JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

DANIEL DUGUÉ

Problèmes actuels de statistique mathématique

Journal de la société statistique de Paris, tome 115 (1974), p. 203-209
<http://www.numdam.org/item?id=JSFS_1974__115__203_0>

II

COMMUNICATIONS

PROBLÈMES ACTUELS DE STATISTIQUE MATHÉMATIQUE

(Communication faite le 21 novembre 1973 devant la Société de statistique de Paris)

Of all the problems now attached to mathematical statistics, it has been chosen to indicate thre broad outlines of the various methods of data analysis in principal components. Then is exposed an example of application to the evolution of the socio-professional composition of population in big French cities during last censuses.

Unter allen Problemen, die sich augenblicklich den mathematischen Statistikern stellen, hat man die grossen Linien der Methoden der Analyse der Gegebenheiten in ihren hauptsachlichen Zusammensetzungen gewahlt. Als Beispiel für die Anwendung nimmt man die Entwicklung der sozio-professionellen Zusammensetzung der Bevolkerung der franzosischen Grosstädte nach den letzten Volkszählungen.

Entre todos los problemos que se planteau hoy a hos que se dedican a la estadistica con tendencia matematica se ha escoyido de indica los lineas grandes de los metodos de analisis de los datos en componentes principales. Se du despues un ejemplo de aplicacion a la evolucion de la composicion socio-profesional de la poblacion de los grandes ciudades francesas en el curso de los ultimos censos.

J'ai hésité avant de choisir un plan : ou bien passer en revue *tous les problèmes* que se posent actuellement les statisticiens de tendance mathématique (avec la quasi-certitude d'en oublier), ou bien choisir un sujet parmi les problèmes actuels et vous en dessiner les grandes lignes.

C'est la seconde option qui a prévalu, elle me paraît même convenir à l'ambiance d'un dîner-débat. Dessiner les grandes lignes ai-je dit. C'est bien le terme propre pour parler de ces méthodes d'analyse des données en composantes principales qui sont fondées aussi bien sur la géométrie des espaces n-dimensionnels que sur le calcul des probabilités. C'est pourquoi on a pu dire que la statistique a dans une certaine mesure débordé le calcul des probabilités.

Un histogramme de fréquence, chacun le sait, est une figure géométrique qui permet de résumer sur une feuille de papier des données à une dimension (fréquence par exemple des individus d'une population ayant un âge compris entre des limites données, fréquence

des individus ayant une taille entre des limites données). Cet histogramme se présente comme une suite de rectangles dont les bases se succèdent sur l'axe horizontal et dont la hauteur mesure la fréquence à étudier.

C'est la représentation expérimentale d'une courbe théorique et l'écart à cette courbe théorique a fait l'objet de plusieurs résultats de calcul des probabilités. Le problème se complique quand au lieu d'une seule grandeur attachée à un individu (l'âge ou la taille) ou étudie un groupe de plusieurs grandeurs (par exemple, âge, revenu, capital) attachées à un même individu. Si à chaque individu sont associés k caractères on pourra représenter l'ensemble des données par les $\frac{k(k-1)}{2}$ tableaux marginaux qui donnent un résumé de l'information, mais pas toute l'information. Pour avoir la totalité de cette dernière il conviendrait de disposer de $p_1\ p_2\ \ldots\ p_k$ nombres si le i^{ieme} caractère a p_i différentes modalités. Ces nombres seraient difficiles à classer d'une manière utilisable. Il serait surtout difficile d'avoir une représentation globale de cet ensemble de données. L'analyse des données va consister tout d'abord à construire la matrice des covariances des k caractères pour l'ensemble de n éléments qui constitue la population examinée, soit comme chacun sait la somme $\frac{1}{n}\sum_{i=1}^{n}(x_r^i - \dot{x}_r)\ (x_s^i - \dot{x}_s) = c_{rs}$, x_r^i, x_s^i étant les mesures du r^{ieme} et du s^{ieme} caractère attaché au i^{ieme} individu, $\dot{x}_r$ et $\dot{x}_s$ étant les moyennes du r^{ieme} et du s^{ieme} caractère.

On obtiendra ainsi la matrice c_{rs} à k lignes et k colonnes.

Cette matrice peut être diagonalisée, ce qui revient à chercher les axes principaux de la forme quadratique définie positive (hyperellipsoïde) qui lui est associée, et à les prendre pour nouveaux axes de coordonnées. Dans ce nouveau système d'axes, les anciens axes auront une certaine représentation et chaque individu sera représenté par un point qui aura encore k coordonnés.

On choisira ensuite les deux axes les plus grands (composantes principales) et on projettera la figure sur le plan de ces deux axes.

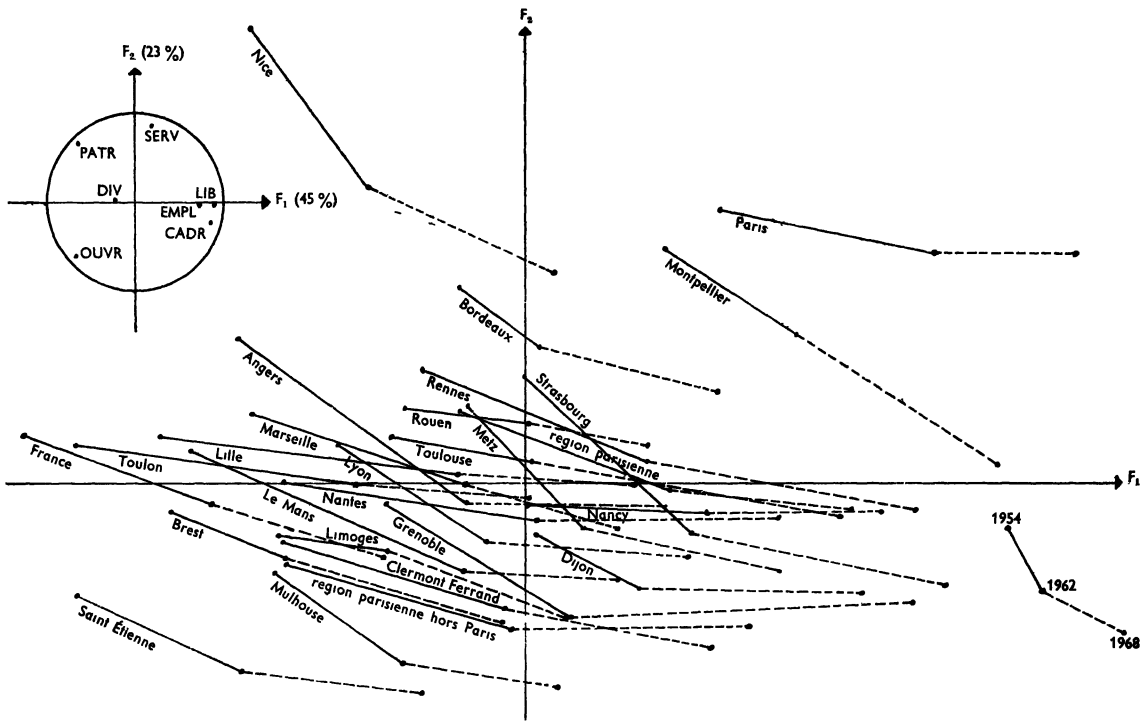
Sur ce plan les anciens axes seront représentés par leur projection et les individus de même. On aura une représentation à deux dimensions d'autant plus proche de la réalité à k dimensions que la somme des deux axes retenus sera une fraction plus importante de la somme totale des axes.

Sur un exemple, il sera facile de lire un résumé de l'information. Examinons le tableau d'analyse en composantes principales de l'évolution de la composition de la population de 24 grandes villes françaises au cours des recensements de 1954, 1962, 1968. La population est groupée en sept catégories (patrons, cadres, ouvriers, employés, personnel de service, professions libérales, divers). Chaque ville est donc représentée par un point ayant sept coordonnées. Il y aura, chaque ville étant représentée 3 fois, 72 points.

Une analyse en composantes principales révèle que l'axe F_1 est 45 % de la somme totale des axes, l'axe F_2 23 %. Leur plan fournit donc 68 % de la variance totale. On aura ainsi une très bonne idée de l'ensemble du phénomène.

Voilà la figure obtenue par projection sur ce plan. Ce travail a été effectué à l'Atelier parisien d'urbanisme (APUR) par une équipe dirigée par notre collègue Morlat. Il est facile de lire les informations sur cette carte.

La ville de Nice a une population en grande partie composée de patrons et de personnel de service avec transfert de 1954 à 1962 puis 1968 vers la région où se projettent les axes employés et cadres. Ce phénomène de tertiarisation est général et affecte toutes les villes.



Une étude de ce genre sur les différents avantages des professions a été effectuée fin septembre 1973 par un hebdomadaire.

Ces méthodes semblent donc maintenant atteindre une fraction importante d'usagers. Il me paraîtrait souhaitable qu'elles se répandent encore plus et j'ai été heureux de les présenter ce soir à la Société de statistique de Paris.

Daniel DUGUÉ

Directeur de l'Institut de statistique des universités de Paris

DISCUSSION

M. GIBRAT. — Je suis resté sur ma faim en vous voyant négliger les carrés latins dont vous aviez promis de parler. Ils m'intéressent parce que leur premier étudiant les avaient choisis parce que, détestant l'humanité, il pensait que ses travaux ne serviraient jamais à rien et il fut bien déçu.

Y a-t-il donc du nouveau?

M. DUGUÉ. — Mon cher président, je suis navré de votre déception. Elle me touche d'autant plus que je connais votre passion pour les carrés latins et la propagande que vous leur avez faite. Malheureusement une conférence comme celle de ce soir est forcément limitée et j'ai voulu mettre notre Société au courant des progrès en analyse des données. Un autre soir nous pourrons aborder le sujet qui me tient à cœur croyez le bien autant qu'à vous.

M. J. STOETZEL. — M. Dugué a centré sa très intéressante communication sur le système de la moyenne et de l'écart type. Les distributions que je rencontre le plus fréquemment sont, non pas normales, mais log-normales et par conséquent je crois devoir utiliser pour valeur centrale et pour mesure de dispersion la médiane et une variable appartenant au système de la médiane, par exemple le rapport des quartiles. Peut-on espérer disposer prochainement d'une théorie mathématique de la variation probabiliste de la médiane et des valeurs associées? C'est ma première question.

En second lieu, l'analyse des données s'intéresse essentiellement à l'association des données. Mais on rencontre d'autres problèmes : hiérarchisation des données, typologie des groupes de données. En ce qui concerne les typologies, on sait que les méthodes actuelles n'apportent pas de résultats universels, indépendants du point de départ choisi. Peut-on espérer obtenir prochainement des techniques typologiques à valeur universelle?

Je terminerai en souhaitant que les statisticiens mathématiciens s'attachent également au problème des épreuves de signification.

R. A. Fisher qui en a montré l'intérêt, n'a envisagé que les cas les plus simples. Certains tests psychologiques ont une structure beaucoup plus compliquée, et méritent que de leur nature logique les statisticiens mathématiciens dérivent des épreuves de signification appropriées.

M. DUGUÉ. — Je suis tout à fait d'accord avec M. J. Stoetzel sur l'importance des médianes et des quartiles. Pourquoi ne s'en sert-on pas davantage? La réponse est simple. Les mathématiciens, et les statisticiens sont un peu mathématiciens, sont des paresseux. Les propriétés de la moyenne et de l'écart type sont tellement plus simples; elle offrent tellement de somptueux développements qui ne sont pas encore complètement épuisés! A vrai dire la loi de Laplace-Gauss qui justifie cette priorité accordée à la moyenne et à l'écart type est assez répandue dans la nature. Malgré tout les autres valeurs types ne devraient pas être négligées : à ce sujet, mentionnons les études de robustesse c'est-à-dire des problèmes qui se posent quand les variables initiales ne sont pas normales mais n'en sont pas « trop » éloignées.

Tout à fait d'accord pour ne pas négliger les autres problèmes à propos des données : hiérarchisation et typologie de groupes.

Merci aussi d'avoir évoqué sir Ronald Fisher qui avec Georges Darmois fut mon maître. Il est certain que beaucoup reste à faire à propos des problèmes de « signification » en particulier dans le cas de variables multidimensionnelles.

M. SCHWARTZ. — 1° L'analyse des données est, sans aucun doute, un procédé ingénieux de *description* d'une situation, mais l'absence d'un procédé permettant de *tester* une hypothèse n'est-elle pas une restriction à son emploi dans la recherche?

2° Si l'on admet sommairement que la recherche procéde en deux étapes; l'idéation (élaboration des hypothèses) et la vérification de ces hypothèses, c'est cette deuxième étape qui, sous forme de tests d'hypothèse, a constitué jusqu'ici la partie principale de la méthode

statistique, qui ne prétendait aucunement aider à l'idéation. L'analyse des données, parce qu'elle peut suggérer des idées, a donc son intérêt mais seulement dans la première étape; c'est dans cette mesure qu'on doit se demander si vraiment la statistique « est en train d'échapper aux probabilités ».

M. Dugué. — Votre première intervention soulève un point important et même très important : il s'agirait en somme d'établir un test analogue à ceux de Kolmogoroff-Smirnoff ou Von Mises-Smirnoff pour la comparaison d'un histogramme à une courbe de probabilité théorique et qui permettrait de comparer une matrice de covariance expérimentale à une matrice de covariance théorique. Aucun résultat définitif n'est acquis sur ce point mais nous sommes très loin de partir de zéro. H. Hotelling qui est un des initiateurs de l'analyse des données a obtenu une loi dite naturellement loi d'Hotelling et qui si elle était exploitée permettrait de dresser les tables que vous souhaitez avec juste raison. Ce problème est signalé aux jeunes chercheurs.

Sur la deuxième intervention il est certain qu'un tableau d'analyse des données peut suggérer un modèle (comme d'ailleurs un histogramme peut suggérer la courbe théorique). Mais ici nous touchons un point délicat : la trop grande facilité matérielle avec laquelle peuvent être analysées des données (je mets en cause les ordinateurs) peut conduire à des erreurs. En introduisant le nombre de dimensions nécessaires on peut arriver à « décrire » n'importe quel nuage, on n'aura pas créé un modèle pour autant.

M. Ferrouillat. — Quels sont les développements futurs à attendre dans les domaines suivants :

— lissages;
— contrôle de fabrication et réception;
— traitement d'une petite masse de données (tests non paramétriques);
— fiabilité (estimation de paramètres).

M. Dugué. — Ce sont des sujets fort intéressants et très étudiés.

Les problèmes d'estimation et de tests non paramétriques retiennent l'attention des jeunes chercheurs. Je n'ai malheureusement pas eu le temps d'aborder tous ces points et j'ai préféré vous donner des informations sur des méthodes en plein développement.

M. Dumas. — M. Dumas attire l'attention sur ce que l'ordinateur donne des résultats exacts en effectuant des opérations que, peut-être, un opérateur calculant lui-même n'aurait pas admises. Il s'ensuit que les résultats obtenus par l'ordinateur ne doivent pas être admis sans qu'une critique approfondie en soit faite et ait montré qu'ils pouvaient l'être.

M. Dugué. — Merci d'avoir souligné, ce que je n'avais pas fait, l'importance de l'ordinateur dans ce domaine. Il est bien évident qu'il ne serait pas possible à un calculateur démuni des moyens modernes de trouver les valeurs et vecteurs propres d'une matrice de corrélation de plus de trois lignes, trois colonnes portant sur des centaines de résultats.

Merci aussi d'avoir rappelé (car ils sont connus) les dangers de l'ordinateur.

M. Gallais-Hamonno. — On vient de poser le problème de l'utilisation de la médiane au lieu de la moyenne. Mais il y a pire! Des économistes sont en train de découvrir qu'un grand nombre de phénomènes — et surtout dans le domaine boursier — suivent une loi de Pareto-Mandelbrodt qui ont une moyenne *mais pas d'écart type*. Or, sans écart type, comment peut-on effectuer des calculs de corrélation?

M. Dugué. — On ne peut évidemment parler de coefficients de corrélation que s'il existe un écart type pour les variables marginales. D'ailleurs dès que l'on s'écarte de variables normales (de Laplace-Gauss selon la terminologie française) le coefficient de corrélation devient suspect. Notre regretté ancien président le professeur Frechet a attiré l'attention sur ce point dans des articles fort intéressants publiés par la revue de l'Institut International de Statistique.

M. Milhaud. — Quand on recueille les valeurs d'une variable biologique chez des sujets dits normaux, on a toujours quelques résultats aberrants. On est tenté de les éliminer en disant qu'ils concernent des malades qui s'ignorent ou qu'ils sont faux.

Mais la décision des seuils au-delà desquels les résultats très élevés ou très bas sont à éliminer est arbitraire.

Or elle a une incidence sur le calcul de la moyenne et sur celui de l'écart type.

Les données recueillies paraissent décrites de façon avantageuse par l'indication de la médiane et par celle de l'écart probable ou médian, qui ne sont que très peu influencés par les résultats extrêmes.

M. Dugué. — Il me paraît très dangereux d'écarter des résultats sous le prétexte qu'ils sont « aberrants » c'est-à-dire en fin de compte en désaccord avec les idées préconçues de l'expérimentation. Cela a d'ailleurs conduit à des erreurs célèbres d'interprétation sur lesquelles je n'insiste pas.

Il va de soi que la médiane à ce point de vue offre bien des avantages.

Simplement elle est plus difficile à manipuler : ce n'est pas un beau modèle mathématique comme la moyenne.

M. Guitton. — Tout à fait d'accord avec la nécessité de l'enseignement statistique à diffuser dans l'opinion.

M. Guitton demande à M. Dugué pourquoi il n'a pas parlé de corrélation. A cet égard, le graphique, le nuage de points est plus parlant qu'un tableau. Précisément l'analyse des données permet de découvrir les axes autour desquels les nuages sont les plus significatifs. La réalité nous dicterait ainsi un modèle d'interprétation. Le problème est de pouvoir identifier, c'est-à-dire donner la signification économique des axes. L'analyse des données nous ferait ainsi découvrir, en dehors de nos préférences, les variables stratégiques de l'économie.

M. Dugué. — Je remercie bien vivement M. Guitton de son intervention. Tout le monde sait combien l'on doit à M. Guitton dans le domaine de l'introduction des mathématiques et de la statistique en économie. Il est certainement d'accord avec moi pour trouver que beaucoup reste à faire. Est-ce sous-estimer le niveau mathématique et statistique de la fraction très cultivée de la population française (professions libérales, hauts fonctionnaires même ceux attachés à l'Éducation nationale, le personnel chargé de la confection des lois) que d'avancer qu'il ne dépasse (s'il atteint) la connaissance des moyens d'effectuer les quatre opérations et la lecture d'un histogramme? Pour ma part je voudrais y ajouter la notion (évidemment très vague) d'intégrale et la lecture d'un tableau d'analyse des données. Que tous les parlementaires puissent utiliser un tel tableau voilà qui me paraitrait devoir figurer dans les programmes de tous les partis politiques.

M. Malinvaud. — Les nouvelles méthodes d'analyse des données sont appelées à un grand avenir; je partage pleinement l'avis du conférencier à ce sujet. Mais nous devons

encore apprendre à les utiliser et à interpréter leurs résultats. A ce sujet je voudrais attirer l'attention sur deux points.

En premier lieu le tableau des données lui-même est souvent intelligible sans le recours à aucune représentation géométrique. Ainsi un utilisateur peut trouver plus naturel de réfléchir directement sur un tableau de données que sur des graphiques assez abstraits qui en sont tirés.

En second lieu représenter géométriquement sur un plan un nuage de points d'un espace à un nombre élevé de dimensions comporte toujours un certain arbitraire.

On peut le comprendre en observant que les composantes principales, lesquelles déterminent toute la représentation, dépendent à la fois des unités de mesure (si l'on part de la matrice des covariances) et de la liste précise des caractères que l'on a observés. Il faut toujours se souvenir de cet arbitraire au moment où on interprète les résultats d'une analyse des données.

M. DUGUÉ. — Sur le premier point je voudrais bien préciser que les « partisans » de l'analyse des données (en tout cas ceux que je connais) ne sont pas sectaires et ne prétendent pas imposer leurs méthodes à l'exclusion de toutes les autres. Il est bien évident que l'ensemble des données multidimensionnelles donne une information plus complète que le résumé, la vue perspective plane qui est une analyse des données. Mais cette vue perspective me paraît ajouter un élément intéressant à la connaissance des phénomènes.

Sur le second, il est bien évident que l'analyse des données à elle seule ne donne aucune explication mais une description en fonction d'une certaine idée, « arbitraire », j'en conviens que l'on se fait de la situation.

Quels sont les facteurs qui interviennent, le phénomène a-t-il n_1 ou n_2 dimensions? Ce sont des points qui doivent être fixés avant l'analyse qui a elle seule ne peut apporter de solution.

M. MASSON. — Il n'existe pas une analyse de données mais de nombreuses analyses de données. Chacune répond à un problème particulier.

Par ailleurs una analyse de données non interprétable ne peut pas être prise en considération.

M. DUGUÉ. — Je suis tout à fait d'accord avec M. Masson. Le temps limité de cette conférence ne m'a permis de parler que de l'analyse des composantes principales. L'interprétation des données est un problème qui est indépendant de leur présentation. Ce qui me paraît important dans l'analyse en composante principale est que, au prix d'une certaine perte d'information, on arrive à résumer un tableau multidimensionnel en un graphique-plan.