Agrégation externe de mathématiques, session 2009
Épreuve de modélisation, option Probabilités et Statistiques

(*public 2009*)

Résumé : On s'intéresse à un modèle de génétique des populations, ce qui nous conduit à l'étude d'une chaîne de Markov. Lorsque le nombre n de gènes considérés est très important, certains calculs sont facilités par des passages à la limite lorsque n tend vers l'infini. On obtient ainsi des équations différentielles dont les solutions sont les limites, lorsque n tend vers l'infini, de certaines caractéristiques de la chaîne probabilités d'absorption, mesure invariante.

Mots clefs : Chaîne de Markov, état absorbant, équation différentielle.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Le problème à modéliser

Nous allons considérer l'évolution d'un caractère génétique d'une plante, disons d'une certaine espèce de maïs pour fixer les idées. De manière à obtenir une description statistique de l'évolution du caractère considéré, chaque nouvelle génération est obtenue par la plantation d'un nouveau champ de maïs.

Le passage d'une génération à la suivante est obtenu de la façon suivante. Issu du champ représentatif de la t-ième génération, on extrait au hasard un échantillon restreint de plantes. Cet échantillon aléatoire sert alors de souche génitrice pour la plantation d'un nouveau champ représentatif de la $(t+1)$-ième génération. L'expérimentateur peut envisager de soumettre chaque génération, que ce soit le champ complet ou l'échantillon prélevé, à un traitement (exposition à des rayons, des produits chimiques, des parasites,...) avant de procéder à la constitution de la génération suivante. On peut alors obtenir des informations sur les effets de *mutation* du traitement.

Le *génotype* d'un individu (génotype : composition génétique d'un individu déterminée par l'ensemble des facteurs génétiques) est déterminé par un grand nombre de gènes, dont la plupart admettent plusieurs formes. Par conséquent, dans une espèce donnée, le nombre des différents génotypes possibles dépasse de loin le nombre des individus. L'étude statistique des groupes génotypiques est donc souvent dépourvue de sens. C'est pourquoi, les études les plus simples portent sur le nombre de gènes de même forme, au sein d'une population.

Le caractère génétique qui est ici observé est donc un gène particulier qui peut prendre deux formes A ou a. Dans de nombreuses situations, A est la forme dominante et a la forme récessive

du gène : les chromosomes sont constitués de paires de gènes issus des parents, ceux de type *AA* ou *Aa* expriment le caractère dominant (A) alors que seuls ceux de type *aa* expriment le caractère récessif (a). Les fluctuations de ce caractère génétique au sein de la population résultent des variations des formes A ou a du gène considéré.

Nous allons étudier l'évolution au cours des générations de la proportion de la forme A du gène.

2. Une dynamique de population

Cette proportion est estimée par la proportion empirique observée sur l'échantillon avant qu'il ne soit éventuellement traité et serve de souche pour la génération suivante. On fixe la taille de tous les échantillons souche tirés à chaque génération de sorte qu'à chaque fois on observe n gènes. Soit $N_n(t)$ le nombre de gènes de type A présents dans le t-ième échantillon ; il y a donc $n-N_n(t)$ gènes de type a dans l'échantillon. Nous conservons la mémoire de la taille de l'échantillon en mettant n en indice de façon à pouvoir le faire tendre vers l'infini un peu plus tard.
En l'absence de traitement, seul l'effet d'*échantillonage* intervient de sorte que le nombre $N_n(t+1)$ de gènes de type A que nous trouverons dans le $(t+1)$-ième échantillon est une variable aléatoire dont la loi, conditionnellement à $N_n(t)=i$, est une binômiale de paramètres n et $p_n(i/n)$:

$$\mathscr{L}\text{oi}(N_n(t+1)\mid N_n(t)=i)=\mathscr{B}(n,p_n(i/n))) \tag{1}$$

avec

$$p_n(i/n)=i/n \tag{2}$$

c'est-à dire la proportion de gènes de type A. Nous venons de construire une chaîne de Markov $(N_n(t))_{t\geq 0}$ sur l'espace d'états $\{0,\ldots,n\}$ dont le mécanisme de transition est décrit par (1) et (2).
Lorsque les champs ou les échantillons prélevés sont traités avant d'engendrer la génération suivante, des effets de mutation peuvent apparaître. Nous modélisons l'effet de la *mutation* de la façon suivante. La mutation transforme à la naissance un gène de type A en type a avec la probabilité α et un gène de type a en type A avec la probabilité β $(0\leq\alpha,\beta\leq 1)$, de sorte que si $N_n(t)=i$, la proportion espérée de gènes de type A à la génération suivante est

$$p_n(i/n)=[i(1-\alpha)+(n-i)\beta]/n=(1-\alpha)i/n+\beta(1-i/n). \tag{3}$$

Dans la pratique n est grand, de sorte que les calculs de mesures invariantes, de probabilités et de temps moyens d'atteinte de certains états sont bien approximés par des passages à la limite lorsque n tend vers l'infini. Puisque dans une telle approche l'espace d'états $\{0,\ldots,n\}$ de N_n varie avec n, il est intéressant de considérer la chaîne de Markov des proportions

$$X_n(t)=N_n(t)/n,\quad t=0,1,2,\ldots$$

vivant dans $\mathscr{X}_n = \{k/n; k = 0, 1, \ldots, n\}$ inclus dans l'ensemble fixe $[0,1]$. Sa matrice de transition est donnée pour tous x, y dans $\mathscr{X}_n$ par

$$P_n(x,y) = \mathbb{P}(X_n(t+1) = y \mid X_n(t) = x) = \mathrm{C}_n^{ny} p_n(x)^{ny} (1 - p_n(x))^{n(1-y)}$$

où C_n^{ny} est le nombre de combinaisons de ny parmi n et p_n est donné en (2) ou (3) selon les mécanismes à décrire.

3. Probabilités d'absorption pour le modèle sans mutation

On se place dans la situation décrite par le mécanisme de pur échantillonage (2). Les états 0 et 1 qui correspondent à des échantillons saturés en gènes de même type, sont des états absorbants de la chaîne de Markov X_n. Une fois que la population est entièrement constituée de gènes du même type, elle ne se modifie plus. Il est intéressant de calculer avec quelles probabilités la population va se fixer en A ou en a, sachant la condition initiale $X_n(0)$. On cherche donc à calculer pour tout x dans $\mathscr{X}_n$

$$u_n(x) = \mathbb{P}(X_n \text{ finit par atteindre } 1 \mid X_n(0) = x).$$

Plus précisément nous allons calculer la limite

$$u(x) = \lim_{n\to\infty} u_n(x),\ x \in [0,1].$$

Pour tout $x \in \mathscr{X}_n$ et tout n nous avons

$$u_n(x) = \sum_{y \in \mathscr{X}_n} P_n(x,y) u_n(y), \tag{4}$$

avec les conditions aux limites $u_n(0) = 0$ et $u_n(1) = 1$.

Théorème 1. *La limite $u(x)$ est la solution de l'équation différentielle*

$$u''(x) = 0,\ x \in]0,1[, \quad u(0) = 0,\ u(1) = 1 \tag{5}$$

c'est-à-dire $u(x) = x$, $0 \le x \le 1$.

Nous allons maintenant donner quelques arguments qui ne constituent pas une preuve complète, mais qui permettent d'identifier $u(x)$ comme la solution de (5).
Par la suite, nous désignerons par $O_{n\to\infty}(n^{-k})$ toute suite réelle v_n telle que $\sup_{n\to\infty} n^k |v_n| < \infty$, toute suite de fonctions $v_n(x)$ telle que $\sup_{n\to\infty} n^k \sup_x |v_n(x)| < \infty$ ainsi que toute suite de variables aléatoires V_n telle que $\sup_{n\to\infty} n^k |V_n| < \infty$ avec une dépendance en l'aléa que nous ne précisons pas.
Pour un usage ultérieur, nous ne supposerons pas dans les calculs qui suivent que $p_n(x) = x$ (comme le stipule (2)), mais seulement que $p_n(x) - x = O_{n\to\infty}(1/n)$.
Notons Z_n une variable aléatoire telle que nZ_n soit de loi $\mathscr{B}(n, p_n(x))$ (donc à valeurs dans $\mathscr{X}_n$). L'égalité (4) s'écrit

$$\mathbb{E}\Big[u_n\big(x + [Z_n - x]\big) - u_n(x)\Big] = 0 \tag{6}$$

Du fait que $Z_n - x = [Z_n - p_n(x)] + [p_n(x) - x]$, nous avons $\mathbb{E}(Z_n - x) = p_n(x) - x$ et $\mathbb{E}(Z_n - x)^2 = p_n(x)(1 - p_n(x))/n + [p_n(x) - x]^2$. Le point important est que grâce au théorème de la limite centrale $Z_n - x$ est de l'ordre de $O(n^{-1/2})\mathcal{N}(0,1) + [p_n(x) - x] = O(n^{-1/2})$. Bien que ce soit plus difficile, on peut montrer grâce à un renforcement de ce théorème que lorsque

$$p_n(x) - x = O_{n\to\infty}(1/n), \tag{7}$$

on a pour tout $k \geq 2$, $\mathbb{E}|Z_n - x|^k = O(n^{-k/2})$. On peut même obtenir le développement limité au second ordre suivant dans (6) :

$$u_n'(x)\mathbb{E}[Z_n - x] + (1/2)u_n''(x)\mathbb{E}[Z_n - x]^2 + O(n^{-3/2}) = 0, \quad x \in \mathscr{X}_n$$

(l'ensemble $\mathscr{X}_n$ étant discret, il faut comprendre u_n' et u_n'' au sens des différences finies : $u_n'(x) = [u_n(x + 1/n) - u_n(x)]/(1/n)$, $u_n''(x) = [u_n'(x + 1/n) - u_n'(x)]/(1/n)$). Après multiplication par n, cela permet le passage à la limite suivant

$$b(x)u'(x) + (1/2)a(x)u''(x) = 0, \quad x \in [0,1] \tag{8}$$

où les coefficients de l'équation différentielle sont donnés par

$$b(x) = \lim_{n\to\infty} n[p_n(x) - x] \quad \text{et} \quad a(x) = \lim_{n\to\infty} p_n(x)(1 - p_n(x)) = x(1-x). \tag{9}$$

Notons que l'existence de la limite $b(x)$ impose (7) et que (7) permet d'obtenir par passage à la limite, l'expression de $a(x)$. Finalement,en tenant compte de (8) avec $b(x) = 0$ et des conditions aux limites, on obtient (5).

4. Probabilité invariante pour le modèle avec mutation

On se place maintenant dans la situation décrite par le mécanisme de mutation (3). Lorsque α et β sont strictement positifs, les états 0 et 1 ne sont plus des états absorbants et la chaîne X_n admet une unique mesure de probabilité invariante $m_n = (m_n(x); x \in \mathscr{X}_n)$ qui nous donne une information importante sur la répartition de la proportion de la forme A du gène au cours du temps. Nous allons identifier la *densité* $m(x), 0 \leq x \leq 1$ de la limite étroite $m(x)\,dx$ de la suite de probabilités (m_n) sur $[0,1]$.

Théorème 2. *La densité limite $m(x)$ est une solution de l'équation différentielle*

$$0 = -\frac{d}{dx}\Big([-\gamma x + \delta(1-x)]m(x)\Big) + \frac{1}{2}\frac{d^2}{dx^2}\Big([x(1-x)]m(x)\Big) \tag{10}$$

lorsque les paramètres de mutation qui apparaissent dans (3) sont de la forme

$$\alpha = \gamma/n, \quad \beta = \delta/n$$

avec $\gamma, \delta > 0$.
Un calcul élémentaire (une fois que l'on introduit le bon facteur intégrant) permet de résoudre cette équation et nous donne

$$m(x) = \frac{C_1 S(x) + C_2}{s(x)\sigma^2(x)}$$

avec $s(x) = \exp\left(-\int^x 2\frac{-\gamma z+\delta(1-z)}{z(1-z)}\,dz\right)$ *et* $S(x) = \int^x s(z)dz$, *où* $\int^x$ *signifie que l'on prend une primitive sans spécifier la constante additive.*

Nous allons donner quelques arguments qui permettent d'identifier la limite $m(x)\,dx$ de la suite (m_n). Pour tout $y \in \mathscr{X}_n$ et tout n nous avons

$$m_n(y) = \sum_{x\in\mathscr{X}_n} m_n(x)P_n(x,y). \tag{11}$$

En représentant $u_n \in \mathbf{R}^{n+1}$ à l'aide d'une matrice colonne, $m_n \in \mathbf{R}^{n+1}$ à l'aide d'une matrice ligne, en notant P_n la matrice de transition de X_n, I_n la matrice identité de $\mathbf{R}^{n+1}$ et en posant $L_n := n(P_n - I_n)$, les relations (4) et (11) s'écrivent

$$L_n u_n = 0 \quad \text{et} \quad m_n L_n = 0.$$

En effet une inspection des arguments de la section précédente nous permet de montrer en remplaçant u_n par $v_n = (v(x); x \in \mathscr{X}_n)$ que pour toute fonction régulière $v(x)$ sur $[0,1]$, nous avons

$$\lim_{n\to\infty} L_n v_n = Lv \text{ où } (Lv)(x) = b(x)v'(x) + (1/2)a(x)v''(x), \quad x \in [0,1]$$

où cette limite signifie que pour toute fonction continue $f(x)$ sur $[0,1]$, nous avons $\lim_{n\to\infty} n^{-1}\sum_{x\in\mathscr{X}_n} f(x)(L_n v_n)(x) = \int_{[0,1]} f(x)(Lv)(x)\,dx$.

Du fait que $m_n L_n = 0$, on a pour tout n et toute fonction régulière v : $m_n L_n v_n = 0$. On peut aussi montrer que si (m_n) admet une limite étroite $m(x)\,dx$, on peut passer à la limite dans la dernière identité, qui nous donne

$$0 = \int_{[0,1]} m(x)(Lv)(x)\,dx = \int_{[0,1]} (L^* m)(x)v(x)\,dx$$

où $(L^*m)(x) = -\frac{d}{dx}(b(x)m(x)) + \frac{1}{2}\frac{d^2}{dx^2}(a(x)m(x))$ s'obtient à l'aide d'intégrations par parties, pour toute fonction régulière v telle que $v(0) = v(1) = 0$. On en déduit $L^*m = 0$ avec $b(x) = \lim_{n\to\infty} n[p_n(x) - x] = -\gamma x + \delta(1-x)$ qui est le résultat désiré.

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

– *Étude de la modélisation*. Vous pourriez approfondir l'explicitation des modèles spécifiés par (1), (2) et (3) les commenter et les critiquer. En particulier, expliquer de manière détaillée (1)–(2) et commenter la décroissance en $1/n$ des paramètres $\alpha = \gamma/n$ et $\beta = \delta/n$. Vous pourriez proposer un schéma approximatif de test statistique de l'hypothèse nulle $(\gamma = \delta = 0)$ contre l'hypothèse alternative $\Big((\gamma,\delta) \neq (0,0)\Big)$ à partir d'une longue observation d'une réalisation de la chaîne X_n.

– *Développements mathématiques*. Vous pourriez compléter, préciser et commenter certains éléments de démonstration. Vous pourriez exposer un développement à partir de (4), ou de tel autre élément mathématique de cet article.
– *Étude numérique*. Vous pourriez effectuer des simulations (avec n modéré) afin d'illustrer l'évolution de la proportion de A et d'évaluer expérimentalement les probabilités d'absorption de chacun des états absorbants partant de x. Vous pourriez comparer ces résultats aux résultats théoriques asymptotiques du Théorème 1. De même, vous pourriez évaluer expérimentalement la probabilité invariante dans le cadre du Théorème 2. Ou encore, mais ce serait alors une autre direction, vous pourriez résoudre numériquement la solution théorique du Théorème 2 pour un choix de paramètres γ et δ strictement plus grands que 1/2, et commenter vos résultats.