

III. DEUX APPLICATIONS CLASSIQUES : les marches aléatoires et les lois de l'hérédité.

Une marche aléatoire à une dimension est le mouvement d'un point matériel qui fait des pas vers l'avant ou vers l'arrière sur un axe, chacune de ces deux possibilités étant choisie au hasard. On peut désigner par $t_0, t_1, t_2, \dots$ les instants successifs où ces pas sont effectués, et par $\varepsilon_0 = t_1 - t_0, \varepsilon_1 = t_2 - t_1, \varepsilon_2 = t_3 - t_2, \dots$ les intervalles de temps entre deux pas successifs. Alors $x_0, x_1, x_2, \dots$ seront les abscisses successives du point ; les amplitudes de chaque pas seront $\alpha_0 = x_1 - x_0, \alpha_1 = x_2 - x_1, \alpha_2 = x_3 - x_2, \dots$. On ne se préoccupera pas de la cinématique intérieure aux intervalles (en fait on fera comme si la vitesse entre deux pas consécutifs était constante).

À chaque pas, le sens (avant ou arrière) peut être choisi par un tirage à pile ou face, par un algorithme de nombres au hasard (par exemple la fonction **random** des langages de programmation usuels), par des collisions avec les molécules d'un liquide, ou tout autre procédé.

Nous allons étudier les marches aléatoires *uniformes*, pour lesquelles les ε_j et les valeurs absolues des α_j sont toutes égales : $\varepsilon_j = \varepsilon$ et $\alpha_j = \pm\alpha$. L'étude des marches aléatoires non uniformes n'est pas plus difficile dans le principe, mais les calculs à faire sont alourdis par la variation des ε_j et α_j : ces deux paramètres deviennent des *variables aléatoires* (voir chapitre **VI**). L'étudiant qui a compris le cas uniforme et assimilé le chapitre **VI** sur les variables aléatoires saura traiter lui-même le cas non uniforme (en fait les propriétés sont les mêmes).

Nous ne donnons ici que des rudiments sur les marches aléatoires, en traitant au moins par exemple les marches uniformes à une dimension. Les cas les plus intéressants de marches aléatoires sont en dimension deux ou trois, et non uniformes. Alors, à chaque pas, au lieu de *deux* choix possibles pour le sens, il y a toutes les directions possibles : cela correspond à une distribution de probabilité sur le cercle en dimension deux, sur la sphère en dimension trois (une telle distribution pouvant bien sûr être discrétisée), avec en outre une distribution de probabilité pour la longueur du pas. Une étude des marches aléatoires de dimensions supérieures à un est mathématiquement assez pénible mais semblable *dans son principe* au cas de la dimension 1.

Une marche aléatoire en dimension trois est un modèle mathématique pour le mouvement brownien. Mais beaucoup de propriétés du mouvement brownien se voient déjà sur les marches aléatoires uniformes de dimension un, d'où l'intérêt scolaire d'une telle étude.

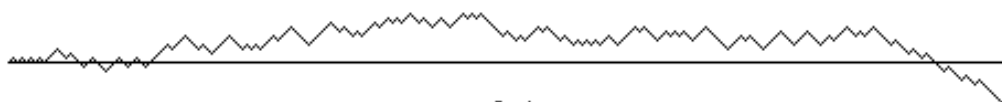
Il faut ajouter à cela que les marches aléatoires, surtout en dimension 1, peuvent aussi servir de modèle à des problèmes de probabilités très simples, tels que la partie de *pile ou face*. Le principal avantage de la modélisation par marche aléatoire est la possibilité de raisonner géométriquement : on peut alors utiliser toutes les ressources de la géométrie euclidienne (voir **III. 3** le principe de symétrie de Désiré André).

Un domaine d'application important et significatif du Calcul des probabilités est aussi la génétique. Depuis qu'on connaît les mécanismes moléculaires de l'hérédité on a pu en expliquer certaines lois empiriques par le Calcul des probabilités. Cela résulte de ce que ces mécanismes moléculaires sont des combinaisons de gènes obéissant à une causalité spatio-temporelle, et sont par conséquent soumis, comme les boules qu'on dispose dans des boîtes, aux différentes règles de dénombrement du chapitre **II**. Nous en donnerons un bref aperçu en présentant deux exemples simples et célèbres de lois de la génétique, celles de Mendel (section **III.5**) et de Hardy-Weinberg (section **III.6**). On verra que le Calcul des probabilités joue un rôle capital en génétique, car c'est lui qui en fait une science exacte (quantitative).

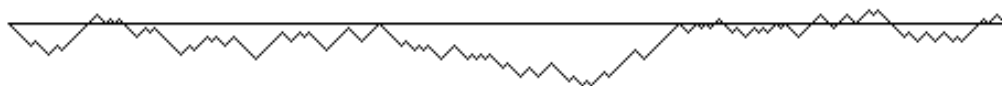
Le contenu de ce chapitre ne prétend à aucune originalité : on pourra en trouver un exposé pratiquement identique dans d'autres ouvrages, notamment celui de William Feller. Mais les applications traitées ici sont tellement typiques que nous les utiliserons fréquemment dans la suite comme exemples pour rendre plus concrète l'introduction de tel ou tel nouveau concept. Il serait gênant pour la cohérence du présent ouvrage de renvoyer à chaque fois le lecteur à d'autres sources.

III. 1. Graphe d'une marche aléatoire.

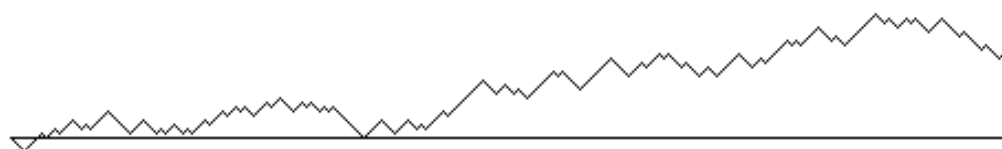
On suppose donc à partir de maintenant que les marches aléatoires sont toujours uniformes. Une marche aléatoire étant avant tout un *mouvement*, on peut représenter l'évolution du point matériel sur un graphique, avec le temps en abscisse et la position sur l'axe en ordonnée. Chaque instant $t_j = j\varepsilon$ peut correspondre à un changement du sens de parcours, ce qui sur un tel graphique se traduit par un changement de pente (voir figures pages suivantes). Dans le graphique on fait comme si la vitesse entre deux changements de direction consécutifs était constante, quoique ce détail soit



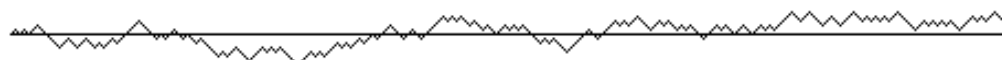
6.1.



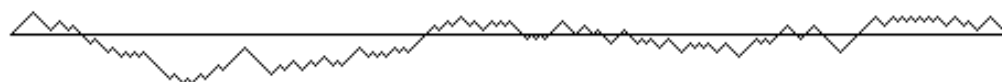
6.2.



6.3.



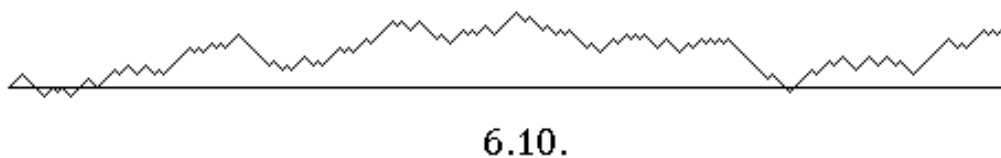
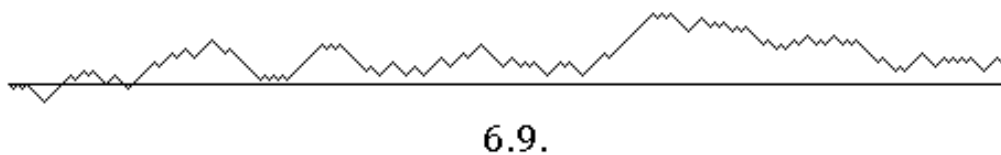
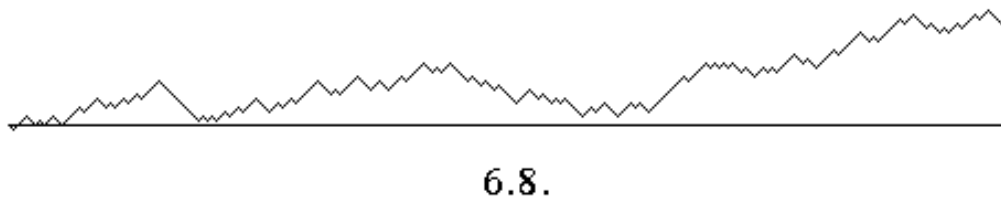
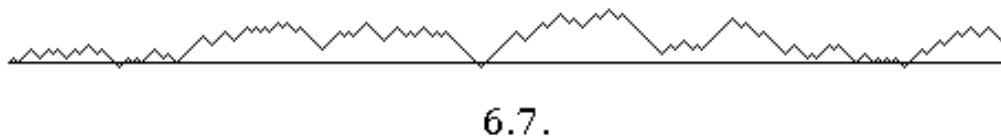
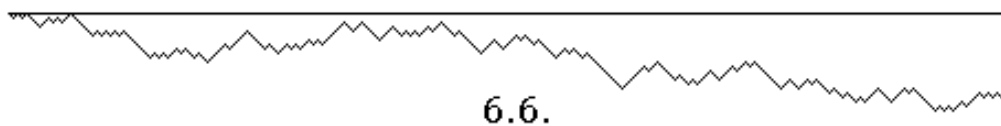
6.4.



6.5.

insignifiant ; par ailleurs, on choisit les unités sur les axes t et x de sorte que la vitesse soit toujours 1 en valeur absolue.

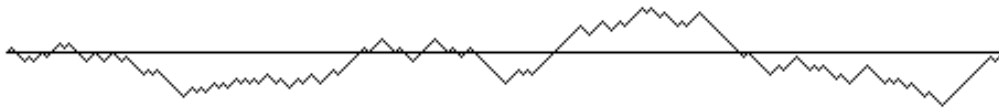
L'espace des épreuves pour une marche aléatoire à n pas est l'ensemble de *tous* les mouvements possibles ; mais chaque mouvement possible est déterminé par la liste des n choix de sens : on peut le représenter par une liste de n signes $+$ ou $-$. L'espace des épreuves Ω est donc isomorphe à celui des parties de pile ou face à n lancers, en particulier son cardinal est 2^n : il y a 2^n marches aléatoires différentes à n pas, qui correspondent à 2^n graphes différents, ou à 2^n suites différentes de $+$ et de $-$. Le hasard pur se traduit



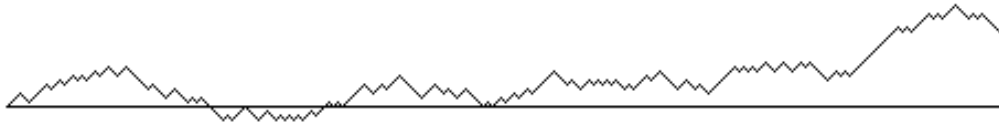
ici par le fait que toutes les marches possibles sont équiprobables.

Cette représentation géométrique sous forme de graphes est bien sûr *logiquement* équivalente à la représentation sous forme de suites formées de + et de -, mais le fait qu'elle soit géométrique permet de poser des problèmes de suites sous une forme imagée qui, comme nous le verrons, peut fournir des méthodes astucieuses de résolution.

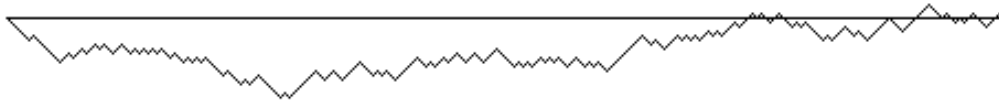
Ainsi une question qui se formule très aisément dans le langage géométrique est “quelle est la probabilité pour qu’une marche aléatoire partie de



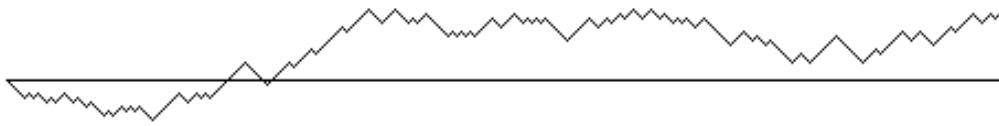
6.11.



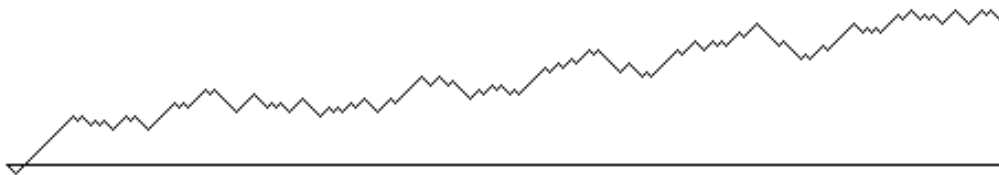
6.12.



6.13.

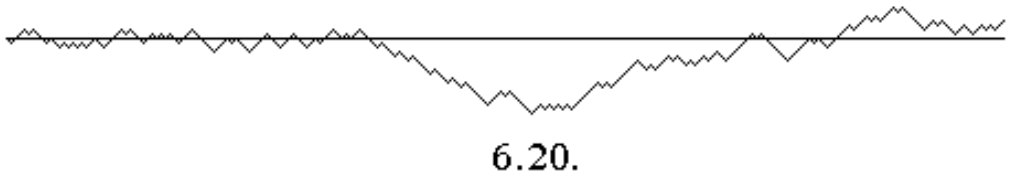
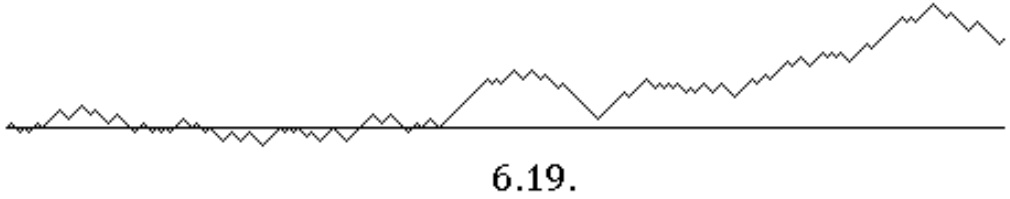
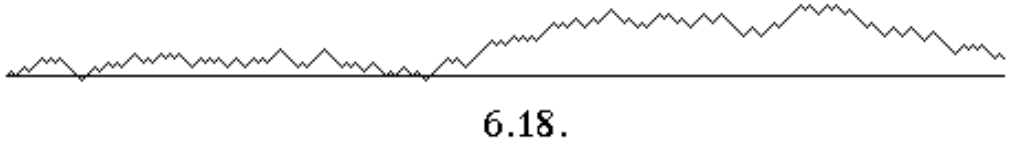
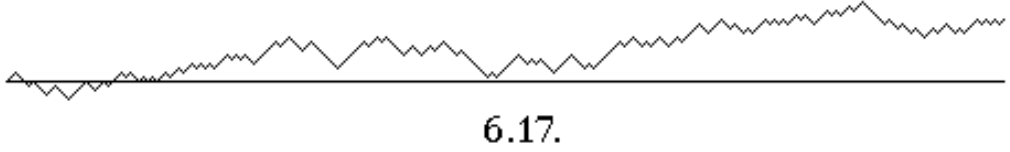
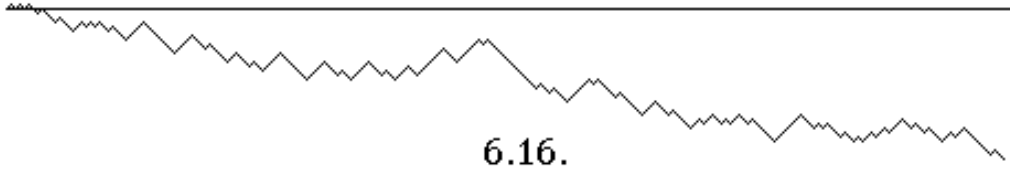


6.14.



6.15.

0 à l'instant 0 aboutisse à x à l'instant n ". Essayez donc de formuler la question équivalente dans le langage des suites de $+$ et de $-$. Pourtant, si on *pose* mieux la question dans le langage géométrique, on la *résoud* mieux en considérant les suites. En effet, si une suite est faite de p signes $+$ et de q signes $-$ (avec évidemment $p+q = n$), la marche correspondante parviendra à l'ordonnée $x = p - q$. Cette valeur ne dépend pas de l'ordre des $+$ et des $-$, mais seulement de leur nombre. Autrement dit, toutes les marches qui ont effectué p pas en avant et q pas en arrière aboutissent au même point x . Leur nombre est le nombre de manières différentes de placer p signes $+$ et



q signes $-$, soit $n!/p!q!$. Si on veut calculer p et q en fonction de x , il suffit de résoudre le système

$$p + q = n$$

$$p - q = x$$

si n et x ont la même parité, cela donne $p = \frac{1}{2}(n+x)$, $q = \frac{1}{2}(n-x)$ (si n et x n'ont pas la même parité il n'y a aucune solution). Ainsi l'événement A : "la marche, partie de 0 à l'instant 0 aboutit à x à l'instant n " est de cardinal $n!/p!q!$ avec $p = \frac{1}{2}(n+x)$, $q = \frac{1}{2}(n-x)$ si n et x ont la même parité, et de

cardinal 0 si n et x n'ont pas la même parité. D'où la probabilité :

$$\mathcal{P}(A) = \begin{cases} 2^{-n} \frac{n!}{p!q!} & \text{si } n \text{ et } x \text{ ont même parité} \\ 0 & \text{si } n \text{ et } x \text{ ont des parités différentes.} \end{cases}$$

III. 2. Le problème du retour à zéro d'une marche aléatoire.

Dans le paragraphe précédent nous avons cherché la probabilité pour qu'une marche de n pas aboutisse (au n^{e} , c'est-à-dire au *dernier* pas) au point d'ordonnée x . Pour $m < n$ on peut aussi chercher la probabilité pour qu'une marche (toujours de n pas) passe en un point donné x au m^{e} pas : soit donc $A_{m,x}$ l'événement : "la marche passe au point x au m^{e} pas". Ce qui a été vu au paragraphe précédent s'applique également. Pour une marche appartenant à $A_{m,x}$ il faut que parmi les m premiers pas il y ait p pas en avant et q pas en arrière avec $p+q = m$ et $p-q = x$; par contre ce qui arrive après le m^{e} pas est indifférent. Il y a donc (du moins si m et x sont de même parité) $m!/p!q!$ possibilités avec les m premiers pas et 2^{n-m} possibilités pour les pas suivants (de $m+1$ à n) de sorte que $\#A_{m,x} = 2^{n-m} \binom{m}{p}$ avec $p = \frac{1}{2}(m+x)$ si x a la même parité que m , et 0 sinon. On voit qu'on obtient pour la probabilité

$$\mathcal{P}(A_{m,x}) = \frac{\#A_{m,x}}{\#\Omega} = \frac{2^{n-m} \binom{m}{p}}{2^n} = 2^{-m} \binom{m}{p}$$

c'est-à dire la même chose que si l'espace des épreuves Ω avait été l'ensemble de toutes les marches à m pas au lieu d'être l'ensemble de toutes les marches à n pas. On voit que le résultat ne dépend pas de la modélisation choisie. Si donc on s'intéresse à ce qui se produit à l'instant m , ou avant, mais que ce qui arrive après est indifférent, il est inutile de considérer des marches ayant plus de m pas.

Un cas particulier intéressant est celui où $x = 0$. Si l'événement $A_{m,0}$ correspondant se produit, on dira aussi : "il y a un retour à zéro à l'instant m ". Cet événement est vide si m est impair, et de probabilité $2^{-2\ell} \binom{2\ell}{\ell}$ si $m = 2\ell$. Ce résultat se déduit par une application immédiate de (II.5.)

Pour les retours à zéro il se pose cependant un autre problème, celui du *premier* retour à zéro. La question est cette fois de trouver la probabilité pour qu'à l'instant $m = 2\ell$ la marche soit revenue à zéro, *et qu'en outre* il n'y ait eu aucun autre retour à zéro avant. Il est clair que la probabilité pour qu'à l'instant m se produise le *premier* retour à zéro est inférieure à la probabilité pour qu'à l'instant m se produise *un* retour à zéro. Mais quelle est sa valeur exacte ? C'est ce que nous nous proposons de calculer

maintenant. On voit que pour tous ces problèmes (retour à zéro ou premier retour à zéro à l'instant m) il est inutile de considérer ce qui se passe après l'instant m , et donc nous prenons pour Ω l'ensemble des marches de m pas, qui contient 2^m éléments.

L'événement "à l'instant m se produit un retour à 0" étant vide si m est impair nous posons $m = 2\ell$; les marches appartenant à cet événement ont en commun que $x_0 = 0$ et $x_{2\ell} = 0$. Par contre les marches appartenant à l'événement P_ℓ : "à l'instant 2ℓ se produit le *premier* retour à 0" vérifient en outre $x_1 \neq 0, x_2 \neq 0, x_3 \neq 0, \dots, x_{2\ell-1} \neq 0$. On ne peut dénombrer P_ℓ par application immédiate d'une formule du chapitre II. Ce type de problème est plus délicat, mais nous allons voir qu'en exploitant la géométrie on peut trouver des méthodes efficaces; en effet, le grand avantage des marches aléatoires est leur sens géométrique: les graphes se situent dans le plan euclidien et par conséquent toute la géométrie euclidienne peut être mise à profit. Or beaucoup de problèmes de probabilités qui au départ n'ont rien à voir avec la géométrie peuvent se modéliser en termes de marches aléatoires: par exemple un jeu de pile ou face avec mises entre deux joueurs P et F , où P reçoit 1 Franc de F chaque fois que sort *pile*, tandis que P donne 1 Franc à F chaque fois que sort *face*. Une marche aléatoire est en effet caractérisée de façon biunivoque par la suite des signes $+$ et $-$ de chacun de ses pas. Si donc on traduit $+$ par *pile* et $-$ par *face* on voit qu'il y a une correspondance bijective entre l'ensemble de toutes les marches possibles et l'ensemble de toutes les parties de pile ou face possibles. L'ordonnée x_j est alors le gain de P à l'instant j . L'événement P_ℓ signifie dans ce cas que le gain de P est resté constamment positif de l'instant 1 jusqu'à l'instant $2\ell - 1$.

Une première étape pour le dénombrement de P_ℓ fait déjà appel à la géométrie: elle consiste à remarquer que par symétrie, à tout graphe tel que $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2\ell-1} > 0$ correspond bijectivement un graphe tel que $x_0 = 0, x_1 < 0, x_2 < 0, x_3 < 0, \dots, x_{2\ell-1} < 0$. Mais d'autre part un graphe tel que $x_0 = 0, x_1 \neq 0, x_2 \neq 0, x_3 \neq 0, \dots, x_{2\ell-1} \neq 0$ ne peut être que de l'un ou l'autre des deux types précédents, en vertu du fait bien connu qu'on ne peut passer du négatif au positif (ou vice-versa) sans passer par 0 (on pourrait passer par exemple de $-\alpha$ à $+\alpha$ avec un pas égal à 2α , mais la marche aléatoire ne fait que des pas égaux à $\pm\alpha$). Par conséquent l'événement $P_{2\ell}$: $x_0 = 0, x_1 \neq 0, x_2 \neq 0, x_3 \neq 0, \dots, x_{2\ell-1} \neq 0, x_{2\ell} = 0$ doit contenir exactement *deux* fois le nombre d'éléments de $P_{2\ell}^{(+)}$: $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2\ell-1} > 0, x_{2\ell} = 0$ ou de $P_{2\ell}^{(-)}$: $x_0 = 0, x_1 < 0, x_2 < 0, x_3 < 0, \dots, x_{2\ell-1} < 0, x_{2\ell} = 0$.

III. 3. Le principe de symétrie de Désiré André.

Cette utilisation des symétries géométriques pour le dénombrement est attribuée historiquement à Désiré André (1887). Elle consiste à établir des correspondances biunivoques entre événements par symétrie ou translation dans le plan (ou dans l'espace). Pour cela il faut bien sûr avoir trouvé auparavant une modélisation géométrique du problème.

Nous allons appliquer une seconde fois ce principe pour calculer le cardinal de $P_{2\ell}^{(+)}$ (et par voie de conséquence, celui de $P_{2\ell}$). On remarquera tout d'abord qu'une marche appartenant à $P_{2\ell}^{(+)}$ vérifie nécessairement $x_{2\ell-1} = +1$ et $x_{2\ell-2} = +2$; en effet on ne peut arriver à $x_{2\ell} = 0$ que par $x_{2\ell-1} = +1$ ou $x_{2\ell-1} = -1$, la seconde possibilité étant exclue dans $P_{2\ell}^{(+)}$. De même, on ne peut arriver à $x_{2\ell-1} = +1$ que par $x_{2\ell-2} = +2$ ou $x_{2\ell-2} = 0$, la seconde possibilité étant elle aussi exclue dans $P_{2\ell}^{(+)}$. Le nombre de marches vérifiant $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2\ell-1} > 0, x_{2\ell} = 0$ est donc égal au nombre de marches vérifiant $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2\ell-3} > 0, x_{2\ell-2} = 2$.



figure 7

Ainsi nous sommes ramenés à dénombrer l'événement $C_{2\ell-2}^2$: $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2\ell-3} > 0, x_{2\ell-2} = 2$. Appelons plus généralement pour n et r quelconques C_{2n}^{2r} l'événement $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2n-1} > 0, x_{2n} = 2r$. C_{2n}^{2r} est ainsi l'ensemble des marches aléatoires issues de 0 à l'instant 0 et aboutissant à $2r$ à l'instant $2n$, et qui restent constamment positives. Or une marche aléatoire issue de 0 à l'instant 0 et aboutissant à $2r$ (avec $2r > 0$) à l'instant $2n$ est ou bien toujours positive, ou bien peut s'annuler au moins une fois entre 0 et $2n$ (ce *ou bien* étant exclusif). C'est pourquoi nous allons considérer les marches aléatoires telles que $x_0 = 0, x_1 = +1, x_{2n} = 2r$ et les décomposer en deux parties disjointes, celles qui restent constamment positives, et celles qui repassent par 0 entre les instants 1 et $2n$.

Introduisons donc les événements :

$$A_{2n}^{2r} : x_0 = 0, x_1 = +1, x_{2n} = 2r$$

$$B_{2n}^{2r} : x_0 = 0, x_1 = +1, x_{2n} = 2r \text{ et il existe un } j, 1 < j < 2n, \text{ tel que } x_j = 0$$

$$D_{2n}^{2r} : x_0 = 0, x_1 = -1, x_{2n} = 2r$$

$$E_{2n}^{2r} : x_0 = 0, x_{2n} = 2r$$

L'événement A_{2n}^{2r} est donc l'ensemble de toutes les marches telles que $x_0 = 0, x_1 = +1, x_{2n} = 2r$, positives ou non; B_{2n}^{2r} est le sous-ensemble de A_{2n}^{2r} des marches qui repassent par 0, et, rappelons-le, C_{2n}^{2r} le sous-ensemble des marches toujours positives; de sorte que $A_{2n}^{2r} = B_{2n}^{2r} \cup C_{2n}^{2r}$, réunion de deux ensembles disjoints, d'où

$$\#A_{2n}^{2r} = \#B_{2n}^{2r} + \#C_{2n}^{2r}$$

Mais comme le montre la figure 7, on peut associer de manière biunivoque à tout graphe de B_{2n}^{2r} un graphe de D_{2n}^{2r} , en vertu de la symétrie par rapport à l'axe des abscisses : pour n'importe quelle marche appartenant à B_{2n}^{2r} , on prend la branche positive entre l'instant 0 et celui du premier retour à zéro (qui a forcément lieu puisqu'on est dans B_{2n}^{2r}), et on la remplace par sa symétrique, en laissant intact ce qui arrive après le premier retour à zéro.

Par conséquent :

$$\#B_{2n}^{2r} = \#D_{2n}^{2r}$$

(ceci constitue le principe de symétrie de Désiré André).

On en déduit alors un moyen de calculer $\#C_{2n}^{2r} = \#A_{2n}^{2r} - \#D_{2n}^{2r}$: en effet les cardinaux de A_{2n}^{2r} et D_{2n}^{2r} peuvent, eux (contrairement à celui de C_{2n}^{2r}), être calculés par application immédiate de (II.5.) : A_{2n}^{2r} équivaut à l'ensemble de toutes les marches issues de +1 à l'instant 1 et aboutissant à $2r$ à l'instant $2n$, qui doivent donc faire p pas en avant et q pas en arrière avec $p+q = 2n-1$ et $p-q = 2r-1$, il y en a $(2n-1)!/(n+r-1)!(n-r)!$; C_{2n}^{2r} équivaut à l'ensemble de toutes les marches issues de -1 à l'instant 1 et aboutissant à $2r$ à l'instant $2n$, qui doivent donc faire p pas en avant et q pas en arrière avec $p+q = 2n-1$ et $p-q = 2r+1$, il y en a $(2n-1)!/(n+r)!(n-r-1)!$. D'où

$$\begin{aligned} \#C_{2n}^{2r} &= \#A_{2n}^{2r} - \#D_{2n}^{2r} \\ &= \binom{2n-1}{n+r-1} - \binom{2n-1}{n+r} \\ &= \frac{(2n-1)!}{(n+r-1)!(n-r)!} - \frac{(2n-1)!}{(n+r)!(n-r-1)!} \quad (III.1.) \end{aligned}$$

$$\begin{aligned}
 &= \frac{(2n-1)!}{(n+r)!(n-r)!} \cdot 2r \\
 &= \frac{(2n)!}{(n+r)!(n-r)!} \cdot \frac{r}{n}
 \end{aligned}$$

Il ne reste plus qu'à voir que dans le cas particulier qui nous intéresse ($r = 1$ et $2n = 2\ell - 2$) on a $\#C_{2\ell-2}^2 = ((2n)!/n!) \cdot (1/2(2n-1))$. D'après ce qui avait été dit plus haut l'événement $P_{2\ell}$ a un cardinal exactement double, de sorte que la probabilité que le *premier* retour se produise à l'instant 2ℓ est

$$\mathcal{P}(P_{2\ell}) = 2^{-2\ell} \frac{(2\ell)!}{\ell!^2} \frac{1}{2\ell-1} \quad (III.2.)$$

on voit qu'elle est $2\ell - 1$ fois plus petite que la probabilité pour qu'il y ait un retour à zéro (non nécessairement le premier) à l'instant 2ℓ .

III. 4. La loi du dernier retour.

Un autre problème qui mérite une étude est le suivant : considérons une marche aléatoire de $2n$ pas⁽¹⁾. Quelle est la probabilité pour que le précédent retour à zéro ait eu lieu à l'instant $2k$? (c'est-à-dire que la marche soit passée par zéro à l'instant $2k$ mais plus jamais ensuite entre $2k$ et $2n$).

Peut-être cette façon de formuler le problème n'est-elle pas assez concrète. Alors imaginons les choses ainsi : supposons qu'on reproduise des milliards de fois une marche aléatoire (par exemple un ordinateur dessine des milliards de graphes du type de la figure 6 à partir d'un programme qui appelle la fonction **random**). Pour chacun de ces graphes on note le dernier instant avant la fin où le graphe est passé par zéro, c'est-à-dire l'instant du dernier retour. Comment se répartissent ces instants ? Il se trouve que leur répartition est inattendue : il est fortement probable que ce dernier retour ait eu lieu à un instant soit tout récent (c'est-à-dire proche de $2n$), soit très ancien (c'est-à-dire proche de 0), et faiblement probable qu'il ait eu lieu à mi-chemin. Pour donner une idée plus quantitative, on peut dire que la probabilité pour que le dernier retour ait eu lieu pendant le premier dixième de la durée totale du parcours est $\frac{1}{5}$, la probabilité pour qu'il ait eu lieu pendant le dernier dixième est aussi $\frac{1}{5}$, et la probabilité pour qu'il ait eu lieu pendant les huit dixièmes intermédiaires est $\frac{3}{5}$.

Avec cette répartition des probabilités on doit s'attendre à ce que pour un échantillon de graphes pris au hasard, un nombre important d'entre eux

⁽¹⁾ Dans le cas impair $2n - 1$ le calcul est un peu différent mais de façon absolument inessentielle ; c'est pourquoi on peut le laisser de côté

aient fluctué tout au début autour de zéro, pour ensuite rester toujours loin de l'axe ; par exemple d'après les chiffres ci-dessus une marche sur cinq environ devrait fluctuer autour de zéro pendant le premier dixième de son parcours, puis demeurer constamment positive ou constamment négative sur les neuf dixièmes restants : un coup d'oeil à la figure 6 montre que tel est bien le cas (il n'y a eu aucun trucage). Une marche sur *dix* devrait ne fluctuer autour de zéro que sur le premier *cinquantième* de son parcours. Ces longs séjours de la marche aléatoire loin de 0 ne sont donc pas l'effet d'une cause inconnue qui favoriserait de tels écarts (par exemple si la fonction **random** du logiciel de calcul était incorrecte), mais bien celui du *pur hasard*.

Pour poser mathématiquement le problème, écrivons l'événement A_k qui, en tant que sous-ensemble de l'ensemble de toutes les marches possibles, exprime le fait que le dernier retour en 0 a eu lieu à l'instant $2k$. Une marche appartenant à A_k vérifie donc $x_0 = 0$ et $x_{2k} = 0$, mais aucune condition n'est requise pour les x_j avec $0 < j < 2k$. Pour une marche aléatoire allant de zéro à $2n$ il y a donc $\binom{2k}{k}$ possibilités différentes pour la partie ayant lieu entre 0 et $2k$. En ce qui concerne la partie entre $2k$ et $2n$, on commencera par remarquer que l'on doit avoir $x_{2k+1} \neq 0, x_{2k+2} \neq 0, x_{2k+3} \neq 0, \dots, x_{2n-1} \neq 0$, et aussi, si on suppose qu'il n'y a pas retour à zéro à l'instant $2n$ lui-même (car alors on aurait $k = n$), $x_{2n} \neq 0$. À cause de l'invariance par translation, il y a exactement autant de possibilités pour cette partie de la marche située entre $2k$ et $2n$ que pour une marche de $2n - 2k$ pas vérifiant $x_0 = 0, x_1 \neq 0, x_2 \neq 0, x_3 \neq 0, \dots, x_{2n-2k-1} \neq 0, x_{2n-2k} \neq 0$. L'ensemble de toutes les possibilités pour les $2n$ pas est alors le *produit* du nombre de possibilités pour les $2k$ premiers pas par le nombre de possibilités pour les $2n - 2k$ pas suivants, puisque à chaque possibilité pour les $2k$ premiers pas on peut ajouter n'importe laquelle des possibilités pour les pas suivants⁽²⁾.

Le problème est donc réduit à dénombrer les marches telles que $x_0 = 0, x_1 \neq 0, x_2 \neq 0, x_3 \neq 0, \dots, x_{2m-1} \neq 0, x_{2m} \neq 0$ pour m quelconque, après quoi on prendra $m = n - k$. Par symétrie leur nombre est exactement le double de celles qui vérifient $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2m-1} > 0, x_{2m} > 0$. Or une marche qui vérifie ces conditions appartient à un et un seul des événements A_m^r : $x_0 = 0, x_1 > 0, x_2 > 0, x_3 > 0, \dots, x_{2m-1} > 0, x_{2m} = 2r$ avec $r = 1, 2, 3, \dots, m$. Leur nombre est donc $\sum_{r=1}^{r=m} \#A_m^r$.

Enfin, pour calculer chacun des $\#A_m^r$, on utilise le principe de symétrie de Désiré André ; il suffit en effet de réutiliser (III.1.) qui nous dit que

$$\#A_m^r = \binom{2m-1}{m+r-1} - \binom{2m-1}{m+r}$$

⁽²⁾ Cette propriété est en fait l'indépendance stochastique, cf chapitre IV

On doit sommer cela de $r = 1$ à $r = m$, ce qui fait que les termes s'annulent mutuellement :

$$\begin{aligned} \sum_{r=1}^{r=m} \#A_m^r &= \binom{2m-1}{m} - \binom{2m-1}{m+1} + \binom{2m-1}{m+1} - \binom{2m-1}{m+2} + \\ &\quad + \binom{2m-1}{m+2} - \binom{2m-1}{m+3} + \binom{2m-1}{m+3} - \dots \\ &= \binom{2m-1}{m} = \frac{1}{2} \binom{2m}{m} \end{aligned}$$

Le nombre de toutes les marches de $2m$ pas ne revenant jamais à zéro est donc le double, soit $\binom{2m}{m}$; enfin, le nombre de marches de $2n$ pas, revenant à zéro à l'instant $2k$ et n'y revenant plus entre $2k+1$ et $2n$ inclus est alors (comme nous l'avions déjà dit plus haut) le produit $\binom{2k}{k} \times \binom{2(n-k)}{n-k}$.

De sorte que la probabilité pour que le dernier retour avant l'instant $2n$ se soit produit à l'instant $2k$ est

$$2^{-2n} \cdot \binom{2k}{k} \cdot \binom{2(n-k)}{n-k} \simeq \frac{1}{\pi \sqrt{k(n-k)}}$$

Si on cherche par exemple la probabilité pour que le dernier retour ait eu lieu *avant* l'instant 2ℓ , il suffit de faire la somme de ces valeurs pour k allant de 1 à ℓ (le dernier retour ne peut avoir lieu à deux instants différents à la fois, donc les événements correspondants sont disjoints et les probabilités s'additionnent). Ainsi

$$\begin{aligned} \mathcal{P}(\text{dernier retour avant } 2\ell) &= \sum_{k=1}^{k=\ell} 2^{-2n} \cdot \binom{2k}{k} \cdot \binom{2(n-k)}{n-k} \\ &\simeq \sum_{k=1}^{k=\ell} \frac{1}{\pi \sqrt{k(n-k)}} \end{aligned}$$

Cette somme discrète peut être interprétée comme la somme de Riemann de l'intégrale $\int_0^x [1/\pi \sqrt{t(1-t)}] dt$ avec $x = \ell/n$. Or cette intégrale peut être calculée par primitives, elle est égale à $\frac{2}{\pi} \arcsin(\sqrt{x})$, de sorte que finalement

$$\mathcal{P}(\text{dernier retour avant } 2\ell) \simeq \frac{2}{\pi} \arcsin\left(\sqrt{\frac{\ell}{n}}\right) \quad (III.3.)$$

Les estimations données au début du paragraphe ont été tirées de cette approximation. Bien sûr il faut être dans le domaine de validité de cette approximation : elle est en principe incorrecte si ℓ ou $n - \ell$ est trop petit,

mais elle donne déjà un résultat correct à 4% près pour k ou $n - k$ supérieur à 3, et à 1% près pour k ou $n - k$ supérieur à 10.

III. 5. Loi de l'hérédité de Mendel

La loi de Mendel concerne la transmission de caractères par l'hérédité. On appelle *phénotypes* ces caractères. Un exemple de phénotype (étudié par Mendel) est la forme, lisse ou ridée, des grains de petits pois. D'autres phénotypes connus sont les groupes sanguins, le facteur Rhésus, la couleur des yeux ou de certaines fleurs, le type albinos, des maladies génétiques telles que la mucoviscidose. Considérons l'expérience suivante, qu'on peut effectuer par exemple sur le *mirabilis*, une plante africaine dont les fleurs s'ouvrent la nuit (appelée pour cela "belle de nuit") : on a sélectionné sur un grand nombre de générations deux variétés de *mirabilis* : l'une à fleurs rouges, l'autre à fleurs blanches. Tout croisement entre deux fleurs rouges donne un descendant à fleurs rouges et tout croisement entre deux fleurs blanches donne un descendant à fleurs blanches. On effectue alors des croisements entre une fleur rouge et une fleur blanche, qui donnent toujours des descendants à fleurs roses. Puis on effectue à nouveau des croisements entre deux de ces descendants à fleurs roses, et on observe parfois une fleur rouge, parfois une blanche, parfois une rose. Si on renouvelle un grand nombre de fois les croisements de fleurs roses, afin de disposer d'un échantillon statistiquement significatif, on observe qu'on obtient une fois sur quatre une fleur rouge, une fois sur quatre une blanche, et une fois sur deux une fleur rose. Ces expériences avaient été effectuées par Gregor Mendel au milieu du *XIX*^e siècle sur des souches de petits pois (pois ridés et pois lisses).

Les conclusions que Gregor Mendel a tirées de ces expériences sont les suivantes. Le premier croisement entre une fleur rouge et une fleur blanche, qui donne une fleur rose, pourrait s'expliquer par le mélange de deux substances. Mais le second croisement, où l'on peut retrouver une fois sur deux une couleur pure, montre que le mélange n'a pas réellement eu lieu, et que les éléments matériels qui ont servi de vecteur aux caractères héréditaires ont dû se transmettre sous forme pure jusqu'à la génération suivante. Gregor Mendel a déduit de ses expériences en 1865 que ces éléments d'hérédité restaient séparés : les différents phénotypes résultaient bien de leur combinaison, mais les éléments devaient se combiner en conservant leur intégrité au cours des générations. La biologie du *XX*^e siècle a identifié ces éléments d'hérédité, qu'on appelle aujourd'hui les gènes.

Le principe de la transmission des caractères est donc le suivant : un gène R est responsable de la couleur rouge, un autre B de la couleur blanche ;

mais chacun des deux parents apporte un gène, et c'est la combinaison RR et non la présence d'un seul gène R qui produit la couleur rouge. De même c'est la combinaison BB qui produit la couleur blanche. Si l'un des parents est rouge et l'autre blanc, la combinaison sera RB ou BR et produira la couleur rose. On appelle *homozygote* une combinaison de deux gènes identiques, et *hétérozygote* une combinaison de deux gènes différents. Ainsi RR et BB sont des combinaisons homozygotes, RB et BR sont des combinaisons hétérozygotes. La théorie de Mendel postule que les mécanismes de la reproduction séparent à nouveau les combinaisons en deux gènes intacts, qui se recombinent autrement pour former la génération suivante. Si les deux parents sont rouges (combinaison RR et RR) il n'y a aucun gène B qui peut apparaître et les enfants seront également rouges. Si les parents sont tous les deux roses il y a quatre possibilités :

1. Les parents sont RB et RB ; alors l'enfant est RR .
2. Les parents sont RB et BR ; alors l'enfant est RB .
3. Les parents sont BR et RB ; alors l'enfant est BR .
4. Les parents sont BR et BR ; alors l'enfant est BB .

On voit que la combinaison fille a été obtenue en retenant la première lettre de chacune des deux combinaisons parentales, mais bien entendu ceci n'est qu'une convention d'écriture ; les parents ne "sont" pas RB ou BR ; ils sont hétérozygotes, et ce sont les hasards de la recombinaison qui décident ce qui sera — dans cette convention — la première lettre. Le mécanisme moléculaire réel de ce processus a été élucidé au milieu du XX^e siècle, c'est la division des chromosomes. La reproduction sexuée fonctionne de la manière suivante. Tous les gènes sont regroupés en chaînes ordonnées appelées chromosomes, qui sont toujours couplés par paires homologues : ces paires sont des chaînes doubles comportant deux chaînons symétriques qui se font face (les chromosomes homologues), de telle sorte qu'à chaque gène de l'un des chaînons correspond un compagnon symétrique sur l'autre chaînon appelé allèle. Les combinaisons de deux gènes qui interviennent dans la théorie de Mendel concernent ces couples d'allèles. Lors de la reproduction les deux chromosomes homologues se séparent, et chaque chromosome ainsi isolé se recombine *au hasard* avec un chromosome isolé de l'autre parent. Chaque parent fournit un seul des deux chromosomes de chaque paire, le choix étant fait au hasard. Ainsi une nouvelle paire sera formée à partir d'un chromosome du père et d'un chromosome de la mère. Il y a donc quatre recombinaisons équiprobables possibles.

L'ensemble des gènes possibles, ainsi que leur place sur les chaînons, est une caractéristique invariable de l'espèce. C'est pourquoi à un endroit donné du chromosome (on appelle cela un *locus*) on trouvera toujours les

mêmes gènes (les allèles du locus) : les gènes pouvant occuper un locus donné sont généralement peu nombreux. Dans les cas les plus simples, il n'y a qu'un seul allèle sur le locus (tout le monde est alors homozygote) ou deux, par exemple R et B , et alors les quatre combinaisons RR , RB , BR , et BB sont possibles. Mais bien entendu ce n'est pas la règle générale : il y a souvent plus que deux allèles. Dans une espèce donnée, on trouvera presque toujours l'une des combinaisons possibles d'allèles au locus correspondant (les exceptions sont les *mutations*) ; mais en un autre locus, on trouvera des combinaisons d'allèles différents. Si les deux parents sont hétérozygotes et s'il n'y a que deux allèles (cas où on croise deux fleurs roses), les quatre recombinaisons possibles de chromosomes donneront *au locus considéré* les quatre combinaisons BB , BR , RB , et RR .

Dans une population naturelle, il n'y a aucune raison que les allèles soient tous exactement aussi répandus : généralement, les uns sont plus rares que les autres (par exemple les mirabilis à fleurs rouges sont plus rares que ceux à fleurs blanches). Toutefois, dans les expériences comme celle décrite plus haut, on a préalablement sélectionné la variété rouge et la variété blanche, de sorte qu'en croisant les fleurs rouges avec les blanches on a créé artificiellement une population comportant exactement autant de gènes R que de gènes B . La règle statistique observée par Mendel s'explique aisément par l'équiprobabilité des quatre recombinaisons de chromosomes : au locus considéré, si les deux parents sont hétérozygotes, on obtiendra (avec probabilité $\frac{1}{4}$ pour chacune) les combinaisons RR , RB , BR , et BB . On appelle *génotype* la combinaison de gènes, le phénotype étant le caractère observable. La couleur rose de la fleur est un phénotype qui correspond indistinctement à l'un ou l'autre des génotypes RB ou BR , donc sa probabilité sera $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Le phénotype rouge correspond au génotype RR . Mais une telle correspondance entre génotype et phénotype est exceptionnelle ; l'immense majorité des génotypes n'ont aucun effet ponctuellement observable sous forme de phénotype ; non que le génotype soit sans effet, mais cet effet est tellement dilué qu'il ne lui correspond aucun caractère manifeste.

Étant donné l'équiprobabilité de ces combinaisons, on peut étendre les raisonnements probabilistes à des situations plus générales, où il n'y a pas exactement autant de gènes R que de gènes B dans la population, ce qui conduit à la loi de Hardy-Weinberg. Pour établir la loi de Mendel on n'a croisé que des fleurs roses, parmi lesquelles il y a autant de BR que de RB . On retrouve alors forcément autant de gènes B que de gènes R dans la population fille. Mais si on croise des fleurs au hasard, à partir d'une population naturelle, sans tenir compte de la couleur, et qu'il y a moins de rouges que de blanches, on n'aura plus cette symétrie.

III. 6. Loi de Hardy-Weinberg.

On établit la loi de Hardy-Weinberg par le Calcul des probabilités; c'est ce qu'a fait le mathématicien Hardy (1900) dont le nom est associé. Prenons par exemple le cas de la mucoviscidose (chez l'homme). Il s'agit d'une maladie génétique, c'est-à-dire due à la présence d'un gène spécifique dans le code génétique de l'individu qui en est atteint. Cette maladie est fortement invalidante: l'un des symptômes, particulièrement pénible, est l'obstruction permanente des bronches par du mucus, entraînant difficultés respiratoires et toux permanentes. Il en va de même dans les canaux excréteurs du pancréas, ce qui bloque le passage de l'insuline. La cause en est la viscosité excessive des sécrétions muqueuses, d'où le nom de la maladie. Parmi les individus possédant le gène responsable, seuls ceux qui portent la combinaison homozygote sont frappés par la maladie; ceux qui possèdent le gène sous la forme hétérozygote sont sains et de ce fait appelés *porteurs sains*.

Appelons M le gène fatidique. Celui-ci peut se trouver combiné à son allèle X (cas hétérozygote XM ou MX), ou à lui-même (cas homozygote MM). Il importe peu ici de savoir s'il n'y a que deux allèles (X représentant alors un gène unique) ou plus (X représentant alors n'importe lequel des autres allèles), puisque les gènes autres que M n'interviennent pas dans la mucoviscidose. La fréquence de la maladie dans la population européenne est d'environ $1/2500$ (un enfant sur 2500 qui naissent en est atteint). Comme pour les croisements entre deux fleurs rouges de mirabilis, deux parents atteints tous les deux de mucoviscidose auraient forcément des enfants également atteints, mais ce cas est évidemment rarissime et peut être négligé. Le cas où l'un des deux parents serait atteint, quoique moins improbable, peut aussi être négligé, d'autant plus que la maladie détourne du mariage. Le cas normal est celui de parents qui sont tous deux hétérozygotes.

Supposons que tous les parents sains s'accouplent au hasard; cette hypothèse signifie concrètement que les parents ignorent s'ils sont porteurs ou non, ou que s'ils le savent cela n'a aucune incidence sur leur fécondité. Cette hypothèse est appelée la *panmixie*; le terme anglo-saxon *random mating* est cependant plus courant. Parmi ces parents "pris au hasard", il y a autant de pères que de mères; soit donc N le nombre de pères (ou de mères). Si parmi les N pères, il y a k porteurs sains, la proportion de porteurs sains est $x = k/N$. Le gène M étant distribué indépendamment du sexe, on aura la même proportion chez les mères, donc k mères hétérozygotes (environ, car la proportion présente toujours des fluctuations statistiques). La proportion x est l'inconnue que justement nous désirons calculer à partir de la fréquence

connue de la maladie. Dans la loi de Mendel le hasard intervenait dans la recombinaison des chromosomes. Ici le hasard intervient en outre dans le choix des parents possibles : on admet que dans la population, *tous les couples de parents sont équiprobables*. Le nombre de tous les couples possibles est $N \times N$. Parmi ces couples, ceux dont les membres sont tous deux hétérozygotes MX sont au nombre de $k \times k$. La proportion de couples dont les membres sont tous deux porteurs sains est donc $k^2/n^2 = x^2$. Mais nous devons tenir compte des deux interventions du hasard ; l'espace des épreuves correspondant n'est pas l'ensemble des couples de parents, mais l'ensemble des combinaisons de chromosomes résultant de la division des paires parentales : chaque parent fournit un chromosome de la paire, donc chaque couple peut produire quatre combinaisons de chromosomes. Par conséquent l'espace Ω des épreuves est de cardinal $4N^2$ (quatre combinaisons de chromosomes pour chaque couple de parents). Si les parents ne sont pas tous les deux porteurs du gène M , *aucune* combinaison ne pourra donner le génotype MM ; mais si les parents sont tous les deux porteurs sains, une seule des quatre combinaisons qu'ils peuvent produire donnera le génotype MM : par conséquent le nombre de possibilités d'avoir le génotype MM est k^2 (une seule combinaison pour chaque couple d'hétérozygotes, plus zéro combinaisons pour les autres couples). La probabilité d'avoir un enfant de génotype MM sera donc $\frac{k^2}{4N^2} = \frac{x^2}{4}$.

Or on sait par les données cliniques que cette proportion est $1/2500$, c'est-à-dire que $\frac{x^2}{4} = \frac{1}{2500}$, d'où $x = 1/25 = 0.04$. On peut donc déduire de la proportion $1/2500$ de cas homozygotes que la proportion de cas hétérozygotes est $1/25$. Un individu sur vingt cinq est porteur sain.

Il est aisé de trouver la formule générale : si pour un gène G quelconque ε est la proportion d'homozygotes GG , la proportion d'hétérozygotes GX ou XG est $x = 2\sqrt{\varepsilon}$, car on doit avoir $\frac{x^2}{4} = \varepsilon$. Cela résulte directement du dénombrement : il y a N pères et N mères donc N^2 couples possibles, qui peuvent donner chacun quatre combinaisons possibles de chromosomes, donc $\#\Omega = 4N^2$; parmi ceux-ci il y a k^2 couples qui peuvent donner chacun une combinaison GG , donc l'événement A : "l'enfant est de génotype GG " a pour cardinal k^2 . La probabilité pour que tous ces couples formés au hasard engendrent un enfant GG est donc

$$\varepsilon = \frac{\#A}{\#\Omega} = \frac{k^2}{4N^2} = \frac{x^2}{4} \quad (III.4.)$$

Cette formule extrêmement simple permet de calculer la fréquence ε de la maladie à partir de la proportion statistique x de porteurs sains. On en déduit que si par exemple on dissuade la moitié des porteurs sains de procréer, on divise par quatre le nombre de cas de mucoviscidose.

Inversement, on obtiendra la proportion inconnue x de porteurs sains à partir de la fréquence ε de la maladie ($x = 2\sqrt{\varepsilon}$). Rappelons encore qu'il s'agit de la fréquence dans une population où les parents se choisissent "au hasard". Il est bien évident qu'une personne qui parmi ses proches (frères ou cousins) compte déjà un cas de mucoviscidose a une probabilité bien plus forte que x d'être porteur sain. Deux parents qui ont déjà mis au monde un enfant atteint sont assurément tous les deux porteurs sains et ont donc une chance sur quatre que le prochain enfant ait aussi la maladie.

La fréquence moyenne ε est variable selon les populations: 1/2500 était sa valeur sur l'ensemble de l'Europe occidentale. Mais elle diffère d'un pays à l'autre; à l'intérieur d'un même pays elle varie selon les régions. En outre, la mucoviscidose est beaucoup plus rare en Asie, par exemple. Le tableau suivant⁽³⁾ en montre quelques exemples:

France:		global	1/2 000
		Finistère Nord	1/1 650
		Morbihan	1/3 500
Royaume Uni:		Angleterre	1/2 350
		Pays de Galles	1/1 650
Italie:		global	1/2 700
Suède:		global	1/5 700
Hawaii:	population de souche européenne		1/3 800
	population de souche polynésienne		1/90 000

On déduit alors immédiatement la proportion de porteurs sains dans ces populations: France 1/22, Suède 1/34, hawaïens de souche polynésienne 1/150, etc. Bien entendu le raisonnement suivi pour obtenir la formule *III.4* suppose une population génétiquement stable et isolée (endogamique), ainsi que l'équiprobabilité des choix de parents (random mating). Ces hypothèses ne sont que très approximativement vérifiées dans les populations humaines réelles. Mais il est possible de les réaliser en laboratoire sur des cultures de fleurs ou de bactéries. En observant la distribution statistique de phénotypes dans de telles cultures artificielles, et en la comparant à des probabilités a priori calculées à partir de *modèles de génotypes*, il devient possible d'étudier scientifiquement l'influence des gènes sur les phénotypes. Cette méthode est à la base de la génétique.

Dans le raisonnement suivi plus haut, nous avons cependant négligé que la naissance d'un GG peut aussi provenir du croisement d'un GX avec un GG , ou du croisement de deux GG . Cette négligence volontaire se justifie,

⁽³⁾ D'après G. Lenoir *La mucoviscidose* Éd Doin.

soit parce que le génotype GG est très invalidant (donc rend l'accès à la procréation difficile ou impossible), soit tout simplement parce que le génotype GG est si rare qu'on peut statistiquement le négliger. Ces deux conditions étaient vérifiées pour la mucoviscidose.

Si le génotype GG n'est ni rare ni invalidant, il faut tenir compte des mariages $GX + GX$ ou $GG + GG$. Supposons toujours que les couples de parents sont choisis au hasard. L'événement A : "naissance d'un GG " est alors la réunion des quatre événements suivants :

$$\begin{aligned} A_1: & GX + GX \rightarrow GG \quad (\text{probabilité } \frac{x^2}{4}); \\ A_2: & GG + GX \rightarrow GG \quad (\text{probabilité } \frac{\varepsilon x}{2}); \\ A_3: & GX + GG \rightarrow GG \quad (\text{probabilité } \frac{x\varepsilon}{2}); \\ A_4: & GG + GG \rightarrow GG \quad (\text{probabilité } \varepsilon^2); \end{aligned}$$

Les probabilités de A_2 , A_3 , et A_4 se calculent par dénombrement exactement de la même façon que pour A_1 . Ces quatre événements étant disjoints, on obtient donc

$$\frac{x^2}{4} + \frac{\varepsilon x}{2} + \frac{\varepsilon x}{2} + \varepsilon^2 = \varepsilon$$

de sorte que x est solution de l'équation du second degré $\frac{x^2}{4} + x\varepsilon + \varepsilon^2 - \varepsilon = 0$, dont la seule solution positive est

$$x = 2(\sqrt{\varepsilon} - \varepsilon) \quad (III.5.)$$

Cette relation est la loi de Hardy-Weinberg. Le raisonnement simplifié précédent donnait $x = 2\sqrt{\varepsilon}$, ce qui est à peu près la même chose si ε est petit (nous avons passé sous silence les événements A_2 , A_3 , et A_4 , qui pour ε petit sont en effet beaucoup moins probables que A_1).

La loi de Hardy-Weinberg n'est évidemment applicable que si les hypothèses d'équiprobabilité sont satisfaites. Pour dénombrer les quatre événements A_1 , A_2 , A_3 , et A_4 , nous avons admis que les gènes se combinaient uniformément. Cela suppose que les parents qui procréent se choisissent "au hasard", et qu'en outre, la recombinaison des chromosomes est parfaitement uniforme. Il est bien clair que les couples ne se forment jamais au hasard, mais le "hasard pur" n'est exigé que pour ce qui concerne les combinaisons des gènes X et G . Si les phénotypes correspondant à ces combinaisons n'ont aucun effet susceptible d'augmenter ou de diminuer l'attirance sexuelle, la fécondité, etc. tout se passera selon le hasard pur, même si on peut trouver des déterminismes sociaux ou psychologiques à la formation des couples. Il est cependant assez évident que si la combinaison homozygote GG produit des symptômes invalidants, ceux-ci auront une réelle influence (négative)

sur la fécondité, ou aboutiront à la mort de l'individu avant la puberté. Seuls les individus porteurs de XX ou de GX seront dépourvus du caractère invalidant et s'accoupleront au hasard. Cela exclut les événements A_2 , A_3 , et A_4 , et dans ce cas notre raisonnement approché devient correct même si ε n'est pas petit. Toutefois supposer ε grand signifie que le caractère invalidant est fréquent, et aucune espèce ayant subi la lutte pour la vie pendant des millénaires ne peut correspondre à une telle hypothèse. D'autre part on peut critiquer l'application de ces hypothèses à l'homme civilisé : celui-ci n'est pas soumis à la lutte pour la vie ; des caractères invalidants peuvent être suffisamment atténués par la médecine pour ne plus détourner de la procréation ; un caractère qui augmente la fécondité peut être compensé par l'usage de contraceptifs, un caractère qui la diminue peut être combattu par des traitements hormonaux, etc.

Le Calcul des probabilités joue un rôle essentiel en génétique ; les méthodes que nous avons mises en oeuvre pour aboutir aux lois de Mendel et de Hardy-Weinberg peuvent être généralisées à des situations moins simples. Pour étudier en laboratoire les mécanismes moléculaires de l'hérédité on analyse les protéines : celles-ci sont formées d'acides aminés qu'on peut isoler par des méthodes adéquates (chromatographie, etc.) Mais la biologie moléculaire ne permet pas de connaître la répartition des gènes dans une population, ni l'influence d'un gène sur le phénotype d'une plante. C'est pourquoi les généticiens cultivent en laboratoire des populations entièrement artificielles (comme le fit Mendel), dans lesquelles l'observation empirique permet de mesurer la répartition statistique de certains phénotypes. L'équiprobabilité des recombinaisons chromosomiques et des croisements de semences permet d'autre part de calculer a priori des probabilités de répartition de génotypes : toute hypothèse sur un génotype peut ainsi être confrontée à l'observation statistique.

En ce qui concerne la méthodologie, le point essentiel est le caractère quantitatif des prédictions : toute hypothèse sur un génotype peut être testée quantitativement à l'aide d'observations statistiques sur les phénotypes correspondants. Cette méthode rigoureuse, initiée par Gregor Mendel dans le cas le plus simple où le phénotype étudié est déterminé par deux allèles seulement, et où le Calcul des probabilités joue un rôle clé, est donc le fondement de la génétique. Pour que les expériences soient statistiquement précises, il faut disposer en laboratoire de populations nombreuses et pouvoir les croiser sur un grand nombre de générations ; donc les espèces étudiées doivent être de petite taille et se reproduire vite ; d'où la préférence pour des espèces particulières ayant ces propriétés : bactérie *Escherichia coli*, mouche drosophile, fleur *Arabidopsis thaliana* (arabette des dames).