HEC–ESCP – Mathématiques II

Mathématiques approfondies

Corrigé du sujet adapté – 2022

Les questions et sous-questions conservent la numérotation de l'énoncé. Les notations $\mathbb{E}$ et $\mathbb{V}$ désignent l'espérance et la variance. La transposée d'une matrice A est notée ${}^tA$.

Préliminaire – Temps d'attente

Une collection comporte n types équiprobables. On note C_i le nombre d'achats requis pour obtenir i types distincts et $X_i = C_i - C_{i-1}$, avec $C_0 = 0$. Les temps X_i sont indépendants.

Corrigé – Préliminaire – Temps d'attente.

1. Décomposition et lois des temps d'attente.

a) Par télescopage,

$$\boxed{\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (C_i - C_{i-1}) = C_n.}$$

b) Après l'obtention de $i-1$ types, il reste $n-i+1$ types nouveaux. Chaque achat indépendant a donc une probabilité $p_i = (n-i+1)/n$ de fournir un type nouveau. Pour $i \geq 2$, la probabilité d'attendre exactement $k \geq 1$ achats est celle de $k-1$ échecs puis d'un succès :

$$\mathbb{P}(X_i = k) = (1-p_i)^{k-1} p_i.$$

Le premier achat fournit nécessairement un premier type, donc $X_1 = 1$. Avec la convention que $\mathcal{G}(1)$ est cette loi certaine,

$$\boxed{X_i \sim \mathcal{G}\left(\frac{n-i+1}{n}\right) \quad (1 \leq i \leq n).}$$

Partie I – Premiers résultats asymptotiques

On pose $H_n = \sum_{k=1}^{n} 1/k$, $S_n = \sum_{k=1}^{n} 1/k^2$ et $u_n = H_n - \ln n$. Les propriétés des temps géométriques donnent les moments de C_n et des bornes de probabilité.

Corrigé – Partie I – Premiers résultats asymptotiques.

2. Convergence et approximation de S_n.

a) Pour $k \geq 2$, les dénominateurs sont positifs et

$$\frac{1}{k} - \frac{1}{k+1} = \frac{1}{k(k+1)} \leq \frac{1}{k^2} \leq \frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}.$$

Ainsi $\boxed{\dfrac{1}{k}-\dfrac{1}{k+1}\leq\dfrac{1}{k^2}\leq\dfrac{1}{k-1}-\dfrac{1}{k}.}$

b) Sommons de $k=2$ à n et ajoutons le terme 1 : les différences télescopent, d'où

$$\boxed{\frac{3}{2}-\frac{1}{n+1}\leq S_n\leq 2-\frac{1}{n}.}$$

c) On a $S_{n+1}-S_n=1/(n+1)^2>0$; la suite est croissante et majorée par 2. Le théorème de la limite monotone donne une limite S. En passant à la limite dans les bornes,

$$\boxed{S_n\longrightarrow S,\qquad \frac{3}{2}\leq S\leq 2.}$$

d) Pour $N>n\geq 1$, on somme l'encadrement de a) entre $n+1$ et N :

$$\frac{1}{n+1}-\frac{1}{N+1}\leq S_N-S_n\leq\frac{1}{n}-\frac{1}{N}.$$

Puis $N\to+\infty$ donne

$$\boxed{\frac{1}{n+1}\leq S-S_n\leq\frac{1}{n}.}$$

e) Le milieu de l'intervalle contenant S est

$$a_n=S_n+\frac{1}{2}\left(\frac{1}{n}+\frac{1}{n+1}\right),\qquad |S-a_n|\leq\frac{1}{2n(n+1)}.$$

On calcule S_n jusqu'à ce que cette dernière borne soit au plus 10^{-7}.

```
def approximation_S():
    n = 1
    somme = 1
    while 1/(2*n*(n+1)) > 10**(-7):
        n = n+1
        somme = somme+1/(n**2)
    return somme+(1/n+1/(n+1))/2
```

Le programme s'arrête à $n=2236$. Le résultat est environ 1,64493407, et la majoration prouvée donne

$$\boxed{|S-a_{2236}|\leq 10^{-7}.}$$

3. Convergence de $H_n-\ln n$.

a) La fonction $t\mapsto 1/t$ est décroissante sur $]0,+\infty[$. Pour $t\in[n,n+1]$, $1/(n+1)\leq 1/t\leq 1/n$. En intégrant,

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

b) Pour $n\geq 2$, la somme des bornes sur $[k,k+1]$, $1\leq k\leq n-1$, donne

$$H_n-1\leq\ln n\leq H_{n-1}\leq H_n.$$

Donc $0\leq u_n\leq 1$. Cela vaut aussi pour $n=1$, où $u_1=1$:

$$\boxed{\forall n\geq 1,\quad 0\leq H_n-\ln n\leq 1.}$$

c) On a

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

La suite est décroissante et minorée par 0. Le théorème de la limite monotone donne

$$\boxed{H_n - \ln n \longrightarrow \gamma, \qquad \gamma \in [0,1].}$$

4. Espérance de C_n.

Chaque temps géométrique a une espérance finie égale à $1/p_i = n/(n-i+1)$; le cas $i = 1$ donne bien 1. La somme est finie, donc C_n a une espérance. Par linéarité, en posant $k = n - i + 1$,

$$\boxed{\mathbb{E}(C_n) = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{k=1}^{n} \frac{1}{k} = nH_n.}$$

5. Variance de C_n.

Les variables X_i ont des moments d'ordre 2 finis et sont indépendantes. Leurs variances sont $(1-p_i)/p_i^2$; pour $i = 1$, cette variance vaut 0. En changeant l'indice dans une somme finie,

$$\mathbb{V}(C_n) = \sum_{i=1}^{n} \frac{1-p_i}{p_i^2} = \sum_{k=1}^{n} \left(\frac{n^2}{k^2} - \frac{n}{k}\right).$$

Ainsi $\boxed{\mathbb{V}(C_n) = n^2 S_n - nH_n.}$

6. Bornes de probabilité.

a) Pour $a > 0$, l'inégalité de Bienaymé–Tchebychev et la question 5 donnent

$$\mathbb{P}(|C_n - nH_n| \geq an) \leq \frac{\mathbb{V}(C_n)}{a^2 n^2} = \frac{S_n - H_n/n}{a^2} \leq \frac{S}{a^2}.$$

Donc $\boxed{\mathbb{P}(|C_n - nH_n| \geq an) \leq S/a^2.}$

b) Posons $D_n = C_n/n - H_n$. On a $C_n/n - \ln n = D_n + u_n$, avec $|u_n| \leq 1$. Si $|D_n + u_n| \geq c > 1$, alors $|D_n| \geq c - |u_n| \geq c - 1$. Par inclusion d'événements et a),

$$\boxed{\mathbb{P}\left(\left|\frac{C_n}{n} - \ln n\right| \geq c\right) \leq \frac{S}{(c-1)^2}.}$$

c) Pour $c = 6$, la borne $S \leq 2$ donne $S/(c-1)^2 \leq 2/25 = 0{,}08$. Ainsi

$$\mathbb{P}\left(\frac{C_n}{n} \in [\ln n - 6, \ln n + 6]\right) \geq 0{,}92.$$

Pour $n = 10^6$, $\ln n \simeq 13{,}81551$, donc l'intervalle de gauche est contenu dans [7,81; 19,92]. Par croissance de la probabilité,

$$\boxed{\mathbb{P}\left(\frac{C_n}{n} \in [7{,}81; 19{,}92]\right) \geq 0{,}92.}$$

Partie II – Convergence vers la loi de Gumbel

Identifier la loi du maximum d'exponentielles, puis celle d'une somme d'exponentielles de paramètres distincts. Un couplage avec des variables géométriques et un encadrement de fonctions de répartition permettront ensuite de retrouver la même limite pour le collectionneur.

Corrigé – Partie II – Convergence vers la loi de Gumbel.

7. Loi du maximum M_n.

Si $x \leq 0$, chaque T_i étant strictement positive presque sûrement, $\mathbb{P}(M_n \leq x) = 0$. Si $x > 0$, l'indépendance des T_i donne

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(T_1 \leq x, \dots, T_n \leq x) \\ &= \prod_{i=1}^{n} \mathbb{P}(T_i \leq x) = (1-e^{-x})^n.\end{aligned}$$

Cette fonction de répartition est continue sur $\mathbb{R}$ et de classe $\mathcal{C}^1$ sur chacun des intervalles $]-\infty, 0[$ et $]0, +\infty[$. Sa dérivée donne donc une densité de M_n ; sa valeur en 0 peut être choisie égale à 0 :

$$\boxed{f_n(x) = \begin{cases} ne^{-x}(1-e^{-x})^{n-1} & \text{si } x > 0, \\ 0 & \text{si } x \leq 0. \end{cases}}$$

8. Une somme d'exponentielles de paramètres distincts.

a) La densité de $Z \sim \mathcal{E}(n+1)$ est $g(s) = (n+1)e^{-(n+1)s}$ pour $s > 0$, et 0 pour $s \leq 0$. Pour $0 < t < x$,

$$\begin{aligned}f_n(t)g(x-t) &= ne^{-t}(1-e^{-t})^{n-1}(n+1)e^{-(n+1)(x-t)} \\ &= (n+1)e^{-(n+1)x}ne^{nt}(1-e^{-t})^{n-1} \\ &= (n+1)e^{-(n+1)x}ne^{t}(e^{t}-1)^{n-1}.\end{aligned}$$

Ainsi

$$\boxed{f_n(t)g(x-t) = (n+1)e^{-(n+1)x}ne^{t}(e^{t}-1)^{n-1}.}$$

b) Montrons par récurrence que $S_n = \sum_{i=1}^n Z_i$ a pour densité f_n. Pour $n = 1$, $Z_1 \sim \mathcal{E}(1)$ et sa densité est bien f_1. Supposons le résultat établi au rang n. La variable Z_{n+1} est indépendante de S_n, puisqu'elle est indépendante de la famille $(Z_1, \dots, Z_n)$. Le théorème de convolution donne une densité h de $S_n + Z_{n+1}$. Pour $x \leq 0$, $h(x) = 0$, car ces variables sont positives. Pour $x > 0$, les deux densités ne contribuent que lorsque $0 < t < x$, donc

$$h(x) = \int_0^x f_n(t)g(x-t)\,\mathrm{d}t = (n+1)e^{-(n+1)x}\int_0^x ne^t(e^t-1)^{n-1}\,\mathrm{d}t.$$

La fonction $t \mapsto (e^t-1)^n$ est de classe $\mathcal{C}^1$ et a pour dérivée $ne^t(e^t-1)^{n-1}$. Par conséquent

$$h(x) = (n+1)e^{-(n+1)x}(e^x-1)^n = (n+1)e^{-x}(1-e^{-x})^n = f_{n+1}(x).$$

La récurrence est établie, et

$$\boxed{\sum_{i=1}^{n} Z_i \text{ a pour densité } f_n \text{ et pour loi celle de } M_n.}$$

9. La densité de Gumbel.

La fonction $f(x) = e^{-x}e^{-e^{-x}}$ est continue et strictement positive sur $\mathbb{R}$. Posons $G(x) = e^{-e^{-x}}$. Cette fonction est de classe $\mathcal{C}^1$, et

$$G'(x) = e^{-e^{-x}}e^{-x} = f(x).$$

De plus, $G(x) \to 0$ lorsque $x \to -\infty$ et $G(x) \to 1$ lorsque $x \to +\infty$. En intégrant $f = G'$ sur un segment puis en faisant tendre ses bornes vers $-\infty$ et $+\infty$, on obtient $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$. Les conditions caractérisant une densité sont donc satisfaites. Sa fonction de répartition est

$$\boxed{G(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = e^{-e^{-x}}.}$$

On appellera ici loi de Gumbel la loi définie par cette densité f et cette fonction de répartition G.

10. Convergence en loi de W_n.

a) D'après la question 8, $\sum_{i=1}^{n} Z_i$ a la même loi que M_n. Pour tout réel x,

$$F_{W_n}(x) = \mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq x + \ln n\right).$$

Si $x + \ln n \leq 0$, cette probabilité est nulle. Sinon, elle vaut $(1 - e^{-(x+\ln n)})^n$. Ainsi

$$\boxed{F_{W_n}(x) = \begin{cases} \left(1 - \dfrac{e^{-x}}{n}\right)^n & \text{si } x > -\ln n, \\ 0 & \text{si } x \leq -\ln n. \end{cases}}$$

b) Fixons $x \in \mathbb{R}$ et posons $a = e^{-x} > 0$. Pour n assez grand, $x > -\ln n$ et $0 < a/n < 1$. Par dérivabilité du logarithme en 1,

$$\frac{\ln(1-h)}{h} \longrightarrow -1 \quad (h \to 0), \qquad n\ln\left(1 - \frac{a}{n}\right) = a\frac{\ln(1-a/n)}{a/n} \longrightarrow -a.$$

Par continuité de l'exponentielle,

$$F_{W_n}(x) \longrightarrow e^{-a} = e^{-e^{-x}} = G(x).$$

La fonction G est continue sur $\mathbb{R}$. La définition de la convergence en loi, par convergence des fonctions de répartition aux points de continuité de la limite, donne donc

$$\boxed{W_n \xrightarrow{\mathcal{L}} L, \qquad L \text{ de loi de Gumbel.}}$$

11. Simulations de V_n.

a) Avec les bibliothèques usuelles importées sous les noms `np` et `rd`, on simule les n variables géométriques indépendantes de la question 1, puis on calcule $C_n/n - \ln n$. La fonction prend un entier $n \geq 1$.

```
import numpy as np
import numpy.random as rd

def simulV(n):
    achats = 0
    for i in range(1, n + 1):
        p = (n - i + 1) / n
        achats = achats + rd.geometric(p)
    return achats / n - np.log(n)
```

Pour $i = 1$, $p = 1$ et la valeur simulée vaut 1, conformément à $X_1 = 1$.

b) Pour un entier $n \geq 1$ fixé, le programme suivant range les 1000 réalisations dans un vecteur :

```
V = np.zeros(1000)
for j in range(1000):
    V[j] = simulV(n)
```

Chaque appel effectue de nouveaux tirages, ce qui permet de simuler des réalisations indépendantes de V_n.

c) Les histogrammes présentent une forme dissymétrique, avec une partie droite plus étendue. Lorsque n augmente, leur forme se rapproche globalement de la courbe de la densité f. Les différences restantes peuvent notamment provenir des 1000 tirages et du choix des classes ; les figures ne constituent pas une preuve de convergence et ne garantissent pas un rapprochement monotone. Elles suggèrent la conjecture

$$\boxed{V_n \xrightarrow{\mathcal{L}} \text{une variable de loi de Gumbel.}}$$

12. Couplage d'une exponentielle et d'une géométrique.

Puisque $p \in]0,1[$, $\ln(1-p) < 0$ et $\alpha = -p/\ln(1-p) > 0$. La variable $V = \lfloor \alpha U \rfloor + 1$ prend ses valeurs dans $\mathbb{N}^*$. Pour tout entier $k \geq 0$,

$$\begin{aligned}\mathbb{P}(V > k) &= \mathbb{P}(\lfloor \alpha U \rfloor \geq k) = \mathbb{P}(\alpha U \geq k) \\ &= \exp\left(-\frac{pk}{\alpha}\right) = \exp(k\ln(1-p)) = (1-p)^k.\end{aligned}$$

Pour $k \geq 1$, la différence de deux probabilités de queue donne

$$\mathbb{P}(V = k) = \mathbb{P}(V > k-1) - \mathbb{P}(V > k) = (1-p)^{k-1} - (1-p)^k = p(1-p)^{k-1}.$$

Ainsi $\boxed{V \sim \mathcal{G}(p).}$

13. Majorations de variance.

a) Pour tout réel $t \geq 0$, $\lfloor t \rfloor \leq t < \lfloor t \rfloor + 1$. Donc

$$0 < V - \alpha U = \lfloor \alpha U \rfloor + 1 - \alpha U \leq 1.$$

En particulier, cette variable possède un moment d'ordre 2. La formule de König–Huygens donne

$$\mathbb{V}(V - \alpha U) = \mathbb{E}((V-\alpha U)^2) - \mathbb{E}(V-\alpha U)^2 \leq \mathbb{E}((V-\alpha U)^2) \leq 1.$$

Ainsi

$$\boxed{V - \alpha U \in [0,1], \qquad \mathbb{V}(V - \alpha U) \leq 1.}$$

b) Si X, Y possèdent des moments d'ordre 2, l'inégalité $2|XY| \leq X^2 + Y^2$ montre que $\mathbb{E}(|XY|) < +\infty$. Les espérances de X et de Y existent aussi, donc leur covariance $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ est bien définie. En développant les carrés des variables centrées,

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X,Y),$$
$$\mathbb{V}(X-Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\,\mathrm{Cov}(X,Y).$$

On additionne ces deux égalités et on utilise $\mathbb{V}(X-Y) \geq 0$, d'où

$$\boxed{\mathbb{V}(X+Y) \leq 2(\mathbb{V}(X) + \mathbb{V}(Y)).}$$

Aucune indépendance de X et de Y n'est nécessaire pour cette inégalité.

c) Écrivons $V - U = (V - \alpha U) + (\alpha - 1)U$. La question b) donne

$$\begin{aligned}\mathbb{V}(V-U) &\leq 2\mathbb{V}(V-\alpha U) + 2\mathbb{V}((\alpha-1)U)\\ &\leq 2 + 2(\alpha-1)^2\mathbb{V}(U).\end{aligned}$$

Comme $U \sim \mathcal{E}(p)$, $\mathbb{V}(U) = 1/p^2$, donc

$$\boxed{\mathbb{V}(V-U) \leq 2 + 2\frac{(1-\alpha)^2}{p^2}.}$$

14. Transfert de la limite par un encadrement de répartitions.

Posons $p_i = (n-i+1)/n$ et $D_n = V_n' - W_n'$.

a) Les variables $X_i' - Y_i$ sont mutuellement indépendantes, puisque chacune est une fonction de Y_i seulement. Pour $i = 1$, $X_1' = 1$ et $Y_1 \sim \mathcal{E}(1)$, donc $\mathbb{V}(X_1' - Y_1) = 1$. Pour $i \geq 2$, on a $0 < p_i < 1$ et

$$\alpha_i = -\frac{p_i}{\ln(1-p_i)}.$$

Les questions 12 et 13 s'appliquent à (Y_i, X_i') et donnent

$$\mathbb{V}(X_i' - Y_i) \leq 2 + 2\frac{(1-\alpha_i)^2}{p_i^2}.$$

Toutes les variances existent. En additionnant les variances des variables indépendantes,

$$\boxed{\mathbb{V}(D_n) \leq \frac{1}{n^2} + \frac{2}{n^2}\sum_{i=2}^{n}\left[\left(\frac{n}{n-i+1}\right)^2(1-\alpha_i)^2 + 1\right].}$$

b) i) Sur $]0,1[$, la fonction

$$\phi(z) = \frac{1}{(1-z)^2}\left(1 + \frac{1-z}{\ln z}\right)^2$$

est continue, puisque ses dénominateurs ne s'annulent pas. Lorsque $z \to 0^+$, $1 - z \to 1$ et $\ln z \to -\infty$, donc $\phi(z) \to 1$. Pour la limite en 1, posons $h = 1 - z \to 0^+$. La fonction logarithme est de classe $\mathcal{C}^2$ au voisinage de 1, avec $\ln'(1) = 1$ et $\ln''(1) = -1$. La formule de Taylor à l'ordre 2 donne

$$\ln(1-h) = -h - \frac{h^2}{2} + h^2\varepsilon(h), \qquad \varepsilon(h) \longrightarrow 0.$$

Par conséquent

$$\frac{1+h/\ln(1-h)}{h} = \frac{\ln(1-h)+h}{h\ln(1-h)} = \frac{-1/2+\varepsilon(h)}{-1-h/2+h\varepsilon(h)} \longrightarrow \frac{1}{2}.$$

En élevant au carré, $\phi(z) \to 1/4$ lorsque $z \to 1^-$. On obtient donc le prolongement continu

$$\boxed{\phi(0) = 1, \qquad \phi(1) = \frac{1}{4}.}$$

ii) Pour $z_i = (i-1)/n$, on a $1 - z_i = p_i$ et $\alpha_i = -(1-z_i)/\ln z_i$. Ainsi

$$\left(\frac{n}{n-i+1}\right)^2 (1-\alpha_i)^2 = \phi(z_i).$$

La continuité sur $[0,1]$ permet d'utiliser les sommes de Riemann :

$$\frac{1}{n}\sum_{i=2}^{n} \phi\left(\frac{i-1}{n}\right) = \frac{1}{n}\sum_{j=1}^{n-1} \phi\left(\frac{j}{n}\right) \longrightarrow \int_0^1 \phi(t)\ \mathrm{d}t =: A.$$

En effet, la somme allant de $j = 1$ à n est une somme de Riemann et l'unique terme supprimé est $\phi(1)/n \to 0$. La fonction est positive ou nulle. Puisque $\phi(0) = 1$ et ϕ est continue en 0, il existe $\delta \in]0,1]$ tel que $\phi(t) \geq 1/2$ sur $[0,\delta]$. Donc $A \geq \delta/2 > 0$. Ainsi

$$\boxed{\frac{1}{n}\sum_{i=2}^{n} \left(\frac{n}{n-i+1}\right)^2 (1-\alpha_i)^2 \longrightarrow A = \int_0^1 \phi(t)\ \mathrm{d}t > 0.}$$

c) Pour $i \geq 2$, $X_i' \sim \mathcal{G}(p_i)$, donc $\mathbb{E}(X_i') = 1/p_i = \mathbb{E}(Y_i)$. Pour $i = 1$, les deux espérances valent 1. Par linéarité,

$$\mathbb{E}(D_n) = \frac{1}{n}\sum_{i=1}^{n} (\mathbb{E}(X_i') - \mathbb{E}(Y_i)) = 0.$$

Notons A_n la somme divisée par n de b) ii). Le majorant de a) devient

$$0 \leq \mathbb{V}(D_n) \leq \frac{1}{n^2} + \frac{2}{n}\left(A_n + \frac{n-1}{n}\right) \longrightarrow 0,$$

car $A_n \to A$. Pour tout $\varepsilon > 0$, l'inégalité de Bienaymé–Tchebychev donne

$$\mathbb{P}(|D_n| > \varepsilon) \leq \frac{\mathbb{V}(D_n)}{\varepsilon^2} \longrightarrow 0.$$

Ainsi $\boxed{V_n' - W_n' \xrightarrow{\mathbb{P}} 0.}$

d) Les variables Y_i/n sont indépendantes et suivent les lois exponentielles de paramètres $np_i = n - i + 1$. En inversant l'ordre des indices, leur somme a donc la même loi que $\sum_{j=1}^{n} Z_j$ de la question 8. Ainsi W_n' a la même loi que W_n et

$$F_{W_n'}(x) \longrightarrow G(x) \qquad (x \in \mathbb{R}).$$

Fixons $x \in \mathbb{R}$ et $\varepsilon > 0$. Sur $\{|D_n| \leq \varepsilon\}$,

$$W_n' \leq x - \varepsilon \implies V_n' \leq x, \qquad V_n' \leq x \implies W_n' \leq x + \varepsilon.$$

Ces inclusions donnent l'encadrement

$$F_{W_n'}(x-\varepsilon) - \mathbb{P}(|D_n| > \varepsilon) \leq F_{V_n'}(x) \leq F_{W_n'}(x+\varepsilon) + \mathbb{P}(|D_n| > \varepsilon).$$

La probabilité d'erreur tend vers 0, et les fonctions de répartition aux deux bornes tendent vers $G(x-\varepsilon)$ et $G(x+\varepsilon)$. Pour tout $\eta > 0$, la continuité de G permet de choisir $\varepsilon > 0$ tel que ces deux valeurs diffèrent de $G(x)$ de moins de $\eta/2$. Puis, pour n assez grand, chacune des deux erreurs de convergence des fonctions de répartition et la probabilité $\mathbb{P}(|D_n| > \varepsilon)$ sont inférieures à $\eta/4$. On obtient alors $|F_{V_n'}(x) - G(x)| < \eta$. Donc $F_{V_n'}(x) \to G(x)$, et

$$\boxed{V_n' \xrightarrow{\mathcal{L}} \text{une variable de loi de Gumbel.}}$$

Enfin, les X_i' sont indépendantes et ont les mêmes lois respectives que les X_i du modèle. Les deux sommes ont donc la même loi, et V_n' a la même loi que V_n. Ainsi

$$\boxed{V_n \xrightarrow{\mathcal{L}} \text{une variable de loi de Gumbel.}}$$

e) La continuité de G et la convergence en loi donnent, pour de grandes valeurs de n,

$$\mathbb{P}(-1{,}17 \leq V_n \leq 3{,}20) \simeq G(3{,}20) - G(-1{,}17) \simeq 0{,}96 - 0{,}04 = 0{,}92.$$

Pour $n = 10^6$, $\ln n \simeq 13{,}816$. L'événement précédent correspond à

$$\frac{C_n}{n} \in [\ln n - 1{,}17, \ln n + 3{,}20] \simeq [12{,}65, 17{,}02].$$

Avec les arrondis annoncés,

$$\boxed{\mathbb{P}\left(\frac{C_n}{n} \in [12{,}65, 17{,}02]\right) \simeq 0{,}92.}$$

L'intervalle est nettement plus resserré que celui obtenu dans la première partie. La méthode par convergence en loi donne toutefois une approximation asymptotique ; elle ne fournit pas, à elle seule, une minoration rigoureuse de la probabilité pour la valeur finie $n = 10^6$. La première partie utilisait au contraire une inégalité valable pour chaque n.

Partie III – Poissonnisation du modèle

Dépendance des nombres de vignettes après un nombre fixé d'achats, indépendance après remplacement de ce nombre par une variable de Poisson et convergence en loi vers la loi de Gumbel. Dans cette partie, $n \geq 2$.

Corrigé – Partie III – Poissonnisation du modèle.

15. Dépendance des nombres de vignettes pour un nombre fixé d'achats.

a) Dans chaque paquet, le type de vignette est choisi uniformément parmi les n types, indépendamment des autres paquets. Pour un type i fixé, l'indicatrice de sa présence dans chacun des m paquets suit donc une loi de Bernoulli de paramètre $1/n$. Ces m indicatrices sont indépendantes et leur somme vaut $A_{i,m}$. Ainsi

$$\boxed{A_{i,m} \sim \mathcal{B}\left(m, \frac{1}{n}\right).}$$

b) Les variables considérées sont bornées et ont donc un moment d'ordre 2. La bilinéarité de la covariance donne

$$\operatorname{Cov}(A_{1,m}, m - A_{1,m}) = \operatorname{Cov}(A_{1,m}, m) - \operatorname{Var}(A_{1,m}) = -\operatorname{Var}(A_{1,m}).$$

D'après la variance de la loi binomiale,

$$\boxed{\operatorname{Cov}(A_{1,m}, m - A_{1,m}) = -\frac{m}{n}\left(1 - \frac{1}{n}\right) < 0.}$$

c) Chaque paquet contient exactement une vignette, donc

$$A_{1,m} + \cdots + A_{n,m} = m.$$

Si les $A_{i,m}$ étaient mutuellement indépendantes, on aurait $\operatorname{Cov}(A_{1,m}, A_{i,m}) = 0$ pour chaque $i \geq 2$. Par bilinéarité,

$$\operatorname{Cov}(A_{1,m}, m - A_{1,m}) = \operatorname{Cov}\left(A_{1,m}, \sum_{i=2}^{n} A_{i,m}\right) = \sum_{i=2}^{n} \operatorname{Cov}(A_{1,m}, A_{i,m}) = 0,$$

ce qui contredit la question b). Ainsi

$$\boxed{A_{1,m}, \ldots, A_{n,m} \text{ ne sont pas mutuellement indépendantes.}}$$

16. Bornes sur la probabilité que la collection reste incomplète.

a) Pour deux événements E, F,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F) \leq \mathbb{P}(E) + \mathbb{P}(F).$$

En appliquant cette inégalité successivement à l'union des événements déjà considérés et au suivant, une récurrence sur r donne

$$\boxed{\mathbb{P}\left(\bigcup_{i=1}^{r} E_i\right) \leq \sum_{i=1}^{r} \mathbb{P}(E_i).}$$

b) La collection est incomplète après m achats si et seulement si au moins un type est absent. Ainsi

$$[C_n > m] = \bigcup_{i=1}^{n} [A_{i,m} = 0].$$

Par a) et par la loi binomiale de la question 15,

$$\mathbb{P}(C_n > m) \leq \sum_{i=1}^{n} \mathbb{P}(A_{i,m} = 0) = n\left(1 - \frac{1}{n}\right)^m.$$

Pour $z \in [0, 1[$, $\ln(1-z) \leq -z$: la fonction $\ln(1-z) + z$ vaut 0 en 0 et sa dérivée est $-z/(1-z) \leq 0$. En prenant $z = 1/n$ puis en exponentiant, on obtient

$$\boxed{\mathbb{P}(C_n > m) \leq n\left(1 - \frac{1}{n}\right)^m \leq ne^{-m/n}.}$$

Cette dernière borne est également vraie pour $m = 0$, puisque $\mathbb{P}(C_n > 0) = 1 \leq n$.

c) Pour tout seuil réel $t \geq 0$, le caractère entier de C_n donne $[C_n > t] = [C_n > \lfloor t \rfloor]$. La question b) entraîne donc

$$\boxed{\mathbb{P}(C_n > t) \leq ne^{-\lfloor t \rfloor/n} \leq ne^{1/n}e^{-t/n},}$$

car $\lfloor t \rfloor \geq t - 1$. En prenant $t = cn\ln n$ avec $c > 0$,

$$\boxed{\forall c > 0, \quad \mathbb{P}(C_n > cn\ln n) \leq e^{1/n}n^{1-c}.}$$

d) Pour $x > -\ln n$, posons $t = n(\ln n + x) > 0$. Par définition de V_n,

$$[V_n > x] = [C_n > t].$$

La question c) donne alors

$$\mathbb{P}(V_n > x) \leq ne^{1/n}e^{-\ln n - x},$$

soit

$$\boxed{\forall x > -\ln n, \quad \mathbb{P}(V_n > x) \leq e^{1/n}e^{-x}.}$$

Corrigé – Partie III – Nombre de paquets suivant une loi de Poisson.

Dans la suite, $N \sim \mathcal{P}(\lambda)$ est indépendante des choix successifs des types de vignettes, où $\lambda > 0$.

17. Loi des nombres de vignettes après N achats.

a) Conditionnellement à $[N = p]$, les types des p vignettes restent indépendants et uniformes, grâce à l'indépendance de N et des choix. Le raisonnement de la question 15 donne donc

$$\boxed{\widetilde{A}_i \mid [N = p] \sim \mathcal{B}\left(p, \frac{1}{n}\right) \quad (p \geq 1).}$$

Lorsque $N = 0$, toutes les variables $\widetilde{A}_i$ valent 0. Pour $k \in \mathbb{N}$, appliquons la formule

des probabilités totales au système complet d'événements $([N=p])_{p\in\mathbb{N}}$. On a

$$\begin{aligned}\mathbb{P}(\widetilde{A}_i=k) &= \sum_{p=k}^{+\infty}\binom{p}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{p-k}e^{-\lambda}\frac{\lambda^p}{p!}\\ &= e^{-\lambda}\frac{(\lambda/n)^k}{k!}\sum_{r=0}^{+\infty}\frac{[\lambda(1-1/n)]^r}{r!},\end{aligned}$$

avec $r=p-k$. Si $a=\lambda(1-1/n)>0$, la somme des probabilités d'une loi $\mathcal{P}(a)$ vaut 1 :

$$\sum_{r=0}^{+\infty}e^{-a}\frac{a^r}{r!}=1.$$

La somme utilisée dans le calcul précédent vaut donc e^a. Ainsi

$$\boxed{\begin{gathered}\mathbb{P}(\widetilde{A}_i=k)=e^{-\lambda/n}\frac{(\lambda/n)^k}{k!},\\ \widetilde{A}_i\sim\mathcal{P}\left(\frac{\lambda}{n}\right).\end{gathered}}$$

b) Fixons $k_1,\dots,k_n\in\mathbb{N}$ tels que leur somme soit p. Pour obtenir ces effectifs dans les p paquets, on choisit successivement les k_1 positions du premier type, puis les k_2 positions du deuxième parmi les positions restantes, et ainsi de suite. Le nombre de suites possibles est

$$\binom{p}{k_1}\binom{p-k_1}{k_2}\cdots\binom{p-k_1-\cdots-k_{n-1}}{k_n}=\frac{p!}{k_1!\cdots k_n!}.$$

Chaque suite de types a une probabilité $1/n^p$. Par conséquent,

$$\boxed{\mathbb{P}_{[N=p]}\left(\bigcap_{i=1}^n[\widetilde{A}_i=k_i]\right)=\frac{p!}{k_1!\cdots k_n!}\frac{1}{n^p}.}$$

La formule reste vraie pour $p=0$: tous les k_i sont alors nuls et la probabilité vaut 1.

c) Fixons maintenant des $k_i\in\mathbb{N}$ quelconques, et posons $p=\sum_i k_i$. On a toujours $\widetilde{A}_1+\cdots+\widetilde{A}_n=N$, donc l'événement où tous les effectifs valent les k_i impose $N=p$. Grâce à b),

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^n[\widetilde{A}_i=k_i]\right) &= e^{-\lambda}\frac{\lambda^p}{p!}\frac{p!}{k_1!\cdots k_n!}\frac{1}{n^p}\\ &= \prod_{i=1}^n\left[e^{-\lambda/n}\frac{(\lambda/n)^{k_i}}{k_i!}\right].\end{aligned}$$

Ainsi

$$\boxed{\mathbb{P}\left(\bigcap_{i=1}^n[\widetilde{A}_i=k_i]\right)=\prod_{i=1}^n\frac{e^{-\lambda/n}(\lambda/n)^{k_i}}{k_i!}.}$$

d) Chaque facteur est la probabilité marginale calculée en a). La factorisation de la loi jointe pour toutes les valeurs possibles caractérise l'indépendance de variables discrètes. Donc

$$\boxed{\widetilde{A}_1,\dots,\widetilde{A}_n \text{ sont mutuellement indépendantes.}}$$

Pour un nombre fixé de paquets, la somme des effectifs est imposée et les effectifs ne sont pas indépendants. Ici, le nombre de paquets varie selon une loi de Poisson ; le calcul explicite montre que les effectifs deviennent indépendants.

18. Probabilité d'une collection complète après N achats.

La collection est complète exactement lorsque chaque type est présent au moins une fois :

$$D_n = \bigcap_{i=1}^{n} [\widetilde{A}_i \geq 1].$$

Par l'indépendance démontrée à la question 17,

$$\mathbb{P}(D_n) = \prod_{i=1}^{n} \mathbb{P}(\widetilde{A}_i \geq 1) = \prod_{i=1}^{n} (1 - e^{-\lambda/n}).$$

Ainsi

$$\boxed{\mathbb{P}(D_n) = (1 - e^{-\lambda/n})^n.}$$

19. Comparaison avec le modèle à nombre fixé de paquets.

a) Les événements $([N = p])_{p\in\mathbb{N}}$ forment un système complet d'événements. La formule des probabilités totales et l'identité admise de l'énoncé donnent

$$\begin{aligned}\mathbb{P}(D_n) &= \sum_{p=0}^{+\infty} \mathbb{P}_{[N=p]}(D_n)\mathbb{P}(N = p) \\ &= \sum_{p=0}^{+\infty} \mathbb{P}(C_n \leq p)\mathbb{P}(N = p).\end{aligned}$$

Donc

$$\boxed{\mathbb{P}(D_n) = \sum_{p=0}^{+\infty} \mathbb{P}(C_n \leq p)\mathbb{P}(N = p).}$$

b) La loi de Poisson donne $\mathbb{E}(N) = \mathbb{V}(N) = \lambda$. Par l'inégalité de Bienaymé–Tchebychev,

$$\mathbb{P}(|N - \lambda| \geq a) \leq \frac{\lambda}{a^2}.$$

Les entiers définis dans l'énoncé vérifient

$$k_1 = \lfloor \lambda - a \rfloor \leq \lambda - a, \qquad k_2 = \lfloor \lambda + a \rfloor + 1 > \lambda + a.$$

On a donc $[|N - \lambda| < a] \subset [k_1 \leq N \leq k_2]$, d'où

$$\boxed{\sum_{p=k_1}^{k_2} \mathbb{P}(N = p) = \mathbb{P}(k_1 \leq N \leq k_2) \geq 1 - \frac{\lambda}{a^2}.}$$

Les hypothèses $a < \lambda$ et $a > \sqrt{\lambda}$ assurent respectivement $k_1 \geq 0$ et $1 - \lambda/a^2 > 0$.

c) La fonction $p \mapsto \mathbb{P}(C_n \leq p)$ est croissante. En restreignant la somme de a) à $k_1 \leq p \leq k_2$,

$$\begin{aligned}\mathbb{P}(D_n) &\geq \sum_{p=k_1}^{k_2} \mathbb{P}(C_n \leq p)\mathbb{P}(N=p)\\ &\geq \mathbb{P}(C_n \leq k_1) \sum_{p=k_1}^{k_2} \mathbb{P}(N=p)\\ &\geq \mathbb{P}(C_n \leq k_1)\left(1-\frac{\lambda}{a^2}\right).\end{aligned}$$

Pour la borne supérieure, on coupe la somme à k_2 :

$$\begin{aligned}\mathbb{P}(D_n) &\leq \mathbb{P}(C_n \leq k_2)\sum_{p=0}^{k_2}\mathbb{P}(N=p) + \sum_{p=k_2+1}^{+\infty}\mathbb{P}(N=p)\\ &\leq \mathbb{P}(C_n \leq k_2) + \mathbb{P}(N > k_2).\end{aligned}$$

Comme $k_2 > \lambda + a$, l'événement $[N > k_2]$ est inclus dans $[|N-\lambda| \geq a]$. En appliquant de nouveau Bienaymé–Tchebychev, on obtient

$$\boxed{\begin{aligned}\mathbb{P}(C_n \leq k_1)\left(1-\frac{\lambda}{a^2}\right) &\leq \mathbb{P}(D_n)\\ &\leq \mathbb{P}(C_n \leq k_2) + \frac{\lambda}{a^2}.\end{aligned}}$$

Corrigé – Partie III – Retour à la convergence en loi.

20. Convergence du collectionneur vers la loi de Gumbel.

a) Soit (c_n) une suite convergente. Elle est donc bornée. Posons

$$\lambda_n = n\ln n + nc_n = n(\ln n + c_n), \qquad a_n = n^{2/3},$$
$$\eta_n = \frac{\ln n + c_n}{n^{1/3}}.$$

On a $\ln n + c_n \to +\infty$, donc $\lambda_n > 1$ pour n assez grand. Les croissances comparées et la borne sur c_n donnent

$$\frac{\lambda_n}{a_n^2} = \eta_n \longrightarrow 0, \qquad \frac{a_n}{\lambda_n} = \frac{1}{n^{1/3}(\ln n + c_n)} \longrightarrow 0.$$

Ainsi, pour n assez grand, $\sqrt{\lambda_n} < a_n < \lambda_n$ et la question 19 s'applique. Notons

$$k_1 = \lfloor \lambda_n - a_n \rfloor, \qquad k_2 = \lfloor \lambda_n + a_n \rfloor + 1.$$

Les propriétés de la partie entière donnent

$$k_1 \geq \lambda_n - a_n - 1, \qquad k_2 \leq \lambda_n + a_n + 1.$$

Par conséquent, par croissance des probabilités cumulées,

$$\mathbb{P}\left(V_n \leq c_n - \frac{1}{n^{1/3}} - \frac{1}{n}\right) \leq \mathbb{P}(C_n \leq k_1),$$

$$\mathbb{P}(C_n \leq k_2) \leq \mathbb{P}\left(V_n \leq c_n + \frac{1}{n^{1/3}} + \frac{1}{n}\right).$$

Le facteur $1-\eta_n$ étant positif pour n assez grand, on peut remplacer les deux probabilités de la question 19 par ces bornes. Enfin, la question 18 donne, avec le paramètre λ_n,

$$\mathbb{P}(D_n) = (1-e^{-\lambda_n/n})^n = \left(1 - \frac{e^{-c_n}}{n}\right)^n.$$

On en déduit

$$\boxed{\begin{aligned}&\mathbb{P}\left(V_n \leq c_n - \frac{1}{n^{1/3}} - \frac{1}{n}\right)(1-\eta_n)\\ &\qquad \leq \left(1 - \frac{e^{-c_n}}{n}\right)^n\\ &\qquad \leq \mathbb{P}\left(V_n \leq c_n + \frac{1}{n^{1/3}} + \frac{1}{n}\right) + \eta_n,\\ &\qquad \eta_n = \frac{\ln n + c_n}{n^{1/3}}.\end{aligned}}$$

b) Notons F_n la fonction de répartition de V_n, et $F(x) = e^{-e^{-x}}$ la fonction de répartition de la loi de Gumbel. Pour tout $y \in \mathbb{R}$ fixé, on a

$$B_n(y) = \left(1 - \frac{e^{-y}}{n}\right)^n \longrightarrow e^{-e^{-y}} = F(y).$$

En effet, pour $0 < z < 1$, on a $z \leq -\ln(1-z) \leq z/(1-z)$, par comparaison de $1/(1-t)$ avec 1 et $1/(1-z)$ dans l'intégrale de 0 à z. Avec $z = e^{-y}/n$, cela donne

$$-\frac{e^{-y}}{1-e^{-y}/n} \leq \ln B_n(y) \leq -e^{-y}.$$

Par encadrement puis continuité de l'exponentielle, on obtient la limite annoncée.
Fixons maintenant $x \in \mathbb{R}$ et $\varepsilon > 0$, et posons $\delta_n = n^{-1/3} + n^{-1}$, qui tend vers 0. Appliquons a) à la suite constante $c_n = x - \varepsilon$. Pour n assez grand, $\delta_n < \varepsilon$, donc $F_n(x - \varepsilon + \delta_n) \leq F_n(x)$ et

$$B_n(x-\varepsilon) - \frac{\ln n + x - \varepsilon}{n^{1/3}} \leq F_n(x).$$

Appliquons ensuite a) à la suite constante $c_n = x+\varepsilon$. Pour n assez grand, $x+\varepsilon-\delta_n \geq x$ et

$$F_n(x) \leq \frac{B_n(x+\varepsilon)}{1 - (\ln n + x + \varepsilon)/n^{1/3}}.$$

Ainsi $F_n(x)$ est encadré par deux suites dont les limites sont respectivement $F(x-\varepsilon)$ et $F(x+\varepsilon)$. La fonction F est continue. Pour toute précision $r > 0$, on peut donc choisir $\varepsilon > 0$ assez petit pour que ces deux limites soient à une distance inférieure à $r/2$ de $F(x)$. Pour n assez grand, chacune des suites encadrantes est elle-même à une distance inférieure à $r/2$ de sa limite. On obtient alors $|F_n(x) - F(x)| < r$. C'est exactement la convergence $F_n(x) \to F(x)$, pour tout réel x.
Par la définition de la convergence en loi,

$$\boxed{\begin{aligned}&V_n \xrightarrow[n\to+\infty]{\mathcal{L}} V,\\ &F_V(x) = e^{-e^{-x}}, \quad V \text{ suit la loi de Gumbel.}\end{aligned}}$$