Du tirage dans les urnes à la théorie de l'évolution

Ludovic Goudenège et Pierre-André Zitt

15 janvier 2015

Objectifs

— Manipulation de la loi binomiale et des probabilités conditionnelles

— Utilisation dans un modèle de biologie.

Introduction

Quand Darwin développe au XIXè siècle la théorie de la sélection naturelle dans *On the origin of species*, il ouvre plus de débats qu'il n'en clôt : l'évolution est majoritairement acceptée, mais la façon dont elle a lieu (Est-elle graduelle ou par sauts ? Les caractères acquis sont-ils héréditaires ? Comment se séparent les espèces ? etc.) est largement débattue. Une de ces questions apparaît au début du XXè quand H. de Vries et C. Correns redécouvrent les travaux de Mendel sur l'hérédité : les mécanismes « rigides » de l'hérédité mendelienne semblaient incompatibles avec les variations continues des caractères observées par les biométriciens ; pour autant une hérédité « souple » (où les caractères de l'enfant sont une moyenne de ceux de ses parents) conduit à une homogénéisation de la population, sur laquelle le mécanisme de sélection ne peut plus opérer...

Notons par ailleurs que le XIX^{e} siècle a également vu apparaître les premiers modèles mathématiques de démographie, avec en particulier les modèles de Malthus (croissance exponentielle des populations) et de Verhulst (où la population croît en se rapprochant asymptotiquement d'une borne). Ces premiers modèles étaient déterministes (suites récurrentes ou équations différentielles) ; de très nombreux autres modèles déterministes plus complexes ont été introduits, et sont toujours utilisés, en dynamique des populations (modèles d'évolution proies/prédateurs, etc.)

Dans la première moitié du XXè siècle, de nombreux scientifiques aux premiers rangs desquels se trouvent R.A. Fisher, J.B.S. Haldane et S. Wright, jettent les bases ce qu'on appellera plus tard la « synthèse évolutive moderne », en particulier en introduisant et en étudiant des modèles de *génétique des populations*. Il s'agit d'étudier la transmission des caractères mendéliens, non pas sur quelques individus et une génération, mais sur une population toute entière, en un temps long. C'est l'un de ces modèles que nous allons étudier.

Pour simplifier on étudie l'évolution d'un seul *locus*, sur lequel un gène est codé, avec deux *allèles* possibles que nous appellerons « bleu » et « rouge » (on peut penser à « lisse » et « ridé », « blond » et « brun », ...). On suppose les individus *diploïdes* : les chromosomes vont par paire et chaque individu a deux copies du gène étudié.

Chaque individu peut donc être *homozygote* (si ses copies sont bleu-bleu ou rouge-rouge) ou *hétérozygote* (si l'une des copies est « bleu » et l'autre « rouge »).

Au début du XXè siècle, le mathématicien Hardy et le médecin Weinberg proposent indépendamment un modèle d'« évolution » de la répartition allélique. Ce modèle postule une population essentiellement infinie d'individus, dans laquelle les proportions initiales d'individus BB, BR et RR sont respectivement x_0, y_0 et z_0. En supposant que chaque individu de la génération 1 prend ses allèles au hasard parmi ceux de la génération précédente, on obtient à la génération 1 des proportions $(x_1, y_1, z_1) = \phi(x_0, y_0, z_0)$ d'individus de chaque type, où, en notant $p_0 = x_0 + y_0/2$ et $q_0 = y_0/2 + z_0$ les proportions respectives de gènes « bleus » et « rouge »,

$$x_1 = p_0^2 \qquad y_1 = 2p_0q_0, \qquad z_1 = q_0^2.$$

On vérifie alors facilement qu'à partir de là, les proportions ne changent plus lors des générations suivantes : en effet, (x_1, y_1, z_1) est un point fixe de ϕ, puisque $p_1 = x_1 + y_1/2 = p_0^2 + p_0q_0 = p_0$. Autrement dit, la variabilité génétique reste inchangée au cours du temps ; on parle d'« équilibre de Hardy–Weinberg ».

Dans des populations de taille finie, il y a nécessairement des fluctuations aléatoires sur la reproduction (certains individus peuvent mourir avant de se reproduire, d'autres peuvent avoir une descendance très nombreuse, etc.). Ces effets, qui ne se voient pas en population infinie (où la loi des grands nombres s'applique), sont essentiels pour expliquer les phénomènes de *dérive génétique* : le fait que les proportions alléliques changent au cours des générations même en l'absence de sélection naturelle.

Le modèle de Wright-Fisher cherche à étudier et quantifier ce phénomène de dérive pour comprendre son importance dans l'évolution.

Question biologique

1. Pour une population finie, a-t-on la même préservation de la diversité allélique ? Si elle disparaît, combien de temps survit-elle ?
2. Pour une population finie, un mutant avantagé peut-il envahir la population ?

1 Le modèle et les questions

1.1 Modèle à une génération

On fait les hypothèses simplificatrices suivantes.

1. les générations se succèdent sans se mélanger ;
2. chaque génération est composée de $N/2$ individus (et donc de N copies du gène) ;
3. la reproduction est aléatoire.

Pour préciser le dernier point, on suppose en fait que chaque individu de la génération $n+1$ choisit, uniformément au hasard, avec remise, et indépendamment les uns des autres, deux individus de la génération n pour être ses « parents ».

On « oublie » alors les individus en se concentrant sur la « population allélique » des N copies du gène.

Question mathématique

Notons $X_0 = k \in \{0, \dots N\}$ le nombre de copies bleues à la génération 0, et X_1 le nombre d'individus bleus à la génération 1.

Si $k = 0$, que vaut X_1 ? Que se passe-t-il si $k = N$?

Plus généralement, quelle est la loi de X_1 ?

Solution — On répète N fois, de façon indépendante, une expérience (choisir une copie au hasard parmi les N présentes à l'origine) qui peut avoir deux résultats : bleu ou rouge. Le nombre de « succès » (tirages bleus) suit donc une loi binomiale, de paramètres N et k/N. □

Quelques mots sur les hypothèses Les hypothèses faites sont naturellement totalement irréalistes. Faisons quelques brèves remarques à ce sujet.

— Ce modèle est suffisamment simple pour pouvoir incorporer facilement des modifications : on peut définir et étudier des modèles similaires prenant en compte la sélection, la mutation, plusieurs *loci*, une reproduction sexuée, etc.

— Il est aussi suffisamment simple pour être étudié, mathématiquement et/ou par simulations, en grand détail. Les outils développés pour ce modèle peuvent parfois être adaptés pour étudier des modèles plus complexes ; dans certains cas les conclusions du modèle simple restent approximativement vraies pour des modèles complexes. Dans tous les cas le modèle simple fournit un point de référence auquel comparer les versions complexes.

1.2 Le modèle de Wright-Fisher

On définit le modèle de Wright-Fisher en itérant simplement le procédé défini au-dessus.

1. Le nombre initial de copies bleues est $X_0 = k \in \{0, \dots, N\}$;
2. si il y a $X_n = j$ copies bleues à la génération n, alors le nombre de bleus à la génération $n+1$ suit la loi binomiale de paramètres N et j/N.

On note $\mathbf{P}_k[\cdot]$ la probabilité décrivant l'expérience partant de k copies bleues au temps 0. En particulier

$$\mathbf{P}_k[X_1 = l] = p_{kl} = \binom{N}{l} \left(\frac{k}{N}\right)^l \left(\frac{N-k}{N}\right)^{N-l}.$$

L'évolution du processus est représentée dans la figure 1.2 pour 5 générations et une population de 6 copies du gènes.

Sur la figure 2, on a représenté le résultat de 3 simulations numériques sur 100 générations pour 50 copies du gènes.

Sur ces simulations, on voit bien que trois cas semblent possibles :

— Le gène bleu s'éteint.

— Le gène rouge s'éteint.

— Les deux copies du gènes survivent.

Lorsqu'il y a extinction d'un gène, on dit qu'il y a absorption.

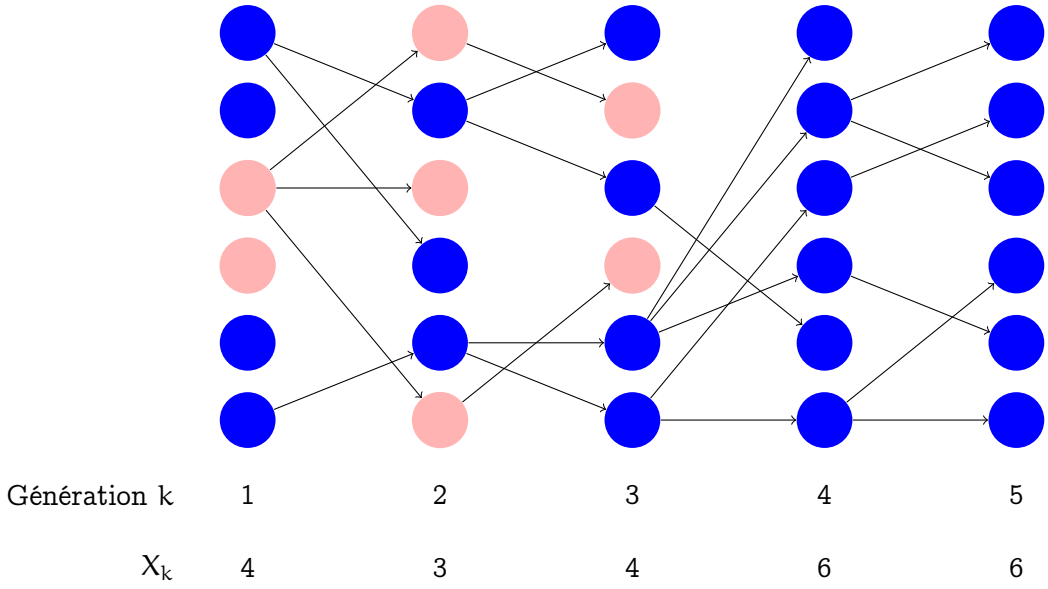
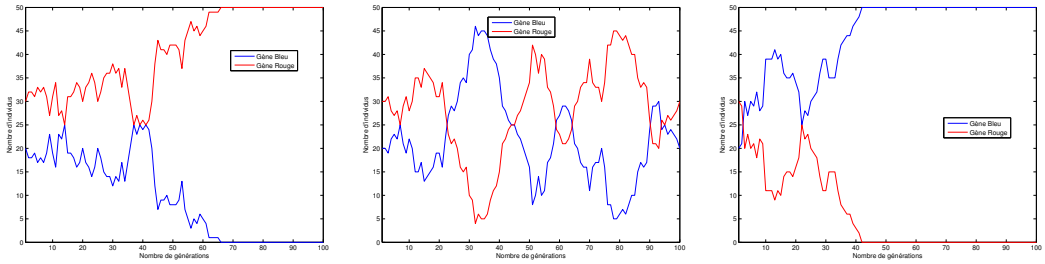


FIGURE 1 – L'évolution du processus sur 5 générations pour $N = 6$.



(a) Extinction gène bleu (b) Les deux gènes survivent (c) Extinction gène rouge

FIGURE 2 – Modèle de Wright-Fischer sur 100 générations.

1.3 Absorption

On note T le premier temps (aléatoire !) où l'un des allèles s'éteint : formellement

$$T = \inf\{n : X_n = 0 \text{ ou } X_n = N\}.$$

Sur le dessin ci-dessus, on a par exemple $T = 4$. Ce temps T prend *a priori* ses valeurs dans $\mathbb{N} \cup \{\infty\}$: il se pourrait que les deux allèles coexistent infiniment longtemps. On va montrer que le temps d'« absorption » est en réalité nécessairement fini.

Exercice

1. Donner $\mathbf{P}_k[T = 1]$ (pour $k = 1, \dots, N-1$).
2. On note $\rho = 1 - 2(1/N)^N$. Montrer que $\mathbf{P}_k[T > 1] \leq \rho$, pour tout $k \notin \{0, N\}$, et plus généralement,
$$\forall n \geq 0, \forall i \in \{1, \dots, N-1\}, \quad \mathbf{P}[T > n+1 | X_n = i] \leq \rho.$$
3. Montrer que, si $A \subset C$ et si C se partitionne en $C = \cup_{i=1}^m B_i$, alors
$$\mathbf{P}[A] = \sum_{i=1}^m \mathbf{P}[A|B_i]\,\mathbf{P}[B_i|C]\,\mathbf{P}[C].$$
4. En posant $A = \{T > n+1\}$, $B_i = \{X_n = i\}$ (pour $i = 1, \cdots, N-1$) et $C = \{T > n\}$, en déduire
$$\mathbf{P}_k[T > n+1] \leq \rho \mathbf{P}_k[T > n].$$
5. En déduire que $\mathbf{P}_k[T > n]$ tend vers 0 quand n tend vers l'infini.

Solution — Comme les événements $\{X_1 = 0\}$ et $\{X_1 = N\}$ sont disjoints,

$$\mathbf{P}_k[T = 1] = \binom{N}{0}(k/N)^0((N-k)/N)^N + \binom{N}{N}(k/N)^N((N-k)/N)^0 = \left(\frac{N-k}{N}\right)^N + \left(\frac{k}{N}\right)^N,$$

qui est minoré par $2(1/N)^N$. On en déduit

$$\mathbf{P}_k[T > 1] = 1 - \mathbf{P}_k[T = 1] \leq \rho.$$

Si $A \subset C$ et si les B_i partitionnent C,

$$\begin{aligned}
\sum_{i=1}^m \mathbf{P}[A|B_i]\,\mathbf{P}[B_i|C]\,\mathbf{P}[C] &= \sum_{i=1}^m \mathbf{P}[A|B_i]\,\mathbf{P}[B_i \cap C] \\
&= \sum_{i=1}^m \mathbf{P}[A|B_i]\,\mathbf{P}[B_i] && (B_i \subset C) \\
&= \sum_{i=1}^m \mathbf{P}[A \cap B_i] \\
&= \mathbf{P}[A \cap C] && (C \text{ union disjointe des } B_i) \\
&= \mathbf{P}[A] && (A \subset C).
\end{aligned}$$

L'événement $\{T > n\}$ correspond à dire « au temps n, il n'y a pas encore eu absorption », ce qui se décompose bien en $\bigcup_{i=1}^{N-1}\{X_n = i\}$. De plus, si $T > n+1$ il est bien supérieur à n, donc $A = \{T > n+1\} \subset \{T > n\} = C$. En appliquant la formule on en déduit :

$$\begin{aligned}\mathbf{P}_k[T>n+1] &= \sum_{i=1}^{N-1} \mathbf{P}_k[T>n+1|X_n=i]\,\mathbf{P}_k[X_n=i|T>n]\,\mathbf{P}_k[T>n] \\ &\geq \rho \sum_{i=1}^{N-1} \mathbf{P}_k[X_n=i|T>n]\,\mathbf{P}_k[T>n] \\ &\geq \rho \mathbf{P}_k[T>n]\end{aligned}$$

puisque $\sum_{i=1}^{N-1} \mathbf{P}_k[X_n=i|T>n]=1$. Par récurrence

$$\mathbf{P}_k[T>n] \leq \rho^n.$$

Comme $\rho<1$, $\mathbf{P}_k[T>n]$ tend vers 0. □

Théorème 1 (Extinction). *En un temps fini (mais aléatoire), l'un des allèles s'éteint.*

Question mathématique

1. Quelle est la probabilité que l'allèle bleu envahisse la population ?
2. En combien de temps le processus est-il absorbé ? En particulier quelle est l'espérance de T ?

2 Probabilité d'invasion : calcul explicite

Pour $k \in \{0,\dots,N\}$ notons :

$$f(k) = \mathbf{P}_k[\text{l'allèle bleu finit par envahir}]$$

la probabilité que le bleu envahisse la population quand on part de k individus bleus au temps 0. Cette probabilité peut être calculée explicitement, d'au moins 3 manières distinctes.

2.1 Résolution d'un système linéaire

Exercice

Écrire la formule des probabilités totales pour un événement A avec la partition $\Omega = \cup_{l=0}^{N}\{X_1=l\}$.

En déduire que, pour tout k,

$$f(k) = \sum_{l=0}^{N} p_{kl} f(l),$$

où l'on rappelle que $p_{kl} = \mathbf{P}_k[X_1=l] = \binom{N}{l}(k/N)^l((N-k)/N)^{N-l}$.

Vérifier que la fonction $f(k)=k/N$ est solution de cette équation.

Solution —

$$\mathbf{P}_k[A] = \sum_l \mathbf{P}_k[A|X_1=l]\,\mathbf{P}_k[X_1=l]\,.$$

Pour $A=$ « bleu envahit », $\mathbf{P}_k[A|X_1=l] = \mathbf{P}_l[A]$ d'où la formule annoncée.

Vérifions maintenant que la fonction $f : k \mapsto k/N$ est solution du système. Pour $k = 0$, $k = N$, l'équation dit $f(0) = f(0)$ et $f(N) = f(N)$, elle est donc vérifiée.

Pour $1 \leq k \leq N-1$, en notant $p = k/N$ et $q = (N-k)/N$,

$$\begin{aligned}\sum_{l=0}^{N} p_{kl} f(l) &= \sum_{l=0}^{N} \binom{N}{l} p^l q^{N-l} \frac{l}{N} \\ &= \frac{1}{N} \sum_{l=0}^{N} \binom{N}{l} p^l q^{N-l} \cdot l \\ &= \frac{1}{N} Np \\ &= k/N = f(k),\end{aligned}$$

où l'on a reconnu à la deuxième ligne l'espérance d'une binomiale de paramètre N et $p = k/N$. □

On peut montrer que si l'on impose $f(0) = 0$ et $f(N) = 1$, la solution de ce système est unique, et que l'on a donc bien trouvé la probabilité d'invasion de l'allèle bleu.

2.2 Une méthode probabiliste

Exercice

Si on part de $X_0 = k$ copies bleues, montrer que l'espérance du nombre X_1 de copies bleues au temps 1 vaut :

$$\mathbf{E}_k[X_1] = \sum_l p_{kl} \cdot l = k.$$

En écrivant

$$\begin{aligned}\mathbf{E}_k[X_2] &= \sum_{l=0}^{N} \mathbf{P}_k[X_2 = l]\, l \\ &= \sum_{l=0}^{N} \sum_{i=0}^{N} \mathbf{P}_k[X_2 = l | X_1 = i]\, \mathbf{P}_k[X_1 = i]\, l\end{aligned}$$

calculer $\mathbf{E}_k[X_2]$.

En déduire que $\mathbf{E}_k[X_n]$ ne dépend pas de n, et que l'on a donc :

$$k = \mathbf{P}_k[X_n = N]\, N + \left(\sum_{l=0}^{N-1} \mathbf{P}_k[X_n = l]\, l \right).$$

Conclure en faisant tendre n vers l'infini.

Solution — $\mathbf{E}_k[X_1]$ est l'espérance d'une binomiale et vaut donc $N \cdot k/N = k$.

$$\begin{aligned}
\mathbf{E}_k[X_2] &= \sum_{l=0}^{N} \mathbf{P}_k[X_2 = l]\, l \\
&= \sum_{l=0}^{N} \sum_{i=0}^{N} \mathbf{P}_k[X_2 = l | X_1 = i]\, \mathbf{P}_k[X_1 = i]\, l \\
&= \sum_{l=0}^{N} \sum_{i=0}^{N} p_{il} p_{ki} l \\
&= \sum_{i=0}^{N} p_{ki} \left(\sum_{l=0}^{N} p_{il} l \right) \\
&= \sum_{i=0}^{N} p_{ki} i \\
&= k.
\end{aligned}$$

L'itération du même raisonnement donne $\mathbf{E}_k[X_n] = k$, d'où la formule :

$$k = \mathbf{P}_k[X_n = N]\, N + \left(\sum_{l=1}^{N-1} \mathbf{P}_k[X_n = l]\, l \right).$$

Quand n tend vers l'infini, $\mathbf{P}_k[X_n = l]$ tend vers 0 puisque $\mathbf{P}_k[X_n = l] \leq \mathbf{P}_k[T > n]$. Donc $\mathbf{P}_k[X_n = N]$ converge, quand n tend vers l'infini, vers k/N. □

2.3 Par la généalogie

On peut justifier la formule donnant la probabilité de fixation de l'allèle bleu d'une troisième manière, en utilisant la généalogie. Enrichissons le processus en gardant la trace de quels individus de la génération $n-1$ ont donné naissance à quels individus à la génération n. Par des arguments similaires à ceux utilisés pour montrer que T est fini, on peut montrer qu'au bout d'un temps aléatoire T', toute la population est issue d'un unique individu de la génération 0 : dans le dessin fait plus haut, ce temps vaut $T' = 5$. Comme tous les choix aléatoires sont faits uniformément et ne dépendent pas des allèles, cet ancêtre commun (aléatoire) est uniformément distribué dans la population initiale. Il y a donc une probabilité k/N qu'il soit bleu, et avec lui toute la population au temps T'.

2.4 Bilan

Nous avons répondu complètement à une première question : dans le modèle élémentaire de Wright-Fisher, il y a bien « dérive génétique », dans le sens où l'un des allèles envahit forcément la population (sans qu'il y ait sélection naturelle !). La probabilité d'invasion de l'allèle bleu est explicite et particulièrement simple : elle est égale à la proportion initiale d'allèles bleus dans la population.

3 Questions complexes : l'approche par simulations

Revenons maintenant au reste des questions biologiques évoquées dans l'introduction :

Question biologique

1. Dans le cas neutre, que vaut le temps d'absorption ?
2. Dans le cas non-neutre où l'un des allèles a un avantage sélectif, quelle est la probabilité qu'il envahisse la population ?

Pour la première question, on peut rester dans le modèle simple ; une première réponse est de chercher l'espérance du temps aléaoire T.

Pour la deuxième question il est aisé de modifier le modèle en rajoutant de la sélection. Une façon de le faire est de remplacer le choix uniforme d'un allèle « ancêtre » par un choix pondéré. Pour cela on fixe un paramètre $s > 1$ qui quantifie l'avantage de l'allèle bleu sur l'allèle rouge. En supposant qu'il y a k copies de l'allèle bleue à la génération n, chaque copie de la génération $n+1$ est alors :

— bleue avec probabilité $\frac{ks}{ks+(N-k)}$,

— rouge avec probabilité $\frac{N-k}{ks+(N-k)}$.

La loi de X_1 est donc encore binomiale, mais de paramètres N et $\frac{ks}{ks+(N-k)} > (k/N)$.

Ce rajout, simple sur le modèle, casse la structure des calculs et empêche de trouver des formules explicites.

Ce problème se pose en fait déjà pour le calcul du temps moyen d'invasion dans le modèle neutre (sans sélection), pour lequel il n'y a pas de formule close. Biologiquement c'est une question importante (y a-t-il coexistence pendant un temps raisonnable ou extinction rapide d'un des allèles ?)

Pour le calcul de la probabilité d'invasion de l'allèle bleu dans le cas sélectif, on peut établir par des techniques relativement sophistiquées [1] l'approximation suivante :

$$\mathbf{P}_k\left[\text{bleu envahit}\right] \approx \frac{1-\exp(-\alpha\cdot p)}{1-\exp(-\alpha}),$$

où $p = k/N$ et $\alpha = 2N(s-1)$. Pour $N = 50$, $s = 1.1$ en partant de 20 copies bleues, on trouve une probabilité d'invasion du bleu de 98% ! Il suffit en fait de partir de plus de 4 copies bleues pour avoir plus de 50% de chances d'envahir la population...

On peut tester la précision de cette formule grâce à des simulations de Monte-Carlo : on répète un grand nombre de fois l'expérience, et on approche la probabilité de succès $f(k)$ par la fréquence empirique des succès. En notant MC le nombre de répétitions :

$$f(k) = \mathbf{E}\left[\mathbf{1}_{\text{bleu envahit}}\right] \approx \mathbf{E}_{MC}\left[\mathbf{1}_{\text{bleu envahit}}\right] := \frac{1}{MC}\sum_{m=1}^{MC} B_m,$$

où B_m vaut 1 si l'allèle bleu envahit lors de la m^e simulation, et 0 sinon. Comme on cherche à évaluer une probabilité, on peut utiliser l'intervalle de confiance classique :

$$I_C(\alpha) = \left[\mathbf{E}_{MC}\left[\mathbf{1}_{\text{bleu envahit}}\right] \pm \frac{1}{2\sqrt{MC}} z_\alpha\right]$$

1. À l'origine cette formule a été établie par des raisonnements directs sur le cas discret ; la technique la plus générale pour expliquer le résultat est l'approximation de l'évolution discrète du nombre de copies bleues par une évolution continue, semblable à un mouvement brownien.

où z_α est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite (et vaut ≈ 1.96 pour $\alpha = 0.05$).

On peut de même approcher le calcul de l'espérance de T par le calcul d'une moyenne empirique

$$\mathbf{E}[T] \simeq \mathbf{E}_{MC}[T] := \frac{1}{MC}\sum_{m=1}^{MC} T_m, \text{ pour } MC \to +\infty,$$

où T_m est le nombre (aléatoire) de générations avant absorption obtenu par la m^e simulation. On trouve $\mathbf{E}_{MC}[T] = 37,78$ pour $MC = 10^2$, $\mathbf{E}_{MC}[T] = 35,5854$ pour $MC = 10^4$.

4 Références

Sur l'historique des modèles en génétique des populations, nous nous sommes appuyés sur les premiers chapitres du livre :

— Warren J. Ewens, *Mathematical Population Genetics I*, Springer-Verlag, 2004.

De nombreux renseignements sont également disponibles sur Wikipedia, un bon point de départ est l'article :

— Modern evolutionary synthesis — wikipedia, the free encyclopedia, 2015. URL http://en.wikipedia.org/w/index.php?title=Modern_evolutionary_synthesis&oldid=640837016 [consulté le 13 janvier 2015].

Pour plus de détails mathématiques (en français) on pourra consulter le chapitre consacré au modèle dans le livre :

— *Recueil de modèles stochastiques* de D. Chafaï et F. Malrieu, disponible en ligne sur http://djalil.chafai.net/#livres.