BAYES - Hearts : a mixture model for count data

Noé Sauzede, Antoine Marques

Mars 2025

1 Contexte et description des données

Les données que nous analysons proviennent de Berry (1987) et concernent l'effet d'un médicament sur les patients souffrant de contractions ventriculaires prématurées (PVCs). Chaque patient est observé avant et après l'administration du traitement, avec les variables suivantes :

- x_i : nombre de PVCs avant traitement.
- y_i : nombre de PVCs après traitement.
- $t_i = x_i + y_i$: somme totale des PVCs observées chez un patient.

Un patient est considéré comme "guéri" si $y_i = 0$, mais il est possible qu'un patient ait $y_i = 0$ simplement par hasard, sans être guéri. L'objectif du modèle est donc de distinguer les patients véritablement guéris de ceux qui ont simplement une faible occurrence post-traitement par hasard, ainsi que l'effet du traitement chez les patients non guéris.

1.1 Paramètres nuisibles et simplification conditionnelle

Le modèle initial suppose :

$$X_i \sim \text{Poisson}(\lambda_i), \quad Y_i \mid \lambda_i \sim \text{Poisson}(\beta\lambda_i)$$

où les paramètres λ_i représentent les taux individuels de PVC pour chaque patient. Ces paramètres individuels constituent des paramètres de nuisance car ils sont spécifiques à chaque individu et n'ont pas d'intérêt direct dans l'analyse globale de l'effet du traitement.

Pour éviter d'estimer ces paramètres nuisibles λ_i, on utilise la distribution conditionnelle de Y_i sachant t_i (le total des PVC observés). Il s'agit d'un résultat classique concernant les lois Poisson :

$$Y_i \mid t_i, \beta \sim \text{Binomiale}\,(t_i, p)\,, \quad p = \frac{\beta}{1+\beta}$$

1.2 Interprétation des paramètres du modèle

Le modèle suppose que le traitement peut agir selon deux mécanismes :

- une proportion θ des patients est **guérie** par le traitement, c'est-à-dire qu'ils présentent systématiquement zéro PVC après traitement ($Y_i = 0$) ;
- une proportion $(1-\theta)$ n'est **pas guérie** et leur nombre de PVC après traitement est réduit selon un facteur multiplicatif β, c'est-à-dire que leur probabilité de PVC post-traitement est $p = \frac{\beta}{1+\beta}$.

Ainsi, les paramètres d'intérêt sont :

- θ: la probabilité qu'un patient soit guéri,
- β: l'effet moyen du traitement chez les patients non guéris (rapport des moyennes post/pré-traitement).

1.3 Écriture de la vraisemblance

La vraisemblance du modèle conditionnel aux totaux (t_i) s'écrit comme un modèle de mélange entre une masse ponctuelle en zéro et une loi binomiale :

$$\mathcal{L}(p, \theta \mid y, t) = \prod_{i=1}^{n} \left[\theta I_{\{y_i=0\}} + (1-\theta) \binom{t_i}{y_i} p^{y_i} (1-p)^{t_i - y_i} \right]$$

1.4 Introduction des variables latentes et vraisemblance complétée

Pour simplifier les calculs d'inférence bayésienne, on introduit les variables latentes Z_i telles que :

$$Z_i = \begin{cases} 1, & \text{si le patient } i \text{ est guéri} \\ 0, & \text{sinon} \end{cases}$$

La vraisemblance complétée (incluant les variables latentes Z_i) devient alors :

$$\mathcal{L}_c(p, \theta \mid y, z, t) = \prod_{i=1}^{n} \theta^{Z_i} (1-\theta)^{1-Z_i} \left[\binom{t_i}{y_i} p^{y_i} (1-p)^{t_i - y_i} \right]^{1-Z_i}$$

1.5 Distributions a priori

Dans un cadre bayésien, nous considérons les lois a priori suivantes sur les paramètres α et δ définis par les transformations logit suivantes :

$$\log \frac{p}{1-p} = \alpha, \quad \log \frac{\theta}{1-\theta} = \delta$$

Nous utilisons des lois peu informatives normales :

$$\alpha \sim \mathcal{N}(0, 10^3), \quad \delta \sim \mathcal{N}(0, 10^3)$$

1.6 Distributions à posteriori conditionnelles

Les lois conditionnelles nécessaires pour le Gibbs sampler sont obtenues par le théorème de Bayes :

1. **Variables latentes** Z_i :

$$Z_i \mid \alpha, \delta, y, t \sim \text{Bernoulli}\left(\frac{\theta}{\theta + (1-\theta)(1-p)^{t_i}} \right)^{I_{y_i=0}} \quad \text{et} \quad Z_i = 0 \text{ si } y_i > 0$$

2. **Paramètre** α :

$$\pi(\alpha \mid \delta, z, y, t) \propto \pi(\alpha) \prod_{i:Z_i=0} \binom{t_i}{y_i} p^{y_i} (1-p)^{t_i - y_i}$$

3. **Paramètre** δ :

$$\pi(\delta \mid \alpha, z, y, t) \propto \pi(\delta) \theta^{\sum Z_i} (1-\theta)^{n - \sum Z_i}$$

Ces deux dernières distributions conditionnelles n'étant pas standard, elles nécessitent l'usage d'un échantillonneur Metropolis-Hastings dans les étapes d'un Gibbs sampler.

Ce cadre bayésien permet d'estimer efficacement les paramètres d'intérêt θ et β, en contournant les difficultés posées par les paramètres nuisibles individuels λ_i.

2 Inférence MCMC

L'estimation utilise un échantillonnage de Monte Carlo par Chaîne de Markov (MCMC) via BUGS. L'algorithme suit ces étapes :

- Initialisation des paramètres α et δ.
- Mise à jour de α et δ via Metropolis-Hastings ou Gibbs.
- Calcul de p et θ via la fonction logit.
- Simulation de y_i conditionnellement à t_i.

Après une phase de burn-in (fixée à 1000 itérations), les échantillons permettent d'estimer les distributions a posteriori de θ et β.

2.1 Résultats obtenus

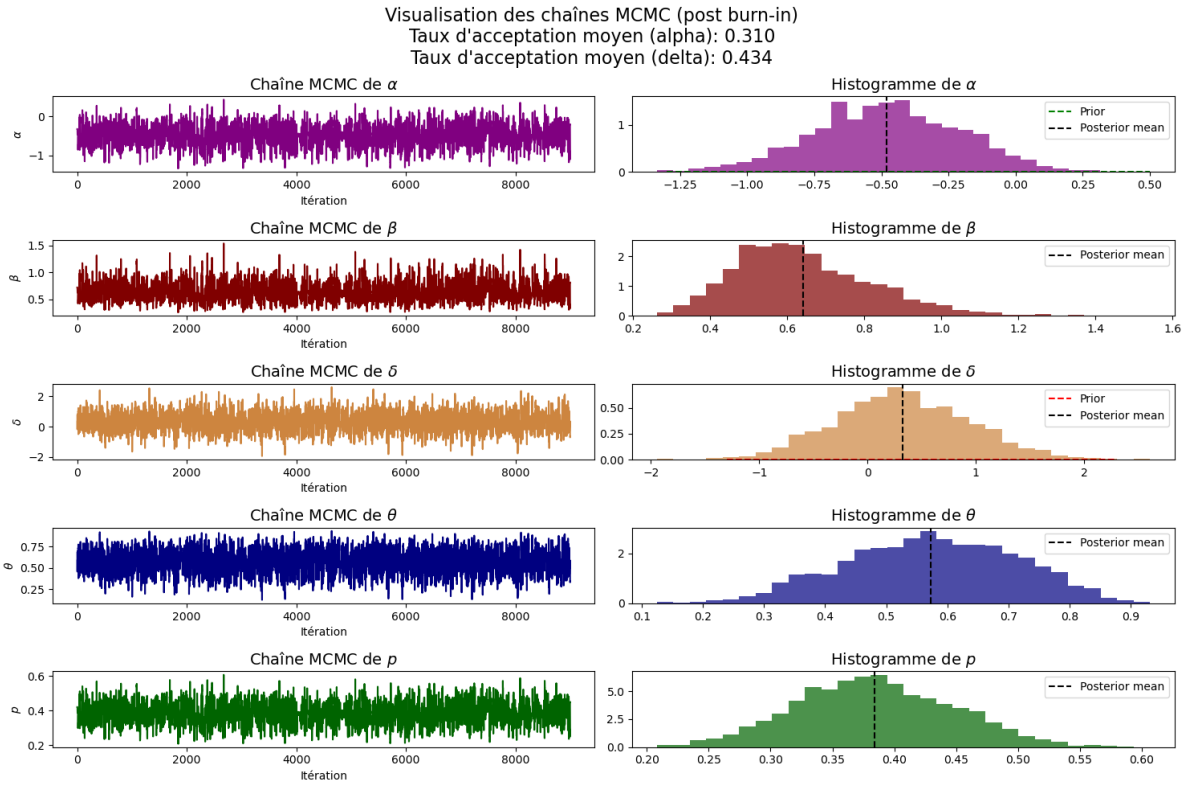

Figure 1: Visualisation des chaînes MCMC pour les paramètres du modèle de mélange.

- Convergence des chaînes
  - Le tracé des chaînes post burn-in (1000 itérations) montrent un bon mélange, une stationnarité apparente et sans tendance marquée ce qui suggère une bonne convergence.
  - Les taux d'acceptation pour α (0.310) et δ (0.434) sont raisonnables, ce qui indique un équilibre entre l'exploration et l'acceptation.
- Comparaison des distributions à priori/posteriori

– La comparaison des distributions à priori (peu informatives) et à posteriori pour les chaînes α et δ montrent que l'information est bien capturé par les lois à posteriori.

Résumé des statistiques descriptives des chaînes MCMC

	mean	sd	MC_error	val2.5pc	median	val97.5pc
alpha	-0.4801	0.2739	0.0029	-1.0288	-0.4803	0.0313
beta	0.6423	0.1772	0.0019	0.3574	0.6186	1.0318
delta	0.32	0.6251	0.0066	-0.8615	0.3048	1.5687
theta	0.5727	0.1405	0.0015	0.297	0.5756	0.8276
p	0.3843	0.0637	0.0007	0.2633	0.3822	0.5078

Figure 2: Résumé des paramètres MCMC estimés après la période de burn-in

Les résultats sont cohérents avec ceux qui nous été présentés dans le sujet.

- Paramètre β
  La valeur moyenne est de 0.6423 avec un intervalle de crédibilité de $[0.3574, 1.0318]$. La valeur moyenne de β suggère un effet bénéfique du traitement d'un point de vue clinique.

- Paramètre θ
  La valeur moyenne de θ est de 0.5727 avec un intervalle de crédibilité de $[0.297, 0.8276]$, nous indique qu'un patient ayant un nombre de PVC réduit a plus d'une chance sur deux d'être guéri.

- Paramètre p
  La valeur moyenne est de 0.3843 avec un intervalle de crédibilité de $[0.2633, 0.5078]$, ce qui suggère que le traitement a un effet signficativement positif sur les patients non guéris.

3 Conclusion et limites

En conclusion, notre approche bayésienne du modèle de mélange permet d'estimer efficacement l'effet du traitement en distinguant les patients guéris de ceux pour qui le nombre de PVC a simplement diminué. Les résultats montrent un effet bénéfique du traitement, avec une probabilité significative de guérison et une réduction moyenne des PVC chez les patients non guéris.
Cependant, la grande amplitude des intervalles de crédibilité des différents paramètres est dûe au faible nombre de données utilisé (seulement 12 paires (x_i, y_i)). Il serait intéressant de confirmer nos résultats avec l'aide de plus de mesures.