Institut royal des Sciences naturelles de Belgique

BULLETIN

Tome XXVII, n° 57.
Bruxelles, septembre 1951.

Koninklijk Belgisch Instituut voor Natuurwetenschappen

MEDEDELINGEN

Deel XXVII, nr 57.
Brussel, September 1951.

DES MÉTHODES STATISTIQUES EN SYSTÉMATIQUE HUMAINE (*),

par Elisabeth DEFRISE-GUSSENHOVEN (Bruxelles).

INTRODUCTION.

Le but de la systématique humaine est non pas simplement de classer les êtres vivants, mais également de détecter « l'évolution en action » (12).

Cette conception nous conduit :

a) à rechercher l'origine de la continuité et de la discontinuité des groupes humains;

b) à décrire ces groupes au point de vue de leur continuité et de leur discontinuité dans le temps et dans l'espace.

a) La génétique constitue la base des recherches sur l'origine de la continuité et de la discontinuité des groupes humains.

Un groupe continu résulte de ce que les homozygotes pour n caractères ont des descendants homozygotes pour ces mêmes caractères, tandis que la discontinuité semble causée principalement par les mutations. Suivies de sélection et d'isolement, les mutations peuvent scinder une population homozygote (pour n caractères) en deux groupes bien différenciés.

Cependant, même la connaissance complète de la formule génétique ne suffirait pas toujours à dévoiler les origines d'une population. En effet, remis en présence, deux groupes différenciés peuvent se refondre et donner une population mélangée.

(*) Communication présentée au IIIme Congrès National des Sciences, Bruxelles, 1950.

Au début, de nombreux caractères liés révèlent cette double origine. Mais, si les couples se forment au hasard, ces liaisons disparaissent au fur et à mesure que les générations se suivent (2). Ainsi, sauf par son histoire, cette population mélangée ne se distingue en rien d'un groupe où, dès leur apparition, les mutants se seraient croisés avec les types primitifs, sans sélection ni isolement.

Pour retrouver les origines d'une population, l'anthropologie devra donc avoir recours à d'autres disciplines que la génétique (sociologie, ethnographie, histoire, préhistoire).

Les mathématiques appliquées aux problèmes de génétique, de sélection et de migration forment un domaine particulier, et combien vaste, où les beaux travaux de F. Bernstein, R. A. Fisher, J. B. S. Haldane, L. S. Penrose, S. Wright peuvent servir de guide; nous n'en parlerons pas ici. Signalons cependant, d'une part, les récents progrès de la génétique animale dans les questions de la transmission héréditaire de certains caractères continus — tels, chez *Drosophila melanogaster,* l'interruption d'une veine de l'aile (1), ou le nombre de poils abdominaux (18) ; d'autre part, en génétique humaine, les tentatives pour séparer les caractères continus mésolabiles et mésostabiles, en se basant, soit sur des taux de croissance (22), soit sur des enquêtes de jumeaux.

b) Description de la continuité et de la discontinuité des populations.

La connaissance de l'hérédité des caractères anthropologiques permettrait une description précise de l'Humanité. Privé de cette connaissance, nous devons cependant décrire les populations avant que certains groupes distincts n'aient eu le temps de se refondre en populations où l'équilibre des gènes arrive à masquer l'hétérogénéité initiale.

En analysant par la biométrie les caractères utilisés en systématique classique et dont la transmission reste inconnue, on arrive à jeter quelque lumière sur la composition des groupes humains. Celle-ci est éclairée dans sa genèse par la paléontologie dont l'emploi exclut également la génétique.

La note actuelle a pour objet de signaler les méthodes statistiques qui facilitent la description de la continuité et de la discontinuité des divers groupes humains.

1. — Groupes. Précisons d'abord ce que nous entendons ici par groupe ou population : c'est un ensemble d'hommes

définis par leur habitat, leur condition sociale, leur langue, leur religion, leur âge, etc..., mais par aucun caractère physique. Ceux-ci servent à décrire le groupe, non pas à le sélectionner. En effet, comme l'étude de l'anthropologie concerne justement les caractères physiques, nous devons choisir nos sujets sur une base non morphologique, afin de ne pas faire une pétition de principe.

Ajoutons encore que les groupes ainsi définis selon des critères objectifs, ne nous sont connus en pratique que par des échantillons. Nous renvoyons le lecteur aux ouvrages de R. A. Fisher (8) et M. G. Kendall (13), pour les questions très importantes d'échantillonnage et d'estimation.

2. — Caractères. Pour la facilité de notre exposé, nous classons les différents caractères utilisés en systématique dans le tableau suivant :

	Caractères discontinus	Caractères continus
Caractères qualitatifs	Groupes sanguins Nature du cheveu 1	Couleur de la peau 2
Caractères quantitatifs	4	3 Taille

Les caractères discontinus qualitatifs dans une population sont donnés avec leurs fréquences : fréquence de gènes (groupes sanguins) ou fréquence de phénotypes (nature du cheveu), suivant que la transmission mendélienne est connue ou non.

Pour faciliter le traitement biométrique, il est avantageux de transformer les caractères qualitatifs continus en caractères quantitatifs continus (en exprimant par un nombre le degré de pigmentation de la peau) (4) et (18). On tient mieux compte ainsi de la réalité biologique qu'en passant à des caractères qualitatifs discontinus par des subdivisions artificielles (en groupant grossièrement les peaux claires et les peaux foncées) (*).

(*) Ajoutons qu'il faut parfois considérer certains caractères qualitatifs continus comme pluridimensionnels : il se peut qu'un seul nombre ne suffise pas pour caractériser la couleur de la peau ; qu'il en faille par exemple un pour la teinte, un autre pour l'intensité.

Les caractères continus qualitatifs sont ainsi assimilés aux caractères continus quantitatifs dont le type est la taille. Bien qu'influencée par le milieu, la taille se transmet des parents aux enfants comme en témoigne une corrélation positive, mais on ignore par quel mécanisme (10).

On peut décrire une population par la fonction de distribution de ses différents caractères continus. A chaque ensemble de valeurs particulières des variables correspond alors une probabilité, faible pour les valeurs extrêmes des caractères, plus forte vers le centre, et il est essentiel de rechercher si la distribution présente plusieurs points de densité maxima, c'est-à-dire de probabilité maxima.

Le nombre variable de pois dans une cosse (20) est un bon exemple de caractère discontinu quantitatif ; on n'en utilise guère en anthropologie. Bien entendu, la nécessité de grouper les données pour la mesure fait qu'en pratique on passe fréquemment des variables continues aux variables discontinues.

3. — Composition d'un groupe. Pour décrire un groupe par un caractère discontinu, tel la nature des cheveux, on donne simplement la fréquence des divers phénotypes observables. Mais lorsqu'il s'agit d'un caractère à hérédité connue (groupes sanguins), on peut passer de la fréquence des phénotypes à celle des gènes. Les relations entre les fréquences des gènes allélomorphes montrent si le groupe a atteint l'équilibre génique vers lequel doit tendre une population panmictique fermée (2). Si cette stabilisation ne s'est pas effectuée, on peut en rechercher la cause : une immigration récente, un isolement religieux ou social avec mariages préférentiels, etc...

La description d'un groupe à l'aide de caractères continus se fait, nous l'avons vu, par la fonction de distribution simultanée. En étendant les résultats de la génétique classique aux caractères continus, très mal connus à ce point de vue, nous admettrons que deux sommets dans la fonction de distribution indiquent un mélange récent ou l'existence de mariages préférentiels, l'état d'équilibre de la population donnant une fonction de distribution à un sommet. Cette hypothèse, assez naturelle, n'a jamais pu être vérifiée.

Il y a une méthode graphique très simple pour déceler un double sommet. On établit le polygone de fréquence pour chaque caractère, en prenant un petit intervalle de groupement ; on retient les variables accusant nettement deux sommets. Placés

deux par deux dans des graphiques de corrélation, ces variables donnent des nuages de points, dans lesquels on tâche de trouver deux zones distinctes de densité maxima. Les renseignements tirés de tels diagrammes suffisent quelquefois et évitent le calcul effectif de la fonction de distribution.

4. — Comparaison de deux groupes.

Méthode graphique. On peut opérer exactement de la même façon pour comparer deux populations. On choisit, parmi les caractères continus, ceux dont les distributions ne se recouvrent pas et on les place deux par deux dans des diagrammes à double entrée. L'éloignement des deux nuages met en relief la différence entre les deux populations.

Tests et mesures de divergence. R. A. Fisher (5), C. C. Seltzer (21) et P. C. Mahalanobis (17) ont attiré l'attention sur la distinction à faire entre un test et une mesure de divergence. Cette distinction est essentielle : elle permet un usage judicieux des formules proposées et met en lumière les défauts de coefficients anciennement employés, comme celui de ressemblance raciale » (19), qui fut utilisé souvent comme mesure de divergence alors qu'il est en réalité un test.

Rappelons, par un exemple simple, en quoi consiste cette différence. Deux échantillons d'effectifs n et n' sont tirés de deux populations ; les moyennes et les déviations standard calculées pour un caractère continu, la taille par exemple, sont respectivement m et m', s et s'.

Les populations sont-elles réellement différentes ou bien l'écart entre les moyennes $d = m - m'$ est-il dû aux hasards de l'échantillonnage ? La réponse est donnée par un test de divergence, par exemple celui de Student. On fait l'hypothèse de travail suivante, commune à tous les tests de divergence : les deux populations sont identiques. L'estimation de leur variance commune est alors $\sigma_e^2 = \dfrac{n s^2 + n' s'^2}{n + n' - 2}$ et le t de Student $= \dfrac{m - m'}{\sigma_e \sqrt{1/n + 1/n'}}$. Tous les manuels de biométrie contiennent des tables (pour le cas de variables normalement distribuées) où l'on trouve, en regard de t, la probabilité correspondante pour $n + n' - 2$ degrés de liberté (9). Une probabilité inférieure à 0,05 indique que la différence d est significative. Si la probabilité est supérieure à 0,05, on admet

l'hypothèse de l'identité des deux populations ; cependant, une probabilité même très élevée n'est pas une preuve formelle d'identité, elle indique simplement qu'on n'a aucune raison de rejeter l'hypothèse adoptée.

La quantité t augmente avec d aussi bien qu'avec l'effectif des échantillons, de sorte que si t est très grand, cela ne signifie pas nécessairement que les populations sont très différentes, mais seulement qu'il y a plus de sujets qu'il n'en faut pour prouver cette différence.

Les populations sont-elles très différentes ? La réponse est donnée par une mesure de divergence, que MAHALANOBIS définit ainsi (17) :

« C'est une estimation quantitative de la différence entre les deux groupes ; elle doit répondre aux conditions suivantes :

— être un scalaire, positif ou nul, indépendant des unités des variables ;
— s'annuler lorsque les deux populations se confondent ;
— être constant, aux erreurs d'échantillonnage près, pour des épreuves successives ;
— augmenter avec la différence des moyennes. »

Et nous ajouterons cette condition : une bonne mesure de divergence, calculée à partir des échantillons, doit avoir une distribution connue, permettant une estimation optima. Dans notre exemple, $\frac{(m-m')^2}{\sigma_e^2}$ est une mesure de divergence ; elle augmente uniquement avec d.

Reprenons la comparaison de deux groupes humains :

A. — Tests de divergence.

Variables discontinues : on applique le test d'homogénéité χ^2 et cela est possible quel que soit le nombre de caractères envisagés simultanément (14) et (15).

Variables continues : nous n'envisageons que les variables continues qui ont une fonction de distribution simultanée normale (13).

Pour une seule variable, la comparaison de deux groupes se base sur le test t de STUDENT. Ce critère a été étendu au cas de plusieurs variables par HOTELLING (13 et 11). Le T^2 de HOTELLING est donc un test de divergence applicable à p variables en corrélation normale.

$$T^2 = \frac{n\ n'}{n + n'}\ s^{ij}\ d_i\ d_j \qquad (i, j = 1, 2, \ldots, p;$$

n et n' sont les effectifs des deux échantillons).

s_{ij} est la covariance des deux populations supposées identiques, estimée à partir des covariances de chaque échantillon.

s^{ij} est le mineur normé de la matrice des covariances s_{ij}.

HOTELLING a calculé la fonction de distribution de T^2 et FISHER (6) a montré qu'un test de signification de T^2 est donné par sa distribution z (8), où

$$e^{2z} = \frac{T^2\ (n + n' - p - 1)}{(n + n' - 2)\ p} \text{ avec } \nu_1 = p \text{ et } \nu_2 = n + n' - p - 1.$$

Pour une seule variable, $T^2 = t^2$ de STUDENT ; le test de HOTELLING généralise bien celui de STUDENT. A l'opposé du « coefficient of racial likeness », T^2 tient compte des corrélations entre les variables. Il a encore sur le C. R. L. l'avantage d'avoir une distribution connue, ce qui en fait un test exact.

B. — Mesure de divergence.

MAHALANOBIS, critiquant le C. R. L., a d'abord proposé une quantité qui n'en différait que par un facteur (16). Ensuite, il y a introduit les coefficients de corrélation. Perfectionnée par BOSE et ROY, voici la mesure de divergence enfin adoptée, répondant aux exigences formulées plus haut. On l'appelle distance généralisée de MAHALANOBIS.

$\Delta^2 = \frac{1}{p}\ \sigma^{ij}\ \delta_i\ \delta_j$ où p est le nombre de variables ; σ_{ij} est l'élément ij de la matrice des covariances supposée identique dans les deux populations ; σ^{ij} est le mineur normé de σ_{ij} ; δ_i est la différence entre les moyennes du i^{ieme} caractère. Comme les paramètres tirés des populations ne sont jamais connus, on calcule une quantité analogue à Δ^2, mais basée sur les données des échantillons.

$$D^2 = \frac{1}{p}\ s^{ij}\ d_i\ d_j.$$

BOSE et ROY (3) ont établi la distribution d'échantillonnage de D^2 ; on peut donc estimer Δ^2 à partir de D^2.

En comparant T^2 et D^2, on voit que $D^2 = \frac{T^2\ (n + n')}{n\ n'\ p}$, de sorte que, contrairement à T^2, la distance généralisée ne dépend pas des effectifs des échantillons.

Examinons de près la distance généralisée de deux populations pour une variable ($p=1$), puis pour deux variables ($p=2$).

$p=1$ $\Delta^2 = \dfrac{\delta^2}{\sigma^2}$ où δ est la différence entre les moyennes et σ la dispersion commune aux deux populations.

$p=2$ $\Delta^2 = \dfrac{1}{2} \cdot \dfrac{1}{(1-\rho_2)} \left(\dfrac{\delta_1^2}{\sigma_1^2} - \dfrac{2\rho\,\delta_1\delta_2}{\sigma_1\sigma^2} + \dfrac{\delta_2^2}{\sigma_2^2} \right)$ où δ_1 et δ_2 sont les différences entre les moyennes, respectivement pour les caractères 1 et 2, σ_1 et σ_2 étant les dispersions et ρ la corrélation, communes aux deux populations.

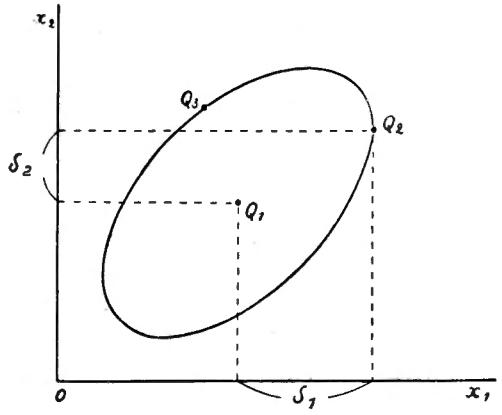

Fig. 1.

Plaçons ces données sur un graphique à deux dimensions (fig. 1). Soit Q_1 l'image des moyennes de la première population, Q_2 celle de la seconde. Parmi les ellipses d'égale probabilité entourant le point Q_1 et contenant un pourcentage donné de la première population, choisissons celle qui passe par Q_2. On montre facilement que tous les points de cette ellipse donnent une même distance généralisée avec la première population : $\Delta^2 = \dfrac{\lambda_2}{2}$, les distances entre les tangentes à l'ellipse parallèles à chacun des axes des coordonnées étant respectivement $2\lambda\sigma_1$ et $2\lambda\sigma^2$. On voit ainsi que la distance généralisée n'est pas une distance géométrique. Par exemple, un point Q_3, géométriquement plus proche de Q_1 que Q_2, est le centre d'une population qui donne avec la population 1 la même distance généralisée que la population 2.

Δ^2 n'a guère été utilisé à ma connaissance. Sous sa première forme, MAHALANOBIS l'a comparé au C. R. L. dans la description de diverses populations et castes de l'Inde (16).

C. — Fonction discriminatoire, pour la comparaison de deux échantillons.

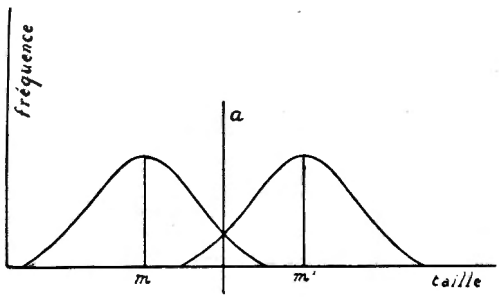

Fig. 2.

Supposons données deux distributions de la taille pour des populations I et II, qui se recouvrent un peu mais sont cependant bien distinctes (fig. 2). Supposons de plus que les dispersions et les effectifs sont égaux. Seules les moyennes diffèrent. Si l'on veut classer, dans une des deux populations, un nouveau sujet dont on ignore l'origine, que fera-t-on ? On le placera dans la population I s'il est à gauche de la droite *a*, et dans la population II s'il est à droite de *a*. En agissant ainsi, on se trompe fatalement pour un certain nombre de sujets, mais on est sûr que le nombre d'erreurs est minimum. On a effectué une bonne discrimination.

Cas de deux variables. Les deux variables ont des variances et des coefficients de corrélation égaux dans les deux populations. Ces restrictions facilitent le raisonnement, mais elles sont abandonnées dans la suite. Représentons sur un diagramme à deux dimensions les points Q_1 et Q_2, images des moyennes des populations et deux ellipses d'égale probabilité, par exemple celles qui contiennent 95 % des sujets de chaque population. Ces ellipses sont égales (fig. 3).

Un nouveau sujet, d'origine inconnue, doit être placé dans l'une ou l'autre population. S'il est à gauche de *a*, on le place dans la première population ; s'il est à droite de *a*, on le place dans la deuxième population. Comme dans le cas d'une seule variable, on a fait une bonne discrimination, parce que le nombre de sujets mal placés est minimum.

J'ai montré (*) que la droite a est l'image géométrique de la fonction discriminatoire de FISHER (6) et (7).

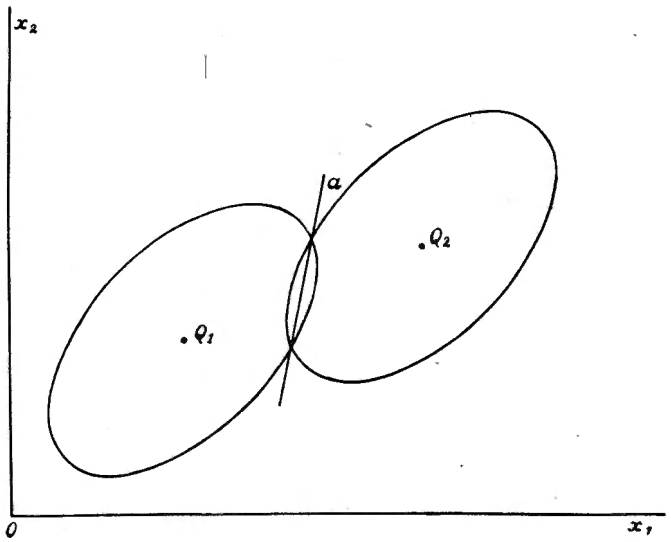

Fig. 3.

FISHER présente cette fonction d'une toute autre façon. Si $x_1, x_2, \ldots x_p$ sont les variables, il cherche des coefficients $b_1, b_2, \ldots b_p$ tels que la différence entre les moyennes de $X = b_1 x_1 + b_2 x_2 + \ldots b_p x_p$ pour les deux populations, divisée par la dispersion de X, soit maxima. WELCH (23) montre que cette fonction de FISHER est la meilleure de toutes les fonctions de discrimination, linéaires ou non (dans le cas de variables normales).

Reprenons un exemple de deux variables, supposons qu'il s'agisse de la longueur x_1 et de la largeur x_2 de la tête. Soit $X = b_1 x_1 + b_2 x_2$ la fonction de FISHER et $I = \frac{x_2}{x_1}$ l'indice céphalique, que l'on peut considérer comme une fonction discriminatoire non linéaire. La discrimination peut se faire à l'aide de X ou de I ; mais celle de X est meilleure, parce que le nombre d'erreurs est moindre

Dans ce problème précis de craniométrie, la fonction discriminatoire est donc plus efficace que l'indice céphalique. Cepen-

(*) Cette démonstration doit paraître dans ce Bulletin.

dant celui-ci a l'avantage sur X de permettre la description d'une population donnée, tandis que X est nécessairement associé à deux populations en présence : I, au contraire de X, a un sens biologique.

5. — Conclusion. En résumé, il semble que la biométrie met actuellement à la disposition de l'anthropologie des méthodes précises. Des progrès devraient venir de l'anthropologie même, où des enquêtes familiales et l'examen de jumeaux univitellins feraient avancer les questions de l'hérédité des caractères continus. Sans doute, dans la mesure où les problèmes biologiques sont posés avec une précision insuffisante, les chercheurs ont raison de n'utiliser que des méthodes simples telles que les méthodes graphiques, qui ont l'avantage d'être plus rapides: des méthodes plus fines qui exigent de longs calculs n'y introduiraient qu'une rigueur illusoire. Mais, dans les problèmes anthropologiques qu'on parvient à poser avec précision, les méthodes exactes dont nous avons parlé, s'appliquent avec toute leur efficacité.

Qu'il me soit permis en terminant de remercier le docteur Twiesselmann qui m'a aidé de ses précieux conseils et de sa grande expérience en anthropologie.

Institut royal des Sciences naturelles de Belgique.

Index bibliographique.

(1) Altdorfer, N., 1950, Congr. Nat. des Sc., sect. biol., Brux.
(—) Ashley Montagu, M. F., 1945, *An introduction to physical anthropology.* (Springfield Illinois, U. S. A.)
(2) Bernstein, F., 1929, *Variations- und Erblichkeitslehre.* (Berlin, p. 67.)
(3) Bose, R. C. & Roy, S. N., 1938, Sankhya, T. 4, p. 19, Calcutta.
(—) Cuénot, L., 1936, *L'espèce.* (Doin, Paris.)
(4) Fisher, R. A., 1928, Trans. Ent. Soc. London, 76, p. 367.
(5) — , 1936, Journ. of the R. Anthr. Inst., 66, p. 57.
(6) — , 1936, Ann. Eug. London, 7, p. 179.
(7) — , 1937, Ann. Eug. London, 8, p. 376.
(8) — , 1948, *Statistical Methods for Research Workers.* (London.)
(9) Fisher, R. A. & Yates, F., 1938, *Statistical Tables.* (Olivers and Boyd, Edinburgh.)
(10) Galton, F., 1885, Journ. of the Anthr. Inst. of Great Britain and Ireland, 15, p. 246.
(—) Hankins, F. M., *La race dans la civilisation.* (Payot, Paris.)
(11) Hotelling, H., 1931, Ann. Math. Statist., 2, p. 360.

(12) HUXLEY, J., 1940, *The new Systematics.* (Oxford Univ. Press, p. 2.)
(13) KENDALL, M. G., 1945, *The advanced Theory of Statistics.* (London.)
(14) LAMOTTE, M., 1948, *Introduction à la biologie quantitative.* (Masson, Paris.)
(15) L'HÉRITIER, Ph., 1949, *Les méthodes statistiques dans l'expérimentation biologique.* (C.N.R.S., Paris.)
(16) MAHALANOBIS, P. C., 1927, Journ. and Proc. Asiat. Soc. of Bengal, 23, p. 301, Calcutta.)
(17) — , 1930, Journ. and Proc. Asiat. Soc. of Bengal, 26, p. 541, Calcutta.)
(18) MATHER, K., 1949, *Biometrical Genetics. The study of continuous variation.* (Methuen. London, p. 10 et p. 38.)
(19) PEARSON, K., 1926, Biometrika, 18, p. 105.
(20) RINGLEB, F., 1937, *Mathematische Methoden der Biologie.* (Leipzig, p. 3.)
(21) SELTZER, C. C., 1937, Amer. Journ. of Phys. Anthr., 23, p. 101.
(—) TOPINARD, P., 1885, *Anthropologie générale.* (Vizot, Paris.)
(—) TWIESSELMANN, F., 1947, Bull. Soc. Anthr. Bruxelles, 68, p. 93.
(22) — , 1949, Mém. de l'Inst. royal des Sc. nat. de Belgique, 2e série, fasc. 35.
(23) WELCH, B. L., 1939, Biometrika, 31, p. 218.

AD. GOEMAERE, Imprimeur du Roi, 21, rue de la Limite, Bruxelles