Arbres aléatoires

DS1

M2 Modélisation statistique, 2022–2023

Exercice 1 :

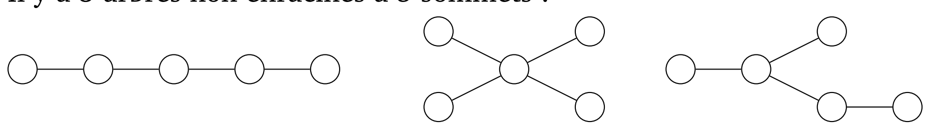
a) Représenter tous les arbres non enracinés à 5 sommets.

b) Pour chaque arbre représenté à la question précédente, combien y a-t-il de façons différentes d'*enraciner* l'arbre ?

c) Représenter tous les arbres enracinés binaires planaires à 7 sommets.

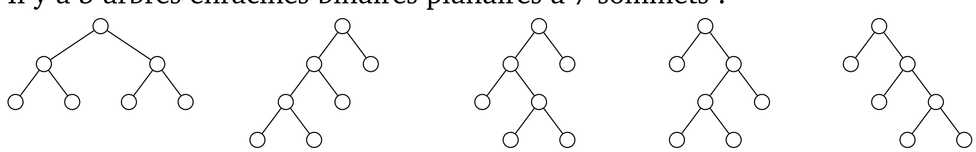
Solution de l'exercice 1.

a) Il y a 3 arbres non enracinés à 5 sommets :



b) En éliminant les symétries, il y a 3 enracinements du premier, 2 enracinements du second et 4 enracinements du troisième arbre représenté ci-dessus.

c) Il y a 5 arbres enracinés binaires planaires à 7 sommets :



Exercice 2 : On considère un arbre aléatoire $T \sim \mathrm{GW}(\mathrm{Poi}(\lambda))$, pour $\lambda > 0$.

a) Pour quelles valeurs de λ l'arbre est-il sous-critique, critique ou sur-critique ?

b) Calculer la fonction génératrice φ de $\xi \sim \mathrm{Poi}(\lambda)$.

c) En déduire que $p_{\mathrm{ext}} = 1/2 \iff \lambda = 2\log(2)$.

d) Pour quelles valeurs de λ a-t-on $p_{\mathrm{ext}} > 1/2$?

Solution de l'exercice 2.

a) Il suffit d'examiner $\mathbb{E}[\xi] = \lambda$. Si $\lambda < 1$ (resp. $= 1$, > 1), T est donc sous-critique (resp. critique, sur-critique).

b) On calcule, pour $z \in [0,1]$,

$$\varphi(z) = \mathbb{E}[z^\xi] = \sum_{n \geqslant 0} e^{-\lambda} \frac{\lambda^n}{n!} z^n = e^{-\lambda} e^{\lambda z} = e^{\lambda(z-1)}.$$

c) On rappelle que la fonction génératrice d'une variable aléatoire à valeurs dans $\mathbb{N}$ ne peut avoir au plus que 2 points fixes dans $[0,1]$. Le point 1 est un point fixe, et p_{ext} est le plus petit point fixe. Ainsi, pour que $p_{\text{ext}} = 1/2$, il faut et il suffit que $1/2$ soit un point fixe de φ. C'est-à-dire

$$p_{\text{ext}} = \frac{1}{2} \iff \frac{1}{2} = e^{\lambda(\frac{1}{2}-1)} \iff \lambda = 2\log(2).$$

d) Puisque $\varphi(0) > 0$ et comme φ a au plus un point fixe dans $]0,1[$, on voit que $\varphi(1/2) > 1/2 \iff p_{\text{ext}} > 1/2$. En effet, la fonction $x \mapsto \varphi(x) - x$ sur $[0,1]$ est toujours positive jusqu'au point p_{ext}, puis négative jusqu'au point 1. Or, $\varphi(1/2) = e^{-\frac{1}{2}\lambda}$ est un terme strictement décroissant en λ, ainsi on a $\varphi(1/2) > 1/2$ si et seulement si $\lambda < 2\log(2)$.

Exercice 3 : On modélise une population de bactéries par un arbre de Yule de paramètre λ.

a) Au bout d'un temps t fixé, on compte $N_t = 42$ individus. Estimer λ (en fonction de t) avec la méthode du maximum de vraisemblance.

b) À votre avis, quelle serait une meilleure façon d'estimer λ si l'on disposait de tout l'arbre généalogique de la population. *On ne demande pas une méthode d'inférence complète et détaillée mais une idée cohérente et argumentée.*

Solution de l'exercice 3.

a) Comme on sait que $N_t \sim \mathrm{Geo}_1(e^{-\lambda t})$, on a, pour tout $n \in \mathbb{N}$:

$$\mathbb{P}(N_t = n) = (1 - e^{-\lambda t})^{n-1} e^{-\lambda t}.$$

En posant $V(\lambda) = (n-1)\log(1 - e^{-\lambda t}) - \lambda t$ le logarithme de cette expression, on cherche à maximiser $V(\lambda)$. On cherche à résoudre $V'(\lambda) = 0$, ce qui équivaut à

$$\begin{aligned} &(n-1)\frac{te^{-\lambda t}}{1 - e^{-\lambda t}} - t = 0 \\ \iff\ & n - 1 = \frac{1 - e^{-\lambda t}}{e^{-\lambda t}} \\ \iff\ & n = e^{\lambda t} \\ \iff\ & \lambda = \frac{\log(n)}{t}. \end{aligned}$$

La fonction $\lambda \mapsto V(\lambda)$ admet un seul point critique, et il est facile de vérifier que c'est un maximum. Ainsi on estime $\hat{\lambda} = \frac{\log(n)}{t}$.

b) Si l'on dispose de tout l'arbre, on a les longueurs de branches. Les branches internes (celles qui sont entre deux nœuds internes) sont i.i.d. de loi $\mathrm{Exp}(\lambda)$, et les branches externes (celles qui soutiennent les feuilles) sont des variables indépendantes de même loi, mais non observées complètement (on connaît seulement une borne inférieure sur leur réalisation – cf. méthodes d'inférence pour données de survie, etc.).

Ainsi, l'on peut calculer une vraisemblance en fonction de λ et de la longueur de chaque branche de l'arbre, puis estimer λ de cette manière avec moins d'incertitude – car on se base sur le même modèle, en utilisant plus d'information.