la pensée

REVUE DU RATIONALISME MODERNE

sciences arts philosophie

MODES DE CONNAISSANCE

MAI-JUIN 1981

N° 220

QUELQUES RÉFLEXIONS SUR LA STATISTIQUE

par Didier DACUNHA-CASTELLE *

DANS ce texte nous nous proposons de faire quelques remarques sur la statistique : un essai de définition, sa place dans la société, sa place dans ou à côté des mathématiques, sa position vis-à-vis de la notion de modèle, retiendront notre attention. Des notes et exemples plus techniques permettront, en fin de texte, au lecteur intéressé d'entrer un peu dans la formulation mathématique de certains problèmes.

STATISTIQUES ET DONNÉES

Le terme de statistique recouvre plusieurs pratiques techniques ou scientifiques. Pour tout un chacun, faire de la statistique c'est d'abord recueillir de manière organisée des données d'origines très diverses : médicale, économique, sociologique.

La forme des données est très variable : quantitative (revenu), qualitative (forme de maladie), ordinale (goûts), etc. Leur organisation est un problème technique important, les moyens informatiques classiques permettent de stocker un nombre très important d'informations diverses, à condition de les coder convenablement. Cette nouvelle possibilité explique à la fois le développement de la statistique et la tendance à privilégier certaines techniques de lecture automatique de données qui sont souvent, nous le verrons, d'une qualité douteuse.

L'organisation des données n'est pas du domaine du statisticien professionnel : gérants de supermarchés, zootechniciens, assureurs, ont l'habitude d'organiser des données assez complexes, en vue de les interpréter mais leur démarche est, au moins au niveau formalisé de routine, quoique, sans doute, ils aient une activité statistique intuitive ou peu formalisée.

Nous n'intégrerons pas, pour plus de clarté, dans l'activité statistique la simple organisation de données et leur gestion automatisée.

UNE DÉFINITION POSSIBLE DE LA STATISTIQUE

Nous appellerons statistique l'activité scientifique (ou technique) qui consiste à donner des diagnostics sur la validité de certains modèles mathématiques représentant des phénomènes divers. Les gens un peu savants, ajouteront volontiers que les phénomènes en question sont aléatoires et ils auront, je pense, quelquefois tort, c'est la technique mathématique qui introduit un passage par l'aléatoire, lorsque celui-ci n'est pas dans la modélisation du phénomène concret (voir plus loin l'exemple médical).

La statistique est avant tout critique de modèles mathématiques simples, dans des domaines allant de la biologie aux sciences humaines et aux processus de production industrielle. Souvent le modèle mathématique est construit à partir d'une théorie très complexe, non mathématisable à ce jour, c'est le cas de la biologie. Mais, pour nombre d'activités, le modèle mathématique est un instrument de travail, éloigné de toute théorie (économie appliquée, contrôle de processus, expérimentation agronomique, etc.). La critique du modèle mathématique est alors un pas utile dans la critique d'hypothèses a-mathématiques.

STATISTIQUE ET SOCIÉTÉ
PROVISIONS, DÉCISIONS, RISQUES

La statistique est sans doute la partie des mathématiques la plus au contact direct, par les médias, d'un large public.

Le point de vue immédiat sur la statistique dépasse celui du simple recueil de données, on est habitué à ce que l'INSEE fasse des prédictions (douteuses), la météorologie des prévisions (mauvaises), les instituts de sondages des sondages (quelquefois bons), etc., donc il y a des gens (pas nécessairement baptisés statisticiens) qui à partir de données diverses font des diagnostics sur certains types de problèmes.

La statistique est peu enseignée, et même les gens d'un très haut niveau mathématique n'en ont qu'une idée fort vague, quoique ses rudiments soient très accessibles. Si un mathématicien professionnel est capable d'assimiler assez rapidement un modèle probabiliste (encore que ceci soit relativement récent), il est totalement incapable, en général, d'expliquer comment l'INSEE fait ses prévisions qui sont ensuite utilisées, déformées à la télévision pour les besoins de la cause.

A un autre niveau, rien n'est fait en général pour que le citoyen soit apte à organiser quelque peu des données socioéconomiques pour en tirer un certain nombre de jugements et favoriser des pratiques autogestionnaires. Ceci est vrai actuellement dans tous les systèmes politiques, y compris les systèmes socialistes, où l'abondance des données chiffrées présentées par les médias aux citoyens, n'est pas, en général, accompagnée des techniques d'interprétation et surtout de la critique même que ces données appellent sur certains modèles de relations

entre les données. Cette remarque est sans doute claire pour ce qui est de la présentation des données économiques, qu'elles soient obtenues par recensement ou par sondage. Elle est fondamentale au niveau de l'instrument de démocratisation qu'est la statistique, élément de validation, donc de contestation de modèles à caractère idéologique, par exemple dans le domaine économique.

Mais des notions comme celle de risque en médecine ou dans le nucléaire nécessitent vraiment de donner à tout un chacun une idée assez précise, très quantitative de la notion de probabilité. Cette idée n'est vraiment naturelle qu'à partir d'une introduction statistique des notions décisives (et intuitives) de la théorie des probabilités, la statistique donnant l'habitude de calculer des probabilités en fonction de la valeur de certains paramètres. Les scénarios à la mode dans les milieux politologues ou des jeunes managers, les gros modèles économétriques de simulation sont des cas très particuliers de raisonnements « statistiques ». Une marque de développement des sociétés est de donner aux individus l'habitude de pondérer leurs décisions par des probabilités dépendant de certains paramètres, usuelle dans la vie de tous les jours : météo, embouteillages, assurances, cette habitude se traduit par les notions de prévision et de contrôle en milieu aléatoire. L'individu, les organisations sociales, l'état, diagnostiquent à partir de modèles comprenant de l'aléatoire. Sans prise sur le futur dans bien des domaines, les sociétés sous-développées n'ont pas ce comportement sauf à des niveaux très particuliers, et de toute manière il n'est jamais « quantitatif ».

LA DÉMARCHE. MODÈLE ET DONNÉES

La démarche statistique lorsqu'elle est formalisée peut se résumer ainsi :

Collection de modèles → recueil de données → calculs mathématiques → diagnostic sur la validité des différents modèles de la collecte.

Les étapes peuvent être ou ne pas être indépendantes, en général le recueil des données est *organisé* de façon à pouvoir faire un traitement mathématique conduisant à un diagnostic (voir plus loin l'exemple de l'agronomie). La statistique vise donc à étudier la validité d'un ou d'une classe de modèles. Il est donc important de comprendre ce que l'on entend par validation, et d'essayer maintenant d'analyser les méthodologies statistiques à cet effet.

Il s'agit de confronter modèle et données expérimentales.

Le modèle, généralement probabiliste, dépend de paramètres inconnus.

Valider un modèle est donc avant tout, indiquer les valeurs des paramètres, compatibles (en un sens à préciser) avec les données expérimentales. Souvent le compatible est pensé comme devant aboutir à un ajustement des paramètres aux données de l'expérience. C'est une vue fausse (en général). L'exemple typique est celui de la prévision économique, un modèle bien ajusté n'est pas en général un modèle donnant une bonne prévision : compatible avec les données doit s'entendre ici par donnant une bonne prévision. Le paragraphe sur la rareté, un peu plus loin, expliquera aussi cette idée de compatibilité en liaison avec la notion de test.

Cette compatibilité peut être prise en un sens plus large.

Ayant des données magmatiques ou astructurées, la statistique à l'aide d'instruments mathématiques peut chercher à les organiser en modèles un peu structu-

rés : *le paramètre est alors une structure des données,* et la statistique vise à valider certaines structurations (voir la remarque sur l'analyse des données). Il y a là un problème épistémologique intéressant. La démarche part de données apparemment inorganisées, mais le simple choix de l'instrument d'observation mathématique, les organise et donc données + statisticien implique modèle (que l'on veuille ou pas). La validation des modèles conduit donc soit à la validation d'une hypothèse préexistante chez le praticien au travail statistique, soit la présentation au praticien d'une nouvelle hypothèse de travail, les deux choses étant bien sûr compatibles.

STATISTIQUE ET DÉCISION

Dans une collection de modèles mathématiques dépendant d'un paramètre Θ et désignés par la valeur de ce paramètre, on peut chercher à valider un seul modèle, qui correspondra à la « vraie » valeur θ_o, inconnue avant d'avoir fait le travail statistique, ici appelé une estimation.

Si l'on valide θ_1 et que la vraie valeur inconnue est θ_o on peut décider de perdre w $(\theta_o, (0_0)$, w petit si θ_o est voisin en un certain sens de $\hat{\theta}$, grand sinon (w $(\theta_o, \hat{\Theta}) = 0$). $\hat{\Theta}$ étant aléatoire car dépendant des données (de l'observation), il en est de même de w $(\theta_o, \hat{\theta})$ et on prend l'habitude de classer les procédures d'estimation par la fonction de risque, qui à chaque θ_o associe la perte moyenne (pour la probabilité correspondant à la valeur θ_o du paramètre).

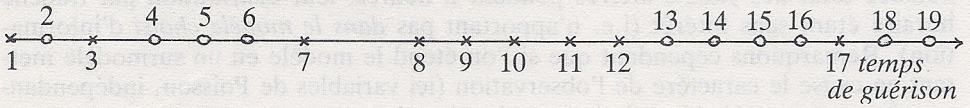
Les notions de pertes, de risques, font de la statistique un cas particulier de la théorie des jeux, le statisticien jouant contre la nature, en essayant de découvrir aux moindres frais (moindre risque) la vraie valeur de θ. Et tout l'arsenal de la théorie des jeux peut être déroulé, les estimations (validations) qui minimisent le risque maximum jouant un rôle appréciable.

Ce point de vue conduit à un vocabulaire décisionnel, une estimation, validation d'un seul modèle, rejet des autres est une décision particulière qui vise à minimiser le risque. Comme en général l'ensemble des risques (qui sont des fonctions) n'est pas totalement ordonné, on se limite à des classes de décisions particulières. Sinon on reste face à de nombreuses décisions, incomparables, toutes admissibles. Voisin de la théorie des jeux, parent de la théorie générale de l'optimisation, ce point de vue est très respectable et très intéressant mathématiquement, il est un bon catalyseur de l'intuition. Son grave défaut est de mystifier un peu l'autre, l'expérimentateur, le médecin, le biologiste ou l'ingénieur en hydraulique. Le statisticien ne peut décider sur le problème réel à la place du praticien, il amène un élément de diagnostic, à partir d'une modélisation mathématique du phénomène, celle du biologiste étant « biologique », celle de l'ingénieur « physique ».

LES ÉVÉNEMENTS RARES

La compatibilité d'un modèle peut être vue de la façon suivante : un modèle ne peut pas être valide si dans ce modèle un événement rare (de très petite probabilité) est réalisé par l'observation. L'événement doit être choisi d'avance.

Par exemple, si l'on compare deux traitements médicaux sur des individus très semblables (toutes choses que l'on a hautement intérêt à préciser, mais passons ici), on ordonne par temps de guérison les individus, les croix désignant ceux traités d'une manière, les o ceux par l'autre



On a l'*impression* que le traitement X est meilleur que le traitement o, et donc qu'il faut rejeter *tout* modèle qui prétend à l'équivalence des traitements et a fortiori à ce que o soit meilleur. A cet effet, on choisit la procédure suivante, on calcule la somme $S = 1 + 3 + 7 + 8 + 9 + 10 + 11 + 12 + 17$ des rangs des X. On montre que tout modèle où les traitements sont équivalents, la probabilité que S soit inférieur à 80 est inférieure à 1/10 (et même moins). Donc ici S vaut 78, on a réalisé l'événement ($S < 80$), trop rare pour être considéré comme *effectivement réalisable.*

La statistique amène à une utilisation souple et assez dialectique de la notion de rareté. Qu'est-ce qu'un événement rare ? Pour le physicien contemporain c'est un événement non réalisable par l'expérience, dans certaines parties de la physique actuelle, il semble que le seuil se situe vers 10^{-20} (à un facteur multiplicatif (10^{3}, 10^{-3}) près !). En tout cas, on considère que les événements de probabilité de l'ordre de 10^{-30} sont non pas rares mais impossibles (personne ne s'assied sur une chaise à l'état gazeux, or la probabilité qu'il en soit ainsi est de cet ordre). A l'opposé, la rareté est décisive dans l'organisation de la vie socio-économique (voir les « dogmes » de l'économie politique, ne pas confondre rareté et pénurie). Mais la rareté organise aussi la vie écologique, certaines espèces très rares organisant complètement la vie d'espèces très nombreuses, et bien sûr la réalisation d'événements rares est décisive de l'évolution biologique.

La statistique est donc toujours une démarche conservative si on la prend en un sens étroit, ne validant pas l'événement rare, elle peut ne pas valider le nouveau modèle, la découverte qui l'intègre à la connaissance du phénomène. Mais c'est à l'expérimentateur de ne pas la prendre en ce sens étroit et décisionnel et de manier le concept de rareté d'un événement avec une quantification liée à sa pratique et au degré réel de quantification de son domaine. De plus, on doit toujours se référer si possible à plusieurs types distincts d'événements rares examinés sur des données, elles aussi distinctes, pour éviter,dans le domaine de la recherche, ces écueils. Il serait très souhaitable que l'on avance dans ce concept de rareté, et de l'importance de tous les domaines des événements de petite probabilité.

STATISTIQUE, INFORMATION, INCERTITUDES

Chaque observation contient de l'information. Cette remarque évidente conduit aux deux grands concepts mathématiques spécifiquement statistiques et fait de la théorie de l'information pour l'essentiel (c'est un avis très personnel) une partie de la statistique.

Le premier concept est quoi retenir d'une observation : c'est celui d'exhaustivité par exemple, si l'on veut estimer le nombre moyen de personnes arrivant au guichet par heure, si l'on observe les arrivées pendant n tranches d'une heure. On aimerait expliquer *clairement* pourquoi on ne peut retenir de l'observation que le nombre total des clients arrivés pendant n heures, leur distribution par tranche horaire étant sans intérêt (i.e. n'apportant pas *dans le modèle choisi* d'information). Remarquons cependant que si l'on étend le modèle en un surmodèle mettant en cause le caractère de l'observation (ici variables de Poisson, indépendantes, de moyenne θ), alors il n'en est plus de même.

Une fois cette *réduction* de l'observation faite, on peut mesurer la quantité d'information qu'elle contient. Il y a un seul type d'information, valablement utile et naturel, c'est l'information de Shannon-Kullback, concept spécifiquement statistique qui est en quelque sorte l'extension à un ensemble de modèles de la notion d'entropie attachée elle, à un seul modèle [1].

On a donc la situation suivante :

Discussion statisticien-praticien
sur les buts de l'expérience
↓
Construction d'un ensemble de modèles
↓
Choix de l'observation et calcul
de sa réduction et de l'information
de l'observation
↓
Réalisation de l'expérience par le praticien
↓
Validation d'un ou d'un ensemble de modèles
↓
Exposé des résultats aux praticiens
↓
Construction d'un nouveau modèle, etc.

Revenons à la réduction de l'observation. Elle se fait si possible sans perte d'information. L'information statistique est toujours calculée à partir du concept de vraisemblance : ayant une observation, ce concept est la traduction mathématique de l'idée suivante : pourquoi ne pas valider le modèle, s'il faut en valider un seul, qui rend cette observation la plus probable. Cette démarche par suite de difficultés mathématiques n'est pas universelle. Cependant, modifiée et vue en un sens élargi, elle n'est pas loin d'être mêlée à toute démarche statistique. Le concept de vraisemblance fonde les concepts de la théorie de l'information, c'est probablement un des concepts mathématiques le plus important dégagé au XX^{e} siècle. Il traduit un concept intuitif très utilisé par tout un chacun dans son comportement journalier, il débouche loin de la statistique classique dans la statistique algorithmique et automatisée (par exemple la reconnaissance d'images).

1. Cf. A ce propos les notes en fin d'article.

CONCLUSION GÉNÉRALE

D'autres sujets mériteraient d'être abordés brièvement ici, l'histoire des statistiques, leur place dans l'enseignement et la recherche, où fait-on de la statistique et pourquoi ?

Pour l'enseignement secondaire, on est loin de compte, la statistique devrait occuper une place importante dans les petites classes, elle est concrète : comparaison de notes de films, séries météorologiques, test de dégustation avec ou sans truquage type verres colorés, simulations avec des calculettes, etc. Cela donnerait un aperçu sur des mathématiques liées à des expériences.

La statistique a connu un essor très grand avec la banalisation de l'ordinateur. Elle est au centre de la réflexion économique de l'individu, elle est utile à toute recherche scientifique, elle est inséparable de l'automatique dans la conduite des processus de production hautement technique, la notion de rareté, celle de risque, sont au centre d'une discussion rationnelle sur le problème du nucléaire, etc. Son enseignement, sa vulgarisation sont donc des instruments de démocratisation, d'accès des gens à la compréhension de phénomènes simples qui engagent leur avenir et celui de leurs enfants. Comme nous le disions au début, c'est sans doute une des meilleures disciplines pour développer l'esprit critique et de contestation, pour réduire à sa juste place le concept de modèle. On voit que de quel côté que l'on tourne les yeux, cela a pas mal d'implications.

NOTES

— *Comment construire un modèle.*

Supposons étudier une série chronologique d'ordre économique X_t, $t = 1 \ldots T$. Comment la modéliser et pour quel but ? Si l'on veut faire de la prédiction, un bon modèle ne sera pas le même en général, que si l'on veut ajuster au mieux, par exemple au sens des moindres carrés.

Supposons se restreindre aux modèles linéaires les plus simples du type

$$X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p} + \varepsilon_t$$

où ε_t est un bruit blanc, suite de variables décorrélées (ou même indépendantes) de même loi, de variance ^{2}. Comment choisir p? p représente ici la taille de la mémoire du phénomène et ε_t l'aléas à l'instant t.

Si l'on met beaucoup de paramètres, on peut ajuster de très près le modèle, c'est-à-dire que $\frac{1}{T}\sum_t \hat{\varepsilon}_t^2 = \frac{1}{T}\sum_t (X_t - \sum_{i=1}^{p} \hat{a}_i X_{t-i})^2 = \hat{\sigma}^2$ est très petit, a_i étant des estimateurs convenables des a_i, par exemple ceux qui minimisent $\hat{\sigma}^2$. Mais ces estimateurs sont très instables, peu robustes [1] et la qualité de la prédiction mauvaise. Il est donc raisonnable

1. La qualité de la prédiction est mesurée en général par la valeur théorique de l'erreur de prédiction, ici σ^2. Comme σ^2 n'est pas connu, on le remplace par son estimation. Mais il faut que le modèle construit pour des prédictions futures soit robuste. Supposons que l'on mette autant de paramètres que d'observations (cas caricatural). Alors, les estimateurs $\hat{a}_i$ n'ont plus aucun sens, si l'on change une donnée, on risque de changer beaucoup les valeurs des estimateurs. C'est la statibilité du modèle vis-à-vis de petites modifications des données, ou de l'absence de certaines que l'on appelle robustesse. Allons plus loin. Supposons que l'on fasse une deuxième observation notée, Y_t, $t = 1 \ldots T$ indépendante de X_t, mais de même structure. On veut que la prédiction obtenue en prenant le modèle, estimée à partir de $X_1 \ldots X_T$ soit la meilleure possible, autrement dit on veut minimiser $\frac{1}{T}\sum_1^T |Y_t - \hat{a}_1 Y_{t-1} \ldots - \hat{a}_p Y_{t-p}|^p$ en fonction de p. On montre alors qu'il ne faut pas prendre p trop grand. Une discussion fine est trop complexe pour être abordée ici.

d'introduire dans la construction des modèles *un principe de parcimonie* qui pénalise un nombre trop grand de paramètres et qui *impose* des modèles relativement simples. Il existe des résultats théoriques et surtout diverses règles empiriques à ce propos — règle présidant à l'identification de modèles simples à l'intérieur d'une classe très large. Par exemple si $L(p, a_1, ap, X_1 \ldots X_T)$ est le logarithme de la vraisemblance on peut choisir p et $a_1 \ldots a_p$ en maximisant $L(p) - \emptyset\ (T, p)$, $\emptyset$ fonction donnée. Divers résultats asymptotiques aident au choix de $\emptyset$. Une solution populaire est $\emptyset\ (T, p) = 2p$.

Pour suivre la mode dans notre pays on ne peut pas terminer sans parler d'analyses de données. Si l'on traite des données sociologiques, par exemple loisirs et catégorie socioprofessionnelle, après sondage, on obtient des tableaux de fréquence de cases, type profession libérale - sports d'hiver, à analyser. Ces tableaux sont apparemment sans structure (bien que de toute manière, il y ait derrière de telles données nombre de structures (idéologiques ?).

L'analyse des données fabrique des instruments d'optique permettant de visualiser suffisamment un nuage de points de R^p, p très grand, pour trouver les anisotropies essentielles de ce nuage. La plus simple des anisotropies consiste à assimiler le nuage à un ellipsoïde (défini par la covariance empirique) et à chercher les axes de cet ellipsoïde. Il s'agit donc mathématiquement de géométrie euclidienne, élémentaire a priori, mais l'interprétation nécessite un savoir-faire qui ne peut s'acquérir que sur le tas. Le statisticien peut faire un diagnostic pour le sociologue (qui ne peut manier cet instrument avec le doigté nécessaire). L'anisotropie découverte donnera lieu à un modèle qui dépendra souvent du choix de l'instrument d'optique choisi.

— *Vraisemblance et information de Kullback.*

Considérons des probabilités sur les entiers N ; c'est-à-dire des nombres p(n), $o \leqslant p \leqslant (n)\ 1$, $\sum_n p(n) = 1$. Si l'on considère une famille $p(n, \theta)$ de probabilités, où θ est un paramètre, la fonction $\theta \rightarrow p(n,\theta)$ s'appelle vraisemblance en n. Si X est une variable aléatoire qui suit la loi $p(n,\theta)$, si l'on observe une valeur X(w) en réalisant X, $p(X(w),\theta)$ s'appelle la vraisemblance observée. S'il existe θ valeur de θ, tel que

$$p(X(w),\hat{\theta}) \geqslant p(X(w),\theta)$$

pour tout θ, $\hat{\theta}$ est dit estimateur du maximum de vraisemblance de θ. Lorsque l'on a deux probabilités, on appelle information de Kullback-Shannon de $p(.,\theta_1)$ par rapport à $p(.,\theta_o)$ la quantité

$$\sum_n \left(\log \frac{p(n,\theta_o)}{p(n,\theta_1)}\, p_n(\hat{\theta}_o)\right)$$

Elle est toujours positive, éventuellement infinie.

$\sum_n \log p(n,\theta_0)\, p_n(\theta_o)$ est elle, l'entropie de $p(.,\theta_o)$.

L'information de Kullback quoique essentiellement dissymétrique en θ_0 et θ_1 joue le rôle d'une distance.

Considérons un k-échantillon d'une loi $p(.,\theta)$, θ est inconnu, on cherche à le découvrir (k-échantillon signifie k répétitions indépendantes de la même expérience). Soit $x_1 \ldots x_k$ les résultats observés. La probabilité empirique $\hat{p}$ associée à l'expérience prend la valeur $\frac{s(n)}{k}$ si l'on observe s(n) fois les valeurs n. On a bien sûr $\sum_n s(n) = k$. Cherchons à choisir θ par le plus proche de θ au sens ou l'information de $p(.,\theta)$ par rapport à p, est minimum. Un calcul élémentaire montre que $\hat{\theta} = \hat{\theta}$. Maximiser la vraisemblance c'est minimiser la distance d'information entre l'observation $\hat{p}$ et la loi inconnue $p(.,\theta)$.

L'entropie de l'observation ne dépend pas de θ Il faut donc minimiser $\sum_n \log p(n,\theta)\, \frac{s(n)}{k}$ qui conduit aux équations $\sum \frac{p'(n,\theta)}{p(n,\theta)}\, \frac{s(n)}{k} = 0$ dont le lecteur se persuadera qu'elles coïncident avec celles donnant le maximum de la vraisemblance.