Excellence Maths

Correction complète

Sujet ENS 2026

Modèle d'échantillonnage avec remise

Correction rédigée avec une attention particulière portée aux justifications probabilistes, aux calculs d'espérance et de variance, ainsi qu'aux résultats asymptotiques du modèle du collectionneur.

Table des matières

Partie I – Questions préliminaires

Pour tout $n \geq 1$, on rappelle

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

1. Soient $a \geq 1$ et $k \geq 2$. La fonction $t \mapsto t^{-a}$ est décroissante sur $\mathbb{R}_+^*$. Ainsi, pour $t \in [k, k+1]$,

$$\frac{1}{t^a} \leq \frac{1}{k^a},$$

d'où

$$\int_k^{k+1} \frac{\mathrm{d}t}{t^a} \leq \frac{1}{k^a}.$$

De même, pour $t \in [k-1, k]$, on a $t \leq k$, donc $t^{-a} \geq k^{-a}$, et par suite

$$\frac{1}{k^a} \leq \int_{k-1}^{k} \frac{\mathrm{d}t}{t^a}.$$

Finalement,

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

2. On pose, pour $n \geq 1$,

$$u_n = H_n - \ln n.$$

(a) Pour $n \geq 1$,

$$\ln\left(\frac{n+1}{n}\right) = \int_n^{n+1} \frac{\mathrm{d}t}{t}.$$

Comme $t \mapsto 1/t$ est décroissante, pour $t \in [n, n+1]$,

$$\frac{1}{n+1} \leq \frac{1}{t} \leq \frac{1}{n}.$$

En intégrant sur $[n, n+1]$, on obtient

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

(b) Pour $n \geq 1$,

$$u_{n+1} - u_n = H_{n+1} - \ln(n+1) - H_n + \ln n = \frac{1}{n+1} - \ln\left(\frac{n+1}{n}\right) \leq 0$$

d'après la question précédente. La suite (u_n) est donc décroissante.

De plus,

$$u_{n+1} - u_n \geq \frac{1}{n+1} - \frac{1}{n} = -\frac{1}{n(n+1)}.$$

Pour encadrer directement u_n, on utilise la question précédente sous la forme

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

En sommant de $k = 1$ à $n-1$, on obtient

$$\ln n \leq \sum_{k=1}^{n-1} \frac{1}{k} \leq H_n,$$

donc $u_n \geq 0$. D'autre part, pour $k \geq 2$,

$$\frac{1}{k} \leq \ln\left(\frac{k}{k-1}\right).$$

En sommant de $k = 2$ à n, on obtient

$$H_n = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \sum_{k=2}^{n} \ln\left(\frac{k}{k-1}\right) = 1 + \ln n.$$

Ainsi,

$$\boxed{0 \leq u_n \leq 1}.$$

La suite (u_n) est décroissante et minorée par 0 ; elle converge donc. Sa limite est notée

$$\boxed{\gamma = \lim_{n\to+\infty}(H_n - \ln n) \in [0,1]}.$$

(c) Comme $H_n = \ln n + u_n$ et que (u_n) converge vers γ, on a $u_n = O(1)$. Or $\ln n \to +\infty$, donc

$$\frac{H_n}{\ln n} = 1 + \frac{u_n}{\ln n} \longrightarrow 1.$$

Ainsi,

$$\boxed{H_n \sim \ln n \qquad (n \to +\infty)}.$$

3. Soit $a > 1$. Pour $k \geq 2$, la question 1 donne

$$\frac{1}{k^a} \leq \int_{k-1}^{k} \frac{\mathrm{d}t}{t^a}.$$

Donc, pour $n \geq 2$,

$$\sum_{k=1}^{n} \frac{1}{k^a} \leq 1 + \int_1^n \frac{\mathrm{d}t}{t^a} = 1 + \frac{1 - n^{1-a}}{a-1} \leq 1 + \frac{1}{a-1} = \frac{a}{a-1}.$$

Cette inégalité reste vraie pour $n = 1$. Ainsi

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

En particulier, pour $a = 2$, la suite (C_n) est croissante et majorée, donc convergente. Enfin,

$$|W_n - W_m| \leq \sum_{k=m+1}^{n} \frac{1}{k^2} \quad (m < n),$$

et la convergence de $\sum 1/k^2$ entraîne que (W_n) est de Cauchy, donc convergente.

4. On note

$$C = \lim_{n\to+\infty} C_n, \qquad W = \lim_{n\to+\infty} W_n.$$

Étudions $C_n + W_n$:

$$C_n + W_n = \sum_{k=1}^{n} \frac{1 + (-1)^k}{k^2}.$$

Les termes impairs sont nuls, et les termes pairs donnent

$$C_n + W_n = \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{2}{(2j)^2} = \frac{1}{2} \sum_{j=1}^{\lfloor n/2 \rfloor} \frac{1}{j^2} = \frac{1}{2} C_{\lfloor n/2 \rfloor}.$$

En passant à la limite,

$$C + W = \frac{1}{2}C,$$

d'où

$$\boxed{W = -\frac{1}{2}C}.$$

5. On cherche maintenant à calculer C.

(a) Soit $f \in \mathcal{C}^1([0,\pi],\mathbb{R})$. Par intégration par parties,

$$\int_0^\pi f(t)\sin(\lambda t)\,\mathrm{d}t = \left[-\frac{f(t)\cos(\lambda t)}{\lambda}\right]_0^\pi + \frac{1}{\lambda}\int_0^\pi f'(t)\cos(\lambda t)\,\mathrm{d}t.$$

Le membre de droite est majoré en valeur absolue par

$$\frac{|f(\pi)| + |f(0)|}{\lambda} + \frac{\pi\|f'\|_\infty}{\lambda},$$

qui tend vers 0 lorsque $\lambda \to +\infty$. Ainsi

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

(b) Pour tous $\alpha, \beta \in \mathbb{R}$,

$$\sin(\alpha+\beta) - \sin(\alpha-\beta) = 2\cos\alpha\sin\beta.$$

Donc

$$\boxed{\cos\alpha\sin\beta = \frac{\sin(\alpha+\beta) - \sin(\alpha-\beta)}{2}}.$$

(c) Pour $n \geq 1$, on pose

$$D_n(t) = \frac{1}{2} + \sum_{k=1}^n \cos(kt).$$

On utilise

$$1 + 2\sum_{k=1}^n \cos(kt) = \sum_{k=-n}^n e^{ikt}.$$

Pour $t \in]0,\pi]$,

$$\sum_{k=-n}^n e^{ikt} = e^{-int}\frac{1 - e^{i(2n+1)t}}{1 - e^{it}}.$$

En réécrivant le quotient sous forme trigonométrique, on obtient

$$\sum_{k=-n}^n e^{ikt} = \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{\sin\left(\frac{t}{2}\right)}.$$

Ainsi

$$\boxed{D_n(t) = \frac{\sin\left(\frac{(2n+1)t}{2}\right)}{2\sin\left(\frac{t}{2}\right)} \qquad (t \in]0,\pi]}.$$

(d) Pour $k \geq 1$, une intégration par parties donne

$$\int_0^\pi t\cos(kt)\,\mathrm{d}t = \left[\frac{t\sin(kt)}{k}\right]_0^\pi - \frac{1}{k}\int_0^\pi \sin(kt)\,\mathrm{d}t.$$

Le terme de bord est nul, donc

$$\int_0^\pi t\cos(kt)\,\mathrm{d}t = -\frac{1}{k}\left[-\frac{\cos(kt)}{k}\right]_0^\pi = \frac{(-1)^k-1}{k^2}.$$

Ainsi

$$\boxed{\int_0^\pi t\cos(kt)\,\mathrm{d}t = \frac{(-1)^k-1}{k^2}}.$$

(e) On a

$$\int_0^\pi tD_n(t)\,\mathrm{d}t = \frac{1}{2}\int_0^\pi t\,\mathrm{d}t + \sum_{k=1}^n \int_0^\pi t\cos(kt)\,\mathrm{d}t.$$

D'après la question précédente,

$$\int_0^\pi tD_n(t)\,\mathrm{d}t = \frac{\pi^2}{4} + \sum_{k=1}^n \frac{(-1)^k-1}{k^2} = \frac{\pi^2}{4} + W_n - C_n.$$

Donc

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

(f) D'après l'expression de D_n,

$$\int_0^\pi tD_n(t)\,\mathrm{d}t = \int_0^\pi \frac{t}{2\sin(t/2)}\sin\left(\frac{(2n+1)t}{2}\right)\,\mathrm{d}t.$$

La fonction

$$f : t \mapsto \frac{t}{2\sin(t/2)}$$

se prolonge en une fonction de classe $\mathcal{C}^1$ sur $[0,\pi]$, avec $f(0)=1$. La question 5(a), appliquée à $\lambda = n+1/2$, donne

$$\int_0^\pi tD_n(t)\,\mathrm{d}t \longrightarrow 0.$$

En passant à la limite dans l'identité de 5(e),

$$0 = \frac{\pi^2}{4} - C + W.$$

Or $W = -C/2$, donc

$$0 = \frac{\pi^2}{4} - \frac{3}{2}C.$$

Ainsi

$$\boxed{C = \sum_{k=1}^{+\infty}\frac{1}{k^2} = \frac{\pi^2}{6}}.$$

Partie II – Premières propriétés du modèle

On fixe $N \in \mathbb{N}^*$ et une loi $\overrightarrow{p} = (p_1, \dots, p_N) \in D$, avec $p_i > 0$ et $\sum_{i=1}^N p_i = 1$. Les variables $(X_n)_{n\geq 1}$ sont indépendantes, de même loi, et

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

On rappelle

$$S_{n,i} = \sum_{k=1}^{n} \mathbf{1}_{\{X_k=i\}}, \qquad Y_n = \sum_{i=1}^{N} \mathbf{1}_{\{S_{n,i}\neq 0\}}.$$

1. Pour $i \in [1, N]$ fixé, les variables $\mathbf{1}_{\{X_k=i\}}$ sont indépendantes et suivent une loi de Bernoulli de paramètre p_i. Donc

$$\boxed{S_{n,i} \sim \mathcal{B}(n, p_i)}.$$

Ainsi

$$\boxed{\mathbb{E}[S_{n,i}] = np_i}, \qquad \boxed{\mathrm{Var}(S_{n,i}) = np_i(1-p_i)}.$$

2. Soient $i \neq j$. On écrit

$$S_{n,i}S_{n,j} = \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathbf{1}_{\{X_k=i\}}\mathbf{1}_{\{X_\ell=j\}}.$$

Si $k = \ell$, le produit est toujours nul car $i \neq j$. Si $k \neq \ell$, les variables X_k et X_ℓ sont indépendantes, donc

$$\mathbb{E}\left[\mathbf{1}_{\{X_k=i\}}\mathbf{1}_{\{X_\ell=j\}}\right] = p_ip_j.$$

Il y a $n(n-1)$ couples (k, ℓ) tels que $k \neq \ell$. Par conséquent

$$\boxed{\mathbb{E}[S_{n,i}S_{n,j}] = n(n-1)p_ip_j}.$$

La covariance vaut alors

$$\mathrm{Cov}(S_{n,i}, S_{n,j}) = n(n-1)p_ip_j - np_i\, np_j = -np_ip_j.$$

Ainsi

$$\boxed{\mathrm{Cov}(S_{n,i}, S_{n,j}) = -np_ip_j < 0}.$$

En particulier, $S_{n,i}$ et $S_{n,j}$ ne sont pas indépendantes : leurs covariances ne sont pas nulles.

3. On considère la matrice de variance-covariance Σ_n de $\overrightarrow{S}_n = (S_{n,1}, \dots, S_{n,N})$.

(a) D'après les questions précédentes,

$$(\Sigma_n)_{ij} = \begin{cases} np_i(1-p_i), & i = j, \\ -np_ip_j, & i \neq j. \end{cases}$$

En notant $\overrightarrow{p}$ le vecteur colonne des probabilités et $D_p = \mathrm{Diag}(p_1, \dots, p_N)$,

$$\boxed{\Sigma_n = n\left(D_p - \overrightarrow{p}\,\overrightarrow{p}^{\top}\right)}.$$

On calcule

$$\Sigma_n\mathbf{1}_N = n\left(D_p\mathbf{1}_N - \overrightarrow{p}\,\overrightarrow{p}^{\top}\mathbf{1}_N\right) = n\left(\overrightarrow{p} - \overrightarrow{p}\sum_{i=1}^{N} p_i\right) = 0.$$

Donc

$$\boxed{\Sigma_n\mathbf{1}_N = 0}.$$

Cette identité traduit le fait que

$$S_{n,1} + \cdots + S_{n,N} = n$$

est déterministe : les fluctuations des effectifs des différentes espèces sont contraintes par une somme totale fixée.

(b) Pour $k \neq \ell$,

$$\mathbb{P}(X_k = X_\ell) = \sum_{i=1}^{N} \mathbb{P}(X_k = i, X_\ell = i) = \sum_{i=1}^{N} p_i^2,$$

donc

$$\boxed{\mathbb{P}(X_k \neq X_\ell) = 1 - \overrightarrow{p}^{\top} \overrightarrow{p} = 1 - \sum_{i=1}^{N} p_i^2}.$$

On note cette quantité

$$\boxed{\mathrm{GS}(\overrightarrow{p}) = 1 - \sum_{i=1}^{N} p_i^2}.$$

Par Cauchy-Schwarz,

$$\left(\sum_{i=1}^{N} p_i\right)^2 \leq N \sum_{i=1}^{N} p_i^2.$$

Comme $\sum_i p_i = 1$,

$$\sum_{i=1}^{N} p_i^2 \geq \frac{1}{N}.$$

Par conséquent

$$\boxed{\mathrm{GS}(\overrightarrow{p}) \leq 1 - \frac{1}{N}}.$$

Le cas d'égalité dans Cauchy-Schwarz impose $p_1 = \cdots = p_N = 1/N$. Donc $\mathrm{GS}(\overrightarrow{p})$ est maximal exactement pour la loi uniforme.

(c) La trace de Σ_n vaut

$$\mathrm{Tr}(\Sigma_n) = \sum_{i=1}^{N} \mathrm{Var}(S_{n,i}) = n \sum_{i=1}^{N} p_i(1 - p_i) = n\left(1 - \sum_{i=1}^{N} p_i^2\right).$$

Ainsi

$$\boxed{\mathrm{Tr}(\Sigma_n) = n\,\mathrm{GS}(\overrightarrow{p})}.$$

La trace mesure la somme des variances des nombres d'individus observés dans chaque espèce. Elle est donc proportionnelle à l'indice de diversité de Gini-Simpson : plus la population est équilibrée, plus la diversité observée dans deux tirages indépendants est grande, et plus la dispersion globale du vecteur des effectifs est élevée.

(d) On suppose ici que $\overrightarrow{p}$ est la loi uniforme sur $[1, N]$.

(i) Dans ce cas,

$$D_p = \frac{1}{N} I_N, \qquad \overrightarrow{p}\,\overrightarrow{p}^{\top} = \frac{1}{N^2} J_N,$$

où $J_N = \mathbf{1}_N \mathbf{1}_N^{\top}$. Donc

$$\boxed{\Sigma_n = \frac{n}{N}\left(I_N - \frac{1}{N} J_N\right)}.$$

En posant

$$M_N = I_N - \frac{1}{N} J_N,$$

on obtient bien

$$\boxed{\Sigma_n = \frac{n}{N} M_N}.$$

(ii) Pour tout $x \in \mathbb{R}^N$,

$$J_N x = (x_1 + \cdots + x_N)\mathbf{1}_N.$$

Ainsi, si $x \in \mathrm{Vect}(\mathbf{1}_N)$, alors $J_N x = Nx$. Si $x \in \mathbf{1}_N^{\perp}$, c'est-à-dire si $x_1 + \cdots + x_N = 0$, alors $J_N x = 0$.

Les valeurs propres de J_N sont donc

$$\boxed{N \text{ avec multiplicité } 1, \qquad 0 \text{ avec multiplicité } N-1}.$$

On en déduit que les valeurs propres de $M_N = I_N - \frac{1}{N} J_N$ sont

$$\boxed{0 \text{ sur } \mathrm{Vect}(\mathbf{1}_N), \qquad 1 \text{ sur } \mathbf{1}_N^{\perp}}.$$

Enfin, celles de $\Sigma_n = \frac{n}{N} M_N$ sont

$$\boxed{0 \text{ avec multiplicité } 1, \qquad \frac{n}{N} \text{ avec multiplicité } N-1}.$$

(iii) La matrice Σ_n est symétrique réelle, donc diagonalisable dans une base orthonormée. Plus explicitement, une base de vecteurs propres est donnée par

$$\mathbf{1}_N$$

associé à la valeur propre 0, complété par une base de l'hyperplan

$$\mathbf{1}_N^{\perp} = \{x \in \mathbb{R}^N : x_1 + \cdots + x_N = 0\},$$

par exemple

$$e_1 - e_N,\ e_2 - e_N,\ \ldots,\ e_{N-1} - e_N,$$

associés à la valeur propre n/N.

4. Étude de $\mathbb{E}[Y_n]$.

(a) Par linéarité de l'espérance,

$$\mathbb{E}[Y_n] = \sum_{i=1}^{N} \mathbb{P}(S_{n,i} \neq 0).$$

Or l'espèce i n'est jamais observée en n tirages avec probabilité $(1-p_i)^n$. Donc

$$\mathbb{P}(S_{n,i} \neq 0) = 1 - (1-p_i)^n.$$

Ainsi

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

(b) Posons

$$h(x) = 1 - (1-x)^n, \qquad x \in [0,1].$$

Pour $n \geq 2$,

$$h''(x) = -n(n-1)(1-x)^{n-2} \leq 0,$$

donc h est concave sur $[0,1]$. Pour $n = 1$, $h(x) = x$ est affine, donc concave également.

Par l'inégalité de Jensen,

$$\frac{1}{N} \sum_{i=1}^{N} h(p_i) \leq h\left(\frac{1}{N} \sum_{i=1}^{N} p_i\right) = h\left(\frac{1}{N}\right).$$

Donc

$$\mathbb{E}[Y_n] \leq N\left(1-\left(1-\frac{1}{N}\right)^n\right).$$

La valeur maximale est atteinte pour la loi uniforme. Pour $n \geq 2$, la concavité est stricte sur $[0,1[$, donc l'égalité impose $p_1 = \cdots = p_N = 1/N$; pour $n = 1$, toutes les lois donnent $\mathbb{E}[Y_1] = 1$.

Ainsi

$$\boxed{\text{Pour } n \geq 2,\ \mathbb{E}[Y_n] \text{ est maximale pour la loi uniforme.}}$$

5. Le graphique compare deux situations avec $N = 4$.

Dans le scénario A, la loi est uniforme : les quatre espèces ont même probabilité d'apparition. L'indice de Gini-Simpson vaut

$$\mathrm{GS}_A = 1 - 4\left(\frac{1}{4}\right)^2 = \frac{3}{4} = 0{,}75,$$

qui est la valeur maximale possible pour $N = 4$. La question 4(b) montre alors que $\mathbb{E}[Y_n]$ est maximal : le nombre moyen d'espèces observées croît rapidement vers 4.

Dans le scénario B, trois espèces sont courantes et une espèce est rare, avec par exemple $p_4 = 0{,}01$. Alors

$$\mathrm{GS}_B \simeq 0{,}67 < 0{,}75.$$

Le nombre moyen d'espèces observées augmente vite au début, car les trois espèces courantes sont rapidement vues, puis la courbe se tasse : la quatrième espèce peut rester longtemps invisible. Cette lecture est cohérente avec la trace de la matrice de covariance, puisque

$$\mathrm{Tr}(\Sigma_n) = n\,\mathrm{GS}(\overrightarrow{p}),$$

et avec le fait que la loi uniforme maximise l'espérance du nombre d'espèces différentes observées.

6. Si N est inconnu, une approximation naturelle à partir d'un échantillon de taille n est

$$\boxed{Y_n,\ \text{le nombre d'espèces distinctes observées.}}$$

On a toujours $Y_n \leq N$, donc cette approximation est un minorant observé de N.

Lorsque la loi est uniforme et que n est suffisamment grand devant N, la question 4(b) montre que $\mathbb{E}[Y_n]$ se rapproche rapidement de N : l'estimation par Y_n devient alors raisonnable.

En revanche, s'il existe des espèces rares, l'approximation peut être très mauvaise : certaines espèces peuvent ne pas apparaître dans l'échantillon, même si la taille n est importante. Dans ce cas, Y_n sous-estime fortement N.

Partie III – Espérance du temps d'observation de toutes les espèces

On définit, pour $i \in [1, N]$,

$$M_i = \min\{n \geq 1 : X_n = i\},$$

le temps de première apparition de l'espèce i.

1. Calcul de $\mathbb{E}[T_N]$ par la méthode du maximum-minimum.

(a) L'instant T_N est le premier instant auquel toutes les espèces ont été observées. Cela se produit lorsque la dernière des premières apparitions a eu lieu. Donc

$$\boxed{T_N = \max_{1 \leq i \leq N} M_i}.$$

(b) Soit $J_k = \{j_1, \ldots, j_k\} \subset [1, N]$ un sous-ensemble contenant exactement k éléments. La variable

$$\min_{j \in J_k} M_j$$

est le premier instant auquel une espèce appartenant à J_k est observée.

À chaque tirage, la probabilité d'obtenir une espèce appartenant à J_k vaut

$$p_{J_k} = \sum_{j \in J_k} p_j.$$

Ainsi

$$\boxed{\min_{j \in J_k} M_j \sim \mathcal{G}(p_{J_k})},$$

où la loi géométrique est portée par $\mathbb{N}^*$. En particulier,

$$\boxed{\mathbb{E}\left[\min_{j \in J_k} M_j\right] = \frac{1}{p_{J_k}}}.$$

(c) Soient $x_1, \ldots, x_n \in \mathbb{R}_+$. Pour $t \geq 0$, on applique la formule du crible aux événements

$$A_i(t) = \{x_i \geq t\}.$$

Alors

$$\mathbf{1}_{\{\max_i x_i \geq t\}} = \sum_i \mathbf{1}_{A_i(t)} - \sum_{i_1 < i_2} \mathbf{1}_{A_{i_1}(t) \cap A_{i_2}(t)} + \cdots + (-1)^{n+1} \mathbf{1}_{A_1(t) \cap \cdots \cap A_n(t)}.$$

En intégrant de 0 à $+\infty$, on obtient

$$\max_{1 \leq i \leq n} x_i = \sum_i x_i - \sum_{i_1 < i_2} \min(x_{i_1}, x_{i_2}) + \cdots + (-1)^{n+1} \min_{1 \leq i \leq n} x_i.$$

Donc

$$\boxed{\max_{1 \leq i \leq n} x_i = \sum_{i=1}^{n} x_i - \sum_{1 \leq i_1 < i_2 \leq n} \min(x_{i_1}, x_{i_2}) + \cdots + (-1)^{n+1} \min_{1 \leq i \leq n} x_i.}$$

(d) En appliquant l'identité précédente à $M_1, \ldots, M_N$, puis en prenant l'espérance, on obtient

$$\mathbb{E}[T_N] = \sum_{i=1}^{N} \mathbb{E}[M_i] - \sum_{1 \leq i_1 < i_2 \leq N} \mathbb{E}[\min(M_{i_1}, M_{i_2})] + \cdots + (-1)^{N+1} \mathbb{E}[\min(M_1, \ldots, M_N)].$$

D'après 1(b),

$$\mathbb{E}[M_i] = \frac{1}{p_i}, \qquad \mathbb{E}[\min(M_{i_1}, \dots, M_{i_r})] = \frac{1}{p_{i_1} + \cdots + p_{i_r}}.$$

Ainsi

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

Comme $p_1 + \cdots + p_N = 1$, le dernier terme vaut $(-1)^{N+1}$.

(e) Pour $q_1, \dots, q_n \in]0,1[$ et $t > 0$, on développe le produit :

$$\prod_{i=1}^{n} (1 - e^{-q_i t}) = 1 - \sum_i e^{-q_i t} + \sum_{i<j} e^{-(q_i+q_j)t} - \cdots + (-1)^n e^{-(q_1+\cdots+q_n)t}.$$

Donc

$$\boxed{1 - \prod_{i=1}^{n} (1 - e^{-q_i t}) = \sum_i e^{-q_i t} - \sum_{i<j} e^{-(q_i+q_j)t} + \cdots + (-1)^{n+1} e^{-(q_1+\cdots+q_n)t}.}$$

(f) En appliquant l'identité précédente avec $q_i = p_i$ et en intégrant terme à terme sur $[0, +\infty[$, on obtient

$$\int_0^{+\infty} \left(1 - \prod_{i=1}^{N} (1 - e^{-p_i t})\right) \mathrm{d}t = \sum_i \frac{1}{p_i} - \sum_{i<j} \frac{1}{p_i + p_j} + \cdots + (-1)^{N+1} \frac{1}{p_1 + \cdots + p_N}.$$

Le membre de droite est exactement la formule obtenue en 1(d). Donc

$$\boxed{\mathbb{E}[T_N] = \int_0^{+\infty} f(\overrightarrow{p}, t)\, \mathrm{d}t},$$

où

$$\boxed{f(\overrightarrow{p}, t) = 1 - \prod_{i=1}^{N} (1 - e^{-p_i t})}.$$

2. On cherche à minimiser $\mathbb{E}[T_N]$ sur D.

(a) Pour $t > 0$, on a

$$f(\overrightarrow{p}, t) = 1 - \exp\left(\sum_{i=1}^{N} \ln(1 - e^{-p_i t})\right).$$

On pose

$$\widetilde{f}(\overrightarrow{p}, t) = \sum_{i=1}^{N} \ln(1 - e^{-p_i t}).$$

Comme la fonction exponentielle est croissante,

$$f(\overrightarrow{q}, t) \le f(\overrightarrow{p}, t) \iff \exp(\widetilde{f}(\overrightarrow{q}, t)) \ge \exp(\widetilde{f}(\overrightarrow{p}, t)) \iff \widetilde{f}(\overrightarrow{q}, t) \ge \widetilde{f}(\overrightarrow{p}, t).$$

Ainsi minimiser $f(\overrightarrow{p}, t)$ revient à maximiser $\widetilde{f}(\overrightarrow{p}, t)$.

(b) Pour $t > 0$, soit

$$g(x) = \ln(1 - e^{-xt}), \qquad x > 0.$$

Alors

$$g'(x) = \frac{t}{e^{xt} - 1},$$

puis

$$g''(x) = -\frac{t^2 e^{xt}}{(e^{xt} - 1)^2} < 0.$$

Donc

$$\boxed{g : x \mapsto \ln(1 - e^{-xt}) \text{ est concave sur } \mathbb{R}_+^*.}$$

(c) Par Jensen, pour tout $\overrightarrow{p} \in D$,

$$\frac{1}{N}\sum_{i=1}^{N} g(p_i) \leq g\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) = g\left(\frac{1}{N}\right).$$

Donc, pour tout $t > 0$,

$$\widetilde{f}(\overrightarrow{p}, t) \leq N \ln\left(1 - e^{-t/N}\right).$$

Le membre de droite est $\widetilde{f}(\overrightarrow{u}, t)$, où

$$\overrightarrow{u} = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right).$$

Ainsi $\widetilde{f}(\overrightarrow{p}, t)$ est maximal pour la loi uniforme, donc $f(\overrightarrow{p}, t)$ est minimal pour la loi uniforme. En intégrant sur $t \in [0, +\infty[$, on obtient

$$\boxed{\mathbb{E}[T_N] \text{ atteint son minimum lorsque } \overrightarrow{p} \text{ est la loi uniforme sur } [1, N].}$$

3. Soit $p_{\min} = \min_{1 \leq i \leq N} p_i$. Prenons i_0 tel que $p_{i_0} = p_{\min}$. Comme $T_N \geq M_{i_0}$,

$$\mathbb{E}[T_N] \geq \mathbb{E}[M_{i_0}] = \frac{1}{p_{\min}}.$$

Si $N \geq 2$, l'inégalité est stricte. En effet, avec probabilité strictement positive, l'espèce i_0 apparaît avant qu'au moins une autre espèce n'ait été observée, et alors $T_N > M_{i_0}$. Donc

$$\boxed{\mathbb{E}[T_N] > \frac{1}{\min_i p_i} \qquad (N \geq 2).}$$

Ainsi, lorsqu'une espèce devient très rare, c'est-à-dire lorsque $p_{\min} \to 0$, on a nécessairement

$$\mathbb{E}[T_N] \to +\infty.$$

Le temps moyen nécessaire pour observer toutes les espèces est alors dominé par l'attente de l'espèce rare.

4. La Partie III complète l'interprétation de la Partie II. Dans le scénario A, la loi uniforme minimise $\mathbb{E}[T_N]$: il est donc normal d'obtenir un temps moyen relativement faible pour observer les quatre espèces, ici environ

$$\mathbb{E}[T_4^A] = 8{,}33.$$

Dans le scénario B, une espèce a une probabilité très faible. La question 3 montre que $\mathbb{E}[T_N]$ est alors au moins de l'ordre de l'inverse de cette petite probabilité. Avec une espèce de probabilité 0,01, on s'attend à une durée moyenne de l'ordre de 100, ce qui correspond à la valeur annoncée

$$\mathbb{E}[T_4^B] = 100{,}22.$$

Les courbes de la Partie II indiquaient déjà que le nombre moyen d'espèces observées stagne longtemps autour de 3 dans le scénario B ; la Partie III explique ce phénomène par le temps d'attente très élevé de l'espèce rare.

Partie IV – Cas uniforme : espérance, variance et concentration

Dans cette partie, la loi $\overrightarrow{p}$ est uniforme sur $[1, N]$. On étudie les variables T_k, et en particulier T_N lorsque $N \to +\infty$.

1. On pose

$$G_k = T_k - T_{k-1}, \qquad k \in [1, N], \qquad T_0 = 0.$$

(a) Par sommation télescopique,

$$G_1 + \cdots + G_k = (T_1 - T_0) + (T_2 - T_1) + \cdots + (T_k - T_{k-1}) = T_k.$$

Donc

$$\boxed{T_k = G_1 + \cdots + G_k}.$$

(b) Étude de $\mathbb{E}[T_k]$.

(i) Supposons que $\ell - 1$ espèces différentes aient déjà été observées. Il reste alors

$$N - (\ell - 1) = N - \ell + 1$$

espèces non observées. Comme la loi est uniforme, la probabilité que le prochain tirage donne une nouvelle espèce est

$$\frac{N - \ell + 1}{N}.$$

Ainsi, conditionnellement au passé, le temps d'attente G_ℓ avant la prochaine nouvelle espèce suit une loi géométrique de paramètre

$$\frac{N - \ell + 1}{N}.$$

Donc

$$\boxed{\mathbb{E}[G_\ell] = \frac{N}{N - \ell + 1}}.$$

Par linéarité de l'espérance,

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

(ii) Si k est fixé et $k < N$, alors pour chaque $\ell \in [1, k]$,

$$\frac{N}{N - \ell + 1} \longrightarrow 1 \qquad (N \to +\infty).$$

Par conséquent

$$\mathbb{E}[T_k] \longrightarrow k.$$

Comme k est fixé,

$$\boxed{\mathbb{E}[T_k] \sim k \qquad (N \to +\infty)}.$$

(iii) Pour $k = N$,

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

En posant $j = N - \ell + 1$, on obtient

$$\mathbb{E}[T_N] = N \sum_{j=1}^{N} \frac{1}{j} = NH_N.$$

D'après la Partie I,

$$H_N \sim \ln N.$$

Donc

$$\boxed{\mathbb{E}[T_N] = NH_N \sim N \ln N}.$$

(iv) Pour un nombre fixé k d'espèces à observer, lorsque N est grand, les collisions sont rares au début : presque chaque tirage donne une nouvelle espèce, donc $\mathbb{E}[T_k] \sim k$.

En revanche, observer toutes les espèces est beaucoup plus long. À la fin, il reste peu d'espèces non vues, donc la probabilité de tirer une nouvelle espèce devient faible. C'est le phénomène classique du collectionneur :

$$\mathbb{E}[T_N] \sim N \ln N.$$

(c) On pose

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

Comme $\mathbb{E}[T_N] = NH_N$,

$$\mathbb{E}[V_N] = \frac{NH_N - N \ln N}{N} = H_N - \ln N.$$

D'après la Partie I,

$$H_N - \ln N \longrightarrow \gamma.$$

Donc

$$\boxed{\lim_{N \to +\infty} \mathbb{E}[V_N] = \gamma}.$$

(d) On admet que les variables $G_1, \ldots, G_N$ sont indépendantes. Pour une loi géométrique de paramètre p sur $\mathbb{N}^*$,

$$\mathrm{Var}(G) = \frac{1-p}{p^2}.$$

Ici, pour $j = N - \ell + 1$,

$$p_\ell = \frac{j}{N}.$$

Donc

$$\mathrm{Var}(G_\ell) = \frac{1 - j/N}{(j/N)^2} = \frac{N-j}{N} \cdot \frac{N^2}{j^2} = \frac{N^2}{j^2} - \frac{N}{j}.$$

Par indépendance,

$$\mathrm{Var}(T_N) = \sum_{\ell=1}^{N} \mathrm{Var}(G_\ell) = \sum_{j=1}^{N} \left(\frac{N^2}{j^2} - \frac{N}{j} \right).$$

Ainsi

$$\boxed{\mathrm{Var}(T_N) = N^2 C_N - NH_N}.$$

Puis

$$\mathrm{Var}(V_N) = \frac{1}{N^2} \mathrm{Var}(T_N) = C_N - \frac{H_N}{N}.$$

Or $C_N \to \pi^2/6$ et $H_N/N \to 0$. Donc

$$\boxed{\lim_{N \to +\infty} \mathrm{Var}(V_N) = \frac{\pi^2}{6}}.$$

2. Pour $\varepsilon > 0$, l'inégalité de Bienaymé-Tchebychev donne

$$\mathbb{P}\left(|T_N - \mathbb{E}[T_N]| \geq \varepsilon N\right) \leq \frac{\mathrm{Var}(T_N)}{\varepsilon^2 N^2}.$$

Comme $\mathbb{E}[T_N] = NH_N$,

$$\mathbb{P}\left(|T_N - NH_N| \geq \varepsilon N\right) \leq \frac{N^2 C_N - NH_N}{\varepsilon^2 N^2} = \frac{C_N - H_N/N}{\varepsilon^2}.$$

Or $C_N \leq C = \pi^2/6$ et $H_N/N \geq 0$. Donc

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

3. D'après la Partie I,

$$0 \leq H_N - \ln N \leq 1.$$

On écrit

$$V_N = \frac{T_N - N\ln N}{N} = \left(\frac{T_N}{N} - H_N\right) + (H_N - \ln N).$$

Si $|V_N| \geq c$ avec $c > 1$, alors, puisque $|H_N - \ln N| \leq 1$,

$$\left|\frac{T_N}{N} - H_N\right| \geq c - 1.$$

Donc

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

En appliquant la question 2 avec $\varepsilon = c - 1$, on obtient

$$\mathbb{P}(|V_N| \geq c) \leq \frac{\pi^2}{6(c-1)^2}.$$

Avec $c = 5$ et l'approximation $\pi^2/6 \simeq 1{,}6$,

$$\mathbb{P}(|V_N| \geq 5) \leq \frac{1{,}6}{16} = 0{,}1.$$

Donc

$$\mathbb{P}(|V_N| < 5) \geq 0{,}9.$$

Or $|V_N| < 5$ équivaut à

$$\ln N - 5 < \frac{T_N}{N} < \ln N + 5.$$

Finalement

$$\boxed{\mathbb{P}\left(\frac{T_N}{N} \in \left]\ln N - 5, \ln N + 5\right[\right) \geq 0{,}9.}$$

4. Dans le scénario C, N est très grand et la loi est uniforme. Les résultats précédents indiquent que le temps nécessaire pour observer toutes les espèces est typiquement de l'ordre de

$$N\ln N.$$

Plus précisément, T_N/N est centré autour de $\ln N$, avec des fluctuations qui restent d'ordre constant d'après la variance limite de V_N.

Cela complète la Partie II : pour n de l'ordre de N, on observe une proportion significative des espèces ; mais pour observer presque toutes, et en particulier toutes les espèces, il faut un effort beaucoup plus important, de l'ordre de $N\ln N$. La Partie III montrait par ailleurs que ce cas uniforme est le plus favorable pour minimiser le temps moyen d'observation complète.

Partie V – Lois exactes dans le cas uniforme

Dans cette partie, la loi $\overrightarrow{p}$ est uniforme sur $[1, N]$.

1. On a

$$T_N > n$$

si et seulement si, après n tirages, toutes les espèces n'ont pas été observées. Cela signifie qu'il existe au moins une espèce i telle que $S_{n,i} = 0$. Donc

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

Par la formule du crible,

$$\mathbb{P}(T_N > n) = \sum_{\ell=1}^{N} (-1)^{\ell+1} \sum_{1 \leq i_1 < \cdots < i_\ell \leq N} \mathbb{P}(S_{n,i_1} = \cdots = S_{n,i_\ell} = 0).$$

Pour un sous-ensemble fixé de ℓ espèces, la probabilité qu'aucune de ces espèces ne soit tirée en n tirages est

$$\left(\frac{N-\ell}{N}\right)^n = \left(1 - \frac{\ell}{N}\right)^n.$$

Il existe $\binom{N}{\ell}$ tels sous-ensembles. Donc

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

2. Soient $n \in \mathbb{N}^*$ et $1 \leq k \leq N$. Pour $J_k = \{j_1, \ldots, j_k\} \subset [1, N]$, on définit

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

(a) L'événement $A_{J_k}^{(n)}$ signifie que les n tirages appartiennent tous à l'ensemble J_k : aucune espèce extérieure à J_k n'est observée.

L'événement $B_{J_k}^{(n)}$ signifie que chacune des k espèces de J_k apparaît au moins une fois dans les n tirages.

Ainsi $A_{J_k}^{(n)} \cap B_{J_k}^{(n)}$ signifie que l'ensemble exact des espèces observées est J_k. Donc

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

Cette union est disjointe, car l'ensemble exact des espèces observées est unique.

(b) Pour un sous-ensemble fixé J_k,

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

Conditionnellement à $A_{J_k}^{(n)}$, les n tirages sont indépendants et uniformes sur l'ensemble J_k, qui contient k espèces. La probabilité que toutes les espèces de J_k soient vues au moins une fois vaut donc, d'après la question V.1 appliquée avec $N = k$,

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

Ainsi

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

(c) Comme l'union de 2(a) est disjointe,

$$\mathbb{P}(Y_n = k) = \binom{N}{k} \mathbb{P}(A_{J_k}^{(n)} \cap B_{J_k}^{(n)}).$$

D'après 2(b),

$$\mathbb{P}(A_{J_k}^{(n)} \cap B_{J_k}^{(n)}) = \left(\frac{k}{N}\right)^n \mathbb{P}(B_{J_k}^{(n)} \mid A_{J_k}^{(n)}).$$

La probabilité conditionnelle d'observer les k espèces d'un ensemble de taille k en n tirages uniformes est

$$\frac{k!}{k^n} S(n,k),$$

où $S(n,k)$ désigne le nombre de Stirling de seconde espèce. En effet, on partitionne les n tirages en k classes non vides correspondant aux espèces observées, puis on affecte les k classes aux k espèces.

Ainsi

$$\mathbb{P}(A_{J_k}^{(n)} \cap B_{J_k}^{(n)}) = \left(\frac{k}{N}\right)^n \frac{k!}{k^n} S(n,k) = \frac{k!}{N^n} S(n,k).$$

Donc

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

Finalement

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

La formule explicite des nombres de Stirling de seconde espèce est

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

Elle est obtenue par inclusion-exclusion : parmi les k^n applications de $[1,n]$ vers un ensemble de taille k, on retire celles qui évitent au moins une valeur.

(d) Pour $n \geq 1$,

$$T_k = n$$

signifie qu'au temps $n-1$, exactement $k-1$ espèces ont été observées, et que le tirage n donne une nouvelle espèce. Ainsi

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

En utilisant la formule de la question précédente,

$$\mathbb{P}(Y_{n-1} = k-1) = \frac{N!}{N^{n-1}(N-k+1)!} S(n-1, k-1).$$

Donc

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

Par simplification,

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

On adopte la convention $S(0,0) = 1$, ce qui donne bien $\mathbb{P}(T_1 = 1) = 1$.

Interprétation combinatoire : pour que $T_k = n$, les $n-1$ premiers tirages doivent utiliser exactement $k-1$ espèces, puis le dernier tirage doit être une nouvelle espèce. On choisit et organise les $k-1$ espèces déjà vues via les partitions comptées par $S(n-1, k-1)$, puis on choisit la nouvelle espèce parmi les $N-k+1$ restantes. Le facteur

$$\frac{N!}{(N-k)!}$$

correspond au choix ordonné des k espèces intervenant dans ce mécanisme.

3. Les résultats de la Partie V apportent des lois exactes, et non seulement des espérances, variances ou bornes asymptotiques.

La formule de $\mathbb{P}(T_N > n)$ donne directement la probabilité de ne pas avoir observé toutes les espèces après n tirages. La loi de Y_n décrit la distribution complète du nombre d'espèces observées après un effort d'échantillonnage fixé. Enfin, la loi de T_k précise la distribution du temps nécessaire pour atteindre k espèces distinctes.

Ces résultats permettent donc de calculer des probabilités fines, des quantiles et des intervalles de confiance. Ils complètent les Parties II à IV : les Parties II et III expliquent les comportements moyens, la Partie IV donne une concentration asymptotique dans le cas uniforme, et la Partie V fournit les formules exactes sous-jacentes.