Arbres aléatoires
DS2

M2 Modélisation statistique, 2022–2023

Exercice 1 :

a) Représenter les quatre arbres enracinés à 4 sommets.

b) Pour chaque arbre de la question (a), combien y a-t-il de représentations planaires différentes de l'arbre ?

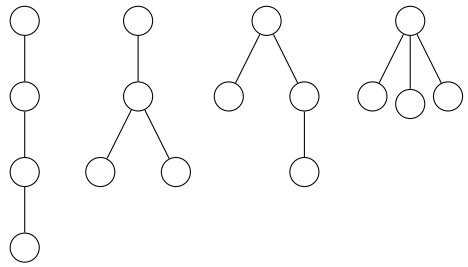
c) Représenter les deux arbres enracinés binaires à 4 feuilles.

d) Pour chaque arbre de la question (c), donner toutes les représentations planaires différentes de l'arbre.

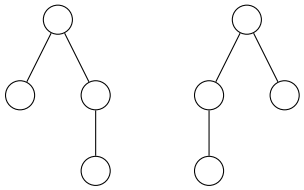
e) Pour chaque arbre de la question (d), combien y a-t-il de représentations ordonnées différentes de l'arbre ?

Solution de l'exercice 1.

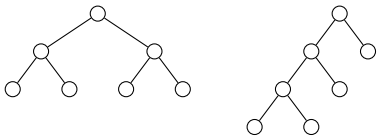
a)



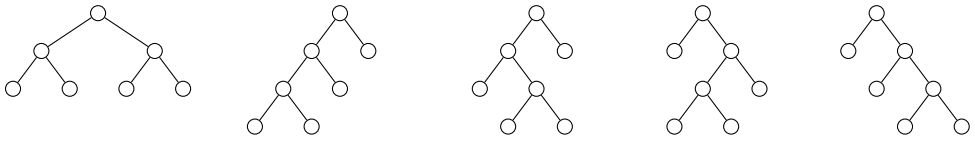
b) Seulement le troisième arbre représenté admet deux représentations planaires distinctes :



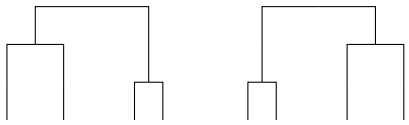
c)



d) Le premier arbre étant symétrique, il n'admet qu'une représentation planaire. Le second en admet quatre, on a donc en tout :



e) Le premier arbre admet deux représentations ordonnées :



Par leur structure, les quatre derniers arbres n'admettent chacun qu'une seule représentation ordonnée.

Exercice 2 : On définit :

- $\mathscr{E}_n$ l'ensemble des arbres enracinés dont tous les sommets sont étiquetés (sans répétition) par $\{1, 2, \dots, n\}$.
- $\mathscr{E}_n^{\mathrm{pl}}$ l'ensemble des arbres enracinés *planaires* dont tous les sommets sont étiquetés (sans répétition) par $\{1, 2, \dots, n\}$.
- $f : \mathscr{E}_n^{\mathrm{pl}} \to \mathscr{E}_n$ la fonction qui "oublie" la structure planaire d'un arbre.

a) Représenter chacun des quatre arbres de $\mathscr{E}_3^{\mathrm{pl}}$ dont la racine porte l'étiquette 1.

b) Des arbres précédents, lesquels ont la même image par la fonction f ?

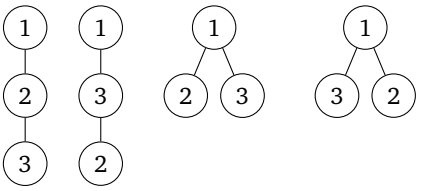
* c) Justifier que dans le cas général, pour $\mathbf{t} \in \mathscr{E}_n$, on a

$$\mathrm{Card}\big(f^{-1}(\mathbf{t})\big) = \prod_{v \in \mathbf{t}} n_v!\,,$$

où comme dans le cours, n_v désigne le nombre d'enfants d'un nœud v.

Solution de l'exercice 2.

a) On a



b) Les deux derniers arbres représentés sont identiques quand on oublie la structure planaire, ils ont donc la même image par f.

c) Dans le cas général, puisque nos arbres sont étiquetés sans répétition, pour chaque nœud de l'arbre $v \in \mathbf{t}$, chaque ordre distinct sur les enfants de v donne un arbre planaire distinct. Ainsi tous les arbres planaires $\tilde{\mathbf{t}}$ satisfaisant $f(\tilde{\mathbf{t}}) = \mathbf{t}$ s'obtiennent en fixant, pour chaque sommet $v \in \mathbf{t}$, un ordre sur les enfants de v, parmi les $n_v!$ ordres possibles.

Finalement, on a $\prod_v n_v!$ façons différentes de poser une structure planaire sur $\mathbf{t}$, c'est donc le cardinal de $f^{-1}(\mathbf{t})$.

Exercice 3 : On rappelle que $X \sim \mathrm{Poi}(1)$ signifie que $\mathbb{P}(X = n) = e^{-1}\frac{1}{n!}$ pour tout $n \in \mathbb{N}$. On considère un arbre aléatoire $T \sim \mathrm{GW}(\mathrm{Poi}(1))$.

a) L'arbre est-il sous-critique, critique ou sur-critique ? Que vaut sa probabilité d'extinction p_{ext} ?

b) Soit $\mathbf{t}$ un arbre planaire fini. Que vaut $\mathbb{P}(T = \mathbf{t})$?

c) On construit $\tilde{T}$ un arbre aléatoire à valeur dans $\bigcup_n \mathscr{C}_n^{\mathrm{pl}}$ en munissant les sommets de T d'étiquettes uniformément choisies (sans répétition) dans les entiers de 1 à N_T = nombre de sommets de T.
Soit $\tilde{\mathbf{t}} \in \mathscr{C}_n^{\mathrm{pl}}$. Que vaut $\mathbb{P}(\tilde{T} = \tilde{\mathbf{t}})$?

d) On reprend la fonction f de l'exercice 2, on fixe $\mathbf{t} \in \mathscr{C}_n$, et l'on choisit $\tilde{\mathbf{t}} \in f^{-1}(\mathbf{t})$. Montrer que

$$\mathbb{P}\big(f(\tilde{T}) = \mathbf{t}\big) = \mathrm{Card}\big(f^{-1}(\mathbf{t})\big)\mathbb{P}(\tilde{T} = \tilde{\mathbf{t}}),$$

et en déduire que $\mathbb{P}\big(f(\tilde{T}) = \mathbf{t}\big) = \frac{e^{-n}}{n!}$.

Solution de l'exercice 3.

a) Si $\xi \sim \mathrm{Poi}(1)$, on a $\mathbb{E}[\xi] = 1$, l'arbre est donc critique, et $p_{\mathrm{ext}} = 1$.

b) D'après la formule du cours,

$$\mathbb{P}(T = \mathbf{t}) = \prod_{v \in \mathbf{t}} \mathbb{P}(\xi = n_v) = \prod_{v \in \mathbf{t}} \frac{e^{-1}}{n_v!}.$$

c) Étant donnée la forme d'arbre de T et conditionnellement à son nombre de sommets N_T, il y a $N_T!$ façons différentes d'attribuer les étiquettes à $\tilde{T}$. Notons $\mathbf{t}$ l'arbre obtenu en oubliant les étiquettes sur les sommets de $\tilde{\mathbf{t}}$. Puisque l'étiquetage est choisi uniformément, on a (on rappelle que $\mathbf{t}$ a n sommets)

$$\mathbb{P}(\tilde{T} = \tilde{\mathbf{t}}) = \frac{1}{n!}\mathbb{P}(T = \mathbf{t}) = \frac{1}{n!}\prod_{v \in \mathbf{t}} \frac{e^{-1}}{n_v!} = \frac{e^{-n}}{n!}\prod_{v \in \mathbf{t}} \frac{1}{n_v!}.$$

d) Toutes les représentations planaires différentes de $\mathbf{t}$ ont la même probabilité selon le modèle de Galton–Watson, parce que l'ensemble des nombres d'enfants reste le même. Ainsi, on a

$$\mathbb{P}(f(\tilde{T}) = \mathbf{t}) = \mathbb{P}(\tilde{T} \in f^{-1}(\mathbf{t})) = \sum_{\hat{\mathbf{t}} \in f^{-1}(\mathbf{t})} \mathbb{P}(\tilde{T} = \hat{\mathbf{t}}) = \mathrm{Card}\big(f^{-1}(\mathbf{t})\big)\mathbb{P}(\tilde{T} = \tilde{\mathbf{t}}).$$

En combinant la question précédente et la question (c) de l'exercice 2, on obtient bien $\mathbb{P}(f(\tilde{T}) = \mathbf{t}) = e^{-n}/n!$.

Exercice 4 : On modélise une population de bactéries par un arbre de Yule de paramètre λ, et on arrête l'évolution de la population quand elle atteint une taille n.

Rappel d'une loi usuelle : si $X \sim \mathrm{Exp}(\mu)$, on a $\mathbb{E}[X] = 1/\mu$ et $\mathrm{Var}(X) = 1/\mu^2$.

a) Construire une variable aléatoire Z qui a la loi du temps qu'on a attendu.
On pourra se donner une suite de variables indépendantes $(X_k)_{k\geqslant 1}$ de lois bien choisies, et construire Z à partir de celles-ci.

b) Donner l'espérance de Z.

c) Donner la variance de Z.

d) Donner la fonction de répartition de Z.

Solution de l'exercice 4.

a) On rappelle que les temps successifs de naissance dans l'arbre de Yule sont séparés par des temps exponentiels indépendants $(X_k)_{k\geqslant 1}$, avec $X_k \sim \mathrm{Exp}(k\lambda)$ pour tout k.
Ainsi, l'on a n individus au bout du temps $Z = X_1 + X_2 + \dots X_{n-1}$ (*j'accepte aussi la solution si l'on prends les n premières variables*).

b) On a donc $\mathbb{E}[Z] = \sum_{k=1}^{n-1} 1/(k\lambda) \sim \log(n)/\lambda$.

c) Similairement, on a donc (car somme de variables deux-à-deux de covariances nulles) $\mathrm{Var}(Z) = \sum_{k=1}^{n-1} 1/(k\lambda)^2 \to \pi^2/(6\lambda^2)$.

d) Avec la construction CPP de l'arbre de Yule arrêté à $n-1$ individus, on voit aussi qu'une construction de Z équivalente est $Z = \max(Y_1, \dots, Y_{n-1})$, où les $Y_1, \dots, Y_{n-1}$ sont i.i.d. de loi $\mathrm{Exp}(\lambda)$. Alors la fonction de répartition se calcule : pour tout $t \geqslant 0$,

$$\mathbb{P}(Z \leqslant t) = \mathbb{P}\Big(\bigcap_{i=1}^{n-1} \{Y_i \leqslant t\}\Big) = \prod_{i=1}^{n-1} \mathbb{P}(Y_i \leqslant t) = (1 - e^{-\lambda t})^{n-1}.$$

Exercice 5 : On considère les données de distances génétiques (fictives) suivantes :

	canard	vélociraptor	alligator	iguane	cobra
canard	0	1	2	15	12
vélociraptor	1	0	4	9	8
alligator	2	4	0	9	7
iguane	15	9	9	0	6
cobra	12	8	7	6	0

a) Appliquer l'algorithme UPGMA pour proposer une phylogénie de ces espèces.

b) Sur deux loci, on observe les bases suivantes :

	canard	vélociraptor	alligator	iguane	cobra
locus 1	A	A	G	A	C
locus 2	A	T	C	C	T

Proposer une décoration (c'est-à-dire placer des séquences génétiques de longueur 2 sur chaque nœud) parcimonieuse de la phylogénie trouvée à la question précédente. Combien de mutations ont été nécessaires pour aboutir aux données observées ?

Solution de l'exercice 5.

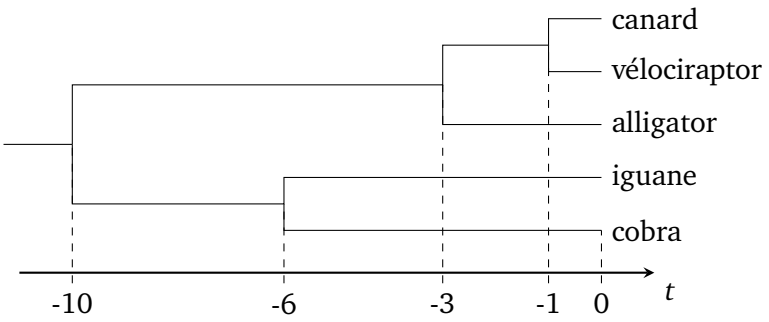
a) Les espèces "canard" et "vélociraptor" sont plus proches (distance 1), on les fusionne, et on calcule les nouvelles distances :

$$d(\mathrm{CV}, \mathrm{alligator}) = \frac{2+4}{2} = 3, \quad d(\mathrm{CV}, \mathrm{iguane}) = \frac{15+9}{2} = 12, \quad d(\mathrm{CV}, \mathrm{cobra}) = \frac{12+8}{2} = 10.$$

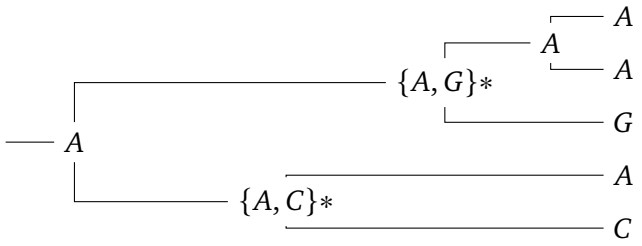
On fusionne ensuite "CV" et "alligator" (distance 3), puis :

$$d(\mathrm{CVA}, \mathrm{iguane}) = \frac{2\times 12+9}{3} = 11, \quad d(\mathrm{CVA}, \mathrm{cobra}) = \frac{2\times 10+7}{3} = 9.$$

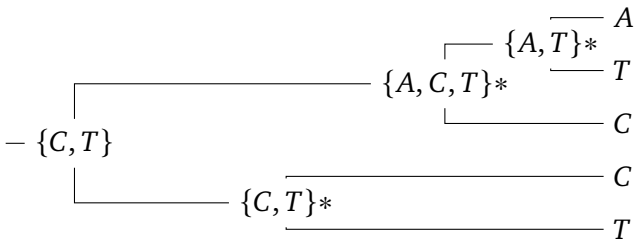
On fusionne ensuite "iguane" et "cobra" (distance 6), qui fusionnent ensuite avec "CVA" à distance $(9+11)/2 = 10$. On se retrouve finalement avec l'arbre phylogénétique suivant :



b) L'algorithme de Fitch nous donne, pour le locus 1 :



et pour le locus 2 :



En tout, on a donc nécessairement cinq mutations (deux pour le locus 1, trois pour le locus 2).