Banque « Agro-Véto »
A BCPST - 0421

MATHÉMATIQUES
MÉTHODES DE CALCUL ET DE RAISONNEMENT
Durée : 2 heures 30 minutes

L'usage d'abaques, de tables, de calculatrice et de tout instrument électronique susceptible de permettre au candidat d'accéder à des données et de les traiter par les moyens autres que ceux fournis dans le sujet est interdit.

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 4 pages numérotées de 1 à 4.

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.

Ce sujet a pour objectif l'étude de quelques propriétés mathématiques de modèles en génétique des populations : le modèle d'évolution de Wright-Fisher (partie I) et l'équilibre de Hardy-Weinberg (partie II). Ce dernier fait intervenir les lois du chi-deux qui seront étudiées en partie III.

Les trois parties sont indépendantes entre elles.

Partie I : modèle d'évolution de Wright-Fisher

Soit $N \in \mathbb{N}^*$. On se donne une variable aléatoire X_0 à valeurs dans $[\![0, 2N]\!]$ et on considère une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans $[\![0, 2N]\!]$ vérifiant, pour tout entier n,

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

Pour un variant génétique biallélique (dont les allèles sont notés A et a) et dans le cadre du modèle de Wright-Fisher, X_n représente le nombre d'allèles de type A à la génération n dans une population finie de taille N.

I. Étude d'un cas particulier

On suppose ici que $N = 1$ et on note, pour tout entier 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 $V_{n+1} = MV_n$.
   Une récurrence n'est pas nécessaire.
2. Prouver que M est diagonalisable et déterminer une matrice P inversible et une matrice D diagonale telle que $M = PDP^{-1}$.

3. Calculer M^n pour tout entier n.
4. En déduire que :
   (a) pour tout entier n, on a $\mathrm{E}(X_n) = \mathrm{E}(X_0)$;
   (b) $\lim\limits_{n\longrightarrow+\infty} \mathbb{P}(X_n \in \{0,2\}) = 1$.

II. Cas général

On suppose désormais que $N \geq 1$. On cherche à généraliser les résultats de la question 4.

5. (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 $\mathrm{E}(X_{n+1}) = \mathrm{E}(X_n)$.
   (c) Interpréter le résultat obtenu.
6. On considère la suite u de terme général $u_n = \mathrm{P}(X_n \in \{0, 2N\})$.
   (a) Soit $n \in \mathbb{N}$ et $k \in [\![1, 2N-1]\!]$.
   Montrer que $\mathrm{P}(X_{n+1} \in \{0,2N\} | X_n = k) \geq 2\left(\frac{1}{2N}\right)^{2N}$.
   (b) En déduire que, pour tout entier n, on a $u_{n+1} \geq u_n + 2(1-u_n)\left(\frac{1}{2N}\right)^{2N}$.
   (c) Soit α un réel appartenant à $]0,1[$. 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 la suite w est convergente et donner sa limite.
   (d) Conclure et interpréter le résultat obtenu.

Partie II : Équilibre de Hardy-Weinberg

Soient p_1, p_2, p_3 trois réels strictement positifs tels que $p_1 + p_2 + p_3 = 1$. On considère N variables aléatoires T_1,..., T_N indépendantes de même loi donnée par :

$$T_1(\Omega) = \{1,2,3\} \quad \mathbb{P}(T_1 = 1) = p_1, \quad \mathbb{P}(T_1 = 2) = p_2, \quad \mathbb{P}(T_1 = 3) = p_3.$$

Cette suite permet de modéliser la répartition d'une population de N individus de type AA (type 1), aa (type 2) ou Aa (type 3) en fonction de leur génotype à un locus génétique donné.

Pour tout $i \in [\![1,3]\!]$, on note N_i la variable aléatoire égale au nombre d'individus de type i. On a donc $N_1 + N_2 + N_3 = N$.

7. Déterminer la loi de N_1.
8. Donner, sans justification, l'espérance et la variance de N_1.
9. Soit $(a,b) \in \mathbb{R}^2$ tel que $a \leq b$. Exprimer la quantité

$$\lim_{N\longrightarrow+\infty} \mathrm{P}\left(a \leq \frac{N_1 - Np_1}{\sqrt{Np_1(1-p_1)}} \leq b\right)$$

   sous forme d'une intégrale.

10. On considère la matrice $W = \begin{pmatrix} \mathrm{V}(N_1) & \mathrm{Cov}(N_1, N_2) \\ \mathrm{Cov}(N_2, N_1) & \mathrm{V}(N_2) \end{pmatrix}$.
   (a) Justifier que W est une matrice diagonalisable.
   (b) Soit $(a, b) \in \mathbb{R}^2$. Prouver que
   $$\mathrm{V}(aN_1 + bN_2) = \begin{pmatrix} a & b \end{pmatrix} W \begin{pmatrix} a \\ b \end{pmatrix},$$
   où l'on a identifié $\mathcal{M}_1(\mathbb{R})$ et $\mathbb{R}$.
   (c) En déduire que les valeurs propres de W sont positives.
   (d) Prouver que les valeurs propres de W sont strictement positives.
   (e) En déduire l'existence deux matrices P et D appartenant à $\mathcal{M}_2(\mathbb{R})$, telles que :
      — D soit diagonale et inversible,
      — P soit inversible d'inverse ${}^tP$,
      — $W = PD^2P^{-1}$.
11. On note $A = D^{-1}P^{-1}$ et on considère les variables aléatoire Y_1, 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 l'on a :
   $$\begin{pmatrix} \mathrm{V}(Y_1) & \mathrm{Cov}(Y_1, Y_2) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{V}(Y_2) \end{pmatrix} = AW\,{}^tA.$$
   En déduire $\mathrm{V}(Y_1)$, $\mathrm{V}(Y_2)$ et $\mathrm{Cov}(Y_1, Y_2)$.
12. (a) Déterminer la variance de $N_1 + N_2$.
   (b) En déduire $\mathrm{Cov}(N_1, N_2)$ la covariance de N_1 et N_2.
   (c) Calculer l'inverse de W.
   (d) Déterminer ${}^tA\,A$ en fonction de W et en déduire que :
   $$Y_1^2 + Y_2^2 = \frac{(N_1 - Np_1)^2}{Np_1} + \frac{(N_2 - Np_2)^2}{Np_2} + \frac{(N_3 - Np_3)^2}{Np_3}.$$

On pourrait prouver que, pour tout $(a, b) \in \mathbb{R}^2$,
$$\lim_{N \longrightarrow +\infty} \mathbb{P}\left(a \leq Y_1^2 + Y_2^2 \leq b\right) = \mathbb{P}\left(a \leq Z_1^2 + Z_2^2 \leq b\right)$$
où Z_1 et Z_2 sont deux variables aléatoires indépendantes suivant la loi normale centrée réduite. Connaître la loi de $Z_1^2 + Z_2^2$ permettrait ainsi de tester la vraisemblance de ce modèle. L'étude de cette loi est l'objet de la partie III.

Partie III : Étude de la loi limite

On considère une variable aléatoire Z suivant la loi normale centrée réduite.

13. Prouver que Z^2 est une variable aléatoire de densité $f_{Z^2} \ : \ t \mapsto \dfrac{1}{\sqrt{2\pi t}} e^{-t/2} \, \mathbf{1}_{\mathbb{R}^{+*}}(t)$.

14. Déterminer l'espérance et la variance de Z^2.

15. On considère la fonction $h \ : \ x \mapsto \displaystyle\int_0^x \frac{1}{\sqrt{t(x-t)}} \mathrm{d}t$.

    On admet que h est correctement définie sur $\mathbb{R}^{+*}$.

    Prouver que h est constante sur $\mathbb{R}^{+*}$. On notera C cette constante.

    On pourra utiliser le changement de variable $t = xu$.

16. Soit Z_1 et Z_2 deux variables aléatoires indépendantes suivant la loi normale centrée réduite.

    On rappelle que, si deux variables aléatoires réelles U_1 et U_2 sont indépendantes de densité respectives f_{U_1} et f_{U_2}, alors $U_1 + U_2$ est une variable aléatoire de densité

    $$x \mapsto \int_{\mathbb{R}} f_{U_1}(x-t) f_{U_1}(t) \, \mathrm{d}t \, .$$

    (a) Déterminer en fonction de C la loi de $Z_1^2 + Z_2^2$.

    (b) En déduire C puis l'espérance et la variance de $Z_1^2 + Z_2^2$.

FIN DU SUJET