Dossier : Statistique et probabilités

La géométrie au service de la statistique

par Jean-Louis Piednoir
avec la collaboration de Jean Aymes

Le présent article a son origine dans une séance de formation organisée dans l'académie de Limoges à la demande de nombreux collègues enseignant en particulier dans les sections ES, en section de techniciens supérieurs du secteur tertiaire ou en D.E.C.F. (diplome d'études comptables et financières). La démarche statistique n'a pas fait partie, en règle générale, de la formation des professeurs de mathématiques et, de ce fait, les collègues se posent de nombreuses questions sur la pertinence de leur enseignement, sur certaines nouveautés dans les programmes comme l'introduction d'un important chapitre de géométrie dans l'espace en Première et Terminale ES (option et spécialité mathématique). Après une présentation générale possible de la démarche statistique, on montrera l'importance de la géométrie euclidienne pour la « fabrication » d'indicateurs statistiques classiques.

A. La démarche statistique

Dans les programmes, les manuels, les divers ouvrages que l'on peut consulter, le mot « statistique » apparaît dans des sens divers. Les situations étudiées dans les chapitres statistiques ont certes un air de famille, mais rien ne met en lumière l'unité de la démarche. On proposera ci-dessous une première présentation unifiée de la méthode statistique. Il s'agit d'un point de vue : celui-ci n'est pas unique.

1. Une première modélisation

Toutes les situations concrètes susceptibles d'une approche statistique sont caractérisées par la structure suivante : on veut mesurer ou au moins comparer sur un objet d'étude diverses caractéristiques. Cet objet d'étude se présente comme un collectif (on dit aussi une population) composé d'éléments (ou individus). Il n'est pas possible, ou bien cela n'a pas de sens, d'effectuer une mesure directe sur le collectif. Par contre, il est possible de procéder à des mesures sur des éléments du collectif. Mais ces mesures sont différentes les unes des autres. Il existe sur les éléments une certaine variabilité.

Prenons quelques exemples pour illustrer cette description :

– Une machine fabrique des pièces en grande série. La production est-elle de bonne qualité ? Sur une pièce donnée, par un essai destructif, on est capable de mesurer sa résistance, à l'usure par exemple.

– Une province donnée de notre beau pays produit des « bléds »[1]. Quelle est sa richesse ? On peut, sur vingt années consécutives, connaître la production.

– Ce commerçant a une boutique prospère. Quelle est sa vente journalière ? Il doit passer sa commande pour la semaine prochaine, connaissant les ventes hebdomadaires passées.

– On veut expérimenter un médicament contre l'hypertension ; avant de le tester, il faut étudier la tension d'une population.

– Quel est le revenu des foyers français ?

On peut formaliser ces situations. On a un collectif P composé d'éléments. On examine les individus ou une partie E d'entre-eux. $E \subset P$ est appelé échantillon. Il peut être impossible d'examiner toute la population : dans le cas des essais destructifs ou celui des malades, cela est évident. Dans d'autres cas, comme celui du revenu des français, cela est théoriquement possible, mais d'un coût excessif.

Sur chaque individu de E, on effectue une mesure ou une observation. Soit X l'ensemble dans lequel les mesures sont prises, x le caractère étudié, x_i la valeur prise par le caractère x pour l'individu i. Si X est un ensemble fini sans structure, x est alors un caractère nominal ; ainsi en est-il de la couleur des cheveux d'une population, si tant est qu'on se mette d'accord sur le nombre de couleurs et leur définition précise. Si $X = \mathbf{R}^k$, x est formé de k caractères numériques. On dispose ainsi d'un ensemble de n mesures $\{x_1, x_2, \ldots, x_n\}$ où n est le nombre d'individus de l'échantillon. Mais les individus ne nous intéressent pas : c'est la population que l'on veut étudier, soit pour la connaître, soit pour mener une action, accepter ou refuser un lot de pièces, pour une commande, etc. Il faudra donc combiner les mesures faites sur les individus pour aboutir à une mesure sur la population. Appelons A (comme action) l'ensemble dans lequel la mesure sur P prend ses valeurs.

(1) Orthographe de l'époque.

On va chercher une application $g_n : X^n \to A$; $a = g_n(x_1,\dots,x_n)$ décrira P. g_n sera un résumé statistique. La statistique est, de ce point de vue, la science qui permet d'inventer des résumés, d'en étudier les propriétés, de permettre des interprétations en fonction des données étudiées, de décrire leur intérêt et leurs limites.

2. Statistique inductive, statistique descriptive

Il est des cas où la population P est inaccessible et n'a d'utilité que conceptuelle. Ainsi en est-il de la tension artérielle : P pourrait être interprété comme l'ensemble de toutes les tensions artérielles possibles de tous les individus concernés par l'étude. Même si elle est théoriquement accessible, il est souvent impossible, pour des raisons pratiques, d'étudier tous les individus (cas des essais destructifs, par exemple). Dans tous ces cas, on étudie un échantillon de la population. On considère alors la présence d'un individu dans l'échantillon comme le résultat d'une épreuve aléatoire. On fait donc intervenir la notion de hasard et on fait appel aux probabilités pour modéliser la situation étudiée. La statistique inductive ou probabiliste ou mathématique se propose de préciser les modèles probabilistes qui permettent d'étudier des carcatéristiques d'une population P à partir des mesures faites sur un échantillon E (sondages, contrôles destructifs, essais thérapeutiques). Certaines techniques de statistique inductive sont au programme de certaines sections de techniciens supérieurs. On se référera aux textes officiels pour plus de précision.

Si E = P, il s'agit de caractériser la population à partir de mesures faites sur tous les individus de la population. Il ne s'agit pas d'induire, mais de décrire. Aussi, cette statistique est-elle appelée descriptive. Dans les cas simples, comme celui de la couleur des cheveux, elle ne repose sur aucune hypothèse *a priori*. Il suffit de compter le nombre de blonds, de bruns, etc. et éventuellement de calculer des proportions. La construction du résumé est immédiate. Il n'en est plus de même dès lors que l'on traite de données structurées par une relation d'ordre ou bien de données numériques.

3. Les qualités d'un résumé

Exprimons-les simplement en termes qualitatifs :

• D'abord, pour des collectifs semblables, la variabilité des résumés doit être beaucoup plus faible que celle relative aux individus de la population. On attend une stabilisation des fluctuations. Naturellement, la statistique est triviale si $x_1 = x_2 = \dots = x_n = a$; a est bien la valeur commune.

• Ensuite le résumé doit être peu sensible à la présence ou à l'absence d'un individu.

• Enfin le résumé doit être pertinent par rapport aux données (les mesures précédentes) et à l'objectif poursuivi.

Exemple : étude du nombre d'enfants par fratrie pour les élèves d'un lycée :

P : les élèves du lycée ; x : la taille de la fratrie de chaque élève.

Présenter les résultats sous la forme d'un diagramme circulaire est un

valeurs de x	1	2	3	4 et plus
effectifs	n_1	n_2	n_3	n_4

résumé statistique : c'est bien une caractéristique de la population des élèves du lycée.

Mais est-ce une bonne représentation ?

Comparons la représentation par un diagramme circulaire et celle par un

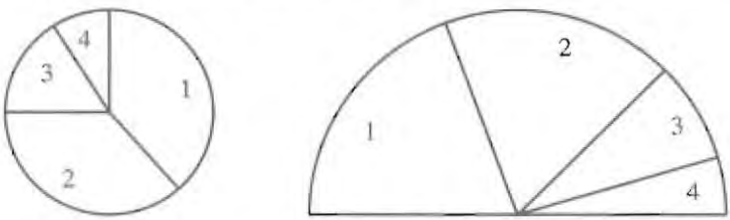


diagramme semi-circulaire. La première ne permet pas de bien voir ce qui se passe car elle ne respecte pas la structure d'ordre sous-jacente, le secteur représentant les fratries de quatre enfants et plus étant à côté de celui représentant les fratries d'un enfant. Par contre, le diagramme semi-circulaire respecte cette structure d'ordre. La remarque ne serait pas valide s'il s'agissait d'un caractère nominal comme la couleur des cheveux.

Un résumé est-il **pertinent** ?

Prenons un exemple historique. Voulant caractériser la production en blé de chaque province, les économistes de la fin du XVIII^e siècle ont essayé de trouver un indice permettant de caractériser la contribution de chaque province à l'alimentation du royaume.

Or il s'agissait d'établir le réseau de voies de communications (routes et canaux) car, à cette époque, il fallait transporter les céréales d'une province à une autre et faire face ainsi aux famines éventuelles de telle ou telle province.

On dispose, pour chaque province, de la quantité annuelle produite pendant n années consécutives : $x_1, x_2, .., x_n$ ($X = \mathbf{R}^+$). Il se peut qu'une année ait été exceptionnellement bonne et on a alors sur un axe la répartition suivante :

ou exceptionnellement mauvaise (aucune récolte) :

Peut-on utiliser la moyenne arithmétique des n productions annuelles pour décrire la production de la province ? La moyenne arithmétique $\frac{1}{n}\sum_{1}^{n} x_i$ de n réels x_i est « attirée » par les valeurs extrêmes $\inf_{1\leq i\leq n} x_i$ et $\sup_{1\leq i\leq n} x_i$: si $\inf_{1\leq i\leq n} x_i \to -\infty$ ou $\sup_{1\leq i\leq n} x_i \to +\infty$, il en est de même de la moyenne. Se doutant que la moyenne ne caractérisait pas bien la région, d'une façon empirique, on adopta comme indice celui obtenu en prenant la moyenne après avoir supprimé la plus grande et la plus petite des observations.

Par rapport à l'objectif poursuivi, ce résumé-là fut perçu plus pertinent que la simple moyenne des observations. Un tel exemple peut nous suggérer d'amener les élèves à réfléchir aux choix des résumés. C'est aussi important que de simplement. les calculer.

La réflexion précédente est illustrée par un article anonyme intitulé « Dissertation sur la recherche du milieu le plus probable » et publié en 1821 dans « Annales de mathématiques pures et appliquées » : « ... il existe certaines provinces de France où, pour déterminer le rendement moyen d'un terrain, on observe ce rendement durant 20 années consécutives, on enlève la plus grande et la plus petite valeur et on prend la dix-huitième partie de la somme des valeurs restantes. »

La pertinence se jauge aussi au regard de l'action qu'on doit mener. Si on étudie les revenus des foyers d'un pays donné, et qu'on recherche une valeur centrale pertinente, que prendra-t-on plutôt :
– la moyenne est liée à la richesse globale du pays, elle paraîtra pertinente à l'économiste ;
– les très hauts revenus tirent la moyenne vers le haut, la médiane sera mieux adaptée pour le sociologue ;
– pour toucher le maximum de personnes, le publicitaire choisira le mode.

B. Les résumés euclidiens

1. Des interprétations géométriques

Dans les situations auxquelles on s'intéresse ici, on envisage X = **R** ou X = $\mathbf{R}^2$ et on cherche à caractériser :
– la ou les valeurs centrales ;
– la ou les valeurs de dispersion ;
– une liaison à peu près linéaire (avec $\mathbf{R}^2$).

On fait une « pétition de principe » : on suppose que l'espace euclidien est adapté à la représentation des données. Cette pétition de principe ne peut être justifiée qu'*a posteriori*, c'est impossible *a priori*.

Géométriquement, dans le cas de données bivariées

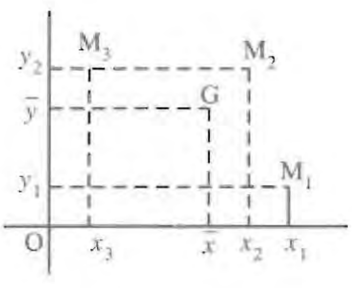
$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix},$$

on peut représenter les données de deux façons distinctes, duales l'une de l'autre :

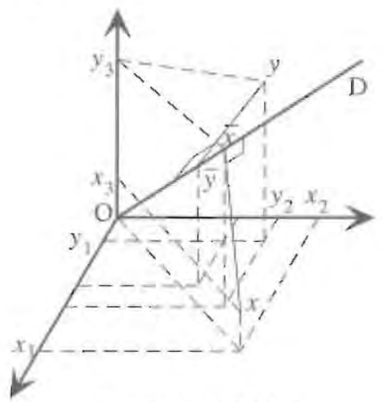
• n points dans un espace à deux dimensions (espace des individus) ;

• deux points dans un espace à n dimensions (espace des variables).

Pour faire les figures, nous avons supposé que $n = 3$:



espace des individus



espace des variables

L'intérêt de ces représentations, outre l'interprétation géométrique des résumés classiques, est d'être généralisables au cas multivarié où, sur un même individu, on effectue k mesures. On a alors dans l'espace des individus n points dans un espace à k dimensions et dans l'espace des variables k points dans un espace à n dimensions. Ceci est étudié dans les classes des lycées préparant au DECF.

Revenons au cas bivarié. L'individu i est représenté dans $\mathbf{R}^2$ par le point de coordonnées (x_i, y_i), la variable x est représentée dans $\mathbf{R}^n$ par le point de coordonnées $(x_1, x_2, \ldots, x_n)$. Soit (D) la droite de coefficient directeur $(1, \ldots, 1)$ et munissons $\mathbf{R}^2$ et $\mathbf{R}^3$ de la structure euclidienne.

Dans $\mathbf{R}^n$, la droite (D) représente les variables qui prennent la même valeur sur tous les individus.

Comme on recherche un indicateur de centralité, notons que, si tous les individus sont identiques, l'indicateur de centralité sera la valeur commune ;

la droite (D) de vecteur directeur (1,1,...,1) est le lieu des points dans l'espace des variables dont toutes les coordonnées sont égales ; la variabilité inhérente à la situation statistique fait que le point x n'est pas sur (D) ; si la structure euclidienne est adéquate, le meilleur indicateur de centralité est le point de la droite (D) le plus proche du point x, c'est-à-dire la projection orthogonale de x sur la droite (D). Un calcul montre que c'est le point dont toutes les coordonnées sont égales à $\frac{1}{n}\sum x_i$, la moyenne. On note $\bar{x}$ ce point.

On voit que, comme indicateur de centralité, la moyenne est fortement liée à la structure euclidienne.

De même y est projeté en $\bar{y}$ sur la droite (D).

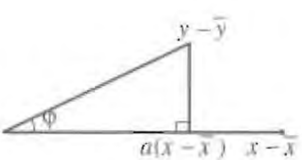
Un caractère est d'autant plus dispersé que x est éloigné de la droite (D) ; la distance du point x au point $\bar{x}$ est un indicateur de la dispersion. Cette distance est $\sqrt{\sum\left(x_i-\bar{x}\right)^2}$; c'est au facteur $\sqrt{n}$ près, fait pour normaliser, pour comparer des populations d'effectifs différents, l'écart-type ; l'écart-type σ_x vérifie alors : $n(\sigma_x)^2 = d^2(x,\bar{x})$.

L'indicateur de dispersion, écart-type, est lui aussi très lié à la structure euclidienne.

Pour des données bivariées, si le nuage de points dans l'espace des individus est aplati autour d'une droite, dans l'espace des variables les vecteurs $x - \bar{x}$ et $y - \bar{y}$ doivent former un angle très faible ou à peu près plat.

Un nuage « rond » ferait que ces vecteurs sont orthogonaux. Une application du produit scalaire permet d'établir que le coefficient de corrélation ρ est égal au cosinus de l'angle des deux vecteurs $x - \bar{x}$ et $y - \bar{y}$: $\rho = \cos(y - \bar{y}, x - \bar{x})$, ce qui indique la liaison entre ce coefficient et la structure euclidienne de l'espace. Ainsi, pour des données bivariées, si le nuage de points dans l'espace des individus est proche d'une droite, $|\rho|$ est voisin de 1 et l'angle de $x - \bar{x}$ et $y - \bar{y}$ est petit ou à peu près plat.

Si la droite de régression des moindres carrés a pour équation $y = ax + b$, alors on a $\mathrm{proj}_{x-\bar{x}}(y-\bar{y}) = a\,(x-\bar{x})$. Le coefficient de régression linéaire apparaît comme un rapport entre la projection de $(y - \bar{y})$ sur $(x - \bar{x})$ et $(x - \bar{x})$.



$\rho = \cos\varphi$

Des considérations simples de géométrie permettent de démontrer ce résultat dans le cas où $n = 3$. Il suffit d'écrire $y_i = ax_i + b + \varepsilon_i$ et de représenter géométriquement le vecteur $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_n)$.

De là l'intérêt de la vision géométrique pour la compréhension statistique de ces indicateurs. Avoir ces schémas de géométrie à l'esprit est susceptible d'aider à l'interprétation : ainsi, un coefficient de corrélation de 0,8 ne peut que conduire à penser à un angle assez faible. De même, un contresens trop classique est révélé : si on a calculé la régression de y en x par la méthode des moindres carrés en minimisant la somme des carrés des résidus[2], il ne peut pas être question de pouvoir « prévoir » une valeur de x à partir d'une valeur de y. La droite de régression de y en x est distincte de la droite de régression de x en y.

Cette image géométrique est destinée à favoriser des interprétations statistiques plus correctes, de distinguer la projection de $y - \bar{y}$ sur $x - \bar{x}$ et celle de $x - \bar{x}$ sur $y - \bar{y}$.

2. Après le lycée... l'analyse de données multidimensionnelle

Pour une même population, sur n individus de cette population, on mesure désormais k caractères (et non seulement deux).

Par exemple, je me suis intéressé il y a quelques années aux notes obtenues à l'écrit par des candidats au bac C ; chaque individu a un vecteur de notes : la note de philo, la note de math, la note de physique, celle de sciences naturelles... Il s'agit de comprendre la structure globale de cette collection de notes. On dispose d'une matrice (n,k) :

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & \cdots & j & \cdots & k \end{array} \leftarrow \text{variables} \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ \ell \end{array} & \left[\begin{array}{ccccc} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & x_{ij} & & \\ & & & & \\ & & & & \end{array}\right] \\ \uparrow & \\ \text{individus} & \end{array}$$

Soit X cette matrice où x_{ij} est la note de l'élève i à la discipline j.

Avec un grand nombre d'individus, c'est parfaitement illisible. Il faut essayer de transformer, de résumer.

(2) On recherche une droite d'équation $y = ax + b$; en fait $y_i = ax_i + b + c_i$ où c_i est un résidu de sorte que la somme des carrés des résidus de 1 à n soit minimale.

On veut chercher à préciser ce qui est le plus important dans la réussite au baccalauréat.

Est-ce la réussite de certains élèves dans les matières littéraires comme on l'a prétendu tant de fois ?

On commence, en général, par regarder ce que donne chaque discipline : moyenne, écart-type pour chaque variable (chaque discipline). C'est ainsi qu'on peut constater que la plus forte dispersion des notes de mathématiques leur fait jouer un rôle plus fort (la philosophie voit ses notes beaucoup plus centrées). L'importance d'ensemble du rôle joué par une discipline ne dépend pas seulement de son coefficient, mais aussi de l'écart-type des notes attribuées. Le fait que la discipline n'a pas de rôle est manifeste si tous les candidats ont la même note (la moyenne).

On peut en outre calculer des coefficients de corrélation entre deux variables... Mais on voudrait aller plus loin.

Reprenons le modèle euclidien en généralisant ce qui a été fait précédemment : représenter la donnée comme n points $(M_i)_{1 \le i \le n}$ dans un espace euclidien à k dimensions (espace des individus), puis comme k points dans un espace euclidien à n dimensions (espace des variables).

En k dimensions, c'est bien peu lisible, mais une dimension, deux dimensions seraient plus commodes à lire. De là une question : quel est le sous-espace à une, deux, trois dimensions le plus proche du nuage de points dans l'espace des individus ?

En termes euclidiens, ces sous-espaces passant nécessairement par le centre de gravité du nuage, on peut centrer X de sorte que, pour tout j de 1 à k, on ait : $\overline{x_j} = \frac{1}{n}\sum_i x_{ij} = 0$.

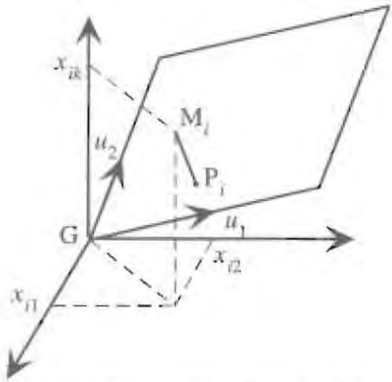
Soit à rechercher le plan vectoriel le plus proche du nuage de points.

Soit (u_1, u_2) une base de ce plan. M est projeté en P sur ce plan : $P_i = \text{proj}_{(u_1,u_2)}(M_i)$.

De la sorte, à deux dimensions, on va pouvoir regarder, obtenir de la lisibilité en balance avec un peu de perte d'information... naturellement(3).

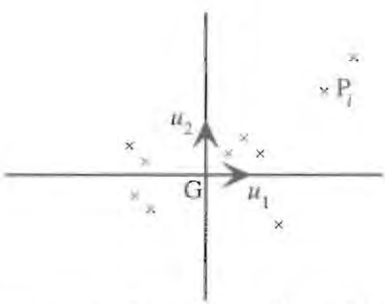
Et dans l'espace des variables une étude duale est possible : rechercher le sous-espace de dimension 1, 2, 3 approchant au mieux le nuage des variables.

(3) En permanence, le statisticien est devant un dilemme : réduire les données, les ordonner, les classer en vue de mettre en évidence la structure du phénomène ou bien perdre le moins d'information possible et, donc, classifier et réduire le moins possible car toute transformation induit une perte information. La tâche du statisticien est de représenter les données avec un minimum de perte d'information... et un maximum d'explication.



espace des variables (faire $n = 3$)

Rappelons qu'algébriquement, il s'agit de diagonaliser la matrice symétrique ${}^t\mathrm{XX}$, donc de trouver ses valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$(4) associées aux vecteurs propres u_1, u_2, ..., u_k dont on ne conserve que les premiers termes (les deux plus grandes valeurs propres). Les logiciels actuellement disponibles font ce calcul directement et produisent la représentation graphique du plan (u_1, u_2) et des points projetés : il s'agit alors de rendre compte qualitativement de ce résultat : comment peut-on l'interpréter ?



La capacité à lire une analyse factorielle est très importante pour débrouiller des données multivariées.

Cette représentation est-elle **pertinente** ?

Un des premiers critères de pertinence est de regarder quel est le pourcentage d'inertie expliquée. La variabilité globale $\sum \mathrm{GM}_i^2$ est la

(4) Dans les études statistiques, ces valeurs propres sont distinctes du fait même de leur provenance de situations réelles ; il faut bien voir que le cas de matrices à valeurs propres multiples est fortement improbable dans la réalité statistique.

somme des valeurs propres $\lambda_1 + \lambda_2 + \ldots + \lambda_k$; elle est à comparer à $\sum GP_i^2$ qui vaut $\lambda_1 + \lambda_2$; le rapport $\dfrac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \ldots + \lambda_k}$ rend compte de l'inertie expliquée par la représentation. Proche de 0,9, ce rapport signale plutôt un nuage peu dispersé autour du plan (u_1,u_2). En trois dimensions il serait en forme de soucoupe volante ; s'il était en forme de fuseau, λ_1 serait nettement plus grand que λ_2.

D'autre part, quels sont les points qui sont bien représentés ? Il est vraisemblable que les points les plus éloignés de G sont les mieux représentés. Plus généralement, les points bien représentés sont tels que l'angle entre le vecteur $\overrightarrow{GM}_i$ et le plan de projection soit petit ; les points les plus proches de G ont toutes chances de ne pas être bien représentés. Encore une fois, l'apprentissage de la lecture d'une représentation graphique dépend de la capacité à se situer vis-à-vis de la représentation géométrique dans l'espace.

Dans l'exemple des résultats au baccalauréat, le premier axe a une interprétation évidente : il est expliqué par le niveau général des élèves – pas besoin d'un gros outil statistique pour mettre en évidence cela – . Par contre, pour le deuxième axe, c'est l'opposition entre réussite en mathématiques et réussite en sciences expérimentales. Un troisième axe met en évidence l'influence de l'opposition entre réussite en sciences physiques et réussite en sciences naturelles. Par contre, l'opposition littéraire - scientifique ou la réussite dans les disciplines littéraires n'entre pratiquement pas dans les facteurs explicatifs de la réussite au baccalauréat C.

Statistiquement, il était erroné de donner crédit à l'idée selon laquelle on pouvait réussir le bac C avec les matières littéraires.

Il est possible que la représentation euclidienne ne soit pas pertinente :
– soit qu'on soit dans l'impossibilité d'interpréter les axes ;
– soit que la forme du nuage ne se prête pas au traitement, par exemple un nuage en forme de banane...

3. Des généralisations

Dans cet exemple, les mesures sont homogènes en ce sens qu'il s'agit de la même grandeur pour chaque caractère (chaque note). Il n'en irait pas de même si on traitait, par exemple, de la longueur du pied, du poids, de la taille : ces mesures sont hétérogènes. Pour pouvoir faire une analyse factorielle, on centre et on réduit chacune des mesures : on retranche $\overline{x_j}$ à x_{ij}

et on divise par l'écart-type σ_j. Ceci fait que tous les individus étudiés appartiennent à une hypersphère de rayon $\sqrt{n}$: $\dfrac{\sum_i \left(x_{ij}-\bar{x}\right)^2}{\sigma_j^2} = n$. C'est-à-dire que toutes les variables interviennent de la même façon sans tenir compte de l'échelle dans lesquelles elles sont mesurées.

Une autre généralisation plus fréquente a lieu dans le cas où on souhaite comparer deux classements. Par exemple, des performances dans une discipline donnée (avec k classes) et les résultats à des tests psycho-génétiques selon ℓ modalités :

$$
\begin{array}{c c l}
 & \begin{array}{ccccc} 1 & \cdots & j & \cdots & k \end{array} & \leftarrow \text{classes} \\
\begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ \ell \end{array} & \left[\begin{array}{ccccc} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & n_{ij} & & \\ & & & & \\ & & & & \end{array} \right] & \\
\uparrow & & \\
\text{modalités} & &
\end{array}
$$

où n_{ij} est le nombre d'individus qui ont la modalité i pour le premier caractère et la modalité j pour le deuxième. Il s'agit d'étudier la structure interne de cette observation.

On introduit les fréquences f_{ij} : $f_{ij} = \dfrac{n_{ij}}{\mathrm{N}}$ où N est l'effectif total, la proportion d'apparitions du couple (i,j) dans l'ensemble de tous les couples apparus, puis $f_{i\bullet} = \sum\limits_j f_{ij}$, $f_{\bullet j} = \sum\limits_i f_{ij}$ qui permettent d'obtenir le tableau des fréquences et ses marges :

$$
\begin{array}{c c c}
 & \begin{array}{ccccc} 1 & \cdots & j & \cdots & k \end{array} & \\
\begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ \ell \end{array} & \left[\begin{array}{ccccc} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & n_{ij} & & \\ & & & & \\ & & & & \end{array} \right] & f_{i\bullet} \\
 & f_{\bullet j} &
\end{array}
$$

Ainsi on peut comparer des profils : par exemple, deux colonnes ou deux lignes en calculant la contribution relative $\frac{f_{ij}}{f_{i\bullet}}$ du couple (i,j) à l'élément i ou celle $\frac{f_{ij}}{f_{\bullet j}}$ du couple (i,j) à l'élément j.

Pour reprendre le traitement précédent, il faut définir une distance entre deux modalités. On adopte en général :

$$d(i,i') = \sqrt{\sum_{j=1}^{n} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2},$$

dite distance du khi-deux[5].

Pour cette distance, deux modalités j ayant le même profil étant fusionnées, ne la font pas changer. Cette propriété, dite de l'équivalence distributionnelle, est utile car elle permet de remédier à l'arbitraire toujours possible des classifications. De là une possibilité de réexploiter la géométrie euclidienne afin de trouver de même des axes factoriels permettant d'interpréter des proximités entre modalités des deux classements. C'est ce qu'on appelle l'analyse des correspondances.

Pourquoi fait-on de la géométrie dans l'espace en ES ?

Ce n'est pas la capacité technique à effectuer les calculs qui est recherchée : ils sont très bien effectués par les logiciels. On fait de la géométrie dans l'espace pour donner aux élèves une vision les rendant capables d'interpréter correctement des résultats alors que l'exécution des calculs est désormais l'enfance de l'art. Ces lycéens vont rencontrer de telles démarches dans diverses voies d'études : économie, sociologie, histoire... Il faut qu'ils soient capables de comprendre de quoi il s'agit. C'est donc comme instrument d'illustration des concepts statistiques que l'apprentissage de la géométrie importe.

Il est important aussi comme constitution d'images mentales qui aideront à apprendre l'algèbre linéaire dont on sait qu'elle est fortement présente dans les études économiques. Ce qu'il est important que les élèves maîtrisent est finalement relativement élémentaire : les notions de points, de droite, de plan, d'intersection, de parallélisme, d'orthogonalité, de projection. Cela ne peut pas être appris d'un coup en Première ou Terminale : un apprentissage

(5) distance qui est euclidienne, sinon la géométrie de représentation est beaucoup plus confuse.

continu et conséquent de la géométrie dans l'espace depuis l'école et particulièrement en classe de Seconde constitue aussi la base de cette formation.

Bibliographie

Analyse des données, brochures APMEP n^{os} 28 et 40 (1980) ; ces deux brochures rassemblent des articles de divers auteurs et sont très pédagogiques par les exemples traités et commentés (ouvrage épuisé).
Gilbert SAPORTA (1990). *Statistiques et analyse de données*, Technip. Ouvrage de référence pour la statistique inductive comme pour l'analyse des données. Les modèles mathématiques sont explicités, de nombreux exemples sont donnés.
Xavier BRY (1995). *Analyses factorielles simples*, Économica. Présentation intuitive de la technique statistique avec, en encarté, le traitement mathématique. Très pédagogique.
Thierry FOUCART (1997). *L'analyse des données, mode d'emploi*, Presses Universitaires de Rennes. À partir de nombreux exemples avec leurs sorties informatiques, on procède à des interprétations ; les parties mathématiques sont réduites.