Processus de Markov en temps continu et génétique des populations.

Feuille d'exercices 3 — 2014–2015

Nicolas Champagnat

La formule d'échantillonnage d'Ewens

Exercice 8 : Estimateur du maximum de vraisemblance pour le taux de mutation θ

Pour $\theta > 0$ fixé, on note $\mathbb{P}_\theta$ la loi du coalescent de Kingman avec infinité d'allèles de taux de mutation $\theta/2$. On note K_n le nombre d'allèles observé dans un échantillon de n individus. On rappelle que dans ce cas la loi du spectre de fréquence est donnée par la formule d'échantillonnage d'Ewens.

1. En introduisant les événements $\{\text{coal}\}$ et $\{\text{mut}\}$ comme dans la preuve de la formule d'échantillonnage d'Ewens, montrer que pour tout $1 \leqslant k \leqslant n$,

$$\mathbb{P}_\theta(K_n = k) = \frac{n-1}{\theta+n-1}\mathbb{P}_\theta(K_{n-1} = k) + \frac{\theta}{\theta+n-1}\mathbb{P}_\theta(K_{n-1} = k-1).$$

En déduire que $K_n = A_1 + \ldots + A_n$ en loi, où $A_1, \ldots, A_n$ sont des v.a. indépendantes et A_i a pour loi la loi de Bernoulli de paramètre $\frac{\theta}{\theta+i-1}$.

2. Calculer $\mathbb{E}_\theta(K_n)$ et $\mathrm{Var}_\theta(K_n)$, où Var_θ désigne la variance sous la loi $\mathbb{P}_\theta$. Donner un équivalent de ces deux quantités lorsque $n \to \infty$.

3. Montrer que $\frac{K_n}{\ln n}$ est un estimateur de θ qui converge en probabilité. Calculer un équivalent de sa variance. Commenter sa vitesse de convergence.

4. Montrer que K_n est une statistique suffisante pour θ, c'est-à-dire que pour tout k, la loi du spectre de fréquence $B_1, \ldots, B_n$ sachant $K_n = k$ ne dépend pas de θ.

5. On note $L_n(\theta, k) = \mathbb{P}_\theta(K_n = k)$. Montrer que

$$L_n(\theta, k) = C\frac{\theta^k}{\theta(\theta+1)\ldots(\theta+n-1)}$$

pour une certaine constante C dépendant de k et n.

6. En dérivant $\ln L_n(\theta, k)$ par rapport à θ, montrer que l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ sachant K_n est donné par la relation

$$k = \sum_{i=0}^{n-1} \frac{\hat{\theta}}{\hat{\theta} + i} = \mathbb{E}_{\hat{\theta}}(K_n)$$

et qu'il est équivalent à l'estimateur de la question **3.** lorsque $n \to +\infty$.

7. On note $I_n(\theta)$ l'information de Fisher de K_n, définie par $I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \ln L_n(\theta, K_n)\right)^2\right]$. Calculer $I_n(\theta)$ et donner un équivalent de $I_n(\theta)$ lorsque $n \to +\infty$.

8. **(théorème de Cramer-Rao)** Soit $\tilde{\theta}$ un estimateur sans biais de θ mesurable par rapport à K_n. En dérivant l'expression $\mathbb{E}_\theta[\tilde{\theta}]$ par rapport à θ et en utilisant l'inégalité de Cauchy-Schwartz, montrer que $\mathrm{Var}_\theta(\tilde{\theta}) \geqslant \frac{1}{I_n(\theta)}$.

9. En déduire qu'il n'existe pas d'estimateur de θ convergeant plus vite que $\frac{K_n}{\ln n}$.