

Une conversation avec Yves Escoufier

Yves Escoufier, Gilbert Saporta

▶ **To cite this version:**

Yves Escoufier, Gilbert Saporta. Une conversation avec Yves Escoufier. Statistique et Société, 2022, 10 (2), pp.81-94. hal-03872079

HAL Id: hal-03872079
https://cnam.hal.science/hal-03872079

Submitted on 25 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Une conversation avec Yves Escoufier



Yves ESCOUFIER[1]
Université de Montpellier

Gilbert SAPORTA[2]
Conservatoire national des arts et métiers, Paris

TITLE
A conversation with Yves Escoufier

RÉSUMÉ
Yves Escoufier est un des fondateurs de l'école française d'analyse des données. Son influence a amplement dépassé les frontières de l'hexagone avec ses travaux sur les tableaux à trois indices, l'opérateur qui porte son nom et le coefficient RV de corrélation entre deux vecteurs.
Très investi dans la communauté des statisticiens, il a été président de l'Association pour la Statistique et ses Utilisations — ancêtre de la Société Française de Statistique — de 1984 à 1986. Par ailleurs, Yves Escoufier a été vice-président de 1989 à 1994, puis président de 1994 à 1999 de l'université de Montpellier 2. Il a été vice-président de l'Institut international de statistique de 1991 à 1993. Il est docteur honoris causa de l'université La Sapienza de Rome (1996).
Au cours de cet entretien réalisé en septembre 2021, Yves Escoufier revient sur sa carrière et nous fait part de ses réflexions.

Mots-clés **:** *analyse factorielle, opérateurs, coefficient RV, STATIS.*

ABSTRACT
Yves Escoufier is one of the founders of the French school of data analysis. His influence has extended far beyond France with his work on three-index tables, the operator that bears his name and the RV coefficient for correlation between two vectors.
Very involved in the community of statisticians, he was president of the Association pour la Statistique et ses Utilisations - ancestor of the Société Française de Statistique - from 1984 to 1986. In addition, Yves Escoufier was vice-president from 1989 to 1994, then president from 1994 to 1999 of the University of Montpellier 2. He was Vice-President of the International Statistical Institute from 1991 to 1993. He holds an honorary doctorate from La Sapienza University in Rome (1996).
In this interview conducted in September 2021, Yves Escoufier looks back on his career and shares his thoughts with us.

Keywords**:** *factor analysis, operators, RV coefficient, STATIS.*

1. escoufieryves@gmail.com / désigné par les initiales YE dans cet entretien.
2. gilbert.saporta@cnam.fr / désigné par les initiales GS dans cet entretien.

© Société Française de Statistique (SFdS), Octobre/October 2022

GS : Comment es-tu devenu statisticien ?

YE : Quand je me suis inscrit en licence de mathématiques, mon idée était d'aller ensuite dans l'industrie. J'avais d'ailleurs écrit à EDF et à la SNCF qui m'avaient répondu : « Quand vous aurez votre licence… ».

Au cours de la licence, j'ai suivi un certificat semestriel de probabilités et statistique qui était le seul certificat à caractère appliqué. Cela m'avait intéressé et j'avais même envisagé de m'inscrire à l'ISUP.

À la fin de la licence, Monique Lafont qui était la professeure de probabilités et statistique, a fait savoir que le centre d'études phytoécologiques (CEPE) que le professeur Louis Emberger venait de créer au CNRS, cherchait un calculateur[3] pour faire des analyses de variance. D'autre part le département de maths créait un DEA et comme j'avais rencontré mon épouse, j'ai donc décidé de rester à Montpellier à la fois pour travailler comme calculateur à mi-temps au CNRS, et pour faire le DEA. C'était en 1964-65.

À la rentrée 1965, on nous a informés qu'il y avait des postes d'assistant à prendre. C'était le moment où l'université rentrait dans ses nouveaux murs, où le nombre d'étudiants augmentait, et il y avait plus de postes d'assistants que de candidats. Pendant un an j'ai été assistant au département de mathématiques.

J'avais demandé à pouvoir continuer ma recherche au CEPE parce que j'y avais découvert les méthodes factorielles et l'ACP, et aussi appris le Fortran. À la rentrée 1966, Jean Falguerettes, qui était l'assesseur du doyen Bernard Charles, créait le département informatique de l'IUT et cherchait des gens qui connaissaient un peu d'informatique. Et puisque je connaissais Fortran, j'étais considéré comme un informaticien. Et c'est comme ça que j'ai débuté ma carrière.

Pour revenir au CEPE, j'y avais donc été embauché pour faire des analyses de variance. Un jour, des chercheurs sont arrivés avec la thèse de Pierre Dagnelie[4] sur l'analyse factorielle en écologie et m'ont demandé de leur expliquer ce qu'il faisait. Et c'est comme ça que j'ai été immergé dans ce type de méthodes. On avait – je crois que ça a été un gros avantage des Français, on en reparlera – fait beaucoup d'algèbre linéaire dans notre formation, si bien que toutes ces méthodes multivariables étaient évidentes pour nous, en tout cas très faciles à aborder. Et je me suis plongé dans le livre de H. Harman[5], puis dans ceux de C. R. Rao[6] et de T. W. Anderson[7].

Mes premières analyses en composantes principales, je les ai faites à la main, sur la machine de bureau avec laquelle d'habitude je faisais des analyses de variance. Ça prenait du temps. J'ai alors appris qu'à Nancy, Richard Tomassone, qui était à ce moment-là à l'École des eaux et forêts, proposait des stages dans lesquels il présentait ces méthodes d'analyse statistique multivariable. Parallèlement, Claude Pair apprenait à les programmer en Fortran. Et c'est là que j'ai commencé à voir l'ensemble des méthodes multivariables et à les programmer. Richard Tomassone – que l'on retrouvera au moment de la création de l'unité de biométrie – a été mon premier contact avec quelqu'un qui connaissait ces méthodes et les enseignait. Sinon, c'étaient les livres. Il n'y avait pas de compétences sur Montpellier.

Côté recherche, j'ai toujours travaillé avec les sciences du vivant. Dès le début, j'ai fait de la statistique avec des chercheurs qui étaient en écologie, en biologie : pour eux la statistique,

3. Un emploi de technicien à l'époque, pas une machine
4. Professeur à la Faculté des Sciences agronomiques de Gembloux, Belgique
5. Harman H. H. (1960), *Modern factor analysis*, Univ. of Chicago Press.
6. Rao C. R. (1965), *Linear statistical inference and its applications*, John Wiley & Sons Inc.
7. Anderson T. W. (1958), *An introduction to multivariate statistical analysis*, John Wiley & Sons Inc.

© Société Française de Statistique (SFdS), Octobre/October 2022

c'était les analyses de variance ; ils comparaient des parcelles selon leur production et on faisait ça à la main, enfin à la machine. Ayant vu qu'ailleurs des chercheurs faisaient de l'analyse multivariable, les gens du CEPE se sont demandé pourquoi on ne ferait pas pareil. Une fois en poste à l'IUT, on a su que j'étais capable de faire des analyses statistiques et il y a eu énormément de demandes de biologistes et d'entreprises locales.

On a eu par exemple un contrat de recherche avec la direction environnement d'Elf Aquitaine. Elf Aquitaine extrayait à Lacq du gaz qui était chargé en soufre. Les paysans autour s'étaient plaints des rejets de soufre. Il y avait un réseau de mesures autour de Lacq pour mesurer la quantité de soufre dans l'air toutes les heures ou les deux heures. La direction est venue nous voir en nous demandant si on pouvait les aider à rationaliser ce réseau parce qu'ils étaient bien conscients qu'on avait mis un capteur ici parce que le maire avait crié fort, un autre là parce que c'était à côté de la maison de monsieur untel, etc. Mais les premières données qu'ils nous ont apportées étaient sous forme de listings car à cette époque-là – c'était le tout début – ce qu'on savait faire, c'était enregistrer sur des *listings*. Le premier travail a été de tout ressaisir. Plus tard on a vu arriver les disques et les bandes magnétiques. C'était pareil avec les Salins du Midi : ils enregistraient des données de météo depuis 100 ans. Quand ils ont eu un ordinateur, au lieu de les mettre dans un cahier, ils les ont mis sur l'ordinateur. Et puis un jour, il y a quelqu'un qui a dit : « Mais ça sert à quoi ? »

Rationaliser le réseau conduisait à un problème de sélection de variables, sujet que je développais. Cela a permis à plusieurs étudiants d'avoir des bourses financées par Elf Aquitaine.

C'étaient les débuts de l'informatique : quand on voulait faire un calcul, on écrivait son programme sur du papier, il y avait des personnels qui le perforaient dans des cartes. On allait porter son paquet de cartes perforées aux techniciens et on retournait une demi-journée après ou une journée après, on voyait le résultat sur du papier, quand il y en avait un.

Jusqu'aux années 1975, quand un biologiste par exemple faisait une régression, il mettait le programme de régression dans sa thèse. Moi, quand j'ai écrit ma thèse de troisième cycle sur l'ACP, il y avait le programme d'ACP dans ma thèse. C'est impensable à l'heure actuelle sous cette forme papier mais à ce moment-là, c'était nécessaire parce que si tu n'avais pas le programme, tu ne pouvais pas utiliser la méthode et cela faisait partie de la preuve que tu maîtrisais la méthode. Cela étant, on demande maintenant de fournir le code R pour s'assurer de la reproductibilité des résultats.

Le département informatique de l'IUT a été un des premiers créés avec ceux de Grenoble et de Paris, parce qu'il y avait à Montpellier l'usine IBM qui fabriquait les ordinateurs 360. Au début quelques ingénieurs d'IBM participaient à l'enseignement. Ensuite s'est créé un centre associé du CNAM avec une filière informatique où les diplômés de l'IUT qui étaient embauchés chez IBM ou dans des entreprises à Montpellier, venaient préparer le titre d'ingénieur du CNAM. Il n'y avait pas de visioconférence à ce moment-là. Paul Namian[8] faisait un cours au CNAM à Paris qui était filmé, on recevait les bobines, on projetait les cours et on discutait avec les étudiants. C'est comme cela que j'ai vécu la naissance de l'informatique à Montpellier.

Au début, la théorie c'était que quand on créait un département d'IUT, on fermait un BTS. Par exemple, quand on a créé le département informatique, on a fermé un BTS mécanographie. Quand on a créé le département gestion des entreprises, on a fermé un BTS de secrétariat. Mais après, ça s'est arrêté. Les lycées techniques ont voulu garder leurs BTS.

8. Premier professeur d'informatique au CNAM

© Société Française de Statistique (SFdS), Octobre/October 2022

Pendant un ou deux ans, je suis intervenu dans un certificat de licence qui s'appelait « Analyse numérique », où je présentais des méthodes statistiques aux côtés d'un professeur de classes préparatoires qui présentait les méthodes de calcul numérique. Il y a eu une volonté portée par la direction de la faculté (Bernard Charles et Jean Falguerettes) de développer les mathématiques appliquées qui n'existaient pas à Montpellier. Comme il n'y avait personne du côté analyse numérique, on a développé la statistique.

Mais j'étais très isolé à Montpellier. Bernard Charles et Jean Falguerettes me disaient : « Il faut que tu fasses des stats », mais j'étais seul. Il n'y avait pas d'équipe, ni de séminaire ; chacun travaillait dans son coin.

L'AFCET (Association Française pour la Cybernétique Économique et Technique)[9] a joué un rôle important en me permettant de rencontrer des statisticiens, des numériciens et des électroniciens : c'était le seul endroit où on entendait de tout, il y avait des informaticiens d'entreprise, qui parlaient de Cobol, d'analyse au sens de l'informatique de gestion, d'aide à la décision, de recherche opérationnelle, etc. Tous ceux et celles qui commençaient à se confronter à l'ordinateur, que ce soit pour le construire ou l'utiliser, s'y retrouvaient[10]. J'y ai également rencontré Paul-Louis Hennequin[11].

GS : Ton nom est associé à un opérateur et à un coefficient. Quelles en furent la genèse ? Peux-tu en rappeler l'usage ?

YE : Mes interlocuteurs savaient que j'étais capable de faire des ACP et ils arrivaient avec leurs données. Les expérimentateurs avaient l'habitude de travailler avec certaines variables, mais ils disaient : « On pourrait en créer d'autres » ou bien « Celle-là, je pense qu'elle apporte la même information que l'autre ». Or, comment répondre mathématiquement à cette interrogation ? La seule solution, c'est que tel paquet de variables amène à telle représentation d'individus. Tel autre paquet de variables amène à telle autre représentation d'individus.

L'idée a été de dire : « Ce qui est caractéristique d'un ensemble de variables, c'est la manière dont elles vont nous permettre de représenter les individus ». Or, qu'est-ce qui fait la représentation des individus ? Ce sont les composantes principales qui sont les vecteurs propres de l'opérateur des produits scalaires entre les individus, noté **W**. On calculait toujours la matrice de variance parce qu'elle a une dimension plus restreinte. Mais ce qui fait la représentation des individus, ce sont les valeurs et vecteurs propres de **W**.

On connaissait déjà un autre opérateur : l'opérateur de projection de la régression. Mais l'ennui de l'opérateur de projection, c'est que s'il engendre l'espace vectoriel, il détruit les distances et ne génère pas la représentation des individus. C'est comme cela que j'ai été amené à travailler sur ce nouvel opérateur.

On retrouve le fait qu'en analyse des données on s'intéresse plus aux individus qu'aux relations entre variables. En statistique classique les individus ne sont là que pour apporter une information sur les variables. Alors que là, la perspective est différente : ce sont les variables qui apportent de l'information sur les individus.

Les types de problèmes étaient également différents : les chercheurs du CEPE étaient toujours capables d'inventer une variable de plus pour décrire le sol. « On a pris l'humidité à 10 cm, est-

9. Créée en 1968, l'AFCET fut très active et influente mais fit faillite et disparut en 1998 après des investissements hasardeux.
10. C'est seulement depuis 1984 qu'existe une section informatique au CNU.
11. Professeur à Clermont-Ferrand (1930-...) ; joua un grand rôle dans la diffusion des probabilités et de la statistique. Il est un des fondateurs de l'école d'été de Saint-Flour.

© Société Française de Statistique (SFdS), Octobre/October 2022

ce que tu veux que je mette aussi l'humidité à 20 cm ?, etc. ». Il s'agissait de faire des tirages dans l'espace des variables, mais comment faire ?

J'avais rencontré dans mes lectures un texte dans lequel H. Hotelling s'interrogeait sur la convergence des valeurs propres et des vecteurs propres de la matrice de variance de groupes de variables tirées au hasard dans une population infinie de variables[12]. Plus tard Mahalanobis, qui mesurait des crânes, du temps où il faisait de la comparaison des castes, écrivait en substance : « Je prends une mesure, puis je tourne d'un degré, il y a bien un moment où il y a trop de variables, où les nouvelles variables n'apportent rien ».

Ma thèse d'état a consisté à essayer de donner un formalisme qui permettait de parler de ce problème et de trouver une solution : comment tirer dans un espace qui est le produit d'un espace aléatoire sur les variables et d'un espace aléatoire sur les individus ? Après, il fallait montrer qu'il y avait des convergences, ce qu'on fait classiquement en mathématiques. Et l'opérateur est venu comme ça. Mais vraiment, ma motivation c'était la comparaison d'ensembles de variables.

Je suis persuadé qu'il y a encore une réflexion à faire sur les familles de variables. Pour les individus, on dit : « Ils sont tirés de manière indépendante », mais ça n'empêche pas qu'ils peuvent se ressembler. L'idée c'est que les variables, c'est pareil, elles peuvent être tirées de manière indépendante, mais elles peuvent se ressembler, elles peuvent être corrélées. Dans sa thèse Paulo Gomes[13] se demandait : « Quand on norme les variables, elles sont toutes sur la sphère unité de $\mathbb{R}^n$, quelle est leur distribution ? ». El Mostafa Qannari[14] a travaillé là-dessus aussi : au lieu de prendre une variable, il prend le projecteur associé à la variable et l'ensemble des projecteurs se trouve sur une autre sphère. Il y a toujours des gens qui s'y intéressent comme Xavier Bry[15] avec la distribution de von Mises-Fisher.

Puisque toutes les méthodes d'analyse statistique multivariables consistent à calculer des valeurs propres et des vecteurs propres, cela veut dire que l'on peut expliciter l'opérateur associé à chaque méthode : l'analyse discriminante, l'analyse canonique et l'analyse des correspondances. Dans chaque cas, ce sont des choix de données et une manière de les utiliser, c'est-à-dire de pondérer les individus et de prendre une métrique sur les variables.

Le coefficient RV est venu un peu après avec l'idée de pouvoir comparer ces opérateurs. Mathématiquement ils sont dans des espaces de Hilbert ; si on prend un produit scalaire, on tombe automatiquement sur la covv (ou covariance vectorielle) qui est le produit scalaire des deux opérateurs. Et en divisant par les normes, on a le RV.

La première étape a donc été de définir l'opérateur, parce que pour comparer des groupes de variables, on ne peut que comparer la manière dont on représente les individus. Le produit scalaire entre opérateurs quantifie la ressemblance des représentations des individus. STATIS[16] puis le triplet **X**, **Q**, **D** sont des enrichissements ultérieurs. Cela parait simple maintenant et on se demande pourquoi on n'y avait pas pensé plus tôt ! Notons qu'on a montré plus tard qu'en prenant des matrices **X**, des métriques **Q** et des poids **D** particuliers, RV s'identifie aux différentes mesures de liaisons entre variables.

12. Hotelling H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, vol. 24, n° 7, pp. 498-520, https://psycnet.apa.org/doi/10.1037/h0070888
13. Gomes P. (1987), *Distribution de Bingham sur la n-sphère : une nouvelle approche de l'analyse factorielle*, Thèse, Montpellier.
14. Qannari E. M., E. Vigneau et Ph. Courcoux (1998), Une nouvelle distance entre variables. Application en classification, *Revue de Statistique Appliquée*, vol. 46, n° 2, pp. 21-32, http://www.numdam.org/item/RSA_1998__46_2_21_0/
15. Bry X. and L. Cucala (2018), Classifying variable-structures: a general framework, *arXiv preprint*, arXiv:1804.08901.
16. Structuration des Tableaux À Trois Indices de la Statistique

© Société Française de Statistique (SFdS), Octobre/October 2022

GS : Parlons donc de STATIS.

YE : À cette époque-là, il y avait tout un courant sur la représentation des matrices de similarité ou de dissimilarité. Joseph Kruskal et Douglas Carroll parlaient de *multidimensional scaling*. Une fois obtenus les opérateurs et le coefficient RV, on pouvait définir une matrice de similarité entre plusieurs tableaux par la matrice des RV entre opérateurs. Elle avait toutes les bonnes qualités, elle était définie positive. Ce que l'on a appelé l'*interstructure* – c'est-à-dire la comparaison des opérateurs entre eux – consistait à calculer les valeurs propres et les vecteurs propres de cette matrice de similarité.

Comme tous les coefficients RV sont positifs, le premier vecteur propre de la matrice de similarité pouvait être choisi avec tous ses coefficients positifs. Prendre la combinaison linéaire des opérateurs définis par ce premier vecteur propre donnait encore un opérateur positif. On pouvait le diagonaliser, ce qui donne le *compromis*. Par des méthodes tout à fait analogues à ce qu'on fait en ACP en projetant des points supplémentaires, on projetait les opérateurs et on avait ce que l'on a appelé l'*intrastructure*. La démarche était assez naturelle : finalement, c'est la même chose que l'ACP, mis à part qu'au lieu de parler de variables, on parle d'opérateurs ; et au lieu de parler de corrélations, on parle de RV.

Toute cette construction était prête quand je suis revenu du Canada, ce qui a conduit à la thèse de Henri l'Hermier des Plantes[17].

GS : Et le schéma de dualité, si typique d'une certaine approche française[18] ?

YE : Quand j'ai rencontré Jean-Pierre Pagès et Francis Cailliez en 1972, ils travaillaient sur ce qui deviendra ensuite leur livre[19]. Ils dessinaient déjà leur schéma de dualité. Ils notaient « E » pour l'espace des individus et « F » pour l'espace des variables. Les variables sont représentées dans F, mais peuvent aussi être représentées comme une application dans E, c'est-à-dire des éléments du dual E* et la matrice **X** permet de passer de E* à F. Aux individus qui sont des points de E, on peut aussi associer des éléments du dual F*. La matrice transposée **X'** permet de passer de F* à E. Du côté de E, il y avait la matrice de variance et la notion de choix d'une métrique **Q**, donc un opérateur **VQ** de E dans lui-même. Mais de l'autre côté, ils ne savaient pas quoi mettre.

Je suis arrivé en disant : « Mais ce qu'on fait avec **VQ** d'un côté, c'est la même chose de l'autre côté avec **WD** », ce qui a donc trouvé automatiquement sa place dans leur schéma.

On peut toujours se demander si cette lecture très algébrique est vraiment utile, mais l'intérêt que j'y trouvais, était de faire apparaître vraiment les choix effectués. Il y a le tableau de données **X**, avec la métrique **Q** que l'on utilise pour calculer les distances entre individus, et **D**, la matrice des poids que l'on donne aux individus pour calculer les variances et les covariances. Tout est explicite. Alors qu'en général, **D** c'est $(1/n)\mathbf{I}_n$ et on n'en parle plus.

Henri Caussinus et Yves Aragon[20] ont traité du cas où les individus étaient auto corrélés – dans l'espace ou dans le temps – avec une métrique **N** à la place de **D** qui décorrélait les individus.

Les utilisateurs, par exemple, réduisaient leurs variables, donc les données changent, ce ne sont plus celles qui ont été observées, ce qui signifie que l'on utilise une métrique particulière

17. L'Hermier des Plantes, H. (1976), *Structuration des tableaux à trois indices de la statistique*, Thèse, Montpellier.
18. Pour une présentation récente en anglais : De la Cruz O. and S. Holmes (2011), The duality diagram in data analysis: Examples of modern applications, *The Annals of Applied Statistics*, vol. 5, n° 4, pp. 2266-2277.
19. Cailliez F. et J.-P. Pagès (1976), *Introduction à l'analyse des données*, SMASH, Paris.
20. Aragon Y. et H. Caussinus (1980), Une analyse en composantes principales pour des unités statistiques corrélées, *in* E. Diday *et al.* (eds.), *Data Analysis and Informatics*, pp. 121-131, North Holland.

© Société Française de Statistique (SFdS), Octobre/October 2022

sur le tableau d'origine.

Au début pour moi, c'était purement un outil de dialogue avec les expérimentateurs qui m'apportaient des données, pour bien expliciter les choix qu'ils faisaient. Et, en fait, ça s'est avéré beaucoup plus riche puisqu'après, on peut voir l'analyse discriminante, l'analyse des correspondances simple ou multiple, comme des analyses en composantes principales de triplets particuliers, etc.

Et ça a été à l'origine de l'ACPVI (Analyse en composantes principales avec variables instrumentales). Si j'ai deux tableaux **X** et **Y** décrivant les mêmes individus munis des mêmes poids, quelle métrique mettre sur **X** pour que les ACP que l'on peut faire sur **X** ressemblent autant que possible à celles que l'on a faites sur **Y** ? On s'aperçoit alors que c'est une autre lecture de la régression, obtenue par la représentation des produits scalaires entre individus et non par la reconstruction de chaque donnée.

J'avais suggéré à John Gower de présenter l'analyse procustéenne dans ce schéma mais je n'ai pas continué ma réflexion là-dessus. Pour lui, c'étaient plutôt des successions de rotations, d'homothéties et de translations.

J'ai repensé récemment à l'analyse canonique : si on prend le projecteur associé à **Y**, que l'on cherche la métrique à mettre sur **X** pour reconstruire ce projecteur, et que l'on fait de même dans l'autre sens, on arrive aux équations de l'analyse canonique.

GS : Comment définis-tu la spécificité de l'école française d'analyse des données ?

YE : Dans cette période de 1965 à 1971, les statisticiens, où qu'ils soient, étaient confrontés au fait qu'il y avait des ordinateurs, que des expérimentateurs, des entreprises et des laboratoires de recherche engrangeaient des données et qu'à un moment donné, on leur a dit : « Sur ces données, il faudrait quand même que vous nous sortiez de l'information ».

Ce qui a fait la spécificité des Français – et le livre de Pagès et Cailliez me paraît un bon exemple – c'est qu'on avait une formation en algèbre qui nous permettait d'écrire de manière précise les manipulations mathématiques qu'on faisait. Dans le livre de Harman déjà cité on ne trouve aucune matrice. Ce sont toujours des sommes de a_i, a_{ij}, x_{ij}...

De l'autre côté de l'Atlantique, John Tukey proposait plutôt des extensions de ce qu'on appelle les méthodes descriptives. Au lieu de faire un histogramme, on faisait un histogramme avec des couleurs différentes. Les Américains jouaient aussi avec les possibilités que donnait l'ordinateur pour enrichir les analyses avec le Bootstrap ou le Jackknife : au lieu de faire l'analyse une seule fois avec l'échantillon, on ré-échantillonne dans l'échantillon, on fait des permutations dans l'échantillon. Ils n'hésitaient pas à faire des calculs multiples que nous n'osions pas faire encore. Mais cela restait très unidimensionnel alors que les Français ont essayé de voir ça globalement avec toujours ce souci de formaliser mathématiquement et d'arriver à quelque chose qui se calcule.

Il y a un article de C. R. Rao[21] dans Sankhya que j'ai cité plusieurs fois, qui explique l'importance de faire du multidimensionnel. Si C. R. Rao en a vu la nécessité, cela veut dire qu'à cette époque-là, ce n'était pas si évident. Cet article prend la peine d'expliquer qu'on doit faire du multidimensionnel quand on a plusieurs variables, et que l'on n'a pas le droit de les étudier les

21. Rao C. R. (1964), The use and interpretation of principal component analysis in applied research, *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 26, pp. 329-358.

© Société Française de Statistique (SFdS), Octobre/October 2022

unes à côté des autres.

J'ai connu beaucoup plus tard les travaux japonais qui étaient complètement indépendants, à l'occasion des congrès qu'Edwin Diday organisait à Versailles et qui ont eu un rôle très important. C'est là qu'on a vu arriver pour la première fois Chikio Hayashi et Kameo Matusita, et, plus tard, Noboru Ohsumi. Ces congrès m'ont également permis de découvrir John Gower, John Nelder, Alfredo Rizzi et d'établir des liens forts avec beaucoup d'échanges, en particulier avec les Italiens qui ont invité certains Français à leur réunion annuelle en Italie et l'année suivante on invitait des Italiens.

GS : Et Jean-Paul Benzécri ?

YE : J'ai croisé une fois Jean-Paul Benzécri à Toulouse dans un congrès de l'AFCET où j'avais présenté une analyse discriminante que j'avais faite avec des médecins de Montpellier. Je crois que par sa personnalité, par le fait qu'il enseignait à Paris, son aura, ses travaux, il a vraiment ouvert la porte à l'analyse des données en France. En plus, il a eu de nombreux étudiants, certains très brillants. Ils avaient les programmes qui permettaient de mettre en œuvre les méthodes, l'analyse des correspondances ou les méthodes de classification.

La difficulté, c'est qu'il n'a jamais fait partie vraiment d'un groupe, il ne s'est jamais intégré dans une communauté ; je pense qu'on ne l'a jamais vu aux réunions de l'ASU. Il était vraiment à part. Son comportement ou sa personnalité ne facilitaient pas les contacts.

GS : Quel rôle jouent les modèles en statistique en général et dans ta conception de la statistique ?

YE : Il faut préciser d'abord de quel modèle on parle : si on parle des lois comme la loi normale ou comme la loi binomiale, comme des modèles possibles d'aléatoire, c'est utile dans la mesure où si on a de bonnes raisons de penser qu'un phénomène peut être décrit par ce genre de modèles, cela donne beaucoup d'informations sur le phénomène. Je dirais que ces modèles sont des outils à la fois de dialogue, de description. Si tu sais qu'un modèle marche bien, c'est un élément de langage. Si tu dis à quelqu'un : « Ma loi suit une loi normale », tous les statisticiens ont compris.

Après, il y a les modèles du type de ceux qu'on a en biologie, comme les modèles de croissance. Jean-Marie Legay à Lyon a pas mal écrit sur ces sujets-là[22]. Il peut y avoir des ambiguïtés sur les objectifs : veut-on reconstruire les résultats, ou veut-on reconstruire le phénomène ? Ce n'est pas toujours clair chez les expérimentateurs. On peut vouloir faire un modèle qui va effectivement reconstruire le résultat, mais sans avoir l'ambition de dire : « Ça explique que le phénomène fonctionne comme ça »[23].

Interpréter des modèles à des niveaux très simples n'est pas toujours évident. Par exemple, après une régression multiple, certains utilisateurs constatent que « telle variable explicative a un signe négatif » et vont expliquer que quand la variable va croître alors la variable expliquée va décroître, sans tenir compte que cette variable est corrélée avec d'autres et que donc son effet ne peut pas être interprété indépendamment des autres variables.

Dans certains modèles, leurs concepteurs ont mis beaucoup d'*a priori* c'est-à-dire de savoirs sur le sujet, sans s'interroger sur la robustesse justement de l'impact de ce qu'ils ont mis dans

22. Legay J.-M. (1986), Méthodes et modèles dans l'étude des systèmes complexes, *Cahiers de la recherche-développement*, vol. 11, pp. 1-6.
23. Cette démarche se retrouve en apprentissage : « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* » (Vapnik, 2006).

© Société Française de Statistique (SFdS), Octobre/October 2022

le modèle. Je pense à certains énormes modèles qui me font un peu peur comme les modèles climatiques, les modèles de production agronomique, etc. Je m'interroge si tous leurs utilisateurs sont bien conscients de la fragilité liée au fait qu'ils y ont mis beaucoup de leur propre savoir ?

Par ailleurs, les prévisions météorologiques marchent relativement bien. Mais elles reposent sur des données énormes. Guy Der Mégréditchian[24] n'est plus là pour nous dire comment ils font, mais je pense qu'ils reconstruisent plus les résultats que les phénomènes. En plus, ils travaillent beaucoup par comparaison avec des situations semblables.

Finalement, je suis un peu mal placé pour parler des modèles dans la mesure où je me suis surtout intéressé à la statistique exploratoire !

GS : Peux-tu nous parler de tes années canadiennes ?

YE : Comme je le disais plus haut, j'étais très seul à faire de la statistique à l'université de Montpellier, et c'était un souci pour moi. À Montréal, il y avait tous les ans une école d'été de mathématique. En 1968, elle était consacrée à la statistique. Et donc j'y suis allé. Je n'en ai pas un grand souvenir : Lucien Le Cam avait fait un cours, et je ne me souviens plus de qui étaient les autres. Mais j'ai rencontré pour la première fois Pierre Robert qui était alors le directeur du département informatique et recherche opérationnelle de l'Université de Montréal.

Je suis revenu à Montpellier, j'ai continué, j'ai passé ma thèse d'État et je voulais aller ailleurs, pour respirer un peu et pour rencontrer des statisticiens. Et il se trouve que dans l'hiver 1970-1971, Pierre Robert est passé en France et son message était : « J'ai des postes au département IRO à Montréal et je cherche des gens qui voudraient venir quelques années ».
J'ai dit « banco », et nous sommes partis avec nos trois enfants dont le plus jeune qui avait un mois.

Partis pour un an, nous sommes restés en fait deux années scolaires.

J'étais donc au département « Informatique et recherche opérationnelle » parce que les probabilistes se trouvaient au département de mathématiques, et les statisticiens au département IRO avec Pierre Robillard, Robert Cléroux, Roch Roy, plus des numériciens et des informaticiens.

Les probabilistes étaient des mathématiciens qui n'étaient pas intéressés par le traitement des données, mais par le développement de modèles. De l'autre côté, les statisticiens étaient confrontés à des données réelles, à la nécessité de les stocker, de les manipuler et d'en dire quelque chose d'opérationnel, de faire des prévisions.

Chaque semaine il y avait un séminaire au département et chaque mois un séminaire commun entre l'Université McGill et l'Université de Montréal. J'ai vu passer tous les grands noms de la statistique et tous les chapitres de la statistique que je ne connaissais pas. Pour moi, ça a été une révélation de découvrir des aspects de la discipline auxquels je ne m'étais jamais confronté et de pouvoir en parler facilement avec des collègues. Du point de vue recherche, ça a été très important, un grand moment d'ouverture.

C'est à Montréal que j'ai rédigé l'article pour Biometrics à partir de mes travaux de thèse : les opérateurs et le RV. Et j'ai eu la chance qu'un des lecteurs, J. Mosiman du National Health Center à Washington, se soit intéressé à cet article ; il a demandé à me voir et l'article a été publié.

24. Statisticien à Météo France (1933-1990)

© Société Française de Statistique (SFdS), Octobre/October 2022

Outre cet aspect recherche et ouverture sur l'ensemble de la statistique, je me souviens de Pierre Robillard qui m'a dit : « Tu vas certainement faire un polycopié comme tous les Français ». Et je lui ai dit : « Oui, et comment tu fais, toi ? » Et lui : « Moi, je repère le dernier livre paru sur le sujet que je dois enseigner, je dis aux étudiants de l'acheter et puis on le lit ensemble ».

Ça voulait dire deux choses :

- En France, les étudiants faisaient grève pour avoir des polycopiés gratuits, et au Canada on leur demandait d'acheter des livres et ils trouvaient cela normal.
- Mais en même temps, cela montrait une position très différente de l'enseignant vis-à-vis du savoir. En France, l'enseignant apporte le savoir. Là-bas, le savoir est dans le livre, l'enseignant est simplement celui qui a quelques années d'avance sur les étudiants et qui peut les aider à mieux le comprendre.

Pierre Robillard ajouta : « Quand il y a un chapitre qui ne me plaît pas ou que je trouve incomplet, à ce moment-là, je complète ». Donc, c'était vraiment une autre manière de vivre le rapport enseignant/enseigné. J'avais trouvé ça intéressant du point de vue pédagogique.

Je crains que cela n'ait pas beaucoup changé en France où chaque enseignant a toujours envie d'avoir son propre document et n'aime pas utiliser les documents des autres. Alors que là-bas, c'était le contraire.

L'attitude des étudiants était différente également. Avec le système majeure/mineure, certains venaient suivre un cours dans leur mineure pour avoir un éclairage sur des aspects particuliers. Ils me demandaient par exemple : « Allez-vous parler de tel modèle d'analyse de variance ? Parce que là où je travaille, on l'utilise et je veux le comprendre ». Ce n'était pas du tout du bachotage, alors que les étudiants français voulaient juste réussir l'examen à la fin de l'année, quel que soit le contenu du cours.

Il y avait une autre différence : les étudiants qui arrivaient à l'université étaient passés par le Cégep[25]. Ils étaient à peu près du niveau deuxième année de licence avec un peu plus de maturité que ceux qui sortent du bac. Et puis peut-être que tout leur système éducatif est plus axé sur l'autonomie de l'individu.

GS : Comment s'est réalisé le passage du CRIG à l'unité de biométrie ?

YE : Après être revenu de Montréal, nous avons créé en 1973 avec deux collègues de l'IUT, Jean Ferrier qui était informaticien et Robert Reix qui était gestionnaire, le « Centre de recherche en informatique et gestion ». On nous a donné une salle dans le bâtiment K de l'IUT. C'était notre laboratoire ! J'avais aussi des étudiants de l'option Analyse des données du DEA de mathématiques où j'intervenais et quelques maîtres-assistants d'alors ont commencé à travailler avec moi, comme Christine Lavit, Alain Delcamp, Jean-François Durand, Henri Aris.

À Montréal, le département IRO avait institué la pratique suivante : chaque année un professeur de statistique était dispensé d'une partie de son enseignement pour être disponible, pour travailler avec les chercheurs des autres disciplines. Je suis allé voir le président d'alors à Montpellier pour lui proposer cette idée et il m'a répondu : « Ce n'est pas possible chez nous. En plus, vous êtes en poste à l'IUT ».

J'ai malgré tout mis en place avec le centre de calcul un système où le jeudi matin, je présentais

25. Collège d'enseignement général et professionnel, établissement d'enseignement public du Québec où est dispensé le premier niveau de l'enseignement supérieur.

© Société Française de Statistique (SFdS), Octobre/October 2022

une méthode, ensuite un technicien du centre de calcul montrait comment on pouvait la mettre en œuvre ; les participants avaient 15 jours pour utiliser la méthode sur leurs données et on discutait des résultats qu'ils avaient obtenus. Cela a duré un an ou deux à titre tout à fait gratuit pour moi, mais il y avait une forte demande de la part des biologistes de la Faculté des sciences, mais aussi des géographes de Paul Valéry, de chercheurs de l'INRA et enfin de l'environnement qui savaient qu'ils pouvaient venir nous demander de l'aide et on essayait de les former un peu.

C'est alors qu'il y eut la volonté de voir s'installer en un seul lieu le CIRAD, l'IRD, le CEMAGREF pour constituer un centre agronomique méditerranéen appelé ultérieurement « Agropolis ». L'INRA, avec Richard Tomassone qui était alors le directeur du département biométrie, a voulu accompagner ce mouvement et créer à Montpellier un laboratoire de biométrie et m'a proposé d'en prendre la direction pendant au moins huit ans. L'unité de biométrie est née en 1982.

Tomassone disait : « Je mettrai progressivement du personnel » ; l'école d'Agronomie disait : « Je donne les locaux » et les enseignants de mathématiques de l'école en feront partie. Moi, j'avais négocié avec la Faculté des sciences que les gens qui travaillaient avec moi seraient logés dans les locaux donnés par l'école d'Agronomie. Et c'est vrai qu'à ce moment-là, ça nous a donné des bourses, des moyens, une grande visibilité et des possibilités d'action importantes.

Ça a été une grande chance pour moi et pour la statistique à Montpellier. Il y a eu le recrutement de Xavier Millau venu de Toulouse, mais j'étais toujours en poste à l'IUT. En 1987, un poste s'est libéré au département de Mathématiques à la Faculté et j'ai été candidat avec le projet de développer des enseignements de statistique pour les biologistes. Il y avait quelques enseignements de statistique dans les filières de biologie et dans les DEA de biologie. Soit on me sollicitait pour intervenir deux heures dans un DEA de biologie – c'était ridicule –, soit l'enseignement était fait par des gens dont la statistique n'était pas la spécialité. J'ai expliqué au département de mathématiques qu'il fallait à tout prix investir ce domaine, qu'il y avait des postes à prendre ou à gagner, et que la formation des étudiants était en cause. On a alors pu recruter sur l'université parce que dès la création de l'unité de biométrie, il y avait l'idée de créer un DEA de statistique pour des mathématiciens et un DESS de méthodes statistiques appliquées à l'agronomie, l'agroalimentaire et la pharmacie, destiné plutôt à des biologistes.

L'unité de biométrie a donné une visibilité et une crédibilité à ce qu'on faisait, a entraîné des créations de postes, des bourses. On a géré beaucoup de bourses CIFRE dans le domaine de l'agroalimentaire, le champagne, la brasserie entre autres. Cela a ouvert beaucoup de débouchés aux étudiants qui sont passés par là et certains sont rentrés au CIRAD, à l'IRD, à l'INRAE.

Parti pour huit ans, je suis resté douze ans. Alain Berlinet m'a succédé, puis Jean-Pierre Vila et quelques années plus tard l'association avec l'université a cessé. L'unité de biométrie s'est restreinte à l'INRA et à l'école d'Agronomie, dénommée maintenant Institut Agro Montpellier[26]. Mais les liens restent forts, les gens se connaissent, se retrouvent dans les séminaires.

GS : Le colloque organisé en 2019 en ton honneur s'intitulait, « Montpellier, berceau de la Data Science »[27]. Quel regard portes-tu sur les relations entre l'informatique, la science des données et la statistique ? Plus particulièrement, penses-tu que l'informatique fasse progresser la statistique ?

YE : Les moyens de calcul permettent d'imaginer des choses qu'on ne pouvait pas imaginer auparavant. Ne serait-ce que l'estimation par noyau par exemple. Aujourd'hui, c'est courant,

26. L'unité de biométrie est devenue en 2009 l'UMR Mathématiques Informatique et STatistique pour l'Environnement et l'Agronomie (MISTEA).
27. C'est à l'occasion du deuxième séminaire franco-japonais organisé à Montpellier par Yves Escoufier en 1992 qu'est apparue la première occurrence de l'expression Data Science. Les actes du colloque en témoignent : Escoufier Y., B. Fichet, L. Lebart, C. Hayashi, N. Ohsumi, and Y. Baba (Eds.) (1995), *Data Science and Its Applications*, Academic Press, Tokyo, Japan.

© Société Française de Statistique (SFdS), Octobre/October 2022

mais on ne pouvait pas l'imaginer dans les années 1965-1970. Même si la théorie existait, les temps de calcul auraient été rédhibitoires.

C'est comme l'analyse factorielle au sens de Harman. Une des raisons pour lesquelles on mettait les spécificités dans la diagonale, c'était parce que diagonaliser une matrice était très compliqué. On cherchait un premier axe, un deuxième axe et pas plus.

Il est clair que les outils modifient la façon de penser et de réfléchir : le Bootstrap, le Jackknife, c'était inimaginable. Quand on mettait une après-midi pour faire une ACP, on n'allait pas la refaire 50 fois !

Il y a 30 ou 40 ans, on n'avait pas les outils qui nous permettaient d'imaginer de faire ces calculs.

Des chercheurs comme Kruskal, Carroll dont on parlait plus haut, avaient une pratique computationnelle bien plus performante que la nôtre, ce qui les a fait avancer, mais on s'est retrouvé sur le *multidimensional scaling* car en réalité on faisait la même chose.

En fait, c'est le problème de la place des données en statistique. J'ai toujours pensé que la statistique avançait en se confrontant à des données nouvelles qui nous amènent à créer des outils nouveaux. Et aujourd'hui, il n'en manque pas des données nouvelles et on a besoin de beaucoup de mathématiques et de mathématiques difficiles avec des algorithmes énormes que l'on ne pouvait même pas imaginer.

Après, il est assez naturel de prendre du recul pour justifier mathématiquement et pour construire informatiquement les solutions. C'est peut-être plus dur d'ailleurs.

GS : Comment réagis-tu au fait que des informaticiens spécialisés en apprentissage redécouvrent les vertus de l'ACP et d'autres méthodes familières aux statisticiens ?

YE : Je commencerai par dire que s'ils redécouvrent la méthode, c'est que la méthode est utile. Ils y arrivent par d'autres chemins, et cela signifie que ce que l'on cherche à obtenir dans une ACP les intéresse aussi. C'est donc un bon point pour la méthode. Il reste que « la redécouverte » peut interroger sur la formation donnée et sur les équipes de recherche constituées.

GS : La statistique se dissout-elle dans la *data science* ?

YE : Tout dépend de ce qu'on appelle la *data science*. Je pense que dans la *data science*, il y a à la fois la collecte des données, leur stockage, leur manipulation. Ensuite, il y a les traitements, donc de la statistique et aussi des approches « intelligence artificielle ».

Ma crainte serait que la *data science* ne devienne le domaine d'exercice de personnes qui n'ont aucune idée en statistique, qui pourraient seulement manipuler des données car on peut toujours en sortir quelque chose. La statistique, parce que c'est le domaine de l'aléatoire, fait germer le doute : « J'ai trouvé ce résultat, quelle confiance je peux lui faire ? ». Et c'est là que la statistique est importante.

Un *data scientist* pourrait être quelqu'un qui connaît la statistique et qui a une bonne pratique informatique. Alors, est-ce que l'inverse existe ? Quelqu'un qui connaît bien l'informatique et qui a une bonne pratique statistique ?

Cela pose la question de la formation. L'image que j'ai, c'est que dans les métiers du traitement de données, les gens arrivent soit par des filières où ils ont fait des mathématiques, des statistiques, des probabilités, soit par des filières où ils ont fait de l'informatique, et qu'il y a des

© Société Française de Statistique (SFdS), Octobre/October 2022

filières informatiques – du moins certaines – qui sont très loin des filières de mathématiques, avec des contenus très différents. Qu'enseigne-t-on de la statistique dans des filières *data science* où on parle surtout informatique ?

Cette diversité d'approches peut être un enrichissement mais il faut que ces approches-là se parlent. L'avantage d'une formation statistique, c'est quand même d'avoir une notion de l'aléatoire. Il est important de pouvoir dire : « Attention, cette méthode, sur ces données, elle nous donne ces résultats, mais sur des données un peu différentes elle peut donner des résultats un peu différents, voire très différents ».

GS : Tu as fait de l'ASU (l'ancêtre de la SFdS) une société à la fois savante et conviviale. Comment cela s'est-il passé ?

YE : Juste avant que je ne parte à Montréal, Bernard Charles avait fait en sorte que se crée à Montpellier une MIAGE – Maîtrise informatique appliquée à la gestion – dans laquelle je donnais un cours de statistique. Si bien qu'en février 1972, je suis revenu 15 jours pour faire mon cours de statistique à la MIAGE.

Pendant ce séjour, Richard Tomassone m'a invité à donner un séminaire à Orsay. J'y ai fait la connaissance de Jean-Pierre Pagès et de son schéma de dualité, comme je le disais plus haut. Jean-Pierre Pagès m'a fait connaître Georges Morlat qui animait les premières réunions de l'ASU avec des universitaires de province comme Marie-Jeanne Laurent-Duhamel, Henri Caussinus et Michel Depaix qui cherchaient à faire savoir que la Statistique universitaire française ne se limitait pas à l'ISUP.

C'est ainsi que j'ai découvert cette association qui était en train de naître[28]. Deux ans après mon retour de Montréal, j'ai organisé en 1975 une réunion de l'ASU à Montpellier où pour la première fois il y a eu des exposés scientifiques[29]. Nous étions à peine 40.

Plusieurs congrès de l'ASU ont eu lieu ensuite à Montpellier, dont un à la Grande Motte en 1984.

GS : La crise sanitaire a révélé l'inculture scientifique de bon nombre de nos gouvernants, communicants, citoyens. Aurais-tu des pistes d'amélioration de ce qu'on appelle la littératie scientifique et plus particulièrement en statistique ?

YE : Quand j'étais président d'université, j'ai eu l'occasion de rencontrer un certain nombre d'hommes politiques. On se présente et quand je dis : « Je suis professeur de mathématiques », le plus souvent, la réponse était : « Oh ! Moi, vous savez, les mathématiques et les sciences, ce n'est pas mon fort ».

Que faut-il en penser ? Est-ce qu'en France, on a trop fait des mathématiques l'objet de sélection pour les écoles ? Je ne sais pas. Mais il y a manifestement une attitude vis-à-vis des mathématiques qui est regrettable. On considère que c'est la discipline clé et en même temps, ils disent presque en s'en vantant : « Moi, ça n'a jamais été mon fort ».

Qu'est-ce que ça veut dire ? Je crois qu'aujourd'hui, on ne peut plus vouloir dominer l'ensemble des disciplines, c'est très compliqué. Alors, à quel niveau, jusqu'où faut-il développer la culture scientifique ?

Récemment, je parlais avec un ami qui s'élevait contre le fait que des élèves de lycée, en ce

28. Les statuts de l'*Association des Statisticiens Universitaires* ont été déposés à Paris en février 1971.
29. C'est aussi à l'occasion de ces 7èmes Journées de Statistique que fut établie la tradition du banquet avec animation musicale.

© Société Française de Statistique (SFdS), Octobre/October 2022

moment, peuvent ne plus faire de maths en terminale. Il disait : « Mais tous les ingénieurs ont besoin de connaître les mathématiques, l'industrie a besoin de gens qui connaissent les mathématiques ». Bien sûr, mais à côté de cela, il y a la psycho, la sociologie, la communication, les relations humaines… Je m'interroge beaucoup sur le niveau de mathématiques qu'il est nécessaire d'y avoir.

Le problème essentiel, me semble-t-il, c'est que chacun sache où s'arrête son propre savoir et soit respectueux du savoir des autres. Après, est-ce qu'on pourrait me reprocher à moi par exemple, de ne pas avoir une culture suffisante en chimie ?

Je trouve que c'est une question très difficile. Quand je participais de près à l'Institut International de Statistique (IIS), cette question de littératie statistique est revenue plusieurs fois.

J'avais fait un exposé dans une réunion de l'IIS que j'avais intitulé : « Que faut-il enseigner et à qui ? ». J'avais présenté l'image ci-dessous qui donne différentes écritures de la moyenne telles qu'on peut les trouver dans les livres.

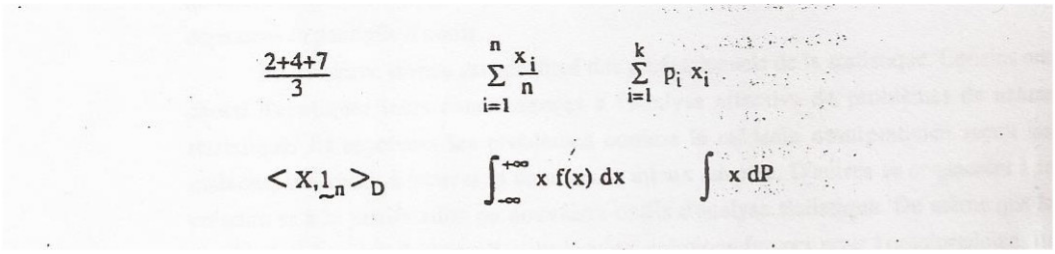


Figure 1 – *Différentes écritures de la moyenne*

On utilise l'une ou l'autre de ces écritures selon à qui on s'adresse. Et donc, cela veut dire qu'il y a des accès différents à des niveaux différents de compétences. Donc quel est le savoir minimal qu'il faut avoir ? Et ce savoir minimal est-il stable ? C'est vrai qu'aujourd'hui, avec tous les moyens de communication, on est abreuvé de chiffres : faut-il insister sur leur acquisition, leur interprétation ? Qu'y a-t-il derrière ? Qui valide une information ?

Comment faire pour intéresser les jeunes à la statistique ? Sylvette Maury qui dirigeait l'Institut de recherche sur l'enseignement des mathématiques (IREM) à Montpellier avait organisé en 1989 un concours pour les lycéens intitulé « À vos Stats »[30] parrainé par l'Insee. L'idée était de faire découvrir l'intérêt qu'il pouvait y avoir à savoir manipuler un peu les outils statistiques.

Il me semble qu'au-delà des techniques, ce qui est important avec la statistique, c'est de sensibiliser les gens à l'aléatoire. Qu'ils sachent qu'un phénomène va avoir plusieurs résultats possibles et que c'est naturel. Quand j'enseignais dans un certificat de biologie, en guise d'introduction, je prenais une craie et je leur disais : « Si je la lâche, elle tombe, c'est un phénomène qui n'est pas aléatoire. Mais en combien de morceaux va-t-elle se casser ? Ça, c'est aléatoire et vous ne pouvez pas me le dire avant, n'est-ce pas ? ».

Il faut prendre conscience de l'aléatoire, puis le décrire et le modéliser, et ensuite selon le nombre d'années d'étude, on va enrichir les moyens de description et de modélisation. Mais cette notion qu'il n'y a pas qu'une issue possible, c'est fondamental.

Note de la rédaction : L'ensemble des publications d'Yves Escoufier est disponible à l'adresse : https://imag.umontpellier.fr/YvesEscoufier/

30. Le concours *À vos Stats* a eu lieu jusqu'en 2004. Sous une forme différente, Eurostat et l'INSEE en France organisent tous les ans depuis 2018 une compétition européenne pour les jeunes collégiens et lycéens.

© Société Française de Statistique (SFdS), Octobre/October 2022