Projet de Simulation et Monte Carlo

Lucie Tournier, Sorelle Methafe & Yang Song

Prévalence de la Listéria

Nous étudions la prévalence de la bactérie *Listeria* dans le lait cru dans plusieurs pays.

Données disponibles : Nombre d'échantillons de lait testés (n_i) et nombre d'échantillons positifs (r_i) pour 91 pays.

Figure: histogramme des taux de prévalence de la listeria dans l'échantillon

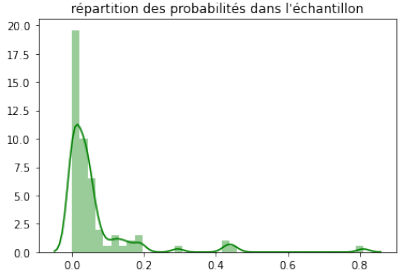


Table of contents

Ici,

- $r_i \rightsquigarrow Binom(n_i, p)$
- la loi a priori $\pi_0(p) : p \rightsquigarrow \mathcal{B}eta(\alpha, \beta)$; $\alpha = \beta = 1$

On cherche à représenter la loi a posteriori de p

Recherche de la loi a posteriori π_N de p

$$\pi_N \propto \mathcal{L}((r_i)_{i=1\dots N}, p)\pi_0(p)$$

avec:

$$\mathcal{L}((r_i)_{i=1\dots N}, p) = \prod_{i=1}^{N} C_{n_i}^{r_i} p^{r_i}(1-p)^{n_i - r_i} \propto p^{N\bar{r}}(1-p)^{N(\bar{n}-\bar{r})}$$

et

$$\pi_0(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

où $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Recherche de la loi a posteriori π_N de p

Ainsi,

$$\pi_N \propto p^{(N\bar{r}+\alpha-1)}(1-p)^{N(\bar{n}-\bar{r})+\beta-1}$$

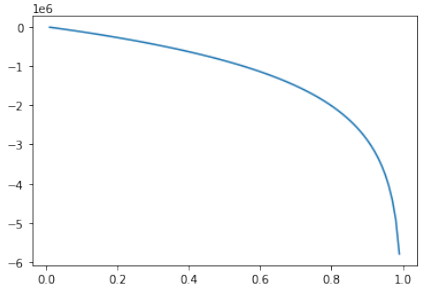
$$p \rightsquigarrow \mathcal{B}eta(N\bar{r}+\alpha, N(\bar{n}-\bar{r}))+\beta)$$

Pour $\alpha = \beta = 1$ et N=91,

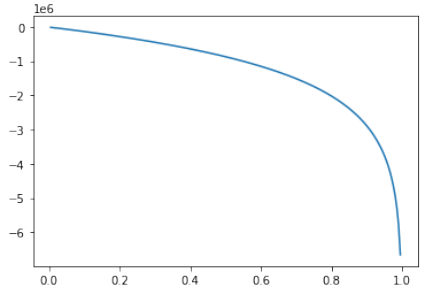
$$p \rightsquigarrow \mathcal{B}eta(91\bar{r}+1; 91(\bar{n}-\bar{r})+1)$$

Ce résultat semble logique étant donné que la loi a priori de p est une loi Bêta et la loi $\mathcal{B}$eta est conjuguée à la loi Binomiale que suivent les données.

Représentation graphique de la loi a posteriori

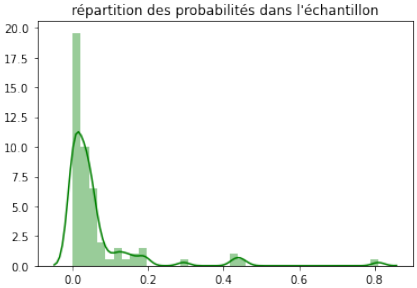


(a) logarithme de la fonction de densité de la loi à posteriori de p



(b) logarithme de la fonction de densité d'une loi Beta de paramètres (1329, 1255727)

Comparaison de la loi a posteriori et des données



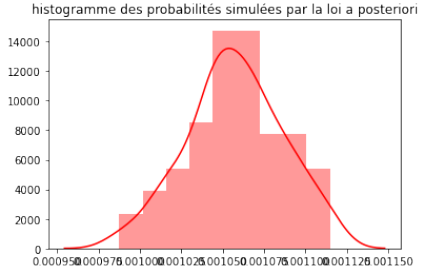


Figure: histogramme de la simulation de la loi à posteriori de p en comparaison avec l'histogramme des r_i/n_i

Présentation du deuxième modèle

ici, on suppose $r_i \rightsquigarrow Binom(n_i, p_i)$

la loi a priori est : $p_i \rightsquigarrow \mathcal{B}eta(\alpha, \beta)$

on pose : $\mu = \frac{\alpha}{\alpha+\beta} \rightsquigarrow \mathcal{U}(0,1)$ et $\kappa = \alpha + \beta \rightsquigarrow Exp(0.1)$.

On cherche à simuler la loi à posteriori du vecteur des p_i ainsi que des paramètres μ et κ avec l'algorithme de Métropolis-Within-Gibbs.

Algorithme de Metropolis-Within-Gibbs

Supposons que nous sommes à l'étape $j+1$ de notre algorithme. Nous notons par p^{j+1} le vecteur constitué des p_i, μ^{j+1} et κ^{j+1} les valeurs des paramètres μ et κ à cette étape.

1. $p^{j+1} = (p_1^{j+1}, ..., p_N^{j+1}) \rightsquigarrow \pi_N(p \backslash (r_i)_{i=1...N}, (n_i)_{i=1...N}, \mu^j, \kappa^j)$
   - pour $k = 1...N, p_k^{j+1} \rightsquigarrow \pi_N(p_k \backslash r_k, n_k, \mu^j, \kappa^j)$
2. $\mu^{j+1} \rightsquigarrow \pi(\mu \backslash (r_i)_{i=1...N}, p^{j+1}, \kappa^j)$
3. $\kappa^{j+1} \rightsquigarrow \pi(\kappa \backslash (r_i)_{i=1...N}), p^{j+1}, \alpha^{j+1})$

Les étapes 2 et 3 utilisent la marche aléatoire de Métropolis.

Lois a posteriori utilisées

- $\pi_N(p, \mu, \kappa|(r_i)_i, (n_i)_i) \propto \pi(\mu)\pi(\kappa)\big(\prod_{i=1}^N \pi(r_i|p_i, n_i)\pi(p_i|\mu, \kappa)\big)$

$$\pi_N(p, \mu, \kappa|(r_i)_{i=1\dots N}) \propto \exp(-0,1\kappa) \prod_{i=1}^N p_i^{(r_i+\mu\kappa-1)}(1-p_i)^{(n_i+\kappa(1-\mu)-r_i-1)}\mathbb{I}_{[0\leq\mu\leq1]}\mathbb{I}_{[0\leq\kappa]}$$

- $\pi_N(p_i|p_{-i}, \mu, \kappa) \propto \pi(p, \mu, \kappa|r_i) \propto p_i^{(r_i+\mu\kappa-1)}(1-p_i)^{(n_i+\kappa(1-\mu)-r_i-1)}$

$$p_i|p_{-i}, \mu, \kappa \rightsquigarrow \mathcal{B}eta(r_i + \mu\kappa, n_i + \kappa(1-\mu) - r_i)$$

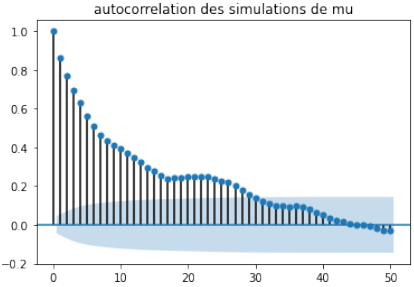
- $\pi(\mu|(r_i)_i, p, \kappa) \propto \prod_{i=1}^N \frac{p_i}{1-p_i}^{\mu\kappa}\mathbb{I}_{[0\leq\mu\leq1]}$
- $\pi(\kappa|(r_i)_i, p, \mu) \propto \exp(0, 1\kappa) \prod_{i=1}^N p_i^{\mu\kappa}(1-p_i)^{\kappa(1-\mu)}\mathbb{I}_{[0\leq\kappa]}$

Algorithme de la marche aléatoire de Métropolis

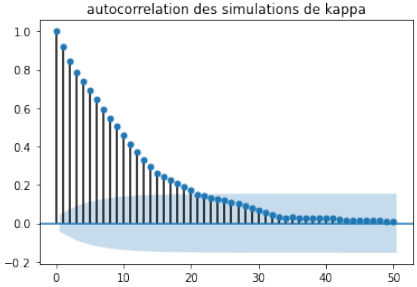
Soit f une fonction proportionnelle à la distribution-cible, c'est-à-dire à la distribution de probabilité recherchée π .

1. **initialisation**: choix de x_0 de la probabilité de transition g (distribution Gaussienne).
2. **A chaque itération t (échantillon courant x_t)**:
   - Tirer aléatoirement un candidat x' pour l'échantillon suivant selon la distribution $g(x'|x_t)$;
   - Calculer le taux d'acceptation $\alpha = \frac{f(x')}{f(x_t)}$. On a $\alpha = \frac{f(x')}{f(x_t)} = \frac{\pi(x')}{\pi(x_t)}$;
   - Accepter ou rejeter:
     - Tirer un nombre aléatoire uniforme $u \in [0, 1]$;
     - Si $u \leq \alpha$, alors accepter le candidat en effectuant $x_{t+1} = x'$
     - Si $u > \alpha$, alors rejeter le candidat en effectuant $x_{t+1} = x_t$.

Fonctions d'auto-corrélation pour les paramètres mu et kappa



(a) fonction d'autocorrélation des mu simulés



(b) fonction d'autocorrélation des kappa simulés

On choisit de ne conserver pour l'estimation qu'un point de simulation sur 30

Burn-in de mu et kappa

Le graphique ci-dessous représentent l'évolution de μ et κ au cours du temps.

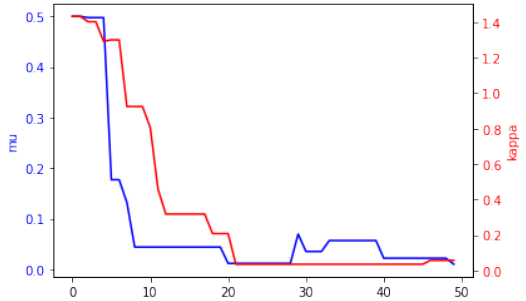


Figure: Burn-in des paramètres μ et κ

Comparaison des taux de prévalence simulés et observés

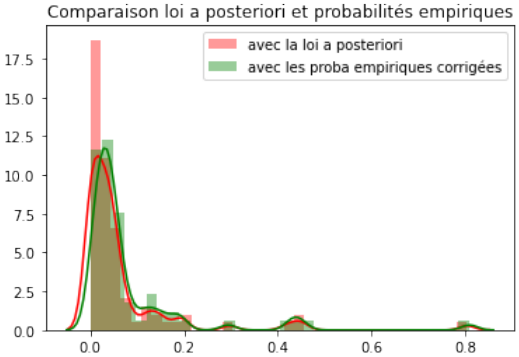


Figure: comparaison de la répartition des taux de prévalence simulés et estimé par $p_i = r_i/n_i$

Présentation du troisième modèle

ici, on suppose $r_i \rightsquigarrow Binom(n_i, p_i)$

p_i suit a priori un mélange de deux lois Beta. Une de paramètre (α_0, β_0) avec une probabilité **q** et une autre de paramètre (α_1, β_1) avec une probabilité **1-q**.

on pose : $\mu = \frac{\alpha}{\alpha+\beta} \rightsquigarrow \mathcal{U}(0,1)$ et $\kappa = \alpha + \beta \rightsquigarrow \mathcal{E}\text{xp}(0.1)$

On introduit ici une variable latente Z_i qui suit une loi de Bernouilli de paramètre q fixé. On suppose que si $z_i = x$, $p_i \rightsquigarrow \mathcal{B}eta(\mu_x\kappa_x, (1-\mu_x)\kappa_x)$

Recherche de la loi π_N des p_i

Ici, on suit l'algorithme de Metropolis-Within-Gibbs ci-dessus. Seulement, on mettra à jour z_i chaque fois et on a:

$p_i|z_i, \mu, \kappa \sim Beta(r_i + \mu_{z_i}\kappa_{z_i}, n_i + \kappa_{z_i}(1-\mu_{z_i}) - r_i)$

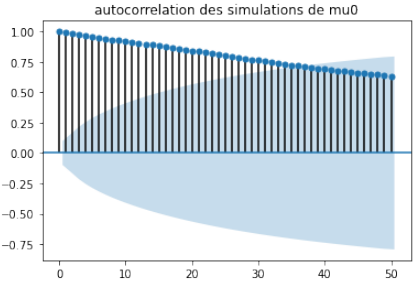
$\pi_{\mu|\kappa,p,z} \propto \prod_i \left(\frac{p_i}{1-p_i}\right)^{\kappa_0\mu_0(1-z_i)+\kappa_1\mu_1 z_i}$

$\pi_{\kappa|\mu,p,z} \propto \left(\prod_i p_i^{\kappa_0\mu_0(1-z_i)+\kappa_1\mu_1 z_i}(1-p_i)^{\kappa_0(1-\mu_0)(1-z_i)+\kappa_1(1-\mu_1)z_i} \exp^{-0.1(\kappa_0(1-z_i)+\kappa_1 z_i)}\right)$

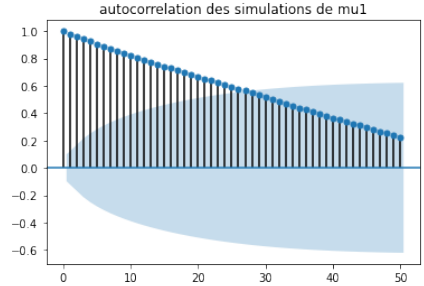
$P(z_i = x|\mu, \kappa, p) \propto p_i^{\mu_x\kappa_x}(1-p_i)^{(1-\mu_x)\kappa_x} q^x (1-q)^{1-x}$

On en déduit que $z_i|\mu, \kappa, p \rightsquigarrow B(q_i)$ avec $q_i = p_i^{\mu_1\kappa_1-\mu_0\kappa_0}(1-p_i)^{(1-\mu_1)\kappa_1-(1-\mu_0)\kappa_0}\frac{q}{1-q}$

Fonctions d'auto-corrélation pour les paramètres μ_0 et μ_1

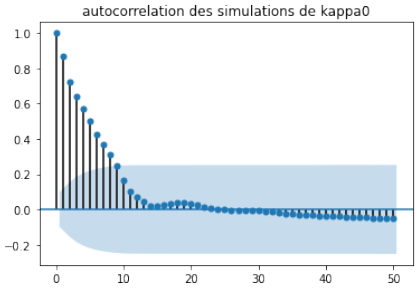


(a) fonction d'autocorrélation des μ_0 simulés

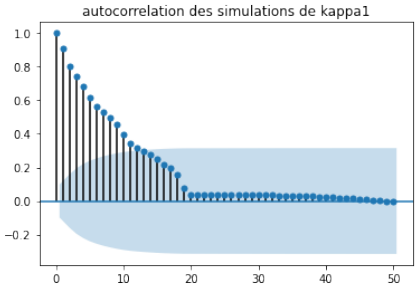


(b) fonction d'autocorrélation des μ_1 simulés

Fonctions d'auto-corrélation pour les paramètres κ_0 et κ_1



(a) fonction d'autocorrélation des κ_0

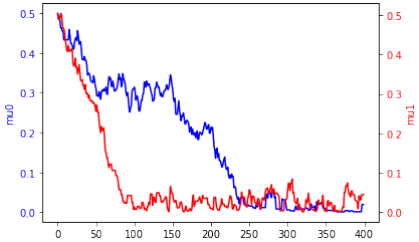


(b) fonction d'autocorrélation des κ_1

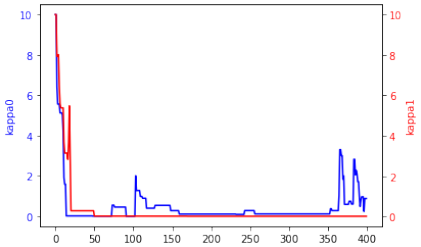
On choisit de ne conserver pour l'estimation qu'un point de simulation sur 40

Burn-in des paramètres μ et κ

Les graphiques ci-dessous représentent l'évolution des paramètres μ_0, μ_1, κ_0 et κ_1 au cours du temps.



(a) Burn-in des μ



(b) Burn-in des κ

Comparaison des taux de prévalence simulés et observés

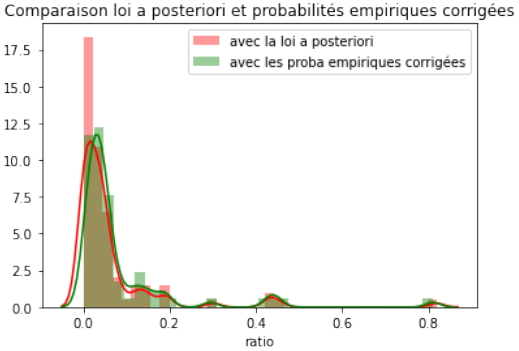
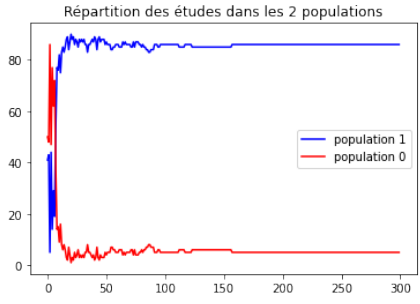
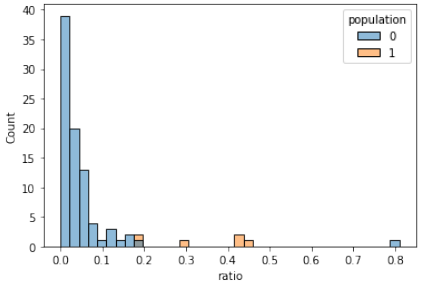


Figure: comparaison de la répartition des taux de prévalence simulés et estimé par $p_i = r_i/n_i$

Répartition des deux lois Beta



(a) proportion des p_i suivant chaque loi Beta



(b) histogramme de la loi simulée

Comparaison des modèles 2 et 3

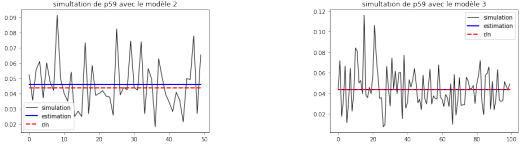


Figure: comparaison des simulations des deux modèles pour p_{59}

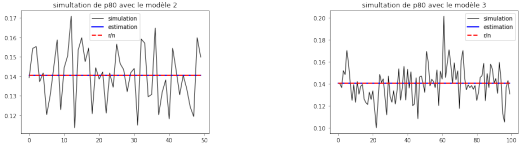


Figure: comparaison des simulations des deux modèles pour p_{80}

Comparaison des modèles 2 et 3

pts de simulation	Modèle 2		Modèle 3	
	erreur $\times 10^{-3}$	temps d'exécution	erreur $\times 10^{-3}$	temps d'exécution
10	3.66	15s	2.89	43s
20	2.39	33s	1.96	1min 27s
50	1.59	1min 19s	1.28	3min 44s
100	1.17	2min 38s	1.11	7min 57s
500	0.46	13min 23s	0.35	38min 24s

Table: écart-type et temps de calcul de l'estimateur du taux de prévalence de la listeria pour le 57ème pays, calculé sur 50 valeurs, pour différentes tailles de simulation