UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne PHILIPPE
Université de Nantes, LMJL

Calcul des lois a posteriori

EXERCICE 1.

On observe $(X_1, ...X_n)$ des temps d'attente entre deux bus que l'on modélise par des variables aléatoire iid suivant une loi exponentielle de paramètre $\theta > 0$. La densité de cette loi exponentielle s'écrit $x \mapsto \theta e^{-x\theta}$

On suppose que la loi a priori du paramètre θ est la loi exponentielle de paramètre 1

1) Ecrire la densité de la loi conditionnelle de $(X_1, ...X_n)$ sachant θ
2) Montrer que la loi a posteriori est une loi Gamma. Préciser les paramètres de la loi Gamma .
3) Quelle est l'expression de $E(\theta|X_1, ..., X_n)$.
4) Donner l'expression de la loi prédictive a priori. (*rappel* c'est la loi marginale de $(X_1, ..., X_n)$).

EXERCICE 2.

Soit N le nombre de personnes qui visitent un magasin un jour fixé. Ce nombre est inconnu et on souhaite l'estimer. La loi a priori choisie pour le paramètre N est la loi de Poisson de paramètre 5.

On observe Y le nombre de clients (le nombre de visiteurs qui effectue un achat). On sait que en moyenne un visiteur sur cinq effectue un achat.

1) Quelle est la loi conditionnelle de Y sachant N
2) Calculer la loi a posteriori du paramètre N.

R On observe $Y = 5$. Tracer sur un même graphique la loi a priori et la loi a posteriori

3) On simule un échantillon $(N_1, ..., N_m)$ de la façon suivante
   1. `Simuler un échantillon` $(Z_1, ..., Z_m)$ `de va iid suivant le loi de poisson de paramètre` 4. `Il est simulé de facon indépendante de l'observation` Y
   2. `Poser` $N_i = Y + Z_i$ `pour tout` $i = 1, ..., n$

Montrer que $(N_1, ..., N_m)$ est un échantillon de va iid suivant la loi a posteriori

4) Simuler un échantillon de taille m= 10000 suivant la loi a posteriori. A partir de cet échantillon donner une approximation de l'espérance, la médiane, les quantiles d'ordre 2,5% et 97,5%.

5) Est ce que le choix de m vous semble pertinent ?

UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne Philippe
Université de Nantes, LMJL

Exercice 3. Modèle Gaussien

On dispose de n observations $X_1, ..., X_n$ que l'on modélise conditionnellement à θ par des variables aléatoires iid suivant la loi gaussienne $\mathcal{N}(\theta, 1)$. On choisit comme loi a priori sur θ la loi gaussienne $\mathcal{N}(0, \tau^{-2})$, $\tau > 0$

1) Ecrire la densité de la loi conditionnelle des observations sachant θ
2) Montrer que la loi a posteriori est la loi Gaussienne

$$\mathcal{N}(\frac{\bar{X}_n}{1+\tau^2/n}, \frac{1}{n+\tau^2})$$

Calcul théorique des régions HPD

3) Pourquoi la région HPD de niveau $1-\alpha$ est un intervalle ?
4) Pourquoi cet intervalle est symétrique autour de la moyenne de la loi a posteriori $\frac{\bar{X}_n}{1+\tau^2/n}$?
5) Montrer que les régions HPD de niveau $1-\alpha$ est égale à

$$\theta \in \left[\frac{\bar{X}_n}{1+\tau^2/n} - \frac{q_{1-\alpha/2}}{\sqrt{n+\tau^2}}; \frac{\bar{X}_n}{1+\tau^2/n} + \frac{q_{1-\alpha/2}}{\sqrt{n+\tau^2}}\right] = I^{HPD}(\tau, \bar{X}_n)$$

où q_α est le quartile d'ordre α de la loi gaussienne standard.

Approximation numérique des régions HPD

6) Simuler un échantillon $X_1, .., X_n$ de taille $n = 10$ suivant la loi normale standard. On conservera cet échantillon dans toute la suite de l'exercice.
7) Simuler un échantillon $\theta_1, ..., \theta_N$ suivant la loi a posteriori.
(Prendre N assez grand par exemple N = 10 000)

- Calculer pour tout i=1,...,N les valeurs de la densité a posteriori

$$\lambda_i = \pi(\theta_i | X_1, ..., X_n)$$

- Trier par ordre croissant les valeurs de λ : on obtient $\lambda_1^* \leq, ... \leq \lambda_N^*$
- On pose $K_{N,\alpha} = \lambda^*_{[N\alpha]}$ où $[\cdot]$ est la partie entière
- Trouver $\mathcal{H} = \{\theta_i : \pi(\theta_i | X_1, ..., X_n) > K_{N,\alpha}\}$
- On pose $l_N = \min(\mathcal{H})$ et $u_N = \max(\mathcal{H})$

8) Justifier que $[l_N; u_N]$ est une approximation de la région HPD.

9) Comparer avec l'intervalle obtenu à la question 5).

Approximation numérique du plus court intervalle de crédibilité

10) Tracer en fonction de β (notation du cours) les bornes de tous les intervalles de crédibilité de niveau 95%

11) Représenter la longueur en fonction de β. et déterminer la valeur de β optimale

12) En déduire une approximation du plus court intervalle de crédibilité

13) Rappeler pourquoi cet intervalle est aussi une approximation de la région HPD.

Comparaison

14) Comparer numériquement les deux approximations de la région HPD avec les bornes théoriques.

15) Faire varier N pour illuster la qualité des approximations

Calcul théorique du niveau fréquentiste

16) Montrer que

$$P_\theta(\theta \in I^{HPD}(\tau, \bar{X}_n)) = \Phi\left(\frac{\theta\tau^2}{\sqrt{n}} + q_{1-\alpha/2}\sqrt{\frac{n+\tau^2}{n}}\right) - \Phi\left(\frac{\theta\tau^2}{\sqrt{n}} - q_{1-\alpha/2}\sqrt{\frac{n+\tau^2}{n}}\right)$$

où Φ est la fonction de répartition de la loi gaussienne standard.

17) Quelle est la limite de cette probabilité quand $n \to \infty$. Commenter.

18) Quelle est la limite de cette probabilité quand $\tau \to 0$. Commenter.

UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne Philippe
Université de Nantes, LMJL

Exercice 4. Modèle non régulier

Soit $\mathbf{X} = (X_1, \ldots, X_n)$ des variables aléatoires indépendantes et identiquement distribuées suivant la loi uniforme sur $[0\ ;\ \theta]$ avec $\theta > 0$ inconnu.

On pose

$$M_n = \max(X_1, \ldots, X_n).$$

1) Ecrire la densité de $(X_1, \ldots, X_n)$ conditionnellement à θ

   Soit (a, b) deux réels tels que $a > 1$ et $b > 0$. On choisit comme loi a priori $\pi_{a,b}$ définie par

$$\pi_{a,b}(\theta) = ab^a \frac{1}{\theta^{a+1}} \mathbb{I}_{[b\ ;\ +\infty[}(\theta).$$

2) Calculer la loi a posteriori de θ.

3) Montrer que l'estimateur de Bayes sous coût quadratique associé à la loi *a priori* $\pi_{a,b}$ vaut

$$\delta_n^{a,b}(\mathbf{X}) = \frac{a+n}{a+n-1} \max(b, M_n).$$

4) Montrer que la région HPD de niveau $1-\gamma$ est égale à

$$\left[\max(b, M_n)\ ;\ \max(b, M_n)\gamma^{-1/(a+n)}\right].$$

5) On suppose que $P_\theta(X_1 > b) > 0$.

   a) Montrer que les variables aléatoires M_n et $\max(b, M_n)$ sont presque sûrement égales à partir d'un certain rang.

   b) En déduire que l'estimateur de Bayes converge presque surement vers la vraie valeur du paramètre.

6) Que se passe-t-il lorsque $P_\theta(X_1 > b) = 0$?

Exercice 5.

Soient $X_1, \ldots, X_n$ des variables aléatoires iid suivant la loi de densité f_θ, où $\theta \in \Theta$. Le paramètre θ est inconnu, on l'estime par une approche bayésienne.

On considère la fonction de coût

$$L(\theta, \delta) = e^{\delta - \theta} - (\delta - \theta) - 1.$$

1) Montrer que la fonction L définie bien une fonction de coût.

2) Montrer que

$$\delta^\pi(X_1, \ldots, X_n) = -\log\left(\mathbb{E}(e^{-\theta}|X_1, \ldots, X_n)\right)$$

est un estimateur de Bayes pour la fonction de coût L.

UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne Philippe
Université de Nantes, LMJL

Exercice 6. Calcul des lois a priori conjuguées

Construire une famille de lois conjuguées pour les modèles suivant :

1) $X_1, ..., X_n$ iid suivant la loi de Poisson $\mathcal{P}(\theta), \theta \in \mathbb{R}_+^*$
2) $X_1, ..., X_n$ iid suivant la loi binomiale $\mathcal{B}(M, \theta), \theta \in]0, 1[$ et M est connu.
3) $X_1, ..., X_n$ iid suivant la loi Gaussienne $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$

Pour chacun des modèles, vous préciserez les paramètres des lois a priori et a posteriori.

Exercice 7. Calcul des lois de Jeffrey

1) Calculer la loi non informative de Jeffrey (si elle exite) dans les situations suivantes :
   a. $X_1, ..., X_n$ iid suivant la loi de Poisson $\mathcal{P}(\theta), \theta \in \mathbb{R}_+^*$
   b. $X_1, ..., X_n$ iid suivant la loi binomiale $\mathcal{B}(M, \theta), \theta \in]0, 1[$ et M est connu.
2) Montrer que l'information de Fisher apportée par un variable gaussienne sur les paramètres (μ, σ^2) est égale à
$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{pmatrix}$$
3) Calculer la loi de Jeffreys (si ellle existe) pour les paramètres θ suivants
   a. $\theta = \mu$ et σ^2 est connu.
   b. $\theta = \sigma$ et μ est connu.
   c. les deux paramètres sont inconnus : $\theta = (\mu, \sigma^2)$.

UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne Philippe
Université de Nantes, LMJL

Exercice 8. 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 la variable aléatoire qui représente le nombre d'étudiants qui dorment plus de 8 heures dans un échantillon de taille $n = 27$.

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

Modèle 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

Modèle B- la loi Beta de paramètres $a = 3.4$ et $b = 7.4$.

Comparaison des deux modèles bayésiens

I-1) Représenter graphiquement les deux lois a priori.

I-2) Calculer la moyenne et la variance des deux lois a priori

I-3) Commenter les résultats obtenus. Les lois apportent-elles la même information a priori ?

I-4) Calculer les deux lois a posteriori.

I-5) Représenter sur un même graphique les lois a priori et les lois a posteriori

I-6) Calculer la moyenne et la variance des loi a posteriori.

I-7) Commenter les résultats obtenus.

Régions de confiance bayésiennes pour le paramètre p

Modèle a priori A (loi discrète)

A-III-1) Construire une région HPD de niveau 95% (ou au moins 95%) ? S'il est différent de 95% quel est le niveau exact de la région HPD.

Modèle a priori B (loi a priori beta))

B-III-1) Calculer le plus court intervalle de crédibilité au niveau 95% à partir d'un échantillon de nombres aléatoires simulés suivant la loi a posteriori. Justifier le choix effectué pour la taille de l'échantillon

B-III-2) Calculer le plus court intervalle de crédibilité au niveau 95% en utilisant la fonction `qbeta`.

B-III-3) Expliquer et justifier les méthodes utilisées

B-III-4) Représenter sur un même graphique la loi a posteriori et les deux intervalles de crédibilité obtenus.

Conclusion

Comparer et commenter les résultats.

Prévision

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

Modèle a priori B (loi a priori beta)

B-IV-1) Trouver les fonctions $a(S), b(S)$ pour que la valeurs s^* simulée à partir de l'algorithme ci-dessous soit une réalisation de la loi prédictive de S^*.
Quelle expression théorique de la loi prédictive permet de justifier cet algorithme ?

1. `Simuler` p `suivant la loi beta de paramètre` $(a(S), b(S))$
2. `Simuler` s^* `suivant la loi binomiale de paramètre` $(20, p)$

B-IV-2) Simuler un échantillon de longueur M suivant la loi predictive de S^*

B-IV-3) A partir de l'échantillons simulé, calculer et représenter sur un même graphique :

(a) une approximation de la loi prédictive,

(b) une approximation du plus court intervalle de prévision de niveau 95 % (ou au moins 95%),

(c) une approximation du prédicteur ponctuel.

Justifier le choix de M.

Ccomparaison.

IV-1) Adapter l'algoritme précedent pour simuler un échantillon suivant la loi prédictive associée au modèle a priori A

IV-2) Représenter sur un même graphique les deux lois prédictives (c'est à dire les lois prédictives pour les modèles A et B)

IV-3) Comparer les plus court intervalle de prévision de niveau 95 % (ou au moins 95%) et les prédicteurs ponctuels.

IV-4) Commenter les résultats

UNIVERSITÉ DE NANTES

Master IS deuxième année

Statistique Bayésienne.

Anne Philippe
Université de Nantes, LMJL

Exercice 9. Fiabilité : modèle exponentiel

Le fichier de données

`http://www.math.sciences.univ-nantes.fr/~philippe/data/duree-de-vie.txt`

contient des durées de fonctionnement de 1000 ampoules.

Modèle : On modélise ces données par des variables aléatoires $X_1, ..., X_n$ iid suivant la loi exponentielle de paramètre $\theta \in \mathbb{R}_*^+$. On suppose que θ suit une loi Gamma.

1) L'information fournit a priori est "θ devrait être proche de $1/2$". On note τ la variance de la loi a priori. Proposer des paramètres pour la loi a priori.
2) On choisit $\tau = 1/2$. Superposer la densité de la loi a priori et les densités a posteriori pour différentes tailles d'échantillon n (par exemple `n in c(2,5,10,100,500,1000)`)
3) Reprendre la question précédente pour différentes valeurs de $\tau = 1/100, 1/10, 10, 100$.
4) Pour les différentes valeurs de τ, représenter l'évolution de la moyenne et de la variance de la loi a posteriori en fonction de n. La valeur $n = 0$ correspond à la loi a priori
5) Ecrire une fonction qui calcule une approximation du plus court intervalle de crédibilité de niveau 95%. (Utiliser la fonction `qgamma`.)
6) Pour les différentes valeurs de τ, représenter les bornes des intervalles en fonction de n le nombre d'observations.
7) Une autre source d'information indique que "θ est autour de 3". Reprendre les questions précédentes et comparer les résultats.