Lycée Pierre-Gilles de Gennes 2024-2025

BCPST2 – Mathématiques

Modélisation mathématique et informatique- 3h

L'usage d'une calculatrice est autorisée pour cette épreuve (mais pas nécessaire).

Chaque candidat est responsable de la vérification de son sujet d'épreuve : pagination et impression de chaque page. Ce contrôle doit être fait en début d'épreuve. En cas de doute, le candidat doit alerter au plus tôt le surveillant qui vérifiera et, éventuellement, remplacera le sujet.

Ce sujet comporte 7 pages numérotées de 1 à 7.
Une annexe des commandes Python utiles se trouve en page 7.

Si, au cours de l'épreuve, un candidat repère ce qui lui semble être une erreur d'énoncé, il le signale sur sa copie et poursuit sa composition en expliquant les raisons des initiatives qu'il a été amené à prendre.

Dans ce problème on étudie un modèle pour décrire la transmission d'allèles et l'évolution des fréquences alléliques au cours du temps dans une population diploïde.
Pour simplifier, on se place dans le cadre suivant :

- on étudie un seul gène sur un locus précis et qui se présente sous deux allèles distincts `A` et `a`,
- la taille de la population est supposée constante au cours du temps égale à $N \in \mathbb{N}^*$; il y a donc $2N$ locus dans la population totale,
- les générations ne se chevauchent pas : à chaque instant k, la k-ième génération meurt et donne naissance aux N individus de la génération suivante,
- la reproduction à l'instant k ne dépend pas des reproductions précédentes.

Chaque individu est de l'un des trois types suivants :

`AA` (type 1) ; `aa` (type 2) ; `Aa` (type 3).

1 Modèle de Wright-Fisher

Dans ce modèle, on néglige les phénomènes de mutation et de sélection.
On s'intéresse aux nombres d'allèles `A` présents sur le locus étudié et pour tout $n \in \mathbb{N}$, on note X_n la variable aléatoire à valeurs dans $[\![0, 2N]\!]$ donnant le nombre d'allèles de type `A` à la génération n dans une population de taille N.
On considère que pour tout entier n :

$$\forall (i,j) \in [\![0,2N]\!]^2, \quad \mathbb{P}(X_{n+1} = j \mid X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

où $\mathbb{P}(A|B)$ désigne la probabilité conditionnelle de A sachant B.

1.1 Étude d'un cas particulier

On suppose dans cette partie que $N = 1$ et on note pour tout $n \in \mathbb{N}$:

$$V_n = \begin{pmatrix} \mathbb{P}(X_n = 0) \\ \mathbb{P}(X_n = 1) \\ \mathbb{P}(X_n = 2) \end{pmatrix}.$$

1. Déterminer une matrice M telle que pour tout entier n on ait : $V_{n+1} = MV_n$.
Indications : une récurrence n'est pas nécessaire ; on pourra utiliser la formule des probabilités totales.

2. **(a)** Déterminer les valeurs propres réelles de M et une base de chaque sous-espace propre.

(b) La matrice M est-elle diagonalisable ? Si oui, déterminer une matrice inversible P et une matrice diagonale D telle que $M = PDP^{-1}$.

3. **Deux méthodes pour le calcul de M^n. Une seule des deux sous-questions peut être traitée.**

(a) Calculer M^n pour tout entier n à l'aide de la question **2.(b)**.

(b) Montrer que par récurrence que pour tout $n \in \mathbb{N}$: $M^n = \begin{pmatrix} 1 & \frac{2^n-1}{2^{n+1}} & 0 \\ 0 & \frac{1}{2^n} & 0 \\ 0 & \frac{2^n-1}{2^{n+1}} & 1 \end{pmatrix}$.

4. En déduire que :

(a) pour tout entier n, on a $E(X_n) = E(X_0)$;

(b) $\lim\limits_{n\to+\infty} \mathbb{P}(X_n \in \{0, 2\}) = 1$.

1.2 Cas général.

On suppose désormais que $N \geq 1$.

5. **Python**. Dans cette question et uniquement dans cette question, on suppose que X_0 suit une loi certaine de paramètre `x0`$\in [\![0, 2N]\!]$.

(a) On considère la fonction suivante :

```
1 def X(N,n,x0):
2     X = x0
3     for k in range(n):
4         X =rd.binomial(2*N,X/(2*N))
5     return X
```

Expliquer soigneusement ce que renvoie cette fonction.

(b) On suppose que `x0`, `n` et `N` ont déjà été déclarés dans le script. Compléter les lignes suivantes pour simuler 1000 fois la variable $X_{\texttt{n}}$ et afficher une valeur approchée de $\mathbb{P}(X_{\texttt{n}} \in \{0, 2\texttt{N}\})$ dans la variable `s` :

```
s = 0
for k in range(____):
        Xn = X(N,n,x0)
        if _______ :
                _______
print(s)
```

(c) On a tracé (voir la figure 1) pour les valeurs de N comprises entre 1 et 50, les valeurs approchées de $\mathbb{P}(X_{1000} \in \{0, 2N\})$ en fonction de N obtenues avec le script précédent.

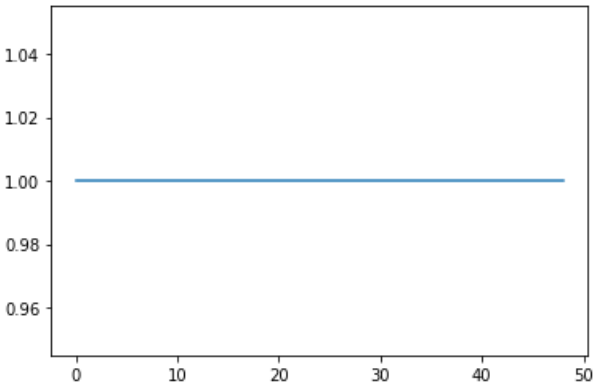

FIGURE 1 – $\mathbb{P}(X_{1000} \in \{0, 2N\})$ pour $N = 1, \dots 50$.

Que peut-on conjecturer sur la valeur de $\lim_{n\to+\infty} \mathbb{P}(X_n \in \{0, 2N\})$? Comment interpréter ce résultat ?

6. **(a)** Soit $i \in [\![0, 2N]\!]$. Donner une interprétation probabiliste de la somme

$$S_i = \sum_{j=0}^{2N} j \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

et en déduire sa valeur.

(b) En déduire que pour tout entier n, on a $E(X_{n+1}) = E(X_n)$.

(c) Interpréter le résultat obtenu.

7. On considère la suite u de terme général $u_n = \mathbb{P}(X_n \in \{0, 2N\})$.

(a) Soit $n \in \mathbb{N}$ et $k \in [\![1, 2N-1]\!]$. Montrer :

$$\mathbb{P}(X_{n+1} \in \{0, 2N\} \mid X_n = k) \geq 2\left(\frac{1}{2N}\right)^{2N}.$$

(b) En déduire que pour tout entier naturel n, on a :

$$u_{n+1} \geq u_n + 2(1 - u_n)\left(\frac{1}{2N}\right)^{2N}.$$

(c) Soit $\alpha \in]0, 1[$. On considère la suite w définie par :

$$w_0 = u_0 \quad \text{et} \quad \forall n \in \mathbb{N},\ w_{n+1} = w_n + \alpha(1 - w_n).$$

Justifier que w est convergente et donner sa limite.

(d) En déduire que u converge et déterminer sa limite.

8. Hétérozygotie. On s'intéresse à la probabilité que deux locus tirés aux hasard (sans remise) à la génération n portent des allèles différents. On note $h(n)$ cette probabilité.

(a) Montrer que pour tout $n \in \mathbb{N}$, on a :

$$h(n) = \sum_{j=0}^{2N} \frac{j(2N-j)}{N(2N-1)} \mathbb{P}(X_n = j).$$

(b) Montrer que pour tout $n \in \mathbb{N}$, $h(n+1) = \left(1 - \frac{1}{2N}\right) h(n)$.

(c) En déduire une expression de $h(n)$ en fonction de n et $h(0)$.

2 Équilibre de Hardy-Weinberg

Soient p_1, p_2 et p_3 trois réels strictement positifs tels que $p_1 + p_2 + p_3 = 1$.
On suppose que pour tout $i = 1, 2, 3$, p_i représente la fréquence du génotype de type i au locus étudié et on note N_i la variable aléatoire de loi $\mathcal{B}(N, p_i)$ donnant le nombre d'individus de type i.

9. Donner sans justification l'espérance et la variance de N_1.

10. On considère la matrice $W = \begin{pmatrix} \mathbb{V}(N_1) & \mathrm{Cov}(N_1, N_2) \\ \mathrm{Cov}(N_2, N_1) & \mathbb{V}(N_2) \end{pmatrix}$.

(a) **Pour les cubes :** justifier que W est diagonalisable.
Pour les autres : on admet que W est diagonalisable.

(b) Soit $(a, b) \in \mathbb{R}^2$. Prouver :

$$\mathbb{V}(aN_1 + bN_2) = \begin{pmatrix} a & b \end{pmatrix} W \begin{pmatrix} a \\ b \end{pmatrix}$$

où l'on a identifié $\mathcal{M}_1(\mathbb{R})$ et $\mathbb{R}$.

(c) En déduire que les valeurs propres de W sont positives.

Indications : on pourra considérer un vecteur propre $\begin{pmatrix} a \\ b \end{pmatrix}$ *et calculer de deux façons* $\begin{pmatrix} a & b \end{pmatrix} W \begin{pmatrix} a \\ b \end{pmatrix}$.

(d) Prouver que les valeurs propres de W sont strictement positives.

(e) En déduire qu'il existe une matrice inversible P et une matrice diagonale D telles que :

$$W = PD^2P^{-1}.$$

(f) Justifier que D est inversible.

On note $A = D^{-1}P^{-1}$ et on considère les variables aléatoires Y_1 et Y_2 telles que :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = A \begin{pmatrix} N_1 - Np_1 \\ N_2 - Np_2 \end{pmatrix}.$$

On admet que :

— $AWA^t = \begin{pmatrix} \mathbb{V}(Y_1) & \mathrm{Cov}(Y_1, Y_2) \\ \mathrm{Cov}(Y_2, Y_1) & \mathbb{V}(Y_2) \end{pmatrix}$

— $P^{-1} = P^t$

où M^t désigne la transposée de la matrice M.

11. Trouver $\mathrm{Cov}(Y_1, Y_2)$, $\mathbb{V}(Y_1)$ et $V(Y_2)$.

12. **(a)** Déterminer la variance de $N_1 + N_2$.

(b) En déduire la covariance de N_1 et N_2.

(c) Montrer que $\det(W) = N^2 p_1 p_2 p_3$ et en déduire W^{-1} (on donnera ses coefficients en fonction de p_1, p_2, p_3 et N).

(d) Déterminer $A^t A$ en fonction de W et en déduire :

$$Y_1^2 + Y_2^2 = \frac{(N_1 - Np_1)^2}{Np_1} + \frac{(N_2 - Np_2)^2}{Np_2} + \frac{(N_3 - Np_3)^2}{Np_3}.$$

Indications : on pourra remarquer que $\begin{pmatrix} Y_1 & Y_2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = Y_1^2 + Y_2^2$.

3 Étude de la loi limite

Soient Z_1 et Z_2 deux variables aléatoires indépendantes de loi normale centrée et réduite. On note $T = Z_1^2 + Z_2^2$.

13. On a tracé sur la figure 2, la fonction de répartition F_T de T.

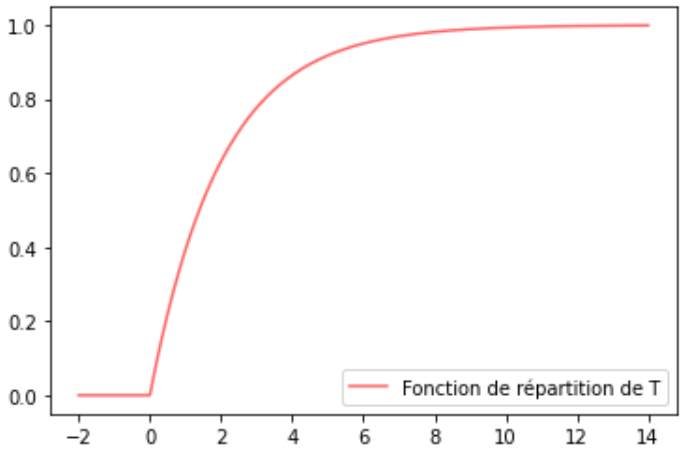

FIGURE 2 – Fonction de répartition de T

(a) Expliquer l'allure de la courbe F_T (variations, limites et valeurs sur $]-\infty, 0]$).

(b) Que peut-on conjecturer sur la nature de T ?

14. **(a)** Écrire une fonction `question_12d(N,p1,p2,p3)` qui prend en argument le nombre d'individus N de la population, les fréquences p_1, p_2, p_3 des trois génotypes et qui simule la variable $Y_1^2 + Y_2^2$ de la question **12.(d)**.

(b) On écrit, à la suite de la fonction précédente, la fonction suivante où `x` est un réel :

```
def mystere(x,N,p1,p2,p3):
    s = 0
    for k in range(1000):
        if question_12d(N,p1,p2,p3)<=x:
            s += 1/1000
    return s
```

Que représente la valeur renvoyée par cette fonction ?

(c) On suppose N, p_1, p_2, p_3 déjà déclarés dans Python et on souhaite tracer la fonction $\texttt{x} \mapsto \texttt{mystere(x,N,p1,p2,p3)}$ entre -2 et 14 avec 1000 points.

Compléter les lignes de codes :

```
x = np.linspace(__ , __ , ___)
L = _______
plt.plot(x,L)
plt.show()
```

(d) On a effectué les tracés pour différentes valeurs de N (voir figures 3 suivantes) :

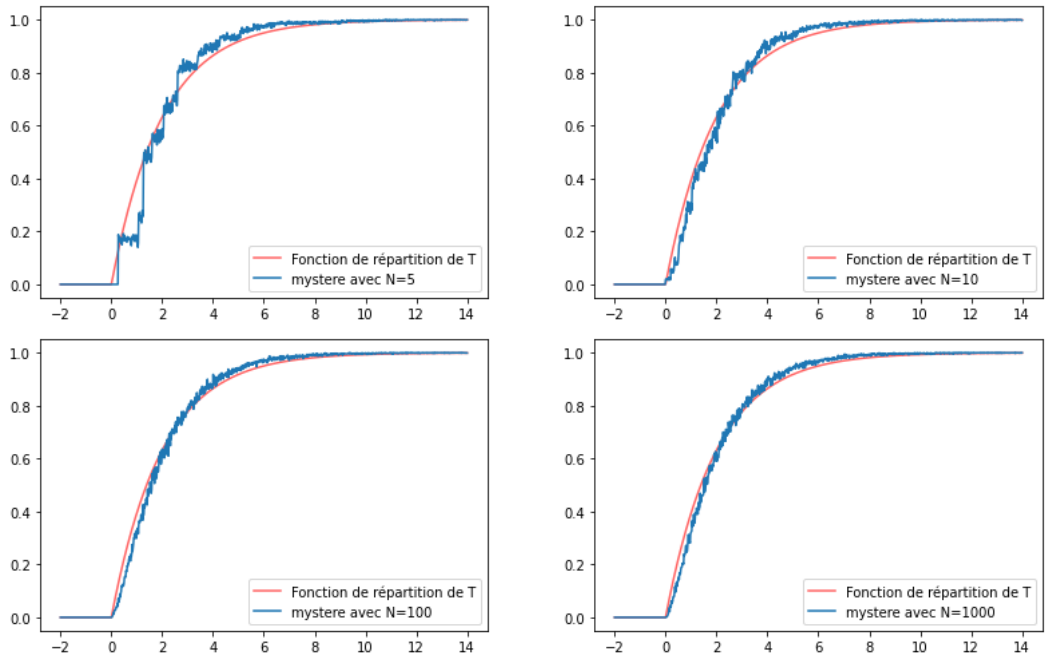

FIGURE 3 – Fonction mystère pour $N \in \{5, 10, 100, 1000\}$.

Que peut-on conjecturer sur $\lim_{N \to +\infty} \mathbb{P}(Y_1^2 + Y_2^2 \leq x)$?

15. Montrer que Z_1^2 est une variable aléatoire à densité de densité $f_1 : t \mapsto \dfrac{1}{\sqrt{2\pi t}} e^{-t/2} 1_{\mathbb{R}^{+*}}(t)$ où $1_{\mathbb{R}^{+*}}$ désigne la fonction indicatrice de $\mathbb{R}^{+*}$ définie par $1_{\mathbb{R}^{+*}}(t) = 1$ si $t \in \mathbb{R}^{+*}$ et $1_{\mathbb{R}^{+*}}(t) = 0$ sinon.

16. Déterminer l'espérance et la variance de Z_1^2.

17. On considère la fonction $h : x \mapsto \displaystyle\int_0^x \frac{1}{\sqrt{t(x-t)}} dt$ et on admet que h est correctement définie sur $\mathbb{R}^{+*}$.

Montrer que h est constante sur $\mathbb{R}^{+*}$.

Indication : on pourra effectuer le changement de variable $t = xu$.

On note C cette constante (on ne cherchera pas à la déterminer).

18. On rappelle que si deux variables aléatoires U_1 et U_2 sont indépendantes et de densités respectives f_{U_1} et f_{U_2} alors $U_1 + U_2$ est une variable aléatoire à densité de densité :

$$x \mapsto \int_{-\infty}^{+\infty} f_{U_1}(x-t) f_{U_2}(t) dt.$$

(a) Déterminer, en fonction de C, la loi de T.

(b) En déduire C puis l'espérance et la variance de T.

4 Annexe des fonctions Python utiles

Dans le module `matplotlib.pyplot` importé sous l'alias `plt` :

```
plt.plot(X,Y)
```

prend en entrée deux vecteurs ou deux listes de même taille, et réalise le tracé des points d'abscisses prises dans X et d'ordonnées prises dans Y. On utilise `plt.show()` pour afficher le tracé.

Dans le module `numpy` importé sous l'alias `np`

```
np.linspace(a,b,n)
```

crée une matrice unidimensionnelle de n coefficients régulièrement espacés dans l'intervalle $[a, b]$.

Dans le module `numpy.random` importé sous l'alias `rd` :

```
rd.binomial(n,p)
```

simule une variable aléatoire suivant une loi de binomiale de paramètres n et p.