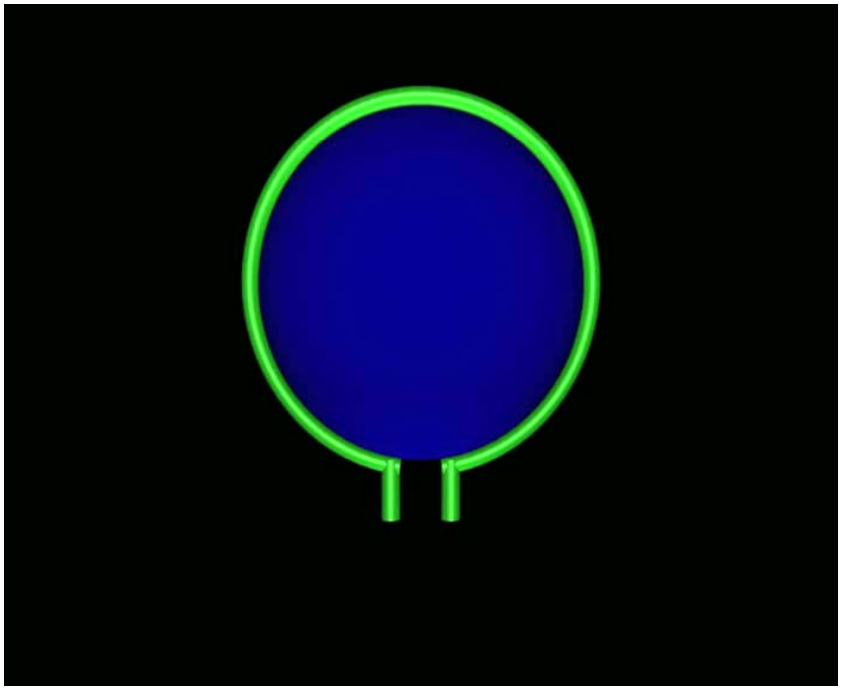
Le cerveau statisticien:
La révolution Bayésienne en sciences cognitives

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

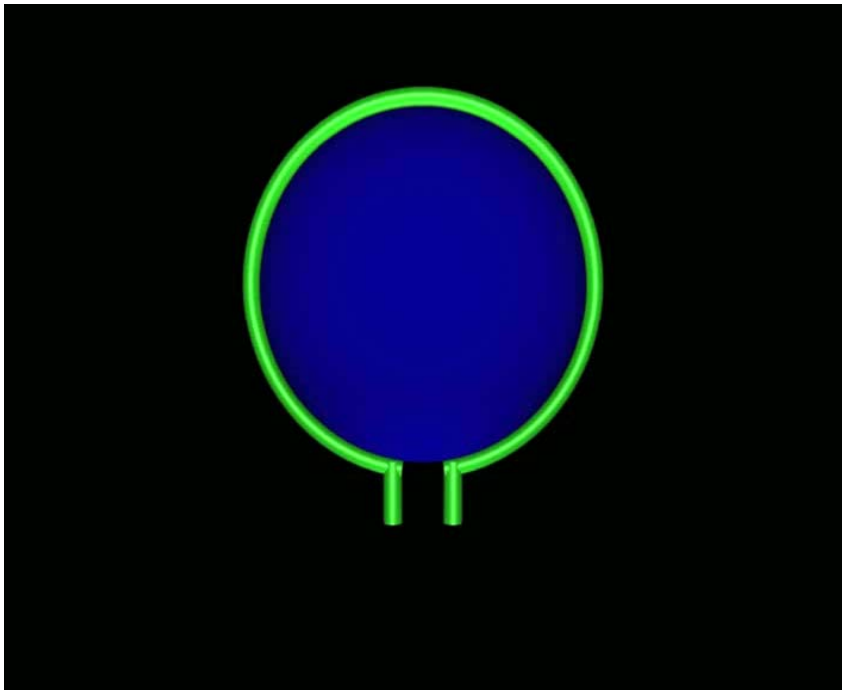
Cours n°1

Introduction au raisonnement Bayésien et à ses applications

Les statistiques intuitives



Evénement improbable



Evénement probable

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proc Natl Acad Sci U S A, 104(48), 19156-19159.*

Les bébés de 12 mois peuvent anticiper la probabilité d'un événement futur.

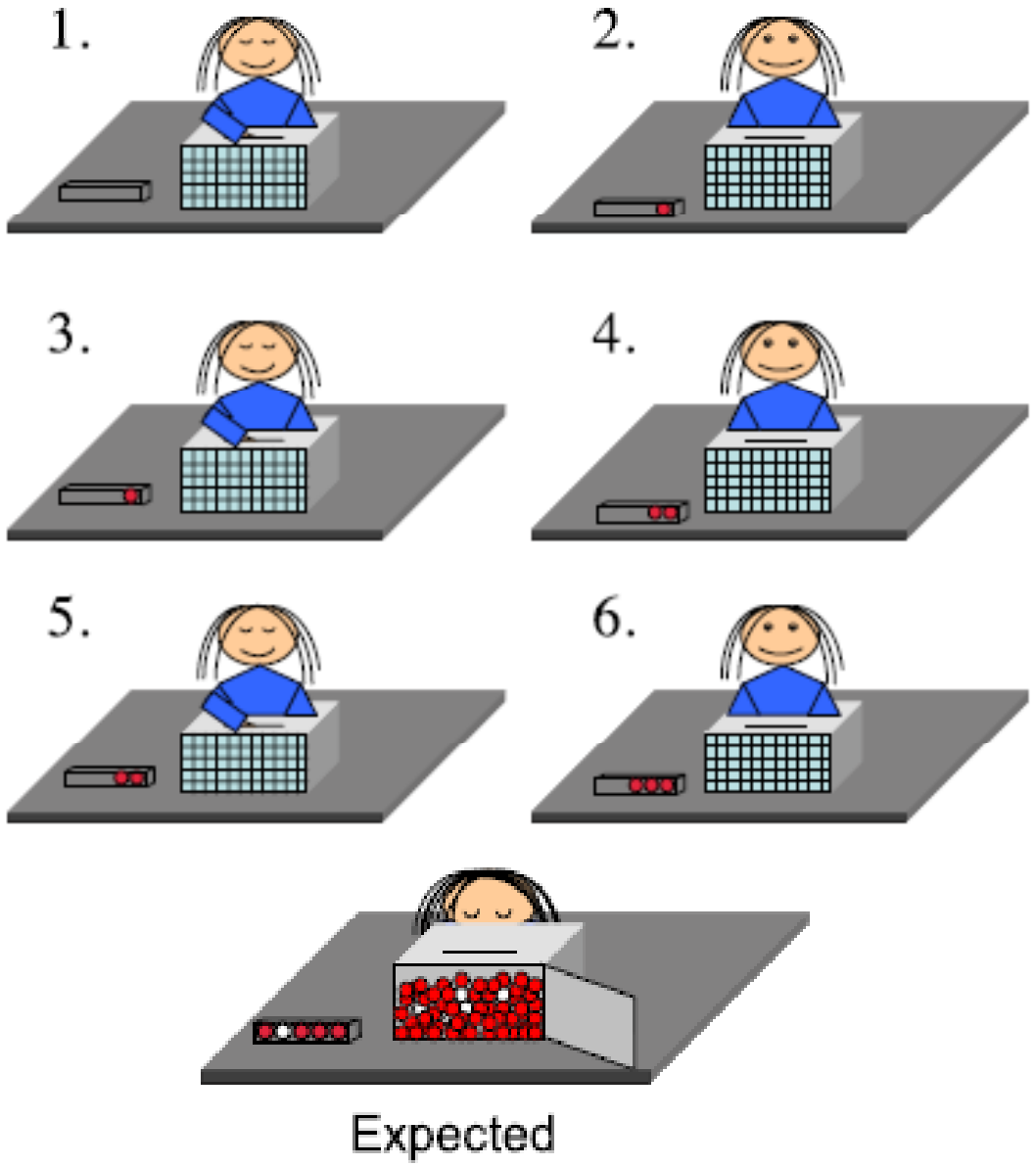
Et inversement, les bébés de 8 mois infèrent le contenu d'une urne à partir de quelques échantillons (Xu et Garcia, *PNAS* 2008).

Nous faisons constamment des inférences probabilistes.

Ces inférences sont accessibles à des enfants de quelques mois.

Hypothèse:
Notre cerveau contient des mécanismes évolués de raisonnement probabiliste.

L'inférence probabiliste chez le bébé de 8 mois



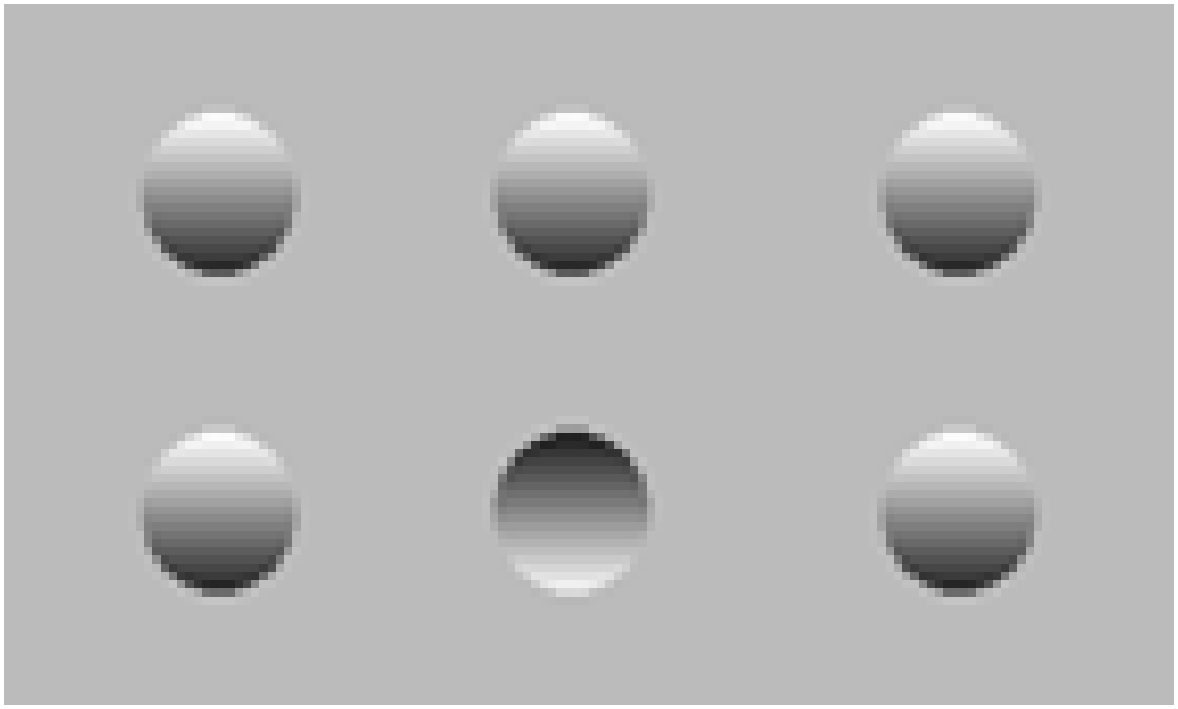
Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proc Natl Acad Sci U S A, 105(13), 5012-5015.*

Nous faisons constamment des inférences probabilistes.

Ces inférences sont accessibles à des enfants de quelques mois.

Hypothèse:
Notre cerveau contient des mécanismes évolués de raisonnement probabiliste

L'inférence dans la perception visuelle



L'inférence statistique fait partie des opérations élémentaires, automatiques et inconscientes de notre cerveau.
Elle s'applique à toutes sortes de domaines de la cognition: perception, action, apprentissage du langage, reconnaissance des mots, inférences sur l'esprit des autres…

Les idées essentielles du cours 2012

L'**inférence Bayésienne** est une théorie mathématique simple qui caractérise le **raisonnement plausible** en présence d'incertitudes.

L'inférence Bayésienne rend bien compte des processus de **perception**: étant donné des entrées ambigües, notre cerveau en reconstruit l'interprétation la plus probable.

Nos **décisions** combinent un calcul Bayésien des probabilités avec une estimation de la valeur probable et des conséquences de nos choix.

L'architecture du cortex pourrait avoir évolué pour réaliser, à très grande vitesse et de façon massivement parallèle, des inférences Bayésiennes.

L'algorithme utilisé pourrait expliquer la manière dont notre cerveau anticipe sur le monde extérieur (**codage prédictif**) et dont il répond à la nouveauté **(propagation des signaux d'erreur).**

Le **bébé** semble dôté, dès la naissance, de compétences pour le raisonnement plausible et l'apprentissage Bayésien, combinant de façon quasi optimale les *a priori* issus de notre évolution et les données reçues du monde extérieur. Ainsi la théorie Bayésienne résoudrait le dilemme classique entre les théories empiristes et nativistes.

L'apprentissage du **langage** , la reconnaissance des mots, et la **théorie de l'esprit** pourraient également être modélisés comme des inférences Bayésiennes.

Aujourd'hui: introduction à la théorie Bayésienne et puissance des algorithmes Bayésiens.

Le révérend Thomas Bayes

T. Bayes.

Pasteur de l'Église presbytérienne et mathématicien britannique (~1701-1761) .

Etudie la logique et la théologie à l'Université d'Edimburgh

Auteur de deux ouvrages publiés de son vivant:

La Bienveillance divine, ou une tentative de preuve que la fin première de la Providence et du Gouvernement divins est le Bonheur de ses créatures

Divine Benevolence :

Or, An ATTEMPT to prove that the

PRINCIPAL END

Of the DIVINE

PROVIDENCE *and* GOVERNMENT

IS THE

Happineſs of his Creatures.

Bellhouse, D. R. (2004). The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth. *Statistical Science, 19(1), 3-43.*

Le révérend Thomas Bayes

Pasteur de l'Église presbytérienne et mathématicien britannique (~1701-1761) .

Etudie la logique et la théologie à l'Université d'Edimburgh

Auteur de deux ouvrages publiés de son vivant:

La Bienveillance divine, ou une tentative de preuve que la fin première de la Providence divine et du Gouvernement est le Bonheur de ses créatures

Une introduction à la doctrine des fluxions, et une défense des mathématiciens contre les objections faites à l'auteur de l'Analyse (défense du calcul différentiel d'Isaac Newton)

Cet ouvrage aurait conduit à son élection à la *Royal Society* le 4 novembre 1742.

Après la mort de Bayes, son ami Richard Price retrouve, dans ses papiers, un *Essai sur la manière de résoudre un problème dans la doctrine des risques* qu'il présente à la *Royal Society*, où il sera publié en 1763. Celui-ci applique la fameuse « règle de Bayes ».

On savait résoudre les problèmes directs (étant donné une urne avec n balles blanches et p noires, quelle est al probabilité de tirer q noires). Le problème que résoud Bayes concerne l'inversion du raisonnement: étant donné un tirage, que peut-on dire sur le contenu de l'urne ? Autrement dit, quelle est la probabilité des tirages suivants?

Notons toutefois que la « règle de Bayes » n'est rien d'autre que la règle du produit en théorie des probabilités, dèjà connue de Bernouilli et de De Moivre. De plus, selon Jaynes, c'est Laplace (1774) et non Bayes qui en perçoit le premier toute la généralité.

Le raisonnement probabiliste, extension de la logique classique

Jaynes, E. T. (2003). *Probability theory: the logic of science.* Cambridge University Press.

En 1854, George Boole publie ses « Investigations des lois de la pensée » , ouvrage dans lequel il introduit la logique binaire que nous appelons « Booléenne ».

La logique Booléenne opère sur des valeurs discrètes: vrai ou faux

L'algèbre de Boole rend compte des syllogismes connus depuis Aristote.

Par exemple: Soit la prémise « Si A est vrai, alors B est vrai »

J'observe que A est vrai, j'en conclus que B est vrai.

J'observe que B est faux, j'en conclus que A est faux.

Cependant la logique classique exclut certains modes de « raisonnement plausible ».

Par exemple,

J'observe que A est faux. Il me semble que B devient moins plausible… bien que la logique classique ne permet aucune conclusion.

De même, j'observe que B est vrai. Il me semble que A devient plus plausible.

Nous utilisons quotidiennement ce type de raisonnement. Par exemple:

Mon fils Olivier tousse. Trois hypothèses: h1=il a la grippe. h2=il a un cancer du poumon. h3=il a une gastro-entérite.

Comment faisons-nous pour conclure que h1 est l'hypothèse la plus plausible?

Le raisonnement probabiliste, extension de la logique classique

Il est nécessaire d'étendre les valeurs de vérité discrètes « vrai » et « faux » à des **plausibilités** continues. Jaynes (2003) introduit 3 critères mathématiques que doivent vérifier ces plausibilités :

1. Les degrés de plausibilité sont représentés par des nombres réels
2. Ces valeurs doivent suivent des règles de bon sens (cf. Laplace: « la théorie des probabilités n'est au fond que le bon sens réduit au calcul »)
    a. Cohérence ou non-contradiction: S'il existe plusieurs manières d'arriver au même résultat, toutes doivent parvenir à la même valeur de plausibilité
    b. Honnêteté: Aucune des données disponibles ne doit être ignorées
    c. Reproductibilité: Des états de connaissance équivalents doivent avoir le même degré de plausibilité

Le théorème de Cox-Jaynes montre que ces règles suffisent à définir, à une fonction monotone près, des règles mathématiques universelles pour la plausibilité p. Ces règles sont les règles habituelles de la **probabilité**.

Ainsi, si l'évolution du cerveau a conduit à des règles cohérentes de manipulation de la plausibilité, ces règles doivent approximer les règles standard de la probabilité.

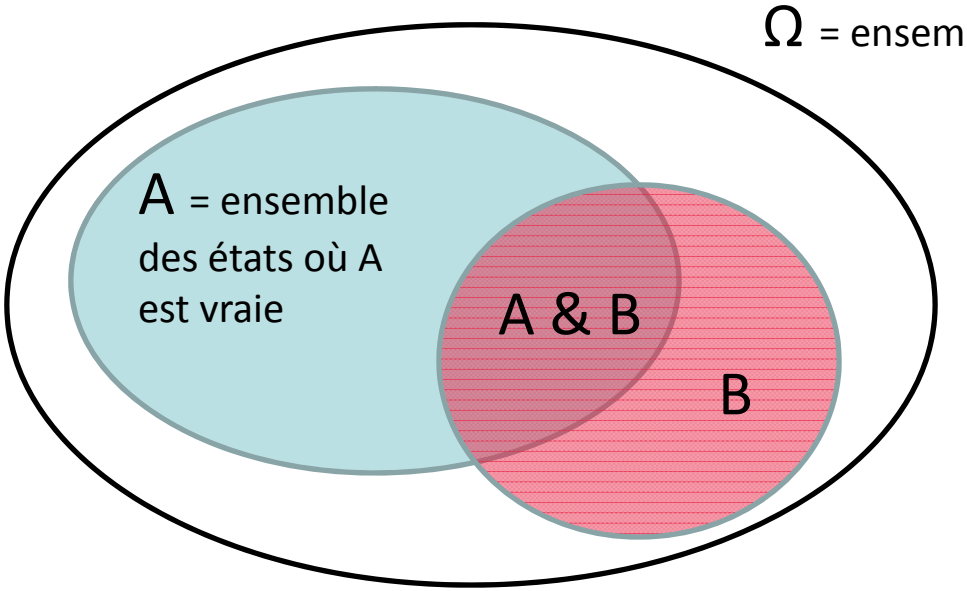
Point très important, ces probabilités ne sont plus interprétées comme les fréquences relatives d'événements (point de vue « fréquentiste »), mais comme des mesures du degré de connaissance subjective (point de vue « Bayésien »).

Le raisonnement probabiliste, extension de la logique classique

Quelques règles fondamentales de la « plausibilité »:

- p(x)=0 si x est fausse
- p(x)=1 si x est vraie
- $p(x_i)=1/n$ si les *n* hypothèses x_i sont mutuellement exclusives et toutes aussi plausibles
- p(non-x) = 1 – p(x)
- la règle fondamentale « de Bayes » : p(A & B) = p(A|B) p(B) = p(B|A) p(A)

On peut motiver cette règle par un diagramme de Venn:



Cependant, ceci n'est qu'un aide-mémoire. Les règles mathématiques qui gouvernent la plausibilité restent valables même pour des assertions A et B qui ne sont pas facilement décomposables en états disjoints plus élémentaires.

Raisonnement « en avant » ou « en arrière »

La théorie des probabilités, telle qu'enseignée à l'école, est surtout utilisée pour calculer la probabilité d'une observation, étant donné certaines hypothèses sur l'état du monde.

Par exemple, soit une urne contenant 3 boules noires et 7 boules blanches. Quelle est la probabilité, lors de deux tirages sans remplacement, de tirer deux boules noires?

$p(H \,\&\, N_1 \,\&\, N_2) = p(N_1|H)\; p(N_2|N_1 \,\&\, H) = (3/10) \times (2/9) = 1/15$

Mais, les données d'observation D et les hypothèses H jouent des rôles strictement symétriques. Rien n'empêche d'utiliser les équations pour inverser le procédé:

Etant donnée l'observation D, quelle est la probabilité de l'hypothèse H?

Application de la règle fondamentale: $p(H \,\&\, D) = p(D|H)\; p(H) = p(H|D)\; p(D)$

D'où $p(H|D) = p(D|H)\; p(H) / p(D)$ ou $\boxed{p(H|D) \;\alpha\; p(D|H)\; p(H)}$

p(H) = *probabilité « a priori » de H* (*prior* en anglais) (mais pas dans le sens Kantien d'indépendant de l'expérience; elle peut résulter d'expériences antérieures)

p(H|D) = *probabilité « a posteriori » de H* (pas nécessairement au sens temporel, mais au sens de la déduction logique, après avoir observé D)

p(D|H), considéré comme une fonction de H, est la *vraisemblance de H*

Un exemple qualitatif de raisonnement Bayésien

Mon fils Olivier tousse. Trois hypothèses:

h1=il a la grippe. h2=il a un cancer du poumon. h3=il a une gastro-entérite.

Règle fondamentale: $p(H|D) \ \alpha \ p(D|H)\ p(H)$

Pour h1 (grippe): tant la probabilité à priori que la vraisemblance sont élevées

Pour h2 (cancer du poumon): la vraisemblance est élevée mais la probabilité à priori est faible

Pour h3 (gastro-entérite): la probabilité à priori est élevée, mais la vraisemblance est faible.

Ainsi h1 est l'hypothèse la plus probable a posteriori .

Nous utilisons un critère « MAP » (*maximum a posteriori* hypothesis) et non un critère de maximum de vraisemblance (ML ou *maximum likelihood*) car la probabilité *a priori* ne peut pas être négligée.

Quelques remarques sur la vraisemblance

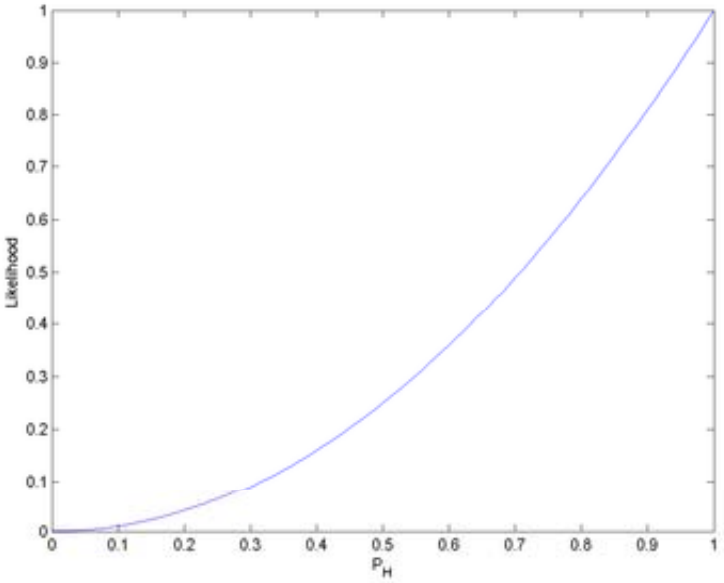
p(D|H), considéré comme une fonction de H, est la *vraisemblance de H* (*likelihood*).

La vraisemblance est donc une fonction qui, étant donné un jeu d'observation D, ordonne les hypothèses H et nous donne une idée sur leurs mérites relatifs.

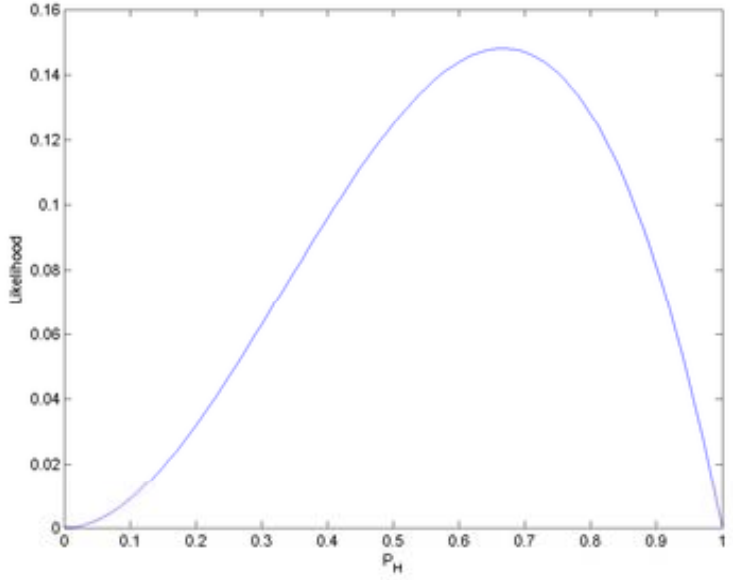
La vraisemblance n'est pas une probabilité! (l'intégrale sur H ne vaut pas 1)

Exemple: jouer à pile ou face avec une pièce truquée:

Vraisemblance de l'hypothèse « la probabilité de pile est p_H » après avoir observé deux fois pile.



Vraisemblance de l'hypothèse « la probabilité de pile est p_H » après avoir observé deux fois pile et fois face



Quelques remarques sur la vraisemblance

Le **facteur de Bayes** qui sépare deux hypothèses ou deux « modèles » M_1 et M_2, est une mesure de leur mérite relatif, le rapport de leurs vraisemblances.

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)}$$

Le **logarithme** de cette valeur est une mesure souvent plus intelligible.

On appelle « évidence » (*weight of evidence* [WOE] ou *log odds* [Turing]) la valeur log(p/(1-p)), qui quantifie la vraisemblance d'une hypothèse par rapport aux autres.

Harold Jeffreys propose une interprétation des valeurs de K (Log K est mesuré en *décibans*)

K	dB	bits	Strength of evidence
< 1:1	< 0		Negative (supports M_2)
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Substantial
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
> 100:1	> 20	> 6.6	Decisive

La méthode de Bayes automatise le rasoir d'Ockham !

On attribue à Guillaume d'Ockham (1285-1347) un principe de raisonnement en réalité énoncé depuis l'Antiquité: « Une pluralité ne doit pas être posée sans nécessité »

« Les entités ne doivent pas être multipliées au delà du nécessaire »

Toutes choses égales par ailleurs, les explications les plus simples doivent être préférées aux plus complexes.

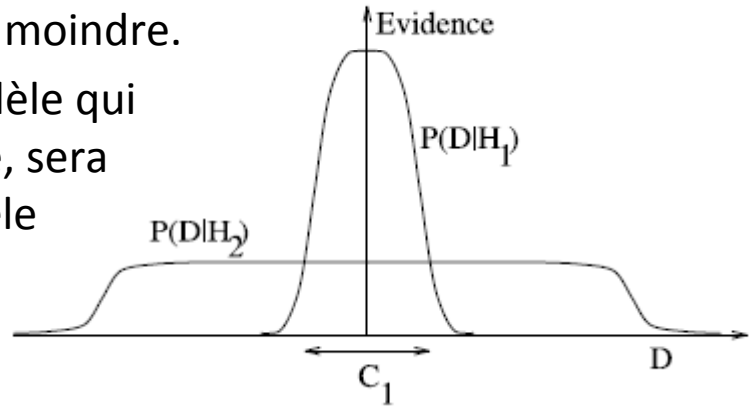
Jaynes explique que le raisonnement Bayésien, en tant que « logique de la science », inclut automatiquement un rasoir d'Ockham, pour plusieurs raisons:

Pour choisir entre deux hypothèses H1 et H2, je compare P(H1|D) et P(H2|D).

Si chacun de ces « modèles » comprend des paramètres libres θ_1 θ_2 …, je dois calculer l'intégrale sur tous les θ_i de P(D|H1, θ_i). Le « volume » de cette intégrale multiple est d'autant plus grand que l'espace des paramètres est vaste. Si les deux modèles atteignent la même vraisemblance, le modèle qui l'atteint avec plus de paramètres libres est donc automatiquement pénalisé par les équations Bayésiennes.

Pour Jeffreys, il est également naturel (?) de supposer que les modèles avec un plus grand nombre de paramètres ont une probabilité *a priori* moindre.

Enfin, la vraisemblance elle-même peut varier. Un modèle qui précisément les données observées, et rien d'autre, sera automatiquement privilégié par rapport à un modèle qui fait cette prédiction, mais également d'autres.



Le « scandale de l'induction » éclairé par la théorie Bayésienne

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

« Pour les scientifiques intéressés par la manière dont les humains parviennent à comprendre le monde, le principal défi est le suivant: Comment nos esprits parviennent-ils à inférer autant à partir de si peu? » (Tenenbaum, *Science*, 2011).

C'est le « scandale de l'induction » discuté par Russell, mais aussi Platon, Aristote, Kant, Peirce…

Les enfants et les adultes réalisent quotidiennement des inférences très sophistiquées alors qu'il parait évident qu'ils n'ont pas assez de données.

Par exemple, tout le monde sait que « corrélation n'est pas causation » -- et pourtant les humains infèrent régulièrement des relations causales sur la base de quelques données qui ne suffiraient même pas à calculer un coefficient de corrélation!

Autre exemple: l'apprentissage du langage.

- Quine et le « gavagai ! »
- Chomsky et la « pauvreté du stimulus »

Le « scandale de l'induction » éclairé par la théorie Bayésienne

Un exemple d'induction rapide (Tenenbaum, *Science*, 2011):

Les objects rouges sont des « tufa ».

Le raisonnement Bayésien peut expliquer l'induction:

- les hypothèses sont des branches de l'arbre
- l'a priori est proportionnel à la hauteur de la branche
- la vraisemblance repose sur l'hypothèse que les exemplaires sont tirés au hasard à l'intérieur de la catégorie