UNIVERSITÉ DE NANTES

CENTRALE NANTES

ECN option mathématiques Parcours S2D

Statistique Bayésienne.

Anne Philippe
Université de Nantes
Laboratoire de Mathématiques Jean Leray

Exercice 1. Estimation d'un proportion

On veut estimer p la proportion des étudiants qui dorment plus de 8 heures par nuit. Les observations sur un échantillon de 27 étudiants sont :

s= 11 étudiants dorment plus de 8 heures
f=16 étudiants dorment moins de 8 heures.

On note S le nombre d'étudiants qui dorment plus de 8 heures dans un échantillon de taille $n = 27$.

On envisage trois lois a priori sur le paramètre $p \in]0, 1[$:

A- La loi discrète définie par

i	b_i	$P(p = b_i)$
1	0.05	0.03
2	0.15	0.18
3	0.25	0.28
4	0.35	0.25
5	0.45	0.16
6	0.55	0.07
7	0.65	0.03

B- Un mélange de loi uniforme (loi a priori de type histogramme) qui admet pour densité

$$h(p) \propto \sum_{i=0}^{9} q_i \, \mathbb{I}_{[a_i, a_{i+1}]}(p)$$

avec $a_i = i/10$ et

i	0	1	2	3	4	5	6	7	8	9
q_i	2	4	8	8	4	2	1	1	1	1

C- La loi Beta de paramètres $a = 3.4$ and $b = 7.4$.

Comparaison des modèles a priori

I-1) Représenter graphiquement les trois lois a priori, calculer la moyenne et la variance de ces lois a priori

I-2) Pour les trois lois a priori proposées, construire une fonction qui retourne la densité de la loi a posteriori. Calculer la moyenne et la variance des 3 lois a posteriori.

I-3) Représenter graphiquement les trois lois a posteriori (superposer avec les lois a priori).

I-4) Commenter les résultats obtenus

Générateur de nombres aléatoires suivant la loi a posteriori

II-1) Pour les modèles A et C, écrire une fonction qui retourne un échantillon de nombres aléatoires simulés suivant la loi a posteriori

Indications Utiliser les fonctions `sample` et `rgamma`.

Résultat du cours .

Mélange de lois Soit $\{f_i, i \in I\}$ une famille de densités et Z une variable discrète à valeurs dans I. Pour tout $i \in I$, on note $P(Z = i) = p_i$. On considère X une variable aléatoire dont la loi conditionnellement à Z admet pour densité f_Z

$$P(X \in A|Z = i) = \int_A f_i(x)\,\mathrm{d}x \qquad \forall\ i \in I.$$

La loi de la variable aléatoire X admet pour densité

$$\sum_{i \in I} p_i f_i(x).$$

On dit que la loi de X est un mélange de lois $\{f_i, i \in I\}$.

II-2) En utilisant le résultat précédent sur les mélanges, construire un générateur de nombres aléatoires suivant la loi a priori du modèle B.

Résultat du cours .

Algorithme d'acceptation rejet [AR] : Soit f et g deux densités de même support. On suppose qu'il existe un réel M tel que $f(x) \leq Mg(x)$.
Soit (X_n) une suite de variables aléatoires réelles iid suivant la loi de densité g et (U_n) une suite de variables aléatoires iid suivant la loi uniforme sur $[0, 1]$. On suppose que les deux suites sont indépendantes.
On définit

$$N = \inf\{j :\ U_j \leq f(X_j)/(Mg(X_j))\}$$

La loi de X_T admet pour densité f.
Ce résultat fournit une méthode pour simuler un échantillon suivant la loi de densité f à partir de nombres aléatoires simulés suivant la loi dite instrumentale de densité g et la loi uniforme.
La moyenne de la variable aléatoire N est indicateur de la performance de l'algorithme.

II-3) Justifier que l'on peut simuler la loi a posteriori du modèle B en utilisant un algorithme [AR] avec les lois instrumentales (g) suivantes
— la loi a priori du modèle
— la loi beta de paramètre (12,17)
Programmer les deux algorithmes. Comparer les performances des deux algorithmes.

II-4) En utilisant une méthode de Monte Carlo, donner une estimation de la moyenne et la variance des trois lois a posteriori. Evaluer à l'aide d'un intervalle de confiance la précision de votre estimation. Comparer avec les résultats obtenus dans la partie I.

II-5) Donner une estimation de la densité de la loi a posteriori à partir d'un échantillon de nombres aléatoires simulés suivant la loi a posteriori. Comparer graphiquement l'estimateur avec la densité calculée dans la partie I.

Régions de confiance bayésiennes pour le paramètre p

III-1) Construire une fonction qui retourne le plus court intervalle de crédibilité de niveau 95% à partir d'un échantillon de nombres aléatoires simulés suivant la loi a posteriori.

III-2) Calculer ces intervalles pour les trois modèles.

III-3) Pour chacun des modèles construire une fonction qui retourne une approximation de la région HPD de niveau 95% à partir d'un échantillon de nombres aléatoires simulés suivant la loi a posteriori.

III-4) Calculer les régions HPD de niveau 95% pour les trois modèles.

III-5) Comparer les intervalles de crédibilité et les régions HPD.

Prévision

On veut prévoir S^* le nombre d'étudiants qui dorment plus de 8 heures dans un groupe de taille 20.

Résultat du cours .

Mélange continu de lois Soit X une variable aléatoire dont la loi admet une densité f par rapport à la mesure de Lebesgue. On suppose que

$$f(x) = \int_{\mathbb{R}^p} h(x|y)g(y)\,\mathrm{d}y$$

où g et $h(\cdot|y)$, pour tout $y \in \mathbb{R}^p$, sont des densités de probabilité.
Pour simuler un nombre aléatoire x suivant la loi de densité f :

1. `on simule` y `suivant la loi de densité` g
2. `on simule` x `suivant la loi de densité` $h(\cdot|y)$
3. `on retourne` x

IV-1) Pour les trois modèles, simuler un échantillon suivant la loi predictive de S^* (loi conditionnelle de S^* sachant S).

IV-2) A partir des échantillons simulés, donner

(a) une approximation de la densité de la loi prédictive,

(b) un intervalle de prévision de niveau 95 %,

(c) un prédicteur ponctuel.