Statistique bayésienne
Feuille 2 : Tests bayésiens. Monte-Carlo.

Robin Ryder

Automne 2019

Nous reprenons des données que nous avons déjà explorées, sur l'activité des députés français. Téléchargez les données et lisez-les dans R :

```
> deputes=read.csv2('/deputes.csv')
> attach(deputes)
```

Nous nous concentrons sur la colonne `questions_orales`, qui indique le nombre de questions orales Y_i posées par chaque député. Nous aimerions savoir si cette variable dépend d'une variable binaire Z_i. Pour Z_i, vous pouvez prendre par exemple le `sexe` du député.

Nous avons donc deux modèles parmi lesquels choisir :

$$\begin{array}{ccc|ccc} & \mathcal{M}_1 & & & \mathcal{M}_2 & \\ & & & Y_i|Z_i = j & \sim_{i.i.d} & \mathcal{P}(\lambda_j) \\ Y_i & \sim_{i.i.d} & \mathcal{P}(\lambda) & \lambda_1 & \sim & \Gamma(a,b) \\ \lambda & \sim & \Gamma(a,b) & \lambda_2 & \sim & \Gamma(a,b) \end{array}$$

1. Explorez rapidement les données. Le modèle de Poisson semble-t-il pertinent ?
2. Tracez les lois a priori et a posteriori des paramètres de chaque modèle ; testez plusieurs lois a priori différentes.
3. Donnez un intervalle de crédibilité à 95% pour les paramètres de chaque modèle.
4. Dans le modèle 2, on pose $r_\lambda = \frac{\lambda_1}{\lambda_2}$ et on veut estimer l'espérance a posteriori de r_λ. Proposer un estimateur de Monte-Carlo de cette quantité.

On aimerait maintenant calculer le facteur de Bayes

$$B_{21} = \frac{m_2(\boldsymbol{y})}{m_1(\boldsymbol{y})} \quad \text{où} \quad m_k(\boldsymbol{y}) = \int_{\Theta_k} L_k(\theta_k|\boldsymbol{y})\pi_k(\theta_k)d\theta$$

Nous proposons plusieurs méthodes pour calculer ce facteur de Bayes ; nous aimerions les comparer. Pour chaque méthode, écrire un code R pour visualiser la convergence de l'estimateur.

5. **Méthode 1 : Monte-Carlo standard**

   Donner une estimation de B_{21} à partir d'un M-échantillon simulé depuis la loi a priori.

6. **Méthode 2 : échantillonnage préférentiel**

   Donner l'espérance et la variance a posteriori pour chaque modèle. Nous allons procéder par échantillonnage préférentiel, avec comme distribution instrumentale g la loi gaussienne de paramètres l'espérance et la variance a posteriori. Donner une approximation de B_{21} dans ce cas.

   Recommencer avec une autre distribution instrumentale. Qu'observez-vous ?

7. **Méthode 3 : valeur analytique**

   Calculer l'expression analytique de B_{21}, et écrire un code R pour calculer sa valeur numérique.

8. Comparez ces méthodes et choisissez la meilleure.

9. Maintenant que vous avez choisi une méthode, interprétez le facteur de Bayes pour décider quel modèle est le meilleur.

10. Répétez ce choix de modèle pour d'autres colonnes : pour Y_i, vous pouvez choisir d'autres variables quantitatives ; pour Z_i, vous pouvez choisir `groupe_sigle`, qui donne l'étiquette politique, ou `nb_mandats`, qui donne le nombre d'autres mandats électifs. Dans ces cas, Z_i peut prendre plus que 2 valeurs.

11. On suppose que notre probabilité a priori pour les modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ vaut 0.5. Quelle est la probabilité a posteriori de chaque modèle ?

12. On étudie l'espérance du nombre de questions qui seront posées par une nouvelle députée. Tirer un échantillon du paramètre λ correspondant dans chacun des trois cas suivants :

    a) On a choisi le modèle $\mathcal{M}_1$.

    b) On a choisi le modèle $\mathcal{M}_2$.

    c) On accorde à chaque modèle sa probabilité a posteriori.

    Calculer l'espérance et la variance empiriques dans chaque cas. Qu'observez-vous ?