

UNIVERSITÉ NICE CÔTE D'AZUR

Modélisation Probabiliste de la Croissance Cellulaire

Méthodes de Monte-Carlo et Chaînes de Markov

SINADINOVIC MARKO

Résumé

Ce travail s'intéresse à la modélisation probabiliste de la division cellulaire à l'aide de simulations de Monte-Carlo. En cherchant d'abord à estimer le taux de division, nous avons identifié un biais de surestimation systématique sur les petits échantillons ($n < 50$) ce qui nous a conduits à proposer une correction de l'estimateur. L'étude de la masse cellulaire a ensuite mis en lumière un phénomène contre-intuitif : la masse moyenne mesurée en continu se révèle être le double de celle observée à la naissance illustrant l'importance du biais d'observation. Enfin la simulation de la croissance de la colonie a confirmé son comportement malthusien. Nos résultats valident numériquement le fait que malgré le hasard des divisions individuelles, l'évolution globale de la population finit par se stabiliser de manière prévisible.

Table des matières

Introduction	1
1 Modélisation Mathématique et Cadre Théorique	2
1.1 Le Processus de Division	2
1.1.1 La durée de vie cellulaire (E_k), le temps inter-divisions	2
1.1.2 L'instant de la division (T_n)	2
1.2 Le Processus de Masse	3
1.2.1 Simplification du problème : du continu au discret	3
1.2.2 Convergence et perte de mémoire	3
1.3 Le Processus de Masse en temps continu	4
1.3.1 L'oubli de la condition initiale	4
1.4 La Dynamique de Population	4
1.4.1 L'arbre de descendance	4
1.4.2 Le Processus de comptage	4
1.4.3 Loi de la population	5
1.4.4 Comportement chaotique ou déterministe?	5
1.5 L'Estimation Statistique	5
1.5.1 L'estimateur naturel	5
1.5.2 L'intervalle de confiance	6
2 Expériences Numériques et Validation	6
2.1 Inférence Statistique du Taux de Division λ	6
2.2 Dynamique de la Masse et Oubli de la Condition Initiale	9
2.3 Explosion Démographique et Processus de Comptage	12
3 Conclusion	15
Bibliographie	16
A Annexe : Mentions Légales et Crédits	18
B Annexe : Compléments Mathématiques et Preuves	18

Introduction : De la Cellule au Modèle Stochastique

Contexte

Pendant les cours de biologie et de sciences du vivant nous avons eu l'occasion de découvrir plusieurs mécanismes liés aux cellules, qu'il s'agisse de plantes ou d'êtres vivants. Pour ceux qui ne se souviennent plus de leurs cours nous avons étudié le mécanisme de division et de croissance cellulaire, appelé cycle cellulaire. En d'autres termes, le cycle cellulaire est le processus fondamental par lequel une cellule grandit, duplique son contenu et se divise pour donner naissance à deux cellules filles. Ce cycle est souvent présenté comme un phénomène périodique, parfaitement réglé, mais lorsqu'on se penche sur le sujet plus en détail, on constate grâce à l'observation microscopique qu'il existe en réalité une part d'aléa. Deux cellules sœurs placées dans des conditions identiques ne se diviseront jamais exactement au même moment. Pour comprendre ce phénomène il est nécessaire d'adopter une approche probabiliste.

Du phénomène biologique à la modélisation

Pour ce faire, nous allons décortiquer le cycle cellulaire pour mieux le comprendre et le traduire mathématiquement. Nous remarquons qu'il existe dans ce cycle une phase prévisible et qui est continue, c'est la phase entre deux divisions. La cellule ne subit pas d'aléas et accumule de la masse de manière déterministe. En complément, on retrouve une phase aléatoire qui est la division de la cellule qui survient brusquement interrompant la croissance de la cellule. Maintenant que nous comprenons le phénomène, nous devons le modéliser avec des outils mathématiques à notre disposition.

Ce comportement hybride mêlant régularité et hasard a déjà été introduit par les mathématiciens. Ce phénomène rentre dans la famille des Processus de Markov Déterministes par Morceaux. Cependant, il est complexe de suivre l'évolution de la cellule à chaque instant t . Pour mener notre modélisation à bien nous allons avoir deux approches différentes. Nous allons utiliser une méthode connue en physique avec les différents modes de descriptions (Lagrangienne, Eulérienne...) nous allons considérer sa taille uniquement au moment où elle se divise. Ensuite nous allons étudier la dynamique de prise de masse à des intervalles de temps très minimes. De ce fait nous pourrions confronter ces deux mondes temporels, le continu et le discret.

Cette approche transforme notre problème en l'étude d'une suite de nombres génération après génération. Puisque la taille future de la cellule ne dépend que de sa taille actuelle et du temps aléatoire entre deux divisions nous nous ramenons à l'étude d'une Chaîne de Markov.

Voir pour comprendre

Maintenant qu'on a réussi à comprendre et à modéliser le phénomène, l'idée est de pouvoir vraiment le manipuler. Comme on travaille avec des mathématiques on a l'habitude de manipuler des lignes de formules et des théorèmes un peu opaques, mais ici on veut éviter de retomber dans des démonstrations trop lourdes qu'on peut déjà trouver dans les cours ou les articles. L'objectif est plutôt de montrer ces résultats sous forme d'expériences numériques, avec quelque chose de visuel et un protocole simple, pour voir concrètement comment les mathématiques se comportent.

Pour ce faire, on commence par poser un cadre mathématique assez minimal mais suffisant pour définir clairement les objets qu'on va utiliser et ceux qu'on va construire. Ensuite on passe aux simulations pour vérifier visuellement ce qu'on attend théoriquement. Et pour finir, on prend un peu de recul pour discuter des limites de notre modélisation, des choix de paramètres et de ce que ça nous apprend vraiment sur le problème.

1 Modélisation Mathématique et Cadre Théorique

Nous cherchons à formaliser le phénomène biologique étudié dans un cadre mathématique suffisamment précis pour permettre ensuite une analyse numérique. L'objectif est de disposer d'outils clairs pour décrire la croissance de la cellule et son mécanisme de division.

Deux questions guident cette modélisation : à quel moment la cellule se divise-t-elle, et comment sa taille évolue-t-elle au cours du temps ? Ces points serviront de base à la construction des objets mathématiques utilisés par la suite.

1.1 Le Processus de Division

1.1.1 La durée de vie cellulaire (E_k) , le temps inter-divisions

Pour décrire cette dynamique temporelle, il est utile de commencer par fixer un repère clair. On note $(T_n)_{n \geq 1}$ la suite des instants successifs de division cellulaire, avec $0 < T_1 < T_2 < \dots < T_n$. On introduit la suite de variables aléatoires $(E_k)_{k \geq 1}$, qui représentent le temps écoulé entre deux divisions (le temps inter-divisions). Par définition :

$$E_1 = T_1 \quad \text{et} \quad E_k = T_k - T_{k-1} \quad \text{pour } k \geq 2. \quad (1)$$

Nous observons d'abord qu'une cellule qui a déjà vécu plusieurs heures n'a pas davantage de chances de se diviser dans les instants qui suivent qu'une cellule beaucoup plus jeune. Autrement dit, le fait qu'une cellule ait déjà survécu pendant un temps t n'influence pas la probabilité qu'elle continue à vivre pendant un temps supplémentaire s . En s'appuyant sur la propriété d'absence de mémoire on peut en déduire que les durées séparant deux divisions successives peuvent être modélisées par une suite de variables aléatoires $(E_k)_{k \geq 1}$ indépendantes et identiquement distribuées (i.i.d), suivant une loi exponentielle : $E_k \sim \mathcal{E}(\lambda)$.

Le paramètre $\lambda > 0$ n'est pas qu'une simple constante. Il correspond au taux de division, autrement dit à la fréquence moyenne des événements. Plus λ est grand, plus l'espérance de vie de la cellule ($\mathbb{E}[E_k] = \frac{1}{\lambda}$) est courte, ce qui traduit un cycle de division plus rapide et plus instable.

1.1.2 L'instant de la division (T_n)

Connaissant la loi qui régit la durée entre deux divisions, il reste à déterminer à quels instants ces divisions se produisent réellement. Pour cela on s'intéresse à l'instant de la n -ième division noté T_n . Cet instant correspond simplement à la somme des durées de vie successives depuis la cellule mère. On peut donc écrire :

$$T_n = \sum_{k=1}^n E_k. \quad (2)$$

Pour mieux comprendre cette division « soudaine », il faut identifier la loi de probabilité qui gouverne T_n . Après quelques calculs nous remarquons que $T_n \sim \Gamma(n, \lambda)$. De plus en exploitant la relation suivante :

$$\lambda T_n = \sum_{k=1}^n \lambda E_k \sim \Gamma(n, 1) \quad (3)$$

notre simulation gagnera en robustesse et utilisera moins de ressources car la loi Gamma est déjà préprogrammée dans les bibliothèques NumPy et SciPy.

Cependant, caractériser l'instant de la n -ième division ne suffit pas pour valider notre modélisation de T_n . Il faut également vérifier que le processus ne génère pas une infinité de divisions en un temps fini. Pour cela,

nous étudions le comportement asymptotique de T_n . La Loi Forte des Grands Nombres garantit qu'un tel emballage est impossible, tandis que le Théorème Central Limite montre que l'erreur d'estimation suit asymptotiquement une distribution gaussienne.

Théorème 1 (Convergences classiques).

$$\begin{aligned} \frac{T_n}{n} &= \frac{1}{n} \sum_{k=1}^n E_k \xrightarrow[n \rightarrow \infty]{p.s.} \mathbb{E}[E_1] = \frac{1}{\lambda} > 0 \\ \sqrt{n} \left(\frac{1}{n} T_n - \frac{1}{\lambda} \right) &\xrightarrow[n \rightarrow \infty]{loi} \mathcal{N} \left(0, \frac{1}{\lambda^2} \right) \end{aligned} \quad (4)$$

Après avoir modélisé le rythme temporel des divisions nous pouvons désormais analyser la dynamique de prise de masse de la cellule.

1.2 Le Processus de Masse

La masse cellulaire notée $(X_t)_{t \geq 0}$ et à valeurs dans \mathbb{R}_+ peut être vue comme l'alternance de deux phases : une accumulation continue de matière suivie d'une division soudaine.

1.2.1 Simplification du problème : du continu au discret

Avant d'étudier le processus continu (X_t) commençons par une version discrète : la masse Z_n mesurée juste après la n -ième division. Entre les divisions n et $n+1$, la cellule croît pendant une durée E_{n+1} atteignant $Z_n + E_{n+1}$ juste avant de se diviser. La division réduit ensuite cette masse de moitié. En généralisant ce procédé par récurrence, nous obtenons l'expression de la masse après n divisions :

$$Z_n = \frac{1}{2}(Z_{n-1} + E_n) = \frac{1}{2^n} Z_0 + \sum_{k=1}^n \frac{1}{2^k} E_{n-k+1} \quad (5)$$

Une question naturelle se pose alors : la masse de la cellule finit-elle par se stabiliser vers un état d'équilibre ?

1.2.2 Convergence et perte de mémoire

L'analyse de la relation de récurrence montre que l'influence de la taille initiale Z_0 s'estompe rapidement à mesure que n augmente en raison du facteur de contraction $1/2$. La suite $(Z_n)_{n \geq 0}$ converge en loi vers une variable aléatoire limite \tilde{Z} dont la distribution est indépendante du point de départ Z_0 :

$$Z_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \tilde{Z} = \sum_{k=1}^{+\infty} \frac{1}{2^k} E_k \quad (6)$$

Théorème 2 (Convergence en loi de $(Z_n)_{n \geq 0}$). La suite des lois des variables $(Z_n)_{n \geq 0}$ converge vers une mesure de probabilité $\tilde{\mu}$ sur \mathbb{R}_+ dont la fonction caractéristique est donnée pour tout $u \geq 0$ par

$$\int_0^\infty e^{iux} \tilde{\mu}(dx) = \prod_{k=1}^\infty \frac{\lambda}{\lambda - 2^{-k} iu}. \quad (7)$$

Pour valider nos futures expériences numériques, nous retenons deux propriétés essentielles de cette loi d'équilibre : $\mathbb{E}[\tilde{Z}] = \frac{1}{\lambda}$ et $\text{Var}(\tilde{Z}) = \frac{1}{3\lambda^2}$

Grâce à ces résultats théoriques nous savons qu'il n'est pas nécessaire de simuler une infinité de divisions

dans nos expériences. Il y aura une stabilisation après un certain nombre d'étapes que nous déterminerons numériquement en confrontant nos histogrammes et erreurs aux valeurs théoriques de l'espérance et de la variance. Avant cela nous devons revenir au temps continu afin de comparer nos résultats discrets et d'analyser la vitesse de convergence

1.3 Le Processus de Masse en temps continu

Nous avons établi que la suite des masses aux instants de division (Z_n) se stabilise. Cependant la prise de masse reste un processus continu noté $(X_t)_{t \geq 0}$ où (X_t) désigne la masse à l'instant t . La question est donc de savoir si la convergence observée pour les valeurs discrètes entraîne une forme de stabilité pour la trajectoire complète et à quelle vitesse cette stabilité s'installe.

1.3.1 L'oubli de la condition initiale

Pour analyser la vitesse de convergence nous devons introduire l'espace \mathcal{P} des mesures de probabilité sur \mathbb{R}_+ ayant un moment d'ordre 1 fini. Nous munissons cet espace d'une distance de couplage D caractérisant la convergence en loi et la convergence des espérances :

$$D(\mathcal{L}(X_n), \mathcal{L}(X)) \rightarrow 0 \iff X_n \xrightarrow{\mathcal{L}} X \quad \text{et} \quad \mathbb{E}[X_n] \rightarrow \mathbb{E}[X]. \quad (8)$$

Cette distance permet de quantifier l'oubli de la condition initiale expliquer le théorème suivant :

Théorème 3 (Oubli exponentiellement rapide de la condition initiale). Soient deux copies $(X_t)_{t \geq 0}$ et $(Y_t)_{t \geq 0}$ du processus, associées aux conditions initiales respectives $X_0 = x$ p.s. et $Y_0 = y$ p.s. En notant \mathcal{L} la loi, alors pour tout $t \geq 0$:

$$D(\mathcal{L}(X_t); \mathcal{L}(Y_t)) \leq |x - y| \exp\left(-\frac{1}{2}\lambda t\right) \xrightarrow[t \rightarrow \infty]{} 0. \quad (9)$$

En réalité, nous comparons l'évolution de deux trajectoires $(X_t)_{t \geq 0}$ et $(Y_t)_{t \geq 0}$ issues de conditions initiales différentes X_0 et Y_0 . Si la distance entre leurs lois tend vers 0 cela signifie que le système finit par oublier d'où il est parti. Autrement dit la masse au temps t ne dépend plus réellement des valeurs initiales. Le théorème indique en plus que cet oubli se fait à la vitesse $\exp(-\frac{\lambda t}{2})$ une propriété que nous illustrerons dans nos simulations en comparant les pentes obtenues numériquement.

1.4 La Dynamique de Population

Nous nous sommes focalisés sur le cycle d'une seule cellule ancêtre mais en réalité on travaille sur toute une colonie de cellules ce qui complexifie l'analyse. Pour que notre modélisation soit complète nous devons travailler sur l'évolution d'une population N_t au cours du temps.

1.4.1 L'arbre de descendance

À chaque nœud de l'arbre (individu) est associée une durée de vie $E_u \sim \mathcal{E}(\lambda)$. Lorsqu'une cellule u se divise, elle disparaît et donne naissance à deux cellules filles, notées u_0 et u_1 . Pour décrire l'état de la colonie à un instant t on note N_t le nombre de divisions survenues avant cet instant. Maintenant que nous savons comment représenter cette population regardons comment N_t évolue au cours du temps.

1.4.2 Le Processus de comptage

Imaginons la colonie à un instant donné comportant k cellules. Chacune possède sa propre horloge interne (processus de division). La prochaine division survient dès qu'une des cellules atteint le moment où elle

doit se diviser. Ainsi le temps d'attente avant la prochaine division correspond au minimum des k durées indépendantes des cellules. Nous disposons d'un lemme essentiel pour optimiser nos simulations : il permet de décrire l'évolution de la population sans suivre chaque cellule individuellement en passant directement au prochain événement de division.

Lemme (Minimum). Si E_1, \dots, E_n sont des variables aléatoires indépendantes de lois exponentielles de paramètres $\lambda_1, \dots, \lambda_n$ alors $\min(E_1, \dots, E_n)$ suit une loi exponentielle de paramètre $\lambda_1 + \dots + \lambda_n$.

Cela signifie qu'avec k cellules identiques de paramètre λ , le temps avant la prochaine naissance suit une loi exponentielle de paramètre $k\lambda$. Plus la colonie est grande, plus les divisions s'accélèrent.

1.4.3 Loi de la population

Théorème 4 (Loi de la taille de la population). Le processus $(N_t)_{t \geq 0}$ suit la même loi que le processus de comptage de tops espacés par des durées aléatoires indépendantes S_1, S_2, S_3, \dots de loi exponentielle de paramètre $\lambda, 2\lambda, 3\lambda, \dots$

Théorème 4 bis (Loi Géométrique). Pour tout temps $t \geq 0$, $N_t \sim \mathcal{G}(e^{-\lambda t})$. La probabilité que la population dépasse une taille n et l'espérance sont données par :

$$\mathbb{P}(N_t \geq n) = (1 - e^{-\lambda t})^n \quad \text{et} \quad \mathbb{E}[N_t] = e^{\lambda t} - 1 \quad (10)$$

1.4.4 Comportement chaotique ou déterministe ?

Cette explosion démographique est-elle chaotique ou suit-elle une tendance stable ? Pour pouvoir répondre à cette question nous devons atténuer la croissance par le facteur théorique exponentielle et voir si notre variable converge vers quelque chose qu'on connaît. Pour cela nous introduisons la variable

$$H_t = e^{-\lambda t}(N_t + 1) \quad (11)$$

En cherchant de montrer vers quelle loi H_t converge nous remarquons que $H_t \xrightarrow{\mathcal{L}} \mathcal{E}(1)$.

1.5 L'Estimation Statistique

La modélisation repose sur le paramètre clé λ qui est le taux de division mais on ne connaît pas ce dernier nous devons l'approcher et le retrouver.

1.5.1 L'estimateur naturel

Si l'on observe n divisions successives on dispose des durées (E_1, \dots, E_n) . Puisque l'espérance d'une loi exponentielle est $\frac{1}{\lambda}$, la moyenne empirique des temps devrait être proche de cette valeur. De ce fait nous allons utiliser l'estimateur du maximum de vraisemblance (l'inverse de la moyenne des temps) noté $\hat{\lambda}_n$:

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^n E_i} = \frac{n}{T_n} \quad (12)$$

Grâce à la convergence de la moyenne empirique cet estimateur tend vers la vraie valeur λ lorsque le nombre d'observations devient grand, ce que nous avons vérifié dans le (1.1.2)

1.5.2 L'intervalle de confiance

Comme nous devons estimer un paramètre $\hat{\lambda}_n$, une estimation ponctuelle ne suffit pas en effet elle ne rend pas compte de l'incertitude liée à la taille de l'échantillon. Il nous faut donc évaluer la marge d'erreur. Pour cela nous construisons un intervalle de confiance de niveau $1 - \alpha$ en utilisant le TCL :

$$I_{n,TCL} = \left[\hat{\lambda}_n - z_{1-\alpha/2} \frac{\hat{\lambda}_n}{\sqrt{n}}, \hat{\lambda}_n + z_{1-\alpha/2} \frac{\hat{\lambda}_n}{\sqrt{n}} \right] \quad (13)$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

De plus comme $\lambda T_n \sim \Gamma(n, 1)$ nous pouvons construire un intervalle de confiance exact au niveau $1 - \alpha$:

$$I_{n,exact} = \left[\frac{q_{\alpha/2}}{T_n}, \frac{q_{1-\alpha/2}}{T_n} \right] \quad (14)$$

Après avoir posé le cadre théorique et compris comment modéliser ces mécanismes, nous passons à la phase numérique. L'objectif est désormais de mener des simulations pour valider les résultats obtenus et en proposer une visualisation.

2 Expériences Numériques et Validation

2.1 Inférence Statistique du Taux de Division λ

Objectif et Motivation

Cette première expérience consiste à estimer le paramètre inconnu λ à l'aide de l'estimateur du maximum de vraisemblance $\hat{\lambda}_n$. L'objectif est d'évaluer sa fiabilité lorsque les données sont limitées et de déterminer selon la taille de l'échantillon la méthode d'intervalle de confiance la plus pertinente.

Protocole Expérimental

Nous fixons $\lambda = 2.0$ et générons par la méthode Monte-Carlo $M = 5000$ jeux de données pour différentes tailles d'échantillon $n \in [0, 2000]$. L'expérience se déroule en trois étapes :

1. Convergence : Les temps de division (E_k) sont simulés par la méthode de la transformée inverse $E_k = -\frac{\ln(U_k)}{\lambda}$ avec $U_k \sim \mathcal{U}([0, 1])$. Pour chaque échantillon nous calculons l'estimateur $\hat{\lambda}_n$ et mesurons l'erreur associée.
2. Normalité : Nous vérifions le TCL en comparant l'histogramme de l'erreur normalisée $Z_n = \sqrt{n}(\frac{T_n}{n} - \frac{1}{\lambda})$ avec la densité gaussienne théorique $\mathcal{N}(0, \frac{1}{\lambda^2})$
3. Robustesse IC : Nous comparons les taux de couverture des intervalles de confiance à $\alpha = 5\%$ obtenus par le TCL $\hat{\lambda}_n \pm z_{1-\alpha/2} \frac{\hat{\lambda}_n}{\sqrt{n}}$ et par loi Gamma $\Gamma(n, 1)$

Résultats et Interprétation

Analyse de la consistance de l'estimateur

La simulation confirme la Loi Forte des Grands Nombres. On a bien la convergence *p.s* de $\hat{\lambda}_n$ vers $\lambda = 2.0$ et l'erreur tend bien vers 0 (Fig. 1). Ce phénomène est aussi vérifié par le Tableau 1.

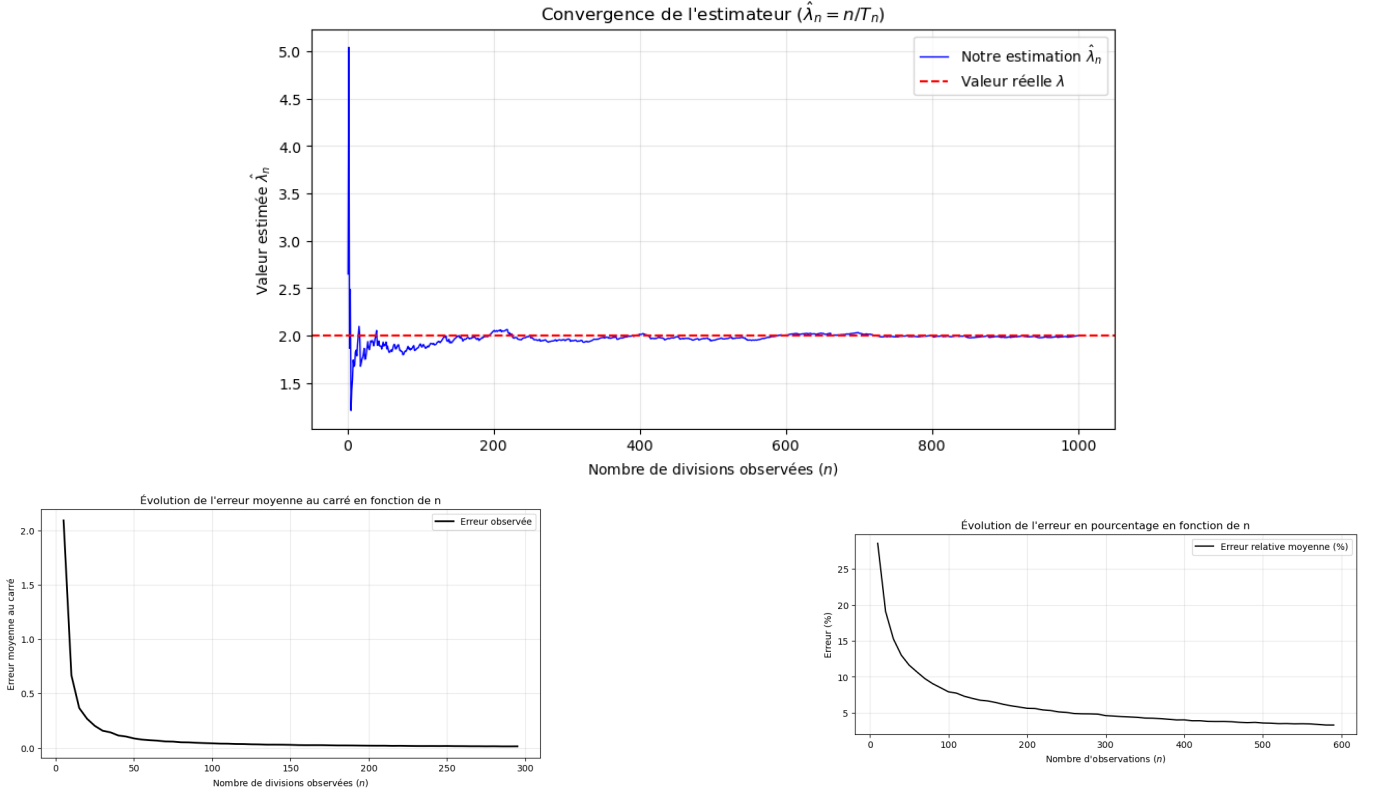


FIGURE 1 – Convergence de l'estimateur et décroissance de l'erreur (Moyenne sur $M = 5000$ essais)

Taille (n)	Vrai λ	Estimateur ($\hat{\lambda}_n$)	Erreur (%)
5	2.0	2.4939	46.54 %
10	2.0	2.2018	28.31 %
50	2.0	2.0399	11.50 %
100	2.0	2.0197	8.01 %
1000	2.0	2.0018	2.53 %
2000	2.0	2.0002	1.81 %

TABLE 1 – Évolution de la précision de l'estimateur (Données simulées)

- Pour $n < 50$: Pour les petits échantillons notre se révèle à la fois instable avec une erreur de 46.54% pour $n = 5$ et biaisé de manière positive avec $\hat{\lambda}_5 \approx 2.49$. De ce fait notre estimateur ne sera pas utilisable pour de petits échantillons.
- Pour $n \approx 100$: L'erreur passe sous les 10% ce qui offre un compromis coût et précision raisonnable dans une expérience biologique.
- Pour $n \geq 1000$: Nous avons une très bonne précision ($< 3\%$) mais il nous faut un jeu de données conséquent.

Cette surestimation est une conséquence directe des propriétés de $\hat{\lambda}_n$. En effet, comme T_n suit une loi Gamma $\Gamma(n, \lambda)$ un calcul direct montre que $\mathbb{E}\left[\frac{1}{T_n}\right] = \frac{\lambda}{n-1}$, et donc $\mathbb{E}[\hat{\lambda}_n] = \frac{n}{n-1}\lambda$. Nous n'avons pas relevé ce détail et avons ainsi oublié ce biais pour les petits échantillons.

Après relecture du cours de Statistique de l'année dernière, nous nous sommes rappelé qu'il est possible de pallier ce problème : pour les petits échantillons on peut corriger l'estimateur en utilisant la version sans biais $\tilde{\lambda}_n = \frac{n-1}{T_n}$.

Pour confirmer la robustesse de ces observations nous avons réalisé deux séries d'expériences indépendantes qui mettent en évidence la fiabilité de l'estimateur corrigé.

Taille n	Expérience 1 (Biais %)		Expérience 2 (Biais %)	
	Ancien	Corrigé	Ancien	Corrigé
5	+25.19 %	+0.15 %	+24.68 %	-0.25 %
10	+11.06 %	-0.04 %	+11.05 %	-0.05 %
20	+5.29 %	+0.02 %	+5.39 %	+0.12 %
50	+1.98 %	-0.06 %	+1.88 %	-0.16 %
100	+1.06 %	+0.05 %	+1.01 %	≈ 0.00 %

TABLE 2 – Comparaison des biais

Validation du TCL et Robustesse des IC

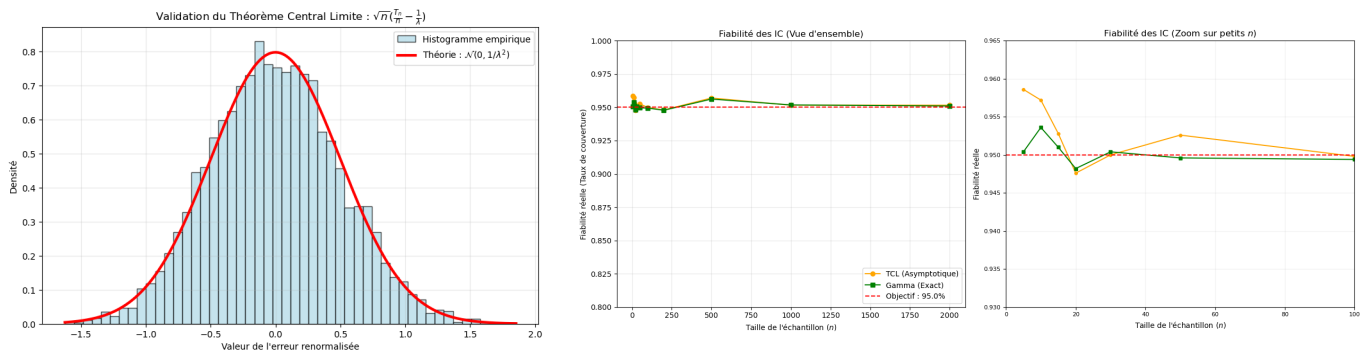


FIGURE 2 – Gauche : Validation du TCL | Droite : Taux de couverture des IC

La superposition de l'histogramme des erreurs avec la densité $\mathcal{N}(0, \frac{1}{\lambda^2})$ visible sur la Figure 2 confirme la validité du Théorème Central Limite.

n	Fiab. TCL	Erreur TCL	Fiab. Gamma	Erreur Gamma
5	95.72 %	-0.72 %	95.02 %	-0.02 %
10	95.20 %	-0.20 %	95.16 %	-0.16 %
50	95.32 %	-0.32 %	95.02 %	-0.02 %
100	95.10 %	-0.10 %	94.82 %	+0.18 %
1000	95.28 %	-0.28 %	95.22 %	-0.22 %
2000	95.26 %	-0.26 %	95.28 %	-0.28 %

TABLE 3 – Comparaison des taux de couverture (Cible : 95%, $M = 5000$)

Cependant, l'analyse des taux de couverture (Tableau 3 et Figure 2 partie droite) montre que la méthode exacte est plus fiable pour les petits échantillons. À $n = 5$ elle atteint une fiabilité quasi-parfaite de 95,02% tandis que l'approximation asymptotique ne couvre que 95,72%. En revanche pour de grandes valeurs de n les deux méthodes deviennent équivalentes.

Conclusion

Cette expérience valide l'estimation par maximum de vraisemblance pour les grands échantillons (convergence vers $\lambda = 2.0$ et IC asymptotiques fiables) mais révèle une surestimation pour les petits n . Une correction de l'estimateur supprime ce biais et s'avère plus robuste.

2.2 Dynamique de la Masse et Oubli de la Condition Initiale

Objectif et Motivation

D'abord nous montrerons que l'influence des conditions initiales s'estompe (perte de mémoire) à une certaine étape et à une vitesse donnée. Ensuite nous vérifierons que la loi limite des masses post-division coïncide avec le résultat théorique. Enfin nous confronterons ces prédictions aux observations en temps continu pour mesurer le biais d'échantillonnage entre les modèles continus et discrets.

Protocole Expérimental

Les simulations sont réalisées avec un taux de division fixé à $\lambda = 2.0$. Le protocole se décompose en trois expériences distinctes :

1. Perte de mémoire : Nous simulons deux trajectoires avec conditions initiales différentes ($x_0 = 0.1$ et $y_0 = 10.0$). Pour une comparaison rigoureuse nous utilisons la même graine aléatoire pour chaque couple de trajectoire imposant des temps de division identiques aux deux cellules. Ensuite nous mesurons le nombre de sauts nécessaires pour que $|X_t - Y_t|$ devienne d'abord petite (≤ 0.5) puis négligeable ($\epsilon < 10^{-4}$).
2. Vitesse de convergence : Afin de valider la décroissance exponentielle sans coût de calcul excessif nous simulons directement le processus de comptage N_t sur $M = 2000$ réalisations. Ensuite nous examinons la décroissance moyenne en représentant la distance en échelle logarithmique et comparons les pentes obtenues.
3. Biais d'observation : Sur une trajectoire longue ($T = 20\,000$), nous comparons deux distributions : la population des masses à la naissance (Z_n) extraite aux instants de saut ensuite la population réelle (X_t) en "semblant de continue" avec un pas de temps $\Delta t = 0.05$. Cette expérience permet de visualiser et de quantifier le décalage de masse entre les temporalités discrète et quasi-continue.

Résultats et Interprétation

Illustration de l'oubli de la condition initiale

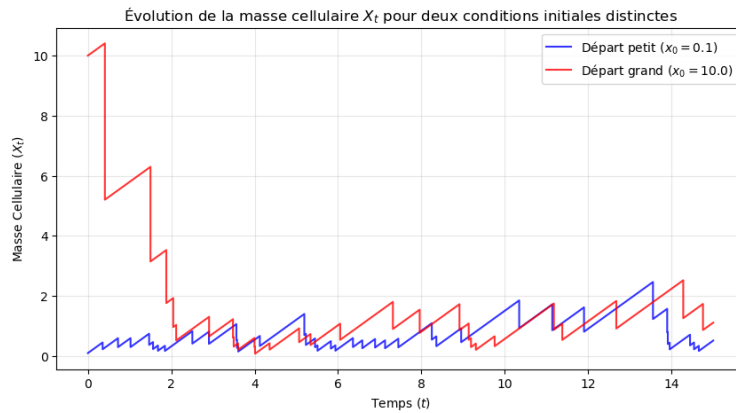


FIGURE 3 – Illustration de la perte de mémoire

Nous avons simulé deux cellules aux conditions initiales opposées avec $x_0 = 0.1$ et $y_0 = 10.0$ ($\Delta_{masse} = 9.9$). La Figure 3 illustre l'oubli progressif de cette condition initiale. En effet on remarque que la trajectoire partant de $y_0 = 10$ rejoint celle partant de $x_0 = 0.1$ en un temps $t \approx 3$. Pour être plus rigoureux nous avons

répété 10 fois cette expériences. En moyenne les trajectoires commencent à se rejoindre à $t \approx 5$ et "fusionnent presque" à $t \approx 17$. Nous remarquons que le nombre d'étapes nécessaires à la convergence semble déterministe alors que le temps de convergence reste stochastique.

Expérience n°	Rapprochement	"Fusion presque parfaite"
1	3.8417 sec	7.3721 sec
2	2.5652 sec	6.1530 sec
3	0.5857 sec	6.5729 sec
4	3.6291 sec	10.6148 sec
5	2.9875 sec	7.5249 sec
6	2.2522 sec	10.4399 sec
7	0.6676 sec	7.0978 sec
8	1.3026 sec	8.6686 sec
9	1.7581 sec	7.8378 sec
10	3.5294 sec	8.9804 sec

TABLE 4 – Données expérimentales de couplage sur 10 simulations

Vitesse de Convergence

Pour valider la nature exponentielle de cette convergence proposée par le théorème 3 nous avons analysé l'évolution de la distance moyenne.

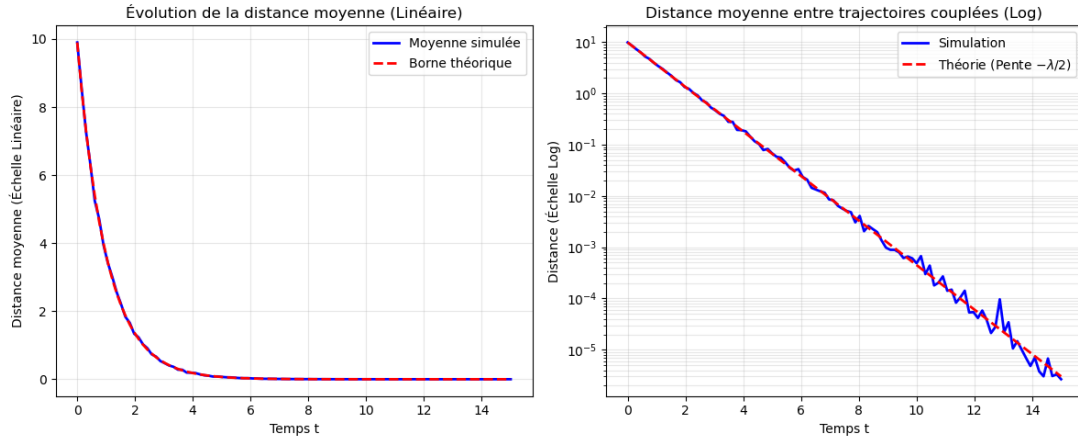


FIGURE 4 – Illustration de la vitesse de la perte de mémoire

La régression linéaire que nous avons fait sur la distance moyenne (échelle log) fournit les résultats suivants qui illustrent parfaitement la vitesse de convergence proposé par le Théorème 3. :

- Pente théorique $-\frac{\lambda}{2}$: -1.0000
- Pente simulée : -1.0049
- Précision : Erreur de 0.49 % ce qui confirme que la vitesse de convergence énoncée par le théorème

Validation de la convergence de la masse en temps discret

Nous nous sommes d'abord intéressés à la chaîne de Markov $(Z_n)_{n \geq 0}$ observée aux instants de naissance (juste après chaque division). Nous voyons que l'histogramme empirique se superpose avec la densité théorique $\tilde{\mu}$ prédite par le Théorème 2 (Figure 5). Cette expérience confirme que la loi stationnaire théorique décrit fidèlement le comportement observé.

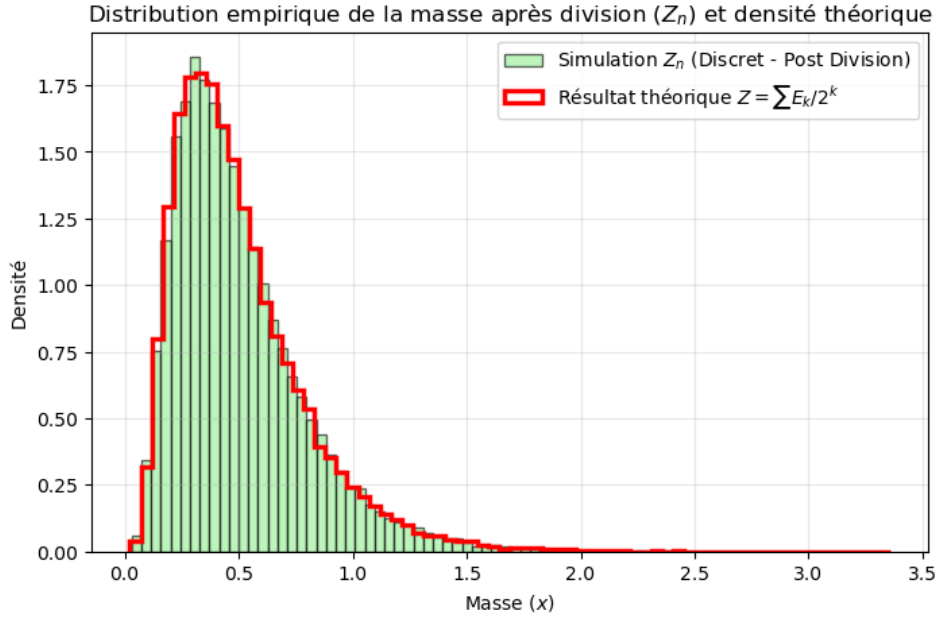


FIGURE 5 – Adéquation entre la distribution simulée des masses post-division Z_n et la densité théorique.

Comme un graphique ne suffit pas, l'analyse numérique des moments montre que les estimateurs de Monte-Carlo convergent précisément vers les valeurs cibles théoriques.

Moment	Simulé	Théorique	Erreur
Espérance ($\mathbb{E}[Z]$)	0.4995	$\frac{1}{\lambda} = 0.5000$	0.10 %
Variance ($\text{Var}[Z]$)	0.0831	$\frac{1}{3\lambda^2} = 0.0833$	0.29 %

Cela montre que le générateur aléatoire est fiable à long terme et ne présente pas de biais. Ainsi nous avons pu illustrer un théorème qui paraissait jusque-là obscur.

Le Biais d'Observation : Continu & Discret

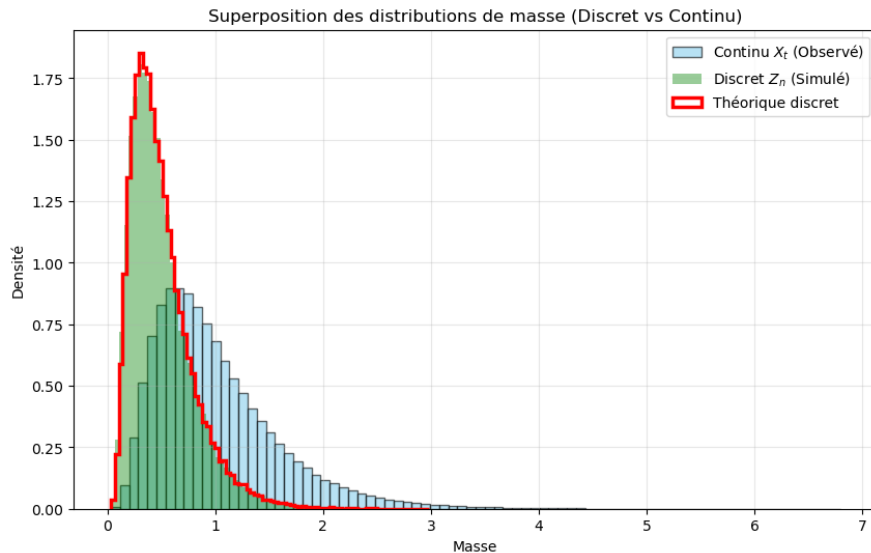


FIGURE 6 – Visualisation du biais d'échantillonnage temporel

En biologie on n'observe pas les cellules uniquement à leur naissance mais à un instant t quelconque. Notre expérience qui consiste à reconstruire le signal continu (X_t) pour comparer la « population réelle » (continue) à la « population des naissances » (discrète) met en évidence un biais important (Figure 6). Ce décalage n'est pas une erreur mais ça illustre le biais d'échantillonnage temporel. Les différentes moyennes calculées

montrent clairement l'ampleur de ce phénomène :

- Moyenne Discrète (Z_n) : 0.5009 (Masse à la naissance)
- Moyenne Continue (X_t) : 1.0053 (Masse moyenne de la population)

Nous remarquons naturellement que : Moyenne Continu (X_t) = 1.0053 $\approx 2 \times$ Moyenne Discret (Z_n)

Les simulations montrent clairement que la loi invariante du processus continu μ diffère de celle de la chaîne discrète $\tilde{\mu}$ et la raison est intuitive. En effet entre deux divisions les cellules croissent donc une observation à un instant choisi au hasard a plus de chances d'attraper une cellule en phase avancée de croissance. En moyenne une cellule observée « au hasard » pèse environ deux fois plus que juste après sa naissance ce qui explique le décalage entre μ et $\tilde{\mu}$.

Conclusion

L'influence des conditions initiales s'efface rapidement. En effet la distance entre les trajectoires décroît de façon exponentielle, conformément au Théorème 3 avec une erreur d'environ 0,49%.

Par ailleurs, la comparaison entre les temporalités discrète et continue met en évidence un biais d'échantillonnage. En temps continu la masse moyenne observée est approximativement le double de celle mesurée juste après la naissance ce qui explique le décalage entre les lois invariantes μ et $\tilde{\mu}$.

2.3 Explosion Démographique et Processus de Comptage

Jusqu'à présent nous avons modélisé la trajectoire d'une cellule unique. Nous changeons désormais d'échelle pour étudier l'arbre généalogique complet.

Objectif et Motivation

D'une part nous voulons illustrer la croissance exponentielle de la population et à vérifier que le nombre d'individus suit une loi géométrique à tout instant t (Théorème 4). D'autre part, nous chercherons à vérifier la convergence de la variable $H_t = N_t e^{-\lambda t}$ vers sa limite théorique qui est $\mathcal{E}(1)$. Cela va permettre de confirmer que le "destin" de la colonie se joue dans les premiers instants et se stabilise.

Protocole Expérimental

Nous fixons le taux de division à $\lambda = 2.0$. Comme la simulation d'une population entière est très coûteuse car nous devons stocker tout l'arbre. De ce fait nous utilisons la perte de mémoire pour accélérer le calcul via le lemme du minimum. Voici les différentes étapes algorithmiques que nous avons mises en place :

1. Simulation de la population : Dans une population de taille n le temps d'attente avant la prochaine division suit une loi $\mathcal{E}(n\lambda)$. On avance le temps d'un pas $\Delta t \sim \mathcal{E}(n\lambda)$ à chaque itération et on incrémente la population globale ($n \leftarrow n + 1$)
2. Dynamique temporelle ($t \in [0, 4]$) : Nous simulons $M = 200$ trajectoires les projetons sur une grille temporelle commune par interpolation ce qui permet de calculer l'erreur moyenne entre la croissance empirique et la courbe théorique $e^{\lambda t} - 1$.
3. Distribution de la population à t fixé : Pour un temps d'observation $T_{obs} = 2.0$ et sur $M = 5000$ échantillons nous comparons l'histogramme des populations finales et ses moments (moyenne, variance) avec la loi Géométrique théorique de paramètre $p = e^{-\lambda T_{observations}}$

4. Convergence en loi de H_t : Nous étudions le comportement à long terme ($T_{lim} = 5.0$) de la variable $H_t = N_t e^{-\lambda t}$. La validation de la convergence en loi vers $\mathcal{E}(1)$ se fait par superposition d'histogrammes/densités et par comparaison des moments empiriques aux valeurs théoriques.

Résultats et Interprétation

Explosion démographique

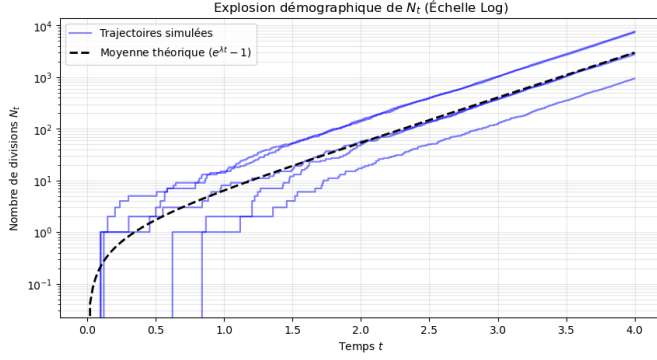


FIGURE 7 – Croissance N_t (Échelle Log)

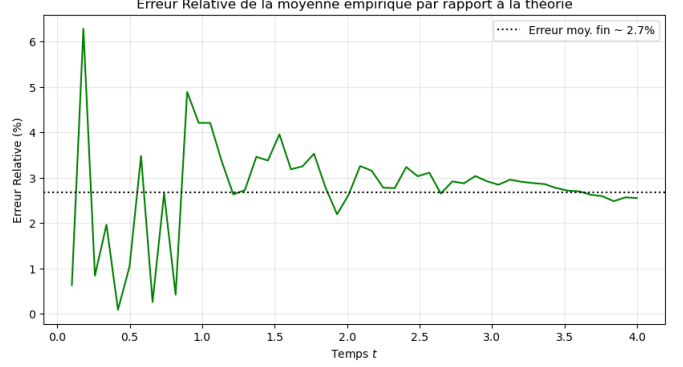


FIGURE 8 – Stabilité de l'erreur relative

La Figure 7 illustre clairement le théorème de croissance exponentielle. En échelle semi-logarithmique les trajectoires simulées oscillent autour d'une droite qui est la moyenne théorique. Cela confirme visuellement que la population croît bien de façon exponentielle. La Figure 8 valide la robustesse de notre modélisation. En effet l'écart entre la moyenne empirique et $e^{\lambda t} - 1$ reste faible et se stabilise autour de 2,7%, ce qui est tout à fait satisfaisant au regard des fluctuations attendues.

Validation de la Loi Géométrique

Pour vérifier que la distribution du nombre d'individus suit la prédiction théorique nous avons montré que la dispersion observée n'est pas arbitraire mais obéit à une loi géométrique. L'alignement des fréquences simulées et théoriques confirme la formule $P(N_t = k) = e^{-\lambda t}(1 - e^{-\lambda t})^{k-1}$ pour $k \geq 1$.

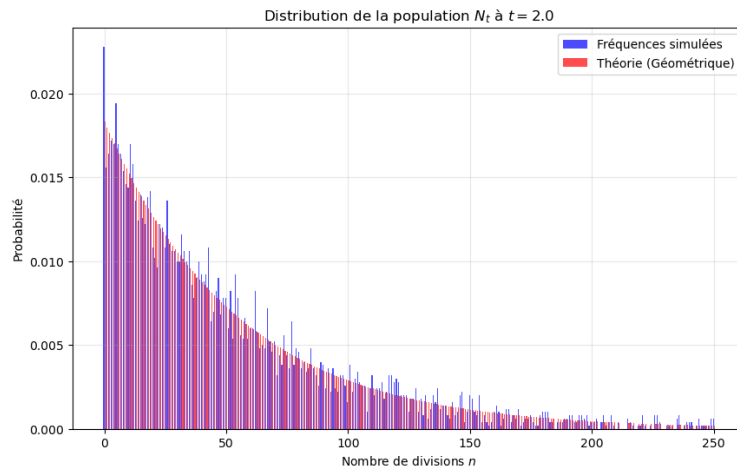


FIGURE 9 – Distribution de la population à $t = 2.0$

Cet alignement visuel est aussi confirmée par une analyse quantitative des moments. L'espérance calculée sur nos 2000 simulations est proche de la valeur théorique comme on peut le constater dans le tableau suivant.

Indicateur	Simulé	Théorique	Erreur Relative
Moyenne ($\bar{N}_{t=2}$)	53.4742	53.5982	0.23 %

Une erreur de seulement 0.23% valide avec précision le Théorème 4. Bien que le processus soit continu dans le temps la loi du nombre d'individus est bien géométrique.

Convergence en loi de la variable aléatoire H_t

Bien que N_t tend vers l'infini, nos simulations montrent que la variable $H_t = N_t e^{-\lambda t}$ converge presque sûrement. En effet la Figure de gauche montre qu'après une phase chaotique chaque trajectoire se stabilise vers une valeur limite aléatoire notée H .

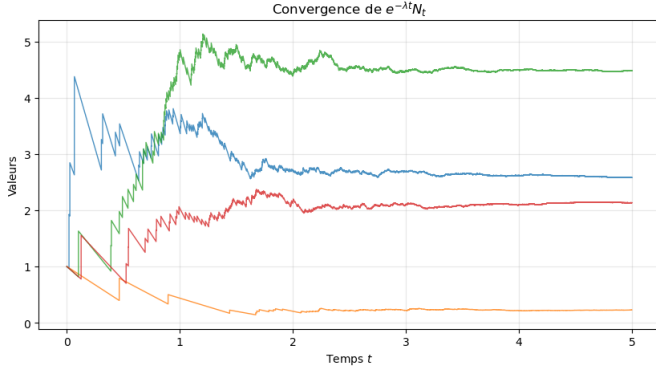


FIGURE 10 – Stabilisation des trajectoires H_t

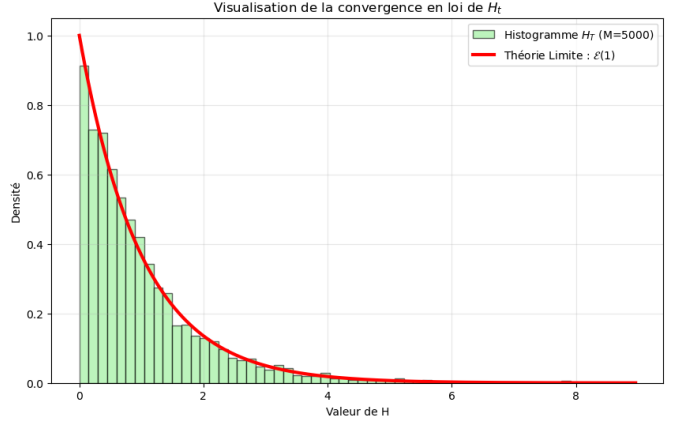


FIGURE 11 – Loi limite $H \sim \mathcal{E}(1)$

Moment (H_T)	Simulé	Cible ($\mathcal{E}(1)$)	Erreur (%)
Moyenne	1.0077	1.0000	0.77 %
Variance	1.0302	1.0000	3.02 %

TABLE 5 – Validation des moments de la variable H

L'histogramme de droite montre que la limite H suit la loi attendue et c'est ce que confirme l'analyse des moments (Tableau 5). Ces résultats apportent une validation quantitative de la convergence en loi $H_t \xrightarrow{\mathcal{L}} \mathcal{E}(1)$.

Conclusion

Cette expérience valide numériquement l'explosion démographique et la structure probabiliste du processus. D'abord, la dynamique moyenne suit bien une croissance exponentielle : l'écart relatif se stabilise autour de 2,7 %, ce qui témoigne de la robustesse de l'algorithme sur le long terme. Ensuite à un instant fixé la population ne se disperse pas au hasard mais obéit à une loi géométrique confirmée par une erreur minime de 0.23% sur l'espérance.

Enfin la variable aléatoire $H_t \xrightarrow{\mathcal{L}} \mathcal{E}(1)$ avec un écart sur la moyenne d'environ 0.77% ce qui signifie que l'aléa du système se « fige » rapidement. Nous pouvons conclure que la taille absolue reste imprévisible mais elle est déterminée à une constante multiplicative près par les premières divisions.

3 Conclusion

Objectif

Le but de ce projet était de modéliser la croissance cellulaire en confrontant la théorie des processus stochastiques à des simulations de Monte-Carlo tout en racontant une histoire et mettant l'accent sur les expériences numériques. En trois expériences numériques nous avons non seulement confirmé les prédictions théoriques mais surtout mis en lumière certains biais statistiques importants pour interpréter des observations biologiques.

Résultats

On a d'abord étudié l'inférence et la correction du biais. L'estimateur du maximum de vraisemblance respecte bien les résultats asymptotiques dès que l'échantillon est suffisamment grand ($n \geq 1000$ avec une erreur $< 1,8\%$) mais il surestime pour de petits effectifs. Dans ces conditions l'estimateur corrigé $\tilde{\lambda}_n$ devient nécessaire pour obtenir des mesures fiables dans la réalité biologique.

Ensuite le contraste entre observation continue et discret (observation au moment de la naissance) révèle un paradoxe d'échantillonnage. L'étude de la dynamique de masse confirme l'oubli de la condition initiale et montre surtout que la masse moyenne observée en continu (1,005) est environ le double de celle mesurée juste après la naissance (0,501). Autrement dit un relevé pris à un instant aléatoire a plus de chances de capter des cellules âgées et plus massives ce qui crée un biais systématique à prendre en compte.

Enfin malgré l'imprévisibilité des divisions individuelles la variable H_t converge vers une loi limite $\mathcal{E}(1)$ avec une bonne précision (écart sur la moyenne $\approx 0,77\%$). De ce fait l'incertitude macroscopique se fige tôt dans l'histoire de la colonie ce qui fait en sorte que la trajectoire absolue reste aléatoire mais sa forme relative est déterminée par une constante aléatoire fixée dès les premières divisions.

Limites

Notre modèle suppose une croissance exponentielle sans limite ce qui est loin de notre réalité. Nous ne disposons pas d'espace infini ni de temps infini, on est vite contraint. Pour rendre la modélisation plus fidèle, on peut modifier notre modèle pour lui donner un facteur d'épuisement avec le temps et sa masse, ce qui se rapproche le plus des vraies expériences biologiques. De plus cette loi exponentielle qui est sans mémoire ne représente pas la réalité du cycle cellulaire car une cellule ne peut pas se diviser immédiatement après sa naissance, mais dans notre modèle c'est tout à fait possible. Nous pouvons intégrer des étapes supplémentaires pour rendre le modèle plus réaliste.

Références

- [1] Jury de l'Agrégation Externe de Mathématiques, *Épreuve de modélisation, option A : Probabilités et Statistiques (Sujet A1)*, Session 2019.
- [2] R. Diel, *Cours et Travaux Pratiques de Méthodes de Monte-Carlo et Chaînes de Markov*, Université Côte d'Azur, 2024-2025.
- [3] A. Godichon-Baggioni, *Statistique Inférentielle (L3)*, Sorbonne Université, basée sur le cours de A. Guyader, 2020-2021.
- [4] Y. Laguel, *Do's and don'ts of mathematical writing*, Université Côte d'Azur, Janvier 2025.
- [5] V. Deshors, *Yule Process Simulation (Implémentation Python)*, GitHub Repository.
Disponible sur : https://github.com/VictorDeshors/Yule_Process

Annexes

A Annexe : Mentions Légales et Crédits

Utilisation de l'IA

Pour ce projet nous avons utilisé l'intelligence artificielle pour corriger les erreurs d'orthographe et pour l'écriture de certains passages du rapport afin de faciliter la lecture notamment pour faire des transitions fluides. Nous l'avons aussi utilisée pour générer ou optimiser certains passages de code de notre simulation (notamment les graphiques). Nous avons pris soin de vérifier rigoureusement tout le contenu généré et de notifier la partie qui est entièrement générée par l'IA pour garantir une totale transparence dans notre démarche de mathématiciens en devenir.

Inspirations

Nous nous sommes inspirés de certains programmes open-source disponibles sur Github, et remercions particulièrement l'auteur du dépôt suivant : https://github.com/VictorDeshors/Yule_Process.

Le cœur du code a été développé par nos soins en suivant les méthodologies des TPs et les conseils de nos professeurs M. Diel et M. Laguel. Le reste du rapport est un travail personnel.

B Annexe : Compléments Mathématiques et Preuves

Cette annexe détaille les justifications théoriques.

Propriété d'absence de mémoire

Le temps inter-divisions suit une loi exponentielle. Cette propriété est appelée *absence de mémoire* et s'exprime de la manière suivante :

$$\forall s, t > 0, \quad \mathbb{P}(T > t + s \mid T > t) = \mathbb{P}(T > s). \quad (15)$$

Loi de l'instant de division (T_n)

On s'intéresse à l'instant de la n -ième division noté T_n . Cet instant correspond simplement à la somme des durées de vie successives depuis la cellule mère. On peut donc écrire :

$$T_n = \sum_{k=1}^n E_k. \quad (16)$$

Comme les variables (E_k) sont indépendantes et identiquement distribuées (i.i.d.), la fonction caractéristique de leur somme s'obtient simplement en multipliant celles de chaque terme. Nous obtenons alors :

$$\begin{aligned} \phi_{T_n}(x) &= \mathbb{E} \left[\exp \left(ix \sum_{k=1}^n E_k \right) \right] = \prod_{k=1}^n \mathbb{E} [e^{ixE_k}] \\ &= \prod_{k=1}^n \left(1 - \frac{ix}{\lambda} \right)^{-1} \\ &= \left(1 - \frac{ix}{\lambda} \right)^{-n}. \end{aligned} \quad (17)$$

Cette expression correspond à la fonction caractéristique d'une loi Gamma, notée $\Gamma(n, \lambda)$.

Effectuons le changement de variable $Y_k = \lambda E_k$. Cherchons la loi de Y_k par calcul direct de sa fonction de répartition :

$$F_{Y_k}(y) = \mathbb{P}(Y_k \leq y) = \mathbb{P}(\lambda E_k \leq y) = \mathbb{P}\left(E_k \leq \frac{y}{\lambda}\right).$$

Or comme $E_k \sim \mathcal{E}(\lambda)$ alors $\mathbb{P}(E_k \leq t) = 1 - e^{-\lambda t}$. On en déduit :

$$F_{Y_k}(y) = 1 - e^{-\lambda \frac{y}{\lambda}} = 1 - e^{-y}.$$

On reconnaît ici la fonction de répartition d'une loi Exponentielle de paramètre 1 : $Y_k \sim \mathcal{E}(1)$.

Par définition, la somme de n variables exponentielles indépendantes de paramètre 1 suit une loi Gamma $\Gamma(n, 1)$. On peut donc conclure que :

$$\lambda T_n = \sum_{k=1}^n \lambda E_k = \sum_{k=1}^n Y_k \sim \Gamma(n, 1).$$

Construction de l'intervalle de confiance exact

Comme $\lambda T_n \sim \Gamma(n, 1)$ nous pouvons construire un intervalle de confiance exact au niveau $1 - \alpha$ en utilisant les quantiles $q_{\alpha/2}$ et $q_{1-\alpha/2}$ de la loi $\Gamma(n, 1)$.

On part de l'encadrement probabiliste :

$$q_{\alpha/2} < \lambda T_n < q_{1-\alpha/2} \iff \frac{q_{\alpha/2}}{T_n} < \lambda < \frac{q_{1-\alpha/2}}{T_n}$$

Ce qui nous donne l'expression finale de l'intervalle :

$$I_{n,exact} = \left[\frac{q_{\alpha/2}}{T_n}, \frac{q_{1-\alpha/2}}{T_n} \right] \quad (18)$$

Récurrence de la masse discrète (Z_n)

La masse Z_n mesurée juste après la n -ième division. Entre les divisions n et $n + 1$, la cellule croît pendant une durée E_{n+1} , atteignant $Z_n + E_{n+1}$ juste avant de se diviser. La division réduit ensuite cette masse de moitié. En répétant ce raisonnement dès $n = 0$:

$$\begin{aligned} Z_1 &= \frac{1}{2}Z_0 + \frac{1}{2}E_1 \\ Z_2 &= \frac{1}{2}Z_1 + \frac{1}{2}E_2 = \frac{1}{4}Z_0 + \frac{1}{4}E_1 + \frac{1}{2}E_2 \end{aligned}$$

En généralisant ce procédé par récurrence, nous obtenons l'expression de la masse après n divisions :

$$Z_n = \frac{1}{2}(Z_{n-1} + E_n) = \frac{1}{2^n}Z_0 + \sum_{k=1}^n \frac{1}{2^k}E_{n-k+1} \quad (19)$$

Convergence du processus H_t

Introduisons la variable $H_t = e^{-\lambda t}(N_t + 1)$ et cherchons de montrer qu'elle converge en loi en utilisant les propriétés des fonctions caractéristiques.

$$\varphi_{H_t}(u) = e^{-\lambda t} e^{iue^{-\lambda t}} \sum_{j=0}^{\infty} \left(e^{iue^{-\lambda t}} (1 - e^{-\lambda t}) \right)^j = \frac{e^{-\lambda t} e^{iue^{-\lambda t}}}{1 - (1 - e^{-\lambda t}) e^{iue^{-\lambda t}}}$$

Effectuons un développement limité à l'ordre 1 lorsque $t \rightarrow \infty$:

$$e^{iue^{-\lambda t}} = 1 + iue^{-\lambda t} + o(e^{-\lambda t}) \quad \text{et} \quad 1 - (1 - e^{-\lambda t}) e^{iue^{-\lambda t}} = e^{-\lambda t} (1 - iu) + o(e^{-\lambda t})$$

D'où :

$$\lim_{t \rightarrow \infty} \varphi_{H_t}(u) = \lim_{t \rightarrow \infty} \frac{e^{-\lambda t} (1 + o(1))}{e^{-\lambda t} (1 - iu) + o(e^{-\lambda t})} = \frac{1}{1 - iu}$$

On reconnaît la fonction caractéristique d'une loi Exponentielle $\mathcal{E}(1)$: $H_t \xrightarrow{\mathcal{L}} \mathcal{E}(1)$.