Institut royal des Sciences naturelles de Belgique

BULLETIN

Tome XXVIII, n° 46.
Bruxelles, août 1952.

Koninklijk Belgisch Instituut voor Natuurwetenschappen

MEDEDELINGEN

Deel XXVIII, nr 46.
Brussel, Augustus 1952.

DISCRIMINATION DE POPULATIONS VOISINES. ÉTUDE BIOMÉTRIQUE,

par Elisabeth Defrise-Gussenhoven (Bruxelles).

TABLE DES MATIÈRES.

INTRODUCTION.

A tous les échelons du règne animal, on trouve des populations difficilement dissociables. Nous nous proposons d'exposer ici des méthodes de discrimination de populations voisines.

Des populations voisines pourront être, selon le cas, des populations appartenant à des genres voisins, des espèces, des variétés ou des races voisines. Les sujets considérés sont toujours supposés très ressemblants, difficilement dissociables.

Il est essentiel de distinguer nettement deux cas, suivant que tous les sujets viennent d'une source unique ou de sources différentes; nous emploierons ici le mot « échantillon » pour désigner un ensemble de sujets provenant d'une source déterminée.

Les sujets de deux échantillons (de provenances distinctes) peuvent fort bien ne présenter aucune différence significative de nature morphologique ou physiologique, alors qu'une telle différence peut exister entre des sujets de provenance unique.

Cette note a pour but de décrire les méthodes graphiques et statistiques qui servent, soit à comparer deux échantillons, soit à découvrir si un échantillon unique provient d'une population mixte.

Dans le premier cas, nous ne ferons guère que rappeler des méthodes classiques; nous nous étendrons plus longuement sur le second cas, qui, à notre connaissance, n'a jamais été traité.

Les critères de discrimination proposés sont essentiellement basés sur la représentation graphique de certaines mensurations. Une fois mis au point, ils permettent de classer de nouveaux spécimens aussi rapidement que n'importe quel autre critère taxonomique, mais avec une plus grande sécurité. Les calculs, introduits là seulement où ils sont indispensables, ne nécessitent pas de connaissances mathématiques approfondies.

Nous avons ainsi tâché de réunir, dans un article directement utilisable par les systématiciens, les méthodes biométriques applicables à différents problèmes concrets qui nous ont été posés, dans le domaine de la discrimination de populations voisines.

§ I. — GÉNÉRALITÉS.

1. Un échantillon ou deux échantillons.

Nous croyons qu'il n'est pas inutile de bien préciser la distinction entre :

Deux échantillons.

Nous dirons ici que deux groupes d'individus sont de provenance différente ou qu'ils constituent deux échantillons lorsqu'ils sont récoltés en des temps ou en des lieux distincts, ou encore lorsqu'ils se distinguent par leur genre de vie.

Exemples.

1. Fossiles de même âge recueillis en des pays distincts, ou bien en une même région mais sur des terrains de composition non identique, ou encore des fossiles d'âges géologiques différents.

2. Insectes habitant la même région, mais se distinguant par leur genre de vie (date de l'hymen, nourriture, etc.), ou bien insectes de régions différentes.

3. Animaux pêchés en deux endroits éloignés ou encore à un même endroit mais à des époques différentes de l'année, ou bien toujours à la même saison mais d'années consécutives.

4. Peuplades habitant la même région, mais au sein desquelles existe une scission gardée intacte, barrière religieuse ou sociale, ou encore le souvenir d'origines distinctes respectées par un système de caste.

etc.

Un échantillon.

Lorsque ni le genre de vie, ni le temps, ni l'espace ne séparent les sujets à analyser, nous sommes en présence d'un échantillon unique. Tous les sujets sont de même provenance.

1. Fossiles récoltés au même endroit et dans une même formation (s. s.).

2. Insectes pris le même jour sur les mêmes plantes d'une même région.

3. Animaux pêchés au même endroit à la même date.

4. Habitants d'une même agglomération, au sein desquels n'existent pas de telles barrières.

etc.

La distinction entre le cas d'un échantillon et le cas de deux échantillons nous semble essentielle dans toute analyse discriminatoire ; elle se reflète à la fois dans les méthodes à utiliser et dans l'interprétation des résultats.

L'identité ou la différence d'origine est une donnée objective à priori, dont il est fondamental de tenir compte, avec toutes ses implications. Ainsi notamment, les sujets d'un échantillon ont eu la possibilité de se croiser, au contraire des sujets d'échantillons distincts.

Soulignons, comme les exemples ci-dessus l'illustrent, que des critères morphologiques ou physiologiques ne peuvent intervenir dans la séparation en deux échantillons. Il n'est pas permis, au moyen de tel ou tel critère physique, d'opérer une coupure en deux lots au sein d'un échantillon unique, et de traiter ces deux lots comme deux échantillons. Cette coupure faite à posteriori par le naturaliste, d'ailleurs plus ou moins subjective et imparfaite, n'a rien de commun (notamment du point de vue génétique) avec la différence de provenance de deux échantillons.

Si nous insistons sur ce point, c'est que la confusion est cependant assez fréquente. Elle s'explique sans doute en partie par le désir du naturaliste d'utiliser les tests classiques de divergence applicables à deux échantillons : il les applique alors parfois à deux lots d'un échantillon unique, ce qui risque de le conduire à des conclusions biologiquement fausses ou dénuées de sens.

Supposons que des souris grises sauvages et des souris blanches soient attrapées au même endroit. En les séparant en deux lots suivant leur couleur, on fait une discrimination superficielle. En effet, comme les deux races primitives ont eu l'occasion de se croiser, on peut craindre que seul le lot des souris blanches forme une race pure pour la teinte (bb), et qu'à côté des grises pures (GG), il y ait des grises hétérozygotes (Gb) (1). C'est pourquoi cet ensemble de souris grises et blanches doit être considéré comme une seule population, notamment pour l'étude d'autres caractères que la couleur.

Bien différent est le cas d'un échantillon de nombreuses souris grises trouvées sans aucune souris blanche, et que l'on compare avec des souris blanches prises en un autre endroit. Ici, on est sûr que les sujets des deux lots ne sont pas proches

(1) Guyénot, E. (1931, p. 44).

parents; on compare réellement deux races de souris : les grises et les blanches.

2. Caractères continus ou discontinus.

Toute discrimination de populations utilise un certain nombre de caractères des sujets donnés. Les méthodes varient suivant qu'il s'agit de caractères continus ou discontinus.

A. — Un *caractère continu* est mesurable (ou repérable) par un nombre qui varie, avec les sujets, d'une façon (pratiquement) continue entre deux valeurs extrêmes; il est représenté par une variable continue x. Par exemple : la taille des hommes.

Donnée directe de l'expérience, l'histogramme (fig. 1) indique la distribution du caractère dans l'échantillon; on en déduit, par le calcul, la distribution probable dans la population totale.

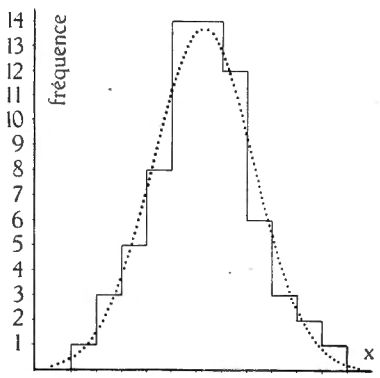

Fig. 1. — ——— histogramme; distribution de x dans l'échantillon. ..... Courbe normale; distribution de x dans la population supposée normale.

L'allure de l'histogramme, le nombre de ses modes, sa moyenne, sa déviation standard sont autant de renseignements sur la composition de l'échantillon.

Rappelons que la variation continue d'un caractère s'explique par l'action simultanée de facteurs mésologiques et de facteurs génétiques, souvent nombreux. Dans une race pure (ou dans une lignée pure chez les plantes), seuls les facteurs extérieurs provoquent la variation continue (2).

Pour envisager simultanément plusieurs caractères continus, on peut procéder de deux façons :

(2) Guyénot, E., (1950, p. 517).

1° On établit la fonction de distribution simultanée de toutes les variables, par exemple x, y, z, t. Cette fonction fournit une description complète des variables et de leurs corrélations.

Malheureusement, il est difficile de l'établir avec précision ; de plus, elle ne se prête pas à une représentation graphique, si commode, notamment pour découvrir la présence de deux modes.

2° Aussi, pour réduire le nombre de variables, on préfère souvent former des fonctions simples telles que $u = \frac{x}{y}, v = z \times t$ (liées à des observations biologiques) et l'on considère u, v comme coordonnées d'un graphique à double entrée. L'information que l'on perd en combinant chaque fois deux (ou éventuellement plusieurs) variables en une seule est compensée par l'efficacité de l'analyse d'un tel graphique.

Il faut toutefois songer à évaluer les erreurs de mesure sur u et sur v qui risquent d'être plus grandes que celles des variables primitives.

Lorsque l'échantillon contient, à côté des adultes, des sujets jeunes, ou si les animaux sont à croissance continue, il est particulièrement utile de former des rapports comme $\frac{x}{y}$.

B. — Un *caractère discontinu* (ou discret) permet de séparer l'ensemble des sujets en un nombre fini (pratiquement petit) de catégories nettement tranchées A_1, A_2, ..., A_n. Ceci peut se ramener à une suite de dichotomies : on distingue d'abord les A_1 des « non A_1 » ; puis, parmi ces derniers, les A_2 et les « non A_2 », et ainsi de suite.

Il nous suffira donc d'envisager des dichotomies, où l'on sépare les sujets en « A » et « non A », suivant qu'ils possèdent ou non le caractère A.

Exemples :

1. Chez l'homme, le sang de certains sujets contient le facteur sérologique P (A), le sang des autres ne contient pas ce facteur (non A) (3).

2. Les deux espèces de Coléoptères Chrysomélides : *Chrysolina menthastri* et *Chrysolina cœrulans*, étudiés par P. Jolivet,

(3) Gates, R. R.,(1946, p. 697).

très ressemblants d'ailleurs, sont, les uns verts (A), les autres bleus (non A).

Il faut toujours qu'entre les deux catégories existe une frontière nette, de sorte que la proportion des cas douteux n'excède en aucun cas un pourcentage très faible qu'il faut s'imposer d'avance d'après la nature du problème. Pour fixer les idées, nous adopterons ici une valeur de 5 %.

En ce qui concerne l'interprétation d'une telle scission, on doit tenir compte du fait que A et non A sont des caractères apparents. S'ils sont caractéristiques d'espèces différentes (4), alors chaque espèce conservera, dans les générations suivantes, l'une le caractère A, l'autre le caractère non A. Mais il n'en est pas toujours ainsi. Bien souvent, on ignore si le caractère

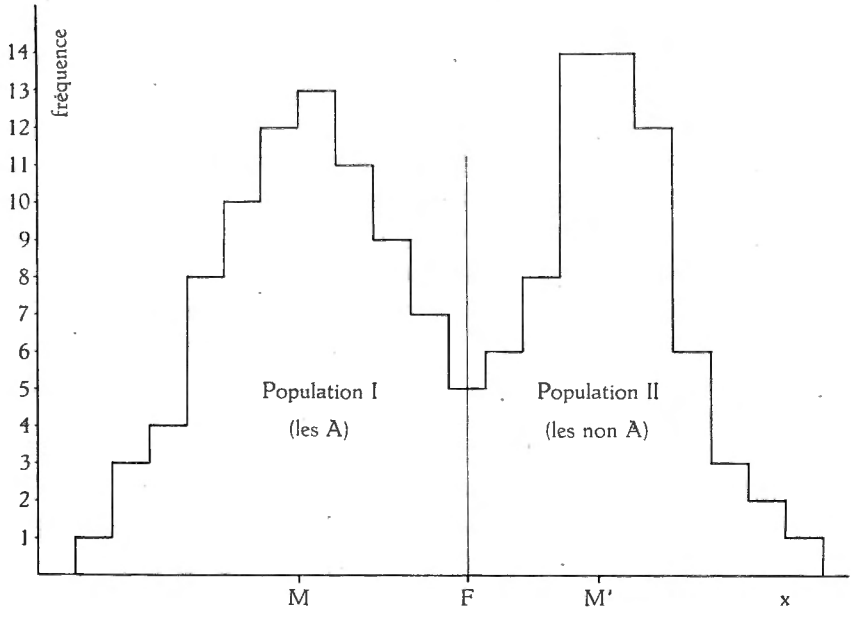

Fig. 2. — Histogramme bimodal; 149 sujets. Par un test de χ^2 on vérifie que les deux modes ne sont pas dus aux hasards de l'échantillonnage; on a donc réellement deux populations. On situe, au jugé, les points moyens M et M'. On calcule σ et σ', en utilisant respectivement la partie de la population I située à gauche de M et la partie de la population II située à droite de M'. Une première approximation situe la frontière F au point de plus basse fréquence. Comme $MF > 1{,}64\ \sigma$ et $M'F > 1{,}64\ \sigma'$, il y a moins de 5 % de sujets mal placés.

(4) Par exemple, lorsque A et non A désignent des appareils chromosomiques non superposables.

envisagé est spécifique, on ne sait même pas s'il est héréditaire ; d'autres fois on ne connaît pas son mode de transmission. Dans ces cas, on doit se contenter d'une simple séparation des phénotypes, les « A » et les « non A », avec le risque que certains descendants des « A » présentent le caractère « non A ».

Un classement de ce genre est superficiel et ne peut être que provisoire.

Il est assez courant de faire dériver une classification discontinue des caractères continus. Cette pratique appelle plusieurs remarques :

1. On ne peut en tout cas effectuer une division en deux catégories à l'aide d'un caractère continu que si celui-ci a une distribution bimodale, dont les deux modes sont séparés par une région de faible fréquence où l'on situe la frontière ; il faut, nous en sommes convenus, qu'il n'y ait pas plus de 5 % de sujets douteux (fig. 2).

On considère une telle distribution bimodale comme résultant de la superposition de deux sortes de spécimens, les A et les non A, se recoupant légèrement.

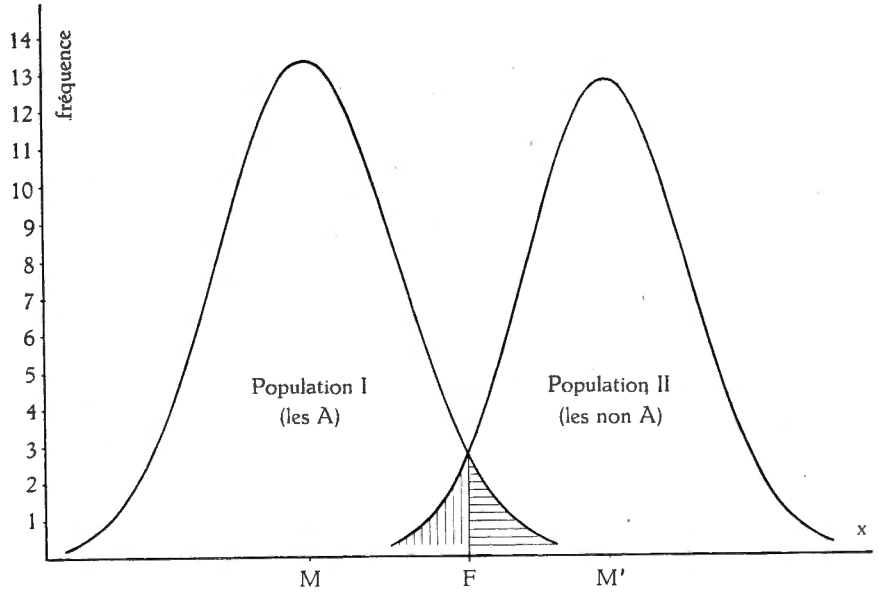

Fig. 3. — Distributions des populations I et II supposées normales. Ici la situation de F est précisée : ce point se trouve au niveau de l'intersection des deux courbes. La région hachurée représente les sujets mal classés (moins de 5 %).

Si chaque groupe de spécimens a une distribution gaussienne (fig. 3), si σ, σ' sont leurs déviations standard respectives, et si M, M' correspondent aux valeurs moyennes, on sait que le nombre de sujets mal classés n'excède pas 5 %, pourvu que

$$MF \geqslant 1{,}64\ \sigma \qquad \text{et} \qquad M'F \geqslant 1{,}64\ \sigma' \quad (\alpha)$$

(les sujets mal classés sont les A placés entre F et M', et les non A placés entre M et F).

Malheureusement, c'est à partir de l'histogramme bimodal que l'on doit juger de la normalité des deux distributions composantes et que l'on doit estimer les effectifs et les valeurs de σ et de σ', ainsi que l'emplacement de M et M', avant de pouvoir vérifier les relations (α). On conçoit que ces opérations ne se font avec une sécurité suffisante que dans les cas plutôt rares d'un nombre de sujets très élevé ou lorsqu'il y a une région de fréquence quasi nulle entre les deux modes.

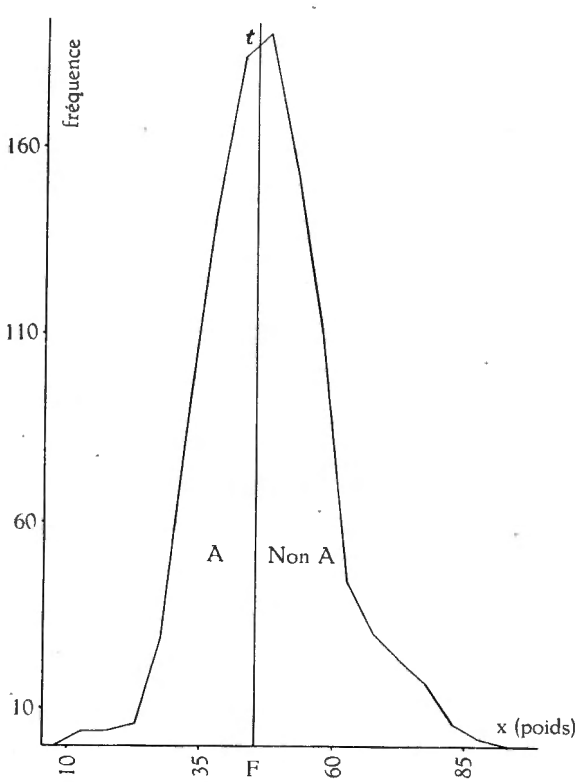

Fig. 4. — Mauvaise séparation (voir aussi fig. 5).

2. Nous venons de voir comme il est hasardeux, même dans un cas de distribution bimodale, de passer du continu au discontinu.

A fortiori, ne peut-on pas scinder en A et non A d'après un caractère continu à distribution unimodale.

En effet, comment justifier du point de vue biologique le choix de la limite F qui sépare les A des non A (fig. 4) ? Cette limite se plaçant à un endroit de haute fréquence, il y aura plus de 5 % de sujets douteux. En outre, certains sujets A peuvent en réalité être plus proches génétiquement de certains non A, à droite de F, que des autres sujets A. Pour s'en convaincre, il suffit de représenter graphiquement quelques lignées pures que E. JOHANNSEN (5) a tirées d'une population de haricots autofécondés (fig. 5).

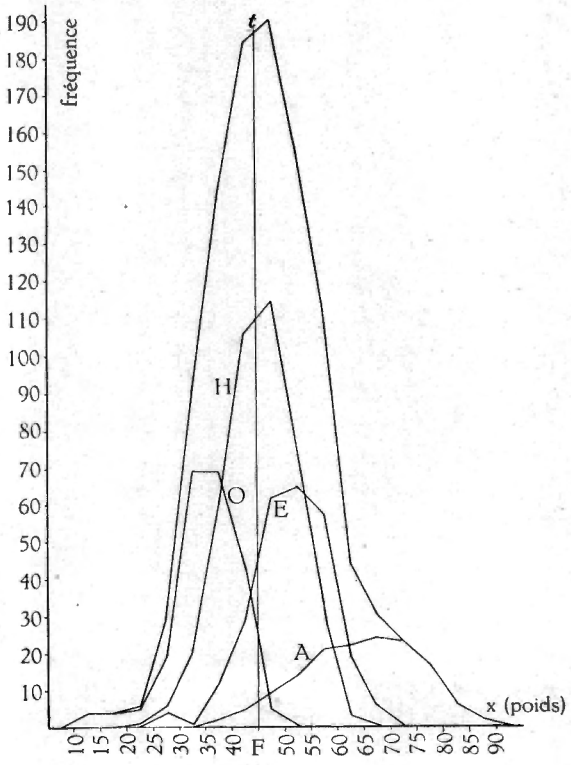

Fig. 5. — A, E, H, O sont les distributions de 4 lignées pures de haricots isolées par E. JOHANNSEN. — *t* est la distribution de l'ensemble de ces 4 lignées pures.

(5) JOHANNSEN, E. (1903, p. 25).

3. On effectue aussi assez souvent des séparations en deux catégories, non pas à l'aide d'un seul caractère continu, mais d'après l'allure globale de la forme des sujets, caractérisée par des adjectifs tels que allongé, arrondi, pointu, bombé, étoilé, etc. Les correctifs « plus ou moins », « plutôt », ... qui accompagnent généralement ces qualificatifs, révèlent d'ailleurs le vague de leur définition (6). Or, il est presque toujours possible de remplacer un terme descriptif de ce genre par des mensurations précises dont la distribution simultanée décrira l'échantillon de façon plus objective.

Comme dans le cas d'une variable, il arrive qu'une telle distribution soit bimodale, et qu'il soit possible de placer une frontière avec moins de 5 % de sujets douteux. Alors, et seulement alors, il est légitime de passer du continu au discontinu, de classer les spécimens en « pointus » et « non pointus », par exemple.

Une classification directe en « pointu » et « non pointu » n'est objective que s'il n'y a aucune confusion possible ; si on n'hésite pas plus d'une fois sur vingt ; si on est sûr qu'un autre ferait la même classification.

Dans tous les autres cas, il est plus sage d'effectuer des mensurations et de considérer un caractère tel que « pointu » comme résultant du jeu de plusieurs variables continues.

§ II. — COMPARAISON DE DEUX ÉCHANTILLONS.

Dans une note sur les Méthodes statistiques en Systématique humaine (7), nous avons indiqué les procédés classiques servant à la comparaison de deux échantillons.

Pour la facilité du lecteur, nous allons les résumer ici ; ils donnent la réponse à trois questions distinctes.

1. Première question : les deux échantillons proviennent-ils de deux populations significativement différentes?

La réponse est donnée par des tests dits d'homogénéité ou de divergence, aussi bien pour les caractères continus que pour les caractères discontinus.

(6) Chez B. H. Burma (1949, pp. 95-96), on trouve une discriminanation basée sur de tels critères descriptifs. Il traite ensuite les deux lots obtenus comme s'ils étaient de provenances différentes !

(7) Defrise, E. (1951, pp. 5-10).

2. DEUXIÈME QUESTION : ces populations sont-elles très différentes ?

La réponse est donnée par la distance généralisée de MAHALANOBIS, dans le cas des caractères continus.

TROISIÈME QUESTION : une différence réelle entre les deux populations étant établie, dans laquelle des deux faut-il placer un nouveau spécimen?

La réponse est donnée par la fonction discriminatoire de FISHER, dans le cas des caractères continus.

PREMIÈRE QUESTION.

1. *Test pour caractères discontinus, du type « A » et « non A ».*

α) Si dans chaque échantillon il y a des A et des non A en proportions différentes, le test de χ^2 (8) dira si cette différence est significative.

Différence significative = les deux populations présentent le caractère A en proportions différentes.

Différence non significative = on n'a aucune preuve pour conclure à la différence des deux populations. En pratique, on admet dans ce cas que les deux populations présentent le caractère A en proportions égales.

β) Naturellement, si l'un des échantillons ne contient que des A, l'autre que des non A, la différence entre les populations est démontrée sans test.

2. *Test pour caractères continus, à distribution de fréquence multivariée normale.*

Grâce à certaines transformations ayant un sens biologique (en passant de x à $\log x$, de y à $\sqrt[3]{y}$, etc.), on peut souvent remplacer une variable non normale par une autre normale, de sorte que l'exigence de la normalité n'est pas une condition tellement restrictive (9).

α) Une variable. Pour comparer les moyennes des deux échantillons, on utilise le test t de STUDENT et pour comparer les déviations standard, le test z de R. A. FISHER (10).

β) Pour plusieurs variables, le test de H. HOTELLING (11)

(8) LAMOTTE, M. (1948, p. 305); L'HÉRITIER, Ph. (1949, p. 46).
(9) QUENOUILLE, M. H., (1950, p. 162).
(10) L'HÉRITIER, Ph., (1949, p. 73).
(11) HOTELLING, H. (1931, p. 360).

indique si les moyennes des deux échantillons sont significativement différentes.

Remarque : lorsque tous les tests indiquent des différences non significatives, on admet, bien qu'on ne puisse le prouver, que les deux populations sont identiques.

Deuxième question.

Pour estimer le degré de divergence de deux populations relativement à une variable normale ou à plusieurs variables liées par une distribution de fréquence multivariée normale, on établit la distance généralisée de P. C. Mahalanobis (12), sorte de distance non géométrique tenant compte des corrélations et dépendant de l'écart entre les valeurs moyennes. Les calculs ne sont possibles que si l'on admet que les variances et les covariances des deux populations sont identiques.

Troisième question.

Dès que la différence entre les moyennes de deux populations multivariées normales est significative (que leur distance généralisée soit faible ou forte), il est important de savoir où classer un nouveau spécimen.

La fonction discriminatoire de R. A. Fisher (traitée au § IV dans le cas de deux variables) est utilisée à cette fin; elle permet en réalité d'atteindre un triple but (13) :

1° Etablir un test de divergence pareil à celui de Hotelling.

2° Classer un nouveau spécimen selon la valeur que ses mesures donnent à la fonction discriminatoire.

3° Evaluer le nombre de mauvaises classifications que l'on fera ainsi, dans l'hypothèse où la probabilité à priori pour un nouveau spécimen d'appartenir à l'une ou l'autre population est la même (14).

§ III. — ANALYSE D'UN ÉCHANTILLON.

1° Si l'on est assuré de la spécificité d'un caractère discontinu donné, on l'utilise d'emblée pour classer les sujets de l'échantillon en deux espèces distinctes (15).

(12) Mahalanobis, P. C. (1930, p. 541) et Fisher, R. A. (1937, p. 378).
(13) Fisher, R. A. (1936, p. 179); id. (1937, p. 376).
(14) Welch, B. L. (1939, p. 218).
(15) Nous adoptons ici la définition d'espèce donnée par L. Cué-

La discrimination est ainsi réalisée. On peut ensuite compléter l'analyse en comparant d'autres caractères des deux espèces par les tests biométriques utilisés dans le cas de deux échantillons.

2° Mais, souvent, la valeur systématique des caractères discontinus n'est pas connue ; alors il ne faut pas s'en servir pour séparer en deux types les spécimens d'un échantillon. On arrive à une meilleure discrimination en commençant l'analyse par des caractères continus.

L'expérience semble montrer qu'une population panmictique fermée suffisamment ancienne est telle que ses caractères continus ont des distributions unimodales. Du moins ne connaissons-nous pas d'exception. On conçoit d'ailleurs que le jeu du regroupement des gênes, l'influence du milieu, toutes ces petites causes régies par le hasard, finissent par donner à la distribution du caractère une allure gaussienne.

Ainsi donc, si une distribution unimodale peut appartenir aussi bien à une population mixte qu'à une population panmictique fermée, une distribution bimodale dénote une population mixte. C'est cette propriété qui est à la base des procédés que nous allons exposer.

Une propriété analogue n'existe pas pour les caractères discontinus. En l'absence de données précises sur leur mode de transmission, il est extrêmement difficile de déduire de leur fréquence si l'on est en présence d'un mélange de deux populations ou d'une population panmictique fermée.

C'est pourquoi une discrimination a plus de chances de se conserver dans les générations suivantes si elle est basée sur des caractères continus, plutôt que sur des caractères discontinus.

Quant à l'interprétation d'une population mixte du point de vue de la systématique, elle n'est pas toujours aisée. Une telle population proviendra parfois du mélange de deux espèces (ou genres) très ressemblants, entre lesquels la barrière empêchant le croisement maintient une distinction morphologique décelable. D'autres fois, on sera simplement en présence de

NOT : L'espèce est une réunion d'individus apparentés ayant même morphologie héréditaire et genre de vie commun, séparée des groupes voisins par quelque barrière, généralement d'ordre sexuel.

deux races dont le croisement est trop récent pour avoir masqué les caractères distinctifs.

1. Analyse graphique d'un échantillon par les caractères continus.

Nous allons maintenant exposer les étapes successives d'une méthode pour effectuer une discrimination au sein d'un échantillon, à l'aide de caractères continus.

On peut distinguer deux stades dans la discrimination :

1° Réussir à décider qu'une population est mixte. Pour cela, il suffira de trouver une distribution à deux modes nettement marqués, soit un simple histogramme, soit une distribution simultanée de deux variables.

2° Dans le cas d'une population mixte, réaliser effectivement la séparation en deux types. Pour atteindre cet objectif, il faudra parvenir à faire apparaître, entre les deux modes, une ligne frontière nette.

Dans les deux cas — que ce soit pour faire apparaître deux modes, ou pour rendre apparente une frontière le plus nette possible, — une distribution bivariée peut réussir là où chacune des deux variables prises séparément aurait été inefficace. Il suffit pour s'en convaincre d'un coup d'œil sur les figures 7 et 8.

Soient x, y, z, t, ... les variables qui désignent les caractères mesurés dans un échantillon.

On construit l'histogramme pour chaque variable en choisissant un intervalle de groupement suffisamment grand pour qu'il n'y ait pas de classes vides, suffisamment petit pour qu'il y ait de 15 à 25 classes.

On retient les variables dont la distribution présente de façon plus ou moins nette deux sommets.

Nous distinguerons trois cas suivant qu'il y a au moins deux variables à distribution bimodale, qu'il n'y en a qu'une seule ou qu'il n'y en a aucune. Dans les deux premiers cas, on est déjà assuré que la population est mixte. Reste seulement à effectuer au mieux la séparation.

1er cas. Il y a au moins deux variables à distribution bimodale.

α) Soient x et y deux distributions bimodales. On construit avec x et y un tableau à double entrée où chaque sujet est figuré par un point de coordonnées x et y (on prend donc ici les

résultats directs des mesures sans faire de groupement). Tous les points forment un nuage qui présentera deux régions distinctes de forte concentration puisque déjà chaque variable x et y prise séparément avait une distribution bimodale. S'il est possible de dissocier le nuage en deux nuages partiels séparés par une ligne frontière nette, la discrimination est réalisée graphiquement (fig. 6) (16). Selon qu'un nouveau sujet se place

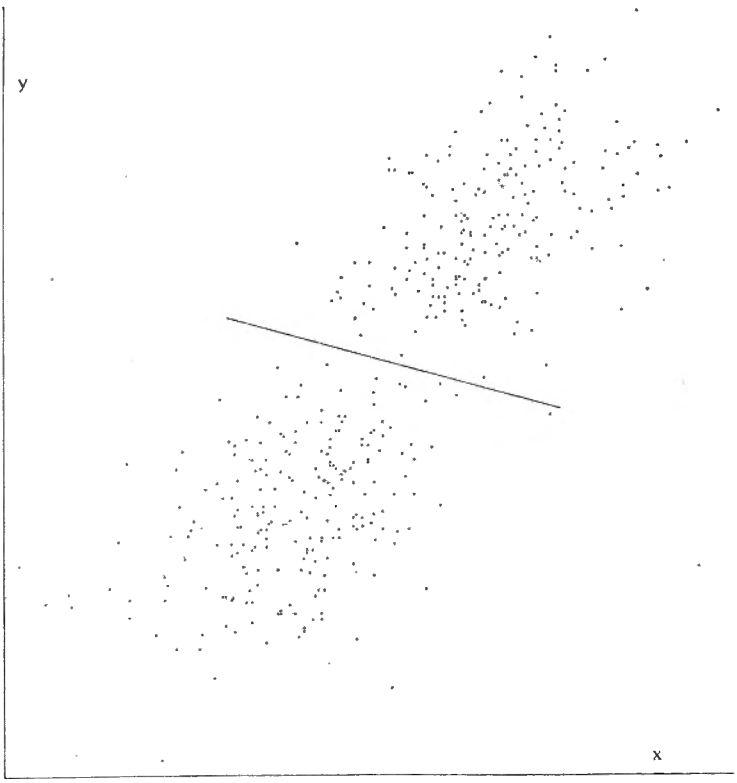

Fig. 6. — Nuage à double concentration, nettement séparable en deux populations par une droite frontière. — x et y ont chacun une distribution bimodale.

d'un côté ou de l'autre de cette frontière, il appartient à l'une ou l'autre population.

Remarquons qu'une discrimination basée sur un tel gra-

(16) Au § IV, nous utiliserons cette frontière comme une première approximation pour en déduire, par le calcul, une meilleure discrimination. Voir, en particulier, la remarque de la page 27.

phique de x et y aura toujours moins de sujets douteux que les discriminations basées sur les histogrammes de x et de y pris séparément. La meilleure discrimination serait donnée par une représentation spatiale à autant de dimensions qu'il y a de variables.

C'est l'impossibilité matérielle d'une telle représentation qui nous oblige à nous borner à des graphiques de deux variables.

β) Si une troisième variable z présente aussi une distribution bimodale, on construit deux nouveaux tableaux : y avec z et z avec x. On a alors en tout trois tableaux, donc chacun présente deux nuages plus ou moins nettement dissociés : en les confrontant, on peut éventuellement corriger les résultats obtenus par la première discrimination.

2me cas. Il y a une seule variable bimodale.

S'il arrive que seule la variable x a une distribution bimodale, on la combine successivement avec chacune des autres jusqu'à ce que l'on trouve un nuage à double concentration, dont les modes sont bien distants (fig. 7). La séparation entre les deux populations a plus de chances d'apparaître nettement sur un tel graphique que sur l'histogramme de x seul : qu'on

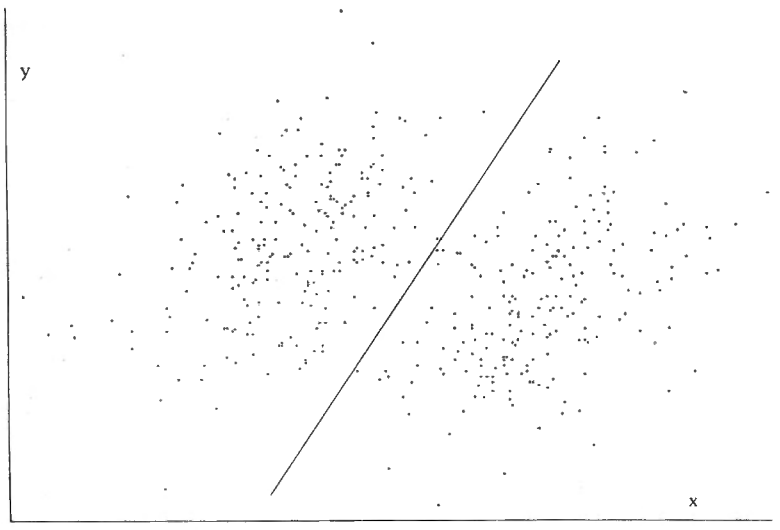

Fig. 7. — Nuage à double concentration, nettement séparable. — x a une distribution bimodale, mais y a une distribution unimodale.

regarde la figure 7 en s'imaginant ce que serait l'histogramme de x.

3me cas. Aucune variable n'a une distribution bimodale.

Examinons enfin le cas le plus défavorable : aucune variable n'a un histogramme à deux sommets.

α) On construit les nuages pour tous les couples de variables, jusqu'à ce que l'on trouve un nuage à double concentration (fig. 8), où l'on essaie alors de tracer une ligne frontière.

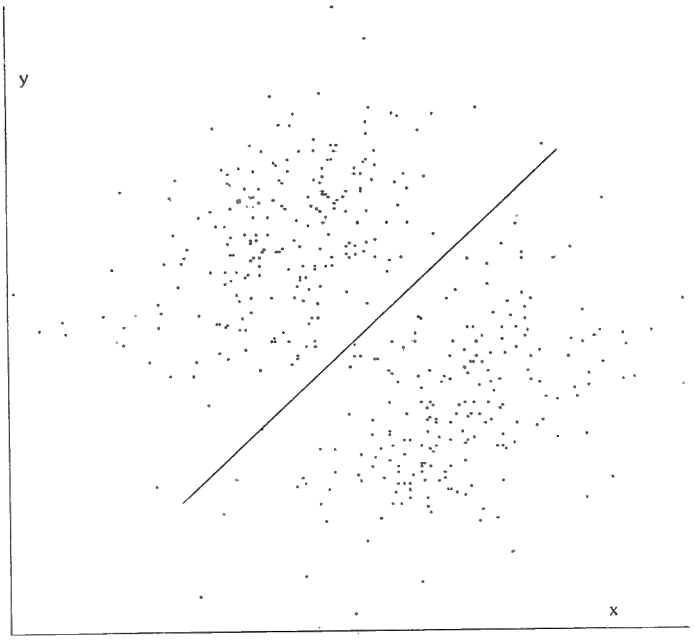

Fig. 8. — Bien que x et y aient chacun une distribution unimodale, on obtient ici un nuage à double concentration, nettement séparable.

β) En cas d'échec, on répète la même opération en adjoignant de nouvelles variables : soit de nouvelles mesures, soit plutôt des fonctions continues simples des premières variables, par exemple $\frac{x}{y}$, $z+t$, $x \times t$, etc.... inspirées par la forme de l'animal.

Si, malgré toutes les tentatives, on n'obtient pas de nuage à double concentration, on admet provisoirement que la population n'est pas mixte.

2. POURSUITE DE L'ANALYSE PAR L'ADJONCTION DES CARACTÈRES DISCONTINUS.

Au début de l'analyse, avant d'utiliser les caractères continus, nous avons signalé la possibilité d'exploitation des caractères discontinus dont la spécificité était certaine. Au stade actuel de l'analyse, nous voulons indiquer les possibilités d'utilisation des caractères discontinus dont la spécificité n'est pas établie. Il s'agira donc de caractères de nature variée : les uns seront dus à des facteurs mésologiques, les autres à des facteurs génétiques. Il y aura sans doute parmi eux des caractères en réalité spécifiques, mais dont la spécificité n'est pas encore reconnue avec certitude. Aussi, le raisonnement qui va suivre est-il suffisamment général pour s'appliquer à tous les caractères discontinus, à condition que leur spécificité n'ait pas été prouvée.

Nous distinguerons trois cas suivant que les nuages fournis par l'analyse graphique des caractères continus sont tous unimodaux, que l'un au moins est à double concentration mais malaisément séparable par une frontière, ou enfin que l'un au moins a été nettement dissocié.

1er cas. Si tous les nuages sont homogènes, on n'a aucune base objective pour conclure que la population est mixte.

On reprend alors les caractères discontinus. Pour chacun d'eux, on examine si l'on trouve un seul type d'individus, ou bien si l'échantillon est divisé en deux classes telles que A et non A.

Si l'on ne trouve jamais qu'une seule modalité, on peut admettre que la population est homogène pour tous les caractères envisagés.

Si, au contraire, on obtient des catégories telles que (A) et (non A), ou encore telles que (A, B), (A, non B), (non A, B), (non A, non B), on doit se borner à noter leur fréquence, sans pouvoir tirer de conclusion. En effet, ces divers types, avec leurs fréquences, pourraient se rencontrer aussi bien dans une population panmictique fermée que dans une population mixte, dans une population hétérozygote aussi bien que dans une population homozygote où certaines influences mésologiques les auraient fait apparaître.

Il arrive cependant que l'on trouve associés plusieurs caractères, par exemple que tous les spécimens se divisent en deux groupes tels que (A, B, C, ...) et (non A, non B, non C, ...). Si l'on est sûr que l'ensemble de ces caractères n'est pas dû

à un seul facteur génétique ou mésologique, on conclut alors à l'existence de deux populations distinctes qui ne s'hybrident pas : la discrimination est effectuée. Toutefois, un pareil cas est assez exceptionnel : n'oublions pas en effet que les populations sont très ressemblantes.

Nous avons dit pourquoi les A et les non A ne forment pas nécessairement des populations distinctes. Ajoutons que même si des tests appliqués à des caractères continus x, y, z, ..., indiquent une différence significative entre le lot des A et celui des non A, cela ne prouve pas que la discrimination en A et non A soit bonne, mais seulement qu'il y a une corrélation, fût-elle légère, entre le caractère A et les caractères x, y, z, ..., intervenant dans les tests. Ainsi, il se peut que dans une certaine population, un test de divergence révèle que les sujets avec des yeux bleus sont plus grands que les sujets avec des yeux bruns. La séparation en yeux bleus et yeux bruns est cependant artificielle si les deux groupes se croisent. Le test ne sert donc ici qu'à montrer la corrélation entre la couleur des yeux et la taille ; il sera intéressant alors d'approfondir l'étude de cette corrélation, d'en rechercher les raisons historiques et biologiques (17).

2me cas. Si l'on a obtenu un nuage à deux modes, on est certainement en présence de deux populations. Mais s'il est impossible de tracer la frontière, en raison du recouvrement partiel des deux populations, s'agit-il de deux espèces ou bien de deux races qui commencent à s'hybrider ?

Seules l'observation de ces populations et certaines expériences pourraient nous l'apprendre.

Cependant, leur description (non pas leur discrimination) se complète éventuellement grâce à tel ou tel caractère discontinu. Si, par exemple, dans les régions du nuage nettement disjointes, on a respectivement des spécimens verts et bleus, les deux populations contiennent principalement l'une des sujets verts, l'autre des sujets bleus. Mais, pas plus que dans le premier cas, la discrimination séparant les bleus des verts n'est complètement satisfaisante tant que l'on ignore le comportement de la couleur dans les hybrides éventuels.

3me cas. Deux nuages, nettement disjoints, où l'on est sûr d'avoir moins de 5 % de sujets douteux, indiquent que l'échantillon comprend deux sortes de sujets, séparables graphique-

(17) Voir par exemple MATHER, K. (1949, p. 7).

ment par une ligne frontière. La population est mixte et l'hybridation, si elle est possible, n'a pas commencé.

Il est intéressant d'examiner alors la répartition des caractères discontinus, car s'il arrive que l'un des nuages partiels ne contienne que des A, et l'autre rien que des non A, une telle circonstance renforcera la valeur de la discrimination et renseignera peut-être sur la nature du caractère A.

Dans le cas où chaque nuage contient des A et des non A, nous pensons que la discrimination basée sur la disjonction des nuages est plus importante du point de vue biologique que la séparation en sujets A et non A. En effet, on doit admettre qu'entre ceux-ci les croisements sont possibles, tant que l'on n'est pas complètement renseigné sur la nature du caractère A. Tandis qu'en séparant les spécimens des deux nuages, on forme deux groupes dont on sait, du moins, qu'ils ne s'hybrident pas.

Ainsi, si une peuplade a donné des graphiques taille/indice céphalique présentant deux nuages nettement disjoints, l'existence de deux populations distinctes est mise en évidence. Au contraire, les groupes sanguins ne définissent pas de telles populations isolées, mais des catégories liées par des mariages.

§ IV. — LA FONCTION DISCRIMINATOIRE DE FISHER.

Après avoir rappelé les points essentiels de la théorie de la fonction discriminatoire de R. A. Fisher (1936 ; 1937), nous calculerons effectivement cette fonction dans le cas de deux variables.

La signification de certains termes, comme distribution multivariée normale, variance intragroupe, covariance, ressortira clairement des formules. D'autre part, la bibliographie indique des ouvrages spécialisés pour le lecteur désireux d'approfondir la question.

1. Généralités.

La fonction discriminatoire de Fisher peut être utilisée si l'on a :

a) soit deux échantillons ;

b) soit un échantillon, mais déjà divisé en deux groupes par les procédés graphiques indiqués au § III.

S'il s'agit de deux échantillons, la fonction discriminatoire de Fisher peut servir d'abord à décider s'ils sont significativement différents (cf. n° 3 ci-dessous).

Si l'on a soit deux échantillons significativement différents, soit deux groupes déjà dissociés au sein d'un échantillon unique, on en déduit l'existence de deux populations ; la fonction discriminatoire de FISHER permet alors de fixer la meilleure frontière entre ces deux populations, pour le classement d'un nouveau spécimen (n° 4) ; il est possible en outre d'estimer la valeur de cette frontière, en calculant le pourcentage de mauvais classements (n° 5).

Sauf au n° 3, nous ne distinguerons pas le cas de deux échantillons et celui de deux groupes au sein d'un échantillon : nous dirons, dans les deux cas, que nous disposons de « deux groupes ».

Si x, y, z, ... sont les variables continues à distribution multivariée normale, ayant mêmes variances et covariances dans chaque groupe, la fonction discriminatoire de FISHER combine linéairement toutes ces variables pour former une variable unique

$$X = b_1 x + b_2 y + b_3 z + \ldots$$

plus sensible que toute autre à l'écart qui existe entre les deux groupes.

Autrement dit, si $\overline{X}_1$, $\overline{X}_2$ désignent respectivement les moyennes de X dans les groupes I et II et σ_x^2 la variance intra-groupe commune, on calcule b_1, b_2, b_3, ... de telle façon que $\dfrac{|\overline{X}_2 - \overline{X}_1|}{\sigma_x}$ soit un maximum.

Selon ses mesures x_i, y_i, z_i, ..., chaque nouveau spécimen fournira une valeur particulière de X, soit X_i, et sera classé dans la population I ou II selon que X_i sera inférieur ou supérieur à $\dfrac{\overline{X}_1 + \overline{X}_2}{2}$.

Cette valeur critique indique la frontière entre les deux populations.

Son choix est tel que le nombre total des sujets mal classés est minimum et que le nombre de sujets I classés dans II et de sujets II classés dans I est égal.

Nous allons effectuer les calculs de la fonction discriminatoire de FISHER

$$X = b_1 x + b_2 y$$

dans le cas de deux variables, sans donner les démonstrations.

X prendra la valeur critique $\frac{\overline{X}_1 + \overline{X}_2}{2}$ pour tous les points x, y, du plan situés sur la droite d'équation

$$b_1 x + b_2 y = \frac{\overline{X}_1 + \overline{X}_2}{2}.$$

Cette droite sera la ligne de meilleure séparation entre les groupes I et II. Selon qu'un nouveau spécimen sera d'un côté ou de l'autre de cette droite, il sera rangé dans la population I ou II.

Remarque. Si l'on a décidé d'opérer avec les variables u et v, où $u = \frac{x}{y}$, $v = z \times t$, il faut d'abord s'assurer que u et v sont distribués normalement (sinon on pourra remplacer v par $\sqrt{v}$, etc...).

La fonction $X = b_1 u + b_2 v$, linéaire en u et v, n'est pas linéaire en x, y, z, t. On adopte cependant $X = b_1 u + b_2 v$, de préférence à la fonction discriminatoire plus sensible calculée directement pour x, y, z, t, en raison de sa représentations graphique commode sur le plan u, v et de la plus grande simplicité des calculs.

2. Calcul de la fonction discriminatoire de Fisher dans le cas de deux variables.

Soient x_1, y_1 les variables d'un spécimen du groupe I et x_2, y_2 celles d'un spécimen du groupe II.

Si les effectifs des deux groupes sont n_1 et n_2, on calcule successivement les moyennes de chaque variable dans chaque groupe

$$\frac{\Sigma x_1}{n_1} = \overline{x}_1, \quad \frac{\Sigma y_1}{n_1} = \overline{y}_1, \quad \frac{\Sigma x_2}{n_2} = \overline{x}_2, \quad \frac{\Sigma y_2}{n_2} = \overline{y}_2$$

et les différences entre les moyennes $d_x = \overline{x}_2 - \overline{x}_1$

$$d_y = \overline{y}_2 - \overline{y}_1$$

Chaque groupe est représenté par un nuage de points sur un graphique de coordonnées x et y. $C_1(\overline{x}_1, \overline{y}_1)$ et $C_2(\overline{x}_2, \overline{y}_2)$ sont les centres de ces nuages.

En posant $n = n_1 + n_2 - 2$, on calcule les quantités A_1, A_2,

B égales à

$$A_1 = n\,\sigma_x^2 = \sum_1^{n1} x_1^2 + \sum_1^{n2} x_2^2 - n_1\,\overline{x_1}^2 - n_2\,\overline{x_2}^2$$

$$(\beta) \qquad A_2 = n\,\sigma_y^2 = \sum_1^{n1} y_1^2 + \sum_1^{n2} y_2^2 - n_1\,\overline{y_1}^2 - n_2\,\overline{y_2}^2$$

$$B = n\,\rho_{xy}\,\sigma_x\,\sigma_y = \sum_1^{n1} x_1 y_1 + \sum_1^{n2} x_2 y_2 - n_1\,\overline{x_1}\,\overline{y_1} - n_2\,\overline{x_2}\,\overline{y_2}$$

σ_x^2, σ_y^2, $\rho_{xy}\,\sigma_x\,\sigma_y$ sont respectivement les variances et la covariance communes aux deux populations, estimées à partir des variances et covariances de chaque groupe (18). On voit que les deux populations ne diffèrent que par leurs moyennes.

La fonction discriminatoire de FISHER est alors

$$X = (A_2 d_x - B d_y)\,x + (A_1 d_y - B d_x)\,y.$$

Les valeurs moyennes de X dans chaque population sont

$$\overline{X}_1 = (A_2 d_x - B d_y)\,\overline{x}_1 + (A_1 d_y - B d_x)\,\overline{y}_1$$

$$\overline{X}_2 = (A_2 d_x - B d_y)\,\overline{x}_2 + (A_1 d_y - B d_x)\,\overline{y}_2$$

3. UTILISATION DE LA FONCTION DISCRIMINATOIRE COMME TEST DE DIVERGENCE DANS LE CAS DE DEUX ÉCHANTILLONS.

Dans le cas d'un échantillon indiquant deux types distincts, il n'est plus nécessaire de faire un test de divergence. Une différence significative entre les deux moyennes est suffisamment prouvée par l'existence de deux nuages totalement dissociés.

Il n'en est pas de même pour deux échantillons. Il se peut que seul un test puisse révéler que la différence entre les deux moyennes est significative.

(18) Si les déviations standard et les coefficients de corrélation sont déjà calculés dans chaque groupe — s_{x_1}, s_{y_1}, r_1 pour le premier, s_{x_2}, s_{y_2}, r_2 pour le second — on peut calculer directement A_1, A_2, B par les formules

$$A_1 = n_1\,s^2_{x_1} + n_2\,s^2_{x_2}$$

$$A_2 = n_1\,s^2_{y_1} + n_2\,s^2_{y_2}$$

$$B = n_1\,r_1\,s_{x_1}\,s_{y_1} + n_2\,r_2\,s_{x_2}\,s_{y_2}$$

On calcule

$D = \overline{X}_2 - \overline{X}_1 = A_2 d_x^2 + A_1 d_y^2 - 2B d_x d_y$, quantité toujours positive ou nulle

et $n' \sigma_x^2 = D(A_1 A_2 - B^2) = n^2 \sigma_x^2 \sigma_y^2 (1 - \rho_{xy}^2) D$,

où n' est le nombre de degrés de liberté valant $n_1 + n_2 - 3$ et où σ_x^2 est la variance intragroupe estimée de X (la variance intragroupe de X est proportionnelle à la somme des carrés des écarts, dans chaque groupe, entre la moyenne de X et les valeurs individuelles de X).

La déviation standard de $\overline{X}_2 - \overline{X}_1$ est $\sigma_x \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, de sorte que le t de STUDENT vaut

$$t = \frac{\overline{X}_2 - \overline{X}_1}{\sigma_x \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Si, pour n' degrés de liberté, les tables de STUDENT (19) indiquent que la valeur de t est significative, il y a une réelle différence morphologique entre les deux échantillons.

4. CLASSEMENT D'UN NOUVEAU SPÉCIMEN ; TRACÉ DE LA DROITE (d) DE MEILLEURE SÉPARATION ENTRE DEUX ÉCHANTILLONS OU ENTRE DEUX TYPES NETTEMENT DISSOCIÉS DÉCOUVERTS DANS UN ÉCHANTILLON UNIQUE.

1° Un nouveau spécimen, de mesures x_i, y_i, ayant comme valeur de X

$$X_i = (A_2 d_x - B d_y) x_i + (A_1 d_y - B d_x) y_i$$

se classe dans la population I ou II suivant que X_i est inférieur ou supérieur à $\frac{\overline{X}_1 + \overline{X}_2}{2}$.

2° Mais on peut éviter de refaire le calcul séparé de X pour chaque nouveau sujet : sur le graphique où figurent les nuages de points relatifs à chaque groupe, on trace la droite (d) de meilleure séparation, d'équation

$$(d) \qquad (A_2 d_x - B d_y) x + (A_1 d_y - B d_x) y = \frac{\overline{X}_1 + \overline{X}_2}{2}.$$

(19) FISHER, R. A. & YATES, F. (1948, p. 32).

Celle-ci peut se mettre sous la forme

$$y - \frac{\bar{y}_1+\bar{y}_2}{2} = m\left(x - \frac{\bar{x}_1+\bar{x}_2}{2}\right) \quad \text{avec} \quad m = \frac{A_2 d_x - B d_y}{B d_x - A_1 d_y}$$

comme coefficient angulaire.

La droite (d) passe par le point M de coordonnées $\frac{\bar{x}_1+\bar{x}_2}{2}$, $\frac{\bar{y}_1+\bar{y}_2}{2}$, milieu de C_1C_2 (fig. 9).

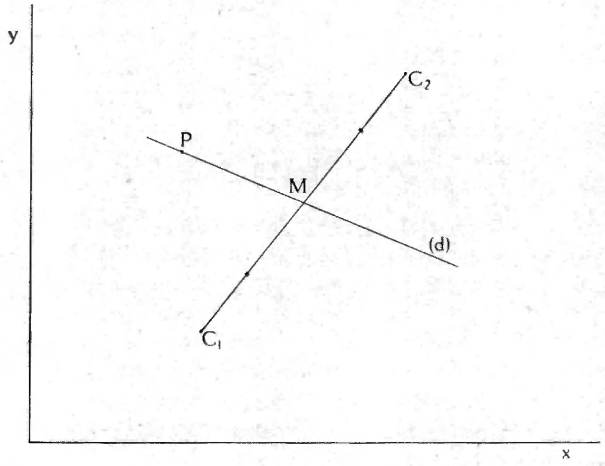

Fig 9. — C_1 $(\bar{x}_1, \bar{y}_1)$ et C_2 $(\bar{x}_2, \bar{y}_2)$ sont respectivement les centres des populations I et II.

M $\left(\frac{\bar{x}_1+\bar{x}_2}{2}, \frac{\bar{y}_1+\bar{y}_2}{2}\right)$ est le milieu de C_1C_2.

Les points M et P (α, β) définissent la droite (d) de meilleure séparation.

Il est facile d'en calculer un deuxième point P de coordonnées α, β :

On donne à x une valeur quelconque α, et on calcule la valeur correspondante de y, soit

$$\beta = m\left(\alpha - \frac{\bar{x}_1+\bar{x}_2}{2}\right) + \frac{\bar{y}_1+\bar{y}_2}{2}$$

En joignant M à P on obtient la droite (d).

Selon qu'un nouveau point x_i, y_i sera du côté de la droite où figure C_1 ou C_2, il appartiendra à la population I ou II.

5. Calcul du nombre de sujets mal classés si l'on prend (d) comme frontière entre les deux populations.

Appliquons cette méthode pour classer 100 nouveaux spécimens dont nous savons seulement qu'ils appartiennent à l'une ou l'autre population.

Le nombre de sujets mal classés sera minimum, mais comment le calculer ?

Les tables de la distribution normale (20) indiquent le pourcentage de sujets d'une population normale situés au delà de la moyenne augmentée de $s = \frac{\overline{X}_2 + \overline{X}_1}{2\sigma_x}$. Ce pourcentage indique le nombre de spécimens mal classés parmi les 100 sujets à répartir dans les deux populations.

Pour en avoir moins de 5 %, il faut que $s \geqslant 1{,}64$.

Le nombre d'erreurs est d'autant plus faible que s est plus grand. On comprend mieux maintenant pourquoi la fonction a été choisie de façon à rendre maximum la quantité $\frac{|\overline{X}_2 - \overline{X}_1|}{\sigma_x} = 2s$, car au maximum de s correspond un minimum de sujets mal classés.

Remarque. Dans le cas d'un échantillon, la séparation préalable en deux groupes, réalisée graphiquement par les procédés du § III, a été utilisée comme une première approximation pour déterminer par le calcul une meilleure frontière : la droite (d). Cependant, si pour la droite (d) obtenue, le nombre de sujets mal classés est inférieur à 5 %, le recouvrement des deux populations est faible et les fluctuations de la première frontière graphique ne peuvent avoir qu'une influence négligeable sur la position de (d).

6. Une propriété de la droite (d).

Pour rendre plus concret l'aspect des nuages de points, entourons chacun d'une ellipse d'égale probabilité, contenant un pourcentage donné de sujets.

Cela est possible puisque chaque population est considérée comme normale. Nous avons supposé en outre que les variances σ_x^2, σ_y^2 et la covariance $\rho_{xy}\ \sigma_x\ \sigma_y$ sont communes (voir p. 24,

(20) Fisher, R. A. & Yates, F. (1948, p. 31).

formule β), de sorte que les éléments de fréquence des distributions sont respectivement

$$df_1 = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2_{xy}}}\, e^{-\frac{L_1}{2}}\, dx\, dy ;$$

$$df_2 = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2_{xy}}}\, e^{-\frac{L_2}{2}}\, dx\, dy ;$$

avec

$$L_t = \frac{1}{1-\rho^2_{xy}}\left\{\frac{(x-\bar{x}_t)^2}{\sigma_x^2} - \frac{2\rho_{xy}(x-\bar{x}_t)(y-\bar{y}_t)}{\sigma_x\sigma_y} + \frac{(y-\bar{y}_t)^2}{\sigma_y^2}\right\}$$

avec $t = 1, 2.$

$L_1=\lambda^2$, $L_2=\lambda^2$ sont les équations de deux ellipses d'égale probabilité, E_1, E_2, contenant chacune le même nombre de sujets de la population à laquelle elle se rapporte. Pour avoir 95 % de sujets dans une ellipse, on prend $\lambda^2 = 5{,}99$; pour avoir 99 %, il faut que $\lambda^2 = 9{,}21$. Ce sont les tables de χ^2 qui indiquent les valeurs de λ^2 correspondant à un pourcentage donné (21), car λ^2 est distribué comme χ^2 avec deux degrés de liberté.

Pour une même valeur de λ^2, les deux ellipses E_1 et E_2 sont égales et leurs axes sont parallèles ; elles ne diffèrent que par leurs centres respectifs C_1 et C_2 (fig. 10).

La droite (d) passe par les deux points d'intersection de E_1 et E_2, I et J.

Démonstration. Tous les points des ellipses E_1 et E_2 ont des fréquences égales. Le lieu des points du plan ayant une fréquence identique dans les deux populations a comme équation $L_1 = L_2$, et doit contenir les points I et J.

A cause de l'hypothèse de l'égalité des variances et de la covariance dans les deux populations, cette équation se réduit aux termes du premier degré en x et y.

Tous calculs faits, on trouve

$$y - \frac{\bar{y}_1+\bar{y}_2}{2} = \frac{\sigma_y^2 d_x - \rho_{xy}\sigma_x\sigma_y d_y}{\rho_{xy}\sigma_x\sigma_y d_x - \sigma_x^2 d_y}\left(x - \frac{\bar{x}_1+\bar{x}_2}{2}\right)$$

(21) Fisher, R. A. & Yates, F. (1948, p. 33).

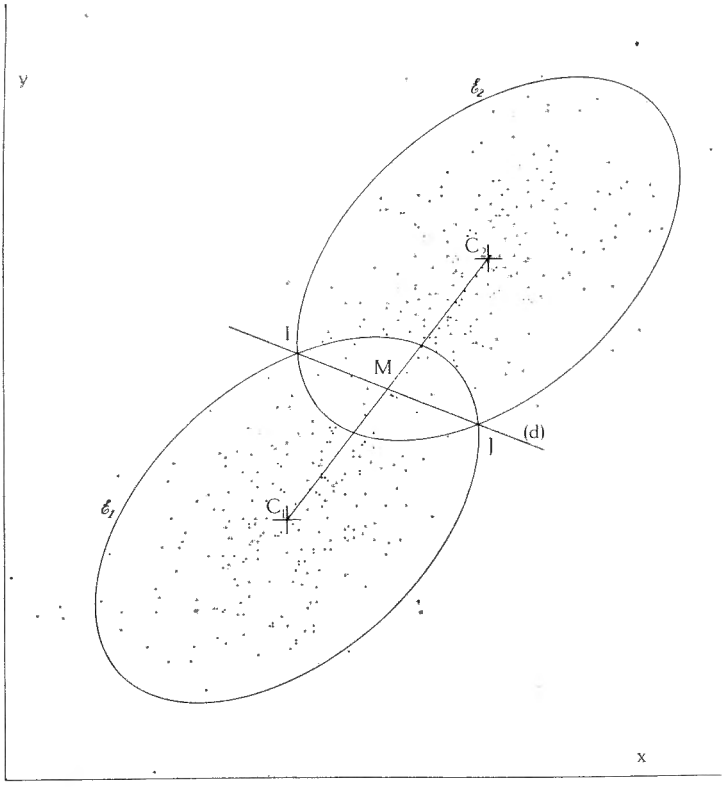

Fig. 10. — L'ellipse E_1 de centre C_1 renferme 95 % des sujets de la population I ($\lambda^2 = 5{,}99$).
L'ellipse E_2 de centre C_2 renferme 95 % des sujets de la population II ($\lambda^2 = 5{,}99$).
Pour obtenir la droite (d), il suffit de joindre IJ. Il y a ici moins de 5% de sujets mal classés.

qui est précisément l'équation de la droite (d). Cette droite est donc le lieu des points d'égale fréquence dans les deux populations et passe par les points I et J.

Généralisation de cette propriété.

Dans l'espace à 3 dimensions, les nuages de points représentant les deux populations — toujours supposées normales à mêmes variances et covariances — sont séparés par un plan, lieu des points d'égale fréquence dans l'une et l'autre population. L'équation de ce plan est donnée par la fonction discriminatoire pour trois variables, où X prend la valeur critique

$$\frac{\overline{X}_1 + \overline{X}_2}{2}.$$

La propriété se conserve de même lorsqu'il y a plus de trois variables.

CONCLUSION.

Il existe des méthodes biométriques objectives et sûres, applicables à ce que l'on peut appeler la systématique à petite échelle, c'est-à-dire celle qui concerne les populations très ressemblantes :

1° des méthodes classiques pour comparer deux ou plusieurs échantillons, qui résultent de théories assez compliquées du calcul des probabilités, mais sont elles-mêmes d'un usage facile.

2° des procédés graphiques, mis au point dans cette note et servant à séparer deux ou plusieurs types au sein d'un échantillon unique ; ils n'offrent aucune difficulté, ni dans la théorie, ni dans les applications.

Cette note contient l'exposé de ces méthodes, en elles-mêmes très simples. Nous avons jugé utile d'y ajouter la discussion de certains points délicats qui doivent être examinés au début de l'enquête biométrique et au moment de tirer les conclusions.

Avant d'aborder l'étude des populations voisines, il faut formuler clairement la question à laquelle on cherche une réponse, car de la nature de cette question dépend le choix de la méthode biométrique. D'autre part, il ne faut pas hésiter, le cas échéant, à abandonner au profit d'un autre tel matériel dont le traitement biométrique exigerait un effort disproportionné à l'importance du problème posé. Cette réflexion préliminaire peut aussi concourir à éviter des conclusions dénuées de sens.

Le choix s'étant fixé sur une méthode convenable, on peut l'appliquer de façon presque mécanique, à condition d'opérer sur un nombre suffisant de spécimens récoltés au hasard (22). L'effectif minimum dépend de la variabilité de la population. On doit adopter pour les mensurations — prises d'après une méthode standardisée — une échelle suffisamment fine, mais en rapport avec les erreurs inévitables de mesure et avec la variabilité de l'ensemble de la population.

Au moment d'interpréter les différences morphologiques éventuellement révélées, on se trouve devant de réelles difficultés — inhérentes, croyons-nous, à de nombreuses discri-

(22) Dans les conditions exposées au § I.

minations en systématique et qui ne tiennent donc pas aux méthodes biométriques employées.

Un obstacle fondamental à une bonne discrimination est l'ignorance profonde où l'on est généralement de la nature biologique des caractères discontinus utilisés. La plupart du temps, l'examen, même approfondi, des spécimens de l'échantillon ne suffit pas ; seule l'observation dans la nature et des élevages expérimentaux aident à comprendre comment les populations vivent, se multiplient et de quelle façon les caractères observés s'acquièrent et se transmettent. C'est dire que très souvent, à cause de l'impossibilité matérielle de tels examens, notamment en paléontologie, le classement basé sur ces caractères reste provisoire et un peu conventionnel.

On peut d'ailleurs se demander s'il est assuré, à priori, que la systématique à une échelle si fine soit possible. Peut-être, en effet, poursuit-on un but inaccessible et doit-on finalement arriver à l'individu en voulant pousser si loin la classification animale. Car existe-t-il réellement entre tous les groupes des barrières nettes que des efforts suffisants finiront toujours par mettre en évidence ? La difficulté de définir clairement le concept de race humaine pourrait nous en faire douter.

Le naturaliste admettrait aisément ce point de vue si son désir d'une nomenclature sans aucune frange d'indétermination ne lui faisait pas considérer comme idéales les formes bien distinctes.

Cependant, un usage rationnel de la biométrie pourrait conduire à un autre genre de classification, aisée et objective, même lorsqu'on n'aurait pas de séparations nettes entre les groupes. Des graphiques, aussi nombreux qu'il le faudrait, fixeraient, sous forme de nuages de points, tous les sujets d'une population ou même de plusieurs populations voisines. Chaque nouveau sujet serait déterminé par la place qu'il occupe sur ces graphiques : il serait ainsi parfaitement situé, sans que des cloisons étanches entre les divers points de forte concentration soient pour cela nécessaires. Au lieu du « type » isolé de la systématique classique, les valeurs moyennes, la variabilité, les coefficients de corrélation des caractères continus de tous ces nuages caractériseraient ces populations.

On aurait ainsi tenu compte de la variation souvent continue d'un groupe à l'autre, alors qu'on regarde fréquemment cette donnée comme gênante, simplement parce qu'elle ne permet pas l'application des méthodes usuelles de la systématique.

Au niveau inférieur de la classification systématique, nous croyons aussi qu'il ne serait pas impossible d'adapter les règles de nomenclature à cette continuité apparente si souvent mise en lumière par la biométrie.

Pour illustrer notre pensée, évoquons, dans un autre domaine, l'exemple des couleurs, autrefois uniquement désignées par des noms, et aujourd'hui repérées dans des graphiques standardisés qui traduisent la continuité effective des couleurs. Il n'en reste pas moins que les noms anciens subsistent pour nommer les tons les plus usuels.

En terminant, je tiens à exprimer tous mes remerciements au Dr F. Twiesselmann pour de nombreuses suggestions et remarques qui m'ont été précieuses. Je dois beaucoup aussi à plusieurs membres de l'Institut royal des Sciences naturelles qui m'ont obligeamment fourni des renseignements sur diverses questions en systématique, et notamment à M. Glibert qui a bien voulu relire mon manuscrit.

RÉSUMÉ.

But : discriminer des populations très ressemblantes.

Distinction entre un et plusieurs échantillons : suivant la provenance, et non d'après des critères morphologiques.

Caractères continus et discontinus. Le passage des premiers aux seconds peut se faire dans certains cas, à condition de prendre de nombreuses précautions. La séparation d'un échantillon en sujets A et non A ne conduit pas nécessairement à deux groupes homogènes.

Comparaison de deux échantillons : divers tests, distance généralisée de P. C. Mahalanobis, fonction discriminatoire de R. A. Fisher.

Analyse d'un échantillon.

1° Si A est un caractère spécifique certain, les A et les non A constituent deux espèces que l'on peut comparer ensuite par les méthodes applicables à deux échantillons.

2° Si l'on ne trouve pas un tel caractère discontinu spécifique, on utilise d'abord les caractères continus avec lesquels on établit des graphiques à deux variables.

α) Si l'on ne trouve aucun nuage à double concentration, il est impossible de prouver que la population est mixte. Si l'on se tourne vers les caractères discontinus (à spécificité non reconnue), un caractère A peut réaliser un partage en deux groupes A et non A, mais qui ne sont pas nécessairement homogènes.

Dans certains cas, plusieurs caractères discontinus peuvent se trouver associés et permettre une bonne discrimination.

β) Si l'on obtient un nuage à double concentration, on a une population mixte. Mais il arrive qu'on ne puisse réaliser la séparation, en raison du recouvrement partiel des deux populations. La description de telles populations se complète heureusement par l'examen de caractères discontinus.

γ) Si, au contraire, on trouve deux nuages nettement séparables, la discrimination est faite; on peut examiner ensuite la répartition des caractères discontinus.

Fonction discriminatoire de Fisher. Cette fonction sert à fixer la meilleure frontière entre deux populations. Elle est calculée ici dans le cas de deux variables. Elle peut servir de test de divergence lorsque les populations ont des provenances distinctes. Dans tous les cas, elle donne l'emplacement de la droite limite (*d*) qui sépare les deux populations. Calcul du nombre de spécimens mal classés. Propriété de la droite (*d*).

Index bibliographique.

Burma, B. H., 1949, *Studies in quantitative paleontology. II. Multivariate analysis - a new analytical tool for paleontology and geology.* (Journal of Paleontology, London, vol. 23, n° 1, 1949.)

Cuénot, L., 1936, *L'espèce.* (Doin, Paris.)

Defrise, E., 1951, *Des méthodes statistiques en systématique humaine.* (Bull. de l'Inst. royal des Sc. natur. de Belgique, tome XXVII, n° 57.)

Fisher, R. A., 1936, *The use of multiple measurements in taxonomic problems.* (Ann. of Eugenics, London, t. 7, p. 179.)

— , 1937, *The statistical utilization of multiple measurements.* (Ann. of Eugenics, London, t. 8, p. 376.)

Fisher, R. A. & Yates, F., 1948, *Statistical tables.* (Oliver and Boyd, Edinburgh.)

Gates, R. R., 1946, *Human genetics.* (Mac Millan, New-York.)

Guyénot, E., 1931, *L'Hérédité.* (Doin, Paris).

— , 1950, *La Variation.* (Doin, Paris.)

HOTELLING, H., 1931, *The generalisation of Student's ratio.* (Ann. Math. Statistics, 2, p. 360.)
HUXLEY, J., 1940, *The new Systematics.* (Oxford University Press.)
JOHANSSEN, E., 1903, *Erblichkeit in Populationen und in Reine Linien.* (Fischer, Jena.)
LAMOTTE, M., 1948, *Introduction à la biologie quantitative.* (Masson, Paris.)
L'HÉRITIER, Ph., 1949, *Les méthodes statistiques dans l'expérimentation biologique.* (CNRS, Paris.)
MAHALANOBIS, P. C., 1930, *On tests and measures of group divergence. I.* (Journ. Asiat. Soc. Bengal, Calcutta, t. 26.)
MATHER, K., 1949, *Biometrical genetics, the study of continuous variation.* (Methuen, London.)
QUENOUILLE, M. H., 1950, *Introductory Statistics.* (Butterworth-Springer, London.)
SCHENK, E. T. & MAC MASTERS, J. H., 1932, *Recommendations for procedure in the description of a new species.* (Edw. Brothers Inc., Ann. Arbor, Michigan.)
WELCH, B. L., 1939, *Note on discriminant functions.* (Biometrika, London, 31, p. 218.)

INSTITUT ROYAL DES SCIENCES NATURELLES DE BELGIQUE.
LABORATOIRE D'ANTHROPOLOGIE ET DE PRÉHISTOIRE.