

La mathématisation de l'hérédité

Hervé Perdry, Université Paris-Sud

Ce court texte présente quelques-uns des aspects des théories de l'hérédité qui se sont prêtées à une mathématisation, à travers les contributions de Mendel, de Galton et Pearson, puis de Fisher ; cette approche historique permettra de présenter la notion, toujours d'actualité, d'*hérédité*.

1 Gregor Mendel

Gregor Mendel est né en 1822 d'une famille de paysans sans fortune, dans ce qui était alors l'Empire d'Autriche. L'imagerie populaire conserve l'image d'un moine qui fit ses expériences sur les pois dans les jardins du monastère de Brno, ce qui en fait un scientifique amateur. Il reçut cependant une éducation scientifique complète à l'université de Vienne ; s'il était bien un scientifique amateur dans la mesure où la science n'était pas son métier, sa formation fut en revanche celle d'un professionnel. Il était préparé à utiliser la méthode expérimentale, et ses cours de botanique l'avaient familiarisé avec l'hybridation des plantes et le phénomène de réapparition de caractères ancestraux dans la descendance des hybrides. Quand il s'installe au monastère de Brno en 1853 après avoir terminé ses études, c'est cette question qu'il est décidé à étudier de façon scientifique.

1.1 Recherches sur l'hybridation des plantes

Mendel choisit de procéder à des expériences sur le pois (*Pisum sativa*) dont la reproduction est facile à contrôler et qui produit des hybrides fertiles. Ses travaux seront communiqués à la Société des Sciences naturelles de Brno en 1865, et publiés dans les actes de cette société, sous le titre *Recherches sur l'hybridation des plantes* [24] (traduction anglaise dans [2]). Nous allons donner ici un rapide aperçu du contenu de ce mémoire. Mendel s'y intéresse presque exclusivement à des caractères discontinus (couleur ou forme des gousses, etc, en tout sept caractères), dont il souligne l'absence de formes intermédiaires.

Cas d'un caractère

La première génération d'hybrides (génération F1 en terminologie moderne) entre deux plantes issues de lignées pures qui diffèrent pour un certain caractère ne présente qu'un seul des deux caractères ancestraux, qu'il nomme le caractère *dominant* : par exemple, un hybride entre des pois à gousses vertes et des pois à gousses jaunes aura des gousses vertes – ce caractère est donc le caractère dominant. L'autre caractère, ici la couleur jaune, est le caractère *récessif*.

Le pois est une plante autogame, c'est-à-dire qu'un individu peut se reproduire avec lui-même. Mendel laisse les hybrides se reproduire par autogamie, et observe à la génération suivante (la génération F2) la réapparition du caractère récessif ; il observe que les individus présentant les caractères récessifs et dominants sont en proportion 1 : 3 (une plante récessive pour trois plantes dominantes, figure 1).

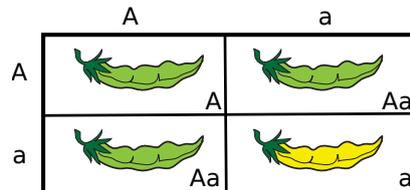


Figure 1 – Descendance d'un hybride. Les lettres en marge représentent les gamètes parentaux (voir plus bas).

Mendel poursuit l'expérience, toujours en laissant les plantes se reproduire par autogamie. Il observe ainsi que dans la descendance d'un individu de la génération F2 présentant le caractère récessif (gousses jaunes) on n'observe plus que ce caractère ; et il observe également qu'un tiers des plantes F2 présentant le caractère dominant ont une descendance exclusivement dominante, tandis que les deux-tiers restant ont, comme l'hybride F1, une descendance où les deux caractères s'observent en proportion 1 : 3.

Mendel comprend que la proportion observée est en fait une proportion 1 : 2 : 1 de formes qu'il note A, Aa et a : les formes A et a sont « constantes » pour les caractères dominants et récessifs, et Aa est la forme hybride où les deux caractères sont présents, le caractère dominant étant seul visible ; il choisit de résumer ces proportions par l'expression formelle

$$A + 2Aa + a$$

qui résume les proportions observées dans la descendance d'un croisement d'une plante de forme A avec une plante forme a. Il montre ensuite que ce principe, aujourd'hui connu sous le nom de loi de ségrégation des caractères, permet de calculer les proportions attendues des trois formes de plantes aux générations suivantes.

Cas de plusieurs caractères

Mendel poursuit ses expériences en hybridant des plantes qui diffèrent par plusieurs caractères, un des parents portant des caractères A, B, etc, et l'autre des caractères a, b, etc. Il obtient à la génération F1 des plantes « dihybrides » (de forme AaBb) qui ne présentent que des caractères dominants ; à la génération F2 il obtient les 4 combinaisons possibles de caractères dans les proportions 1 : 3 : 3 : 9 (soit une plante sur 16 présentant les deux caractères récessifs, 3 plantes récessives pour le premier caractère et dominantes pour le second, etc ; figure 2).

Mendel explique cette répartition en remarquant comme auparavant que certaines des plantes qui présentent un caractère dominant portent les deux caractères, le dominant

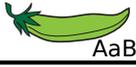
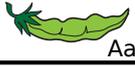
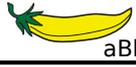
	AB	Ab	aB	ab
AB	 AB	 ABb	 AaB	 AaBb
Ab	 ABb	 Ab	 AaBb	 Aab
aB	 AaB	 AaBb	 aB	 aBb
ab	 AaBb	 Aab	 aBb	 ab

Figure 2 – Descendance d'un dihybride.

étant le seul observé. Les proportions de chacun des types sont obtenues en combinant formellement les deux expressions $A + 2Aa + a$ et $B + 2Bb + b$ pour obtenir

$$AB + 2ABb + Ab + 2AaB + 4AaBb + 2Aab + aB + 2aBb + ab.$$

La figure 2, où les capitales A et B correspondent aux caractères dominants « gousse verte » et « gousse gonflée », associés aux caractères récessifs « gousse jaune » et « gousse étranglée », illustre le raisonnement. Cette figure peut s'obtenir en éclatant les quatre cases de la figure 1, où seule la couleur de la gousse est importante (le caractère A/a) en quatre cases où on fait varier la forme de la gousse (le caractère B/b) selon un motif analogue.

Ce résultat est connu sous le nom de loi de ségrégation indépendante des caractères.

Mécanisme proposé

Mendel propose un mécanisme simple pour expliquer ses résultats : les cellules reproductrices émises par les plantes (en termes modernes, les gamètes ; Mendel utilise les mots *Keimzellen* et *Pollenzellen*, cellules germinales et cellules du pollen) ont un caractère A ou a ; et les cellules reproductrices des hybrides présentent toutes les combinaisons possibles de caractères ancestraux dans des proportions égales.

Le cas le plus simple est celui des hybrides Aa qui émettent des gamètes A et a dans les proportions 1 : 1 et l'appariement aléatoire des gamètes fait le reste. Le même mécanisme explique ce qui est constaté pour les dihybrides, avec quatre types gamétiques AB, Ab, aB et ab (voir les marges des figures 1 et 2).

Mendel teste cette hypothèse avec succès par une expérience appelée aujourd'hui *rétrécissement* : il s'agit de croiser un hybride avec un de ses parents (ou toute autre plante de forme constante, qui n'émet des gamètes que d'un seul type). Ainsi, le croisement d'une plante Aa croisée avec une plante a produit, selon l'hypothèse de Mendel, des plantes Aa et a (qui présentent respectivement les caractères dominant et récessif) en proportions 1 : 1. De même, le croisement d'une plante $AaBb$ avec une plante ab produit des plantes $AaBb$, aBb , Aab et ab dans les proportions 1 : 1 : 1 : 1. Ici encore, ces quatre types de plantes sont reconnaissables par l'observation des caractères présentés.

Caractères continus

Ces résultats concernent des caractères discontinus, sans forme intermédiaire. Mendel rapporte également des expériences sur les haricots (*Phaseolus multiflorus*), pour lesquels il a notamment considéré la couleur des fleurs. Il est dérouté par les résultats : il observe en effet un continuum de variations (du blanc au violet) et de trop rares retours à la forme récessive ou supposée telle (le blanc).

Il note pourtant :

Même ces résultats énigmatiques, cependant, peuvent probablement s'expliquer par les lois qui régissent *Pisum* si nous supposons que la couleur des fleurs et des graines de *Ph. multiflorus* est la combinaison de deux couleurs entièrement indépendantes ou plus, qui se comportent individuellement comme n'importe quel autre caractère constant de la plante. Si la couleur de fleur A est la résultante des caractères $A_1 + A_2 + \dots$, qui produit la couleur violette, alors par hybridation avec le caractère distinct de couleur blanche a , on obtient un individu hybride $A_1a + A_2a + \dots$ (...). Selon les hypothèses précédentes, ces caractères hybrides sont indépendants et vont par conséquent se développer de façon indépendante. On voit alors facilement que la combinaison indépendante de tels caractères produirait une série complète de couleurs. Si par exemple, $A = A_1 + A_2$, alors aux hybrides A_1a et A_2a correspondent les séries

$$A_1 + 2A_1a + a$$

$$A_2 + 2A_2a + a$$

dont les membres se combinent de neuf façons différentes, chacune désignant une couleur différente :

$$\begin{array}{lll} 1 A_1A_2 & 2 A_1aA_2 & 1 A_2a \\ 2 A_1A_2a & 4 A_1aA_2a & 2 A_2aa \\ 1 A_1a & 2 A_1aa & 1 aa \end{array}$$

Les nombres qui précèdent chaque combinaison indiquent combien de plantes de la couleur correspondante font partie de la série. Le total étant de 16, toutes les couleurs doivent en moyenne apparaître parmi 16 plantes, mais, on le voit, en proportions inégales.

Il est difficile de ne pas voir, dans ce court passage, une esquisse du modèle polygénique que Fisher développera en 1918 : il aurait suffi que Mendel assigne aux différentes combinaisons énumérées ci-dessus une nuance dépendant du nombre de caractères dominants et récessifs portés pour l'avoir complètement formulé.

1.2 Postérité

Le mémoire de Mendel est étonnant de clarté et modernité ; cela s'explique en partie par le fait que le modèle proposé par Mendel, qui correspond à une certaine réalité biologique, nous est familier puisqu'il est utilisé et enseigné de nos jours. Dans le texte qui précède, il n'y a presque qu'un changement qui serait fait par un biologiste moderne : c'est l'utilisation

de la notation AA au lieu de A, pour les plantes ne portant que le caractère A. On parle aujourd'hui des deux *allèles* A et *a* du gène considéré ; la notation AA indique le fait qu'on a reçu un allèle A de chacun de nos parents. Le caractère observé est appelé *phénotype*.

Les travaux de Mendel sont passés inaperçus de son vivant. C'est peut-être en partie dû au fait qu'ils étaient présentés comme des travaux sur l'hybridation des plantes, et non sur les lois de l'hérédité, sujet qui intéressait plus particulièrement ceux qui l'auraient lu avec le plus d'intérêt ; mais surtout, Mendel étant devenu en 1868 le père supérieur de son couvent, il fut absorbé par cette tâche et n'eut guère le loisir de donner davantage de publicité à sa théorie. Il mourut en janvier 1884 d'une insuffisance rénale, et ses travaux ne furent redécouverts qu'en 1900 par Hugo de Vries, Karl Correns et Erich von Tschermak qui réalisaient des expériences similaires. Mendel n'émet aucune hypothèse sur la nature matérielle des caractères transmis ; l'hypothèse que les chromosomes en constituaient le support physique fut vite émise, et c'est à Thomas Morgan, qui établit la première carte génétique – celle du génome de la drosophile –, qu'on doit la confirmation expérimentale de ce fait.

La loi de ségrégation indépendante des caractères est fautive en générale : elle n'est vraie que pour des caractères *non liés* – c'est le cas si les gènes correspondant à ces caractères sont sur des chromosomes distincts. Dans le cas contraire, la loi reste approximativement vraie si les gènes sont suffisamment éloignés les uns des autres. C'est globalement le cas des sept caractères étudiés par Mendel, à l'exception de deux (la longueur de la tige et la forme des gousses) [35], pour lesquels il est possible qu'il n'ait pas réalisé d'expérience avec des doubles hybrides.

La note finale est plus négative : en 1936, Ronald Fisher réanalyse les résultats de Mendel [8], et montre que les proportions observées par Mendel dévient trop peu de celles prédites par la théorie par rapport aux déviations aléatoires attendues. L'accumulation de déviations trop petites, d'expérience en expérience, est accablante. Voici donc Gregor Mendel soupçonné d'avoir manipulé ses données. Il faut cependant nuancer un peu. Mendel n'avait aucune idée des ordres de grandeurs des déviations attendues ; dépourvu de l'outil mathématique (la statistique du χ^2) nécessaire à leur analyse, il a pu écarter de bonne foi de ses hybrides F2 des échantillons à ses yeux suspects de contamination par une fertilisation extérieure ; il a également pu réaliser plusieurs expériences et choisir de ne rapporter que celle qui correspondait le mieux à la théorie, méthode qui fournit des déviations en accord avec celles qui sont observées [34]. Les exigences de rigueur expérimentale en biologie ne pouvaient pas en 1850 être ce qu'elles sont de nos jours, et cette erreur ne peut être jugée avec la sévérité qui serait de mise à présent. Il faut noter en outre que ces critiques ne portent que sur la collecte ou le traitement des données issues des expériences ; la conception des expériences est impeccable, et les rétrocroisements restent un des outils des expérimentateurs en génétique animale ou végétale.

2 Sir Francis Galton

Francis Galton est né en 1822 (la même année que Mendel) dans une famille anglaise aisée ; c'est un enfant précoce et brillant. Son père le destine à la médecine, mais les études médicales lui déplaisent, sa préférence allant aux mathématiques ; il fit en troisième année

d'études à Cambridge une dépression sévère, liée à des résultats moins bons qu'espérés dans cette matière [31]. Il se tourne alors à nouveau brièvement vers la médecine, jusqu'à la mort de son père en 1844, qui le rend financièrement indépendant. Il voyage en Afrique et au Moyen-Orient, tout d'abord sans but scientifique, puis sous le patronage de la Société royale de géographie. Il devient alors un auteur prolifique, publiant chaque année plusieurs articles et ouvrages sur une foule de sujets : météorologie, avalanches, voyages...

Comme beaucoup de ses contemporains, il est passionné par l'ouvrage de Charles Darwin (qui se trouve être son cousin *) publié en 1859, *L'Origine des espèces*. Il consacre dès lors une part croissante de son activité à la réflexion sur les lois de l'hérédité, ce qui le conduira à fonder une nouvelle science, l'« eugénisme », qu'il définit comme la science qui traite des façons d'améliorer l'espèce humaine [21].

2.1 Les lois de l'hérédité

Galton réalisera une série d'expériences pour tester la théorie de la pangenèse de Darwin, qui postulait que tous les organes émettent des *gemmules* qui s'agrègent entre elles avant d'être transmises à la descendance. Il transfuse des lapins gris avec du sang de lapins blancs, dans l'espoir que la descendance des lapins gris présente des traits hybrides ; l'expérience n'est pas concluante [11] : le sang des lapins blancs ne transporte pas de gemmules.

Après cet échec, Galton élaborera sa propre théorie de l'hérédité [5, 12–14], formulant tout d'abord une théorie biologique, avant de se concentrer sur la recherche d'une loi mathématique de l'hérédité. Galton cherche une loi qui explique à la fois l'hérédité des traits continus (comme la stature) et celle des traits discrets (la couleur des yeux), et en particulier pour ces derniers la réapparition de caractères ancestraux qu'il appelle l'*atavisme*.

Une théorie de l'hérédité : la stirpe

Dans *Une théorie de l'hérédité* [14], article publié en 1875, Galton propose d'appeler *stirpe* (du latin *stirpes*, racine), l'ensemble des germes présents dans l'œuf fertilisé et qui sont à l'origine du développement de l'organisme. Il énonce quatre postulats qui lui semblent nécessaires à une théorie organique de l'hérédité :

1. l'organisme est la juxtaposition d'un grand nombre d'unités quasi-indépendantes, qui dérivent de germes distincts ;
2. la stirpe contient une multitude de germes, bien plus nombreux et divers que les unités organiques qui en seront dérivées, de sorte que très peu de ces germes sont finalement développés ;
3. les germes qui ne sont pas développés conservent leur vitalité et contribuent à la formation de la stirpe de la descendance de l'individu ;
4. la structure de l'organisme découle des affinités mutuelles des germes, au sein de la stirpe et au cours du développement.

*ou plus précisément son demi-cousin, la mère de Francis Galton, Frances Darwin, étant la demi-sœur de Robert Darwin, le père de Charles Darwin. Leur grand-père commun, Erasmus Darwin, est un médecin et naturaliste célèbre.

Galton estime le demi-écart interquartile de la stature dans la population, il est de 1,7 pouce ; celui de la stature M du parent-moyen est estimé à 1,19 pouce, ce qui est cohérent avec une valeur théorique (en supposant l'indépendance des tailles des deux parents) de $\frac{1}{\sqrt{2}}1,7 = 1,21$ pouce ; et il estime le demi-écart interquartile dans les fratries à 1,5 pouce*.

À partir de la Table 1, Galton met en évidence la *regression towards mediocrity*, en français « régression vers la moyenne » : l'écart entre la stature Y d'un individu et la stature moyenne μ tend à être moins important que celui qu'on observe entre M et μ . Plus précisément, on a approximativement (les notations sont de nous) †

$$E(Y - \mu | M) = \frac{2}{3}(M - \mu).$$

Il s'agit de l'espérance de Y conditionnellement à M (dans le texte : la déviation filiale vaut *en moyenne* seulement les deux-tiers de la déviation du parent-moyen). Toujours à partir de cette table, Galton estime également la « régression de la stature du parent-moyen » :

$$E(M - \mu | Y) = \frac{1}{3}(Y - \mu).$$

Pour estimer le coefficient de régression de la stature des enfants, Galton répartit les parents-moyens en catégories (la stature est arrondie au pouce le plus proche), puis il calcule la moyenne des statures de tous les enfants d'une catégorie ; on reporte les quantités obtenues sur un graphe (figure 3), et, les points étant à peu près alignés, on y fait passer « au jugé » une droite dont on détermine ensuite la pente (figure 3).

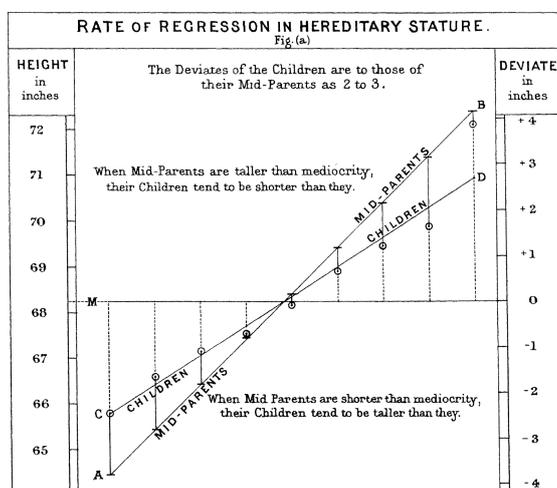


Figure 3 – La droite de régression (figure IX de [16])

*Le demi-écart interquartile Q est la mesure de dispersion utilisée par Galton. Dans le cas de la loi normale on a $Q = 0,76 \times$ l'écart-type ; les mesures ci-dessous correspondent à des écart-types de 2,52 pouces (population), 1,76 pouce (parent-moyen), 2,22 pouces (fratries).

†Pour ce coefficient de $\frac{2}{3}$, Galton dit avoir tout d'abord fait une estimation de $\frac{3}{5}$, mais avoir préféré $\frac{2}{3}$ qui est plus simple. Est-ce par qu'il recherche une loi naturelle qu'il pense devoir être parcimonieuse ?

Voici une interprétation moderne des résultats obtenus : la loi jointe du vecteur $(Y - \mu, M - \mu)$ est à peu près une loi normale bivariée centrée, de matrice de variance

$$2,52^2 \times \begin{pmatrix} 1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{2} \end{pmatrix},$$

ce qui correspond aux régressions $E(Y - \mu|M) = \frac{2}{3}(M - \mu)$ et $E(M - \mu|Y) = \frac{1}{3}(Y - \mu)$. En outre, la variance de Y conditionnellement à une valeur donnée de M est

$$\text{var}(Y) - \left(\frac{2}{3}\right)^2 \text{var}(M) = \frac{14}{18} \times 2,52^2 \simeq 2,22^2,$$

ce qui est cohérent (l'ordre de grandeur de l'erreur est d'un millièmètre de pouce) avec la valeur de l'écart-type à l'intérieur des fratries estimée (indirectement) par Galton (voir la note en bas de la page 8).

Pour traiter cette question en 1886, celui-ci détermine tout d'abord que les courbes de niveau de la loi jointe de Y et M sont des ellipses dont il estime certaines caractéristiques. Il s'adresse ensuite à James Dickson, professeur à Cambridge, auquel il soumet les caractéristiques de ces ellipses, ainsi que la pente de la régression de Y sur M , en lui demandant de retrouver la pente de la régression de M sur Y ; Dickson répond que cette pente est 0,34, ce qui est compatible avec l'estimation de Galton citée plus haut. Tout ceci convainc Galton de la cohérence de ses résultats.

Ainsi Galton a eu l'intuition de la possibilité de relier sa loi de « régression vers la moyenne » aux caractéristiques d'une loi normale bivariée. Des résultats généraux à ce sujet avaient déjà été énoncés par Bravais [4]; il faudra attendre quelques années pour que Yule, Edgeworth, et bien sûr Pearson (cf par exemple [28]) reprennent et étendent ces résultats, posant des bases mathématiques solides à la théorie de la régression linéaire.

Contribution de chaque ancêtre À partir des constatations faites ci-dessus, Galton veut estimer la contribution de chaque ancêtre à la stature de l'individu. En effet, il faut démêler dans ce qui précède ce qui est réellement imputable aux parents, de ce qui est imputable à une partie de la stirpe transmise par les parents mais provenant d'ancêtres plus lointains, et qu'on n'observe qu'indirectement chez les parents – Galton ne fait pas explicitement référence à la stirpe dans le texte, mais on ne comprend sa démarche que si on a ce modèle à l'esprit.

Voici comment il procède (je suis ci-après le texte d'assez près sans toutefois toujours le traduire littéralement, cf [18] pp 134–136) :

Si un parent-moyen a un écart de D à la stature moyenne, ses enfants ont un écart, en moyenne, de $\frac{2}{3}D$, ceci indépendamment de la contribution des ancêtres plus éloignés. D'autre part, un écart de D chez un individu implique un écart $D' = \frac{1}{3}D$ chez son parent-moyen, qui lui-même implique un écart de $\frac{1}{3}D' = \frac{1}{9}D$ chez le parent-moyen de ce parent-moyen, c'est-à-dire chez le grand-parent-moyen de l'individu considéré; soit, dans la totalité de la lignée de D , des écarts dont la somme est $D + \frac{1}{3}D + \frac{1}{9}D + \dots = \frac{3}{2}D$.

Si on suppose que la contribution de chaque ancêtre est « taxée » également, pour qu'une accumulation de contributions ancestrales dont la somme est $\frac{3}{2}D$

produise un héritage effectif de $\frac{2}{3}D$, il faut que chaque contribution ait été réduite par un facteur de $\frac{4}{9}$, puisque $\frac{4}{9} \times \frac{3}{2} = \frac{2}{3}$.

Une autre possibilité est que la contribution de chaque ancêtre soit taxée de façon répétée à chaque transmission, et qu'une proportion $\frac{1}{r}$ seulement soit transmise à chaque fois. Dans ce cas l'héritage effectif serait $(\frac{1}{r} + \frac{1}{3r} + \frac{1}{9r^2} + \dots)D = \frac{3}{3r-1}D$, et pour qu'il soit égal à $\frac{2}{3}D$ il faut prendre $\frac{1}{r} = \frac{6}{11}$.

Selon le modèle choisi, les particularités du parent-moyen contribuent pour $\frac{4}{9}$ aux particularités de l'enfant, ou pour $\frac{6}{11}$. Ces valeurs diffèrent peu de $\frac{1}{2}$, et leur moyenne est proche de $\frac{1}{2}$, donc on peut accepter ce résultat. Ainsi l'influence pure et simple du parent-moyen peut être considérée comme égale à $\frac{1}{2}$, celle du grand-parent à $\frac{1}{4}$, etc. Par conséquent l'influence d'un parent individuel est de $\frac{1}{4}$, d'un grand-parent de $\frac{1}{16}$, etc.

On voit que dans le dernier paragraphe, Galton n'envisage plus que le second modèle, celui où la contribution des ancêtres décroît selon une série géométrique. Le moins qu'on puisse dire est que l'argumentaire manque de rigueur. Galton en est peut-être conscient ; il conclut par ces quelques mots :

Il serait cependant hasardeux, sur cette fragile base, d'étendre cette séquence avec confiance à des générations plus éloignées.

La loi de l'hérédité de 1897

En 1897, dans *La contribution moyenne de chacun des ancêtres à l'héritage total de leur descendance* [19], Galton n'hésite plus : la loi qu'il n'a ébauchée qu'avec précaution dans *Natural Inheritance* lui semble à présent suffisamment confirmée, en particulier par l'analyse de nouvelles données sur la couleur du pelage des bassets.

Il postule donc que les parents d'un individu contribuent à eux deux (en moyenne) pour moitié à l'héritage total de leur enfant, soit (en moyenne) un quart chacun ; que les quatre grands-parents contribuent (en moyenne) pour un quart, soit un seizième chacun ; etc. Avec des notations modernes, on pourra écrire

$$Y_1 = \frac{1}{4} \underbrace{(Y_2 + Y_3)}_{\text{deux parents}} + \frac{1}{16} \underbrace{(Y_4 + Y_5 + Y_6 + Y_7)}_{\text{quatre grands-parents}} + \frac{1}{64} \underbrace{(Y_8 + \dots + Y_{15})}_{\text{huit bisaïeuls}} + \dots \quad (1)$$

La séduction de cette loi aux yeux de Galton tient beaucoup au fait que la somme des contributions ancestrales vaut 1, ce qui permet notamment d'appliquer la loi tant au phénotype qu'à sa déviation d'avec la moyenne. En effet, on a

$$1 = \frac{1}{4} \underbrace{(1 + 1)}_{\text{deux termes}} + \frac{1}{16} \underbrace{(1 + 1 + 1 + 1)}_{\text{quatre termes}} + \frac{1}{64} \underbrace{(1 + \dots + 1)}_{\text{huit termes}} + \dots$$

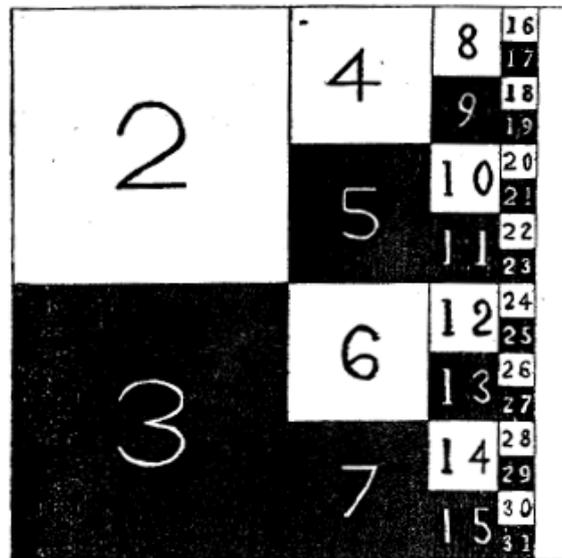


Figure 4 – Un diagramme de l’hérédité. Il faut mentalement compléter le carré par une multitude de carrés de plus en plus petits. Figure de Meston [25], reproduite par Galton [20].

Dans une lettre à Nature de 1898 [20], Galton reproduit la figure 4 créée par Meston [25], qui lui paraît propre à illustrer et à populariser cette loi. L’intérêt de cette figure apparaît si on considère que Galton veut en fait élucider la composition de la stirpe, même s’il n’en fait plus mention : elle représente la stirpe d’un individu, mosaïque des stirpes de ses ancêtres.

Bien que l’égalité énoncée ci-dessus paraisse déterministe, une variabilité subsiste : un individu n’est pas totalement déterminé par l’ensemble de ses ancêtres – il faut expliquer les différences entre les membres d’une même fratrie. Pour Galton ces différences proviennent de variations dans la valeur des proportions $\frac{1}{4}$, $\frac{1}{16}$, etc, qui donnent la contribution de chacun des ancêtres à la stirpe de l’individu, et ne sont que des valeurs moyennes ; en pratique elles peuvent s’écarter de ces valeurs (mais leur somme restera égale à 1).

Application aux traits discrets Pour Galton, l’hérédité de traits discrets comme la couleur des yeux [15, 18] ou celle du pelage des bassets [19], est également déterminée par la loi (1). Il suppose implicitement que ces caractères ne se développent qu’à partir d’un seul des germes de la stirpe. La couleur des yeux d’un individu donne des informations sur la façon dont la composition de celle-ci diffère de la composition moyenne de la population ; Galton utilise la loi (1) pour calculer cette composition, puis la probabilité que deux parents aux yeux clairs aient un enfant aux yeux clairs, etc. Pour alléger l’exposé, je ne développerai pas ces calculs peu convaincants malgré l’élégance de l’idée de départ.

2.2 La contribution de Pearson

J’ai choisi de ne consacrer que ce court paragraphe à Karl Pearson : un des fondateurs des statistiques modernes, eugéniste, défenseur acharné des conceptions de Galton, il en aurait mérité davantage.

Quand Galton publie son article de 1897 [19], Pearson se saisit immédiatement de cette loi qu'il appelle *Galton's Law of Ancestral Heredity*. Dans un article [29] publié en janvier 1898 et dédié à Galton (*A New Year's Greeting to Francis Galton*), il l'écrit sous une forme similaire à l'équation (1), et l'interprète comme la régression de Y_1 sur l'ensemble des valeurs du trait chez ses ancêtres, donc comme une espérance conditionnelle. Ainsi pour Pearson, les coefficients sont fixés, et si la valeur de Y_1 n'est pas totalement déterminée par les autres Y_i , cela provient de l'existence d'une variance résiduelle. La lecture de Pearson change d'emblée le sens de la loi de Galton.

La ré-interprétation va plus loin. Pearson introduit discrètement un paramètre libre en considérant la loi plus générale :

$$Y_1 = \frac{1}{2}\gamma r(Y_2 + Y_3) + \frac{1}{4}\gamma r^2(Y_4 + Y_5 + Y_6 + Y_7) + \frac{1}{8}\gamma r^3(Y_8 + \dots + Y_{15}) + \dots$$

tout en conservant la contrainte $\sum_{i \geq 1} \gamma r^i = 1$, c'est-à-dire $(1 + \gamma)r = 1$ *. Cette variante correspond à des proportions différentes de celles supposées par Galton dans la constitution du matériel héréditaire, ce que Galton et Pearson à sa suite appellent « la taxe sur l'héritage ».

Si les lois marginales sont gaussiennes, de même espérance μ et variance σ^2 , s'il n'y a ni homogamie, ni hétérogamie (c'est-à-dire qu'il n'y a pas de corrélation, positive ou négative, entre les phénotypes des individus qui forment un couple), cette équation suffit à décrire la loi jointe des Y_i . Pearson montre que la corrélation entre deux individus ancêtre l'un de l'autre et distants de ℓ générations est

$$r_\ell = c \left(\frac{r(1 + \gamma)}{2} \right)^\ell = c \left(\frac{1}{2} \right)^\ell,$$

où c dépend également de γ et r :

$$c = \frac{r^2 - 3r + 2}{r^2 - 2r + 2}. \quad (2)$$

On vérifie que c prend toutes les valeurs entre 0 et 1 quand r va de 1 à 0 ; en pratique on peut donc choisir une valeur arbitraire pour c . De plus, c est le coefficient de la régression sur le parent-moyen cher à Galton.

Pearson calcule également les corrélations entre collatéraux (frères, cousins, etc). Ce sont ces coefficients de corrélation qui l'intéressent en pratique ; dans des travaux ultérieurs (cf par exemple [32]) il s'attachera à estimer leurs valeurs pour diverses mesures anthropométriques.

2.3 Postérité

Le contraste entre Mendel et Galton, tous deux à la recherche des lois de l'hérédité, est frappant. La pauvreté de Mendel fut en grande partie la cause de ce qu'il ne put diffuser et faire reconnaître son travail ; l'aisance financière de Galton lui laissa au contraire tout loisir de donner une large publicité à ses théories.

*Étonnamment, Pearson fait ici une erreur de calcul dans [29], qu'il corrige deux ans plus tard dans [30].

Il est difficile de se faire une idée de la célébrité atteinte par Galton de son vivant ; il était considéré comme un des plus grands scientifiques de son temps. Sa réputation est restée importante dans certains milieux ; de façon générale, on le présente comme pionnier des statistiques. C'est indubitable, mais il faut préciser : des statistiques descriptives. Comme dit plus haut, Galton estime les coefficients de régression graphiquement. Il y a peu de mathématiques dans l'œuvre de Galton, et quand il fait appel à l'aide d'un mathématicien comme Dickson, certains indices permettent de douter qu'il comprenne le détail des calculs qu'il reçoit en réponse.

Outre la définition du coefficient de régression, on lui doit celle de la corrélation (dite aujourd'hui corrélation de Pearson), qu'il définit comme le coefficient de régression d'une grandeur sur une autre, les deux grandeurs ayant été exprimées en nombre d'écart-types * [17]. Il est également un pionnier de la biométrie ; son seul prédécesseur illustre est Adolphe Quételet, qui avait notamment remarqué que les mesures anthropométriques sont réparties selon une loi normale. Galton est le premier à s'intéresser de façon systématique aux corrélations entre apparentés, ouvrant une voie de recherche féconde. Est-ce bien suffisant pour ranger Galton au premier rang des savants de son temps, au même titre par exemple que Darwin, Cantor, Maxwell ?

Revenons à la loi de l'hérédité. La façon dont Galton prétend la déduire des données expérimentales laisse rêveur — l'argumentation de Swinburne [36], qui remarque que Galton avait déjà énoncé cette loi en 1865 [9] sans produire d'arguments pour la démontrer, et n'a fait par la suite que chercher à confirmer son intuition première est très convaincante.

Cependant la version de Pearson de la loi de Galton produit des résultats compatibles avec les corrélations observées entre apparentés, et avec les prédictions obtenues sous le modèle polygénique que Fisher proposera en 1918 — le tout évoque la conception classique selon laquelle une nouvelle théorie scientifique inclut la précédente comme cas particulier. Mais comme on l'a vu Pearson réinterprète la loi de Galton : tout d'abord, il lit l'égalité énoncée par Galton comme une espérance conditionnelle ; ensuite il introduit discrètement un paramètre qui permet d'obtenir des corrélations différentes selon le trait étudié, ce qui contredit l'idée première de Galton. Enfin, les traits discrets sont beaucoup mieux expliqués par les lois de Mendel, et les espoirs de Galton d'expliquer du même coup le phénomène de l'atavisme resteront vains.

Mais Galton reste également dans l'histoire comme le fondateur de l'eugénisme. Dès la première page d'*Hereditary Genius* [10], son premier livre sur l'hérédité, le décor est planté :

« De même qu'il est facile d'obtenir des races de chiens ou de chevaux particulièrement douées sur le plan de la course, ou sur tout autre plan, il serait faisable, par des mariages judicieux pendant plusieurs générations consécutives, de produire une race d'hommes extrêmement douée.

L'eugénisme se préoccupe donc du moyen d'améliorer les qualités d'une population, en éliminant le mauvais matériel héréditaire (déficients mentaux, pauvres, vagabonds, alcooliques, voleurs...) au profit du bon. Les moyens d'arriver à cette fin sont variés, on le sait. Galton se contente quant à lui de préconiser de bannir socialement les mariages peu

*En fait, Galton utilise là aussi l'écart interquartiles pour standardiser les mesures.

souhaitables d'un point de vue eugénique, et de faire de l'eugénisme une forme de nouvelle religion laïque ; à son crédit, il recommande prudemment de ne pas se hâter de prendre des mesures vouées à l'échec et qui risqueraient de discréditer la nouvelle science [21].

Galton considère également comme nécessaire de préserver « la pureté de la race », ou de n'y introduire que des individus de valeur génétique supérieure, sous peine de décadence [10]. Cette question restera une des grandes préoccupations des eugénistes – on se référera par exemple aux travaux de Pearson sur l'immigration des juifs en Angleterre [33], qui envisage la question sous l'angle de la valeur génétique de ceux-ci. Il paraît impossible d'absoudre Galton et les eugénistes de leur responsabilité dans le succès que connut l'idée d'« hygiène raciale ». On répondra qu'il faut replacer les choses dans leur contexte ; il est sans doute vrai que Galton reproduit les préjugés de son époque, mais, ce faisant, il les pare d'un vernis de respectabilité scientifique, et les renforce.

3 Enfin Fisher vint

3.1 Biométriciens et mendéliens

La redécouverte des lois de Mendel en 1900 survient alors qu'une controverse scientifique oppose d'un côté les biométriciens Karl Pearson et Raphael Weldon, et de l'autre, William Bateson. Cette controverse porte sur la nature des mécanismes de l'évolution et de la spéciation. Pour Bateson, la spéciation est un phénomène discontinu ; une nouvelle espèce apparaît quand « une mutation » crée des individus qui ne peuvent plus s'hybrider avec l'ancienne espèce. Pour Weldon, qui entraîne à sa suite son collègue Pearson, l'évolution est un phénomène graduel et continu ; la spéciation nécessite que deux populations soient isolées l'une de l'autre et évoluent dans des directions différentes. La querelle n'est pas que scientifique, c'est également un affrontement de personnalités, une détestation réciproque de Weldon et Bateson.

Dès sa redécouverte, Bateson se fait le champion du mendélisme, qui est la théorie rêvée pour la modélisation de phénomènes discontinus dans l'hérédité ; de leur côté, les biométriciens rejettent ces idées qu'ils jugent incapables d'expliquer l'hérédité des traits continus. Le mendélisme devient sinon l'enjeu principal de la controverse, du moins le plus visible ; le compromis paraît longtemps impossible et on est sommé de choisir son camp.

Du même coup, les premiers généticiens seront également souvent critiques à l'égard des théories eugénistes soutenues par les biométriciens : Bateson ne rejette pas en bloc toutes les thèses des eugénistes, mais il leur demande [3] avec malice si, plutôt que de prôner la stérilisation des criminels, il ne conviendrait pas de s'attaquer « aux fournisseurs de l'armée et à leurs complices, les patriotes de salle de rédaction » ; en 1925, Thomas Hunt Morgan, un autre pionnier de la génétique, insiste sur le rôle de l'environnement dans les traits comportementaux ou mentaux, et affirme que les preuves apportées par les eugénistes dans ce domaine sont très insuffisantes (cf [26], pp 198–207).

Cette hostilité réciproque entre biométriciens et généticiens eut pour effet qu'il fallut attendre 1918 pour que Ronald Aylmer Fisher fasse la synthèse entre les deux théories dans *The Correlation between Relatives on the Supposition of Mendelian Inheritance* [7]*.

*La bibliographie donne la date de 1919, qui est l'année de parution en volume dans les *Transactions*

3.2 Le modèle de Fisher

Fisher suppose qu'un grand nombre r de facteurs mendéliens indépendants contribuent à la valeur mesurée, qui en est la somme. Chacun de ces facteurs mendéliens est de la forme

$$X = \begin{cases} u & \text{si le génotype est AA} \\ v & \text{si le génotype est Aa} \\ w & \text{si le génotype est aa} \end{cases} \quad (3)$$

et la valeur mesurée est

$$Y = X_{(1)} + \dots + X_{(r)} + E = G + E,$$

où $G = X_{(1)} + \dots + X_{(r)}$ est l'effet total du génome ; sa loi est approximativement normale. Le terme E , indépendant de G , est supposé normal également ; il est la résultante des effets environnementaux.

Composante additive et composante de dominance

La démarche de Fisher est la suivante : le but étant de calculer la corrélation entre apparentés, on commence par s'intéresser à la corrélation entre parent et enfant, pour un unique facteur mendélien X de la forme (3).

Notons p et $q = 1 - p$ les fréquences des allèles A et a ; sous l'hypothèse d'indépendance des allèles parentaux (que nous appelons aujourd'hui le modèle d'Hardy-Weinberg), les trois génotypes ont pour fréquences p^2 , $2pq$ et q^2 . Dans le cadre mendélien, les probabilités conjointes pour les génotypes parent-enfant sont les suivantes :

	AA	Aa	aa
AA	p^3	p^2q	0
Aa	p^2q	pq	pq^2
aa	0	pq^2	q^3

Table 2 – Probabilités conjointes des génotypes parent-enfant

Notons X^p et X^e les valeurs de X chez le parent et l'enfant. Leur covariance est alors

$$\begin{aligned} \text{cov}(X^p, X^e) &= p^3 u^2 + 2p^2 q uv + pq v^2 + 2pq^2 vw + q^3 w^2 - (p^2 u + 2pqv + q^2 w)^2 \\ &= pq(p(v - u) + q(w - v))^2 \end{aligned} \quad (4)$$

Cette factorisation miraculeuse incite à considérer le cas particulier des facteurs mendéliens additifs, c'est-à-dire le cas $v - u = w - v = a$; on calcule alors $\text{cov}(X^p, X^e) = pqa^2$, $\text{var}(X) = 2pqa^2$, et on a $\text{cor}(X^p, X^e) = \frac{1}{2}$.

Pour traiter le cas général, Fisher décompose les facteurs mendéliens en une partie additive et un facteur résiduel :

of The Royal Society of Edinburgh. La date de 1918, souvent attribuée à cet article, correspond à une publication en fascicule séparé.

La contribution des facteurs mendéliens imparfaitement additifs se décompose, à des fins statistiques, en deux parties : une partie additive qui reflète la nature génétique sans distorsion et est à l'origine des corrélations observées ; et un résidu qui se comporte à peu près de la même façon que l'erreur arbitraire introduite dans les mesures.

On peut donner une interprétation géométrique de cette décomposition. En identifiant l'espace des variables aléatoires de la forme (3) à un espace euclidien de dimension 3, avec le produit scalaire $\langle X, Y \rangle = E(XY)$, X' est la projection de X sur le plan engendré par les variables aléatoires X_1 , définie par $u = v = w = 1$, et X_{012} , définie par $u = 0, v = 1$ et $w = 2$. Ce plan est bien l'ensemble des facteurs additifs. On en construit une base orthonormale (X_1, X_a) en orthogonalisant la base (X_1, X_{012}) : on obtient

$$X_a = \frac{1}{\sqrt{2pq}} (X_{012} - 2qX_1) = \begin{cases} \frac{1}{\sqrt{2pq}}(0 - 2q) & \text{si le génotype est AA} \\ \frac{1}{\sqrt{2pq}}(1 - 2q) & \text{si le génotype est Aa} \\ \frac{1}{\sqrt{2pq}}(2 - 2q) & \text{si le génotype est aa} \end{cases} \quad (5)$$

On peut compléter cette base en une base orthonormale de l'espace des facteurs mendéliens en y ajoutant

$$X_d = \begin{cases} \frac{q}{p} & \text{si le génotype est AA} \\ -1 & \text{si le génotype est Aa} \\ \frac{p}{q} & \text{si le génotype est aa} \end{cases} \quad (6)$$

On peut réinterpréter le fait que $\langle X_1, X_a \rangle = 0$ et $\|X_a\|^2 = 1$ en $E(X_a) = 0$ et $\text{var}(X_a) = 1$, c'est-à-dire que X_a est centrée et réduite ; il en va de même pour X_d . Sur cette base, X s'écrit

$$X = \mu X_1 + \alpha X_a + \delta X_d$$

avec

$$\begin{aligned} \mu &= p^2 u + 2pq v + q^2 w \\ \alpha &= \sqrt{2pq}(p(v - u) + q(w - v)) \\ \delta &= pq(u + w - 2v) \end{aligned} \quad (7)$$

On a bien sûr $E(X) = \langle X_1, X \rangle = \mu$ et $E(X^2) = \|X\|^2 = \mu^2 + \alpha^2 + \delta^2$, donc $\text{var}(X) = \alpha^2 + \delta^2$; la variance de X a été décomposée en variance additive et en variance de dominance.

Finalement, chacun des r facteurs mendéliens considérés s'écrit $X_{(i)} = \mu_i X_1 + \alpha_i X_{ia} + \delta_i X_{id}$, et leur somme est

$$G = \alpha_1 X_{1a} + \dots + \alpha_r X_{ra} + \delta_1 X_{1d} + \dots + \delta_r X_{rd}.$$

En posant $\tau_a = \sum_i \alpha_i^2$ et $\tau_d = \sum_i \delta_i^2$, on a $\text{var}(G) = \tau = \tau_a + \tau_d$.

Corrélation entre apparentés

Revenons à la corrélation entre apparentés, et d'abord à la corrélation parent-enfant. La variance du phénotype Y est $\text{var}(Y) = \text{var}(G) + \text{var}(E) = \tau + \sigma^2 = \tau_a + \tau_d + \sigma^2$. On note avec

des exposants p et e les différents termes du modèle chez le parent et l'enfant considérés. D'après les équations (4) et (7) on a, pour chacun des facteurs mendéliens impliqués,

$$\text{cov}(X^p, X^e) = \frac{1}{2}\alpha^2.$$

C'est ce résultat qui motivait la décomposition en composantes additives et dominantes. Si on écrit $X^p = \mu + \alpha X_a^p + \delta X_d^p$ et $X^e = \mu + \alpha X_a^e + \delta X_d^e$, ce résultat est équivalent à

$$\begin{aligned} \text{cov}(X_a^p, X_a^e) &= \frac{1}{2} & \text{cov}(X_a^p, X_d^e) &= 0 \\ \text{cov}(X_d^p, X_a^e) &= 0 & \text{cov}(X_d^p, X_d^e) &= 0 \end{aligned}$$

Ces valeurs peuvent bien entendu être calculées directement à partir des probabilités de la table 2. Les facteurs mendéliens étant supposés deux à deux indépendants, on a alors $\text{cov}(G^p, G^e) = \frac{1}{2}(\alpha_1^2 + \dots + \alpha_r^2) = \frac{1}{2}\tau_a$. En supposant l'indépendance des effets environnementaux E^p et E^e , on calcule

$$\text{cor}(Y^p, Y^e) = \frac{1}{2}h^2$$

où

$$h^2 = \frac{\tau_a}{\tau_a + \tau_d + \sigma^2}$$

est l'*héritabilité restreinte* ; c'est le rapport de la variance génétique additive sur la variance totale du phénotype. L'*héritabilité large* est

$$H^2 = \frac{\tau_a + \tau_d}{\tau_a + \tau_d + \sigma^2}.$$

On généralise facilement ce résultat à d'autres formes d'apparentement. Fisher envisage les ascendances directes — un individu et ses grands-parents, arrière-grands-parents, etc ; il envisage également le cas des germains (frères ou sœurs), des cousins germains, des doubles cousins germains, dressant à chaque fois une table analogue à la table 2.

Dans un formalisme moderne, on peut caractériser une relation d'apparentement entre deux individus par les probabilités $\zeta_0, \zeta_1, \zeta_2$ de chacun des trois « états IBD » possibles, qu'on peut définir comme ceci : deux allèles d'un même gène sont IBD (pour *Identical By Descent* : identique par origine) si ils sont hérités d'un ancêtre commun ; en une région du génome donnée, l'état IBD des deux individus est IBD = 0 s'ils n'ont pas d'allèles IBD, IBD = 1 s'ils ont exactement un allèle IBD, et IBD = 2 s'ils ont deux allèles IBD.

On utilise souvent la paramétrisation équivalente ϕ et ψ (table 3) :

- ϕ est le coefficient d'apparentement, défini comme la probabilité pour que deux allèles d'un même gène autosomal, tirés au hasard chez chacun d'eux, soit hérités d'un ancêtre commun. On a $\phi = \frac{1}{4}\zeta_1 + \frac{1}{2}\zeta_2$.
- $\psi = \zeta_2$ est la probabilité pour que les deux individus partagent deux allèles IBD.

Si IBD = 0, les valeurs $X^{(1)}$ et $X^{(2)}$ d'un facteur mendélien chez les deux individus considérés sont indépendantes ; si IBD = 2, on a $X^{(1)} = X^{(2)}$; et le cas IBD = 1 correspond au cas parent-enfant. On a donc pour les facteurs X_a et X_d

$$\begin{aligned} \text{cov}(X_a^{(1)}, X_a^{(2)}) &= \frac{1}{2}\zeta_1 + \zeta_2 = 2\phi & \text{cov}(X_a^{(1)}, X_d^{(2)}) &= 0 \\ \text{cov}(X_d^{(1)}, X_a^{(2)}) &= 0 & \text{cov}(X_d^{(1)}, X_d^{(2)}) &= \zeta_2 = \psi \end{aligned}$$

Relation	ζ	ψ	ϕ
Parent/enfant	$(0, 1, 0)$	0	$\frac{1}{4}$
Grand-parent/petit-enfant	$(0, \frac{1}{2}, 0)$	0	$\frac{1}{8}$
Germain	$(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	$\frac{1}{4}$	$\frac{1}{4}$
Cousins germains	$(\frac{3}{4}, \frac{1}{4}, 0)$	0	$\frac{1}{16}$

Table 3 – Valeurs de ϕ et ψ pour quelques relations d'apparentement.
(Les germains sont les membres d'une fratrie)

et $X^{(1)}$ et $X^{(2)}$ est

$$\text{cov}(X^{(1)}, X^{(2)}) = 2\phi\alpha^2 + \psi\delta^2.$$

La covariance entre les composantes génétiques chez les deux individus est

$$\text{cov}(G^{(1)}, G^{(2)}) = 2\phi\tau_a + \psi\tau_d,$$

et on a, toujours en supposant l'indépendance des effets environnementaux,

$$\text{cor}(Y^{(1)}, Y^{(2)}) = 2\phi \frac{\tau_a}{\tau_a + \tau_d + \sigma^2} + \psi \frac{\tau_d}{\tau_a + \tau_d + \sigma^2} = 2\phi h^2 + \psi(H^2 - h^2).$$

Études familiales et études de jumeaux

Les héritabilités h^2 et H^2 peuvent donc être estimées à partir de la corrélation observée entre apparentés – par exemple, entre germains d'une part, entre parents et enfants d'autre part. Une des limitations les plus évidentes du modèle de Fisher est l'hypothèse d'absence de corrélation entre les effets environnementaux : si on considère deux apparentés proches, de phénotypes Y^a et Y^b avec

$$Y^a = G^a + E^a$$

$$Y^b = G^b + E^b,$$

en l'absence de corrélation gène-environnement, la covariance de Y^a et Y^b est

$$\text{cov}(Y^a, Y^b) = \text{cov}(G^a, G^b) + \text{cov}(E^a, E^b).$$

Les calculs qui précèdent et les procédures d'estimation de l'héritabilité qui en découlent supposent le terme $\text{cov}(E^a, E^b)$ nul; s'il est positif, ce qui est plausible, l'estimation de $\text{cov}(G^a, G^b)$ est biaisée vers le haut, et celle de l'héritabilité également.

Les études de jumeaux sont la solution classique pour minimiser ce problème. Dans le cas de jumeaux monozygotes, on a $\phi = \frac{1}{2}$ et $\psi = 1$, d'où on tire la valeur de la corrélation phénotypique :

$$r_{MZ} = h^2 + (H^2 - h^2) + \rho_{MZ}$$

où ρ_{MZ} est la corrélation entre les environnements des jumeaux monozygotes. De même, la corrélation phénotypique entre jumeaux dizygotes est

$$r_{DZ} = \frac{1}{2}h^2 + \frac{1}{4}(H^2 - h^2) + \rho_{DZ},$$

où ρ_{DZ} est la corrélation entre les environnements des jumeaux dizygotes. Si on suppose que ces termes de corrélations environnementales sont égaux, $\rho_{MZ} = \rho_{DZ}$, on a

$$2(r_{MZ} - r_{DZ}) = h^2 + \frac{3}{2}(H^2 - h^2).$$

Si on suppose en outre que le terme de dominance est nul, c'est-à-dire $H^2 = h^2$, on a $2(r_{MZ} - r_{DZ}) = h^2$. Cette formule est connue sous le nom de *formule de Falconer*.

Si il y a un terme de dominance non nul, $2(r_{MZ} - r_{DZ}) = H^2 + \frac{1}{2}(H^2 - h^2)$ surestime l'héritabilité large. En outre, l'hypothèse $\rho_{MZ} = \rho_{DZ}$ est très critiquable : il y a beaucoup de raisons, biologiques, comportementales ou sociétales, qui peuvent être cause qu'elle n'est pas vérifiée ; si $\rho_{MZ} > \rho_{DZ}$, on a là encore un biais positif. La présence de biais dus à la présence d'environnement partagé entre apparentés reste donc possible (et même probable quand on s'intéresse à des traits comportementaux ou cognitifs).

4 Postérité

Le modèle polygénique additif de Fisher reste hégémonique dans l'étude des traits quantitatifs. L'héritabilité est notamment très utilisée en génétique animale, et plus spécialement dans le cadre de l'élevage, car elle sert à prédire le succès d'une procédure de sélection artificielle.

L'intérêt historique de l'article de 1918 n'est pas limité à la génétique quantitative : c'est dans cet article que Fisher introduit le terme même de variance pour désigner le carré de l'écart-type. et met pour la première fois l'emphase sur l'intérêt de sa décomposition additive, qui donnera plus tard naissance à l'analyse de variance (dite anova). Dès l'introduction, Fisher met le lecteur en garde contre les interprétations erronées des rapports de variance :

Il est souhaitable d'une part que les idées élémentaires à la base du calcul des corrélations soient clairement comprises, et facilement exprimées dans le langage ordinaire, et d'autre part que les phrases peu rigoureuses sur le « pourcentage de causalité », qui obscurcissent la distinction essentielle entre l'individu et la population, soient soigneusement évitées.

Malgré l'emphase mise ainsi sur la variance, le but de Fisher reste de calculer des coefficients de corrélations, et de retrouver les résultats des biométriciens. C'est Sewall Wright qui le premier notera h le ratio de l'écart-type de l'effet du génotype sur l'écart-type total du phénotype [39, 40], sans envisager précisément le problème de l'additivité des effets. De nos jours, l'héritabilité, définie comme un rapport de variance, parfois hélas comprise comme « un pourcentage de causalité », a éclipsé la question de la corrélation entre apparentés.

La plus importante innovation récente, due à Peter Visscher et son équipe, [41, 42] consiste à estimer une héritabilité à partir de données génétiques couvrant le génome d'un

grand nombre d'individus non apparentés. Très rapidement, voici comment fonctionne la méthode. Il y a sur le génome des millions de sites polymorphes ; si on dispose du génotype de deux individus en plusieurs centaines de milliers, voire plusieurs millions de tels sites, on peut calculer la corrélation génotypique de ces deux individus. Cette corrélation joue le même rôle que le coefficient 2ϕ pour des individus apparentés. On peut alors analyser un grand échantillon d'individus pris au hasard dans la population comme s'il s'agissait de membres d'une seule (immense) famille, en utilisant les corrélations génotypiques comme on utiliserait les coefficients 2ϕ .

Visser et ses collaborateurs ont montré que l'application de cette procédure produit des estimations « plausibles ». Un des avantages avancés est que la méthode permettrait d'obtenir des estimations de l'héritabilité d'un phénotype en s'affranchissant des biais dus à l'environnement partagé dans les familles. Cependant d'autres biais sont possibles, notamment l'existence d'une variation géographique dans les fréquences alléliques : en appliquant cette méthode sans plus de précautions aux coordonnées géographiques du lieu de naissance, on estime que leur héritabilité, dans la population française, est égale à 1, alors qu'il s'agit bien sûr d'une variable purement environnementale, dont l'héritabilité est nulle [6].

De plus, les critiques qui peuvent être faites à la notion d'héritabilité ne s'arrêtent pas à la question de l'environnement partagé : d'autres hypothèses du modèle additif sont très discutables, notamment l'indépendance entre les facteurs génétiques et environnementaux et l'absence d'interactions entre différents gènes ou entre gènes et environnement (les effets de certains facteurs mendéliens peuvent être différents selon l'environnement).

Pour illustrer ce dernier point, et la façon dont il rend difficile l'interprétation de l'héritabilité, on peut considérer l'exemple de la lèpre. L'héritabilité de cette maladie infectieuse a été estimée à $h^2 = 0,71$ [1, 23] dans une population exposée à l'agent qui en est responsable, *Mycobacterium tuberculosis*. Cette valeur élevée montre qu'il existe des facteurs génétiques modulant de façon importante la réponse à l'infection ; on sait pourtant qu'il serait erroné d'en conclure qu'il s'agit d'une « maladie génétique ».

De façon générale, le fait qu'un trait ait une héritabilité élevée n'exclut pas que des facteurs environnementaux jouent un rôle important voire primordial dans la construction de celui-ci. Nombre d'argumentaires à destination du grand public (concernant l'autisme, le QI, etc), voire des chercheurs, sont pourtant fondés sur cette conception erronée.

Bibliographie

- [1] Laurent Abel, Jacques Fellay, David W Haas, Erwin Schurr, Geetha Srikrishna, Michael Urbanowski, Nimisha Chaturvedi, Sudha Srinivasan, Daniel H Johnson, and William R Bishai. Genetics of human susceptibility to active and latent tuberculosis : present knowledge and future perspectives. *The Lancet Infectious Diseases*, 18(3) :e64–e75, 2018.
- [2] William Bateson. *Mendel’s principle of heredity, A defence*. Cambridge University Press, Cambridge, 1902.
- [3] William Bateson. Commonsense in racial problems. *The Eugenics review*, 13(1) :325, 1921.
- [4] Auguste Bravais. *Analyse mathématique sur les probabilités des erreurs de situations d’un point*. Imprimerie royale, Paris, 1844.
- [5] Michael Bulmer. The development of Francis Galton’s ideas on the mechanism of heredity. *Journal of the History of Biology*, 32(2) :263–292, 1999.
- [6] Claire Dandine-Roulland, Céline Bellenguez, Stéphanie Debette, Philippe Amouyel, Emmanuelle Génin, and Hervé Perdry. Accuracy of heritability estimations in presence of hidden population stratification. *Scientific reports*, 6, 2016.
- [7] R A Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2) :399–433, 1919.
- [8] Ronald A Fisher. Has Mendel’s work been rediscovered? *Annals of science*, 1(2) :115–137, 1936.
- [9] Francis Galton. Hereditary talent and character. *Macmillan’s magazine*, 12(157-166) :318–327, 1865.
- [10] Francis Galton. *Hereditary genius*. Macmillan and Company, 1869.
- [11] Francis Galton. Experiments in pangeneses, by breeding from rabbits of a pure variety, into whose circulation blood taken from other varieties had previously been largely transfused. *Proceedings of the Royal Society*, 19 :393–410, 1871.
- [12] Francis Galton. On blood-relationship. *Proceedings of the Royal Society*, 20 :394–402, 1872.
- [13] Francis Galton. On blood-relationship. *Nature*, 6 :173–176, 1872.
- [14] Francis Galton. A theory of heredity. *Contemporary Review*, 27 :80–95, 1875.
- [15] Francis Galton. Family likeness in eye-colour. *Proceedings of the Royal Society of London*, 40(242-245) :402–416, 1886.
- [16] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15 :246–263, 1886.
- [17] Francis Galton. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279) :135–145, 1888.
- [18] Francis Galton. *Natural Inheritance*. Macmillan and Co., London and New York, 1889.
- [19] Francis Galton. The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society of London*, 61(369-377) :401–413, 1897.
- [20] Francis Galton. A diagram of heredity. *Nature*, 57(1474) :293, 1898.
- [21] Francis Galton. Eugenics : Its definition, scope, and aims. *American Journal of Sociology*, 10(1) :1–25, 1904.

Bibliographie

- [22] Francis Galton and James Douglas Hamilton Dickson. Family likeness in stature. *Proceedings of the Royal Society of London*, 40(242-245) :42–73, 1886.
- [23] Annette Jepson, Amanda Fowler, Winston Banya, Mahavir Singh, Steve Bennett, Hilton Whittle, and Adrian VS Hill. Genetic regulation of acquired immune responses to antigens of mycobacterium tuberculosis : a study of twins in West Africa. *Infection and immunity*, 69(6) :3989–3994, 2001.
- [24] Gregor Mendel. Versuche über Pflanzen-Hybriden. *Actes Soc. Hist. Nat. Brünn*, 3 :3–47, 1865.
- [25] A. J. Meston. The Galton Law of Heredity and how breeders may apply it. Published by the author, Pittsfield, Mass., 1898.
- [26] Thomas Hunt Morgan. *Evolution and Genetics*. Princeton University Press, Princeton, 1925.
- [27] Hermann Joseph Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703) :118–138, 1932.
- [28] Karl Pearson. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A.*, 187 :253–318, 1896.
- [29] Karl Pearson. Mathematical contributions to the theory of evolution. On the law of ancestral heredity. *Proceedings of the Royal Society of London*, 62(379-387) :386–412, 1898.
- [30] Karl Pearson. Mathematical contributions to the theory of evolution. On the law of reversion. *Proceedings of the Royal Society of London*, 66(424-433) :140–164, 1900.
- [31] Karl Pearson. *The life, letters and labours of Francis Galton*. Cambridge University Press, Cambridge, 1914.
- [32] Karl Pearson and Alice Lee. On the laws of inheritance in man : I. Inheritance of physical characters. *Biometrika*, 2(4) :357–462, 1903.
- [33] Karl Pearson and Margaret Moul. The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children. *Annals of Human Genetics*, 1(1) :5–54, 1925.
- [34] Ana M Pires and João A Branco. A statistical model to explain the Mendel—Fisher controversy. *Statistical Science*, pages 545–565, 2010.
- [35] James B Reid and John J Ross. Mendel’s genes : Toward a full molecular characterization. *Genetics*, 189(1) :3–10, 2011.
- [36] Richard G Swinburne. Galton’s law—formulation and development. *Annals of science*, 21(1) :15–31, 1965.
- [37] August Weismann. *Die Continuität des Keimplasma’s als Grundlage einer Theorie der Vererbung*. Gustav Fischer, Jena, 1885.
- [38] August Weismann. *Das Keimplasma. Eine Theorie der Vererbung*. Gustav Fischer, Jena, 1892.
- [39] Sewall Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, 6(6) :320–332, 1920.
- [40] Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7) :557–585, 1921.
- [41] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7) :565–569, 2010.
- [42] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA : a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88(1) :76–82, 2011.