Début de l'épreuve

Le sujet porte sur un modèle d'échantillonnage avec remise dans une population avec N espèces différentes, pour une valeur de $N \in \mathbb{N}^*$ fixée. Nous allons étudier différentes quantités d'intérêt sur ce modèle, telles que le nombre d'espèces observées sur un échantillon de taille $n \in \mathbb{N}^*$, le nombre minimal d'individus à échantillonner afin d'observer toutes les espèces, ainsi que le comportement de cette dernière variable aléatoire lorsque $N \to +\infty$.

Le sujet comprend 7 pages, numérotées de 1 à 7, et contient 5 parties.
La **Partie I** ne fait aucune référence à des probabilités, certains résultats obtenus dans cette partie seront nécessaires dans la **Partie IV**. Les **Parties II - V** portent sur le même modèle. Toutes les définitions et notations relatives au modèle sont données dans la section **Modèle**, les parties **II** à **V** peuvent donc être traitées indépendamment, sauf mention explicite du contraire. Tous les résultats intermédiaires donnés par l'énoncé peuvent être admis pour traiter la suite du sujet, à condition de le mentionner explicitement sur la copie. Il est recommandé de lire attentivement le sujet. Il est demandé de veiller au soin de la présentation, ainsi qu'à la rigueur et à la concision des raisonnements.

Notations et rappels

– On utilise les notations usuelles $\mathbb{N}, \mathbb{R}$ pour l'ensemble des nombres naturels et réels, mais aussi $\mathbb{N}^*, \mathbb{R}^*$ pour ces ensembles privés de 0. Pour $m \leq n$ deux entiers naturels, on note $[\![m, n]\!] := \{k \in \mathbb{N} : m \leq k \leq n\}$.
– Pour A sous-ensemble de Ω, on appelle indicatrice de A l'unique fonction $\mathbb{1}(A) : \Omega \to \{0, 1\}$, qui vaut 1 pour tout $\omega \in A$ et 0 sinon.
– Deux suites réelles (u_n) et (v_n), avec $v_n \neq 0$ $\forall n \in \mathbb{N}$, sont dites équivalentes lorsque $n \to \infty$ si la suite $(\frac{u_n}{v_n})$ tend vers 1. On utilise la notation $u_n \sim v_n$.
– Pour tout $n \in \mathbb{N}^*$, on note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées à coefficients réels de taille n. La matrice identité dans $\mathcal{M}_n(\mathbb{R})$ est notée $\mathbf{I}_n$. La transposée d'une matrice $M \in \mathcal{M}_n(\mathbb{R})$, est notée M^T et sa trace, définie comme la somme de ces éléments diagonaux, est notée $\mathrm{Tr}(M)$. Pour $n \in \mathbb{N}^*$, on considère l'espace vectoriel $\mathbb{R}^n$ et on identifie cet espace avec $\mathcal{M}_{n,1}(\mathbb{R})$, l'ensemble de vecteurs colonnes de taille n. Dans $\mathbb{R}^n$, on note $\overrightarrow{1_n}$ le vecteur ne comportant que des 1.
– Pour $D \subset \mathbb{R}^n$ et une fonction $\varphi : D \to \mathbb{R}$, on dit que φ atteint son maximum sur D en $\vec{x}^* \in D$, si $\forall \vec{x} \in D$, $\varphi(\vec{x}^*) \geq \varphi(\vec{x})$. On dit qu'elle atteint son minimum si l'inégalité est vérifiée dans l'autre sens ($\leq$).
– Toutes les variables aléatoires de cet énoncé sont définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On note respectivement $\mathbb{E}, \mathbb{V}\mathrm{ar}$ et $\mathbb{C}\mathrm{ov}$, l'espérance, la variance et la covariance dans cet espace.
– Pour tout $p \in [0, 1]$, on dira qu'une variable aléatoire G est de loi géométrique de paramètre p, si $\mathbb{P}(G = k) = (1-p)^{k-1}p$, pour tout $k \in \mathbb{N}^*$.

– On admettra les deux résultats suivants, qui pourront être utilisés uniquement si l'énoncé le précise explicitement. Il est conseillé de ne pas lire ces deux résultats tout de suite, mais seulement lorsqu'ils seront invoqués dans l'énoncé.

Formule générale du crible. *Soit $n \in \mathbb{N}^*$, et soit $(A_i)_{1 \leq i \leq n}$ une famille d'évènements de $\mathcal{F}$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\ell=1}^{n} \left((-1)^{\ell+1} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_\ell \leqslant n} \mathbb{P}\left(\bigcap_{j=1}^{\ell} A_{i_j}\right)\right),$$

où $\sum_{1 \leq i_1 < i_2 < \ldots < i_\ell \leq n}$ est la somme sur tous les éléments de l'ensemble $\{(i_1, \ldots, i_\ell) \in [\![1, n]\!]^\ell : i_1 < \ldots < i_\ell\}$.

Théorème 1. *Soit I un intervalle de $\mathbb{R}$, et une fonction $\varphi : I \to \mathbb{R}$. La fonction φ est dite **concave** si $\forall x, y \in I$ et tout $\lambda \in [0, 1]$, $\varphi(\lambda x + (1-\lambda)y) \geq \lambda\varphi(x) + (1-\lambda)\varphi(y)$. Nous avons les deux propositions suivantes,*

(i) Si φ est de classe $\mathscr{C}^2(I)$ (admettant des dérivées continues jusqu'à l'ordre 2), φ est concave si et seulement si φ'' est à valeurs négatives ou nulles.

*(ii) **Inégalité de Jensen**. La fonction φ est concave si et seulement si pour tout $n \in \mathbb{N}$, $n \geq 2$, pour tous $x_1, \ldots, x_n \in I$ et pour tous réels $\lambda_1, \ldots, \lambda_n \in [0, 1]$ tels que $\sum_{i=1}^n \lambda_i = 1$, on a l'inégalité suivante*

$$\varphi\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geq \sum_{i=1}^{n} \lambda_i \varphi(x_i).$$

Partie I - Questions préliminaires

Pour tout $n \in \mathbb{N}^*$, on définit

$$H_n = \sum_{k=1}^{n} \frac{1}{k}, \qquad C_n = \sum_{k=1}^{n} \frac{1}{k^2}, \qquad \text{et} \qquad W_n = \sum_{k=1}^{n} \frac{(-1)^k}{k^2}.$$

1. Soit $a \geq 1$ et $k \geq 2$. Montrer que

$$\int_k^{k+1} \frac{dt}{t^a} \leq \frac{1}{k^a} \leq \int_{k-1}^{k} \frac{dt}{t^a}.$$

2. On étudie le comportement de $(H_n)_{n\geq 1}$. Pour tout $n \geq 1$, on pose $u_n = H_n - \ln n$.

(a) Montrer que pour tout entier $n \geq 1$, on a

$$\frac{1}{n+1} \leq \ln\left(\frac{n+1}{n}\right) \leq \frac{1}{n}.$$

(b) Montrer que pour tout entier $n \geq 1$, $0 \leq u_n \leq 1$. En déduire que la suite (u_n) converge vers une certaine limite $\gamma \in [0, 1]$, qui est connue sous le nom de *constante d'Euler-Mascheroni*.

(c) Donner un équivalent de H_n lorsque $n \to \infty$.

3. Montrer que pour tout $a > 1$,

$$\sum_{k=1}^{n} \frac{1}{k^a} \leq \frac{a}{a-1}.$$

En déduire que les suites $(C_n)_{n\geq 1}$ et $(W_n)_{n\geq 1}$ sont convergentes.

4. Soient $C = \lim_n C_n$ et $W = \lim_n W_n$. Montrer que $W = -\frac{1}{2}C$.

Suggestion : étudier la suite $(C_n + W_n)$.

5. On cherche à calculer C.

(a) Soit $f : [0, \pi] \to \mathbb{R}$ une fonction de classe $\mathscr{C}^1$ sur $[0, \pi]$. Montrer que

$$\lim_{\lambda \to +\infty} \int_0^{\pi} f(t) \sin(\lambda t)\, \mathrm{d}t = 0$$

(b) Vérifier que pour tout $\alpha, \beta \in \mathbb{R}$, nous avons $\cos(\alpha)\sin(\beta) = \dfrac{\sin(\alpha+\beta) - \sin(\alpha-\beta)}{2}$.

(c) Pour $n \in \mathbb{N}^*$, on pose $D_n(t) = \frac{1}{2} + \sum_{k=1}^{n} \cos(kt)$. En écrivant $\cos(kt) = \Re e\left(e^{ikt}\right)$ pour tout $k \in [\![1, n]\!]$, et à l'aide des formules d'Euler, vérifier que, pour $t \in]0, \pi]$, on a

$$D_n(t) = \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{2\sin\left(\frac{t}{2}\right)}$$

(d) Calculer, pour tout entier $k \geq 1$,

$$\int_0^{\pi} t\cos(kt)\mathrm{d}t$$

(e) Vérifier alors que pour tout entier $n \geq 1$,

$$\int_0^{\pi} tD_n(t)\mathrm{d}t = \frac{\pi^2}{4} - C_n + W_n$$

(f) Déduire des questions précédentes que $\lim_n C_n = \dfrac{\pi^2}{6}$.

Tournez la page S.V.P.

Modèle

Échantillonnage avec remise dans une population à N espèces

On fixe $N \in \mathbb{N}^*$. On considère l'ensemble $[\![1, N]\!]$ correspondant aux espèces des individus dans la population. On note $\mathcal{D} = \{\vec{q} \in]0,1[^N : \sum_{i=1}^N q_i = 1\}$ l'espace des lois de probabilité dans $[\![1, N]\!]$, avec $q_i > 0, \forall i \in [\![1, N]\!]$. On fixe $\vec{p} = (p_1, \ldots, p_N)^\mathsf{T} \in \mathcal{D}$, le vecteur colonne dont la i-ème coordonnée p_i, est la probabilité d'observer l'espèce i. Dans certaines questions, on précisera que l'on considère le cas où $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$ (soit $p_i = 1/N$ pour tout $i \in [\![1, N]\!]$). Si ce n'est pas précisé, on considère $\vec{p}$ quelconque.

On considère une suite $(X_n)_{n\geq 1}$, de variables aléatoires à valeurs dans $[\![1, N]\!]$, indépendantes et de même loi $\vec{p}$, correspondant aux espèces des individus observés au cours des tirages successifs. Les individus échantillonnés sont remis dans la population. Ainsi, pour $n \geq 1$, et $i \in [\![1, N]\!]$,

$$\mathbb{P}(X_n = i) = p_i. \tag{1}$$

On considère la suite associée $(\overrightarrow{S_n})_{n\geq 1}$, à valeurs dans $\mathbb{N}^N$, dont le n-ième terme est noté $\overrightarrow{S_n} = (S_{n,1}, \ldots, S_{n,N})$, et est défini de la manière suivante, pour $i \in [\![1, N]\!]$,

$$S_{n,i} = \sum_{k=1}^{n} \mathbb{1}\left(X_k = i\right). \tag{2}$$

Le vecteur aléatoire $\overrightarrow{S_n}$ donne le nombre d'individus de chaque espèce échantillonnés au cours de n tirages.

On définit une autre suite de variables aléatoires $(Y_n)_{n\geq 1}$ à partir de $(\overrightarrow{S_n})_{n\geq 1}$ comme suit, pour $n \geq 1$,

$$Y_n = \sum_{i=1}^{N} \mathbb{1}\left(S_{n,i} \neq 0\right). \tag{3}$$

La variable aléatoire Y_n quantifie le nombre d'espèces différentes observées jusqu'au n-ième tirage.

Enfin, on considère les variables aléatoires $(T_k)_{1\leq k\leq N}$, définies à partir de la suite précédente par

$$T_k = \min\{n \geq 1 : Y_n = k\}, \tag{4}$$

qui correspondent au nombre minimal de tirages pour observer k espèces différentes.

Example : échantillonnage des oiseaux dans la forêt

On considère une forêt dans laquelle vivent N espèces d'oiseaux. Un dispositif capture successivement des images de ces oiseaux et identifie leur espèce. Nous allons utiliser le modèle présenté ci-dessus pour étudier plusieurs scénarios, qui seront évoqués par la suite.

– Scénario A : la forêt uniforme avec $N = 4$, où toutes les espèces sont aussi courantes les unes que les autres, soit $p_1 = p_2 = p_3 = p_4 = 0,25$.

– Scénario B : la forêt asymétrique avec $N = 4$, où trois espèces sont courantes et l'une est très rare, avec $p_1 = p_2 = p_3 = 0,33$ et $p_4 = 0,01$.

– Scénario C : la forêt tropicale, où N est connu et très grand.

Partie II

Dans cette partie on va étudier des propriétés des suites $(\overrightarrow{S_n})_{n\geq 1}$, définie par (2), et $(Y_n)_{n\geq 1}$, définie par (3).

1. Déterminer, pour tout $i \in [\![1, N]\!]$, et tout $n \geq 1$, la loi de $S_{n,i}$. Calculer $\mathbb{E}[S_{n,i}]$ et $\mathbb{V}\mathrm{ar}(S_{n,i})$.
2. Pour $n \geq 1$, et pour tout $(i, j) \in [\![1, N]\!]^2$, $i \neq j$, calculer $\mathbb{E}\left[S_{n,i}S_{n,j}\right]$ à l'aide de la définition (2), puis en déduire la valeur de $\mathbb{C}\mathrm{ov}(S_{n,i},\ S_{n,j})$. Les variables $S_{n,i}$ et $S_{n,j}$ sont-elles indépendantes ?

3. Pour $n \geq 1$, on va s'intéresser à la matrice de variance-covariance du vecteur aléatoire $\overrightarrow{S_n}$, que l'on notera $\mathbf{\Sigma}_n$, appartenant à $\mathcal{M}_N(\mathbb{R})$, et définie comme suit

$$\mathbf{\Sigma}_n = \begin{pmatrix} \mathbb{V}\mathrm{ar}\,(S_{n,1}) & \mathbb{C}\mathrm{ov}\,(S_{n,1}, S_{n,2}) & \dots & \mathbb{C}\mathrm{ov}\,(S_{n,1}, S_{n,N}) \\ \mathbb{C}\mathrm{ov}\,(S_{n,2}, S_{n,1}) & \mathbb{V}\mathrm{ar}\,(S_{n,2}) & \dots & \mathbb{C}\mathrm{ov}\,(S_{n,2}, S_{n,N}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}\,(S_{n,N}, S_{n,1}) & \mathbb{C}\mathrm{ov}\,(S_{n,N}, S_{n,2}) & \dots & \mathbb{V}\mathrm{ar}\,(S_{n,N}) \end{pmatrix}.$$

   (a) Donner $\mathbf{\Sigma}_n$ et calculer $\mathbf{\Sigma}_n \overrightarrow{1_N}$. Interpréter le résultat.

   (b) Montrer que pour $k, \ell \in [\![1, n]\!]$, $k \neq \ell$, $\mathbb{P}(X_k \neq X_\ell) = 1 - \vec{p}^{\,\mathsf{T}} \vec{p}$. Cette probabilité est appelée l'*indice de diversité de Gini-Simpson*, que nous allons noter $\mathrm{GS}(\vec{p})$. Montrer, en utilisant l'inégalité de Cauchy-Schwarz, que $\mathrm{GS}(\vec{p})$ atteint son maximum sur $\mathcal{D}$ lorsque $\vec{p}$ est la **loi uniforme**.

   (c) Que représente la trace de $\mathbf{\Sigma}_n$ dans ce modèle ? Vérifier que $\mathrm{Tr}(\mathbf{\Sigma}_n) = n\mathrm{GS}(\vec{p})$. Interpréter ce résultat.

   (d) On suppose pour la suite de cette question que $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$.

      (i) Vérifier que $\mathbf{\Sigma}_n = \frac{n}{N}\mathbf{M}_N$, où $\mathbf{M}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{J}_N$, et $\mathbf{J}_N = \overrightarrow{1_N}\,\overrightarrow{1_N}^{\mathsf{T}}$.

      (ii) Déterminer les valeurs propres de $\mathbf{J}_N$, puis en déduire celles de $\mathbf{M}_N$ et de $\mathbf{\Sigma}_n$.

      (iii) Montrer que la matrice $\mathbf{\Sigma}_n$ est diagonalisable. Déterminer une base de $\mathbb{R}^N$ de vecteurs propres associés à cette matrice.

4. Pour tout entier $n \geq 1$,

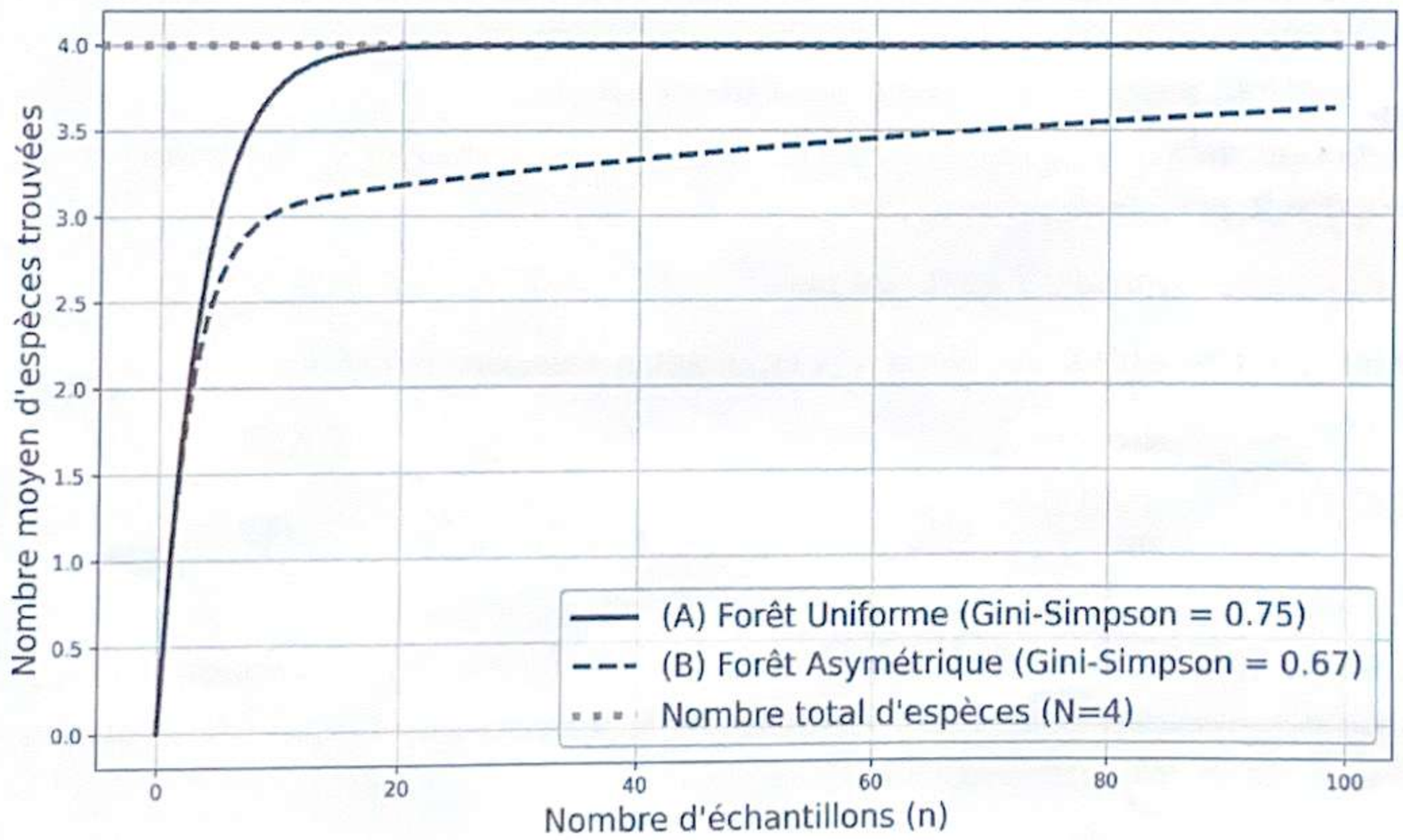
   (a) Montrer que $\mathbb{E}[Y_n] = \sum_{i=1}^{N} (1 - (1 - p_i)^n)$.

   (b) En utilisant le **Théorème 1**,

      (i) Montrer que la fonction $h : x \mapsto -(1-x)^n$ est concave sur $]0, 1[$.

      (ii) Montrer que $\mathbb{E}[Y_n]$ atteint son maximum sur $\mathcal{D}$ lorsque $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$.

5. La figure ci-dessous montre les courbes d'accumulation du nombre moyen d'espèces trouvées en fonction de n (l'effort d'échantillonnage) pour les scénarios A et B de l'introduction. Interpréter ce graphique en vous appuyant sur les résultats des questions **3.(b)-(c)** et **4.(b)**.



6. Si N est inconnu, par quelle variable aléatoire pourrions-nous l'approximer à partir d'un échantillon de taille n ? Comment se comporte cette approximation lorsque $\vec{p}$ est la **loi uniforme** ? Lorsqu'il y a des espèces plus rares que d'autres ?

Tournez la page S.V.P.

Partie III

Dans cette partie, nous allons étudier le comportement de l'espérance $\mathbb{E}[T_N]$ selon la loi de probabilité $\vec{p}$.

1. Dans cette question, nous allons calculer l'espérance de T_N par la méthode du *maximum-minimum*. On définit les variables aléatoires $M_i = \min\{n \geq 1 : X_n = i\}$ pour $i \in [\![1, N]\!]$.

(a) Montrer que $T_N = \max\limits_{1 \leq i \leq N} M_i$.

(b) Montrer que, pour tout $k \in [\![1, N]\!]$, et tout sous-ensemble $J_k = \{j_1, \ldots, j_k\} \subset [\![1, N]\!]$ contenant exactement k éléments, la variable aléatoire $\min\limits_{j \in J_k} M_j$ est de loi géométrique de paramètre $p_{J_k} = \sum_{j \in J_k} p_j$.

(c) Soit $n \in \mathbb{N}^*$ et $(x_i)_{1 \leq i \leq n}$ un sous-ensemble de valeurs dans $\mathbb{R}_+$. Montrer l'identité suivante,

$$\max_{1 \leq i \leq n} x_i = \sum_{1 \leq i \leq n} x_i - \sum_{1 \leq i_1 < i_2 \leq n} \min_{i \in \{i_1, i_2\}} x_i + \sum_{1 \leq i_1 < i_2 < i_3 \leq n} \min_{i \in \{i_1, i_2, i_3\}} x_i + \ldots + (-1)^{n+1} \min_{1 \leq i \leq n} x_i.$$

Suggestion : Montrer d'abord l'identité lorsque $x_i \in [0, 1]$, $1 \leq i \leq n$, en considérant une variable aléatoire U de loi uniforme dans $[0, 1]$ et en utilisant la ***formule du crible****. Généraliser ensuite le résultat pour $x_i \in \mathbb{R}_+$, $1 \leq i \leq n$.*

(d) Montrer que

$$\mathbb{E}[T_N] = \sum_{1 \leq i \leq N} \frac{1}{p_i} - \sum_{1 \leq i_1 < i_2 \leq N} \frac{1}{p_{i_1} + p_{i_2}} + \sum_{1 \leq i_1 < i_2 < i_3 \leq N} \frac{1}{p_{i_1} + p_{i_2} + p_{i_3}} + \cdots + (-1)^{N+1} \frac{1}{p_1 + \cdots + p_N}.$$

(e) Soit $n \in \mathbb{N}^*$ et $(q_i)_{1 \leq i \leq n}$ un sous-ensemble de valeurs dans $]0, 1[$. Montrer l'identité suivante, pour tout $t > 0$,

$$1 - \prod_{i=1}^{n} \left(1 - e^{-q_i t}\right) = \sum_{1 \leq i \leq n} e^{-q_i t} - \sum_{1 \leq i_1 < i_2 \leq n} e^{-(q_{i_1} + q_{i_2})t} + \ldots + (-1)^{n+1} e^{-(q_1 + \ldots + q_n)t}.$$

Suggestion : Considérer une suite $(E_i)_{i=1,\ldots,n}$ de variables aléatoires indépendantes, avec E_i de loi exponentielle de paramètre q_i et utiliser la ***formule du crible****.*

(f) Montrer que

$$\mathbb{E}[T_N] = \int_0^\infty f(\vec{p}, t)\mathrm{d}t, \quad \text{où} \quad f(\vec{p}, t) = f(p_1, \ldots, p_N, t) = \left(1 - \prod_{i=1}^{N} \left(1 - e^{-p_i t}\right)\right).$$

2. Pour cette question, nous aurons besoin du **Théorème 1.**

Soit f la fonction définie à la question **1.(f)**. On cherche à trouver le minimum de f sur l'ensemble $\mathcal{D}$, soit résoudre le problème suivant,

$$\text{trouver } \vec{q} \in \mathcal{D}, \text{ tel que } f(\vec{q}, t) \leq f(\vec{p}, t), \ \forall \vec{p} \in \mathcal{D}, \forall t \geq 0. \tag{*}$$

(a) Montrer que résoudre le problème (*) équivaut à résoudre le problème suivant,

$$\text{trouver } \vec{q} \in \mathcal{D}, \text{ tel que } \widetilde{f}(\vec{q}, t) \geq \widetilde{f}(\vec{p}, t), \ \forall \vec{p} \in \mathcal{D}, \ \forall t > 0, \tag{**}$$

$$\text{où} \quad \widetilde{f}(\vec{p}, t) = \widetilde{f}(p_1, \ldots, p_N, t) = \sum_{i=1}^{N} \ln\left(1 - e^{-p_i t}\right).$$

(b) Pour tout $t > 0$, montrer que la fonction $g : x \mapsto \ln\left(1 - e^{-xt}\right)$ est concave sur $\mathbb{R}_+^*$.

(c) Déduire des résultats précédents, en utilisant le **Théorème 1**, que $\mathbb{E}[T_N]$ atteint son minimum lorsque $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$.

3. Montrer que $\mathbb{E}[T_N] \geq \dfrac{1}{\min\limits_{1 \leq i \leq N} p_i}$. Que se passe-t-il avec $\mathbb{E}[T_N]$ lorsqu'il y a une espèce très rare ?

4. Interprétez les résultats de cette partie et leurs liens avec ceux de la **Partie II** en vous appuyant sur les scénarios A et B de l'introduction, pour lesquels nous avons, respectivement $\mathbb{E}[T_4^A] = 8,33$ et $\mathbb{E}[T_4^B] = 100,22$.

Partie IV

Dans cette partie $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$. Nous allons étudier l'espérance et la variance des variables $(T_k)_{1\leq k\leq N}$, définies dans (4) et le comportement asymptotique de T_N lorsque $N \to \infty$.

1. Soit $G_k = T_k - T_{k-1}$, pour $k \in [\![1, N]\!]$, avec $T_0 = 0$.

(a) Vérifier que pour tout $k \in [\![1, N]\!]$,

$$T_k = G_1 + \ldots + G_k.$$

(b) Nous allons étudier le comportement de $\mathbb{E}[T_k]$ pour $k \in [\![1, N]\!]$.

(i) Expliquer informellement pourquoi pour tout $\ell \in [\![1, N]\!]$, la variable aléatoire G_ℓ suit la loi géométrique de paramètre $\frac{N-\ell+1}{N}$. En déduire que pour tout $k \in [\![1, N]\!]$,

$$\mathbb{E}[T_k] = \sum_{\ell=1}^{k} \frac{N}{N-\ell+1}.$$

(ii) Pour tout k fixé, tel que $k < N$, donner un équivalent de $\mathbb{E}[T_k]$ lorsque $N \to \infty$.

(iii) En utilisant les résultats de la **Partie I.2**, donner un équivalent de $\mathbb{E}[T_N]$ lorsque $N \to \infty$.

(iv) Interpréter les résultats de **(ii)** et **(iii)**.

(c) Soit

$$V_N = \frac{T_N - N \ln N}{N}.$$

Vérifier que $\lim_{N\to\infty} \mathbb{E}[V_N] = \gamma$, où γ est la constante d'Euler-Mascheroni définie dans **Partie I.2.**

(d) On admet que les variables $(G_k)_{1\leq k\leq N}$ sont indépendantes. En déduire que

$$\mathbb{V}\mathrm{ar}\,(T_N) = N^2 C_N - N H_N,$$

où C_N et H_N correspondent aux sommes définies dans la **Partie I.** Vérifier que

$$\lim_{N\to\infty} \mathbb{V}\mathrm{ar}\,(V_N) = \frac{\pi^2}{6}.$$

Par la suite, nous pourrons utiliser l'approximation $\frac{\pi^2}{6} \approx 1,6$.

2. Pour tout $\varepsilon > 0$, montrer que

$$\mathbb{P}\left(|T_N - N H_N| \geq \varepsilon N\right) \leq \frac{\pi^2}{6\varepsilon^2}.$$

3. Utiliser l'encadrement de la question **I.2.(b)**, pour montrer que pour tout réel $c > 1$, on a

$$\mathbb{P}\left(|V_N| \geq c\right) \leq \mathbb{P}\left(\left|\frac{T_N}{N} - H_N\right| \geq (c-1)\right).$$

En utilisant de nouveau l'inégalité de la question **IV.2**, montrer que

$$\mathbb{P}\left(\frac{T_N}{N} \in]\ln N - 5, \ln N + 5[\right) \geq 0.9.$$

4. Interpréter les résultats obtenus dans cette partie en lien avec le scénario C et les **Parties II** et **III.**

Tournez la page S.V.P.

Partie V

Dans cette partie $\vec{p}$ est la **loi uniforme** sur $[\![1, N]\!]$. Ici, nous allons déterminer la loi des suites des variables aléatoires $(Y_n)_{n\geq 1}$, définie par (3), et $(T_k)_{1\leq k\leq N}$, définie par (4).

1. Montrer que, pour tout entier $n \geq 1$,

$$\{T_N > n\} = \bigcup_{i=1}^{N} \{S_{n,i} = 0\}.$$

En déduire, à l'aide de la **formule du crible**, que

$$\mathbb{P}(T_N > n) = \sum_{\ell=1}^{N} (-1)^{\ell+1} \binom{N}{\ell} \left(1 - \frac{\ell}{N}\right)^n.$$

2. Soit $n \in \mathbb{N}^*$, et soit k un entier tel que $1 \leq k \leq N$, on notera $J_k = \{j_1, \ldots, j_k\} \subset [\![1, N]\!]$ les sous-ensembles contenant exactement k éléments. On considère les évènements $A_{J_k}^{(n)}$ et $B_{J_k}^{(n)}$ définis par

$$A_{J_k}^{(n)} = \bigcap_{i=1}^{n} \{X_i \in J_k\}, \quad B_{J_k}^{(n)} = \bigcap_{j\in J_k} \left(\bigcup_{i=1}^{n} \{X_i = j\} \right).$$

(a) Donner une interprétation des évènements $A_{J_k}^{(n)}$ et $B_{J_k}^{(n)}$, en déduire que pour tout $n \geq 1$,

$$\{Y_n = k\} = \bigcup_{J_k \subset [\![1,N]\!]} \left(A_{J_k}^{(n)} \cap B_{J_k}^{(n)} \right),$$

où $\bigcup_{J_k \subset [\![1,N]\!]}$ est la réunion pour tous les sous-ensembles J_k de $[\![1, N]\!]$ contenant k éléments.

(b) Montrer à l'aide de la question **V.1**, que pour tous les sous-ensembles J_k considérés, et tout $n \geq 1$,

$$\mathbb{P}\left(A_{J_k}^{(n)}\right) = \left(\frac{k}{N}\right)^n,$$

$$\mathbb{P}\left(B_{J_k}^{(n)} \,\middle|\, A_{J_k}^{(n)}\right) = 1 - \sum_{\ell=1}^{k} (-1)^{\ell+1} \binom{k}{\ell} \left(1 - \frac{\ell}{k}\right)^n,$$

où $\mathbb{P}(A \mid B)$ est la probabilité conditionnelle de A sachant B, pour tout $A, B \in \mathcal{F}$.

(c) Montrer à partir des questions précédentes que pour tout $n \geq 1$,

$$\mathbb{P}(Y_n = k) = \frac{1}{N^n} \frac{N!}{(N-k)!} \mathcal{S}(n, k),$$

où $\mathcal{S}(n, k)$ correspondent aux *nombres de Stirling de seconde espèce*, donnés par

$$\mathcal{S}(n, k) = \frac{1}{k!} \sum_{\ell=0}^{k} (-1)^{k-\ell} \binom{k}{\ell} \ell^n,$$

et définis comme le nombre de façons de partitionner un ensemble de n objets distincts en k sous-ensembles non vides.

(d) Montrer que pour pour tout $n \geq 1$,

$$\mathbb{P}(T_k = n) = \mathbb{P}(Y_{n-1} = k-1) \frac{N-k+1}{N},$$

et déduire de cette égalité que

$$\mathbb{P}(T_k = n) = \frac{1}{N^n} \frac{N!}{(N-k)!} \mathcal{S}(n-1, k-1),$$

où, par convention, $\mathcal{S}(0, 0) = 1$. Donner une interprétation combinatoire de ce résultat par rapport au modèle et à la définition des nombres de Stirling de seconde espèce.

3. Qu'apportent ces résultats par rapport à ceux obtenus dans les **Parties II - IV** ?

Fin de l'épreuve