Random Models of Dynamical Systems Introduction to SDE's TP 2 : Numerical simulation of stochastic differential equations Wright–Fisher diffusion approximation

Baptiste CALLARD & Malo TANNÉ

10 janvier 2022

Table des matières

I Introduction

I.1 Motivations et objectifs

L'objet de ce TP est de montrer l'efficacité de l'approximation du principe de diffusion. Nous allons faire deux études d'approximation d'un temps d'arrêt en utilisant ou non le principe de diffusion. Puis nous justifierons l'utilisation de cette approximation, plutôt qu'une méthode plus directe.
L'objectif est d'illustrer ce propos sur un modèle relativement simple : transmission d'allèles dans une population. Considérons le cas d'un gène avec deux allèles A et a. La population possédera N$<$ *cste* individus. Pour toutes les générations k, un individu hérite son allèle de ses parents de la génération k-1. La transmission de l'allèle est modélisée par un tirage qui se fait de manière aléatoire et uniforme (avec remplacement) parmi ses parents.

I.2 Modélisation du phénomène étudié

De manière plus formelle, on note $Y_k^N \in \{0, 1, ..., N\}$, le nombre d'allèles de type A présent dans la génération k. Ce processus peut être modélisé par une chaîne de Markov de la forme :

$$\pi_{i,j}^N = \mathbb{P}[Y_k^N = j | Y_{k-1}^N = i] = \binom{N}{j} (\frac{i}{N})^j (1 - \frac{i}{N})^{N-j} \ , j \in \{0, ..., N\}$$

Plutôt que de travailler avec le nombre d'allèles, on préfère raisonner en terme de proportion car cela est plus compréhensible et interprétable. On effet, cela permet de faire des comparaisons relativement à une population donnée. En effet, sinon si l'on dit qu'il y a 5000 personnes qui ont des allèles A mais que l'on ne donne pas la population totale alors cela permet du sens. On introduit donc le processus aléatoire normalisé $X_k^N = \frac{Y_k^N}{N}$. $X_k^N \in \{0, \frac{1}{N}, ..., \frac{N-1}{N}, 1\}$. On introduit également le processus de Markov linéaire et continue suivant :

$$X^N(t) = X_{k-1}^N + (X_k^N - X_{k-1}^N)(N\ t - (k-1)) \text{ qui interpole le point } X_k^N \text{ à l'instant } t_k^N = \frac{k}{N}$$

II Simulation du processus de Markov

Nous allons dans un premier temps simuler des trajectoires du processus de Markov entre l'instante t = 0 et t = T. Pour ce faire nous utilisons étant donné N la probabilité de transité de l'état i vers l'état j : π_{ij}^N.

La méthode est la suivante, on se donne un état initial.

Connaissant l'état occupé, nous pouvons tirer aléatoirement le prochain état occupé grâce aux probabilités de la matrice de transition associé à l'état actuellement occupé. Ainsi, nous sommes capables en itérant ce procédé d'obtenir une trajectoire de la variable aléatoire $Y_1^N, ..., Y_{Nt}^N$. Ensuite, il suffit de normaliser le processus pour obtenir le processus $X_1^n, ..., X_{Nt}^N$.

Nous pourrons ensuite obtenir le processus linéaire et continue par morceaux $X^N(t)$ sur l'intervalle [0,T] en faisant une interpolation linéaire sur les différents intervalles.

Nous avons représenté le processus discret sans l'interpolation puis le processus "continue" avec l'interpolation linéaire. Nous voulions vérifier et observer que les deux trajectoires se confondent.

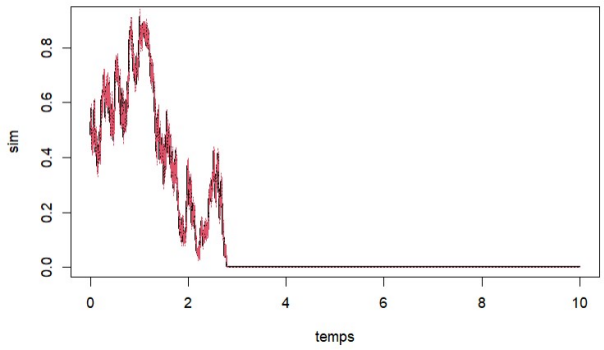
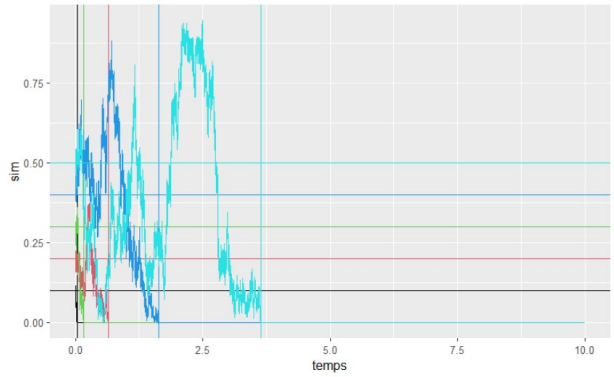


FIGURE 1

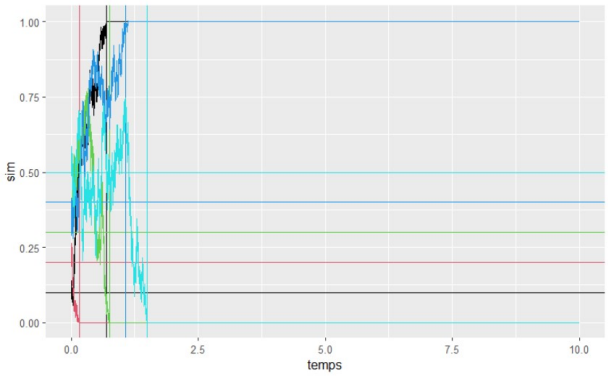
Nous pouvons voir que le processus continue $X^N(t)$ permet d'avoir un processus plus riche en gardant la même trajectoire pour les points du processus initiale $(X_k^N)_{k\in\{0,\dots,Nt\}}$. Cela pourra être utile dans la suite par exemple pour détecter précisément le temps pour lequel le processus atteint un état donné. En effet, plus nous possédons de points et plus l'on est précis pour donner le temps pour atteindre un bord par exemple.

Nous allons maintenant tracer plusieurs trajectoires pour T = 10. Cela nous permettra d'introduire les prochaines sections ainsi que l'enjeu de ce TP. Ainsi, nous allons observer l'influence de la condition initiale sur le temps pour atteindre les états absorbants 0 ou 1.

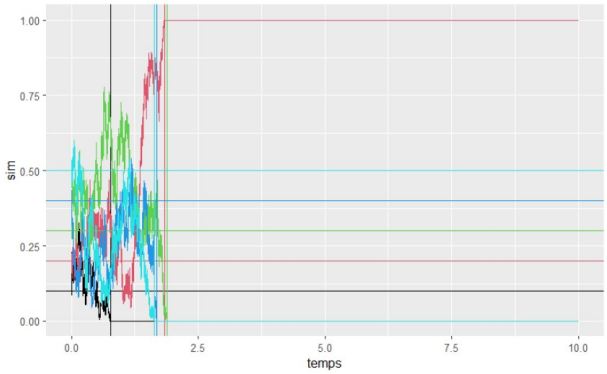
Dans les graphiques suivants, les lignes verticales correspondent au moment où un bord est atteint. Les lignes horizontales quant à elles correspondent à la condition initiale de la trajectoire associée (elles servent à mieux visualiser d'où commencent les trajectoires).



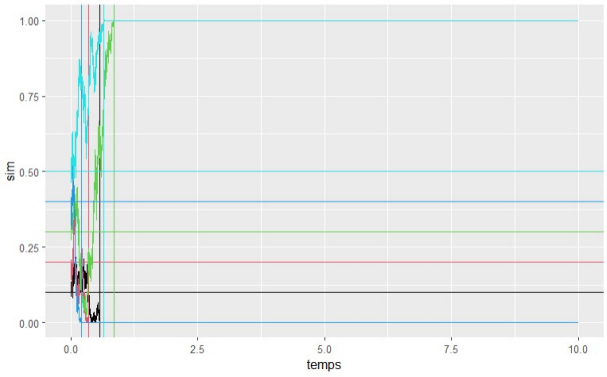
(a) Avec N = 1000



(b) Avec N = 2000



(c) Avec N = 5000



(d) Avec N = 10000

FIGURE 2 – En noir $x_0 = 0.1$, en rouge $x_0 = 0.2$, en vert $x_0 = 0.3$, en bleu $x_0 = 0.4$ et en bleu clair $x_0 = 0.5$

Nous pouvons voir qu'en fonction de la condition initiale, le processus à tendance à rejoindre le bord qui lui est le plus proche. De plus, il met plus de temps à atteindre un bord s'il se situe avec une population homogène, avec autant d'allèles a que A ($x_0 = 0.5$). Par contre si on a une population avec une grande majorité de A (resp a) alors on aura tendance à atteindre rapidement le bord 1 (resp 0).
Le fait de prendre une discrétisation plus fine (ie N grand) fait que l'on détecte plus tôt si le processus atteint l'état stationnaire. En effet, on ne pourra le détecter que les $\frac{1}{N}$ et donc si N est petit alors on le détectera que potentiellement plus tard. Donc pour les maillages grossiers, on sur-estime systématiquement les temps d'arrêt. Il faut donc prendre une valeur de N suffisamment grande pour ne pas faire une étude biaisée.

III Comportement asymptotique et SDE

III.1 Comportement asymptotique

Nous avons vu lors des TDs que lorsque $N \to +\infty$ alors le processus de Markov $X^N(t)$ interpolant X_k^N à l'instance $t_k^N = \frac{k}{N}$ convergeait en distribution vers la solution de la SDE :

$$X(t) = X(0) + \int_0^t \sqrt{X(s)(1 - X(s))}dB(s)$$

Remarque : Le coefficient de diffusion de cette SDE ne satisfait pas les conditions suffisantes permettant d'avoir l'existence et l'unicité. Cependant, en utilisant des résultats plus généraux, il est possible de prouver que la solution existe et qu'elle est bien unique.

III.2 Résolution numérique de la SDE

Nous allons résoudre la SDE en utilisant le schéma d'Euler perturbé par une variable aléatoire obtenue par la simulation d'un mouvement Brownien.

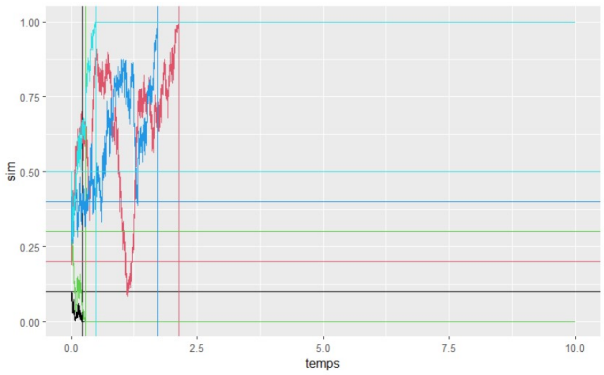
Ainsi, le schéma d'Euler pour notre SDE prend la forme :

$X_k = X_{k-1} + \sqrt{X_{k-1} * (1 - X_{k-1})} * B_k$ avec B_k un tirage d'une variable aléatoire d'une mouvement Brownien

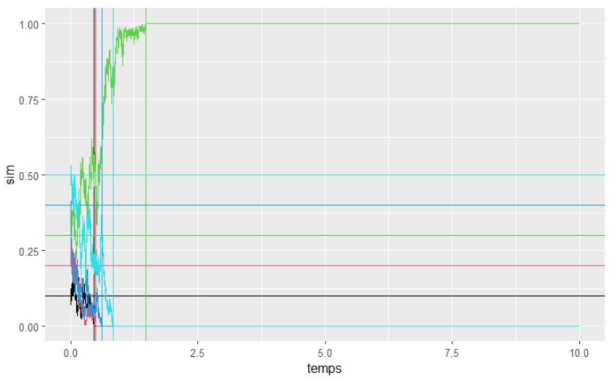
Nous allons maintenant observer des trajectoires de ce processus sur l'intervalle [0,T=10] avec un pas de temps de h. Pour les différentes expériences, les incréments seront tirés indépendamment.

Étude en fonction de l'état initial et de la discrétisation

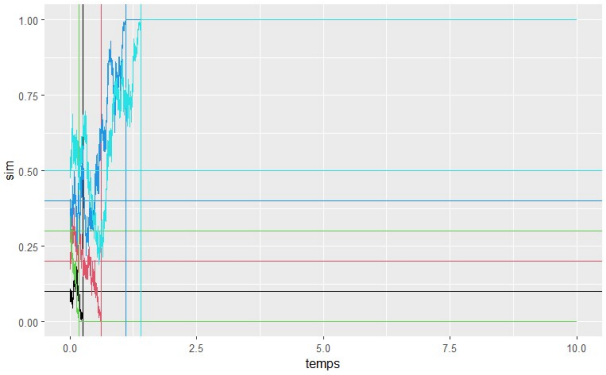
Nous allons observer les trajectoires pour différentes conditions initiales. Nous serons particulièrement attentifs au temps pour atteindre les proportions limites, c'est-à-dire 0 ou 1. Nous allons aussi regarder l'évolution en faisant varier le pas de temps. De plus on a que h = T/N.



(a) Avec h = 10-4



(b) Avec h = 10-5



(c) Avec h = 10-6

FIGURE 3 – En noir $x_0 = 0.1$, en rouge $x_0 = 0.2$, en vert $x_0 = 0.3$, en bleu $x_0 = 0.4$ et en bleu clair $x_0 = 0.5$

Nous avons ensuite regardé la répartition des temps d'arrêts pour atteindre un bord pour une condition initiale x_0= 0.3 pour 10.000 trajectoires et N = 10^5.

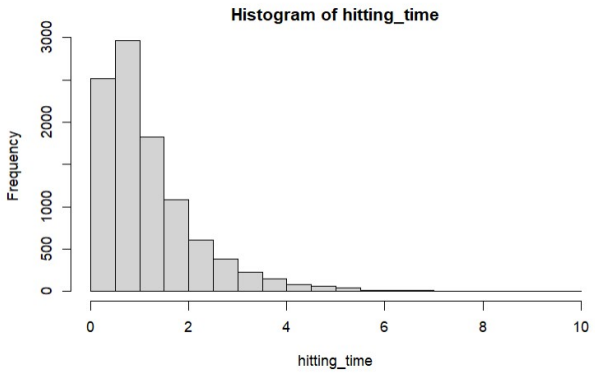


FIGURE 4

En gardant les autres valeurs identiques et en posant $x_0 = 0.1$

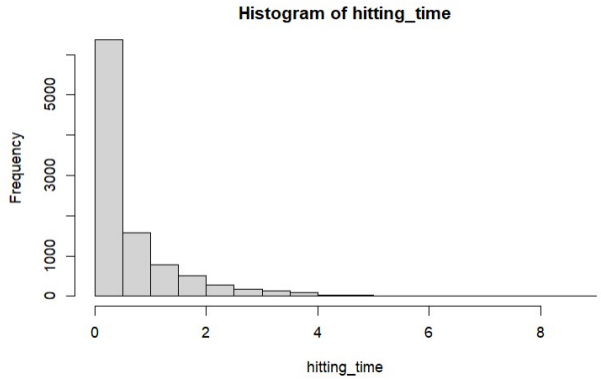


FIGURE 5

En gardant les autres valeurs identiques et en posant $x_0 = 0.5$

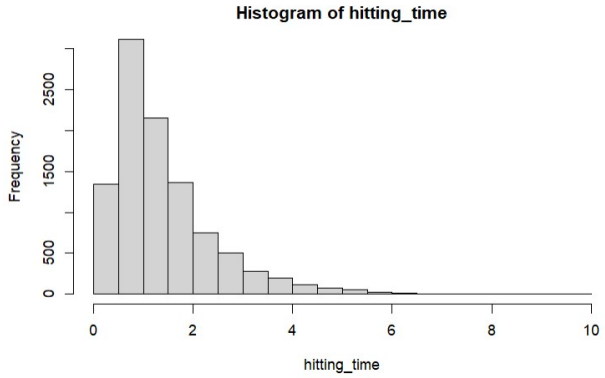


FIGURE 6

Cela confirme l'observation que l'on pouvait faire dans la partie précédente. En effet, si l'on commence par une condition initiale proche d'un bord alors le temps pour l'atteindre sera en général plus court que pour une condition initiale proche de 0.5. Les cas où le bord n'est pas atteint existe, mais il est très rare. En effet, ce n'est jamais le cas pour $x_0 = 0.1$. En revanche pour $x_0 = 0.4$ et $x_0 = 0.5$ il existe des trajectoires qui n'atteignent pas le bord en T=10. Par exemple pour $x_0 = 0.5$ on a 0.1% des trajectoires qui n'atteignent pas de bord en T = 10.

Nous allons mener une étude plus précise dans la partie suivante sur le temps moyen pour atteindre un bord en fonction de la condition initiale.

IV Approximation du temps d'arrêt avec l'approximation du principe de diffusion

IV.1 Estimateur du temps d'arrêt moyen

Nous allons chercher dans cette partie à estimer le temps à partir duquel l'intégralité de la population possède le même allèle (a ou A). On introduit τ^N :

$$\tau^N = inf\{k \geq 0 : Y_k^N = 0 \ ou \ Y_k^N = N\}/N$$
$$\tau^N = inf\{k \geq 0 : X_k^N = 0 \ ou \ X_k^N = 1\}/N$$
$$\tau^N = inf\{t \geq 0 : X^N(t) = 0 \ ou \ X^N(t) = 1\}$$
$$\tau^N \approx inf\{t \geq 0 : X(t) = 0 \ ou \ X(t) = 1\} = \tau$$

Clairement τ *et* τ^N sont des temps d'arrêt. En effet,

$$\{\tau^N \leq t\} = \{\exists \ t_0 \leq t \ tel \ que \ X^N(t_0) = 0 \ ou \ X^N(t_0) = 1\} \in \mathcal{F}(t)$$

De même,

$$\{\tau \leq t\} \in \mathcal{F}(t)$$

Le principe de diffusion donne une approximation du temps moyen pour atteindre ce temps d'arrêt.
D'où

$$\mathbb{E}_{0,x}[\tau^N] \approx \mathbb{E}_{0,x}[\tau]$$

De plus, nous connaissons la valeur exacte de

$$\mathbb{E}_{0,x}[\tau] = -2[x.log(x) + (1-x)log(1-x)]\forall x \in [0,1]$$

Ainsi, nous allons utiliser cette approximation pour obtenir une bonne approximation du temps moyen pour atteindre les valeurs $\{0,1\}$. En commençant en x_0 en t=0 et en utilisant l'approximation de la méthode de Monte Carlo. En pratique cela consiste au procédé itératif suivant :

On résout M fois la SDE. On stocke pour toutes les solutions les temps pour atteindre les états X(t)=0 ou X(t)=1.

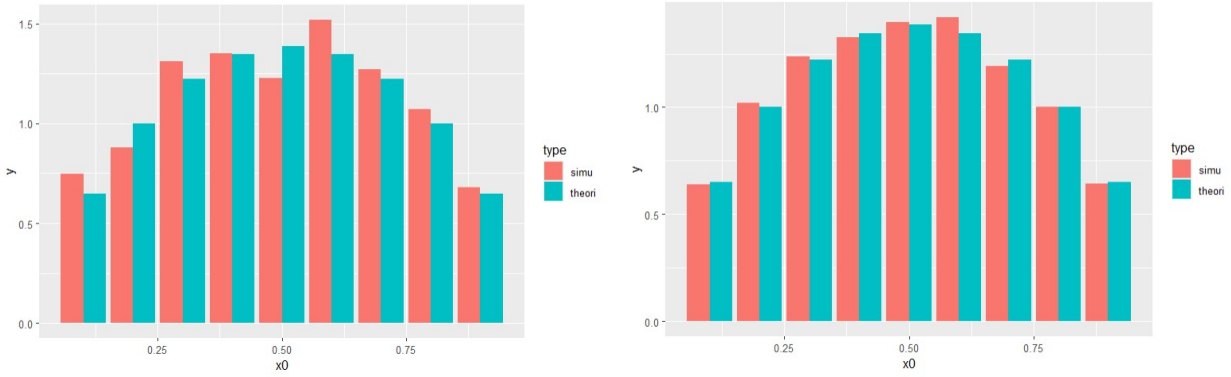
Puis, on en déduit grâce à une moyenne empirique un estimateur de la moyenne théorique $\mathbb{E}_{\mathbb{0},\varkappa}[\tau]$:

$$MHT_x^{h,M} = \frac{1}{M}\sum_{i=1}^{M} \bar{\tau}_i^h$$

IV.2 Estimation du temps d'arrêt moyen

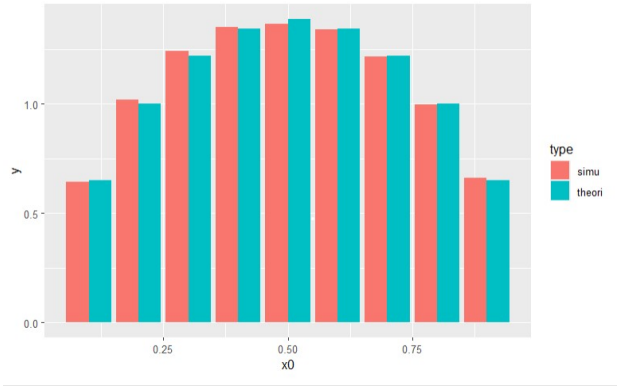
On va utiliser la procédure évoquée plus tôt du temps d'arrêt moyen $\mathbb{E}_{\mathbb{0},\varkappa}[\tau]$ grâce à $MHT_x^{h,M}$. Nous allons faire cette étude en faisant varier une fois de plus la condition initiale, différents pas de temps pour la discrétisation de la méthode d'Euler et pour des tailles d'échantillon M variables (pour la méthode de Monte Carlo).
Nous commençons à regarder pour quelques valeurs de conditions initiales puis nous faisons varier M. On fixe N = 10000.

En abscisse des Bar-plot, on retrouve les conditions initiales. Nous sommes allé de 0.1 à 0.9 avec un pas de 0.1.



(a) M = 100

(b) M = 1000

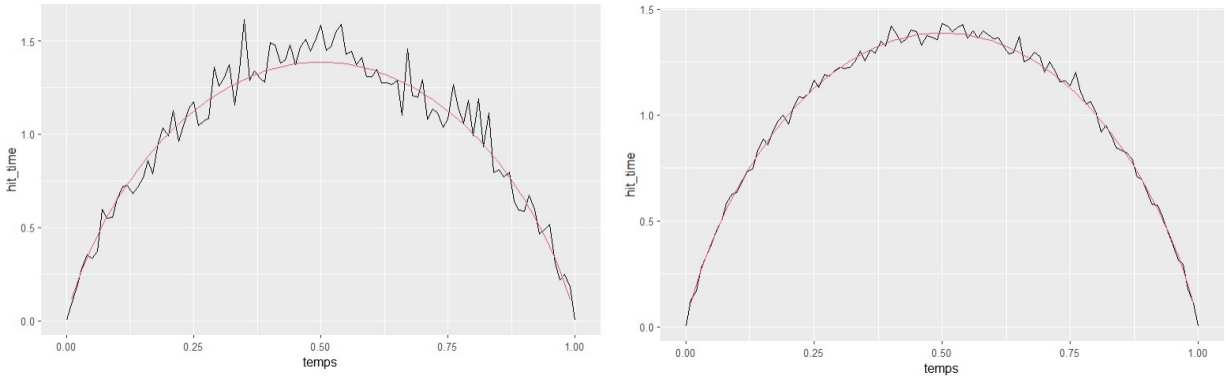


(c) M = 10000

FIGURE 7 – barplot en faisant varier M

On peut clairement voir que plus M est grand et meilleure est la convergence vers la valeur théorique. C'est logique car c'est une résultat valable quand M tend vers ∞. On remarque que pour M $\approx$1000, la convergence est déjà très satisfaisante.

On peut également comparer avec l'espérance théorique. Pour cela on fixe à nouveau N = 10000. Nous avons calculé la moyenne empirique de m expériences de Monte Carlo pour des conditions initiale $x_o \in \{0, 0.01, 0.02, ..., 0.99,$ Le temps d'arrêt moyen en partant d'une condition initiale est donné par la courbe continue rouge sur les graphiques suivants. En noir, nous avons représenté la moyenne des temps d'arrêt empiriques avec des échantillons de M = 100 et M = 1000.



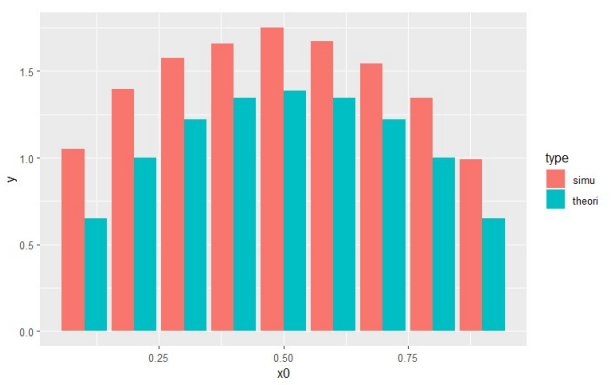
(a) M = 100

(b) M = 1000

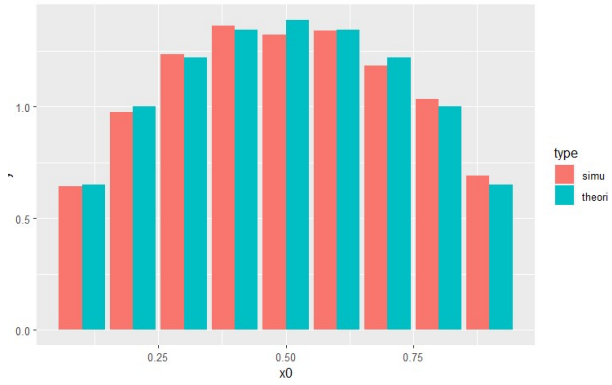
FIGURE 8 – comparaison à l'espérance théorique en faisant varier M

Pour plus de conditions initiales, nous voyons que l'on a bien convergence vers la moyenne théorique quand

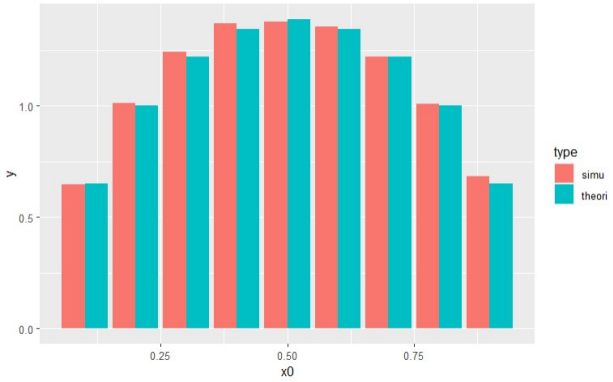
M augmente (pour toutes les conditions initiales).
On peut regarder ce qui se passe en fixant m = 1000 et en faisant varier N :



(a) N = 100



(b) N = 1000



(c) N = 10000

FIGURE 9 – barplot en faisant varier N

Nous voyons que si N n'est pas assez grand alors on détecte très mal le temps d'arrêt. On le sur-estime, c'est compréhensible car plus la discrétisation est fine et plus l'on détectera finement le moment où l'on n'est plus dans [0,1]. Ainsi si pour toutes les trajectoires on sur-estime le temps d'arrêt alors la moyenne sera aussi biaisée. Cela montre la nécessité de prendre un N suffisamment grand (comme évoqué dans la première section). Nous voyons que pour N = 1000, nous avons déjà de bon résultats en revanche pour N = 100 la subdivision est trop grossière.

En combinant les deux études précédents, on voit qu'en prenant N = 1000 et M = 1000 les temps de calcul restent raisonnables et l'on a des résultats de convergences qui semblent cohérents.

V Approximation du temps d'arrêt de l'intégrale de Wright-Fisher

Finalement, nous pouvons voir que la convergence vers la valeur explicite et théorique de l'espérance du temps arrêt est vérifiée pour des M et N convenables. En revanche, nous avons pour cela du faire beaucoup d'approximations. En effet, une première approximation se situe lorsque l'on dit que $\tau^N \approx \tau$ est vérifiée en faisant tendre N vers ∞. Puis nous utilisons le Schéma d'Euler qui constitue une seconde source d'approximation, mais il faut faire tendre h vers 0. Enfin, nous utilisons la méthode de Monte Carlo pour trouver $MHT_x^{h,M}$ avec un M qui tend vers ∞. Ainsi, pour que les approximations soient vérifiées, il faut que : N et M tendent vers ∞ et h tend vers 0.

V.1 Estimateur du temps d'arrêt moyen

On pourrait se dire que l'approximation du temps d'arrêt $\mathbb{E}_{\mathbb{0},x}[\tau^N]$ pourrait se faire directement en utilisant l'approximation de Monte Carlo pour l'intégrale de Wright Fisher(sans l'approximation de diffusion). Cela semble avantageux en terme de d'exactitude, car la méthode de diffusion comporte deux types d'approximation en plus que la méthode plus directe. En pratique, étant donné une condition initiale x_0 au temps t=0 et un échantillon de taille M, la procédure consiste à faire :

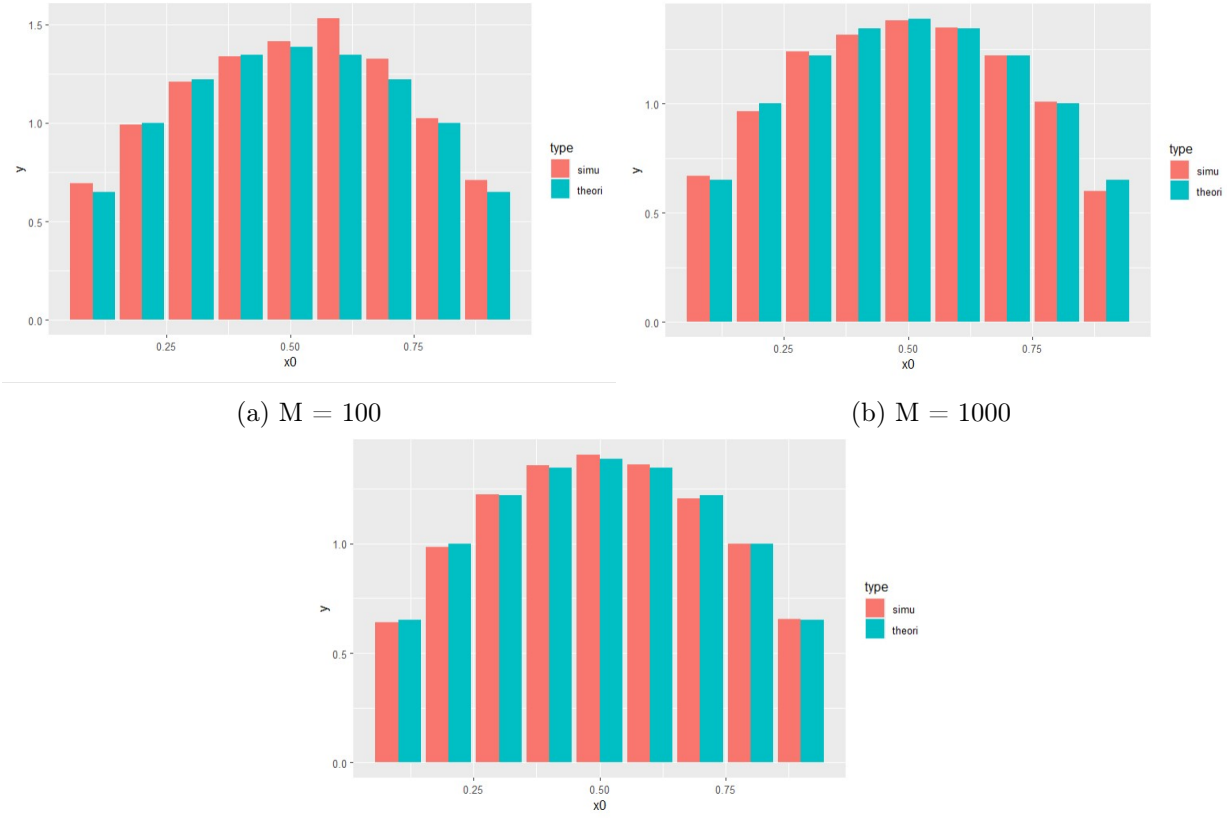
— $\forall i \in \{1, ..., M\}$ on simule la réalisation du processus de Markov et on relève la valeur d'atteinte d'un bord c'est à dire l'état 0 ou 1 par le processus $X_i^N(t)$.

— La moyenne empire donne un estimateur du temps d'arrêt moyen :

$$MHT_x^{N,M} = \frac{1}{M}\sum_{i=0}^{M} \tau_i^N$$

V.2 Estimation du temps d'arrêt moyen

On fait varier M pour un N fixé à 1000 [1].



(a) M = 100

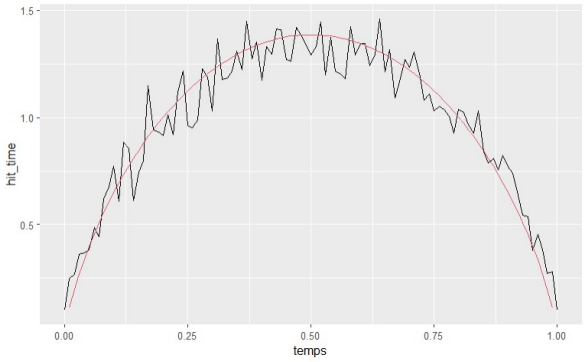
(b) M = 1000

(c) M = 10000

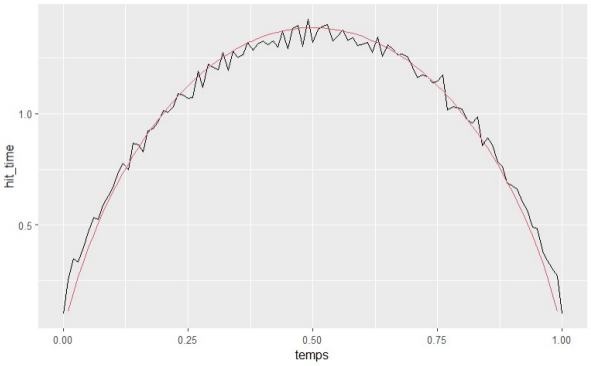
FIGURE 10 – barplot en faisant varier M

Quand M augmente, nous avons une meilleure convergence.
Comme attendu, plus l'échantillon est important et plus l'estimateur est proche de la valeur explicite du temps d'arrêt $\mathbb{E}_{0,x}[\tau]$.

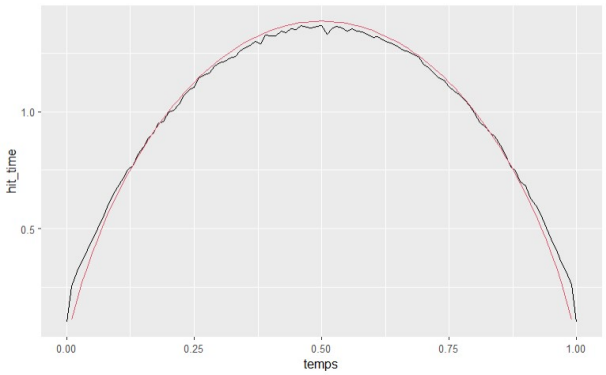
1. et pas à 10000 comme l'étude précédente car les temps sont trop long et l'ordinateur ne peut pas calculer



(a) M = 100



(b) M = 1000



(c) M = 10000

FIGURE 11 – comparaison à l'espérance théorique en faisant varier M

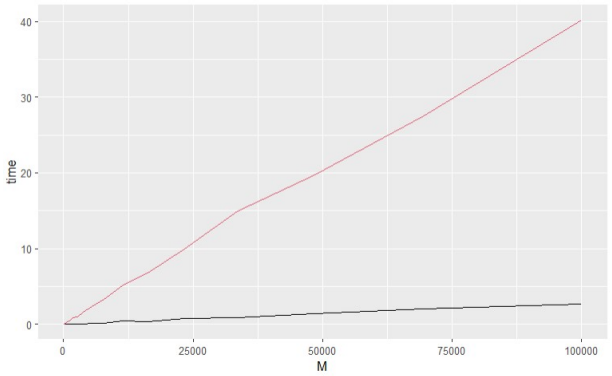
Nous observons le même résultat que précédemment, quand M augmente la convergence est meilleure.

VI Comparaison de temps de calcul des deux méthodes d'approximation du temps d'arrêt moyen

Nous avons précédemment évoqué les avantages de la méthode qui fait une approximation de l'intégrale de Wright-Fisher avec une population finie par "Wright–Fisher model with a finite population size N". En effet, en comparaison de la deuxième méthode, on fait moins d'approximations et donc moins d'erreurs pour un nombre d'itérations équivalents.

Cependant, le fait d'avoir une méthode précise n'est pas la seule chose qui nous intéresse. En effet, le temps de calcul et le stockage sont des paramètres importants en simulation.

C'est pourquoi dans notre étude, nous allons comparer les temps de calcul des deux méthodes. Au premier abord, sans avoir fait aucune étude, on s'attendrait à observer que la méthode avec le plus d'approximation soit la méthode la plus rapide. Cela justifierait les approximations.
Pour N = 100 et M allant de 10^2 à 10^5.



(a) En rouge **Modèle de Wright-Fisher avec une population de taille finie N** et en noir **diffusion limitée de Wright-Fisher**

FIGURE 12 – comparaison des temps de calcul

Nous voyons que la moyenne empirique $MHT_x^{h,M}$ de **diffusion limitée de Wright-Fisher** est bien plus rapide que le modèle de**Modèle de Wright-Fisher avec une population de taille finie N**. Ainsi, ce la montre l'utilité d'utiliser le procédé de diffusion.

VII Conclusion

Pour conclure sur ce TP, nous nous sommes intéressés à l'évolution de la présence d'allèles dans une population au cours des générations. Nous nous sommes rendu compte que nous atteignons un état absorbant en temps fini peu importe la proportion initiale des gènes dans la population. Soit nous avions une proportion de 0% ou 100% de A. Ainsi, l'étude du temps pour atteindre un de ces états en fonction de la condition initiale était logique. Nous avons étudié deux méthodes pour calculer l'espérance du temps d'arrêt. Nous avions une méthode plus directe de l'intégrale de Wright Fisher qui repose sur le processus de Markov. Nous avons également étudié une méthode qui consiste à résoudre une SDE grâce à Euler. Cette deuxième méthode fait plus d'approximation. En revanche, nous obtenons des résultats très satisfaisant en un temps bien plus rapide que la première méthode. Ainsi, nous avons bien mis en avant l'intérêt d'utiliser la simulation d'SDE avec l'approximation de diffusion.