Statistique et analyse des données

G. Morlat

De vieux outils pour appréhender l'avenir

Statistique et analyse des données, tome 1, n° 1 (1976), p. 3-11.

<http://www.numdam.org/item?id=SAD_1976__1_1_3_0>

D E B A T .....

S T A T I S T I Q U E E T A N A L Y S E D E D O N N E E S

DE VIEUX OUTILS POUR APPREHENDER L'AVENIR...

Par G. MORLAT

A l'heure où fleurissent de nouvelles techniques pour nous aider à comprendre,dominer,influencer l'évolution de nos sociétés : prospective, analyse de systèmes, évaluation technologique - et d'aucuns n'hésitent guère à ajouter, comme dans un inventaire de Prévert : méthodes Delphi,Monte Carlo, impacts croisés,analyses multifactorielles,brainstorming..... - il n'est pas sans intérêt de jeter en arrière un long regard - sur trois siècles - cinquante ans, dix ou quinze ans - sur l'histoire de la probabilité et de la statistique. Certes, les disciplines classiques doivent être remises en cause, cela ne peut qu'être fécond. L'analyse des systèmes permettra aux chercheurs de la jeune génération d'établir de meilleurs modèles de la structure, du fonctionnement et des influences réciproques du monde naturel, des sociétés humaines et des outils techniques les plus variés que celles-ci se sont donnés. Pour étudier efficacement le monde où nous vivons, la classification des sciences d'Auguste Comte s'est avérée depuis longtemps insuffisante: bien sûr il faut toujours faire des classifications,mais il faut toujours aussi les réviser. Cela dit, il n'est peut-être point encore temps de considérer comme périmés, ou frappés de désuétude, les outils incomparables que sont le calcul des probabilités et la statistique. Il serait au contraire extrêmement fâcheux d'oublier que probabilité et statistique ne sont pas encore passées dans nos moeurs. Il y a plus de cinquante ans qu'Emile Borel écrivait de manière fort pertinente : " Le calcul des probabilités est une des branches les plus attrayantes et les moins ardues de la mathématique. C'est simplement pour des raisons de tradition, l'on n'ose écrire de routine, que les éléments de ce calcul ne figurent pas aux programmes de l'enseignement secondaire, où ils remplaceraient avantageusement bien des matières qui y subsistent pour le seul motif que personne ne se donne la peine de les supprimer."

Son appel n'a guère été entendu : il est resté longtemps parole dans le désert, et si l'on a considérablement élagué dans notre enseignement des proliférations de la géométrie, on a cherché à mettre à la place des mathématiques isolées, pas assez reliées au monde et aux problèmes concrets.Parmi d'autres chapitres des mathématiques appliquées,le calcul des probabilités et la statistique devraient certainement contribuer, avec les sciences expérimentales mises à jour, la technologie (y compris manipulation d'une lime, d'un fer à souder et d'une truelle), le maniement des ordinateurs, et les sciences humaines, à former ce tronc commun d'une culture mieux adaptée à notre temps que les lettres grecques et latines, détrônées mais non remplacées.

L'ORIGINE AMBIGUE DES PROBABILITES ET LES CONTROVERSES SUBSEQUENTES

Sur les sources de calcul des probabilités au dixseptième siècle avec PASCAL, le chevalier de MERE, et d'autres gentilshommes, beaucoup a été écrit; que les jeux de hasard, et notamment les jeux de dés,aient été l'occasion, voilà une grande source d'ambiguïté : il s'agissait bien d'apprendre à parier, à jouer au mieux, et même si un homme n'avait eu le droit, de par un décret royal, de jouer qu'un seul coup aux dés dans son existence, il ne semble pas que les arguments de PASCAL en eussent été diminués de quelque façon. Et cependant, lorsque Jacques BERNOUILLI eût établi plus tard la loi des grands nombres, d'autres retinrent que la fréquence d'un résultat observé dans une suite d'expériences assez longue,pouvait donner une idée de la probabilité de ce résultat. D'où la tentation de parler de probabilités " objectives". D'où les controverses qui divisèrent environ deux siècles plus tard,probabilistes et statisticiens Les uns pensaient, avec Von Mises, que la probabilité trouve sa seule justification dans la possibilité (au moins théorique) d'observer les résultats d'une longue suite d'expériences exécutées dans des conditions homogènes. D'autres étaient convaincus que c'était là un domaine trop restreint, et que la notion de probabilité était en droit de mesurer le degré de confiance d'un sujet donné vis-à-vis d'une éventualité incertaine pour lui (c'est la probabilité subjective). Pour les premiers,on peut parler de la probabilité de pile avant de lancer une pièce de monnaie en l'air, et sans doute aussi pendant qu'elle tournoie, mais on n'en a plus le droit dès lors qu'elle est à plat sur la table, même si vous n'avez point encore vu quel côté elle montre, parce que je l'ai recouverte de ma main.

Mais, pour les seconds, la probabilité de " pile" conserve tout son sens, si j'ai posé avec précaution une pièce de monnaie, là, sur la table sans la lancer, et en choisissant délibérément le côté qui paraîtra,alors que vous, du fond de la salle, ne pouviez le discerner. Que voilà une byzantine querelle,serions-nous tentés de penser aujourd'hui. Mais pourtant, nous sommes bien incapables d'apprécier combien de personnes peuvent être troublées par ces difficultés, en fonction de la manière dont elles ont été enseignées pour ce qui regarde les probabilités. La présentation axiomatique donnée par KOLMOGOROFF vers 1935 aurait dû en bonne logique, faire disparaître le problème comme par enchantement : force est de constater que ce ne fut point le cas.

L'AXIOMATISATION ET LA PROBABILITE ABSTRAITE.

Bien que POTERIN du MOTEL eut écrit, quelque vingt ans plus tôt, des définitions sensiblement équivalentes, c'est dont à KOLMOGOROFF que revient le mérite d'avoir été entendu, lorsqu'il proposa de définir à peu près comme suit la probabilité :

Soit une famille de parties d'un ensemble E, qu'on appellera des

évènements; si cette famille contient l'ensemble vide et si elle est fermée pour les opérations de réunion, d'intersection et de complémentation, on dira que c'est une algèbre d'évènements (si de plus elle est encore fermée pour des réunions dénombrables, ce sera une σ-algèbre, et les choses seront plus simples ultérieurement, mais on peut considérer cela comme un point technique).

Dès lors, on peut, sans courir le risque de rencontrer plus tard des paradoxes sur son chemin, définir une application p de notre famille (algèbre) d'évènements dans le segment (0,1) de la droite réelle vérifiant les conditions :

$$p(\emptyset) = 0 \quad p(E) = 1$$

$$p(A \cup B) = p(A) + p(B) \quad \text{si } A \cap B = \emptyset$$

c'est tout simplement une telle application p qu'on nomme une mesure de probabilité, ou plus brièvement une probabilité (abstraite).

A partir de là, on peut développer une théorie mathématique simple et élégante, qu'on nomme le calcul des probabilités. Mais on ne se souciera guère, à ce niveau, de savoir à quels objets du monde réel peut s'appliquer cette théorie.

LES INTERPRETATIONS CONCRETES

Il faudra bien pourtant s'en soucier si l'on veut se servir de l'outil, et l'on n'aura pas grande peine à découvrir de nombreux phénomènes concrets, dont la probabilité au sens de KOLMOGOROFF fournit un modèle adéquat. Parmi bien d'autres, en voici trois catégories :

Il y a d'abord la comptabilité d'un ensemble quelconque d'objets ou de matière ("counting measure"). Si l'on prend pour unité la surface du territoire de la FRANCE, et pour évènements les divers ensembles concevables de propriétaires fonciers (les ménages, les administrations, les personnes physiques de plus de trente cinq ans, ma concierge, ...) alors l'application qui fait correspondre à chacun de ces "évènements" (propriétaires ou groupes de propriétaires) la surface de son patrimoine, est bien une probabilité, comme on le vérifiera sans peine. Nous l'appellerons une probabilité comptable. Cette notion pourrait aussi bien s'appliquer au poids de viande des bovins vivant sur un territoire, ou à la proportion des citoyens français appartenant à tel ou tel ensemble de catégories professionnelles, etc....

On citera ensuite la probabilité-fréquence, couramment appelée de façon un peu abusive, probabilité objective; c'est la notion à laquelle voulaient naguère se restreindre une classe importante de probabilistes et de statisticiens.

En troisième lieu, la probabilité subjective, ou vraisemblance.

On verra plus loin que la théorie de la décision dans l'incertain permet de dire que c'est là une probabilité si leschoix que peut effectuer un sujet sont cohérents dans un sens très précis. En présence d'une famille de propositions incertaines,toute probabilité vérifiant les axiomes de KOLMOGOROFF peut être valablement prise en considération.

Par exemple, si je possède un jeu de trente deux cartes, dont j'extrais l'une d'elle, et que je m'intéresse seulement aux diverses propositions

" c'est un roi "

" c'est un coeur "

et à celles que l'on obtient par les opérations logiques de réunion ou de complémentation, on constate aisément qu'on est en présence d'une algèbre comprenant seize évènements. Les atomes (évènements élémentaires) étant au nombre de quatre (partition du jeu de 32 cartes) :

- roi de coeur
- roi de trèfle carreau ou pique
- coeur, sauf le roi
- autres cartes

Attribuer la probabilité 1/4 à chacune de ces éventualités constitue une option parfaitement légitime : c'est une probabilité subjective au même titre que n'importe quelle autre application numérique vérifiant les axiomes de KOLMOGOROFF (Savoir si l'on a quelque intérêt à utiliser une telle probabilité lorsqu'on joue au poker est une autre question).

Nous sommes ainsi en présence de trois interprétations concrètes de la probabilité : comptable, objective,subjective.

En présence d'une théorie statistique particulière,il peut être utile de se demander quelle est l'interprétation de la probabilité qui est sous-jacente à cette théorie. C'est d'abord cette question que nous nous poserons en passant en revue, dans ses grandes lignes, l'histoire des theories statistiques.

LA PERIODE ANTIQUE

En statistique, le découpage de l'histoire en périodes antique, classique et moderne, s'éloigne quelque peu de la convention couramment admise pour ce qui concerne la litterature,l'art ou les civilisations : la période antique de la statistique, en effet, va jusqu'à l'an 1900, très précisément. Cette période qu'on peut faire commencer au choix avec les philosophes de l'antiquité (au sens ordinaire) ou avec PASCAL et ses contemporains, est caractérisée par le fait qu'on ne distinguait pas toujours clairement entre ce qui est du calcul des probabilités, et ce qui est de la statistique - bien que les deux disciplines se fussent d'abord développées en s'ignorant à peu près complètement:

le calcul des probabilités avec des mathématiciens, puis des physiciens spécialement au dixneuvième siècle, une certaine forme de statistique avec les démographes et les actuaires, plus tard des économistes.

La distinction que nous faisons ici entre calcul des probabilités et statistique est la suivante : toute démarche purement déductive est du ressort du calcul des probabilités - tout raisonnement qui part des données observées appartient au domaine de la statistique.

Ainsi donc, les démographes et les actuaires qui construisaient et utilisaient des tables de mortalité, faisaient incontestablement une certaine forme de statistique - mais la distinction n'apparaissait guère pour d'autres - tels CONDORCET, LAPLACE, plus tard QUETELET - qui traitaient des problèmes d'inférence comme de simples applications du calcul des probabilités,utilisant parfois avec une certaine hésitation d'ailleurs, la formule de BAYES . L'interprétation donnée à la probabilité était bien entendu subjective. On a même pu considérer qu'un certain abus de ces applications avait conduit à l'époque de QUETELET, à discréditer l'usage des probabilités dans le raisonnement inductif. C'est d'ailleurs sans doute l'emploi inconsidéré de la formule de BAYES qui a conduit des scientifiques vers la fin du dixneuvième siècle, à la recherche des méthodes qui ne donneraient pas lieu à discussion, concernant le choix des probabilités a priori - et qui devaient donc conduire aux mêmes résultats tous les observateurs.

LA STATISTIQUE CLASSIQUE

Dans cette voie, une étape fut franchie - et avec quel brio - par Karl PEARSON, publiant en 1900, dans le Philosophical Magazine, son célèbre mémoire sur le test de Khi Deux,dont le titre vaut d'être cité intégralement : " on a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling ".

C'est le point de départ de la période classique de la statistique, un peu plus d'un demi-siècle,puisqu'il convient d'arrêter cette période entre 1950 et 1960 - où l'on a vu se développer avec un succès considérable, et à peu près exclusivement dans les pays anglo-saxons une statistique qu'on disait tout naturellement à cette époque là, " moderne ".

Ce qui caractérise cette " statistique mathématique" classique, c'est d'abord que ses rapports avec le calcul des probabilités sont définitivement établis (même si la notion de probabilité n'est pas encore au début du siècle sortie des controverses) :

le domaine propre de la statistique, c'est la confrontation entre des observations et des lois de probabilité, qu'il s'agisse de préciser laquelle convient le mieux (problème d'estimation) ou de se demander s'il faut rejeter celle qu'on avait préalablement retenue,pour en adopter une autre (test d'hypothèse).

une autre caractéristique de la statistique classique,c'est le choix exclusif de l'interprétation de la probabilité comme fréquence. Ce choix ne s'est pas imposé d'emblée à tous les statisticiens (créateurs)de l'époque - témoin les efforts de R.A. FISHER pour prôner la " probabilité fiducielle ",

sorte de vraisemblance attachée aux paramètres, et issue des seules observations, ne résultant donc pas d'une probabilité a priori. La plupart des statisticiens anglais et américains n'ont pas suivi FISHER dans cette voie - et J. NEYMAN a sans doute été le premier à proclamer qu'à ses yeux, le seul emploi légitime de la probabilité c'était le concept de fréquence, ou de probabilité objective. La quasi totalité des statisticiens de l'époque a suivi ce point de vue : on peut reconnaître que les techniques élaborées à ce moment, et les succès pratiques auxquels elles ont conduit (statistique industrielle, contrôle des fabrications, recherche agronomique et biologique, etc...) confortaient pleinement cette attitude. Mais, ces techniques formaient un arsenal quelque peu hétéroclite, et le besoin d'une synthèse conduisit, dans les années quarante, Abraham WALD à proposer le couronnement de la statistique classique, avec la théorie des fonctions de décision. statistiques. A vrai dire l'objectif était double : proposer cette théorie générale, dont les techniques qui venaient d'être développées durant cinquante ans apparaîtraient comme des cas particuliers - et intégrer du même coup les techniques séquentielles, développées avec succès par le même auteur pour répondre à des problèmes posés par des fabrications de guerre aux ETATS-UNIS.

Sur ces deux points la théorie proposée par WALD a pleinement réussi. Mais elle a également atteint un autre objectif, qui n'était sans doute pas recherché par son auteur. En effet, cherchant à caractériser des " fonctions de décision admissibles", WALD retrouve la famille des " fonctions de décision de BAYES " - c'est-à-dire celles que le statisticien utiliserait s'il était disposé à se donner une distribution de probabilité a priori des paramètres intervenant dans le " modèle statistique" - et si de plus son objectif était de minimiser le coût moyen des conséquences de ses décisions. Bien sûr,WALD a insisté sur le fait que les probabilités a priori(subjectives!) n'ont pour lui aucun sens, et qu'elles ne constituent qu'un instrument technique commode, pour la sélection des meilleures règles de décision. Il n'en reste pas moins que le résultat général qu'il a établi, concernant le " privilège des règles bayesiennes ", a pu contribuer à convaincre quelques statisticiens, d'abandonner l'ostracisme auquel étaient vouées, pendant la période classique, les méthodes bayesiennes.

LES THEORIES DE LA DECISION DANS L'INCERTITUDE

A vrai dire, un petit nombre de statisticiens étaient toujours restés bayesiens et le succès de la statistique classique ne les avaient pas ébranlés : le plus notable est certainement de FINETTI, dont les thèses sur la probabilité subjective ont été reprises par les théoriciens de la décision : une autre composante de ces théories se trouve dans l'oeuvre de VON NEUMANN, concernant les jeux et le comportement économique. En simplifiant beaucoup, on peut dire que de FINETTI avait montré que si les conséquences de nos décisions sont mesurables en termes d'utilité numérique, alors la probabilité (subjective) résulte d'une exigence de cohérence, de " rationalite ", de nos décisions.

Von NEUMANN a montré que si l'on admet l'existence de probabilités pour décrire les évènements incertains, alors de même, des règles de rationalité permettent d'établir l'existence d'une fonction numérique d'utilité. Il revenait à SAVAGE ("Foundation of Statistics")d'aller plus loin) et de montrer que si l'on peut décrire un problème de décision dans l'incertain en termes d'états (de la nature), de conséquences (pour la personne qui choisit) et d'actes (applications de l'ensemble des états dans l'ensemble des conséquences), alors un nombre réduit de postulats de cohérence, ou de rationalité concernant les préférences entre les actes, permettent d'établir simultanément l'existence de la probabilité (subjective), de la fonction d'utilité sur les conséquences et la règle de la maximisation de l'espérance mathématique des utilités.

LES TENDANCES MODERNES DE LA STATISTIQUE

Nous avons vu très brièvement plus haut ce que furent en statistique la période de l'antiquité (jusqu'à 1900) et la période classique (1900-1950). Il nous reste à dire quelques mots des tendances de la statistique dans les dernières vingt cinq années.

On ne saurait prétendre que rien n'a été fait dans l'esprit de la statistique mathématique classique : une bonne part du contenu des "Annals of mathematical statistics " témoigne d'une activité intense, sur la lancée des statisticiens classiques. Certains travaux apparaissent quelque peu académiques,transposant dans un langage mathématique plus abstrait, ou généralisant des résultats acquis 10 ou 20 ans plus tôt - d'autres éclairent des zônes d'ombre qui subsistaient. Quelques chapitres importants se développent cependant, comme la robustesse, l'estimation des distributions de probabilité, ou l'étude des valeurs extrêmes. Et sans doute de tels travaux permettront de mieux répondre aux questions que posent de nombreux praticiens.

Une tendance plus spécifique de ce récent quart de siècle réside dans le développement des méthodes " néo-Bayésiennes". Inspirées peut-être par le théorème central établi par Abraham WALD (dont ce n'était guère l'objectif, nous l'avons vu, en établissant les privilèges formels des règles de décision Bayésiennes), confortés par les résultats observés par la théorie de la décision dans l'incertain, un certain nombre de statisticiens sont revenus aux techniques Bayésiennes de la période antique. On peut dire qu'ils prônent les techniques anciennes, la différence étant qu'ils savent mieux pourquoi ils le font. Ils ont à la fois de ce fait plus de liberté dans le choix des distributions a priori, et de meilleurs arguments pour les justifier. Ces méthodes néo-Bayésiennes, qui prennent donc au départ le contrepied des motivations des statisticiens classiques (éliminer comme entachées de subjectivisme les probabilités a priori) ont été particulièrement pronées et utilisées par des statisticiens économistes, tels Rober SCHLAIFER à HARVARD et Jacques DREZE à LOUVAIN, parmi beaucoup d'autres.

On peut noter ici que dans les universités françaises, l'enseignement donné aux économistes, s'appuie encore, et parfois exclusivement, sur la statistique classique, que la génération actuelle d'enseignants a eu l'occa -

sion d'apprendre à l'époque où elle apparaissait comme moderne.Et les méthodes Bayésiennes ne sont citées que comme une curiosité, ou un appendice méthodologique auquel on consacre peu de développements. Il ne s'agit là bien sûr que d'un trait général de nos enseignements, l'Université d'AIX MARSEILLE constituant une exception remarquable, où les méthodes néo Bayésiennes sont en honneur.

Si l'on suit l'argument des néo Bayésiens, selon lequel la statistique classique n'a pas de bases logiques solides - ou plutôt n'a pas de bases logiques du tout, puisque, selon de FINETTI, " on a retiré le sable (les probabilités a priori) et la statistique a été construite sur rien" - alors on ne doit pas manquer de se poser une question troublante,concernant les raisons pour lesquelles la statistique classique a réussi, et réussit encore souvent, à donner à de nombreux praticiens des réponses satisfaisantes. SAVAGE, après Etienne HALPHEN, répond qu'il faut en chercher l'explication dans le fait que la distribution de probabilité a posteriori d'un paramètre devient rapidement à peu près indépendante de sa distribution a priori, dès que les observations sont un peu nombreuses, et pourvu que la distribution a priori, ne soit pas trop extraordinaire.
Nous ne sommes pas sûr que cela suffise, car la statistique classique réussit souvent en présence d'observations peu nombreuses. Il faudrait se demander par quel mécanisme s'établit la décision de traiter par une méthode statistique (classique) un problème qui s'y prête bien. Qualitativement, il n'est pas difficile d'entrevoir ce que peut être ce mécanisme. Mais, une étude empirique de la pratique statisticienne pourrait tenter un jour quelque observateur.

L'ANALYSE DES DONNEES

Sur les tendances modernes de la statistique, il reste à dire l'essentiel : c'est que vers 1950 sont apparus des instruments de calcul d'une puissance telle que les rêves les plus fous des statisticiens du passé, (par exemple analyses factorielles portant sur des centaines,voire des milliers de variables) se sont transformés en quelques années en une pratique quotidienne.

C'est vers 1955 qu'on avait pu lire sous la plume de John TUKEY, un article prophétique - qui avait paru quelque peu obscur comme il convient à une prophétie - " The future of Data Analysis ".

Aujourd'hui, l'obscurité est dissipée, et tous les statisticiens savent que dans les domaines où peut être recueillie une information abondante, rien ne s'oppose à l'examen simultané de toutes les liaisons qui peuvent apparaître : examen qu'on nomme traditionnellement " analyse des correspondances ", mais qui possède plutôt les traits distinctifs d'une synthèse, comme l'observe judicieusement Jean Paul BENZECRI.

On sait aussi que le riche arsenal des techniques de traitement des données multidimensionnelles (description par l'analyse en composantes principales, regression, correspondance,classification, analyse discriminante, etc...) permet d'extraire des observations des structures cachées, ou de préciser des structures soupçonnées ou connues par ailleurs.

Certaines de ces techniques, comme la régression ou l'analyse descriminante,peuvent répondre directement à des fins de prévision ou de décision.

Affaiblissement des hypothèses et des modèles posés a priori, rôle plus grand des observations plus nombreuses et plus complexes prises en compte, tout ceci, permis par les capacités de calcul disponibles fait que l'ensemble " ordinateur-techniques d'analyse des données" constitue un instrument d'observation du monde réel, d'une puissance notablement accrue. On lira le récit passionnant que donne Jean Paul BENZECRI de l'" Histoire et Préhistoire de l'analyse des Données " (Cahiers d'Analyse des Données).

ET L'AVENIR ?

Tels sont les traits dominants de l'histoire de la statistique, qui a eu l'ambition de formaliser les méthodes pour décrire et mettre en ordre les faits observés, découvrir ou mettre à l'épreuve des lois auxquelles obéissent les phénomènes, prévoir leur évolution et fonder la-dessus de bonnes décisions. La formalisation était poussée fort loin - jusqu'à la prise de décision - dans le programme de la statistique classique; les techniques d'analyse de données qui prennent aujourd'hui le pas font la part meilleure aux faits observés, dans leur complexité; elles peuvent pousser moins loin la formalisation (et garder une efficacité accrue en se limitant au stade descriptif, avec le minimum d'hypothèses) - elles ne peuvent d'ailleurs pousser la formalisation aussi loin que le faisait la statistique classique, faute d'outils analytiques adéquats, maniables et assez généraux, en matière de modèles probabilistes multidimensionnels.

L'analyse des données prendra-t-elle un jour ce chemin ? Ce n'est pour l'instant qu'un sujet d'interrogation pour le statisticien, qui ne dispose guère, à propos de sa propre discipline, du recul nécessaire pour en apprehender l'avenir.