Prévalence du listeria dans le lait cru

Claire He, Robin Guillot, Mathilde Binet

ENSAE Paris

May 19, 2021

Plan

Données

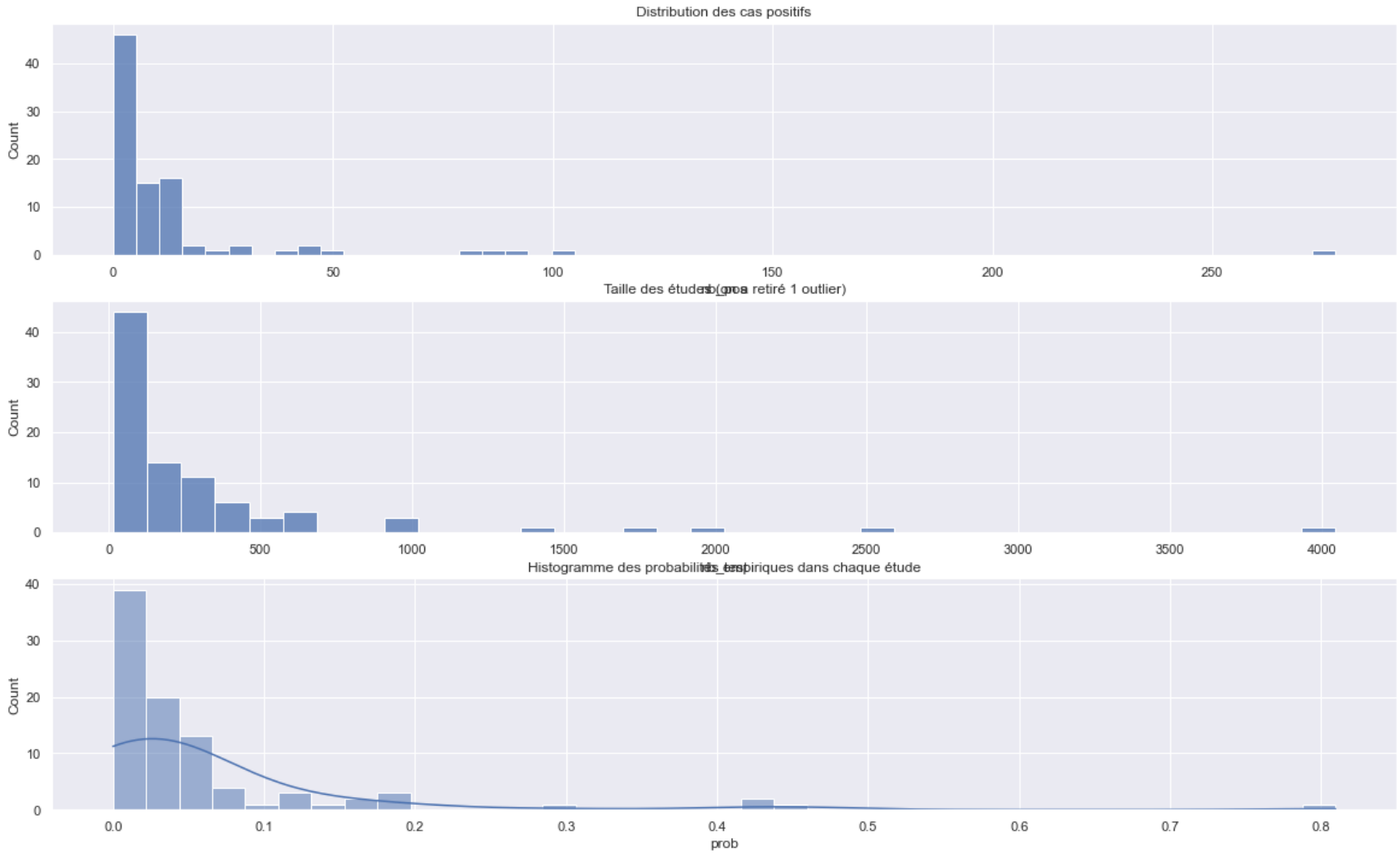


Figure 1: Histogrammes sur les données `listeria`

Problème

Objectif

Caractériser au mieux la loi que suivent les probabilités de positivité au listeria dans chacune des 91 études. Pour cela, une **approche bayésienne** est adoptée : une loi a priori est donnée pour ces probabilités ; puis cette loi est "améliorée" via les données, en caractérisant au mieux la loi a posteriori.

Démarche

Déroulement de la démarche

- Cas 1 : Probabilité constante de présence dans le listeria — la loi a posteriori peut être calculée explicitement.
- Cas 2 : loi a priori suivant une loi $\mathcal{B}eta$ — simulation de la loi a posteriori avec utilisation de méthodes MCMC
- Cas 3 : loi a priori suivant un mélange de lois $\mathcal{B}eta$ — adaptation de la méthode précédente
- Comparaison des modèles obtenus

Loi a posteriori dans un cas simple

On suppose dans un premier temps que la probabilité de présence de listeria dans toutes les études est constante p.

Modèle binomial

On suppose $r_i \sim \mathcal{B}(n_i, p)$, de fonction de masse :

$$f(r_i, p) = \binom{n_i}{r_i} p^{r_i} (1-p)^{n_i - r_i}$$

Loi a priori

$p \sim \mathcal{B}eta(\alpha, \beta)$ avec $\alpha, \beta = 1$ ce qui revient à avoir $p \sim \mathcal{U}([0,1])$ de densité $\mathbb{1}_{[0,1]}(p)$

Loi a posteriori dans un cas simple

Loi a posteriori

La densité de la loi a posteriori est alors (n=91 études) :

$$\begin{aligned}
\pi(p|r_1,\ldots,r_n) &\propto \prod_{i=1}^{n} f(r_i,p)\mathbb{1}_{[0,1]}(p) \\
&\propto \prod_{i=1}^{n} p^{r_i}(1-p)^{(n_i-r_i)}\mathbb{1}_{[0,1]}(p) \\
&\propto p^{\sum_{i=1}^{n} r_i}(1-p)^{\sum_{i=1}^{n}(n_i-r_i)}\mathbb{1}_{[0,1]}(p) \\
&\sim \mathcal{B}eta(\sum_{i=1}^{n} r_i + 1, \sum_{i=1}^{n}(n_i-r_i)+1)
\end{aligned}$$

Paramètres obtenus via les observations : $\mathcal{B}eta(1329, 1255727)$.

Loi a posteriori : Démarche et commentaires

Démarche

Simulation de la loi Bêta :

- Simulation de lois Gamma : algorithme de rejet-acceptation [1]
- Utilisation de la propriété suivante :
  Soient X et Y deux variables indépendantes distribuées suivant des lois $\Gamma(a, 1)$ et $\Gamma(b, 1)$:
  Alors $U = X + Y$ et $V = X/(X + Y)$ sont deux variables indépendantes distribuées suivant des lois $\Gamma(a + b, 1)$ et $\mathcal{B}eta(a, b)$.

Commentaires

Coût computationnel pour des paramètres "grands" - notre cas : 1:08:00 pour 500 simulations en fig. 2.

[1] voir annexe pour le calcul de la constante

Loi a posteriori : simulations à la main

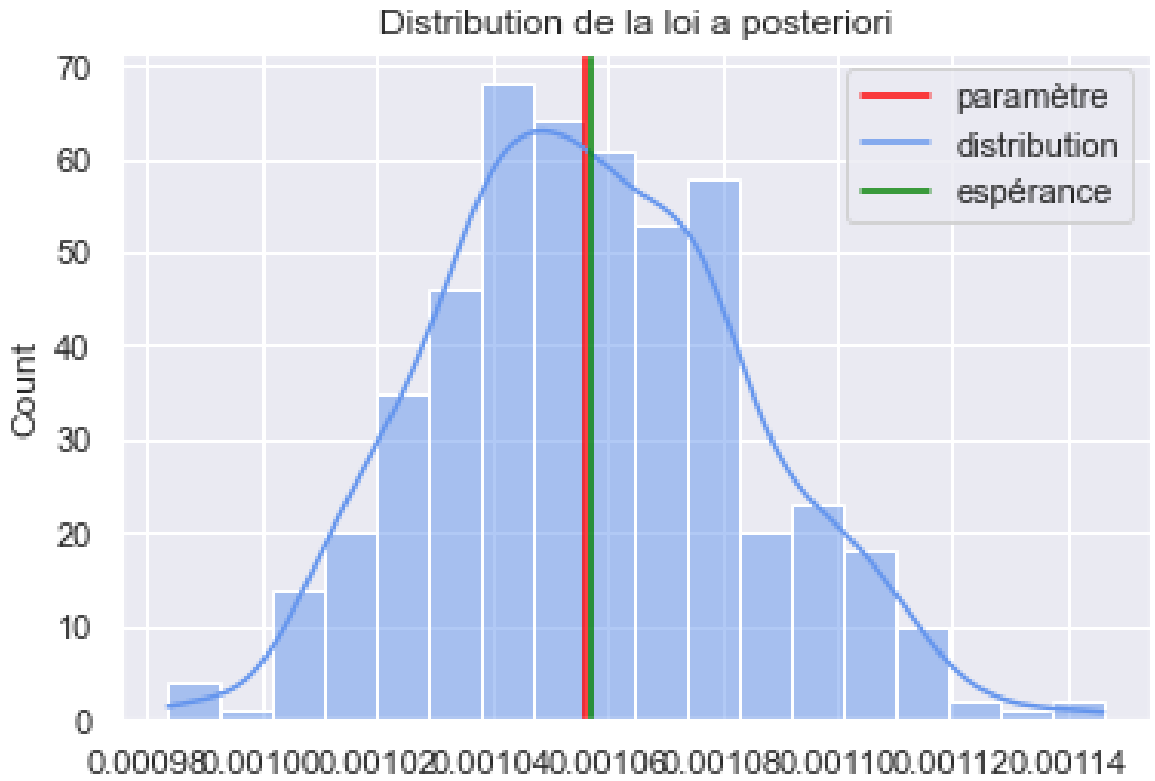


Figure 2: Distribution de la loi a posteriori pour 500 simulations de $\mathcal{B}eta(1329, 1255727)$

Loi a posteriori : simulations avec `scipy.stats`

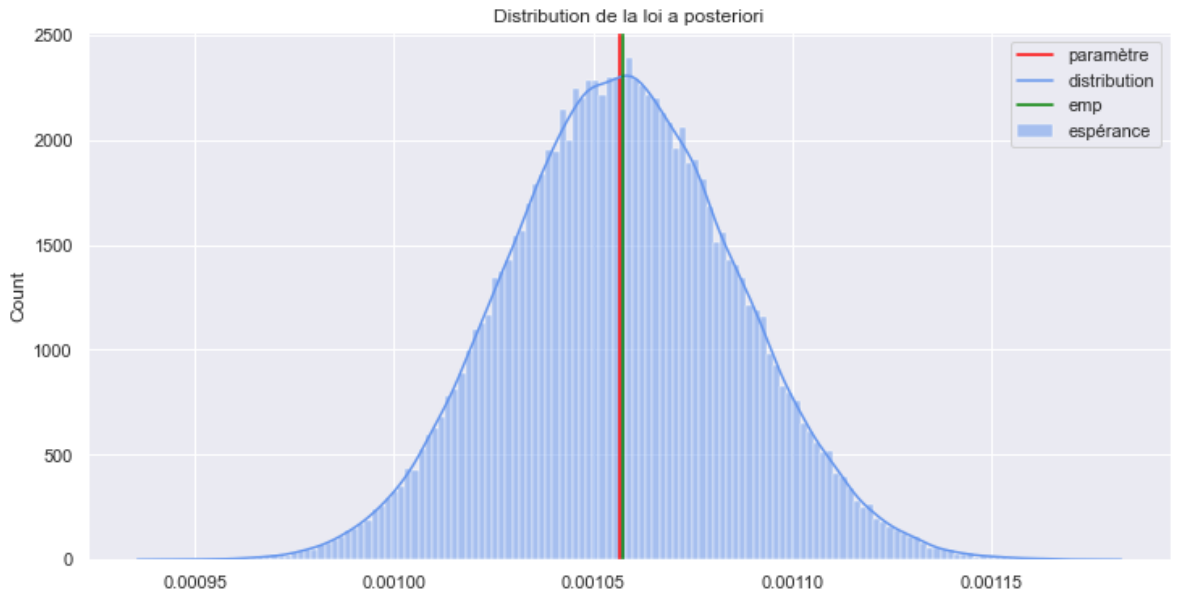


Figure 3: Distribution de la loi a priori

Comparaison données empiriques et modèle

Problème du modèle

19 études avec 0 cas de listeria détectés. La loi a posteriori permet d'estimer l'effet agrégé mais elle ne permet pas de capter les effets sur chaque étude. On omet les études de cas nuls (qui posent problème).

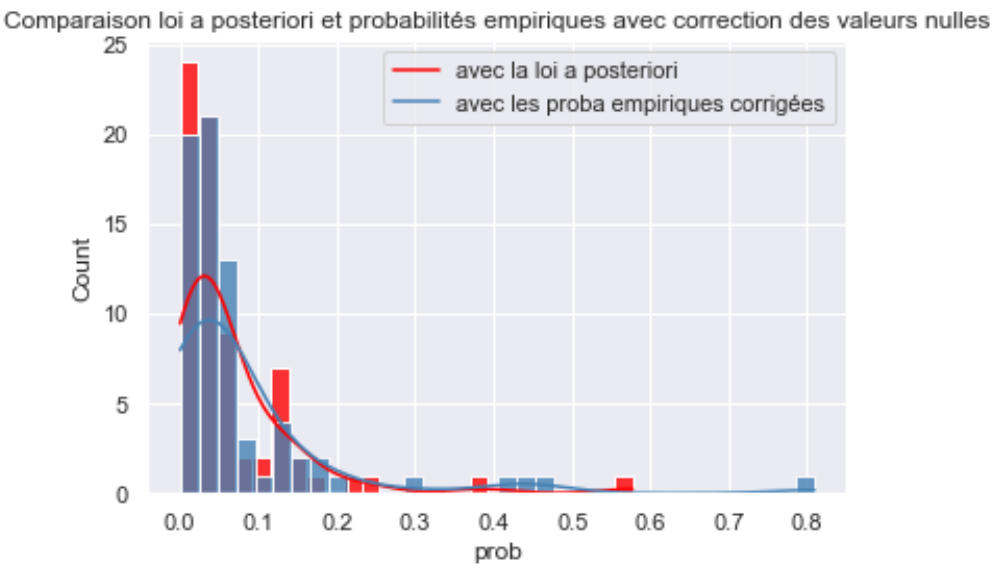


Figure 4: Comparaison entre les données empiriques et le modèle 1

Cas 2 : probabilité de présence de listeria variable

On suppose cette fois que la probabilité de présence de listeria p_i varie selon l'étude i

Modèle binomial

On suppose toujours : $r_i \sim \mathcal{B}(n_i, p_i)$, de fonction de masse :

$$f(r_i, p_i) = \binom{n_i}{r_i} p_i^{r_i} (1 - p_i)^{n_i - r_i}$$

Loi a priori

$\forall i, p_i \sim \mathcal{B}eta(\alpha, \beta)$, de densité

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha - 1} (1 - p_i)^{\beta - 1} \mathbb{1}_{[0,1]}(p_i)$$

Probabilité de présence de listeria variable

Reparamétrisation

$\mu = \alpha/(\alpha+\beta) \sim \mathcal{U}([0,1])$ et $\kappa = \alpha + \beta \sim \mathcal{E}(0.1)$
En inversant : $\alpha = \kappa\mu$ et $\beta = \kappa(1-\mu)$
On supposera que : κ **et** μ **sont indépendantes**
Donc on a la densité jointe de (μ, κ) suivante :

$$p(\mu,\kappa) = 0.1e^{-0.1\kappa}\mathbb{1}_{]0;\infty[}(\kappa)\mathbb{1}_{[0,1]}(\mu)$$

Et donc p_i a pour densité (sachant μ et κ) :

$$p(p_i|\mu,\kappa) = \frac{\Gamma(\kappa)}{\Gamma(\kappa\mu)\Gamma(\kappa(1-\mu))} p_i^{\kappa\mu-1}(1-p_i)^{\kappa(1-\mu)-1}\mathbb{1}_{[0,1]}(p_i)$$

Cas 2 : Loi a posteriori

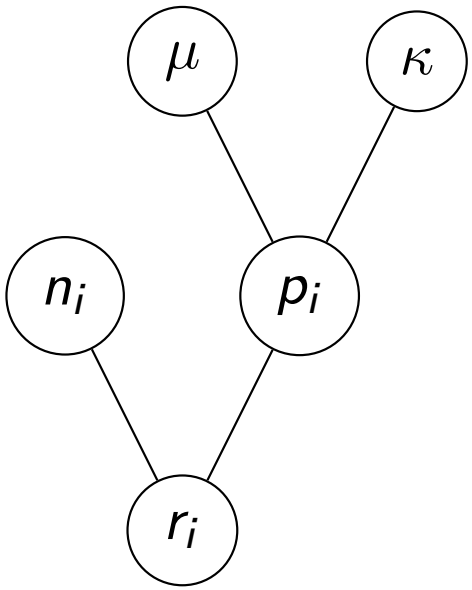
Loi a posteriori

La densité de la loi a posteriori est alors :

$$\begin{aligned}
\pi(\mu, \kappa, p_1, ..., p_n | r_1, \ldots, r_n) &\propto \prod_{i=1}^{n} f(r_i, p_i) p(p_i | \mu, \kappa) p(\mu, \kappa) \\
&\propto e^{-0.1 n \kappa} \left(\frac{\Gamma(\kappa)}{\Gamma(\kappa\mu)\Gamma(\kappa(1-\mu))} \right)^n \\
&\times \prod_{i=1}^{n} p_i^{r_i + \kappa\mu - 1} \\
&\times (1 - p_i)^{(n_i - r_i) + \kappa(1-\mu) - 1} \mathbb{1}_{[0,1]}(p_i) \\
&\times \mathbb{1}_{]0;\infty[}(\kappa) \mathbb{1}_{[0,1]}(\mu)
\end{aligned}$$

Cas 2 : Loi a posteriori

Contrairement au cas précédent, la loi a posteriori ne correspond pas à une loi usuelle connue. On va alors mettre en oeuvre un algorithme de **Metropolis within Gibbs** pour simuler cette loi a posteriori, en s'appuyant sur les différentes lois (connues) de nos paramètres $(p_i)_i, \kappa$ et μ.



Cas 2 : Metropolis within Gibbs

Démarche du Gibbs Sampler

On doit connaître les lois conditionnelles. Le principe du Gibbs sampler est de générer une chaîne de Markov invariante pour la distribution à partir des lois conditionnelles.

1. INPUT: $X_{n-1} = (\mu^{(n-1)}, \kappa^{(n-1)}, p^{(n-1)})$
2. On génère $\mu^{(n)}|\kappa^{(n-1)}, p \sim p(\mu|\kappa, p^{(n-1)}, r)$
3. On génère $\kappa^{(n)}|\mu^{(n-1)}, p \sim p(\kappa|\mu, p^{(n-1)}, r)$
4. On génère pour tout i, $p_i^{(n)}|\mu^{(n-1)}, \kappa^{(n-1)}, p_{-i}^{(n-1)} \sim p(p_i|\mu, \kappa, p_{-i}, r) \sim \mathcal{B}eta(\mu\kappa + r_i, n_i - r_i + \kappa(1-\mu)$
5. OUTPUT: X_n

Metropolis-step

Démarche du Metropolis-Hastings Random Walk

Pour μ et κ, nous ne connaissons pas les lois conditionnelles. On peut en effet bien générer à l'étape 4. la loi Bêta. La démarche du Gibbs sampler va donc "coincer" aux étapes 2. et 3. On va utiliser un algorithme de **Metropolis-Hastings** avec marche aléatoire pour simuler les lois conditionnelles correspondant à μ et κ.

Cas 2 : Metropolis within Gibbs

Démarche

1. INPUT: $(\mu^{(n-1)}, \kappa^{(n-1)})$
2. On introduit l'incrément $\mu' = \mu^{(n-1)} + \sigma_\mu \varepsilon$ de la marche aléatoire avec $\varepsilon \sim \mathcal{N}(0,1)$, de même $\kappa' = \kappa^{(n-1)} + \sigma_\kappa \varepsilon$
3. On calcule la probabilité d'acceptation pour μ en passant au log : $\min(0, m^{(n-1)})$ où $m^{(n-1)}$ est déduite du log de la densité de l'a posteriori précisée précédemment, idem pour κ : $\min(0, k^{(n-1)})$[2].
4. On accepte selon la probabilité précédente : alors $\mu^{(n)} = \mu'$ ou on rejette $\mu^{(n)} = \mu^{(n-1)}$. Idem $\kappa^{(n)} = \kappa'$ ou $\kappa^{(n)} = \kappa^{(n-1)}$.
5. OUTPUT: $(\mu^{(n)}, \kappa^{(n)})$

[2] voir annexe

Cas 2 : MCMC

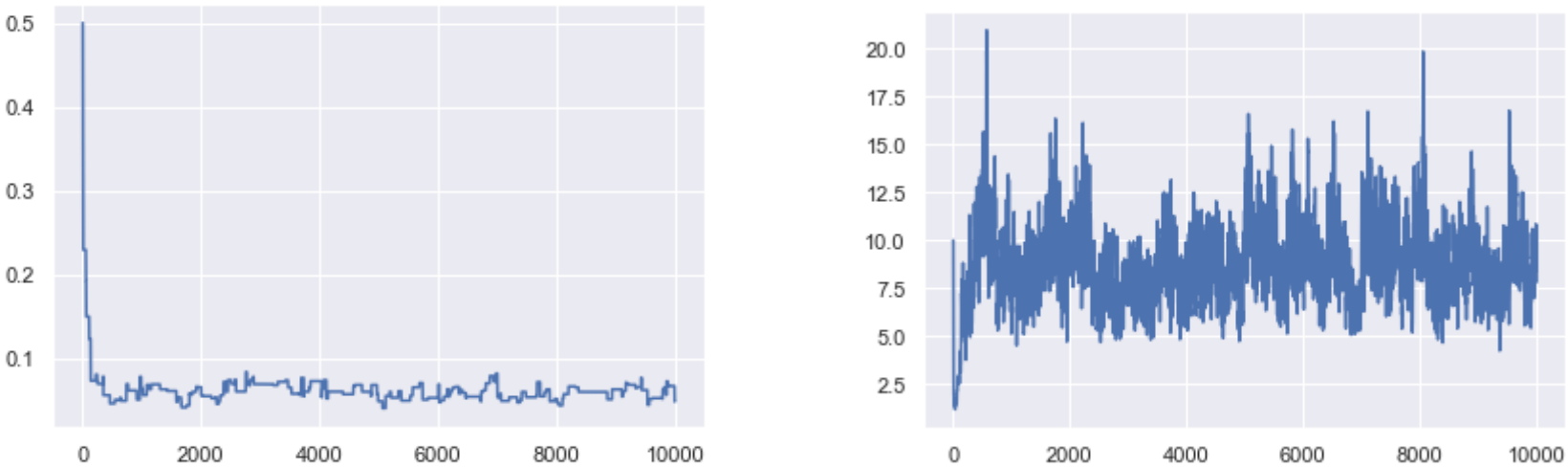
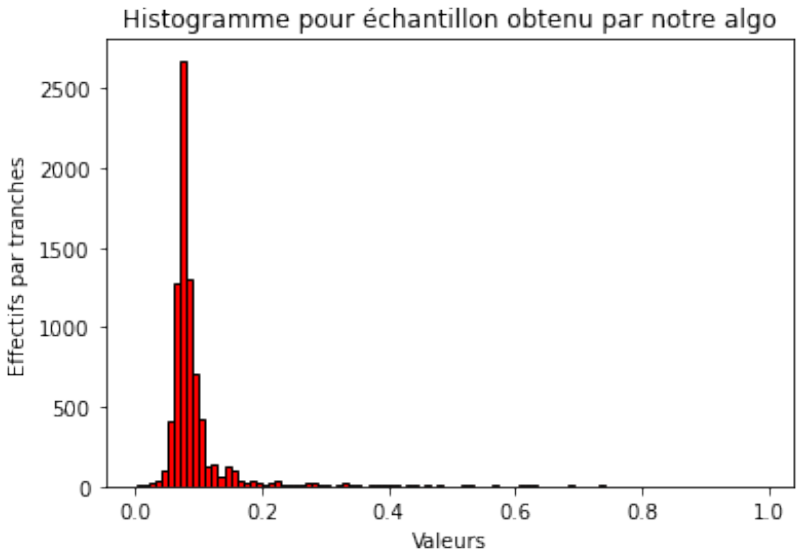
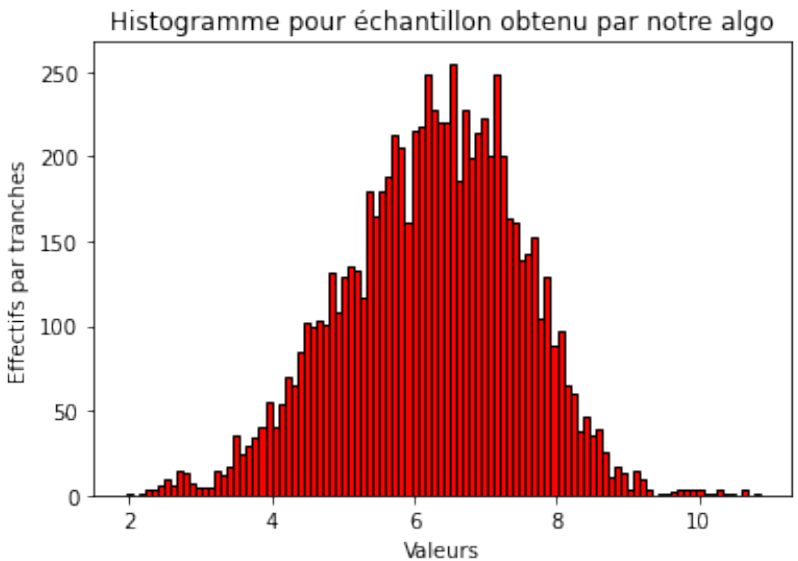


Figure 5: MCMC traces pour μ et κ



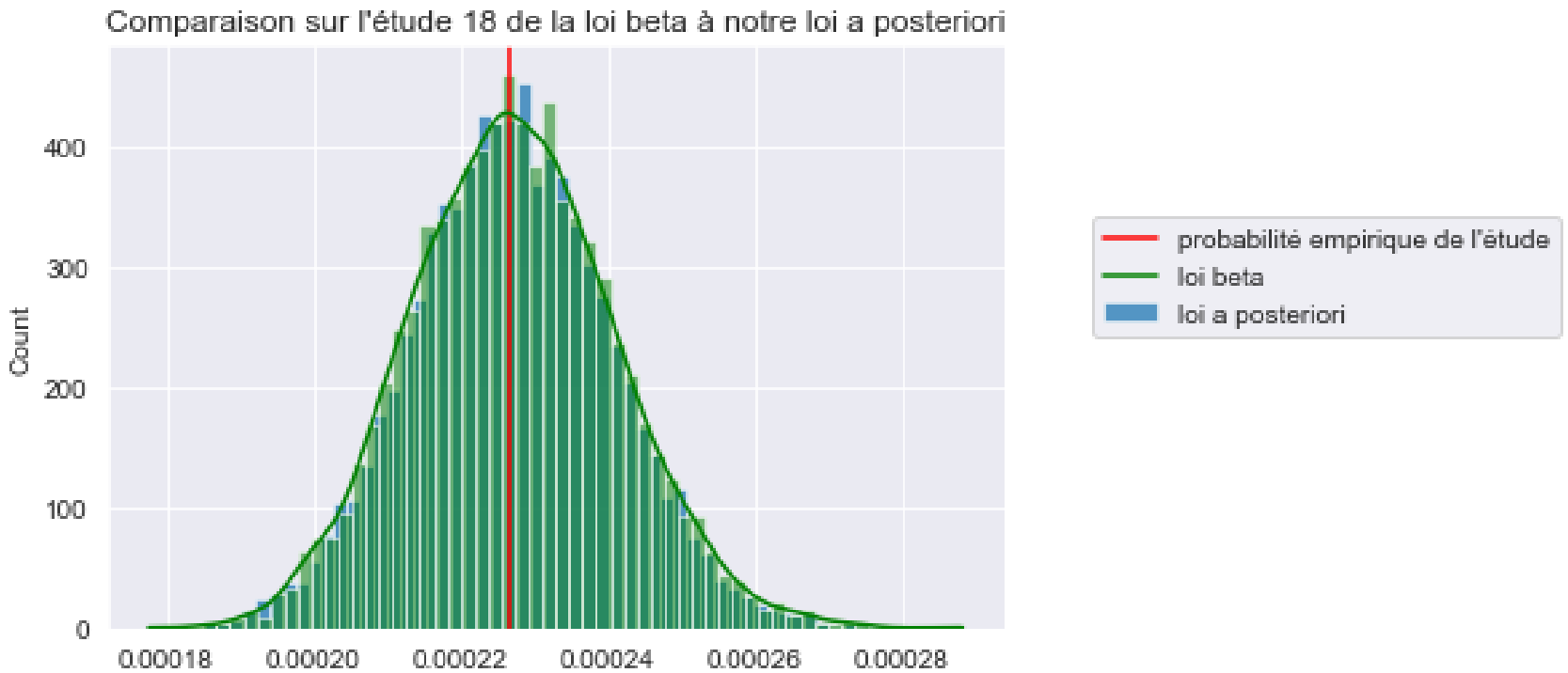


Loi testée	stat.	p-value
$\mathcal{N}(0.096, 0.069^2)$	0.006	0.884
$\mathcal{N}(6.24, 1.24)$	0.007	0.871

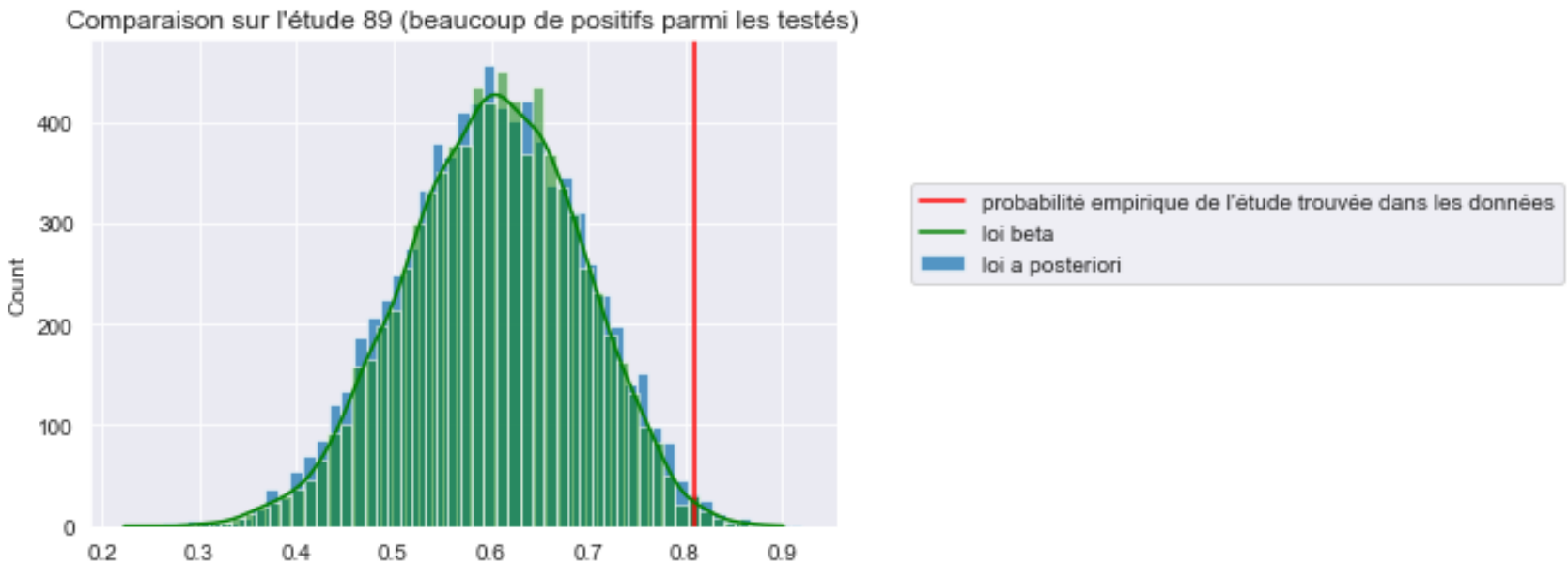
Table 1: Tests de Kolmogorov Smirnov sur les a posteriori de μ puis κ

Les lois a posteriori de μ, κ, sont symétriques mais ce ne sont pas des lois normales.

Distributions a posteriori



Distribution a posteriori



Cas 3 : Probabilité de présence variable et modèle de mélange

Modèle de mélange de l'a priori

Dans ce troisième cadre, on suppose que l'a priori est un mélange de deux lois bêta. Comme on ne connaît pas l'attribution des observations à l'une des lois Beta, on crée une variable latente w_i qui attribue la probabilité a posteriori de chaque observation d'être dans une des deux populations.

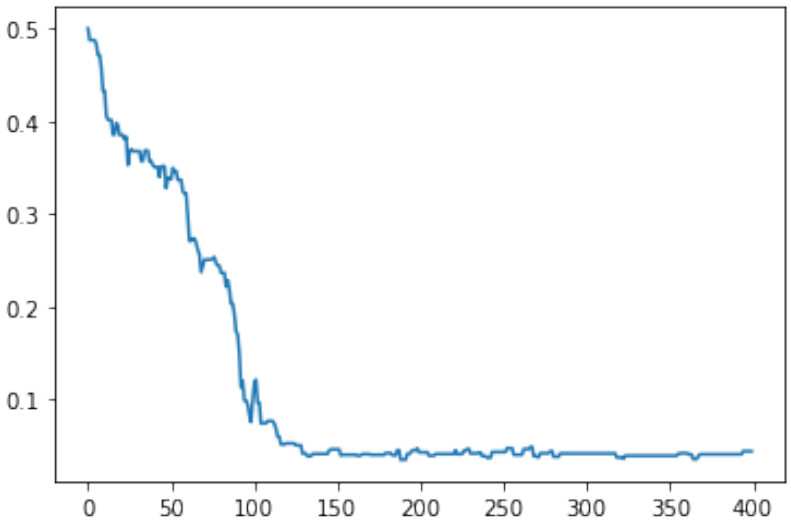
$$p_i \sim w_i \times \mathcal{B}(\mu_1, \kappa_1) + (1 - w_i) \times \mathcal{B}(\mu_0, \kappa_0)$$

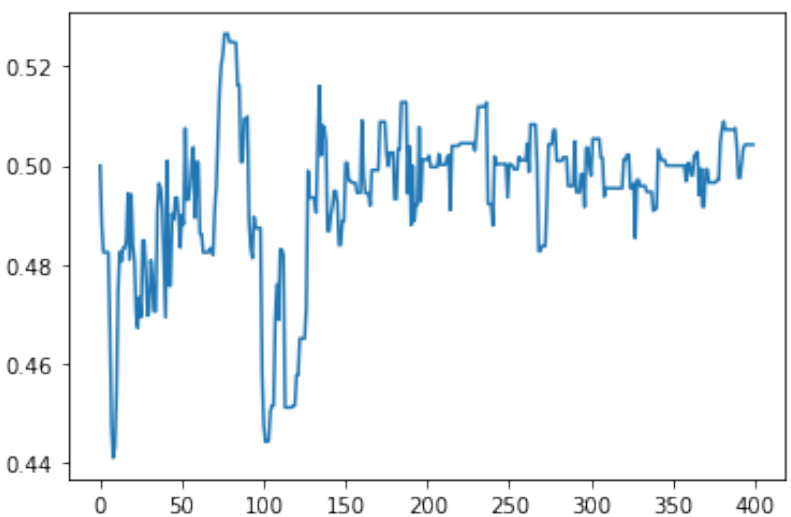
Démarche

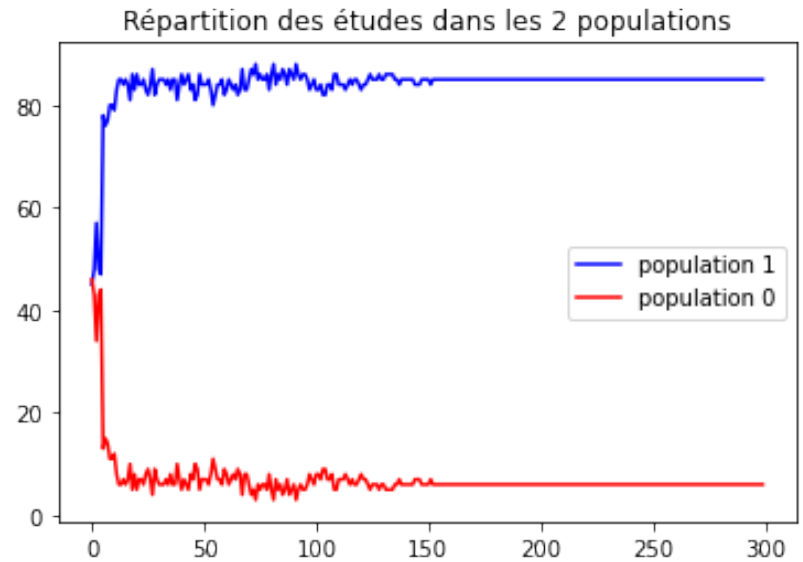
Ainsi, à chaque étape n :

1. On actualise les $(w_i)_i$, ce qui va séparer à l'étape$n$ notre échnatillon d'études en deux sous-populations
2. Pour toutes les études i dont on a obtenu w_i=1, algorithme de Metropolis within Gibbs de paramètres μ_0 et κ_0, ie on fait :
   - actualisation des $(p_i)_i$ concernés
   - actualisation de μ_0
   - actualisation de κ_0
3. Pour toutes les études i dont on a obtenu w_i=1, algorithme de Metropolis within Gibbs de paramètres μ_1 et κ_1, ie on fait :
   - actualisation des $(p_i)_i$ concernés
   - actualisation de μ_1
   - actualisation de κ_1

Résultats







A gauche, trace MCMC pour mu_0 et mu_1 dans le modèle de mélange.

Résultat

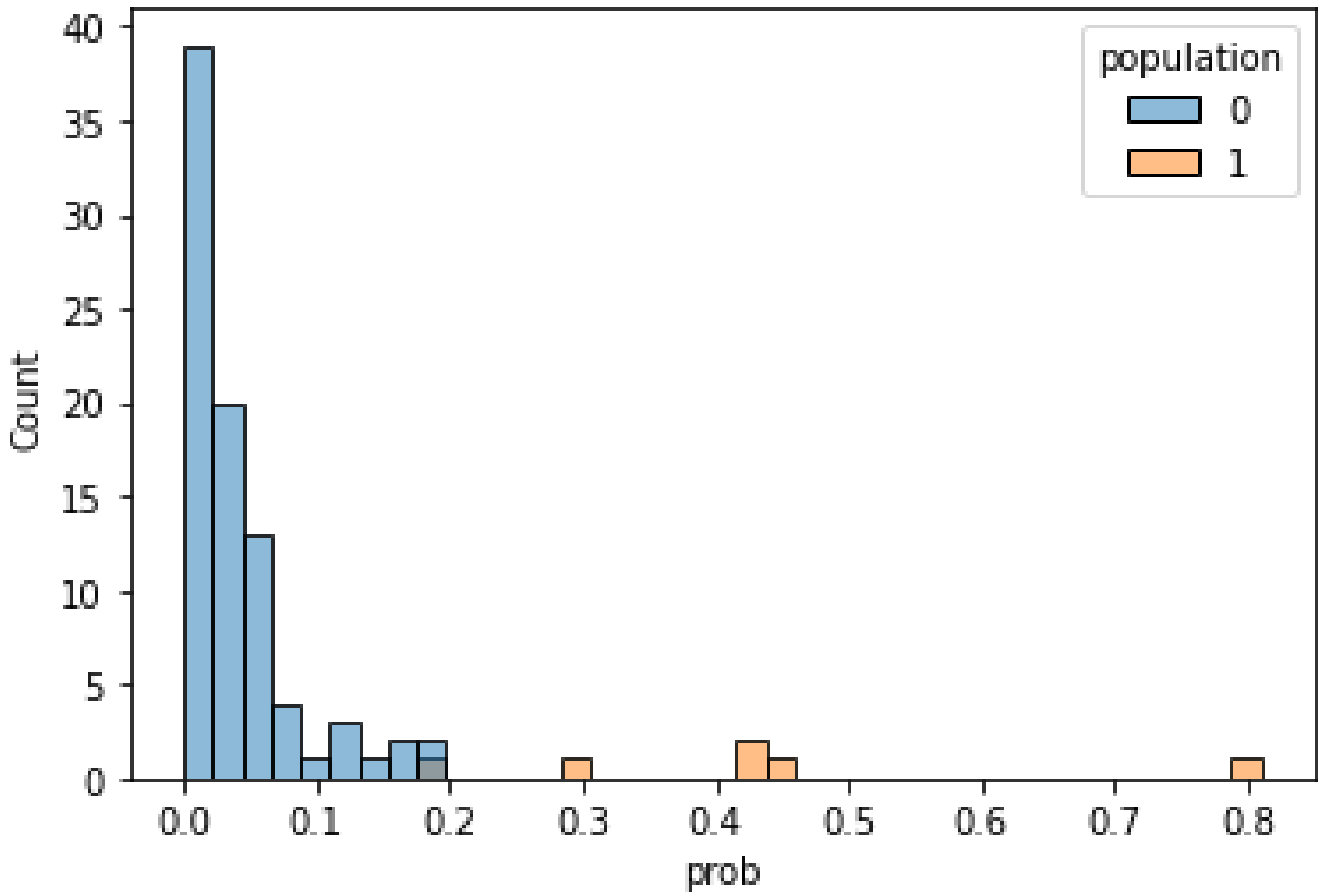


Figure 6: Résultat dans l'étude avec modèle de mélange

Critères de convergence

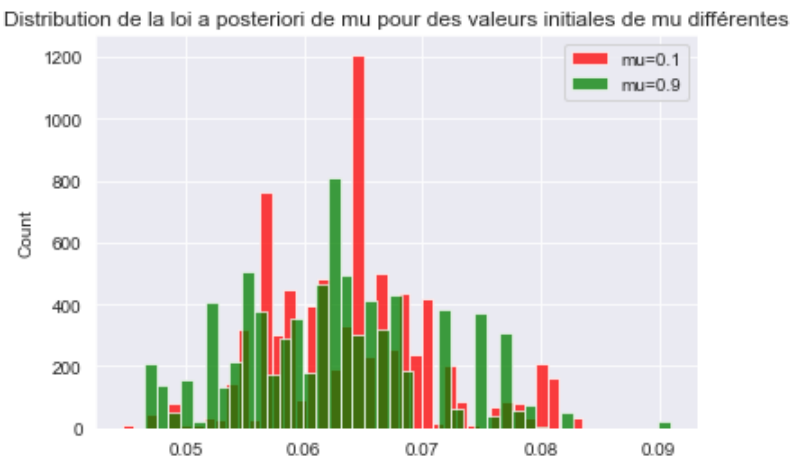
Maintenant qu'on a simuler nos lois a posteriori, on peut faire varier σ_1 et σ_2 associés à μ et κ dans le MHRW et analyser certains critères de convergence :

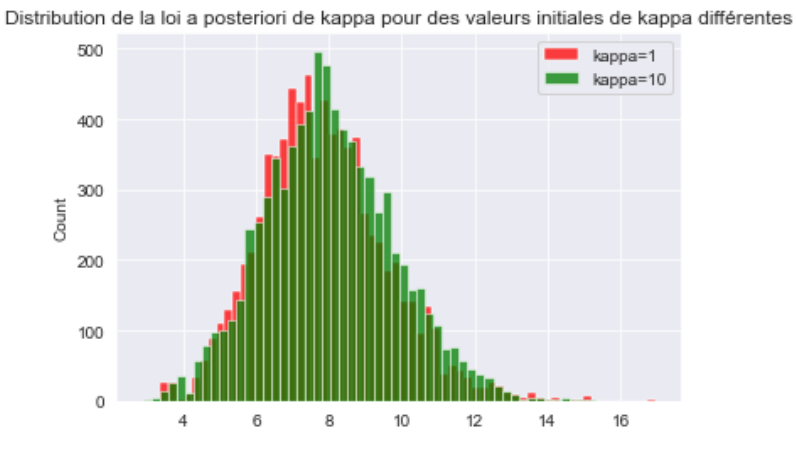
- trace du paramètre d'intérêt : le paramètre doit fluctuer aléatoirement ;
- l'histogramme des valeurs générées ne dépend pas des valeurs initiales ;
- taux d'acceptation pas trop élevés ni trop faibles ;
- tracer les fonctions d'autocorrélation : celles-ci doivent décroître rapidement
- utiliser le DSCM.

Critère des valeurs initiales

Remarque

La distribution simulée pour la loi a posteriori ne doit pas dépendre de la valeur initiale prise par le paramètre. On vérifie effectivement ce critère :





Fonctions d'autocorrélation modèle 2

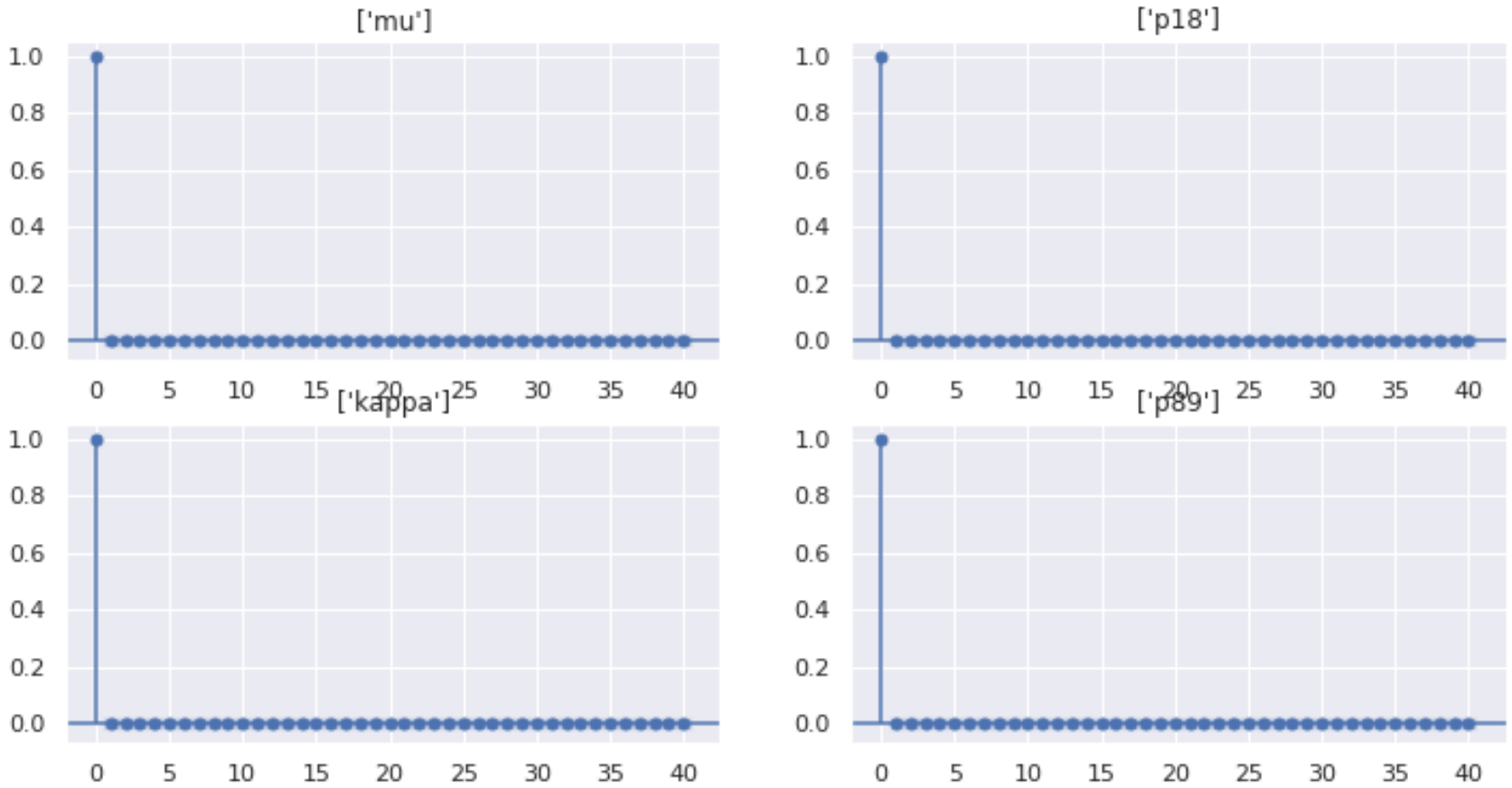
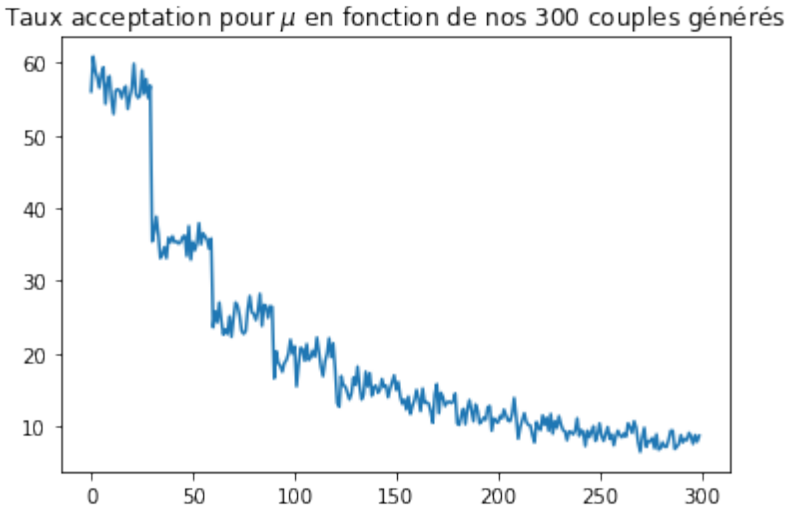
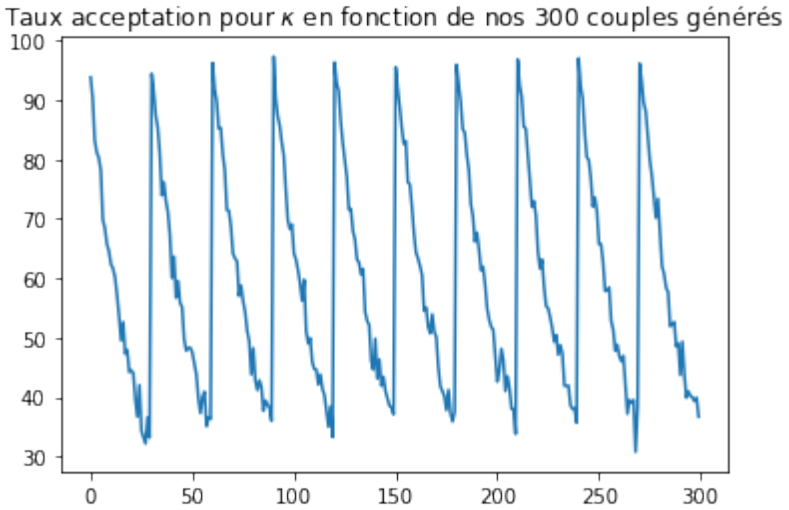
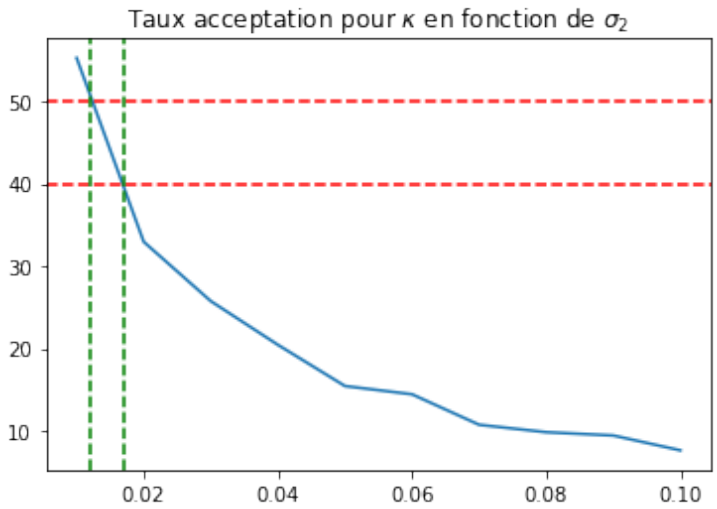


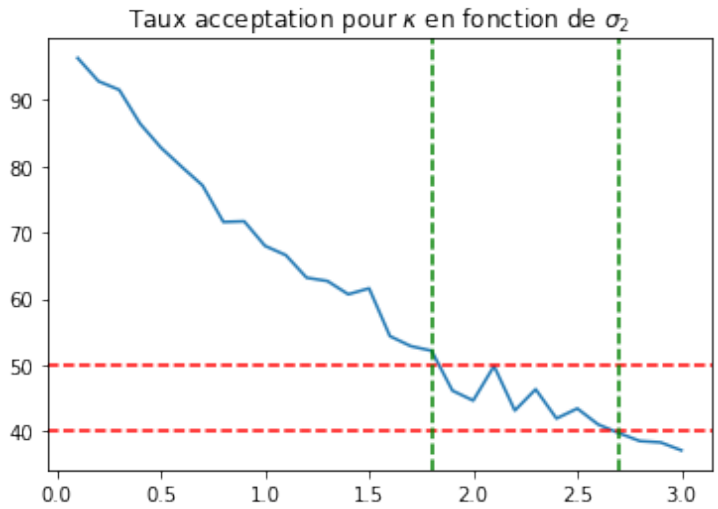
Figure 7: Fonctions d'autocorrélation pour les études 18, 89 et pour les lois a posteriori associées à μ et κ

Taux d'acceptation (40-50 %)

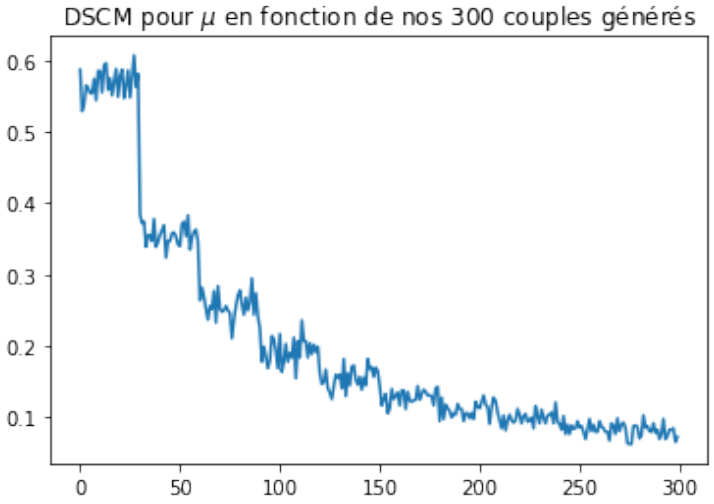


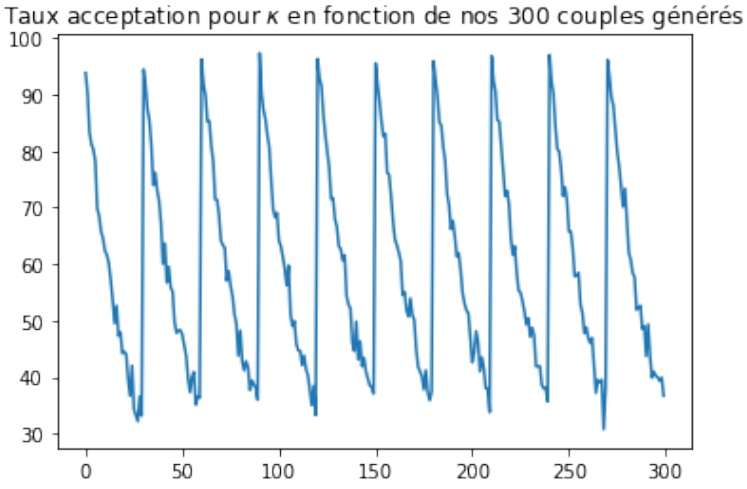


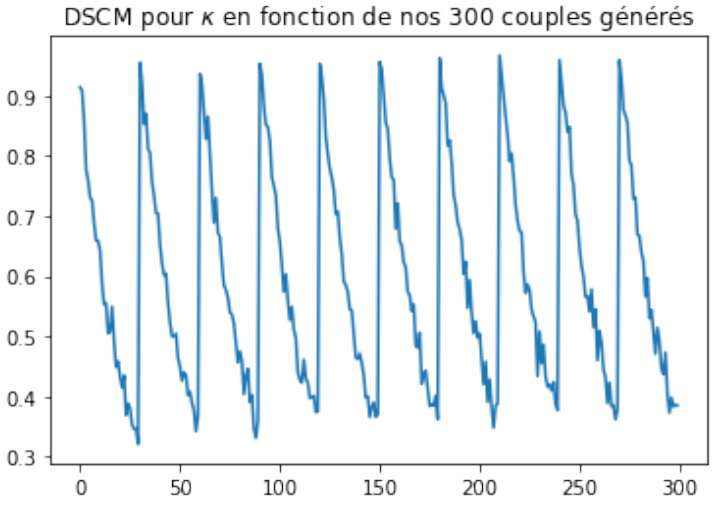


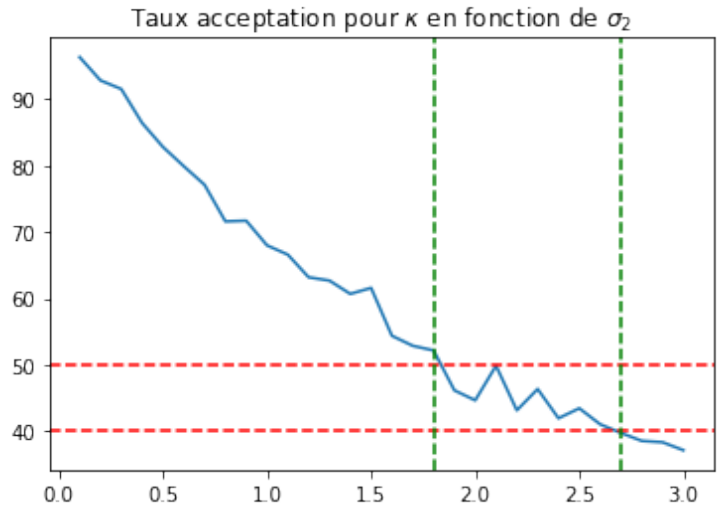


DSCM


DSCM pour μ en fonction de nos 300 couples générés
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300


Taux acceptation pour κ en fonction de nos 300 couples générés
100
90
80
70
60
50
40
30
0
50
100
150
200
250
300


DSCM pour κ en fonction de nos 300 couples générés
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
50
100
150
200
250
300


Taux acceptation pour κ en fonction de σ_2
90
80
70
60
50
40
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Conclusion sur les modèles

- Modèle 1 : simpliste, satisfaisant pour l'observation agrégée mais n'explique pas chaque étude
- Modèle 2 & 3 : approche bayésienne affinée - prennent en compte les variations proba possibles dues aux caractéristiques des pays d'étude

Annexe

- Calcul de la constante de rejet-acceptation de la loi Gamma
- Calcul des lois conditionnelles
- Table des valeurs pour σ_1 et σ_2

Calcul de la constante de RA - loi Gamma

Lois conditionnelles