Master 2 Biostatistiques - UE Bayes

Daniel Commenges, François Caron et Mélanie Prague

23 janvier 2013 - 3h

Les documents de cours et les calculatrices sont autorisées.
Les exercices sont indépendants les uns des autres.

Exercice 1 (Durée conseillée : 45 minutes)

Les observations $Y_i, i = 1, \ldots, n$ sont indépendantes et identiquement distribuées (iid) suivant une loi Normale de paramètre μ et σ^2. La densité de la loi de Normale est : $f_Y^{\theta,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$. On considérera σ^2 connu.

(1) Écrire la vraisemblance et la logvraisemblance de l'échantillon $Y_i, i = 1, \ldots, n$, en faisant apparaître $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

(2) Écrire la dérivée première et seconde de la log-vraisembance par rapport à θ et l'information de Fisher pour θ.

(3) Quel est le prior de Jeffrey pour θ; est-ce une densité propre ou impropre ?

(4) En prenant ce prior, écrire le numérateur de la loi a posteriori de θ. En déduire la distribution a posteriori de θ

(5) On observe un deuxième échantillon $Y_i, i = n+1, \ldots, 2n$ iid de même loi que le premier échantillon. Quelle est la distribution a posteriori de θ en prenant un a priori uniforme ? Faire le calcul de deux façons:

- **(a)** en considérant que l'on a un échantillon iid de taille $2n$
- **(b)** en utilisant la distribution a posteriori obtenue pour le premier échantillon comme distribution a priori pour le second échantillon. (On utilisera le résultat de l'exemple 1 du document de cours appliqué au cas $n > 1$).

Exercice 2 (Durée conseillée : 1 heure 30)

Propriétés utiles:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta = \begin{cases} \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} & \text{si } a > 0, b > 0 \\ \infty & \text{sinon} \end{cases}$$

Soit $y_i \in \mathbb{N}$, $i = 1, \ldots, n$ des données indépendemment et identiquement distribuées selon une loi binomiale négative de paramètres $\theta \in [0, 1]$ et $r \in \mathbb{N}_*$. La fonction de masse de la loi binomiale négative est donnée par

$$\text{NB}(y_i; \theta, r) = \frac{(y_i + r - 1)!}{y_i!(r-1)!}(1-\theta)^r \theta^{y_i},\ y_i \in \{0, 1, 2, \ldots\}$$

On suppose dans la suite que r est connu.

(1) Ecrire la vraisemblance des données.

(2) Déterminer l'a priori de Jeffreys pour le paramètre θ. Cet a priori est-il propre ou impropre ?

(3) Calculer la distribution a posteriori correspondante de θ. Cette distribution est-elle propre ou impropre ?

On considère maintenant une distribution a priori beta de paramètres $a > 0$ et $b > 0$ pour θ

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \tag{1}$$

(4) Montrer que la loi beta est un a priori conjugué pour la loi binomiale négative. Déterminer la loi a posteriori du paramètre θ sachant $y_1, \ldots, y_n$.

(5) Calculer la vraisemblance marginale $p(y_1, \ldots, y_n)$.

On souhaite tester l'hypothèse

$$H_0 : \theta = \frac{1}{2} \text{ vs } H_1 : \theta \sim U(0,1)$$

où $U(0,1)$ désigne la distribution uniforme sur $[0,1]$.

(6) Déterminer le facteur de Bayes B_{01}.

(7) On obtient une valeur du facteur de Bayes de $B_{01} = 0.5$. Que pouvez-vous en conclure ?

On s'intéresse en particulier à la variable indicatrice c définie par

$$c = \begin{cases} 1 & \text{si } \theta > \frac{1}{2} \\ 0 & \text{sinon} \end{cases}$$

(8) Indiquer comment obtenir $p(c = 1|y_1, \ldots, y_n)$ et $p(c = 0|y_1, \ldots, y_n)$ à partir de la distribution a posteriori de θ.

On obtient $p(c = 1|y_1, \ldots, y_n) = \frac{1}{3}$. On souhaite obtenir une estimée ponctuelle de c. On considère la fonction de coût $L(c, \delta)$ asymétrique définie par

$L(c,\delta)$	$\delta = 0$	$\delta = 1$
$c = 0$	0	1
$c = 1$	2	0

(9) Calculer l'estimateur bayésien $\widehat{c}$ de c correspondant à cette fonction de masse a posteriori et cette fonction de coût.

Exercice 3 (Durée conseillée : 45 minutes)

Dans cet example, nous considérons les taux de mortalité dans 12 hôpitaux pratiquant des chirurgies cardiaques sur les enfants. Les données sont présentées Table 1 page 4. Le nombre de morts r_i de l'hôpital i est modélisé par une loi binomiale correspondant à la réalisation successive de n_i épreuves de Bernoulli de probabilité p_i. Nous nous proposons dans un premier temps d'adopter un modèle sans effet aléatoire où les probabilités de mortalité sont indépendantes entre les hôpitaux. De plus, nous choisissons un a priori non informatif sur les p_i: sur ce type de paramètres, Winbugs conseille de prendre une $Beta(1.0, 1.0)$.

(1) Ecriture du code Winbugs

(a) Ecrire le code de spécification du modèle. En Winbugs, la loi binomiale s'écrit $dbin(p, n)$ et la loi Beta s'écrit $dbeta(a, b)$.

(b) Quelles variables doivent être fournies dans "Load Data"? Ecrire le code de spécification des données.

(c) Quelles variables doivent être initialisées ? Ecrire le code d'initialisation des variables.

Un modèle plus réaliste pour la modélisation de ces données de chirurgie est de supposer que les taux de décès entre les hôpitaux sont quelque part similaires, il n'existe qu'une variabilité inter-hôpitaux. Ceci est équivalent à spécifier un modèle à effets aléatoires pour les probabilités de décès p_i. Dans ce cas là, une façon standard de prendre un a priori est de prendre une loi normale sur le logit du taux de décès. Le code Winbugs associé est donnée ici :

```
model
    {
   for( i in 1 : N ) {
   b[i] ~ dnorm(mu,tau)
   r[i] ~ dbin(p[i],n[i])
   logit(p[i]) <- b[i]
   }
    mu ~ dnorm(0.0,1.0E-4)
    tau ~ dgamma(1.0E-4, 1.0E-4);
}
```

(2) Commentaires sur les a prioris.

(a) Comment appelle-t-on le type d'a priori choisis pour mu et tau?

(b) Dans quel cas sont-ils conseillés par Winbugs?

(c) Pourquoi sont-il conseillés par Winbugs? Quelles propriétés de leurs distributions nous intéressent ?

(3) Nous donnons certaines sorties de Winbugs (page 5 - ne pas considerer la Table 2 page 4 ici). L'épaisseur d'échantillonnage est de 1 ($n_{epaisseur}$).

(a) Quel est l'intérêt d'échantillonner simultanément plusieurs chaînes ?

(b) Quelle est la longueur de la phase de burn-in utilisée ($n_{burn-in}$)? Quelle est la longueur de la phase d'échantilonnage utilisée ($n_{echantillonage}$)?

(c) Commentez chaque diagnostic de convergence pour μ et τ. Si besoin, proposez une solution pour améliorer les résultats. Justifiez quels diagnostics et quels arguments vous poussent à chaque conclusion.

(d) Dans l'état actuel, pouvez-vous conclure à la convergence ? Avec quel paramétrage proposeriez-vous de refaire tourner le modèle: $n_{burn-in}$, $n_{echantillonage}$, $n_{epaisseur}$?

(4) L'analyse est menée de nouveau avec un meilleur calibrage. Les résultats sont désormais interprétables (voir les résultats Table 2 page 4).

(a) Quelle est la probabilité en population de décès dans tous ces hopitaux ?

(b) Quel est l'écart-type des effets aléatoires sur les probabilités de décès dans ces hôpitaux?

(c) Proposer une solution pour classer ces hopitaux. L'hôpital H(numéro 8) est-il significativement moins bon que l'hôpital D (numéro 4)?

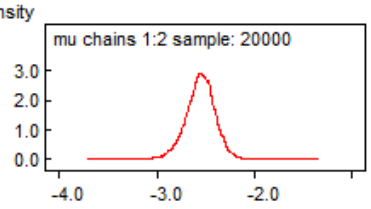
Hôpital	Nb d'opérations	Nb de morts
A	47	0
B	148	18
C	119	8
D	810	46
E	211	8
F	196	13
G	148	9
H	215	31
I	207	14
J	97	8
K	256	29
L	360	24

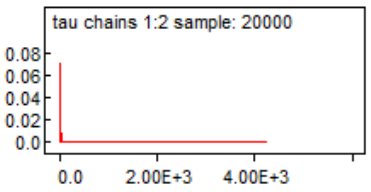
Table 1: Données de mortalité chez des enfants opérés pour des chirurgies cardiaques dans 12 hôpitaux.

Noeud	moyenne	écart-type	2.5%	médiane	97.5%
μ	-2.554	0.1524	-2.877	-2.547	-2.27
τ	10.68	0.2001	1.673	6.882	37.11
p[1]	0.05339	0.01949	0.0183	0.05288	0.09332
p[2]	0.1032	0.02207	0.06643	0.1011	0.1527
p[3]	0.07053	0.01718	0.0404	0.06958	0.1073
p[4]	0.05937	0.007905	0.04484	0.05904	0.07583
p[5]	0.05181	0.01327	0.02797	0.05114	0.08004
p[6]	0.06956	0.01475	0.04355	0.0686	0.1014
p[7]	0.06671	0.01582	0.03879	0.0655	0.1008
p[8]	0.123	0.02208	0.08326	0.1219	0.1696
p[9]	0.06993	0.01463	0.04388	0.0691	0.1017
p[10]	0.07854	0.01998	0.04476	0.07686	0.1232
p[11]	0.102	0.0175	0.07154	0.1008	0.1397
p[12]	0.06857	0.0118	0.04716	0.06799	0.09341

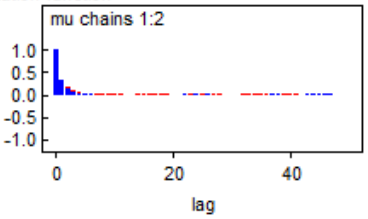
Table 2: Sorties statistiques WinBUGS

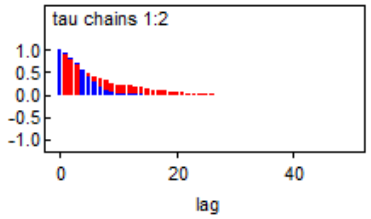
Kernel density



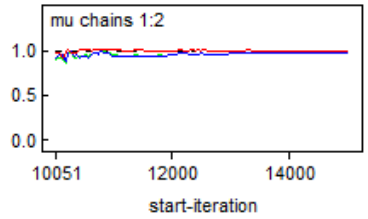


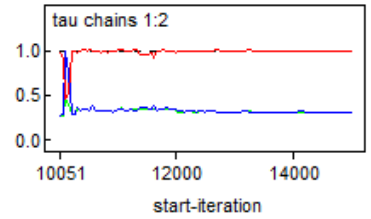
Autocorrelation function



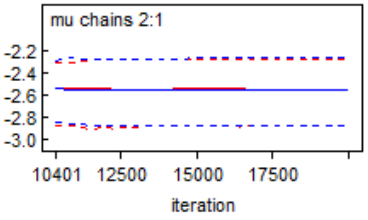


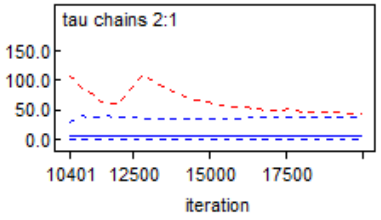
Gelman Rubin statistic



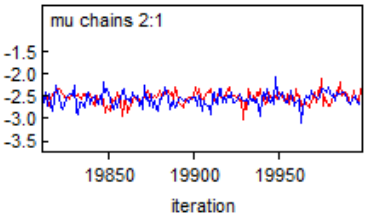


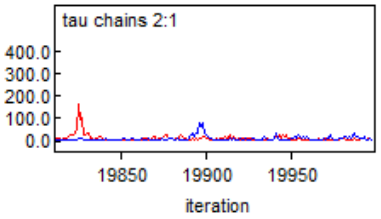
Running quantiles





Dynamic trace





Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu	-2.554	0.1538	0.001588	-2.878	-2.548	-2.267	10001	20000
tau	12.65	68.23	1.515	1.637	6.993	40.75	10001	20000