LA STATISTIQUE VUE PAR UN MATHEMATICIEN

Jean-Pierre KAHANE[1]

Ce texte est celui de la conférence inaugurale prononcée par Jean-Pierre Kahane, mathématicien, membre de l'Académie des sciences de Paris, à l'occasion du *Deuxième colloque francophone international sur l'enseignement de la statistique* (CFIES'2010 – Bruxelles – 8, 9 et 10 septembre 2010).

La rédaction de *Statistique et Enseignement* remercie l'auteur de l'avoir autorisée à reproduire ce texte dans sa rubrique « Libres propos ».

Bruxelles, 8 septembre 2010

Vous avez voulu, pour introduire ce colloque international sur l'enseignement de la statistique, le point de vue d'un mathématicien qui n'est pas statisticien. C'est bien mon cas. Mais l'espèce du mathématicien-non-statisticien est faite d'individus très différents, et vous auriez une toute autre conférence en vous adressant à quelqu'un d'autre.

En Angleterre par exemple, il existe une forte tradition de statistique intégrée aux mathématiques. Le département de mathématiques de l'Université de Cambridge s'appelle département de mathématiques et de statistique. Les probabilistes anglais ont baigné dans la statistique au cours de leurs études.

Vous savez bien que ce n'est pas le cas en France. Quand j'étais jeune, au CNRS, il n'y avait pas de recherche en statistique, et les probabilités faisaient partie de la section de physique théorique. Paul Lévy, le grand probabiliste français de l'époque, était un point singulier. Les probabilités comme domaine de recherche ne se sont vraiment développées en France qu'à partir des années 1950. Elles se sont fort bien développées, vous le savez bien, puisque la première médaille Fields attribuée à un probabiliste l'a été à un Français, Wendelin Werner. La statistique a pris un bon départ, mais n'est pas encore en France au niveau atteint par les probabilités.

Mon expérience personnelle est d'avoir découvert la statistique par l'enseignement et par l'histoire, puis par le contact avec des statisticiens et en particulier ceux d'Orsay, puis via l'Académie des sciences, et enfin récemment par ses relations avec les mathématiques qui me sont le plus familières. C'est donc de cela que je vais vous parler.

Mon expérience de l'enseignement de la statistique date du début des années 80. Les étudiants qui commençaient leurs études universitaires en se destinant à la biologie ou à la géologie avaient un premier cycle qui s'appelait CBBG. Ils avaient fait très peu de mathématiques jusque-là, et la plupart considéraient que c'était une matière dure, et leur point faible. Par rapport à la biologie, qui était la matière principale pour la plupart d'entre eux, les

[1] Professeur émérite à l'Université Paris-Sud, Orsay, membre de l'Académie des sciences de Paris et ancien président de la Commission de Réflexion sur l'Enseignement des Mathématiques en France, jean-pierre.kahane@math.u-psud.fr

© Société Française de Statistique (SFdS), Décembre/December 2010

mathématiques faisaient figure de discipline de service. J'ai aimé enseigner dans ces conditions. Les données et les objectifs étaient clairs, j'en avais discuté avec les collègues des autres disciplines ; les étudiants savaient peu de choses, ils avaient peu de temps à consacrer aux mathématiques, et ils devaient se familiariser avec les fonctions de plusieurs variables, les équations différentielles, les probabilités, et les estimateurs et tests d'hypothèse en statistique. L'objectif en statistique était donc très limité et très clair. Il fallait trouver les chemins les meilleurs pour la réalisation du programme, c'est à quoi se réduisait pour l'essentiel mon travail de mathématicien. Grâce à mon collègue John Hubbard, qui était chargé d'une autre section, j'ai découvert à cette occasion l'usage des calculettes dans l'enseignement, je me suis bien amusé, et j'ai su que certains élèves avaient trouvé le goût des mathématiques avec les probabilités et la statistique.

J'aimais bien, comme illustration de l'approximation des lois binomiales par les lois de Gauss et comme exemple de test d'hypothèse, raconter la découverte de Laplace concernant les naissances de garçons et de filles à Paris au 18^{e} siècle. Elle est excellemment décrite dans l'essai philosophique sur les probabilités, et déjà dans la dernière leçon de Laplace à l'Ecole normale de l'an III, consacrée aux probabilités. Je vous y renvoie plutôt que de la déflorer si vous ne la connaissez pas. Plus tard, j'ai découvert que Nicolas Bouleau faisait lire l'essai philosophique sur les probabilités à ses élèves de l'Ecole des Ponts comme complément de son propre cours. C'est une lecture qui a passionné les gens cultivés du début du 19^{e} siècle, et qui reste de très grand intérêt. C'est d'ailleurs la préface du monumental traité qui s'appelle « théorie analytique des probabilités ».

Laplace voulait que l'enseignement des probabilités se répande dans les lycées et fasse partie de l'instruction publique. C'est de façon très significative qu'il consacre sa dernière leçon aux mille élèves de l'Ecole normale de l'an III aux probabilités et à leurs usages, en particulier en statistique. Nous avons un retard de 150 ans sur la recommandation de Laplace pour les probabilités, et de 200 ans pour ce qui est de la statistique. Mais vous êtes réunis ici pour relever le flambeau.

A l'époque de Laplace, il y a d'autres références en France pour la statistique. Je pense à Legendre et à Fourier.

Dans l'excellente Histoire de la statistique de Stephen Stigler (The history of statistics, the measurement of uncertainty before 1900, Harvard University Press, 1986), Legendre occupe tout le premier chapitre, comme créateur de la méthode des moindres carrés. Stigler le cite, le commente, et reproduit les premières pages de l'« appendice » des « Nouvelles méthodes pour la détermination des orbites des comètes », qui date de 1805. Son commentaire est dithyrambique. Je traduis : « Pour la force et la clarté de l'exposition cette présentation est insurpassée ; elle doit être comptée comme une des plus claires et des plus élégantes introductions à une nouvelle méthode de la statistique dans l'histoire de la statistique ». Il vaut la peine en effet de lire Legendre. Je résume. On trouve, en géodésie et en astronomie, pour la détermination de certaines grandeurs, des équations surabondantes. Elles sont généralement incompatibles, et cependant elles correspondent à des mesures faites avec soin. Au lieu de résoudre le système $A = 0$, $B = 0$, $C = 0$, etc., on cherche à rendre minimum la somme des carrés des premiers membres. On obtient ainsi un nouveau système d'équations, avec autant d'équations que d'inconnues. On le résout, cela donne des valeurs pour A, B, C… ; s'il y en a d'aberrantes, on supprime les équations correspondantes et on refait le calcul, facilement comme le montre Legendre. A la fin de son exposé, Legendre donne l'interprétation géométrique : le point dont la somme des carrés des distances à des points donnés est la plus

© Société Française de Statistique (SFdS), Décembre/December 2010

petite possible est le centre de gravité. « On voit donc que la méthode des moindres carrés fait connaître, en quelque sorte, le centre autour duquel viennent se ranger tous les résultats fournis par l'expérience, de façon à s'en écarter le moins possible. » Puis vient l'application à la détermination du méridien terrestre.

Pour Stigler, la méthode des moindres carrés est le début de la statistique comme science : une méthode générale, issue de questions importantes concernant le traitement de données numériques nombreuses et susceptibles d'erreurs. Elle a été immédiatement adoptée, et un peu plus tard revendiquée par Gauss, qui en a publié sa version en 1809. Son interprétation probabiliste a été une source d'inspiration pour Laplace et pour Gauss.

Ici une parenthèse : les œuvres de Legendre sont unanimement appréciées, mais elles n'ont pas été réunies pour publication, il n'est pas facile d'y avoir accès. Il y a une photo de Legendre au tout début du livre de Stigler, mais c'est d'un contemporain, un boucher qui est passé de la Montagne à la contre-révolution. Pauvre Adrien-Marie Legendre !

Stigler ne parle guère de Fourier, sauf à l'occasion quand il s'agit d'un autre personnage. Cependant Fourier a joué un rôle dans l'histoire de la statistique. Quand, en 1815, il a quitté Grenoble pour Paris, il a trouvé comme situation la direction du bureau des statistiques. Il y était bien préparé par ses rapports avec Laplace. En 1817, juste après avoir été réélu et nommé à l'Académie des sciences, où il avait été déjà élu mais récusé par le roi, il a participé à une commission chargée d'examiner la proposition faite par un anonyme d'offrir un capital pour la fondation d'un prix destiné à la statistique ; il s'est révélé ensuite que cet anonyme s'appelait Montyon, et c'est ainsi qu'est né le prix Montyon de statistique. C'est Fourier qui a présenté le rapport. D'abord il précise : il s'agira des recherches statistiques. Et il précise la portée du sujet : « La statistique, cultivée et enseignée dans plusieurs Etats du Nord de l'Europe, a fait pendant le dernier siècle des progrès remarquables. Cette science emprunte ses éléments des branches les plus diverses de nos connaissances ; son objet est très étendu. Il consiste surtout à recueillir et à exposer avec ordre les faits qui intéressent directement l'économie publique. Le désir d'encourager une étude aussi utile et de la ramener, autant que possible, à des principes constants, ne peut qu'être approuvé et partagé par l'Académie des sciences. » Le rapport poursuit dans la même veine : l'Etat, la société civile, l'agriculture, le commerce, l'industrie peuvent bénéficier d'un traitement des données qu'ils fournissent au moyen de principes communs. « On a découvert en effet par des observations réitérées quelques principes constants qui peuvent servir dans un grand nombre de cas à comparer entre eux et même à vérifier les résultats des recherches statistiques. On peut aussi déterminer exactement le nombre des observations nécessaires pour procurer un degré suffisant de certitude. »

On reconnaît aisément l'influence de Laplace, dont la position éphémère comme ministre de l'Intérieur de Napoléon avait exactement la signification qu'indique Fourier.

Avec Legendre et Fourier nous voyons deux aspects et deux sources de la statistique dans les années 1800. D'un coté l'astronomie, avec toutes les recherches entraînées par la navigation et le problème de la détermination des longitudes, et la géodésie avec toutes les questions liées au système métrique et à la mesure de la Terre ; de l'autre, les besoins de l'Etat, la collecte des données nécessaires à la conscription et à l'établissement de l'impôt, et toutes les questions qui se posent à la société civile. La stature de Laplace domine les deux aspects.

Ainsi la France n'avait pas un mauvais départ en statistique. Cependant, au cours du 19^{e} siècle, c'est ailleurs en Europe qu'on vit les progrès marquants. Le premier nom qui émerge

© Société Française de Statistique (SFdS), Décembre/December 2010

est celui du Belge Quetelet. C'est Fourier qui enseigna à Quetelet la théorie des probabilités et ses applications. Mais Quetelet mena des travaux dans tous les domaines des statistiques : l'astronomie, la météorologie, les recensements de population, les naissances et les morts, les caractères physiques et mentaux des individus, la criminalité, les pratiques judiciaires, et de façon générale tout ce qui était susceptible de mesures, de calculs et d'erreurs. Il était entreprenant dans tous les domaines et contribua à la fondation des principales associations de statisticiens en Europe, et en particulier en Angleterre.

Et c'est en Angleterre qu'à la fin du 19^{e} siècle la statistique prit sa forme actuelle, avec son application aux sciences sociales et les concepts qui en résultaient. Les études de Francis Galton sur l'hérédité l'ont conduit à la notion de régression et à celle de corrélation, qui affranchissait la science du recours à la causalité. Elles ont fortement influencé ses successeurs, Francis Edgeworth et Karl Pearson. Avec Pearson, la statistique introduit de nouveaux objets en probabilités, en particulier les distributions non symétriques qu'on appelle maintenant distributions Gamma. Avec Edgeworth, la statistique devint matière d'enseignement, et il est instructif de voir les tables des matières traitées dans ses cours à King's College et à University College à Londres en 1885 et 1892 que Stigler donne en appendice à son livre.

Personnellement, ce sont les arbres de Galton-Watson, les processus de naissance et de mort, dont je connais le mieux certains prolongements. Les cascades multiplicatives de Benoît Mandelbrot en sont un prolongement.

J'abandonne l'histoire, dont je ne connais bien que le tout début, et je reviens à mes motivations pour m'intéresser à la statistique.

Ma première motivation vient d'Orsay. Le premier enseignement de probabilités a été donné par Jacques Neveu en 1967 à l'occasion d'une réforme de la licence, très contestée à l'époque. Nous en avions établi un programme à Orsay que nous avions appelé PLA2 : probabilités, logique, algèbre et analyse. Comme nous n'avions pas de probabiliste ni de logicien à Orsay, nous avions fait appel à Neveu et à Krivine. La suite a été la nomination de Didier Dacunha-Castelle et de Jean Bretagnolle, et la constitution d'une équipe de probabilités qu'ils ont très rapidement transformée en équipe de probabilités et statistique. La statistique s'est développée à Orsay de belle manière, avec Pascal Massart en particulier, et j'ai eu l'occasion de voir que certaines grandes questions de la statistique étaient de grandes questions mathématiques. D'autre part recherche et enseignement étaient très liés aux applications à la biologie, avec une interaction forte avec l'INRA. En matière d'enseignement, une collaboration exemplaire s'est développée entre statisticiens et biologistes pour assurer conjointement certains cursus de biologie ; Elisabeth de Turckheim, qui assurait cette collaboration comme statisticienne, a été l'une des inspiratrices de l'étude de la Commission internationale de l'enseignement mathématique, ICMI, sur les mathématiques comme discipline de service. A cette époque, le milieu des années 80, je présidais cette commission, et le sujet même de l'étude reposait sur le rôle des mathématiques dans la formation des physiciens, des ingénieurs et des biologistes. En matière de biologie, la statistique était en première ligne, et le type d'enseignement assuré par Elisabeth de Turckheim conserve, je crois, sa valeur exemplaire.

Ma seconde motivation vient de la petite collection dont j'étais responsable aux PUF au début des années 70. Les premiers livres de la collection, le cours d'arithmétique de Jean-Pierre Serre et la théorie axiomatique des ensembles de Jean-Louis Krivine, avaient connu un grand succès. Je désirais élargir le spectre de la collection vers les mathématiques appliquées,

© Société Française de Statistique (SFdS), Décembre/December 2010

et j'ai fait appel à Jean-Pierre Raoult pour la statistique. Son livre, structures statistiques, est très original : il s'efforce de mettre de l'ordre dans la forêt de la statistique, sans perdre le sens et la saveur des problèmes divers auxquels elle répond. Mais pour autant il donne accès à de beaux théorèmes ; je pense aux liens entre statistique et analyse convexe illustrés par David Blackwell, ce grand mathématicien américain qui vient de mourir, centenaire, et qui a été le premier noir à occuper une place importante en recherche parmi les mathématiciens américains ; le New York Times lui a consacré un bel article nécrologique.

Ma troisième motivation est la première dans l'ordre chronologique, mais au départ elle concerne moins directement la statistique. C'est ma fascination pour le processus de Wiener, que Wiener appelait « the fundamental random function » et que Paul Lévy, suivant Einstein, a exploré et popularisé sous le nom de « mouvement brownien ». C'est un objet mathématique fondamental, comme l'indiquait Norbert Wiener, et d'une richesse inépuisable ; les médailles Fields à Wendelin Werner et à Stanislas Smirnov attestent la vigueur des concepts qui lui sont attachés. Je me suis aperçu tardivement qu'on le trouve sous-jacent à la statistique contemporaine, comme la courbe en cloche était sous-jacente à la statistique du 19^{e} siècle. Le cas typique est la modélisation des cours de la Bourse par Bachelier en 1900, et celle des stratégies financières par Black et Scholes en 1973. J'avais été impressionné par l'usage du mouvement brownien dans les problèmes de recherche de points extrémaux sur des variétés, disons, du point le plus bas dans une variété plongée dans un espace euclidien. Cela consiste à remplacer la descente la plus rapide, qui aboutit bien à un trou, mais pas en général au trou le plus bas, par une descente bruitée, suivant une idée empruntée à la chimie ; la méthode s'appelle le recuit simulé, et elle m'a été révélée au cours des années 90 par Robert Azencott, avec ses résultats très fins sur la bonne manière d'opérer.

Mon élection tardive à l'Académie des sciences, en 1998, a été l'occasion d'aviver mon intérêt pour la statistique. Un rapport sur la statistique était en cours de préparation, sous la direction de Paul Malliavin, dans le cadre des rapports sur les sciences et les technologies établis par l'Académie des sciences à l'initiative de son secrétaire perpétuel Jean Dercourt. La règle est la présentation à l'Académie, pour adoption, par un membre qui n'a pas été partie prenante de la rédaction. J'ai été chargé de cette présentation, et ce que j'en ai dit me paraît toujours valable. Le rapport est riche d'informations sur la statistique et les statisticiens, avec des insuffisances et des lacunes qu'il m'incombait comme présentateur de signaler. Mais les recommandations qu'il exprime, et que je reprends intégralement pour conclure mon rapport, me paraissent très pertinentes et toujours d'actualité ; je cite :

« Les exigences de la société en matière de qualité des produits et de contrôle des risques accroissent la demande de traitements statistiques dans les entreprises.

Le groupe recommande aux entreprises et aux établissements d'enseignement supérieur de mettre en oeuvre une réflexion concertée sur les métiers de la statistique et sur les compétences qu'ils requièrent, afin d'adapter les formations aux besoins professionnels.

En France, à la différence d'autres pays européens, les citoyens n'ont pas une formation suffisante à la prise en compte du mode de pensée statistique. Pour améliorer cette situation, des initiatives récentes ont été prises dans le cadre d'une réforme de l'apprentissage des mathématiques dans l'enseignement primaire et secondaire.

Le groupe souligne l'opportunité de ces réformes et encourage les responsables de l'enseignement à en assurer la mise en œuvre, en particulier par un effort de formation initiale et continue des professeurs des lycées et collèges. »

© Société Française de Statistique (SFdS), Décembre/December 2010

Dans les initiatives dont parle le rapport, et dans celles qui ont suivi, vous reconnaissez sans doute le rôle de Claudine Schwartz. La part de la statistique dans l'enseignement progresse. Ce qui est en panne en France actuellement est la formation continue des professeurs.

Au cours de l'année 2010, l'Académie a accueilli comme membre associé David Donoho, qui est un pilier du département de statistique de l'Université Stanford aux Etats-Unis, et un novateur en statistique et en traitement des données. La proposition de l'élire venait de Wendelin Werner, et ce fut l'occasion pour moi d'entrer de plus près dans certains de ses travaux. Je vous ai dit au début de cette conférence que j'avais découvert récemment les rapports de la statistique avec les mathématiques qui me sont le plus familières. C'est en particulier de l'oeuvre de Donoho et de son élève le Français Emmanuel Candès qu'il s'agit.

Donoho avait montré la portée en statistique des ondelettes d'Yves Meyer. Je ne vous parlerai pas des ondelettes, mais d'un sujet bien différent qui touche à l'analyse harmonique et à la théorie de l'échantillonnage ; c'est une remise en cause des idées et pratiques bien établies concernant la dimension des échantillons nécessaires pour reconstituer un signal ou une image.

Au congrès international des mathématiciens à Madrid en 2006, Emmanuel Candès a donné une conférence invitée sur un sujet neuf qui s'appelle en anglais « compressive sensing » ou « compressed sensing », et qu'Yves Meyer propose de traduire par « échantillonnage parcimonieux ». On peut en donner une idée de bien des manières, tant il a d'aspects et d'applications. Je me borne au début de la présentation de Candès, parce qu'elle fait écho à ce que j'ai dit de Legendre et des moindres carrés. Pour Legendre, il s'agissait de déterminer au mieux un petit nombre de paramètres à l'aide d'un grand nombre d'équations, provenant des observations, et la méthode qui s'imposait était une approximation dans l'espace des suites de carré intégrable, l^2. Pour Candès la situation est inversée : il s'agit de reconstruire un vecteur appartenant à un espace de grande dimension, N, à partir de résultats de mesures dont le nombre, K, est petit par rapport à N. Qu'on pense à l'imagerie médicale pour avoir une illustration de cette situation. Au lieu d'être surabondant, le système d'équations est indéterminé. A priori, on ne voit pas comment en tirer une solution. Ici intervient une hypothèse vérifiée dans beaucoup de cas concrets : le vecteur recherché n'a qu'un nombre S de composantes non nulles, mais on ne sait pas lesquelles. Par exemple, il s'agit d'un signal possédant S fréquences significatives, et S est petit par rapport à N. Selon un remarquable théorème de Candès, Romberg et Tao datant de 2006, il suffit de choisir au hasard K coefficients de Fourier, K étant grand par rapport au produit $S \times \mathrm{Log}\, N$, pour retrouver exactement le signal avec une probabilité voisine de 1 (la probabilité dépendant du rapport $K/(S \times \mathrm{Log}\, N)$). Et comment faire pour retrouver le signal ? Ici vient une surprise : on dispose de K équations linéaires entre N variables ; elles définissent une variété affine dans l'espace R^N ; le point qui convient dans cette variété affine est celui qui est le plus proche de l'origine dans la norme de l^1, c'est-à-dire la somme des valeurs absolues de composantes. Il ne s'agit plus d'approximation dans l^2 , mais d'approximation dans l^1. Comme j'ai passé une bonne partie de ma vie à étudier les algèbres de convolution l^1 ou L^1, constituées de suites ou de fonctions sommables, j'apprécie naturellement leur retour sous un angle que j'étais loin de soupçonner. J'avais également étudié les séries de Fourier lacunaires, et je savais bien qu'elles étaient parfaitement définies par les valeurs que prennent leurs sommes sur un intervalle, si petit soit-il. Mais je n'avais aucune idée d'un algorithme permettant le calcul des coefficients à partir de la somme restreinte à un intervalle, sous la seule hypothèse de la rareté

© Société Française de Statistique (SFdS), Décembre/December 2010

des fréquences intervenant dans la série. Le thèorème de Candès, Romberg et Tao donne un procédé incroyablement efficace et élégant.

S'il y a une leçon à tirer de cette vision très partielle et partiale de la statistique, il me semble que c'est ceci. Les données statistiques fournies par les observations et les mesures dans tous les domaines aujourd'hui, de la médecine à l'astrophysique, sont une mine de problèmes pour la statistique comme science. Beaucoup de ces problèmes sont de nature mathématique. Beaucoup sont attaquables à partir de théories connues, mais quelques-uns renouvellent ces théories elles-mêmes ; ils sont une fontaine de jouvence pour les mathématiques. Dans la conception que j'ai des progrès à venir des mathématiques, ce qui vient d'ailleurs occupe une grande place, longtemps dissimulée aux mathématiciens français de ma génération. Fourier en est un annonciateur. Et j'ai plaisir à voir que l'analyse de Fourier, au contact avec la statistique, donne à cette conception de nouvelles justifications.

Il me reste à vous souhaiter bon travail. Il est bien nécessaire, mais pas évident du tout, de faire passer dans l'enseignement l'esprit de la statistique. Est-ce que c'est en la détachant du reste des mathématiques ou au contraire en la liant aux autres parties du programme mathématique ? La réponse ne va pas de soi. Mais vous concevez où va ma préférence.

Bon travail donc, et merci de votre attention.

© Société Française de Statistique (SFdS), Décembre/December 2010