Statistiques Bayésiennes

Hicham Janati

hjanati@insea.ac.ma

INSEA

1. Introduction

2. Les Bayésiens vs Les fréquentistes

3. Rappels de probabilités (exemples)

4. Loi a posteriori et modèles conjugués

5. Estimateur de Bayes

Mes objectifs:

1. Synthétiser l'information et vous la présenter: vous faire gagner du temps

2. Corriger les lacunes que vous pouvez avoir

3. Vous montrer des applications concrètes des statistiques Bayésiennes en Stats / ML / Finance

INSEA

Ce que je vous demande

1. Interactivité: Interrompez-moi SVP si quelque chose n’est pas clair
2. Les cours contiennent des “pauses interactives / mini-exercices” : soyez actifs
3. Adoptez un esprit critique: la qualité du cours dépend de votre participation

Organisation

- On va alterner entre cours / TD / TP (Python)
- 1 test par semaine portant sur la séance en cours et les séances d’avant
- Tests (30%) + Examen (70%) qui évaluent votre compréhension
- Ponctualité: je commence à 8h30 - je finis à 10h20
- Les slides sont un support de cours qui m’aident à visualiser — non pas un substitut de cours

INSEA

Programme

- **Chapitre I. Fondements du modèle bayésien**
    1. Introduction: Les Bayésiens vs Les fréquentistes, rappels
    2. Modèle bayésien: Loi a priori et loi a posteriori, exemples
    3. Modèles à priors conjuguées

Maîtriser > couvrir tout le programme

- **Chapitre 2. Méthodes de Monte-Carlo**
    1. Introduction aux méthodes Monte-Carlo
    2. Méthodes d'échantillonnage (sampling) Markov-Chain Monte-Carlo (MCMC)
- **Chapitre 3. Applications et thématiques avancées**
    1. Modèles Bayésiens hiérarchiques
    2. Bayesian machine learning

INSEA

Pourquoi ce cours ?

- Modélisation des risques / incertitudes (Actuariat)
- Modélisation probabiliste en médecine / biostats
- Machine learning (Réseaux de neurones Bayésiens, modèles génératifs, Optimisation)

Bayesian statistics in a nutshell

Qu’est-ce que les statistiques Bayésiennes ?

L’approche Bayésienne est souvent contrastée avec l’approche Fréquentiste:

Approche Fréquentiste:

- On cherche à estimer de manière précise un paramètre considéré comme constant à partir des données
- On cherche une valeur avec un intervalle de confiance, un test statistique

Approche Bayésienne:

- On modélise le paramètre comme une variable aléatoire avec une distribution a priori que l’on met à jour après avoir “vu” les données
- On cherche **une distribution** sur le paramètre qui donne un niveau d’incertitude sur toutes ses valeurs possibles

Des différences mathématiques avec une portée philosophique

INSEA

Frequentist vs Bayesian (1)

Une femme enceinte se demande: Quel est la probabilité que mon nouveau-né soit de sexe masculin ?

The **Frequentist**: “Notons le sexe par une variable aléatoire binaire suivant la loi de Bernoulli $X \sim \mathcal{B}(p)$. Il me faut des observations indépendantes et identiquement distribuées (i.i.d) binaires $x_1, \ldots, x_n \sim \mathcal{B}(p)$. Je pourrais ensuite inférer le paramètre p en utilisant l’estimateur: $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ”

The **Bayesian**: “En considérant p comme un paramètre quelconque, le Fréquentiste omet complètement l’idée qu’il s’agit d’une probabilité dans un contexte spécifique. Sans données, n’importe qui aurait répondu: 0.5. L’incertitude sur la valeur du paramètre p ne devrait pas être uniforme sur tout l’intervalle [0, 1]. Cette **croyance à priori** devrait être incluse dans l’analyse statistique.”

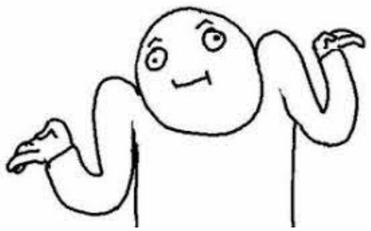
Frequentist vs Bayesian (2)

The **Frequentist**: “Je regarde les données et j’en tire des conclusions objectives.”

The **Bayesian** says: “J’ai une croyance à priori. Je regarde les données pour mettre à jour mes croyances.”

ON NE PEUT PAS INCLURE SES OPINIONS. LA SCIENCE DOIT RESTER OBJECTIVE.

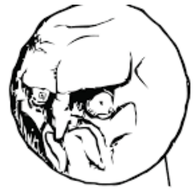
The **Frequentist**



Oui, mais parfois, comme scientifique, on peut avoir des croyances ou des hypothèses a priori qu’on doit inclure dans notre analyse car on ne peut pas se fier aux données à 100%.

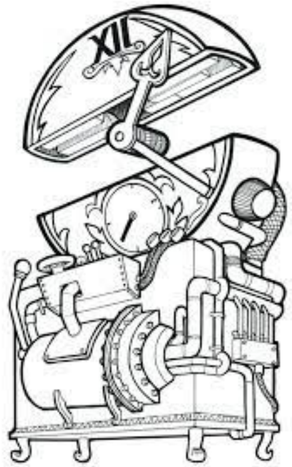
The **Bayesian**

The **Frequentist**



Frequentist vs Bayesian (3)

Il fait nuit. Deux statisticiens discutent si le soleil a disparu.



The **Frequentist:** D'après les observations données par ma machine astronomique, j'ai testé l'hypothèse si les particules détectées proviennent du soleil. La probabilité que cela se produise par hasard (p-val) est de 0,03. Avec une valeur $p < 0,05$, je conclus que le soleil a disparu.

The **Bayesian:** Je parie $100 qu'il est toujours là.

Source: Adapted from xkcd.com/1132/

Frequentist vs Bayesian (3)

Pourquoi le bayésien est-il convaincu que le soleil n'a pas disparu ?

Le **Bayésien** croyaient auparavant que la disparition du soleil était extrêmement rare.

Mais cela ne signifie pas que les **Bayésiens** peuvent « croire » tout ce qu’ils veulent.

Mais cela ne signifie pas que les **Bayésiens** peuvent « croire » tout ce qu'ils veulent.

"Steve est timide mais c'est une personne très serviable, organisée et ordonnée. Il a un besoin d'ordre et de structure. C'est aussi une personne réservée."

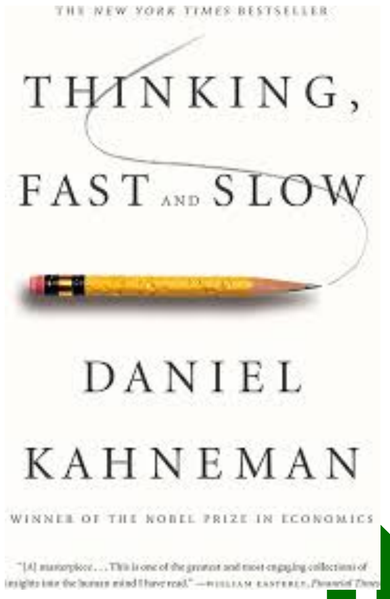
Steve est-il plus susceptible d'être un bibliothécaire ou un agriculteur ?

20% des agriculteurs sont timides

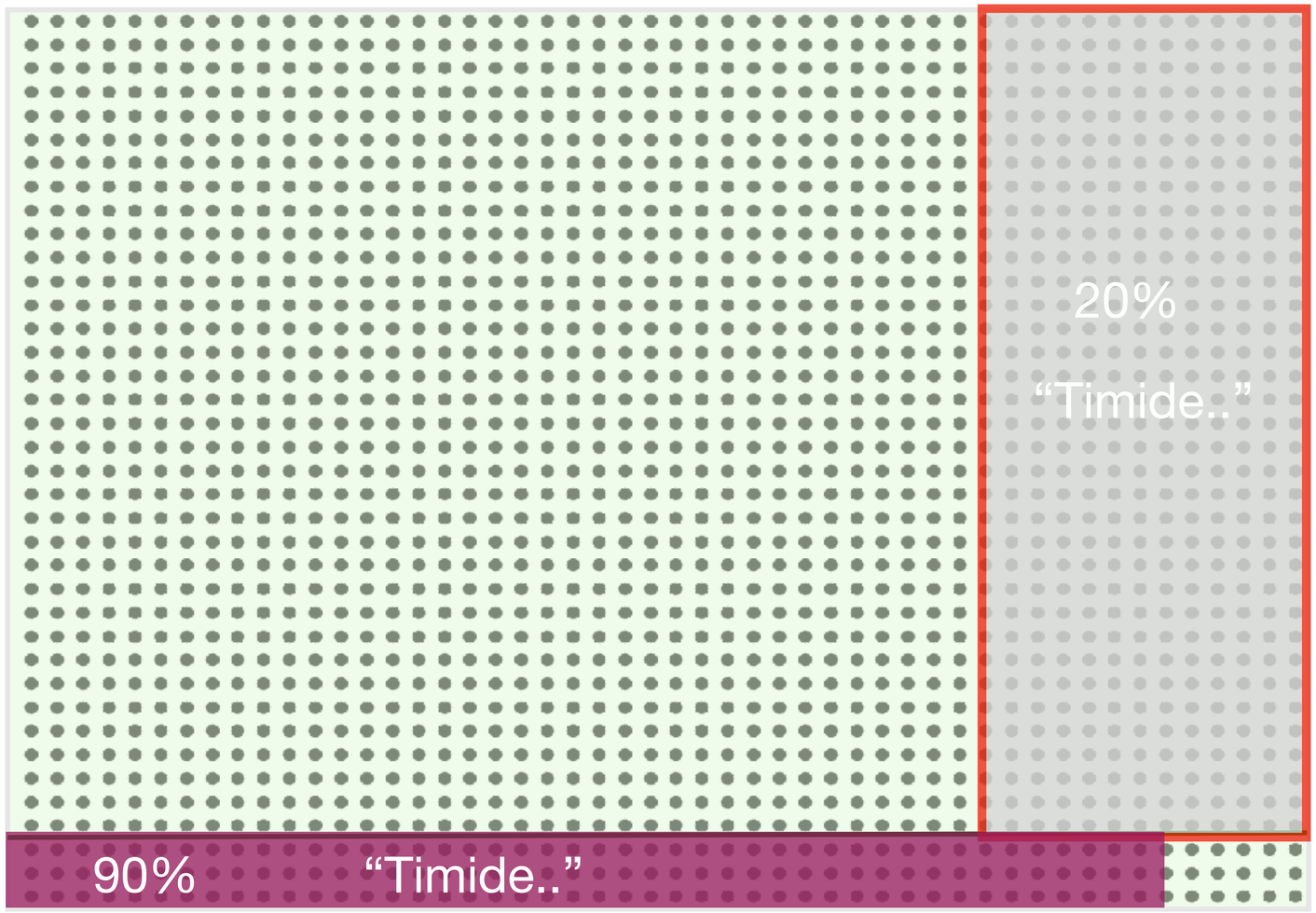
90% des bibliothécaires sont timides

Cela confirme-t-il votre jugement "intuitif" ?

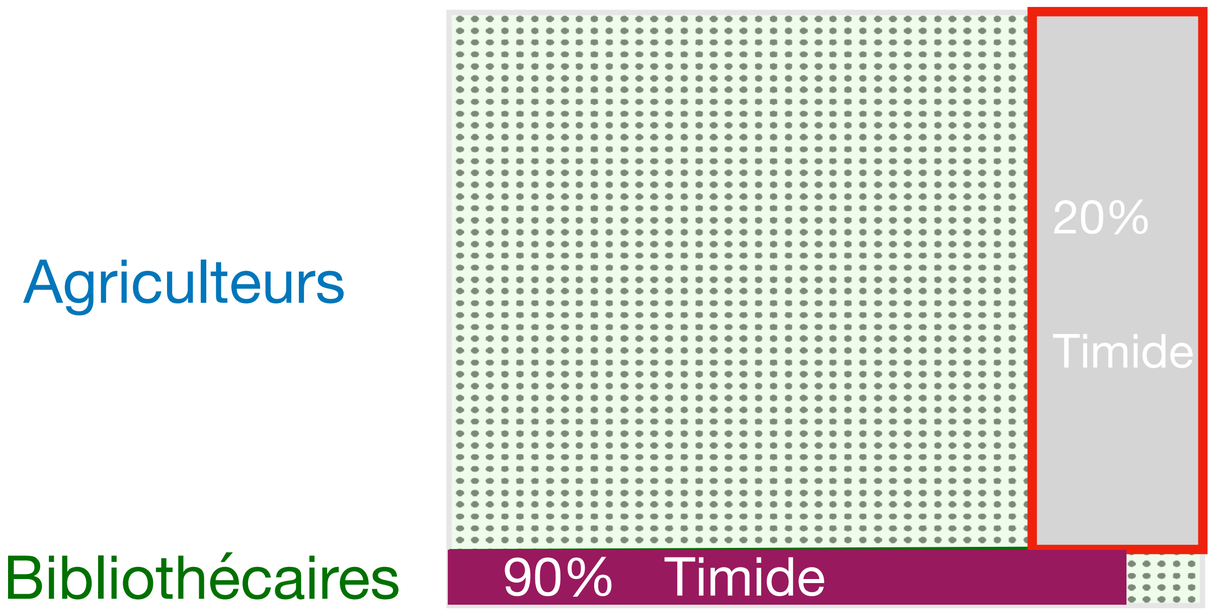
Adapted from:



Il y a 9 fois plus d'Agriculteurs que de Bibliothécaires:



Il y a 9 fois plus d'Agriculteurs que de Bibliothécaires (aux US)



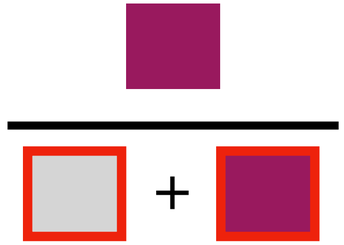
Steve est **Timide**, quelle est la probabilité qu'il soit un bibliothécaire ?

Deux méthodes possibles:

1. Formelle: On calcule la probabilité d'être un Bibliothécaire sachant qu'on est **Timide:**

$$\mathbb{P}(\mathbf{B}|\mathbf{T})$$

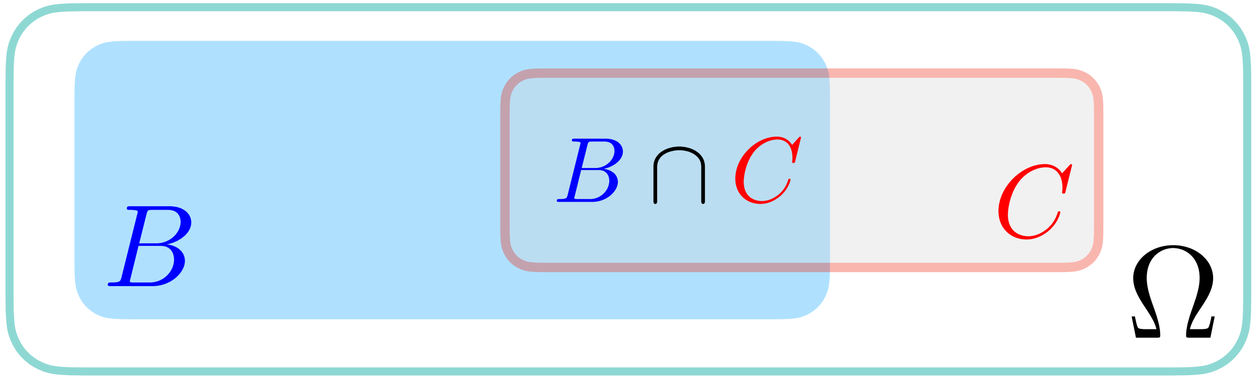
2. Graphique: On estime la fréquence des **B timides** parmi tous les **Timides**



Théorème de Bayes

Soit $(\Omega, \mathcal{P}, \mathcal{A})$ un espace probabilisé. Soient $B, C \in \mathcal{A}$, alors:

$$\mathbb{P}(C|B) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)}$$



Ensemble de tous les événements possibles

"Probabilité de C sachant B" =

"La probabilité de B et C mais restreinte à B: comme si B était devenu tout l'espace"

En appliquant deux fois: $\mathbb{P}(C|B) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|C)\mathbb{P}(C)}{\mathbb{P}(B)}$

INSEA

Théorème de Bayes

Soit $(\Omega, \mathcal{P}, \mathcal{A})$ un espace probabilisé. Soient $B, C \in \mathcal{A}$, alors:

$$\mathbb{P}(C|B) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)}$$

On a: $\mathbb{P}(C|B) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)}$

Et: $\mathbb{P}(B|C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}$

Donc: $\mathbb{P}(B \cap C) = \mathbb{P}(B|C)\mathbb{P}(C)$

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

"Inversion des probabilités"
(Nom original donné par Bayes en 1763)

Appliquons cela pour calculer $\mathbb{P}(\mathbf{B}|\mathbf{T})$

$$\mathbb{P}(\mathbf{B}|\mathbf{T}) = \frac{\mathbb{P}(\mathbf{T}|\mathbf{B})\mathbb{P}(\mathbf{B})}{\mathbb{P}(\mathbf{T})}$$

Peut-on calculer ces quantités ?

90% des bibliothécaires sont timides donc: $\mathbb{P}(\mathbf{T}|\mathbf{B}) = 0.9$

Il y a 9 fois plus d'**A** que de **B** donc $\mathbb{P}(\mathbf{A}) = 9\,\mathbb{P}(\mathbf{B})$

Or l'espace est restreint aux A et B (il n'y a pas d'autres possibilités) donc

$$\mathbb{P}(\mathbf{A}) + \mathbb{P}(\mathbf{B}) = 1$$

Ainsi: $\mathbb{P}(\mathbf{A}) = 0.9$ $\mathbb{P}(\mathbf{B}) = 0.1$

$\mathbb{P}(\mathbf{T}) = \quad ?$

Suite au tableau

Loi des probabilités totales

Soit $(\Omega, \mathcal{P}, \mathcal{A})$ un espace probabilisé. Soit $(A_i)_{i\in\mathbb{N}}$ une partition de Ω, c'est-à-dire que $A_i \cap A_j = \varnothing$ pour $i \neq j$ et $\bigcup_{i\in\mathbb{N}} A_i = \Omega$. Soit $B \in \mathcal{A}$, alors :

$$\mathbb{P}(B) = \sum_{i=1}^{+\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Quelle est l'intuition de cette loi ? Autrement dit, d'où vient-elle ?

$$\sum_{i=1}^{+\infty} \mathbb{P}(\mathbf{B}|A_i)\mathbb{P}(A_i) = \sum_{i=1}^{+\infty} \mathbb{P}(\mathbf{B} \cap A_i) = \mathbb{P}(\cup_{i=1}^{n} B \cap A_i) = \mathbb{P}(\mathbf{B} \cap \cup_{i=1}^{n} A_i) = \mathbb{P}(\mathbf{B} \cap \Omega) = \mathbb{P}(\mathbf{B})$$

Bayes

Car éléments disjoints

Car partition (faire un schéma)

Corollaire : Soit $\mathbf{X}$ une variable aléatoire discrète prenant ses valeurs dans $\mathbb{N}$. Alors :

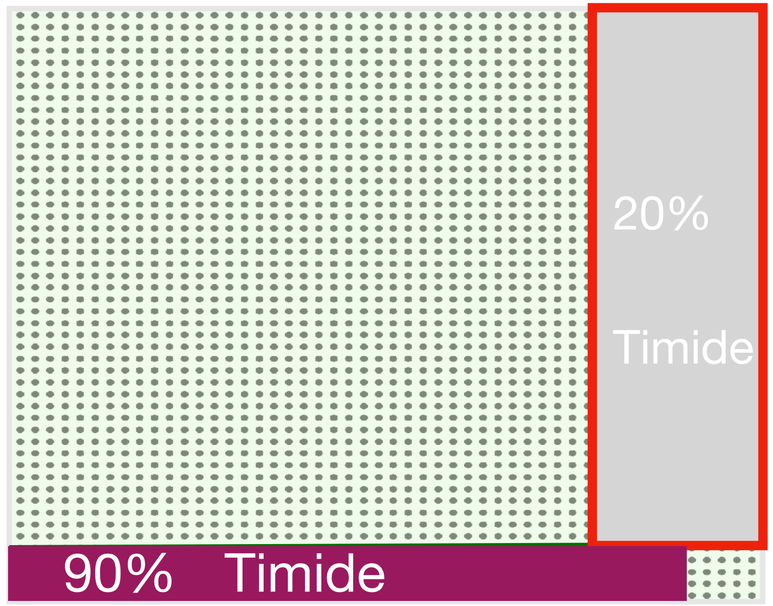
$$\mathbb{P}(B) = \sum_{i=1}^{+\infty} \mathbb{P}(B|\mathbf{X} = i)\mathbb{P}(\mathbf{X} = i)$$

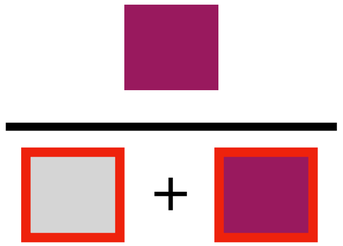
Les **A** et les **B** forment une partition, donc:

$$\mathbb{P}(\mathbf{T}) = \mathbb{P}(\mathbf{T}|\mathbf{A})\mathbb{P}(\mathbf{A}) + \mathbb{P}(\mathbf{T}|\mathbf{B})\mathbb{P}(\mathbf{B})$$

Conclusion: $$\mathbb{P}(\mathbf{B}|\mathbf{T}) = \frac{\mathbb{P}(\mathbf{T}|\mathbf{B})\mathbb{P}(\mathbf{B})}{\mathbb{P}(\mathbf{T})} = \frac{\mathbb{P}(\mathbf{T}|\mathbf{B})\mathbb{P}(\mathbf{B})}{\mathbb{P}(\mathbf{T}|\mathbf{A})\mathbb{P}(\mathbf{A})+\mathbb{P}(\mathbf{T}|\mathbf{B})\mathbb{P}(\mathbf{B})}$$

$$= \frac{0.9\times0.1}{0.2\times0.9+0.9\times0.1} = \frac{1}{3}$$





Vous voulez savoir si vous faites partie des **1%** de la population humaine qui sont des génies. Vous achetez un test de QI qui fait l'affaire et WOW, votre test est positif. Sur l'étiquette il est écrit : « **95 %** de précision ».

Quelle est la probabilité que vous soyez un génie ?

Vous voulez savoir si vous faites partie des **1%** de la population humaine qui sont des génies. Vous achetez un test de QI qui fait l'affaire et WOW, votre test est positif. Sur l'étiquette il est écrit : « **95 %** de précision ».

Qu'est-ce que la précision ?

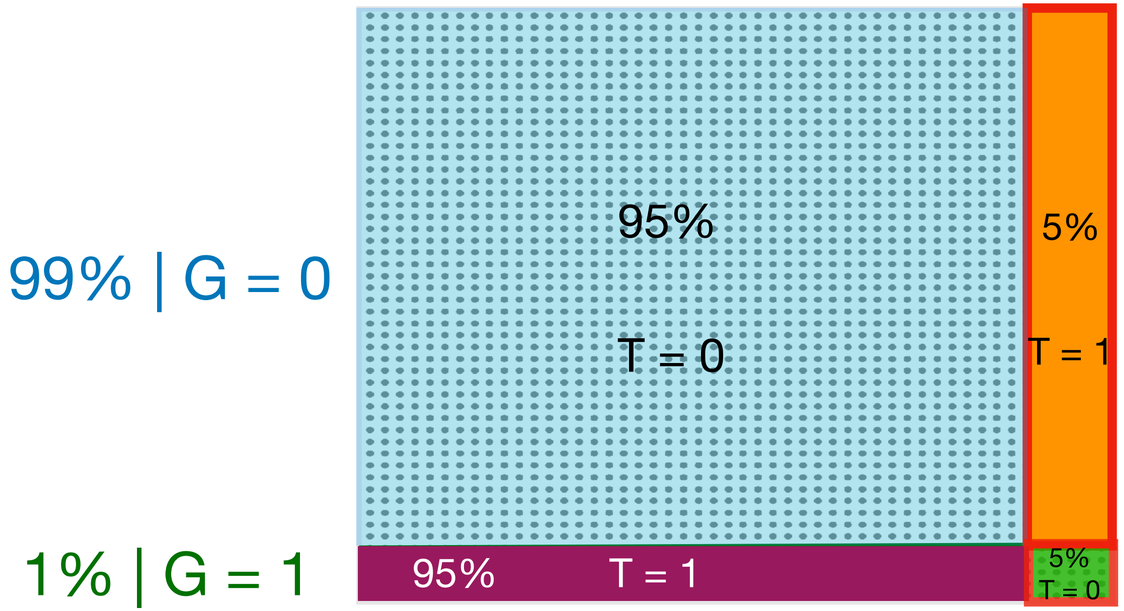
Quelle est la probabilité que vous soyez un génie ?

On note G et T les variables aléatoires binaires: "génie": G = 1. "test positif": T = 1

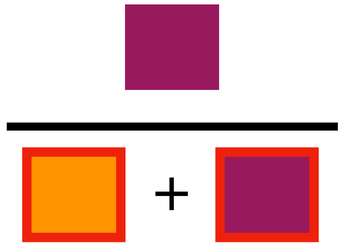
Une précision de 95% implique deux faits:

Si on est un génie, alors le test est positif 95% des fois: $\mathbb{P}(T = 1|G = 1) = 0.95$

Si on n'est pas un génie, alors le test est positif (se trompe) 5% des fois donc $\mathbb{P}(T = 1|G = 0) = 0.05$



Nous cherchons la probabilité d'être un **génie** parmi les **tests positifs** :



$$= \frac{0.01 \times 0.95}{0.99 \times 0.05 + 0.01 \times 0.95} \approx 0.161$$

Formellement, avec Bayes + Probabilités totales:

$$\mathbb{P}(G=1|T=1) = \frac{\mathbb{P}(T=1|G=1)\mathbb{P}(G=1)}{\mathbb{P}(T=1)} = \frac{\mathbb{P}(T=1|G=1)\mathbb{P}(G=1)}{\mathbb{P}(T=1|G=1)\mathbb{P}(G=1)+\mathbb{P}(T=1|G=0)\mathbb{P}(G=0)}$$

INSEA

1. Pour le moment, on a appliqué le théorème de Bayes
2. Utilisé des probabilités conditionnelles
3. Pas encore défini ce qu'est un modèle Bayésien

Voyons cela en le contrastant au modèle Fréquentiste

Soit un phénomène dont on observe des variables aléatoires i.i.d $X_1, \ldots, X_n$ modélisées par une loi paramétrée par θ.

Modèle Fréquentiste:

1.

2.

3.

Soit un phénomène dont on observe des variables aléatoires i.i.d $X_1, \ldots, X_n$ modélisées par une loi paramétrée par θ.

Modèle Bayésien:

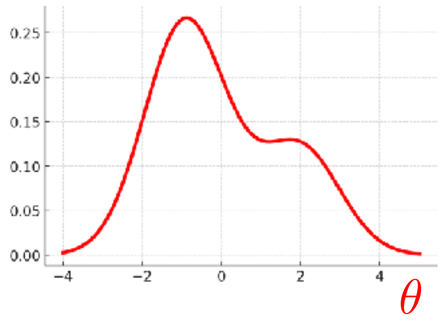
1.

2.

3.

4.

loi a posteriori



En statistiques Bayésiennes, les variables aléatoires sont très souvent continues:

La fonction de densité de probabilité d'une variable aléatoire continue (a priori) θ est notée f_θ.
On suppose que la densité des données $f_{\mathbf{X}}$ est bien définie. La distribution a posteriori de $\theta \mid \mathbf{X}$ est donnée par :

$$f_{\theta|\mathbf{X}} = \frac{f_{\mathbf{X},\theta}}{f_{\mathbf{X}}} = \frac{f_{\mathbf{X}|\theta}\, f_\theta}{f_{\mathbf{X}}}$$

Comment trouver la loi marginale ?

Loi des probabilités totales: $f_{\mathbf{X}} = \int f_{\mathbf{X}|\theta}\, f_\theta \, \mathrm{d}\theta$

On obtient une fonction divisée par son intégrale: l'intégrale du rapport = 1

$$f_{\theta|\mathbf{X}} = \frac{f_{\mathbf{X}|\theta}\, f_\theta}{\int f_{\mathbf{X}|\theta}\, f_\theta \, \mathrm{d}\theta}$$

Constante de normalisation indépendante de θ

$$= \mathrm{cst} \times f_{\mathbf{X}|\theta}\, f_\theta$$

$$\propto f_{\mathbf{X}|\theta}\, f_\theta$$

Symbole qui signifie: "proportionnel à"

Le bayésien multiplie la vraisemblance des données par la loi a priori pour "mettre à jour" ses croyances sur les valeurs probables de θ

INSEA

Exemple 1: “Quel est la probabilité que mon nouveau-né soit de sexe masculin ?”

On note θ cette probabilité. On définit une variable aléatoire binaire X désignant le sexe masculin avec $\mathbb{P}(X = 1) = \theta$.

Ainsi, X suit une loi de Bernoulli $\mathcal{B}(\theta)$ et on a pour $k \in \{0, 1\}$ $\mathbb{P}(X = k) = \theta^k(1 - \theta)^{1-k}$.

Soit $X_1, \ldots, X_n$ des variables i.i.d $\sim \mathcal{B}(\theta)$ pour les quelles on observe n valeurs $x_1, \ldots, x_n$.

1. Détaillez l’approche Fréquentiste en calculant la vraisemblance et son maximum.

2. Le Bayésien considère que θ est une variable aléatoire suivant une loi a priori π. Quelle est la formule pour trouver la distribution de $\theta|X_1, \ldots X_n$?

3. $\theta|X_1, \ldots X_n$ suit la loi a posteriori. Trouver sa distribution en prenant une loi a priori uniforme.

   On rappelle que la densité d’une loi $\text{Beta}(a, b)$ est donnée par :

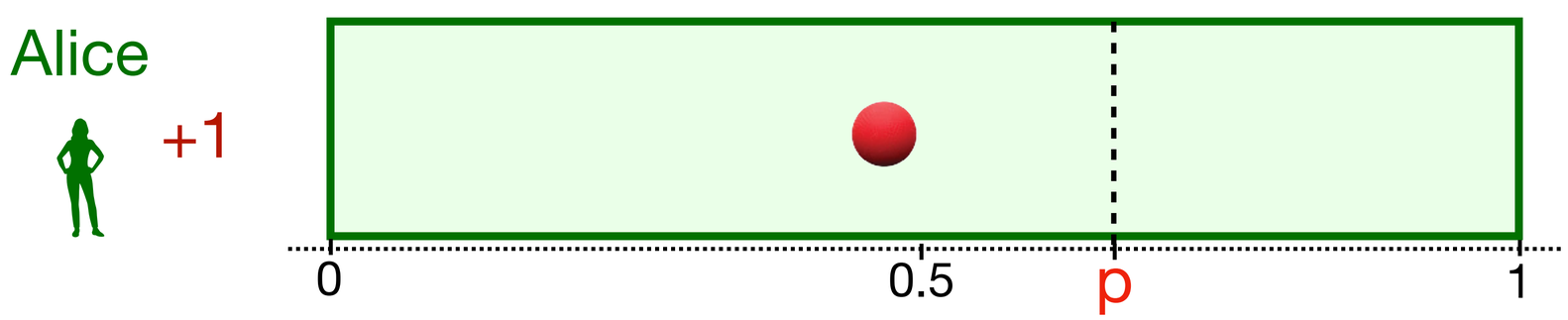
$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}, \quad \text{pour } x \in [0, 1].$$

   Sa moyenne est donnée par $\frac{a}{a+b}$. Pour $a, b > 1$, son maximum est atteint en $\frac{a-1}{a+b-2}$.

4. Comparez avec l’approche Fréquentiste.

La table de Bayes

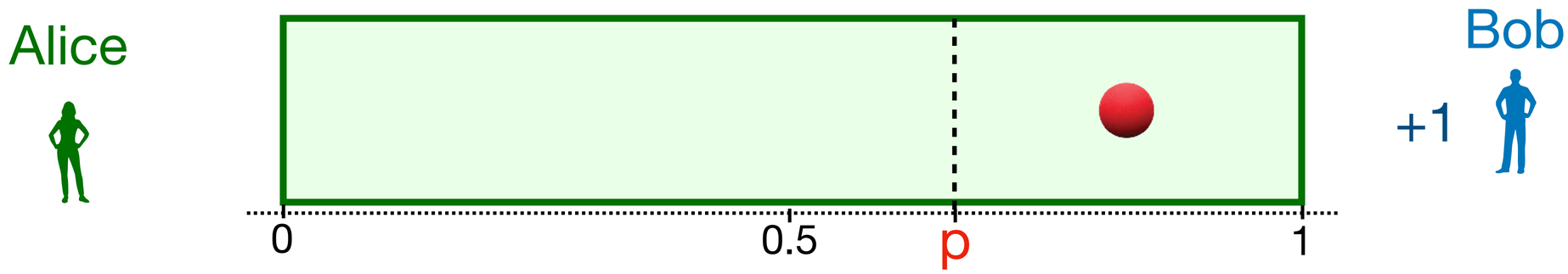
Exemple 2:



1. Sur une table de billard, il y a une séparation abstraite (coordonnée p) invisible et inconnue par les joueurs.
2. À chaque tirage, une balle est jetée (uniformément entre 0 et 1) sur la table.
3. Si elle tombe à gauche de p, Alice gagne un point

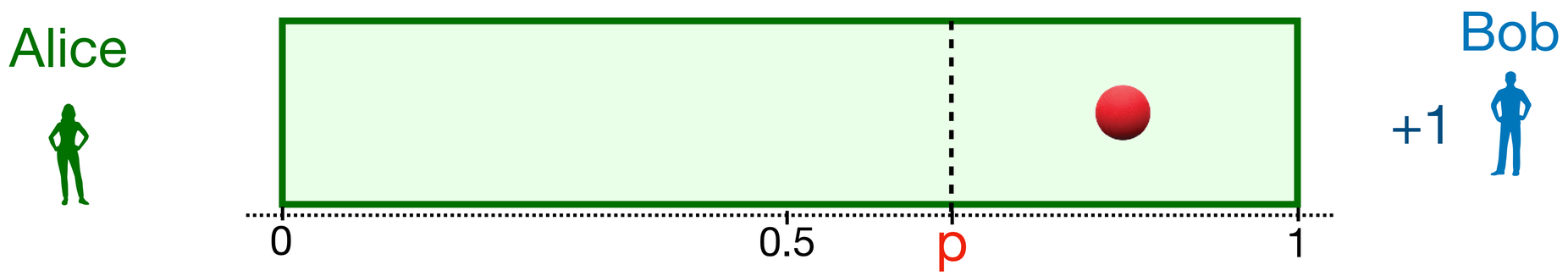
La table de Bayes

Exemple 2:



1. Sur une table de billard, il y a une séparation abstraite (coordonnée p) invisible et inconnue par les joueurs.
2. À chaque tirage, une balle est jetée (uniformément entre 0 et 1) sur la table.
3. Si elle tombe à gauche de p, Alice gagne un point
4. Si elle tombe à droite de p, Bob gagne un point

Exemple 2:



1. Sur une table de billard, il y a une séparation abstraite (coordonnée p) invisible et inconnue par les joueurs.
2. À chaque tirage, une balle est jetée (uniformément entre 0 et 1) sur la table.
3. Si elle tombe à gauche de p, Alice gagne un point
4. Si elle tombe à droite de p, Bob gagne un point
5. Le premier à 6 points gagne. Le score est Alice 5 - 3 Bob. Quelle est la probabilité que Bob gagne ?

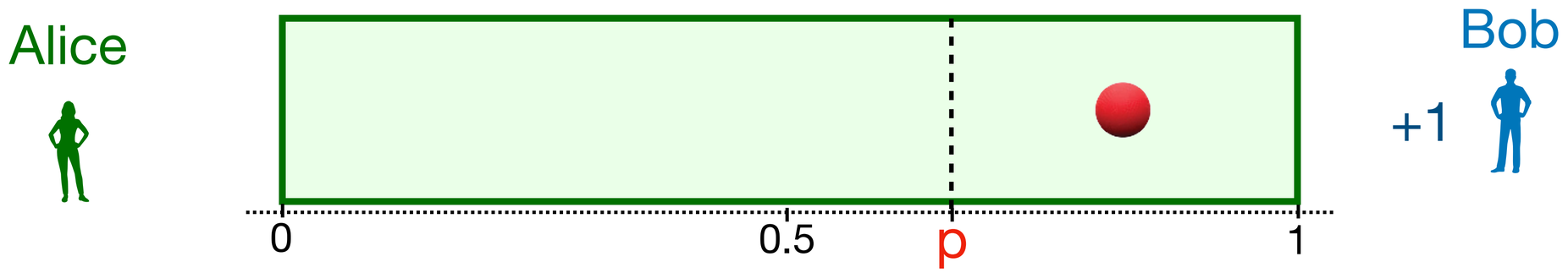
1. Approche fréquentiste

1. Quelle est la probabilité qu'Alice gagne un point lors d'un tirage? En déduire une estimation intuitive fréquentiste de p.
2. Notons A le score d'Alice et B celui de Bob. Quelle distribution permet de modéliser: $\mathbb{P}(A = a, B = b)$?
3. En déduire la vraisemblance du modèle. Le maximum correspond-il à votre estimation de p ?
4. Quelle est la probabilité que Bob gagne la partie ?

1. La probabilité qu'Alice gagne un point est p. Elle a gagné 5 parmi 8 tirages, une estimation de p est 5/8.

4. Pour que Bob gagne, il faut qu'Alice perde 3 fois de suite. La probabilité est donc $(1-p)^3 = (3/8)^3 \sim 0.052$

Exemple 2:



1. Sur une table de billard, il y a une séparation abstraite (coordonnée p) invisible et inconnue par les joueurs.
2. À chaque tirage, une balle est jetée (uniformément entre 0 et 1) sur la table.
3. Si elle tombe à gauche de p, Alice gagne un point
4. Si elle tombe à droite de p, Bob gagne un point
5. Le premier à 6 points gagne. Le score est Alice 5 - 3 Bob. Quelle est la probabilité que Bob gagne ?

2. Approche Bayésienne

1. On suppose une loi a priori uniforme pour p. Déterminez la loi a posteriori $\mathbb{P}(p|A=5, B=3)$.
2. En utilisant la loi a posteriori, proposez un estimateur de la probabilité que Bob gagne sachant le score A = 5, B = 3.

3. Simulation

Simulez ce jeu en Python et estimez cette probabilité empiriquement. Quelle approche est la plus précise ?

Alice

Bob

La distribution a posteriori est donnée par:

$$\mathbb{P}(p|A=5,B=3) = \frac{\mathbb{P}(A=5,B=3|p)\mathbb{P}(p)}{\mathbb{P}(A=5,B=3)} \quad (1)$$

$$\propto \mathbb{P}(A=5,B=3|p)\mathbb{P}(p) \quad (2)$$

$$= \binom{8}{5} p^5(1-p)^3 \mathbb{1}_{[0,1]}(p) \quad (3)$$

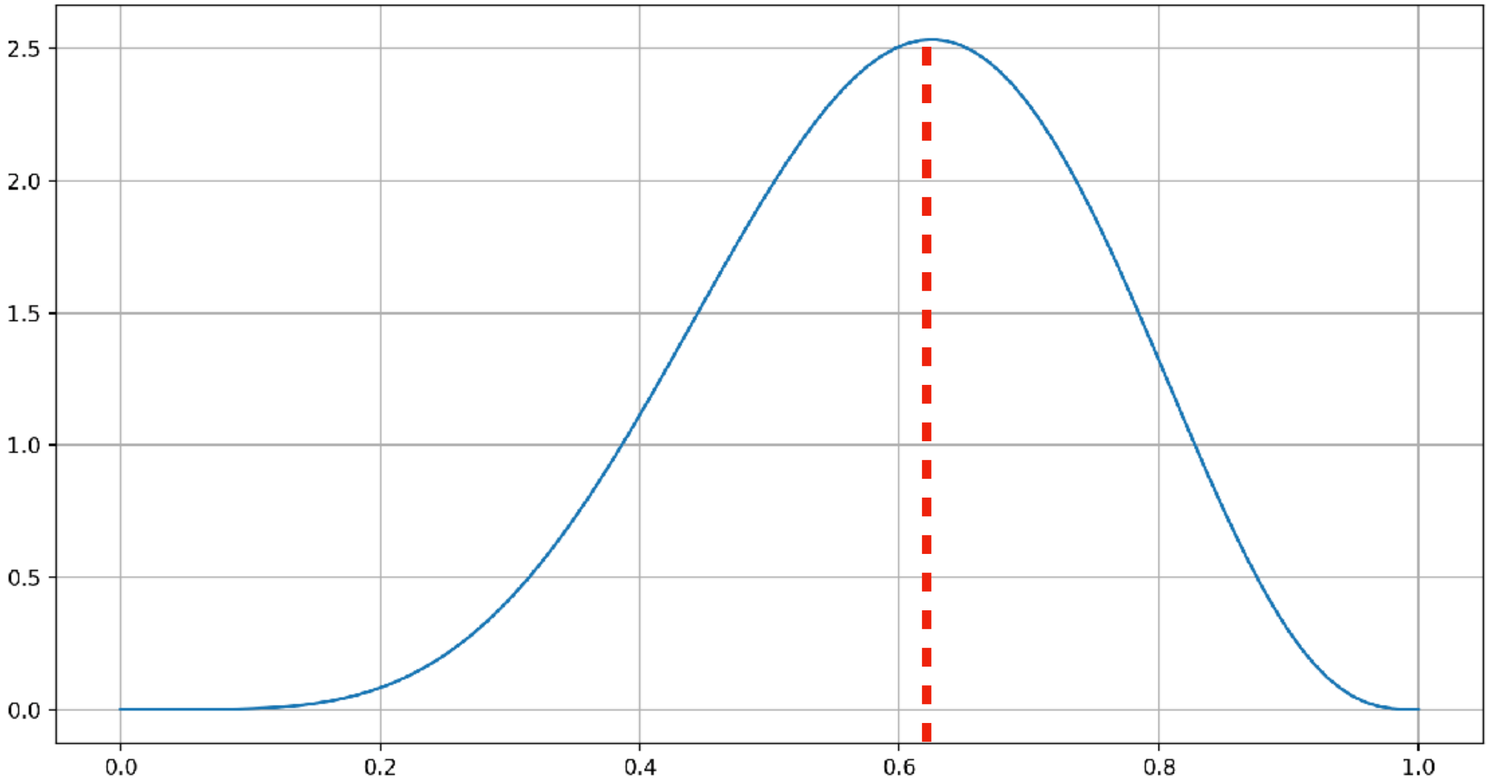
$$\propto p^5(1-p)^3 \quad (4)$$

On reconnait la loi Beta(6, 4).
Sa constante de normalisation est: $\frac{\Gamma(6)\Gamma(4)}{\Gamma(10)} = \frac{5!3!}{9!}$

$$\mathbb{P}(p|A=5,B=3) = \frac{9!}{5!3!} p^5(1-p)^3$$

Après avoir vu les donnés, la distribution a posteriori est une Beta(6, 4):

$$\mathbb{P}(p|A=5, B=3) = \frac{9!}{5!3!}p^5(1-p)^3$$



Identique au maximum de vraisemblance, car l'a-priori est constante !

Quelle serait ici la valeur du (MAP): maximum a posteriori ? Le mode d'une Beta(6, 4) est $\frac{6-1}{6+4-2} = \frac{5}{8} = 0.625$.

Et celle de la moyenne a posteriori ? La moyenne d'une Beta(6, 4) est $\frac{6}{6+4} = 0.6$

Bayésien: Probabilité conditionnelle en utilisant la loi des probabilités totales:

Distribution a posteriori

$$\mathbb{P}(\text{Bob Gagne}|A=5,B=3)=\int \mathbb{P}(\text{Bob Gagne}|A=5,B=3,p)\mathrm{d}\mathbb{P}(p|A=5,B=3)$$

Au lieu d'estimer p et substituer dans $(1-p)^3$, le Bayésien calcule une "moyenne a posteriori"

Bayésien: Probabilité conditionnelle en utilisant la loi des probabilités totales:

$$\mathbb{P}(\text{Bob Gagne}|A=5,B=3)=\int \mathbb{P}(\text{Bob Gagne}|A=5,B=3,p)\mathrm{d}\mathbb{P}(p|A=5,B=3)$$

Distribution a posteriori

Au lieu d'estimer p et substituer dans $(1-p)^3$, le Bayésien calcule une "moyenne a posteriori"

Dans l'exemple de Alice - Bob on a:

1. Modélisé θ comme une **variable aléatoire** avec une loi a priori $\mathbb{P}(\theta)$.

2. Calculé la loi a posteriori $\mathbb{P}(\theta|X_1, \ldots, X_n)$:

$$\mathbb{P}(\theta|X_1, \ldots, X_n) \propto \mathbb{P}(X_1, \ldots, X_n|\theta)\mathbb{P}(\theta)$$

"Modèle facile" car on a obtenu une loi a posteriori classique (Beta) dont l'intégrale est connue, sinon n'on aurait pas pu obtenir la densité a posteriori

3. Estimé la quantité $(1-\theta)^3$ en prenant la moyenne a posteriori:

$$\mathbb{E}_{\theta|X_1,\ldots,X_n}\left[(1-\theta)^3\right] = \int_0^1 (1-\theta)^3 f_{\theta|X_1,\ldots,X_n}(\theta)\mathrm{d}\theta$$

4. Ainsi, le bayésien utilise la loi a posteriori pour estimer n'importe quelle quantité: $\varphi(\theta)$ avec

$$\mathbb{E}_{\theta|X_1,\ldots,X_n}\left[\varphi(\theta)\right]$$

1. Pour quel type de modèle obtient-on une loi a posteriori facilement ?
2. Pourquoi la moyenne a posteriori, non pas le maximum ou la médiane a posteriori ?

INSEA

Choix de l'a-priori

1. Pour quel type de modèle obtient-on une loi a posteriori facilement ?

$$f_{\theta|\mathbf{x}} = \frac{f_{\mathbf{x}|\theta}\, f_\theta}{\int f_{\mathbf{x}|\theta}\, f_\theta \,\mathrm{d}\theta} \propto f_{\mathbf{x}|\theta}\, f_\theta$$

1) Prendre une loi a priori uniforme:

1. Une loi a priori uniforme sur Θ: f_θ constante.
2. La loi a posteriori est alors donnée par le modèle $f_{\mathbf{x}|\theta}$ mais en fonction de θ.
3. Comment faire si le domaine de θ n'est pas borné, $\mathbb{R}$ par exemple ?
4. On peut considérer une loi a priori impropre c-à-d à "densité" non-intégrable $\int_\Omega f_\theta = +\infty$ mais...
5. ... à condition que la loi a posteriori obtenue soit propre : $\int_\Omega f_{\mathbf{x}|\theta} f_\theta \mathrm{d}\theta < +\infty$.

Exemple:

On considère $X_1, \ldots, X_n$ i.i.d $\sim \mathcal{N}(\mu, \sigma^2)$ avec σ^2 connue.
Trouver la loi de $\mu | X_1, \ldots, X_n$ en prenant une loi a priori impropre $\pi(\mu) \propto 1$ sur $\mathbb{R}$.

1. Pour quel type de modèle obtient-on une loi a posteriori facilement ?

2) Prendre une loi a priori conjuguée:

On dit qu'une loi *a priori* $\pi(\theta)$ est **conjuguée** pour un modèle $\mathbb{P}(X_1, \dots, X_n \mid \theta)$ si la loi *a posteriori* $\mathbb{P}(\theta \mid X_1, \dots, X_n)$ appartient à la même famille de distributions que $\pi(\theta)$.

Quelques exemples:

Modèle	**Loi a priori**	**Vraisemblance**	**Loi a posteriori**
Beta-Bernoulli	$\theta \sim \text{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$	$X \mid \theta \sim \text{Bernoulli}(\theta)$, $\mathbb{P}(X \mid \theta) = \theta^X(1-\theta)^{1-X}$	$\theta \mid X \sim \text{Beta}(\alpha + X, \beta + 1 - X)$
Beta-Binomial	$\theta \sim \text{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$	$X \mid \theta \sim \text{Bin}(n, \theta)$, $\mathbb{P}(X \mid \theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X}$	$\theta \mid X \sim \text{Beta}(\alpha + X, \beta + n - X)$
Gaussian-Gaussian	$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \propto \exp\left(-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right)$	$X \mid \theta \sim \mathcal{N}(\theta, \sigma^2)$, $\mathbb{P}(X \mid \theta) \propto \exp\left(-\frac{(X-\theta)^2}{2\sigma^2}\right)$	$\theta \mid X \sim \mathcal{N}(\mu_n, \sigma_n^2)$, $\mu_n = \frac{\sigma_0^2 X + \sigma^2 \mu_0}{\sigma_0^2 + \sigma^2}$, $\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}$
Gamma-Poisson	$\theta \sim \text{Gamma}(\alpha, \beta) \propto \theta^{\alpha-1}e^{-\beta\theta}$	$X \mid \theta \sim \text{Poisson}(\theta)$, $\mathbb{P}(X \mid \theta) = \frac{\theta^X e^{-\theta}}{X!}$	$\theta \mid X \sim \text{Gamma}(\alpha + X, \beta + 1)$

Risque fréquentiste minimax

2. Pourquoi la moyenne a posteriori, non pas le maximum ou la médiane a posteriori ?

Idée: développer un critère pour comparer ces estimateurs

D'abord, en analyse fréquentiste:

1. On suppose qu'il existe un vrai paramètre θ tel que les données $X \sim \mathbb{P}_\theta(X)$.
2. On détermine une fonction T (estimateur) appliquée aux données pour estimer θ par $T(X)$.
3. On peut quantitifer sa qualité en calculant une fonction de perte (loss function) $\mathcal{L}(T(X), \theta)$.
4. Choix classiques de $\mathcal{L}$: la perte quadratique $\mathcal{L}(a, b) = (a - b)^2$ ou en valeur absolue: $\mathcal{L}(a, b) = |a - b|$.
5. $\mathcal{L}(T(X), \theta)$ dépend des données, on calcule une perte moyenne dite "Risque": $\mathcal{R}(T, \theta) \overset{\text{def}}{=} \mathbb{E}_X \left[\mathcal{L}(T(X), \theta)\right]$
6. Ce Risque dépend de θ, on veut un estimateur T qui soit bon pour tous les θ.
7. On quantifie sa **pire performance** par le plus grand risque $\mathcal{R}(T) \overset{\text{def}}{=} \max_\theta \mathcal{R}(T, \theta)$. "worst-case" analysis
8. Le meilleur estimateur T est celui qui minimise ce **plus grand risque** $\min_T \max_\theta \mathcal{R}(T, \theta)$.

"critère mini-max"

Risque bayésien et risque a posteriori

Le fréquentiste cherche l'estimateur T qui minimise: $\max_\theta \mathcal{R}(T,\theta) = \max_\theta \mathbb{E}_X\left[\mathcal{L}(T(X),\theta)\right]$

Que devrait-être le critère du Bayésien ?

Le Bayésien ne maximise pas, il "moyenne" avec la loi a priori:

$$\mathcal{R}_\pi(T) \overset{\text{def}}{=} \int \mathcal{R}(T,\theta)\pi(\theta)\mathrm{d}\theta = \int \left(\mathbb{E}_X\left[\mathcal{L}(T(X),\theta)\right]\right)\pi(\theta)\mathrm{d}\theta \quad = \mathbb{E}_{\theta\sim\pi}\left(\mathbb{E}_X\left[\mathcal{L}(T(X),\theta)\right]\right)$$

Risque de Bayes pour la loi a priori π

Le Bayésien cherche l'estimateur T qui minimise: $\mathcal{R}_\pi(T)$ Un tel estimateur est dit: "Estimateur de Bayes"

Qu'obtient-on si on intègre la perte par rapport a la loi a-posteriori ?

$$\rho_\pi(T,X) \overset{\text{def}}{=} \int \mathcal{L}(T(X),\theta)\mathrm{d}\mathbb{P}(\theta|X)$$ On intègre par rapport à θ uniquement, ce risque dépend des données !

Risque a posteriori

Et si on le minimise par rapport à T, on obtient une fonction en X: un estimateur !

Théorème

L'estimateur: $\arg\min_T \rho_\pi(T,X)$, s'il existe, est un estimateur de Bayes: il minimise $\mathcal{R}_\pi(T)$ pour la même perte $\mathcal{L}$.

Preuve facile: appliquer le théorème de Bayes puis changer l'ordre de l'intégrale double avec le théorème de Fubini (notes UPMC)

Estimateur de Bayes

Corollaire 1

On considère la perte quadratique $\mathcal{L}(a,b) = \|a-b\|^2$ et $\theta \in \Theta \subset \mathbb{R}^d$ avec une loi a priori π sur Θ. On suppose que la loi π admet un moment d'ordre 2. Un estimateur de Bayes pour la loi π est donné par la moyenne a posteriori:

$$\widehat{\theta}(X) = \int \theta dP(\theta|X)$$

Preuve (au tableau)

Corollaire 2

On considère la perte en valeur absolue $\mathcal{L}(a,b) = |a-b|$ et $\theta \in \Theta \subset \mathbb{R}$ avec une loi a priori π sur Θ. On suppose que la loi π admet un moment d'ordre 1. Un estimateur de Bayes pour la loi π est donné par la médiane a posteriori:

$$\widehat{\theta}(X) = med(P(\theta|X))$$

Preuve facile: il suffit d'écrire la définition de l'intégrale de la valeur absolue et de la découper à la médiane

Si la perte n'est pas précisée, l'expression "Estimateur de Bayes" désigne toujours la moyenne a posteriori

Estimateur de Bayes asymptotique

Que devient la loi a posteriori lorsque la quantité des données tend vers l'infini ?

Théorème (Bernstein-von Mises)

Soit $X_1, \ldots, X_n$ i.i.d. $\sim P_{\theta_0}$ avec $\theta_0 \in \Theta \subset \mathbb{R}^d$ tel que P_θ un modèle régulier (densité $\in \mathcal{C}^\infty$, $\theta_0 \in \text{int}(\Omega)$).

On note l'estimateur du maximum de vraisemblance par $\widehat{\theta}_{MV}$.

Si la loi a priori est régulière et non nulle en θ_0 et la matrice d'information de Fisher $I(\theta_0)$ est inversible, alors:

La loi a posteriori $\theta | X_1, \ldots, X_n$ est asymptotiquement normale centrée en $\widehat{\theta}_{MV}$:

$$\sqrt{n}\left(\theta - \widehat{\theta}_{MV}\right) | X_1, \ldots, X_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

*Preuve très avancée: Voir 10.2 dans l'ouvrage **Asymptotic Statistics** de van der Vaart (1998).*

Application

On rappelle que l'information de Fisher en dimension 1 est donnée par: $I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$

1. Analysez l'une de ces expressions pour comprendre son intuition.
2. Simulez en Python un modèle Beta-Bernoulli et vérifier le comportement asymptotique donné par le théorème BvM.

Résumé

1. On suppose $X_1, \dots, X_n$ des observations i.i.d suivant une loi paramétrée $\mathbb{P}_\theta$.
2. Le modèle Bayésien considère que θ est une **variable aléatoire** à modéliser par une loi a priori $\mathbb{P}(\theta) = \pi$, la distribution des données (vraisemblance) devient alors conditionnelle: $\mathbb{P}(X|\theta)$.
3. Le modèle Bayésien met à jour la distribution sur θ avec les données et obtient la loi a posteriori $\mathbb{P}(\theta|X_1, \dots, X_n)$ avec le théorème de Bayes: $\mathbb{P}(\theta|X_1, \dots, X_n) \propto \mathbb{P}(X_1, \dots, X_n|\theta)\mathbb{P}(\theta)$
4. Une loi de densité proportionnelle à g est dite impropre si $\int g = +\infty$.
5. On peut très bien considérer des lois a priori impropres (ex. uniforme sur $\mathbb{R}$: $\pi(\theta) \propto 1$ sur $\mathbb{R}$) ...
6. ... si la loi a posteriori est propre c-à-d que **la constante de normalisation** $\int \mathbb{P}(X_1, \dots, X_n|\theta)\mathbb{P}(\theta)\mathrm{d}\theta$ est finie.
7. On dit que la loi a priori est conjuguée pour le modèle $X|\theta$ si elle appartient à la même famille de la loi a posteriori.
8. L'estimateur de Bayes correspond à la moyenne de cette distribution appelée moyenne a posteriori.
9. BvM: La loi a posteriori est asymptotiquement normale de moyenne $\widehat{\theta}_{MV}$ et de variance $I(\theta)^{-1}$.
10. Ainsi, si $n \to +\infty$, la loi a priori devient négligeable: on retrouve l'approche fréquentiste.
11. Calculer cet estimateur nécessite de connaître la constante de normalisation pour avoir une densité.
12. Si le modèle est conjugué, alors ce calcul est déjà connu.
13. Si ce n'est pas le cas, la constante de normalisation est très souvent intractable (difficile à calculer).
14. Avec des échantillons $\theta_1, \dots, \theta_m \sim \theta|X$, on peut estimer une moyenne a posteriori empirique $\frac{1}{m}\sum_{i=1}^m \theta_i$.
15. Pour générer de tels échantillons, on utilise des méthodes dites de Monte Carlo.

Abus de notation: $\mathbb{P}$ d'une variable aléatoire désigne sa densité (continue ou discrète)