Application des fonctions discriminantes à des problèmes biométriques

Richard Tomassone

APPLICATION
DES
FONCTIONS DISCRIMINANTES
A DES
PROBLÈMES BIOMÉTRIQUES

PAR

R. TOMASSONE

Ingénieur des Eaux et Forêts
9e Section de la Station de Recherches de Nancy

I — INTERET DE L'ANALYSE STATISTIQUE A PLUSIEURS VARIABLES

I. 1 — But de l'article.

La donnée d'un seul caractère suffit quelquefois à définir un être mathématique ; mais ce cas est déjà très rare, à peu près exceptionnel : pour définir un cercle dans un plan, il faut connaître la position de son centre et la valeur de son rayon. Le numéro atomique d'un élément de la classification de Mendeleiev suffit pour caractériser l'élément ; mais dès qu'on analyse des caractères macroscopiques il faut plusieurs données pour reconnaître un corps simple. Ces simples constatations de bon sens permettent de supposer que lorsqu'on aborde des problèmes relevant du domaine biologique, vouloir définir un être vivant, que ce soit une plante ou un animal, par un seul caractère est un point de vue bien utopique. Les botanistes et les entomologistes sont d'ailleurs absolument conscients de cette nécessité puisqu'ils collectionnent le maximum de caractères représentatifs pour établir des clés de reconnaissance (1).

Les caractères sont quelquefois liés entre eux ; il est tout à fait logique de rendre compte de ces liens quand on décrit un élément : c'est précisément un des buts des méthodes statistiques à p variables (2). A un élément on n'associe plus la valeur d'une variable mais l'ensemble des p variables qui ont servi à mesurer les p caractères.

(1) D'un point de vue statistique, ces clés de reconnaissance utilisent, sans le dire, les propriétés de la loi binomiale (quelquefois multinomiale) puisqu'à chaque niveau de la clé, on classe les individus comme possédant ou ne possédant pas un certain caractère. On peut aborder ces classifications d'une façon tout à fait « moderne » en utilisant la théorie des ensembles : étant donné 2 ensembles A et B, on cherche les caractères appartenant à A et non à B (donc A — B), et ceux appartenant à B et non à A (donc B — A), c'est-à-dire que l'on cherche à rendre maximum la différence symétrique de A et B. On voit tout l'intérêt de ces notions si on songe aux applications possibles sur machine électronique qui peuvent utiliser les variables booléennes (une expression booléenne est une valeur logique : VRAI ou FAUX, elle peut prendre 2 valeurs). Les caractères communs à A et B ne sont pas utiles pour reconnaître les individus. A ce sujet cf. LUBISCHEW (A.A.) 1962, op. cit.

(2) On dit quelquefois analyse multivariable.

On peut, dans un premier stade, séparer ces méthodes en deux groupes :

a) celles qui permettent de diminuer le nombre des variables : on fait une transformation, en général linéaire, des variables initiales et on ne garde dans les transformées que celles dont la contribution à la variation totale est la plus importante : c'est l'objet de l'analyse des composantes principales (1).

b) celles qui permettent de séparer des ensembles, que nous nommerons à partir de maintenant populations ; puis une fois ces ensembles différenciés de trouver un critère valable, pour rattacher un élément inconnu à l'un d'eux : c'est l'objet de l'étude des fonctions discriminantes. Comme toujours en calcul des probabilités, on ne dira qu'un élément appartient à une population qu'avec une certaine probabilité (2).

Nous nous intéressons dans cet article au second de ces groupes. L'intérêt que nous portons à ce problème n'a rien de bien nouveau puisqu'en 1759 ADANSON disait : « S'il y a dans la nature un système que nous puissions saisir, il ne peut être fondé que sur l'ensemble des rapports des caractères tirés de toutes les parties et qualités d'une plante ». Plus près de nous, FISHER, en utilisant les travaux auxquels il a lui-même grandement contribué, a donné aux études biométriques une orientation nouvelle, grâce aux méthodes statistiques modernes. Depuis, les progrès rapides des moyens de calcul, et la grande capacité de mémoire des calculatrices électroniques, permettent d'aborder les problèmes et de les résoudre rapidement, alors qu'on ne pensait même pas s'y attacher il y a quelques années (4).

Dans un premier stade, ces méthodes peuvent permettre de retrouver des résultats déjà connus, mais après, elles peuvent les éprouver, les confirmer ou les infirmer, et enfin noter jusqu'où elles sont valables (5).

I. 2 — **Les outils mathématiques nécessaires.**

Le problème mathématique que l'on doit résoudre est suffisam-

(1) De plus, les variables transformées sont non corellées, donc indépendantes dans le cas de distributions normales.

(2) Il ne faut pas être choqué par le terme « avec une certaine probabilité », les physiciens eux-mêmes parlent de la probabilité qu'a un électron de se trouver sur une orbite privilégiée, alors que les calculs qu'ils font demandent souvent une grande précision. La nature probabiliste des phénomènes n'empêche tout de même pas ces physiciens de construire des réacteurs atomiques...

(3) Cf. FISCHER R.A. (1936), op. cit.

(4) Un petit ordinateur moderne peut avoir 60 000 chiffres en mémoire, en lire 300 par seconde, et faire une multiplication en 15 millisecondes.

(5) On peut en particulier espérer arrêter la prolifération des sous-espèces et des variétés...

ment complexe, nous le verrons au chapitre suivant, pour qu'il soit nécessaire d'employer des outils adéquats. C'est en particulier la raison de l'utilisation du calcul matriciel qui arrive à synthétiser au maximum les relations entre deux ensembles de variables. Dans le domaine biologique, ce n'est pas une utilisation entièrement nouvelle, puisqu'on l'a déjà employé pour représenter des populations aussi bien végétales qu'animales (1).

Mais nous verrons que la complexité du problème ne doit pas rebuter le botaniste ou l'entomologiste puisque l'ensemble des calculs peut être programmé sur machine électronique.

I. 3 — **Exemples d'utilisation.**

Il y a toujours eu un fossé assez large entre la théorie et son application ; ce fossé est d'autant plus large que les sciences auxquelles elle s'applique sont éloignées des spéculations mathématiques. Pourtant il existe à l'heure actuelle de très nombreux champs d'application (2). Nous allons en donner un très bref aperçu, car très souvent le problème posé par le psychologue (3) est formellement identique à celui du botaniste :

— *en psychologie:* on a réussi à séparer différents groupes de personnes en les soumettant à des tests. Une personne absolument quelconque est de nouveau soumise aux mêmes tests : à quel groupe appartient-elle ?

— *en archéologie:* on possède une série de squelettes sur lesquels on fait un grand nombre de mesures, on peut se demander s'ils appartiennent à la même race (4). On peut ensuite étudier quelles sont les meilleures mesures à prendre (5) ; puis essayer de rattacher un squelette isolé à un des groupes précédemment définis (6).

— *en médecine:* le diagnostic est quelquefois simple, mais des maladies dont les manifestations cliniques sont très voisines demandent des critères spéciaux, il faut faire de nombreux examens au laboratoire dont la synthèse est grandement simplifiée par l'utilisation des fonctions discriminantes.

I. 4 — **Applications forestières déjà étudiées.**

Il nous a paru bon d'exposer les méthodes qui ne sont connues sous leur aspect théorique et pratique que dans des ouvrages de lan-

(1) Cf. JOLLY G.M. (1963), op. cit.
(2) Cf. KENDALL M.G. (1961), op. cit., p. 7-9.
(3) Cf. FAVERGE J.M. (1950), op. cit.
(4) Cf. STOESSIGER B.N. and MORANT G.M. (1932), op. cit.
Cf. WOO T.L. and MORANT G.M. (1932), op. cit.
(5) On peut prendre plus de quarante mesures différentes sur un seul crâne, n'en prendre que certaines permettrait un gain de temps appréciable.
(6) Cf. MARTIN E.S. (1936), op. cit.

gue anglaise. L'emploi, au moins en France, semble être beaucoup plus l'objet d'un vœu qu'une réalité (1). Dans le domaine strictement forestier, les travaux les plus importants ont été effectués en Grande-Bretagne; nous noterons en particulier:

— Une analyse sur *Pinus contorta* qui a confirmé la division entre provenance côtière et provenance intérieure, mais qui a suggéré qu'une classification meilleure existait en tenant compte des variations proprement botaniques et probablement technologiques (2).

— Sur deux variétés de Peupliers *Populus serotina* et *Populus gelrica* en mesurant 3 caractères: longueur, largeur et angle d'attache des feuilles, on peut différencier les races dans la plupart des cas (3).

II — PRINCIPE DE LA METHODE CAS DE DEUX POPULATIONS

II. 1 — Généralisation de la comparaison de deux moyennes.

a) Lorsqu'on veut étudier une population trop grande pour qu'on puisse examiner chacun des éléments: plantes appartenant à un même génotype par exemple (4), on choisit un échantillon aussi représentatif que possible de la population; si cet échantillon comprend n éléments, nous l'appellerons, suivant l'usage, un n-échantillon. Si sur chaque élément on mesure un caractère, on peut avoir une estimation de la densité de répartition de la population en traçant l'histogramme du n-échantillon relatif à ce caractère. Généralement, on calcule deux valeurs attachées au n-échantillon:

— sa moyenne qui est un paramètre de position,
— sa variance (ou son écart type) qui est un paramètre de dispersion.

Lorsqu'on a une distribution normale, le grand intérêt provient du fait que ces deux paramètres suffisent pour la définir entièrement.

b) Si on désire ensuite comparer deux populations, on peut d'une façon très générale comparer les deux distributions et se rendre compte si les 2 échantillons diffèrent entre eux par autre chose que des erreurs d'échantillonnage. Le problème ainsi posé est déjà

(1) Cf. DUFRENOY J. (1963), op. cit.
(2) Cf. JEFFERS J.N.R. (1961), op. cit.
(3) Cf. JEFFERS J.N.R. (1960), op. cit.
(4) Suivant le langage employé dans les traités de mathématiques modernes, cette population constitue un ensemble: par exemple l'ensemble des plantes ayant les mêmes caractères.

très difficile à résoudre théoriquement (1) ; aussi se contente-t-on dans les cas usuels de comparer les moyennes et les variances. Pour plusieurs populations, on s'attache généralement à « tester » les hypothèses suivantes :

— H_0 : les populations ont les mêmes moyennes et les mêmes variances.
— H_1 : les populations ont les mêmes variances, sans s'occuper des moyennes.
— H_2 : *étant donné* l'égalité des variances, les moyennes sont égales.

Le test de l'hypothèse H_1 peut s'effectuer de plusieurs façons en particulier à l'aide du test de BARTLETT (2).

Celui de l'hypothèse H_2 est résolu au moyen de l'analyse de variance.

c) Si on mesure sur un même élément plusieurs caractères, soit p, on peut évidemment faire p comparaisons entièrement distinctes, mais on néglige un aspect important du problème car les caractères mesurés ne sont généralement pas indépendants, et une analyse séparée ne tient aucun compte de la liaison qui existe. La solution logique consiste donc à utiliser tous les caractères réunis.

Pour avoir une idée plus précise du problème posé, il est bon d'avoir une représentation géométrique : si on considère p axes orthogonaux où chacun se rapporte à un des caractères mesurés (on définit ici un espace que nous noterons E_p), une observation sur un élément est définie par un point dont les coordonnées sont les p mesures (3). De cette façon, un n-échantillon est naturellement défini par n points de l'espace, ces n points constituent un nuage ; si le n-échantillon est suffisamment représentatif de la population, le nuage donne une estimation de la distribution de la population suivant les p caractères examinés.

Pour séparer deux populations à partir de leur nuage, il faut donc s'assurer que les « différences » ne sont pas uniquement dues à des variations d'échantillonnage. Evidemment par « différence »

(1) La loi de KOLMOGOROV-SMIRNOV étudie la façon dont une loi d'échantillon s'approche de la loi véritable. Le maximum de l'écart suit une loi analytique assez complexe. Dans l'état actuel des travaux théoriques, les possibilités d'application semblent assez réduites, voir à ce sujet :
— pour l'aspect théorique : DUGUE D. (1958), op. cit.
— pour l'aspect pratique : FERIGNAC P. (1962), op. cit.

(2) Pour l'utilisation pratique de ce test, cf. VESSEREAU A. (1960), op. cit., p. 157-158.

(3) Il revient évidemment au même de définir une observation par un vecteur observation dont les coordonnées sont les p mesures, cf. à ce sujet :
— KENDALL M.G. and MORAN P.A.P. (1963), op. cit.
— KUIPER W.H. (1960), op. cit.

il faut entendre ici différence de volumes dans l'espace E_p, ce qui est l'extension logique de la différence linéaire qui correspond par exemple, dans le cas où on mesure les hauteurs moyennes de diverses provenances, à la différence de hauteur de deux d'entre elles. Nous allons voir que le problème à une dimension s'étend aisément, au moins d'une façon intuitive, à celui à p dimensions :

— la moyenne d'une population est remplacée par les p moyennes dont une image est donnée par la position du centre de gravité G du nuage.

— la variance est remplacée par une matrice de dispersion (1) qui est le tableau carré symétrique formé par les quantités :

(II,1) $$c_{ij} = \text{cov}\ (x_i, x_j)\ (2)$$

(II,2) $$\mathbf{C} = \begin{vmatrix} \text{var}\ (x_1) & \text{cov}\ (x_1, x_2) & \dots & \text{cov}\ (x_1, x_p) \\ \text{cov}\ (x_2, x_1) & \text{var}\ (x_2) & \dots & \text{cov}\ (x_2, x_p) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \text{cov}\ (x_p, x_1) & \text{cov}\ (x_p, x_2) & \dots & \text{var}\ (x_p) \end{vmatrix}$$

L'intérêt de la distribution normale à une dimension se retrouve dans la distribution normale à p dimensions : le point G et la matrice **C** la définissent entièrement (3).

II. 2 — **Position du problème.**

a) Nous supposons qu'on veuille mettre en évidence les différences entre deux populations possibles A_1 et A_2. Sur chaque élément des deux populations, on mesure p caractères. En outre, on admet que les densités de probabilités des deux populations sont connues, soit f_1 et f_2. Si on connaît les probabilités a priori pour qu'un élément appartienne à A_1, ou à A_2, on peut calculer par une simple application du principe de Bayes les probabilités a posteriori, c'est-à-dire après avoir mesuré les p caractères de l'élément, pour que l'individu appartienne à A_1 ou à A_2 (4).

(1) On l'appelle aussi matrice des variances et covariances.

(2) Cf. Annexe I pour l'explication des notations employées et quelques rappels des propriétés du calcul matriciel qui nous sont nécessaires. Une matrice sera toujours notée avec un caractère gras **C**, de même pour un vecteur **X**.

(3) Cf. Annexe II pour un rappel des propriétés de la loi normale à p dimensions.

(4) Cette éventualité n'est pas impossible, en particulier dans les problèmes où le sexe de l'élément considéré intervient. Cf. MARTIN E.S. (1936), op. cit.

Ces probabilités a priori ne sont en général pas connues ; il faut donc songer à aborder le problème d'une autre façon : on peut en particulier s'imposer les conditions suivantes :

a — si un élément appartient à A_1, la probabilité pour qu'on le classe dans l'autre population doit être la même que la probabilité de le classer dans A_1 s'il avait appartenu à A_2 ; ce qui s'écrit (1) :

$$\int_{a_1} f_2 \, dv = \int_{a_2} f_1 \, dv \qquad \text{(II,3)}$$

b — cette probabilité doit être minimum (2).

Les conditions *a* et *b* permettent de trouver la frontière de la région A_1 par :

$$\frac{f_1}{f_2} = l \qquad \text{(II,4)}$$

où l est un paramètre qui doit satisfaire à l'équation (II,3). Les points pour lesquels le rapport f_1/f_2 est supérieur à l sont affectés à la population A_1, et s'il est inférieur à la population A_2 (3). Enfin, la probabilité pour qu'un classement soit mauvais est minimum et donnée aussi par la valeur d'un des deux membres de (II,3).

Le problème est donc résolu de façon entièrement théorique, sans aucune hypothèse sur les distributions f_1 et f_2.

b) Pour pouvoir l'appliquer pratiquement, nous supposerons que les 2 distributions sont p-normales et ne diffèrent que par leurs valeurs moyennes (4). La démonstration à partir des résultats énoncés au paragraphe précédent, assez délicate, est renvoyée à la fin de l'Annexe II. Mais dans ce cas particulier, on peut aborder le

(1) L'intégrale est une intégrale multiple dans un espace à p dimensions dont l'élément de volume est dv ; a_1 est la région de l'espace correspondant à A_1, a_2 celle correspondant à A_2 ; évidemment a_2 est le complémentaire de a_1, si l'espace est normé

$$a_2 = 1 - a_1$$

(2) L'ensemble de ces deux conditions est donc un cas particulier du problème envisagé à la note (1) de la page 1.

(3) Voir à ce sujet WELCH B.L. (1939), op. cit.

(4) Un test approché permet de s'assurer que les matrices de dispersion sont bien identiques, cf. Annexe III. Dans ce cas, on trouve une fonction discriminante linéaire ; si les matrices sont différentes, on peut étudier le cas où tous les coefficients de corrélation partiels entre les variables sont égaux : on réduit le nombre de variables à deux en utilisant ce que les Anglais appellent la composante de grandeur (« size component ») et celle de forme (« shape component »). Cf. BARTLETT M.S. and PLEASE N.W. (1963), op. cit.

problème d'une façon plus intuitive et qui, nous allons le voir, aboutit au même résultat: on cherche une combinaison linéaire des variables telle que si:

(II,5) $$Y = \Sigma_i \, l_i \, x_i \quad (1)$$

le rapport:

(II,1) $$t = \frac{|Y_1 - Y_2|}{\sqrt{\text{var } (Y_1 - Y_2)}}$$

soit maximum (2).

Pour chaque élément d'une population, les quantités x_i sont mesurées, et on cherche les l_i en fonction de ces valeurs. Rendre t maximum est équivalent à rendre t^2 maximum, t^2 étant une fonction de $l_1, l_2, \ldots l_p$. Donc pour chaque l_i on doit avoir:

(II,7) $$\frac{\delta}{\delta l_i}\left(\frac{A^2}{B}\right) = 0 \qquad \text{où:}$$

(II,8)
$$\begin{cases} \cdot\ A = Y_1 - Y_2 = \Sigma_i \, l_i \, d_i \\ \cdot\ B = \text{Var } (Y_1 - Y_2) = \Sigma_{ij} \, l_i \, l_j \, c_{ij} \quad (3) \\ \cdot\ d_i = x_{1i} - x_{2i} = \text{différence des valeurs de la même variable (i) pour les 2 populations.} \end{cases}$$

L'équation (II,7) est équivalente à:

(II,9) $$2 \cdot \frac{\delta A}{A \, \delta l_i} = \frac{\delta B}{B \, \delta l_i} \qquad \text{avec:}$$

(II,10)
$$\begin{cases} \dfrac{\delta A}{\delta l_i} = d_i \\ \dfrac{\delta B}{\delta l_i} = 2 \, \Sigma_j \, l_j \, c_{ij} \end{cases}$$

(1) Quand il n'y aura aucun risque de confusion et pour ne pas alourdir le texte, la notation $\Sigma_i \, l_i \, x_i$ signifiera:

$$\sum_{i=1}^{i=p} l_i \, x_i$$

(2) Y est évidemment une variable normale, et le test se ramène à celui de la comparaison de 2 moyennes par un test t de STUDENT. Le problème du paragraphe II,1,c a été simplifié puisqu'on ne s'intéresse plus qu'à une composante linéaire de la variation.

(3) La sommation est ici étendue aux deux indices i et j.

donc en portant ces valeurs dans (II,9)

$$\Sigma_j \, l_j \, c_{ij} = \frac{B}{A} \, d_i \qquad \text{(II,11)}$$

Nous remarquons tout d'abord que nous ne voulons connaître les quantités l_i qu'à un coefficient de proportionnalité près (1). D'autre part, on peut considérer les p valeurs de d_i et l_i comme les composantes de deux vecteurs colonnes **D** et **L**, et alors (II,11) s'écrit :

$$\mathbf{C\,L = D} \qquad \text{(II,12)}$$

qui se résout immédiatement par :

$$\mathbf{L = C^{-1}\ D} \qquad \text{(II,13)}$$

où $\mathbf{C^{-1}}$ est la matrice inverse de **C**.

La fonction discriminante correspondant au vecteur observation **X** s'écrit sous forme de produit scalaire

$$Y = \mathbf{L'\,X = X'\,L} \qquad \text{(II,14)}$$

Connaissant A, il ne reste plus qu'à calculer B :

$$B = \Sigma_{ij} \, l_i \, l_j \, c_{ij} = \Sigma_i \, l_i \, d_i \qquad \text{(II,15)}$$

$$B = Y_1 - Y_2 \qquad \text{(II,16)}$$

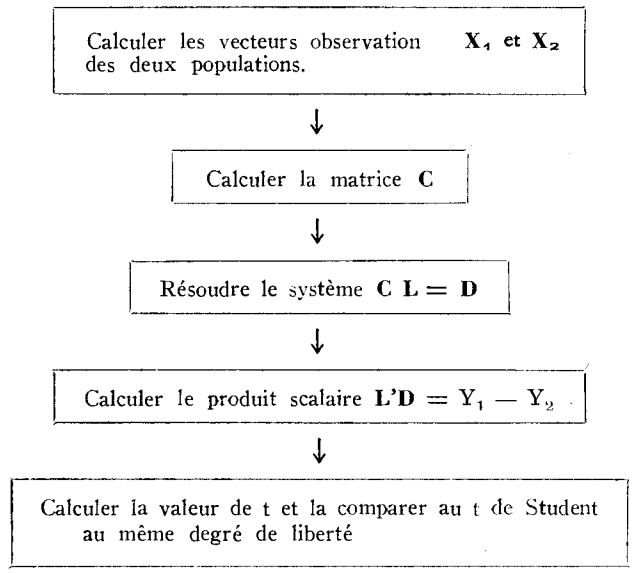
Y_1 est calculé dans (II,5) en prenant les valeurs des variables correspondant à la population A_1 (2).

(1) Si on multiplie tous les l_i par une même valeur quelconque, la valeur de t dans (II,6) est inchangée.

(2) Cette façon d'aborder le problème s'apparente évidemment à l'étude de la régression d'une variable Y en fonction de p variables x_i ; les quantités l_i sont les coefficients de régression de Y en x_i dont on peut tester la valeur par rapport à zéro. Cf. BARTLETT M.S. (1939), cité par KENDALL M.G. (1961), op. cit.

II. 3 — **Organisation des calculs.**

On peut présenter la suite des calculs sous forme d'un organigramme utile pour la programmation de l'ensemble du problème sur machine électronique (1) :



II. 4 — **Exemple.**

(M. BOUVAREL et LEMOINE exp T 21 - IV - 1960).

On voulait mettre en évidence les différences qui peuvent exister entre deux populations de pin sylvestre voisines, mais situées dans des stations à microclimats très différents : les deux peuplements étaient situés à une distance de 5 kilomètres dans le massif de la Sainte-Baume (Var) :

— Sainte-Baume I (B 1) : dans une station sèche à pin d'Alep
— Sainte-Baume II (B II) : dans une station humide à hêtre.

Une expérience avec 4 répétitions avait été effectuée sur des semis et on avait mesuré les 4 caractères suivants :

— x_1 : hauteur des plants (cm)
— x_2 : longueur de l'aiguille la plus longue (cm)
— x_3 : diamètre au collet (cm)
— x_4 : poids total des plants après séchage (g).

(1) Voir à ce sujet ARBONNIER P. (1962), op. cit.

Au lieu d'utiliser la matrice **C**, nous utilisons la matrice $\mathbf{Q} = n_E\ \mathbf{C}$ qui est obtenue directement dans les calculs (1):

(II.17)
$$\mathbf{Q} = \begin{vmatrix} 557{,}775 & 66{,}450 & 5{,}132 & -14{,}121 \\ & 81{,}611 & 1{,}075 & -5{,}812 \\ & & 10{,}983 & 0{,}654 \\ & & & 7{,}680 \end{vmatrix}$$

D'autre part, les moyennes des 2 provenances sont:

(II,18)

	B I	B II	D = B II — B I
x_1	9,599	10,448	0,849
x_2	4,459	5,254	0,795
x_3	1,870	1,872	0,002
x_4	0,607	0,714	0,107

On a calculé la matrice inverse $\mathbf{Q}^{-1}$ (multipliée par 10^7)

(II,19)
$$\mathbf{Q}^{-1} \times 10^7 = \begin{vmatrix} 20448 & -14569 & -9760 & 27402 \\ & 140272 & -11707 & 80363 \\ & & 922478 & -105360 \\ & & & 1422256 \end{vmatrix}$$

et le vecteur **L**

(II,20)
$$\mathbf{L} \times 10^7 = \begin{vmatrix} 8690 \\ 107723 \\ -27022 \\ 239124 \end{vmatrix}$$

Il y avait $n_1 = 35$ mesures pour B I et $n_2 = 38$ mesures pour B II, et en tenant compte des 4 répétitions $n_E = 65$, la variance

(1) Nous omettons volontairement la partie symétrique de la matrice. Les unités ne sont pas homogènes, mais ceci est sans importance puisqu'on s'intéresse à un rapport sans dimensions. n_E est le nombre de degrés de liberté de la variance résiduelle.

de la différence des 2 valeurs Y_1 et Y_2 correspondant aux moyennes de B_1 et B_2 vaut :

$$\text{Var}\ (Y_1 - Y_2) = \frac{Y_1 - Y_2}{n_E} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \tag{II,21}$$

$$\sqrt{\text{Var}\ (Y_1 - Y_2)} = \sqrt{\frac{|Y_1 - Y_2|}{K}} \qquad \text{où} \tag{II,22}$$

$$K = \sqrt{\frac{n_1 \times n_2 \times n_E}{n_1 + n_2}} \tag{II,23}$$

et finalement :

$$t = K \sqrt{|Y_1 - Y_2|} \tag{II,24}$$

ici $\quad |Y_1 - Y_2| = 11854{,}75896 \times 10^{-6}$

$$\sqrt{|Y_1 - Y_2|} = 0{,}10888$$

$$K = 34{,}4129 \qquad \text{donc}$$

$$t = 3{,}75$$

cette valeur est nettement significative au seuil 1 % (1).

Lorsqu'on avait examiné les caractères un par un, seul x_2 était significatif, avec une valeur de t égale à 2,90 ; le gain est donc assez net.

III — CAS DE PLUSIEURS POPULATIONS RATTACHEMENT D'UN ELEMENT INCONNU A UNE POPULATION DEJA DEFINIE

III. 1 — Cas de plusieurs populations.

L'extension à l'étude de plusieurs populations ne pose pas de difficulté d'ordre théorique si on compare toutes les populations deux à deux. Si k est le nombre des populations, on a pour chaque comparaison des frontières définies par des égalités du type (II,3). Pour simplifier le problème, il faut supposer qu'il existe une matrice de

(1) Il faut tout de même remarquer que ce test n'est qu'approximatif. Cf. KENDALL (M.G.), 1961, op. cit., p. 158-163.

dispersion commune et que les k distributions sont p normales et ne diffèrent que par leurs valeurs moyennes (1).

III. 2 — **Exemple.**

(M. POURTET Dossier Les Barres 63-9).

Il s'agissait de comparer 14 arbres appartenant à différents clones de peupliers, trois caractères avaient été mesurés :

x_1 : longueur des chatons (cm)
x_2 : nombre de fleurs par chaton
x_3 : nombre d'étamines d'une fleur.

Pour les différents arbres, les moyennes suivantes avaient été obtenues :

(III,1)

Clones	x_1	x_2	x_3
1	5,145	67,15	21,95
2	6,695	111,10	33,20
3a	8,235	107,70	34,65
3b	9,085	116,40	31,95
3c	8,165	112,95	34,55
3d	9,345	112,45	33,95
4	8,945	106,20	32,60
5	9,430	119,10	32,95
6	8,315	105,70	32,50
7	8,455	103,20	32,85
8	9,455	112,55	35,35
9a	8,755	102,80	33,00
9b	8,765	115,65	32,55
9c	8,240	110,85	33,05

L'étude analogue à celle faite pour les 2 provenances de pin sylvestre a conduit à tracer un tableau (cf. fig. 1) où chaque clone est représenté par son numéro et où les traits simples et doubles traduisent, à la façon d'un arbre généalogique, les liens plus ou moins grands entre les arbres (2) :

— s'il n'y a aucun trait, les arbres sont différents au sens statistique au seuil 5 %.

(1) Cette façon de résoudre le problème de séparation des k populations n'est évidemment qu'une première approche qui manque d'unité, il faudrait pour faire une étude plus homogène :

— définir une fonction discriminante commune,

— définir à l'aide de cette fonction un test à intervalles multiples plus valable que le test t de Student. Cf. à ce sujet RIVES M. (1959), op. cit.

Mais si la deuxième proposition est logique et n'enlève rien de sa précision au problème, par contre la première risque de masquer une partie des différences qui peuvent exister entre les populations.

(2) L'étude théorique de ces schémas, qui constituent des graphes, peut être envisagée. Cf. BERGE (C.) 1963, op. cit.

— un trait simple : les arbres sont différents au seuil 10 % mais pas à 5 (1).

— un trait double : les arbres ne sont pas différents au seuil 10 %.

On peut introduire ainsi une certaine notion de distance statistique fondée sur une valeur du t de Student (2).

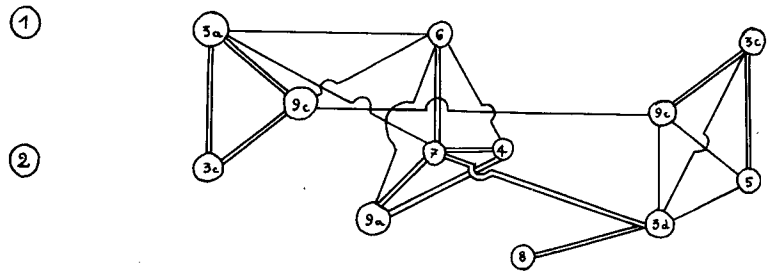


FIG. 1.

Liaison entre 14 arbres appartenant à des clones de peupliers (cf. § III,2)
— Pas de trait : aucune liaison.
— Un trait : liaison faible.
— Deux traits : liaison forte.

On peut raisonnablement penser, d'après la figure 1, qu'il existe 5 groupes dont au moins deux sont très différents des autres (I et II) :

(III,2)

Groupe	Clones	Caractères x_1	x_2	x_3
I	1	5,145	67	22
II	2	6,695	111	32
III	3a, 3c, 9c	8,213	111	34
IV	4, 9a	8,850	105	33
V	3b, 9b	8,925	116	32

De plus, on voit d'après (III,2) que si le groupe I est très différent des autres, c'est le caractère x_1 qui permet de différencier le groupe II des suivants ; et le caractère x_2 est le plus important

(1) Le nombre assez faible de mesures pour chaque arbre (20) ne permet pas d'avoir un seuil plus faible. Les résultats ne sont donnés ici que pour expliquer une méthode et non pour tirer des conclusions d'ordre botanique.

(2) Cette distance n'a rien à voir avec la distance de deux variables définie de façon axiomatique. Cf. DUGUE D. (1958), op. cit., p. 9.

pour la séparation des groupes III, IV, V. Ainsi, l'étude a permis, non seulement de différencier des groupes, mais de connaître mieux les caractères importants de différenciation (1).

III. 3 — **Rattachement d'un élément inconnu à une population préalablement définie.**

Lorsqu'on a réussi à définir des populations relativement homogènes comme au paragraphe précédent, le problème qui se pose immédiatement est le suivant: on a mesuré sur un individu les p valeurs des variables qui ont servi à séparer les populations: à quelle population a-t-il le plus de chance d'appartenir? Le problème est résolu simplement en calculant pour chaque population j (2):

$$Z_j = \mathbf{X}' \, \mathbf{C}^{-1} \, \mathbf{X}_j - 1/2 \, \mathbf{X}'_j \, \mathbf{C}^{-1} \, \mathbf{X}_j \quad (3) \qquad \text{(III,3)}$$

où $\mathbf{X}_j$ est le vecteur colonne dont les coordonnées sont les moyennes de la j^{me} population.

$\mathbf{X}$ est le vecteur colonne dont les coordonnées sont les valeurs observées sur l'individu.

On voit d'après (III,3) qu'on compare l'individu à toutes les populations et qu'on peut donc écrire k équations du type:

$$Z_j = \Sigma_i \, b_{ij} \, x_i + b_{oj} \qquad \text{(III,4)}$$

la valeur Z_j maximum donne l'indice de la population à laquelle l'individu inconnu appartient le plus sûrement.

Sur la dernière équation, on voit qu'il suffit pour résoudre immédiatement le problème de connaître les quantités b_{ij} pour toutes les populations: il faut donc calculer:

1) la quantité constante b_{oj} par:

$$b_{oj} = - 1/2 \, \mathbf{X'j} \, \mathbf{C}^{-1} \, \mathbf{Xj} \qquad \text{(III,5)}$$

(1) Evidemment on a toujours supposé qu'on donnait à chaque caractère la même importance, donc le même « poids »; il pourrait ne pas toujours en être ainsi, par exemple dans des études de résistance à un facteur climatique et de croissance où il faut d'abord qu'une espèce possède le premier caractère puis le second.

(2) j est l'indice mobile qui peut prendre les valeurs 1 à k

(3) On doit, en plus des hypothèses déjà admises, supposer que toutes les populations ont la même probabilité d'apparition; la quantité Z_j s'obtient directement à partir de l'équation donnée en Annexe (AII)

2) les quantités b_{ij} qui sont les coordonnées de k vecteurs (i = 1, 2... p ; j = 1, 2... k)

(III,6) $$\mathbf{L_j} = \mathbf{C}^{-1}\,\mathbf{X_j}$$

III. 4 — **Exemple.**

(M. POLGE exp. hêtre CHRETIENNETTE E 3 B).

L'étude portait sur des peuplements de hêtre répartis en 6 populations correspondant à la forme de la cime et à la place dans les étages de végétation (1).

Pour chaque élément de ces populations, on a mesuré sur des arbres debout :

x_1 : le diamètre de l'arbre (cm)

x_2 : le couple de torsion à la prise d'une carotte (cm × kg)

x_3 : la densité d'un échantillon (g/dm^3)

x_4 : le retrait volumétrique (%).

On désirait savoir si ces 4 caractères étaient valables pour définir chacune des catégories : si un échantillon quelconque permet de retrouver la population dont il est issu, on peut raisonnablement penser que les 4 critères sont bons (2). Les 6 catégories correspondaient aux moyennes :

(III,7)

Catégories	x_1	x_2	x_3	x_4
A_1	55,64	332,14	543,71	20,13
A_2	52,37	356,25	566,25	20,89
B_1	42,77	353,85	567,00	21,57
B_2	32,19	384,38	591,38	22,20
C_2	33,36	381,82	576,91	22,15
C_3	28,62	362,50	578,50	22,05

(1) Pour plus de détails sur ces catégories, cf. AYRAL P. et ABADIE J. (1956), op. cit.

(2) On connaît ici la population dont provient l'individu, car le fait même d'avoir mesuré le diamètre de l'arbre sur pied suppose qu'on l'a approximativement classé dans le peuplement ; on peut toutefois penser que la méthode permet de le classer de façon plus objective.

S'il s'agit d'un arbre abattu, ce qui ne modifie en rien le raisonnement, la méthode réapparaît avec tout son intérêt.

La matrice de dispersion commune a pour valeur :

$$\text{(III,8)} \qquad \mathbf{C} = \begin{vmatrix} 44{,}3183 & -\ 5{,}5502 & 9{,}2306 & -\ 4{,}0228 \\ & 897{,}3037 & 285{,}6493 & 9{,}1731 \\ & & 2091{,}7832 & 10{,}3972 \\ & & & 1{,}3700 \end{vmatrix}$$

et son inverse :

$$\text{(III,9)} \qquad \mathbf{C}^{-1} = \begin{vmatrix} 0{,}03186315 & -\ 0{,}00067117 & -\ 0{,}00055719 & 0{,}10229049 \\ & 0{,}00124476 & -\ 0{,}00012013 & 0{,}00939146 \\ & & 0{,}00052070 & -\ 0{,}00478306 \\ & & & 1{,}12951638 \end{vmatrix}$$

Les coefficients de l'équation (III,4) forment le tableau suivant (1) :

(III,10)

Coefficients / Populations	b_1	b_2	b_3	b_4	b_o
A_1	3,305	0,498	0,115	28,949	- 497,28297
A_2	3,250	0,535	0,123	29,591	- 524,30050
B_1	3,014	0,545	0,125	29,351	- 512,86655
B_2	2,708	0,593	0,137	29,151	- 521,63956
C_2	2,750	0,590	0,130	29,259	- 520,04948
C_3	2,600	0,568	0,136	28,472	- 493,39780

Par exemple pour obtenir les b correspondant à A_1 : pour b_{11} on multiplie les éléments de la première ligne de (III,7) par ceux de la première ligne de (III,9) ; pour les b_{21} on conserve ceux de la première ligne (III,7) mais on les multiplie par ceux de la deuxième de (III,9), et ainsi de suite jusqu'à la quatrième ligne.

Quant au terme constant, c'est la moitié du produit des éléments correspondant à A_1 dans les 2 tableaux (III,7) et (III,10) :

$$497{,}28297 = 1/2\ (3{,}305 \times 55{,}64 + 0{,}498 \times 332{,}14 + 0{,}115 \times 543{,}71 + 28{,}949 \times 20{,}13)$$

(1) Dans cet exemple, j peut varier de 1 à 6.

Sur un échantillon inconnu, on a les 4 mesures suivantes :

45,00 ; 325,00 ; 570,00 ; 23,29

Les valeurs des Z_j pour les 6 populations en découlent :

A_1	A_2	B_1	B_2	C_2	C_3
553,064 ;	*555,109 ;*	554,723 ;	549,962 ;	550,992 ;	548,835.

On peut classer cet échantillon dans le groupe A_2 en considérant toutefois qu'il est assez voisin de B_1. De façon plus précise, on peut le classer avec des probabilités décroissantes dans les groupes suivants :

$$A_2 — B_1 — A_1 — C_2 — B_2 — C_3$$

III. 5 — **Conclusion : Champ possible d'utilisation.**

Il nous paraît maintenant utile de citer les champs possibles d'utilisation de cette méthode dans un domaine plus purement forestier : dans les problèmes de taxonomie son intérêt est évident ; on peut en outre compléter l'étude de classification par une étude plus économique en cherchant dans le cas particulier de l'amélioration des plantes celles à croissance maximum et à plus grande résistance (1).

Un autre aspect non moins intéressant pourrait être celui du classement des communautés végétales. En particulier, les peuplements forestiers pourraient être classés en tenant compte de tous les facteurs de station : facteurs climatiques, teneur du sol en divers éléments, pourcentage des diverses essences, croissance et qualité technologique des essences économiques.

On peut même penser que ces études pourraient se montrer très utiles pour l'aménagement des forêts où le classement des diverses parcelles pourrait être fait à partir de critères reconnus valables d'une façon entièrement objective.

(1) Cf. ELSTON R.C. (1963), op. cit.

ANNEXE I — CALCUL MATRICIEL NOTATIONS ET OPERATIONS ELEMENTAIRES

Cette annexe n'a d'autre but que celui de montrer l'utilisation extrêmement pratique du calcul matriciel et de préciser les notations utilisées dans le texte (1).

AI. 1 — Généralités.

On appelle matrice **A** un tableau de coefficients disposés suivant p lignes et q colonnes ; a_{ij} est l'élément de la ligne i et de la colonne j : i et j sont des indices mobiles qui varient respectivement de 1 à p et de 1 à q.

On a donc un tableau (2) :

$$(AI,1) \qquad \underset{(p,\ q)}{\mathbf{A}} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{vmatrix} = (a_{ij})$$

Un vecteur n'est qu'une matrice particulière, le vecteur colonne correspond à q = 1 :

$$(AI,2) \qquad \underset{(p,\ 1)}{\mathbf{X}} = \begin{vmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{vmatrix}$$

(1) Pour une étude détaillée mais visant surtout à montrer l'intérêt pratique du calcul matriciel, nous signalons deux ouvrages :

J. VIGNAL (1962), op. cit.

A. MONJALLON (1961), op. cit.

(2) Nous noterons ici au-dessous de la matrice ou du vecteur deux nombres entre parenthèses qui correspondent au nombre de lignes et de colonnes ; cette notation peut être fort utile pour un débutant, en particulier lorsqu'on aborde la notion de produit matriciel.

On appelle matrice transposée de **A**, et on note **A'**, la matrice obtenue en échangeant les lignes et les colonnes :

(AI,3) $$\mathbf{A'} = (a'_{ij}) \text{ avec } a'_{ij} = a_{ji}$$

En particulier le vecteur transposé **X'** est le vecteur ligne

(AI,4) $$\underset{(1,\,p)}{\mathbf{X'}} = (x_1\ x_2\ \ldots\ x_p)$$

AI. 2 — **Produit de 2 matrices.**

a) On définit le *produit* de 2 matrices **A** et **B** (à la condition que le nombre de lignes de **B** soit égal au nombre de colonnes de **A**) de la façon suivante :

(AI,5) $$\underset{(p,\,r)}{\mathbf{C}} = \underset{(p,\,q)}{\mathbf{A}} \times \underset{(q,\,r)}{\mathbf{B}}$$

où l'élément c_{ik} de **C** s'écrit :

(AI,6) $$\mathbf{c_{ik}} = \Sigma_j\ a_{ij}\ b_{jk}$$

Pour obtenir le terme d'indice i, k du produit matriciel **A** × **B**, on fait la somme des éléments de la ligne i de **A** par les éléments de la colonne k de **B**.

Si p = r, le produit **B** × **A** est possible, mais cette opération n'est en général pas commutative : **B** × **A** est en général différent de **A** × **B**. C'est pour cette raison qu'on doit parler du produit à gauche de **A** par **B**, ou du produit à droite. Par exemple **X'X** est un nombre, alors que **XX'** est une matrice.

b) *Matrices carrées :* dans ces cas p = q, et on peut alors définir le déterminant de **A** qu'on note d (**A**). Si ce dernier n'est pas nul, on peut en outre définir la matrice inverse de **A** : $\mathbf{A}^{-1}$ telle que :

$$\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{I}$$

où **I** est la matrice unité formée par des 1 sur la diagonale principale et des zéros partout ailleurs.

AI. 3 — **Les calculs élémentaires interprétés par le calcul matriciel.**

En voici 2 exemples :

a) *Système d'équation linéaire* — L'égalité :

(AI,7) $$\underset{(p,\,1)}{\mathbf{Y}} = \underset{(p,\,q)}{\mathbf{A}}\ \underset{(q,\,1)}{\mathbf{X}}$$

traduit l'ensemble des p équations (j varie de 1 à p)

$$(AI,8) \qquad Y_j = a_{j1}\, x_1 + a_{j2}\, x_2 + \ldots\ldots a_{jq}\, x_q$$

On obtient également les x en fonction des y *si* **A** est une matrice carrée (p = q) et si d (**A**) est différent de 0 par:

$$(AI,9) \qquad \mathbf{X} = \mathbf{A}^{-1}\ \mathbf{Y}$$

La connaissance des éléments de $\mathbf{A}^{-1}$ entraîne celle des x_i

b) *Produit scalaire de 2 vecteurs* — c'est le nombre (1) défini par:

$$(AI,10) \qquad \underset{(1,1)}{s} = \underset{(1,p)}{\mathbf{X}'}\ \underset{(p,1)}{\mathbf{X}} = x^2_1 + x^2_2 + \ldots\ldots x^2_p = \Sigma_i\, x^2_i$$

en particulier si A est une matrice carrée:

$$(AI,113) \qquad \underset{(1,p)}{\mathbf{X}'}\ \underset{(p,p)}{\mathbf{A}}\ \underset{(p,1)}{\mathbf{X}} = \Sigma_{ij}\ a_{ij}\ x_i\ x_j$$

Par exemple si on a n mesures d'une même grandeur, on peut définir un vecteur dont les coordonnées sont les n mesures.

Si **G** est le vecteur représentatif de la moyenne, la variance d'une mesure a pour valeur:

$$(AI,12) \qquad V = \frac{1}{n-1}\ (\mathbf{X}-\mathbf{G})'\ (\mathbf{X}-\mathbf{G})$$

AI. 4 — **Programmation sur machine électronique.**

Nous n'insisterons pas davantage sur ces quelques notations de calcul matriciel mais on peut à ce stade là en voir tout l'intérêt:

— d'abord l'extrême concision des formules.

— d'un point de vue pratique et dans l'optique d'une utilisation de machine électronique, il suffit de donner à la machine l'ordre suivant (2):

45 N OA OB OC

(1) On le retrouve symboliquement puisque c'est une matrice à 1 ligne et 1 colonne.

(2) Cf. ARBONNIER P. (1962), op. cit., p. 214: il s'agit ici du code de programmation de la Faculté de Nancy sur calculatrice 650 IBM.

A est l'adresse du 1er élément de la matrice

B est l'adresse du 1er mémoire de travail

C caractérise le système (dimensions de la matrice).

La matrice est inversée, son déterminant calculé (1) et on passe à l'ordre situé dans la mémoire N

(1) On a intérêt à connaître la valeur du détérminant. Cf. Annexe III.

ANNEXE II — **LOI NORMALE A p DIMENSIONS**

AII. 1 — **Rappel.**

La loi normale à une variable dépend de deux paramètres, sa valeur moyenne μ et son écart type σ ; la loi élémentaire a pour expression (1) :

$$\text{(AII,1)} \quad \text{Prob.}\,(x \leqslant X < x + dx) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

La loi normale à p dimensions ou loi p-normale est une extension de la loi précédente qui prend une forme simple en utilisant la notation matricielle. Si $\mathbf{C}^{-1}$ est une matrice carrée (p,p) et $\mathbf{X}$ un vecteur (p,1) dont les composantes sont les p variables aléatoires X_1, X_2, ... X_p. La loi normale élémentaire de l'ensemble de ces variables s'écrit (2) :

$$\text{(AII,2)} \quad \frac{\sqrt{d\,(\mathbf{C}^{-1})}}{(2\pi)^{p/2}} \exp\,(-1/2\ \mathbf{X}'\,\mathbf{C}^{-1}\,\mathbf{X})\ dX_1\ dX_2\ \ldots\ dX_p$$

La matrice inverse $\mathbf{C}$ de $\mathbf{C}^{-1}$ est la matrice de dispersion définie en (II,1).

(1) Le symbole exp (a) est équivalent à e^a.

(2) Les variables sont supposées centrées.

Pour une étude approfondie de la loi normale, cf.

KENDALL M.G. and STUART A. (1958), op. cit., p. 347-348.

CRAMER H. (1954), op. cit., p. 310-319.

Pour une étude plus rapide mais néanmoins suffisante pour un biométricien, cf.

VESSEREAU A. (1960), op. cit., p. 55-59 et p. 466-475.

La loi p-normale constitue assez souvent une bonne approximation pour les distributions réelles, cf.

WOO T.L. and ELDERTON E.M. (1932), op. cit.

Cas de deux dimensions: dans le cas où p = 2, nous allons retrouver des résultats classiques; on a alors:

$$((\text{AII,3}) \qquad \mathbf{C} = \begin{vmatrix} \sigma^2_x & \rho\ \sigma_x\ \sigma_y \\ \rho\ \sigma_x\ \sigma_y & \sigma^2_y \end{vmatrix}$$

On peut vérifier en utilisant la propriété $\mathbf{C}^{-1}\ \mathbf{C} = \mathbf{I}$ que:

$$(\text{AII,4}) \qquad \mathbf{C}^{-1} = \begin{vmatrix} \dfrac{1}{\sigma^2_x\ (1 - \rho^2)} & \dfrac{-\rho}{\sigma_x\ \sigma_y\ (1 - \rho^2)} \\ \dfrac{-\rho}{\sigma_x\ \sigma_y\ (1 - \rho^2)} & \dfrac{1}{\sigma^2_y\ (1 - \rho^2)} \end{vmatrix}$$

et la distribution liée de x et y a pour expression:

$$(\text{AII,5}) \qquad \text{Prob}\ (x \leqslant X < x + d\,x,\quad y \leqslant Y < y + d\,y) =$$

$$\frac{dx\ dy}{2\ \pi\ \sigma_x\ \sigma_y \sqrt{1 - \rho^2}} \exp \left[\frac{-1}{2\ (1 - \rho^2)} \left[\frac{(x - \mu_x)^2}{\sigma^2_x} - \frac{2\ \rho\ (x - \mu_x)\ (y - \mu_y)}{\sigma_x\ \sigma_y} + \frac{(y - \mu_y)^2}{\sigma^2_y} \right] \right]$$

AII. 2 — **Application aux fonctions discriminantes.**

Nous démontrons ici le résultat trouvé au paragraphe II,2 b en utilisant une méthode plus générale. Nous avons vu que la frontière entre les régions a_1 et a_2 était définie par l'équation (II,A); dans le cas où les populations sont p-normales aux constantes près on a:

$$(\text{AII,6}) \qquad \begin{cases} f_1 = \exp\ [-1/2\ (\mathbf{X} - \mathbf{m}_1)'\ \mathbf{C}^{-1}\ (\mathbf{X} - \mathbf{m}_1)] \\ f_2 = \exp\ [-1/2\ (\mathbf{X} - \mathbf{m}_2)'\ \mathbf{C}^{-1}\ (\mathbf{X} - \mathbf{m}_2)] \end{cases}$$

où $\mathbf{m}_1$ et $\mathbf{m}_2$ sont les vecteurs représentatifs des centres de gravité de A_1 et A_2.

On peut évidemment écrire que le logarithme népérien du rapport est constant pour obtenir la surface frontière :

(AII,7)

$$\text{Log}\,\frac{f_1}{f_2} = -\,1/2\left[\,(\mathbf{m_2}-\mathbf{m_1})'\,\mathbf{C}^{-1}\,\mathbf{X} + \mathbf{X}'\,\mathbf{C}^{-1}\,(\mathbf{m_2}-\mathbf{m_1}) + \mathbf{m}'_1\,\mathbf{C}^{-1}\,\mathbf{m_1} + \mathbf{m}'_2\,\mathbf{C}^{-1}\,\mathbf{m_2}\right]$$

La matrice $\mathbf{C}^{-1}$ est symétrique et si on pose $\mathbf{L} = \mathbf{C}^{-1}\,(\mathbf{m}_2 - \mathbf{m}_1)$, on retrouve bien comme fonction discriminante la quantité donnée en (II,14).

En comparant l'élément inconnu, de densité de répartition f, à la première population on trouve, aux termes constants près :

(AII,8) $$\text{Log}\,\frac{f}{f_1} = Z_1 = \mathbf{L'X} - 1/2\;\mathbf{m}'_1\,\mathbf{C}^{-1}\,\mathbf{m}_1$$

et en le comparant à la seconde :

(AII,9) $$\text{Log}\,\frac{f}{f_2} = Z_2 = \mathbf{L'X} - 1/2\;\mathbf{m}'_2\,\mathbf{C}^{-1}\,\mathbf{m}_2$$

Si $|Z_1| > |Z_2|$, $f_1 > f_2$ on dit que l'élément inconnu appartient à la première population.

ANNEXE III — UN TEST GLOBAL D'HOMOGENEITE DES VARIANCES ET DES COVARIANCES

Dans les problèmes statistiques à p variables on suppose qu'il existe une matrice de dispersion commune: cette supposition n'a rien d'évident: des tests ont été étudiés pour éprouver cette hypothèse. Nous nous proposons donc de comparer un ensemble de matrices de dispersion et de tester leur homogénéité; ce problème constitue l'extension logique de la comparaison globale d'un groupe de variances, pour celle-ci il existe plusieurs tests, le meilleur paraît être celui de Bartlett (1).

Il est évident que dans les problèmes à plusieurs variables la difficulté est plus grande; on peut en outre se demander si un critère unique est suffisant pour donner une réponse à un problème aussi complexe. Le critère qui va être exposé peut donner une réponse assez valable (2).

AIII. 1 — **Le critère de vraisemblance généralisé de Wilks** (3).

Nous rappelons les notations déjà utilisées ci-dessus:

- nombre de populations examinées: k
- nombre d'éléments d'une population: n
- nombre de caractères mesurés: p

donc le nombre total d'individus examinés est:

$$N = nk$$

Chaque population possède une matrice $\mathbf{C}_j$; le critère ne s'intéresse qu'au déterminant de cette matrice (4) que nous noterons D_j ($j = 1, 2, \ldots k$).

(AIII,1) $$D_j = d\ (\mathbf{C_j})$$

(1) Pour l'utilisation de ce test, cf. VESSEREAU A. (1960), op. cit., p. 157-158.

(2) Cf. BISHOP D.J. (1939), op. cit.

(3) Cf. WILKS J.J. (1932), op. cit.

(4) On donne quelquefois à ce déterminant le nom de variance généralisée.

On appelle déterminant moyen D :

$$D = \frac{1}{k} \Sigma_j D_j \qquad \text{(AIII,2)}$$

On démontre que la quantité l définie par :

$$l^2 = \frac{\left[[D_1 \times D_2 \times \ldots D_k] \right]^{\frac{1}{k}}}{D} \quad ; \quad 0 \leqslant l \leqslant 1 \qquad \text{(AIII,3)}$$

suit approximativement une distribution B (distribution Bêta) (1) de paramètres a et b dont des valeurs approchées sont :

$$\text{(AIII,4)} \quad \begin{cases} a = k\,(n - p) - \frac{1}{100}\,(k - 1)\,(90 - 39\,p + p^2) \\ b = \frac{1}{4}\,p\,(p + 1)\,(k - 1) \end{cases}$$

L'hypothèse d'égalité des matrices de dispersion est rejetée si, au seuil de probabilité choisi, la valeur calculée de l est inférieure à celle lue sur les tables.

AIII. 2 — **Cas de grands échantillons.**

On démontre (2) que si n est assez grand (au moins supérieur à 20), la quantité — 2 N Log l est distribuée comme un χ^2 à 2 b degrés de liberté (3).

Ce résultat est intéressant car aux seuils couramment choisis et avec un grand nombre de mesures, les valeurs données par la table des Bêta sont peu précises.

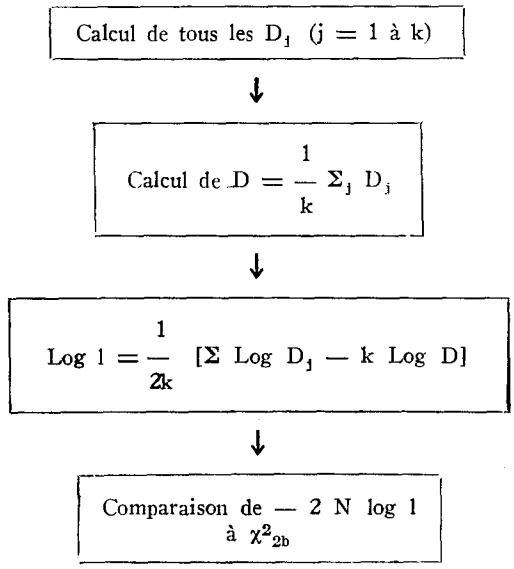
(1) B est la fonction eulérienne de première espèce, pour une étude théorique, cf. CRAMER H. (1954), op. cit., p. 243-244. Etant donné son importance dans de nombreux problèmes statistiques, elle a été tabulée, cf. PEARSON, op. cit.

(2) Cf. BISHOP D.J. (1939), op. cit., p. 45.

(3) Le logarithme noté Log. est un logarithme népérien (à base e = 2,71...) ; celui noté log est le logarithme à base 10 usuel.

Ainsi les différentes étapes du calcul mises sous forme d'organigramme utilisable pour la programmation sur machine électronique sont les suivantes :

(AIII,5)



AIII. 3 — **Exemple.**

(M. POURTET déjà cité § III,2)

Dans cet exemple $p = 3$; $n = 20$; $k = 14$

les valeurs des 14 déterminants sont les suivantes :

58380
414623
2379008
1674598
569238
1733453
1763263
3084479
803435
4728333
1050833
1470889
2448142
790495

qui permettent de calculer un déterminant moyen égal à 1640661, d'où on déduit :

$$N = 280$$
$$\text{Log } 1 = -0{,}18393$$
$$2\,b = 78$$

ainsi on obtient un χ^2 calculé = 103,0 alors que $\chi^2_{0,05} = 101{,}9$ et $\chi^2_{0,01} = 106{,}6$.

On peut donc admettre que les 14 déterminants, différents au seuil 5 %, ne le sont pas au seuil 1 %.

BIBLIOGRAPHIE

ARBONNIER (P.). — (1962) : Le forestier à l'heure de l'électronique. R.F.F., n° 3.

AYRAL (P.) et ABADIE (J.). — (1956) : Méthode de calcul du volume des peuplements sur pied dans les places d'essai de sylviculture. Annales E.N.E.F., Tome XV - 1.

BARTLETT (M.-S.). — (1939) : The standard errors of discriminant function coefficients. J. Roy. Statist. Soc. Supp. 6-169.

BARTLETT (M.-S.) and PLEASE (N.-W.). — (1963) : Discrimination in the case of zero mean differences. Biometrika L, p. 17/21.

BERGE (C.). — (1963) : Théorie des graphes et ses applications. Dunod, Paris.

BISHOP (D.-J.). — (1939) : On a comprehensive test of homogeinity of variances and covariances in multivariate problem. Biometrika XXXI, p. 31-55.

CRAMER (H.). — (1954) : Mathematical Methods of Statistics ; Princeton 2e édition.

DUFRENOY (J.). — (1963) : Commémoration du Bicentenaire de la publication par Adanson de « familles de plantes ». Cahier des Ingénieurs agronomes, n° 175.

DUGUE (D.). — (1958) : Traité de Statistique Théorique et Appliquée. Masson. Paris.

ELSTON (R.-C.). — (1963) : A weight-free index for the purpose of ranking or selection with respect to several traits at a time. Biometrics, Vol. 19, n° 1, p. 85-97.

FAVERGE (J.-M.). — (1950) : Introduction aux méthodes statistiques en psychologie appliquée. P.U.F. Paris.

FERIGNAC (P.). — (1962) : Test de Kolmogorov - Smirnov sur la validité d'une fonction de distribution. Revue de Statistique Appliquée. Vol. X, n° 4.

FISHER (R.-A.). — (1936) : The use of multiple measurements in taxonomic problems. Ann. Eug. London 7, 179-188.

JEFFERS (J.-N.-R.). — (1960) : Experimental Design and analysis in Forest Research. Uppsala.

JEFFERS (J.-N.-R.). — (1961) : An analysis of Variability in *Pinus contorta*. Statistics Section Paper n° 23. Forestry Commission - Alice Holt.

JOLLY (G.-M.). — (1963) : Estimates of population parameters from multiple recapture data with both death and dilution deterministic model. Biometrika, Vol. L 1.2, p. 113-128.

KENDALL (M.-G.). — (1961) : A course in multivariate analysis. Griffin, Londres, 2e édition.

KENDALL (M.-G.) and MORAN (P.-A.-P.). — (1963) : Geometrical Probability. Griffin, Londres.

KENDALL (M.-G.) and STUART (A.). — (1958): The Advanced Theory of Statistics (3 vol.). Griffin, Londres.

KUIPER (W.-H.). — (1960): Random variables and random vectors. Bull. Inst. Agr. Stat. Rech. Gembloux. Hors Série. Vol. I, p. 344-355.

LUBISCHEW (A.-A.). — (1962): On the use of discriminant functions in taxonomy. Biometrics, 18-4, p. 455-477.

MARTIN (E.-S.). — (1936): A study of an Egyptian series of mandibles with special reference to mathematical methods of sexing. Biometrika, XXXVIII, p. 149-172.

MONJALLON (A.). — (1961): Initiation au calcul matriciel. Vuibert. Paris, 3e édition.

RIVES (M.). — (1959): Sur la comparaison des moyennes dans les essais variétaux. Annales INRA, 9e année, n° 3, p. 357-376.

STOESSIGER (B.-N.) and MORANT (G.-M.). — (1932): A study of the crania in the vaulted ambulatory of Saint-Leonard's church Hythe. Biometrika, XXIV, p. 135-202.

VESSEREAU (A.). — (1960): Méthodes statistiques en biologie et en agronomie. Baillère. Paris, 2e édition.

VIGNAL (J.). — (1961): Calcul matriciel. Vuibert. Paris.

WELCH (PH.-D.). — (1939): Note on discriminant functions. Biometrika, XXXI, p. 218-220.

WILKS (S.-S.). — (1932): Certain generalisations in the analysis of variance. Biometrika, XXIV, p. 471-494.

WOO (T.-L.) and ELDERTON (E.-M.). — (1932): On the normality or want of normality in the frequency distributions of cranial measurements. Biometrika, XXIV, p. 45-54.

WOO (T.-L.) and MORANT (G.-M.). — (1932): A preliminary classification of Asiatic Races based on cranial measurements. Biometrika, XXIV, p. 108-134.

Tables of the Incomplete Beta function, edited by the Biometrika Office under the direction of PEARSON (K.), 1948.

RÉSUMÉ

Comme toute science en pleine croissance, la biologie a besoin pour s'affirmer, d'un outil précis et objectif. La résolution des problèmes biométriques fait de plus en plus appel à des notions de statistiques mathématiques peu habituelles dans ce domaine. L'analyse statistique à plusieurs variables, ou analyse multivariable, est l'un de ces outils et l'un des plus puissants.

Quel que soit l'aspect sous lequel est étudiée l'analyse multivariable, on conserve l'idée de base suivante : un élément quelconque d'une population (par exemple les arbres appartenant à un même clone) est défini par un ensemble de caractères. Alors que chacun des caractères pris isolément suffit rarement à établir des critères valables de reconnaissance, l'ensemble, avec les liens qui peuvent exister entre ces caractères, permet une analyse plus fine.

Un des champs d'application de l'analyse multivariable est constitué par l'étude des fonctions discriminantes qui font l'objet de cet article : on cherche la meilleure relation linéaire des caractères d'origine pour séparer des populations. Le critère ainsi utilisé est absolument objectif. Après un rapide exposé théorique dans le cas de deux populations, les résultats obtenus sont appliqués à des données expérimentales. Les résultats sont ensuite étendus à la séparation de plusieurs populations. L'aspect complémentaire est alors étudié : une fois les populations isolées, il est nécessaire, lorsqu'on retrouve un individu quelconque, de pouvoir le rattacher à une des populations préalablement définie. Un exemple est traité de cette façon.

Enfin, d'autres possibilités d'utilisation sont suggérées.

En annexe au texte, les résultats nécessaires à la compréhension de l'article sont rappelés ; ils concernent : le calcul matriciel et la loi normale à p dimensions. En outre, on utilise un test d'ensemble d'homogénéité des matrices de dispersion.

SUMMARY

Like every science in full growth, biology needs an accurate and objective tool to assert itself. To solve biometric problems, it is more and more necessary to resort to notions of mathematical statistics, little used in that field. Multivariate analysis is one of these tools and one of the most powerful.

Whichever the aspect under which multivariate analysis is studied, the following basic idea remains: any element of a population (for exemple trees belonging to the same clone) is defined by a group of characters. Whereas each separate character seldom unables to set up reliable determining criterions, this group together with the relationships which may occur between characters, makes a more subtle analysis possible.

One field of application of multivariate analysis consists in the study of discriminant functions which are dealt with in this article: the best linear relationship between original characters is searched for, so as to separate populations. The criterion thus utilized is entirely objective. After a brief theoretical account of a case with two populations, the results obtained are applied to experimental data. The results are further extended to the seperation of several populations. Then, the complementary aspect is studied; once the populations have been isolated, it is necessary, when finding an individual, to be able to relate it to one of the populations formerly determined. An example is treated in this way.

Finally, other possible uses are suggested. The results necessary to the understanding of the article are reviewed and appended to the text: they deal with matricial calculation and multivariate normal distribution. Besides, a comprehensive test of homogeinity of dispersion matrix is used.

ZUSAMMENFASSUNG

Wie jede in voller Entfaltung stehende Wissenschaft, so gebraucht die Biologie ein genaues und objektives Werkzeug um sich zu behaupten. Die Lösung der biometrischen Aufgaben wird immer mehr von vielfachmathematischen Kenntnissen, wenig in diesem Fache benutzt, Gebrauch machen. Die mehrfache Korrelationsanalysis, oder vielfache Analysis, ist eines-und gewaltigsten-von diesen Werkzeugen.

Welch auch die Ansicht sei, mit der die mehrfache Korrelationsanalysis ergründet wird, behält man immer den nachstehenden Grundsatz: ein etwaiges Einzelwesen einer Bevölkerung (z.B. die Bäume die einem selben Klon angehören) ist durch einen Zusammenschluss von Kennzeichen festgesetzt. Alsdann eine jede einzelgenommene Eigenschaft selten genügt gültige Erkennungskriterien aufzubringen, deren Gesamtheit, mit all den möglichen Verwandschaften unter diesen Kennzeichen, gestattet aber eine genaue Zergliederung.

Das Studium der Trennfunktionen, als eines der Anwendungsgebieten der Korrelationsanalysis, bildet den Inhalt dieses Artikels: man sucht das geeigneste Linearverhältnis betreffs der ursprünglichen Charakteren um die Populationen zu trennen. Das so angewendete Unterscheidungsverfahren ist gänzlich objektiv. Nach einem kurzen theoretischen Uberblick im Falle zweier Populationen, werden die erreichten Resultaten an experimentalen Angaben geprüft, alsdann an der Trennung von mehreren Populationen erstreckt. Dann kommt die ergänzende Studie: nach erreichter Isolierung der Populationen, wenn man auf ein etwaiges Einzelwesen stösst, ist es allererst notwendig es an eine der vorher bestimmten Populationen zuteilen zu können. Ein Beispiel ist auf diese Art ergründet.

Endlich sind noch andere Anwendungsmöglichkeiten angegeben. Zum Textanhang, und dies zur Verständigung des Artikels, sind die notwendigen Begriffe wiederholt; sie betreffen die Matrizenberechnung und die mehrdimensionale Normalverteilung. Ausserdem benutzt man einen Homogenitätstest der Dispersionsmatrizen.

Imprimerie Georges Thomas - Nancy

Dépôt légal IV-1963 - N° 615