Documents et calculatrices non autorisés. La clarté et la présentation de la copie seront prises en compte dans la notation.

Les méthodes classiques en génétique des populations sont essentiellement prospectives : partant de données collectées, on développe des modèles permettant de prédire par exemple l'évolution du polymorphisme génétique.

En 1982, les travaux de J. Kingman notamment, ont permis de développer une approche rétrospective : partant des données collectées, la théorie de la *coalescence* propose de reconstruire l'histoire d'un échantillon de n gènes donné par un arbre généalogique remontant au premier ancêtre commun des gènes de cet échantillon, appelé n-coalescent.

L'objet de cette épreuve est l'étude du modèle du n-coalescent, depuis sa construction jusqu'à l'étude de ses propriétés basiques.

- Dans la **partie I**, on introduit le modèle de Wright-Fisher, qui est un modèle simple en temps discret de transmission de gènes et sa simulation informatique.
- La **partie II** introduit la terminologie utilisée dans toute la suite et a pour but de calculer la loi discrète du temps de coalescence d'un échantillon de gènes actuels donné.
- La **partie III** met en place un théorème d'approximation des lois géométriques utilisé dans la **partie IV**.
- La **partie IV** introduit une approximation continue du temps de coalescence étudié **partie II**.
- La **partie V** traite de la simulation du modèle de coalescence.
- Les **parties VI** et **VII** étudient des variables aléatoires liées à la géométrie des arbres généalogiques produits par le modèle de coalescence semblables à ceux de la figure 1b.

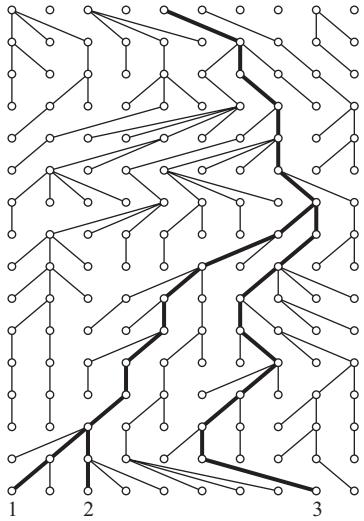

Fig. 1a – Processus de Wright-Fisher et lignées ancestrales en temps discret de trois gènes.

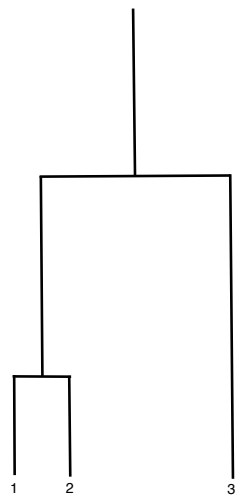

Fig. 1b – Limite d'échelle du processus : on obtient un 3-coalescent en temps continu.

I Le modèle de Wright-Fisher (1930-1931)

Le *modèle de Wright-Fisher* est un modèle simple de reproduction mis au point afin de décrire la dynamique de transmission de gènes au sein d'une population donnée, de génération en génération. Dans la suite, le terme *population* désignera une population de gènes.

La variable temporelle est notée i et est supposée prendre des valeurs entières positives.

Soit $N > 0$ un entier fixé.

A – Hypothèses du modèle

(H1) *Population finie :* À la date initiale $i = 0$, la population initiale, notée $\mathscr{P}_0$, est une *liste* constituée de $2N$ gènes numérotés $1, \dots, 2N$: $\mathscr{P}_0 = (1, 2, \dots, 2N-1, 2N)$.

(H2) *Population à effectif constant :* À chaque instant $i > 0$, la nouvelle population est une *liste* $\mathscr{P}_i$ de $2N$ gènes :

$$\mathscr{P}_i = \big(g_i(1), g_i(2), \dots, g_i(2N-1), g_i(2N)\big).$$

Ainsi, le gène en position j de la liste $\mathscr{P}_i$ est noté $g_i(j)$, ou simplement $g(j)$ afin de ne pas alourdir les notations.

(H3) *Absence d'avantage adaptatif, et pas de structuration de la population :* pour tout entier $i \geq 0$, la population $\mathscr{P}_{i+1}$ est obtenue en tirant successivement et au hasard avec remise $2N$ gènes de la population $\mathscr{P}_i$.

(H4) *Indépendance des reproductions :* Pour tout entier $i \geq 0$, les tirages dans $\mathscr{P}_i$ sont supposés mutuellement indépendants.

La suite des listes $(\mathscr{P}_i)$ obtenue à partir de $\mathscr{P}_0$, appelée *processus de Wright-Fisher*, est représentée graphiquement de la manière suivante sur la figure 2.

B– Illustration

Pour fixer les idées, on a pris ici $N = 2$.

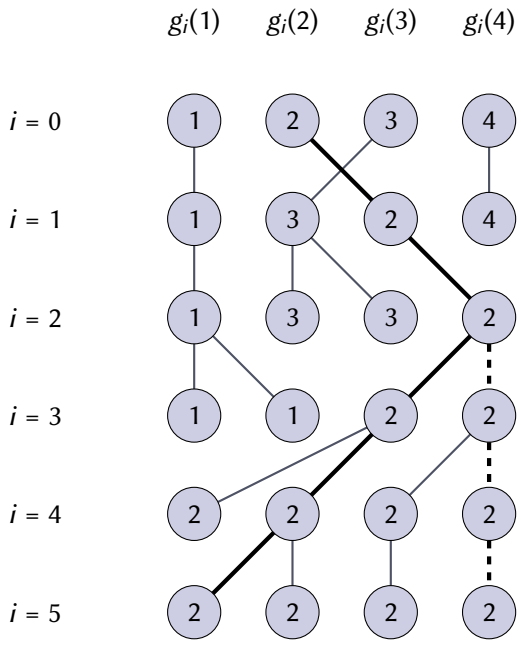

Fig. 2 – *Un exemple de processus de Wright-Fisher*

- La ligne i indique la composition de la liste $\mathscr{P}_i$. Ainsi, on lit par exemple : $\mathscr{P}_1 = (1, 3, 2, 4)$, $\mathscr{P}_2 = (1, 3, 3, 2)$.
- La cellule à l'intersection de la ligne i et de la colonne j est donc $g_i(j)$, et le numéro reporté dans la cellule indique la copie du gène de génération 0 qui est transmis.
- Entre deux générations successives i et $i+1$, la présence d'un segment reliant le gène $g_{i+1}(j)$ à un gène g de la génération précédente i exprime que le résultat du j-ème tirage dans $\mathscr{P}_i$ est g.

Exemple : Sur la figure 2, on lit que $\mathscr{P}_3 = (1, 1, 2, 2)$ et $\mathscr{P}_4 = (2, 2, 2, 2)$, tandis que les segments liant les individus de $\mathscr{P}_4$ à ceux de $\mathscr{P}_3$ précisent que :

- $g_4(1) = g_3(3) = 2$
- $g_4(2) = g_3(3) = 2$
- $g_4(3) = g_3(4) = 2$
- $g_4(4) = g_3(4) = 2$

On souhaite simuler informatiquement sous Python le processus de Wright-Fisher.

```
from random import choice
```

La fonction `choice` du module `random` prend en entrée une liste et renvoie en sortie un item de cette dernière choisi équiprobablement, sans toutefois la modifier.

1. Donner le script d'une fonction `next_P(L)` prenant en entrée une liste `L` et renvoyant en sortie une liste de même longueur que `L` obtenue par tirages successifs avec remise d'éléments de `L`.
2. Proposer alors en Python à l'aide de la fonction `next_P`, le script d'une fonction `WF(N,i)` prenant en entrée deux entiers `N` et `i`, et renvoyant en sortie la liste `P=[P[0],P[1],...,P[i]]` des populations $\mathscr{P}_0, \dots, \mathscr{P}_i$ obtenues par le processus de Wright-Fisher, la liste `P[0]` étant égale à celle des entiers de 1 à $2N$.

II Loi de l'âge du premier ancêtre commun de lignées données

Le processus de Wright-Fisher modélise des relations de généalogie entre les individus des populations $\mathscr{P}_i$ en adoptant la terminologie suivante :

Déf. 1. Père d'un gène :

- Le résultat du j-ème tirage dans la liste $\mathscr{P}_i$ s'appelle le *père* du gène $g_{i+1}(j)$.
- *Exemple :* sur la figure 2, le père du gène $g_4(1)$ est $g_3(3)$.

Déf. 2. ℓ-aïeul d'un gène :

- Le 0-*aïeul* d'un gène g est le gène g lui-même.
- Le père de g est appelé le 1-*aïeul* g.
- Le père du père du gène g est appelé 2-*aïeul de* g.
- De façon générale, si $\ell \geq 1$ est un entier, le ℓ-*aïeul de* g est le père de son $(\ell-1)$-aïeul.
- *Exemple :* Sur la figure 2, le gène $g_5(1)$ de génération 5 a pour père $g_4(2)$, pour 2-aïeul le gène $g_3(3)$, et pour 3-aïeul le gène $g_2(4)$, *etc.*

Def. 3. Lignée ancestrale d'un gène de génération i : Soit $i \geq 0$, et g un gène de $\mathscr{P}_i$.

- La suite L de longueur $i+1$, $L = (g, a_1, a_2, \dots, a_i)$, où a_ℓ est le ℓ-aïeul de g $(1 \leq \ell \leq i)$ s'appelle la *lignée ancestrale* ou *lignée du gène* g_i.
- *Exemple :* sur la figure 2, le gène $g_5(1)$ de génération 5 a pour lignée ancestrale :

$$L = \big(g_5(1), g_4(2), g_3(3), g_2(4), g_1(3), g_0(2)\big) \quad \text{(indiquée en traits gras sur la figure 2).}$$

Def. 4. Âge du premier ancêtre commun de deux gènes :

- si g et g' sont deux gènes distincts de génération i donnée, le *premier ancêtre commun de* g *et* g'*, si il existe, est le* ℓ*-aïeul de* g *et* g' *pour lequel* l'entier ℓ est *minimal.*
- L'entier ℓ s'appelle *âge du premier ou du plus récent ancêtre commun de* g *et* g'.
- *Exemple :* Sur la figure 2
  - les gènes $g_5(1)$ et $g_5(4)$ ont pour lignées ancestrales respectives
    $L = \big(g(1), g(2), g(3), \mathbf{g(4)}, g(3), g(2)\big)$ (en traits gras sur la figure)
    $L' = \big(g(4), g(4), g(4), \mathbf{g(4)}, g(3), g(2)\big)$ (en traits gras pointillés sur la figure)
    Le premier ℓ-aïeul commun à g et g' est leur 3-aïeul. L'âge de leur premier ancêtre commun est 3.

— En revanche, les gènes $g_3(1)$ et $g_3(4)$ de génération 3 n'ont pas d'ancêtre commun.

Def. 5. Coalescence des lignées de deux gènes.

- Si g et g' sont deux gènes distincts de génération i donnée, leurs lignées ancestrales sont dites *coalescentes* si g et g' possèdent un ancêtre commun. Le *temps de coalescence de leurs lignées* est l'âge de leur premier ancêtre commun.
- *Exemple :* les lignées L et L' de l'exemple de la définition 4 sont coalescentes. Le temps de coalescence de ces lignées est 3.

Dans cette partie , on fixe un instant $i_0 > 0$ supposé aussi grand que nécessaire, et on raisonne en partant de la population $\mathscr{P}_{i_0}$ que l'on peut penser comme la population actuelle.

1. On fixe deux gènes distincts g et g' de la population $\mathscr{P}_{i_0}$.

a) Calculer la probabilité p_2 de l'évènement F : «Le gène g' a le même père que le gène g.»

On note T_2 la variable aléatoire égale au temps de coalescence des lignées de g et g', si celui existe, et égale à 0 sinon. On admet que T_2 est bien une variable aléatoire réelle discrète.
Pour tout entier i tel que $1 \leq i \leq i_0$, on note D_i l'évènement : «les gènes g et g' ont des i-aïeux distincts».

b) Soit $k \geq 1$ un entier. Exprimer l'évènement $[T_2 = k]$ à l'aide des évènements D_i.

c) En déduire que T_2 suit une loi géométrique de paramètre p à préciser.

d) Comment qualifier mathématiquement l'évènement D : «les gènes g et g' n'ont pas d'ancêtre commun» ?

e) Rappeler les valeurs des espérance $E(T_2)$ et variance $V(T_2)$.

f) Pour le génome humain qui contient environ 2×10^4 gènes, la plupart des échantillons de gènes étudiés possèdent un ancêtre commun de moins de 200 000 ans. En prenant la durée d'une génération égale à 20 ans, commenter la valeur de $E(T_2)$ dans ce modèle.

Afin de généraliser le résultat de la question **1.**, fixons un entier $k \geq 2$, ainsi qu'une sous-liste $\mathscr{G}$ de k gènes deux à deux distincts de $\mathscr{P}_{i_0}$. En notant T_k la variable aléatoire égale au premier temps de coalescence (s'il existe) d'au moins 2 lignées d'éléments de $\mathscr{G}$, et égale à 0 sinon, on peut montrer que T_k suit une loi géométrique de paramètre $p_{k,N} = 1 - q_{k,N}$, où

$$q_{k,N} = \frac{2N \times (2N-1) \times \cdots \times (2N-k+1)}{(2N)^k} = \prod_{\ell=1}^{k-1} \left(1 - \frac{\ell}{2N}\right).$$

2. Vérifier que le résultat obtenu en question **1. c)** pour le paramètre p est correct en le comparant avec la valeur donnée ci-dessus de $p_{2,N}$.

III Un théorème limite sur les lois géométriques

Cette partie est indépendante des autres.

Dans cette partie, on souhaite prouver le théorème suivant :

Théorème. *Soit $\alpha > 0$, $(p_n)_{n\geq 1}$ une suite de réels de $]0,1[$ telle que :*

$$p_n \underset{n\to\infty}{\sim} \frac{\alpha}{n},$$

et soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires définies sur un espace probabilisé $(\Omega, \mathscr{T}, P)$ de lois géométriques sur $\mathbf{N}^\star$ de paramètre p_n.
Soit Y_n la variable aléatoire définie par $Y_n = \dfrac{X_n}{n}$. Alors :

$$\forall t \geq 0 \quad \lim_{n\to+\infty} P(Y_n > t) = e^{-\alpha t} \qquad (\star)$$

Autrement dit, pour n assez grand, la loi de Y_n est approximativement une loi exponentielle de paramètre α.

1. Justifier en quoi le fait qu'une variable aléatoire Y vérifiant : $\forall t \geq 0 \quad P(Y > t) = e^{-\alpha t}$ suit une loi exponentielle de paramètre α.
2. Démontrer l'assertion $(\star)$ dans le cas particulier où $t = 0$.
3. Soit $n \in \mathbf{N}^\star$ et $t > 0$. On rappelle que la notation $\lfloor x \rfloor$ désigne la partie entière du réel x, et qu'elle est définie comme l'unique entier tel que $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$.
   Montrer que : $P(Y_n \leq t) = P(X_n \leq \lfloor nt \rfloor)$.
4. Établir un encadrement de $\lfloor x \rfloor$ par des expressions affines de x, et valable pour tout réel x. Calculer ensuite un équivalent simple de $\lfloor nt \rfloor$ quand $n \to +\infty$.
5. En déduire que $\lim_{n\to+\infty} P(Y_n > t) = e^{-\alpha t}$ (*on pourra étudier* $u_n = \ln P(Y_n > t)$).

IV Approximation continue du temps de coalescence d'un groupe de k gènes donné

Cette partie utilise des résultats des parties II et III

Dans cette partie, on souhaite adapter le modèle de Wright-Fisher à des populations d'effectif grand. Pour cela, on étudie les quantités introduites en **partie II** quand $N \to +\infty$.

1. Quelle est la limite de la probabilité $P(F)$ calculée en **Partie II 1a)** quand $N \to +\infty$? Que dire alors de la pertinence du modèle de Wright-Fisher si $N \to +\infty$ pour analyser la coalescence de lignées ?

Afin d'obtenir une limite non triviale du modèle, on prend dans ce qui suit une limite d'échelle quand $N \to +\infty$, c'est-à-dire en liant les variables T_k définies dans la **partie II** avec la variable N.

2. Pour tout entier $k \geq 2$, on définit le polynôme R_k par :

$$\forall h \in \mathbf{R} \quad R_k(h) = \prod_{\ell=1}^{k-1}(1-\ell h) = (1-h)(1-2h) \times \cdots \times \big(1-(k-1)h\big).$$

À partir d'une relation simple liant R_{k+1} à R_k, montrer par récurrence que pour tout entier $k \geq 2$, le développement limité à l'ordre 1 en 0 de R_k est :

$$R_k(h) = 1 - \frac{k(k-1)}{2}h + o(h) \quad h \to 0.$$

3. Exprimer le paramètre $q_{k,N}$ défini en fin de **question 1.** de la **partie II** en fonction du polynôme R_k et déduire de ce qui précède pour tout entier $k \geq 2$ un équivalent quand $N \to +\infty$ de $p_{k,N}$.

4. En déduire à l'aide du **théorème** énoncé en **partie III** que si $N \to +\infty$, la loi de la variable aléatoire $Y_k = \frac{T_k}{2N}$ est approximativement une loi exponentielle de paramètre λ_k à préciser.

V Simulation du n-coalescent en temps continu

Cette partie est, dans une assez large mesure, indépendante des autres.

1. Soit U une variable aléatoire de loi uniforme sur $]0, 1]$ et λ un réel strictement positif.

a) Montrer que la variable $Z = -\frac{1}{\lambda}\ln(U)$ suit une loi exponentielle de paramètre λ.

b) Proposer en `Python`, à partir du script ci-dessous, une fonction `Y(k)` prenant en entrée un entier k et simulant une variable aléatoire de loi exponentielle de paramètre $\frac{k(k-1)}{2}$.

```
from numpy import log
from random import random()
```

Soit $n \geq 2$ un entier. On souhaite construire informatiquement l'arbre généalogique des lignées ancestrales d'une sous-liste $\mathscr{G}$ de n gènes donnés en simulant les évènements de coalescence des lignées jusqu'au plus récent ancêtre commun de ces n gènes. Cet arbre est appelé **n-coalescent en temps continu** : il est le résultat de la limite d'échelle obtenue en **partie IV 4.**, illustrée en figure 1.

Notons Y_n la variable aléatoire égale à l'instant du premier évènement de coalescence d'au moins 2 lignées parmi les n lignées de gènes de $\mathscr{G}$. De la même façon, on note Y_k $(k = n-1, n-2, \dots, 2)$ la variable aléatoire égale au premier instant de coalescence d'au moins 2 lignées parmi k lignées des gènes de $\mathscr{G}$ pour lesquelles il n'a pas encore été observé d'instant de coalescence en remontant l'arbre.

- On suppose que les variables $Y_2, \dots Y_n$ sont mutuellement indépendantes et que pour tout entier k dans $\{2, \dots, n\}$, la variable Y_k suit une loi exponentielle de paramètre $\frac{k(k-1)}{2}$.
- Les résultats de la **partie IV** indiquent que dans le modèle en temps continu, si à un instant donné, dans un ensemble de k lignées $(2 \leq k \leq n)$, il y a coalescence, ce ne peut être *qu'entre deux lignées à la fois* : dit autrement, la coalescence simultanée de trois lignées ou plus est quasi-impossible.

L'algorithme proposé est alors le suivant :

Algorithme : simulation d'un n-coalescent en temps continu.

1. Commencer avec $k = n$ et $\mathscr{G}$ l'ensemble de k gènes numérotés $0, 1, \dots, (n-1)$.
2. Simuler le temps d'attente Y_k de coalescence de 2 lignées parmi les k lignées des gènes de $\mathscr{G}$.
3. Choisir équiprobablement un couple de gènes (i, j) $(i < j)$ parmi les couples de gènes possibles parmi les k gènes de $\mathscr{G}$.
4. Fusionner les gènes i et j en le gène i, et décrémenter k.
5. Si $k > 1$ reprendre depuis 2., sinon arrêter.

2. **a)** Écrire une fonction `Ltemps(n)` qui prend en entrée un entier `n` et qui renvoie en sortie la liste `Lt` de réalisations des variables $Y_n, Y_{n-1}, \dots, Y_2$ dans cet ordre.

b) Écrire une fonction `paires(L)` prenant en entrée une liste `L` d'entiers, et renvoyant en sortie la liste des couples `(i,j)` d'éléments de `L` tels que `i` $<$ `j`.

La commande `L.remove(item)` modifie la liste `L` en lui retirant l'élément `item`. L'implémentation de l'algorithme conduit alors au script de la fonction suivante simulant un n-coalescent.

```
1  def coalescent(n):
2      Lgen = list(range(n))
3      lignees= []
4      while(len(Lgen))>1:
5          Lpaires = paires(Lgen)
6          (i,j) = choice(Lpaires)
7          lignees.append((i,j))
8          Lgen.remove(j)
9      durees = Ltemps(n)
10     return lignees,durees
```

c) les assertions suivantes sont-elles vraies ou fausses ? Justifier par une réponse précise.

A- La commande `coalescent(7)[0][0][0]` renvoie un message d'erreur.

B- La commande `coalescent(7)[0][0]` renvoie un tuple.

C- La commande `coalescent(7)[1][0]` renvoie un objet de type entier.

D- La boucle `while` peut ne jamais s'arrêter.

3. La figure 3 donne une représentation graphique d'un 7-coalescent réalisé à partir des sorties renvoyées par une exécution de la commande `(###)` ci-desous :

```
console
In [1]: lignees,durees = coalescent(7)          ###
```

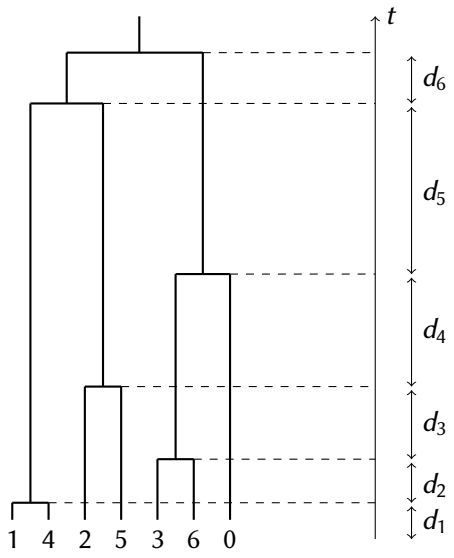

Fig. 3 – *Représentation d'un 7-coalescent. Les gènes sont numérotés* $0, 1 \dots, 6$.

a) Donner la valeur de la sortie `lignees` renvoyée par la commande `(###)` et ayant conduit à cet arbre généalogique.

b) Indiquer sur la copie le lien entre les nombres d_i de la figure 3 et les éléments de la sortie `durees` générée par la commande `(###)`.

4. Il est connu que dans la transmission des gènes, des mutations apparaissent aléatoirement. Pour tenir compte de ce phénomène dans le modèle, on introduit un paramètre $\theta > 0$ appelé *taux de mutation* du modèle. Notons alors X_θ une variable aléatoire de loi exponentielle de paramètre $\frac{\theta}{2}$. Cette variable simule le temps d'attente de la survenue d'une mutation en remontant le n-coalescent partant d'une lignée donnée. On définit aussi pour tout entier $k \in \{2, \dots, n\}$, une variable aléatoire M_k par $M_k = \min(Y_k, X_\theta)$. On suppose les variables $X_\theta, Y_2, \dots, Y_n$ mutuellement indépendantes.

a) Que représente la variable M_n dans le modèle ?

b) Soit $k \in \{2 \dots n\}$. Montrer que la variable M_k est une variable aléatoire à densité et en donner une densité.

VI Hauteur moyenne du n-coalescent

Cette partie est, dans une assez large mesure, indépendante des autres.

Soit $n \geq 2$ un entier et $Y_2, \dots, Y_n$ des variables aléatoires mutuellement indépendantes. On suppose que pour tout entier k dans $\{2, \dots, n\}$, Y_k suit une loi exponentielle de paramètre $\lambda_k = \dfrac{k(k-1)}{2}$. On définit une variable aléatoire H_n par : $H_n = \displaystyle\sum_{k=2}^{n} Y_k$, et appelée *hauteur du n-coalescent*.

1. **a)** Interpréter la variable H_n en termes d'ancêtre commun.

b) Justifier que pour tout entier $n \geq 2$, la variable H_n admet une espérance et donner son expression sous forme d'une somme qu'on ne cherchera pas à calculer.

c) Simplifier l'expression $\dfrac{1}{k-1} - \dfrac{1}{k}$ pour tout entier $k \geq 2$, et déduire que lorsque n tend vers l'infini, $E(H_n)$ converge vers 2.

d) Interpréter ce résultat par rapport à la donnée initiale d'une liste de $2N$ gènes étudiés.

e) Donner les valeurs numériques de $E(H_{10})$ et de $E(H_{50})$.

f) En pratique, pour estimer l'âge du plus récent ancêtre commun d'une population de gènes donnée, on collecte un échantillon de n individus porteurs de ce gène en vue d'une analyse statistique. Selon ce modèle, le fait de prendre un échantillon 5 fois plus gros apporte-t-il un gain informatif important ?

On rappelle que pour des variables aléatoires mutuellement indépendantes $X_1 \dots, X_n$ admettant des variances, la variable $S_n = X_1 + \cdots + X_n$ admet aussi une variance, et $V(S_n) = V(X_1) + \cdots + V(X_n)$.

2. **a)** Justifier que pour tout entier $n \geq 2$, H_n admet une variance.

b) En admettant que la somme de la série $\displaystyle\sum_{k\geq 1} \frac{1}{k^2}$ est $\dfrac{\pi^2}{6}$, montrer que la suite $\left(V(H_n)\right)_{n\geq 2}$ est convergente et que sa limite est : $V_\infty = \dfrac{4}{3}\left(\pi^2 - 9\right)$.

Indication : on pourra commencer par développer $\left(\dfrac{1}{k-1} - \dfrac{1}{k}\right)^2$ *et utiliser la question* ***1.c)***.

c) Le tableau ci-dessous donne en pourcentage la contribution du terme $V(Y_2)$ dans la valeur de $V(H_n)$ pour différentes valeurs de n :

n	2	3	4	5	6	10	15	20
$V(H_n)$	1,000	1,111	1,139	1,149	1,153	1,158	1,159	1,159
Contribution de $V(Y_2)$ (%)	100	80,0	73,4	70,2	68,3	64,9	63,5	62,7

Comment interpréter ces données en termes de hauteur des n-coalescents ?

VII Longueur totale moyenne du n-coalescent

Cette partie est, dans une assez large mesure, indépendante des autres.

Les branches du n-coalescent sont par définition les segments allant d'un des n gènes au premier instant de coalescence de sa lignée, ou tout segment compris entre deux instants consécutifs de coalescence d'une lignée (voir figure 3).

Soit $n \geq 2$ un entier et $Y_2, \dots, Y_n$ des variables aléatoires mutuellement indépendantes. On suppose que pour tout entier k dans $\{2, \dots, n\}$, Y_k suit la loi exponentielle de paramètre $\dfrac{k(k-1)}{2}$. On définit une variable aléatoire L_n par :

$$L_n = \sum_{k=2}^{n} kY_k,$$

appelée longueur totale du n-coalescent.

1. **a)** Que représente géométriquement la variable L_n sur l'arbre généalogique d'un n-coalescent semblable à celui de la figure 3 en termes des branches du coalescent ?

b) Montrer que pour tout entier $n \geq 2$, L_n admet une espérance et donner son expression sous forme d'une somme qu'on ne cherchera pas à calculer.

c) Donner la limite de $E(L_n)$ quand $n \to +\infty$.

d) Montrer que pour tout entier $k \geq 2$:

$$\int_{k-1}^{k} \frac{\mathrm{d}t}{t+1} \leq \frac{1}{k} \leq \int_{k-1}^{k} \frac{\mathrm{d}t}{t}.$$

e) En déduire que $E(L_n) \underset{n\to+\infty}{\sim} 2\ln n$.

Dans le cas extrême d'un échantillon de n gènes partant d'un ancêtre commun, et ayant ensuite des lignées distinctes, l'ordre de grandeur de la longueur moyenne de l'arbre serait de $2n-1$ si l'on accepte l'âge du plus récent ancêtre commun estimé par le modèle du n-coalescent.

2. **a)** Justifier cet ordre de grandeur en vous appuyant sur le résultat de la question **1c)** de la **partie VI**.

b) Calculer la limite de la suite de terme général $\tau_n = \dfrac{E(L_n)}{2n-1}$.

c) On dit que le ratio τ_n mesure l'histoire commune d'un échantillon de n gènes. Que signifie cette assertion dans l'exemple d'un échantillon de $n = 10$ gènes ? *On donne* $\tau_{10} \simeq 31\%$.

— Fin du sujet —