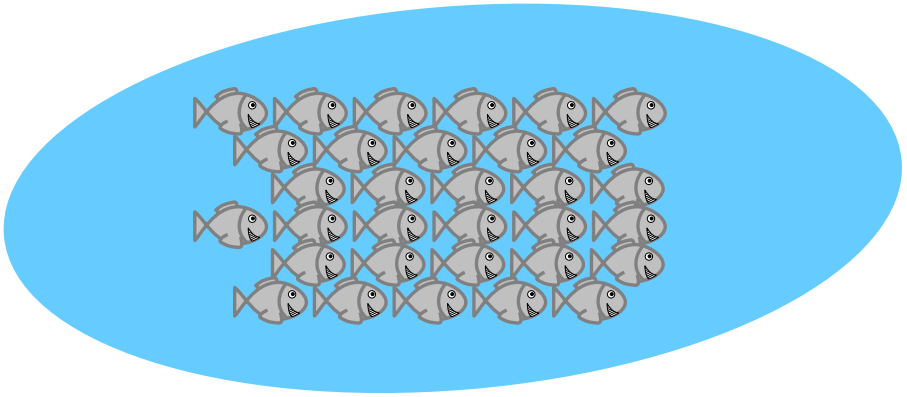
♦ Thème 3.1. La biodiversité et son évolution

I. — Estimation d'une abondance par la méthode CMR

Dans un milieu donné, on considère une population d'animaux d'une certaine espèce. On souhaite déterminer combien d'animaux de cette espèce sont présents dans ce milieu. C'est ce qu'on appelle l'**abondance** de cette espèce dans le milieu considéré.

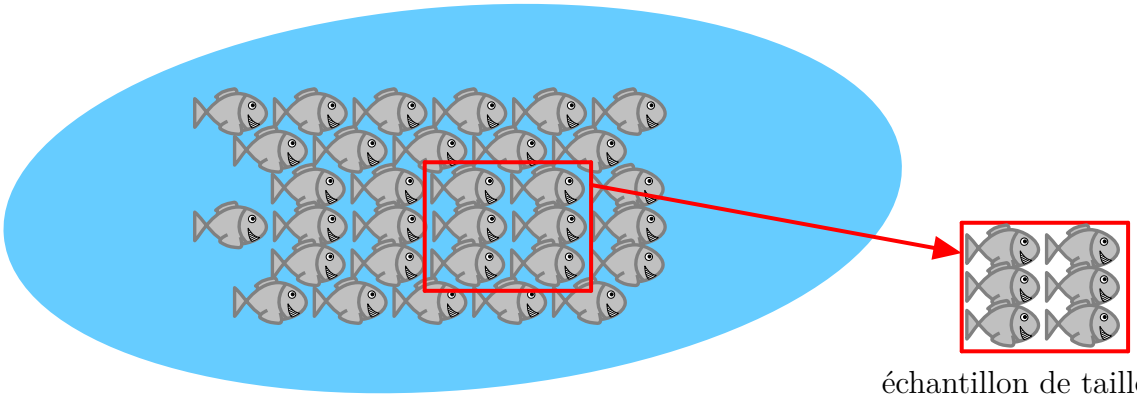
Par exemple, on souhaite estimer l'abondance de truites dans un lac donné c'est-à-dire le nombre N de truites qui vivent dans ce lac.



Il est impossible de toutes les compter une à une. On peut utiliser une méthode d'estimation appelée méthode « Capture-Marquage-Recapture » (en abrégé CMR).

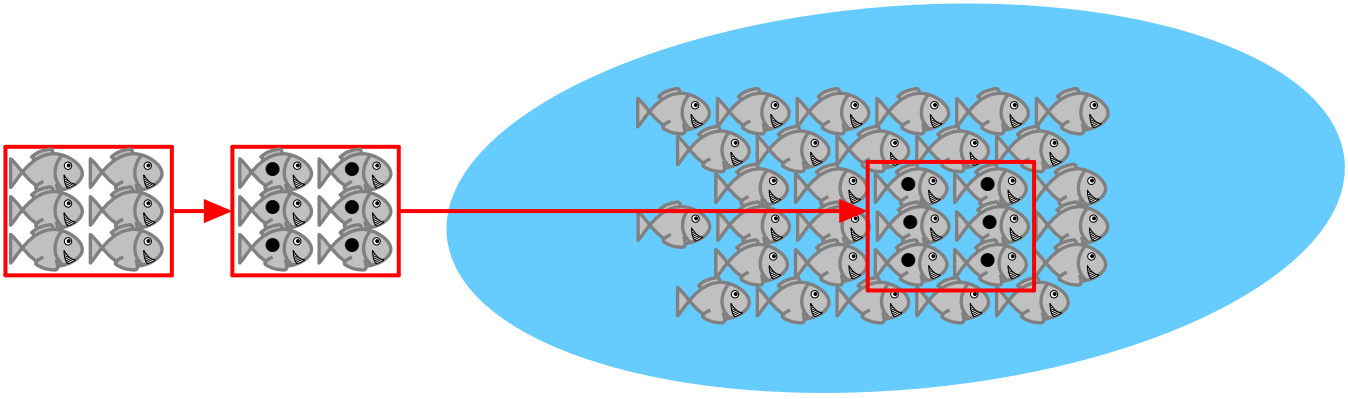
Son principe est le suivant.

1ère étape : Capture. On capture un certain nombre M de truites. C'est ce qu'on appelle un échantillon et M est la taille de cet échantillon.

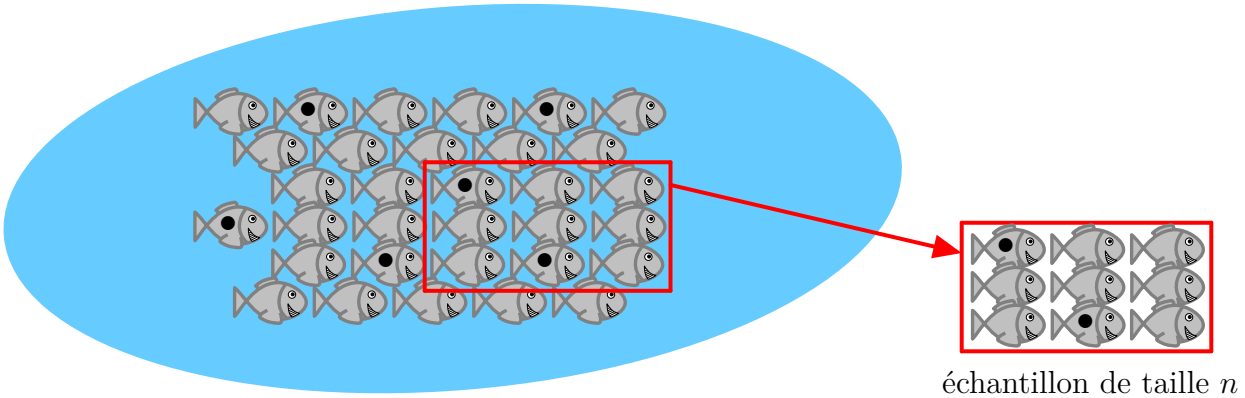


échantillon de taille M

2ième étape : Marquage. On marque les truites capturées (à l'aide d'une bague, d'une marque de couleur, d'un transpondeur...) puis on relâche les truites dans le lac.



3ième étape : Recapture. Après un certain temps, on effectue une nouvelle capture dans le lac d'un échantillon de taille n quelconque et on compte le nombre m de truites marquées dans ce nouvel échantillon.



échantillon de taille n

En faisant l'hypothèse que la proportion de truites marquées est la même dans le nouvel échantillon que dans la population totale, on a l'égalité

$$\frac{M}{N} = \frac{m}{n}.$$

On en déduit que

$$N = \frac{n \times M}{m}.$$

Imaginons, par exemple, qu'on capture 150 truites pour le premier échantillon et qu'on capture 50 truites pour le second échantillon dont 8 sont marquées. On peut estimer le nombre total de truites dans le lac à

$$N = \frac{50 \times 150}{8} \approx 938.$$

Conditions d'application et limites de la méthode

Pour que la méthode soit efficace, il est nécessaire que :

1. la population étudiée n'évolue pas (ou peu) entre les deux captures donc il ne faut pas que des individus puissent la quitter ou y entrer, par exemple à l'occasion de flux migratoires. De même, la temps écoulé entre le marquage et la recapture doit être assez court pour éviter les naissances et les décès mais suffisamment important pour assurer un brassage uniforme des individus marqués dans l'ensemble de la population.
2. les animaux marqués ne soient pas affectés par le marquage (que ce soit dans leur comportement ou leur espérance de vie) et les marques ne soient pas perdues. Ainsi, les deux premières étapes (capture et marquage) doivent être aussi courtes que possibles et il faut utiliser soit des marques indélébiles soit un double marquage ;
3. la probabilité de capturer un animal marqué doit être la même que n'importe quel animal de la population.

On peut mettre en évidence certaines limites de la méthode :

1. il n'est pas toujours facile de capturer suffisamment d'individus pour obtenir un résultat satisfaisant et certaines catégories, comme les jeunes, sont souvent plus difficiles à capturer donc sous-représentés dans les échantillons ;
2. le marquage peut affecter le comportement des animaux, les rendre vulnérables face aux prédateurs ou affecter leur place hiérarchique dans leur société ;
3. un animal capturé lors de la constitution du premier échantillon peut se montrer par la suite plus méfiant et avoir une probabilité plus faible d'être recapturé qu'un individu non marqué.

II. — Fluctuation d'échantillonnage et intervalle de confiance

1) Fluctuation d'échantillonnage

Dans la méthode CMR, on fait l'hypothèse que la proportion $\frac{m}{n}$ d'animaux marqués dans le second échantillon est égale à (ou en tout cas proche de) la proportion $\frac{M}{N}$ d'animaux marqués dans l'ensemble de la population. La proportion $\frac{m}{n}$ s'appelle la **fréquence observée** d'animaux marqués dans l'échantillon.

Dans les faits, cette fréquence dépend de l'échantillon. Voici, par exemple, les résultats obtenus lors de 10 recaptures de 50 truites après avoir marquées 150 truites lors de la première capture.

capture	1	2	3	4	5	6	7	8	9	10
Nombre de truites marquées dans l'échantillon	8	10	7	5	6	4	10	4	3	7
Proportion de truites marquées dans l'échantillon	0,16	0,2	0,14	0,1	0,12	0,08	0,2	0,08	0,06	0,14
Estimation de l'abondance	937	750	1071	1500	1250	1875	750	1875	2500	1071

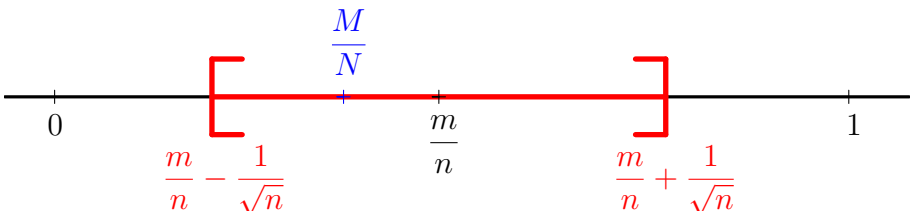
Cette variation est ce qu'on appelle la **fluctuation d'échantillonnage**.

Une question naturelle est alors la suivante : puisque $\frac{m}{n}$ varie, cette proportion est-elle toujours une bonne approximation de $\frac{M}{N}$?

La réponse est non. On peut très bien, par exemple, avoir un échantillon qui ne contient pas d'animaux marqués, ce qui impliquera que $\frac{m}{n} = 0$ alors que $\frac{M}{N} \neq 0$. Cependant, on peut montrer que, dans la plupart des cas, $\frac{m}{n}$ est bien une bonne approximation de $\frac{M}{N}$ et qu'elle est d'autant meilleure que n est grand.

2) Intervalle de confiance

Plus précisément, on peut montrer que, dans 95% des cas, la **marge d'erreur** c'est-à-dire l'écart entre la fréquence observée $\frac{m}{n}$ et la proportion $\frac{M}{N}$ dans la population totale est inférieure à $\frac{1}{\sqrt{n}}$. Autrement dit, pour 95% des échantillons, la proportion $\frac{M}{N}$ appartient à l'intervalle $\left[\frac{m}{n} - \frac{1}{\sqrt{n}} ; \frac{m}{n} + \frac{1}{\sqrt{n}}\right]$.



On dit que l'intervalle

$$\left[\frac{m}{n} - \frac{1}{\sqrt{n}} ; \frac{m}{n} + \frac{1}{\sqrt{n}}\right]$$

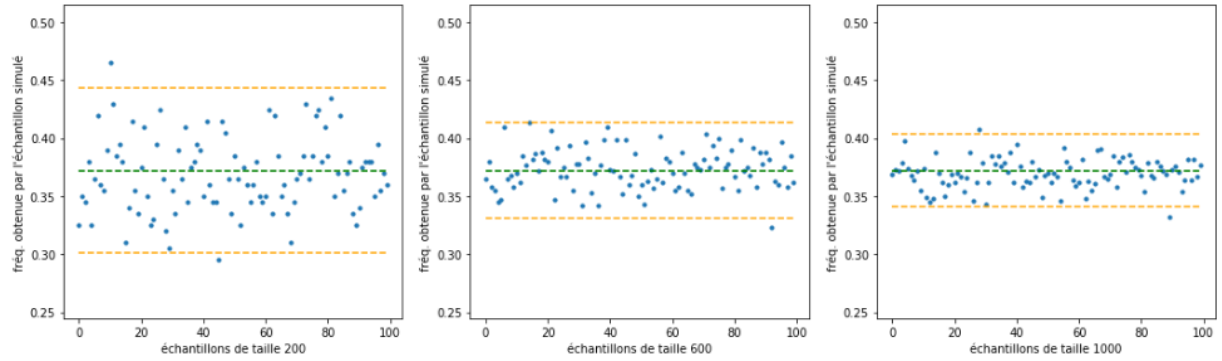
est **un intervalle de confiance** au niveau de confiance 95% de la proportion $\frac{M}{N}$.

Si on reprend l'exemple précédent où $M = 150$, $n = 50$ et $m = 8$ alors on peut affirmer, avec un niveau de confiance 95% que la proportion $\frac{M}{N}$ d'animaux marqués dans la population totale est comprise entre $\frac{8}{50} - \frac{1}{\sqrt{50}} \approx 0{,}018$ et $\frac{8}{50} + \frac{1}{\sqrt{50}} \approx 0{,}302$. On en déduit que N est compris entre $\frac{150}{0{,}302} \approx 497$ et $\frac{150}{0{,}018} \approx 8333$.

On voit qu'on obtient une estimation très imprécise. Cela vient du fait que la taille de l'échantillon ($n = 50$) est trop petite.

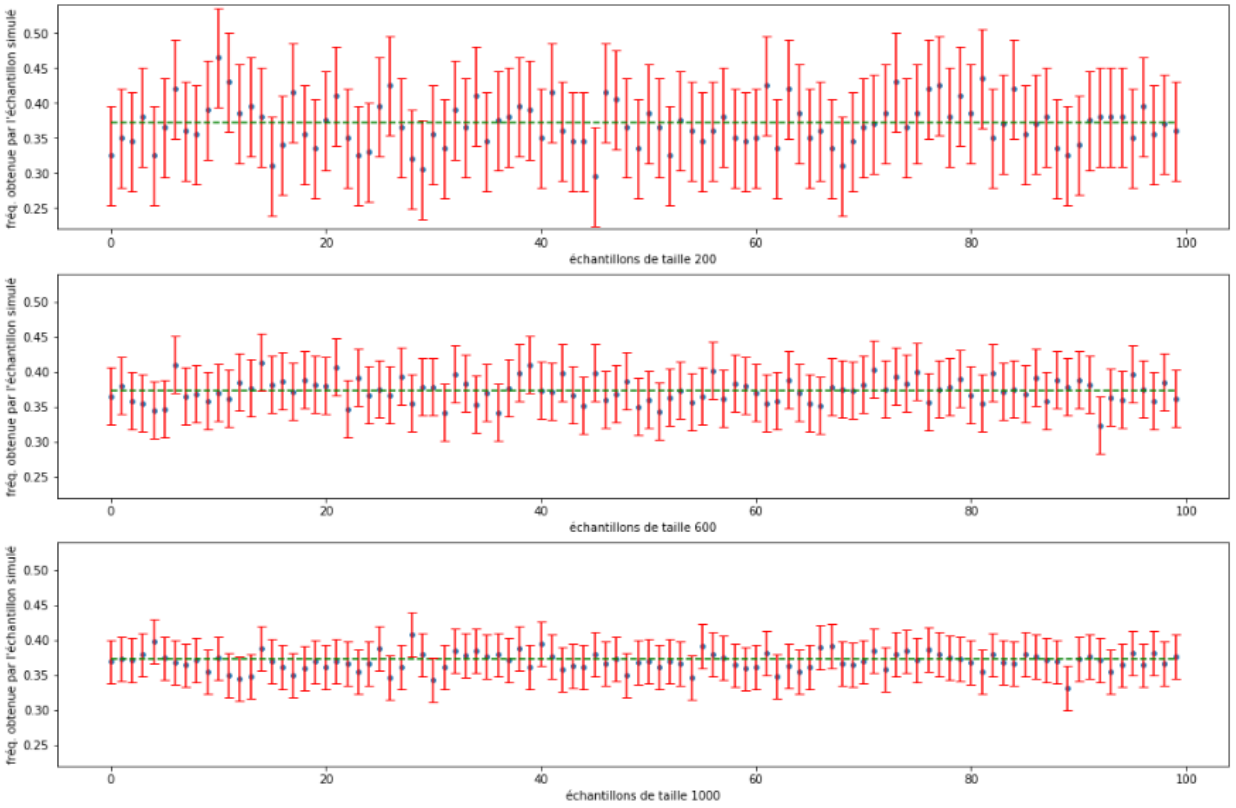
L'amplitude de l'intervalle de confiance, c'est-à-dire l'écart entre les deux bornes de l'intervalle est $\frac{2}{\sqrt{n}}$. Cette amplitude diminue lorsque n augmente. Autrement dit, plus on prend un échantillon de taille n grande, plus on aura un encadrement précis. Remarquons cependant qu'augmenter la valeur de n ne change pas le niveau de confiance qui reste égal à 95%.

Sur les graphiques ci-dessous, on a représenté des simulations de 100 recaptures différentes pour des échantillons de tailles $n = 200$, $n = 600$ et $n = 1000$. Les pointilles centraux représentent la proportion réelle de $\frac{M}{N}$ et les pointillés supérieurs et inférieurs délimitent un écart de plus ou moins $\frac{1}{\sqrt{n}}$ par rapport à $\frac{M}{N}$.



On constate que plus n augmente, plus les valeurs obtenues se resserrent autour de $\frac{M}{N}$ mais qu'il y a toujours des valeurs qui s'écartent de plus de $\frac{1}{\sqrt{n}}$.

Sur les graphiques ci-dessous, on a représenté les mêmes situations mais en traçant les intervalles de confiance autour de chaque proportion.



On constate que plus n augmente plus la taille des intervalles diminue mais qu'il y a toujours des intervalles qui ne contiennent pas la proportion $\frac{M}{N}$ (symbolisée par la ligne en pointillés).

Dans 95%, la proportion $\frac{M}{N}$ appartient à l'intervalle $\left[\frac{m}{n} - \frac{1}{\sqrt{n}} ; \frac{m}{n} + \frac{1}{\sqrt{n}}\right]$ donc la marge d'erreur est inférieur à $\frac{1}{\sqrt{n}}$. Ainsi, pour obtenir une approximation avec une marge d'erreur inférieur à une valeur E, il suffit de choisir n de telle sorte que $\frac{1}{\sqrt{n}} \leqslant E$ c'est-à-dire $n \geqslant \frac{1}{E^2}$.

Par exemple, pour avoir une estimation avec une marge d'erreur inférieure ou égale à 1%, il faut prendre un échantillon de taille au moins $\frac{1}{0{,}01^2} = 10000$.

III. — Modèle de Hardy-Weinberg

1) Fréquences génotypiques et fréquences alléliques

Dans une population, on considère un caractère déterminé par un gène présent sur deux chromosomes homologues c_1 et c_2. On suppose que ce gène peut prendre deux formes différentes appelées **allèles** : l'allèle A et l'allèle a.

Le **génotype** d'un individu est la composition allélique c'est-à-dire la donnée des deux allèles présents sur ses gènes. Dans notre cas, on a les génotypes ci-contre.

c_1 \ c_2	A	a
A	AA	Aa
a	aA	aa

Ainsi, il y a 3 génotypes différents : 2 allèles A (AA), 1 allèle A et 1 allèle a (Aa ou aA) et 2 allèles a (aa). Les individus ayant un génotype AA ou aa sont dit homozygotes et les autres, ayant un génotype Aa, sont appelés hétérozygotes.

Dans une population donnée, on peut définir deux types de fréquences :
les fréquences génotypiques : la fréquence d'un génotype est la proportion d'individus possédant ce génotype ;
les fréquences alléliques : la fréquence d'un allèle est la proportion de gènes portant cet allèle.

Par exemple, supposons que dans une population de 1200 individus, il y en ait 350 qui possèdent le génotype AA, 560 le génotype Aa et 290 le génotype aa.
Alors les fréquences génotypiques sont :

$$f(AA) = \frac{350}{1200} = \frac{7}{24} \qquad f(Aa) = \frac{560}{1200} = \frac{7}{15} \qquad f(aa) = \frac{290}{1200} = \frac{29}{120}.$$

Pour ce qui est des fréquences alléliques, on remarque que chaque individu possède deux gènes donc le nombre total de gènes est $2 \times 1200 = 2400$. Ensuite, chaque individu du génotype AA possède deux allèles A et chaque individu du génotype Aa possèdent un allèle A donc la fréquence de l'allèle A est

$$f(A) = \frac{2 \times 350 + 560}{2400} = \frac{21}{40}.$$

De même, chaque individu du génotype aa possède deux allèles a et chaque individu du génotype Aa possèdent un allèle a donc la fréquence de l'allèle a est

$$f(a) = \frac{2 \times 290 + 560}{2400} = \frac{19}{40}.$$

Remarquons que, comme il n'y a que 3 génotypes possibles, $f(AA) + f(Aa) + f(aa) = 1$ et, comme il n'y a que deux allèles possibles, $f(A) + f(a) = 1$.

Considérons une population de taille N. Notons p la fréquence de AA, q la fréquence de Aa et r la fréquence de aa.
Alors, le nombre d'individus ayant un génotype AA est $p \times N$, le nombre d'individus ayant un génotype Aa est $q \times N$ et le nombre d'individus ayant un génotype aa est $r \times N$.
Comme chaque individu possède deux gènes, le nombre total de gènes est $2N$. Chaque individu ayant le génotype AA apporte 2 allèles A et chaque individu ayant un génotype Aa apporte 1 allèle A donc le nombre total d'allèles A dans la population est $2 \times p \times N + 1 \times q \times N$. La fréquence allélique de A est donc

$$f(A) = \frac{2 \times p \times N + q \times N}{2 \times N} = \frac{(2p+q)N}{2N} = \frac{2p+q}{2} = \frac{2p}{2} + \frac{q}{2} = p + \frac{1}{2}q.$$

De la même manière, la fréquence allélique de a est

$$f(a) = \frac{2 \times r \times N + q \times N}{2 \times N} = \frac{(2r+q)N}{2N} = \frac{2r+q}{2} = \frac{2r}{2} + \frac{q}{2} = r + \frac{1}{2}q.$$

2) Évolution sur plusieurs générations

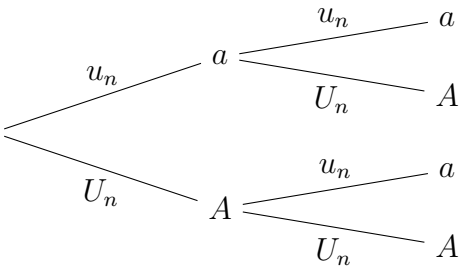
On s'intéresse maintenant à l'évolution des fréquences génotypiques sur plusieurs générations.

On numérote les générations en considérant que la génération initiale est la génération 0. On note, pour tout entier n :

- p_n la fréquence de AA
- q_n la fréquence de Aa
- r_n la fréquence de aa
- U_n la fréquence de A
- u_n la fréquence de a

à la génération n. Ainsi, par exemple, q_0 est la fréquence du génotype Aa à la génération initiale et U_3 est la fréquence de l'allèle A à la génération 3.

Lors d'une reproduction sexuée, chaque parent va transmettre au descendant un de ses deux allèles. En supposant que la population est suffisamment grande, que les couples se forment au hasard et que la transmission des allèles se fait également au hasard, on peut modéliser la transmission des allèles d'une génération à l'autre comme le choix de deux allèles dans l'ensemble des allèles de la population. Autrement dit, on peut calculer les fréquences génotypiques de la génération $n+1$ à l'aide des fréquences alléliques de la génération n :



On en déduit que $p_{n+1} = U_n^2$, $q_{n+1} = u_n U_n + U_n u_n = 2u_n U_n$ et $r_{n+1} = u_n^2$. Or, d'après ce qui précède, $U_n = p_n + \frac{1}{2}q_n$ et $u_n = r_n + \frac{1}{2}q_n$ donc on a les relations suivantes entre les fréquences génotypiques de deux générations successives :

$$p_{n+1} = \left(p_n + \frac{1}{2}q_n\right)^2 \qquad q_{n+1} = 2\left(p_n + \frac{1}{2}q_n\right)\left(r_n + \frac{1}{2}q_n\right) \qquad r_{n+1} = \left(r_n + \frac{1}{2}q_n\right)^2.$$

Lorsqu'on étudie numériquement l'évolution de ces fréquences, on s'aperçoit que les fréquences génotypiques sont constantes à partir de la génération 1 et que les fréquences alléliques sont constantes dès la génération 0 et ceci quelles que soient les fréquences initiales.

	A	B	C	D	E	F
1	n	p_n	q_n	r_n	U_n	u_n
2	0	0,4	0,1	0,5	0,45	0,55
3	1	0,2025	0,495	0,3025	0,45	0,55
4	2	0,2025	0,495	0,3025	0,45	0,55
5	3	0,2025	0,495	0,3025	0,45	0,55
6	4	0,2025	0,495	0,3025	0,45	0,55
7	5	0,2025	0,495	0,3025	0,45	0,55
8	6	0,2025	0,495	0,3025	0,45	0,55
9	7	0,2025	0,495	0,3025	0,45	0,55

En effet, on a, pour tout entier naturel n,

$$U_{n+1} = p_{n+1} + \frac{1}{2}q_{n+1} = U_n^2 + \frac{1}{2} \times 2u_n U_n = U_n^2 + u_n U_n = U_n(U_n + u_n) = U_n$$

car $U_n + u_n = 1$. Dès lors, $u_{n+1} = 1 - U_{n+1} = 1 - U_n = u_n$.

Ainsi, les fréquences alléliques sont constantes dès la génération initiale.

De plus, pour tout entier naturel n, $p_{n+1} = U_n^2$ donc p_n est constante à partir de $n = 1$ et, de même pour q_n et r_n.

Ainsi, les fréquences atteignent un état d'équilibre appelé **équilibre de Hardy-Weinberg**. Ce phénomène a été découvert en 1908 de manière indépendante par le mathématicien anglais Godfrey Hardy et le médecin allemand Wilhelm Weinberg.

On a vu précédemment qu'il est facile d'exprimer les fréquences alléliques en fonction de fréquences génotypiques. Dans le cadre du modèle d'Hardy-Weinberg, il est également simple d'exprimer les fréquences génotypiques à l'aide des fréquences alléliques (au moins à partir de la deuxième génération) : on a en effet, dans ce cadre,

$$f(AA) = f(A)^2 \qquad f(Aa) = 2f(A)f(a) \qquad f(aa) = f(a)^2.$$

3) Écart par rapport au modèle

La validité du modèle de Hardy-Weinberg repose sur plusieurs hypothèses.

- La taille de la population doit être très grande.
- Le caractère étudié doit être lié à des cellules diploïdes c'est-à-dire des cellules contenant des paires de chromosomes homologues.
- La population étudiée doit être panmictique (c'est-à-dire dotée d'un système de reproduction sexuée où les fécondations se font au hasard). En particulier, les couples doivent se former au hasard (panmixie) et la transmission des allèles doit également se faire au hasard (pangamie).
- Il ne doit y avoir dans la population ni sélection, ni mutation, ni migration.
- Il ne doit pas y avoir de reproduction intergénérationnelle.
- Les génotypes sont tous viables et féconds.

Lorsqu'on constate des écarts entre les fréquences observées et l'équilibre prévu par le modèle d'Hardy-Weinberg, cela signifie qu'une ou plusieurs des conditions précédentes ne sont pas remplies. Cela est dû principalement à des **forces évolutives** qui vont faire évoluer les fréquences génotypiques. Les principales forces évolutives sont :

- **la dérive génétique** : de petits écarts initiaux dans la transmission peuvent entraîner une grande disparité dans la répartition des allèles après plusieurs générations voire la disparition pure et simple d'une allèle ; cette dérive a d'autant plus de risques de se produire que la taille population initiale est faible ;
- **la sélection naturelle** : les individus qui sont les plus aptes à survivre peuvent influencer les proportions d'allèles transmis à la génération suivante ;
- **la mutation génétique** : des mutations aléatoires d'allèles peuvent s'opérer et se transmettre à la génération suivante, modifiant ainsi les proportions de chaque allèle.

Par exemple, les îles Galapagos abritent treize espèces différentes de pinsons qui se différencient par la taille de leur corps, ainsi que par la forme et la taille de leur bec. Ces trois caractères, et notamment la taille du bec, se transmettent fortement d'une génération à la suivante.

Deux biologistes, Peter et Rosemary Grant, ont suivi l'évolution sur trente ans des populations de pinsons sur l'île de Daphne Major et ont pu détecter sur cette période des événements sélectifs importants.

Par exemple, la fin des années 1970 a été marquée par une sécheresse importante, entraînant une raréfaction des graines de petite taille et qui a coïncidé avec une sélection très forte des individus à gros bec chez l'espèce G. Fortis. Chez cette espèce, la consommation de petites graines est préférée quand celles-ci sont abondantes, alors qu'en cas de sécheresse la consommation de graines plus grosses, accessibles uniquement aux individus à gros bec, est favorisée. L'épisode de sécheresse, en causant la raréfaction des graines de petite taille, a entraîné une mortalité plus importante des individus à petit bec alors que les individus à gros bec ont survécu en plus grand nombre. Ces derniers se sont donc en moyenne plus reproduits ce qui a entraîné un déplacement de caractère à la génération suivante. On a donc ici un exemple de l'effet de la sélection naturelle sur la répartition des génotypes.