



# Rapport de stage

## Application R Shiny

### par l'exemple du modèle de Wright-Fisher

26 Avril-23 Juillet 2021

Pauline SPINGA

20180998

Université Évre/ Paris-Saclay  
Laboratoire de Mathématiques et Modélisation d'Evry (LaMME)

#### **Encadrement:**

Tuteur de stage  
Vincent RUNGE  
Tuteur enseignant  
Carène RIZZON

10 juin 2021

# Table des matières

<b>1</b>	<b>Travail réalisé</b>	<b>1</b>
1.1	Modèle de Wright-Fisher . . . . .	1
1.1.1	Présentation du modèle . . . . .	1
1.1.2	Modèle sans effet de sélection . . . . .	1
1.1.3	Notations [1] . . . . .	2
1.1.4	Coefficient d'hétérozygotie . . . . .	2
1.1.5	Temps de fixation . . . . .	3
1.1.6	Modèle avec sélection . . . . .	6
1.1.7	Temps de fixation avec sélection . . . . .	6
1.2	App Shiny . . . . .	8
1.2.1	Sans sélection (onglet GenPop) . . . . .	9
1.2.2	Avec sélection (onglet Select GenPop) . . . . .	12
1.2.3	Étude du temps de fixation avec et sans sélection . . . . .	13
1.2.4	Mise à disposition . . . . .	17
<b>2</b>	<b>Bilan et perspectives</b>	<b>18</b>
	<b>Bibliographie</b>	<b>19</b>
<b>A</b>	<b>Annexe</b>	<b>A-1</b>
A.1	Espérance et Variance de $X_{t+1}$ . . . . .	A-1
A.2	Espérance et Variance de $p_{t+1}$ . . . . .	A-1
A.3	Coefficient d'hétérozygotie sans remise . . . . .	A-1
A.4	Coefficient d'hétérozygotie avec remise . . . . .	A-2
A.5	Espérance et Variance de $Z_t$ . . . . .	A-2
A.6	Formule de Taylor . . . . .	A-3
A.7	Espérance et Variance $Z_t$ avec sélection . . . . .	A-3
A.8	Temps de fixation avec sélection . . . . .	A-4
A.9	Application Shiny- Courbes avec effet de sélection . . . . .	A-7
A.10	Variance du temps de fixation . . . . .	A-9

# Table des figures

1.1	Panel . . . . .	8
1.2	Onglets . . . . .	9
1.3	Évolution du nombre d'allèles $A$ en fonction du temps. . . . .	9
1.4	Évolution du coefficient $H$ en fonction du temps . . . . .	10
1.5	Évolution du nombre d'allèles $A$ en fonction du temps . . . . .	10
1.6	Évolution du coefficient $H$ en fonction du temps . . . . .	11
1.7	Histogramme du temps de fixation pour $N$ petit . . . . .	11
1.8	Histogramme du temps de fixation pour $N$ grand . . . . .	12
1.9	Valeur de l'effet de sélection . . . . .	12
1.10	Temps de fixation moyen en fonction de la taille de la population initiale (exprimée en nombre d'allèles) avec $X_0 \in \{0, \dots, 1000\}$ et $s = 0.25$ . . . . .	13
1.11	Pas de temps choisi . . . . .	13
1.12	Nombre de simulations réalisées . . . . .	14
1.13	Temps de fixation en fonction des probabilités initiales . . . . .	14
1.14	Temps de fixation avec sélection en fonction des probabilités initiales . . . . .	15
1.15	Temps de fixation avec un faible effet de sélection en fonction des probabilités initiales . . . . .	16
1.16	Paramètres réseau . . . . .	17
1.17	R-studio server . . . . .	17
A.1	Évolution du nombre de $A$ en fonction du temps . . . . .	A-7
A.2	Histogramme du temps de fixation . . . . .	A-7
A.3	Évolution du nombre de $A$ en fonction du temps . . . . .	A-8
A.4	Histogramme du temps de fixation . . . . .	A-8
A.5	Variance sans effet de sélection . . . . .	A-9
A.6	Variance avec effet de sélection . . . . .	A-9

# Introduction

Ce rapport présente le travail réalisé lors de mon premier mois de stage de fin d'année pour la filière L3 DL SDV-Info. Les objectifs étaient de réaliser une application R-shiny, de la déployer sur un serveur et de rédiger un document explicatif – à mettre à disposition des chercheurs – de la procédure à suivre pour le déploiement.

Je réalise mon stage au Laboratoire de Mathématiques et Modélisation d'Evry (LaMME) sous le tutorat de Mr Vincent RUNGE. Le travail effectué est basé sur un cahier des charges comportant les indications suivantes :

1. L'application doit modéliser un problème biologique : la dérive génétique ;
2. Le modèle de génétique des populations imposé est le modèle de Wright-Fisher ;
3. Nous nous sommes intéressés tout particulièrement au temps de fixation.

**Remarque:** L'ensemble des résultats utilisés pour l'étude du modèle sans effet de sélection ont déjà été démontrés dans la littérature. L'objectif étant dans un premier temps de les comprendre, de les redémontrer et de les modéliser en R-Shiny.

Dans un second temps, nous avons étudié le modèle de Wright-Fisher avec l'ajout d'un effet de sélection. Cela a nécessité un travail théorique original que nous n'avons pas trouvé dans la littérature. Les objectifs ont été de :

- Déterminer une formule du temps de fixation lors de l'ajout d'un effet de sélection ;
- Comparer les résultats théoriques avec les simulations numériques ;
- Comparer les modèles avec et sans effet de sélection.

Le rapport est divisé en deux parties :

- Travail réalisé :
  - Modèle biologique utilisé et étude théorique
  - Présentation de l'application R-Shiny
- Bilan et perspectives : les perspectives présenteront le travail à réaliser lors de la seconde partie de mon stage.

# Travail réalisé

---

## 1.1 Modèle de Wright-Fisher

### 1.1.1 Présentation du modèle

Le modèle de Wright-Fisher est un des modèles les plus simples, en génétique des populations. Il permet de modéliser la transmission des allèles d'individus diploïdes ou haploïdes au cours des générations et d'en déterminer le temps de fixation. En effet, la fréquence des allèles peut être modifiée aléatoirement dans une population en évolution au cours du temps. La durée au bout de laquelle l'un des allèles disparaît définit le temps de fixation. Ce remaniement des allèles peut être expliqué par le fait que certains individus se reproduisent plusieurs fois et à l'inverse que d'autres se ne reproduisent pas. Au cours des générations, un allèle va donc être favorisé et se transmettre préférentiellement. Ainsi, plus la population initiale est petite et plus le phénomène de dérive génétique est rapide. [2]

Pour notre étude nous poserons les hypothèses suivantes :

- La population est de taille finie et constante au cours des générations à  $M$  individus, soit  $N = 2M$  allèles ;
- Pas de mutation au cours des générations ;
- Les générations sont indépendantes : les individus de la génération  $t + 1$  sont tous issus de la génération  $t$ .

### 1.1.2 Modèle sans effet de sélection

Dans un premier temps, nous considérerons le modèle simple de Wright-Fisher pour lequel il n'y a pas d'effet de sélection. Ainsi, tous les allèles peuvent être transmis de manière équiprobable et aucun des deux allèles ne confère un avantage sélectif.

Soit l'étude d'une population de  $M$  individus diploïdes d'allèles  $A$  et  $a$ . Au temps  $t = 0$ , l'échantillon est composé de  $N$  allèles et on suppose que l'on connaît le nombre d'allèles  $A$ .

Ainsi, afin de modéliser chaque nouvelle génération  $t$ , on réalise  $N$  tirages parmi les allèles au temps  $t - 1$ . Dans notre cas, nous considérerons que les tirages sont faits avec remise. On réalise des tirages tant qu'il ne reste pas qu'un allèle unique dans la population. Le temps de fixation correspondra alors au nombre de générations avant uniformisation de la population.

### 1.1.3 Notations [1]

On note  $X_t$ , le nombre de  $A$  au temps  $t$ . La proportion de  $A$  au temps  $t$  est alors notée :

$$p_t = \frac{X_t}{N},$$

avec  $X_t \in \{0, \dots, N\}$ .

À la génération  $t + 1$ , tirer un allèle représente un schéma de Bernoulli dont le succès est de tirer un  $A$  de probabilité  $p_t$  et l'échec, tirer un  $a$  de probabilité  $1 - p_t$ . On répète ce schéma  $N$  fois,  $X_t$  suit donc une loi binomiale :

$$X_{t+1} \sim \mathcal{B}(N, p_t).$$

Ainsi, pour tout  $t \in \mathbb{N}$  et pour tout  $i, j$  dans  $\{1, \dots, N\}$

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}.$$

**Remarque:**  $(X_t)_{t \in [0, N]}$  est une chaîne de Markov [3] où  $p_{ij}$  représente la probabilité de transition de l'état  $i$  à l'état  $j$ .

On a alors :  $\mathbb{E}[X_{t+1}] = Np_t$  et  $\text{Var}[X_{t+1}] = Np_t(1 - p_t)$ .

Quant à  $p_{t+1}$ , on a :  $\mathbb{E}[\frac{X_{t+1}}{N}] = p_t$  et  $\text{Var}[\frac{X_{t+1}}{N}] = \frac{1}{N}p_t(1 - p_t)$ .

*Démonstration.* (cf. annexe A.1 et A.2). □

### 1.1.4 Coefficient d'hétérozygotie

Afin de mieux comprendre l'effet de fixation d'un allèle dans une population, nous allons définir un coefficient d'hétérozygotie. Son évolution montre qu'une population finie de deux allèles tend à devenir homogène.

Si nous tirons deux allèles dans la population au temps  $t$ ,  $H_t$  désigne la probabilité que ces allèles soient différents. L'objectif est de calculer son espérance  $h_t$  puis sa limite afin de savoir si nous pouvons espérer obtenir un temps de fixation fini.

Dans notre étude nous considérerons 2 tirages aléatoires, indépendants et avec remise. Dans la littérature [1, 4] il a été montré qu'il était également possible de considérer 2 tirages avec remise. Le coefficient  $H_t$  obtenu ne diffère que d'une constante dépendant de la taille de la population. (cf. annexe A.3)

Soit  $n = 2$ , le nombre d'allèles tirés dans la population de  $N$  allèles au temps  $t$ . Il existe donc  $n! = 2$  permutations possibles.  $H_t$  est alors défini par [2] :

$$H_t = 2p_t(1 - p_t).$$

L'espérance  $h_t$  s'exprime alors :

$$h_t = \mathbb{E}[H_t \mid p_{t-1}] = h_{t-1} \left(1 - \frac{1}{N}\right).$$

*Démonstration.* (cf. annexe A.4) □

Par récurrence, on en déduit que  $\forall t \in \mathbb{N}$  :

$$h_t = h_0 \left(1 - \frac{1}{N}\right)^t.$$

À  $t = 0$ ,  $X_t = 0$  ou  $X_t = N$ , on a  $h_t = 0$ .

Donc  $\forall X_0 \in \{1, \dots, N-1\}$ ,  $h_0 > 0$

On a alors  $\left(1 - \frac{1}{N}\right) \in ]0, 1[$ , donc  $\lim_{t \rightarrow +\infty} h(t) = 0$

**Remarque:** On obtient le même résultat lorsque  $H_t$  est défini en considérant 2 tirages sans remise.

Biologiquement, cela signifie donc qu'au cours des générations, la population tend à être homozygote et que le temps de fixation est donc sans doute fini.

### 1.1.5 Temps de fixation

On étudie à présent notre donnée d'intérêt : le temps de fixation. Nous avons démontré précédemment que la population tendait à devenir homozygote, plus ou moins rapidement, selon le nombre d'allèles  $N$  de départ. Notre objectif est de démontrer que l'espérance du temps de fixation est toujours finie, et d'en trouver une formule explicite, quelle que soit la taille initiale de la population.

Pour cela, deux approches sont envisageables :

1. L'approche discrète permet d'estimer les temps de fixation pour toutes les quantités initiales d'allèles  $A$  (notées  $i$ , tel que  $i \in \{0, \dots, N\}$ ). Cette approche sera alors exprimée sous forme matricielle.
2. L'approche continue, qui consiste quant à elle à considérer la proportion de  $A$  comme une valeur continue au cours du temps. Cette approche nous mènera à la résolution d'une équation différentielle.

#### Cas discret

- On pose  $T$  la variable aléatoire représentant le temps de fixation minimum afin de fixer un allèle, telle que :

$$T = \min \{t \mid X_t = 0 \text{ ou } X_t = N\}$$

- On pose  $m_i$ , l'espérance du temps de fixation d'un allèle sachant sa quantité  $X_0 = i$  au temps  $t = 0$ .

On a alors [1, 4] :

$$\begin{aligned}
 m_i &= \sum_{j=0}^N \mathbb{P}(X_1 = j \mid X_0 = i) \mathbb{E}(T \mid X_0 = i, X_1 = j) \\
 &= \sum_{j=0}^N \mathbb{P}(X_1 = j \mid X_0 = i) \mathbb{E}(T \mid X_1 = j) \\
 &= \sum_{j=1}^{N-1} p_{ij} (1 + m_j) \\
 &= 1 + \sum_{j=1}^{N-1} p_{ij} m_j (*)
 \end{aligned}$$

avec  $p_{ij}$  la probabilité de passage de  $i$  à  $j$  allèles  $A$  et  $\sum_{j=0}^N p_{ij} = 1$ .

**Remarque:** Pour  $X_0 = 0$  et  $X_0 = N$ , le temps de fixation est nul, soit  $m_0 = m_N = 0$ . Cela nous permet d'enlever les termes pour  $j = 0$  et  $j = N$  dans (\*).

On détaille ici l'avant-dernière égalité de la preuve précédente :

$$\begin{aligned}
 \mathbb{E}[T \mid X_0 = i, X_1 = j] &= \mathbb{E}[T \mid X_1 = j] \\
 &= \mathbb{E}[T + 1 \mid X_0 = j] \\
 &= 1 + \mathbb{E}[T \mid X_0 = j] \\
 &= 1 + m_j
 \end{aligned}$$

### Écriture matricielle

On a :

- $m_i$  et  $m_j$  de dimension  $(N - 1) \times 1$  : temps moyen sachant  $i$  le nombre de  $A$  au temps 0 et  $j$  le nombre de  $A$  au temps 1 ;
- $p_{ij}$  de dimension  $(N - 1) \times (N - 1)$ , la matrice des probabilités de passage de  $i$  à  $j$  ;

$$\begin{pmatrix} m_i(1) \\ m_i(2) \\ \dots \\ m_i(N-1) \end{pmatrix} = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,N-1} \\ p_{2,1} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ p_{N-1,1} & \dots & \dots & p_{N-1,N-1} \end{bmatrix} \begin{pmatrix} m_j(1) \\ m_j(2) \\ \dots \\ m_j(N-1) \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}.$$

Le temps de fixation  $m_i$ , pour tout  $i \in \{1, \dots, N - 1\}$  s'exprime :

$$m_i = (I_m - P_{ij})^{-1} \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}.$$

### Cas continu

On choisit maintenant d'écrire  $m_i$  comme une fonction de  $p$ , proportion initiale de  $A$ . On pose  $Y$  la variable aléatoire représentant la proportion de  $A$  au temps  $t$ , telle que :



- $Y_t = \frac{1}{N}X_t$
- Au temps  $t + \Delta t$ ,  $Y_{t+\Delta t} = \frac{1}{N}X_{t+\Delta t}$

Au pas de temps  $\Delta t$ , la proportion de  $A$  varie d'une quantité  $\Delta p$  telle que, on note  $Z$  la variable aléatoire représentant cette variabilité sur  $p$  :

$$Z_t = Y_{t+\Delta t} - Y_t = \Delta p_t$$

L'espérance de  $T$ , s'écrit alors

$$m(p) = \mathbb{E}[T(Y_0 \mid Y_0 = p)]$$

De plus, on a :

- $\mathbb{E}[Z_t] = 0$
- $\text{Var}[Z_t] = \frac{p(1-p)}{N}$

*Démonstration.* cf. annexe A.5

$m(p)$  s'écrit alors [1, 4, 5] :

$$\begin{aligned} m(p) &= \sum_{\Delta p} \mathbb{P}(Y_{\Delta t} = p + \Delta p \mid Y_0 = p) m(Y_{\Delta t} \mid Y_{\Delta t} = p + \Delta p, Y_0 = p) \\ &= 1 + \sum_{\Delta p} \mathbb{P}(Y_{\Delta t} = p + \Delta p \mid Y_0 = p) m(p + \Delta p) \end{aligned}$$

*Démonstration.* cf. annexe A.6

On utilise la formule de Taylor à l'ordre 2 et les égalités suivantes :

$$\mathbb{E}[Z_t] = 0 \text{ et } \mathbb{E}[\Delta p^2] = \frac{p(1-p)}{N}, \text{ pour obtenir } m''(p).$$

$$\begin{aligned} m(p) &= 1 + \sum_{\Delta p} \mathbb{P}(Y_{\Delta t} = p + \Delta p \mid Y_0 = p) \left( m(p) + \frac{m'(p)}{1!} \Delta p + \frac{m''(p)}{2!} \Delta p^2 + O(\Delta p^3) \right) \\ &\Leftrightarrow m''(p) = -2N \frac{1}{p(1-p)}. \end{aligned}$$

Afin d'intégrer  $m''(p)$ , on décompose  $\frac{1}{p(1-p)}$  en une somme :

$$\frac{1}{p(1-p)} = \frac{1}{p} + \frac{1}{1-p}.$$

On obtient :

- $m'(p) = -2N[\ln(p) - \ln(1-p)] + C_0$
- $m(p) = -2N[p \ln(p) + (1-p) \ln(1-p)] + C_1 p + D$

On sait que  $X_0 = 0$  et  $X_0 = N$  sont deux états absorbants. On a donc :

- $m(0) = 0 \Leftrightarrow D=0$
- $m(1) = 0 \Leftrightarrow C_1=0$

## Conclusion

Le temps de fixation moyen peut donc s'exprimer comme une fonction de  $p$  telle que :

$$m(p) = -2N \left[ p \ln(p) + (1-p) \ln(1-p) \right].$$

### 1.1.6 Modèle avec sélection

Au sein d'une population, on observe une modification aléatoire de la transmission des allèles au cours des générations. C'est ce qu'on appelle la dérive génétique induisant une évolution de la biodiversité de la population initiale. Selon les conditions (environnementales, climatiques), certains allèles peuvent conférer ou non à l'individu porteur un avantage sélectif. La transmission de ces allèles est alors favorisée ou diminuée au cours des générations. Il s'agit de la sélection naturelle. Au cours du temps, dérive génétique et sélection naturelle contribuent donc à la biodiversité des individus et à la spéciation.

Lors notre étude, nous considérerons que l'avantage sélectif conféré par un allèle  $A$  d'intérêt est la viabilité de l'individu ou sa meilleure capacité de reproduction. Ainsi, un individu porteur de cet allèle a  $(1+s)$  fois plus de chance de survivre qu'un individu porteur d'un allèle  $a$  [1].

À la génération  $t$ , la proportion de  $A$  est toujours notée  $p_t = \frac{X_t}{N}$ . À la génération  $t+1$ , la probabilité de transmission de l'allèle  $A$  est augmentée telle que :

$$p_{t+1} = \frac{(1+s)X_t}{(1+s)X_t + N - X_t}$$

Il s'agit à présent d'estimer le temps de fixation de cet allèle en prenant en compte l'effet de sélection.

On a alors :

$$X_{t+1} \sim \mathcal{B}\left(N, \frac{(1+s)X_t}{(1+s)X_t + N - X_t}\right)$$

### 1.1.7 Temps de fixation avec sélection

#### Cas continu

De même que pour le modèle sans effet de sélection, on pose  $Y$  la variable aléatoire représentant la proportion de  $A$  au temps  $t$ , telle que :

$$Y_t = \frac{X_t}{N}$$

De la même manière :  $Z_t = Y_{t+1} - Y_t$

$$\begin{cases} \mathbb{E}\left[Z_t \mid Y_t = \frac{i}{N}\right] = \frac{is(N-i)}{N(is+N)} = p \frac{s(1-p)}{1+ps} \\ \text{Var}\left[Z_t \mid Y_t = \frac{i}{N}\right] = \frac{(1+s)i(N-i)}{N(is+N)^2} = \frac{1}{N} \frac{p(1+s)(1-p)}{(1+ps)^2} \end{cases} \quad (1.1)$$

*Démonstration.* cf. annexe A.7

On utilise à nouveau la formule de Taylor à l'ordre 2 et on obtient :

$$m(p) = 1 + \sum_{\Delta p} \mathbb{P}(Y_{t+\Delta t} = p + \Delta p \mid Y_0 = p) \left( m(p) + \frac{m'(p)}{1!} \Delta p + \frac{m''(p)}{2!} \Delta p^2 + O(\Delta p^3) \right)$$

On sait que :

$$\text{Var}[Z_t] = \mathbb{E}[Z_t^2] - \mathbb{E}[Z_t]^2$$

Ainsi,

$$\mathbb{E}[Z_t^2] = \frac{p(1-p)}{(1+ps)^2} \left[ \frac{1}{N}(1+s) + ps^2(1-p) \right]$$

D'après la littérature [1], on admet que le coefficient  $s$  est de l'ordre de  $\frac{\alpha}{N}$ . En remplaçant  $\mathbb{E}[Z_t]$  et  $\mathbb{E}[Z_t^2]$  dans l'expression de  $m(p)$ , on obtient :

$$0 = 1 + p \frac{s(1-p)}{1+ps} m'(p) + \frac{p(1-p)}{(1+ps)^2} \left[ \frac{1}{N}(1+s) + ps^2(1-p) \right] \frac{m''(p)}{2}$$

D'après annexe A.8 [6], on obtient  $m(p)$  :

$$m(p) = K - \frac{Ce^{-2\alpha p}}{2\alpha} + \frac{N}{\alpha} \left[ e^{-2\alpha p} Ei(2\alpha p) - e^{-2\alpha(p-1)} Ei(2\alpha(p-1)) + \ln(1-p) - \ln(p) \right]$$

avec :

- $\alpha \neq 0$
- $\gamma \approx 0.5772157$  (constante d'Euler-Mascheroni)
- 

$$\begin{cases} C = \frac{-2N}{e^{-2\alpha} - 1} \left[ 2\gamma + e^{2\alpha} \left( -Ei(-2\alpha) - e^{-4\alpha} Ei(2\alpha) \right) \right] \\ K = \frac{C}{2\alpha} - \frac{N}{\alpha} [\gamma - e^{2\alpha} Ei(-2\alpha)] \end{cases}$$

## 1.2 App Shiny

L'application shiny réalisée permet de simuler des populations modélisées par Wright-Fisher pour un ensemble de paramètres choisis par l'utilisateur :

- *Population size* : le nombre **d'individus** étudiés ;
- *Number of alleles A* : le nombre d'allèles *A* observés dans la population initiale ;
- *Generations number* : le nombre de populations initiales tirées pour les paramètres choisis ; (cf Figure 1.1)

The screenshot shows a Shiny app interface with the following elements:

- Population size : (nb\_A + nb\_a)/2**: A text input field containing the value 10.
- Number of alleles A (nb\_A)**: A text input field containing the value 10.
- Generations number:**: A slider control with a range from 1 to 1,000. The current value is 10, indicated by a blue label.
- Binwidth**: A slider control with a range from 1 to 30. The current value is 3, indicated by a blue label.
- Select the type of graph**: Three radio buttons are present: ☐ CoeffH, ☐ A, and ☒ Both.
- Display plot**: A button to trigger the plot display.
- Simu number:**: A slider control with a range from 1 to 1,000. The current value is 100, indicated by a blue label.
- Step**: A text input field containing the value 2.
- Add a selection effect**: A text input field containing the value 1.

FIGURE 1.1: Panel

3 types de graphes sont alors observables :

- *A* : l'évolution du nombre d'allèles *A* en fonction du temps ;
- *CoeffH* : l'évolution du coefficient d'hétérozygotie en fonction du temps ;
- L'histogramme des valeurs de temps de fixation obtenus pour l'ensemble des simulations (affiché par défaut et dont la largeur des barres est modifiable en changeant le paramètre **Binwidth**).

L'allure des courbes obtenues dépend du modèle choisi : avec ou sans sélection. Il est alors possible de visualiser ces modèles indépendamment en sélectionnant l'onglet d'intérêt. (cf Figure 1.2)

## Wright Fisher

GenPop Multi GenPop Select GenPop Select Multi GenPop

FIGURE 1.2: Onglets

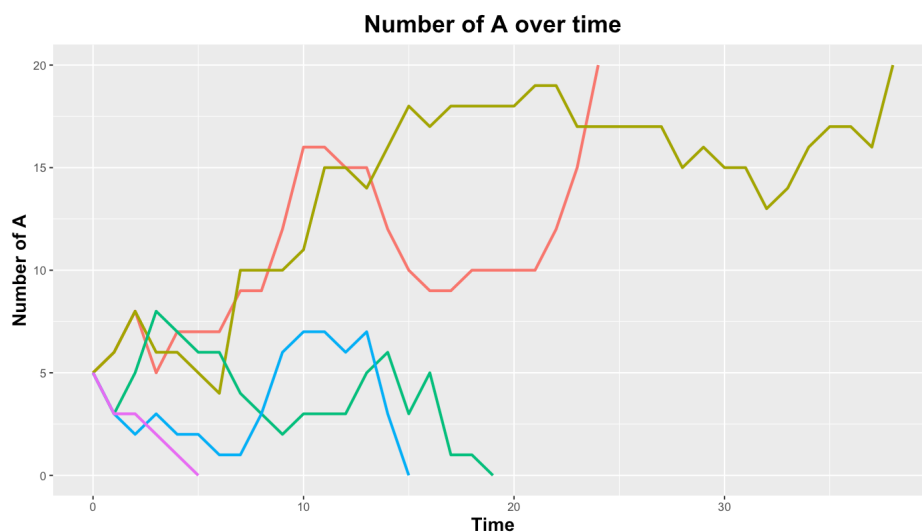
### 1.2.1 Sans sélection (onglet GenPop)

#### Petite population initiale

On fixe les paramètres initiaux suivants :

- *Population size* : 10 individus (soit 20 allèles)
- *Number of alleles A* : 5
- *Generations number* : 5

On obtient les graphes suivants :



This plot shows alleles number over time for an initial population of 10 individuals and 5 alleles A.

FIGURE 1.3: Évolution du nombre d'allèles *A* en fonction du temps.  
*Generations number* : 5

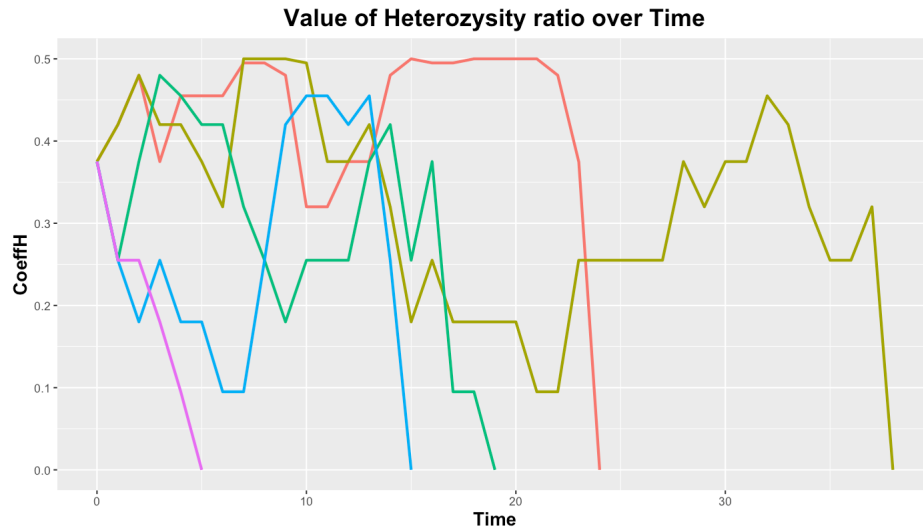


FIGURE 1.4: Évolution du coefficient  $H$  en fonction du temps  
Generations number : 5

**Remarque:** Lors de l'étude théorique, nous avons démontré que le coefficient d'hétérozygotie tendait vers 0 au cours du temps. Ceci nous a permis de poser l'hypothèse d'un temps de fixation fini. On constate effectivement que toutes les courbes représentant le nombre de  $A$  au cours du temps finissent par devenir constantes à 0 ou  $N$  et que le coefficient d'hétérozygotie se fixe bien à 0. Les deux graphes sont donc cohérents et confirment nos hypothèses.

### Grande population initiale

La taille de la population est augmentée, mais la proportion initiale de  $A$  est conservée :

- Population size : 100 individus (soit 200 allèles)
- Number of alleles  $A$  : 50

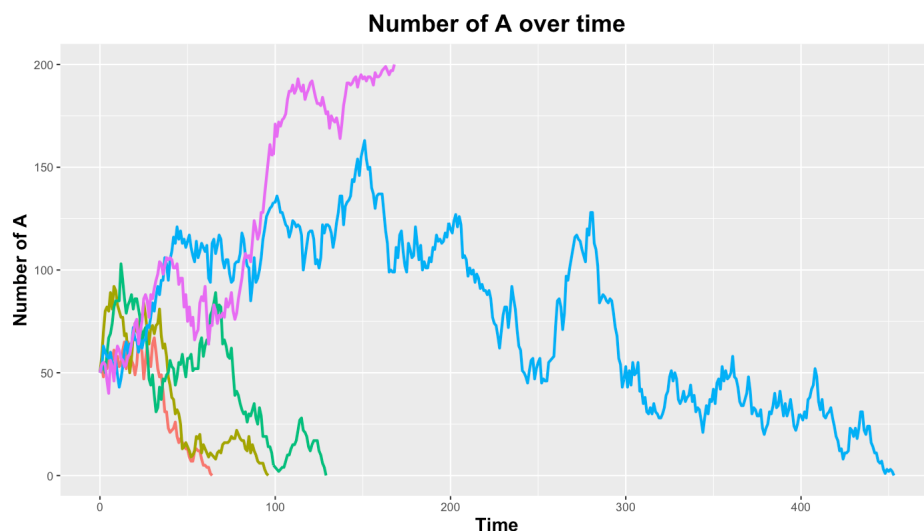
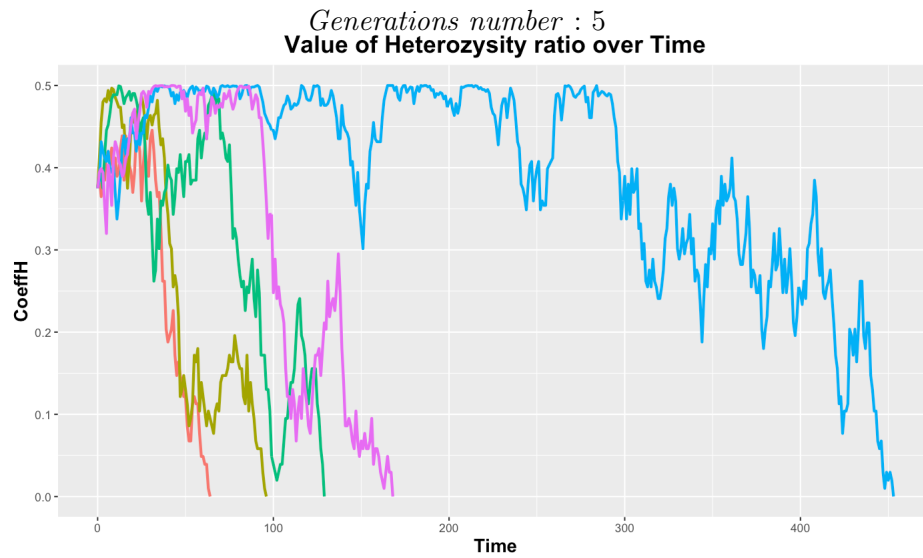


FIGURE 1.5: Évolution du nombre d'allèles  $A$  en fonction du temps



This plot shows the Heterozygosity Ratio over time for an initial population of 100 individuals and 50 alleles A.

FIGURE 1.6: Évolution du coefficient H en fonction du temps  
*Generations number : 5*

### Histogramme des temps de fixation

Avec les paramètres fixés ci-dessus, on constate que les temps de fixation diffèrent selon la taille de la population initiale. Le nombre de générations a été augmenté à 1000 afin d'obtenir des résultats plus précis.

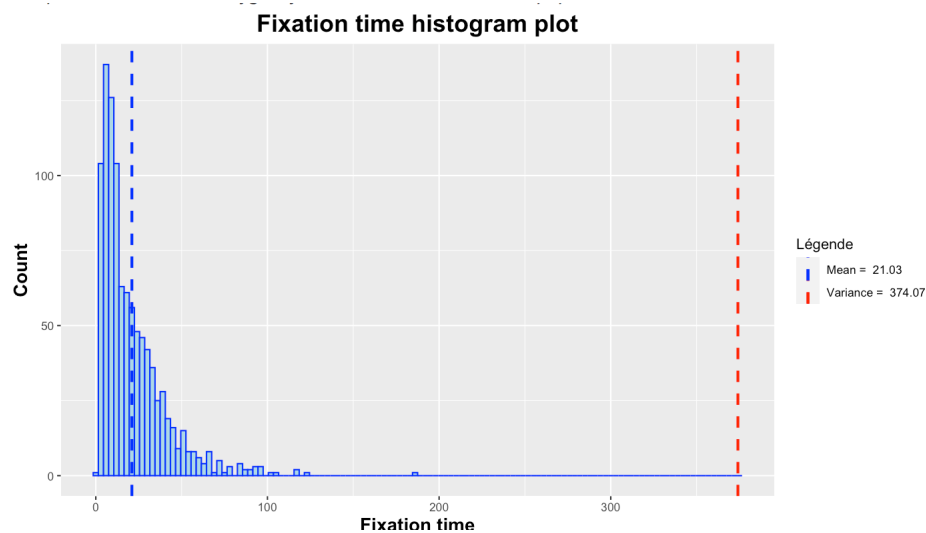


FIGURE 1.7: Histogramme du temps de fixation pour  $N$  petit  
*Population size : 10*  
*Number of alleles A : 5*  
*Generations number : 1000*

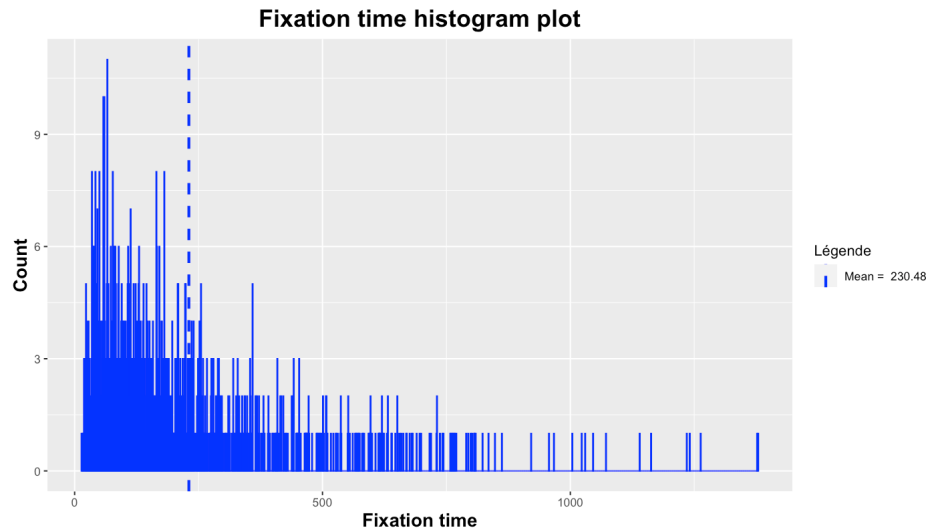


FIGURE 1.8: Histogramme du temps de fixation pour  $N$  grand  
*Population size* : 100  
*Number of alleles A* : 50  
*Generations number* : 1000

Le temps de fixation dépend de la taille de la population : plus celle-ci est grande et plus le temps l'est aussi. Nos simulations sont donc cohérentes avec les résultats théoriques.

### 1.2.2 Avec sélection (onglet Select GenPop)

L'application permet d'ajouter un effet de sélection : paramètre **Add a selection effect**. Cet effet de sélection représente plus particulièrement la valeur  $\alpha$ , précédemment explicitée dans le rapport. Pour rappel, l'effet de sélection  $s$  est proportionnel à  $\alpha$ , tel que  $s = \frac{\alpha}{N}$ .

Dans notre exemple :

- Les paramètres précédents (*Population size*, *Number of alleles A*, *Generations number*) sont conservés.
- $\alpha = 5$
- $s = 0.25$

Add a selection effect

FIGURE 1.9: Valeur de l'effet de sélection

Cela signifie, qu'à chaque génération, la proportion d'allèles  $A$  est multipliée par 1,25.

Comme pour le modèle sans sélection, le temps de fixation est lié à la taille de la population initiale (cf Annexe A.9). Les résultats sont résumés dans le graphique ci-dessous.



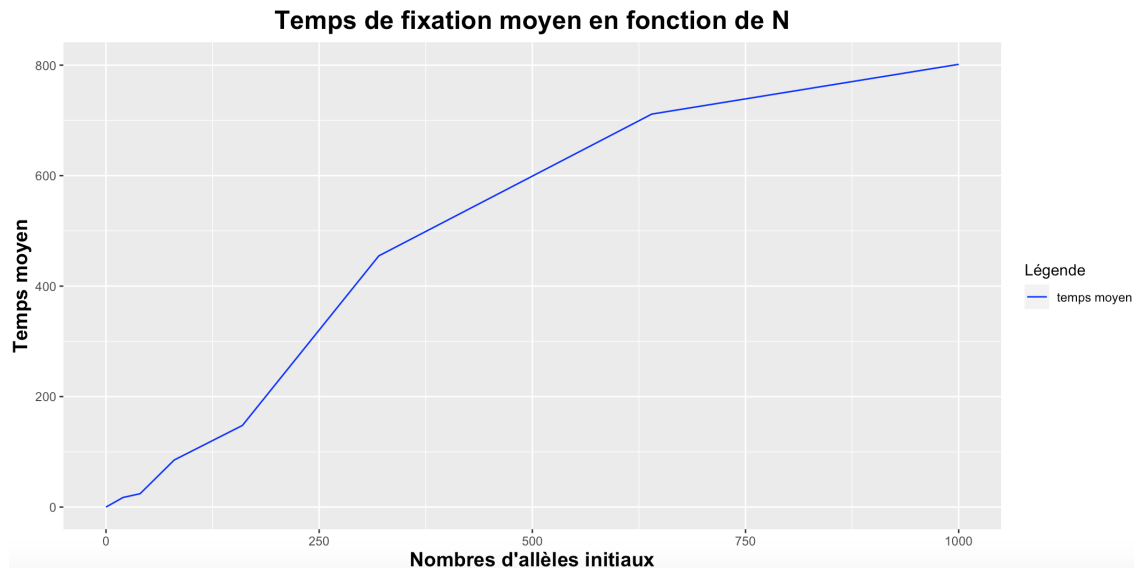


FIGURE 1.10: Temps de fixation moyen en fonction de la taille de la population initiale (exprimée en nombre d'allèles) avec  $X_0 \in \{0, \dots, 1000\}$  et  $s = 0.25$ .

On constate que plus la population initiale est importante et plus le temps de fixation est élevé.

### 1.2.3 Étude du temps de fixation avec et sans sélection

Les résultats obtenus avec nos simulations démontrent que plus l'effet de sélection est fort et plus nos courbes convergent vite. Nos simulations ont été réalisées avec un jeu de paramètres fixés : une proportion initiale de  $A$  égale à 25%. Il s'agit à présent de :

- Regarder l'allure de la courbe représentant le temps de fixation pour toutes les probabilités initiales possibles ;
- Pour chaque modèle, comparer les résultats théoriques continus et discrets avec les simulations ;
- Comparer les deux modèles.

À présent, toutes les proportions initiales de  $A$  sont calculées à partir de la taille de la population initiale (*Population size*, fixée à 100) et du pas choisi (*Step*, fixé à 1). Cela signifie plus particulièrement que le temps de fixation sera estimé pour chaque valeur de  $A$  variant de 0 à 200, par pas de 1.



FIGURE 1.11: Pas de temps choisi

Il est également possible de modifier le nombre de simulations réalisées avec le paramètre *Simu number*, fixé à 100 pour notre exemple. Pour chaque proportion initiale, 100 populations

différentes seront tirées et simulées. Ce paramètre est finalement identique à *Generation Number*, mais leur dissociation permet de rendre indépendant les onglets *GenPop-Select GenPop* et *Multi GenPop-Select Multi GenPop*

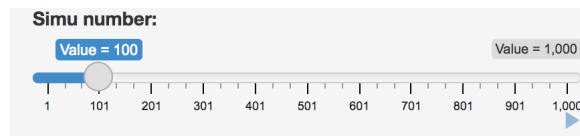


FIGURE 1.12: Nombre de simulations réalisées

### Sans sélection (Onglet Multi GenPop)

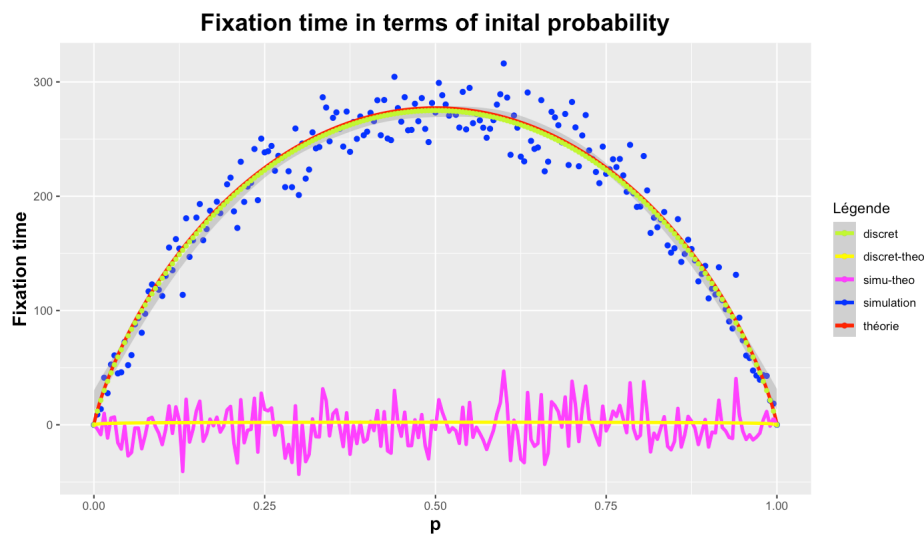


FIGURE 1.13: Temps de fixation en fonction des probabilités initiales

Population size : 100

Simu number : 100

Step :1

On constate que :

- Les résultats théoriques continus (courbe rouge) et discrets (points verts) sont cohérents, de par la superposition des deux courbes. De plus, la courbe jaune, proche de 0, représente la différence entre les deux résultats. Pour rappel, le résultat continu représente la formule, démontrée, du temps de fixation :

$$m(p) = -2N[p \ln(p) + (1 - p) \ln(1 - p)].$$

Le résultat discret a été obtenu par l'inversion de matrices. **Remarque:** On s'attend à ce que le résultat discret soit moins précis pour de très grandes valeurs de  $N$ , dû à des limites numériques lors de l'inversion des matrices.

- Nos simulations (points bleus) semblent corrects de par leur allure qui se rapprochent des résultats théoriques et la courbe de tendance comprise dans l'intervalle de confiance à 95 % (représentée par une zone légèrement grisée) qui se superpose également avec les résultats théoriques. De plus, la courbe rose, représentant l'écart entre nos simulations et la théorie, est centrée autour de 0.

- La courbe représentant le temps de fixation, est symétrique par rapport à la droite d'équation  $x=\frac{1}{2}$ . Le temps de fixation est donc maximal pour une probabilité initiale de  $\frac{1}{2}$ . Plus la probabilité initiale est proche de 0 ou de 1, plus le temps de fixation attendu est faible. Ce résultat est cohérent, puisque si les proportions d'allèles initiales sont déséquilibrées, il y a déjà un allèle plus dominant, tendant à rendre la population homozygote pour cet allèle.

### Avec sélection (Onglet Select Multi GenPop)

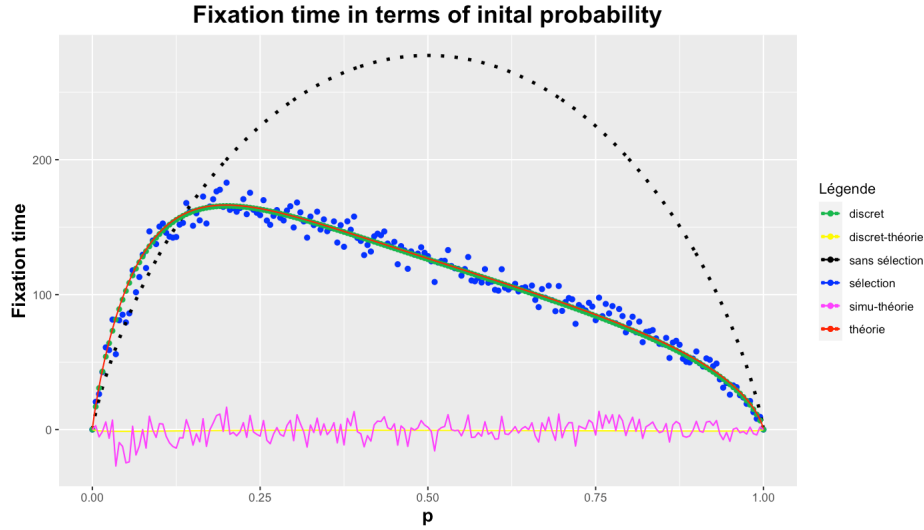


FIGURE 1.14: Temps de fixation avec sélection en fonction des probabilités initiales

Population size : 100

Simu number : 100

Step : 1

$\alpha : 5$

### Remarque: Utilisation du package [7]

On constate que :

- Les résultats théoriques continus (courbe rouge) et discrets (points verts) sont cohérents, de par la superposition des deux courbes. De plus, la courbe jaune, proche de 0, représente la différence entre les deux résultats. Pour rappel, le résultat continu représente la formule du temps de fixation que nous avons déterminé :

$$m(p) = K - \frac{C e^{-2\alpha p}}{2\alpha} + \frac{N}{\alpha} \left[ e^{-2\alpha p} Ei(2\alpha p) - e^{-2\alpha(p-1)} Ei(2\alpha(p-1)) + \ln(1-p) - \ln(p) \right]$$

Le résultat discret a également été obtenu par l'inversion de matrices. Ces résultats nous permettent de confirmer la validité de la formule trouvée, ou du moins sa cohérence.

- Nos simulations (points bleus) semblent également cohérentes avec à la théorie, pour les mêmes raisons que celles énumérées pour le modèle sans effet de sélection.
- La courbe en pointillés noirs représente le résultat théorique sans effet de sélection. Avec un effet de sélection, les temps de fixation sont réduits avec un décalage du pic vers la gauche. Ce résultat correspond à celui attendu (à démontrer mathématiquement par la suite). Entre  $p = 0$  et  $p = p_{pic}$ , où  $p_{pic}$  est la proportion  $p$  pour laquelle le pic est atteint, les courbes avec et sans sélection évoluent de la même façon. À partir de la proportion  $p_{pic}$ ,

bien inférieure à  $\frac{1}{2}$ , la quantité de  $A$  est suffisante (de par son effet de sélection), pour faire décroître rapidement le temps de fixation.

De plus, on s'attend à ce que, lors d'un effet de sélection proche de 0, le modèle avec sélection tende à se rapprocher de celui sans effet de sélection. On s'attend alors à ce que les courbes théoriques et simulées se superposent à la courbe en pointillés noirs (courbe théorique du modèle sans sélection).

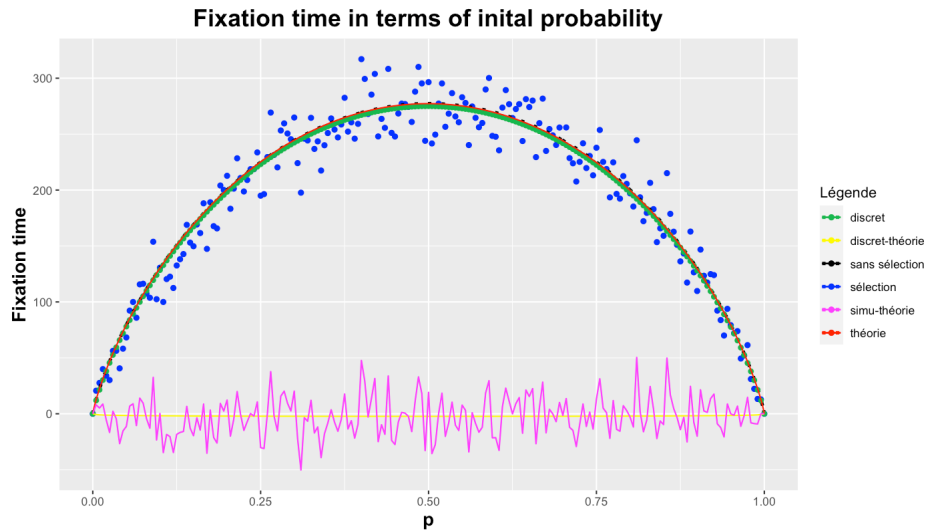


FIGURE 1.15: Temps de fixation avec un faible effet de sélection en fonction des probabilités initiales

*Population size* : 100

*Simu number* : 100

*Step* : 1

$\alpha$  : 0.01

Le résultat obtenu correspond à celui attendu, ce qui confirme notre étude théorique et les hypothèses posées.

### 1.2.4 Mise à disposition

En parallèle, j'ai eu l'occasion d'apprendre à déployer l'application sur le serveur du laboratoire avec l'aide de Franck Samson. J'ai notamment appris à :

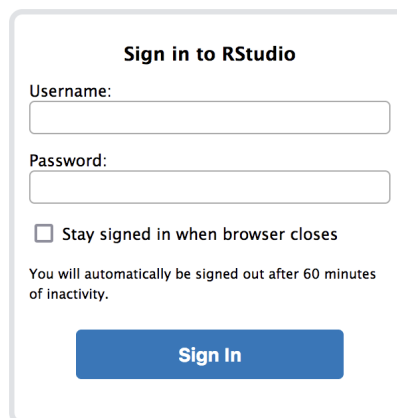
- Manipuler le terminal en lignes de commandes ;
- Reconfigurer les paramètres réseau de mon navigateur ;



The image shows a browser's network settings page. At the top, 'Configuration manuelle du proxy' is selected with a radio button. Below it, there are three sections: 'Proxy HTTP' with an empty text box and a 'Port' dropdown set to '0'; a checkbox 'Utiliser également ce proxy pour FTP et HTTPS' which is unchecked; 'Proxy HTTPS' with an empty text box and a 'Port' dropdown set to '0'; and 'Hôte SOCKS' with a text box containing 'localhost' and a 'Port' dropdown set to '0'. At the bottom, there are two radio buttons: 'SOCKS v4' (unchecked) and 'SOCKS v5' (checked).

FIGURE 1.16: Paramètres réseau

- Accéder au serveur R-studio et à la machine de calcul ;



The image shows the 'Sign in to RStudio' login page. It has a title 'Sign in to RStudio' in bold. Below it are two input fields: 'Username:' and 'Password:'. There is a checkbox labeled 'Stay signed in when browser closes'. Below the checkbox, a note says 'You will automatically be signed out after 60 minutes of inactivity.' At the bottom is a blue button labeled 'Sign In'.

FIGURE 1.17: R-studio server

- Accéder aux applications déposées ;

# Bilan et perspectives

---

Au cours de ce premier projet j'ai donc pu étudier un problème biologique avec une approche mathématiques et informatique. L'ensemble des objectifs fixés quant à la réalisation de l'application R-Shiny et la détermination d'une formule pour le temps de fixation avec sélection ont été réalisés. Les études théoriques associées et l'apprentissage de Shiny m'ont permis d'acquérir de nouvelles connaissances dans les trois domaines. Le travail réalisé est disponible sur GitHub : <https://github.com/PaulineSpinga/wrightfisher>. L'application Shiny est également disponible à l'adresse suivante : [http://192.168.216.252:3838/users/pspinga/wright\\_fisher/](http://192.168.216.252:3838/users/pspinga/wright_fisher/) (accessible uniquement depuis l'IBGBI).

Afin de finaliser ce travail, plusieurs objectifs sont encore à réaliser tels que :

- Déterminer le maximum du temps de fixation avec sélection et retrouver la formule du temps de fixation sans sélection à partir du celle avec effet de sélection lorsque celui-ci tend vers 0 ;
- Déterminer la formule de la variance du temps de fixation sans sélection (cf Annexe A.10) ;
- Mettre à disposition des chercheurs la méthodologie à adopter pour mettre une application en ligne sur le serveur du laboratoire.

Pour la suite de ce stage, nous allons travailler sur une application shiny pour la détection de ruptures dans les séries temporelles ou poursuivre les développements théoriques trouvés sur le modèle de Wright-Fisher.

# Bibliographie

- [1] J. ANGST, “Modèles aléatoires en biologie,” in *Master de modélisation des systèmes biologiques*, Semestre d’Automne 2011-2012.
- [2] J.-H. SMITH-LACROIX, “Modèles de wright-fisher et n-coalescent,” in *Essai*, Août 2005.
- [3] S. Méléard, “Modèles aléatoires en ecologie et evolution,” in *Ecole Polytechnique*, 2009.
- [4] D. C. et Florent Malrieu, “Modèles markoviens en biologie,” Mars 2007.
- [5] A. Etheridge, “Diffusion process models in mathematical genetics.”
- [6] E. Masina, “A review on the exponential-integral special function and other strictly related special functions,” in *The Exponential Integral and linked functions*, March 2017.
- [7] V. Goule, “Exponential integral and incomplete gamma function,” in *Package ‘expint’*, December 2019.

# Annexe

---

## A.1 Espérance et Variance de $X_{t+1}$

Soit  $X_t$ , le nombre de  $A$  au temps  $t$ . Au temps  $t + 1$ ,  $X_{t+1} = \sum_{i=1}^N X_{i,t+1}$ . L'espérance et la variance de  $X_{t+1}$  s'expriment alors :

$$\mathbb{E}[X_{t+1}] = \mathbb{E}\left[\sum_{i=1}^N X_{i,t+1}\right] = \sum_{i=1}^N \mathbb{E}[X_{i,t+1}] = Np_t,$$

$$Var[X_{t+1}] = Var\left[\sum_{i=1}^N X_{i,t+1}\right] = \sum_{i=1}^N Var[X_{i,t+1}] = Np_t(1 - p_t) = X_t\left(1 - \frac{X_t}{N}\right).$$

## A.2 Espérance et Variance de $p_{t+1}$

Soit  $\frac{X_t}{N}$ , la fréquence de  $A$  au temps  $t + 1$ .  
L'espérance et la variance de  $p_{t+1}$  s'expriment alors :

$$\begin{aligned}\mathbb{E}\left[\frac{X_{t+1}}{N}\right] &= \frac{\mathbb{E}[X_{t+1}]}{N} = p_t, \\ Var\left[\frac{X_{t+1}}{N}\right] &= \frac{1}{N^2} Var[X_{t+1}] = \frac{1}{N} p_t(1 - p_t).\end{aligned}$$

## A.3 Coefficient d'hétérozygotie sans remise

La probabilité d'obtenir un individu hétérozygote au temps  $t$  en tirant 2 allèles parmi  $N$ , avec remise est donc défini selon la formule suivante :

Pour rappel, on a  $X_t$ , le nombre de  $A$  et  $p_t = \frac{X_t}{N}$ , la proportion de  $A$ , au temps  $t$ .

$$H_t = \frac{\binom{X_t}{1}\binom{N-X_t}{1}}{\binom{N}{2}} = \frac{2X_t(N - X_t)}{N(N - 1)} = \frac{N}{N - 1} 2p_t(1 - p_t)$$



## A.4 Coefficient d'hétérozygotie avec remise

$$\begin{aligned}
h_t &= \mathbb{E}[H_t \mid p_{t-1}] \\
&= \mathbb{E}[2p_t(1 - p_t) \mid p_{t-1}] \\
&= 2\mathbb{E}[p_t(1 - p_t)] \\
&= 2\mathbb{E}[p_t - p_t^2 \mid p_{t-1}] \\
&= 2(\mathbb{E}[p_t \mid p_{t-1}] - \mathbb{E}[p_t^2 \mid p_{t-1}]) \\
&= 2(\mathbb{E}[p_t \mid p_{t-1}] - \text{Var}(p_{t-1} \mid p_{t-1}) - \mathbb{E}[p_t \mid p_{t-1}]^2) \\
&= 2(p_{t-1} - \frac{1}{N}p_{t-1}(1 - p_{t-1}) - p_{t-1}^2) \\
&= 2p_{t-1}(1 - p_{t-1})(1 - \frac{1}{N}) \\
&= H_{t-1}(1 - \frac{1}{N})
\end{aligned}$$

## A.5 Espérance et Variance de $Z_t$

$$\begin{aligned}
\mathbb{E}[Z_t] &= \mathbb{E}[Y_{t+\Delta t} - Y_t] \\
&= \mathbb{E}[Y_{t+\Delta t}] - \mathbb{E}[Y_t] \\
&= \frac{1}{N}\mathbb{E}[X_{t+\Delta t}] - \frac{1}{N}\mathbb{E}[X_t] \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\text{Var}[Z_t] &= \text{Var}[Y_{t+\Delta t} - Y_t \mid Y_t] \\
&= \text{Var}[Y_{t+\Delta t} - Y_t] \\
&= \text{Var}[Y_{t+\Delta t}] + \text{Var}[Y_t] \\
&= \text{Var}[Y_{t+\Delta t}] \\
&= \frac{1}{N^2}\text{Var}[X_{t+\Delta t}] \\
&= \frac{p(1 - p)}{N}
\end{aligned}$$

**Remarque:**  $Y_t$  est une valeur connue, donc  $\text{Var}[Y_t] = 0$

## A.6 Formule de Taylor

$$\begin{aligned}
m(p) &= 1 + \sum_{\Delta p} \mathbb{P}(Y_{\Delta t} = p + \Delta p \mid Y_0 = p) \left( m(p) + \frac{m'(p)}{1!} \Delta p + \frac{m''(p)}{2!} \Delta p^2 + O(\Delta p^3) \right) \\
&= 1 + m(p) + \mathbb{E}[\Delta p] m'(p) + \mathbb{E}[\Delta p^2] \frac{m''(p)}{2} + \mathbb{E}[O(\Delta p^3)] \\
&\Leftrightarrow 0 = 1 + \frac{1}{N} p(1-p) \frac{m''(p)}{2} \\
&\Leftrightarrow m''(p) = \frac{-2N}{p(1-p)}
\end{aligned}$$

**Remarque:** On utilise :  $\mathbb{E}[\Delta p] = \mathbb{E}[Z_t]$  et  $\mathbb{E}[\Delta p^2] = \mathbb{E}[Z_t^2]$ .

$p \in [0;1]$ , le terme  $O(\Delta p^3)$  est donc négligeable.

## A.7 Espérance et Variance $Z_t$ avec sélection

$$\begin{aligned}
\mathbb{E}\left[Z_t \mid Y_t = \frac{i}{N}\right] &= \mathbb{E}\left[Y_{t+1} \mid Y_t = \frac{i}{N}\right] - \mathbb{E}\left[Y_t \mid Y_t = \frac{i}{N}\right] \\
&= \mathbb{E}[Y_{t+1}] - \frac{i}{N} \\
&= \frac{(1+s)i}{is+N} - \frac{i}{N} \\
&= \frac{Nsi - i^2s}{N(is+N)} \\
&= \frac{is(N-i)}{N(is+N)} \\
&= p \frac{s(1-p)}{1+ps}
\end{aligned}$$

$$\begin{aligned}
Var\left[Z_t \mid Y_t = \frac{i}{N}\right] &= Var\left[Y_{t+1} \mid Y_t = \frac{i}{N}\right] + Var\left[Y_t \mid Y_t = \frac{i}{N}\right] \\
&= Var\left[Y_{t+1} \mid Y_t = \frac{i}{N}\right] \\
&= Var\left[Y_{t+1}\right] \\
&= \frac{1}{N} \frac{(1+s)i}{is+N} \left(1 - \frac{(1+s)i}{is+N}\right) \\
&= \frac{1}{N} \frac{(1+s)i}{is+N} \frac{N-i}{is+N} \\
&= \frac{1}{N} \frac{(1+s)i(N-i)}{(is+N)^2} \\
&= \frac{1}{N} \frac{p(1+s)(1-p)}{(1+ps)^2}
\end{aligned}$$

## A.8 Temps de fixation avec sélection

Pour rappel on avait :

$$0 = 1 + p \frac{s(1-p)}{1+ps} m'(p) + \frac{p(1-p)}{(1+ps)^2} \left[ \frac{1}{N}(1+s) + ps^2(1-p) \right] \frac{m''(p)}{2}$$

On pose  $s = \frac{\alpha}{N}$  avec  $\alpha \neq 0$

De plus on sait que  $\forall x \in \mathbb{R} \setminus \{-1\}$ ,  $f(x) = \frac{1}{1+x}$  admet un développement limité en 0 à l'ordre  $n$  tel que :

$$f(x) = \frac{1}{1+x} = 1 - x + x^2 - \dots + (-1)^n x^n + O(x^{n+1})$$

A l'ordre  $n = 1$ , on a alors :

$$f(x) = \frac{1}{1+x} \simeq 1 - x + O(x^2)$$

$\mathbb{E}[Z_t]$  s'exprime alors :

$$\begin{aligned} \mathbb{E}[Z_t] &= p \frac{s(1-p)}{1+ps} \\ &= p(1-p) \frac{\frac{\alpha}{N}}{1+p\frac{\alpha}{N}} \\ &= p(1-p) \frac{\alpha}{N} \frac{1}{1+p\frac{\alpha}{N}} \end{aligned}$$

On pose  $x = p\frac{\alpha}{N}$ . On utilise alors le DL de  $f(x)$  à l'ordre 1 et on déduit que :

$$\begin{aligned} \mathbb{E}[Z_t] &\simeq p(1-p) \frac{\alpha}{N} (1 - p\frac{\alpha}{N}) \\ &\simeq p(1-p) \frac{\alpha}{N} - O\left(\frac{\alpha}{N}\right)^2 \\ &\simeq p(1-p) \frac{\alpha}{N} \end{aligned}$$

**Remarque:** Le terme  $p^2(1-p)(\frac{\alpha}{N})^2$  est approximé par un terme négligeable de l'ordre de  $O\left(\frac{\alpha}{N}\right)^2$

De la même façon, on exprime à nouveau  $\mathbb{E}[Z_t^2]$ .

On sait que  $\forall x \in \mathbb{R} \setminus \{-1\}$ ,  $g(x) = \frac{1}{(1+x)^2}$  admet un développement limité en 0 à l'ordre 1 tel que :

$$g(x) = \frac{1}{(1+x)^2} \simeq 1 - 2x + O(x^2)$$

On a donc :

$$\begin{aligned}
\mathbb{E}[Z_t^2] &= \frac{p(1-p)}{(1+ps)^2} \left[ \frac{1}{N}(1+s) + ps^2(1-p) \right] \\
&\simeq \frac{p(1-p)}{(1+ps)^2} \left[ \frac{1}{N}(1+s) \right] + O\left(\frac{\alpha}{N}\right)^2 \\
&\simeq \frac{p(1-p)}{N} \frac{(1+s)}{(1+ps)^2} \\
&\simeq \frac{p(1-p)}{N} \frac{(1+\frac{\alpha}{N})}{(1+p\frac{\alpha}{N})^2}
\end{aligned}$$

On pose à nouveau  $x = p\frac{\alpha}{N}$ . On utilise alors le DL de  $g(x)$  à l'ordre 1 et on déduit que :

$$\begin{aligned}
\mathbb{E}[Z_t^2] &= \frac{p(1-p)}{N} \left( 1 - \frac{2\alpha}{N}p \right) \\
&\simeq \frac{p(1-p)}{N} - O\left(\frac{\alpha}{N^2}\right) \\
&\simeq \frac{p(1-p)}{N}
\end{aligned}$$

En remplaçant dans notre expression, on obtient :

$$1 + p(1-p)\frac{\alpha}{N}m'(p) + \frac{p(1-p)}{2N}m''(p) = 0$$

Il s'agit de résoudre l'équation différentielle ci-dessus afin de déterminer  $m(p)$ .

$$\begin{aligned}
\Leftrightarrow 0 &= \frac{2N}{p(1-p)} + 2\alpha m'(p) + m''(p) \\
\Leftrightarrow 0 &= e^{2\alpha p} \frac{2N}{p(1-p)} + 2\alpha e^{2\alpha p} m'(p) + e^{2\alpha p} m''(p) \\
\Leftrightarrow 0 &= 2N \frac{e^{2\alpha p}}{p(1-p)} + [e^{2\alpha p} m'(p)]' \\
&\Leftrightarrow [e^{2\alpha p} m'(p)]' = -2N \frac{e^{2\alpha p}}{p(1-p)} \\
&\Leftrightarrow e^{2\alpha p} m'(p) = \int -2N \frac{e^{2\alpha p}}{p(1-p)} dp
\end{aligned}$$

On calcule alors  $f(p) = \int -2N \frac{e^{2\alpha p}}{p(1-p)} dp$  :

$$f(p) = 2N (e^{2\alpha} \text{Ei}(2\alpha(p-1)) - \text{Ei}(2\alpha p)) + C$$

$m(p)$  s'exprime alors :

$$\Leftrightarrow e^{2\alpha p} m'(p) = f(p)$$

$$\Leftrightarrow m'(p) = e^{-2\alpha p} f(p)$$

$$\Leftrightarrow m(p) = K + \int e^{-2\alpha p} f(p) dp$$

$$\Leftrightarrow m(p) = K + 2N \int e^{-2\alpha p} \left( e^{2\alpha} \text{Ei}(2\alpha(p-1)) - \text{Ei}(2\alpha p) + \frac{C}{2N} \right) dp$$

$$\Leftrightarrow m(p) = K - \frac{C e^{-2\alpha p}}{2\alpha} + \frac{N}{\alpha} \left[ e^{-2\alpha p} \text{Ei}(2\alpha p) - e^{-2\alpha(p-1)} \text{Ei}(2\alpha(p-1)) + \ln(1-p) - \ln(p) \right]$$

avec :

- $\alpha \neq 0$
- $\gamma \approx 0.5772157$  (constante d'Euler-Mascheroni)
- $\text{Ei}(2\alpha p) = - \int_{t=-2\alpha p}^{t=+\infty} \frac{e^{-t}}{t} dt$

Il s'agit de déterminer l'expression des constantes  $C$  et  $K$ , sachant que :

- $m(0) = 0$
- $m(1) = 0$

Pour cela, on utilise [6] :

$$\text{Ei}(x) = \gamma + \ln|x| + \sum_{k=1}^{+\infty} \frac{x^k}{kk!}$$

En remplaçant  $p$  par 0 et 1, on obtient le système suivant :

$$\begin{cases} 0 = K - \frac{C}{2\alpha} + \frac{N}{\alpha} (\gamma - e^{2\alpha} \text{Ei}(-2\alpha)) \\ 0 = K - \frac{C e^{-2\alpha}}{2\alpha} + \frac{N}{\alpha} (-\gamma + e^{-2\alpha} \text{Ei}(2\alpha)) \end{cases}$$

$$\Leftrightarrow \begin{cases} C = \frac{-2N}{e^{-2\alpha} - 1} \left[ 2\gamma + e^{2\alpha} (-\text{Ei}(-2\alpha) - e^{-4\alpha} \text{Ei}(2\alpha)) \right] \\ K = \frac{C}{2\alpha} - \frac{N}{\alpha} [\gamma - e^{2\alpha} \text{Ei}(-2\alpha)] \end{cases}$$

## A.9 Application Shiny- Courbes avec effet de sélection

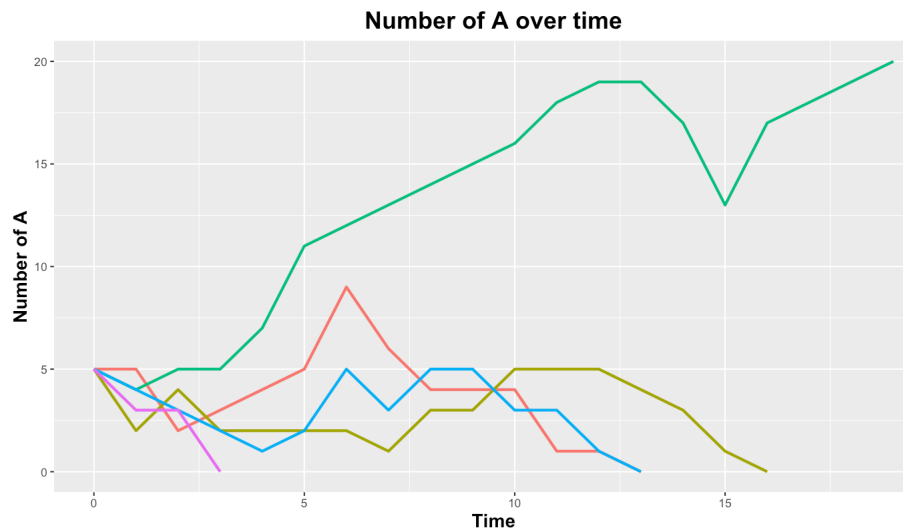


FIGURE A.1: Évolution du nombre de  $A$  en fonction du temps

*Population size : 10*

*Number of  $A$  : 5*

*Generation numbers : 5*

$\alpha : 5$

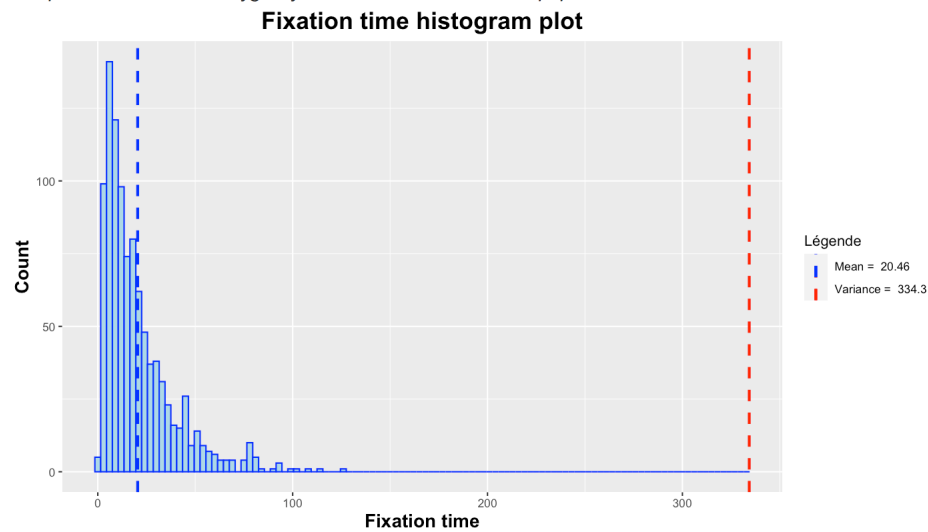


FIGURE A.2: Histogramme du temps de fixation

*Population size : 10*

*Number of alleles  $A$  : 5*

*Generations number : 1000*

$\alpha : 5$

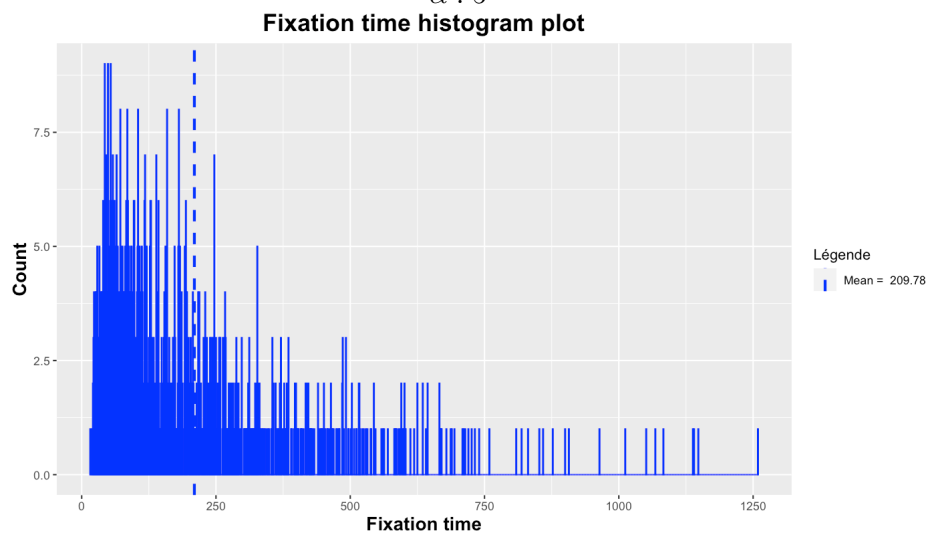
FIGURE A.3: Évolution du nombre de  $A$  en fonction du temps*Generation numbers : 5* $\alpha : 5$ 

FIGURE A.4: Histogramme du temps de fixation

*Population size : 100**Number of alleles  $A$  : 50**Generations number : 1000* $\alpha : 5$ 

**Remarque:** Pour  $N$  grand, la variance vaut 35027,78. Cette valeur étant trop importante, elle n'est pas représentée sur l'histogramme afin de pouvoir mieux observer la distribution du temps de fixation.

## A.10 Variance du temps de fixation

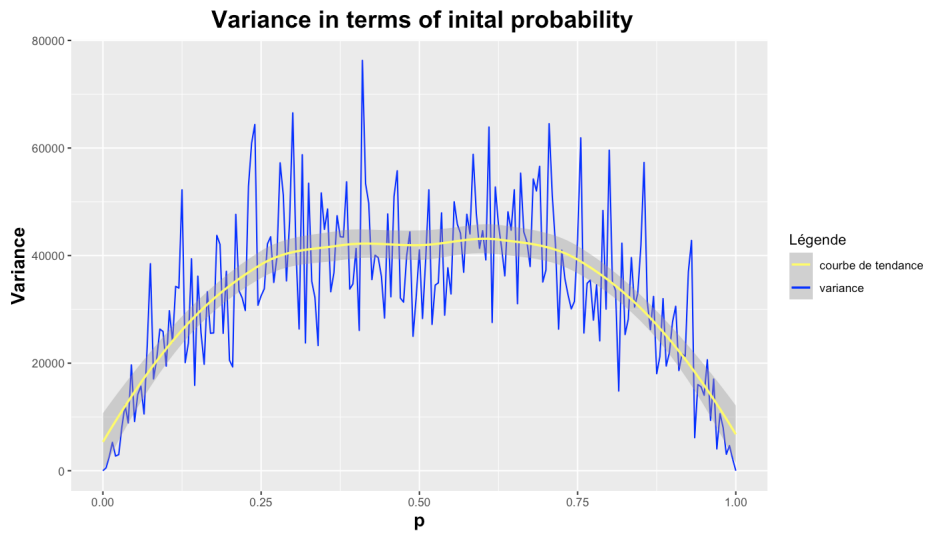


FIGURE A.5: Variance sans effet de sélection

*Population size : 100*

*Number of alleles  $A$  : 50*

*Simu Number : 100*

*Step : 1*

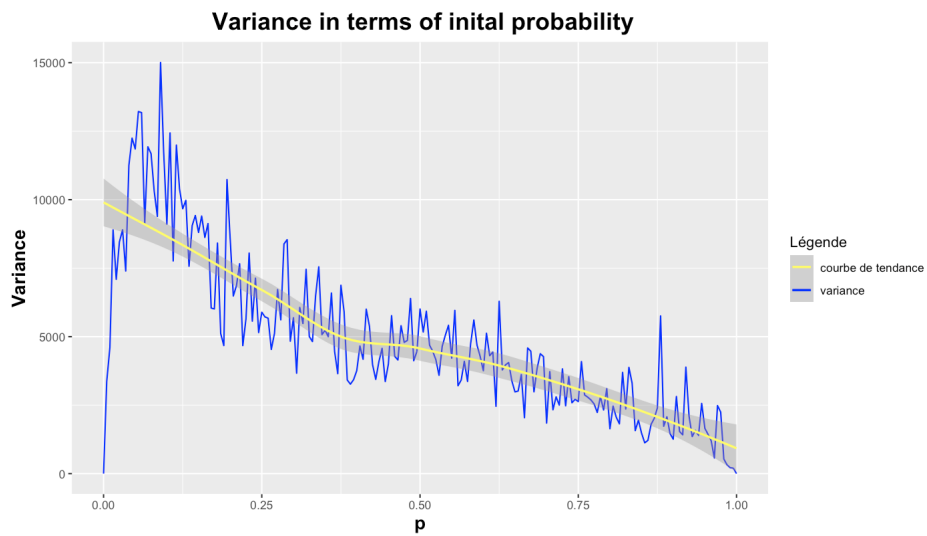


FIGURE A.6: Variance avec effet de sélection

*Population size : 100*

*Number of alleles  $A$  : 50*

*Simu Number : 100*

*Step : 1*

$\alpha : 5$