l’évolution de la biodiversité génétique : le principe de hardy-weinberg

publié le 17.12.21 |par karim zayana

l'improbable contribution à la biologie de godfrey harold hardy, spécialiste anglais de théorie des nombres de la première moitié du xx^{e} siècle.

ce texte revient sur le principe d'équilibre de hardy-weinberg [1] tel qu'il est énoncé dans le programme d'enseignement scientifique de terminale [2] : en l'absence de forces évolutives et pour peu que le hasard joue correctement son rôle, la structure génotypique d'une grande population de même espèce se stabilise dès la première descendance, c'est-à-dire à partir de la génération f_1 [1].

godfrey harold hardy : mathématicien britannique, 1877 — 1947. il s'est illustré en théorie des nombres et en analyse, entretenant une collaboration fructueuse avec ses collègues john edensor littlewood et edward maitland wright, ainsi qu'avec le mathématicien indien srinivasa ramanujan. c'est de manière tout à fait anecdotique, au cours d'un repas au trinity college de cambridge avec son ami et partenaire de cricket le généticien reginald crundal punnett, que le problème de la structure génotypique d'une population lui fut soumis. il résolut en quelques minutes la question de sa stabilisation théorique et, se pensant insuffisamment légitime dans un domaine aux frontières du sien, hésita même à publier ce résultat ! [3][4]

wilhelm weinberg : médecin allemand, 1862 — 1937. c'est parallèlement aux travaux de hardy qu'il postula lui aussi le principe d'équilibre auquel il donne également son nom.

ce principe tout théorique est rarement respecté dans la réalité, précisément parce que son cadre d'application est trop étroit. mais avant d'envisager ses limites, nous réaliserons la démonstration formelle de ce principe en l'accordant avec le langage des probabilités vu au lycée [5]. nous raisonnerons avec un nombre p d'allèles quelconque plutôt que de nous limiter au seul cas où $p=2$. plus réaliste, ce choix nous permettra, en outre, de suivre plus aisément le fil de l'exposé. interviennent en particulier la notion de probabilité conditionnelle[2], le concept d'indépendance et la formule des probabilités totales. nous verrons qu'une fois les conventions d'écriture, certes un peu fastidieuses, bien posées, la preuve coule de source. les hypothèses mathématiques nécessaires à sa mise en œuvre s'imposeront d'elles-mêmes. elles fixeront *de facto* les contours du modèle. assez nombreuses, nous les porterons *en italiques* à mesure que nous les rencontrerons.

1. prérequis

un gène est une unité d'information constituée par une séquence d'adn [6][7]. il est responsable d'un caractère héréditaire se traduisant par une spécificité anatomique ou physiologique chez une espèce donnée[3]. un même gène admet en général plusieurs variants, mettons p, comme autant de déclinaisons possibles d'une matrice commune g. appelons-les g_1, g_2,..., g_p et signalons dès maintenant qu'on pourrait

aussi les noter en lettres minuscules, la raison en sera clarifiée tout à l'heure.

par exemple, le gène g déterminant le groupe sanguin chez l'homme est porté par le chromosome 9 et peut apparaître sous trois formes : $g_1 = a$, $g_2 = b$ ou $g_3 = o$ (ce dernier figurant provisoirement en majuscule).

- le variant $g_1=a$ induit la synthèse d'une certaine molécule a sur la membrane des hématies (ou globules rouges), les cellules qui renferment l'hémoglobine ;
- le variant $g_2=b$ induit la synthèse d'une certaine molécule b sur la membrane des hématies ;
- le variant $g_3=o$ n'induit rien de son propre fait, mais n'inhibe rien non plus.

dans son patrimoine génétique, un être humain détient deux exemplaires[4] de chaque chromosome, dits homologues car portant les mêmes gènes. en particulier, il possède deux exemplaires du chromosome 9 : l'un provient de son père, l'autre de sa mère. de son père il a reçu une première version, ou premier allèle, du gène g : potentiellement a, b ou o. de sa mère, il a reçu une seconde version, ou second allèle, du gène g : potentiellement encore a, b ou o. soit une combinatoire de $\binom{3}{2}=6$ paires d'allèles, ou génotypes, possibles :

$$\{a,a\},\{b,b\},\{o,o\},\{a,b\},\{a,o\},\{b,o\}$$

nous avons noté les paires entre accolades comme c'est de circonstance en mathématiques. toutefois, en biologie, elles apparaîtraient différemment, par exemple $a\parallel b$, voire ab tout court en lieu de $\{a,b\}$. nous opterons pour $a\!\parallel\! b$, car les mots « parallélisme » et « allèle » ont une racine grecque commune et car le parallélisme est une relation symétrique tandis que l'autre écriture, ab, laisse penser que l'ordre d'exposition compte.

les individus $a\!\parallel\! a$, $b \!\parallel\! b$ et $o \!\parallel\! o$ sont homozygotes pour le gène g, à la différence des individus $a\!\parallel\! b$, $a\!\parallel\! o$ et $b\!\parallel\! o$, dits hétérozygotes. un individu homozygote, par exemple de type $a\!\parallel\! a$, ne produira que des cellules reproductrices, les gamètes (spermatozoïdes ou ovules), possédant l'allèle a. un individu hétérozygote, par exemple de type $a\!\parallel\! b$, produira *à parité* des gamètes portant l'allèle a ou b : *on stipule à cet endroit une hypothèse d'équiprobabilité* dite loi de ségrégation (ou de disjonction) de mendel, loi que nous évoquerons encore un peu plus loin.

un individu possédant au moins un allèle a (respectivement b) produit à la surface de ses globules rouges des molécules a (respectivement b), susceptibles, en cas de transfusion, de déclencher une réaction immunitaire chez un individu ne possédant pas l'allèle a (respectivement b). par contre, la molécule o n'induit pas de réaction immunitaire chez les individus ne possédant pas cet allèle. ainsi, du point de vue des conséquences médicales (en cas de transfusion) qu'ils produisent, il est possible de distinguer 4 traits de caractères, déterminés par 6 génotypes différents : le trait a issu de $a \!\parallel\! a$ ou de $a\!\parallel\! o$, le trait b issu de $b\!\parallel\! b$ ou de $b\!\parallel\! o$, le trait o issu de $o\!\parallel\! o$ uniquement et le trait ab issu de $a\!\parallel\! b$ uniquement. ces traits sont qualifiés de phénotypes. en biologie, on les note plus volontiers entre crochets, soit [a], [b], [o] et [ab], et on les qualifie présentement de groupes.
la nature dominante ou non d'un allèle ne remet pas en question la règle d'équiprobabilité posée plus haut. cela tient au mécanisme de la méiose, le processus de division cellulaire qui donne naissance aux gamètes et sur lequel la loi de mendel se fonde. concrètement, un individu $a\!\parallel\! o$ a autant de chances de produire un gamète avec l'allèle dominant a qu'un gamète avec l'allèle récessif o. évidemment, un individu $a\!\parallel\! a$ ne peut produire qu'un gamète avec l'allèle a, un individu $o\!\parallel\! o$ qu'un gamète avec l'allèle o. au regard des probabilités,

écrire en majuscule un certain variant g_i ou l'écrire en minuscule, g_i, n'a donc aucune incidence. aussi, nous l'indiquerons toujours en majuscule dans ce qui suit. et nous *considérerons que l'allèle que produit un individu qui en aurait deux distincts est une issue de bernoulli [5] de paramètre $\frac{1}{2}$.*

le modèle construit par hardy et weinberg pour expliquer la stabilité de la structure génétique d'une population au cours des générations repose sur des hypothèses supplémentaires. la première d'entre elles est la *panmixie*. elle consiste à considérer que les *individus s'accouplent sans tenir aucun compte de leur génotype ni même de leur phénotype (qui en est la partie visible)*. autrement dit, cette hypothèse consiste à considérer que les individus choisissent au hasard leur partenaire sexuel, ce qui est bien entendu rarement le cas en réalité. le modèle d'hardy-weinberg fait également *l'hypothèse de pangamie :* les gamètes que chaque partenaire a produits se rencontrent au hasard.

bref, de la panmixie, de la loi de discrimination de mendel, et de la pangamie, retenons que tout se passe comme si les deux futurs parents étaient juste tirés au sort dans la population mâle et femelle sans qu'aucun des deux tirages n'influence l'autre (ils ne « cherchent » pas à se ressembler ni à être complémentaires par exemple), et comme si les allèles qu'ils transmettent ensuite étaient eux aussi tirés au sort (parmi deux possibles pour chaque partenaire) sans qu'aucun des deux tirages n'influe non plus sur l'autre. du point de vue des probabilités, nous poserons qu'au moment d'une fécondation, les *événements* « le père transmet l'allèle g_i » *et* « la mère transmet l'allèle g_j » *sont indépendants* [6]. du point de vue de la biologie, c'est cette « grande loterie » du vivant, répétée sur tous les gènes de tous les chromosomes, qui assure que chaque individu est un être inédit.

2. le principe de hardy-weinberg

soit une population « zéro », f_0, composée d'individus d'une même espèce. une expérience aléatoire peut consister à les tirer au sort. aussi, nous introduisons les probabilités suivantes, que nous *supposerons indifférentes au sexe, mâle ou femelle*. pour tous i,j dans $\{1,\dots,p\}$:

- $p_{\{i,j\},0} = \mathbb{p}\big(\text{« un individu a le génotype $g_i\!\parallel\! g_j$ »}\big)$. lorsque $i \neq j$, notons qu'un individu $g_i\!\parallel\! g_j$ peut être d'allèle g_i par son père et g_j par sa mère, ou vice versa et que, vu des probabilités, père et mère sont interchangeables.
- $p_{i,0}=\mathbb{p}\big(\text{« un individu transmet dans son gamète l'allèle g_i »}\big)$

à l'origine, la population étudiée pourrait réunir différents individus très homogènes, essentiellement homozygotes pour le gène g. il est donc possible qu'au sein du grand groupe ainsi reconstitué, les densités soient inéquitablement distribuées, avec des probabilités homozygotes $p_{\{i,i\},0}$ significatives et des probabilités hétérozygotes $p_{\{i,j\},0}$, $i\neq j$, insignifiantes.

nous admettons que la population se régénère à un rythme régulier *par vagues successives* f_1, f_2,...f_n,... *étanches les unes aux autres* : une génération donnée se reproduit au même moment quand les générations passées ne sont, elles, plus fécondes. on dit que les *générations successives sont discrètes et non chevauchantes*, hypothèse bien sûr discutable. dès lors, nous définissons à l'avenant pour la génération f_n les probabilités $p_{\{i,j\},n}$ et $p_{i,n}$ quand $i,j \in \{1,\dots,p\}$.

principe de hardy-weinberg

dans un cadre idéal, les probabilités p_i sont stables dès le départ, tandis que les probabilités $p_{\{i,j\}}$ se stabilisent à partir de la première génération filiale, f_1. ainsi la répartition génotypique de la population se fige-t-elle à l'image du premier brassage qui se sera produit.

on démontre ce principe en passant d'un degré n de filiation au degré $n+1$ pour les p_i, puis au degré $n+2$ pour les $p_{\{i,j\}}$, $n\geq 0$. trois étapes nous seront nécessaires.

2.1. étape 1 : calcul de $p_{\{i,j\},n+1}$

tout d'abord, compte-tenu des définitions ci-avant, de la similarité des rôles du père et de la mère, des propriétés d'indépendance déjà soulevées, quand $i,j\in \{1,\dots,p\}$ avec $i\neq j$,

$$\begin{align*} p_{\{i,j\},n+1} & = \mathbb{p}(\text{« un individu de f_{n+1} est du génotype $g_i \!\parallel\! g_j$ »})\\ &=\mathbb{p}(\text{« le père transmet l'allèle }g_i\text{ »}\cap \text{« la mère transmet l'allèle }g_j\text{ »}) \\ &\quad +\mathbb{p}(\text{« le père transmet l'allèle }g_j\text{ »}\cap \text{« la mère transmet l'allèle }g_i\text{ »})\\ &= 2\mathbb{p}(\text{« le père transmet l'allèle g_i »}\cap \text{« la mère transmet l'allèle g_j »})\\ &= 2\mathbb{p}(\text{« le père transmet l'allèle g_i »})\times\mathbb{p}(\text{« la mère transmet l'allèle g_j »})\\ &= 2 p_{i,n} p_{j,n}. \end{align*}$$

de même, si $i\in \{1,\dots,p\}$,

$$p_{\{i,i\},n+1} = p_{i,n}^2$$

2.2. étape 2 : calcul de $p_{i,n+1}$ et première stabilité annoncée

prenons $i\in \{1,\dots,p\}$. d'après la formule des probabilités totales,

$$\begin{align*} p_{i,n+1} &= \mathbb{p}(\text{« un individu de f_{n+1} transmet l'allèle g_i »})\\ &= \sum_{k=1}^p \mathbb{p}(\text{« l'allèle transmis est g_i »} \mid \text{« l'individu est de génotype $g_i\!\parallel\! g_k$ »})\\ & \qquad\qquad\qquad\qquad\qquad\times \mathbb{p}(\text{« l'individu est de génotype $g_i\!\parallel\! g_k$ »}) \end{align*}$$

quand $k\neq i$, la première probabilité vaut $\frac{1}{2}$, et la seconde $p_{\{i,k\},n+1} = 2p_{i,n} p_{k,n}$ d'après l'étape 1. quand $k=i$, la première probabilité vaut 1 et la seconde $p_{\{k,k\},n+1}=p_{k,n}^2$ toujours d'après l'étape 1. la somme se simplifie donc en $$\label{eq2} p_{i,n+1} = \sum_{k=1}^p p_{i,n} p_{k,n}=p_{i,n}\sum_{k=1}^p p_{k,n}= p_{i,n}$$ $\leadsto$ ceci prouve la première stabilité annoncée.

2.3. étape 3 : calcul de $p_{\{i,j\},n+2}$ et deuxième stabilité annoncée

comme à l'étape 1, pour tous $i,j\in \{1,\dots,p\}$, $i \neq j$,

$$p_{\{i,j\},n+2} = 2p_{i,n+1}\, p_{j,n+1}$$

mais alors, d'après l'étape 2, suivie de l'étape 1, $$p_{\{i,j\},n+2} = 2p_{i,n} p_{j,n} = p_{\{i,j\},n+1}$$ le raisonnement est analogue quand

$i=j$.

$\leadsto$ ceci prouve la deuxième stabilité annoncée.

une autre façon de mener la preuve consistait à suivre un variant quelconque v, symbolisant un certain g_i où $i\in\{1,\dots,p\}$, à désigner sous l'appellation générique $\overline{v}$ tous les autres dont on aurait fait masse, puis à ne manipuler que des probabilités du type $p_{v\parallel v}$, $p_{\overline{v} \parallel \overline{v}}$, $p_{v \parallel \overline{v}}$, p_v et $p_{\overline{v}}$. ceci nous affranchit des sommations et de leur indexation, mais conduit à des informations un peu moins fines en termes de distributions.

notre démonstration contient par ailleurs plusieurs implicites :

- que de sa naissance à son âge de procréer, une génération *n'ait pas été perturbée par des facteurs extérieurs. migrations, décès prématurés ou mutations* peuvent en effet en modifier la structure. tel génotype supportera moins tel climat, tandis que des mutations (à faible probabilité sur une courte échelle de temps) peuvent altérer ou modifier les gènes et créer un nouveau réservoir de variants, etc.
- que *tous les individus d'une génération soient également viables et fertiles*. savoir que tel génotype ne serait pas fécond changerait certainement la donne.

enfin, si les probabilités « zéro » correspondent bien à des fréquences empiriques, c'est-à-dire à des proportions calculées sur une population effectivement existante, les probabilités concernant les générations suivantes ne sont que prospectives. c'est par la grâce de la loi des grands nombres, dans *l'hypothèse d'une très grande population* — échantillon statistique dont la taille n tend vers l'infini, qu'elles coïncideraient, asymptotiquement, avec les fréquences génotypiques et alléliques observées lorsque se réalise l'expérience.

3. conclusion

en pratique, nous étudions un échantillon de population. nous y mesurons des effectifs que nous rangeons dans une table, par exemple selon que le variant v est présent ou non comme dans le tableau suivant.

génotypes	$v \!\parallel\! v$	$v \!\parallel\! \overline{v}$
effectifs	$n_{v\parallel v}$	$n_{v \parallel \overline{v}}$

effectifs à l'aune d'un variant v d'un gène
tableau d'effectifs d'un échantillon de n individus d'une même espèce classés selon leur génotype. l'attention est ici portée sur un variant v d'un certain gène g.

les fréquences $\hat{p}$ et $\hat{q}$ des allèles v et $\overline{v}$ dans cet échantillon valent par conséquent :

$$\hat{p} = \frac{2n_{v\parallel v} + n_{v\parallel\overline{v}}}{2n} et \hat{q} = \frac{2n_{\overline{v}\parallel\overline{v}} + n_{v,\overline{v}}}{2n}.$$

avec $n=n_{v\parallel v}+n_{v\parallel \overline{v}}+n_{\overline{v}\parallel \overline{v}}$.

en considérant que $\hat{p}$ et $\hat{\overline{p}}=\hat{q}$ sont des estimateurs des probabilités $p=p_{v}$ et $q = p_{\overline{v}}$ qu'un individu choisi au hasard transmette l'allèle v ou l'allèle $\overline{v}$ à l'équilibre de hardy-weinberg, nous attendons les probabilités génotypiques $p_{v\parallel v}=\hat{p}^2$, $p_{v\parallel \overline{v}}=2\hat{p}\hat{q}$ et $p_{\overline{v}\parallel \overline{v}}=\hat{q}^2$, et donc les effectifs génotypiques renseignés dans le tableau ci-après.

génotypes	$v \!\parallel\! v$	$v \!\parallel\! \overline{v}$
effectifs attendus	$\hat{p}^2 n$	$2\hat{p}\hat{q}

résultats attendus
tableau des effectifs génotypiques attendus à l'équilibre de hardy-weinberg.

sur des données expérimentales, on peut comparer visuellement les deux tableaux. mais on peut aussi, plus rigoureusement, utiliser un indicateur de conformité. typiquement $$\begin{aligned} \chi_c^2 &= \sum_{\text{génotypes}} \frac{(\text{effectif attendu}-\text{effectif réel})^2}{\text{effectif attendu}}\\ &= \frac{(\hat{p}^2 n - n_{v\parallel v})^2}{\hat{p}^2 n}+\frac{(2\hat{p}\hat{q} n - n_{v \parallel \overline{v}})^2}{2\hat{p}\hat{q} n}+\frac{(\hat{q}^2 n - n_{\overline{v} \parallel \overline{v}})^2}{\hat{q}^2 n}, \end{aligned}$$ censé suivre une loi du χ^2 à $3-1=2$ degrés de liberté, ce qui permet d'élaborer un test dont l'hypothèse nulle h_0 [8] affirmerait par exemple que la population étudiée répond à la loi de hardy-weinberg.

concrètement, on observera plutôt des écarts au modèle de hardy-weinberg : à de grandes échelles de temps, les populations, de tailles aussi grandes soient-elles, n'ont pas de structure allélique stable. ceci démontre que des forces évolutives sont à l'œuvre [9][10] : sélection, migrations, mutations, dérive génétique. il reste ensuite à les déterminer, mais aussi à reconsidérer les hypothèses formulées quant à la panmixie et la pangamie par exemple.

on pourra faire l'analogie avec la physique et son principe d'inertie : quand aucune force résultante ne s'applique à un mobile, ce dernier suit un mouvement rectiligne uniforme. tout écart à ce modèle prouve, *a contrario*, la présence d'interventions extérieures.

4. remerciements

l'auteur remercie tout particulièrement jean-françois beaux, professeur de svt en classe de spéciale bcpst au lycée henri iv, bruno anselme, professeur de svt en classe de spéciale bcpst au lycée fénelon, pascal combemorel, responsable éditorial du site ens-dgesco planet-vie, vincent pantaloni, inspecteur d'académie – inspecteur régional de mathématiques dans l'académie d'orléans-tours, astrid zayana, étudiante ingénieur à l'agro nancy (ensaia) et gaëlle guillotin, inspectrice d'académie – inspectrice

pédagogique régionale de svt dans l'académie de versailles pour leur relecture minutieuse (et bienveillante !).

CRÉDITS

AUTEUR(S)/AUTRICE(S)

karim zayana

inspecteur général de l'éducation, du sport et de la recherche, professeur invité à télécom paris.

MISE EN LIGNE

victor rabiet

responsable éditorial de culturemath depuis septembre 2020.

PARTENAIRE(S)

cet article a initialement été publié sur culturemath, l'homologue de planet-vie pour les mathématiques.

l'évolution de la biodiversité génétique : le principe de hardy-weinberg

notes

1. f_1 pour *first filial generation*, première descendance de la population d'origine, f_0. sur le même modèle, f_2 est la *second filial generation*, f_3 la *third filial generation*, etc.

2. dans tout le texte, la probabilité d'un événement a conditionnellement à un événement b sera notée $\mathbb{p}(a \mid b)$ comme c'est l'usage dans la littérature, et non $\mathbb{p}_b(a)$, comme c'est l'habitude dans les manuels scolaires, afin de ne pas surcharger l'espace alloué aux indices qui sera déjà occupé par divers compteurs.

3. l'être humain possède environ 20 000 gènes.

4. on dit que les cellules de l'être humain (exception faite de ses cellules sexuelles) sont diploïdes. beaucoup d'autres êtres vivants fonctionnent ainsi. mais pas tous, comme les abeilles mâles (les faux bourdons) et la majorité des bactéries par exemple.

5. jakob bernoulli, mathématicien suisse et premier d'une longue lignée de savants, 1654 — 1705.

6. attention toutefois. pour $g \neq g'$, les événements « le père transmet l'allèle g_i » et « le père transmet l'allèle g'_j » ne sont en revanche pas indépendants. certains gènes « voyagent » ensemble, portés qu'ils sont par un même chromosome. et tout cela peut être complexifié par les phénomènes de *crossing-over* (enjambement), quand des morceaux de chromosomes homologues s'échangent.

bibliographie

1 anselme, b. . (2015). *biomathématiques*. dunod.

2 programme de l'enseignement scientifique de la classe de terminale de la voie générale. (2019). *bulletin officiel spécial n°8 du 25 juillet*. consulté à l'adresse https://www.education.gouv.fr/pid285/bulletin_ officiel.html?pid_bo=39051

3 vulpiani, a. . (2021). *hasard, probabilité et complexité*. nouvelles visions des sciences. spartacus-idh. consulté à l'adresse https://spartacus-idh.com/liseuse/084/

4 crow, j. . (1999). hardy, weinberg and language impediments. *genetics*.

5 programme d'enseignement de spécialité de mathématiques de la classe de première de la voie générale. (2019). *bulletin officiel spécial n°1 du 22 janvier*. consulté à l'adresse https://www.education.gouv.fr/bo/19/special1/mene1901632a.htm

6 programme de l'enseignement de spécialité des sciences de la vie et de la terre de première des voies générale et technologique. (2019). *bulletin officiel spécial n°1 du 22 janvier*. consulté à l'adresse https://cache. media.eduscol.education.fr/file/sp1-men-22-1-2019/54/2/ spe648_annexe_1063542.pdf

7 programme de l'enseignement de spécialité des sciences de la vie et de la terre de terminale des voies générale et technologique. (2019). *bulletin officiel spécial n°8 du 25 juillet*. consulté à l'adresse https://cache. media.eduscol.education.fr/file/spe8_menj_25_7_2019/11/4/ spe252_annexe_1159114.pdf

8 couty-fredon, f. ., debord, j. ., & fredon, d. . (2018). *probabilités et statistiques*. dunod.

9 samadi, p. . (2021). *la théorie de l'évolution*. champs sciences. flammarion.

10 harari, y. . (2015). *sapiens*. albin michel.