Université de Montpellier — Année 2021-2022
M2 Biostatistique - HAX918X

Examen terminal - vendredi 29 octobre 2021
Durée 2h00 - Documents non autorisés

Exercice 1 (6 pts)

On souhaite estimer

$$\theta = \int_0^1 \int_0^1 \left(\frac{1 + x^2 + y^2}{1 + \sin(xy)} \right) \mathrm{d}x \mathrm{d}y \,.$$

1 (2 pts) Proposer un estimateur $\hat{\theta}$ de θ par une méthode de Monte Carlo.

2 (2 pts) Sans approximation asymptotique, déterminer le nombre de simulation nécessaire pour que l'erreur absolue relative de $\hat{\theta}$ soit inférieure à 1% avec une probabilité supérieure ou égale à 99%.

3 (2 pts) Répondre à la même question que précédemment en utilisant une approximation asymptotique.

Exercice 2 (6 pts)

On considére la variable aléatoire X admettant pour densité

$$f_X(x) \propto \exp(-|x|)\mathbb{I}_{[-2,2]}(x)\,.$$

1 (2 pts) Proposer une méthode de simulation exacte de X, basée sur l'utilisation d'un algorithme acceptation-rejet.

2 (2 pts) Proposer une méthode de simulation exacte et directe de X (directe dans le sens où elle ne nécessite pas l'utilisation d'un algorithme itératif).

3 (2 pts) Proposer une méthode de simulation approchée de X, basée sur l'utilisation de chaînes de Markov.

Exercice 3 (8 pts)

On considère le problème consistant à estimer une taille $N > 0$ de population. C'est notamment le cas en écologie lorsque l'on s'intéresse à des espèces en voie d'extinction. Un processus inférentiel consiste à capturer et garder des individus sur une zone pendant un temps donné. On note $n^+ > 0$ le nombre d'individus capturés et p la probabilité de capture. Si l'on suppose que l'événement capturer l'individu i est indépendant de celui consistant à capturer l'individu j, alors, clairement, $n^+ \sim \mathcal{B}(N, p)$. N est le paramètre d'intérêt et p est un paramètre de nuisance.

On se place dans le paradigme bayésien et on suppose que

$$\pi(N,p) \propto \mathbb{I}_{\{1,\dots,S\}}(N) \times \mathbb{I}_{]0,1[}(p)\,,$$

où S est un hyper-paramètre fixé correspondant à une borne supérieure sur la taille de population à estimer.

1 (4 pts) Donner la loi a posteriori de N, $\pi(N|n^+)$, indication

$$\int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad ; \quad \Gamma(n) = (n-1)!\,.$$

Une extension logique au modèle binomial précédent est le modèle de capture-marquage-recapture qui considère deux périodes de capture plus une étape de marquage :

— n_1 individus d'une population de taille N sont capturés, c'est-à-dire échantillonnés sans remise ;

— ces individus sont marqués, c'est-à-dire identifiés par une étiquette numérotée (oiseaux, poissons), un collier (pour les mammifères), ou un autre dispositif (comme le numéro de sécurité sociale pour les sans-abri ou le numéro d'identification personnel pour les personnes âgées), et ils sont ensuite relâchés dans la population ;

— un deuxième échantillonnage similaire (encore une fois sans remise) est effectué avec n_2 individus capturés ;

— m_2 individus sur les n_2 portent la marque d'identification et sont donc caractérisés comme ayant été capturés dans les deux expériences.

Clairement,

$$n_2 - m_2 | n_1, m_2 \sim \mathcal{B}(N - n_1, p)\,,$$
$$m_2 | n_1 \sim \mathcal{B}(n_1, p)\,,$$
$$n_1 \sim \mathcal{B}(N, p)\,.$$

On note $n_c = n_1 + n_2$ le nombre total de capture sur les deux périodes incluant le nombre d'individus recapturés et $n^+ = n_1 + (n_2 - m_2)$ le nombre d'individus différents capturés.

2 (2 pts) Donner la vraisemblance et montrer qu'elle ne dépend de n_1, n_2 et m_2 qu'au travers de n_c et n^+, justifiant ainsi du fait que le couple (n_c, n^+) est une statistique exhaustive pour (N,p).

3 (2 pts) On se place dans le paradigme bayésien et on suppose que

$$\pi(N,p) \propto \left(\frac{1}{N}\right) \mathbb{I}_{\mathbb{N}^*}(N)\mathbb{I}_{]0,1[}(p)\,.$$

Montrer que l'on peut utiliser l'échantillonneur de Gibbs pour simuler des réalisations suivant

$$\pi(N, p | n_c, n^+)\,.$$

1

Correction examen Statistique bayésienne HAX918X – 29/10/2021

Exercice 1

1] $$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1+X_i^2+Y_i^2}{1+\min(X_i,Y_i)}\right]$$

où $(X_1,Y_1), \ldots, (X_N,Y_N) \overset{iid}{\sim} U_{[0,1]} \otimes U_{[0,1]}$

2] $$\mathbb{P}\left(\frac{|\hat{\theta}-\theta|}{|\theta|} \leq 0.01\right) \geq 1 - \frac{\mathbb{V}(\hat{\theta})}{\theta^2 10^{-4}}$$

$$\mathbb{V}(\hat{\theta}) = \frac{1}{N}\left[\int_0^1\int_0^1 \frac{(1+x^2+y^2)^2}{(1+\min(x,y))^2}\,dx\,dy - \theta^2\right]$$

$$\mathbb{P}\left(\frac{|\hat{\theta}-\theta|}{|\theta|}\leq 10^{-2}\right)\geq$$

$$1-\frac{10^4}{N}\left[\frac{\int_0^1\int_0^1 \frac{(1+x^2+y^2)^2}{(1+\sin(xy))^2}\,dx\,dy}{\theta^2}-1\right]$$

Et

$$1-\frac{10^4}{N}\left[\frac{\overbrace{\int_0^1\int_0^1 \frac{(1+x^2+y^2)^2}{(1+\sin(xy))^2}\,dx\,dy}^{a}}{\underbrace{\left[\int_0^1\int_0^1 \frac{(1+x^2+y^2)}{(1+\sin(xy))}\,dx\,dy\right]^2}_{b}}-1\right]\geq 0.99$$

$$\iff N\geq 10^6\left[\frac{a}{b}-1\right]$$

On cherche alors une borne supérieure à $\left(\frac{a}{b}-1\right)$!

$\forall x, y \in [0,1]$ (3)

$$(1+x^2+y^2)^2 \leq 3^2$$

$$(1+\sin(xy))^2 \geq 1^2$$

$$\Rightarrow a \leq 9$$

$\forall x, y \in [0,1]$

$$(1+x^2+y^2) \geq 1$$

$$(1+\sin(xy)) \leq 1+\sin(1)$$

$$\Rightarrow b \geq \frac{1}{(1+\sin(1))^2}$$

Ainsi $\left[\frac{a}{b} - 1\right] \leq 9\left[1+\sin(1)\right]^2 - 1$

Et donc,

$$N \geq \left\{9(1+\sin(1))^2 - 1\right\} 10^6$$

4

3] $\mathbb{P}\left(\dfrac{|\hat{\theta}-\theta|}{|\theta|} \leq 10^{-2}\right)$

$= \mathbb{P}\left(|\hat{\theta}-\theta| \leq 10^{-2}|\theta|\right)$

$= \mathbb{P}\left(\dfrac{|\hat{\theta}-\theta|}{\sqrt{\mathbb{V}(\hat{\theta})}} \leq \dfrac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right)$

$\# \mathbb{P}\left(|\mathcal{N}(0,1)| \leq \dfrac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right)$

[#] pour N suffisamment grand
approximation asymptotique
TCL

Ainsi,

$\mathbb{P}\left(\dfrac{|\hat{\theta}-\theta|}{|\theta|} \leq 10^{-2}\right) \# F_{\mathcal{N}(0,1)}\left[\dfrac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right] - F_{\mathcal{N}(0,1)}\left[\dfrac{-10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right]$

$$\Longleftrightarrow \mathbb{P}\left(\frac{|\hat{\theta}-\theta|}{|\theta|} \leq 10^{-2}\right)$$

$$\# \quad 2\, F_{N(0,1)}\left(\frac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right) - 1$$

Ainsi,

$$2\, F_{N(0,1)}\left[\frac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}}\right] - 1 \geq 0.99$$

$$\Longleftrightarrow \frac{10^{-2}\theta}{\sqrt{\mathbb{V}(\hat{\theta})}} \geq F^{-1}_{N(0,1)}(0.995)$$

$$\Longleftrightarrow N \geq 10^{4}\left[\frac{a}{b} - 1\right] F^{-1}_{N(0,1)}(0.995)$$

En utilisant le fait que

$$\left[\frac{a}{b} - 1\right] \leq 9(1+\sin(1))^2 - 1$$

On obtient

$$N \geq 10^{4}\, F^{-1}_{N(0,1)}(0.995)\left\{9(1+\sin(1))^2 - 1\right\}$$

Borne très $<$ à celle de la question précédente !

Exercice 2

1] Par exemple, on peut simuler suivant une loi de Laplace de paramètres (0,1) et accepter cette simulation dès qu'elle tombe dans l'intervalle [-2,2].
On peut aussi utiliser une loi de proposition différente : par exemple, une loi uniforme sur l'intervalle [-2, 2].

2] Calculons avant tout la constante de normalisation de $f_X(\cdot)$

$$\int_{-2}^{2} e^{-|x|}\,dx = 2\int_0^2 e^{-x}\,dx$$

$$= 2\left[-e^{-x}\right]_0^2 = 2(1 - e^{-2})$$

Nous obtenons ainsi

$$f_X(x) = \frac{e^{-|x|}\, \mathbb{1}_{[-2,2]}(x)}{2(1-e^{-2})}$$

Ainsi, si $x \in [-2,0]$,

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-2}^{x} f_X(x)\, dx$$

$$= \int_{-2}^{x} \frac{e^{x}}{2(1-e^{-2})}\, dx = \left[\frac{e^{x}-e^{-2}}{2(1-e^{-2})} \right]$$

Si $x \in [0,2]$,

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-2}^{0} \frac{e^{x}}{2(1-e^{-2})}\, dx + \int_{0}^{x} \frac{e^{-x}}{2(1-e^{-2})}\, dx$$

$$= \left[\frac{1}{2} + \frac{1-e^{-x}}{2(1-e^{-2})} \right]$$

Bien sûr, $F_X(x) = 0$ si $x \leq -2$ et $F_X(x) = 1$ si $x \geq 2$.

Nous n'avons plus qu'à inverser la fonction de répartition et l'appliquer à une simulation uniforme.

Si $p \leq 1/2$ alors $x \in [-2, 0]$

Soit $p \in [0, 1/2]$ t.q.

$$p = \frac{e^x - e^{-2}}{2(1-e^{-2})} \iff 2p(1-e^{-2}) + e^{-2} = e^x$$

$$\iff x = \log\left[2p(1-e^{-2}) + e^{-2}\right]$$

Si $p \geq 1/2$ alors $x \in [0, 2]$

Soit $p \in [\frac{1}{2}, 1]$ t.q.

$$p = \frac{1-e^{-x}}{2(1-e^{-2})} + \frac{1}{2}$$

$$\iff (2p-1)(1-e^{-2}) = 1 - e^{-x}$$

$$\iff -x = \log\left[1-(2p-1)(1-e^{-2})\right]$$

$$\iff x = -\log\left[1-(2p-1)(1-e^{-2})\right]$$

Donc, on simule $U \sim U_{[0,1]}$

si $U \leq 1/2$ alors $X = \log\left[2U(1-e^{-2}) + e^{-2}\right]$

si $U \geq 1/2$ alors $X = -\log\left[1-(2U-1)(1-e^{-2})\right]$

3] Pour simuler des réalisations suivant

$$f_X(x) \propto e^{-|x|} \mathbb{1}_{[-2,2]}(x)$$

on peut utiliser l'algorithme de Metropolis-Hastings avec par exemple comme noyau de proposition uniforme la loi $\mathcal{U}_{[-2,2]}$ (noyau indépendant)

Au cœur de l'algorithme, nous avons

$y \sim \mathcal{U}_{[-2,2]}$

$u \sim \mathcal{U}_{[-1,1]}$

si $\dfrac{e^{-|y|}}{e^{-|x^{(t-1)}|}} < u$ alors $x^{(t)} = y$

sinon $x^{(t)} = x^{(t-1)}$

10

Exercice 3

1] $m^+ \sim \mathcal{B}(N, p)$; $m^+ \geq 0$

$$\pi(N, p) \propto \mathbb{1}_{\{1,\dots,S\}}(N) \, \mathbb{1}_{]0,1[}(p)$$

$$\pi(N, p \mid m^+) \propto \mathbb{1}_{\{m^+ \leq N\}}(N) \, C_N^{m^+} p^{m^+} (1-p)^{N-m^+} \pi(N, p)$$

$$\Rightarrow \pi(N \mid m^+) \propto C_N^{m^+} \int_0^1 p^{m^+} (1-p)^{N-m^+} \, dp \; \mathbb{1}_{\{m^+,\dots,S\}}(N)$$

$$\Rightarrow \pi(N \mid m^+) \propto C_N^{m^+} \frac{\Gamma(m^+ + 1)\,\Gamma(N - m^+ + 1)}{\Gamma(N+2)} \mathbb{1}_{\{m^+,\dots,S\}}(N)$$

$$\Rightarrow \pi(N \mid m^+) \propto \frac{N! \; m^+! \; (N-m^+)!}{(N-m^+)! \cdot m^+! \; (N+1)!} \mathbb{1}_{\{m^+,\dots,S\}}(N)$$

$$\Rightarrow \pi(N \mid m^+) \propto \left(\frac{1}{N+1}\right) \mathbb{1}_{\{m^+,\dots,S\}}(N)$$

2] $M_2 - m_1 \mid M_1, m_2, N, p \sim B(N - M_1, p)$

$m_2 \mid M_1, N, p \sim B(M_1, p)$

$M_1 \mid N, p \sim B(N, p)$

$$f(M_1, M_2, m_2 \mid N, p) = f(M_2 \mid M_1, m_2, N, p)\, f(m_2 \mid M_1, N, p)\, f(M_1 \mid N, p)$$

$$\ell(N, p \mid M_1, M_2, m_2)$$

$$\propto C_{N-M_1}^{M_2 - m_2}\, p^{M_2 - m_2} (1-p)^{N - M_1 - M_2 + m_2}\, \mathbb{1}_{\{N \geq M^+\}}(N)$$

$$C_{M_1}^{m_2}\, p^{m_2} (1-p)^{M_1 - m_2}$$

$$C_N^{M_1}\, p^{M_1} (1-p)^{N - M_1}\, \mathbb{1}_{\{N \geq M_1\}}(N)$$

$$\ell(N, p \mid M_1, M_2, m_2)$$

$$\propto \frac{N!}{(N - M_1 - M_2 + m_2)!}\, p^{M_1 + M_2} (1-p)^{2N - M_1 - M_2}\, \mathbb{1}_{\{N \geq M^+\}}(N)$$

$$\propto \frac{N!}{(N - M^+)!}\, p^{M_c} (1-p)^{2N - M_c}\, \mathbb{1}_{\{N \geq M^+\}}(N)$$

Ainsi le couple (m_c, m^+) est une statistique exhaustive pour le paramètre (N, p)

3) Nous avons

$$\pi(p|N, m_c, m^+) \propto l(N,p|m_c, m^+)\, \pi(N,p)$$

$$\Rightarrow \pi(p|N, m_c, m^+) \propto p^{m_c}(1-p)^{2N-m_c}\, \mathbb{1}_{]0,1[}(p)$$

$$\Rightarrow p\,|\,N, m_c, m^+ \sim \text{Beta}(m_c+1, 2N-m_c+1)$$

Aussi,

$$\pi(N|p, m_c, m^+) \propto l(N,p|m_c, m^+)\, \pi(N,p)$$

$$\Rightarrow \pi(N|p, m_c, m^+) \propto \frac{(N-1)!}{(N-m^+)!}\, p^{m_c}(1-p)^{2N-m_c}\, \mathbb{1}_{\{N \geq m^+\}}(N)$$

$$\Rightarrow \pi(N|p, m_c, m^+) \propto C_{N-1}^{N-m^+}\left[1-(1-p)^2\right]^{m^+}\left((1-p)^2\right)^{N-m^+}\, \mathbb{1}_{\{N \geq m^+\}}(N)$$

$\Rightarrow N|p, m_c, m^+$ suit une loi Binomiale Negative de paramètres $(m^+, 1-(1-p)^2)$

(nombre de tentatives pour obtenir m^+ succès avec probabilité de succès $1-(1-p)^2$!

Disposant des 2 lois
conditionnelles complètes,
nous pouvons utiliser
l'échantillonneur de Gibbs qui
donne successivement des
simulations suivant ces 2 lois !