Introduction à la statistique bayésienne

TP - Exercice 10

Matthieu TROTREAU

2025-01-16

Contents

Introduction

On veut estimer la proportion p d'étudiants qui dorment plus de 8 heures par nuit. Les observations sur un échantillon de 27 éudiants sont : s = 11 étudiants dorment plus de 8 heures f = 16 étudiants dorment moins de 8 heures.

On note S la variable aléatoire qui représente le nombre d'étudiants qui dorment plus de 8 heures dans un échantillon de taille $n = 27$. On va comparer les résultats de 3 modèles différents donnés par leur loi a priori.

Comparaison des 3 modèles bayésiens

Représentations graphiques des lois a priori

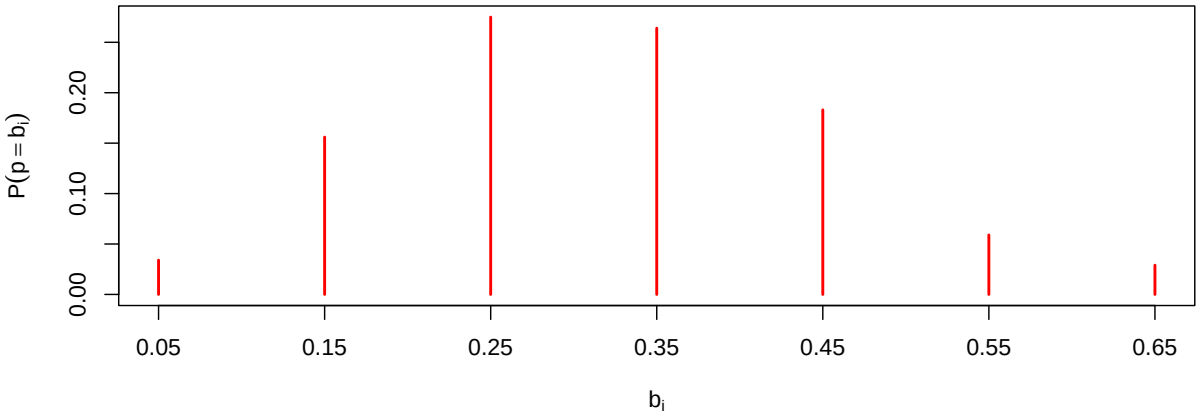


Figure 1: Représentation de la loi a priori du modèle A

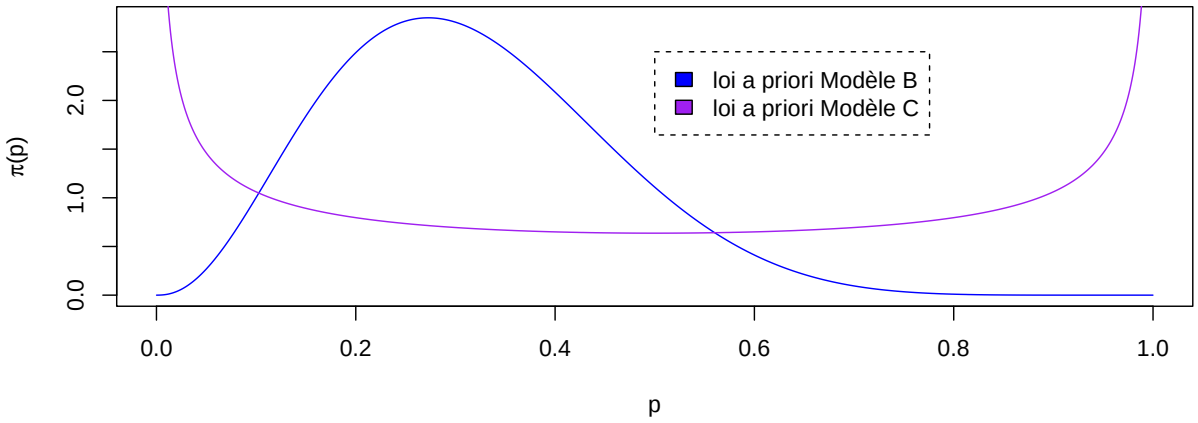


Figure 2: Représentation des lois a priori pour les modèles B et C

Moyennes et variances des lois a priori

On calcule la moyenne de la loi a priori du premier modèle a l'aide de la formule suivante :

$$\sum_{i=1}^{7} b_i \mathbb{P}(p = b_i)$$

Pour les modèles B et C on sait que ces lois a priori sont des lois beta de paramètres respectifs $(3.4, 7.4)$ et $(\frac{1}{2}, \frac{1}{2})$

Le modèle A a pour moyenne 0.316 et pour variance 0.017962. Le modèle B a pour moyenne 0.3148148 et pour variance 0.0182802. Le modèle C a pour moyenne 0.5 et pour variance 0.125.

Les moyennes et variances des modèles A et B sont proches, on peut donc dire que ces deux modèles apportent des informations a priori similaires. De plus la variance de ces deux modèles étant proche de 0 on peut dire qu'en utilisant ces lois a priori on fait confiance à l'information qu'elles apportent.

Le modèle C quant à lui possède une moyenne plus élevée, de même pour sa variance. On considère donc que l'on fait moins confiance à l'information de cette loi a priori. Ce qui fait écho au choix d'une loi a priori non informative pour ce dernier modèle.

Marginales des différents modèles

L'expresion de la marginale du modèle A est la suivante :

$$\mathbb{P}(S = s) = \sum_{i=1}^{7} \mathbb{P}(p = b_i) \binom{27}{s} b_i^s (1 - b_i)^{27-s}$$

Les modèles B et C admettant des lois $\beta(a, b)$ en lois a priori la marginale est donnée est par l'expression suivante :

$$\mathbb{P}(S = s) = \binom{27}{s} \int_0^1 \frac{p^{a-1}(1-p)^{b+27-s-1}}{\beta(a, b)} dp$$

On obtient les facteurs de Bayes suivants : $B_{A/B} = 0.9973497$, $B_{B/C} = 2.8543286$, $B_{A/C} = 2.8467637$.

Les facteurs de Bayes nous permettant de comparer les 3 modèles on peut dire que les modèles A et B sont extrèmement proches. On donnera un léger avantage au modèle B. De plus le modèle C est considéré moins bon que les deux autres, ce modèle est dit moins vraisemblable.

Lois a posteriori

La loi a posteriori du modèle A est donnée par l'expression exacte suivante :

$$\mathbb{P}(p = b_i | S = s) = \frac{\mathbb{P}(p = b_i)\binom{27}{s} b_i^s (1 - b_i)^{27-s}}{\sum_{k=1}^{7} \mathbb{P}(p = b_k)\binom{27}{s} b_k^s (1 - b_k)^{27-s}}$$

Dans le cas de notre experience on sait que $S = 11$.

Ensuite on peut déterminer l'expression exacte de la loi a posteriori pour les modèles B et C, cette expression faisant intervenir la densité de la marginale. Cependant on dira seulement que les lois a posteriori de ces deux modèles sont des lois $\beta(a + s, b + 27 - s)$ avec a et b les paramètres des lois a priori.

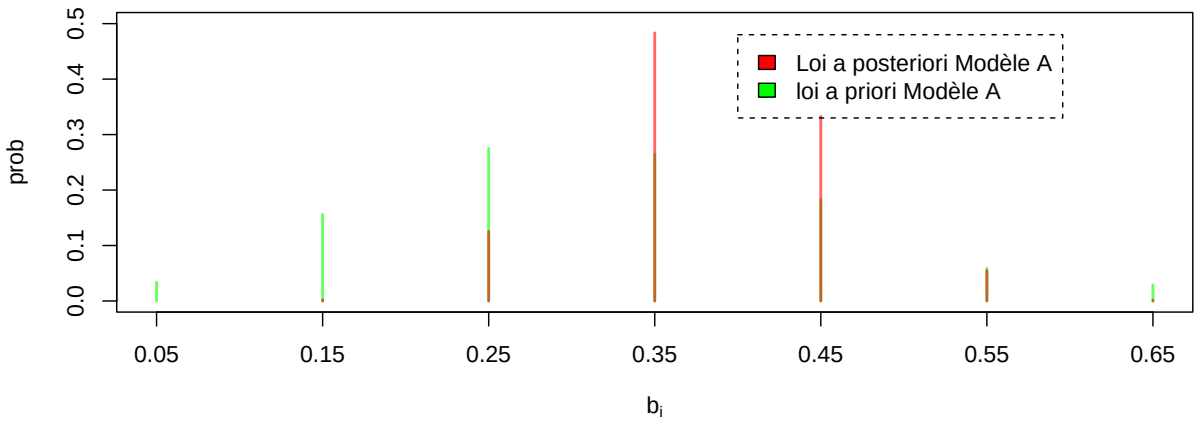


Figure 3: Représentation loi a priori et loi a posteriori pour le modèle A

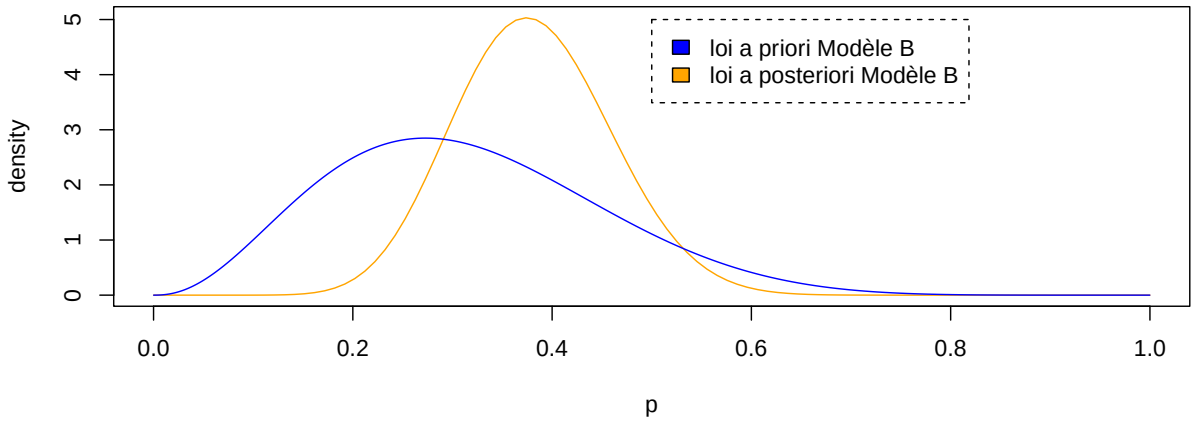


Figure 4: Représentation loi a priori et loi a posteriori pour le modèle B

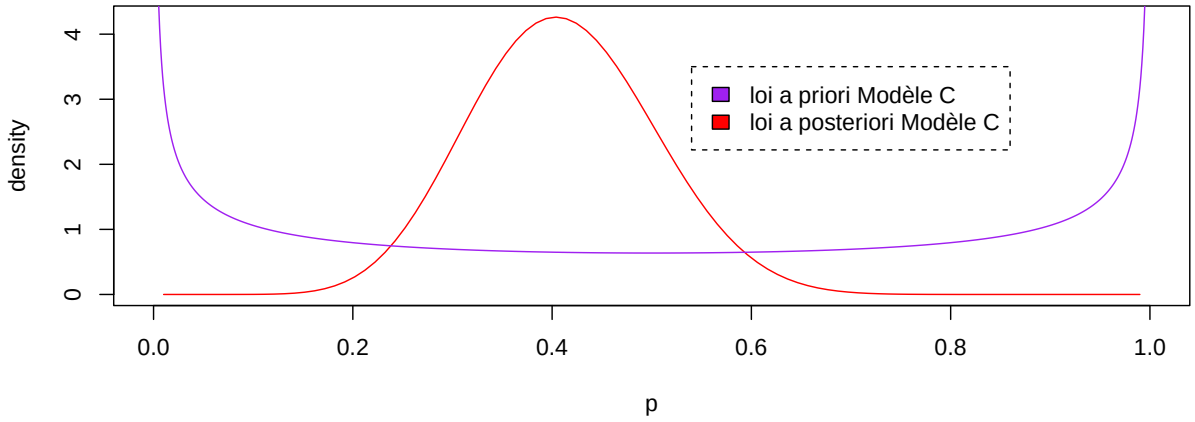


Figure 5: Représentation loi a priori et loi a posteriori pour le modèle C

On a évoqué précédemment que les lois a posteriori des modèles B et C sont des lois beta dont on connaît les paramètres. On applique donc les formules connues de l'espérance et la variance pour des lois beta. La moyenne et la variance du modèle A sont déterminées à l'aide des expressions classiques pour les loi discrètes.

On a les résultats suivants $\mu(\sigma^2)$:
Modèle A : 0.381485 (0.0061243)
Modèle B : 0.6153846 (0.0262985)
Modèle C : 0.6969697 (0.0422406)

Pour les lois a posteriori, les modèles B et C sont proches au niveau de la moyenne, le modèle A quant à lui s'éloigne des deux autres. On observe le même phénomène avec la variance du modèle A qui est très faible par rapport aux deux autres modèles. Mais on a la variance du modèle C 1.6 fois supérieure à celle du B.

On rappelle que l'on avait observé des résultats similaires pour les modèles A et B sur les lois a priori. Pour les lois a posteriori ce sont cette fois-ci les modèles B et C qui sont proches.

Régions de confiance bayésiennes pour le paramètre p

Modèle a priori A

```
loi_post_A_ordered = loi_post_A[order(loi_post_A, decreasing = T)]
cumsum_post_A = cumsum(loi_post_A_ordered)
cumsum_post_A
```

```
## [1] 0.4768594 0.8113229 0.9415095 0.9951603 0.9977506 1.0000000 1.0000000
```

On remarque que 94% de la probabilité se concentre sur les 3 valeurs les plus probables parmi 7. On atteint 99% avec la 4e valeur. On doit donc décider d'inclure, ou non, cette 4e valeur. À première vue, une région dont la probabilité d'appartenance du paramètre est de 0.99 paraît une bonne idée. Cependant cette 4e valeur représenterait environ 5% des 99% tandis que le reste serait concentré sur les 3 autres valeurs. Chosir cette région de niveau $1 - 0.99$ reviendrait à exclure les 3 valeurs dont la probabilité est proche de 0 mais en conserver une dont la probabilité est d'environ 0.05. Je préfère donc conserver une région HPD de niveau exact 94.1509545% qui contient seulement les valeurs de réalisations les plus pertinentes.

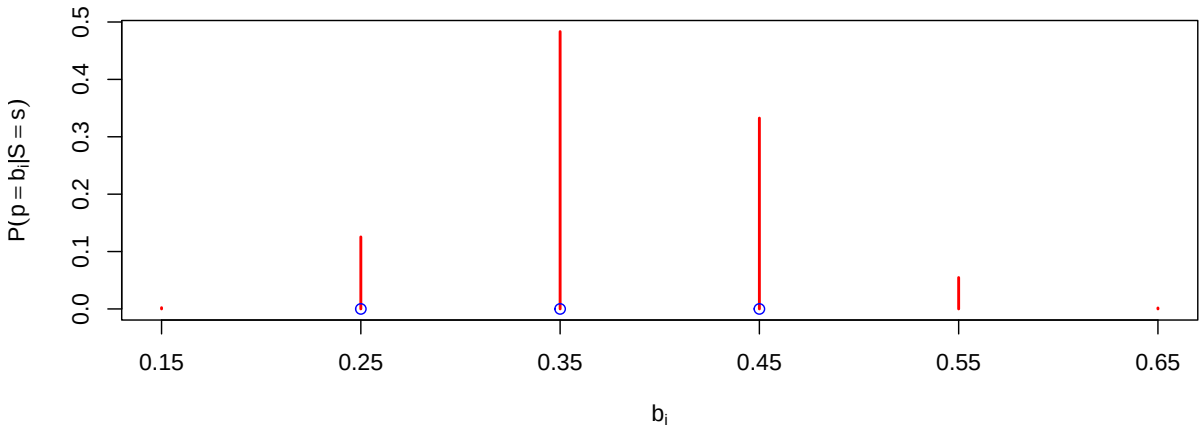


Figure 6: Représentation de la loi a posteriori avec la région HPD pour le modèle A

La région HPD est représentée par les cercles au niveau de l'axe des abscisses.

Modèle a priori B

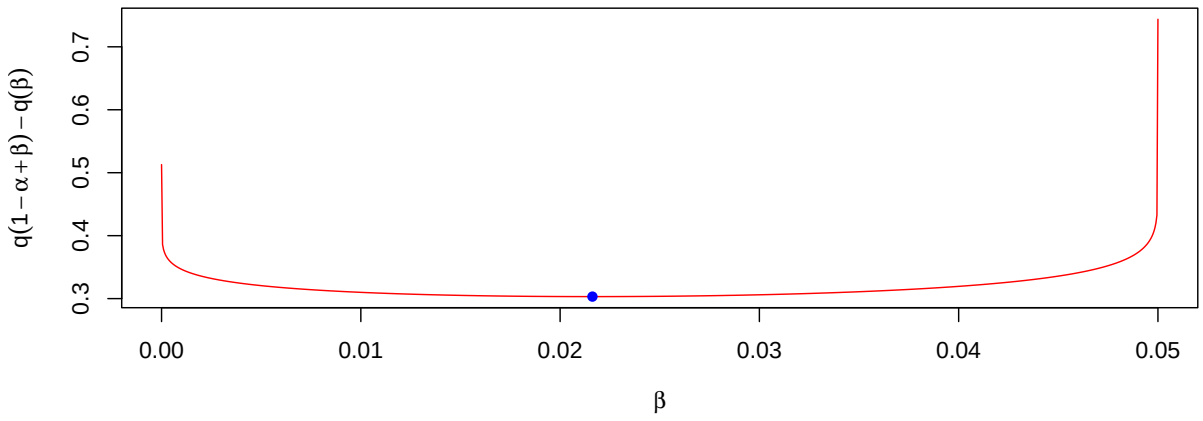


Figure 7: Représentation de la longueur de la région HPD selon β pour le modèle B, avec $\alpha = 0.05$

On trouve un β optimal égal à 0.0216216.

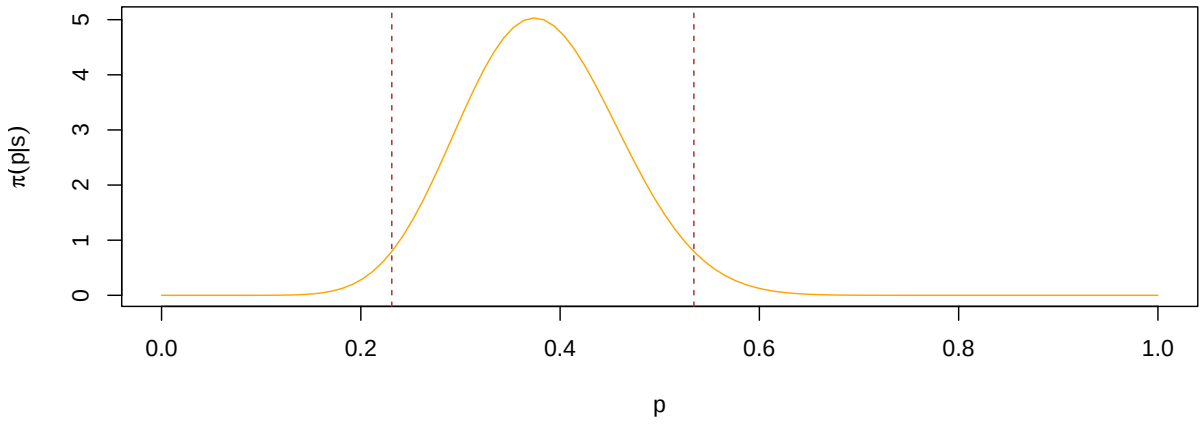


Figure 8: Densité de la loi a posteriori avec la région HPD pour le modèle B

On a une loi a posteriori unimodale donc la région HPD coïncide avec le meilleur intervalle de crédibilité. Pour le modèle B cet intervalle est donc : [0.2310868 , 0.5342663].

Modèle a priori C

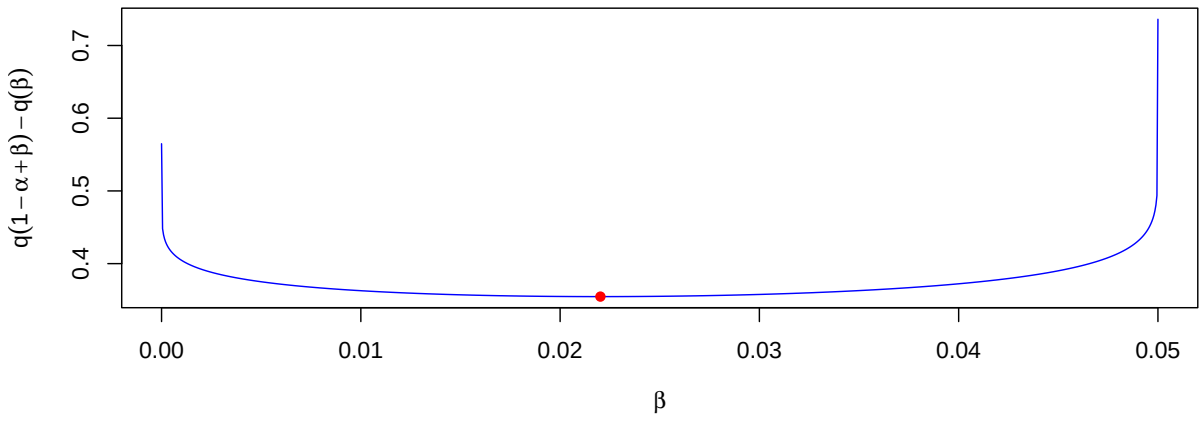


Figure 9: Représentation de la longueur de la région HPD selon β pour le modèle C, avec $\alpha = 0.05$

On trouve un β optimal égal à 0.022022.

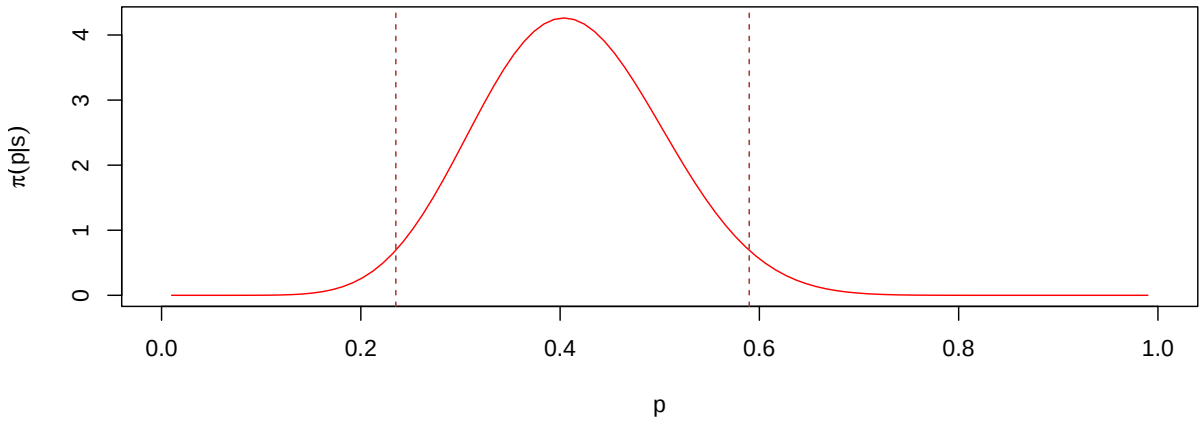


Figure 10: Densité de la loi a posteriori avec la région HPD pour le modèle C

Pour le modèle C la région HPD est donc : [0.2351351 , 0.5897916].

Conclusion

On a deux régions HPD similaires pour les modèles B et C. Cependant la plus courte est celle du modèle B. De plus, les 3 valeurs de la région HPD pour le modèle A appartiennent aux régions HPD des deux autres

modèles. Du fait d'une région HPD plus courte on pourrait choisir de favoriser le modèle B ici aussi. On notera donc que le modèle B paraît meilleur au niveau des facteurs de Bayes ainsi que des régions HPD.

Prévision

Modèle a priori B

On a une loi a posteriori qui correspond à une $\beta(a+s, b+27-s)$. Les fonctions $a(S)$ et $b(S)$ sont donc les suivantes : $a(S) = a + S$, $b(S) = b + 27 - S$, avec a et b les paramètres de la loi Beta a priori. En l'occurence pour le modèle B on a $a = 3.4$ et $b = 7.4$.

On justifie cet algorithme à l'aide de l'expression suivante :

$$\pi(s^*|s) = \int_0^1 f(s^*|p)\pi(p|s)dp$$

En effet, on a une seule observation de S on peut alors considérer que "l'ensemble" de nos observations est iid selon $\pi(.|p)$. Ceci implique que $\pi(s^*|p, s) = f(s^*|p)$, ce qui nous ramène à l'expression précédente de $\pi(s^*|s)$.

On choisit $M = 10000$ afin d'avoir une certaine stabilité au niveau de la simulation. En effet les valeurs des probabilités pour s^* dans {6,7,8,9} étant relativement proches on peut avoir des répartitions assez différentes pour plusieurs simulations si M n'est pas pris suffisamment grand.

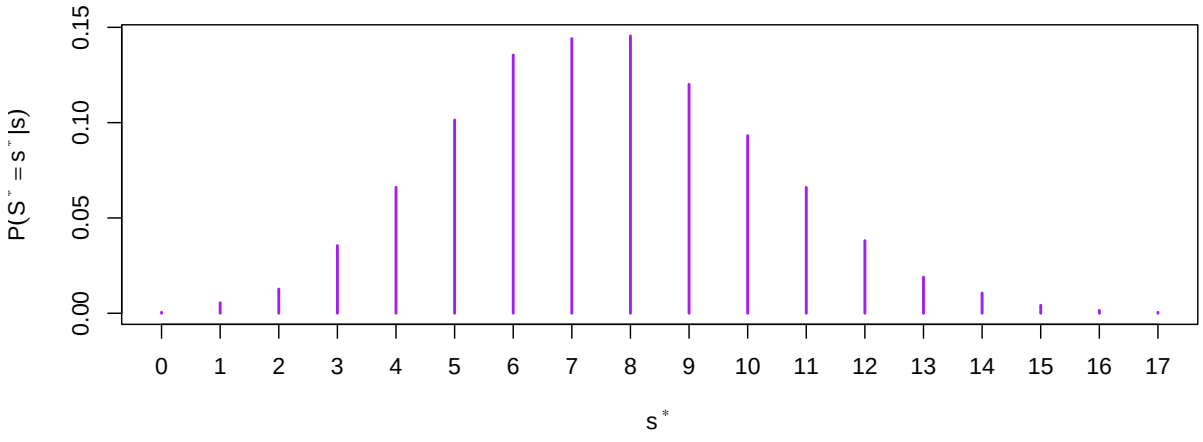


Figure 11: Représentation de la loi prédictive pour le modèle B

```
table_prev_Mod_B_ordered = table_prev_Mod_B[order(table_prev_Mod_B, decreasing = T)]
cumsum_prev_B = cumsum(table_prev_Mod_B_ordered)
cumsum_prev_B
```

```
##      8      7      6      9      5     10      4     11     12      3     13
## 0.1455 0.2896 0.4251 0.5452 0.6466 0.7398 0.8059 0.8719 0.9100 0.9455 0.9645
##      2     14      1     15     16      0     17
## 0.9772 0.9878 0.9933 0.9975 0.9990 0.9995 1.0000
```

On choisit de conserver un niveau exact de 96.45% de manière à obtenir une probabilité d'appartenance au moins égale à 0.95. La somme cumulée atteignant cette fois-ci 0.9645 et non près de 0.99 comme précédemment.

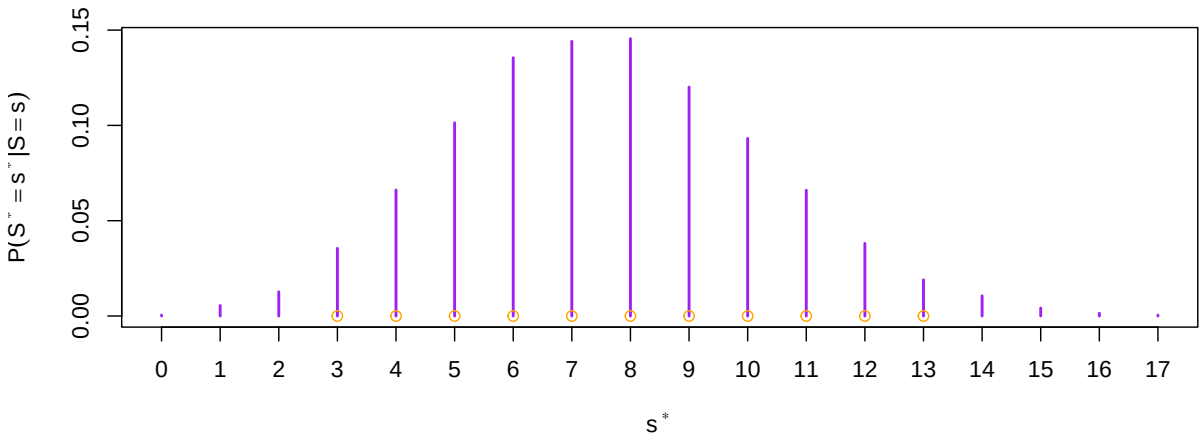


Figure 12: Représentation de la loi prédictive avec sa région HPD pour le modèle B

```
table_prev_Mod_B_DF = as.data.frame(table_prev_Mod_B, stringsAsFactors = F)
supp_prev_Mod_B = as.numeric(table_prev_Mod_B_DF$Prev_Mod_B)

Best_pred_ponct_modB = sum(supp_prev_Mod_B*table_prev_Mod_B_DF$Freq)
```

La meilleure approximation du prédicteur ponctuel pour l'erreur L^2 est égale à $\mathbb{E}[S^*|S]$ qui correspond donc à l'espérance de notre loi prédictive. Pour le modèle B on a $\hat{S^*} = 7.5815$.

Modèle a priori C

Les fonctions $a(S)$ et $b(S)$ sont les mêmes que précédemment, avec $a = b = \frac{1}{2}$.

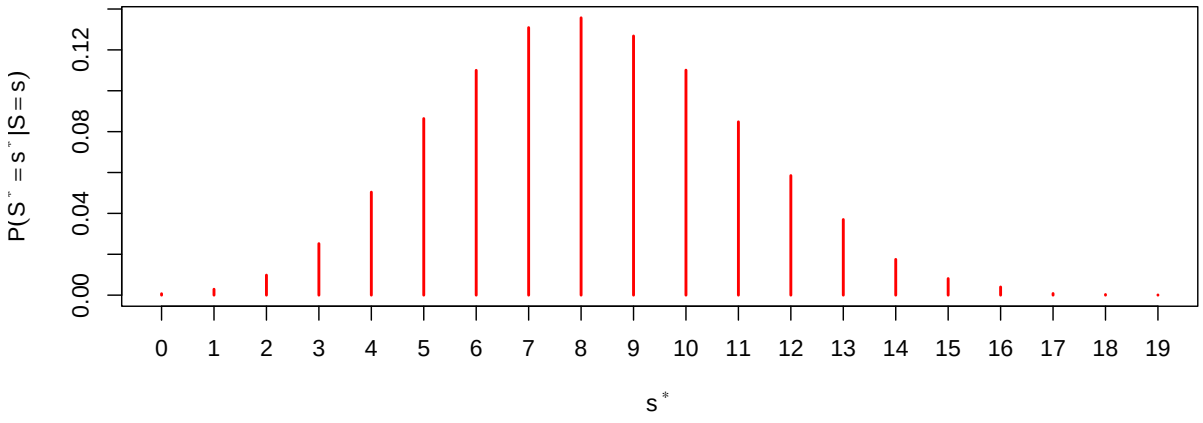


Figure 13: Représentation de la loi prédictive pour le modèle C

```
table_prev_Mod_C_ordered = table_prev_Mod_C[order(table_prev_Mod_C, decreasing = T)]
cumsum_prev_C = cumsum(table_prev_Mod_C_ordered)
cumsum_prev_C
```

```
##      8      7      9     10      6      5     11     12      4     13      3
## 0.1357 0.2666 0.3934 0.5035 0.6135 0.6999 0.7847 0.8432 0.8936 0.9306 0.9558
##     14      2     15     16      1     17      0     18     19
## 0.9733 0.9831 0.9912 0.9952 0.9981 0.9989 0.9996 0.9999 1.0000
```

On obtient une région HPD exactement de niveau 95.58% pour ce modèle.

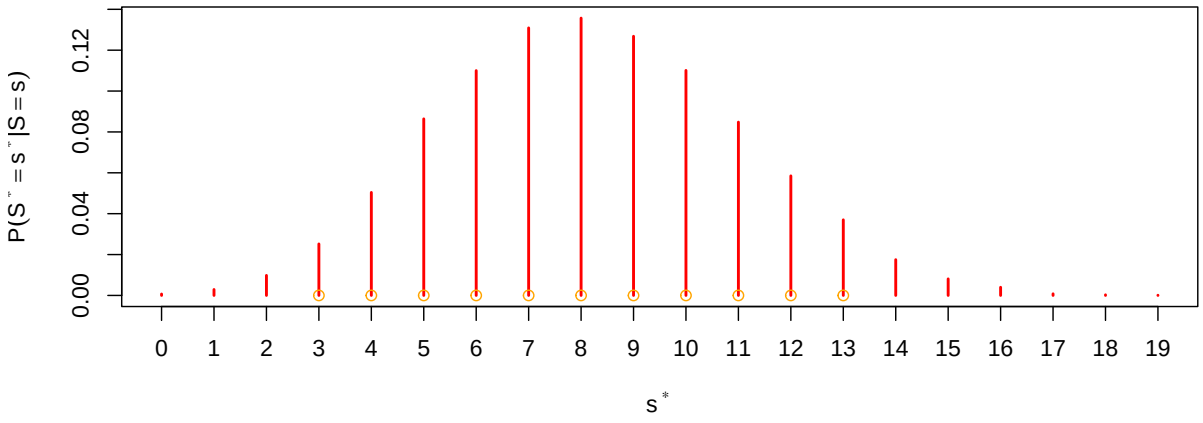


Figure 14: Représentation de la loi prédictive avec sa région HPD pour le modèle C

```
table_prev_Mod_C_DF = as.data.frame(table_prev_Mod_C, stringsAsFactors = F)
supp_prev_Mod_C = as.numeric(table_prev_Mod_C_DF$Prev_Mod_C)

Best_pred_ponct_modC = sum(supp_prev_Mod_C*table_prev_Mod_C_DF$Freq)
```

Pour le modèle C on a $\hat{S}^* = 8.203$.

Modèle a priori A

Pour le modèle a priori A la loi a posteriori déterminée n'étant pas une loi Beta il nous faut modifier l'algorithme au niveau de la génération de p.

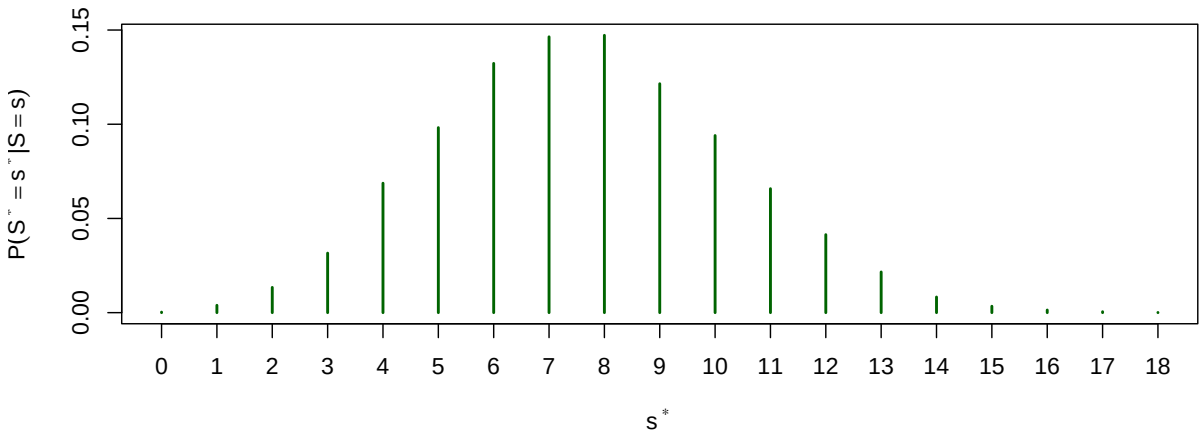


Figure 15: Représentation de la loi prédictive pour le modèle A

```
table_prev_Mod_A_ordered = table_prev_Mod_A[order(table_prev_Mod_A, decreasing = T)]
cumsum_prev_A = cumsum(table_prev_Mod_A_ordered)
cumsum_prev_A
```

```
##      8      7      6      9      5     10      4     11     12      3     13
## 0.1472 0.2936 0.4259 0.5474 0.6456 0.7396 0.8083 0.8741 0.9155 0.9471 0.9687
##      2     14      1     15     16     17      0     18
## 0.9821 0.9904 0.9943 0.9977 0.9991 0.9996 0.9999 1.0000
```

Comme pour les modèles B et C on choisit une probabilité d'appartenance du paramètre au moins égale à 0.95. Donc ici le niveau de notre région HPD est de niveau exact 96.87%

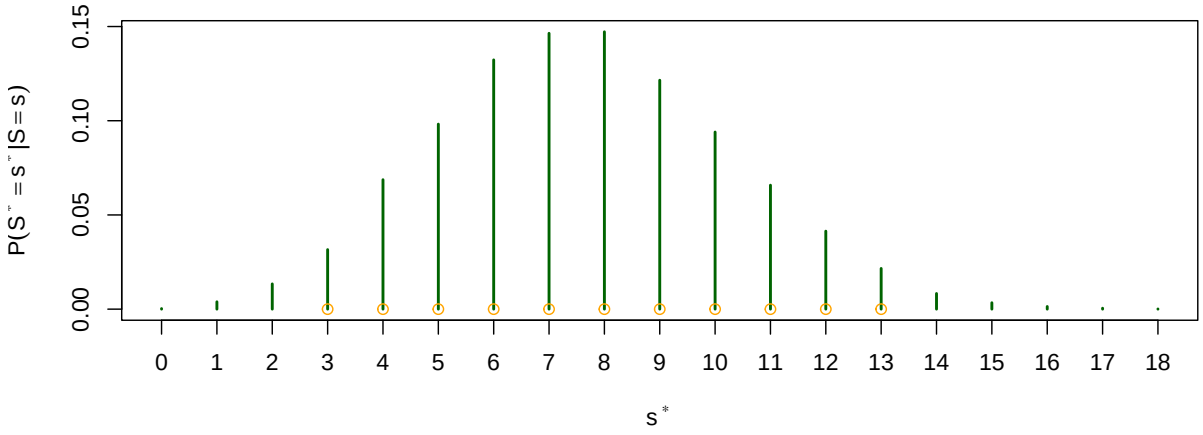


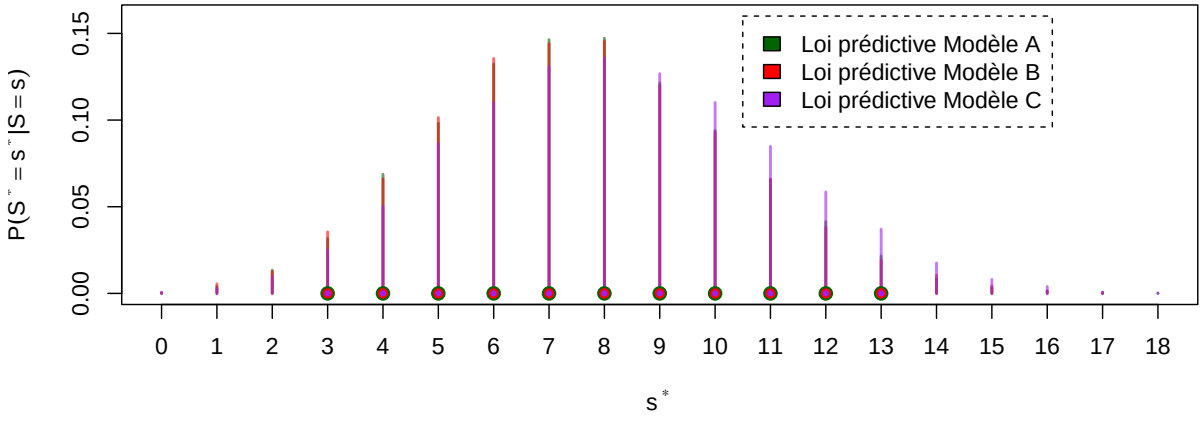
Figure 16: Représentation de la loi prédictive avec sa région HPD pour le modèle A

```
table_prev_Mod_A_DF = as.data.frame(table_prev_Mod_A, stringsAsFactors = F)
supp_prev_Mod_A = as.numeric(table_prev_Mod_A_DF$Prev_Mod_A)

Best_pred_ponct_modA = sum(supp_prev_Mod_A*table_prev_Mod_A_DF$Freq)
```

Pour le modèle A on a $\hat{S}^* = 7.6223$.

Comparaison



Les régions HPD de la loi prédictive, pour ces différents modèles, sont représentées par des cercles, de tailles différentes pour la lisibilité, au niveau de l'axe des abscisses. On remarque donc que la région HPD de niveau au moins 95% est identique pour les 3 modèles.

On rappelle les valeurs des différents prédicteurs ponctuels :
Modèle A : $\hat{S}^* = 7.6223$
Modèle B : $\hat{S}^* = 7.5815$
Modèle C : $\hat{S}^* = 8.203$

Ces prédicteurs sont quasiment égaux pour les modèles A et B. Ces deux modèles nous prédisent donc environ 7.6 élèves qui dorment plus de 8 heures parmi un groupe de 20. Le modèle C quant à lui donne une prévision d'environ 8.2 élèves.

Commentaires des résultats

On a des prédictions identiques pour les 2 premiers modèles, tandis que le modèle C donne une prévision légèrement plus élevée. Dans le cadre des lois prédictives il est difficile de déterminer le modèle le plus intéressant tant les 3 nous donnent des résultats similaires.

Conclusion

Du fait des différents résultats sur les facteurs de Bayes ainsi que les régions HPD pour les lois a posteriori, et une absence de démarcation de l'un des modèles pour les lois prédictives, je privilégierais le modèle B pour le problème étudié.