Probabilité des données vs. probabilité du modèle

Avantages de la décision bayésienne en psychologie

Yvonnick Noël

Université de Rennes 2

Markstein, 28-30 Juin 2011

Sommaire

Une expérience de télékinésie

- Quelqu'un affirme pouvoir faire sortir le même côté d'une pièce (que ce soit pile ou face) plus souvent que le hasard ne le laisserait attendre, par la « force de sa pensée ».
- Vous le mettez à l'épreuve : sur 200 lancers d'une pièce équilibrée, à l'aide d'une machine, 115 sont tombés sur le côté « pile ».
- Diriez-vous que cette personne dispose d'un « pouvoir psychique » ?

Modélisation

- On ne peut répondre à cette question qu'en construisant un **modèle de probabilité** pour la situation.
- Le modèle le plus simple consiste à considérer que le sujet n'a pas de pouvoir. La probabilité de voir le côté pile sortir à chaque essai est alors **connue** : $\pi_0 = \frac{1}{2}$.
- Si ce modèle est vrai, la probabilité que le nombre X d'apparitions de « pile » soit k ($k = 0, ..., 200$) est **binomiale** :

$$P(X = k|\pi_0) = C_N^k \pi_0^k (1 - \pi_0)^{N-k}$$

ou en abrégé $X \sim B(N, \pi_0)$.

Notion de valeur p

- Pour mettre à l'épreuve ce modèle, il est classique de calculer la valeur p : c'est la probabilité d'observer un résultat **au moins aussi élevé**, selon ce modèle.
- On a ici :

$$
\begin{aligned}
p &= 2 \times P(X \geq 115|\pi_0) \\
&= 2\left[P(X = 115|\pi_0) + P(X = 116|\pi_0) + ... + P(X = 200|\pi_0)\right] \\
&= 2 \sum_{k=115}^{200} C_{200}^{k} \left(\frac{1}{2}\right)^{200} \\
&\approx 0.04
\end{aligned}
$$

Conclusion

- Cette valeur est plutôt **petite** si on la juge par rapport à un seuil arbitraire de $\alpha = 0.05$ par exemple.
- Dans cette situation, on **rejette l'hypothèse** de départ (équiprobabilité des faces), avec moins de 5 chances sur 100 de se tromper.
- Je suis quand même un peu perturbé d'avoir à donner du crédit au « pouvoir psychique »...

La notion d'hypothèse nulle

- C'est Fisher (1925, 1935) qui cristallise et formalise la pratique de la **valeur** p déjà en vigueur (Pearson publie le principe du test de χ^2 en 1900).
- Elle s'appuie sur la définition d'une **hypothèse nulle**, c'est-à-dire celle qui doit être « nullifiée » (rejetée). C'est en général la négation de l'hypothèse à laquelle s'intéresse le chercheur.
- Le résultat de la procédure est soit le rejet de l'hypothèse nulle, soit la **suspension du jugement**, l'hypothèse nulle ne pouvant être acceptée positivement.

Notion d'erreur de décision

- Dans cette approche fishérienne, il n'y a donc qu'**un seul type d'erreur** : l'erreur de rejet de l'hypothèse nulle alors qu'elle était vraie.

Une erreur sur le long terme

- Pour Fisher, la valeur p est essentiellement un **indice de décision** pour rejeter ou non une hypothèse nulle sur un jeu de données particulier.
- Neyman & Pearson (1933) proposent un cadre de réflexion plus fréquentiste sur les tests : il s'agit de minimiser la **probabilité de se tromper à long terme**, sur un grand nombre d'essais du même type.

Erreurs de type I et II

- Neyman & Pearson définissent une **hypothèse de référence**, qui est celle qui intéresse le chercheur et qu'ils notent H_0 (ce n'est donc pas l'hypothèse nulle fishérienne !).
- La procédure de test doit mener à l'une ou l'autre des deux décisions : rejeter **ou accepter** cette hypothèse de référence.
- Il y a donc **deux types d'erreur** de décision, inversement liés :
  - rejeter H_0 alors qu'elle est vraie (erreur de type I ou de type α),
  - accepter H_0 alors qu'elle est fausse (erreur de type II ou de type β).

Une approche comparative

- Dans l'approche fishérienne, on cherche à cumuler des preuves contre l'hypothèse nulle, **sans égard pour la nature de l'alternative**.
- Dans l'approche de Neyman & Pearson :
  - on cherche à opposer et confronter **plusieurs hypothèses** (souvent deux)
  - on **accepte l'une ou l'autre**, selon l'appartenance ou non d'une statistique à un intervalle critique (par exemple $[0; 0.05]$ pour la valeur p).

Définir l'hypothèse de référence (I)

- Au fait, dans notre exemple du lancer de pièce, **quelle est notre hypothèse de référence** ?
- Si je suis un adepte des sciences occultes, mon hypothèse est de référence est que la probabilité π de voir « pile » apparaître est supérieure à $\frac{1}{2}$: $\pi > \pi_0$.
- Le cadre fishérien me convient : je cherche assidûment **tout élément de contradiction** quel qu'il soit avec l'hypothèse nulle...

Définir l'hypothèse de référence (II)

- Si je suis rationaliste, l'hypothèse de référence est l'équiprobabilité des faces :
  - le **cadre fishérien** me pose un problème car je ne peux pas valider l'hypothèse nulle !
  - le **cadre Neyman-Pearson** me conduit à (si je persiste à utiliser la valeur p) :
    - fixer un seuil de décision élevé ($\alpha = 0.20$ par exemple),
    - exiger d'avoir une valeur p **supérieure au seuil** pour donner du crédit à ma théorie (en minimisant ainsi, sans la connaître, la probabilité de me tromper).

Pratiques actuelles

- Les pratiques actuelles semblent **mêler les deux approches** sans plus les distinguer :
  - on ne distingue plus dans le choix des termes H_0 (Neyman-Pearson) et hypothèse nulle (Fisher).
  - on définit son hypothèse de référence par négation de l'hypothèse nulle (Fisher), mais on évoque parfois l'erreur de type II (Neyman-Pearson).
  - la valeur p utilisée comme statistique de décision (Fisher) est parfois appelée « risque » (Neyman-Pearson).
  - ...

Limites des théories classiques (I)

L'approche par valeur p a suscité, et depuis longtemps, de **nombreuses critiques** :

- Le choix de la valeur seuil α est **arbitraire** dans l'approche fishérienne (voir aussi la distinction unilatérale ou bilatérale).
- Calculer la probabilité d'un résultat **au moins aussi extrême** revient à utiliser des évènements qui ne se sont pas produits !
- La valeur p ne permet pas de mesurer le **degré de désaccord** avec H_0 (effet taille d'échantillon).
- La valeur p est un critère qui n'est pas approprié pour **valider** H_0.
- Le raisonnement est **peu intuitif** : plutôt que de calculer $P(D|M)$, on préférerait calculer la probabilité que le modèle soit vrai, au vu des données $P(M|D)$.

Comparaison de deux modèles

- Dans l'exemple du lancer du dé, nous souhaitons **comparer deux modèles** :

$$M_0 : X \sim B(N, \frac{1}{2})$$
$$M_1 : X \sim B(N, \pi),\ \pi \neq \frac{1}{2}$$

- Nous souhaitons **sélectionner le meilleur des deux**, sans en privilégier un *a priori*.
- Nous allons chercher à calculer $P(M_0|D)$ et $P(M_1|D)$, ou $\frac{P(M_1|D)}{P(M_0|D)}$.

Notion de vraisemblance

Définition

On appelle **vraisemblance** d'un modèle statistique M la probabilité $P(D|M)$ d'obtenir les données empiriques telles qu'elles sont d'après ce modèle.

- La vraisemblance est un indice naturel de la **qualité d'ajustement** d'un modèle aux données.
- **Attention** : elle n'est cependant pas à confondre avec la probabilité que le modèle soit vrai !

Rappels de probabilité

- Pour deux évènements A et B, probabilité conditionnelle, conjointe et marginale, sont liées par les formules :

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

- Par conséquent :

$$P(A|B)P(B) = P(B|A)P(A)$$

- et (formule d'**inversion de Bayes**) :

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

Calcul de la probabilité a posteriori

- Nous ne connaissons pas $P(M_m|D)$, $m \in \{0,1\}$.
- Mais nous savons par la **formule d'inversion de Bayes** que :

$$P(M_m|D) = \frac{P(D|M_m)P(M_m)}{P(D)}$$

- avec :
  - $P(D|M_m)$ est la **vraisemblance** du modèle,
  - $P(M_m)$ est la **probabilité a priori** que le modèle m soit correct.
  - $P(D)$ est la **probabilité marginale** d'observer les données telles qu'elles sont.
  - $P(M_m|D)$ est la **probabilité a posteriori** du modèle.

Rapport a posteriori

Le choix de modèle s'appuie sur le **rapport des probabilités a posteriori** :

$$\begin{aligned}\frac{P(M_1|D)}{P(M_0|D)} &= \frac{P(D|M_1)P(M_1)/P(D)}{P(D|M_0)P(M_0)/P(D)} \\ &= \frac{P(D|M_1)}{P(D|M_0)} \times \frac{P(M_1)}{P(M_0)} \\ &= B_{10} \times \frac{P(M_1)}{P(M_0)}\end{aligned}$$

Rapport de vraisemblances

- Incertitude sur les modèles :

$$P(M_1) = P(M_0) = \frac{1}{2}$$

- On a alors :

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{P(D|M_1)}{P(D|M_0)}$$

- Nous déciderons donc de rejeter M_0 si :

$$B_{10} = \frac{P(D|M_1)}{P(D|M_0)} > 1$$

Vraisemblance de M_0

Par définition de la binomiale, la vraisemblance de M_0 est :

$$P(D|M_0) = C_N^k \pi_0^k (1 - \pi_0)^{N-k}$$

avec :

$$\pi_0 = \frac{1}{2}$$
$$N = 200$$
$$k = 115$$

soit

$$P(D|M_0) = C_{200}^{115} \left(\frac{1}{2}\right)^{200} \approx 0.006$$

Vraisemblance de M_1

- Par définition de la binomiale, la vraisemblance de M_1 s'écrirait :

$$P(D|M_1, \pi) = C_N^k \pi^k (1-\pi)^{N-k}$$

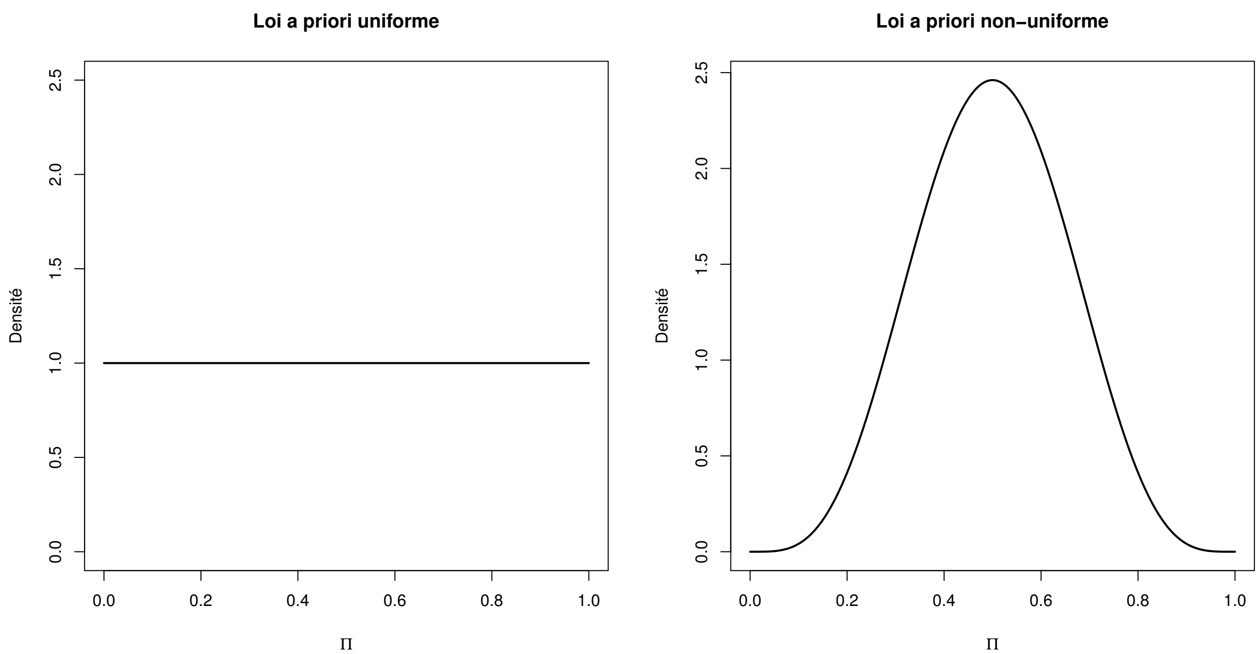
si π, **paramètre inconnu**, était fixé.

- Dans l'approche bayésienne, on traite un paramètre inconnu comme une **variable aléatoire**, en définissant sur lui une densité de probabilité a priori $f(\pi)$.

- On a alors, par le **théorème des probabilités totales** :

$$P(D|M_1) = \int_0^1 P(D|M_1, \pi) f(\pi) d\pi$$

Distribution a priori sur le paramètre



Vraisemblance intégrée

- On a alors :

$$\begin{aligned} P(D|M_1) &= \int_0^1 P(D|M_1,\pi)f(\pi)d\pi \\ &= C_N^k \int_0^1 \pi^k(1-\pi)^{N-k}d\pi \\ &= \frac{1}{N+1} \end{aligned}$$

Facteur de Bayes

- On peut donc calculer le **facteur de Bayes** :

$$B_{01} = \frac{P(D|M_0)}{P(D|M_1)} = \frac{C_N^k \pi_0^k (1-\pi_0)^{N-k}}{1/(N+1)} = \frac{0.006}{0.004975} \approx 1.2$$

- **Interprétation** : le rapport de probabilité a augmenté de 20% en faveur de M_0 **en prenant connaissance des données**.

- Ce n'est pas énorme, mais cela signifie que l'examen des données **ne fait pas** pencher la balancer du côté de M_1.

Schéma de la démarche bayésienne

Dans la démarche bayésienne, on examine la direction et l'amplitude du **changement d'avis** induit par l'examen des données :

1. Avant	2. Examen des données	3. Après
$\frac{P(M_0)}{P(M_1)} = 1$		$\frac{P(M_0 \mid D)}{P(M_1 \mid D)} = 1.2$

B_{01}	Evidence
1 à 3	Négligeable
3 à 20	Positive
20 à 150	Forte
>150	Très forte

Probabilités a posteriori

- Pour chaque modèle ($m = 0, 1$), on peut calculer les **probabilités a posteriori** :

$$\begin{aligned} P(M_m \mid D) &= \frac{P(D \mid M_m)P(M_m)}{P(D)} \\ &= \frac{P(D \mid M_m)P(M_m)}{P(D \mid M_0)P(M_0) + P(D \mid M_1)P(M_1)} \\ &= \frac{P(D \mid M_m)}{P(D \mid M_0) + P(D \mid M_1)} \end{aligned}$$

- Soit :

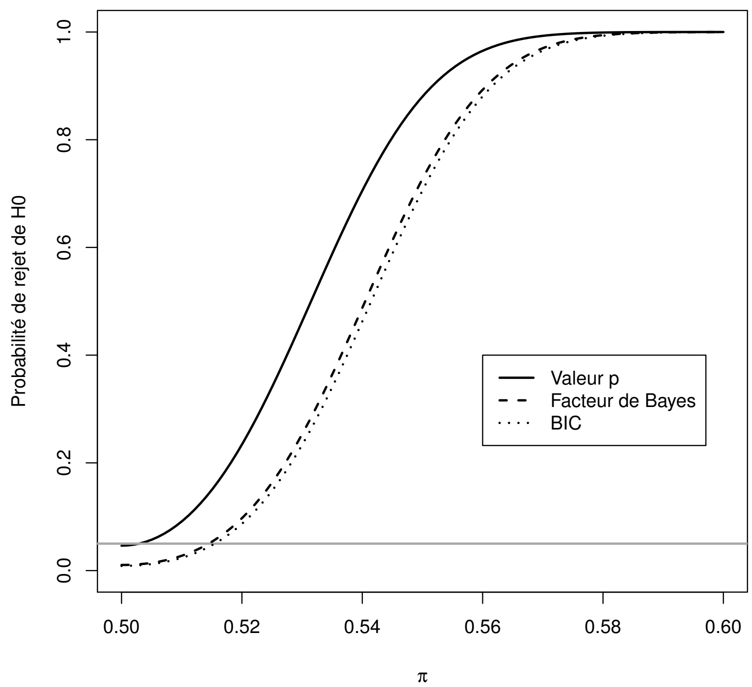
$$P(M_0 \mid D) = \frac{0.006}{0.006 + 0.004975124} \approx 0.545$$

$$P(M_1 \mid D) = \frac{0.004975124}{0.006 + 0.004975124} \approx 0.455$$

En résumé

- On note qu'il s'agit bien de la **probabilité que le modèle soit vrai**.
- S'il faut choisir entre les deux modèles, nous dirons donc que M_0 est **plus probablement vrai**.
- On cherchera en outre à quantifier l'amplitude de l'effet ou poids de l'**évidence** par le facteur de Bayes (>3, Jeffreys, 1961).

Consistance du facteur de Bayes



Logiciel

- Les calculs sont réalisés par la librairie AtelieR sous R :

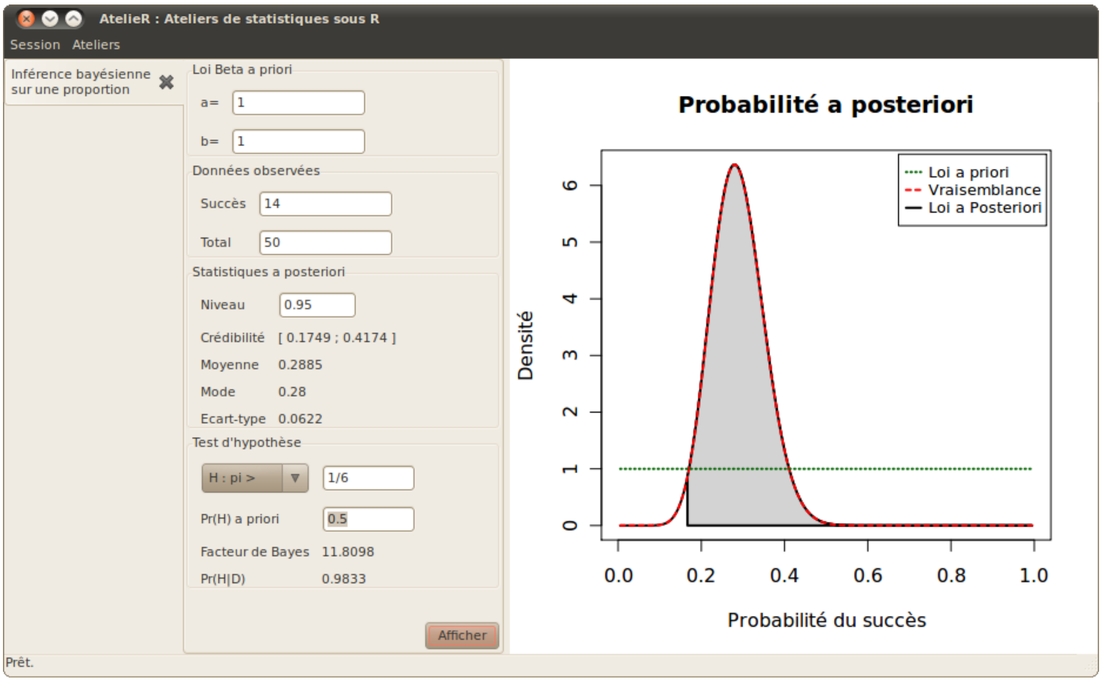
  https://r-forge.r-project.org/projects/atelier/

- Cliquez sur le lien à droite « R packages »

Identification de criminel

Dans les enquêtes criminelles, on présente parfois des photos de suspects. Certains artifices de présentation peuvent-ils orienter la réponse de sujets qui n'ont en réalité absolument rien vu (Busey & Loftus, 2006) ?

Application sous R



Antécédents d'abus sexuels et délinquance

- Smith & Ireland (2005) étudient l'apparition de **toxicomanie** en relation avec la **maltraitance** chez les filles.
- 3 groupes sont distingués : i) celles qui n'ont pas été maltraitées, ii) celles qui ont été victimes d'abus physique non sexuel, et iii) celles qui ont été victimes d'abus sexuels.
- On fait dans cette étude **deux hypothèses** :
  1. il y a bien un lien entre maltraitance et le devenir toxicomane chez les filles.
  2. l'abus n'a cet effet que s'il est de nature spécifiquement sexuelle.

Données et modèle (Smith & Ireland, 2005)

Les données recueillies sont les suivantes :

Groupe	1. Non	2. Physique	3. Sexuel
#Toxic.	48	20	12
Total	145	51	15
Freq.	0.331	0.392	0.800

On cherche un modèle sur :

Groupe	1. Non	2. Physique	3. Sexuel
Pr(Toxic.)	π_1	π_2	π_3

Données et modèle (Smith & Ireland, 2005)

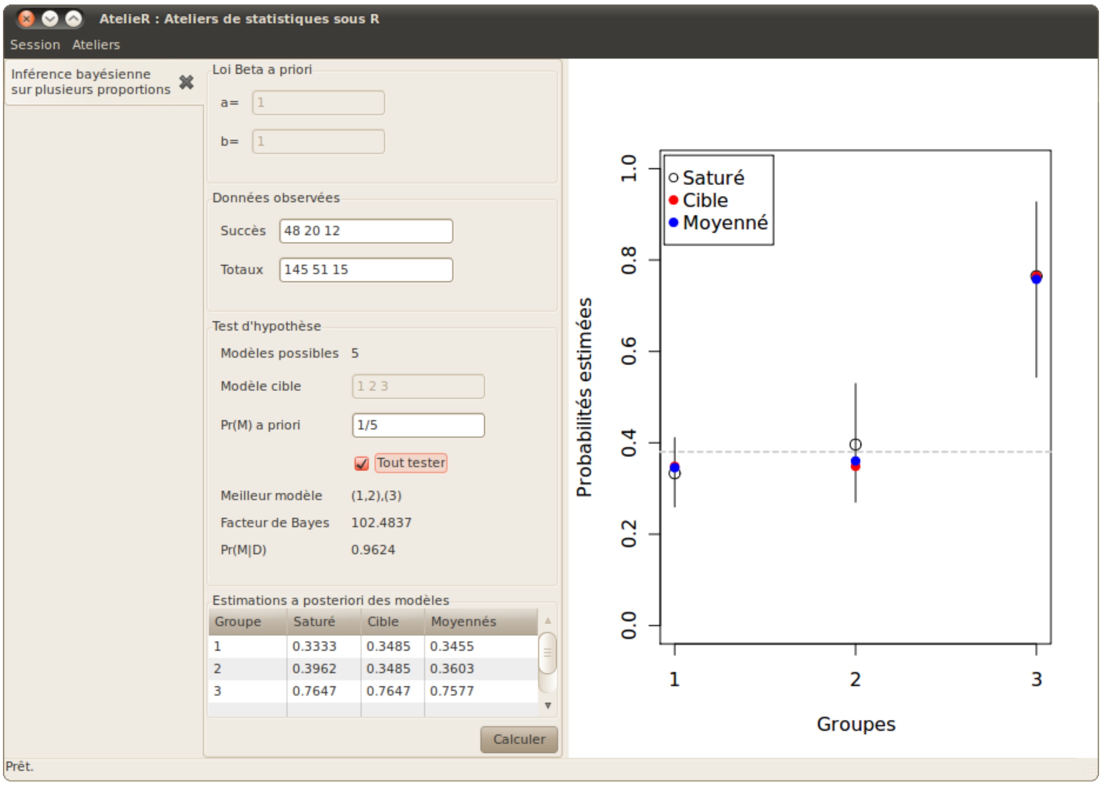
Les données recueillies sont les suivantes :

Groupe	1. Non	2. Physique	3. Sexuel
#Toxic.	48	20	12
Total	145	51	15
Freq.	0.331	0.392	0.800

Le modèle théorique attendu est :

Groupe	1. Non	2. Physique	3. Sexuel
Pr(Toxic.)	π_{12}	π_{12}	π_3

Application sous R



Lien des facteurs de Bayes

- Quand plus de deux modèles sont possibles, il est suffisant d'examiner les facteurs de Bayes de ces modèles contre M_0 car :

$$B_{21} = \frac{P(D|M_2)}{P(D|M_1)} = \frac{P(D|M_2)}{P(D|M_0)} \times \frac{P(D|M_0)}{P(D|M_1)} = \frac{B_{20}}{B_{10}}$$

Analyse de tables de contingence

Cooper & al. (2004) compare 3 modes d'accompagnement de fumeurs qui essaient d'arrêter de fumer avec des patches à la nicotine :

- A. Message général d'encouragement à l'arrêt, et **traitement par patches** de six semaines.
- B. Accompagnement A + **12 appels téléphoniques** et **6 courriers personnalisés** pendant un an.
- C. Accompagnement B + **quatre entretiens de face à face** avec un éducateur de santé.

On mesure le **degré d'adhésion** des patients au traitement : complète, partielle, ou aucune (à partir du nombre de patches effectivement utilisés au bout de 7 semaines de traitement).

Données

La répartition des sujets par condition et niveaux d'adhésion est la suivante :

Cond. / Adhés.	Complète	Partielle	Aucune	Total
A	46	93	42	181
B	53	82	28	163
C	46	86	9	141
Total	145	261	79	485

Modèles possibles

	M_s			
Cond.	Comp.	Part.	Non	$\sum$
A	π_1	π_2	π_3	1
B	π_4	π_5	π_6	1
C	π_7	π_8	π_9	1

Modèle de la différence des 3 accompagnements

Modèles possibles

	M_1			
Cond.	Comp.	Part.	Non	$\sum$
A	π_{14}	π_{25}	π_{36}	1
B	π_{14}	π_{25}	π_{36}	1
C	π_7	π_8	π_9	1

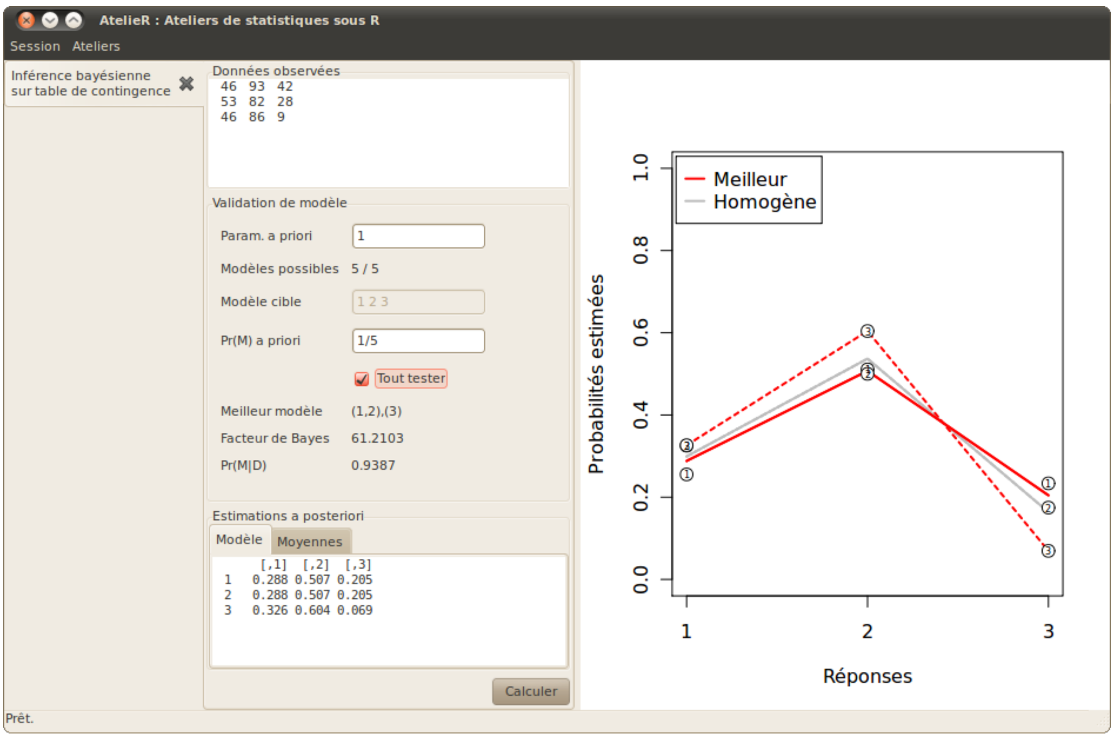
Modèle de l'absence de différence entre accompagnements A et B

Modèles possibles

	M_h			
Cond.	Comp.	Part.	Non	$\sum$
A	π_{147}	π_{258}	π_{369}	1
B	π_{147}	π_{258}	π_{369}	1
C	π_{147}	π_{258}	π_{369}	1

Modèle de l'absence de toute différence

Application sous R



Bibliographie

- Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists : A tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60, 158-189.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Maas, H. L. J. van der. (2011). Why psychologists must change the way they analyze their data : The case of psi. *Journal of Personality and Social Psychology*, 100, 426-432.