B. POUJARDIEU et J. MALLARD*

INRA Station d'Amélioration génétique des animaux
BP 27 31326 Castanet Tolosan Cedex

*INRA-ENSA Laboratoire de Génétique Animale
65 rue de Saint Brieuc 35000 Rennes

Les bases de la génétique quantitative

Les méthodes d'estimation de l'héritabilité et des corrélations génétiques

Résumé. *Pour sélectionner, il est nécessaire de connaître les valeurs de certains paramètres génétiques tels que les héritabilités et les corrélations génétiques. Leur estimation directe ne peut se faire puisque le déterminisme génétique des caractères d'importance économique est infiniment trop complexe pour que l'on puisse observer les valeurs génétiques additives.*
Il est par contre possible d'observer les covariances entre des individus apparentés, dues aux fractions de génome qu'ils possèdent en commun. On peut alors calculer à partir de ces paramètres statistiques les paramètres génétiques recherchés.
Les estimateurs utilisés pour obtenir ces covariances entre apparentés sont complexes. Ils doivent porter sur des populations de très grande taille, et donc réparties sur une aire vaste et hétérogène. De plus, les dispositifs ainsi réalisés ne peuvent être orthogonaux. Les méthodes employées n'ont pas de propriétés statistiques permettant de conclure à l'absolue suprématie de l'une par rapport aux autres. Leur mise en œuvre nécessite des procédures numériques compliquées, gourmandes de ressources informatiques. Mais, au bout du compte, les valeurs estimées donnent satisfaction au sélectionneur.

Pourquoi estimer ces paramètres

Nous sommes placés entre deux mondes bien différents. Celui, imaginaire et conceptuel, de la génétique quantitative permet de décrire la transmission héréditaire de caractères dont le déterminisme, génétique et environnemental, a été décrit. Celui, bien réel, des populations animales chez lesquelles on souhaite améliorer la moyenne de certains caractères que l'on sait présenter une importance économique, mais dont on est loin de connaître le déterminisme génétique.

C'est dans le premier que l'on peut définir sans équivoque la valeur génétique additive d'un individu pour un caractère donné. Rappelons que cette valeur est ainsi construite que la descendance d'un individu aura en espérance (= en moyenne sur tous les descendants possibles issus de tous les accouplements possibles pour cet individu et répartis dans toutes les conditions possibles du milieu) la moitié de cette valeur additive.

On conçoit aisément que c'est dans le second que ce paramètre a un intérêt opérationnel considérable. Retenir comme reproducteur celui qui a la plus grande valeur génétique additive rend maximales les chances d'avoir une descendance la plus productive possible.

Malheureusement, les caractères économiquement intéressants sont extrêmement synthétiques : la croissance, la viabilité embryonnaire, la qualité de la viande, etc. Les métabolismes qui assurent leur expression sont extraordinairement complexes, interconnectés et globalement régulés, mettant en jeu un nombre de gènes inconnus, mais sans doute très grand, interagissant entre eux. La belle modélisation du fonctionnement du petit opéron lactose de *E. Coli*, réalisée par Chevalet, illustre bien la complexité du problème. Il faut donc renoncer au rêve d'observer les valeurs génétiques additives vraies dans une population réelle, faute de pouvoir ne serait-ce qu'écrire le déterminisme génétique sous jacent. Ne parlons pas de son éventuel traitement.

Faute de pouvoir les calculer exactement, nous sommes contraints d'en rechercher des valeurs approchées, que nous appellerons des index de sélection. Il s'agit d'utiliser des mesures effectuées sur les individus constituant la population pour réaliser une prédiction de cette valeur. Nous avons vu dans l'article de Mallard qu'il était possible de prédire une valeur A inconnue, par une mesure C statistiquement liée, à condition que l'on connaisse les différents moments d'ordre 2 de la distribution conjointe de ces deux variables (pour la compréhension de cette phrase, se reporter au texte de Mallard).

La mise en oeuvre de l'indexation suppose ainsi que l'on connaisse toute une série de variances et de covariances, et en particulier les paramètres génétiques évoqués dans les articles de Minvielle et Ollivier. Les plus importants sont les héritabilités des caractères et les corrélations génétiques qui les lient, mais les variances de dominance et d'épistasie sont également nécessaires, surtout quand on s'intéresse au résultat du croisement entre populations.

Ces statistiques sont bien évidemment inconnues, de sorte qu'il faut se débrouiller pour en obtenir un ordre de grandeur aussi "ressemblant" que possible. En termes plus orthodoxes, il faut les estimer.

Nous prendrons l'exemple de σ_A^2, variance génétique additive, qui constitue le numérateur de l'héritabilité ($h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2)$)pour exploiter la démarche conceptuelle qui conduit à son estimation.

1 / Principe de l'estimation

1.1 / Paramètres génétiques et covariances entre apparentés

Les valeurs génétiques A ont été créées dans l'univers théorique de la génétique quantitative. Personne ne peut se targuer d'en avoir observé une dans la réalité : l'estimation directe de sa variance n'est pas possible. Par contre, Minvielle a démontré que, sous certaines hypothèses génétiques, les covariances entre certains types d'apparentés étaient des fonctions simples - linéaires - des diverses variances génétiques. Rappelons en particulier que :

$CovPF = \sigma_A^2/2$
$CovDF = \sigma_A^2/4$
$CovF = \sigma_A^2/2 + \sigma_D^2/4$
$CovMF = \sigma_A^2/2$

La notation CovPF, condensé de covariance entre un père et un fils, représentant la covariance entre la mesure d'un caractère chez un individu et la mesure du même caractère réalisé chez son fils. DF, F, MF représentent respectivement des couples de demi-frères, de frères et mère-fille.

Il est donc équivalent d'estimer les variances génétiques ou les covariances entre apparentés :

$\sigma_A^2 = 2\ CovPF = 2\ CovMF = 4\ CovDF$
$\sigma_D^2 = 2\ (CovF - CovDF)$ (1)

1.2 / Principe de l'estimation

Les populations animales réelles sont structurées par un réseau très dense d'apparentements. Tout individu a un père, il est bien rare qu'il n'ait pas - surtout chez les espèces polytoques - des frères, des demi-frères, etc. On peut donc sans peine extraire de la population un grand nombre de couples d'apparentés d'un type donné. Sur la figure 1, on peut voir le nuage des points représentant les valeurs conjointes des mesures des couples père-fils dans une population. CovPF exprime l'aplatissement du nuage de points (voir l'article de Mallard). Ces données renferment donc toute l'information suffisante pour réaliser une estimation de ce paramètre (nous discuterons dans la partie suivante des problèmes statistiques posés).

Admettons que nous disposions de l'estimation de diverses covariances entre apparentés dans une population réelle. Si nous supposons que le déterminisme génétique du caractère considéré est exactement celui dont le développement a conduit aux formules (1), nous en déduisons les estimations des paramètres génétiques dont nous avons besoin. C'est à ce niveau que se réalise la jonction entre les modèles théoriques et la réalité des populations animales.

1.3 / Retour sur les hypothèses sous-jacentes

Il reste bien sûr en suspens le Si. Nous avons souligné dans l'introduction qu'il est bien peu réaliste, tant les modèles les plus complexes de génétique quantitative paraissent simplistes face à la réalité biologique. Mais ces formules (1) vont de surcroît nous permettre de porter un jugement sur le bien fondé de cette hypothèse. On y lit par exemple que $CovPF = CovMF = 2\ CovDF$

Si les estimations de ces quatre termes sont proches, on en conclura que cette hypothèse, pour aussi irréaliste qu'elle soit, rend compte de façon satisfaisante de la réalité observée. La première égalité par exemple indique que le fait de n'avoir inclus que des gènes autosomaux dans le modèle est a posteriori acceptable. De même, la comparaison de CovF et 2CovDF permettra de porter un certain jugement sur l'importance des effets de dominance (σ_D^2, DF représente des demi-frères ayant une mère en commun).

En définitive, la pratique de la sélection a depuis longtemps tranché. Toute la sélection est basée sur l'acceptation des hypothèses de la génétique quantitative. Or, cette sélection est efficace et les simulations montrent qu'il y aurait peu à gagner - en efficacité de la sélection - à utiliser des modèles plus complexes, conduisant à des développements extrêmement lourds.

Cela ne veut bien sûr pas dire que, dans la réalité, les 10^5 gènes d'un organisme additionnent des effets infinitésimaux, les phénomènes d'interaction entre loci restant négligeables. On sait bien que c'est faux. Mais globalement, sur un plan statistique, la résultante de l'extrême complexité biologique est peu différente - à l'intérieur d'une population donnée - des conséquences du modèle polygénique.

Cette approximation est un peu de même nature que l'assimilation d'une courbe quelconque à sa tangente au voisinage du point de tangence. En un autre point, une approximation linéaire pourra également être définie, mais la tangente sera différente. De même, on retrouvera des phénomènes d'interaction - dominance et épistasie - quand on s'intéressera à plusieurs populations, notamment au travers du croisement (voir l'article de Brun).

2 / Les problèmes statistiques posés

Nous venons de démontrer qu'il "suffit" d'estimer des covariances entre apparentés. Prenons le cas d'une covariance père-fils. Trois problèmes d'ordre statistique se conjuguent pour rendre inextricable ce "il suffit" : la précision des estimations, la non indépendance des données, le déséquilibre.

2.1 / Précision des estimateurs

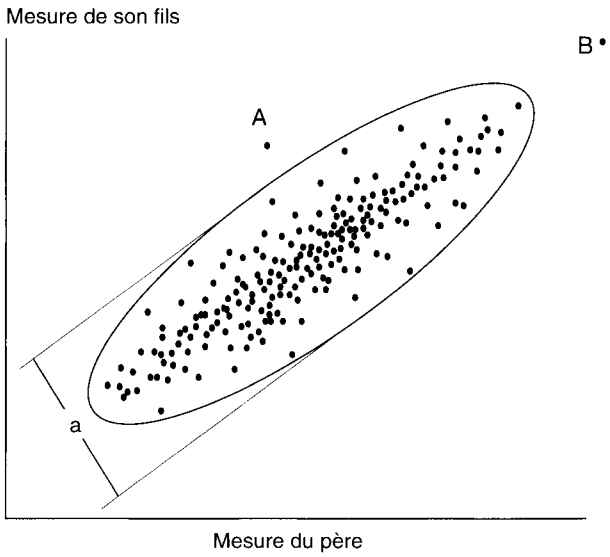
Appelons N le nombre d'objets utilisables pour estimer dans la population le paramètre statistique recherché. Ce sera par exemple le nombre de mesures réalisées si l'on désire connaître la variance ou le nombre de couples représentés sur la figure 1 pour estimer CovPF.

N est forcément fini, de sorte que l'estimation tirée de cet échantillon limité est imparfaite. Si nous jouons à pile ou face - avec une pièce non truquée - la

probabilité de pile est de 1/2 ; mais si nous effectuons seulement N = 9 lancers, le pourcentage de cas où pile apparaît n'est pas de 1/2 exactement. Nous savons bien par contre que si N devient très grand, une valeur différente de 1/2 nous fera suspecter l'honnêteté de la pièce ou du lanceur.

De même, imaginons que, sur la figure 1, nous ayons retrouvé un couple père-fils - noté A - oublié dans la population. Son report va diminuer l'aplatissement du nuage et donc l'estimation de la covariance. B aurait eu l'effet inverse.

Figure 1. *Principe de l'estimation d'une covariance entre apparentés. Chaque point représente un couple de mesures d'un père et de son fils. La covariance mesure l'étirement du nuage de points, elle est inversement liée à l'épaisseur "a".*



Cette incertitude sur l'estimation, liée à un échantillon de taille limitée, est bien connue quand il s'agit de l'estimation d'une moyenne. Nous allons la matérialiser par l'exemple suivant. Nous disposons d'une population infinie dans laquelle l'histogramme d'une mesure est centrée sur 10 et a un écart type de 10 (elle est de plus supposée normale dans ce cas). Nous extrayons un échantillon de taille limitée (N = 40) d'individus indépendants et estimons moyenne et variance. Nous trouverons des valeurs différentes de 10. Recommençons avec un autre échantillon de même taille : le hasard dans la constitution de ce nouvel échantillon donnera des valeurs différentes. Sur la figure 2 ont été reportés deux petits segments qui renferment toutes les valeurs trouvées comme estimation de la moyenne et de la variance. Plus exactement, ce segment renferme 95 % de toutes ces estimations : on a continué de considérer que 5 % constituent une quantité de cas négligeable en statistique, de sorte qu'on assimile 95 % et "toutes les valeurs". La longueur de ces segments porte le nom d'intervalle de confiance des estimations de la moyenne et de la variance. Plus N est grand, plus ils sont petits.

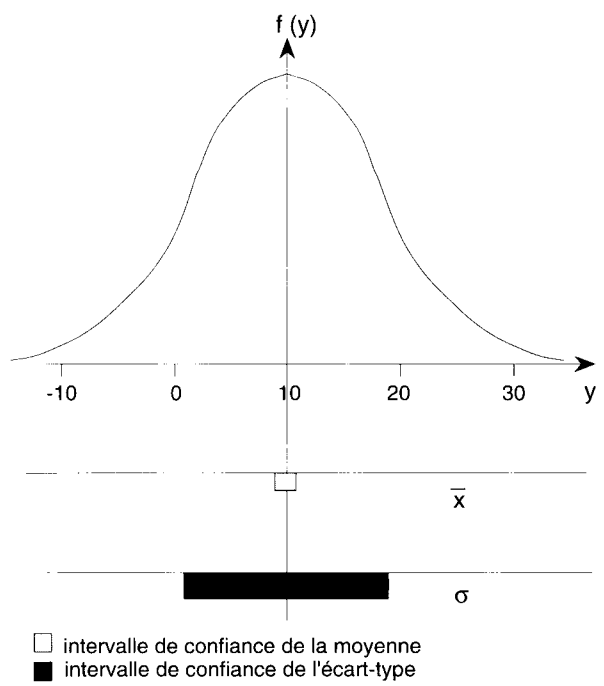
L'intervalle de confiance d'une moyenne est une notion qui nous est familière. On voit sur cet exemple que pour N = 40 il est raisonnablement réduit (tout expérimentateur raisonnable s'estimera bien doté si chaque case élémentaire de son dispositif atteint cet effectif). Par contre cet exemple montre que les moments d'ordre 2 sont beaucoup plus difficiles à estimer. Il faudrait plus de 400 données pour que l'intervalle de confiance de σ atteigne la taille de celui de la moyenne de 40 données.

Cela se comprend intuitivement : créons le nuage de la figure 1 en accumulant un à un des points supplémentaires : on voit rapidement se dessiner une masse centrale, où la plupart des nouveaux points s'accumule. Il est aisé de déterminer "à l'oeil" une valeur centrale qui très rapidement se stabilise. Quelques points éloignés de ce centre apparaissent de temps en temps, constituant un halo. Les moments d'ordre 2 chiffrent l'étendue de ce halo. On conçoit qu'il faudra disposer d'un contour précis pour cela, donc de beaucoup de points dans cette zone floue. Or justement, ce sont des cas rares qui amènent des points dans ces parages.

En résumé, pour obtenir des estimations pas trop imprécises des variances et des covariances entre apparentés, il faudra disposer d'un nombre de répétitions considérable. Par exemple, il faudra sur la figure 1 plusieurs centaines de couples père-fils pour obtenir une valeur acceptable de CovPF. Et en fait, on recherche des effectifs de plusieurs milliers. Rappelons enfin que héritabilités et corrélations génétiques sont des rapports, les incertitudes d'estimation sur les numérateurs et dénominateurs se cumulant pour rendre leurs estimations encore plus imprécises.

L'estimation des paramètres génétiques nécessite donc l'accumulation d'un jeu très important de données donc de gros fichiers et une grosse puissance de calcul.

Figure 2. *Longueur des intervalles de confiance pour l'estimation d'une moyenne et d'un écart-type. L'estimation porte sur un échantillon de 40 données y, tirées de la population de moyenne 10 et d'écart-type 10 représentée au-dessus.*



2.2 / La non indépendance des données

Pour estimer la variance dans le cas illustré par la figure 2, nous avons utilisé un estimateur, c'est-à-dire une formule, encore relativement bien connue où Σ représente une sommation sur l'ensemble de l'échantillon, x la mesure réalisée sur chaque individu.

$$\sigma^2 = \frac{\Sigma x^2 - (\Sigma x)^2 / N}{N - 1} \qquad (2)$$

Dans les cas très simples où tous les objets utilisés pour l'estimation (individus si on estime une variance, couples de données pour les covariances) sont indépendants entre eux, cette formule s'impose comme la meilleure possible parmi tous les estimateurs (quadratiques) imaginables. Elle procure des estimations de variance minimum, c'est-à-dire une longueur aussi petite que possible à l'intervalle de confiance (le segment cd de la figure 2). Cet estimateur est non biaisé, c'est-à-dire que la vraie valeur se situe exactement au milieu du segment. En d'autres termes, les incertitudes liées à l'échantillonnage d'un groupe limité d'objets jouent également dans les deux sens (remarque : cette propriété de non biais paraît en général secondaire à l'utilisateur mais revêt une grande importance aux yeux de la statistique classique... peut être parce qu'elle est la contrainte la plus facile à prendre en compte dans la construction des estimateurs).

Malheureusement cette condition d'indépendance des données n'est jamais respectée. N'avons-nous pas dit que les populations animales étaient structurées par des liens de parenté ? Si un individu dispose d'un génome exceptionnel et donc exprime un niveau intéressant du caractère, son frère, qui possède la moitié de ses gènes, a de fortes chances d'être lui aussi doté de bonnes performances. Ces mesures de ces deux individus apparentés ne sont pas indépendantes puisque la connaissance de la valeur de l'un donne déjà une idée de ce que sera la valeur de l'autre. N'a-t-on pas créé les formules "tel père tel fils", "se ressembler comme des frères" ou "comme des jumeaux" ?

Le milieu d'exploitation des animaux crée lui aussi son petit réseau de dépendance entre les données. Reprenons l'exemple de la figure 2 et imaginons que les 40 objets d'un échantillon soient divisés arbitrairement en 2 sous-groupes de 20. Chacun de ces lots est élevé dans deux bâtiments différents. Autant le premier est luxueux, autant le second est froid, venteux, sale, placé en bordure d'autoroute. Tout individu de ce deuxième lot aura des performances très inférieures. La variance au sein de l'échantillon s'est vue augmentée par l'application de ces deux traitements différents ; on parle d'un "effet bâtiment".

Or, du paragraphe 2.1 nous avons retenu la nécessité de disposer d'effectifs considérables, qu'il est particulièrement utopique d'imaginer élever dans des milieux rigoureusement identiques. Les bovins, par exemple, sont répartis dans une poussière d'élevages, les poulets regroupés en cases différentes selon leur date de naissance, etc.

La formule (2) est donc particulièrement inadaptée à son usage en génétique quantitative. Illustrons cela par un exemple : pour estimer une variance génétique additive, nous disposons de deux jumeaux (leur génome est pratiquement identique). La formule 2 intégrera les deux informations indépendantes (deux degrés de liberté). En fait, il n'y a qu'une seule valeur génétique (répétée deux fois) : le calcul d'une variance n'a aucun sens. La formule utilisée aurait dû tenir compte d'un coefficient de parenté de 1/2 entre ces jumeaux pour ne les considérer comme porteurs que d'une seule information (donc ne leur donner qu'un degré de liberté collectivement).

Pour résoudre cette difficulté, nous allons décrire les données par un modèle et utiliser les estimations qui lui sont adaptés.

Ecrire un modèle signifie exprimer toutes les dépendances, qu'elles soient dues au milieu ou à la génétique, de chaque individu avec tous les autres : l'individu A est frère de B et C et non apparenté à D. Par contre, il a été élevé dans le même bâtiment que C et D mais pas B. Cette description par le menu devient vite inextricable, de sorte que l'on utilise une sorte de langage codé nommé modèle linéaire.

On constitue des sous-populations d'individus parfaitement homogènes quant aux relations existant entre les individus. Nommons-les cases élémentaires. A l'intérieur d'une case, les individus ne sont pas identiques, leur variation permettant de chiffrer la variance résiduelle. On réalise ensuite des regroupements de ces cases en ensembles plus vastes homogènes quant à leurs relations entre eux. On continue jusqu'à ce que toutes les relations aient été décrites.

Supposons par exemple qu'en aviculture, deux familles A et B issues de deux mères différentes soient réparties dans deux bâtiments différents X et Y. La partie de famille A élevée dans X (AX) est une case élémentaire. De même que AY, BX et BY. A l'intérieur de AX tous les individus sont frères, élevés dans un même milieu. Les cases AX et AY sont reliées par le fait que tous les individus ont la même mère : on dit qu'ils partagent le même niveau du facteur mère. AX et BY partagent le même bâtiment, c'est-à-dire ont le même niveau du facteur bâtiment.

On utilise enfin une équation symbolique pour décrire toutes ces relations. C'est le modèle stricto sensu. On écrira :
U_{AX} Jules = Effet mère A + Effet bâtiment X + Résiduelle de Jules

On décrit ainsi la valeur U de l'individu Jules, descendant de la mère A et placé dans le bâtiment X, comme la somme des effets des facteurs que nous avons identifié. Deux individus, Jules et Arthur partageant le même bâtiment X verront apparaître dans cette équation la même valeur de l'effet bâtiment X, ce qui matérialise la dépendance de ces deux données. A l'opposé, les résiduelles de deux individus appartenant à la même case AX différeront partiellement du fait que leur résiduelle est différente.

Pour finir, on écrit plus prosaïquement l'équation ci-dessus avec un jeu d'indices.
$U_{ijk} = m + M_i + B_j + Z_{ijk}$

où i et j nomment les niveaux des différents facteurs, mère et bâtiment, et k numérote les individus intra case. ijk = 115 représenterait ainsi Jules, le 5ème descendant de la 1ère mère (A) placé dans le premier bâtiment (X). Et on résume cette écriture en disant que les données sont organisées selon un dispositif factoriel croisant les effets des facteurs mère (2 niveaux) et bâtiment (2 niveaux).

Ayant ainsi décrit la structuration des données, il

ne reste plus qu'à choisir, parmi toutes les formules (on se limite à la classe des estimations quadratiques) incorporant l'ensemble des données, celle qui est la mieux adaptée à l'estimation du paramètre recherché compte tenu de cette présentation particulière des données.

Il est facile d'écrire la forme la plus générale des fonctions quadratiques des données. Osons le faire pour l'estimation de la covariance père-fils : $Y_P'AY_F$ où Y_P' et Y_F sont les listes de toutes les mesures réalisées des pères et des fils et A un tableau de valeurs numériques comprenant autant de lignes et de colonnes qu'il y a de couples père-fils. Pour un jeu de valeurs remplissant A, on obtient l'une des fonctions quadratiques possibles des données. On voit à l'évidence qu'il y en a beaucoup : N^2 si N est le nombre de couples père fils... soit 10^6 pour un effectif raisonnable de 1 000 couples !

Il est encore facile d'écrire les équations qui permettent de rechercher la valeur de A donnant le meilleur estimateur. On demandera que cet estimateur soit non biaisé, de variance minimum. Finalement il est facile de trouver la formule donnant la valeur A_0 recherchée. C'est l'affaire de quelques lignes de calcul matriciel.

Il faut à ce point distinguer deux cas, selon que le dispositif est "orthogonal" ou non. Imaginons que dans chaque cellule élémentaire du modèle, l'effectif d'individus soit égal et qu'il y ait autant de niveaux d'un facteur croisé dans chacun des niveaux de l'autre. Ce dispositif est dit équilibré ou "orthogonal". Dans ce cas, les formules qui donnent la valeur A_0 sont particulièrement simples (disons plutôt que l'on peut en trouver d'infiniment pires). Elles ont été développées pour chacun des types de modèles que l'on peut rencontrer et des paramètres statistiques dont on souhaite l'estimation. Les programmes correspondants sont aisément accessibles dans les grandes bibliothèques statistiques (SAS...).
Hélas, ce cas idyllique ne se rencontre pas en génétique.

2.3 / Le cas des dispositifs non orthogonaux

Dans certains cas, le déséquilibre peut être peu important. Un individu qui meurt dans une case élémentaire de 40 n'oblige pas à bouleverser la forme de l'estimateur, c'est-à-dire la valeur de A_0. On va donc utiliser les mêmes types de formules, tout en sachant que, si les estimateurs restent non biaisés, leur variance n'est plus tout à fait minimum.

Par contre, dans le cas où la moitié de certaines cases disparaît, ou certaines autres entières ne sont pas réalisées, la médiocrité de ces estimateurs devient insupportable. Cela se produit très facilement quand les effectifs N sont petits, c'est-à-dire chez les animaux peu prolifiques (familles d'apparentés réduites) et de grosse taille (effectifs réduits des unités de production). Si N = 2, un seul petit mort divise par 2 l'effectif ; et avec un peu de malchance, la case entière disparaît.

On est alors amené à rechercher directement la valeur de A_0. En écriture matricielle, elle n'est pas plus compliquée. Mais on n'y retrouve plus dans leur développement les simplifications qu'apportait l'équilibre. Les calculs deviennent rapidement irréalisables (avec une précision numérique acceptable) directement sur les ordinateurs pourtant puissants dont nous disposons. On a recours à des procédures de résolutions itératives qui posent des problèmes de précision et de temps de calcul entre lesquels on doit trouver des compromis.

Mais le plus extraordinaire est que ces formules conduisant à la valeur de A_0 contiennent la vraie valeur du paramètre que l'on cherche à estimer. Comprenons nous bien : nous disposons des données de la figure 1 pour estimer le paramètre CovPF. Elles sont réparties dans un dispositif non orthogonal. Nous recherchons la formule $Y_P'AY_F$ qui permettra d'utiliser ces couples de données père-fils pour réaliser la meilleure estimation du paramètre CovPF. Et le calcul nous dit "dis-moi quelle est la valeur de CovPF et je te dirai alors quelle est la meilleure forme de A". C'est le serpent qui se mord la queue.

Il n'existe plus dans ce cas d'estimateur optimal mais seulement des localement optimaux. A chaque valeur des vrais paramètres (ceux que l'on cherche à estimer) correspond un estimateur adapté.

Comme la vraie valeur est par essence inconnue, on fournira au calcul un ordre de grandeur bâti à partir de données bibliographiques, d'estimations réalisées dans des conditions comparables faites sur des populations du même genre. On utilise souvent des estimations réalisées lors de générations antérieures.

L'expérience et des simulations montrent que l'écart à l'optimalité ainsi introduit reste négligeable tant que l'erreur n'est pas trop importante. On peut en général se tromper du simple au double sans que la variance de l'estimateur n'augmente plus que de quelques pourcents. On dit que cette méthode est très robuste.

On est tout de suite tenté par l'itération de cette méthode : en injectant une valeur grossièrement approchée, on a bâti un estimateur presque optimal qui fournit une valeur estimée, plus proche vraisemblablement de la vraie valeur. On va recommencer le même calcul en injectant cette première estimation comme idée a priori.

On doit pouvoir comprendre intinctivement que le A ainsi construit dépend des données Y_P' et Y_F et donc que $Y_P'AY_F$ n'est plus une simple fonction quadratique des données. De sorte que dès la première itération, l'estimation perd toutes les propriétés dont on l'avait dotées : il n'est plus ni quadratique, ni non biaisé, ni de variance minimum.

Il a par contre la propriété de converger (même si théoriquement des statisticiens grincheux s'irritent de cette aptitude non démontrée théoriquement). C'est-à-dire qu'après un nombre réduit d'itérations, l'estimation devient quasiment égale à l'idée a priori injectée. C'est une situation intinctivement satisfaisante pour l'utilisateur, mais qui n'a guère de propriétés statistiques.

On a pu par contre démontrer que ce point stationnaire est identique à celui issu d'une autre école de pensée de l'estimation, très différente de celle des moindres carrés (on a recherché une variance minimum de l'estimateur) : celle du maximum de vraisemblance. C'est démontré. Cela laisse le statisticien perplexe. Nous ne développerons pas ces concepts trop éloignés de notre sujet. Nous l'avons cité pour montrer que les statistiques restent un monde mal maîtrisé sur ses confins puisqu'en réalisant des opé-

rations peu licites (l'itération) à partir d'une certaine conceptualisation, on en rejoint une autre, radicalement différente et qui lui a été longtemps farouchement opposée.

2.4 / Le déséquilibre

C'est dans la structuration par le réseau des apparentements que le comble du déséquilibre est atteint. En première approximation, les ancêtres communs, ne remontent pas à plus de une ou deux générations. Le nombre d'apparentements est limité (frères, père-fils, cousins...). Mais dans les faits, on rencontrera des 1/2 cousins à la mode de Bretagne et autres liens de parenté trops complexes pour seulement porter un nom.

Dans ces conditions, le nombre d'apparentements devient prohibitif, et on ne pourra estimer toutes les covariances entre apparentés. D'autre part, un couple d'apparentés d'un type donné ne sera plus représenté que par un effectif très réduit. Si par exemple, sur 1000 couples père-fils, la moitié est issue de la même mère (fils issus d'un accouplement mère-fils), une autre partie est en plus 1/2 oncle-neveu ...etc, il restera très peu de couples vraiment père-fils. Pas assez en tout cas pour une estimation correcte.

On doit adopter alors une modélisation totalement différente ; mais qui a l'inconvénient d'introduire beaucoup plus tôt le modèle de la génétique quantitative dans la modélisation statistique.

Toute relation de parenté entre deux individus peut être parfaitement décrite au niveau d'un locus par un jeu de 15 paramètres, les coefficients d'identité (ce nombre devient gigantesque quand on s'intéresse à plusieurs loci). Dans les cas où les individus ne sont pas consanguins (la pratique de lélevage évite au maximum l'accouplement d'apparentés trop proches), on peut se contenter d'un seul coefficient : le coefficient de parenté. Si, de plus, on suppose que le caractère étudié a un déterminisme génétique purement additif (variances de dominance et d'épistasie nulles), les covariances CovApp entre deux animaux liés par une parenté ΦApp quelconque s'écrivent
$\text{CovApp} = 2\ \Phi\text{App}\ \sigma_A^2$

Si l'on admet toutes ces hypothèses génétiques avant d'écrire le modèle statistique décrivant les données, on gardera l'écriture classique pour expliciter la présence d'effets du milieu. Mais la structure de parenté sera décrite par la seule matrice de parenté (tableau incluant tous les coefficients de parenté de chaque individu avec tous les autres). Chaque individu constitue de ce fait une cellule élémentaire à lui tout seul, et cette matrice décrit toutes les liaisons.

Les problèmes statistiques et calculatoires posés par cette autre conception ne sont ni pires ni différents en nature que ceux déjà évoqués. La seule différence est qu'on n'obtient pas des estimations des covariances entre apparentés, mais les valeurs qu'elles auraient si les hypothèses génétiques admises (1 locus, pas de consanguinité, additivité pure) étaient respectées. A l'inverse, cela évite de négliger au niveau du modèle toutes sortes d'apparentements mineurs qui viendraient par trop compliquer une modélisation classique.

Conclusion

Dans la réalité, les dispositifs destinés à l'estimation de paramètres statistiques sont complexes. Ils portent obligatoirement sur des effectifs importants, de sorte que l'on ne peut réduire l'hétérogénéité du milieu où ils sont implantés. Enfin, il n'est pas possible, même en triant un sous-ensemble des données, d'obtenir un dispositif orthogonal.

Les estimateurs les plus généraux adaptés à ces cas complexes sont théoriquement connus depuis un quart de siècle. Ils sont pourtant bien déconcertants : ils ne sont que localement optimaux, ce qui conduit à itérer leur calcul.... mais en perdant leurs propriétés d'optimalité, et en en retrouvant d'autres bien différentes. Les valeurs ainsi obtenues sont cependant considérées par les utilisateurs comme d'excellentes estimations, les plus dignes de confiance.

Enfin, les calculs (que nous avons escamotés) de ces estimateurs sont extrêmement lourds. La résolution directe suppose l'inversion de matrices ayant autant de lignes et de colonnes qu'il y a d'individus (N) dans la population. Et nous avons dit qu'il y en avait obligatoirement beaucoup. En pratique, on doit recourir à des solutions numériques, très souvent itératives, qui cherchent à réaliser un compromis entre temps de calcul, volume de mémoire utilisée, précision numérique. C'est encore, et sans doute pour longtemps malgré la fantastique augmentation de puissance des ordinateurs, un domaine de recherche ouvert où beaucoup reste à faire.

Malgré toute cette panoplie statistique et informatique, malgré des tailles de populations qui atteignent parfois plusieurs millions, les estimateurs des moments d'ordre 2 des loci restent imprécis. Il n'est pas rare que l'intervalle de confiance d'une héritabilité atteigne 10 % de sa valeur. On en verra des exemples dans les articles suivants.

L'usage montre cependant que cette précision est suffisante. Les paramètres génétiques que l'on calcule à partir des covariances entre apparentés observées et estimées permettent le calcul d'index remarquablement efficaces permettant d'accumuler un progrès génétique conséquent.