*J. MALLARD*

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

## Les bases de la génétique quantitative

# Populations et variabilité

***Résumé.*** *Cet article présente de façon aussi imagée et peu calculatoire que possible un certain nombre de concepts de base de statistique. Sont définies successivement les notions de population et de distribution des caractères, de moyenne, de variance et de corrélation. Le principe de la prédiction par régression est développé, ainsi que celui de techniques de représentation de distributions multidimensionnelles.*

Cet article vise à rappeler que Généticiens Moléculaires et Quantitatifs vivent dans deux mondes de pensée fort différents. L'un, mécaniste et déterminé, l'autre, statistique et relatif. La différence est d'autant plus insidieuse que, bien souvent, les mêmes mots désignent des concepts essentiellement différents.

Cette incompatibilité d'humeur peut, par exemple, être illustrée par la notion de gène. Ce mot si familier représente pour les uns une molécule de composition connue, avec des zones spécalisées assurant des fonctions précises. Cela s'isole, se synthétise, se bricole, peut se manger si vraiment on a faim. Pour le quantitatif, c'est un nom donné à l'interprétation de certains types de ségrégations de caractéristiques observables, à l'intérieur de la descendance de certaines familles (les lois de Mendel). Ce sont des notions très éloignées. Songez au travail acharné (5 ans de compétition mondiale) qu'a représenté l'isolement moléculaire du gène de la mucoviscidose, pourtant connu depuis longtemps par sa transmission mendélienne. Au passage, soulignons un point qui sera développé plus loin : le généticien quantitatif a eu besoin d'examiner plusieurs individus (la descendance d'un couple par exemple) pour définir son concept de gène, alors qu'un seul génome suffit à caractériser une configuration moléculaire. En d'autres termes, un Généticien Quantitatif est avant tout un statisticien.

La statistique est à la fois très simple (encore qu'un peu abstraite) au niveau des concepts manipulés, et très compliquée dans leur formulation mathématique. En général, c'est le très compliqué qui s'impose à l'imaginaire du Biologiste : des pages de signes cabalistiques, des grands prêtres au front plissé, des piles branlantes de listings jaunis.

Mais l'informatique fournit des programmes extrêmement conviviaux qui intègrent les formules et procédures de calcul les plus complexes de façon parfaitement transparente pour l'utilisateur . Il devient de ce fait bien plus important d'appréhender les concepts, de savoir formuler ses questions et interpréter les résultats des calculs.

## 1 / La notion de population

L'objet manipulé par la statistique est relativement abstrait : la population. C'est une collection d'objets, pas forcément identiques, mais que l'on accepte de réunir en fonction d'une caractéristique qu'ils ont en commun. On peut concevoir par exemple la population des habitants d'une ville, ou celle des porcs vivant en Bretagne. Deux points sont importants dans cette définition.

Il s'agit d'une collection. Elle doit contenir au moins deux objets pour que la description qui en sera donnée soit possible, et en fait un très grand nombre pour qu'elle soit précise. Et c'est la collection qui est décrite, pas l'un ou l'autre des individus constituants.

Bien sûr, dans une deuxième étape, on pourra attribuer à l'un des objets constituant la population les caractéristiques de cette population. Par exemple, si nous admettons que la population des vaches, en France, produit entre 3 et 10 000 litres de lait par an, et que, au détour d'un chemin, nous rencontrons un paysan et sa vache, il vaut mieux éviter, dans une éventuelle conversation, de parler de 500 litres.

Ces objets sont réunis parce que nous leur avons reconnu une caractéristique commune, qui revêt à nos yeux un sens.Il faut bien voir que cette définition est arbitraire et capitale.

C'est l'utilisateur qui décide, en fonction de ses idées et du but qu'il poursuit, des contours de la population. Il est tout aussi légitime de parler de la population des vaches Pie-Noires Bretonnes ou des vaches Pie-Noires de France ou des vaches dans le monde. Ce ne sont pas les mêmes personnes qui le feront (un nostalgique bretonnant, un technocrate

ministériel, un haut cadre de la FAO), ni pour les mêmes finalités.

Et pourtant, c'est une décision fondamentale. Si on refuse de réunir deux individus dans une même population, on s'interdit de les comparer. Peut-on dire qu'un chou est préférable à un navet ? Non, sauf, par exemple, à les considérer tous deux comme des sources d'un certain élément nutritif (voilà leur caractéristique commune) dont l'un s'avèrera plus riche.

Il y a dix ans, coexistaient en France deux types de vaches Pie-Noires : les FFPN, bien de chez nous, et des croisées Holstein résultant de l'importation massive de reproducteurs et de semence d'Amérique du Nord. Question en apparence futile : s'agit-il de la même population ; va-t-on réunir les livres généalogiques et les fichiers de contrôle de performances ; va-t-on les indexer ensemble ?

Garder deux populations conduit à demander à un éleveur à quelle race doit appartenir le taureau qu'il recherche, puis à lui proposer les meilleurs géniteurs. En cas de population unique, ce choix préalable n'est pas proposé. La comparaison des niveaux génétiques met en concurrence tout le monde, ce qui fait apparaître - compte tenu des critères retenus - la très nette supériorité du type Holstein. Les gens qui ont opté pour la population unique ont sciemment accéléré l'Holsteinisation du cheptel Français.

# 2 / Représentation monodimensionnelle d'une population

## 2.1 / L'histogramme

Supposons qu'une même mesure (nommée C sur la figure 1) soit effectuée sur tous les individus d'une population. En dehors de la caractéristique commune qui les inclut dans la population, ils n'ont aucune raison d'avoir la même valeur de c. Nous représentons cette collection de mesures par un dessin. L'abscisse représente les valeurs numériques obtenues par C.

L'ordonnée, $f(c_0)$, est la fréquence (le nombre d'individus présentant cette caractéristique relativement au nombre total d'individus) avec laquelle on observe une valeur $C_0$ donnée dans la population. L'histogramme que l'on obtient ainsi dépeint complètement la population sous le rapport de la mesure c. Un dessin n'est pas facile à manipuler. On va le décrire avec des concepts, que l'on apelle les moments...dont les premiers sont d'ailleurs parfaitement intuitifs.

## 2.2/ Le moment d'ordre 1 : la moyenne

Parlons des Pygmées. La taille des Pygmées, c'est-à-dire l'ensemble des tailles de tous ces gens, est représentée par leur histogramme. Mais si je demande quelle est la taille d'un Pygmée, on me dira qu'en gros, ils mesurent autour d'un mètre. Nous venons de redécouvrir la notion d'ordre de grandeur, de valeur la plus représentative de toutes les autres.

La quasi totalité des gens s'accorde à la quantifier par la moyenne des observations. Ce n'est pas la seule possibilité. Il existe par exemple des statistiques "non paramétriques", où c'est la médiane (valeur telle qu'il y ait autant d'observations de part et d'autre) qui matérialise cette notion de valeur centrale.

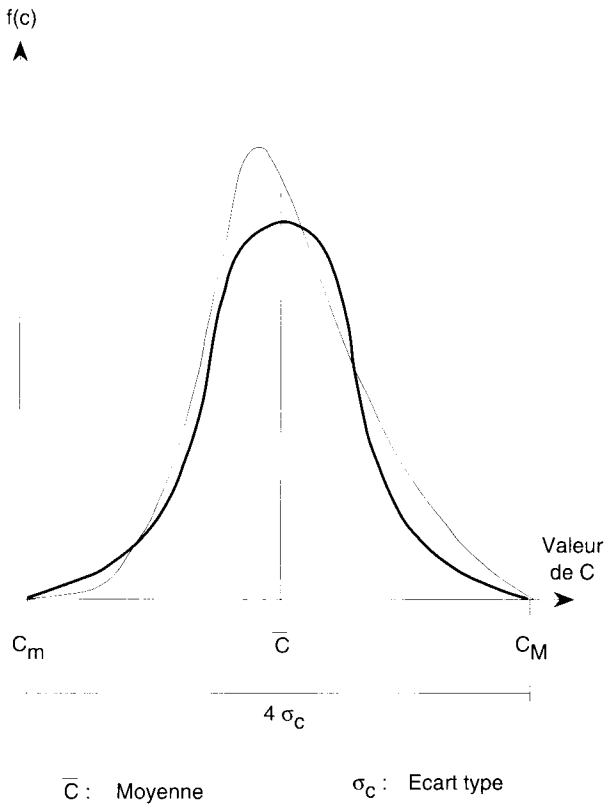
## 2.3 / Le moment d'ordre 2 : variance et écart-type

Si nous souhaitons préciser la description de la population, il est probable que nous donnerons les valeurs extrêmes : "celà va de tant pour les plus petits à tant pour les plus grands". C'est la notion de gamme de variation.

La mesure de cet intervalle est l'écart-type (ou la variance qui est sa puissance carrée). Son calcul est simple dans le cas où tous les objets constituant la population sont indépendants. Il se complique énormément si cette hypothèse très simplificatrice ne peut être admise. Le travail d'équipes puissantes, les moyens actuels de la grande informatique, ne sont pas encore parvenus à une solution définitive et "optimale" de l'estimation d'une variance dans une population très structurée (comme le sont les populations en sélection). Plusieurs des articles suivants évoquent ces problèmes.

Cela n'empêche pas l'écart-type d'avoir toujours la même signification : il chiffre le degré d'étalement de la distribution. Certes il peut paraitre abstrait. Mais, pour peu que l'histogramme soit à peu près symétrique, la gamme de variation ($C_mC_M$ sur la figure 1) recouvre à peu près quatre écarts-types .

**Figure 1.** *Répartition des valeurs de C.*



## 2.4 / Les autres "moments" de l'histogramme

Moyenne et écart-type ne suffisent pas à résumer la forme exacte de l'histogramme. Les deux courbes de la figure 1 diffèrent par autre chose que moyenne et variance. D'autres concepts peuvent améliorer cette description : le degré de symétrie (la courbe fine est visiblement moins symétrique), l'aplatissement, et une infinité d'autres . En fait, dans la quasi-totalité des cas, l'utilisateur se satisfait des deux premiers (On trouvera un contre-exemple d'utilisation de moments d'ordre supérieur dans l'article de Pascale Leroy).

C'est sans doute l'une des raisons expliquant avec quelle facilité on admet qu'une distribution, parfois de forme très exotique, est conforme à l'équation de la loi normale. Le test de normalité est basé sur l'examen des moments d'ordre égal ou supérieur à 3, qui sont extrêmement difficiles à apprécier. (Quelques dizaines d'individus suffisent à se faire une bonne idée de la moyenne. Pour appréhender l'écart-type, il faut observer les valeurs extrêmes de l'hitogramme qui sont justement les moins fréquentes, donc un très grand nombre d'individus (dix fois plus que pour la moyenne). Pour les moments d'ordre supérieur, cela empire régulièrement).

En résumé , la moyenne et l'écart-type des mesures d'un caractère dans une population constituent une description, en général jugée parfaitement satisfaisante, de cette population.

# 3 / Représentation bidimensionnelle

## 3.1 / Les nuages de points

Imaginons que chaque individu appartenant à la population est caractérisé par les valeurs numériques de deux caractères (C et G sur la figure 2) . Il peut donc être représenté par un point dans le système d'axes (C,G). Après avoir placé tous les individus, la population constitue le nuage de points de la figure 2. On notera que la répartition des points n'est pas uniforme : il y en a une densité plus forte le long d'une "dorsale" constituant l'axe d'étirement de la figure. Si on ajoute, perpendiculairement au plan (G,C), un axe des fréquences de chaque couple (G,C), on obtient la figure 3a, équivalent tridimensionnel de la figure 1.

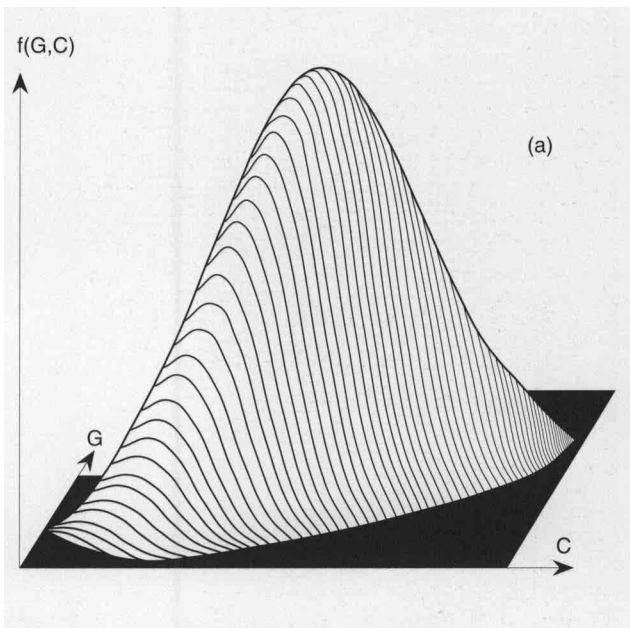
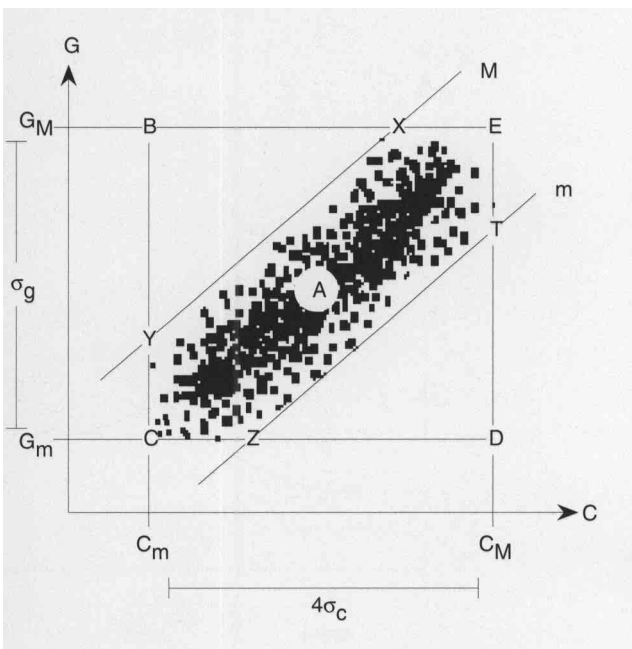


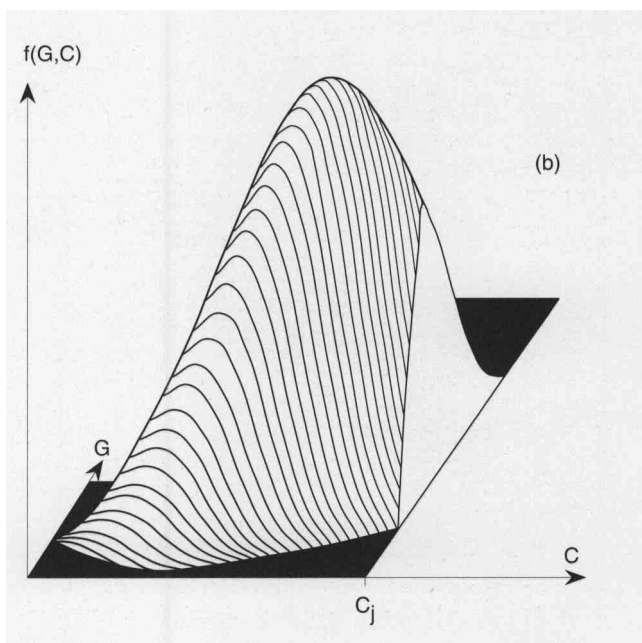

**Figure 3.** *(a) Représentation tridimensionnelle d'un nuage de points.*

*(b) Histogramme des valeurs possibles de G chez des individus ayant la mesure $C_j$.*

**Figure 2.** *Description d'un nuage de points.*

Les coordonnées de **A** sont les moyennes de C et G dans la population.
$C_m$ $C_M$ $G_m$ $G_M$ sont les valeurs extrêmes observées pour C et G.
$\sigma_c$ et $\sigma_g$ sont les écarts-types des deux mesures.



## 3.2 / La corrélation

Le résumé de cette figure consiste en certains paramètres dont nous connaissons déjà la plupart. Les moyennes de C et de G sont les coordonnées d'un point, noté A , qui représente la position "en gros" de la population.

Nous savons de même calculer les écarts-types des deux variables, qui chiffrent l'étendue du nuage de points dans le sens des deux axes G et C, c'est-à-dire les segments ($C_m$ $C_M$) et ($G_m$ $G_M$). En termes à peine abusifs, on exprime ainsi que le nuage est placé à l'intérieur du rectangle B C D E, dont les sommets ont pour coordonnées les valeurs des moyennes plus ou moins deux écarts-types.

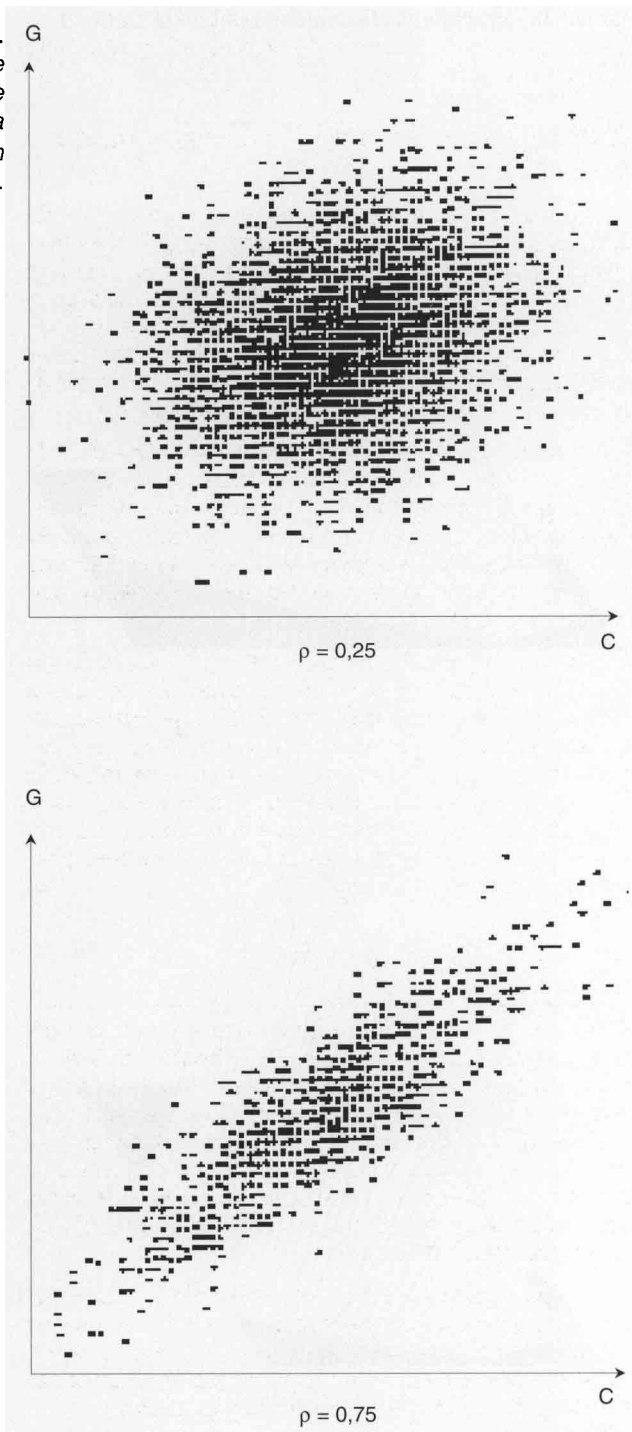



**Figure 4.** *Exemple de nuages de points. ρ est la corrélation entre G et C.*

La figure 4 représente deux nuages de points obtenus dans deux populations différentes. Moyennes et écarts-types de chaque variable sont égaux pour les deux populations. Il nous manque le moyen d'exprimer cette différence pourtant évidente : le deuxième est beaucoup plus étiré. Le paramètre chiffrant l'aplatissement du nuage de points est appelé corrélation.

C'est un paramètre compris entre -1 et 1. Il est nul quand le nuage de points est globuleux. On dit que les mesures C et G sont indépendantes (nous reviendrons sur cette notion dans le paragraphe suivant). Quand la corrélation s'écarte de zéro, le nuage s'étire dans une direction, de pente négative ou positive selon le signe de la corrélation. La figure 4a correspond à une corrélation proche de zéro (0,25), 4b à une valeur nettement plus forte (0,75).

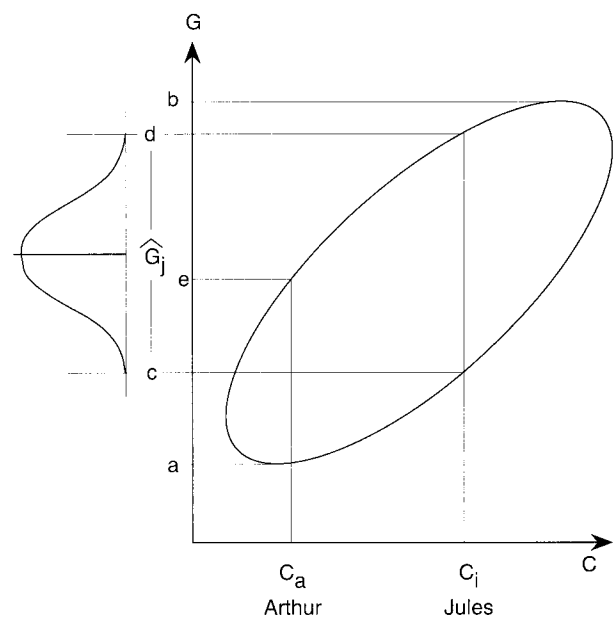
La valeur de la corrélation enferme le nuage entre deux droites D1 et D2 dont la pente est proportionnelle à la corrélation et l'écart d'autant plus petit que la corrélation est forte. La connaissance de cinq chiffres (moyennes, écarts-types et corrélation) permet de tracer le polygone EXYCZT qui décrit fort bien les contours du nuage de points sur la figure 2.

## 3.3 / La prédiction par régression

Revenons sur cette notion de liaison statistique. Considérons un individu, nommé Jules, et intéressons nous à la valeur prise par sa mesure de G. A priori, elle est quelconque, de moins l'infini à plus l'infini. Mais si nous savons que Jules appartient à la population dépeinte sur la figure 5, cela implique que la valeur $G_j$ de Jules, comme celle de tout individu de la population, est contenue sur le segment (a,b). Elle reste inconnue, puisque quelconque sur ce segment, mais beaucoup moins incertaine qu'avant.

Imaginons maintenant que nous connaissions la valeur exacte de la mesure $C_j$ de Jules (mais toujours

**Figure 5.** *Prédiction de G par régression sur C. L'ellipse représente le contour extrême du nuage de points. La courbe sur la gauche représente l'ensemble des valeurs de G possibles pour Jules quand on connaît sa valeur $C_j$.*



pas $G_j$). Il existe dans la population plusieurs individus qui partagent cette même valeur $C_j$. La forme du nuage de points nous indique que leurs valeurs pour G sont incluses dans la gamme (c,d). La mesure de $C_j$, en réduisant de (a,b) à (c,d) la gamme des possibles pour $G_j$, a permis de réaliser une prédiction statistique de G par C.

Rappelons que ce segment (c,d) n'est pas uniforme. La figure 3b en donne la représentation tridimensionnelle : c'est un histogramme que l'on voit sur la tranche de cette figure et qui est reporté en le couchant sur la gauche de la figure 5. La prédiction de G par C est cet histogramme représentant les seules valeurs possibles. Nous allons le résumer comme nous savons le faire: la moyenne ($\hat{G}j$) prend le nom de prédiction de G par C. L'étalement de (c,d) sera par contre exprimée un peu différemment par un "coefficient de détermination" : la longueur de (c,d) (comprenez l'écart-type de la distribution) est exprimée en valeur relative par rapport à la variation totale (a,b). Il exprime ainsi la réduction d'incertitude sur la variable G apportée par la connaissance de mesures statistiquement liées.

Cette prédiction permettra de faire un classement anticipé, sur des valeurs de G que l'on ne connaît pas encore. Si Arthur est un autre individu de la même population ayant la mesure $C_a$, il a moins de chances que Jules d'avoir une meilleure valeur de G. Le contraire est bien sûr possible... mais moins probable. On ne le saura que plus tard quand G sera mesuré. Et si on doit choisir tout de suite entre les deux, c'est Jules le meilleur. Bien sûr, plus le nuage de points sera aplati (corrélation élevée) plus la longueur du segment (c, d) sera petite et meilleure sera la prédiction.

Appelez G une valeur génétique, et remplacez C par un certain nombre de mesures phénotypiques réalisées et vous venez de commencer à comprendre ce qu'est l'indexation des reproducteurs.

# 4 / Représentation multi-dimensionnelle

Quand les individus de la population sont caractérisés par plus de deux variables, le nuage de points correspondant occupe un espace de dimension supérieure à trois, évidemment non visualisable. Diverses méthodes, appelées multivariables, ont été imaginées pour fournir la "meilleure" visualisation dans notre espace réel (deux voire trois dimensions). L'acception du mot "meilleur" dépend bien sûr du but recherché. Nous allons présenter deux méthodes.

## 4.1 / L'Analyse en Composantes Principales (ACP)

Prenons un stylo. Nous voulons faire de cet objet tridimensionnel une photographie plane, la plus représentative possible. D'instinct, nos placerons la plus grande longueur de l'objet perpendiculairement à l'objectif ; puis le ferons tourner autour de son axe jusqu'à ce que la barrette ne soit plus cachée par le corps. Nous avons réalisé une sorte d'ACP.

L'ACP consiste à rechercher une première direction selon laquelle l'étalement du nuage de points est maximum (c'est la première composante principale) ; puis, ayant fixé cette direction, à faire tourner le nuage jusqu'à obtenir perpendiculairement à elle l'étalement maximum (deuxième composante) ; et ainsi de suite. La projection du nuage de points sur le plan défini par les deux (ou trois si on fait une représentation tridimensionnelle) premiers axes fournit la représentation la meilleure de la diversité de la population.

Souvent d'ailleurs, on utilise le point de vue ainsi défini pour représenter non les individus, mais les variables initiales (le système d'axes ayant servi à construire le nuage de points). La figure obtenue s'interprète de façon très intuitive à l'aide de quelques règles simples : les variables groupées au centre du dessin ne sont pas interprétables. Mais pour celles de la périphérie, proximité des points et liaison statistique sont synonymes. Deux points confondus indiquent une corrélation de 1 entre les variables ; placés à deux extrémités opposées du dessin, ils dénotent une liaison voisine de -1, etc.

## 4.2 / L'Analyse Factorielle Discriminante (AFC)

Mettons deux cahiers peu épais l'un au dessus de l'autre. Cette fois, nous voulons montrer qu'il y a effectivement plusieurs sous populations - pardon, cahiers - dans la pile. L'ACP est particulièrement inadaptée, puisque, recherchant l'étalement maximum, elle ne visualise que la couverture du premier. Nous prendrons dans ce cas une photo de la tranche des cahiers, selon un axe parallèle à la table qui les supporte.

La méthode consiste à rechercher une direction (le premier axe discriminant) sur lequel les nuages de points représentant chacune des sous-populations ont des intersections aussi réduites que possible ; puis, cet axe restant définitivement fixé, à en rechercher une seconde, perpendiculaire, qui permet la meilleure discrimination ; et ainsi de suite.

Une fois encore, les formules d'algèbre linéaire qui réalisent les calculs correspondants sont déclarées "triviales" (comprenez qu'il convient de mépriser quelqu'un avouant qu'elles sont compliquées) par les mathématiciens, ce qui prouve leur mauvaise foi. Mais l'existence de logiciels très conviviaux n'exige pour comprendre ces méthodes et les utiliser que des connaissances conceptuelles et intuitives guère plus élaborées que ce qui précède.

# Conclusion

La valeur génétique, telle qu'elle est vue par les Généticiens Quantitatifs, est une abstraction statistique comme vous le démontreront les articles de Minvielle et Ollivier qui suivent. Elle ne peut être définie qu'à l'intérieur d'une population d'animaux dont on a décidé de comparer entre eux les potentiels génétiques. Elle n'a aucun sens en dehors de cette population puisque toute statistique est la description d'une population donnée.

La description statistique de la population est obtenue au travers de l'estimation d'un certain nombre de paramètres : moyennes, variances, corrélations. Ils constituent une description jugée en général suffisante du nuage de points représentant la population.

L'indexation des reproducteurs consiste à utiliser la forme et la structure du nuage de points (c'est-à-dire les paramètres statistiques) pour réduire la gamme des valeurs que peuvent prendre certains paramètres (les valeurs génétiques), en utilisant des observations réalisées (les valeurs phénotypiques). Elle se traduit par deux chiffres : une valeur espérée, et la précision de la prédiction.

## Références bibliographiques

Dagnelie P., 1985. Théorie et méthodes statistiques.Ed : Les Presses Agronomiques de Gembloux, Gembloux (Belgique), Volumes 1 et 2 : 378 et 463 pp.