L'@nalyse des données : histoire, bilan, projets, …, perspective

Jean-Paul BENZÉCRI

In memoriam : Pierre BOURDIEU

Pierre Bourdieu, né le 1er août 1930 à Denguin (Pyrénées-Atantiques), décédé à Paris le 23 janvier 2002.

Il y a un demi-siècle, Pierre Bourdieu et moi-même étions élèves à l'Ecole Normale Supérieure de la rue d'Ulm ; lui dans la section des Lettres, et moi en Sciences. Or, le savoir ne se sectionne pas ! C'est pourquoi, depuis lors, nous ne nous sommes jamais longtemps perdus de vue, l'un l'autre…

La dernière lettre reçue de Bourdieu me posait une question trop difficile : c'était, ou presque : qu'est-ce que l'Analyse des Données ? Je lui fis une réponse élusive.

Maintenant que mon ami n'est plus de ce monde, je lui dois de rassembler mes esprits, sinon pour lui répondre, au moins pour attester que je n'ai pas de réponse à donner qui me satisfasse.

Quand, en 1954-55, je séjournais à Princeton, il n'y avait d'ordinateur (on disait *computer*), ni à l'Université, ni à *l'Institute for Advanced Studies*. Un étudiant pouvait consacrer une année à fabriquer un calculateur analogique destiné à résoudre des problèmes techniques d'une catégorie particulière. Et il en était de même au Laboratoire de Physique de l'Ecole Normale Supérieure.

Je pensais que les ordinateurs ne pouvaient être que des merveilles ; mais que, bien peu y ayant accès, il était sage de recourir à des simplifications mathématiques radicales, afin de renfermer les calculs dans les limites du possible.

C'est ainsi que, de 1959 à 1960, étant au Groupe de Recherche Opérationnelle de la Marine Nationale, je ne me faisais pas de scrupule de réduire à une loi normale toute donnée multidimensionnelle, collaborant parfois avec un camarade pour des simulations graphiques.

Pourtant, quand, sur le projet de la Traduction Automatique des Langues Naturelles, linguistique, logique et mathématique entreprirent de collaborer en ayant l'ordinateur pour outil …, il apparut que, dans la voie frayée par Louis Guttman † et Chikio Hayashi †, le principe d'équivalence distributionnelle proposé par le linguiste Zelig Harris † devait régler l'analyse des données statistiques.

Alors, en donnant forme géométrique à cette analyse, on aboutirait à la recherche des axes principaux d'inertie d'un nuage de points munis de masse ; problème classique, en dimension 3, mais à traiter ici en une dimension, *n*, quelconque. Ce qui requiert, impérativement, des diagonalisations de matrices carrées *n* x *n*, calcul infaisable sans une machine, dès que *n* dépasse 3 (ou 4 …).

Vers 1963, diagonaliser une matrice 7 x 7, était, pour un laboratoire modestement équipé, une tâche considérable. Par la suite, la Classification Ascendante Hiérarchique demanda des calculs encore plus lourds que ceux des diagonalisations. Mais la puissance des machines croissant avec l'efficacité des algorithmes, notre carrière de statisticien se développa … ; en mettant au service d'ambitions croissant sans cesse, des techniques dont le progrès défiait tous les rêves !

Vers 2000, sur un micro-ordinateur tel que ceux offerts à la clientèle des marchés, on peut, en quelques minutes, classer plusieurs milliers d'individus. Plus exactement, il faut quelques minutes pour les algorithmes de classification et d'analyse factorielle … Mais la conception des données, leur mise en forme, l'examen des résultats prennent non seulement des heures mais des mois …

Il n'y a plus, à strictement parler, de problème de calcul ; mais le problème même de l'Analyse des données subsiste ; d'autant plus vaste que, le calcul ne mettant point de borne à la recherche, on n'a point d'excuse pour s'arrêter dans la collecte des données et la méditation. Relativement à 1960 …, le rapport de difficulté, entre projets intellectuels et calculs, est inversé.

Il s'en faut de beaucoup que les principes qui nous paraissent s'imposer soient admis de tous.

Quant à la philosophie des nombres, la distinction entre qualitatif et quantitatif ne nous semble pas être toujours bien comprise. En bref, il ne faut pas dire :

- grandeur numérique continue ≈ donnée quantitative ;
- grandeur à un nombre fini de modalités ≈ donnée qualitative ;

car au niveau de l'individu statistique (e.g. le dossier d'un malade), une donnée numérique : l'âge ou même : la pression artérielle ou la glycémie, n'est généralement pas à prendre avec toute sa précision, mais selon sa signification ; et, de ce point de vue, il n'y a pas de différence de nature entre âge et profession.

Et surtout, pour comparer un individu à un autre, il faut considérer, non deux ensembles de données primaires, par exemple deux ensembles de 100 nombres réels, un point de IR^{100}, à un autre point de IR^{100}, entre lesquels des ressemblances globales ne se voient pas,

mais la synthèse de ces ensembles aboutissant à quelques gradations, ou à des discontinuités, à des diagnostics …

Quant au calcul, les algorithmes d'analyse factorielle (dont on a dit que le coût numérique est celui d'une diagonalisation de matrice) et de classification ascendante hiérarchique, jouant sur des données codées suivant le principe global d'équivalence distributionnelle (de profil), l'emportent en efficacité sur le jeu des contiguïtés entre individus, pris en compte par les algorithmes d'approximation stochastique, souvent effectués, aujourd'hui, suivant le schéma des réseaux de neurones.

Tel est le Monde vu par un statisticien géomètre après quarante ans de pratique.

Est-il permis d'assimiler le monde à ce qu'on a vu et vécu ? Prétention commune, aussi condamnable que le refus de rêver – au moins (faute de mieux) - sur l'avenir !

Premièrement : la question de reconnaître l'ordre de ce qui est dans les éléments que saisissent les sens (ou les outils qui arment les sens) est peut-être celle-même de la Philosophie dans toute sa noblesse. On a dit, en latin … que toute connaissance commence par ce qui est sensible dans les objets de la nature : *« Omnis cognito initium habet a naturalibus… vel : a sensibilibus* ». Au-delà de cette connaissance, il n'y a que le mystère ; et la révélation même, donnée par Dieu, se médite, à l'exemple de ce qu'on a pu connaître par le jeu naturel de la raison.

Il faut ici que le statisticien, le géomètre, le sociologue soient modestes ! En cherchant ce qu'on a toujours dû chercher, chaque génération ne peut avoir fait plus que sa part : la question subsiste.

Deuxièmement : on voit sur l'exemple des mathématiques, que le calcul nouveau dont la vitesse dépasse celle du calcul, de notre génération dans un rapport aussi inimaginable aujourd'hui que ne l'était, il y a un demi-siècle, le rapport de la complexité des calculs que nous avons pu faire à celle des calculs antérieurs … on voit, dis-je, qu'il ne faut pas trop vite affirmer, d'une manière méprisante, que la pensée ne peut que devenir paresseuse quand l'outil devient plus puissant.

D'une part, afin de résoudre des problèmes de calcul intégral, on a inventé des « fonctions spéciales » ; et, chemin faisant, on a créé l'analyse des fonctions de variable complexe (ou, du moins, approfondi cette analyse). De même, pour l'intégration des équations aux dérivées partielles, laquelle demande la théorie des espaces fonctionnels. Aujourd'hui, tous les calculs pratiques semblent être réduits au jeu banal des méthodes les plus simples, sur des réseaux de points arbitrairement denses…

En somme, le problème pratique provoque (ou, du moins, aiguillonne) le développement des idées théoriques ; et le perfectionnement des outils rend paresseuse la spéculation théorique.

Cependant, le mouvement inverse existe aussi. On remarque des coïncidences et on donne à ces coïncidences forme de lois, mises en circulation, avant que le développement d'idées théoriques appropriées permette de démontrer ces lois.

La théorie des fonctions analytiques doit beaucoup au désir de démontrer le grand théorème de Fermat : $x^n + y^n = z^n$ n'a pas de solution pour des entiers $n > 2$.

Or Fermat n'a pu conjecturer qu'après avoir calculé … remarqué … essayé de renverser sa remarque ou de la confirmer ; et cela, avec les moyens de calcul de son temps.

Voici un exemple trouvé par internet : le résumé d'un article de Théorie physique des hautes énergies :

hep-th/9811173 19 Nov 1998 ; @ http://xxx.lanl.gov/

résumé qui intéressera ceux mêmes qui ne sont pas mieux avertis que moi de la théorie des fonctions zeta généralisées et de l'analyse des diagrammes de Feynman (de la Théorie quantique des Champs) faite en dénombrant des nœuds.

Au prix de calculs formidables, on aboutit au résultat que des fractions dont le numérateur, « a », peut être de l'ordre de un million, n'ont pas de dénominateur, « b », supérieur à 9.

Avec, à la fin du résumé cette conclusion.

Nos résultats sont sûrs, numériquement ; mais il semble bien difficile de les démontrer par l'analyse.

« Sûrs » étant à comprendre en termes statistiques : si l'on procède par tirage aléatoire, il n'y a pour ainsi dire pas de tels ensembles de fractions, etc. … présentant les propriétés arithmétiques qu'on a trouvées.

On sera peut-être encore plus surpris de trouver dans ce même domaine de la Physique des hautes énergies des conclusions telles que la suivante : le résultat de la présente étude ne fait aucun doute … Cependant, pour décider de telle autre question …., il faudrait des calculs mille fois plus complexes ; calculs présentement inabordables.

Mais comme, répétons-le, la rapidité des instruments de calcul a, en un demi-siècle, été plusieurs fois multipliée par mille, cette conclusion n'est pas à prendre ironiquement.

Présentement, la physique théorique des hautes énergies progresse, principalement en constituant des corpus de phénomènes rares parmi des ensembles immenses de cas ordinaires. La seule observation d'un de ces cas ordinaires demande des appareils de détection où jouent simultanément des millions de petits détecteurs élémentaires…

Finalement toute la physique est subordonnée au progrès de cette branche particulière, de physique du solide à très basse énergie, qui produit les outils de détection et de calcul. [Les « puces », dont le nom est connu de tous, n'étant que la moindre chose en la matière…].

Quant aux problèmes de l'avenir dans les domaines mieux connus de tous que la physique théorique et l'analyse des fonctions zeta … mais interdits jusqu'à présent à tout traitement numérique satisfaisant : analyse des images ou même seulement des sons, de la musique et de la parole ; configurations météorologiques saisies globalement dans toutes leurs dimensions. Alors que, de par le seul fait des mouvements du terme, il n'y a là pas de population statistique, au sens usuel du terme, puisque chaque jour, de par les variables astronomiques, ne peut différer des autres jours qui le suivent ou le précèdent … voici ma première impression.

Les praticiens foncent dans les analyses, les transformations de Fourier d'images, etc. sans savoir ce qu'ils cherchent.

Je suis assez satisfait des idées que je me suis faites sur la parole (même si j'ai plutôt un programme de recherches que des résultats suffisants : voir « Analyse spectrale et analyse statistique de la voix humaine parlée », *Les Cahiers de l'Analyse des Données*, Vol. XIII, 1988, n°1, pp. 99-130). Je dois avouer (ne le dites à personne, je vous en prie !) que l'analyse des données n'y est pour rien. Ce qu'il faut, ce à quoi je me targue d'avoir quelque peu réussi… c'est voir ce qui dans les objets étudiés, en l'espèce des sons, est pertinent. Voilà ce dont on doit d'abord s'enquérir dans tous corpus de dimension et de complexité « astronomiques ».

Je le répète : le statisticien doit être modeste …. Le travail de ma génération a été exaltant … une nouvelle analyse est à inventer, maintenant que l'on a, et parfois à bas prix, des moyens de calcul dont on ne rêvait même il y a trente ans … Même si les voies explorées, jusqu'ici dans certains domaines sur lesquels il ne serait pas charitable d'insister, offrent à notre ironie une facile matière …

A la mémoire de Pierre Bourdieu, je devais présenter ces excuses pour n'avoir pas répondu à sa dernière lettre.

Moi qui dois tant à sa familière et indulgente amitié,

et suis pénétré de respect pour l'exemple que nous laisse la leçon, faite le 27 mars 2001, au terme de son enseignement au Collège de France.

Jean-Paul Benzécri

.

OUT–4102–76
CECM–98–120
hep-th/9811173
19 November 1998

Determinations of rational Dedekind-zeta invariants of hyperbolic manifolds and Feynman knots and links

J. M. Borwein[a)] and D. J. Broadhurst[b)]

Abstract We identify 998 closed hyperbolic 3-manifolds whose volumes are rationally related to Dedekind zeta values, with coprime integers a and b giving

$$\frac{a}{b}\,\mathrm{vol}(\mathcal{M}) = \frac{(-D)^{3/2}}{(2\pi)^{2n-4}}\,\frac{\zeta_K(2)}{2\zeta(2)}$$

for a manifold $\mathcal{M}$ whose invariant trace field K has a single complex place, discriminant D, degree n, and Dedekind zeta value $\zeta_K(2)$. The largest numerator of the 998 invariants of Hodgson Weeks manifolds is, astoundingly, $a = 2^4\times 23\times 37\times 691 = 9,408,656$; the largest denominator is merely $b = 9$. We also study the rational invariant a/b for single-complex-place cusped manifolds, complementary to knots and links, both within and beyond the Hildebrand–Weeks census. Within the censi, we identify 152 distinct Dedekind zetas rationally related to volumes. Moreover, 91 census manifolds have volumes reducible to pairs of these zeta values. Motivated by studies of Feynman diagrams, we find a 10-component 24-crossing link in the case $n = 2$ and $D = -20$. It is one of 5 alternating platonic links, the other 4 being quartic. For 8 of 10 quadratic fields distinguished by rational relations between Dedekind zeta values and volumes of Feynman orthoschemes, we find corresponding links. Feynman links with $D = -39$ and $D = -84$ are missing; we expect them to be as beautiful as the 8 drawn here. Dedekind-zeta invariants are obtained for knots from Feynman diagrams with up to 11 loops. We identify a sextic 18-crossing positive Feynman knot whose rational invariant, $a/b = 26$, is 390 times that of the cubic 16-crossing non-alternating knot with maximal D_9 symmetry. Our results are secure, numerically, yet appear very hard to prove by analysis.

[a)] CECM, Simon Fraser University, Burnaby, B.C. V5A 1S6, Canada;
jborwein@cecm.sfu.ca; http://www.cecm.sfu.ca/~jborwein
[b)] Physics Department, Open University, Milton Keynes MK7 6AA, UK;
D.Broadhurst@open.ac.uk; http://physics.open.ac.uk/~dbroadhu