JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

PHILIPPE PILIBOSSIAN

Contribution des méthodes statistiques au développement des sciences de l'éducation

Journal de la société statistique de Paris, tome 130, n° 3 (1989), p. 149-161

<http://www.numdam.org/item?id=JSFS_1989__130_3_149_0>

CONTRIBUTION DES MÉTHODES STATISTIQUES AU DÉVELOPPEMENT DES SCIENCES DE L'ÉDUCATION

Philippe PILIBOSSIAN
Université Pierre-et-Marie-Curie (Paris VI)
UER de Mathématiques [1]

C'est le texte intégral de la conférence prononcée au Congrès de l'Éducation, dans le cadre du IIe Congrès Mondial Basque qui a eu lieu en octobre 1987 à Bilbao (Pays basque). L'auteur qui n'est pas un spécialiste de l'analyse des données mais un mathématicien-statisticien, présente, sans utiliser de formules mathématiques, dans une première partie, les différentes rubriques de l'analyse des données, celles de base : l'analyse en composantes principales, l'analyse factorielle des correspondances, l'analyse des proximités, la classification automatique..., et celles qui sont en plein développement : le traitement d'image, l'analyse en scènes, la reconnaissance des formes, les systèmes experts... Dans une seconde partie, il donne quelques exemples d'applications d'études effectuées en France dans le domaine des sciences de l'éducation.

On trouvera en annexe : une bibliographie multilangues mise à jour fin novembre 1988, les références des articles cités et une liste des revues éditées en France, spécialisées en analyse des données.

It is the whole text of the conference pronounced at the Educational Congress within the framework of the II World Basque Congress which took place in October 1987 in Bilbao (Basque Country).

The author, who is not an Exploratory Data Analysis specialist but a mathematician-statistician, presents in the first part, without using any mathematical formulas, different titles of Exploratory Data Analysis, the basic ones of which are: Principal Component Analysis, Correspondance Analysis, Multidimentional Scaling, Cluster Analysis, and those which are being fully developed, i.e. Image Processing, Scenic Analysis, Pattern Recognition, Expert Systems.

In the second part, the author gives some examples of study applications carried out in France in the field of Educational Sciences.

In the annex, one will find the following: a multi-lingual bibliography updated as at end November 1988, references to articles quoted, as well as a list of magazines published in France specializing in Exploratory Data Analysis.

ABRÉVIATIONS UTILISÉES

AD Analyse des Données
ACP Analyse en Composantes Principales
AFC Analyse Factorielle des Correspondances
AP Analyse des proximités
CA Classification Automatique
CH Classification Hiérarchique

1. 4 place Jussieu, 75230 Paris Cedex 05.

Journal de la Société de statistique de Paris, tome 130, n° 3, 1989.

INTRODUCTION

Depuis l'existence de la civilisation humaine, le problème de la *formation* des jeunes générations a été une des principales préoccupations des hommes. Pendant des siècles les sciences humaines s'étant enrichies l'*information* s'amplifia à un point tel que plusieurs sciences s'entremêlèrent pour donner naissance à d'autres sciences. Par conséquent, il devint impossible pour l'homme de tout savoir, il fut contraint de *choisir* son avenir, sa profession, ses idées, sa vie suivant son origine sociale, ses intérêts, sa situation démographique, etc.

Le thème de la *formation* et de *l'orientation* des jeunes, étroitement lié à l'ÉDUCATION est un thème éternel et toujours actuel. Alors il est naturel de lier une science moderne la STATISTIQUE, à cette science vieille comme le monde, qu'est l'ÉDUCATION.

En France actuellement plus que quatre millions d'enfants fréquentent l'école primaire, plus de cinq millions sont inscrits dans un établissement secondaire, et plus d'un million font des études supérieures. Derrière ces chiffres subsiste une grande diversité dans le type d'études menées. Selon l'origine sociale les probabilités d'accéder à l'enseignement supérieur long ou à l'enseignement technique court varient de façon considérable. L'institution scolaire n'a pas la même signification pour tous les Français, pourtant sur un point au moins une certaine unanimité semble se dégager; l'importance d'une bonne préparation à la vie professionnelle, même si elle ne se fait pas de la même façon pour tous, cela dépend de la diversité des origines sociales et des expériences professionnelles. Devant la diversité et l'énormité des données les professionnels de l'Éducation font de plus en plus appel aux techniques sophistiquées d'avant-garde de la Statistique et récemment de l'Informatique.

Le but de mon exposé est de vous faire connaître l'évolution de la Statistique, en vous décrivant d'abord brièvement l'école anglo-saxonne, ensuite le développement de l'école française de l'Analyse des données, en dernier lieu, je vous donnerai quelques exemples d'utilisation des techniques statistiques à certains problèmes particuliers dans le domaine de l'Éducation.

1. *La statistique classique : école anglo-saxonne*

Le terme de STATISTIQUE semble avoir été introduit en 1748 par l'allemand Achenwall, mais il faut attendre la seconde moitié du XIX^{e} siècle, en Angleterre, pour voir se former une méthodologie statistique, c'est-à-dire une théorie bien formalisée de l'*inférence*, du raisonnement qui permet, à partir des données observées, de tirer des conclusions sur les lois de probabilités des populations. Depuis 1900 s'est développée ce qu'on appelle aujourd'hui LA STATISTIQUE MATHÉMATIQUE. Le début de cette période a été dominé par deux grands statisticiens : K. Pearson (1857-1936), auteur du célèbre test universel du χ^2 et Sir R.A. Fisher (1890-1962) qui a mis les bases de la plupart des problèmes de la Statistique mathématique étudiés jusqu'à maintenant.

Dans la pléiade des savants qui ont fortement contribué au développement de cette science, on peut citer quelques noms : Yules, E.S. Pearson, Student, Edgeworth, Neymann, Hotelling, Polya, Wald, Van Neumann, Von Mises, Wiener, Kendall, Wilks, Cramer, ... sans oublier les deux grands théoriciens soviétiques Kolmogorov et Smirnov.

En France, ce sont Émile Borel (1871-1951) et Georges Darmois (1888-1960) qui ont été les premiers innovateurs en Statistique mathématique.

En Statistique classique on utilise un modèle définissant la population-parente de la variable aléatoire dont la caractérisation probabiliste est donnée par sa fonction de répartition. La réalisation d'un nombre élevé d'épreuves aléatoires permet d'obtenir un échantillon, « image se voulant fidèle » de la population-parente. On considère en toute rigueur que les observations sont indépendantes,

identiquement distribuées (iid), extraites au hasard de la population-parente. Dans ce cadre rigide on peut établir de nombreux résultats sur le comportement statistique des variables d'échantillonnage. La procédure est la suivante : on formule une hypothèse sur la caractérisation probabiliste de la population-parente et l'on en tire les lois de distribution de diverses caractéristiques d'échantillonnage calculées à partir de l'échantillon ainsi généré.

La principale hypothèse est ici bien sûr la normalité de la population-parente et l'on peut dire que tous les résultats de la Statistique classique sont obtenus sous cette hypothèse. Pour le mathématicien l'hypothèse de normalité permet une utilisation commode d'un outil mathématique peu compliqué et débouchant sur des résultats de formulation simple. Heureusement à part cette commidité mathématique, il existe un autre argument beaucoup plus sérieux qui plaide en faveur de l'adoption de l'hypothèse de normalité. Cet argument a été formulé sous des formes multiples et représente ce que l'on appelle *« la loi des grands nombres »*. Ce principe dit que tout phénomène régi par des causes multiples, dont aucune n'est dominante suit une loi asymptotiquement normale.

Les grands problèmes de la statistique peuvent être scindés en plusieurs rubriques :

1° Étude de la représentabilité de l'échantillon, propriétés statistiques des variables d'échantillonnage, qualité des estimations et différents types de convergence des variables d'échantillonnage vers leurs analogues théoriques.

2° Étude de l'adéquation des conjonctures vérifiées sur l'échantillon à celle vérifiée sur la population. Ici il s'agit de tous les tests permettant de vérifier les hypothèses ayant trait aux valeurs centrales, ou aux paramètres de dispersion, ou même aux fonctions de répartition.

3° Étude des interliaisons statistiques entre variables aléatoires débouchant sur la possibilité de prévision statistique. C'est le domaine de la recherche des interprétations statistiques de relations à caractère causal.

4° Le passage de l'étude des grandeurs aléatoires unidimensionnelles (v.a. réelle), à celles des vecteurs et des champs aléatoires (variable multivariée).

5° La *statistique non paramétrique (distribution free)* permet également d'obtenir des résultats à caractère probabiliste très importants (tests non paramétriques) sans qu'il soit nécessaire d'adopter des contraintes de normalité.

6° Les *séries chronologiques (ou temporelles)*, dont l'objet est l'étude de l'évolution dans le temps des variables aléatoires (tendance générale ou trend, périodicité, cycle, ...).

7° Un chapitre important de la statistique classique et qui est en plein développement à l'heure actuelle, concerne également la validation des contraintes générales dans le cas où les hypothèses de base ne sont plus valables. « Que se passe-t-il, par exemple, quand pour l'application d'un test statistique, les hypothèses de base sont violées. On entre ici dans le domaine de la *robustesse* (Hubert) et de la *stabilité* de l'inférence statistique.

2. *L'analyse des données : école française*

L'essor de l'*Analyse des données* (AD) est lié à la conjoncture de deux faits importants. D'une part le développement de la technique a entraîné l'accumulation d'énormes masses d'informations que l'on ne pouvait plus traiter d'après les méthodes traditionnelles. D'autre part, l'avènement de l'ère des ordinateurs a donné aux statisticiens un nouvel outil de calcul extrêmement puissant.

Il fallait donc s'affranchir des contraintes rigides de la Statistique classique puisqu'elles devenaient invérifiables pour les masses de données disponibles. Il fallait aussi perfectionner les méthodes de la Statistique descriptive (par de nouvelles représentations graphiques) afin de pouvoir dans un sous-espace approprié « visualiser » la structure des inter-relations statistiques pour un nombre de variables élevées (plusieurs centaines).

En plus, il était nécessaire d'apporter un outil mathématique à des disciplines (psychologie, sociologie, linguistique, ...) où le mode de raisonnement était traditionnellement qualitatif. Cela a nécessité le développement d'un appareil mathématique approprié permettant le traitement de variables qualitatives ordonnées et non-ordonnées.

Le développement des méthodes de l'*Analyse des données* est relativement récent. Il est important de préciser que le terme ANALYSE DES DONNÉES, en français n'a pas la signification de la traduction littérale DATA ANALYSIS, en anglais, mais est plus conforme à l'appelation de J.W. Tukey, EXPLORATORY DATA ANALYSIS. Le terme Analyse des Correspondances a été pour la première fois utilisé par J.P. Benzecri en 1963, lors d'un cours au Collège de France. L'*Analyse Factorielle des Correspondances* (AFC), qui est l'instrument le plus puissant de l'Analyse des données, a été mis au point deux ans plus tard par J.P. Benzecri et Mme B. Escoffier.

L'Analyse des données diffère principalement de la Statistique classique par son approche. Sans faire d'hypothèse *a priori* sur les variables aléatoires étudiées, elle se préoccupe de nous fournir des représentations graphiques simplifiées, d'une interprétation commode, mais à l'aide d'outils mathématiques (géométrie classique, analyse, calcul matriciel, etc.) bien connus depuis plus d'un siècle, mais aisément utilisables aujourd'hui grâce à l'ordinateur.

A ses principaux chapitres de ses débuts, Analyse en Composantes Principales (ACP), Analyse Factorielle des Correspondances (AFC), Classification Automatique (CA), d'autres sont venus d'ajouter. Aujourd'hui on peut distinguer les rubriques suivantes :

2.1. *Analyse en Composantes Principales (ACP)*

L'idée essentielle formulée par K. Pearson en 1905 concernait le principe qui doit être mis à la base de la condensation de l'information. Pearson constatait que parmi les transformations linéaires des variables, les plus informatives sont celles dont la variabilité est maximum. Cette idée de base s'est avérée extrêmement fructueuse et l'on découvrit par la suite que les *composantes* dites *principales*, combinaisons linéaires à variance maximum dans la famille des transformations orthogonales, possèdent de très nombreuses propriétés déterminant leur caractère optimum parmi toutes les transformations orthogonales possibles. A l'heure actuelle grâce à l'ordinateur permettant de réaliser l'opération essentielle de diagonalisation de la matrice des covariances, l'ACP est devenue un outil puissant permettant l'étude de la structure statistique des échantillons de variables quantitatives.

2.2. *Analyse Factorielle des Correspondances (AFC)*

Déjà au début du siècle l'analyse factorielle classique avait été utilisée par les psychologues (Spearman en 1905) pour « expliquer » des résulats des facteurs cachés (mémoire, intelligence, etc.).

L'Analyse Factorielle des Correspondances est l'outil principal pour étudier les inter-relations entre variables qualitatives. Le concept mathématique de cette technique se trouve dans l'introduction de l'ouvrage de Benzecri (1973); c'est l'utilisation d'une métrique appropriée dans l'espace des variables qualitatives, la *distance du khi-deux* (distance entre deux variables qualitatives calculée par rapport à une partition donnée de l'événement certain). On obtient ainsi la matrice d'interdistances entre variables qualitatives. On peut alors appliquer les techniques d'Analyse des Proximités permettant d'obtenir dans tout sous-espace de dimension inférieure les points figuratifs correspondant aux points initiaux dans un espace de dimension élevée, associés à la matrice des interdistances et cela de façon optimale.

2.3. *Analyse des Proximités (AP)*

Il est souvent nécessaire d'étudier la structure topologique d'un nuage de points appartenant à un espace de dimension élevée. Pour cela on ne dispose parfois que de la matrice des interdistances de ces points et l'on voudrait représenter le nuage correspondant dans un espace de dimension inférieure, en l'occurence de plan. Cela permet de déceler, grâce à l'étonnante faculté humaine de la perception visuelle, les interrelations (de *proximité* ou *d'éloignement*) d'un seul coup d'œil, alors que l'esprit humain est incapable de synthétiser la topologie d'un nuage de points décrit par la matrice de ces interdistances.

2.4. *Classification Automatique (CA)*

Le principe général de la *Classification Automatique*, déjà formulé par Aristote, est que les éléments appartenant à la même classe sont en moyenne plus proches que les éléments appartenant à des classes différentes.

Depuis toujours la pensée scientifique s'était fixée pour tâche dans de très nombreuses disciplines de mettre de l'ordre dans une multitude d'objets de nature les plus diverses. C'est ainsi qu'on a établi des classifications des espèces vivantes du monde animal ou végétal (Buffon), des éléments chimiques (Mendeleev), des maladies (Claude Bernard), des sciences (Auguste Comte), des langues (Adjarian), des régimes climatiques, etc.

Ce processus était réalisé de manière subjective par de savants illustres qui ont réussi à dégager dans chaque discipline les principes particuliers qui étaient à la base de la classification. Or, cette nécessité de classer les objets de nature arbitraire, s'est avérée tellement impérieuse que s'est élaborée en tant que discipline indépendante la science de la classification, c'est-à-dire une théorie permettant de façon objective de réaliser cette opération sans intervention humaine (celle des spécialistes). Aujourd'hui on dispose de programmes appropriés permettant de réaliser cette opération de façon parfaitement objective à partir des données fournies par les spécialistes.

Sur le plan méthodologique deux idées principales distinctes sont à la base des algorithmes correspondants ce qui nous permet de définir deux familles de classification :

2.4.1. *Classification autour des Centres mobiles*

Ce sont des procédures d'agrégation autour des *centres mobiles, Méthodes des Étalons,* connues en France sous le nom de *Méthodes des Nuées Dynamiques* (E. Diday).

Chaque classe est assimilée à un *étalon* ou élément représentatif de la classe. Les étalons sont initialisés au départ soit par affectation subjective directe, soit par tirage aléatoire. Le fichier est ensuite relu et, grâce à l'introduction d'une métrique appropriée, chaque élément du fichier est affecté à la classe correspondante, à partir du principe du minimum de la distance à l'étalon correspondant. Chaque étalon est alors modifié en fonction de l'évolution de l'effectif des classes. Les étalons sont l'une des caractéristiques centrales de la population (valeur moyenne, médiane, mode). Bien que le plus logique serait de prendre les valeurs modales pour définir les étalons, la commodité mathématique fait que l'on choisit le plus souvent la valeur moyenne.

On distingue aussi deux types d'algorithmes correspondant au nombre de classes fixé ou non fixé à l'avance.

Le principe de la méthode est qu'en effectuant un certain nombre d'itérations (relecture complète du fichier avec affectation de classes) l'influence du choix initial subjectif ou arbitraire des étalons s'atténue et l'algorithme converge vers une structure stable. Cette convergence est beaucoup plus lente pour les algorithmes avec un nombre de classes non fixé.

2.4.2. *Classification Hiérarchique (CH)*

Le principe de la *Classification Hiérarchique* est l'introduction de deux catégories de distances : distance entre éléments et distance entre classes. L'affectation d'un objet à une classe est basée sur le principe du minimum de la distance entre éléments, puis entre classes. L'idée de base de la méthodologie consiste à regrouper dans une même classe les deux éléments ou les deux classes les plus proches au sens de la métrique adéquate. Le principal intérêt de la CH (Jambu) est de fournir un arbre de classification, c'est une structure graphique hiérarchique permettant de matérialiser la structure des interconnections réciproques entre les classes.

2.5. *La prévision statistique*

L'un des principaux objectifs des statisticiens est d'estimer la valeur future de variables aléatoires, ou encore de prévoir la réalisation de certains événements. Pour cela, suivant la nature des variables explicatives on fait recours à la Régression (variables quantitatives) ou à l'Analyse Discriminante (variables qualitatives).

2.5.1. *La Régression*

Dans le cadre d'un modèle de génération simultanée d'un couple prédicteur-prédictand (X, y) au cours d'une épreuve aléatoire, on recherche dans une famille paramétrique donnée de fonctions $G = \{g_\alpha(X)\}$, la « meilleure » approximation (au sens d'un critère de qualité Q donné) d'une relation physique $y = \varphi(X)$ inconnue φ, mais supposée existante. Les différentes familles de régression correspondent aux choix appropriés :

1° de la famille $G = \{g_\alpha(X)\}$ (linéaire, quadratique, polynômiale ou autres);

2° du critère de qualité $Q = Q(\varphi, \widehat{\varphi})$ retenue (erreur quadratique moyenne, erreur absolue moyenne, erreur minimax, erreur quadratique des distances orthogonales, ...).

Sur le plan pratique le problème le plus important reste celui du choix des variables les plus informatives dans l'ensemble des nombreux prédicteurs potentiels, et celui de la *stabilité* de la qualité des résultats obtenus sur le *fichier* dit d'*apprentissage* lors du passage à un *fichier test.*

L'utilisateur est concerné par la qualité qu'il obtiendra sur le *fichier test* et de nombreuses méthodes ont été élaborées pour définir une évaluation objective de cette qualité (Reconnaissance glissante, Jackknife Chaotisation, Bootstrap). A l'heure actuelle les plus en vogues sont les méthodes de reéchantillonnage (Bootstrap), introduite par Efron pour obtenir des estimations objectives dans le cas de petits fichiers.

2.5.2. *L'Analyse Discriminante*

L'une des constantes de l'activité humaine est la prise de décision et souvent elle s'accompagne d'une incertitude sur les conséquences de telle ou telle décision. En ce sens l'*Analyse Discriminante* est la théorie de prise de décision optimale dans un environnement incertain. L'originalité réside dans le caractère qualitatif du prédictand (prévision par classes) et dans l'évaluation du critère de qualité; celui-là est défini grâce à l'introduction d'une matrice des coûts des décisions erronées; la décision optimale étant celle qui minimise l'espérance mathématique des coûts. Ces méthodes, peuvent être scindées suivant le degré de connaissance de la caractérisation probabiliste des prédicteurs en méthodes paramétriques (pour lesquelles on connaît la loi de distribution du vecteur des prédicteurs X) et non paramétriques (pour lesquelles aucune hypothèse n'est faite sur la loi de X).

De multiples applications sont faites dans des domaines très divers : prévision de phénomènes dangereux, problème de faillite d'entreprise, diagnostic médical, prospection géologique, etc.

2.6. *Traitement d'images, Analyse des Scènes*

Une part importante de l'information qui envahit le monde moderne est livrée sous forme d'images qui décrivent différents objets. L'exemple le plus frappant est celui des photo-satellites qui fournissent une information considérable et on se trouve devant le problème de l'utilisation de ces informations. Elles ne peuvent plus être traitées par les méthodes traditionnelles après numérisation à cause du volume considérable des données correspondantes (une image est classiquement définie par un tableau 256×256). Il est donc nécessaire d'élaborer des méthodes automatiques de *traitement d'images* basées sur les principes de la Statistique spatiale. Elle requiert l'utilisation de la notion de fonction d'autocorrélation et de structure *(variogramme)*, et comporte l'étude de la texture des figures correspondantes. Ce traitement ne peut être effectué, que par l'ordinateur et l'on développe de plus en plus des ordinateurs spécialisés dans le traitement d'images. Les applications sont multiples : météorologie, géologie, géomorphologie, géographie, applications militaires, etc.

Dans ce domaine un chapitre intéressant est constitué par l'*Analyse des Scènes*, où l'on développe des méthodes permettant d'analyser les différentes composantes des images dans leurs rapports mutuels. Le programme est destiné à remplacer la synthèse réalisée traditionnellement par un opérateur humain (nephanalyse automatique, surveillance d'images, radar, etc.).

2.7. *Reconnaissance des formes*

Dans de nombreux domaines pratiques il est aujourd'hui nécessaire d'élaborer des procédures automatiques réalisant l'opération d'*identification de formes* connues. Dans ce domaine l'Analyse des données apporte une contribution intéressante pour dégager dans une collection d'objets décrits par un ensemble de caractérisations numériques, les paramètres synthétiques qui permettent ensuite de les identifier. Ces objets peuvent être très complexes tels dans le cas de la reconnaissance de la parole et plus généralement la reconnaissance d'un signal perturbé par un bruit.

2.8. *Systèmes Experts*

Depuis une décennie se développent des méthodes dites d'identification automatique s'apparentant au principe du comportement humain devant le processus de décision. La première étape consiste à formaliser la base des règles élémentaires dont l'expert humain se sert pour formuler ses décisions. Cette base de règles est ensuite introduite dans un *moteur d'inférence* où elle est organisée, ordonnée, hiérarchisée afin d'élaborer au cours d'un processus intéractif d'interrogation de l'utilisateur la réponse finale après élimination de toutes les situations non conformes aux règles. L'avantage de cette procédure est de faire participer le décideur humain au processus de prise de décision, alors que pour bien d'autres méthodes le processus de décision lui apparaît comme une boîte noire à laquelle il ne participe pas. Le décideur retrouve ainsi un cheminement de la pensée qui lui est familier et dans lequel il est partie prenante. La généralisation des micro-ordinateurs a pour effet de multiplier les applications possibles des Systèmes Experts, que l'on élabore dans les domaines les plus divers, mais qui pour l'instant, s'ils ne sont pas encore capables de concurrencer les méthodes traditionnelles, permettent une meilleure formalisation du processus de prise de décision.

Depuis 1965, l'AD s'est beaucoup développée et J.P. Benzecri a de nombreux élèves aussi bien en France qu'à l'étranger parmi lesquels le plus méritant bien que le plus modeste est Pierre Cazes.

Cependant, je pense utile de vous signaler que d'autres statisticiens indépendamment de son école sont connus dans le monde pour leurs travaux poussés dans le domaine de l'AD.

Voici quelques noms : aux USA : J.W. Tukey, Sokal, F. Mosteller, C.R. Rao; en URSS : Aïvazian, Enioutrov, Bouliguine, Michkine, Dorofeiouk; Angleterre : Critckley; Japon : Hayashi, Tanaka, Oshumi; RSA : M. Greenacre; Italie : Lauro, Ricci; Belgique : Libert; Pays-Bas : De Leeuw; Israël : Guttman; Portugal : Bacelar-Nicolau; Argentine : Carreiro; Brésil : Da Fouseca-Genevois; Espagne : F. Galléco et Pays Basque : Y. Yurraendi et même en France : Der Megreditchian, Caussinus, M. Masson, Michaud, Marcotorchino, Tomassone, Escoufier, Schektman.

3. *Exemples d'application à l'éducation*

Il existe des domaines variés et nombreux où la Statistique et l'Analyse des données ont pu apporter une contribution importante : linguistique, archéologie, géologie, géographie, écologie, météorologie, urbanisme, médecine, sociologie, psychologie, et dans une moindre mesure en : économie.

On peut trouver de nombreux exemples d'application principalement dans les ouvrages de J.P. Benzécri, dans *les Cahiers de l'Analyse des données* et dans une série d'article de R. Gibrat dans le *Journal de la Société de Statistique de Paris.*

Nous avons choisi ici quelques exemples dans le domaine de l'Éducation.

3.1. *Les objectifs de l'enseignement, une synthèse. L. Haenster (1986)*

Une enquête effectuée par le Centre de Recherche pour l'Étude et l'Observation des Conditions de vie (CREDOC), en 1985, dans la région parisienne dont le niveau d'équipement est plus élevé, avait pour but de faire une synthèse des opinions des Français en matière d'enseignement, de leurs intérêts. Les questions portaient sur le rôle des établissements d'enseignement, les débouchés et la culture générale.

On a appliqué une AFC, avec comme principale préoccupation de vérifier la corrélation entre intériorisation sociale des contraintes et choix entre culturel et professionnel.

On a pu distinguer sept classes (ou zones), dont toutes ne présentent pas un intérêt évident du point de vue de l'analyse des attentes en matière d'éducation. On a mis en valeur l'opposition entre, d'une part, les partisans d'études longues, de l'intérêt culturel ou intellectuel des enseignements au niveau secondaire et d'autre part, ceux qui ont acquis essentiellement des diplômes techniques de niveau inférieur au Baccalauréat. Cette dualité des attitudes ne reflète pas totalement la diversité des réponses. Le facteur qui explique le mieux ce type de choix est le niveau du diplôme; son effet est renforcé si l'on tient compte du diplôme du père.

3.2. *Recherche d'un scalogramme sur les réponses de 1 300 élèves à une batterie d'épreuves de mathématique — F. Murtagh (1981)*

L'épreuve de *Drumcondra Criterion Referenced Mathematics Test* (DCRMT) a été appliquée en 1974-1975 en Irlande à 1 300 élèves. Les questions portent sur tous les points importants du programme de mathématiques, les réponses étant sous forme de choix multiples, avec une réponse exacte. L'épreuve est notée suivant un ensemble de critères ou qualités (portant sur la compréhension, le calcul, la solution du problème, etc.). Le test a été appliqué deux fois avec un intervalle de six mois sur la même population. Les résultats ont indiqué qu'il y avait accroissement global des connaissances entre les deux passations de l'épreuve, mais qu'il subsiste beaucoup d'élèves d'un niveau très bas.

L'auteur de l'étude se pose une question « dans quelle mesure peut-on ordonner les critères? Un ordre de facilité-difficulté, commun aux diverses applications de l'épreuve, impliquerait un matériel digne d'intérêt pour l'examen des connaissances mêmes ».

On a appliqué d'abord un algorithme de Classification Automatique Ascendante Hiérarchique et ensuite un d'Analyse Factorielle des Correspondances. L'auteur examine en détail les différentes distributions des garçons et des filles, ainsi que les variations de niveau des élèves qui ont passé deux fois l'épreuve.

L'étude comporte de nombreux graphiques qui pourraient éclairer l'interprétation dans le domaine de l'enseignement.

3.3. *Autres exemples*

Faute de temps, je citerais uniquement quelques exemples portant sur l'éducation :

1° Les lycéens du second cycle : par J. Goudard et Y. Grelet (1977).

2° Statistique de l'enseignement en Grèce, par M. Meimaris (1978).

3° Orientation des bacheliers tunisiens dans l'enseignement supérieur, par N. Gnichi (1979).

4° Orientation des élèves après la classe de troisième par D. Trancard (1981).

5° Diplômes et emplois. Les emplois précaires, par J. Affichard, D. Meure et F. Audière (1984).

6° Flexibilité polyvalence et mobilité par Sylvestre, Hollard et Kiffer (1986).

7° Évolution de données de sommeil hebdomadaire... chez des enfants..., par P. Koch et C. Taillard (1986).

8° Enquête épidémiologique sur la consommation de produit psychotropes et la santé des adolescents, par F. Fay et H. Ralambondrainy (1987).

Cet après-midi, j'écouterai avec vous d'autres exemples qui seront présentés par des collègues ici-même.

Mais par rapport à d'autres disciplines l'utilisation de la Statistique dans les études de l'Éducation reste peu courant. Je tiens à souligner que les organisateurs de ce Congrès ont eu la bonne idée et le mérite de consacrer une journée à la méthodologie statistique. Désormais, la voie étant déjà largement ouverte, je souhaite que les spécialistes de l'Éducation fassent beaucoup plus souvent recours aux nouvelles techniques de la Statistique mathématique, de l'Analyse des données, ainsi qu'à ceux de l'Informatique dans leurs recherches, afin d'y apporter un peu plus d'objectivité et d'ordre.

CONCLUSION

Il est difficile de faire un survol complet de la Statistique mathématique et de l'Analyse des données en une heure. On m'excusera, j'espère, d'avoir omis de citer tous les mathématiciens et praticiens qui ont contribué au développement de la Science statistique. J'aurais souhaité vous exposer en particulier les nouveautés mises au point en Statistique mathématique en France et dans les autres pays et qui trouvent des applications dans divers domaines et même dans de nouveaux secteurs inattendus. Mais j'ai essayé d'être le plus concis et objectif que possible; j'ai sciemment évité l'emploi des formules dans une discipline relevant des Mathématiques.

Je ne veux pas non plus m'attarder sur les querelles d'écoles inévitables pendant la naissance d'une méthodologie : d'une part entre les statisticiens d'école mathématique et les partisans de l'Analyse des données, et d'autre part, entre les économètres (qui seraient « de droite ») et les partisans de l'AD

(qui seraient « de gauche »). Voici ce que Michel Volle (1978) dit à ce sujet : *« Il est ... plus pertinent de voir dans l'Analyse des données une sorte de retour aux sources avec des moyens bien supérieurs aux travaux des pionniers comme* A. Quételet (1796-1874) et Sir F. Galton (1822-1917) *qui ... établissent les premiers fondements de la Statistique mathématique sur un volumineux travail expérimental. Il est impossible de savoir vers quelles constructions théoriques nous mène l'AD actuelle; elle annonce cependant, à coup sûr, du nouveau. »*

Il existe un autre terrain glissant dont il faut attirer l'attention du public : c'est l'utilisation abusive et erronée des méthodes statistiques. Un non statisticien, au nom de la Statistique ou de l'Analyse des donnée tire parfois les conclusions qui lui conviennent, sans se préoccuper de la validité des méthodes. D'un autre côté, on demande souvent à un statisticien de faire une étude complète dans un domaine qui lui est complètement inconnu. Dans ce cas il ne peut qu'apporter son outil pour éclairer le raisonnement du praticien, mais c'est au spécialiste qu'il appartient de donner une interprétation concrète des résultats de l'analyse statistique.

En guise de conclusion, permettez-moi de citer un de mes anciens professeurs, plus tard ami Georges Morlat : *« L'Analyse des données doit rendre service partout où l'on se soucie d'accumuler des observations... Les services rendus montrent bien que l'analyse des données constitue aujourd'hui, et de loin, la partie la plus immédiatement rentable de la statistique... Cela permet, selon les cas, de découvrir dans les phénomènes étudiés des structures directement visibles sur les résultats de l'analyse, alors qu'elles ne l'étaient pas sur les données originelles, ou de retrouver en les précisant des structures que l'on soupçonnait déjà pour telle ou telle raison. »*

(Introduction du livre de Cailliez Pagès (1976)).

ANNEXES

A : Livres

Alvazian S. (1977). — Étude statistique des dépendances (traduit du russe). Éditions Mir.

Alvazian S.; Enukov I. et Mechalkine L. (1983). — Prikladnava statistika, 3 vol. (en russe) Moscou. Finansv i statistika.

Alvazian S.; Enukov I. et Mechalkine L. (1986). — Éléments de modélisation et traitement primaire des données (traduit du russe). Éditions Mir.

Anderberg M.R. (1973). — Cluster Analysis for Applications Academic Press.

Barbut M. et Frey L. (1971). — Technique ordinale en analyse des données, tome 1 : Algèbre et combinatoire, Hachette.

Benzecri J.-P. (1982). — Histoire et préhistoire de l'analyse des données, Dunod.

Benzecri J.-P. et collab. (1973, 1984). — L'Analyse des données. 1. La Taxinomie, 4e éd., Dunod.

Benzecri J. P. et collab. (1973, 1980). — L'Analyse des données. 2. L'Analyse des correspondances, 3e éd., Dunod.

Benzecri J. P. et Benzecri F. et collab. (1980, 1984). — Pratique de l'analyse des données. 1. Analyse des correspondances : Exposé élémentaire, 2e éd. augmentée, Dunod.

Benzecri J.-P.; Bastin Ch.; Bourgarit Ch. et Cazes P. (1980). — Pratique de l'analyse des données. 2. Abrégé théorique : Études de cas modèles, Dunod.

Benzecri J.-P. et coll. (1981). — Pratique de l'analyse des données. 3. Linguistique et lexicologie, Dunod.

Benzecri J.-P. et coll. (1986). — Pratique de l'analyse des données. 5. Sciences économiques, Dunod.

Bertier P. et Bouroche J. M. (1977, 1982). — Analyse des données multidimentionnelles. 3e éd. PUF.

Bouroche J.-M. et Saporta G. (1980, 1987). — L'Analyse des données. 3e éd., Coll. Que sais-je? 1854. PUF.

Der Megreditchian G. (1988). — Le Traitement statistique des données multidimentionnelles, 2 tomes, École Nat. Météo., Toulouse.

Cailliez F. et Pages J.-P. (1976). — Introduction à l'analyse des données, SMASH.

Compstat (1974, 1976, 1978, 1980, 1982, 1984, 1986). — Proceeding in Computational Statistics, 8 vol. Physica Verlag.

Dagnelie (1975). — Analyse statistique à plusieurs variables. Presses Agronomiques.

Diday E. (1981). — Optimisation en classification automatique, 2 vol. INRIA.

Diday E. (éditeur) (1986). — Data Analysis and Informatics IV, North Holland.

Diday E.; Hayushi C. et Jambu M. (éditeurs) (1988). — Proceedings of the Japanese French Scientific Seminar, March 24 28, 1987. Academic Press.

Diday E.; Jambu M.; Lebart L.; Pages J. P. et Tomassone R. (éditeurs) (1984). — Data Analysis and Informatics, III, North-Holland.

Diday E.; Lebart L.; Pages J. P. et Tomassone R. (éditeurs) (1980). — Data Analysis and Informatics, North-Holland.

Diday E.; Lemaire J.; Pouget J. et Testu F. (1982). — Éléments d'analyse de données, Dunod.

Diday E. et coll. (1985). — Metodi analiza dannih (en russe), Moscou, Finansy i statistika.

Epron B. (1982). — The Jackknife, the Bootstrap and Other Resampling Plans, Soc. Indust. Appl. Math.

Erikson B.H. et Nosanchuk T.A. (1977). — Understanding Data, McGraw-Hill.

EUROSTAT (Luxembourg) (1984). — Développements récents dans l'analyse de grands ensembles de données, CR du Séminaire de Luxembourg, 1983. Informations de l'Eurostat.

Fenelon J. P. (1981). — Qu'est-ce que l'analyse des données? LEFUNEN.

Foucart T. (1982). — Analyse factorielle, Programmation sur micro-ordinateur avec nouveaux programmes, 2e éd., Masson.

Foucart T. (1984). — Analyse factorielle et tableaux multiples, Masson.

Foucart T. et Lafaye J.-Y. (1983). — Régression linéaire sur micro-ordinateurs, Masson.

Freeman D.H. (1987). — Applied Categorical Data Analysis, Dekker.

Gaul W. et Schader M. (éditeurs) (1986). — Classification as a Tool of Rechearch (congrès), North-Holland.

Gnanadesikan R. (1977). — Methods for Statistical Data Analysis of Multivariate Observations, Wiley.

Greenacre M. (1984). — Theory and Application of Correspondence Analysis, Academic Press.

Guigudu J.-L. (1977). — Méthodes multidimentionnelles, Analyse des données et choix à critères multiples, 2e éd. complétée, Dunod.

HAND (1981). — Discrimination and Classification, Wiley.
HARTIGAN J.A. (1975). — Clustering Algorithms, Wiley.
HOAGLIN D.C.; FREDERICK M. et TUKEY J.W. (1983). — Understanding Robust and Exploratory Data Analysis, Wiley.
HOAGLIN D.C.; FREDERICK M. et TUKEY J.W. (1985). — Exploratory Data Tables, Trends and Shapes, Wiley.
HUBERT P.J. (1981). — Robust Statistics, Wiley.
JAMBU M. (1979). — Classification automatique pour l'analyse des données. 1. Méthodes et algorithmes, Dunod.
JAMBU M. (1988). — Ierarxitcheskiy klaster Analiz i sootvetstruya (en russe), Moscou, Finansv i statistika.
JAMBU M. et LEBEAUX M.O. (1979). — Classification automatique pour l'analyse des données. 2. Logiciels, Dunod.
JAMBU M. et LEBEAUX M.O. (1983). — Clusster Analysis for Data Analysis, North-Holland.
JOLLIFFE I.T. (1986). — Principal Component Analysis, Springer Verlag.
KOOPMANS L.H. (1981). — Introduction to Contemporary Statistics, Ouxbury Press.
KOTZ S. et JOHNSON N.L. (1982-1983). — Encyclopedia of Statistical Science, vol. 2, pp. 579-582, vol. 3, pp. 479-582, Wiley.
LAGARDE de J. (1983). — Initiation à l'analyse des données, Dunod.
LEBART L. et FENELON J.-P. (1975). — Statistique et informatique appliquées, 3e éd., Dunod.
LEBART L. et MORINEAU A. (1982). — SPAD : Système portable pour l'analyse des données, CESIA.
LEBART L.; MORINEAU A. et FENELON J.-P. (1979). — Traitement des données statistique. Méthodes et programmes, Dunod
LEBART L.; MORINEAU A. et FENELON J.-P. (1985). — Tratamento Estatistico de Datos (en espagnol), Marcambo, Barcelona.
LEBART L.; MORINEAU A. et TABARD N. (1977). — Technique de la description statatistique, Méthodes et logiciels pour l'analyse des grands tableaux, Dunod.
LEBART L.; MORINEAU A. et WARWICK K.M. (1984). — Multivariate Descriptive Statistical Analysis, Wiley.
LERMAN I.C. (1970). — Les Bases de la classification automatique, Gauthier-Villard.
LERMAN I.C. (1981). — Classification et analyse ordinale des données, Dunod.
MARCOTORCHINO J.F. et MICHAUD P. (1979). — Optimisation en analyse ordinale des données, Masson.
MASSON M. (1980). — Méthodologies générales de traitement statistique de l'information de masse, 1re partie : Théorie, 2e partie : Pratique, CEDIC/Fernand Nathan.
MATUSITA K. (éditeur) (1985). — Statistical Theory and Data Analysis, Proceeding of the Pacific Area Statistical Conference, North-Holland.
MONEIL D.R. (1977). — Interactive Data Analysis, Wiley.
MOSTELLER F. et TUKEY J.W. (1977). — Data Analysis and Regression, Addison-Wesley.
MOSTELLER F. (éditeur) (1973). — Statistics by Examples, 4 vol., Addison-Wesley.
MURTAGH F. et HECK A. (1987). — Multivariate Data Analysis, Reidel Publ. Comp.
NISHISATO S. (1980). — Analysis of Categorical Data, Dual Scaling and its Applications, Univ. Toronto Press.
ROMEDER J.-M. (1971). — Méthodes et programmes d'analyse discriminante, Dunod.
ROUX M. (1985). — Algorithmes et classification, Masson.
SAPORTA G. (1979). — Théories et méthodes de la statistique, Technip.
SOKAL R.R. et SNEATH P.H.A. (1973, 1980). — Principles of Numerical Taxonomy, 2e éd., Freeman.
TOMASSONE R.; DAUZART M.; DAUDIN J.J. et MASSON J.P. (1988). — Discrimination et classement, Masson.
TORENS-IBERN J. (1972). — Modèle et méthodes de l'analyse factorielle, Dunod.
TUKEY J.W. (1977). — Exploratory Data Analysis, Addison-Wesley.
VAN RIJCKEVORSEL et al. (1988). — Component and Correspondence Analysis : Dimension Reduction by Functional Approximation, Wiley.
VELLEMAN P.F. et HOAGLIN D.C. (1981). — Applications, Basics and Computation of Exploratory Data Analysis, Ouxbury Press.
VOLLE M. (1978, 1984). — Analyse des données, 3e éd. Economica.

B. Articles cités

Adames (1985). — Analyse d'un fichier de la banque des données statistiques de l'UNESCO, le fichier par âge, sexe et pays des élèves immatriculés. Cahiers Analyse Données, 4, pp. 53-74.

Affichard J.; Menu D. et Audier F. (1984). — Diplômes et emplois, les emplois précaires. Formations et Emplois, CEREQ, Documentation française.

Aludaat K. (1983). — Les étudiants jordaniens à l'étranger de 1975 à 1981 : relations culturelles et relations économiques. Cahiers Analyse Données, 8, pp. 293-310.

Facy F. et Ralambondrainy H. (1986). — Enquête épidémiologique sur la consommation de produits psychotropes et la santé des adolescents. Colloque sur l'AD et Recherche en Santé et Sécurité du travail, 6 8 oct. 1986, Montréal.

Gibrat R. (1978). — L'Analyse des Données (1re partie). Journal Soc. Statist. Paris, 119, pp. 201-228.

Gibrat R. (1978). — L'Analyse des Données (2e partie). Les Sciences humaines : impasses, échecs et succès. Journal Soc. Statist. Paris, 119, pp. 312-331.

Gibrat R. (1979). — L'Analyse des Données (3e partie). Causalité et Analyse des Données en Médecine. Journal Soc. Statist. Paris, 120, pp. 224-243.

Gnichi N. (1979). — Orientation des bacheliers tunisiens dans l'enseignement supérieur. Cahiers Analyse Données, 4, pp. 301-312.

Goudard J. et Grelet Y. (1977). — Les lycées du second cycle : comparaison entre filles et garçons. Cahiers Analyse Données, 2, pp. 273-291.

Haeusler L. (1986). — Le système d'enquêtes sur les conditions de vie et aspirations des Français. L'Éducation, vague d'automne 1985. CREDOC Juillet 1986.

Koch P. et Taillard C. (1986). — Évolution de données de sommeil hebdomadaire au cours du premier trimestre de l'année scolaire 1984, chez les enfants de CM1 à terminale. Cahiers Analyse Données, 9, pp. 229 234.

Meimaris M. (1978). — Statistique de l'enseignement en Grèce : étude des différents établissements d'enseignement supérieur suivant l'origine socioprofessionnelle de leurs étudiants. Cahiers Analyse Données, 3, pp. 355 365.

Murtagh F. (1981). — Recherche d'un scalogramme sur les réponses de 1 300 élèves à une batterie d'épreuves de Mathématiques. Cahiers Analyse Données, 6, pp. 297 318.

Sylvestre; Hollard et Kiffer (1986). — Flexibilité, polyvalence et mobilité, CEREQ, Formations et Emplois, Documentation française.

Trancart D. (1981). — L'orientation des élèves après la classe de troisième : vœux, aptitudes et condition sociale. Cahiers Analyse Données, 6, pp. 19-38.

C. Revues spécialisées publiées en France

— Cahiers de l'Analyse des Données depuis 1976, dir. J.-P. Benzécri.
— Statistique et Analyse des Données depuis 1976, dir. E. Van Cutsem.
— Revue de Statistique appliquée, depuis 1953, dir. F. Cazes.
— Journal de la Société statistique de Paris, depuis 1860, dir. P. Damiani.
— Mathématique et Sciences humaines depuis 1962.
— La Météorologie, dir. Lebaux.
— Rapports de l'INRIA.
— Formation et Emplois, Documentation française éditée par le CEREQ.