Bull. Soc. Roy. Belge Anthrop. Préhist., 81 : 125-156, 1970

LES MÉTHODES DE L'ANALYSE MULTIVARIÉE EN ANTHROPOLOGIE (*)

par

André LEGUEBE

Institut royal des Sciences naturelles de Belgique

INTRODUCTION

Une recherche anthropologique passe nécessairement par une série d'étapes : la première consiste à poser une question, à définir le problème qu'on envisage de traiter, la seconde nous amène à choisir le ou les échantillons sur lesquels on prélèvera les données correspondant aux variables sélectionnées, la troisième nous conduit à établir entre les données recueillies des relations qui expriment la ou les hypothèses qui peuvent fournir une réponse au problème posé, la dernière doit permettre de prendre une décision en faveur ou contre les hypothèses avancées, ou indiquer dans quel sens la recherche doit être poursuivie.

Nous nous proposons de présenter dans cet article un tableau général des différentes méthodes statistiques mises à notre disposition en essayant de faire ressortir les relations qui les unissent tant du point de vue théorique que du point de vue pratique.

Nous avons affaire aux éléments d'un ensemble, les hommes appartenant à telle population ou les populations occupant telle aire géographique par exemple. Il ne nous est généralement pas possible de définir les ensembles en énumérant tous les éléments qui les constituent : on les définit donc à l'aide d'une propriété caractéristique qui doit permettre de classer sans ambiguïté chaque élément comme appartenant ou n'appartenant pas à l'ensemble.

D'autre part, les mesures ne porteront que sur une partie des éléments de l'ensemble, un échantillon, et les paramètres que nous

(*) Communication présentée le 26 janvier 1970.

calculerons, correspondront donc à des *estimations* des paramètres de l'ensemble de la population.

Les mesures se rapportent soit à des variables à caractère discontinu soit à des variables à caractère continu. Sous le rapport de ces dernières (taille, indice céphalique, etc...), il est possible de ranger les éléments selon un ordre de grandeur déterminé. Sous le rapport des premières, les caractères descriptifs en sont des exemples, on peut seulement ranger chaque élément dans une des catégories sans pouvoir nécessairement établir une hiérarchie des catégories.

Les variables à caractère continu sont du point de vue statistique beaucoup plus faciles à traiter et ce sont celles que nous envisagerons dans la suite de cet exposé. La mesure d'une variable est affectée d'une erreur de mesure qui doit être distinguée des variations interindividuelles ou même intra-individuelles.

Pour des raisons pratiques, j'établirai dans l'ensemble des problèmes trois subdivisions consacrées respectivement à :

1. l'analyse des caractéristiques d'un seul échantillon.
2. la comparaison de deux ou plusieurs échantillons.
3. l'établissement de classifications :
   31. soit qu'il faille fixer la position d'un individu ou d'une population par rapport à des populations définies a priori
   32. soit qu'on veuille hiérarchiser des éléments ou structurer un ensemble, c'est-à-dire qu'on cherche à constituer à l'intérieur de cet ensemble des groupes qui n'auront pas été définis a priori.

1. **Analyse d'un échantillon**

Si nous recueillons pour chacun des éléments constituant l'échantillon les valeurs de différentes variables, il nous est possible de considérer chacune des variables séparément ou d'étudier les relations de ces variables par groupes de deux, de trois ou dans leur ensemble.

11. Une seule variable.

Les valeurs de la variable pour les divers individus composant l'échantillon peuvent être résumées par la *moyenne* et la *variance* (ou

sa racine carrée : écart-type) qui valent respectivement :

$$\bar{x} = \frac{\Sigma x}{N} \qquad \text{et} \qquad s_x^2 = \frac{\Sigma (x - \bar{x})^2}{N}$$

Il est important de s'assurer au moyen des tests adéquats de la normalité de la distribution, la condition de normalité étant souvent essentielle à l'utilisation des méthodes dont nous parlerons dans la suite.

Si la distribution n'est pas normale, il faudra, pour utiliser cette variable, lui appliquer une transformation qui la rendra normale. La notion de normalité d'une distribution peut être étendue à plusieurs variables.

La variance, tirée de la somme des carrés des différences entre les observations et la moyenne de l'échantillon, mesure l'étalement des observations autour de la moyenne ou la moyenne du carré des distances des observations à la moyenne.

Une technique importante applicable à l'étude d'une variable est *l'analyse de la variance* : elle permet notamment de fractionner la variation totale de façon à tester l'influence exercée par diverses causes (mesures effectuées par plusieurs observateurs, échantillon prélevé à deux endroits différents, sexe, classe sociale, etc...) et à exercer un contrôle préalable de l'effet d'une autre variable.

12. Deux variables (x, y).

L'étude de la relation existant entre deux variables (x, y) exige que soient précisées la nature des variables et la forme de la relation existant entre celles-ci.

Sous le rapport de la nature, on parlera, de façon relative d'ailleurs, de variables dépendante et indépendante d'une part, et de variables interdépendantes d'autre part. Les variations de la variable aléatoire (y) résultent des variations non aléatoires de la variable indépendante (x) qui joue le rôle de cause de la variabilité observée pour la première : les variations de la variable indépendante, au contraire, ne sont pas conditionnées par celle de la variable dépendante. Les deux variables seront interdépendantes quand elles exercent l'une sur l'autre des actions réciproques ou quand leurs variations traduisent l'effet de l'action d'une ou de plusieurs causes communes.

Sous le rapport de la forme, on peut envisager une relation li-

néaire, $y = a + bx$, forme la plus généralement traitée et à laquelle nous nous limiterons, mais il existe d'autres formes (exponentielle, logarithmique, quadratique, hyperbolique, cubique, quartique).

Nous utiliserons, à titre d'exemple, les mensurations de la longueur (x) et de la largeur (y) de la tête mesurées sur un échantillon de Sango, caractérisé par :

Sango	Moyenne	Variance	Écart-type
largeur de la tête	150,31	26,051	5,104
longueur de la tête	193,17	36,845	6,070

Dans le cas d'une relation linéaire de dépendance, y dépendant de x, la *droite de régression* de y en x, dont l'équation est

$$y - \bar{y} = \frac{r.s_y}{s_x} \cdot (x - \bar{x})$$

permet de prédire quelle sera la valeur de y correspondant à une valeur de x déterminée ; cette droite est telle que la somme des carrés des distances des points à la droite, parallèlement à l'axe des x, est minimum (fig. 1 a).

On aura donc recours à cette relation si on désire déduire la valeur la plus probable de la largeur de la tête d'un Sango à partir de la valeur de la longueur. Pour prédire la valeur de la longueur à partir de la connaissance de la valeur de la largeur, on se servira de la droite de régression de x (longueur) en y (largeur) (fig. 1b).

Dans le cas d'une relation linéaire d'interdépendance entre deux variables, plutôt que de considérer les droites de régression de y en x (fig. 1a) et de x en y (fig. 1b), on y substitue *la droite de Teissier* (1948) :

$$\frac{y - \bar{y}}{s_y} = \frac{x - \bar{x}}{s_x}$$

qui est ajustée à l'ensemble des points de l'échantillon de façon à ce que la somme des produits des distances de chaque point à la droite, parallèlement à l'axe des x et à l'axe des y, soit minimum (moindres rectangles) (fig. 1c). L'emploi de la droite de Teissier est avantageux pour étudier l'évolution de deux caractères ou lorsque les coordonnées sont des logarithmes ; elle ne tient toutefois pas compte du fait qu'une partie de la variation de chacune des variables est liée aux

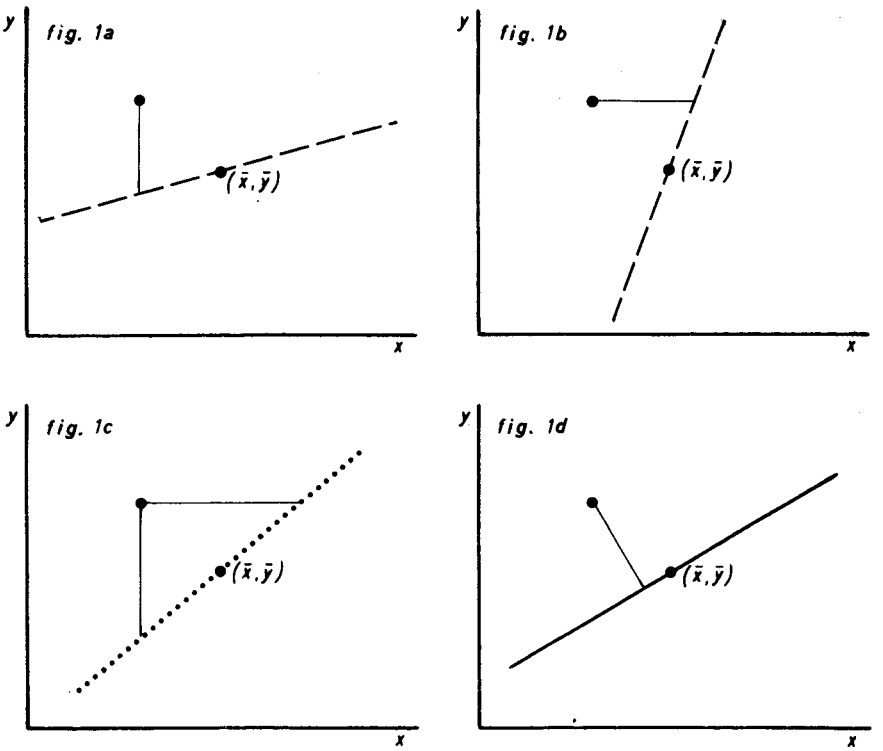

FIG. 1

variations de l'autre, c'est-à-dire que la variation de l'une des variables entraîne une variation déterminée de l'autre. C'est *la covariance,* cov (x,y) dont la valeur est égale à la somme des produits, pour chacun des N points, des distances de chaque coordonnée à sa moyenne, somme divisée par le nombre de sujets N, soit :

$$\text{cov } (x, y) = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{N}$$

Si les deux variables varient dans le même sens, la covariance est positive ; si l'une des variables augmente lorsque l'autre diminue, la covariance est négative ; la covariance ne prend sa pleine signification que si on l'envisage en relation avec les variances des deux variables, soit :

$$r = \frac{\text{cov } (x,y)}{s_x \cdot s_y}$$

r étant le coefficient de corrélation ($-1 \leqslant r \leqslant 1$).

En tenant compte de ces paramètres, il est possible de dessiner les ellipses équiprobables théoriques relatives à la population, ayant comme centre le point $\bar{x}$, $\bar{y}$: le grand axe de ces ellipses est tel que la somme des carrés des distances des points à ce grand axe (perpendiculaire abaissée de chaque point sur l'axe) est minimum (fig. 1d).

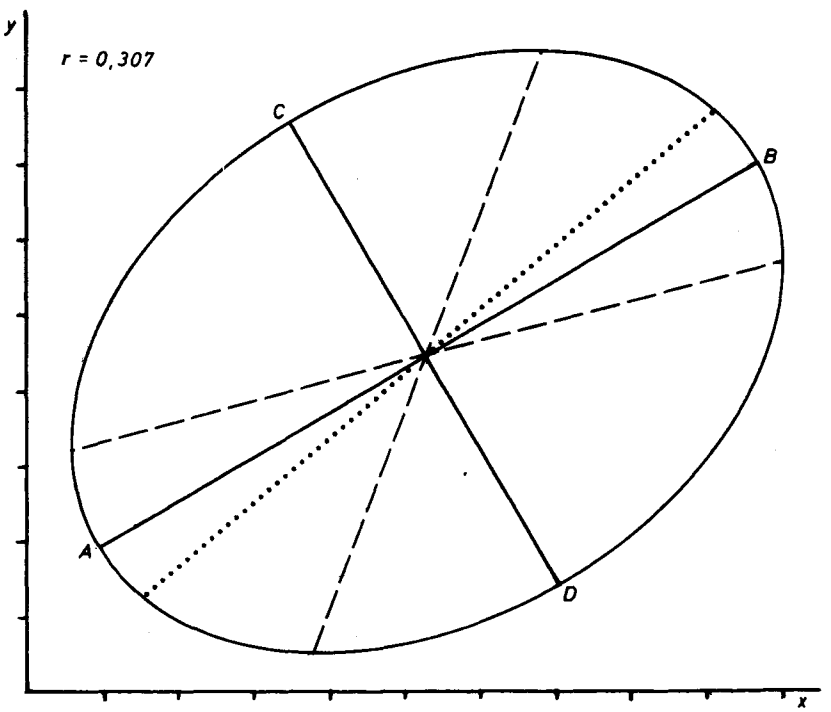

FIG. 2

Les ellipses équiprobables (Defrise, 1955) ont les propriétés suivantes :

a) chaque ellipse contient un pourcentage déterminé de sujets de l'échantillon.

b) tous les points situés sur une même ellipse sont également « distants » du centre de l'ellipse en ce sens que les individus A, B, C, D ont la même probabilité de faire partie de l'échantillon considéré (fig. 2).

c) l'ellipse nous renseigne non seulement sur les dimensions du sujet considéré mais sur sa forme : ainsi les sujets A et B ont des mesures très différentes mais ils ont même forme alors que les sujets C et D qui ont des mesures plus voisines ont cependant des formes plus différentes.

On peut observer sur la figure 2 représentant l'ellipse équiprobable contenant 95 % des Sango de l'échantillon, la position des différentes droites définies plus haut ; on notera aussi que la forme de l'ellipse équiprobable dépend des unités choisies sur les axes des coordonnées.

13. Trois variables.

La distinction entre régression, relation fonctionnelle et structure est essentielle et Kendall a souligné la confusion introduite par l'emploi du terme d'« erreur » pour qualifier les discordances entre modèle et observation (Quenouille, 1952 ; Kendall, 1957, p. 52 ; Dagnelie, 1967).

A. *Une variable aléatoire considérée comme dépendant de deux variables indépendantes.*

Il s'agit donc d'envisager la *régression* de la première variable, y, par rapport aux deux autres, x_1 et x_2, c'est-à-dire de déterminer la relation qui fournisse la meilleure prédiction de y en fonction des valeurs de x_1 et x_2.

La relation la plus simple qui se puisse établir dans ce cas s'exprime par une fonction linéaire correspondant à un plan :

$$y = a + bx_1 + cx_2$$

mais on ne peut pas toujours négliger l'existence de relations entre les deux variables dites indépendantes, auquel cas la fonction doit comporter des termes en $x_1.x_2$, $x_1^2.x_2$, etc. correspondant à l'interaction des variables indépendantes.

La carte donnant la distribution géographique des fréquences relatives d'un caractère ou des fréquences géniques d'un allèle (variable dépendante y) en fonction des coordonnées géographiques (x_1 et x_2) constitue une représentation graphique d'une telle relation.

B. *Trois variables aléatoires* (x, y, z).

Dans ce cas, on peut étendre les méthodes d'analyse utilisées dans le cas de deux variables et envisager les corrélations deux à deux

des trois variables ; ces *corrélations totales* sont au nombre de trois (r_{xy}, r_{xz}, r_{yz}).

Il est également possible d'éliminer l'influence exercée par une des variables sur la corrélation existant entre les deux autres : on obtiendra ainsi une *corrélation partielle*. Il y a trois corrélations partielles : $r_{xy.z}$, $r_{yz.x}$ et $r_{xz.y}$.

Finalement, on peut estimer au moyen d'une *corrélation multiple* la relation qui unit les variations d'une des variables aux variations des deux autres. Il en résulte trois corrélations multiples : $R_{x.yz}$, $R_{y.xz}$ et $R_{z.xy}$.

Il n'est toutefois plus possible de donner une représentation graphique tenant compte des trois variables simultanément et il devient difficile d'embrasser d'un seul coup d'œil le grand nombre de relations établies.

14. Plus de trois variables (p = nombre de variables).

Au fur et à mesure que le nombre de variables augmente on se trouve rapidement devant l'impossibilité non pas d'effectuer les calculs nécessaires, mais de les interpréter puisqu'il y a

$$C_p^2 = \frac{p\,!}{(p-2)\,!2\,!}$$ corrélations totales.

On peut évidemment avoir recours à un biais consistant par exemple à combiner deux ou même trois variables sous forme d'un indice (indice céphalique, indice de robustesse) qui sert lui-même de nouvelle variable. L'utilisation de ce biais a parfois conduit à des résultats très significatifs mais les indices n'en restent pas moins sujets à de nombreuses critiques : il est certain que la simplification résultant de leur emploi s'accompagne de la perte d'une partie de l'information contenue dans les données primitives et introduit un cause d'erreur supplémentaire pour la suite des opérations. La multiplicité des variables offre une autre possibilité qui n'est signalée que pour la forme car elle n'a pas encore reçu d'application en anthropologie : il s'agit des coefficients de corrélation canonique qui mesurent l'intensité des liaisons existant entre deux groupes de variables (Kendall, 1957, p. 68 à 85 ; Dagnelie, 1967).

La matrice des variances et des covariances et la matrice des corrélations expriment les relations existant entre une série de variables interdépendantes et se présentent respectivement de la façon suivante :

$$\begin{vmatrix} \text{var } (x_1) & \text{cov } (x_1, x_2) & \dots & \text{cov. } (x_1, x_p) \\ \text{cov } (x_2, x_1) & \text{var } (x_2) & \dots & \text{cov } (x_2, x_p) \\ \dots & \dots & \dots & \dots \\ \text{cov } (x_p, x_1) & & \dots & \text{var } (x_p) \end{vmatrix}$$

avec $\text{cov } (x_i, x_j) = \text{cov } (x_j, x_i)$

ou

$$\begin{vmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & \dots & \dots & 1 \end{vmatrix}$$

avec $r_{ij} = r_{ji}$

Pour étudier ces matrices, deux techniques statistiques ont été proposées : l'analyse des composantes principales et l'analyse factorielle.

A. *L'analyse des composantes principales.*

Considérons la figure 2 : la position de chaque point est déterminée par ses coordonnées x_i et y_i, qui sont les mesures des deux variables du sujet représenté par un point. Il nous est possible d'effectuer un changement d'axes et de substituer à l'axe des x le grand axe de l'ellipse de façon à ce que la variance devienne maximum. Le second axe choisi sera alors le petit axe de l'ellipse, perpendiculaire au grand axe : nous constatons qu'en rapportant la position des points à ces deux nouveaux axes, nous avons éliminé toute corrélation entre les deux nouvelles variables. Il apparaît également que, si les variables initiales avaient présenté entre elles une corrélation égale à 1 (points tous situés sur une droite), la première transformation aurait suffi à inclure toute la variance et les deux variables initiales auraient été ramenées à une seule. Tel est le principe de la méthode des composantes principales.

Cette méthode est essentiellement valable quand les mesures de toutes les variables sont normées, c'est-à-dire, exprimées par leur écart à la moyenne divisé par l'écart-type, mais elle n'implique aucune autre condition relative aux variables.

Elle vise à transformer linéairement (Hotelling, 1933 ; Pearce et Holland, 1961), les p variables en m autres variables — les composantes principales — m étant plus petit que p, telles que ces m

variables suffisent à expliquer la plus grande partie de la variabilité ; il importe toutefois de se rappeler que p composantes seraient nécessaires si on voulait restituer la totalité de la variabilité, sauf dans les cas où la matrice est de rang inférieur au nombre de variables. La première transformation est effectuée de façon à ce que la première composante qui en résulte, ait une variance aussi grande que possible. La seconde sera choisie de façon à n'avoir pas de corrélation avec la première et à posséder une variance aussi grande que possible : c'est pour cette raison que l'analyse des composantes principales est dite axée sur la considération de la variance. Ainsi progressivement, on extrait des composantes qui sont orthogonales et en ordre de grandeur décroissant en espérant que les quelques premières composantes suffiront à expliquer la plus grande partie de la variance observée.

La projection sur une composante du vecteur représentant un caractère est dénommée saturation en cette composante. La composante (ou racine caractéristique) est donnée par la somme des carrés des saturations de tous les vecteurs ; le rapport de la valeur de la composante au nombre total de variables, multiplié par cent représente le pourcentage du total de la variance expliqué par cette composante.

Le très gros obstacle auquel se heurte l'analyse des composantes principales réside dans l'interprétation des composantes : le plus souvent, les composantes ne peuvent pas être rapportées à des caractéristiques physiques définies et elles nous apparaissent comme des abstractions d'ordre purement mathématique qui conduisent, ainsi que l'a précisé Kendall (1957, p. 37) à un modèle hypothétique.

En outre, le fait de mesurer les variables au moyen d'échelles différentes peut avoir des répercussions considérables sur la nature des résultats obtenus (Seal, 1964, p. 116-120).

Toutefois, cette méthode peut être avantageusement utilisée pour sélectionner, parmi un grand nombre de caractères, ceux qui suffisent à représenter la majeure partie de la variance totale. C'est ainsi qu'il a été mis en évidence (Coblentz, 1968) que, parmi cinquante caractères métriques de la main, huit seulement suffisaient à traduire la variation, la valeur des autres pouvant être déterminée à partir des valeurs des huit premiers. Il s'imposerait toutefois de vérifier si le choix auquel conduit l'analyse d'un échantillon déterminé reste valable dans le cas d'échantillons d'individus

dont le sexe, la race, l'âge, la condition sociale sont différents.

En effet, Moeschler (1968), appliquant ce type d'analyse à trois échantillons de femmes genevoises du canton de Vaud groupées par catégories d'âge, a obtenu des facteurs qui, dans les trois échantillons, n'expliquent pas le même pourcentage de la variance : il a tenté d'interpréter les différents facteurs en tirant parti du rôle qu'y jouaient les mensurations de différents segments du corps et des influences auxquelles la croissance de ces divers segments est sensible.

B. *L'analyse factorielle.*

L'analyse factorielle a pour but de trouver des « facteurs » qui, par leur intervention, expliquent les corrélations observées entre les différentes variables. Supposant que les régressions sont linéaires, elle s'oriente vers la considération de la covariance et aboutit à déterminer un nombre aussi petit que possible de nouvelles variables hypothétiques qui suffisent à restituer la matrice originelle (Lawley et Maxwell, 1963 ; Cattell, 1965).

Dans l'analyse factorielle, on part donc d'un modèle hypothétique et on vérifie s'il s'ajuste aux données.

Considérons le cas de deux variables et supposons que, pour une raison déterminée, nous trouvions justifié d'exprimer toute la covariance de x et y par les variations d'un seul facteur qui sera donc un facteur commun aux deux variables : ceci est réalisable au moyen d'un changement d'axe. Il reste évidemment une partie de la variance de chacune des deux variables non expliquée par ce facteur commun. Dans le cas de p variables, on suppose donc que chacune des p variables initiales est analysable en $k<p$ facteurs communs qui ne sont pas en corrélation et un facteur spécifique à chaque variable. On dispose d'une grande liberté sous le rapport du choix des facteurs et il en résulte que l'analyse factorielle peut conduire à nombre de solutions différentes selon les hypothèses de départ.

On aboutit à une expression de la i^e variable :

$$x_i = \sum_{p=1}^{k} l_{ij} \cdot f_j + c_i \qquad \begin{matrix} (i = 1, 2, \ldots p) \\ (j = 1, 2, \ldots k) \end{matrix}$$

où nous avons k : nombre de facteurs ($k \leqslant p$).

f_j : facteur commun (général ou de groupe).

c_i : source de variation n'affectant que la variable x_i (facteur spécifique et erreur,

dont on suppose qu'ils sont indépendants des autres facteurs).

l_{ij} : saturation du j^e facteur dans la i^e variable.

et dans laquelle on peut introduire différents types de facteurs :

a) des facteurs généraux qui figurent dans toutes les variables.
b) des facteurs de groupe qui ne se rencontrent que dans certaines variables.
c) des facteurs spécifiques qui sont propres à une variable.

Quand un seul facteur peut suffire à expliquer toutes les corrélations, les saturations sont déterminées de façon unique mais dès qu'il est fait appel à plusieurs facteurs, ni les facteurs, ni les saturations ne sont définis de façon unique puisque toute transformation orthogonale des facteurs (rotation) constituera également une solution satisfaisante. Le caractère arbitraire de la rotation des facteurs est un élément qui rend difficile l'interprétation des résultats : on a tenté de l'éliminer par le principe d'économie imposant d'avoir recours au plus petit nombre possible de facteurs. L'emploi de facteurs non indépendants différents selon les auteurs, a encore accru ces difficultés.

D'autre part, la possibilité d'interpréter les résultats dépend de la fixation préalable du nombre de facteurs auxquels on aura recours et de la connaissance des variables qui auront zéro ou approximativement zéro comme saturations de différents facteurs.

Plusieurs applications de l'analyse factorielle à des problèmes d'anthropologie ont été réalisées notamment par Burt et Banks (1947), Howells (1951 et 1953), Schreider (1955 et 1963), Schwidetzky (1960), Landauer (1962), Solow (1966) (Chtetsov, 1960).

Schreider (1955), a, en particulier, cherché à l'utiliser pour analyser les différences de variabilité des caractères métriques d'une population alors que, pour chaque caractère, on observe que les coefficients de variation relative de différentes populations tombent dans une marge plutôt étroite ; l'auteur a montré que les coefficients de variation ne sont pas tous indépendants et qu'il existe notamment une solidarité des dimensions transverses de la tête et du tronc en même temps que de la taille.

Soulignons que l'analyse factorielle, telle que nous venons de la décrire est sans rapport avec les expériences 2^n factorielles imaginées par Fisher pour tester l'effet dû à chacun de n facteurs et aux interactions entre deux ou plusieurs de ces n facteurs, méthodes parfois appelées analyses factorielles.

2. Comparaison de deux ou plusieurs échantillons

La comparaison de deux ou plusieurs échantillons peut être envisagée à plusieurs points de vue que nous aborderons successivement.

21. Méthodes de comparaisons multiples.

Le but de ces tests, généralement passés dans la pratique courante, est de déterminer la probabilité que les différences observées entre plusieurs échantillons soient attribuables aux hasards de l'échantillonnage.

A) *Une variable.*

En analyse univariée, le problème consiste à comparer les moyennes et les variances qui caractérisent la distribution d'une variable dans deux ou plusieurs populations.

On a recours entre autres au test de Student, à la méthode des contrastes et à la méthode de Dunnett, pour la comparaison des moyennes (Dagnelie, 1965), au test de Fisher pour la comparaison des variances, à l'analyse de variance ou au test de χ^2. Ce dernier test peut également être employé pour comparer des distributions.

B) *Deux ou plusieurs variables* (Rao, 1948).

En analyse multivariée, nous devons considérer le cas de p variables mesurées pour k populations comprenant n sujets, n n'étant pas nécessairement le même pour toutes les populations.

Chaque population est caractérisée par

a) les moyennes des p variables donnant le vecteur des moyennes (centroïde).

b) la variance généralisée ou dispersion représentée par le déterminant de la matrice des variances et des covariances.

Il y a lieu de tester deux hypothèses :

a) H_1 : s'il y a homogénéité des dispersions dans les différentes populations indépendamment de la valeur des moyennes.

b) H_2 : si, les dispersions étant égales, les moyennes sont égales ce qui correspond en analyse multivariée à l'extension de l'analyse de la variance.

L'hypothèse H_1 est vérifiée au moyen du critère M de Box (Cooley et Lohnes, 1962) auquel on applique un test de signification faisant appel au rapport de variance F. Pour l'hypothèse H_2, on applique le critère Λ de Wilks qui n'est applicable que si l'hypothèse H_1 est vérifiée : le Λ est toutefois peu sensible à l'hétérogénéité des variances généralisées.

Le lambda est égal au rapport de

a) la somme des matrices des carrés et des produits des déviations à l'intérieur de chaque population ($|\mathbf{W}|$)

b) la somme des carrés et des produits des déviations pour l'ensemble des populations considérées ($|\mathbf{T}|$)

et sa signification est testée au moyen du χ^2 (approximation de Bartlett) ou du F de Rao (1952, p. 258).

Le calcul du critère Λ de Wilks doit précéder tout recours à l'utilisation de l'analyse discriminante.

Enfin, l'utilisation d'un rapport de variance permet de vérifier si l'adjonction d'une ou plusieurs variables apporte des informations concernant la différence entre les groupes et de tester la signification de la différence entre deux corrélations multiples.

Pearson et Wilks (1933) ont appliqué cette technique aux mesures de la longueur et de la largeur de 600 crânes (30 échantillons de 20 crânes, chaque échantillon appartenant à une population différente) et ont montré l'hétérogénéité raciale des variances et des corrélations de ces deux mensurations.

Si, pour chaque échantillon, on a p variables, on disposera de p sommes de carrés et de $\frac{p\,(p-1)}{2}$ sommes de produits sur lesquelles l'analyse de la dispersion (P. C. Mahalanobis) conduira à déterminer la part de la dispersion totale due à différentes catégories (Rao, 1952, p. 258 ; Kendall, 1957, p. 130). Outre cela, une analyse de covariance qui envisage l'influence que peut avoir une ou plusieurs variables sur un ensemble d'autres en raison de l'existence de relations de dépendance, permet d'éliminer les influences étrangères à l'aide des techniques de régression (Rao, 1952, p. 264).

Par exemple, l'étude des valeurs de quatre mensurations de crânes égyptiens, datant de quatre époques différentes (Barnard, 1935 ; Kendall, 1957, p. 137) a montré que les séries étaient hétérogènes, que les quatre variables considérées participaient à l'hétérogénéité

et que la régression de l'ensemble des variables, par rapport au temps, n'était pas linéaire ou qu'il y avait des sources supplémentaires de variation, indépendantes du temps.

22. L'ANALYSE DISCRIMINANTE ET L'ANALYSE CANONIQUE.

L'analyse discriminante a notamment pour objet de déterminer la fonction qui offre la meilleure ligne de séparation, tenant compte des valeurs de p variables, entre deux ou plusieurs populations *définies a priori.*

Si nous considérons la figure 3 où sont représentées les moyennes $\bar{x}_1$ et $\bar{x}_2$ de deux populations normalement distribuées avec un écart-type commun, sous le rapport de la variable envisagée X, nous constatons que le meilleur critère de séparation entre les deux populations correspond à la valeur x_i et que l'adoption de ce critère entraîne le classement erroné d'un certain nombre de sujets, ceux situés en tête de la distribution 2 et ceux situés en queue de la distribution 1.

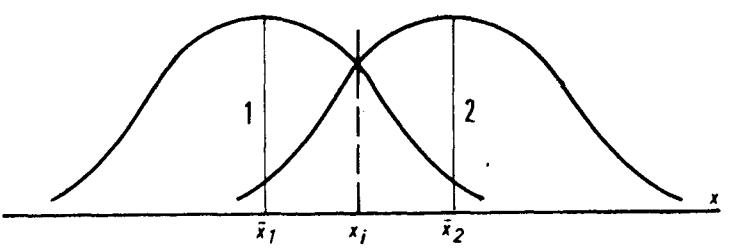

FIG. 3

Si deux mesures ont été effectuées sur les sujets de deux échantillons (Defrise, 1952), sous réserve que les variances et covariances soient égales pour les deux populations, la fonction discriminante sera une droite d'équation :

$$Y = a_1x_1 + a_2x_2$$

pour laquelle la discrimination obtenue est meilleure ou, au pis, égale à celle réalisée au moyen de l'une quelconque des deux variables et telle que le rapport de la différence prise en valeur absolue des Y des 2 populations à l'écart-type des Y soit maximum.

De même, en ayant recours à un nombre p de variables de plus en plus élevé, on obtient :

a) pour 3 variables, un plan : $Y = a_1x_1 + a_2x_2 + a_3x_3$

b) pour p variables, un hyperplan : $Y = a_1x_1 + a_2x_2 \ldots\ldots + a_px_p$

L'expérience montre toutefois que, après utilisation d'un certain nombre de variables, variant selon les cas, l'adjonction de variables supplémentaires n'augmente pratiquement plus le pouvoir de discrimination de la fonction. Le choix des variables joue donc un rôle important (Howells, 1966, p. 9).

La notion de fonction discriminante introduite en 1936 par R. A. Fisher a fait l'objet de nombreuses applications en anthropologie notamment par Bronowski et Long (1952), Pons (1955), Steel (1962), Howells (1966), Defrise (1966), Boulinier (1968, 1969) pour n'en citer que quelques-unes. Son application soulève certainement des problèmes et on ne doit pas s'attendre à ce que les critiques (Marshall, 1969) soient très sensibles aux avantages et au caractère d'objectivité des méthodes statistiques (Giles, 1966).

Quand le nombre de populations est supérieur à deux, avoir recours à une seule fonction impliquerait le sacrifice d'une partie du pouvoir discriminant des données mais il est possible d'établir autant de fonctions indépendantes qu'il y a de populations moins une si le nombre de variables est supérieur ou égal au nombre de populations.

En faisant appel à l'analyse canonique, il est possible de traiter k échantillons comptant respectivement N_1, N_2, ...N_k sujets dont on a mesuré p variables et d'arriver à un arrangement qui exprime les relations existant entre les divers échantillons. Son but est de trouver une série de fonctions linéaires des données primitives, chacune de ces fonctions ne présentant aucune corrélation avec les autres et séparant les groupes *définis a priori* aussi bien que possible, compte tenu de la séparation déjà réalisée par chacune des fonctions précédentes (Oxnard, 1969). L'analyse canonique se prête à l'utilisation de mesures de types différents mais elle n'est applicable, à l'heure actuelle, que si les matrices des corrélations dans les différentes populations sont approximativement identiques.

Après avoir vérifié que les vecteurs moyens des ensembles sont différents, on les compare suivant les axes canoniques : le premier de ces axes est disposé selon la plus grande variabilité intergroupes des moyennes, et ainsi de suite pour les suivants par ordre de variabilité décroissante, les axes étant toujours orthogonaux. On exprime ensuite les vecteurs des ensembles sous leur forme canonique et on porte leur valeur sur des graphiques dont les ordonnées correspondent aux variables canoniques prises deux à deux. On peut en-

tourer chaque point représentatif d'un échantillon au moyen d'un cercle de rayon proportionnel à l'inverse de la racine carrée de l'effectif de l'échantillon, cercle qui définit un domaine de confiance du pourcentage déterminé. Deux ou trois fonctions épuisent généralement les possibilités de discrimination : les résultats peuvent donc faire aisément l'objet d'une représentation graphique.

Ashton et ses collaborateurs (1965) ont appliqué cette méthode d'analyse à l'étude des Primates et particulièrement à la différenciation des divers genres sous le rapport de la morphologie de la ceinture scapulaire en relation avec le mode de locomotion.

23. Le calcul des distances.

Le calcul des distances est une notion qui répond au désir des anthropologistes de caractériser par une mesure unique les différences existant entre des populations pour plusieurs caractères métriques : les diverses approches de cette notion tentées depuis le début du siècle ont été passées en revue par Huizinga (1962, 1965) et Hiernaux (1964) mais la seule solution correcte a été élaborée par P. C. Mahalanobis à partir de 1936 (Mahalanobis, 1949). C'est la distance généralisée ou D^2 qui offre l'avantage de tenir compte des corrélations existant entre les caractères utilisés mais dont l'emploi est soumis à des conditions bien définies et notamment à l'obligation de transformer les variables originales exprimées en écarts-type en variables indépendantes les unes des autres (Rao, 1952) :

$$D^2 = \sum_{i=1}^{p} d_i^2$$

d représentant la différence entre les moyennes des variables obtenues par transformation. L'allure manifestée par l'accroissement du D^2 au fur et à mesure que l'on tient compte d'un plus grand nombre de variables dépend de l'ordre d'introduction des variables ; il est possible de tester la signification de la valeur de D^2 (Rao, 1952) et de saisir l'accroissement du D^2 dû à l'adjonction de chaque variable, donc d'apprécier le pouvoir de différenciation de chaque variable pour un ordre d'introduction des variables déterminé.

La longueur et la complexité des opérations réclamées par le calcul du D^2 ont incité certains auteurs à défendre le recours à des expressions approchées, d'autant plus fermement qu'on observe

des corrélations très élevées entre les différentes valeurs ainsi qu'on le voit dans le tableau 1. Hiernaux (1964) a toutefois clairement montré que la meilleure de ces distances, C^2_H, surestime la distance morphologique et qu'elle n'est pas valable quand ce sont les valeurs absolues des distances qui importent.

La notion de distance généralisée de Mahalanobis a été appliquée par plusieurs chercheurs notamment par P. C. Mahalanobis, D. N. Majumdar et C. R. Rao (1949), Hiernaux (1956), Mukherjee, Rao et Trevor (1955), D. N. Majumdar et C. R. Rao (1960) et B. Hanna (1962). Hiernaux a utilisé les distances entre des populations du Ruanda, de l'Urundi et du Kivu (caractères biométriques et groupes sanguins) pour mettre en évidence l'action simultanée de mécanismes évolutifs comme l'adaptation au milieu ou les processus de flux génique.

Plus récemment, Schwidetzky (1967) a réalisé une analyse des crânes néolithiques d'Europe au moyen du C^2_H de Penrose, les valeurs individuelles des variables n'étant généralement pas publiées. Signalons enfin l'application du coefficient de divergence de Clark par Spuhler (1954) et les tentatives faites en vue d'établir des distances reposant sur les caractères sérologiques ou descriptifs (Womble, 1951 ; Sanghvi, 1953 ; Hiernaux, 1965 ; Howells, 1966 ; Karve et Malhotra, 1968).

3. **Les problèmes de classification**

L'ensemble des problèmes que nous avons abordés jusqu'à présent s'est limité à définir les caractéristiques de populations à partir d'échantillons et à différencier autant que possible ces populations les unes des autres. Il est bien évident que les résultats obtenus ne prendront leur complète signification que si ils peuvent nous aider à ordonner les populations humaines de façon objective et en accord avec l'état actuel de nos connaissances dans le domaine de la biologie.

Il s'impose ici de distinguer deux catégories de problèmes (Sokal et Sneath, 1963 ; Dagnelie, 1968) :

1) soit qu'on cherche à situer un individu, plusieurs individus ou un nouvel ensemble d'individus par rapport à un ensemble défini au préalable ou à les classer parmi d'autres ensembles également définis au préalable.

2) soit qu'on désire hiérarchiser des groupes préalablement

TABLEAU 1

Corrélations entre les différentes expressions de la distance

	D^2	DD	DDR	C^2_H	C^2_Q
CD	0,864 (2)	—	—	—	—
$(CD)^2$	0,899 (2)	—	—	—	—
DD	0,956 (2)	—	0,98 (2) 0,951 (4)	0,94 (3)	0,88 (3)
$(DD)^2$	0,970 (2)	—	—	—	—
CRLR	—	0,81 (4)	—	—	—
C^2_H	0,990 (1)	—	—	—	—
	0,958 (1)	—	—	—	—
	0,970 (2)	0,94 (3)	0,97 (3)		0,91 (3)
C^2_R	0,992 (1)	—	—	—	—
	0,987 (1)	—	—	—	—
C^2_Q	0,957 (1)	0,88 (3)	0,90 (3)	0,91 (3)	—
C^2_Z	0,274 (1)	0,44 (3)	0,49 (3)	0,53 (3)	0,07
Δsy	0,933 (5)	—	—	0,959 (5)	0,13 (3)
Δg	0,959 (5)	—	—	0,991 (5)	—

(1) Penrose, 1954.
(2) Hiernaux, 1964.
(3) Huizinga, 1965.
(4) Czekanowski, 1932.
(5) Hiernaux, 1965.

D^2 : distance généralisée de Mahalanobis.
CD : coefficient de divergence de Clark (1952).
$(CD)^2$: id. porté au carré en tenant compte du fait que CD est la racine carrée d'une somme de carrés de différences.
DD : différence moyenne (durchschnittliche Differenz de Czekanowski).
$(DD)^2$: méthode approximative pour transformer DD en une somme de carrés de différences absolues.
DDR : différence moyenne réduite.
CRL : coefficient of racial likeness (Pearson).
CRLR : « reduced » coefficient of racial likeness.
C^2_H : coefficient de Heincke, moyenne de la somme des différences de moyennes exprimées en écarts-type (Penrose, 1954), qui peut être réparti en deux composantes C^2_Q et C^2_Z.
C^2_R : distance dérivée de C^2_H et tenant compte, dans une certaine mesure, des corrélations (Penrose, 1954).
C^2_Q : composante résultant de la différence de dimensions.
C^2_Z : composante résultant de la différence de forme.
Δsy : différence systémique (Womble, 1952).
Δg : distance générale (Hiernaux, 1965) : permet l'utilisation simultanée de moyennes métriques, de fréquences géniques et de pourcentages de caractères descriptifs.

définis ou établir des groupements à l'intérieur d'un ensemble d'éléments, non structuré.

3.1. CLASSIFICATION ENVISAGÉE PAR RAPPORT À DES ENSEMBLES DÉFINIS AU PRÉALABLE.

a) J'ai précédemment traité divers aspects de ce problème (Leguebe, 1970) en rappelant les travaux de M^me^ E. Defrise qui a appliqué les principes de l'analyse multivariée à la comparaison d'un individu à un ensemble de référence (pour un essai antérieur, cf. Morant, 1930 ; Rao, 1948) et à la comparaison de la distance généralisée entre deux individus avec celle existant entre deux sujets pris au hasard dans la même population qu'eux ; ces méthodes ont été appliquées à l'étude de divers fémurs fossiles par comparaison avec un ensemble de fémurs actuels, à la distribution des chromosomes dans les cellules humaines en métaphase et à l'analyse du caractère héréditaire de divers aspects de la morphologie au moyen d'échantillons de jumeaux. Citons également dans des domaines voisins les travaux de Knussmann (1967), Patterson et Howells (1967) et Howells (1969b).

b) Pour choisir celui des deux ensembles auquel il faut affecter de nouveaux individus, avec un risque d'erreur aussi petit que possible, on se sert de la fonction discriminante de Fisher sous réserve que les deux populations de référence aient des effectifs suffisamment fournis et aient mêmes variances et mêmes covariances.

Dans le cas où il y a plus de deux populations, le problème est plus complexe surtout si les points moyens des différentes populations ne sont pas alignés (Rao, 1952 ; Cyffers, 1965).

L'analyse canonique a été appliquée à des problèmes de paléontologie humaine notamment pour essayer de préciser, pour un ossement fossile, de quel groupe systématique défini a priori, l'homologue présentait le plus d'affinités sous le rapport des mensurations.

De même, des taux d'excrétion de quatre acides aminés (cystine, lysine, ornithine et arginine) de sujets cystinuriques homozygotes, de leurs parents hétérozygotes et d'individus normaux (Crawhall *et al.*, 1969), on a déduit au moyen de l'analyse canonique deux fonctions non en corrélation et rendant maximum la distance entre ces groupes d'individus préalablement définis :

$$y_1 = a_1.\ln(cys) + b_1.\ln(lys) + c_1.\ln(orn) + d_1.\ln(arg).$$

$$y_2 = a_2.\ln\ (cys) + b_2.\ln(lys) + c_2.\ln(arn) + d_2.\ln(arg).$$

Au moyen de ces deux fonctions, on tente de déterminer, en fonction des taux d'excrétion des quatre acides aminés, à quel génotype appartiennent d'autres individus.

3.2. Structuration et hiérarchisation d'éléments sans recours à une définition à priori de groupements.

Considérons que nous avons n éléments (populations ou individus) dont la position est fixée dans un espace p dimensionnel par les valeurs des p variables qui servent à caractériser chacun des éléments. Les rapports existant entre les éléments sont donnés par exemple par les distances entre les éléments ; on pourrait aussi utiliser une mesure du degré de similitude. Le problème consiste donc à grouper les éléments en constellations (Rao, 1948 et 1952).

L'éventail des méthodes proposées est très large et, ici encore, nous constaterons que le choix de la méthode à adopter est fonction du but poursuivi et que le résultat obtenu dépend des critères adoptés pour la définition des constellations. Il est possible soit d'appliquer à l'ensemble des éléments une suite de subdivisions, soit d'effectuer des groupements des éléments.

A. *Groupement par divisions successives* (Edwards et Cavalli-Sforza, 1965).

Ce procédé est basé sur l'analyse de la variance : l'ensemble des éléments est divisé en deux groupes A et B tels que la somme des carrés des distances des éléments à la moyenne générale est répartie en deux termes (fig. 4) :

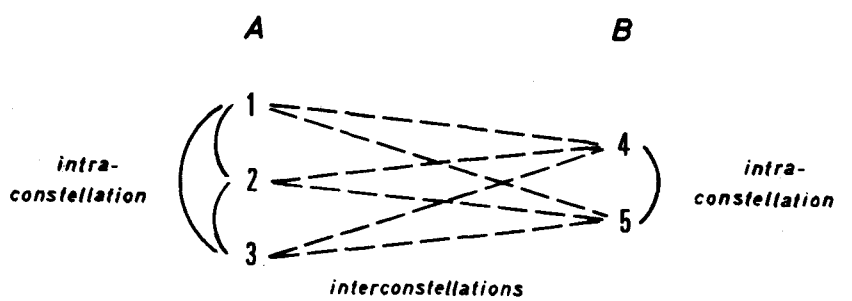

Fig. 4

a) un premier terme qui comporte la somme des carrés des distances des éléments de la constellation A à leur moyenne et la somme des carrés des distances des éléments de la constellation B à leur moyenne (variances intra-constellation).

b) un second terme relatif à la somme des carrés des distances entre les éléments des deux constellations A et B.

La répartition est réalisée par un ordinateur qui détermine celle des $2^{n-1} - 1$ solutions pour laquelle :

la Σ D^2 intra-constellation est minimum.

la Σ D^2 interconstellations est maximum.

On procède de même sur chacun des groupes obtenus et on aboutit à un dendrogramme ou phénogramme qui figure dans un plan la classification réalisée : on obtient ainsi une hiérarchisation des éléments sans toutefois former de groupes distincts puisqu'on procède à une dichotomisation progressive de l'ensemble.

L'attention de Cavalli-Sforza et Edwards (1967) s'est aussi portée sur les possibilités d'extension de cette méthode aux aspects de la phylogénèse.

B. *Groupement par accroissements successifs.*

a) Une possibilité utilisée par Howells (1969a) consiste à établir une première paire au moyen des deux éléments présentant la distance la plus faible entre eux, de recalculer les distances pour les éléments restants et la moyenne de la paire formée, de constituer une seconde paire au moyen des éléments présentant alors la distance la plus faible et ainsi de suite jusqu'à ce qu'on ait épuisé tous les éléments et constitué un dendrogramme.

b) Rao (1952) utilise, comme point de départ du groupement, deux éléments étroitement associés qui formeront l'ébauche d'une première constellation : il recherche alors le troisième élément dont le D^2 moyen par rapport aux deux premiers est le plus faible et ainsi de suite, jusqu'à ce que le D^2 moyen d'un élément paraisse élevé si on le compare aux précédents ou si son introduction dans la constellation provoque un changement appréciable du D^2 moyen intra-constellation. On applique alors le même traitement aux éléments restants.

Cette manière de procéder, bien qu'introduisant un jugement subjectif, a pour effet de ne pas seulement hiérarchiser les éléments

de l'ensemble comme dans le cas précédent, mais de constituer des constellations à l'intérieur de l'ensemble.

c) Van den Driessche (1965) a précisé des critères permettant de décider objectivement si un élément doit ou non entrer dans une constellation.

On établit, pour chaque élément, la liste des distances à tous les autres éléments par ordre de grandeur croissante et on calcule *l'étendue* qui correspond à la différence entre le D^2 maximum et le D^2 minimum pour tous les éléments.

La distance la plus faible donne les éléments de l'ébauche de la première constellation, soit 1 et 2 : pour que ces éléments puissent entrer dans une constellation, il faut que leur distance soit plus faible que l'étendue. On ajoute à cette ébauche un troisième élément, soit 3, celui qui accroît de façon minimale le premier D^2 des distances 1-3 et 2-3. Pour que cet élément 3 soit intégré dans la première constellation, il faut que la distance moyenne intra-constellation (somme des D^2 entre éléments d'une constellation, divisé par le nombre de D^2) soit plus petite que toute distance entre un des éléments de la première constellation et tout autre élément non inclus dans la première constellation, distances désignées par D^2 intra-extra-constellation.

Si cette condition n'est pas remplie l'élément est rejeté de la première constellation considérée comme complète et on prend, parmi ceux qui restent, comme ébauche de la deuxième constellation, les deux éléments qui présentent la valeur la plus faible de D^2.

Cette ébauche sera acceptée et on y ajoutera d'autres éléments aussi longtemps que les deux conditions suivantes seront remplies :

$\bar{D}^2$ interconstellations $> \bar{D}^2$ intra-constellation.

D^2 intra-extra-constellation $> \bar{D}^2$ intra-constellation.

Les $\bar{D}^2$ interconstellations sont égales à la somme des distances de tous les éléments de la première constellation aux éléments de la seconde, divisée par le nombre de distances : elles doivent être calculées chaque fois qu'on ajoute un nouvel élément à une constellation.

Le résultat final se présente sous forme d'une matrice symétrique dont les éléments de la diagonale sont les distances moyennes intra-constellation et dont les autres sont les distances moyennes inter-constellations.

d) Hiernaux (1967) a fait remarquer que d'autres critères plus rigoureux pouvaient être adoptés :

1. soit qu'aucun élément ne puisse être plus semblable à un élément extérieur à sa constellation qu'à un élément faisant partie de celle-ci.
2. soit que toute distance à l'intérieur d'une constellation doive être inférieure à toute distance entre populations non classées dans la même constellation.

L'adoption de l'un ou l'autre de ces critères conduit à des constellations différemment constituées et le fait n'est pas surprenant puisque le nombre de solutions différentes auquel peut donner lieu la classification de N éléments en m classes (Dagnelie, 1968) est approximativement égal à $\frac{m^N}{m\,!}$ ce qui donne :

m = \ N =	10	20	50	100
2	511	5×10^5	$\backsimeq 10^{15}$	$\backsimeq 10^{30}$
3	9.330	58×10^7	$\backsimeq 10^{23}$	$\backsimeq 10^{47}$
4	34.105	45×10^9	$\backsimeq 10^{29}$	$\backsimeq 10^{59}$

Le calcul de toutes les solutions dépasse pratiquement les possibilités des ordinateurs et il n'est pas pensable qu'une équipe d'anthropologistes puisse interpréter une telle masse de résultats.

Conclusion

Les avantages que nous offre le recours aux ordinateurs pour traiter de nombreuses variables, ne sont pas à rechercher essentiellement dans une augmentation de l'efficience des moyens de calcul mis à notre disposition.

Certes, si les résultats soumis à notre examen sont plus complets, ils sont aussi plus complexes mais surtout, ils ont exigé que soient très strictement définies toutes les opérations auxquelles nous avons soumis les données et toutes les implications sous-jacentes à ces opérations. Ils font donc de plus en plus appel à l'exercice de notre esprit de finesse, à une connaissance aussi large que possible des phénomènes biologiques, à une intuition qui doit s'exercer avec d'autant plus d'acuité et de réflexion que notre décision sera suivie

de la mise en œuvre de moyens de calcul puissants, coûteux et difficiles à contrôler (Seal, 1964, p. 183). Ce sont les statisticiens eux-mêmes qui nous font entendre la voix de la sagesse puisque C. R. Rao (1952) a jugé utile de placer en exergue de son livre cette phrase de R. A. Fisher : « Le statisticien n'est plus semblable à l'alchimiste dont on attendait qu'il transformât en or le matériau quelconque qu'on lui présentait. Il est plus proche du chimiste qui analyse ce matériau, en détermine la composition et en extrait les constituants. Aussi, il serait absurde de le complimenter pour la qualité des résultats ou de lui adresser des reproches pour leur insuffisance. S'il est maître de sa technique, la qualité du résultat dépend uniquement de la valeur des données qui lui sont soumises. Un matériel déterminé contient une quantité limitée d'informations : la tâche du statisticien se limite à la mettre en évidence ».

BIBLIOGRAPHIE

La bibliographie ne vise pas à être exhaustive : le lecteur trouvera des données plus complètes dans les ouvrages marqués d'un astérisque.

A. **Ouvrages généraux.**

*Anderson, T. W.
1958 An introduction to multivariate statistical analysis. New York, Wiley and sons, 374 p.

*Cooley, W. W. et P. R. Lohnes.
1962 Multivariate procedures for the behavioral sciences. New York, Wiley and sons, 211 p.

Dagnelie, P.
1966 Introduction à l'analyse statistique à plusieurs variables. *Biométrie-Praximétrie*, 7 (1) : 43-66.

*Kendall, M. G.
1957 A course in multivariate analysis. London, Charles Griffin, 185 p.

Kendall, M. G. et A. Stuart.
1958 The advanced theory of statistics.
1966 London, Charles Griffin ; 3 vol. : 433 p., 676 p., 552 p.

*Rao, C. R.
1952 Advanced statistical methods in biometric research. New-York, John Wiley and sons, 390 p.

Roy, S. N.
1957 Some aspects of multivariate analysis. Calcutta, *Statistical Publ. Soc. Indian stat. ser.*, 3 : 214 p.

*Seal, H. L.
1964 Multivariate statistical analysis for biologists.
London, Methuen and Co, 207 p.
*Sokal, R. R.
1965 Statistical methods in systematics.
Biol. Rev., Cambridge, **40** : 337-391.

B1. **Relations entre variables**

Dagnelie, P.
1967 La corrélation multiple, la corrélation partielle et la corrélation entre groupes de variables.
Biométrie-Praximétrie, **8** : 3-25.
Defrise-Gussenhoven, E.
1955 Ellipses équiprobables et taux d'éloignement en biométrie.
Bull. Inst. roy. Sci. Nat. Belgique, **31** (26) : 1-31.
Quenouille, M. H.
1952 Associated measurements.
London, Butterworths Sci. Publ., 242 p.
Teissier, G.
1948 La relation d'allométrie. Sa signification statistique et biologique.
Biometrics, **4** (1) : 14-53.

B2. **Analyse des composantes principales.**

Coblentz, A.
1968 Les liaisons des caractères métriques de la main.
Bull. Mém. Soc. Anthrop. Paris, 12e sér., **3** : 331-345.
Hotelling, H.
1933 Analysis of a complex of statistical variables into principal components.
J. Educ. Psychol., **24** : 417-441, 498-520.
Kraus, B. S. et S. Chil Choi.
1958 A factorial analysis of the prenatal growth of the human skeleton.
Growth, **22** : 231-242.
Moeschler, P.
1968 Biométrie des femmes de Genève.
L'Anthropologie, Paris, **72** (5/6) : 489-516.
Pearce, S. C. et D. A. Holland.
1961 Analyse des composantes, outil de recherche biométrique.
Biométrie-Praximétrie, **2** : 159-177.

B3. **Analyse factorielle.**

Burt, Sir Cyril et C. H. Banks.
1947 A factor analysis of body measurements for British adult males.
Ann. Eugen., London, **13** : 328-256.

*Cattell, R. B.
1965 Factor analysis : an introduction to essentials.
Biometrics, **21** : 190-210, 405-435.
*Chtetsov, V. P.
1960 Faktornii Analiz v Anthropologii.
Voprosy Anthropologii, **3** : 106-111.
Howells, W. W.
1951 Factors of human physique.
Am. J. phys. Anthrop. N.S. **9** : 159-191.
1953 Correlations of brothers in factor scores.
Am. J. phys. Anthrop. N.S. **11** : 121-140.
Kanda, S., K. Kurisu.
1967 Factor analytic studies on the Japanese skulls.
1968 Factor analysis of Japanese skulls. Part 2, Part 3.
Medical J. Osaka Univ. **18** : 1-9 ; 315-318 ; 319-330.
Landauer, C. A.
1962 A factor analysis of the facial skeleton.
Human Biol., **34** : 239-253.
*Lawley, D. N. et A. E. Maxwell.
1963 Factor analysis as a statistical method.
London, Butterworths, 117 p.
Schreider, E.
1955 Emploi de l'analyse factorielle dans l'étude de la variabilité biologique.
In : *L'analyse factorielle et ses applications*, Paris, p. 253-262.
1963 Les liaisons anthropométriques dans l'espèce humaine. Étude comparée de onze populations : corrélations et analyses factorielles.
L'Anthropologie, Paris, **67** : 49-84.
Schwidetzky, I.
1959 Faktoren des Schädelbaus bei der vorspanischen Bevölkerung der Kanarischen Inseln.
Homo, **10** : 237-248.
Solow, B.
1966 The pattern of craniofacial associations. A morphological and methodological correlation and factor analysis study on young adult males.
Acta Odont. Scandinav., **24** : supplem. 46.
Thurstone, L. L.
1947 Factorial analysis of body measurement.
Am. J. phys. Anthrop., N.S. **5** : 15-28.

C1. Méthodes de comparaisons multiples.

Barnard, M. M.
1935 The secular variation of skull characters in four series of Egyptian skulls.
Ann. Eugen. London, **6** : 352-372.

DAGNELIE, P.
1965 A propos de quelques méthodes de comparaisons multiples de moyennes.
Biométrie-Praximétrie, **6** (3/4) : 115-124.

PEARSON, E. S. et S. S. WILKS.
1933 Method of statistical analysis appropriate for k samples of two variables.
Biometrika, **25** : 353-378.

RAO, C. R.
1948 Tests of significance in multivariate analysis.
Biometrika, **35** : 58-79.

RULON, P. J. et W. D. BROOKS.
1961 On statistical tests of group differences.
Cambridge, Mass., Educ. Research Corpor.

C2. Analyse discriminante et analyse canonique.

ASHTON, E. H., M. J. R. HEALEY, C. E. OXNARD et T. F. SPENCE.
1965 The combination of locomotor features of the primate shoulder girdle by canonical analysis.
J. Zool., **147** : 406-429.

BOULINIER, G.
1968 La détermination du sexe des crânes humains à l'aide des fonctions discriminantes.
Bull. Mém. Soc. Anthrop. Paris, XII^e s., **3** : 301-316.
1969 Variations avec l'âge du dimorphisme sexuel des crânes humains adultes.
Bull. Mém. Soc. Anthrop. Paris, XII^e s., **4** : 127-128.

BRONOWSKI, J. et W. M. LONG.
1952 Statistics of discrimination in anthropology.
Amer. J. phys. Anthrop., N.S. **10** : 385-394.

DEFRISE-GUSSENHOVEN, E.
1952 Discrimination de populations voisines. Étude biométrique.
Bull. Inst. roy. Sci. nat. Belgique, **28** (46) : 1-34.
1966 A masculinity-femininity scale based on a discriminant function.
Acta genet., Basel, **16** : 198-208.

FISHER, R. A.
1936 The use of multiple measurements in taxonomic problems.
Ann. Eugen., London, **7** : 179-188.

GILES, E.
1966 Statistical techniques for sex and race determination. Some comments of defense.
Am. J. phys. Anthrop., 25 (1) : 85-86.

HOWELLS, W. W.
1966 The Jomon population of Japan : a study by discriminant analysis of Japan and Ainu crania.
Papers Peabody Mus. Arch. Ethnol., **57** (1) : 1-43.

MARSHALL, D. S.
1969 Book review of « Craniometry and multivariate analysis » by W. W. Howells et J. M. Crichton.
Hum. Biol., **41** (2) : 290-295.

OXNARD, C. E.
1969 Mathematics, shape and function : a study in primate anatomy.
Am. Scientist, **57** : 75-96.

PONS, J.
1955 Discriminacion sexual en femures, pelvis y esternones.
Trab. Inst. B. de Sahagun Antrop. Etnol., **14** (4) : 137-159.

STEEL, F. L. D.
1962 The sexing of long bones, with reference to the St Bride's series of identified skeletons.
J. Roy. Anthrop. Inst. Great Britain and Ireland, **92** (2) : 212-222.

C3. **Calcul de distances.**

CLARK, P. J.
1952 An extension of the coefficient of divergence for use with multiple characters.
Copeia, **2** : 61-64.

CZEKANOWSKI, J.
1932 'Coefficient of racial likeness' und 'Durchschnittliche Differenz'.
Anthrop. Anz., **9** : 227-249.

HANNA, B. C.
1962 The biological relationships among Indians of the Southwest. Analysis of morphological traits.
Am. J. phys. Anthrop., **20** : 499-508.

HIERNAUX, J.
1956 Analyse de la variation des caractères physiques humains en une région de l'Afrique Centrale : Ruanda-Urundi et Kivu.
Ann. Mus. roy. Congo Belge, sér. in-8°, *sci. Homme, Anthrop.*, **3** : 1-131, 8 pl.
1964 La mesure de la différence morphologique entre populations pour un ensemble de variables.
L'Anthropologie, Paris, **68** (5/6) : 559-568.
1965 Une nouvelle mesure de distance anthropologique entre populations, utilisant simultanément des fréquences géniques, des pourcentages de traits descriptifs et des moyennes métriques.
C. R. Acad. Sci., Paris, **260** : 1748-1750.

HOWELLS, W. W.
1966 Population distances : biological, linguistic, geographical and environmental.
Current Anthropology, **7** (5) : 531-540.

HUIZINGA, J.
1962 From DD to D^2 and back. The quantitative expression of resemblance.
Proc. Kon. Nederl. Akad. Wetensch., **C 65** (4) : 1-12.

HUIZINGA, J.
1965 Some more remarks on the quantitative expression of resemblance (distance coefficients).
Proc. Kon. Nederl. Akad. Wetensch., **C 68** (1) : 69-80.
KARVE, I. et K. C. MALHOTRA.
1968 A biological comparison of eight endogamous groups of the same rank.
Current Anthropology, **9** (2/3) : 109-124.
MAHALANOBIS, P. C.
1949 Historical notes on the D^2 statistic.
Sankhya, **9** : 237-240.
MAHALANOBIS, P. C., D. N. MAJUMDAR et C. R. RAO.
1949 Anthropometric survey of the United Provinces, 1941.
Sankhya, **9** : 89-324.
MAJUMDAR, D. N. et C. R. RAO.
1958 Race elements in Bengal.
Bombay, Asia Publ. House, 200 p.
MUKHERJEE, R., C. R. RAO, J. C. TREVOR.
1955 The ancient inhabitants of Jebel Moya.
Cambridge, Cambridge Univ. Press. 123 p.
PENROSE, L. S.
1954 Distance, size and shape.
Ann. Eugen., London, **18** (4) : 337-343.
SANGHVI, L. D.
1953 Comparaison of genetical and morphological methods for a study of biological differences.
Am. J. phys. Anthrop. N.S., **11** : 385-404.
SCHWIDETZKY, I.
1967 Vergleichend-statistische Untersuchungen zur Anthropologie des Neolithikums.
Homo, **18** (3) : 133-207.
SPUHLER, J. N.
1954 Some problems in the physical anthropology of the American Southwest.
Am. Anthropologist, **56** : 604-625.
WOMBLE, W. H.
1951 Differential systematics.
Science, **114** : 315-322.

D. Méthodes de classification.

CAVALLI-SFORZA, L. L. et A. W. F. EDWARDS.
1967 Phylogenetic analysis : models and estimation procedures.
Am. J. hum. Genet., **19** (3) : 233-257.
CRAWHALL, J. C., P. PURKISS, R. W. E. WATTS et E. P. YOUNG.
1969 The excretion of amino acids by cystinuric patients and their relatives.
Ann. hum. Genet., **33** (2) : 149-169.

CYFFERS, B.
1965 Analyse discriminatoire.
Rev. Stat. appl., **13** (2) : 29-43 ; **13** (3) : 39-65.
DAGNELIE, P.
1968 Introduction aux problèmes et aux méthodes de classification numérique.
Biométrie-Praximétrie, **9** (2) : 87-111.
DAY, M. H.
1967 Olduvai hominid 10 : a multivariate analysis.
Nature, London, **215** : 323-324.
DAY, M. H. et B. A. WOOD.
1968 Functional affinities of the Olduvai hominid 8 talus.
Man, **3** (3) : 440-455.
EDWARDS, A. W. F. et L. L. CAVALLI-SFORZA.
1965 A method for cluster analysis.
Biometrics, **21** : 362-375.
HIERNAUX, J.
1967 La diversité humaine en Afrique subsaharienne.
Bruxelles, Éd. Inst. Sociol., 261 p.
HOWELLS, W. W.
1969a Multivariate analysis of human crania.
379 p.
1969b The use of multivariate techniques in the study of skeletal populations.
Am. J. phys. Anthrop., **31** (4) : 311-314.
KNUSSMANN, R.
1967 Das proximale Ende der Ulna von *Oreopithecus bambolii* und seine Aussage über dessen systematische Stellung.
Z. Morph. Anthrop., **59** : 57-76.
LEGUEBE, A.
1970 L'utilisation de l'analyse multivariée en anthropologie.
L'Anthropologie, Paris, **74** (sous presse).
MORANT, G. M.
1930 Studies on paleolithic man.
Ann. Eugen., London, **4** : 109-215.
PATTERSON, B. et W. W. HOWELLS.
1967 Hominid humeral fragment from early Pleistocen of northwestern Kenya.
Science, **156** : 64-66.
RAO, C. R.
1948 The utilization of multiple measurements in problems of biological classification.
J. Roy. Stat. Soc., London, **B10** (2) : 159-203.
*SOKAL, R. R. et P. H. A. SNEATH.
1963 Principles of numerical taxonomy.
San Francisco and London, W. H. Freeman, 359 p.

VAN DEN DRIESSCHE, R.
1965 La recherche des constellations de groupes à partir des distances généralisées D^2 de Mahalanobis.
Biométrie-Praximétrie, **6** : 36-47.

Adresse de l'auteur : A. LEGUEBE
Institut royal des Sciences naturelles de Belgique
rue Vautier, 31
1040-Bruxelles.