École Polytechnique **Promotion X08**
Programmes Éco-Sciences et Maths Appliquées

MAP 563 Modèles Aléatoires pour l'Écologie et l'Évolution (MAEE)
Vincent Bansaye, Amandine Véber, Sylvie Méléard

Génétique des populations, modèles de Wright–Fisher et Moran, coalescent
PC8 et PC9 : 7 et 14 mars 2011

A) Coalescent de Kingman et nombre de sites polymorphes

Le coalescent de Kingman est un processus dans lequel on suit les lignées d'une généalogie asexuée dans le sens rétrospectif du temps, et où chaque paire de lignées, indépendamment, fusionne (*coalesce*) à taux constant 1. On appelle T_n le temps au bout duquel les n lignées n'en forment plus qu'une, c'est-à-dire le temps que mettent les n lignées à rejoindre leur *plus récent ancêtre commun*.

Partie I : temps moyen jusqu'au premier ancêtre commun

I.1. Calculer $\mathbb{E}(\mathrm{T}_n)$.

I.2. Que vaut la limite $\lim_{n\to\infty} \mathbb{E}(\mathrm{T}_n)$. Interprétation ?

Partie II : loi du nombre de sites polymorphes

On suppose que le coalescent de Kingman représente l'histoire généalogique d'un chromosome. Conditionnellement au coalescent, on jette un nuage ponctuel de Poisson de paramètre $\theta/2$ sur ses branches, chaque point correspondant à une mutation qui est supposée toucher à chaque fois un nouveau site de la séquence ADN du chromosome. Cela signifie que

– sur une longueur L de l'arbre généalogique, le nombre de mutations suit une loi de Poisson de paramètre $\theta \mathrm{L}/2$,
– les nombres $\mathrm{Z}_1, \ldots, \mathrm{Z}_k$ de mutations présentes sur k portions *disjointes* de l'arbre sont indépendants.

On note S_n le nombre de sites polymorphes de l'échantillon, c'est-à-dire existant à l'état ancestral dans au moins une des n séquences et à l'état mutant dans au moins une autre.

II.1. On note N_j le nombre de mutations apparues dans l'arbre généalogique lorsque celui-ci était composé de j ancêtres. Calculer la fonction génératrice de N_j. Quelle est la loi de N_j ?

II.2. Calculer l'espérance et la variance du nombre S_n de sites polymorphes.

II.3. Quelle est la fonction génératrice de S_n ?

Partie III : estimation du taux de mutation à partir du nombre de sites polymorphes

On cherche à estimer le taux de mutation θ à partir du nombre S_n de sites polymorphes dans l'échantillon.

III.1. On note $\mathrm{H}_n = \sum_{k=1}^{n-1} 1/k$. Proposer un estimateur $\hat{\theta}_n$ de θ tel que $\mathbb{E}(\hat{\theta}_n) = \theta$.

III.2. Calculer la variance de $\hat{\theta}_n$. A quelle vitesse $\mathbb{E}[(\hat{\theta}_n/\theta - 1)^2]^{1/2}$ tend vers 0 ? Commenter.

III.3. Calculer la limite de $\log \mathbb{E}\left[e^{\mathbf{i}t\sqrt{\mathrm{H}_n}(\hat{\theta}_n - \theta)}\right]$ lorsque $n \to \infty$.

III.4. En déduire que $\sqrt{\frac{\log n}{\hat{\theta}_n}}\left(\hat{\theta}_n - \theta\right) \overset{\text{loi}}{\to} \mathrm{Z}$ lorsque $n \to \infty$, où Z suit une loi gaussienne standard.

B) Coalescent de Kingman et nombre d'haplotypes

On se place toujours dans le même cadre que dans la partie précédente. On appelle K_n le nombre d'haplotypes de l'échantillon, c'est-à-dire le nombre de séquences différentes. Pour obtenir certaines propriétés de K_n, on va utiliser sa relation avec le modèle suivant appelé « processus du restaurant chinois » : des individus numérotés $1, 2, \ldots, n$, arrivent successivement dans une salle de restaurant contenant une infinité de tables infiniment longues. Le premier individu s'assied à la première table. Ensuite, pour tout entier $k \geqslant 1$ l'individu $k+1$ s'assied à côté d'un convive déjà attablé (choisi uniformément au hasard) avec probabilité $1/(k+\theta)$, ou occupe une nouvelle table, avec probabilité $\theta/(k+\theta)$.

B)0. Montrer que le nombre A_n de tables occupées lorsque n convives se sont installés a la même loi que K_n.

Partie I : nombre d'haplotypes

I.1. Montrer que

$$\mathrm{A}_n = \sum_{i=1}^{n} \varepsilon_i,$$

où les $(\varepsilon_i)_{i=1,\ldots,n}$ sont des variables de Bernoulli indépendantes dont on précisera les probabilités de succès respectives.

I.2. En déduire $\mathbb{E}(\mathrm{K}_n)$ et montrer que

$$\mathrm{Var}(\mathrm{K}_n) = \sum_{k=0}^{n-1} \frac{k\theta}{(k+\theta)^2}.$$

I.3. Donner un équivalent de $\mathbb{E}(\mathrm{K}_n)$ lorsque $n \to \infty$ et, en étudiant la différence $\mathrm{Var}(\mathrm{K}_n) - \mathbb{E}(\mathrm{K}_n)$, en déduire un équivalent de $\mathrm{Var}(\mathrm{K}_n)$.

Partie II : estimation du taux de mutation à partir du nombre d'haplotypes

On cherche à estimer θ à partir du nombre K_n d'haplotypes dans l'échantillon.

II.1. Proposer un estimateur $\tilde{\theta}_n$ de θ basé sur K_n et vérifiant $\lim_{n\to\infty} \mathbb{E}(\tilde{\theta}_n) = \theta$.

II.2. Calculer la variance de $\tilde{\theta}_n$ et donner sa vitesse de décroissance lorsque $n \to \infty$. Comparer avec l'estimateur de la question **III.1**.

C) Modèle de Moran

Le modèle de Moran est un modèle d'évolution d'une population de taille fixe N comportant des individus de type « résident » et d'autres de type « mutant ». Lorsque N est grand, la proportion de mutants dans la population peut être approchée par la diffusion suivante :

$$dY_t = s_N Y_t(1 - Y_t)dt + \sqrt{\frac{2 + s_N}{N} Y_t(1 - Y_t)}\, dB_t,$$

où $s_N > -1$ est l'avantage sélectif des mutants.

1. Quel est le générateur de la diffusion Y ?

2. Décrire qualitativement le comportement de cette approximation diffusion lorsque N tend vers l'infini, suivant que $s_N = O(1)$, $s_N = O(1/N)$ ou $s_N = o(1/N)$.

3. Calculer la probabilité de fixation d'une sous-population de Nx mutants.

D) Généalogie de BGW conditionnée à être de taille constante

On considère la généalogie d'un processus de BGW conditionné à être de taille constante N et ayant pour loi de reproduction la loi de Poisson de paramètre m. Ce modèle est intimement lié au modèle de Wright-Fisher. On va voir que dans ce cas la répartition du nombre de descendants suit une loi multinomiale.

On note $Z_1, \ldots, Z_N$ le nombre de descendants des individus $1, \ldots, N$. Les $Z_1, \ldots, Z_N$ sont donc i.i.d. de loi de Poisson de paramètre m et on ramène ensuite la taille de la population à la valeur souhaitée en conditionnant à ce que leur somme soit égale à N.

1. Quelle est la loi de $Z_1 + \ldots + Z_N$?

2. Pour tous entiers $k_1, \ldots, k_N$ vérifiant $k_1 + \ldots + k_N = N$, calculer la probabilité

$$\mathbb{P}(Z_1 = k_1, \ldots, Z_N = k_N \,|\, Z_1 + \ldots + Z_N = N).$$

3. Conclure.