BUERI L., LESCANNE S. & SCHNEIDER T.

Modélisation mathématique

Comment l'ataxie spastique s'est répandue dans la population ?

I. PRESENTATION DU PROBLEME

L'ataxie est une **maladie génétique héréditaire** qui provoque des problèmes de coordination (difficultés pour écrire, parler, marcher), ainsi que des raideurs et faiblesses musculaires dans les membres. Les capacités intellectuelles ne sont pas toutefois pas affectées.

Très souvent vers 30 ans, les malades ont besoin d'une canne pour se déplacer, et vers 50 ans ils sont en fauteuil roulant. Néanmoins, **l'espérance de vie est presque identique aux personnes non malades**.

Le gène responsable de cette maladie a été découvert en 2000. Il est situé sur le **13ème chromosome.**

La population de la région du Saguenay Lac-Saint-Jean, au Québec, **comporte une fréquence très élevée de cette maladie héréditaire**. Or elle est très rare, voire inexistante, dans les autres populations d'origine européenne dont descendent les habitants de cette région québécoise (descendants de pionniers européens du XVIIème siècle).

Comment expliquer que ces deux populations (les européens restés en Europe, et leurs descendants québécois) voient la maladie disparaître pour l'une, et se répandre pour l'autre ?

Nous présenterons trois modèles pour trouver une explication. Pour chaque modèle, nous présenterons ses conditions d'application, puis son fonctionnement. Pour finir, nous donnerons les limites de chacun, et évaluerons si l'explication apportée est satisfaisante.

Données sur la maladie :

Aujourd'hui, à Saguenay Lac-Saint-Jean, **1 personne sur 22 est porteuse (saine) et 1 sur 1 932 est malade on a donc 20 275 sur 21 252 personnes saines**.

Le gène porteur de la maladie est récessif et se situe sur l'autosome n°13.

Les malades sont féconds et engendrent normalement une descendance.

II. LE MODELE DE HARDY-WEINBERG

Les conditions d'application du modèle :

- Les couples se forment au hasard (panmixie).
- La population est infinie (afin de pouvoir considérer que les fréquences génotypes sont égales à leurs probabilités respectives).
- Il n'y a pas de mutation génétique.
- Il n'y a pas de sélection gamétique lors de la fécondation.
- Il n'y a pas de sélection zygotique (les trois génotypes ont la même espérance de vie).
- Il n'y a pas de différence de fertilité et il n'y a pas de migrations.

Le gène à l'origine de la maladie étant récessif, une personne saine possède deux gènes « sains », une personne porteuse possède un gène « malade » et un gène « sain », et une personne malade possède deux gène « malades ».

Comme chaque parent fourni aléatoirement un de ses deux gènes pour leur enfant, à partir des lois de la génétique, on peut dresser les trois tableaux des **probabilités conditionnelles** :
Chaque entrée représente un parent et les résultats sont donnés sous forme d'un pourcentage.

	M	P	S
M	100	50	0
P	50	25	0
S	0	0	0

Probabilités pour un couple d'avoir un enfant Malade

	M	P	S
M	0	50	100
P	50	50	50
S	100	50	0

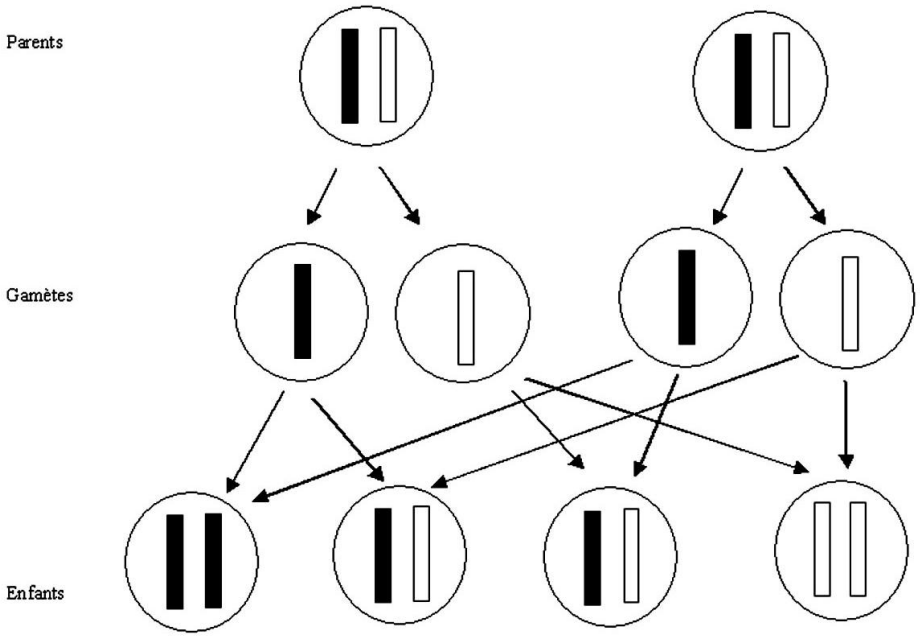
Probabilités pour un couple d'avoir un enfant Porteur

	M	P	S
M	0	0	0
P	0	25	50
S	0	50	100

Probabilités pour un couple d'avoir un enfant Sain

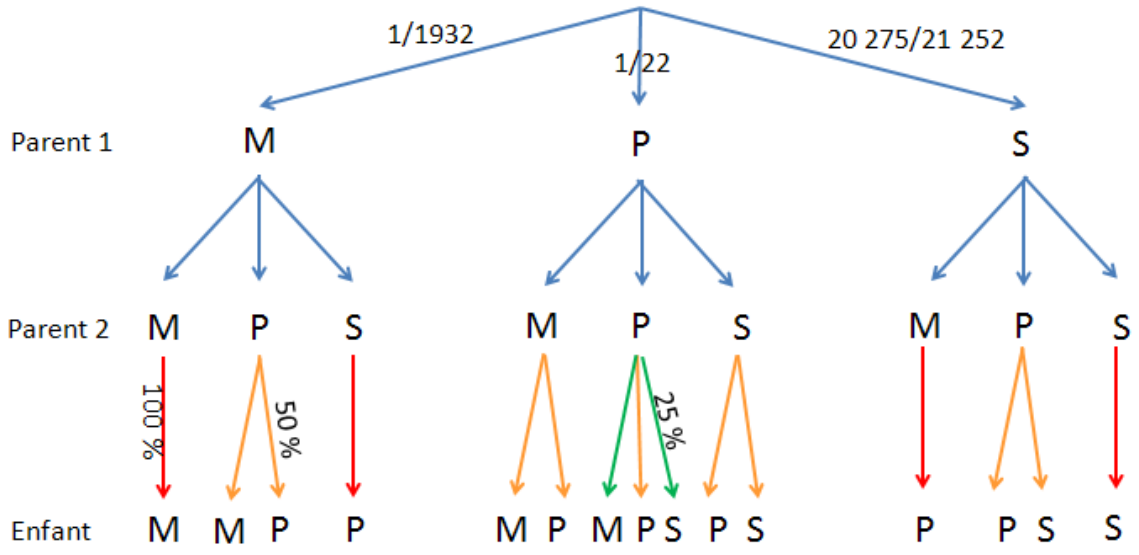
Par exemple, dans le tableau de gauche, si un parent est malade et l'autre porteur, il 50% de chance que l'enfant soit malade.
Autre exemple, dans le tableau de droite, si les deux parents sont porteurs, ils sont 25% de chance d'avoir un enfant sain.

Cela se retrouve par le schéma page suivante.



Les deux parents sont porteurs (un gène récessif en noir et un gène dominant en blanc). Les enfants sont malades dans un quart (25%) des cas, porteurs dans la moitié (50%) des cas, et totalement sains dans le dernier quart (25%) des cas. Ces trois valeurs se retrouvent dans chaque tableau à l'intersection des lignes et colonne « parent porteur ».

A partir de ces tableaux on construit un **arbre pondéré** :



On note M_n la proportion de malades à la nième génération. On fait de même avec P_n et S_n.

Par lecture de l'arbre, on aboutit aux relations de récurrences suivantes :

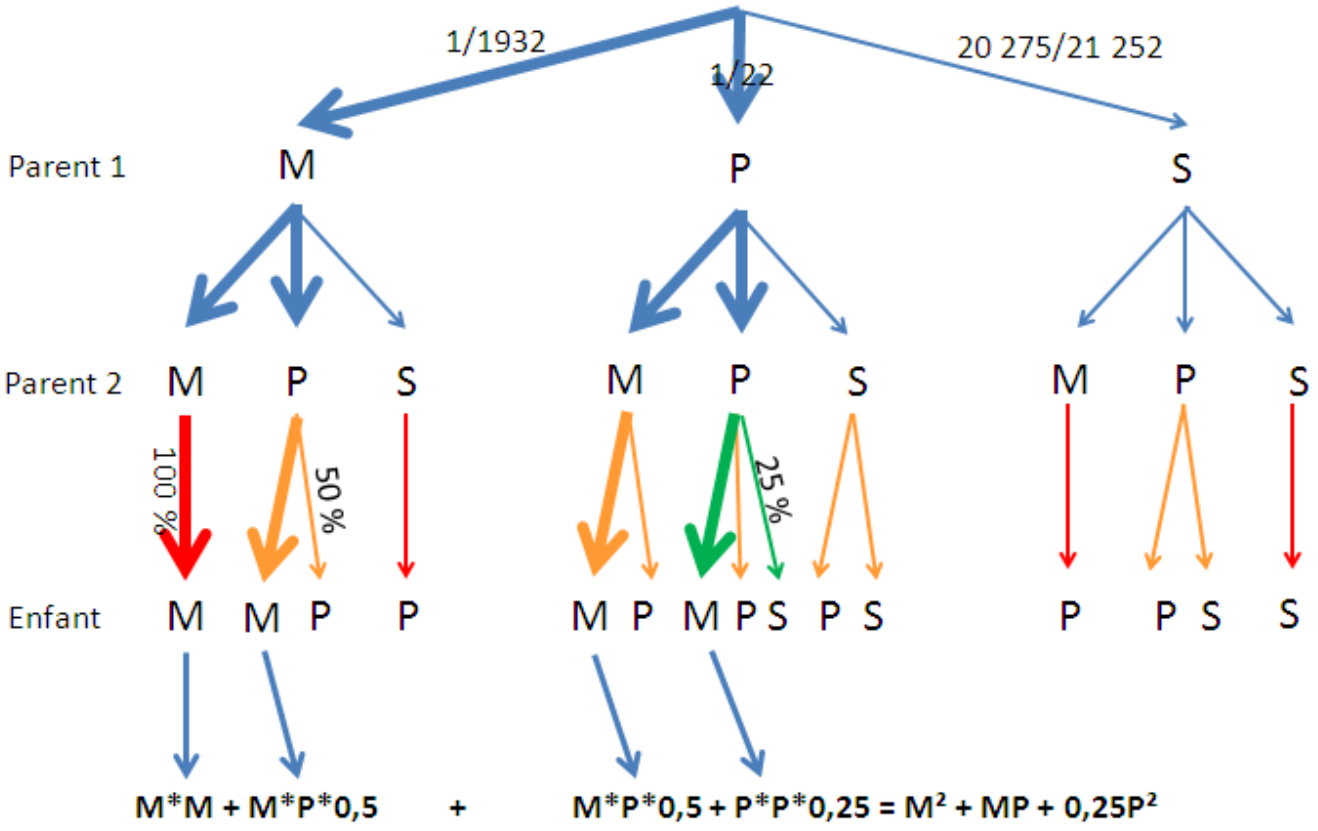
$$M_{n+1} = {M_n}^2 + M_n P_n + 0{,}25{P_n}^2$$

$$P_{n+1} = 0{,}5M_n P_n + M_n S_n + 0{,}5P_n M_n + 0{,}5{P_n}^2 + S_n M_n + 0{,}5S_n P_n$$

$$S_{n+1} = 0{,}25{P_n}^2 + 0{,}5P_n S_n + 0{,}5S_n P_n + {S_n}^2$$

Exemple (avec les probabilités observées aujourd'hui) :

Pour calculer M_{n+1}, on additionne les chemins aboutissant à M :



En factorisant et en utilisant le fait que $S_n = 1 - M_n - P_n$, on aboutit aux relations de récurrence :

$$M_{n+1} = \left(\frac{1}{2}P_n + M_n\right)^2$$

$$\text{®} \quad P_{n+1} = 2\left(\frac{1}{2}P_n + M_n\right)\left(1 - \frac{1}{2}P_n - M_n\right)$$

$$S_{n+1} = \left(1 - \frac{1}{2}P_n - M_n\right)^2$$

En prenant pour valeur initiale[1] : $M_0 = \frac{1}{1\,932}$, $P_0 = \frac{1}{22}$, $S_0 = 1 - (M_0 + P_0) = \frac{20\,275}{21\,252}$, on observe à partir de $n = 1$ un état d'équilibre aux valeurs :

$M_n = 0{,}000\,540\,324$, $P_n = 0{,}045\,409\,094$, $S_n = 0{,}954\,050\,582, \forall n \geq 1$

Propriété :
Avec ce modèle, on observe que la fréquence des malades se stabilise dès la première génération d'enfants.

[1] Données issues de www.uqac.ca (Université du Québec à Chicoutimi).

Démonstration :
Démontrons par récurrence que : « la proportion de malades se stabilise dès la première génération d'enfants à la valeur $M = \left(\frac{1}{2}P_0 + M_0\right)^2$ ».
Initialisation : Calculons M_1
D'après la première égalité de ®, on a $M_1 = \left(\frac{1}{2}P_0 + M_0\right)^2 = M$. CQFD
Posons $p = \frac{1}{2}P_0 + M_0$ et $q = 1 - p$ et regardons toutes les égalités. Ainsi, ® devient :

$$M_1 = p^2$$
$$P_1 = 2pq$$
$$S_1 = q^2$$

Hérédité : montrons que $\forall n \geq 1, M_n = M \quad M_{n+1} = M$

$$M_{n+1} = \left(\frac{1}{2}P_n + M_n\right)^2 = p^2$$

La relation ® devient $P_{n+1} = 2\left(\frac{1}{2}P_n + M_n\right)\left(1 - \frac{1}{2}P_n - M_n\right) = 2pq$

$$S_{n+1} = \left(1 - \frac{1}{2}P_n - M_n\right)^2 = q^2$$

Limite du modèle : les deux premières conditions ne sont pas vérifiées. En effet, l'hypothèse de la population infinie ne convient pas car la population de cette région est d'environ 200 000 habitants. Pour ce qui est de la panmixie (formation aléatoire des couples), elle n'est pas valable car la formation des couples dépend de critères sociaux que nous ne sommes pas capables de mesurer.
Il faut donc cherche un modèle plus pertinent.

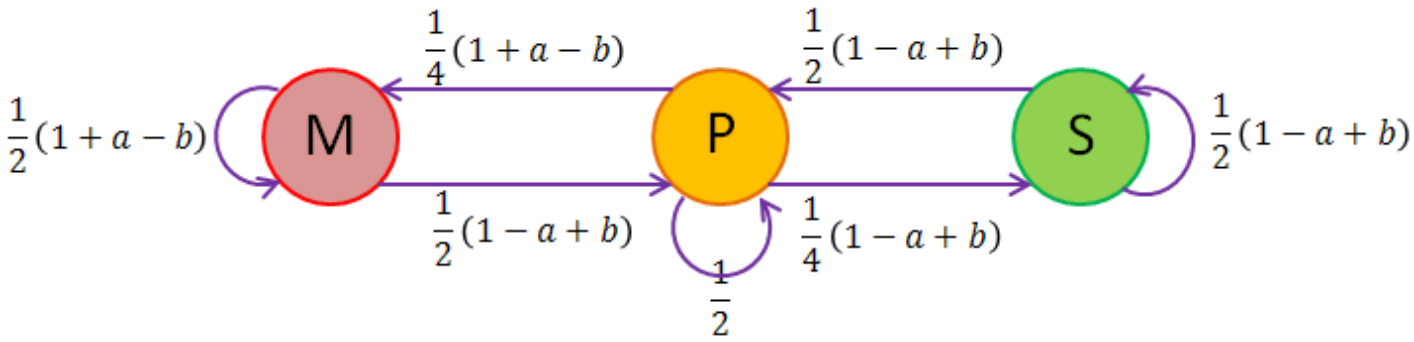
III. LES MARCHES ALEATOIRES (HOMOGENE ET NON HOMOGENE)

Dans ce modèle, on regarde le « changement d'état » d'UN parent à SON enfant. Plus précisément, la probabilité qu'un état (malade, porteur ou sain) devienne à la génération suivante dans un autre état sachant que son parent est malade, porteur ou sain.

Les conditions sont les mêmes que dans le modèle précédent ; en particulier, la population est toujours considérée comme infinie.

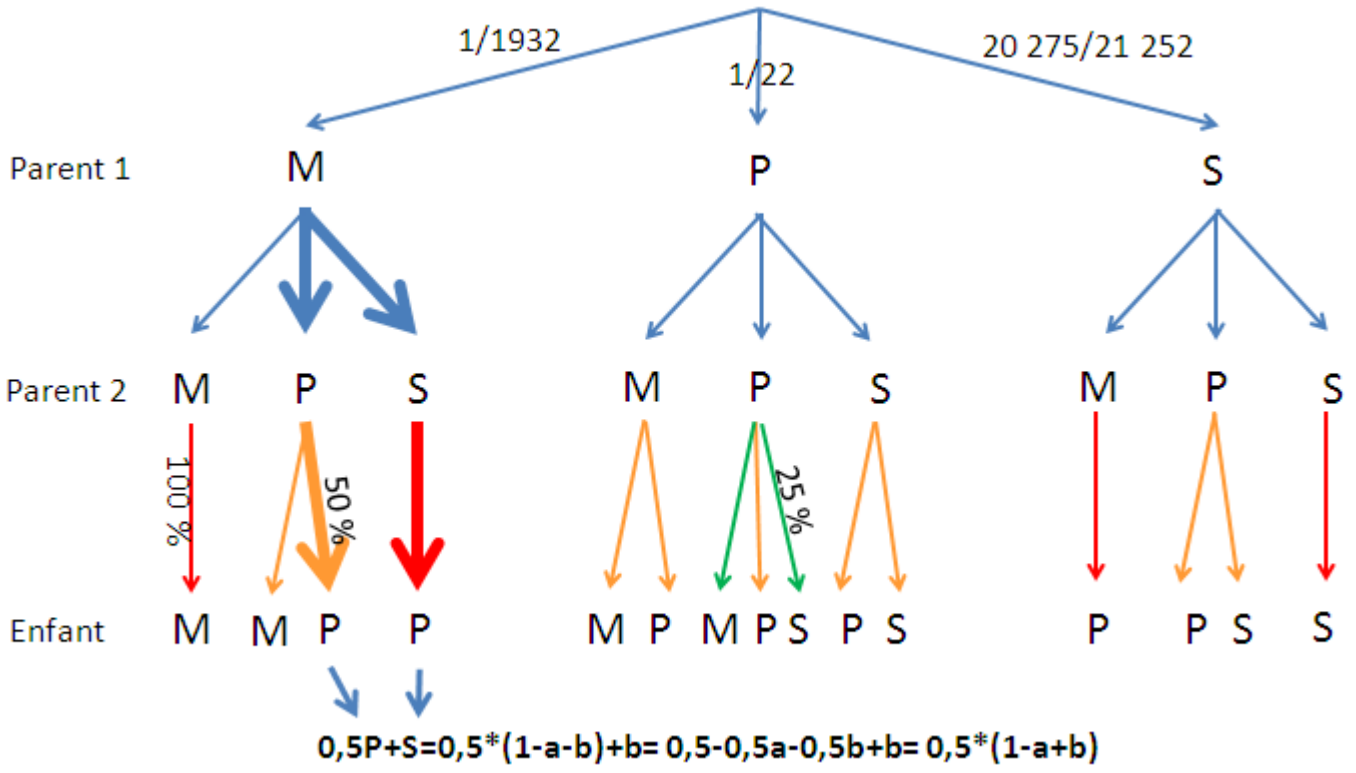
Le **graphe** modélisant la situation :

a est la proportion de malades et *b* la proportion de sains.
Les flèches indiquent qu'un changement d'état est possible et avec quelle probabilité.
Par exemple, la flèche de M vers P signifie que la probabilité qu'un parent malade ait un enfant porteur est de $\frac{1}{2}\left(1 - a + b\right)$.
L'absence de flèche indique que ce changement d'état est impossible. Par exemple, un parent malade ne pourra jamais avoir une descendance saine.



Exemple :

Pour savoir quelle est la probabilité d'avoir un enfant porteur sachant qu'on est malade est de avec M = a S = b et P = 1 – a – b



Soit A la **matrice de transition** associée au graphe et μ_n la matrice qui indique la proportion de malades à la génération n.

$$A = \begin{matrix} \frac{1}{2}\ 1+a-b & \frac{1}{2}\ 1-a+b & 0 \\ \frac{1}{4}\ 1+a-b & \frac{1}{2} & \frac{1}{4}\ 1-a+b \\ 0 & \frac{1}{2}\ 1+a-b & \frac{1}{2}\ 1-a+b \end{matrix}$$

Où a est la proportion de malades et b la proportion de sains.

C'est bien une matrice de transition car la somme des coefficients sur chaque ligne est égale à 1.

On note $\mu_0 = \begin{pmatrix} M_0 & P_0 & S_0 \end{pmatrix}$, le vecteur « proportions initiales » de la génération 0.
On obtient $\mu_1 = \begin{pmatrix} M_1 & P_1 & S_1 \end{pmatrix}$, le vecteur proportions de la génération 1 en faisant le calcul $\mu_0 \times A$.
Et de manière générale, on a la **relation de récurrence** entre une génération et la suivante :

$$\mu_{n+1} = \mu_n \times A$$

Ainsi, on a $\mu_n = \mu_0 \times A^n = \begin{pmatrix} M_n & P_n & S_n \end{pmatrix}$, le vecteur « proportions à la nième génération ».

Pour observer, on effectue les calculs dans le langage python :

Voici le code sous python 3.2.

```
μ = [1/1000,1/22,20978/22000]
a = 1/1000
b = 20978/22000

A = [[0.5*(1+a-b),0.5*(1-a+b),0] , [0.25*(1+a-b),0.5,0.25*(1-a+b)] , [0,0.5*(1+a-
b),0.5*(1-a+b)]]

for i in range (100) :

    μ =
[μ[0]*A[0][0]+μ[1]*A[1][0]+μ[2]*A[2][0],μ[0]*A[0][1]+μ[1]*A[1][1]+μ[2]*A[2][1],μ[0]*A
[0][2]+μ[1]*A[1][2]+μ[2]*A[2][2]]
    print (i+1,μ)
```

On observe la stabilité des fréquences à partir de la 10ème génération.

```
1  [0.0005629834710743785, 0.046328578512396620, 0.9531084380165290]
2  [0.0005629834710743775, 0.046328578512396586, 0.9531084380165289]
3  [0.0005629834710743771, 0.046328578512396565, 0.9531084380165288]
4  [0.0005629834710743767, 0.046328578512396550, 0.9531084380165287]
5  [0.0005629834710743766, 0.046328578512396540, 0.9531084380165286]
6  [0.0005629834710743764, 0.046328578512396530, 0.9531084380165284]
7  [0.0005629834710743764, 0.046328578512396520, 0.9531084380165283]
8  [0.0005629834710743763, 0.046328578512396520, 0.9531084380165282]
9  [0.0005629834710743763, 0.046328578512396516, 0.9531084380165282]
10 [0.0005629834710743762, 0.046328578512396510, 0.9531084380165282]
11 [0.0005629834710743761, 0.046328578512396510, 0.9531084380165282]
12 [0.0005629834710743761, 0.046328578512396510, 0.9531084380165282]
13 [0.0005629834710743761, 0.046328578512396510, 0.9531084380165282]
14 [0.0005629834710743761, 0.046328578512396510, 0.9531084380165282]
```

Ici, on fait l'hypothèse que *a* et *b* sont constants (chaîne de Markov homogène). Cette hypothèse n'est pas réaliste car justement, les proportion changent et ne se stabilisent qu'à la 10ème génération !

On réécrit donc le programme en prenant en compte, à chaque génération, les nouvelles valeurs de *a* et *b* (*a* : proportion de malades et *b* : la proportion de sains).

```
μ = [1/1000,1/22,20978/22000]
a = 1/1000
b = 20978/22000

for i in range (100) :

    A = [[0.5*(1+a-b),0.5*(1-a+b),0] , [0.25*(1+a-b),0.5,0.25*(1-a+b)] , [0,0.5*(1+a-
    b),0.5*(1-a+b)]]

    μ = [μ[0]*A[0][0]+μ[1]*A[1][0]+μ[2]*A[2][0] , μ[0]*A[0][1]+μ[1]*A[1][1]+μ[2]*A[2]
    [1] , μ[0]*A[0][2]+μ[1]*A[1][2]+μ[2]*A[2][2]]

    a = μ[0]
    b = μ[1]
    print (i+1,μ)
```

On observe cette fois de la stabilité à la 48ème génération !
En prenant une petite période de 20 ans par génération, il faut attendre presque un millénaire avant d'observer la stabilité. Cette échelle de temps est beaucoup trop longue pour l'humanité.

Propriété :
Avec ce modèle, on observe que les fréquences des malades se stabilisent dès la 10ème génération avec le modèle homogène et dès la 48ème avec le modèle non homogène.

Notre modèle doit être à nouveau pensé pour expliquer pourquoi cette maladie s'est répandue dans la population sur une durée beaucoup plus courte.

IV. AVEC LA LOI BINOMIALE (LA DERIVE GENETIQUE ET LE MODELE DE WRIGHT-FISHER) :

On s'inspire de la notion de dérive génétique :

La dérive génétique est l'évolution d'une population causée par des phénomènes aléatoires. Du point de vue génétique, c'est la modification de la fréquence d'un allèle au sein d'une population, indépendamment des mutations, de la sélection naturelle et des migrations. La dérive génétique est causée par des phénomènes aléatoires (comme les rencontres des spermatozoïdes et des ovules dans le cas d'une reproduction sexuée).

Les effets de la dérive génétique sont d'autant plus importants que la population est petite, car les écarts observés par rapport aux fréquences alléliques y seront d'autant plus perceptibles. Cette situation peut se produire dans une situation d'insularisation écologique. C'est le cas ici de notre population avec la migration des colons au XVIIème siècle. Ils se retrouvent isolés des populations européennes dont ils sont issus et ne se sont pas mêlés aux populations autochtones déjà présentes sur le sol américain (c'est ce qu'on appelle l'effet fondateur).

Les conditions d'application du modèle :

- La population est **fixe mais finie**.
- Il n'y a pas de mutation génétique.
- Il n'y a pas de sélection gamétique lors de la fécondation.
- Il n'y a pas de sélection zygotique (les trois génotypes ont la même espérance de vie).
- Il n'y a pas de différence de fertilité et il n'y a pas de migrations.

La différence fondamentale avec nos modèles précédents, c'est que la **population** est maintenant **finie**, et plus infinie. Par contre, on la supposera fixe (Dans la réalité, en trois siècle, la population à Saguenay-Lac Saint-Jean est passée de quelques milliers d'habitants à environ 280 000 en 2014).

Le modèle :

Soit une population de N personnes. Il y a donc $2N$ allèles pour cette génération (car une personne possède deux allèles de ce gène). Imaginons que nous plaçons ces $2N$ allèles dans une urne.

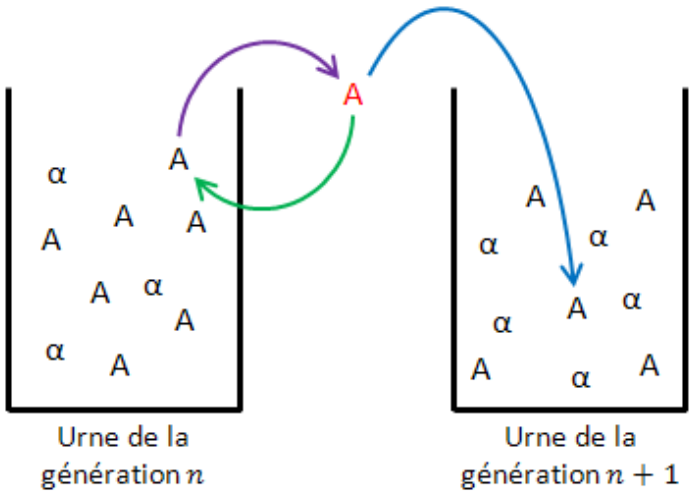
Comme il y aura N personnes à la génération suivante (hypothèse de population constante), il faut donc effectuer $2N$ tirages avec remises

Il faut noter qu'on s'intéresse ici à la fréquence allélique, peu importe qui porte effectivement ces allèles.

Exemple :

Le cas pour $N = 5$ ci-dessous. L'allèle A est tiré et placé dans l'urne de la génération suivante (transmission à la descendance), avant d'être replacé pour un nouveau tirage (possibilité d'engendrer de nouveaux enfants).

Dans cet exemple, la fréquence de l'allèle α passe ainsi de 0,3 (urne n) à 0,5 (urne n+1), même si on ne sait pas qui est porteur. On ne peut donc pas savoir combien il y a de malades, ni de porteurs sains, mais juste de nombre d'allèles.



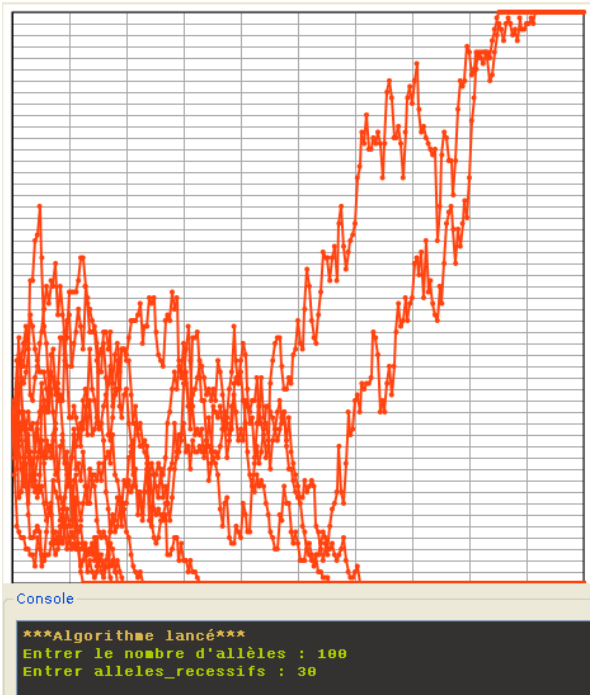
Exemple pour 10 allèles

A chaque tirage, il n'y a que deux issues (allèle malade ou allèle sain), et les tirages sont considérés comme indépendants. Nous pouvoir donc appliquer le modèle de **loi binomiale** de paramètres $2N$ et p, où p est la probabilité de tirer un allèle malade.

Pour simuler cela, on écrit donc un programme qui simule dans le cadre d'une loi binomiale, le nombre d'allèle malade de génération en génération.

```
1   VARIABLES : nombre_alleles, alleles_recessifs, compteur_alleles, generation, alea,
                PCC, boucle, recessifs_depart
2   DEBUT_ALGORITHME
3     LIRE nombre_alleles
4     LIRE alleles_recessifs
5     recessifs_depart PREND_LA_VALEUR alleles_recessifs
6     POUR boucle ALLANT_DE 1 A 10                    #pour faire 10 simulations
7      DEBUT_POUR
8       alleles_recessifs PREND_LA_VALEUR recessifs_depart
9       POUR generation ALLANT_DE 1 A 250
10       DEBUT_POUR
11        alea PREND_LA_VALEUR random()
12        compteur_alleles PREND_LA_VALEUR 0
13        PCC PREND_LA_VALEUR 0
14        TANT_QUE (PCC<alea) FAIRE
15         DEBUT_TANT_QUE
16          PCC PREND_LA_VALEUR PCC +
                        ALGOBOX_LOI_BINOMIALE(nombre_alleles,alleles_reces
                        sifs/nombre_alleles,compteur_alleles)
18          compteur_alleles PREND_LA_VALEUR compteur_alleles+1
19         FIN_TANT_QUE
20        TRACER_SEGMENT (generation-1,alleles_recessifs)->
                                              (generation,compteur_alleles-1)
21        alleles_recessifs PREND_LA_VALEUR compteur_alleles-1
22        TRACER_POINT (generation,alleles_recessifs)
23       FIN_POUR
24     FIN_POUR
25  FIN_ALGORITHME
```

Cet algorithme calcule l'évolution du pourcentage d'allèles récessifs au cours du temps.



On observe à gauche que deux simulations (deux courbes) sont absorbées vers l'état « tout le monde malade ». Cela signifie que pour ces 2 populations il n'y aurait que des allèles récessifs (pourcentage de 100%) donc tout le monde est malade.

On observe aussi que les autres simulations sont absorbées vers l'état « tout le monde sain ». Cela signifie que pour ces populations le pourcentage d'allèles récessifs est nul. Il n'y aurait donc plus de malade pour ces populations.

Remarque : Algobox ne peut effectuer des calculs de loi binomiale au–delà de 100 allèles.
On ne peut donc tester notre cas où il y aurait des milliers d'allèles.

Néanmoins, cela permet de comprendre les différentes évolutions possibles.

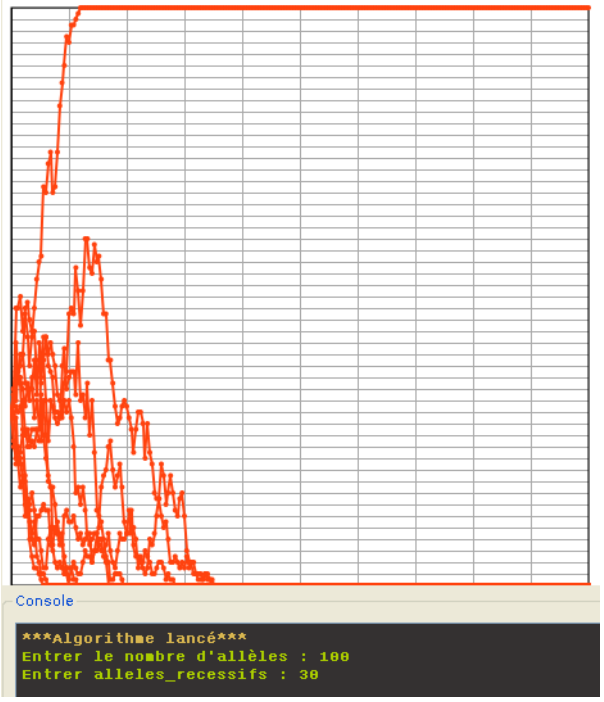
D'autres exemples de simulations :

On peut aussi changer le nombre d'allèles récessifs (entre 0 et 100). Plus on se rapproche de 0 plus on a de populations avec un pourcentage d'allèles récessifs proche de zéro et inversement quand on se rapproche de 100.

Une simulation est absorbée vers l'état « tout le monde malade », et les neuf autres sont absorbées vers l'état « tout le monde sain ».

En recommençant les simulations, on pourrait établir un tableau de données statistiques donnant la fréquence des états absorbés vers l'état « tout le monde malade » en fonction de la proportion initiale d'allèles récessifs.

Ceci n'a pas été fait faute de temps.



V. CONCLUSION

Nous avons donc étudié trois modèles pour tenter d'expliquer comment la maladie s'est répandue dans la population québécoise de Saguenay-Lac-Saint-Jean.

D'abord le modèle de Hardy-Weinberg. Mais il n'est pas satisfaisant car on aurait dû observer une stabilité des fréquences des malades dès la première génération.

Ensuite, le modèle des marches aléatoires. La version non-homogène semble mieux tenir compte de la réalité, mais sa durée d'application n'est pas réaliste. Il faudrait que la population reste isolée pendant un millier d'années !

Enfin, le modèle de Wright-Fisher. Ce modèle semble le plus réaliste.

La maladie se serait répandue dans la population par un phénomène de dérive génétique. Les premiers colons sont restés en entre eux, ne se mêlant pas à la population locale. Par un brassage aléatoire des allèles, la fréquence des allèles malades a augmenté, augmentant ainsi le nombre de malades au sein de cette population. Ce scénario n'était pas le plus probable, mais pas impossible.

Pour la population européenne, il y a eu plus de brassage. La fréquence des allèles malade a d'autant plus diminué, réduisant en conséquence la probabilité du scénario vers un état absorbant « tout le monde malade ».

Nos modèles ne sont pas parfaits, mais apportent une explication partielle. En effet, nous ne pouvons prendre en compte tous les paramètres (comme les paramètres sociaux), et nos moyens de calculs étaient limités.