STATISTIQUE
Prise de vue

Le mot « statistique » désigne à la fois un ensemble de données d'observation et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation.

Au cours de l'histoire, la collecte d'observations et la méthodologie de leur emploi se sont développées de façons largement indépendantes. Aujourd'hui, le recueil de statistiques est une activité importante, indispensable à la gestion des sociétés modernes. Leur traitement bénéficie des moyens offerts par l'alliance des calculateurs électroniques et de l'algèbre linéaire ; et l'on voit apparaître, dans l'analyse des données, des méthodes extrêmement puissantes pour « faire parler les chiffres ». Les raisonnements par lesquels on peut tirer, à partir des observations, des conclusions concernant les lois de probabilité des phénomènes (inférence ou induction statistique) sont codifiés par la statistique mathématique.

Dans la diversité des significations attachées au mot, il convient de souligner au moins la distinction entre les deux définitions suivantes :

– statistique : activité qui consiste à réunir des données, concernant en particulier la connaissance de la situation des États ou des sociétés humaines (c'est le « budget des choses » de Napoléon) ;

– statistique : méthode de traitement et d'interprétation des observations, de passage de celles-ci aux lois des phénomènes et aux modèles théoriques susceptibles de les représenter (c'est l'« inférence statistique » des statisticiens classiques, qu'on a eu quelque raison d'assimiler à l'induction formalisée).

En réalité, ces deux sens du mot statistique ont naturellement entre eux des liens étroits. Il serait vain de recueillir des données, si ce n'était pour les traiter et les interpréter en vue d'éclairer les actions humaines ou de faire progresser la connaissance des phénomènes. Inversement, la manière de recueillir des données peut et doit être influencée d'abord par les méthodes de traitement ultérieures et ensuite par l'utilisation pratique de ces données ou des produits qui en sont dérivés.

De plus, la statistique, en tant qu'activité consistant à recueillir des données, ne concerne pas seulement la connaissance des États et des sociétés, mais aussi tous les phénomènes particuliers pouvant faire l'objet d'études en laboratoire ou dans la nature. La méthodologie statistique, quant à elle, s'applique indifféremment dans ces diverses circonstances.

Dans le présent article, les deux sens du mot statistique – recueil de données ou méthodologie de traitement et d'induction – sont présents tour à tour ou simultanément.

I - Historique

On attribue souvent la création du terme « statistique » à un professeur de Göttingen, G. Achenwall, qui aurait en 1746 créé le mot *Statistik*, dérivé de la notion *Staatskunde*. En fait, l'activité correspondante de recueil de données permettant de connaître la situation des États remonte à une fort lointaine antiquité. On cite, d'une part, l'empereur chinois Yao, organisant le recensement des productions agricoles en 2238 avant J.-C., et, d'autre part, l'institution du cadastre et du cens chez les Égyptiens, en 1700 avant J.-C. L'importance sociale de la statistique était reconnue, puisqu'il advint que le pharaon Amasis édicta la peine de mort contre ceux qui refusaient de déclarer leurs nom, profession et moyens de subsistance. Un tableau d'ensemble de l'activité statistique dans ce sens particulier, au cours des différentes périodes de l'histoire, mériterait d'être dressé ; mais on ne pourra donner ici que quelques points de repère. On notera que le recueil de statistiques n'implique pas la connaissance de l'écriture : des planchettes à encoches (encore en usage dans certaines professions) ou des cordes à nœuds peuvent servir de supports à l'information statistique ; on assure que les *quipos* des Incas constituaient des systèmes particulièrement sophistiqués, permettant de recueillir et de tenir à jour leurs statistiques de récoltes sur des cordes de couleur.

R. Horwath (*Revue de l'Institut international de statistique*, 1972) signale le rôle de précurseurs des commerçants de la république de Venise, rassemblant, aux XIIIe et XIVe siècles, dans leurs *Relazioni*, de nombreuses données sur le commerce extérieur, qui furent utilisées pour la politique commerciale des régents. Au début du XVIIe siècle, ce sont les frères Elzévir, aux Pays-Bas, qui publient, sous le titre *Respublica Elzeviriana*, une sorte d'encyclopédie en soixante volumes contenant des informations sur l'économie et le commerce des États. Les travaux de W. Petty en Angleterre (cadastre, statistiques commerciales) sont bien connus, ainsi qu'en France les enquêtes ordonnées par Colbert et par Vauban (mémoires des intendants).

Au cours de cette période, les recensements de population et de ressources sont restés à un niveau purement descriptif, et c'est seulement au XVIIIe siècle que s'est répandue l'idée (introduite au siècle précédent par John Graunt, en Angleterre) que les statistiques recueillies en matière démographique pouvaient servir de base à des prévisions : tables de mortalité de P. G. Wargentin en Suède, de A. Deparcieux en France). Pour ce faire, les fréquences observées sur des populations assez nombreuses étaient purement et simplement assimilées à des probabilités, approximation légitime lorsqu'on dispose d'un grand nombre d'observations.

Parallèlement, le calcul des probabilités avait été développé par des mathématiciens, de Pascal et Fermat au XVIIe siècle jusqu'à Laplace au XIXe siècle, à peu près sans rapports réels avec l'activité statistique. Pourtant, Thomas Bayes, dans les *Philosophical Transactions*, avait donné, dans deux mémoires publiés après sa mort par les soins de son ami R. Price en 1764 et en 1765, le théorème et la formule qui portent son nom, en quoi il est permis de voir la naissance de l'induction formalisée. Mais la plupart des applications qui en furent faites d'abord, et pendant près de cent cinquante ans, méritent d'être regardées plutôt comme des applications de la théorie des probabilités que comme des exemples d'inférence statistique. Il s'agit notamment des travaux de Condorcet et de Laplace, qui sont parmi les plus connus, où l'on trouve des exemples assez variés de calculs de « probabilité des causes » : jugements des tribunaux, résultats obtenus dans des jeux de société, probabilité des témoignages, etc.

Adolphe Quételet fut certainement le premier à concevoir que la statistique pouvait être fondée sur le calcul des probabilités, et son œuvre extrêmement variée a donné à cette discipline une impulsion considérable. Elle concerne aussi bien l'anthropométrie que l'économie et les sciences sociales. Outre son rôle dans les services officiels de statistique de son pays, la Belgique, Quételet fut l'initiateur d'une coopération internationale, en créant des Congrès internationaux de statistique, de 1853 à 1876.

Il faut attendre les premiers statisticiens anglais, autour de 1900, pour voir apparaître réellement une méthodologie statistique, c'est-à-dire une théorie bien formalisée de l'inférence, du raisonnement qui permet, à partir des données observées, de tirer des conclusions sur les lois de probabilité des phénomènes. C'est la statistique mathématique, qui s'est développée entre 1900 et 1950 et dont les succès ont semblé imposer, au cours de cette période, une interprétation particulière du concept de probabilité : l'interprétation objectiviste ou fréquentiste (cf. *infra*, chap. 5).

À partir des années cinquante, ce point de vue a été mis en doute par les statisticiens néo-bayésiens, qui ont mis l'accent sur le fait que l'inférence statistique ne saurait s'appuyer seulement sur l'information contenue dans les données d'observation, mais doit aussi nécessairement prendre en compte la connaissance a priori des modèles probabilistes. Vers la même époque, l'apparition de calculateurs puissants a donné naissance aux méthodes d'analyse des données multidimensionnelles, qui ont connu une grande vogue, parfaitement justifiée par leur efficacité. Ces méthodes concernent plus la description que l'induction ; elles font peu de place aux hypothèses a priori et permettent de décrire, de classer et de simplifier des données ; enfin, les résultats auxquels elles conduisent peuvent suggérer des lois, des modèles ou des explications des phénomènes, mais ils ne permettent pas de porter un jugement, d'apprécier d'une manière formalisée la confiance que doivent inspirer ces lois ou ces modèles, comme c'était l'ambition de la statistique mathématique classique.

II - Production de statistiques

Statistiques générales et statistiques particulières

Par une simplification pas trop abusive, il est parfois commode de distinguer deux catégories de situations, dans lesquelles on est amené à recueillir des observations et à leur faire subir des traitements statistiques.

La première catégorie concerne des problèmes particuliers qui conduisent à mener des expériences, des observations ou des enquêtes en vue de répondre à une question scientifique précise ou d'éclairer une décision bien définie. On peut alors se faire d'avance une idée, au moins approchée, de tout le processus qui va de la conception de l'expérience ou de l'enquête jusqu'à la décision qui sera prise ou la conclusion qui sera tirée après traitement des données. Dans cette catégorie entrent des situations aussi diverses que les expériences scientifiques, les enquêtes de marché, le contrôle des fabrications industrielles et certains sondages d'opinion. En général, les données recueillies dans ces circonstances sont utilisées uniquement pour répondre au besoin explicitement formulé au début de l'expérience ou de l'enquête. Nous dirons qu'il s'agit là de « statistiques particulières ».

Dans la seconde catégorie, celle des « statistiques générales », on recueille des données qui seront utilisées par un grand nombre de personnes en vue de répondre à des objectifs parfois très divers, ce qui entraîne plusieurs conséquences importantes : il n'est pas concevable de formaliser globalement tout le processus statistique, du recueil à l'utilisation, puisque celle-ci est multiforme et mal connue a priori ; d'autre part, le recueil, le traitement préliminaire et la présentation des données font alors l'objet d'une activité quasi autonome, celle des services statistiques des organismes internationaux, des États ou des entreprises.

Dans la pratique, la distinction entre statistiques particulières et statistiques générales n'est pas tranchée de manière parfaite ni définitive. Jusqu'en 1970 environ, les statistiques de pollution entraient dans le domaine des statistiques particulières, car elles n'intéressaient qu'un nombre extrêmement limité de gens qui pouvaient être amenés à recueillir des observations pour leurs besoins propres. La situation a rapidement changé, et on va voir se développer les statistiques de pollution et d'environnement comme activité autonome, fournissant la matière première nécessaire à un grand nombre d'études, ainsi qu'à l'orientation des décisions des pouvoirs publics. Il faut souligner encore que l'apparition de calculateurs de plus en plus puissants, et surtout dotés de mémoires de grande capacité, accélère la transformation de certaines statistiques particulières en statistiques générales ; c'est le phénomène des banques de données.

Quant à la production des statistiques particulières, la seule chose sur laquelle il convient d'insister est la nécessité de tenir compte, au moment du recueil des observations, de l'usage qui en sera fait ; à cela répond la théorie des plans d'expérience ou des plans d'enquête. Mais on ne saurait songer à décrire la production de ces données qui concernent à peu près toutes les activités humaines. On esquissera seulement la description du processus de production des statistiques générales, au sens large indiqué plus haut, sans prétendre à l'exhaustivité.

Statistiques géophysiques

On signalera d'abord les statistiques concernant le milieu physique dans lequel nous vivons : statistiques stellaires, recueillies par les observatoires, ou données géophysiques, climatologiques, hydrologiques.

En France, par exemple, fonctionne un réseau d'observations météorologiques qui compte plusieurs centaines de postes où sont mesurées chaque jour diverses grandeurs telles que la pression atmosphérique, la température, la direction et la vitesse du vent, les précipitations et parfois la durée d'ensoleillement ou la nébulosité. Ces observations sont regroupées par l'établissement central de la Météorologie nationale et font l'objet de publications régulières, mensuelles, trimestrielles et annuelles. De même, les débits des cours d'eau français sont mesurés chaque jour en plus d'un millier de points et sont réunis dans divers annuaires hydrométriques.

Il existe une grande variété de statistiques géophysiques : eaux souterraines, phénomènes volcaniques, mesures océanographiques, magnétisme terrestre, constante *g* de la pesanteur font l'objet de mesures de plus en plus nombreuses, régulières et centralisées.

Statistiques démographiques

Pour l'établissement des statistiques concernant les populations humaines, l'observation n'est pas chose aisée, et tous les procédés dont on dispose ne fournissent, pour les populations importantes des États modernes, que des résultats approximatifs. L'instrument privilégié est le recensement, qui permet en principe de connaître l'état de la population d'un pays à un instant donné. Mais il est pratiquement impossible de « photographier » une population de façon instantanée, même en mobilisant des milliers d'agents recenseurs et en consentant aux coûts considérables de ce genre d'opération.

Si l'on veut se référer à un jour précis, on rencontre des causes d'erreurs telles qu'on a trouvé meilleur, lors de recensements récents, d'y renoncer et d'étaler les opérations sur une durée de l'ordre du mois. Il est vrai que la connaissance de l'état d'une population à un instant donné ne présente en fait guère d'intérêt. Dans les États développés, on a constaté que les recensements les mieux faits conduisent à des erreurs de l'ordre de un ou deux pour cent sur la population globale (et à des erreurs plus importantes ou plus faibles, selon les cas, pour des catégories particulières de population). Dans des États moins développés, et surtout moins urbanisés, il arrive qu'un recensement soit moins précis qu'une enquête par sondage.

Les questions posées lors d'un recensement concernent, d'une part, des caractéristiques des individus : âge, sexe, niveau d'instruction, profession, d'autre part, des informations concernant les ménages, en particulier le logement, et enfin des questions touchant à des problèmes sociaux d'actualité tels que les déplacements entre le domicile et le travail, comme ce fut le cas des recensements les plus récents en France.

Une tendance à l'inflation des questionnaires se manifeste, parce qu'il existe une grande quantité de phénomènes sur lesquels le recensement pourrait apporter des informations précieuses. Ainsi, dans un pays comme la France, des données biométriques aussi élémentaires que la taille et le poids des individus sont à peu près complètement inconnues, au sens de la connaissance statistique, qui impliquerait des données valables en fonction des caractères démographiques et sociologiques fondamentaux : âge, sexe, catégorie sociale, zones géographiques. Peut-être connaît-on mieux de telles caractéristiques pour certains animaux domestiques, ovins ou bovins, car cette connaissance présente un intérêt économique. Pour les êtres humains, l'intérêt en est sans doute moins évident : il s'agit par exemple de l'interprétation des disparités constatées, selon les régions ou les catégories sociales, en matière de taux de mortalité ou de morbidité ; des données biométriques ont aussi de la valeur pour l'industrie de la confection vestimentaire.

Si la périodicité des recensements varie, entre quatre ans et une dizaine d'années, selon les époques, sans que les décisions prises plus ou moins solennellement au niveau international puissent être longtemps observées, cela est dû en partie au coût énorme de ces opérations, et en partie au fait que, malgré la puissance des calculatrices utilisées, les services spécialisés parviennent parfois à grand-peine à exploiter les données d'un recensement avant d'entreprendre le suivant. En effet, si la quantité d'informations demandées dans un questionnaire s'accroît modérément avec le temps, la variété des exploitations et des diverses présentations de résultats qui sont nécessaires à la recherche scientifique ou pour les décisions des autorités publiques s'accroît de façon infiniment plus rapide. Cependant, les recensements constituent la source privilégiée de connaissances sur les populations humaines, et leurs résultats ont une valeur inappréciable.

Des informations complémentaires sur les populations sont fournies par l'état civil, les fichiers électoraux et diverses autres sources administratives. Les statistiques concernant les naissances sont sans doute celles qui fournissent les meilleurs résultats, puisque, dans les pays développés du moins, il est extrêmement rare qu'un nouveau-né ne soit pas enregistré à l'état civil, encore que des phénomènes de mortalité périnatale rendent moins rigoureux qu'on ne pourrait le penser les chiffres disponibles dans ce domaine : problème des faux mort-nés, par exemple ; des difficultés notables surgissent dans l'interprétation dès qu'on veut dépasser le stade du nombre global des naissances (problème des lieux de naissance). Les statistiques de décès sont encore assez bonnes quand on s'en tient aux données globales. Mais les observations portant sur les causes de décès, si importantes pour l'orientation de la politique sanitaire, se heurtent à d'énormes difficultés, malgré les efforts consentis pour informer les médecins, responsables des déclarations à ce sujet : intervention fréquente de causes multiples, divergences de diagnostic, dénominations différentes selon les catégories de personnes décédées (on meurt plus souvent de cirrhose du foie dans les classes aisées et d'alcoolisme dans les classes laborieuses),

fréquence énorme des décès de cause indéterminée ou par « sénescence ». Est-ce à dire qu'il s'agit là de statistiques inutilisables ? Bien loin de là, tous les spécialistes de l'épidémiologie ou de l'économie médicale le savent, l'interprétation de telles données est délicate mais fructueuse. Voici un autre exemple des incertitudes des statistiques de causes de décès : elles font apparaître chaque année en France onze mille quatre cents suicides, dont quelque deux tiers concernent des individus de sexe masculin ; on a émis l'hypothèse que ce chiffre serait sous-estimé de moitié. Il est actuellement tout à fait impossible, sur le plan scientifique, de dire où se situe la vérité. Mais il est aussi très instructif de méditer sur ce que signifie le mot de « vérité » dans le cas d'espèce.

Statistiques économiques et sociales

Les statistiques concernant les activités économiques et la consommation, et tous les phénomènes qui y sont liés directement, sont peut-être les mieux connues du public ; c'est plus particulièrement celles-là qu'évoque l'expression de statistiques générales.

Elles sont établies par centralisation et élaboration progressive des observations faites au niveau des ateliers, des usines, des entreprises, des syndicats professionnels et, le cas échéant, au niveau de divers échelons administratifs pour les statistiques de production. La tâche est relativement aisée pour les activités très concentrées, voire de quasi-monopole ou de monopole (mines de charbon, électricité, chemins de fer, production automobile, industrie pétrolière par exemple). Les productions sont plus faciles à observer et les statistiques sont d'autant plus sûres quand il s'agit d'éléments ayant une valeur importante : on connaît sans doute exactement la production d'avions ou de calculettes (encore que des difficultés de définition puissent se présenter) et avec une excellente précision la production d'automobiles. En revanche, il est extrêmement difficile de compter les voyageurs dans un réseau de transports urbains ou les objets pris en charge par la poste (plus de dix milliards d'objets par an en France), parce que l'observation attentive d'un tel objet a un coût prohibitif en regard de la valeur de cet objet. La situation privilégiée de l'industrie électrique constitue un cas exceptionnel, car les kilowatts-heures peuvent être aisément comptabilisés en bloc, en des points particuliers du réseau allant de la production à la consommation ; cette situation se retrouve, quoique de façon moins nette, dans l'industrie pétrolière. Pour les productions très dispersées, ainsi que pour la plupart des consommations finales, on ne peut procéder que par enquêtes. D'autres aspects de l'activité économique entrent aussi dans la production des statistiques générales : c'est le cas des phénomènes financiers et de l'emploi. Ce dernier point pose des problèmes de définition et même de conceptualisation (chômage partiel, emplois demandés). Il est à noter enfin que le secteur de la production agricole française a longtemps offert au statisticien des difficultés liées au statut social particulier des producteurs.

Chaque État moderne possède des organismes chargés de mesurer l'activité économique sous ses différents aspects. Le premier bureau de statistique a été créé en France, en 1800, par Napoléon, et des organismes similaires ont vu le jour vingt ou trente ans plus tard aux Pays-Bas et en Angleterre. Cela s'appela plus tard la Statistique générale de la France, devenue en 1941 le Service national des statistiques puis, en 1946, l'Institut national de la statistique et des études économiques (I.N.S.E.E.). Cet institut centralise maintenant l'ensemble des statistiques générales intéressant les autorités publiques, coordonne l'activité statistique spécialisée des différentes administrations, élabore des indices (indices de production, de prix, d'emploi, etc.) et publie les données essentielles dans le *Bulletin mensuel de statistiques* et dans l'*Annuaire statistique*, les plus pertinentes étant présentées à l'intention d'un public plus large, depuis 1960 environ, dans les excellents *Tableaux économiques de la France* (ouvrage familièrement connu sous le nom d'*Annuaire statistique de poche*).

Pour des raisons pratiques, l'I.N.S.E.E. ne peut centraliser et publier que des statistiques d'intérêt assez général, et les divers ministères et services publics recueillent et publient de leur côté les données relatives à leurs domaines d'activité. Ces publications sont extrêmement nombreuses, et la liste ci-après est loin d'être complète ; elle n'est destinée qu'à en montrer la grande variété : *Statistique agricole*, *Recensement de la production forestière française*, *Statistique des pêches maritimes*, *Recensement général du vignoble*, *Annuaire de statistique industrielle*, *Les Établissements industriels et commerciaux en France*, *Statistique annuelle des Charbonnages de France*, *Activité de l'industrie pétrolière*, *Statistiques de la production et de la consommation d'électricité*, *Statistiques officielles de*

l'industrie gazière, *Statistique de l'industrie minérale*, *Statistiques de la construction*, *Statistiques du commerce extérieur de la France*, *Statistique générale des gens de mer*, *Les Accidents corporels de la circulation routière*, *Statistiques postales*, *Statistiques du cinéma français*, *Les Élections législatives*, *Tableaux de l'Éducation nationale*, etc.

On pourrait citer des dizaines d'autres publications analogues ; un ample inventaire est donné dans le *Répertoire des sources statistiques françaises*, publié par l'I.N.S.E.E. L'ensemble des données susceptibles de décrire le fonctionnement économique et social d'une nation constitue une masse énorme, si l'on note que le répertoire que nous venons de citer est un volume de quatre cents pages, qui recense plusieurs milliers de publications. De longue date, on a cherché à condenser les principales informations économiques, qui sont les plus nombreuses. Deux voies sont offertes : celle des indices et celle de la comptabilité nationale.

Un indice a pour objet de résumer, d'agréger un ensemble d'observations plus ou moins homogènes, de façon à mesurer un concept global qui n'est pas directement accessible. La définition précise et le mode de calcul des indices soulèvent des problèmes théoriques délicats, souvent liés à leur caractère de statistiques générales : on a souligné avec raison que la structure d'un indice devrait être déterminée par la fonction qu'on assigne à cet instrument, mais la plupart des indices sont utilisés à des fins très diverses par les économistes aussi bien que par l'administration, sans parler des journalistes. La structure des indices a reposé longtemps sur des bases largement empiriques, et l'emploi des méthodes d'analyse des données multidimensionnelles paraît ouvrir la voie à des améliorations méthodologiques importantes. Les indices les plus largement utilisés sont destinés à mesurer la croissance économique – indices de production industrielle ou de production nationale brute (P.N.B.) – ou l'évolution du niveau des prix ou de l'emploi.

Les indices permettent de résumer un aspect homogène de l'activité économique ou sociale d'une collectivité. Il est difficile de relier entre eux les indices de production, de consommation des ménages, d'emploi, de prix, etc. pour obtenir une description plus synthétique . Celle-ci est l'objet de la comptabilité nationale, qui couronne en quelque façon l'ambition du statisticien. Les concepts primitifs se trouvent déjà dans les tableaux économiques de Quesnay, mais c'est seulement la familiarisation des économistes avec l'algèbre linéaire (matrices *input output* de Leontief) et surtout le fait de disposer d'un appareil statistique perfectionné qui ont permis à la comptabilité nationale de devenir un instrument efficace, à la fois pour une meilleure connaissance des mécanismes économiques et pour une aide à la décision des pouvoirs publics.

Statistiques internationales

Sur le plan international, divers organismes se préoccupent de centraliser et d'harmoniser les statistiques nationales. Le premier qu'il convient de citer est le Bureau de statistique des Nations unies qui produit un certain nombre de publications ; son *Annuaire statistique* est le document de base pour les statistiques mondiales. Il présente des séries aussi comparables que possible concernant la population (effectifs, taux de natalité, mortalité, etc.), la main-d'œuvre (emploi, chômage), l'agriculture, les forêts, les pêches, la production industrielle (indices), les industries extractives (houille, lignite, gaz, pétrole, fer, étain), les industries manufacturières (aliments, textiles, papier, caoutchouc, produits chimiques, matériaux de construction, métaux, matériel de transport), la construction, l'énergie, la consommation, le commerce intérieur et extérieur, la balance des paiements, les transports (trafic voyageurs et marchandises par fer, air, mer), les salaires et les prix, le revenu national, les finances, les statistiques d'instruction. Sur le plan régional, l'O.C.D.E. (Organisation de coopération et de développement économiques) et l'Office statistique des Communautés européennes élaborent et publient les statistiques correspondant à leur domaine d'action.

Pour rendre comparables les statistiques des différents pays, les difficultés sont grandes et beaucoup d'efforts seront nécessaires pour rendre possible l'établissement des comptes de la planète, qu'il faudra bien aborder un jour.

III - Analyse des données

Mis en présence de données, le statisticien peut se voir assigner une assez grande variété d'objectifs ; la statistique lui offre les méthodes adaptées à ces objectifs. Ceux-ci peuvent être, par exemple, de présenter les observations, ou de les traiter de telle façon que soit suggérée une explication des phénomènes, ou encore de répondre à une question précise concernant une théorie scientifique (

les observations sont-elles en accord avec la théorie ?) ou des aspects non directement observables de la réalité, ou enfin d'éclairer une décision particulière. Certains de ces objectifs peuvent être atteints par des méthodes de traitement des données n'impliquant apparemment aucune hypothèse sur les phénomènes étudiés : ces méthodes sont groupées sous l'expression d'analyse des données. D'autres objectifs impliquent que le statisticien introduise, sous forme d'un modèle probabiliste, les théories qu'il veut vérifier ou préciser (inférence statistique) ou les conséquences possibles du choix d'une action (décision statistique).

En analyse des données, il s'agit donc de « décrire » un ensemble de données. Dans la période classique de la statistique mathématique (première moitié du XXe siècle), on ne disposait de méthodes efficaces que pour des statistiques unidimensionnelles ou bidimensionnelles ; c'est seulement grâce aux ordinateurs que l'on peut traiter, sans trop les appauvrir, des tableaux d'observations de dimensions quelconques.

Statistique descriptive

Pour un phénomène unidimensionnel, une statistique est un ensemble de n mesures $\{x_1, x_2, ..., x_n\}$; on dit que c'est un n-échantillon. On peut le représenter aussi bien comme un n-uple de points de R que comme un point de R^n. Les méthodes statistiques élémentaires (statistique descriptive) s'attachent à décrire de tels objets. On définit ainsi des caractéristiques de valeur centrale : d'une part, la moyenne arithmétique :

$$\bar{x} = \frac{1}{n} \sum x_i,$$

la médiane (c'est-à-dire la valeur telle que la moitié des valeurs de l'échantillon lui soit inférieure ou égale), la moyenne des valeurs extrêmes, etc., et d'autre part des caractéristiques de dispersion : ainsi la variance, ou carré de l'écart type s, est la moyenne des carrés des écarts entre les valeurs de l'échantillon et la moyenne arithmétique :

$$s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

Les choix d'une caractéristique de valeur centrale et d'une caractéristique de dispersion doivent être cohérents : ils sont liés au choix d'une distance dans R^n permettant de comparer deux échantillons. Ainsi, avec la distance euclidienne classique, si on cherche le point de coordonnées toutes égales $\underline{t} = (t, t, ..., t)$, le plus proche de l'échantillon $\underline{x} = (x_1, x_2, ..., x_n)$, on trouve qu'il faut prendre $t = \bar{x}$, et que la distance entre $\underline{t}$ et $\underline{x}$ est alors $\sqrt{n}s$.

La représentation graphique d'un échantillon se fait à l'aide de l'histogramme, ou polygone des fréquences. On utilise aussi le diagramme des fréquences cumulées, ou fonction de répartition.

Pour le cas d'un échantillon d'un couple de deux variables, nous aurons par exemple :

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n).$$

Un tel échantillon peut être considéré comme un n-uple de points dans R^2 ou comme un couple de points dans R^n. Les caractéristiques usuelles « au second ordre » d'un tel échantillon sont les valeurs moyennes $\bar{x}$ et , les écarts types s_x et s_y, et le coefficient de corrélation :

$$r = \frac{\operatorname{cov}(x, y)}{s_x s_y},$$

quotient de la covariance :

$$\operatorname{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

par le produit $s_x s_y$ des écarts types.

Le coefficient de corrélation n'est pas affecté par une transformation affine sur l'une ou l'autre des variables x et y. On montre sans peine qu'il est compris entre − 1 et + 1. D'un point de vue géométrique, lorsqu'on représente l'échantillon par un couple de points de R^n, le coefficient de corrélation est le cosinus de l'angle formé par les vecteurs centrés :

$$r = \cos[(..., x_i - \bar{x}, ...), (..., y_i - \bar{y}, ...)].$$

Analyse en composantes principales

Passons maintenant au cas d'une statistique de dimension p (ou échantillon d'un p-uple de variables), n étant toujours l'effectif de l'échantillon. Les observations sont alors présentées sous la forme d'un tableau à p lignes et n colonnes (on dit qu'il y a n individus, sur chacun desquels ont été mesurés p caractères) :

$$X = \begin{pmatrix} x_1^1 \dots x_i^1 \dots x_n^1 \\ \dots\dots\dots \\ x_1^j \dots x_i^j \dots x_n^j \\ \dots\dots\dots \\ x_1^p \dots x_i^p \dots x_n^p \end{pmatrix}$$

$X_i \in \mathrm{R}^p$ est le vecteur des p caractères associés à l'individu i et $X^j \in \mathrm{R}^n$ est le vecteur des n observations du caractère j. Dans un langage géométrique, on représente le tableau X de deux manières : soit n points (individus) dans R^p (où chaque coordonnée est associée à un caractère), soit p points (caractères) dans R^n (où chaque coordonnée est associée à un individu). Dans R^p, on peut généraliser ce qui a été dit plus haut pour le cas de deux caractères en définissant le vecteur des caractères moyens :

$$G = (\bar{x}^1, \bar{x}^2, ..., \bar{x}^p)$$

et la matrice des covariances :

$$V = \| \operatorname{cov}(X^j, X^k) \|,$$

dont les termes diagonaux sont les variances des caractères. On définit de la même façon la matrice des coefficients de corrélation.

On peut donc caractériser « au second ordre » le tableau de données X, pour n individus et p caractères, par le couple (G, V) composé d'un vecteur et d'une matrice, ou, ce qui est parfois plus parlant, par le triplet (G, S, R) en notant S le vecteur des écarts types des caractères et R la matrice de corrélation. Mais cela ne permet pas, en général, de se faire une idée des ressemblances entre les individus, ou des parallélismes entre les caractères (sauf à les considérer seulement deux par deux). Cela tient à ce que nous ne pouvons donner un support concret aux espaces ayant plus de deux dimensions. Une technique fort répandue, et efficace, consiste à chercher dans R^p (espace des individus où nous avons « représenté » un nuage de n points) un sous-espace à deux dimensions, dans lequel les projections orthogonales (avec la métrique euclidienne usuelle par exemple ; mais d'autres variantes sont possibles) des « points individus » forment un nuage aussi voisin que possible du nuage initial. Ce résultat est obtenu par l'analyse en composantes principales, qui donne d'ailleurs aussi bien la suite des sous-espaces de dimension 1, 2, 3, ..., p − 1 ayant la propriété désirée (nuage projeté le plus proche du nuage initial). On montre qu'il s'agit d'une famille de sous-espaces engendrés par les vecteurs propres de la matrice V des covariances (ou de la matrice $V \circ M$ si l'on choisit une métrique M quelconque). De façon plus précise, le sous-espace de dimension k qui donne le nuage projection le plus proche du nuage initial est engendré par les vecteurs propres correspondant aux k plus grandes valeurs propres de la matrice V. Dans la pratique, on se borne souvent à prendre $k = 2$ ou $k = 3$.

La proximité entre le nuage initial et le nuage projeté est repérée par le quotient entre l'inertie (variance généralisée) de l'un et l'autre nuage. Si ce quotient est voisin de l'unité, les individus sont approximativement situés dans le sous-espace fourni par l'analyse, qui décrit alors à peu près exhaustivement le tableau initial des données. On appelle « facteurs principaux » les nouvelles variables (combinaisons linéaires des variables initiales) constituées par les vecteurs propres qui engendrent les espaces principaux. Les corrélations entre les variables originales et les facteurs principaux permettent parfois de donner à ces derniers une interprétation.

Exemple d'application

Pour illustrer la méthode de l'analyse en composantes principales, la figure montre les résultats obtenus dans une étude faite par l'Atelier parisien d'urbanisme et portant sur la composition socioprofessionnelle des actifs, dans vingt-quatre grandes villes françaises, lors des trois recensements de 1954, 1964 et 1968. Les catégories socio-professionnelles retenues sont réparties en sept postes et excluent les agriculteurs et salariés agricoles, car il s'agit ici d'étudier l'évolution des sociétés urbaines.

La structure en sept postes est la suivante :

– patrons de l'industrie et du commerce (patr.) où se regroupent les industriels, les artisans, les patrons pêcheurs, les gros et les petits commerçants ;

– professions libérales et cadres supérieurs (lib.) : professions libérales, professeurs, professions littéraires et scientifiques, ingénieurs et cadres administratifs supérieurs ;

– cadres moyens (cadr.) : instituteurs et professions intellectuelles diverses, services médicaux et sociaux, techniciens et cadres administratifs moyens ;

– employés (empl.), qui correspondent à la fois aux employés de bureau et aux employés de commerce ;

– ouvriers (ouvr.) : contremaîtres, ouvriers qualifiés et spécialisés, mineurs, marins, pêcheurs, apprentis ouvriers et manœuvres ;

– personnel de service (serv.) : gens de maison, femmes de ménage et autres ;

– divers (div.) : artistes, clergé, police, armée.

Le tableau des données initiales comporterait donc 7 lignes (caractères) et 72 colonnes (individus : 24 villes × 3 recensements). Les résultats sont représentés sur la figure, qui montre la position des villes et leur évolution, dans le premier plan principal. On obtient dans ce plan un peu plus des deux tiers de l'inertie totale du nuage : 68 p. 100. Le cercle des corrélations, à gauche du graphique, montre comment chaque catégorie socio-professionnelle est liée à chacun des deux premiers facteurs. Ce type d'analyse permet de décrire avec précision un phénomène qualitativement bien connu des sociologues, qui est la « tertiarisation » de la population des villes françaises. Un examen attentif du graphique montre ou suggère un nombre considérable d'observations, concernant les ressemblances entre les différentes villes ou le parallélisme de leur évolution.

Régression linéaire

La régression linéaire répond à un problème moins purement descriptif, plus structuré : on dispose d'un tableau de données comprenant $p + 1$ lignes et n colonnes, et il s'agit d'examiner comment on peut expliquer l'une des variables (correspondant par exemple à la première ligne) à l'aide d'une combinaison linéaire des p autres lignes. La représentation géométrique la plus commode consiste à se placer dans l'espace R^n des caractères. Les p caractères « explicatifs » engendrent un sous-espace vectoriel W, dans lequel il s'agit de trouver le point le plus proche possible du caractère « à expliquer ». Si l'on écrit le tableau de données sous la forme :

$$Z = \begin{pmatrix} y_1 \dots y_i \dots y_n \\ x_1^1 \dots x_i^1 \dots x_n^1 \\ \dots\dots\dots \\ x_1^p \dots x_i^p \dots x_n^p \end{pmatrix}$$

on peut dire qu'il s'agit de rechercher la projection ŷ de y dans W, espace engendré par les x_i dans R^n. Cela suppose qu'on a choisi une distance euclidienne, donc une norme dans R^n. On peut alors mesurer la qualité de la représentation de y par le quotient :

$$r = \frac{\|\hat{y}\|}{\|y\|},$$

qui est le coefficient de corrélation totale entre y et l'ensemble des x_j.

La plupart des nombreuses techniques présentées dans la littérature statistique sous le nom de régression (régression multiple, régression polynomiale, etc.) sont des cas particuliers de ce modèle.

Analyse canonique

L'analyse canonique généralise, d'une certaine façon, la régression linéaire, puisqu'elle vise à confronter deux groupes de variables, mesurées les unes et les autres pour *n* individus, et à trouver jusqu'à quel point il est possible de prévoir l'un des groupes à partir de l'autre. Le tableau des données se présente avec *p* + *q* lignes et *n* colonnes, sous la forme :

$$Z = \begin{pmatrix} x_1^1 \dots x_n^1 \\ \dots\dots \\ x_1^p \dots x_n^p \\ y_1^1 \dots y_n^1 \\ \dots\dots \\ y_1^q \dots y_n^q \end{pmatrix}$$

On se place dans l'espace R^n des caractères, et l'on étudie les propriétés du couple des sous-espaces engendrés par les *p* caractères *x* et par les *q* caractères *y*, respectivement. Résumons ce qui peut être dit sur les possibilités de prévision des *y* à partir des *x* (on pourra dire symétriquement des choses en tout point analogues en ce qui concerne les possibilités de prévision des *x* à partir des *y*). Soit W_1 le sous-espace vectoriel engendré dans R^n par les caractères *x*, et W_2 le sous-espace vectoriel engendré par les caractères *y*. Si ces sous-espaces sont confondus, il est clair qu'on peut « prévoir » exactement les *y*, connaissant les *x* : pour tout individu, si l'on connaît les *x*, on reconstitue les *y* par des combinaisons linéaires appropriées. Dans le cas contraire, on montre qu'on peut décomposer W_2, par exemple, en somme directe de trois sous-espaces vectoriels, à savoir :

$$\mathbf{W}_2 = (\mathbf{W}_1 \cap \mathbf{W}_2) \oplus \mathbf{W}_{22} \oplus \mathbf{W}_{222},$$

où W_{222} est un sous-espace vectoriel orthogonal à W_1 et W_{22}, le sous-espace supplémentaire, dans W_2, de la somme directe des sous-espaces W_2 et W_{22}. On dit alors que :

– des caractères (combinaisons linéaires des caractères *y*) représentés dans $W_1 \cap W_2$ sont exactement prévisibles à partir des caractères *x* ;

– des caractères représentés dans W_{222} sont tels que la connaissance des caractères *x* n'apporte sur eux aucune information ;

– des caractères représentés dans W_{22} sont « partiellement prévisibles » à partir des caractères *x*.

Mathématiquement, la décomposition dont on vient de parler s'effectue à l'aide de projecteurs de W_2 sur W_1 et sur le sous-espace orthogonal de W_1 dans R^n, par exemple. On étudie l'opérateur produit de ces deux projecteurs, et on trouve que les sous-espaces vectoriels engendrés par les vecteurs propres de cet opérateur correspondant aux valeurs propres nulles, comprises entre 0 et 1, égales à l'unité, coïncident en fait respectivement avec les trois sous-espaces vectoriels suivant lesquels nous avons décomposé W_2. On aboutit ainsi à un algorithme programmable ; dans les programmes usuels, des sorties graphiques permettent de décrire les phénomènes dans des plans de projection situés soit dans W_1, soit dans W_2, de façon un peu analogue aux sorties obtenues dans l'analyse en composantes principales (cercle de corrélation de la figure).

Analyse factorielle discriminante

Dans l'analyse en composantes principales, on avait en vue la description d'un tableau de données de dimensions (*p*, *n*) pour *p* caractères et *n* individus, les deux ensembles I des individus et J des caractères n'ayant aucune structure particulière. Dans les techniques examinées ensuite, l'ensemble J des caractères était l'objet d'une dichotomie : 1 caractère et *p* caractères pour le cas de la régression ; *p* caractères et *q* caractères dans le cas de l'analyse canonique.

Dans l'analyse factorielle discriminante, c'est sur l'ensemble des individus que l'on se donne une partition (on dit aussi, de façon équivalente, qu'on possède, outre les *p* caractères quantitatifs du tableau de données initial, un caractère qualitatif, avec un nombre fini de modalités). L'objet de l'analyse est alors de rechercher si ce caractère qualitatif supplémentaire possède une influence sur l'ensemble des *p* variables mesurées et de déterminer, le cas échéant, des caractères discriminants, c'est-à-dire des caractères induisant sur l'ensemble I des individus une partition aussi proche que possible de celle que définit la variable qualitative initiale. L'analyse factorielle discriminante se ramène à une analyse en composantes principales effectuée sur l'ensemble des centres de gravité des diverses classes d'individus correspondant aux modalités de la variable qualitative, l'espace vectoriel des individus étant muni de la métrique définie par la matrice inverse de la matrice des covariances.

Analyse factorielle des correspondances

Les tableaux de données qui ont été l'objet des techniques passées en revue jusqu'ici contenaient des caractères mesurés sur les individus. On est souvent en présence de tableaux un peu différents, dont le contenu est formé par les fréquences avec lesquelles sont observées les modalités de deux phénomènes, ces modalités étant représentées respectivement par les lignes et les colonnes du tableau. Il s'agit de « tableaux de contingence » du même type que le tableau représenté ci-dessus , mais qui peuvent être d'assez grandes dimensions.

yeux \ cheveux	blonds	châtains	noirs	roux	total
bleus	1 768	807	189	47	2 811
gris ou verts	946	1 387	746	53	3 132
marron	115	438	288	16	857
total	2 829	2 632	1 223	116	6 800

Tableau de correspondance
Tableau de correspondance entre la couleur des yeux et la couleur des cheveux pour un ensemble de 6 800 hommes du pays de Bade (d'après Ammon)
.

En présence de tels tableaux, la statistique classique nous donnait, par le test du χ^2 (« test du khi-deux »), le moyen de savoir s'il existe une liaison entre les phénomènes étudiés, mais ne permettait guère de décrire cette liaison, ce qui est précisément l'objet de l'analyse factorielle des correspondances.

Appelons de manière générale I et J les ensembles des modalités des deux phénomènes étudiés, et soit k_{ij} l'effectif des observations dans la case (i, j), soit k_i, les effectifs totaux par ligne, et k_j les effectifs totaux par colonne. S'il y a *p* colonnes, on peut représenter, dans l'espace R^p, le contenu du tableau par un nuage de *n* points, les coordonnées du *i*-ième point étant k_{ij}/k_i. Sur ce nuage , on fera une analyse en composantes principales, mais en prenant comme distance entre deux points *i* et *i'* la « distance du khi-deux » :

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{k_{.j}} \left[\frac{k_{ij}}{k_i} - \frac{k^{i'j}}{k_{i'}} \right]^2.$$

Le choix de cette distance est justifié par le fait qu'elle ne dépend que du profil des colonnes du tableau. L'analyse permet, dans le plan des deux premiers axes factoriels, une représentation simultanée, souvent fort suggestive :

– des ressemblances entre les colonnes du tableau,

– des ressemblances entre les lignes,

– de la proximité entre lignes et colonnes.

À titre d'exemple des résultats auxquels peut conduire l'analyse factorielle des correspondances, nous donnons dans la figure ci-contre le graphique obtenu dans l'étude du tableau rapprochant les causes de décès aux catégories socioprofessionnelles. Les auteurs (L. Lebart et J.-P. Fénelon) ont retenu huit causes de décès, pour les hommes de 46 à 54 ans, décédés entre 1956 et 1960, et treize catégories socioprofessionnelles. Les résultats de l'analyse sont assez éloquents pour se passer de commentaires.

Méthodes de classification

On rencontre très fréquemment des problèmes de classification : étant donné un ensemble d'objets possédant certains caractères, il s'agit de les grouper en classes d'objets voisins. En réalité, on peut être plus ou moins exigeant sur les critères de ressemblance entre objets situés dans la même classe, et on est conduit à rechercher souvent simultanément une famille de classifications. D'autre part, les caractères donnés sur les objets et les similarités entre les caractères peuvent se présenter de diverses façons : parfois on définit une distance entre les objets ou bien seulement un indice de similarité possédant des propriétés plus ou moins fortes, parfois on peut seulement dire qu'un couple est plus distant qu'un autre couple. C'est pourquoi il existe des méthodes de classification assez nombreuses, et il n'est pas aisé d'en donner une description brève et cohérente. Indiquons toutefois quelques notions importantes.

Une ordonnance est une relation d'ordre total sur les distances entre tous les couples de deux objets (cf. ensembles ORDONNÉS). R. N. Shepard a montré que, sous des conditions peu restrictives, si on connaît l'ordonnance d'un ensemble de points assez nombreux dans R^n, on peut reconstituer le nuage de ces points à une similitude près. Cette propriété montre qu'on peut rechercher alors une classification à partir des seules informations fournies par une ordonnance.

Une hiérarchie est une famille de partitions ordonnée complètement par la relation « plus fine que » (partitions emboîtées) et contenant les partitions triviales extrêmes. À une hiérarchie on associe un « arbre de classification » qui permet, avec quelques précautions techniques et des hypothèses supplémentaires (niveaux relatifs des nœuds non situés sur un même chemin), de trouver, par sectionnement, toutes les classifications cohérentes entre elles.

Les algorithmes, qui permettent de déterminer une classification à partir des éléments caractérisant un ensemble d'objets (distance, indices de similarité, éventuellement indices multidimensionnels), se classent en algorithmes descendants (on divise progressivement l'ensemble de tous les objets ; les méthodes de segmentation en sont un exemple) et en algorithmes montants (on part de la partition la plus fine et on agglomère progressivement les objets en groupes d'objets, puis les groupes entre eux ou avec les objets non encore agglomérés). La plupart des algorithmes utilisés en pratique sont établis sur des bases largement empiriques, les analyses complètes conduisant souvent à des calculs d'un coût prohibitif pour des ensembles d'effectif assez grand.

On peut aussi fonder une classification soit sur une analyse en composantes principales qui permet de représenter, de façon au moins approchée, les distances entre les objets et donc de déceler des groupements naturels, soit sur une analyse factorielle des correspondances, s'il s'agit de classification d'ensembles de modalités des variables qualitatives sur lesquelles opère une telle analyse.

IV - Inférence statistique classique

Les méthodes que nous avons décrites au chapitre précédent sous le titre d'analyse des données concernent la description des tableaux d'observations statistiques et ne font intervenir aucune hypothèse sur l'origine de ces données. La statistique classique était au contraire une statistique probabiliste, fondée sur la prise en considération des lois de probabilité auxquelles obéissent les phénomènes naturels, objets d'observation. On décrira maintenant les principes et les méthodes de la statistique classique.

Théorie de l'échantillonnage

Tout, dans la statistique classique, repose sur l'étude des distributions des échantillons. Pour la statistique classique, un échantillon est défini dans le cas le plus simple comme un ensemble d'épreuves indépendantes et de même loi. Il convient d'abord d'étudier la distribution de probabilité de tels échantillons, dont on donnera quelques exemples. Soit x une variable aléatoire suivant une loi de Laplace-Gauss de moyenne m, d'écart type σ, et soit :

$$(x_1, x_2, \ldots, x_n)$$

un n-échantillon, au sens qui vient d'être indiqué au chapitre 3, sous le titre *Statistique descriptive.* La moyenne arithmétique $\bar{x}$ de cet échantillon suit une loi de Laplace-Gauss, de moyenne m et d'écart type $\sigma/\sqrt{n}$. La variance, ou plus exactement la quantité $(ns^2)/\sigma^2$, qui

est proportionnelle à la variance de l'échantillon, suit une loi dite loi de χ^2 à $n - 1$ degrés de liberté, dont la densité s'écrit, en posant $(ns^2)/\sigma^2 = u$ et $n - 1 = \nu$,

$$f(u) = \frac{1}{\Gamma(\nu)} e^{-u} u^{\nu-1};$$

cette loi est dite aussi loi gamma. La quantité :

$$t = \frac{\bar{x} - m}{s} \sqrt{n-1}$$

suit une loi de probabilité qui ne dépend d'aucun des paramètres de la loi de Laplace-Gauss initiale ; c'est la loi de Student, de densité :

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\sqrt{\pi}\,\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}.$$

De même, étant donné un échantillon d'une loi de Laplace-Gauss à deux variables, soit :

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

on sait établir la loi de probabilité des grandeurs servant habituellement à caractériser les moyennes et dispersions, ainsi que la liaison entre x et y. On montre que si m_x, m_y, σ_x, σ_y et ρ sont les paramètres de la loi du couple (x, y), alors le couple $(\bar{x},)$ suit une loi de Laplace-Gauss de paramètres $(m_x, m_y, \sigma_x/\sqrt{n}, \sigma_y/\sqrt{n}, \rho)$ et se trouve indépendant du triplet (s_x, s_y, r) dont on sait par ailleurs calculer la loi de distribution exacte (c'est une expression analytique assez compliquée).

Une autre loi de distribution d'échantillonnage couramment utilisée dans la statistique classique est celle du quotient de deux variances indépendantes (ou de deux χ^2) ; elle est connue sous le nom de loi de Fisher-Snedecor. Un certain nombre d'autres lois, que suivent les grandeurs caractérisant des échantillons de variables gaussiennes, ont été étudiées et tabulées. Les premiers statisticiens, au début du siècle, utilisaient de tels résultats à l'aide d'un principe de raisonnement extrêmement simple : les grandeurs calculées à partir d'un échantillon étant choisies de façon quelque peu intuitive, on convenait que des valeurs de probabilité très petites ne devaient pas se produire, ou plus précisément qu'il était raisonnable de « faire comme si » des valeurs de probabilité très petites ne s'étaient pas produites. La statistique mathématique a donné peu à peu une forme cohérente à ce type de raisonnement. Nous donnerons comme exemples quelques-unes des théories développées à cet effet.

Théorie de l'estimation

Supposons qu'une grandeur observable x suive une loi de probabilité dépendant d'un paramètre θ et possédant une densité $f(x, \theta)$; on dispose d'un échantillon de valeurs données, soit $x_1, x_2, \ldots, x_n$, et on veut se faire une idée de la valeur du paramètre inconnu θ. On peut préciser cet énoncé de plusieurs façons. La plus simple conduit à poser le problème de l'estimation ponctuelle : on veut, au vu des observations, choisir une valeur qu'on attribuera à θ. Ce qu'il faut déterminer est donc une application de R^n dans R ; une telle application est appelée un estimateur. On a étudié systématiquement les propriétés des estimateurs, et on convient souvent qu'un bon estimateur doit être :

– *sans biais*, c'est-à-dire que son espérance mathématique doit être égale à la vraie valeur du paramètre θ ;

– *convergent*, c'est-à-dire qu'il doit converger en probabilité (ou en moyenne quadratique) vers θ lorsque n tend vers l'infini ;

– *de variance minimum* (on dit estimateur « efficace »).

On a montré que, avec des hypothèses assez générales, on obtient un estimateur convergent, asymptotiquement sans biais et efficace lorsqu'on utilise la méthode du maximum de vraisemblance : on prend pour valeur estimée $\hat{\theta}$ la valeur de θ qui rend maximum la densité de probabilité de l'échantillon (ou son logarithme) :

$$\max \sum_i \ln f(x_i, \theta);$$

on peut alors calculer la variance (au moins asymptotique) de l'estimateur ainsi obtenu.

Dans certains cas, la situation est encore meilleure, en ce sens que, pour *n* fini, on obtient un estimateur de variance minimum : il est impossible de faire mieux. Cela se produit en particulier si l'on a un estimateur exhaustif : la loi de distribution conditionnelle des observations, étant donné la valeur prise par l'estimateur, ne dépend pas du paramètre à estimer.

Par exemple, lorsqu'on doit estimer la moyenne d'une loi de Laplace-Gauss, dont l'écart type est connu, on montre que la moyenne arithmétique des observations constitue l'estimateur du maximum de vraisemblance ; c'est d'ailleurs un estimateur exhaustif.

Un autre point de vue, qui concerne le problème de l'estimation, conduit à rechercher pour un paramètre non pas une valeur estimée mais un ensemble de valeurs, un intervalle, dans lequel « il y ait de bonnes chances » que se trouve le paramètre inconnu. Si l'on veut, dans la perspective de la statistique classique, donner un sens précis à cet énoncé, il faut observer que le paramètre θ, pour être inconnu, n'est pas un élément aléatoire (on n'est pas disposé à lui donner une loi de probabilité). C'est l'intervalle, construit à partir des observations, qui est aléatoire ; c'est ce que Neyman a appelé intervalle de confiance. On sait en effet, dans un certain nombre de cas, construire de tels intervalles, qui ont une probabilité donnée, choisie d'avance, de recouvrir la vraie valeur, inconnue, du paramètre, que celui-ci reste fixe ou varie d'une façon quelconque au cours des expériences. On sait même construire des intervalles de confiance pour un paramètre lorsque la loi de probabilité des observations contient aussi d'autres paramètres auxquels on ne s'intéresse pas et qui sont dits paramètres nuisibles. Par exemple, si l'on s'intéresse à la moyenne d'une loi de Laplace-Gauss d'écart type inconnu et si l'on dispose de quatre observations, on peut affirmer que l'intervalle :

$$(\bar{x} - 3{,}2\,s\,;\ \bar{x} + 3{,}2\,s)$$

recouvre la valeur inconnue de la moyenne avec une probabilité de 95 p. 100, quelles que soient par ailleurs les valeurs de la moyenne et de l'écart type de la loi.

Test du χ^2

Un autre exemple, justement célèbre et d'usage courant, du mode d'inférence classique est le test du χ^2. C'est en fait, historiquement, la première pierre de la statistique, posée par Karl Pearson, en 1900, dans un article du *Philosophical Magazine* intitulé (et ce titre vaut d'être médité par les statisticiens) : « *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* » ; on voit avec quelle prudence et quelles précautions l'auteur introduit son célèbre critère.

Le test du χ^2 a été le plus souvent utilisé soit comme test d'ajustement d'une loi de probabilité à un échantillon d'observations supposées indépendantes et de même loi, soit comme test de liaison dans un tableau croisé de fréquences de deux phénomènes qualitatifs (ou variables ayant fait l'objet d'un regroupement en classes), soit parfois comme test d'homogénéité. En fait, c'est essentiellement un test d'un modèle multinomial. La possibilité de bâtir ce test résulte des considérations suivantes :

a) La loi de probabilité de la somme des carrés de v variables gaussiennes, réduites et indépendantes, est connue sous le nom de loi de χ^2 et se calcule aisément : c'est la loi gamma citée plus haut à propos de la distribution de la variance d'un échantillon d'une loi normale, où v est le nombre de degrés de liberté.

b) La loi de distribution de la forme quadratique intervenant en exposant dans la densité d'une loi de Laplace-Gauss à *n* variables est la même que la loi précédente en prenant pour v le rang de la forme quadratique.

c) La loi de distribution multinomiale à *k* catégories converge, quand *n* tend vers l'infini, vers une loi de Laplace-Gauss dégénérée de rang ($k - 1$). Rappelons que la loi multinomiale est la distribution, entre les catégories, de *n* objets, chaque objet ayant les probabilités $p_1, p_2, ..., p_k$ d'appartenir à chacune des catégories ; on a donc la distribution d'un vecteur $(x_1, x_2, ..., x_k)$ dont les composantes sont assujetties à la condition $\Sigma x_i = n$.

d) La forme quadratique qui apparaît dans l'expression de la loi de Laplace-Gauss, limite pour *n* grand de la loi multinomiale, a pour expression :

$$\sum \frac{(x_i - np_i)^2}{np_i}.$$

Dès lors, en rapprochant les quatre propriétés que l'on vient d'énoncer et en supposant que *n* est assez grand pour qu'on puisse utiliser la loi limite de l'expression ci-dessus, on peut dire qu'elle suit approximativement une loi de χ^2, avec $n - 1$ degrés de liberté. De façon empirique, on dit que, si une valeur, obtenue en fait à partir d'observations qui étaient supposées provenir d'une loi multinomiale, dépasse des valeurs données comme très improbables par la loi de χ^2, c'est l'indice que le modèle probabiliste multinomial doit être suspecté.

Le test de χ^2 a été beaucoup utilisé. Dans son emploi comme test d'ajustement d'une loi de probabilité continue, et particulièrement pour de petits échantillons, on a proposé de le remplacer par des tests fondés sur la distribution d'une distance entre les fonctions de répartition théorique et empirique : cette distance peut être définie comme une moyenne quadratique de l'écart entre les deux fonctions (test de Smirnov) ou comme le maximum de la valeur absolue de cet écart (test de Kolmogorov).

Ce dernier test est utilisable également pour juger si deux échantillons peuvent être considérés comme provenant de la même loi de probabilité (homogénéité de deux échantillons). Dans cet emploi, le test ne fait pas intervenir la loi de probabilité inconnue (il est fondé sur la distribution de l'écart maximum entre les deux fonctions de répartition). On dit que c'est un test « non paramétrique » (*distribution-free*).

Théorie générale des tests

Comme nous l'avons indiqué à propos de χ^2, les tests utilisés sur des bases empiriques s'appuient sur le principe de raisonnement suivant : si des observations avaient une probabilité très faible, cette probabilité étant calculée à l'aide d'une hypothèse (loi de probabilité) particulière, alors cette hypothèse est vraisemblablement fausse. Cette façon de raisonner a été très féconde, et, à une certaine époque, Émile Borel a proposé à peu près ceci : « Axiome unique pour les applications du calcul des probabilités : les événements de probabilité inférieure à un dix-millième ne se produisent pas. » Naturellement, il a fallu reconnaître par la suite qu'un tel axiome n'est pas admissible, bien que des propositions voisines aient été admises par certains scientifiques. D'une part, des miracles se produisent tous les jours. D'autre part, on ne peut pas faire œuvre scientifique en adoptant à chaque instant les hypothèses qui expliquent le mieux les données observées, et donc excluent les miracles. En fait, on a constaté, vers les années trente, qu'une théorie cohérente des tests contraignait à prendre en compte non seulement l'hypothèse testée, mais aussi celle qu'il faudrait bien mettre à la place si les observations conduisaient à rejeter la précédente. Partant de là, J. Neyman a eu le grand mérite d'avoir proposé, sous l'expression de théorie générale des tests, une construction cohérente dont nous esquisserons seulement les grandes lignes, au plan conceptuel, laissant de côté toutes les difficultés techniques, et elles sont fort nombreuses.

Une situation de test implique qu'on prenne en compte les observations (un point *x* d'un espace quelconque, souvent R^n dans les applications), l'hypothèse testée H (qui est une loi de probabilité particulière sur R^n), une hypothèse alternative K (qui est une autre loi de probabilité sur R^n). Le problème consiste, à partir de ces éléments, à décider d'avance quels sont tous les résultats possibles *x* de l'observation qui conduiront à rejeter l'hypothèse H. Ce faisant, on peut commettre deux espèces d'erreurs : rejeter H alors qu'elle est « vraie », ou la conserver alors qu'elle est « fausse ». Si *f* et *g* sont les densités de probabilité du point observé *x* sous H et K respectivement et si W (région critique) est l'ensemble des points conduisant à rejeter H, alors la probabilité de l'erreur de première espèce est égale à :

$$\alpha = \int_W f \, d\mu$$

et la probabilité de l'erreur de seconde espèce est égale à :

$$\beta = 1 - \int_W g \, d\mu ;$$

il ne faut pas perdre de vue toutefois que ces probabilités sont calculées dans des hypothèses différentes.

Formalisant et précisant quelque peu une attitude usuelle, Neyman propose comme principe de choix d'un test (ou, ce qui est équivalent, principe de choix d'une région critique) : fixer α à une valeur faible choisie a priori (seuil du test) et rendre β minimum. Ce point de départ a permis de rendre apparemment solide, dans la statistique classique, la théorie des tests. Des difficultés techniques apparaissent, cependant, lorsque H et K ne sont plus des hypothèses simples, mais des hypothèses composites (familles de lois de probabilité).

La principale critique qu'on peut faire à cette théorie porte sur le fait qu'elle ne prend pas en compte la confiance que l'on peut avoir a priori dans l'hypothèse que l'on teste ; en vérité, c'était précisément une des motivations des promoteurs de la théorie que d'éliminer tout recours à un tel concept. On peut maintenir ce point de vue ; mais, si l'on veut expliquer une pratique universellement acceptée (et à bon droit), il faudrait ajouter qu'on s'interdit de tester les hypothèses dans lesquelles on a toute confiance. En effet, si quelqu'un a obtenu quinze fois pile au cours d'une série de quinze tirages à pile ou face qu'il a exécutés lui-même avec honnêteté et une pièce de monnaie non truquée, il refusera certainement de mettre en doute l'hypothèse que la probabilité de pile est 1/2 et préférera considérer qu'un miracle s'est produit, comme il s'en produit à chaque instant.

Théorie des décisions statistiques

La théorie des décisions statistiques a joué un rôle remarquable par rapport à la statistique classique ; élaborée par Abraham Wald, dans les années quarante, pour parachever l'œuvre de J. Neyman et des statisticiens anglo-saxons, elle offre en effet un cadre logique unificateur à ce qui pouvait apparaître comme une collection de problèmes hétéroclites : l'estimation, les tests, les plans d'expérience peuvent être aisément présentés comme des cas particuliers de recherche d'une « fonction de décision statistique ». De plus, sa théorie a permis à Wald non seulement d'unifier ce qui existait déjà, mais également de faire progresser des techniques usuelles d'une importance pratique assez grande, les tests séquentiels.

Il convient de dire aussi que la théorie des décisions statistiques a contribué à susciter la critique des fondements logiques de la statistique classique, dont il sera question au chapitre suivant.

Donnons une description sommaire de la théorie élaborée par A. Wald ; on notera que son auteur était influencé par la théorie des jeux développée à la même époque par J. von Neumann et O. Morgenstern.

Soit un espace probabilisable (E, B), et une famille P de lois de probabilité sur cet espace. Le triplet (E, B, P) est appelé une structure statistique. Il sera commode de repérer une loi de probabilité appartenant à P par un paramètre θ, prenant ses valeurs dans Θ. Soit par ailleurs un ensemble D de décisions. Wald appelle fonction de décision statistique, ou stratégie, une application δ de E dans D. En fait, il introduit, sous le nom de stratégie mixte, des transitions de E dans D, où D est une σ-algèbre sur D ; une telle généralisation a un grand intérêt technique ; on se bornera ici aux applications δ, qui forment un ensemble Δ.

Quelles raisons le statisticien peut-il avoir de choisir telle ou telle fonction de décision ? Ces raisons seront contenues, pour la théorie, dans le choix d'une fonction de perte, qui est une application de Θ × D dans R représentant l'ensemble des coûts impliqués lorsqu'une décision *d* est prise, alors que la loi à laquelle obéissent les observations a pour paramètre θ. On peut alors calculer, pour une fonction de décision δ, l'espérance de la fonction de perte : c'est une application de Θ × Δ dans R que Wald appelle fonction de risque. Le but recherché par le statisticien, dans le choix d'une méthode, est de donner à la fonction de risque des valeurs aussi faibles que possible ; la difficulté fondamentale provient du fait que le statisticien choisit librement δ, mais n'a aucun contrôle sur θ (paramètre choisi par la « nature », dit Wald, par analogie avec la théorie des jeux). La fonction de risque n'induit sur Δ qu'un préordre partiel, généralement assez pauvre. Wald propose alors d'utiliser un critère, toujours inspiré par la théorie des jeux, permettant de déterminer dans tous les cas une fonction de décision optimale : c'est le critère du minimax, qui consiste à rendre minimale, par le choix de δ, la valeur maximale de la fonction de risque lorsque θ varie dans Θ.

Nous avons dit que la théorie des fonctions de décision statistique a joué un rôle unificateur : on peut voir en effet que les problèmes les plus classiques de l'inférence statistique en sont des cas particuliers.

On aura un problème d'estimation si l'on suppose que l'ensemble D des décisions est isomorphe à l'ensemble Θ des paramètres : on appellera alors T l'ensemble des décisions et on nommera arbitrairement un de ses éléments, *t*, une « valeur estimée du paramètre » ; on supposera de plus que la fonction de perte est une mesure de l'écart entre θ et *t*. D'ailleurs, si cette fonction est de la forme $c(t-\theta)^2$ et si l'on ajoute la condition, assez naturelle, que l'estimateur (fonction de décision) soit sans biais, alors on constate que les exigences de la théorie des décisions statistiques (minimiser la fonction de risque) recoupent parfaitement les règles de la statistique classique, puisqu'elles se traduisent par la recherche d'estimateur de variance minimum. Il est remarquable que ce problème puisse être résolu sans rencontrer les difficultés inhérentes à la présence de θ, pour la plupart des problèmes courants d'estimation, car le préordre induit sur l'ensemble des estimateurs par la fonction de risque est en général un préordre complet.

Pour voir comment la théorie générale des tests s'insère également dans la théorie des décisions statistiques, il suffit d'envisager le cas où l'ensemble des décisions se réduit à deux éléments d_1 et d_2 et où Θ est muni d'une dichotomie, c'est-à-dire d'une partition en deux classes $\Theta = \Theta_1 \cup \Theta_2$. Interprétant les décisions « faire comme si $\theta \in \Theta_i$ », décisions que l'on nommera d_i, on est conduit à définir une fonction de perte qui est nulle aux points (d_i, θ) avec $\theta \in \Theta_i$ et égale à l'unité quand $\theta \notin \Theta_i$. On montre alors aisément que la fonction de risque n'est autre que le couple des probabilités d'erreur de première et de seconde espèce définies dans la théorie des tests.

V - Fondements logiques

Dans sa théorie des fonctions de décision statistique, Wald avait introduit des solutions bayésiennes : ce sont les stratégies qui minimisent la valeur moyenne de la fonction de risque, au sens d'espérance par rapport à une distribution de probabilité a priori sur Θ. L'avantage technique est évident : cela permet, si l'on a choisi une distribution a priori particulière, d'obtenir un préordre complet sur Θ et donc de donner sans ambiguïté une solution à tout problème de décision statistique. Mais Wald, imprégné des idées de la statistique classique, n'adoptait pas cette interprétation et considérait qu'il n'est pas permis en général de choisir une distribution a priori sur Θ ; pour lui, les stratégies bayésiennes n'étaient qu'un outil intermédiaire, intéressant parce que l'ensemble de toutes les stratégies de cette nature (pour toutes les distributions a priori concevables) forme, sous des conditions assez larges, une classe complète de stratégies non dominées (dans le préordre induit par la fonction de risque) ; il suffit donc de rechercher une stratégie optimale dans cette classe, quelle que soit au reste la définition de l'optimum.

Depuis 1950 environ, d'autres statisticiens ont adopté un point de vue différent et proposent l'idée que, dans tel problème statistique concret, il convient de choisir une distribution a priori des paramètres inconnus des lois de probabilité qui traduise au mieux la connaissance imparfaite des phénomènes ; ce point de vue a trouvé un écho particulièrement large auprès des économistes. Nous allons donner quelques indications sur les racines et le développement de cette tendance, dite néo-bayésienne, de la statistique moderne.

Méthodes néo-bayésiennes

Si l'on veut bien accorder un sens à la notion de distribution de probabilité des paramètres, les problèmes d'inférence statistique se présentent sous un aspect bien différent de celui auquel s'attachent les statistiques classiques. Si $f(x, \theta)$ est la loi de probabilité des observations pour θ fixé, et si $\phi(\theta)$ est la densité a priori du paramètre θ, l'estimation du paramètre θ, lorsqu'on connaît les observations, conduit à rechercher la loi a posteriori de θ conditionné par *x*. On sait que le résultat est fourni par la formule de Bayes, soit :

$$\frac{\phi(\theta) f(x, \theta)}{\int \phi(\theta) f(x, \theta)\, d\theta}.$$

Sur le plan des principes, il n'y a pas grand-chose à ajouter, puisque cette formule exprime complètement, dans l'optique bayésienne, ce qu'on doit penser de θ, au vu des observations *x*. Mais en pratique, avec les expressions des lois d'observation

habituellement utilisées (qui appartiennent souvent à ce qu'on appelle des familles exponentielles), une expression du type donné ci-dessus n'a pas nécessairement une forme analytique simple, selon l'expression choisie pour $\phi(\theta)$. Deux possibilités s'offrent et sont effectivement en usage.

D'une part, on peut se borner à utiliser des expressions analytiques pour $\phi(\theta)$ qui conduisent à des expressions simples : c'est la théorie des familles de lois de probabilité conjuguées. Il s'agit de prendre pour $\phi(\theta)$ une expression telle que la distribution a posteriori conserve la même forme analytique, ses paramètres étant seuls modifiés. Ainsi, lorsque la loi des observations est une loi binomiale et si θ est le paramètre de fréquence de cette loi, on montre que la distribution conjuguée est la loi eulérienne β-incomplète. Il peut paraître un peu gênant, sur le plan des principes, de choisir les distributions a priori en fonction de la commodité analytique. Sur le plan pratique, cette objection est de portée limitée, car la loi β-incomplète choisie au départ dépend de deux paramètres et peut avoir une grande diversité de forme. Une autre objection tient à ce qu'on ne connaît pas de méthode générale pour trouver des familles de distributions conjuguées.

Une façon différente d'employer la méthode bayésienne pour l'estimation des paramètres consisterait à renoncer au calcul analytique et à demander que l'opinion a priori sur les paramètres soit exprimée sous forme numérique. L'emploi d'une calculatrice permet d'obtenir aisément la distribution a posteriori. On se prive ainsi de la possibilité d'obtenir des résultats généraux et synthétiques.

Interprétation concrète de la probabilité

Contrastant avec l'embarras des anciens auteurs, la définition du concept de probabilité ne soulève aujourd'hui plus aucune difficulté du point de vue mathématique. Cependant, des divergences de vue subsistent, concernant les objets du monde réel dont la probabilité, avec ses règles de calcul, est propre à fournir un modèle valable. En gros, on distingue deux tendances :

– Les objectivistes, ou fréquentistes, considèrent que la probabilité fournit un modèle idéalisé du comportement des fréquences avec lesquelles peuvent se produire certains résultats d'expériences effectuées dans des conditions suffisamment stables ; ils considèrent tout autre emploi pratique de la probabilité comme entaché d'arbitraire. C'est le point de vue qui a présidé au développement des théories sur l'inférence statistique classique.

– Les subjectivistes pensent que la probabilité est une mesure du degré de confiance que nous pouvons avoir vis-à-vis de tout événement incertain ; ce qui englobe le cas de la fréquence, mais aussi bien d'autres situations.

Le point de vue fréquentiste a dominé longtemps la statistique mathématique, depuis sa naissance en 1900, et demeure la règle dans certains domaines d'application, parfois peut-être avec raison (contrôle des fabrications industrielles) ; inversement, la critique et l'étude des fondements logiques se sont le plus souvent exercées en faveur du point de vue subjectiviste (B. de Finetti, J. L. Savage).

Théorie de la décision dans l'incertain

Ainsi, la théorie axiomatisée de la décision dans l'incertain, telle que la présente Savage, constitue l'une des meilleures justifications philosophiques de la probabilité subjective. Esquissons les grandes lignes de cette théorie.

Une situation de décision dans l'incertain est caractérisée par la donnée d'un ensemble S d'états, d'un ensemble K de conséquences, et les actes sont des applications de S dans K. Il s'agit de caractériser les préférences (cohérentes ou rationnelles) qu'un agent peut exercer sur l'ensemble des actes ; les sous-ensembles de S sont appelés événements. Les postulats de la théorie sont les suivants :

– Il existe un préordre complet sur l'ensemble F des actes.

– Si deux actes *f* et *g* comportent les mêmes conséquences pour tout état n'appartenant pas à un événement B, l'ordre de préférence entre *f* et *g* ne dépend que des conséquences qu'ils comportent pour les états appartenant à B.

– L'ordre de préférence entre deux conséquences ne dépend pas des états qui se réalisent.

– L'ordre de préférence sur les actes induit un préordre sur les événements (préordre de probabilité).

– Il existe des partitions de cardinal arbitrairement grand de S en événements presque équivalents : cela signifie qu'aucune réunion de $r + 1$ événements de la partition ne peut être plus probable qu'une réunion de r événements élémentaires.

Avec quelques précautions techniques, on montre alors qu'il existe une fonction, dite fonction d'utilité, sur l'ensemble K et une distribution de probabilité sur S telles que le préordre de préférence sur les actes soit le préordre induit par l'espérance mathématique de l'utilité.

En somme, la théorie montre que, si les choix sont cohérents (au sens des postulats), la probabilité subjective apparaît comme une nécessité logique. Cela semble être un résultat de même nature, mais de portée plus générale, que le théorème de Wald concernant les stratégies bayésiennes.

Il ressort aussi que cela apporte un argument extrêmement fort en faveur de la probabilité subjective, ce qui encourage à revenir à l'emploi des probabilités a priori dans l'inférence statistique, qui avait été proscrit par les statisticiens classiques. Il faut reconnaître cependant que le lien théorique avec les décisions statistiques ne semble pas parfaitement élucidé – Savage lui-même a achoppé sur ce point – et que, pour des raisons encore quelque peu obscures, les méthodes de la statistique classique se révèlent être bien adaptées à un certain nombre de situations pratiques. Mais les travaux de Finetti et Savage ont encouragé les statisticiens à revenir aux méthodes bayésiennes (dont on s'était détourné après les applications parfois un peu imprudentes des auteurs du XIX^{e} siècle : cf. chap. 1). Les économistes sont certainement, entre tous les praticiens, les plus attirés par les méthodes néo-bayésiennes ; cela paraît lié au fait que l'emploi de probabilités dans le domaine économique est, de toute façon, irrémédiablement entaché d'un caractère subjectif : il n'existe guère, dans ce domaine, d'expériences se déroulant dans des conditions rendant crédible l'interprétation objectiviste, qui suppose une longue suite d'expériences à conditions égales, et, de ce fait, les économistes sont moins que d'autres enclins à adopter le parti fréquentiste. Quoi qu'il en soit, un domaine d'investigation intéressant s'offre à la statistique appliquée : il s'agit de définir des critères permettant de distinguer les champs d'applications préférentiels respectifs de l'inférence classique et des méthodes néo-bayésiennes.

Georges MORLAT

Bibliographie

- M. ANDREFF, *Statistique : traitement des données d'échantillon*, 2 vol., Presses univ. de Grenoble, 1993-1994
- J.-R. BARRA, *Notions fondamentales de statistique mathématique*, Dunod, Paris, 1971
- J.-P. BENZECRI, *L'Analyse des données* : *l'analyse des correspondances, la taxinomie*, *ibid.*, 2 vol., 1982-1984
- J.-M. BOUROCHE & G. SAPORTA, *L'Analyse des données*, coll. Que sais-je ?, P.U.F., Paris, 5^{e} éd. 1992
- D. DACUNHA-CASTELLE & M. DUFLO, *Probabilités et statistiques*, 2 vol., Masson, Paris, 2^{e} éd. 1993-1994
- Y. DODGE, *Statistique : dictionnaire encyclopédique*, Dunod, 1993
- P. G. HOEL, *Statistique mathématique*, 2 vol., Armand Colin, Paris, 1991
- I.N.S.E.E., *Répertoire des sources statistiques françaises*, 2 vol., Paris, 1983 ; *Annuaire statistique de la France, 1993, résultats de 1992*
- M. G. KENDALL, A. STUART & J. K. ORD, *Kendall's Advanced Theory of Statistics*, 2 vol., Oxford Univ. Press, 5^{e} éd. 1992
- L. LEBART & J.-P. FÉNELON, *Statistique et informatique appliquées*, Dunod, 3^{e} éd. 1975
- T. M. PORTER, *The Rise of Statistical Thinking, 1820-1900*, Princeton Univ. Press, 1986
- D. G. REES, *The Foundations of Statistics*, Chapman & Hall, New York, 1987
- C. ROBERT, *L'Analyse bayésienne*, Economica, Paris, 1992
- S. M. STIGLER, *The History of Statistics, the Measurement of Uncertainty before 1900*, Belknap Press of the Harvard Univ. Press, 1986
- A. WALD, *Statistical Decision Functions*, Wiley, New York, 1950, rééd. Chelsea, New York, 1971
- H. WESTERGAARD, *Contributions to the History of Statistics*, Agathon Press, Inc., New York, 1986

Revues

- *Bulletin signalétique du C.N.R.S. n^{o} 110*, Paris
- *Journ. Soc. Statist. Paris*, Paris, depuis 1859
- *Biometrika*, Londres, depuis 1901
- *Rev. Statist. appliquée*, Paris, depuis 1953
- *Publications de l'Institut de statistique de l'Univ. de Paris*, dep. 1957

- *Canadian Journal of Statistics*, Montréal, dep. 1973
- *Publications du laboratoire de statistique et probabilités*, Toulouse, dep. 1980
- *Statistique mathématique et probabilités*, Economica, Paris, 1994.