JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. DELAPORTE

Recherche statistique de facteurs indépendants

Journal de la société statistique de Paris, tome 96 (1955), p. 162-175

<http://www.numdam.org/item?id=JSFS_1955__96__162_0>

III

RECHERCHE STATISTIQUE DE FACTEURS INDÉPENDANTS

UTILISATION A LA RECHERCHE DES FACTEURS DE CROISSANCE, DE PRODUCTION INDUSTRIELLE, D'INTELLIGENCE, ETC.

Parmi les méthodes d'analyse des liaisons statistiques qui existent entre plusieurs caractères, il en est une peu connue malgré ses grandes possibilités de résolution des problèmes : c'est l'Analyse Factorielle.

Je voudrais, ce soir, vous montrer ce qu'est l'Analyse Factorielle, en quoi elle diffère des autres méthodes d'analyse des corrélations et, sur des exemples, vous faire comprendre les très grandes possibilités de cette méthode pour résoudre des problèmes que l'analyse classique ne permet pas d'étudier. Rappelons ce qui l'a fait naître.

A la fin du XIX^{e} siècle, les psychologues étaient en présence de deux conceptions opposées de l'intelligence, toutes deux incompatibles avec le résultat des tests. Selon la première, l'ensemble des facultés psychologiques de l'esprit serait décomposable en un petit nombre de facultés distinctes, indépendantes les unes des autres. C'est ainsi qu'un test de mémoire mesurerait la mémoire de l'individu et rien d'autre, un test d'association mesurerait le pouvoir d'association et rien d'autre. Un autre test d'association est donc une nouvelle mesure de ce même pouvoir d'association qui doit être indépendant de la mémoire. Donc après correction de l'effet d'atténuation dû aux erreurs de mesure, deux tests différents de mémoire auraient entre eux une corrélation parfaite et la corrélation entre deux tests s'adressant à des facultés distinctes serait nulle, donc leur coefficient de corrélation serait nul, aux erreurs d'échantillonnage près.

Selon la deuxième conception, l'ensemble des facultés psychologiques de l'esprit peut être groupé en une entité unique : l'intelligence, que les individus ont à des degrés divers. Les individus d'un groupe se classeront toujours dans le même ordre quels que soient les tests qu'on leur fait subir : un test d'attention ou un test de mémoire ou un test de représentation dans l'espace devront donner le même classement hiérarchique des individus entre eux, donc la corrélation entre deux tests devra être parfaite, aux erreurs d'observation près.

Les résultats des tests psychologiques montraient que la corrélation entre deux tests ne pouvait statistiquement ni être considérée comme nulle, ni montrer une liaison aussi étroite que la liaison fonctionnelle. Il y avait donc désaccord entre ces deux conceptions de l'intelligence et le résultat des tests. C'est alors que le psychologue anglais C. Spearman [1] (1) chercha une concep-

---

(1) Les numéros entre crochets renvoient à la bibliographie.

tion de l'intelligence qui soit en meilleur accord avec les observations. Ayant remarqué que les coefficients de corrélation qui existent entre des tests s'adressant à des facultés distinctes de l'esprit ne peuvent statistiquement être considérés ni comme nuls, ni comme montrant une corrélation parfaite et, d'autre part, qu'il existait un ordre hiérarchique entre les coefficients de corrélation des divers tests, il proposa le schéma suivant :

Pour Spearman, le résultat de chaque test s'adressant à une seule aptitude d'un individu est formé par la somme de deux parties indépendantes appelées facteurs :

1° un facteur général G commun à tous les tests, mais particuliers à l'individu;

2° un facteur spécifique S du test, particulier à l'aptitude testée et à l'individu.

Ce facteur général a été, à l'origine, considéré comme *facteur d'intelligence général de l'individu.*

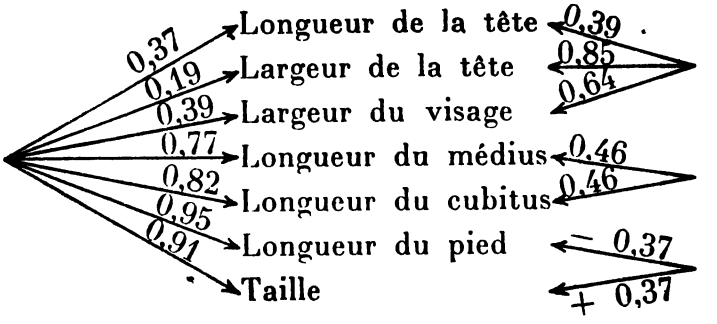
Les premiers résultats d'application furent d'abord très satisfaisants, d'autant plus satisfaisants qu'on travaillait souvent sur de petits nombres d'individus, ainsi les erreurs d'échantillonnage permises par la Statistique mathématique étaient plus grandes et empêchaient de pouvoir rejeter statistiquement la théorie.

Des analyses faites sur de plus grands nombres d'individus et de tests montrèrent cependant que le seul facteur général ne suffisait souvent pas à expliquer la liaison statistique qui existe entre les tests; par exemple, entre un test de mémoire des syllabes et un test de mémoire des nombres on constate que toute la liaison n'est pas entièrement expliquée par le facteur général. Il fallut alors admettre l'existence de facteurs supplémentaires qu'on appela *facteurs de groupe*, communs à deux ou à plusieurs tests et indépendants à la fois du facteur général et des facteurs spécifiques.

L'Analyse Factorielle se présente ainsi comme l'analyse des corrélations qui existent entre plusieurs caractères mesurés sur des individus ou des objets, en supposant que ces liaisons sont explicables par la superposition de plusieurs facteurs inconnus indépendants entre eux. Le principe de l'Analyse Factorielle donné par Spearman pour la Psychologie expérimentale est évidemment applicable à tous les domaines où il est possible de mesurer des individus, des objets ou des faits.

Pour faire mieux comprendre la signification des facteurs de groupe, prenons un exemple d'application de l'Analyse Factorielle à la Biométrie. Mac Donnel avait calculé les coefficients de corrélation qui existent entre 8 mesures de longueur faites sur 3.000 hommes adultes. Essayons de les analyser par une superposition de facteurs statistiquement indépendants [2]. On constate tout d'abord que l'hypothèse que chaque caractère mesuré serait la somme du facteur général de l'individu et d'un facteur spécifique ne convient pas, on doit donc rejeter l'hypothèse de la structure simple de représentation des liaisons par le seul facteur général. L'analyse statistique montre qu'il est nécessaire d'introduire en outre 3 facteurs de groupe indépendants entre eux et indépendants des autres facteurs. La décomposition se fait alors de la manière

suivante en indiquant sur chaque flèche la valeur du coefficient de corrélation qui unit le facteur aux caractères :



Cette décomposition peut s'interpréter de la manière suivante :

Le facteur général qui unit tous les caractères entre eux est un facteur de développement général du squelette, un facteur d'homothétie des individus entre eux : si un individu a un membre très développé, il aura aussi l'ensemble du corps plus développé que la moyenne.

Le deuxième facteur est seulement lié aux trois mesures de la tête, il est donc naturel de le considérer comme *facteur de groupe de la tête.* Comme il est, par hypothèse, indépendant du facteur général, sa signification est la suivante : si, parmi les individus ayant un même facteur général de développement, l'un a une dimension de la tête plus grande que ne le ferait prévoir ce facteur d'homothétie générale, les autres dimensions de la tête seront aussi plus grandes, sans que ceci influe sur les autres parties du corps.

Pour expliquer l'ensemble des intercorrélations qui existent entre ces sept caractères, on est obligé de faire intervenir un facteur de groupe propre à la longueur du médius et à la longueur du cubitus. Il s'agit donc d'un *facteur de groupe du bras* qu'on peut interpréter d'une manière analogue à celui trouvé pour la tête : pour deux individus ayant la même valeur du facteur d'homothétie générale, si l'un d'eux a une dimension du bras plus grande que celle de l'autre individu, les autres dimensions du bras seront aussi plus grandes.

On observe enfin un troisième facteur de groupe propre à la longueur du pied et à la taille, qui semble être un *facteur de groupe des jambes.*

Les valeurs des coefficients de corrélation observées entre chacun des caractères et le facteur général montrent immédiatement que la longueur du pied et la taille sont les dimensions les plus liées au facteur général d'homothétie. La largeur du visage, la longueur de la tête et surtout la largeur de la tête ont de faibles coefficients de corrélation avec le facteur général et sont donc peu liées à l'homothétie générale des individus entre eux. Au contraire, la largeur du visage et surtout la largeur de la tête sont très fortement liées au facteur de groupe de la tête, c'est dire qu'elles expriment surtout le développement particulier à la tête.

Maintenant que nous voyons plus clairement ce que l'on veut obtenir de l'Analyse Factorielle, examinons schématiquement comment se fera l'analyse de Statistique mathématique. Notons tout de suite qu'il s'agit d'une méthode très générale applicable aussi bien aux individus qu'à la recherche industrielle,

à la formation de nombres indices de prix ou de nombres indices de l'activité des affaires.

ANALYSE STATISTIQUE EN FACTEURS INDÉPENDANTS

Nous avons vu que chaque caractère d'un individu est supposé formé par la somme d'un facteur général, d'un facteur spécifique et, s'il y a lieu, d'un ou de plusieurs facteurs de groupe, tous statistiquement indépendants entre eux, chacun de ces facteurs étant affecté d'un coefficient convenable.

Comme *a priori* on ne connaît ni le nombre des facteurs de groupe ni les caractères x auxquels ils s'étendent, on est conduit à considérer que les k caractères d'un individu j peuvent s'exprimer par le système suivant d'équations linéaires en fonction du facteur général, des facteurs de groupe et des facteurs spécifiques :

$$
\begin{aligned}
{}_jx_1 &= \alpha_1\,{}_jG + \beta_1\,{}_jB + \gamma_1\,{}_jC + \ldots + \lambda_1\,{}_jS_1\\
{}_jx_2 &= \alpha_2\,{}_jG + \beta_2\,{}_jB + \gamma_2\,{}_jC + \ldots + \lambda_2\,{}_jS_2\\
{}_jx_3 &= \alpha_3\,{}_jG + \beta_3\,{}_jB + \gamma_3\,{}_jC + \ldots + \lambda_3\,{}_jS_3\\
{}_jx_4 &= \alpha_4\,{}_jG + \beta_4\,{}_jB + \gamma_4\,{}_jC + \ldots + \lambda_4\,{}_jS_4\\
&\ldots\\
{}_jx_i &= \alpha_i\,{}_jG + \beta_i\,{}_jB + \gamma_i\,{}_jC + \ldots + \lambda_i\,{}_jS_i\\
&\ldots\\
{}_jx_k &= \alpha_k\,{}_jG + \beta_k\,{}_jB + \gamma_k\,{}_jC + \ldots + \lambda_k\,{}_jS_k
\end{aligned}
$$

${}_jG$ étant le facteur général de l'individu j, ${}_jB$ ${}_jC$ ... étant les facteurs de groupe de l'individu j.

${}_jS_1$ ${}_jS_2$ ... les facteurs spécifiques de l'individu j pour chacun des caractères x_1, x_2, etc.

$$\alpha_1, \alpha_2, \ldots \quad \beta_1, \beta_2, \ldots \quad \gamma_1, \gamma_2, \ldots \quad \ldots \quad \lambda_1, \lambda_2, \ldots$$

désignent les coefficients de chacun des facteurs.

Tous les coefficients et facteurs formant les seconds membres du système d'équations sont inconnus, leur nombre dépassant de beaucoup le nombre des données du problème, il y a indétermination.

Pour lever cette indétermination, il est nécessaire de faire quelques hypothèses complémentaires, Nous supposerons tout d'abord que le premier facteur, le facteur général G, est commun à l'ensemble des k caractères et qu'il explique le plus possible des intercorrélations entre ceux-ci. Il faut, en outre, que la distribution des k caractères soit une loi de Laplace-Gauss à k variables liées.

L'analyse statistique se fait en estimant successivement les coefficients α et le facteur général, puis les coefficients β non nuls et le premier facteur de groupe B, puis les coefficients γ non nuls et le deuxième facteur de groupe, etc. jusqu'à explication de toutes les intercorrélations entre les caractères étudiés.

RECHERCHE DES COEFFICIENTS α ET DU FACTEUR GÉNÉRAL [3]

Une condition est nécessaire et suffisante pour que les caractères 1, 2, ... f, g, h, i, ... k soient liés entre eux par le seul facteur général est que leurs coefficients de corrélation théorique ρ vérifient les $k - 1$ suites d'égalités

$$\frac{\rho_{e1}}{\rho_{f1}} = \frac{\rho_{e2}}{\rho_{f2}} = \ldots = \frac{\rho_{eh}}{\rho_{fh}} = \frac{\rho_{ei}}{\rho_{fi}} = \ldots = \frac{\rho_{ek}}{\rho_{fk}} = \frac{\alpha_e}{\alpha_f} \qquad (h, i \neq e, f)$$

construites sur la base d'un caractère f en prenant successivement $e = 1, 2, \ldots e, g, \ldots k$.

On ne connaît pas les coefficients de corrélation théoriques ρ mais seulement leurs valeurs empiriques r calculées sur les mesures faites sur un échantillon de n individus (1). Des abaques donnent en fonction de r_{eh} et r_{fh} les limites possibles du quotient α_e / α_f avec une probabilité P donnée à l'avance. Une représentation graphique de ces intervalles permet de voir s'il existe une partie commune à tous ces intervalles ou quels sont les coefficients de corrélation qu'il faut éliminer pour que les autres forment des quotients ayant tous une partie commune et soient donc représentables par le facteur général sans facteur de groupe.

On obtient alors les estimations a de α à écart type minimum par la formule

$$a_h = r_{Gh} = \frac{\pm \sum_{ef} \left[\left(\frac{r_{he}\, r_{hf}}{r_{ef}} \right)^{\frac{1}{2}} \cdot W_{h\ ef} \right]}{\sum_{ef} W_{h.\ ef}} \qquad (e \neq f \neq h)$$

où

$$W_{h\ ef} = [(\varphi_{he}^2 + \varphi_{hf}^2 + \varphi_{ij}^2 + \varphi_{he}^2\, \varphi_{ij}^2)\,(1 + 3\,\varphi_{ij}^2) + 5\,\varphi_{ef}^4]^{-1}$$

les $\varphi_{he}, \ldots$ étant les coefficients de variation des coefficients de corrélation $r_{he}, \ldots$ Les sommes $\sum_{ef}$ sont étendues à toutes les combinaisons de k variables prises 2 à 2.

La loi de distribution de a_h autour de α_h tend vers une loi de Laplace-Gauss d'écart type

$$\sigma_{r_{Gh}} = \frac{1}{2}\, r_{Gh} \left(\sum_{ef} W_{h\ ef} \right)^{-\frac{1}{2}}$$

L'estimation du facteur général d'un individu s'obtient par une équation de régression multiple en fonction des caractères les plus étroitement liés au facteur général, c'est-à-dire ayant les plus fortes valeurs absolues de r_{Gh}.

RECHERCHE DES COEFFICIENTS β ET DU PREMIER FACTEUR DE GROUPE

Si les caractères ef sont liés par le seul facteur général, les résidus empiriques $r'_{ef} = r_{ef} - a_e\, a_f$ doivent êtres statistiquement compatibles avec l'hypothèse

(1) Si les erreurs de mesures sont indépendantes de la valeur du caractère mesuré et entre elles, on peut montrer que le critère d'égalité des quotients des coefficients de correlation reste valable.

de nullité du résidu théorique $\rho'_{ef} = 0$. Sinon, les résidus r'_{ef} sont les coefficients de corrélation entre des caractères liés par les facteurs de groupe. Le facteur général de ces résidus est le premier facteur de groupe; on obtiendra donc les caractères auxquels s'étend le premier facteur de groupe B, les estimations b_h des coefficients $\beta_h = \rho_{Bh}$ et du facteur de groupe B d'un individu en appliquant aux premiers résidus r'_{ef} la méthode précédente de recherche du facteur général.

Les seconds résidus $r''_{ef} = r_{ef} - a_e\, a_f - b_e\, b_f$ permettront par une nouvelle application de la méthode de dégager le deuxième facteur de groupe. Cette méthode sera appliquée jusqu'à explication de toutes les intercorrélations entre les caractères étudiés.

LES APPLICATIONS DE LA MÉTHODE

Reprenons notre premier exemple d'application de l'Analyse Factorielle aux 8 mesures faites sur des hommes. L'utilisation de la méthode d'analyse statistique précédente conduit à former le diagramme des intervalles communs aux quotients de coefficients de corrélation. Ceux-ci montrent que l'ensemble des coefficients de corrélation est explicable par la présence d'un facteur général et de facteurs spécifiques, à l'exception des coefficients de corrélation entre longueur et largeur de la tête, longueur de la tête et largeur du visage, largeur de la tête et largeur du visage, longueur du médius et longueur du cubitus, longueur du pied et taille. L'ensemble des coefficients de corrélation non éliminés sert à estimer les paramètres mais sans donner immédiatement d'indication sur la décomposition en facteurs de groupe. Ces estimations permettent de reconstituer d'une manière statistiquement satisfaisante les coefficients de corrélation entre chacun des caractères, à l'exception des coefficients de corrélation qui avaient dû être éliminés pour la détermination du facteur général.

Ces résidus non nuls sont alors considérés comme des coefficients de corrélation empiriques existant entre les caractères après élimination du facteur général. On forme alors un nouveau diagramme des intervalles communs qui montre tout d'abord que des intervalles communs existent à la condition de rejeter les coefficients de corrélation résiduels entre longueur du médius et longueur du cubitus d'une part, longueur du pied et taille d'autre part. Le facteur général qui subsiste est alors celui qui a été désigné comme premier facteur de groupe. L'examen des 2es résidus permet de mettre en évidence la position du 2^{e} et du 3^{e} facteurs de groupe.

Nous ne reviendrons pas sur l'interprétation à donner à chacun de ces facteurs, remarquons cependant qu'aucun de ceux-ci ne s'identifie avec l'un des caractères mesurés. L'application de cette méthode d'analyse statistique a montré tout d'abord que l'ensemble des caractères mesurés pouvait être représenté par la superposition de facteurs statistiquement indépendants, ensuite de dégager chacun de ces facteurs inconnus. Les caractères mesurés ont nécessairement été choisis d'une manière plus arbitraire, habituellement le choix s'est porté sur les caractères qu'il était le plus facile de mesurer. Cependant, les facteurs qui sont obtenus ne dépendent que très peu de ce choix

arbitraire, c'est ainsi que le facteur général d'homothétie entre les individus aurait été aussi bien obtenu si, au lieu des mesures retenues, on avait par exemple choisi : le tour de la tête, la distance du menton aux yeux, la largeur des épaules, etc., ou encore si, par examen aux rayons X par exemple, on avait pu mesurer les longueurs des os formant le squelette. D'une manière analogue, le facteur propre au développement de la tête aurait encore été obtenu au moyen d'un autre ensemble de mesures de la tête. C'est dire que les facteurs obtenus par Analyse Factorielle présentent par rapport à toutes les autres méthodes d'analyse, l'avantage d'être presque *indépendants du choix arbitraire des caractères à mesurer.*

Parfois, les hypothèses que nous avons dû faire ne sont pas statistiquement vérifiées par les caractères : si la loi de distribution de l'ensemble de ceux-ci n'est pas laplace-gaussienne par exemple, on pourra souvent faire subir à chacun une transformation fonctionnelle telle qu'elle ramène à une loi de Laplace-Gauss. Parfois, encore, l'effet des facteurs n'est pas additif; une transformation fonctionnelle permettra encore de les ramener à la forme additive, hypothèse initiale.

Par exemple, lors d'une étude de *Maia squinado*, l'araignée de mer, adulte. G. Teissier [4] a constaté que la décomposition pouvait se faire en 3 facteurs indépendants, le 1er d'homothétie générale, le 2e de développement de l'ensemble des appendices, le 3e propre aux diverses parties de la pince et des 2 appendices les plus voisins de celle-ci. Cependant, ces facteurs ne sont pas additifs, mais le 1er facteur doit être multiplié par une fonction linéaire des facteurs de groupe. L'analyse s'obtient alors en faisant subir une transformation logarithmique telle qu'elle ramène la décomposition à une forme linéaire.

Autre exemple. — *Exemple d'application aux résultats d'épreuves sportives.* Une étude des performances sportives de 359 élèves du Lycée Saint-Louis, candidats aux Écoles de Saint-Cyr et de Navale, a été faite en y joignant la taille, le poids corporel de l'individu et son âge. L'Analyse Factorielle a donné les résultats suivants [5 et 6].

	Coefficient de corrélation avec	
	Facteur général	Facteur de groupe
Saut en hauteur	0,66	0,26
Saut en longueur	0,81	—
Course de 60 mètres	0,79	—
Course de 600 mètres	0,62	—
Lancement du poids	0,64	0,35
Grimper	0,55	— 0,18
Développer	0,55	0,30
Taille	0,04	0,59
Poids corporel	0,09	1,00
Age	0,25	—

On observe ainsi que le facteur général est essentiellement un facteur de réussite dans l'ensemble des épreuves sportives, celles de saut en longueur et de course 60 mètres étant celles qui donnent le plus d'informations sur ce facteur. L'âge n'est que peu lié à cette réussite sportive, la taille et le poids corporel en sont pratiquement indépendants pour le groupe d'individus étudié. Le facteur de groupe semble s'identifier avec le poids du corps, fortement lié

à la taille, qui agit favorablement pour les épreuves de lancement du poids, développer, saut en hauteur, mais est, au contraire, défavorable pour grimper, ce qui s'explique facilement. Il n'a pas été possible de pousser au delà l'analyse en facteurs de groupe à cause du nombre trop faible des individus étudiés.

Application à la sélection professionnelle des Secrétaires.

Quelles qualités doit présenter une bonne secrétaire? Doit-elle être surtout bonne en sténographie, en dactylographie, ponctuelle, discrète, avoir une bonne mémoire, un physique agréable ou de bonnes connaissances musicales? Chacune de ces qualités peut faire l'objet d'une note, mais comment apprécier globalement les qualités d'une secrétaire d'après 10 ou 15 caractères notés : une moyenne arithmétique de ces notes ne convient pas, car on ne peut mélanger l'absentéisme avec les qualités professionnelles par exemple. On peut prendre une moyenne pondérée, mais il faudrait pouvoir déterminer les poids respectifs à donner de telle sorte que cette note globale corresponde véritablement aux qualités d'une secrétaire. J. Bongard [7] a demandé à un grand nombre d'utilisateurs de secrétaires, parmi une liste de qualités, quelles étaient celles, deux par deux, qui leur semblaient les plus utiles. Il a alors procédé à l'Analyse Factorielle de ces résultats et a ainsi déterminé le facteur général de ces réponses qui est le facteur général d'aptitude à être secrétaire. L'équation qui donne le facteur général d'un individu présente les poids qu'il y a lieu d'affecter à chacun des caractères pour mesurer l'aptitude d'une personne à être secrétaire.

Application à la construction d'un indice d'activité industrielle.

E. C. Rhodes [8], en Angleterre, puis F. Brambilla et M. Savini [9], en Italie, se sont penchés sur le problème de la construction d'un indice de l'activité industrielle. Il est relativement facile d'obtenir un assez grand nombre de séries de mesures telles que l'emploi, la consommation d'électricité, l'activité du bâtiment ou bien un nombre indice des prix de gros, un nombre indice du coût de la vie, un nombre indice de l'épargne, un nombre indice du commerce extérieur, etc., qui toutes doivent être liées à l'activité industrielle. Il est beaucoup plus délicat de savoir quel poids affecter à chacun d'eux pour former le nombre indice de l'activité industrielle. En prenant de nombreuses séries de nombres indices, on peut rechercher par Analyse Factorielle s'il est possible d'en faire une décomposition en facteurs statistiquement indépendants. E. C. Rhodes a ainsi analysé les 14 séries composant l'indice de l'activité industrielle de l' « Economist » pendant plusieurs années. Il a ainsi montré qu'il existait un facteur général et des facteurs de groupe entre ces séries. La construction de l'indice de l'activité industrielle peut alors se faire en utilisant comme poids ceux qui donnent la meilleure estimation du facteur général; les poids ainsi calculés sont nettement différents de ceux utilisés par l' « Economist ». Les facteurs de groupe sont, par définition, statistiquement indépendants du facteur général. Doit-on les faire entrer dans l'indice ou doit-on les considérer comme exprimant un phénomène nettement distinct de l'activité industrielle que l'on se proposait d'étudier? Rhodes ajoute que si l'on prenait des observations pendant une autre période, on trouverait, semble-t-il, que le centre de gravité des affaires s'est déplacé et que, par conséquent, le système de poids à utiliser pour l'indice s'est modifié. Il semble que la connaissance des

phénomènes économiques progresserait sensiblement si l'on recherchait les facteurs indépendants qui les constituent.

Étudions maintenant une fabrication chimique. Outre son hétérogénéité propre, chacune des matières premières a une composition variable dans le temps qui provient des différences de provenance ou des irrégularités de filons exploités. Les conditions dans lesquelles s'opèrent les réactions chimiques varient légèrement, ainsi que les éléments extérieurs : température, humidité de l'air, etc. Tout ceci entraîne l'irrégularité des caractéristiques de qualité du produit fini. Tous les caractères retenus dans une telle étude sont nécessairement choisis parmi les plus faciles à mesurer, donc arbitrairement. Par l'Analyse Factorielle on pourra rechercher si l'ensemble de ces caractères est décomposable en quelques facteurs statistiquement indépendants sur lesquels on essaiera d'agir en cours de fabrication pour stabiliser les caractéristiques de qualité du produit fini et pour les améliorer.

Dans ce bref exposé, j'ai essayé de vous montrer, par quelques exemples d'application, ce qu'est l'Analyse Factorielle, les conditions générales à imposer pour éviter l'indétermination de la solution et la nature des résultats qu'on peut en espérer.

Quoique assez délicat à manier, ce procédé d'analyse statistique peut s'appliquer à de nombreux caractères différents, 10 à 20 ou 30, et les résultats sont plus précis si le nombre de ces caractères est grand, ainsi que le nombre des objets ou individus étudiés. Les mesures doivent être faites avec précision, mais nous avons vu que le critère reste valable dans le cas de mesures entachées d'erreurs.

J'espère surtout vous avoir fait saisir la fécondité de cette méthode d'Analyse Factorielle qui permet de mettre en lumière les facteurs de structure du phénomène étudié en se dégageant du choix nécessairement arbitraire des caractères qu'on a pu mesurer.

P. Delaporte.

DISCUSSION

M. René Casse demande si, d'une manière analogue à l'Analyse Factorielle des résultats d'épreuves sportives, il avait été fait un travail d'Analyse Factorielle portant sur les épreuves intellectuelles de divers examens, en particulier du baccalauréat.

M. Flaus fait remarquer que l'Analyse Factorielle des épreuves d'examen d'entrée à de grandes écoles telles que Polytechnique et Centrale permettrait peut-être des résultats intéressants grâce à l'homogénéité plus grande de ces épreuves par rapport à celle du baccalauréat.

M. Delaporte répond que des études poussées des résultats d'examen du baccalauréat ont été faites en particulier par H. Laugier et D. Weinberg : Recherches sur la solidarité et l'indépendance des aptitudes intellectuelles d'après les notes des examens écrits de baccalauréat (Paris, Imprimerie Chantenay, 1938), mais cette étude a révélé le peu de précision de la notation. D'autre part, le petit nombre des épreuves rend peu précise une Analyse Factorielle.

M. Casse voudrait entrevoir ainsi une solution au problème qui se pose fréquemment à lui en tant que professeur : il est toujours choqué du système qui consiste à prendre la moyenne arithmétique donnée à des élèves des lycées pour déterminer leur passage d'une classe à une autre; les conseils de professeurs ayant à en décider sont inénarrables.

M. Delaporte indique que des études ont été faites sur la pondération apportée involontairement par les notations des professeurs (voir *Biotypologie*, t. VI, n° 4, décembre 1938, p. 271-273).

M. F. Bastenaire. — Je pense qu'il est très important de saisir la différence entre l'analyse factorielle selon la conception de Spearman que M. Delaporte nous a très clairement exposée, et selon la conception de Hotelling.

Si l'on considère l'ensemble des k caractères mesurés sur un individu comme un vecteur dans un espace à k dimensions, et si l'on suppose ce vecteur distribué dans cet espace selon une loi de Laplace-Gauss, l'objet de Hotelling se réduit à représenter cette distribution statistique par rapport à un autre système d'axes orthogonaux. Les nouvelles coordonnées du vecteur peuvent alors se trouver réduites en nombre, si le rang de la distribution est inférieur à k (ou s'il peut être considéré comme tel) mais les k' nouvelles coordonnées sont toutes en général des fonctions linéaires de toutes les variables initiales au nombre de k. Il n'en résulte guère de simplification à part que les nouvelles variables sont indépendantes.

L'objet de la conception de Spearman, paraît être au contraire, par l'introduction du facteur général et des facteurs de groupes, de représenter le vecteur observé sous la forme d'une somme de vecteurs devant appartenir à certains sous-espaces déterminés. L'effet d'un facteur de groupe par exemple, peut être représenté par un vecteur de direction fixe mais orthogonale à un certain nombre des axes du système et par suite parallèle à un « hyperplan » de coordonnées.

La représentation de Spearman possède donc un sens physique que n'a pas celle de Hotelling, mais elle constitue une hypothèse *a priori* (de même par exemple qu'en régression, l'hypothèse de linéarité par rapport aux paramètres), tandis qu'à cet égard, la représentation de Hotelling est plus générale.

Je voudrais demander à M. Delaporte si l'idée que je me fais de la différence entre les deux méthodes est bien exacte.

M. Delaporte répond à M. Bastenaire que la schématisation sommaire qu'il vient de donner pour les méthodes d'Analyse Factorielle est correcte.

Les méthodes qui découlent des hypothèses de Spearman introduisent des conditions propres au facteur général et aux facteurs de groupe. Celles-ci se traduisent dans l'hyperespace par les conditions que doivent satisfaire les nouveaux axes de coordonnées représentant chacun des facteurs.

Ces conditions de choix des axes entraînent des conditions de possibilité de représentation par un schéma de type spearmanien. Ces conditions sont celles que nous avons vues précédemment sous la forme d'égalité des rapports des coefficients de corrélation.

M. le Président R. Hénon demande comment s'obtient pratiquement cette vérification d'égalité des rapports de coefficients de corrélation.

En réponse à M. le Président Hénon, M. Delaporte indique que des abaques ont été construits pour donner, en fonction du coefficient de corrélation empirique du numérateur et de celui du dénominateur, entre quelles limites doit se trouver la valeur des rapports de coefficients de corrélation théoriques avec une quasi-certitude P donnée.

On forme alors pour chaque suite d'égalité des rapports de coefficients de corrélation un *graphique d'intervalles communs*. Chacun de ces graphiques s'obtient en portant en ordonnée les valeurs limites des rapports de coefficients de corrélation théoriques à lire sur l'abaque et en abscisse successivement chacun des caractères qui figurent à la fois au numérateur et au dénominateur du rapport de coefficients de corrélation.

Pour chacun de ces rapports on obtient ainsi un intervalle et pour l'ensemble des intervalles correspondant à une suite d'égalités on regarde s'ils ont une partie commune. S'ils ont une partie commune, le schéma de décomposition de Spearman en un facteur général et K facteurs spécifiques est acceptable. S'ils n'ont pas de partie commune, on trouve les caractères qu'il faut éliminer pour que les autres soient représentables selon le schéma de Spearman. Les caractères éliminés sont ceux qui nécessiteront l'introduction de facteurs de groupe.

M. R. Risser. — M. Delaporte, dans sa très intéressante conférence, nous a montré que parmi les méthodes d'analyse des liaisons stochastiques entre plusieurs caractères, il en est une dénommée Analyse factorielle qui, malgré les multiples possibilités de son application, est assez peu connue.

Il nous a rappelé tout d'abord que c'est Spearman qui a donné en 1904 une théorie relative à la structure interne des aptitudes, rendant ainsi compte des corrélations existant entre elles, et a de plus remarqué que si cette structure est réalisée, elle nécessite l'existence de relations particulières entre les mesures de ces aptitudes.

Dans cette théorie des deux facteurs, qui a fait l'objet durant un demi-siècle de multiples recherches, on admet que toute aptitude A résulte de la mise en œuvre simultanée de deux fonctions indépendantes, l'une commune à toutes les aptitudes appelée facteur général G, l'autre spéciale à l'aptitude considérée et désignée facteur spécifique S_A.

La grandeur de ces facteurs variant avec les individus, et chaque aptitude A requérant l'intervention en proportions définies de chacun des facteurs G et S_A, il s'ensuit que le résultat a de la mesure de l'aptitude A à l'aide d'un test pour un sujet i, est exprimé par :

$$a^{(i)} = m_a\, g^{(i)} + n_a\, S_a^{(i)};$$

Spearman suppose que les facteurs spécifiques S_a, S_b, S_c..., sont indépendants entre eux.

En sommant les produits $a^{(i)}\, b^{(i)}$ relatifs à N sujets, dans le cas de deux aptitudes A et B, on voit apparaître le coefficient de corrélation r_{ab}. Si maintenant, on considère avec Spearman quatre aptitudes A, B, C, D, auxquelles correspondent les m_i et S_i, on constate que :

$$r_{ab}\, r_{cd} = r_{acbd} = r_{adbc} = m_a\, m_b\, m_c\, m_d$$

et les différences tétrades $\Delta = r_{ab} \cdot r_{cd} - r_{ac} \cdot r_{bd}$.

M. G. Darmois a étudié la théorie des deux facteurs au point de vue des lois de probabilité, afin de déceler l'origine et la portée des relations qui sont à la base de la théorie de Spearman.

Considérant n aptitudes (1) $x_i = m_i g + \lambda_i s_i$, il montre que la nullité des tétrades se déduit des conditions nécessaires résultant de l'identité

$$E\left[e^{i(u_1 x_1 + \ldots + u_n x_n)}\right] = \Phi(m_1 u_1 + \ldots + m_n u_n) \cdot \varphi_1(u_1) \ldots \varphi_n(u_n)$$

pour les moments du deuxième ordre

$$(2)\ E(u_1 x_1 + \ldots + u_n x_n) = (m_1 u_1 + \ldots + m_n u_n)^2 + \lambda_1 u_1^2 + \ldots + \lambda_n u_n^2$$

en supposant les mesures standardisées, et par suite

$$E(x_i x_j) = r_{x_i} r_{x_j} = m_i m_j$$

La condition (2) n'est pas suffisante pour affirmer la structure type de Spearman.

Le problème le plus important paraît être la conservation de g au sens complet; il n'a de solutions que si les S_i suivent toutes des lois de Gauss.

La théorie des deux facteurs de Spearman est une des plus simples, car elle peut suffire dans les limites des erreurs usuelles d'échantillonnage; comme en réalité, les combinaisons de facteurs sont plus complexes, Kelley a étudié des schémas statistiques plus complexes tels que :

- - Cas de trois variables représentées par des combinaisons de deux facteurs généraux sans facteurs spécifiques, lorsque les coefficients de corrélation multiple

$$r_{a(bc)} = r_{b(ca)} = r_{c(ab)} = 1.$$

— Cas de deux deux facteurs spécifiques dépendants sur un groupe de quatre variables (a, b, c, d).

— Cas de deux facteurs spécifiques dépendants sur un groupe de cinq aptitudes.

Analysant les schémas de Kelley, Holzinger a fait observer que, des vérifications expérimentales conduisant à des tétrades mesurables et prouvant simplement que certains facteurs spécifiques n'étaient pas indépendants, il y avait lieu de faire intervenir de nouveaux facteurs.

On se rend ainsi compte de l'introduction par M. Delaporte de facteurs de groupes, et que les K caractères d'un individu j peuvent s'exprimer par un système d'équations linéaires en fonction du facteur général G, de facteurs de groupes $_jB$, $_jC$... et de facteurs spécifiques $_jS_i$, soit :

$$x_i = \alpha_i\, {}_jG + (\beta_i\, {}_jB + \gamma_i\, {}_jC + \ldots) + {}_j\delta_i\, {}_jS_i,$$

où ($_jG$) est le facteur général de l'individu j, ($_jB$, $_jC$...,) sont les facteurs de groupe de l'individu j, et ($_jS_1$, $_jS_2$...) les facteurs spécifiques de l'individu j pour les caractères (x_1, x_2,...), envisagés par certains statisticiens et psychologues.

En raison du nombre de coefficients et facteurs qui sont inconnus, apparaissant dans les seconds membres du système d'équations, le conférencier utilise un certain nombre d'hypothèses complémentaires, dont la plus importante d'ailleurs concerne le facteur général G, considéré comme commun aux

K caractères; une autre hypothèse est afférente à la distribution des K carac tères qui doit suivre une loi de Laplace-Gauss à K variables liées.

Pour que les K caractères soient liés par le seul facteur général G, M. Delaporte introduit les égalités :

$$\frac{\rho_{eh}}{\rho_{fh}} = \frac{\rho_{ei}}{\rho_{fi}} = \ldots = \frac{\alpha_e}{\alpha_f} \text{ (avec } h, i \neq f, e),$$

les ρ étant des coefficients de corrélation théorique, auxquels il substitue les valeurs empiriques r calculées sur les mesures effectuées sur l'ensemble de n individus.

L'utilisation ingénieuse des limites possibles des quotients $\frac{\alpha_e}{\alpha_f}$ en fonction des r_{eh} et r_{fh}, permet de déterminer une partie commune aux intervalles correspondants, et dans le cas contraire de procéder à l'élimination de certains coefficients de corrélation, en vue d'une détermination de nouveaux coefficients ayant alors une partie commune et susceptibles de faire intervenir une représentation de facteur de groupe.

M. Delaporte nous a montré comment l'on utilise les résidus $r'_{ef} = r_{ef} - a_e\, a_f$, en vue de la détermination des coefficients B, et des résidus $r''_{ef} = r_{ef} - a_e\, a_f - b_e b_f$, en vue de celle des coefficients C...

Sa méthode diffère donc notablement des méthodes basées soit sur le minimum de Σa_i^2, soit sur le minimum de Σa_i, qui font appel à la méthode des multiplicateurs deLagrange, et celà toutes les fois que l'on se propose d'analyser les expériences du type

$$x_i = a_i\, G + b_i\, B_i + \gamma_i\, C \text{ (avec } i = 1, 2, 3), \text{ ou encore } x_i = a_i\, G + \Sigma\, \lambda_j\, B_j. \quad \text{avec } i > 3.$$

Le type d'analyse factorielle étudié par M. Delaporte est évidemment plus général que celui où l'on fait intervenir pour chacun des caractères un facteur général G commun à tous les caractères, un facteur de groupe B pour m d'entre eux, et un facteur de groupe C pour les n autres, de telle façon qu'en définitive il n'apparaisse que K facteurs de groupe. En définitive, la contribution théorique de notre collègue, complétée par l'énoncé d'exemples judicieux, nous montre que son processus d'analyse factorielle permet de déceler les éléments de structure d'un phénomène, alors même que le nombre des caractères est assez grand.

Pourrais-je lui soumettre une suggestion; lui serait-il possible dans un prochain article de compléter son étude théorique par la présentation de tous les calculs afférents aux résultats d'épreuve sportive auxquels il a fait allusion dans sa communication. En l'occurence, il me semble qu'il nous rendrait à tous un très grand service, ce dont nous le remercions à l'avance.

M. Delaporte indique que l'ensemble des graphiques à construire et des calculs numériques à effectuer pour l'analyse factorielle des résultats d'épreuves sportives, auxquels il a fait allusion est trop long pour être publié dans le *Journal de la Société de Statistique de Paris*, mais que les résultats ont été déjà publiés dans Biotypologie (5 et 6).

BIBLIOGRAPHIE

1. — C. Spearman, *The Abilities of Man*, London, Macmillan, 1932.
2. — P. Delaporte, *Une méthode d'analyse des corrélations et son application.* Comptes-rendus Acad. Sciences, Paris, t. 209, 1939, p. 142.
3. — P. Delaporte, *Nouvelle méthode de Statistique mathématique pour l'estimation des facteurs et de leur écart type en Analyse Factorielle. Colloque sur l'Analyse Factorielle et ses applications.* Centre National de la Recherche Scientifique, Paris, 1955.
4. — G. Teissier, *Un essai d'Analyse Factorielle : les variants sexuels de Maia squinado.* Biotypologie, t. VI, n° 2, juin 1938, p. 73-96.
5. — F. Pierre, *Analyse Factorielle des aptitudes sportives.* Biotypologie, t. XII, nos 3-4, décembre 1951, p. 15-21.
6. — P. Delaporte, *Analyse Factorielle de quelques résultats d'épreuves sportives.* Biotypologie, t. XII, nos 3-4, décembre 1951, p. 22-25.
7. — F. Bongard, *L'Analyse Factorielle et la pondération des questionnaires de notation professionnelle. Le choix des hommes.* Chap. VIII, p. 83-91, les Éditions d'Organisation, Paris.
8. — E. C. Rhodes, *The construction of an Index of Business Activity.* J. Royal Stat. Soc., vol. 100, 1937, p. 18-66.
9. — F. Brambrilla e M. Savini. *Contributo per la costruzione di un indice di attività economica. Istituto per gli studi di ecocomica : Econometrica.* Milano-Roma, 1948, 47 p.