Agrégation externe de mathématiques, session 2021
Épreuve de modélisation, option A : Probabilités et Statistiques

(*A1*)

Résumé : On se propose de construire une famille de variables aléatoires dépendant d'un paramètre pour modéliser le nombre d'*espèces* par *genre*. Un genre est un ensemble d'espèces ayant des caractérisques communes. La construction de ces variables aléatoires repose sur celle de processus à temps continu modélisant l'évolution au cours du temps du nombre de genres et du nombre d'espèces dans chacun de ces genres. Les variables aléatoires cherchées sont alors obtenues en regardant le nombre d'espèces dans un genre choisi au hasard parmi ceux présents aujourd'hui.
L'objectif de ce texte est de présenter ces modèles et d'étudier certaines de leurs propriétés.

Mots clefs : Loi exponentielle, espérance conditionnelle, martingales.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

En 1924, des statisticiens écossais observent sur des données que la distribution du nombre d'*espèces* par *genre* est à queue lourde, c'est-à-dire que la probabilité que ce nombre soit supérieur à x ne décroît pas exponentiellement vite vers 0 quand $x \to +\infty$. Un genre est un ensemble d'espèces ayant des caractérisques communes. Par exemple, la truite fario, la truite arc-en-ciel et le saumon de fontaine sont différentes espèces qui font partie du genre « truite ». Ces statisticiens ont cherché à construire une famille de modèles aléatoires avec lesquels comparer leurs données.

Pour cela, ils raisonnent de la façon suivante. De nouvelles espèces sont régulièrement créées suite à des mutations du génome. Plus rarement, des mutations plus conséquentes sont à l'origine de nouveaux genres. Ces mutations sont a priori indépendantes. Un modèle logique pour l'évolution des genres et des espèces dans les genres satisfera donc les trois points suivants :

1. Initialement, il y a un genre avec une espèce. Cette espèce va muter pour donner une nouvelle espèce. Il y aura alors 2 espèces dans le genre, espèces qui vont à leur tour muter de façon indépendante, et ainsi de suite.
2. Indépendamment, le genre va muter pour donner un nouveau genre. Puis ces deux genres vont muter à leur tour, indépendamment l'un de l'autre, etc.

3. Une fois qu'un nouveau genre est créé, il contient une espèce, qui va ensuite évoluer suivant le schéma 1., indépendamment de la création de nouveaux genres et de l'évolution des espèces dans ceux-ci.

La description mathématique de ces modèles à temps continu se trouve dans la Section 2. La Section 1 est consacrée à la description de l'évolution du nombre d'espèces dans un genre donné. Enfin dans la Section 3, on étudiera la distribution du nombre d'espèces dans un genre choisi au hasard aujourd'hui.

1. Évolution du nombre d'espèces dans un genre

Dans cette partie, on travaille avec un genre donné. On cherche à modéliser l'évolution du nombre d'espèces dans ce genre par un processus à temps continu $X = (X_t, t \geq 0)$ à valeurs dans $\mathbb{N}$, dépendant d'un paramètre $\lambda \in]0, +\infty[$. Il se construit de la façon suivante :

— $X_0 = 1$

— au bout d'un temps T_1 de loi exponentielle de paramètre λ, l'espèce initiale donne naissance à une nouvelle espèce. On a alors deux espèces dans le genre, et on pose

$$X_t = 1, \quad \text{si } t < T_1, \quad X(T_1) = 2. \tag{1}$$

— les deux espèces présentes au temps T_1 évoluent ensuite indépendamment l'une de l'autre et indépendamment du passé, suivant le même schéma : chaque espèce donne naissance à une nouvelle espèce au bout d'un temps de loi exponentielle de paramètre λ. Et ainsi de suite. On note X_t le nombre d'espèces présentes dans le genre au temps t.

On peut étendre ce modèle au cas où $i \in \mathbb{N}^*$ espèces sont présentes au temps 0 dans le genre, en décidant que chacune de ces i espèces initiales évoluent indépendamment les unes des autres. On note $X^{(i)}$ le processus obtenu ainsi. On a donc $X^{(1)} \overset{\text{loi}}{=} X$ et $X^{(i)}$ est la somme de i processus indépendants, tous de même loi que X.

Propriétés de Markov et de branchement (admises). Le modèle est construit de sorte que pour chaque temps s déterministe et toute fonction $f : \mathbb{N} \to \mathbb{R}^+$,

— $\mathbb{E}[f(X_{t+s})|\mathscr{F}_s] = \mathbb{E}[f(X_{t+s})|X_s]$, où $\mathscr{F}_s$ désigne la tribu engendrée par les variables $X_u, 0 \leq u \leq s$ et t est un réel positif

— la loi de $(X_{t+s}, t \geq 0)$ sachant $X_s = i$ est la même que celle de $(X_t^{(i)}, t \geq 0)$.

Ces propriétés restent vraies si on remplace s par un des temps de saut du processus. On peut en déduire la proposition suivante.

Proposition 1. *Posons $T_0 = 0$ et notons $(T_i)_{i\geq 1}$ la suite des temps de saut du processus $(X_t)_{t\geq 0}$. Alors la suite $(T_i - T_{i-1})_{i\geq 1}$ est une suite de variables aléatoires indépendantes, et pour tout $i \geq 1$, la variable aléatoire $T_i - T_{i-1}$ suit la loi exponentielle de paramètre λi.*

Théorème 2. *Pour tout $t \geq 0$, la variable aléatoire X_t suit la loi géométrique de paramètre $e^{-\lambda t}$.*

On rappelle qu'une variable aléatoire Y suit la loi géométrique de paramètre $p \in]0, 1[$ si, pour tout $n \in \mathbb{N}^*$, $\mathbb{P}(Y = n) = p(1-p)^{n-1}$.

Démonstration. Soit un réel $t \geq 0$. On voit que $\mathbb{P}(X_t = 1) = e^{-\lambda t}$.

Fixons à présent un entier $j \geq 2$. Avec les notations de la proposition précédente, on peut écrire

$$\begin{aligned}
\mathbb{P}(X_t = j) &= \mathbb{P}(T_{j-1} \leq t < T_j) \\
&= j!\lambda^j \int_{\{x_1+\cdots+x_{j-1} \leq t < x_1+\cdots+x_j\} \cap (\mathbb{R}_+)^j} e^{-\lambda x_1} e^{-2\lambda x_2} \cdots e^{-j\lambda x_j} \, \mathrm{d}x_1 \cdots \mathrm{d}x_j \qquad (2) \\
&= j!\lambda^j \int_{\{0 \leq y_1 \leq y_2 \leq \cdots \leq y_{j-1} \leq t < y_j\}} e^{-j\lambda y_j} e^{\lambda y_1} e^{\lambda y_2} \cdots e^{\lambda y_{j-1}} \, \mathrm{d}y_1 \cdots \mathrm{d}y_j \\
&= (j-1)!\lambda^{j-1} e^{-\lambda j t} \int_{\{0 \leq y_1 \leq y_2 \leq \cdots \leq y_{j-1} \leq t\}} e^{\lambda y_1} e^{\lambda y_2} \cdots e^{\lambda y_{j-1}} \, \mathrm{d}y_1 \cdots \mathrm{d}y_{j-1}.
\end{aligned}$$

On montre ensuite que

$$(j-1)!\lambda^{j-1} \int_{\{0 \leq y_1 \leq y_2 \leq \cdots \leq y_{j-1} \leq t\}} e^{\lambda y_1} e^{\lambda y_2} \cdots e^{\lambda y_{j-1}} \, \mathrm{d}y_1 \cdots \mathrm{d}y_{j-1} = (e^{\lambda t} - 1)^{j-1}, \qquad (3)$$

ce qui permet de conclure. □

Le théorème 2 permet de décrire le comportement en temps grand du processus X.

Théorème 3. *Lorsque* $t \to \infty$*,*

$$e^{-\lambda t} X_t \xrightarrow{\text{loi}} E \qquad (4)$$

où E *suit la loi exponentielle de paramètre* 1*.*

Démonstration. Soit un réel $x \geq 0$. D'après le théorème précédent,

$$\mathbb{P}(e^{-\lambda t} X_t \geq x) = (1 - e^{-\lambda t})^{\lceil x e^{\lambda t} \rceil - 1} \to e^{-x} \qquad (5)$$

(où $\lceil u \rceil$ désigne la partie entière supérieure d'un réel u) lorsque $t \to +\infty$. On reconnaît la queue de distribution d'une variable aléatoire suivant la loi exponentielle de paramètre 1. □

La convergence du théorème 3 est en fait plus forte.

Théorème 4. *Lorsque* $t \to \infty$*,*

$$e^{-\lambda t} X_t \xrightarrow{\text{p.s.}} E. \qquad (6)$$

Démonstration. Pour tout $t \geq 0$, on a

$$\mathbb{E}[X_t] = e^{\lambda t}. \qquad (7)$$

On en déduit que $\left(e^{-\lambda t} X_t \, ; \, t \geq 0\right)$ est une martingale, c'est-à-dire que pour $0 \leq s \leq t$

$$\mathbb{E}[e^{-\lambda t} X_t | \mathscr{F}_s] = e^{-\lambda s} X_s, \qquad (8)$$

($\mathscr{F}_s$ est la tribu engendrée par les variables $X_u, 0 \leq u \leq s$). De façon analogue au cas discret, une (sur-)martingale positive et continue à droite converge presque sûrement. D'où la conclusion. □

2. Évolution du nombre de genres

On suppose toujours que le nombre d'espèces dans un genre donné évolue suivant le modèle à un paramètre de la section précédente avec un paramètre $\lambda > 0$ quelconque. On dira alors que le nombre d'espèces dans un genre donné évolue suivant le modèle $M(\lambda)$.

Le modèle proposé par les statisticiens écossais est alors le suivant :

1. Initialement, il y a un genre, avec une espèce. Le nombre d'espèces dans ce genre évolue suivant le modèle $M(\lambda)$.
2. Indépendamment, chaque genre produit un nouveau genre à un taux constant $\mu > 0$. Autrement dit, le nombre de genres dans le système suit le modèle $M(\mu)$.
3. Au moment où un genre est créé, il a une espèce. Puis le nombre d'espèces dans ce genre va évoluer suivant le modèle $M(\lambda)$, indépendamment de la création de nouveaux genres et de l'évolution des espèces dans les autres genres.

3. Nombre d'espèces dans un genre choisi au hasard aujourd'hui

On se place maintenant au temps présent, disons t, et on choisit un genre uniformément au hasard parmi ceux présents au temps t. On note T_t le temps qui s'est écoulé entre l'apparition de ce genre et le temps t.

Proposition 5. *T_t converge en loi, lorsque $t \to \infty$, vers une loi exponentielle de paramètre μ.*

Démonstration. Pour tous $t \geq 0$, notons N_t le nombre de genres présents au temps t. Par hypothèse, le processus $(N_t, t \geq 0)$ a la même dynamique que le processus $(X_t, t \geq 0)$ de la section 1, de paramètre μ.

On remarque alors que, pour $0 \leq u < t$,

$$\mathbb{P}(T_t \geq u | N_s, 0 \leq s \leq t) = \frac{N_{t-u}}{N_t}, \tag{9}$$

donc,

$$\mathbb{P}(T_t \geq u) = \mathbb{E}\left[\frac{N_{t-u}}{N_t}\right] = \mathbb{E}\left[\frac{N_{t-u}}{\tilde{N}_{u,1} + \cdots + \tilde{N}_{u,N_{t-u}}}\right], \tag{10}$$

où, sachant N_{t-u}, les $\tilde{N}_{u,i}$ (pour $i \geq 1$) sont des v.a. indépendantes de même loi que N_u. En utilisant la loi forte des grands nombres et le théorème de convergence dominée, on obtient

$$\mathbb{E}\left[\frac{N_{t-u}}{\tilde{N}_{u,1} + \cdots + \tilde{N}_{u,N_{t-u}}}\right] \to \frac{1}{\mathbb{E}[N_u]} \tag{11}$$

lorsque $t \to +\infty$. On conclut alors avec le théorème 2 : $\mathbb{P}(T_t \geq u) \xrightarrow[t \to +\infty]{} e^{-\mu u}$. □

On supposera dans la suite que t est très grand et donc que le genre choisi au hasard aujourd'hui est apparu il y a un temps distribué suivant une loi exponentielle de paramètre μ. On note alors N le nombre d'espèces dans ce genre. Le théorème 2 nous permet d'établir le résultat suivant.

Proposition 6. *Pour tout entier $n \geq 1$,*

$$\mathbb{P}(N=n) = \int_0^\infty \mu e^{-\mu t}\, e^{-\lambda t}\, (1-e^{-\lambda t})^{n-1} \mathrm{d}t,$$

$$= \int_0^\infty e^{-(1+\rho)u} (1-e^{-\rho u})^{n-1} \mathrm{d}u \tag{12}$$

avec $\rho = \lambda/\mu$. On en déduit que $\mathbb{E}[N]$ est finie si et seulement si $\rho < 1$, et que dans ce cas $\mathbb{E}[N] = 1/(1-\rho)$.

En posant $x = ne^{-\rho u}$, on trouve l'égalité

$$\mathbb{P}(N=n) = \frac{n^{-1-\rho^{-1}}}{\rho} \int_0^n x^{\rho^{-1}} \left(1 - \frac{x}{n}\right)^{n-1} \mathrm{d}x, \tag{13}$$

qui montre que $\mathbb{P}(N=n)$ est asymptotiquement équivalente à $\Gamma(1+\rho^{-1}) n^{-1-\rho^{-1}}/\rho$, où Γ désigne la fonction Gamma d'Euler. Par ailleurs, le changement de variable $z = x/n$ dans (13) donne la relation $\mathbb{P}(N=n) = B(1+1/\rho, n)/\rho$, où B désigne la fonction Bêta d'Euler.

C'est cette famille de lois indexée par le paramètre ρ que les statisticiens écossais proposent pour décrire le nombre d'espèces dans un genre choisi au hasard. Le tableau de la figure 1 fournit un exemple de données qu'ils avaient en leur possession.

x	Nombre observé de genres à x espèces	Nombre calculé de genres à x espèces
1	125	124.61
2	35	45.48
3	28	24.32
4	17	15.39
5	15	10.73
6	9	7.96
7	8	6.17
8	6	4.94
9 à 11	13	10.35
12 à 14	3	6.61
15 et +	34	36.44
	Total = 293	Total = 293

FIGURE 1. Serpents : nombre, observé et calculé, de genres de chaque taille. Les valeurs calculées correspondent à $1/\rho = 0.74$ (valeurs arrondies à 10^{-2} près).

Suggestions et pistes de réflexion

► *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

— *Développements mathématiques :*
- Vous pouvez développer les preuves des théorèmes 2, 3, 4 et des propositions 5 et 6.
- Dans la description du processus $(X_t, t \geq 0)$ de la Section 1, il est implicite que X_t est bien défini pour tout $t \geq 0$ mais ceci n'est pas démontré. Précisément, si l'on note (T_n) la suite croissante des temps de sauts de ce processus et $T_\infty = \lim_{n\to+\infty} T_n$ alors X_t n'est défini que pour $t < T_\infty$. Pouvez-vous montrer que $\mathbb{P}(T_\infty = +\infty) = 1$?
- Comment pourrait-on estimer ρ ?

— *Étude numérique :*
- Vous pouvez tester l'adéquation de la loi proposée aux valeurs observées dans le tableau de données.
- Vous pouvez simuler le modèle proposé dans la Section 1. et illustrer les comportements asymptotiques des théorèmes 3 et 4.

— *Modélisation :*
- En comparant leur modèle aux données, avec un paramètre ρ adéquat, les statisticiens écossais ont conclu que ce modèle est « bon ». Êtes-vous d'accord ? Critiquez le modèle.
- Pouvez-vous expliquer pourquoi la loi de N ne dépend que du paramètre ρ ? Discutez le rôle de ce paramètre.