M2 BBS **2017-2018**
Atelier Génétique Statistique

TD1

Génétique des populations : estimation de θ

On s'intéresse ici à l'estimation du paramètre $\theta = 2N\mu$ d'une population, où N est la taille (haploïde) de la population et où μ est le taux de mutation par site et par méiose. Dans la plupart des espèces, μ est un paramètre connu donc l'estimation de θ nous permet indirectement d'accèder à l'estimation de N. Pour estimer θ, supposons que l'on dispose d'un échantillon de n séquences d'ADN de longueur L provenant de la population d'intérêt. On suppose que L est suffisamment petit pour qu'il n'y ait pas eu de recombinaison dans la généalogie de ces n séquences.

Exercice 1. Soit S_n le nombre de sites polymorphes dans l'échantillon. On définit l'estimateur de Watterson par:

$$\theta_W = \frac{1}{L} S_n (\sum_{k=1}^{n-1} \frac{1}{k})^{-1}$$

Nous allons étudier les propriétés de cet estimateur. On suppose pour cela un modèle à infinité de sites (chaque nouvelle mutation se produit sur un nouveau site).
1) Montrer que $S_n = \sum_{k=2}^{n} Y_k$, où Y_k est le nombre de mutations ayant lieu pendant que les n individus de l'échantillon ont k ancêtres communs.
2) Montrer que $\mathbb{E}[Y_k] = \frac{\theta L}{k-1}$.
3) En déduire que θ_W est un estimateur sans biais de θ, c'est à dire que $\mathbb{E}[\theta_W] = \theta$.
4) Montrer que $\mathrm{Var}(S_k) = \frac{\theta L}{k-1} + \frac{(\theta L)^2}{(k-1)^2}$.
5) Montrer que les Y_k sont indépendants et en déduire une expression de $\mathrm{Var}(\theta_W)$.
6) Montrer que $\mathrm{Var}(\theta_W) \to 0$ quand $n \to +\infty$. Cette propriété est importante car elle implique qu'en prenant des échantillons très grands on peut en théorie estimer θ de manière aussi précise qu'on le souhaite.
7) On suppose maintenant qu'on dispose non plus de un locus mais de p locus indépendants de taille L. Chacun fournit un estimateur $\theta_W^{(i)}$ de θ, et on estime globalement θ par :

$$\hat{\theta} = \frac{1}{p} \sum_{i=1}^{p} \theta_W^{(i)}$$

Quelle est la variance de cet estimateur global? Quelle est sa limite quand $p \to +\infty$? Vaut-il mieux augmenter le nombre de séquences n ou le nombre de locus p ?
8) Pour l'espèce humaine, le taux de mutations par site et par méiose est environ de

$2 * 10^{-8}$, et on estime que la taille efficace est de l'ordre de 10 000 haploïdes. Que vaut θ? Calculer l'écart type attendu de θ_W pour un échantillon de taille $n = 100$ et un locus de taille $L = 1000$.

Exercice 2. Un autre estimateur classique est l'estimateur de Tajima (également appelé diversité nucléotidique).

$$\theta_T = \pi_n = \frac{1}{L}\frac{1}{n(n-1)}\sum_{i \neq j}\Pi(i,j)$$

où $\Pi(i,j)$ représente le nombre de différences entre les séquences i et j.
1) En utilisant ce qui a été fait dans l'exercice 1, montrer que $\mathbb{E}[\Pi(i,j)] = L\theta$ pour toute paire (i,j). En déduie que θ_T est un estimateur sans biais de θ.
2) Le calcul de la variance est plus compliqué ici car les statistiques $\Pi(i,j)$ ne sont pas indépendantes. On admet (Tajima, 1983) que

$$\mathrm{Var}(\theta_T) = \frac{1}{L}\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$$

Reprendre la question **8)** de l'exercice 1. Quelle est la limite de $\mathrm{Var}(\theta_T)$ quand $n \to +\infty$? Qu'en déduisez-vous?
3) Montrer que

$$\theta_T = \frac{n}{n-1}\frac{1}{L}\sum_{l=1}^{L} H_l$$

où $H_l = 2p_l(1-p_l)$ est l'hétérozygotie au site l (p_l fréquence de l'allèle 1). On retrouve ainsi que l'estimateur de Tajima est égal à l'hétérozygotie moyenne du locus étudié (à un facteur $\frac{n}{n-1}$ près qui est très proche de 1).