Inférence statistique

Biologie et Modélisation

M. Bailly-Bechet

Université Claude Bernard Lyon 1 – France

Table des matières

Table des matières

Modèle

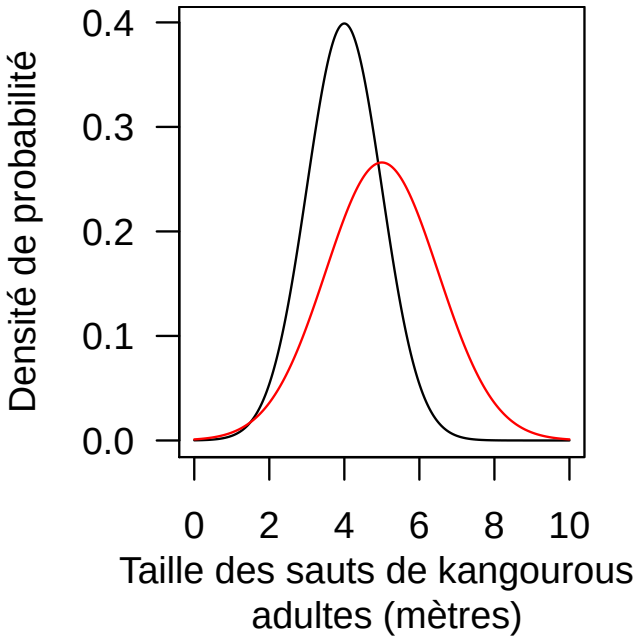
- Un modèle est une formulation mathématique de la réalité. Il est très souvent idéalisé, et ne représente que *partiellement* la réalité.
- Choisir un modèle dans une certaine situation revient à choisir un ensemble d'hypothèses, et donc une forme d'équation pour le problème.
- Il existe toujours un choix au niveau des modèles : les choses ne *sont* pas, on les *définit* en fonction d'hypothèses.
  Rarement, on peut avoir plusieurs formulations mathématiques non équivalentes d'hypothèses similaires, et donc plusieurs modèles.

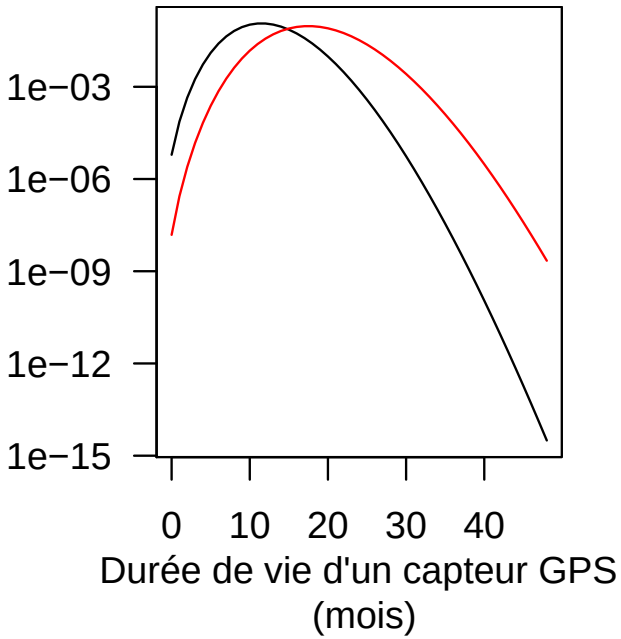
Exemples

- Pour modéliser la taille des sauts chez des kangourous adultes, on dira qu'elle est proportionnelle à la taille de leurs pattes, dont on peut supposer qu'elle suit une loi normale, car elle est multifactorielle et que le théorème central limite s'applique.
- Pour modéliser le temps de vie d'un capteur GPS sur le dos d'un chamois, on va dire q'uà chaque instant, le capteur a une petite probabilité de prendre un choc et d'être détruit. Si on fait le calcul explicite, on se rend compte que la loi du temps de survie suit une loi exponentielle.

Paramètres

- Un paramètre est une variable numérique qui peut prendre un ensemble de valeurs.
- Une loi normale a deux paramètres : la moyenne et la variance.
- Une loi exponentielle a un seul paramètre, qui est à la fois sa moyenne et sa variance.





Que cherche-t-on à faire ?

- Si on cherche à trouver quel modèle et quelles hypothèses correspondent le mieux à la réalité, on fait de la *sélection de modèle*.
- Si on veut trouver la ou les meilleures valeurs des paramètres dans un modèle donné, on fait de l'*inférence statistique*.
- Si on veut vérifier si une valeur particulière des paramètres est conforme avec des données, on réalise des *tests statistiques*.

Dans la suite, on donne deux méthodes d'inférence – la sélection de modèle est un sujet très complexe et n'est pas abordée ici.

Table des matières

Modèle à un paramètre

- On a des données que l'on cherche à modéliser.
- On prend un cas simple : le modèle est choisi, il n'a qu'un seul paramètre.
- Les différentes valeurs du paramètre correspondent à différentes *hypothèses*.

Notion de vraisemblance

La *vraisemblance* d'une hypothèse, c'est la probabilité que les données soient observées dans le cadre de cette hypothèse.
Exemple : je jette une pièce et j'obtiens face.

- L'hypothèse "Cette pièce fait toujours pile", a une vraisemblance de 0 : il n'y a aucune chance de voir une pièce qui fait toujours pile. . . faire face.
- L'hypothèse "Cette pièce fait toujours face", a une vraisemblance de 1 : si cette hypothèse est vraie, on est certains de tirer face à chaque fois avec la pièce.
- L'hypothèse "Cette pièce est non truquée", a une vraisemblance de 0.5 : en effet avec une pièce non truquée la probabilité d'obtenir face est de 0.5.

Quelle hypothèse correspond le mieux à la réalité ?

On veut choisir parmi toutes les hypothèses (toutes les valeurs du paramètre) laquelle correspond le mieux à ce que l'on a observé. Comment choisir entre les hypothèses ? Dans l'exemple précédent, quelle est l'hypothèse qui explique le mieux les données ?

Quelle hypothèse correspond le mieux à la réalité ?

On veut choisir parmi toutes les hypothèses (toutes les valeurs du paramètre) laquelle correspond le mieux à ce que l'on a observé. Comment choisir entre les hypothèses ? Dans l'exemple précédent, quelle est l'hypothèse qui explique le mieux les données ?

C'est l'hypothèse "Cette pièce fait toujours face" – contre-intuitif, n'est-ce-pas ?

Maximum de vraisemblance

Le principe du maximum de vraisemblance dit que la valeur du paramètre qui doit être choisie est celle pour laquelle la vraisemblance est maximale.

Mathématiquement, on a :

$$\theta^\star = \text{argmax}_\theta P(obs|\theta, \mathcal{M}) \tag{1}$$

On peut démontrer (Cours de Biostatistiques-MIV, L3) que cette manière de choisir nous donnera la valeur "réelle" si le nombre d'observations est très grand. Dans le cas précédent, $n = 1\ldots$

Un exemple médical

Un patient souffre des symptômes suivants [1] :

- Défaillance rénale
- Paralysie partielle

On cherche à trouver un diagnostic pour ces symptômes, mais plusieurs correspondent.

1. Inspiré de la série Dr. House, saison 2, épisode "Le Rasoir d'Occam" – données médicales non vérifiées

Diagnostics

Deux diagnostics sont possibles :

- Un parasite cérébral pourrait expliquer facilement la paralysie et, éventuellement, expliquer la défaillance rénale. On considère que la probabilité d'avoir ces 2 symptômes avec un parasite cérébral est de 1/100.
- D'autre part, un lupus (maladie auto-immune) expliquerait parfaitement la défaillance rénale, mais assez mal la paralysie. Ici aussi, la probabilité d'avoir ces symptômes suite à un lupus est de 1/100.

Quel est le diagnostic le plus vraisemblable ?

Diagnostics (II)

On apprend ensuite plusieurs éléments :

Vancomycine Le patient a subi un traitement à la vancomycine contre le parasite cérébral. Ce traitement a déclenché chez lui une allergie profonde et a du être arrêté très tôt. La probabilité que ce nouveau symptôme soit causé par un lupus est de 1/10, alors que la probabilité que ce soit le parasite qui le cause est de 1/1000.

Stéroïdes Un traitement aux stéroïdes destiné à éradiquer le lupus provoque des crises de démence chez le patient. La probabilité que les crises de démence soient causé par le lupus est quasi nulle, alors que le parasite cérébral les cause dans 1 cas sur 100.

Maladie de Wilson Une étude montre que la maladie de Wilson, une maladie génétique très rare, pourrait causer dans certains cas l'ensemble des symptômes observés, avec une probabilité de 1 pour un million.

Quel diagnostic est le plus vraisemblable à chaque étape du traitement ? Est ce que le fait que la maladie génétique soit rare influe sur votre analyse ? Que pensez-vous du diagnostic "Le patient a un parasite cérébral *et* un lupus" ?

Table des matières

Formule de Bayes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

d'où :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

On écrit souvent cette dernière formule :

$$P(\theta|obs) = \frac{P(obs|\theta)P(\theta)}{\sum_\theta P(obs|\theta)P(\theta)} \tag{4}$$

Concept clef 1 : *a priori*

En statistiques bayesiennes, on considère, en plus des données récoltées dans le cadre d'une expérience, un *a priori* sur le paramètre θ que l'on cherche à estimer. C'est le terme $P(\theta)$.

Cela peut permettre d'inclure dans les analyses des résultats précédents, formels ou non. En pratique, toute la difficulté consiste à estimer de manière correcte nos *a priori*.

Concept clef 2 : distributions

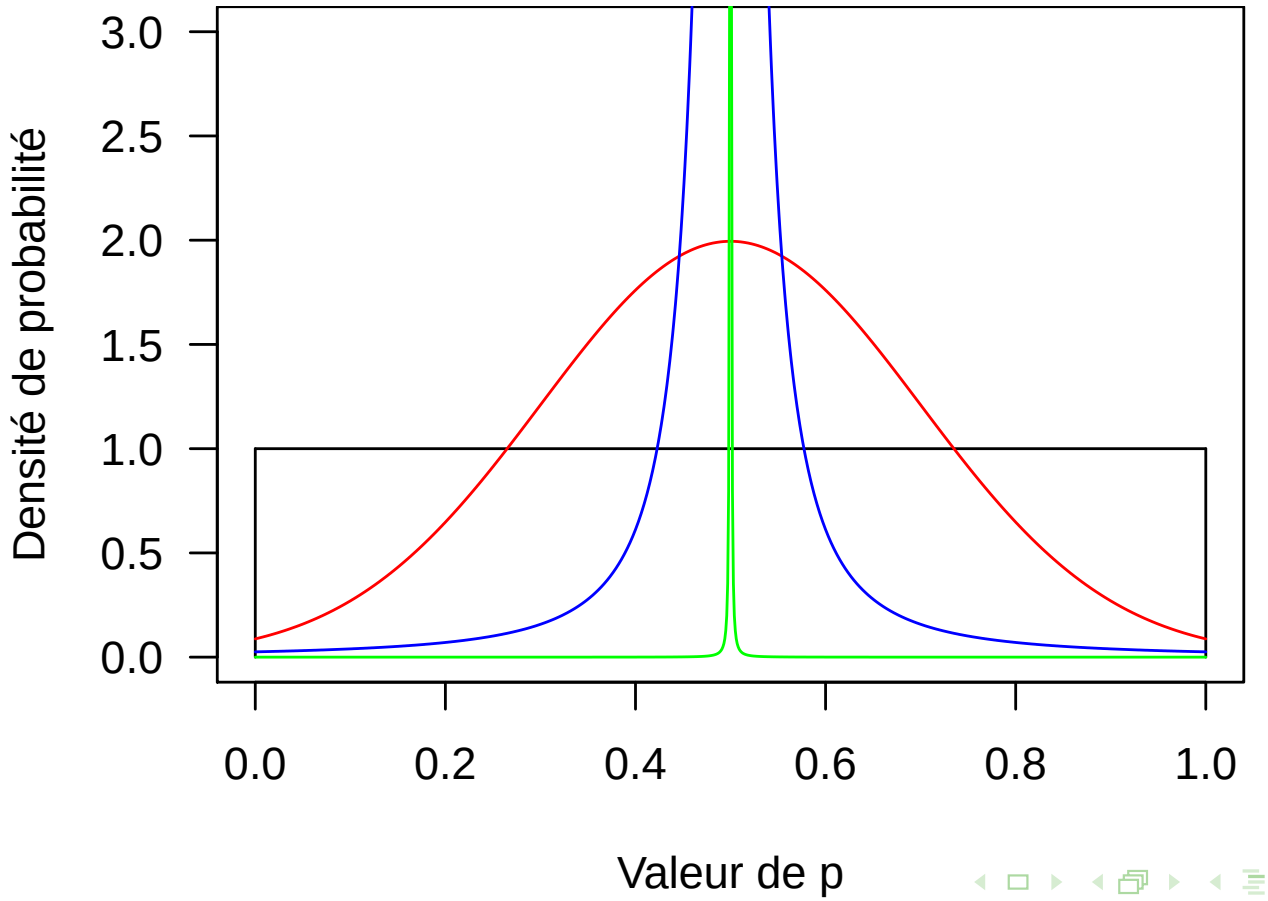
En statistiques bayesiennnes, on va chercher non pas la "meilleure valeur" d'un paramètre, comme dans le maximum de vraisemblance, mais on va estimer la distribution de probabilité de ce paramètre. On peut imaginer que l'on ne cherche pas à estimer la valeur du paramètre sur un cas, mais sur un grand ensemble de cas pour lesquelles les valeurs divergent. Si on fait suffisamment d'expériences, la distribution de probabilité du paramètre peut être tellement pointue qu'en pratique on considèrera une unique valeur.

Le retour de l'exemple de la pièce

On considère souvent qu'une pièce, *a priori*, a 50% de chances de faire pile et 50% de chances de faire face. Avant même de la regarder. Dans les statistiques bayesiennes, cela veut dire que l'*a priori* que l'on a est centré sur 0.5, et que l'on va préférentiellement utiliser une des distributions présentées ci-après. Par contre, le niveau de certitude que l'on a a l'avance dans le fait que la pièce va faire pile une fois sur deux sera représenté par la largeur de la distribution.

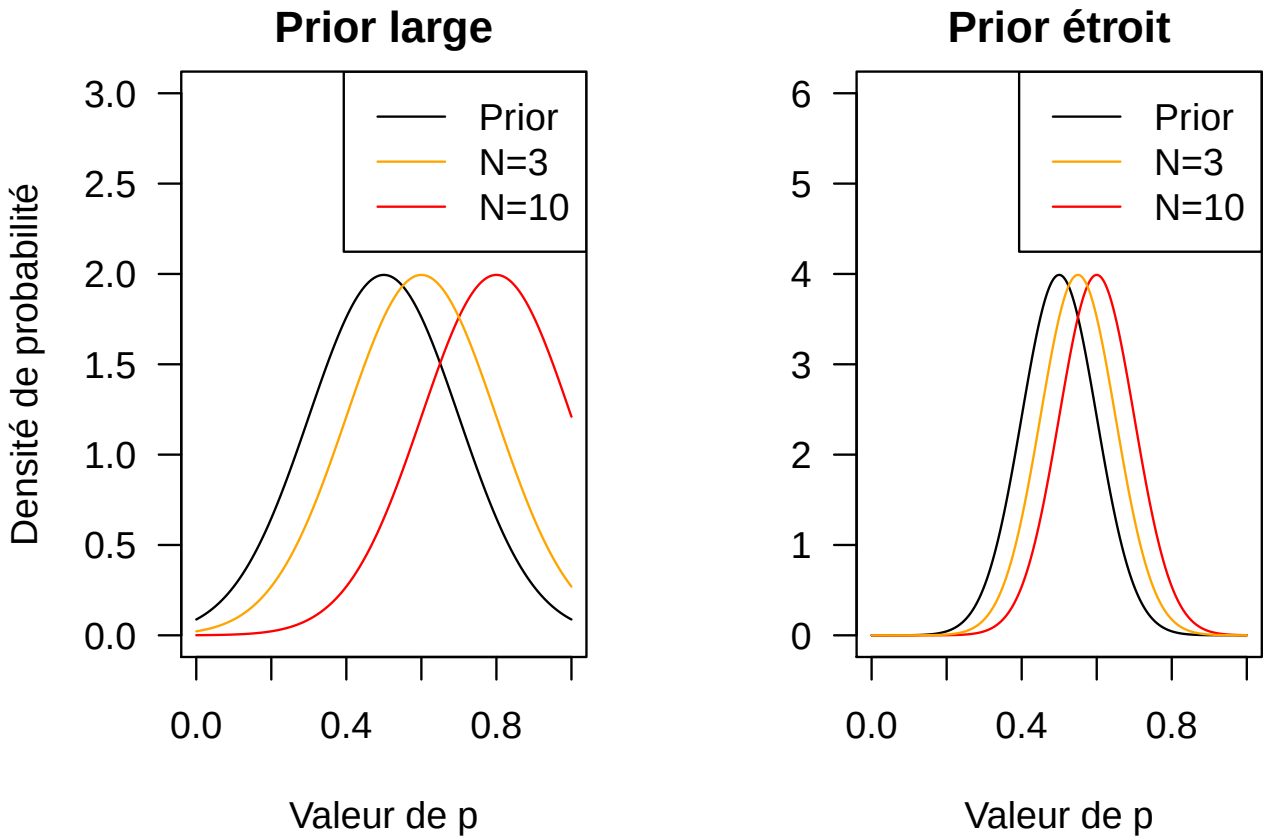
Distribution de probabilité d'un paramètre : exemple de la pièce

Voici plusieurs exemples de distributions de probabilité symétriques pour la probabilité qu'aurait une pièce de faire face :



Effet de l'*a priori*

Voila deux cas avec des *a priori* différents, et les résultats de deux expériences où on lance N fois une pièce, qui fait toujours face. p est la probabilité estimée que la pièce fasse face :



Limites du modèle bayesien

- Aucune expérience ne pourra modifier des certitudes absolues, comme un *a priori* $P(\theta = \theta_0) = 0$.
- Le résultat sera toujours un mélange entre la distribution du paramètre *a priori* et les résultats expérimentaux : il faut peser soigneusement les deux parties.
- La manière de travailler bayesienne impose souvent de travailler numériquement, sauf dans quelques cas bien particuliers (i.e. un *a priori* normal donne un *a posteriori* normal).
- Le contexte bayesien pose des problèmes théoriques de fond, comme le fait de savoir si on a de telles analyses ont un sens ou pas (paradoxes bayesiens).