I Modèle de Wright-Fisher

1. En utilisant la fonction `choice`, on propose le script suivant :

```
from  random import choice
def next_P(L):
    return [choice(L) for _ in L]
```

2. La construction de `P[k+1]` se fait en tirant successivement avec remise dans `P[k]`. Ainsi en partant de `P[0]` intialisée à la liste $(1, 2, \dots, 2N)$ on propose :

```
def WF(N,i):
    Lgenes = list(range(1,2*N+1))
    P = [Lgenes]
    for _ in range(i):
        P.append(next_P(P[-1]))
    return P
```

II Loi de l'âge du premier ancêtre commun

1. **a)** Puisque chaque gène choisit équiprobablement son père dans un ensemble de $2N$ gènes, et que g a un unique père, la probabilité que g' choisisse le père de g est $1/2N$. Ainsi $p_2 = P(F) = 1/2N$.

b) Soit $k \geq 1$. Dire que l'évènement $[T_2 = k]$ est observé, c'est dire que les i-aïeux de g et g' sont distincts jusqu'au rang $k-1$ mais que que leurs k-aïeux sont égaux, ainsi $[T_2 = k] = \bigcap_{i=1}^{k-1} D_i \cap \overline{D_k}$.

c) On en déduit par la formule des probabilités composées que :

$$P(T_2 = k) = P(D_1) \times P(D_2|D_1) \times \cdots \times P\left(\overline{D_k} \middle| \bigcap_{j=1}^{k-1} D_j\right). \qquad (1)$$

Le raisonnement de la question **1a)** s'applique aux $k-1$ premiers facteurs de (1) en considérant pour le i-ème facteur les i-aïeux de g et g', qui sont bien distincts sous l'hypothèse $D_1 \cap \cdots \cap D_i$. Ainsi : $\forall k \in \mathbf{N}^\star \quad P(T_2 = k) = (1-p_2)^{k-1}p_2 = (1-1/2N)^{k-1}(1/2N)$.
D'après la question précédente, on lit que T_2 suit une loi géométrique de paramètre $p = 1/2N$.

d) Puisque par défintion : $D = \bigcap_{k\in\mathbf{N}^\star} \overline{T_2 = k}$, et que par ailleurs $P(T_2 \in \mathbf{N}^\star) = 1$ par définition de la loi géométrique, on déduit, puisque $\overline{D} = [T_2 \in \mathbf{N}^\star]$, que $P(D) = 0$: D est quasi-impossible.

e) D'après le cours $E(T) = 1/p = 2N$ et $V(T_2) = (1-p)/p^2 = 2N(2N-1)$.

f) D'après les données, le temps de coalescence moyen pour deux gènes donnés est de l'ordre de $200000/20 = 10000$ générations. Ce qui amène à poser $E(T_2) = 2N = 10000$, c'est-à-dire autant pour l'effectif de $\mathscr{P}_0$. L'ordre de grandeur pour le nombre de gènes reste donc correct. Notons néanmoins que le modèle de reproduction est haploïde ici.

2. Si $k = 2$, $p_{2,N} = 1-(1-1/2N) = 1/2N = p$, ce qui est conforme au résultat obtenu à la question **2.c)**.

III Un théorème limite sur les lois géométriques

1. Soit $t \geq 0$. Puisque $[Y \leq t] = \overline{[Y > t]}$, on déduit que : $P(Y \leq t) = 1 - e^{-\alpha t}$. Puisque deux variables aléatoires ayant la même fonction de répartition suivent la même loi, il suffit de vérifier que $P(Y \leq t) = 0$ pour tout réel t négatif. Or la relation donnée dans la question donne pour $t = 0$ que Y prend quasi-certainement ses valeurs dans $\mathbf{R}_+^\star$. Ainsi, la fonction de répartition de Y est nulle sur $\mathbf{R}_-$. Cela prouve bien que Y suit bien une loi exponentielle de paramètre α.

2. Pour tout entier $n \geq 1$, X_n suit une loi géométrique, donc l'évènement $Xn/n > 0$ est certain, c'est-à-dire : $\forall n \geq 1 \quad P(Y_n > 0) = 1$. On peut passer à la limite dans cette relation puisqu'elle est vraie pour tout entier n, ce qui donne $\lim\limits_{n\to+\infty} P(Y_n > 0) = 1 = e^{-\alpha 0}$. Pour $t = 0$, le théorème est donc vrai.

3. Soit $t > 0$. Comme $n \geq 1 > 0$, on a l'égalité d'évènements : $[Y_n \leq t] = [X_n \leq nt]$. Or, par définition de la partie entière, on a l'égalité d'évènements : $[X_n \leq nt] = [X_n \leq \lfloor nt \rfloor] \cup [\lfloor nt \rfloor < X_n \leq nt]$. Comme ces deux derniers évènements sont incompatibles, en passant aux probabilités, on obtient :
$P(Y_n \leq t) = P(X_n \leq \lfloor nt \rfloor) + P(X_n \in]\lfloor nt \rfloor, nt])$. Puisque la variable X_n ne prend que des valeurs entières et que l'intervalle $]\lfloor nt \rfloor, nt]$ ne contient aucun entier, $P(X_n \in]\lfloor nt \rfloor, nt]) = 0$. D'où le résultat.

4. Par définition de la partie entière, pour tout entier n, $nt - 1 < \lfloor nt \rfloor \leq nt$. En divisant ces inégalités par $nt > 0$ (t est non nul), et en faisant tendre n vers l'infini, le théorème des gendarmes nous donne que $\lfloor nt \rfloor \underset{n\to\infty}{\sim} nt$.

5. Comme suggéré par l'énoncé, étudions $u_n = \ln P(Y > nt)$. D'après **3.**, en passant aux évènements contraires, $u_n = \ln P(X_n > \lfloor nt \rfloor)$. Puisque X_n suit une loi géométrique de paramètre p_n, et que $\lfloor nt \rfloor$ est un entier positif, $u_n = \ln\left((1-p_n)^{\lfloor nt \rfloor}\right) = \lfloor nt \rfloor \ln(1-p_n)$. Comme $p_n \underset{n\to\infty}{=} o(1)$, par équivalent usuel, produit d'équivalents, et **4.**, on trouve que $u_n \underset{n\to\infty}{\sim} nt \times (-p_n)$. Enfin par hypothèse sur la suite (p_n), et encore par produit d'équivalents (puisque $\alpha t \neq 0$), on trouve $u_n \underset{n\to\infty}{\sim} nt \times (-\frac{\alpha}{n})$. Par composition de *limites*, on en déduit que $P(Y_n > t) = \exp(u_n)$ converge vers $e^{-\alpha t}$.

IV Approximation continue du temps de coalescence

1. On a vu que $P(F) = \dfrac{1}{2N} = o(1) \quad N \to \infty$: en limite simple, la probabilité que deux gènes donnés aient un ancêtre commun est nulle. Cette limite du modèle ne permet pas de construire une quelconque généalogie de gènes.

2. Il suffit de remarquer que pout tout entier $k \geq 1$ et tout réel h : $R_{k+1}(h) = R_k(h)(1-kh)$. Procédons ensuite par récurrence sur l'entier k. Il est clair qu'au rang $k = 2$, $R_2(h) = 1 - h$, donc *a fortiori* $R_2(h) = 1 - h + o(h) \quad h \to 0$.

Soit maintenant $k \geq 2$ un entier naturel fixé tel que $R_k(h) = 1 - k(k-1)h/2 + o(h) \quad h \to 0$. Montrons alors que $R_{k+1}(h) = 1 - k(k+1)h/2 + o(h) \quad h \to 0$. Par hypothèse de récurrence, et en développant le produit $R_{k+1}(h) = R_k(h)(1-kh)$, la précision $o(h)$ dans le calcul permet de ne retenir que les termes de degré au plus 1 en développant :

$$R_{k+1}(h) = \left(1 - \frac{k(k-1)}{2}h + o(h)\right)(1-kh) = 1 - \frac{k(k-1)}{2}h - kh + o(h) = 1 - \frac{k(k+1)}{2}h + o(h).$$

Ce qui achève la récurrence.

3. Soit $k \geq 2$. On remarque que pour tout entier $N \geq 1$, $q_{k,N} = R_k\left(\frac{1}{2N}\right)$. Si $N \to \infty$, on constate que $h \underset{N\to\infty}{=} 1/2N = o(1)$, donc d'après la question précédente : $p_{k,N} = 1 - q_{k,N} = \frac{k(k-1)}{2\times 2N} + o\left(\frac{1}{N}\right) = \frac{k(k-1)}{2\times 2N}(1+o(1)) \underset{N\to\infty}{\sim} \frac{k(k-1)}{2\times 2N}$.

4. Comme T_k suit une loi géométrique de paramètre $p_{k,N}$ et que $p_{k,N} \underset{N\to\infty}{\sim} \frac{\alpha}{2N}$ où $\lambda_k = \frac{k(k-1)}{2}$, on peut conclure d'après le **théorème** appliqué avec $\alpha = \lambda_k$ que la loi de $T_k/2N$ est une approximation de la $\mathscr{E}(\alpha)$, où $\lambda_k = \frac{k(k-1)}{2}$.

V Simulation du n-coalescent en temps continu

1. **a)** Soit F_Z la fonction de répartition de Z. Comme U prend quasi-certainement ses valeurs dans $]0,1]$, par propriétés du logarithme, Z prend quasi-certainement ses valeurs dans $\mathbf{R}_+$, donc F_Z est nulle sur $\mathbf{R}_-$.

Soit maintenant $t > 0$, et calculons $F_z(t)$. On a l'égalité d'évènements suivantes, parce que $\lambda > 0$, et par stricte croissance de l'exponentielle : $[Z \leq t] = U \geq e^{-\lambda t} = \overline{U < e^{-\lambda t}}$, d'où en passant aux probabilités : $F_Z(t) = P\left(\overline{U < e^{-\lambda t}}\right) = 1 - P(U < e^{-\lambda t})$. Comme U suit une loi uniforme sur $]0,1]$, et que $0 \leq e^{-\lambda t} \leq 1$, $F_Z(t) = 1 - P(U < e^{-\lambda t}) = 1 - e^{-\lambda t}$.

Finalement, Z a la même fonction de répartition qu'une variable de loi exponentielle de paramètre λ. D'après le cours, on en déduit que Z suit aussi une loi exponentielle de paramètre λ.

b) Il suffit de simuler la variable Z avec $\lambda = k(k-1)/2$:

```
1 def Y(k):
2     param = k*(k-1)/2
3     u = random()
4     return (-1/param)*log(u)
```

2. **a)** En utilisant les propriétés de range :

```
1 def Ltemps(n):
2     return [Y(k) for k in range(n,1,-1)]
```

b) On construit une liste en compréhension :

```
1 def paires(L):
2     return [ (i,j) for i in L for j in L if i<j]
```

c) **A-** Faux : `coalescent(7)[0]` est une liste (ligne 7 du script). Ses items sont des tuples `(i,j)` (ligne 7) d'entiers (lignes 6 et 5). Donc `coalescent(7)[0][0][0]` renvoie un entier.

B- Vrai : comme dit en **A**, la commande renvoie un tuple.

C- Faux : la commande renvoie un élément de la liste `durees`, c'est-à-dire une réalisation de Y_k, donc un flottant.

D- Faux : chaque passage dans la boucle while supprime un élément de la liste `Lgen`, donc la suite des longueurs décroît vers 1 : la condition de la ligne 4 ne sera plus vraie au bout de `n-1` passages puisqu'initialement, `Lgen` est de longueur `n`.

3. **a)** Sur l'arbre, on lit que le premier évènement de coalescence a lieu entre les gènes 1 et 4, puis entre 3 et 6, ainsi de suite : `lignees` est donc la liste `[(1,4),(3,6),(2,5),(0,3),(1,2),(0,1)]`.

b) Par définition du temps de coalescence, d_i =`durees[i-1]`, $1 \leq i \leq 6$.

4. **a)** La variable M_n mesure le temps au bout duquel, partant des n gènes, survient le premier évènement de mutation ou coalescence.

b) Soit $k \in \{2 \dots n\}$. Puisque les variables X_θ et Y_k sont des variables exponentielles, elles sont quasi-certainement positives. Soit alors t un réel positif. Par définition de minimum :

$$[M_k > t] = [Y_k > t] \cap [M_k > t].$$

Par mutuelle indépendance de M_k et X_θ on a donc :

$$P(M_k > t) = P(Y_k > t) \times P(X_\theta > t).$$

Or, d'après le cours, pour des lois exponentielles, on a :

$$P(M_k > t) = \exp(-k(k-1)t/2) \times \exp(-\theta/2) = \exp(-\lambda t)$$

où $2\lambda = k(k-1) + \theta$. Ainsi, d'après la question **1.** de la **partie II**, on déduit que M_k suit une loi exponentielle de paramètre $\lambda = (k(k-1) + \theta)/2$.

VI Hauteur moyenne du coalescent

1. **a)** La variable H_n est la somme des temps de coalescence jusqu'à ce qu'il ne reste plus qu'une seule lignée. Elle mesure donc l'âge du plus récent ancêtre commun d'un échantillon de n gènes donnés.

b) Les variables Y_k admettent toutes une espérance. Comme l'ensemble des variables aléatoires admettant une espérance est un espace vectoriel, H_n admet une espérance, et par linéarité de cette dernière, $E(H_n) = \sum_{k=2}^{n} E(Y_k) = 2\sum_{k=2}^{n} \frac{1}{k(k-1)}$ connaissant la valeur de l'espérance d'une loi exponentielle, d'après l'indispensable cours.

c) Par réduction au même dénominateur et telescopage, il vient que pour tout entier $n \geq 2$:

$$E(H_n) = 2\left(\sum_{k=2}^{n} \frac{1}{k-1} - \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right).$$

Ce qui prouve bien que la suite $(E(H_n))_{n\geq 2}$ converge vers 2.

d) Interprétation : quelque soit la taille de l'échantillon, en moyenne, l'âge du plus récent ancêtre commun est légèrement inférieur à 2, et ce, quelque soit la taille de la population initiale dès lors qu'elle est suffisamment grande.

e) Un calcul numérique donne $E(H_{10}) = 1,80$ et $E(H_{50}) = 1,96$.

f) Partant d'une collecte de $n = 10$ gènes, la multiplication de la taille de l'échantillon par 5 engendre en moyenne un gain relatif d'environ 8,9% ($\frac{1,96-1,80}{1,8}$) sur l'estimation de l'âge du plus récent ancêtre commun, ce qui est faible pour les efforts de collecte nécessaires. Autrement dit, l'essentiel de l'information en termes de hauteur d'arbre sur un 50-coalescent est contenu dans un 10-coalescent.

2. **a)** D'après le cours, les variables de lois exponentielles admettent des moments d'ordre 2, donc des variances. En utilisant le résultat rappelé dans l'énoncé, on en déduit que H_n admet une variance.

b) Soit $n \geq 2$. Toujours d'après le résultat rappelé dans le sujet et connaissant par le cours la valeurs de $V(Y_k)$, on obtient : $V(H_n) = \sum_{k=2}^{n} V(Y_k) = \sum_{k=2}^{n} E(Y_k)^2 = 4\sum_{k=2}^{n} \frac{1}{(k(k-1))^2}$. Ensuite on remarque que pour tout entier $k \geq 2$,

$$\frac{1}{(k(k-1))^2} \overset{\textbf{1.c)}}{=} \left(\frac{1}{k-1} - \frac{1}{k}\right)^2 = \frac{1}{k^2} - 2\frac{1}{k(k-1)} + \frac{1}{(k-1)^2}.$$

D'après les séries de référence et la question **1.c)**, le membre de droite est combinaison linéaire de termes généraux de séries convergentes. On en déduit que la suite $(V(H_n))$ est convergente, et par calcul direct en sommant la relation précédente de $k = 2$ à $k = +\infty$:

$$S = \sum_{k=2}^{+\infty} \frac{1}{(k(k-1))^2} = \left(\frac{\pi^2}{6} - 1\right) - 2 + \frac{\pi^2}{6} = \frac{\pi^2}{3} - 3 = \frac{1}{3}\left(\pi^2 - 9\right),$$

ce qui donne bien le résultat puisque la limite de $(V(H_n))$ est $4S$.

c) Une application numérique donne que $V_\infty \simeq 1,159$. La première ligne du tableau indique que la série converge rapidement vers V_∞ puisque $V(H_{10}) \simeq V_\infty$. En outre, la deuxième ligne indique que le terme $V(Y_2)$ contribue au moins aux deux tiers dans la variance de la hauteur de l'arbre de coalescence : la première fourche rencontrée en descendant le coalescent est donc celle de hauteur la plus variable d'une simulation à une autre pour un nombre n de gènes fixé.

VII Longueur moyenne du n-coalescent

1. **a)** D'après la définition de branches, L_n mesure la longueur totale des branches constituant le n-coalescent puisqu'avant le k-ème instant de coalescence, il y a $n - k + 1$ branches, donc k branches de longueur Y_k, pour $k \in \{2 \dots n\}$.

b) Encore par linéarité de l'espérance et comme justifié en **VI. 1b)**, L_n admet une espérance et :

$$E(L_n) = \sum_{k=2}^{n} kE(Y_k) = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\frac{1}{k-1} = 2\sum_{k=1}^{n-1}\frac{1}{k}.$$

c) Il est connu que la série harmonique diverge vers $+\infty$. Ici, on a la suite des sommes partielles de cette dernière au facteur $2 > 0$ près. On en déduit que $E(L_n)$ tend vers $+\infty$.

d) Soit $k \geq 2$ et $t \in [k-1, k]$. Des inégalités $t \leq k$ et $t \geq k-1$, la dernière se réécrivant $t+1 \geq k$, on tire par décroissance de la fonction inverse sur $\mathbf{R}_+^\star$ que :

$$\forall t \in [k-1, k] \quad \frac{1}{t+1} \leq \frac{1}{k} \leq \frac{1}{t}.$$

On conclut par croissance de l'intégrale en intégrant ces inégalités sur le segment $[k-1, k]$

e) Fixons maintenant $n \geq 2$ et sommons les inégalités de la question précédente de $k = 2$ à $k = n-1$. Il vient par la relation de Chasles :

$$\int_1^{n-1} \frac{\mathrm{d}t}{t+1} \leq \sum_{k=2}^{n-1}\frac{1}{k} \leq \int_1^{n-1}\frac{\mathrm{d}t}{t}$$

Après multiplication par 2 membre à membre (ce qui ne change par le sens des inégalités), et primitivation à vue notamment, on obtient :

$$\forall n \geq 2 \quad 2\ln n + 2(1-\ln 2) \leq E(L_n) \leq 2\ln(n-1) + 2.$$

Réécrivant dans le membre de droite que $\ln(n-1) = \ln n + \ln\left(1 - \frac{1}{n}\right)$, on obtient après division par $\ln n > 0$ que :

$$\forall n \geq 2 \quad 1 + \frac{1-\ln 2}{\ln n} \leq \frac{E(L_n)}{2\ln n} \leq 1 + \frac{1}{\ln n} + \frac{\ln\left(1-\frac{1}{n}\right)}{\ln n}.$$

Par quotient de limites et le théorème des gendarmes, on obtient bien l'équivalent annoncé.

2. **a)** La hauteur moyenne du n-coalescent, ou, ce qui revient au même, l'espérance de l'âge du plus récent ancêtre commun est, d'après **VI 1c)**, égale à $E(H_n) = 2(1-1/n)$. Si on accepte cet âge moyen dans le cas particulier extrême de n lignées restant distinctes jusqu'à leur unique ancêtre commun de génération 0, la longueur totale de l'arbre généalogique $L_\bullet$ serait, puisqu'il y a exactement n branches de longueur $E(H_n)$, égale à $L_\bullet = nE(H_n) = n(2-1/n) = 2n-1$.

b) D'après les croissances comparées en $+\infty$, $\tau_n \underset{n\to\infty}{=} o(1)$.

c) Si les lignées des n gènes restaient effectivement distinctes jusqu'au premier ancêtre commun, le ratio τ_n vaudrait 1, et cela signifie que ces gènes n'ont aucune histoire commune. Le fait que $\tau_{10} \simeq 31\%$ indique qu'environ 69% de la généalogie de ces 10 gènes leur est commune.