QUE SAIS-JE ?

Histoire de la statistique

JEAN-JACQUES DROESBEKE
Professeur à l'Université Libre de Bruxelles

PHILIPPE TASSI
Directeur scientifique de Médiamétrie
Professeur à l'ENSAE

ISBN 2 13 043014 7

Dépôt légal — 1re édition : 1990, août

© Presses Universitaires de France, 1990
108, boulevard Saint-Germain, 75006 Paris

CHAPITRE PREMIER

L'HISTOIRE ? UNE (CERTAINE) HISTOIRE

Le mot *Histoire*, qui date de la Renaissance, provient du grec ancien *historiè* qui, chez Hérodote, a le sens d' « enquête ». Définir l'histoire est un problème ardu. Nous sommes sensibles à la conception de Pierre Salmon selon qui « l'histoire est une reconstitution intelligible et critique du passé vécu par les hommes en société » (*Histoire et Critique*, Editions de l'Université de Bruxelles, 1987, p. 20). Toute reconstitution, si objective soit-elle sur le fond, est subjective par la forme et le langage utilisé. Elle l'est encore davantage quand on doit l'inscrire dans les cent vingt-huit pages d'un « Que sais-je ? » !

Aussi avons-nous choisi délibérément d'écrire *une certaine histoire*, celle qui peut intéresser l'enseignant qui utilise la méthodologie statistique, le chercheur qui écrit un article ou participe à un congrès, l'utilisateur soucieux de ne pas se limiter aux aspects techniques d'une étude.

Notre intérêt s'est tout d'abord porté sur la statistique descriptive (chapitre II). Le chapitre III est consacré à une présentation très succincte des fondements de *la probabilité*, en portant notre attention sur les aspects mathématiques de ce concept.

Le chapitre IV porte sur *les méthodes de sondage*. Leur histoire et leur usage constituent à nos yeux un

thème suffisamment important pour le mettre ainsi en évidence.

La statistique mathématique est traitée dans les chapitres V et VI. Le premier de ceux-ci aborde *l'émergence de l'inférence statistique (estimation et tests)*. Le second est centré sur la *statistique non paramétrique et la robustesse*. Nous avons voulu éviter de trop aborder l'époque contemporaine, reprenant à notre compte les propos que Quételet tenait dans la préface de son ouvrage intitulé *Sciences mathématiques et physiques au commencement du XIXe siècle* : « Nous sommes trop près des événements pour pouvoir les juger avec tout le discernement et toute l'impartialité nécessaires. » Ce principe sera à nouveau appliqué dans les chapitres suivants.

Les *séries chronologiques* font l'objet du chapitre VII, étant donné la spécificité et l'importance de leur étude. Un autre thème que nous ne pouvions pas éviter est *l'analyse multivariée* et, tout particulièrement, l'*analyse des données* au sens français du terme. Il constitue la matière du chapitre VIII. Bernard Fichet nous a fait don de ce chapitre. Nous l'en remercions chaleureusement.

Enfin, le chapitre IX est dédié à *quelques grands noms* auxquels il est souvent fait référence. Nous n'en avons choisi que huit : Fisher, Galton, Gini, Gosset, Neyman, Pearson (I et II) et Quételet. D'autres noms, plus anciens, sont évoqués dans les chapitres précédents. Les plus récents nous pardonneront de ne pas les citer ici. Ils savent que leur qualité et leur mérite sont connus de tous.

Pour conclure, nous n'avons pas voulu aborder l'*histoire des statistiques*, publiques, officielles ou privées. Nous renvoyons le lecteur intéressé aux ouvrages publiés chez Economica et intitulés *Pour une histoire de la Statistique* (tomes 1 et 2 : 1987).

CHAPITRE II

LA STATISTIQUE DESCRIPTIVE

I. — Introduction

Dans sa quinzième lettre à SAR le duc régnant de Saxe-Cobourg et Gotha intitulée *La statistique est-elle un art ou une science ?*, Adolphe Quetelet écrit :

« ... toutes les sciences d'observation, à leur début, ont subi les mêmes phases ; c'étaient des arts, car elles se bornaient à grouper d'une manière plus ou moins heureuse des collections de faits appartenant à un même ordre de choses ; et c'est par le rapprochement et l'étude de ces faits qu'elles se sont élevées ensuite au rang où on les voit briller aujourd'hui. Pourquoi se montrer plus exigeant envers la statistique ? »

Il est exact que la statistique a d'abord été descriptive et a consisté au début à observer des faits. Nous évoquerons d'ailleurs l'origine du mot *statistique* dans le chapitre IV. Mais on peut se demander quand est apparu le besoin d'ordonner des observations, de les représenter dans des tableaux ou des graphiques, de rechercher des valeurs typiques qui porteront plus tard le nom de paramètres de position, de dispersion ou de forme, d'étudier simultanément deux ou plusieurs caractéristiques, de construire des outils spécifiques. C'est à ces questions que nous voulons tenter de répondre dans ce chapitre.

II. — Représentation graphique

La représentation graphique « quantitative » trouve son origine dans la construction de cartes géographi-

thème suffisamment important pour le mettre ainsi en évidence.

La statistique mathématique est traitée dans les chapitres V et VI. Le premier de ceux-ci aborde *l'émergence de l'inférence statistique (estimation et tests)*. Le second est centré sur la *statistique non paramétrique et la robustesse*. Nous avons voulu éviter de trop aborder l'époque contemporaine, reprenant à notre compte les propos que Quételet tenait dans la préface de son ouvrage intitulé *Sciences mathématiques et physiques au commencement du XIXe siècle* : « Nous sommes trop près des événements pour pouvoir les juger avec tout le discernement et toute l'impartialité nécessaires. » Ce principe sera à nouveau appliqué dans les chapitres suivants.

Les *séries chronologiques* font l'objet du chapitre VII, étant donné la spécificité et l'importance de leur étude. Un autre thème que nous ne pouvions pas éviter est *l'analyse multivariée* et, tout particulièrement, l'*analyse des données* au sens français du terme. Il constitue la matière du chapitre VIII. Bernard Fichet nous a fait don de ce chapitre. Nous l'en remercions chaleureusement.

Enfin, le chapitre IX est dédié à *quelques grands noms* auxquels il est souvent fait référence. Nous n'en avons choisi que huit : Fisher, Galton, Gini, Gosset, Neyman, Pearson (I et II) et Quételet. D'autres noms, plus anciens, sont évoqués dans les chapitres précédents. Les plus récents nous pardonneront de ne pas les citer ici. Ils savent que leur qualité et leur mérite sont connus de tous.

Pour conclure, nous n'avons pas voulu aborder l'*histoire des statistiques*, publiques, officielles ou privées. Nous renvoyons le lecteur intéressé aux ouvrages publiés chez Economica et intitulés *Pour une histoire de la Statistique* (tomes 1 et 2 : 1987).

CHAPITRE II

LA STATISTIQUE DESCRIPTIVE

I. — Introduction

Dans sa quinzième lettre à SAR le duc régnant de Saxe-Cobourg et Gotha intitulée *La statistique est-elle un art ou une science ?*, Adolphe Quetelet écrit :

« ... toutes les sciences d'observation, à leur début, ont subi les mêmes phases ; c'étaient des arts, car elles se bornaient à grouper d'une manière plus ou moins heureuse des collections de faits appartenant à un même ordre de choses ; et c'est par le rapprochement et l'étude de ces faits qu'elles se sont élevées ensuite au rang où on les voit briller aujourd'hui. Pourquoi se montrer plus exigeant envers la statistique ? »

Il est exact que la statistique a d'abord été descriptive et a consisté au début à observer des faits. Nous évoquerons d'ailleurs l'origine du mot *statistique* dans le chapitre IV. Mais on peut se demander quand est apparu le besoin d'ordonner des observations, de les représenter dans des tableaux ou des graphiques, de rechercher des valeurs typiques qui porteront plus tard le nom de paramètres de position, de dispersion ou de forme, d'étudier simultanément deux ou plusieurs caractéristiques, de construire des outils spécifiques. C'est à ces questions que nous voulons tenter de répondre dans ce chapitre.

II. — Représentation graphique

La représentation graphique « quantitative » trouve son origine dans la construction de cartes géographi-

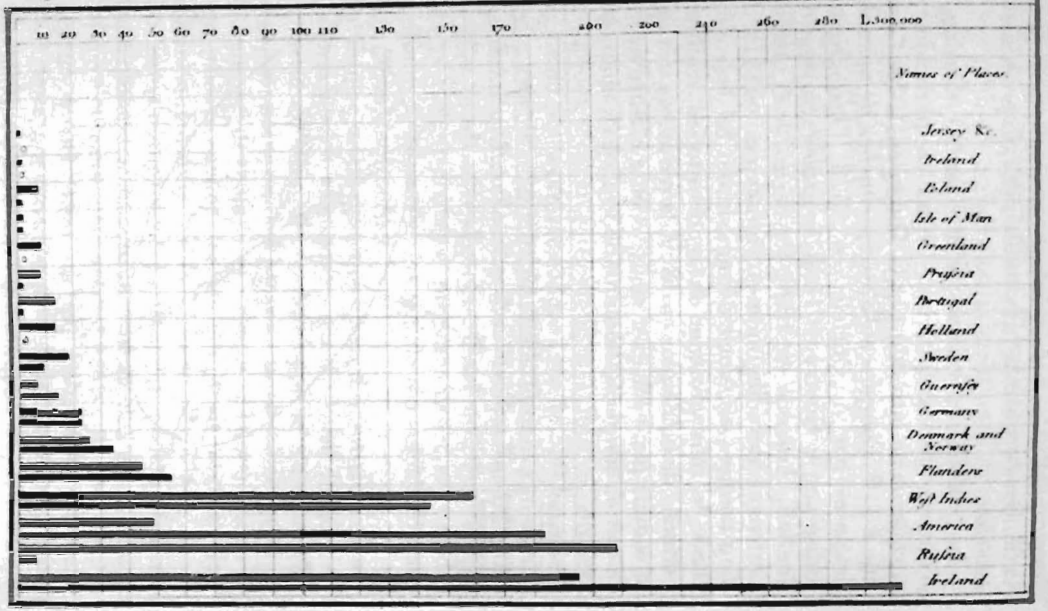
ques. Les plus anciennes d'entre elles qui nous sont parvenues datent d'environ 6 000 ans, gravées sur des tablettes d'argile en Mésopotamie. Les graphiques statistiques sont eux plus récents.

Leur apparition est liée à l'utilisation de coordonnées dans un plan. Déjà abordé par Léonard de Vinci pour analyser la vitesse d'objets en mouvement, ce problème a surtout été traité par René Descartes, en 1637, dans un appendice du *Discours de la Méthode* intitulé *La Géométrie.* A cette époque, l'intérêt d'une représentation dans un système d'axes coordonnés ne s'impose pas immédiatement, même si Edmund Halley en fait usage avec succès en 1686 pour représenter des mesures barométriques en fonction de l'altitude.

Parmi les utilisateurs de représentations « cartésiennes », il faut distinguer Jean-Henri Lambert (1728-1777) qui naquit à Mulhouse, ville libre alors associée à la Suisse. Ce mathématicien s'est illustré par des études très diversifiées. Sa contribution à la statistique s'est principalement développée dans ses études sur la théorie des erreurs, présentées en 1760 dans un ouvrage intitulé *Photometria* (cf. chapitre III). Sa correspondance avec Daniel Bernoulli et son approche des probabilités non additives sont aussi dignes d'intérêt. Mais, pour ce qui nous concerne dans ce chapitre, il faut souligner la qualité exemplaire des représentations graphiques de Lambert, ce qui a vraisemblablement joué un rôle important dans ses réflexions.

Près de 25 ans plus tard, un autre personnage va marquer la statistique descriptive de son empreinte : William Playfair (1759-1823). Ce dernier introduit plusieurs outils importants. Dans son ouvrage intitulé *The commercial and political atlas*, publié à Londres en 1786, Playfair présente 44 graphiques de grande qualité. 43 d'entre eux concernent des séries chronologiques (cf. chapitre VII). Le 44e est le premier *diagramme en barre* connu :

Exports and Imports of **SCOTLAND** to and from different parts for one Year from Christmas 1780 to Christmas 1781



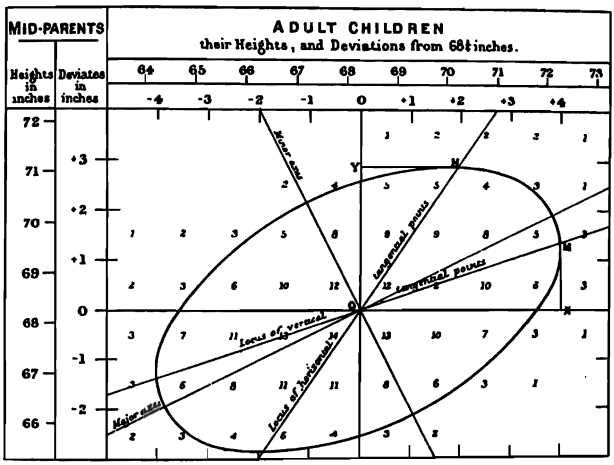
The Upright divisions are Ten Thousand Pounds each. The Black Lines are Exports the Ribbed lines Imports

Published as the Act directs June 7th 1785 by Wm Playfair

Neele sculp. 352 Strand London

Un autre de ses ouvrages, intitulé *The Statistical breviary* (1801), constitue aussi une référence essentielle. On y trouve la présentation des *diagrammes en secteurs*.

Le XIXe siècle voit apparaître ou se développer d'autres outils graphiques. Citons en particulier le *cartogramme* introduit par C. Dupin en 1819, l'*ogive* due à J.-B. Fourier en 1821, les *courbes de mortalité* dessinées par A. Quételet en 1828, l'*histogramme* utilisé par A. Guerry en 1833, la *grille logarithmique* due à L. Lalanne en 1843, les cartes de C. Minard contenant des diagrammes statistiques (1861), les *diagrammes semi-logarithmiques* de S. Jevons en 1863, le *stéréogramme* ou représentation tridimensionnelle de G. Zeuner en 1869, la *pyramide des âges* introduite par F. Walker en 1874 et la célèbre *surface de corrélation* de Galton en 1885 (cf. § II, chap. IX).



Le XXe siècle n'est pas en reste. On y constate périodiquement des développements qui vont de la construction de diagrammes statistiques sophistiqués

aux dessins multivariés de H. Chernoff et aux diagrammes d'analyse exploratoire de J. Tukey. Nous ne pouvons ici qu'inviter le lecteur intéressé à consulter la bibliographie concernée dans laquelle nous conseillons les articles et ouvrages mentionnés à la fin de ce livre.

III. — Valeurs typiques d'une série statistique univariée

Qui dit « valeur typique d'une série » pense au problème du choix d'un paramètre résumant l'information contenue dans une série d'observations. Il semble que les premiers paramètres de position qui aient été utilisés soient, assez naturellement, le *mode* (valeur apparaissant le plus fréquemment) et le « *milieu de l'intervalle défini par les valeurs extrêmes* » (en anglais : *midrange*). Comme beaucoup de concepts scientifiques, c'est en astronomie qu'il faut rechercher l'origine de ces paramètres. Il y a près de 2 500 ans, les Babyloniens ont établi des procédés destinés à mesurer les mouvements du soleil et des planètes sur base de relevés à intervalles réguliers. Bien que l'on sache qu'ils en étaient arrivés à proposer des « valeurs de compromis », nous n'avons pas de témoignage quant à leur façon de procéder.

Par contre, nous possédons plus d'informations sur la démarche des astronomes grecs. Ainsi, Ptolémée, astronome du IIe siècle, a développé le système qui porte son nom en se basant sur les découvertes d'Aristarque de Samos et surtout d'Hipparque, né quatre siècles avant lui et qui avait obtenu des résultats importants sur la position des étoiles et la périodicité de leur retour. Le fait de posséder plusieurs valeurs observées ont conduit ces astronomes à proposer des valeurs uniques accompagnées de mesure de variation basée, semble-t-il, sur l'étendue des observations. Le choix systématique d'une moyenne ou d'une médiane n'était cependant pas encore d'actualité.

Le recours à des observations « nombreuses » d'une même quantité pour en estimer la valeur apparaît dans l'œuvre de Tycho Brahé (1546-1601) qui constitua un ensemble de données sur le mouvement des planètes tel qu'il permet à Képler de formuler ses lois sur le sujet (Tychonis Brahe Dani (1602), *Opera Omnia*, Tomus II). Il apparaît clairement que Tycho Brahé a recours à une *moyenne arithmétique* pour éliminer des erreurs d'observations. En 1722, Roger Cotes, dans une annexe de son ouvrage *Harmonia Mensurarum*, considère le problème de connaître la position exacte d'un point pour lequel il dispose de 4 observations qui ne sont pas toutes aussi fiables l'une que l'autre. Il propose dès lors d'utiliser une *moyenne pondérée* dont les coefficients sont inversement proportionnels à la « dispersion » des erreurs d'observations. Mais dès cette époque l'étude de ces paramètres va s'effectuer dans un contexte probabiliste et plus particulièrement dans celui de la théorie des erreurs. Nous reviendrons sur cette théorie dans le chapitre III. Signalons cependant, dès à présent, que le premier à défendre l'usage de la moyenne arithmétique, en calculant sa distribution, semble être Thomas Simpson qui, en 1756, rédigea « A letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations in practical astronomy », *Phil. Trans. Roy. Soc. Lond.*, vol. 49, p. 82-93.

Remarquons encore que la *médiane* voit naître son intérêt à la même époque, en 1757, sous la plume de Roger Joseph Boscovich. Notons enfin que la *moyenne géométrique* et la *moyenne harmonique* ont été introduites par William Stanley Jevons en 1874, dans *The principles of science* paru chez Macmillan, à Londres.

L'idée de variabilité n'est pas récente. Nous en avons parlé à propos des astronomes grecs. Elle est aussi intimement liée à la théorie des erreurs que nous évoquerons dans le chapitre suivant. C'est au XVIIIe siè-

cle que les astronomes (encore eux !) eurent pour objectif d'utiliser des mesures expérimentales afin de déterminer la position d'objets dans le ciel. Ils en arrivèrent ainsi, « tout naturellement », à étudier la *distribution des erreurs*. Un précurseur en la matière est Galileo Galilée qui, en 1632, s'intéresse à la détermination de la distance entre la terre et une étoile nouvelle. Il se penche à ce propos sur 78 mesures d'élévation de cette étoile fournies par 13 observateurs répartis à plusieurs endroits. Dans son écrit intitulé *Dialogo sopra i due massimi sistemi del mondo : Tolemaico e Copernicano*, publié à Florence, Galilée écrit :

« Etant donné que ces observateurs sont capables, qu'ils ont tous fait des erreurs, et que leurs erreurs doivent être corrigées pour nous permettre d'avoir la meilleure information possible à partir de ces observations, nous nous consacrerons à appliquer les modifications minimales et les corrections les plus petites possibles, juste assez pour ôter les observations de l'impossible et les remettre dans le possible... »

La variance naît au XIXe siècle avec les moindres carrés (cf. § IV) ; l'écart type suit. Mais les termes usités actuellement ne sont pas si anciens. Ce n'est qu'en 1893 que Karl Pearson introduit le terme *standard deviation* , noté σ.

Carl Friedrich Gauss propose, en 1816, l'usage de l'expression suivante pour mesurer les variations des erreurs de mesure en astronomie :

$$\varepsilon_k = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^k\right)^{1/k}$$

où x_i désigne la valeur absolue de l'écart de la i-ème observation à la moyenne arithmétique, n est le nombre d'observations et k un nombre naturel. Nous retrouvons ainsi l'écart type $\sigma = \varepsilon_2$ préféré par Gauss à toutes les autres mesures alors que ε_1 avait reçu les faveurs de Laplace (cf. § IV).

Au XIXe siècle, on utilisera plusieurs mesures de

dispersion apparentées à l'écart type : le *module* C ($= \sigma\sqrt{2}$), l'*erreur moyenne* (C/π), le *carré moyen* ($C^2/2 = \sigma^2$)... Parmi les autres mesures de variation introduites à cette époque, citons en particulier celle basée sur l'écart entre les quartiles introduites par A. Quételet en 1846 et reprise par F. Galton en 1875.

Notons enfin que les principales mesures d'asymétrie et d'aplatissement, appelées couramment *paramètres de forme*, apparaissent, avec les coefficients b_1, g_1, b_2 et g_2, dans le cadre des travaux de K. Pearson et R. A. Fisher (cf. chapitre IX). Il est cependant intéressant de noter que Karl Pearson introduit en 1895 un paramètre basé sur la position relative de la moyenne μ et du mode M, et défini par $(\mu - M)/\sigma$. Il constate en outre, empiriquement, que pour les distributions de type III qu'il introduit il obtient la relation suivante :

$$\mu - M \approx 3\,(\mu - m)$$

où m désigne la médiane. Il s'ensuivit l'idée que les trois paramètres μ, m et M se présentaient toujours dans cet ordre ou dans l'ordre opposé. La recherche de conditions permettant d'affirmer ce résultat a suscité beaucoup d'intérêt, jusqu'à l'aube des années 1980 (cf. par exemple, W. R. Van Zwet (1979), Mean, median, mode II, *Statist. Neerl.*, vol. 33, p. 1-5).

IV. — Les ajustements

Le problème de l'ajustement d'un ensemble de points représentés dans un système d'axes coordonnés par une droite (et plus généralement par une courbe) occupe une place essentielle dans le développement de la statistique.

Au XVIIIe siècle, Leonhard Euler (dans *Inégalités du mouvement de Saturne et de Jupiter*, Prix de l'Académie royale des Sciences, 1748, Paris) et Johan Tobias Mayer (Abhandung über die Umwälzung des Mondes um seine Axe, *Kosmog. Nachr. und Samm.*, vol. 1,

p. 52-183) développent indépendamment l'un de l'autre la *méthode des moyennes* permettant d'ajuster des points observés par une droite. Cette méthode consiste à partager l'ensemble des observations en deux groupes et à imposer dans chacun de ceux-ci que la somme des *résidus* (écarts entre observations et droite) soit égale à zéro. Cette méthode pouvait être étendue à un ajustement à plus de deux paramètres en divisant l'ensemble des observations en autant de groupes qu'il y avait de paramètres à déterminer. C'est ainsi que Mayer utilise 27 observations de la position d'un point situé sur la surface lunaire pour déterminer les trois coefficients d'une courbe d'ajustement en divisant ces observations en 3 groupes de 9 valeurs. Le fait de devoir choisir des groupes souligne le caractère subjectif de cette démarche. Remarquons en passant que, dans le cadre de ces travaux, Euler fut aussi le premier à s'intéresser au problème de la minimisation du résidu maximum, inaugurant ainsi le principe du *minimax*.

Quelques années plus tard, en 1755, Roger Joseph Boscovich, jésuite d'origine serbe, est envoyé par le pape Benoît XIV, avec son collègue Christopher Maire, pour effectuer des mesure astronomiques. Dans leur rapport, écrit en latin et dont la traduction française (parue en 1770) porte le titre précis de *Voyage Astronomique et Géographique dans l'Etat de l'Eglise, entrepris par l'Ordre et sous les Auspices du Pape Benoît XIV, pour mesurer deux degrés du méridien, et corriger la carte de l'Etat ecclésiastique*, les auteurs tentent de déterminer la meilleure valeur de l'ellipticité de la terre à partir de 5 observations de degrés du méridien. Dans ce contexte, Boscovich propose, deux ans plus tard, un ajustement linéaire en imposant deux conditions :

1) la somme des corrections positives et négatives (résidus) doit être nulle ;

2) la somme de toutes les (valeurs absolues des) corrections, positives et négatives, doit être aussi petite que possible.

Boscovich présente en outre, à la fin de la version française du rapport dont il est fait mention ci-dessus, une méthode géométrique de résolution pratique de son ajustement. Cette idée sera reprise par Pierre Simon, marquis de Laplace, en 1789, sous une forme qualifiée par ce dernier de « plus simple » et désignée sous le terme *méthode de situation* , dans son mémoire intitulé *Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule.*

Les travaux de Boscovich vont être éclipsés par la *méthode des moindres carrés* qui apparaît près de 50 ans plus tard. Il est peut-être encore intéressant de noter que Boscovich s'interrogea aussi sur l'intérêt de calculer une moyenne arithmétique d'une série dont on a exclu la plus petite et la plus grande observation. Par ailleurs, il fallut attendre un siècle pour voir Francis Ysidro Edgeworth recommander l'usage simultané de la médiane (plutôt que celui de la moyenne arithmétique) et de l'estimateur de Boscovich de la pente (plutôt que celui des moindres carrés) pour réduire l'influence des valeurs discordantes.

Le premier texte paru faisant mention de *la méthode des moindres carrés* est dû à Adrien-Marie Legendre qui, dans ses *Nouvelles méthodes pour la détermination des orbites des Comètes* publiées en 1805 chez Courcier, à Paris, présente un appendice sur ce sujet. On y lit le texte significatif suivant :

« De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre *minimum* la somme des quarrés des erreurs. Par ce moyen, il s'établit entre les erreurs une sorte d'équilibre qui empêchant les extrêmes de prévaloir, est très-propre à faire connoître l'état du systême le plus proche de la vérité. »

En 1806, Carl Friedrich Gauss écrit dans « II Comet vom Jahr 1805 », *Monatliche Correspondenz zur Beförderung der Erd- und Himmelskunde*, vol. 14, p. 181-186 :

« Je n'ai pas encore vu les travaux de Legendre. Je n'ai intentionnellement pas pris la peine de le faire afin que le travail sur ma méthode reste ma propre idée. Au travers de quelques mots sur *la méthode des moindres quarrés* que de la Lande indique dans la dernière *Histoire de l'Astronomie 1805*, j'en arrive à la supposition qu'un théorème fondamental que j'ai moi-même utilisé depuis 12 ans dans beaucoup de calculs et que j'utiliserai aussi dans mon livre, a aussi été employé par Legendre. »

Indépendamment de Legendre et Gauss, il faut citer une autre référence relative à cette méthode : elle concerne Robert Adrain qui, en 1808, publie un article intitulé « Research concerning the probabilities of the errors which happen in making observations », *Analyst*, vol. 1, p. 93-109, et qui introduit aussi les moindres carrés.

C'est avec l'apparition de la « loi normale » que cette méthode va pouvoir être justifiée. Nous présenterons dans le chapitre III les raisons pour lesquelles « les moindres carrés » allaient devenir, pour longtemps, *la méthode* d'ajustement.

V. — Corrélation et Régression

La paternité de la corrélation a donné lieu à une littérature intéressante. On constate ainsi que Gauss aurait pu introduire ce concept dans l'étude de la distribution normale s'il n'avait pas basé son étude sur la considération de variables indépendantes. Il est fait, par ailleurs, référence à Auguste Bravais (1811-1863) qui, en 1846, publia un article intitulé « Sur les probabilités des erreurs de situation d'un point », *Mém. Acad. roy. Sci. Inst. France*, vol. 9, p. 255-332. Ce scientifique fut considéré comme le père de la corrélation par Karl Pearson en 1895 ; ce dernier

revint sur cet avis en 1920, faisant à Bravais des reproches semblables à ceux qu'il adressait à Gauss. Sans vouloir sous-estimer l'importance des travaux de Gauss et Bravais, nous devons cependant souligner le rôle essentiel joué par Francis Galton, à plusieurs titres (nous évoquerons son œuvre dans le chapitre IX). Ses travaux l'ont amené de façon tout à fait spécifique à souligner la nécessité d'une mesure de corrélation dans l'analyse des séries bivariées. Mais c'est par le concept de régression qu'il débute. Le 9 février 1877, Galton fait un exposé au Royal Institution of Great Britain, intitulé *Typical laws of heredity in man*. Dans le texte de son allocution (publiée dans les *Proc. R. Inst. G. Brit.*, vol. 8, p. 282-301), Galton écrit :

« Reversion is the tendency of the ideal mean filial type to depart from the parental type, reverting to what may be roughly and perhaps fairly described as the average ancestral type. If family variability had been the only process in simple descent that affected the characteristics of a sample, the dispersion of the race from its mean ideal type would indefinitely increase with the number of generations, but reversion checks this increase, and brings it to a standstill. »

Galton exprime ensuite le désir de construire un *coefficient de réversion r* qui indique la réduction de variabilité de la famille. Peu après, cette réversion se mutera en *régression*.

En 1888, dans un article intitulé « Correlations and their measurement chiefly from anthropometric data », *R.S. Proc.*, vol. 45, p. 135-145, il écrit :

« It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be *nil*. Between these two extremes are an endless number of intermediate cases, and it will be shown how the closeness of co-relation in any particular case admits of being expressed by a single number. »

Ce nombre est « notre » coefficient de corrélation *r*. Dans le même article, Galton utilise le terme de *partial co-relation* annonçant la corrélation multiple.

En 1896, Karl Pearson reprend les concepts de Galton pour leur donner la forme que nous connaissons actuellement, dans un article intitulé « Mathematical contributions to the theory of evolution, III : Regression, heredity and panmixia », *Phil. Trans.*, Ser. A, vol. 187, p. 253-318. Dans ce même article, Pearson établit la théorie de la corrélation à trois variables. Dans ce contexte, il faut aussi souligner les contributions de F. Y. Edgeworth et G. U. Yule.

D'autres *mesures d'association* allaient naître au début du XXe siècle. Citons à ce sujet le coefficient de *corrélation de rang* introduit en 1904 par Charles E. Spearman dans une étude psychologique consacrée à l'intelligence. Le coefficient TAU (τ) de Kendall date de 1938. Son origine se trouve cependant dans les écrits de Fechner (1897), Lipps (1906) et Deuchler (1914), tous trois scientifiques allemands. Par ailleurs, G. U. Yule consacre en 1900 une étude sur les mesures d'association concernant des variables dichotomiques tandis que Karl Pearson introduit en 1904 la statistique « classique » du chi-deux (ou chi-carré). Notons encore que la *causalité* allait être abordée par Serwall Wright en 1918 via la *path analysis* et que la notion de corrélation allait s'étendre vers d'autres domaines comme l'analyse des séries chronologiques (cf. chapitre VII) et l'analyse multivariée (cf. chapitre VIII).

VI. — Les indices

La réflexion sur les nombres-indices est née voici trois siècles environ en Angleterre du besoin de mesurer par un et un seul nombre les variations d'un ensemble de prix entre deux dates. En effet, si les indices sont utilisés dans de nombreux domaines, tels que la production industrielle, la démographie, l'agri-

culture, etc., l'histoire nous apprend que les premières applications ont été l'étude du niveau général des prix ou, parallèlement, du pouvoir d'achat. Comment combiner les modifications relatives de prix et de quantités pour obtenir un indice interprétable, pour les prix, comme une mesure d'une variation du niveau des prix, et, pour les quantités, comme une mesure du bien-être ?

La première réflexion sur le sujet semble remonter au *Chronicon Preciosum*, publié en 1707 par l'évêque W. Fleetwood, dans le contexte suivant : les statuts d'un collège d'Oxford, créé vers 1450, exigeaient que tout étudiant dispose d'un revenu annuel maximum de cinq livres, sous peine d'exclusion. Que signifiait cette contrainte en 1700, compte tenu de la dépréciation monétaire ? W. Fleetwood tente de répondre à cette question en étudiant, depuis 1450, l'évolution des prix du blé, de la viande, des boissons et du drap à vêtements, et, plus précisément, en évaluant le prix en 1700 des quantités que l'on pourrait obtenir vers 1450. Ses conclusions fixent un niveau de revenu équivalent à 30 livres pour 1700, ce qui correspond à une inflation moyenne de 7,4 % par décennie.

Fleetwood ne se pose pas le problème de l'agrégation des indices élémentaires pour les quatre biens qu'il a retenus, tout simplement parce que ces indices prennent la même valeur. Cette question apparaît en 1738, lorsque le Français Dutot désire évaluer l'évolution des prix en France de 1515 à 1735. Il recherche les prix d'un grand nombre d'articles existant à ces deux dates et mesurés par la même unité, soit $(p_k(0), k = 1 \text{ à } K)$ et $(p_k(1), k = 1 \text{ à } K)$, $p_k(0)$ (resp. $p_k(1)$) représentant le prix unitaire du bien k en 1515 (resp. en 1735). L' « indice des prix » de Dutot est défini par :

$$\sum_{k=1}^{K} p_k(1) / \sum_{k=1}^{K} p_k(0).$$

Sur les données recueillies, qui seront parfois critiquées par des auteurs de la même époque, l'indice vaut 22, soit une « inflation » moyenne par décennie de l'ordre de 15 %. A noter que le célèbre indice « Dow-Jones » est calculé selon une formule proche de celle de Dutot.

En 1764, Gian Rinaldo Carli, professeur d'astronomie à Padoue et Milan, publie un traité sur la dépréciation monétaire depuis la découverte de l'Amérique, plus précisément de 1500 à 1750. Contrairement à Dutot, et avec des notations analogues (0 pour 1500 et 1 pour 1750), son indicateur de mesure est la moyenne des indices élémentaires :

$$\frac{1}{K} \sum_{k=1}^{K} p_k(1)/p_k(0).$$

L'indice de Carli va être utilisé en Angleterre par George Shuckburgh Evelyn ; ce dernier publie, en 1798, les valeurs de cet indicateur de 1050 à 1800 (après extrapolation), ce qui fournit une dépréciation moyenne par décennie de l'ordre de 4 % sur la période considérée, ainsi d'ailleurs que de 1450 à 1700 (chiffre à comparer avec précaution avec ceux de Fleetwood car les biens retenus ne sont pas les mêmes, sans parler des incertitudes de mesure).

Proche de la formule de Carli, on voit apparaître au début du XIXe siècle un autre indicateur qui est la moyenne géométrique des indices élémentaires. Dans le même esprit, C. Walsh propose, en 1901, dans le cas où l'on a T instants d'observations, de considérer le produit des facteurs $(p_k(1)/p_k(0))^{\alpha_k}$, où les coefficients α_k dépendent des quantités aux T dates.

Si la notion de moyenne pondérée apparaît chez Arthur Young (1812) dans sa critique des travaux de Evelyn, le « père » de la théorie des indices peut être considéré comme étant Joseph Lowe (1822). Celui-ci introduit l'idée de comparer sur deux années consécu-

tives 0 et 1 la valeur d'un panier fixe de K biens consommés en quantités $(q_1, \ldots, q_k)$ aux prix $p(0)$ et $p(1)$, ce qui conduit à l'indicateur :

$$\sum_{k=1}^{K} q_k\, p_k(1) / \sum_{k=1}^{K} q_k\, p_k(0).$$

Si les quantités sont prises en l'année 0, cette formule prendra plus tard le nom d'indice de Laspeyres (1864, 1871) ; si elles sont prises en la date 1, on retrouve l'indice dit de Paasche (1874, 1878).

M. Drobisch propose une moyenne simple des formules de Laspeyres et de Paasche. Plus généralement, en 1887, A. Marshall reprend la formulation de Lowe, et prend pour q_k la moyenne des quantités en 0 et 1.

La pratique a privilégié les indices de Laspeyres et de Paasche, le premier en particulier. L'indice de Laspeyres est très utilisé pour les divers indices élaborés par les Instituts de statistiques officielles. En effet, on peut l'écrire sous la forme $\sum_k (p_k(1)/p_k(0))\ \alpha_k(0)$, où $\alpha_k(0) = p_k(0)\ q_k(0) / \sum_j p_j(0)\ q_j(0)$ représente le poids de la dépense en bien k dans la dépense totale de l'année 0. Les coefficients α_k (0) sont donc calculables une fois pour toutes tant que la composition de l'indice n'est pas modifiée.

Il ne saurait être question de citer de façon exhaustive toutes les formules d'indices qui ont été proposés dans la littérature statistique. En 1922, Irwing Fisher tente une classification des indices existant selon des propriétés définissant l'indice « idéal ». Son célèbre ouvrage (*The making of index numbers : a study of their varieties, tests and reliability*, Boston, Houghton Miffin, 1922) recense 134 formules distinctes. Tout au plus peut-on mentionner pour leur originalité les travaux de F. Divisia (1925), statisticien français, qui se place en temps continu ; si les vecteurs prix et

quantités à la date t sont notés $p(t)$ et $q(t)$, l'indice de Divisia des quantités $Q(t)$ est défini par l'équation différentielle :

$$d \text{ Log } Q\ (t) = \sum_{k=1}^{K} \left[\frac{p_k(t)\ q_k(t)}{\sum_{k=1}^{K} p_k(t)\ q_k(t)} d \text{ Log } q_k(t) \right].$$

Pour conclure, comment évolue la réflexion actuelle sur les nombres indices ? On peut, sans nul doute, la situer dans le contexte de l'analyse économique à travers la théorie de l'agrégation et la recherche sur les fonctions d'agrégation. Les travaux de H. Theil (*Economics and Information Theory*, Amsterdam, North Holland, 1967) et de W. Diewert (« Exact and superlative index numbers », *Journal of Econometrics*, 1976, p. 115-145) sont très représentatifs de cette approche, potentiellement intéressante.

CHAPITRE III

LA PROBABILITÉ

I. — Introduction

Vouloir décrire l'histoire de la probabilité dans un chapitre de quelques pages peut sembler impossible. Aussi voulons-nous préciser tout de suite notre objectif. Parler d'inférence statistique sans évoquer l'évolution du concept de probabilité est un non-sens. On ne peut comprendre l'un sans mesurer l'importance de l'autre. C'est pourquoi, tout en sachant que notre présentation sera forcément partielle et partiale, nous croyons essentiel d'introduire ce chapitre dans cet ouvrage, en nous concentrant sur les développements historiques qui allaient conduire à l'inférence statistique.

II. — La préhistoire

A l'origine, la probabilité est liée aux jeux de hasard. Parmi les plus anciens, on cite couramment le jeu de l'astragale ([1]), ancêtre de notre jeu de dés. On relève des témoignages de ce jeu dans l'Antiquité, notamment dans la civilisation égyptienne, sous la première dynastie (3500 ans avant J.-C. — cf. [David, 1955]). Le dé le plus ancien que nous connaissons provient de

([1]) : Ce petit os du talon présente une structure qui sera idéalisée plus tard par le dé, tout particulièrement chez le mouton.

Mésopotamie et remonte au début du troisième millénaire. Ces jeux revêtent bien sûr un aspect ludique. Mais leur usage à des fins religieuses est important. Ne devaient-ils pas, en effet, permettre de deviner les volontés des dieux ? La naissance du christianisme allait bouleverser leur évolution. Ils n'étaient pas compatibles avec l'idée de la toute-puissance divine et leur usage fut condamné par l'Eglise. On y fera à nouveau discrètement allusion vers la fin du Moyen Age (cf., par exemple, *le jeu de Saint-Nicolas*, de Bodel).

Plusieurs écrits peuvent être catalogués dans la période « préhistorique » et contiennent des idées qui sont à la base du développement de la probabilité. Il en est ainsi pour le *Liber de ludo aleæ* de Cardano (écrit vers 1526 mais publié à Lyon en 1663) et *Sulla scoperta dei dadi* de Galilée (qui ne fut publié qu'en 1656) dans lesquels on trouve des énumérations de façons d'obtenir des résultats recourant à des permutations. Il faut aussi citer les travaux de Fra Luca dal Borgo (ou Paccioli) qui, dans son ouvrage publié en 1494 et intitulé *Summa de arithmetica, geometria, proportioni et proportionalita*, annonce le *problème des partis* traité ultérieurement en partie par Tartaglia et résolu par Pascal et Fermat (voir ci-dessous). Par ailleurs, en 1606, Képler se livre à des spéculations sur l'apparition d'une nouvelle étoile. Mais ces travaux ne permirent pas encore de développer à la fin du XVIe siècle ce que certains ont appelé « les lois de la chance ». L'idée était cependant dans l'air.

III. — Blaise, Pierre, Christian, Jacques et les autres

C'est à Blaise Pascal (1623-1662) et Pierre de Fermat (1601-1665) que l'on attribue la naissance de la probabilité. Le premier n'est plus à présenter : il a marqué son temps d'une empreinte impressionnante.

Le second a excellé dans de nombreux domaines des mathématiques, en particulier en théorie des nombres et en analyse. Nous évoquerons brièvement leur correspondance ci-dessous, à propos du célèbre problème des *partis* (1654). Au cours de sa période mondaine, Pascal rencontre le chevalier de Méré (1607-1684), grand amateur de jeux de hasard. Ce dernier lui soumet le problème suivant : lorsqu'on lance plusieurs fois deux dés, « combien faut-il de coups au moins pour qu'on puisse parier avec avantage que, après avoir joué ces coups, on aura amené sonnez ? » (c'est-à-dire double six). L'usage que fit le chevalier de la réponse de Pascal semble justifier le jugement mitigé de ce dernier sur son interlocuteur : « Il a très bon esprit, mais il n'est pas géomètre ; c'est, comme vous le savez, un très grand défaut. »

Attribuer la naissance d'un concept est parfois une tâche malaisée. En ce qui concerne la probabilité, nous avons déjà mentionné l'existence de premières tentatives avant 1654. Par ailleurs, d'autres travaux voient le jour vers la même époque. Christiaan Huyg(h)ens (1629-1695) rédige entre 1650 et 1660 un ouvrage intitulé *de ratiociniis in aleæ ludo* qui pose le principe de l'espérance mathématique et introduit l'échantillonnage avec ou sans remise. Sa réflexion sur le concept de probabilité semble avoir eu plus d'influence dans le siècle qui suivra que celle de Pascal et de Fermat. Elle marque notamment Pierre de Montmort (1678-1719) dans son *Essai d'analyse sur les jeux de hasard* (1708) ainsi que Jacques Bernoulli et Abraham de Moivre dont nous parlerons dans la suite. Il faut aussi citer dans ce contexte les réflexions de Leibniz.

Nous avons souligné l'importance accordée à l'œuvre de Pascal et à sa correspondance avec Fermat. Retenons en particulier l'*adresse à l'Académie parisienne* qui date de septembre ou octobre 1654, c'est-à-dire avant la nuit du Mémorial (23 novembre). Pascal n'était pas encore entré dans la dernière phase de son

existence, beaucoup plus austère celle-là, à Port-Royal. L'Académie parisienne fut fondée en 1635 par Marin Mersenne. Elle fut très vite considérée comme le lieu de rencontre privilégié des savants français et étrangers. C'est là que Blaise Pascal expose la plupart de ses travaux scientifiques. L'adresse de 1654, écrite en latin, fait état de ces derniers (cf. œuvres complètes de Pascal). Plus particulièrement, il y signale l'existence d'un « ... *traité tout à fait nouveau, d'une matière absolument inexplorée jusqu'ici, savoir : la répartition du hasard dans les jeux [...]. Ainsi, joignant la rigueur des démonstrations de la science à l'incertitude du hasard, et conciliant ces choses en apparence contraires, elle peut [...] s'arroger à bon droit ce titre stupéfiant* : La Géométrie du Hasard ».

Le problème des partis est exposé dans *le traité du triangle arithmétique* rédigé la même année. Le troisième *usage de ce traité* est intitulé « *pour déterminer les partis qu'on doit faire entre deux joueurs qui jouent en plusieurs parties* ». Les problèmes abordés par Pascal dépassent largement une simple question de jeux. Outre les principes de la théorie des probabilités, on y trouve la source de réflexions juridiques importantes pouvant se développer dans les décisions de partage. Ce problème peut être résumé comme suit : si un jeu de hasard, qui se déroule en plusieurs parties, est interrompu avant la victoire de l'un des joueurs, comment peut-on faire un partage (parti) de l'argent misé ?

On trouve un exposé de la réflexion de Pascal dans les *lettres* adressées à Fermat. La lettre du 29 juillet 1654 contient la proposition de Pascal à la solution du problème des partis. Bien qu'elle ne soit pas la première traitant de ce problème (deux lettres antérieures, l'une de Pascal, l'autre de Fermat, abordaient cette question mais ont été perdues), on y constate la différence d'approche des deux savants. La démarche

de Fermat repose sur les combinaisons, celle de Pascal procède de proche en proche, par récurrence.

On ne peut terminer ce paragraphe sans citer un phénomène historique assez peu fréquent : il concerne l'importance qu'une famille a prise dans le développement des Sciences : les Bernoulli. Ses membres sont originaires de Bâle. La première génération qui fait parler d'elle, dans le domaine qui nous intéresse, est composée de quatre frères dont l'aîné se prénomme *Jacques* (1654-1705). Suivant la volonté de son père, il poursuit des études de théologie, mais complète ensuite sa formation en se tournant vers les mathématiques. Son frère *Jean* (1667-1748) suit un chemin analogue, après avoir étudié la médecine — toujours selon la volonté de leur père. En ce qui concerne la théorie de la probabilité, Jacques obtient des résultats plus importants que son frère. On lui doit un traité essentiel intitulé *Ars conjectandi*, dans la lignée du travail de Huyg(h)ens. C'est lui qui est à l'origine du premier théorème limite : *la loi des grands nombres* (cf. § V). Leur neveu *Nicolas* (1687-1759), ainsi que le fils de Jean, *Daniel* (1700-1782), sont les points forts de la deuxième génération. Guidé par ses oncles, Nicolas applique certains de leurs principes dans le domaine juridique. Il est surtout connu par sa correspondance prolifique, notamment avec de Moivre, Euler, Leibniz et Montmort. Daniel, quant à lui, se spécialise dans la botanique et la physique. Son traité essentiel concerne l'hydrodynamique, rédigé à Saint-Pétersbourg. En probabilité, il est surtout connu par le problème de l'espérance mathématique infinie qui le conduit à formuler le paradoxe qui porte le nom de cette ville. Il développe aussi sa recherche dans une théorie probabiliste des erreurs d'observation en astronomie. Un dernier membre de cette famille contribue, mais dans une moindre mesure, à la théorie des probabilités. Il s'agit de *Jean* (1744-1807), neveu de Daniel. Une telle famille méritait qu'on la citât !

IV. — La théorie des erreurs

Nous avons déjà fait allusion à cette théorie dans le chapitre II. Nous n'en avions alors considéré que l'aspect « descriptif ». Elle consiste à décrire *la loi de probabilité des erreurs* (différences entre la vraie valeur d'une quantité et les mesures fournissant des valeurs observées). Elle fut introduite dans le but de montrer l'utilité de prendre la moyenne arithmétique de valeurs observées pour « estimer » un paramètre. Nous avons déjà souligné le rôle joué par Simpson dans ce problème. Plus précisément, ce dernier introduit la *loi uniforme discrète* en 1756 dans la lettre adressée au président de la Royal Society que nous avons déjà mentionnée plus haut (cf. § III du chap. II) : si p_x est la probabilité associée à la valeur x, cette loi est définie par

$$p_x = \frac{1}{2a + 1}$$

$$x = -a, -a + 1, ..., 0, ..., a - 1, a$$

où a est un entier positif. Simpson envisage aussi une *loi triangulaire discrète des erreurs* (*discrete isosceles triangle distribution*) :

$$p_x = \frac{(a + 1) - |x|}{(a + 1)^2}$$

$$x = -a, -a + 1, ..., 0, ..., a - 1, a$$

C'est en recourant à la *fonction génératrice* utilisée par Abraham de Moivre (1667-1754) pour résoudre des problèmes liés à des jets de dés que Simpson obtient, pour ces deux distributions, *la loi de probabilité de la somme de n erreurs indépendantes* (et par conséquent, de la moyenne arithmétique de ces derniè-

res). Ensuite, par passage à la limite, il introduit la distribution de la moyenne arithmétique d'erreurs admettant une *loi triangulaire continue (continuous isosceles triangle distribution)* définie par la fonction de densité

$$f(x) = \frac{a - |x|}{a^2} \qquad -a \leq x \leq a.$$

Vingt ans plus tard, en 1776, Joseph-Louis, comte de Lagrange (1736-1813), rédige un *Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résout différents problèmes relatifs à cette matière* (cf. *Œuvres de Lagrange*, t. II, p. 173-234, Paris, Gauthier-Villars, 1868). On y retouve les résultats de Simpson via le même outil (fonction génératrice) introduit par de Moivre (sans référence à ces auteurs). Mais il convient d'ajouter d'autres lois *continues* étudiées par Lagrange :

a) *distribution uniforme :*

$$f(x) = \frac{1}{b - a} \qquad a \leq x \leq b$$

(où a et b sont deux réels tels que $a < b$) ;

b) *distribution parabolique :*

$$f(x) = \frac{3(a^2 - x^2)}{4a^3} \qquad -a \leq x \leq a$$

(cette loi allait être généralisée à un intervalle $[a, b]$ par Laplace en 1781) ;

c) *distribution cosinusoïdale :*

$$f(x) = \frac{1}{2} \cos x \qquad -\pi / 2 \leq x \leq \pi / 2.$$

Peu avant ces travaux de Lagrange, J.-H. Lambert (1728-1777) avait consacré une partie de son traité *Photometria* (publié en 1760) à l'étude de courbes de densité continues, symétriques et unimodales. A cette

occasion, il introduit ce qui allait s'appeler ultérieurement la *méthode du maximum de vraisemblance* en recherchant la position du centre d'une distribution continue symétrique. Cinq ans plus tard, dans un ouvrage intitulé *Beyträge zum Gebrauche der Mathematik und deren Anwendung*, il propose une loi *semi-circulaire*, dont la fonction de densité est

$$f(x) = (2/\pi\, a^2)\, \sqrt{a^2 - x^2} \qquad -a \leq x \leq a\,[2].$$

Dans la préface de ce dernier ouvrage, Lambert introduit l'expression « die Theorie der Fehler » (théorie des erreurs) qui allait perdurer dans la suite.

Toutes les lois mentionnées ci-dessus concernaient des variables prenant leurs valeurs dans un intervalle limité. Le premier à proposer une loi où le domaine de définition de la variable est la droite réelle, fut Pierre Simon, Marquis de Laplace (1749-1827) qui, en 1774, considéra la fonction de densité

$$f(x) = (k/2)\, e^{-k|x|} \qquad -\infty < x < \infty$$

où k est un paramètre réel positif. Cette loi, appelée *double exponentielle*, est encore dénommée *première loi des erreurs de Laplace*. Elle prenait en compte les écarts des observations à la médiane, pris en valeurs absolues. Laplace considéra plus tard le carré des écarts à la moyenne. Nous y reviendrons à propos de la loi normale.

Remarquons que Laplace introduisit aussi, dans un mémoire publié en 1777, la loi définie par

$$f(x) = \frac{1}{2a} \log \frac{a}{|x|} \qquad -a \leq x \leq a.$$

Puisque nous parlons de Laplace, il nous semble intéressant d'en citer quelques traits. Disciple de d'Alembert, il devient professeur à l'Ecole royale militaire. Son œuvre est vaste : il suffit pour s'en

[2] La loi semi-circulaire se retrouve aussi dans un mémoire publié en 1778 par Daniel Bernoulli, sans référence à l'œuvre de Lambert.

rendre compte de noter que ses *Œuvres complètes* ont été publiés entre 1878 et 1912 en quatorze volumes. Inspecteur militaire en 1783, il fait la connaissance de Bonaparte en assumant ses fonctions. Le Premier Consul le nomme d'ailleurs ministre de l'Intérieur en 1799... mais pour six semaines seulement avant que ce dernier n'exerce des fonctions importantes au Sénat en 1803. Devenu comte en 1806, Laplace se retourne contre Napoléon en 1814... et devient marquis sous Louis XVIII. Membre de l'Académie des Sciences en 1785, il participe à la constitution de l'Institut en 1795 et est élu à l'Académie française en 1816.

Outre ses travaux de mathématiques et son œuvre d'astronome traduite dans sa *Mécanique céleste*, Laplace est l'auteur d'une *Théorie analytique des Probabilités* et d'un *Essai philosophique sur les probabilités* qui auront beaucoup de succès. Une de ses premières découvertes en probabilité date de 1774 : Laplace obtient le « théorème de Bayes ». Il semble qu'il n'ait pas eu connaissance du résultat contenu dans un article posthume du presbytérien anglais Thomas Bayes (1702-1761), paru près de dix ans plus tôt. Les travaux de Laplace ne se situèrent pas tous d'un point de vue bayésien, mais il en fit un usage diversifié, en particulier dans l'étude de problèmes combinatoires et en démographie. Son influence fut grande, notamment sur A. Quételet et S.-D. Poisson, qui s'inspirèrent fortement de son œuvre.

Mais revenons à la théorie des erreurs et en particulier à celle qui allait marquer la suite de l'histoire de la statistique : la *loi normale* ([3]). Nous lui devons bien un paragraphe.

([3]) Notons que certains ont vu dans Laplace l'auteur de cette loi. Il est vrai que Laplace fut le premier à évaluer explicitement, en 1774, l'intégrale d'Euler :

$$\int_0^\infty \exp[-u^2/2]\, du = \sqrt{\pi/2},$$

mais ce ne fut pas, à cette époque, dans l'intention de construire une loi des erreurs.

V. — La loi normale

La distribution normale est souvent attribuée à Laplace et Gauss, dont elle porte le nom. Son origine est cependant plus ancienne et repose sur les travaux de Jacques Bernoulli (1654-1705) dont nous avons parlé plus haut. Dans la quatrième partie de son œuvre, *Ars Conjectandi*, Bernoulli s'interroge sur le calcul de la probabilité d'avoir face avec une pièce mal équilibrée. Il trouve une réponse à sa question en montrant que si l'on répète un grand nombre de fois l'expérience consistant à lancer cette pièce, la fréquence d'apparition de face se rapproche d'une quantité p susceptible d'être qualifiée de probabilité d'obtenir ce résultat. Il s'agit de la forme élémentaire de la *loi des grands nombres* .

C'est en travaillant sur le problème posé par Bernoulli que Abraham de Moivre (1667-1754) découvre une formule approchée des « probabilités binomiales ». Dans un supplément aux deuxième et troisième éditions de son œuvre principale *Doctrine of chances* et à certaines copies de *Miscellanea Analytica*, intitulé *Approximatio ad summam terminorum binomii* $\overline{a + b/n}$ *in seriem expansi* et daté du 12 novembre 1733, de Moivre montre tout d'abord que le terme modal d'une distribution binomiale symétrique

$$a_0 = \binom{n}{k} (1/2)^n$$

est approché, pour n grand, par $2/\sqrt{2\,\pi\,n}$. Le terme a_t, distant du mode d'un écart t, est, quant à lui, approximativement égal à $a_0 \exp(-2\,t^2/n)$. Il déduit aussi des résultats semblables pour le cas non-symétrique : de Moivre peut donc être considéré comme l'inventeur du premier *théorème limite central* (cf. § VII).

Mais voyons le rôle joué par Gauss et Laplace dans l'usage de cette loi. Carl Friedrich Gauss (1777-1855)

fut un mathématicien et un astronome brillant. Durant ses études et sa fonction de directeur de l'observatoire de Göttingen, il produit un nombre impressionnant de résultats dans des domaines très variés : philologie et théorie des nombres, astronomie, géodésie et géométrie différentielle, géomagnétisme, électromagnétisme, topologie, fonctions de variables complexes,...

La loi normale est introduite par Gauss en recourant à la méthode des moindres carrés. Gauss reprend à cette occasion le problème « d'ajustement » qui avait déjà été abordé par Boscovich, Euler, Laplace, Legendre et Mayer, suscité ici par l'étude des orbites planétaires. Cette analyse ne sera publiée qu'en 1809 dans *Theoria motus corporum cœlestium in sectionibus conicis solem ambientium* ; Gauss prétend cependant s'être intéressé à ce problème dès 1794. L'apport innovateur de Gauss se situe dans le caractère probabiliste de son approche. C'est pourquoi nous décrivons ci-dessous la façon dont il procéda, en utilisant une notation simplifiée et plus actuelle. Considérant que l'erreur $\Delta = x - \mu$ (faite en mesurant par la valeur observée x un paramètre inconnu μ) admet une fonction de densité $f(\Delta)$, la loi combinée pour n mesures $x_1, x_2, \ldots, x_n$ indépendantes est donnée par

$$\Omega = \prod_i f(x_i - \mu).$$

En supposant que toutes les valeurs possibles de μ sont également probables *a priori*, Gauss montre que la distribution a posteriori de μ, étant donné $x_1, x_2, \ldots, x_n$, est proportionnelle à Ω. Il en conclut que la valeur la plus probable pour μ correspond à celle qui maximise ce produit, c'est-à-dire telle que

$$\sum_{i=1}^{n} \frac{d \log f(x_i - \mu)}{d\mu} = 0.$$

Partant alors du principe

« ... de regarder comme axiome l'hypothèse selon laquelle, si une quantité a été déterminée par plusieurs observations directes effectuées dans les mêmes conditions et avec le même soin, la moyenne arithmétique des valeurs observées constitue la valeur la plus probable, si pas rigoureusement, du moins encore très proche, de sorte que il est toujours le plus sûr d'y adhérer »,

et tenant compte de ce que la moyenne arithmétique des valeurs observées est solution de l'équation $\Sigma\ (x_i - \mu) = 0$, il en résulte que

$$\frac{d \log f(x_i - \mu)}{d\mu} = k\ (x_i - \mu) \qquad i = 1, 2, ..., n$$

$f(x - \mu)$ est donc proportionnelle à
$\exp [-(1/2)k(x - \mu)^2]$, où k est une quantité positive. Posant $k = 2h^2$ (où $h > 0$) et se rappelant que $x - \mu$ est l'erreur Δ, Gauss obtient la loi

$$f(\Delta) = \frac{h}{\sqrt{\pi}}\, e^{-h^2\Delta^2} \qquad -\infty < \Delta < \infty$$

dans laquelle il signale que « h peut être considéré comme la mesure de précision des observations ».

A la même époque, en 1810, Laplace introduit sa « seconde loi des erreurs » en montrant que la distribution des moyennes arithmétiques de n erreurs indépendantes, ayant la même précision ou des précisions distinctes, peut être approchée par la distribution normale, avec une erreur d'approximation qui tend vers zéro quand n croît indéfiniment. C'est donc par un *théorème limite central* que Laplace obtient cette distribution. Notons encore que cette loi fut aussi obtenue par Adrain, en 1809, désireux, comme Gauss, de justifier la méthode des moindres carrés.

La littérature contient de très nombreuses « démonstrations » de la loi normale. Les unes utilisent des hypothèses spécifiques sur la façon de combiner les erreurs [Herschell (1850), Donkin (1857), Kelvin et Tait (1867), ...]. Les autres considèrent que l'erreur

est une somme algébrique d'un grand nombre « d'erreurs élémentaires » et recourent alors à l'emploi d'un théorème limite central [Young (1819), Hagen (1837), Bessel (1838), Tait (1865), ...]

Les premières tables donnant des valeurs de probabilités « normales » datent de 1899 et sont dues au physicien français Kramp. Remarquons enfin que la distribution fut qualifié de *normale* par Pearson, en 1893, et est souvent connue sous cette dénomination.

VI. — D'autres lois de probabilité

La loi normale connaît une vogue certaine au XIXe siècle. Néanmoins, plusieurs critiques voient le jour quant aux hypothèses sous-jacentes et à l'universalité de cette loi. C'est ainsi que Siméon-Denis Poisson considère, en 1824, la loi définie par la fonction de densité

$$f(x) = \frac{1}{\pi(1 + x^2)} \qquad -\infty < x < \infty$$

qui sera appelée dans la suite « loi de Cauchy ». Il constate en particulier que la moyenne arithmétique de n erreurs indépendantes distribuées selon cette loi ne tend pas vers une loi normale.

Faisons à nouveau une parenthèse pour situer Siméon-Denis Poisson. Ce dernier naît le 21 juin 1781 à Pithiviers et décède le 25 avril 1840 à Paris. D'origine modeste, il est encouragé à faire des études. C'est ainsi qu'il est admis en 1798 major à l'Ecole polytechnique de Paris. Soutenu par Laplace, il entre dans le corps académique de cette grande école et, en 1806, remplace Fourier comme professeur. Il obtient en outre d'autres fonctions parmi lesquelles une chaire de mécanique à la Sorbonne en 1816. Cette année est aussi celle de son élection à l'Académie des Sciences de Paris.

Ses publications scientifiques sont nombreuses et

diverses. Plus de trois cents articles lui sont dus tant en mathématiques, qu'en physique et en astronomie. Outre son *Traité de mécanique*, publié en 1811, il contribue, trois ans avant sa mort, à la théorie des probabilités par son ouvrage intitulé *Recherches sur la probabilité des jugements en matière criminelle et en matière civile* (1837). Dans cet ouvrage, la distribution de Poisson est introduite comme distribution limite de la loi de Pascal ([4]) et de la loi binomiale ([5]), *via* la relation existant entre ces deux dernières. Il faut cependant souligner le fait que la probabilité fournie par l'expression

$$p_x = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \ldots; \quad \lambda > 0$$

apparaît en 1830 dans l'œuvre de Poisson. Mais, en réalité, l'approximation de la distribution binomiale est contenue dans la *doctrine des chances* publiée en 1718 par de Moivre. L'Histoire n'a pas voulu lui reconnaître cet enfant ! Ne jetons pas la pierre à Poisson : il n'est pas responsable de cette situation. D'ailleurs, n'a-t-il pas étudié lui-même la *distribution de Cauchy* plus de vingt ans avant Cauchy ? Et ne lui doit-on pas aussi une version de la *loi des grands nombres* qui ne porte pas son nom mais celui de Chebychev (qui, il faut le reconnaître, démontra cette loi « plus rigoureusement ») ?

Parmi les travaux reposant sur des critiques à l'égard de la loi normale, il faut souligner les propositions introduisant des « corrections » étudiées tout particulièrement par l'école scandinave (Gram, Charlier, Thiele...).

D'autres lois méritent d'être citées. Nous parlerons des distributions du χ^2, de Student, de Fisher-Snédecor dans le chapitre IX. Il faut aussi citer le

([4]) $p_x = p(1-p)^x \quad x = 0, 1, 2, \ldots; \quad 0 < p < 1.$
([5]) $p_x = C_n^x p^x (1-p)^{n-x} \quad x = 0, 1, 2, \ldots, n; \quad 0 < p < 1.$

système de Pearson des courbes de densité que nous évoquerons aussi dans ce chapitre.

VII. — Approximations et lois limites

Il nous est difficile de présenter ici la riche histoire des théorèmes asymptotiques. Nous en avons déjà évoqué certains. Il en est ainsi de la *loi des grands nombres* introduite par J. Bernoulli en 1713 (cf. § V). Le *théorème limite central* , évoqué dans le même paragraphe, concerne de façon générale la convergence vers la loi normale d'une somme d'un nombre croissant de variables aléatoires. Ce problème a été traité dans le cas de variables binomiales de paramètre p par de Moivre en 1718 pour $p = 1/2$ et par Laplace en 1812 pour des valeurs quelconques de p. Par ailleurs, en étudiant le comportement asymptotique de proportions au sein de schémas où la probabilité de succès d'un évènement est variable, Poisson suscite, en 1837, la réflexion sur le comportement d'une somme de variables indépendantes, mais non identiquement distribuées. Signalons encore la qualité des démonstrations du théorème limite central effectuées par Augustin Cauchy (1789-1857) et le résultat obtenu par Jules Bienaymé (1796-1878), que l'on appelle aujourd'hui l'inégalité de Bienaymé-Chebychev. Notons enfin que l'étude rigoureuse d'une somme de variables aléatoires arbitraires est principalement due à Chebychev (1821-1894) qui est le fondateur de l'école de Saint-Pétersbourg au sein de laquelle s'illustrèrent notamment Markov (1856-1922) et Liapunov (1857-1918). De nombreuses versions du théorème limite central peuvent être trouvées dans de bons ouvrages de probabilité, incluant les travaux de Bernstein, Lévy, Lindeberg, Liapunov, Khinchin, Feller... On y trouvera aussi référence à l'étude des différents types de *convergence* .

Remarquons, pour terminer, que le nom de *théo-*

rème limite central a été proposé par G. Polya qui, dans un article paru en 1920, parle de *central limit theorem of probability theory.*

VIII. — Théorie moderne des probabilités

Les fondements de la probabilité mathématique sont dus à Kolmogorov (1933). Ce dernier va développer sa théorie axiomatique dans son ouvrage intitulé *Grundbegriffe der Wahrscheinlichkeitsrechnung*, en s'inspirant du développement de la *théorie de la mesure* .

En 1881, A. Harnack introduit la notion d'ensemble discret, puis, en 1882, celle d' « égalité en général », qui « mesure » le concept d'égalité presque sûre. La « mesure » d'un ensemble apparaît ensuite dans de nombreux travaux ; les définitions sont différentes selon les auteurs (Cantor, 1884 ; Peano, 1887 ; Jordan, 1892). Mais c'est Emile Borel qui va introduire la notion d'ensemble de mesure nulle (1894) et la théorie des ensembles mesurables (1897). Travaillant sur **R**, et, plus précisément sur [0,1], il privilégie la classe des ensembles ouverts, qu'il « mesure » à partir de la longueur des intervalles, et il montre que l'on peut définir sur cette classe (qui portera plus tard le nom de classe « borélienne ») une « mesure » μ vérifiant les propriétés de σ — additivité et de différence.

La théorie de la mesure de Borel va alors ouvrir la voie à l'intégrale de Lebesgue, ou intégrale généralisée, qui va permettre à l'intégrale de Riemann et aux probabilités de dépasser certains contre-exemples gênants. En effet, jusqu'à la fin du XIXe siècle, la seule façon de définir l'intégrale d'une fonction f sur un segment $[a,b]$ consistait à utiliser les sommes de Riemann-Darboux. Néanmoins, à partir de 1870, l'intégrale de Riemann s'était vue proposer des particularités où elle était inemployable. En 1901, H. Le-

besgue donne une théorie de l'intégration plus générale que Riemann, à partir du concept de mesure introduit par Borel. Sous l'impulsion de J. Radon, les concepts de mesure et d'intégration quitteront vite $\mathbf{R}$ et $\mathbf{R}^n$; ainsi apparaîtra la théorie de la mesure abstraite définie sur un ensemble Ω quelconque muni d'une tribu (1913). C'est cette notion de mesure abstraite qui a permis le développement de la théorie mathématique des probabilités. Par exemple, pour définir dans tous les cas l'espérance E (X) d'une variable aléatoire X, il fallait l'intégrale de Lebesgue et ses extensions. L'achèvement de la construction des théories de la mesure et de l'intégration peut être daté de 1930, avec les théorèmes de décomposition de Lebesgue-Nikodym et l'existence des densités.

Ainsi, la théorie de la mesure et de l'intégration ont jeté les bases d'une formalisation homogène et cohérente des probabilités, en y introduisant la rigueur du raisonnement mathématique ; rappelons en effet que Keynes, en 1921, écrivait à propos des probabilités que « les savants y décèlent un relent d'astrologie ou d'alchimie », et que von Mises est encore plus direct en affirmant (1919) que « le calcul des probabilités n'est pas une discipline mathématique ».

La probabilité étant un cas particulier de mesure se voit donc, à partir des années trente, appliquer une grande partie des résultats de mesure et d'intégration. Plusieurs auteurs lui donnèrent ses lettres de noblesse. Citons, en particulier, la *théorie de l'addition des variables aléatoires* de Lévy (1937), *Random variables and probability distribution* de Cramer (1937), *Distributions limites de sommes de variables aléatoires indépendantes* (en russe — 1949) de Gnedenko et Kolmogorov, *An introduction to probability theory and its applications* de Feller (1950) et *Probability Theory* de Loève (1955).

CHAPITRE IV

LES SONDAGES

I. — Les origines

L'apparition du mot « statistique » lui-même est relativement récent, surtout en comparaison de l'activité correspondante de recueil des données, qui remonte à la plus haute antiquité. On associe souvent la création du terme « statistique », issu du latin « statisticum » : qui a trait à l'Etat, à l'école allemande de Gottingue, et plus particulièrement à Gottfried Achenwall (1746). Il est plus probable cependant que ce dernier a seulement été le premier à proposer un enseignement traitant de statistique. L'emploi du mot est plus ancien puisque l'on possède une *Biblioteca Statistica* datant de 1701 et un *Microscopium Statisticum* de 1672. En remontant plus encore dans le temps, le mot « statistique » appartient au langage administratif français colbertien : Jean Meyer mentionne l'existence d'une *Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne* établie par Claude Bouchu, intendant de Bourgogne, de 1666 à 1669 [*Colbert*, Paris, Hachette, 1981].

Par contre, l'apparition du besoin « statistique » de posséder des données chiffrées précède sa dénomination de plusieurs millénaires. A son origine, il est le fait de chefs d'Etat — ou de ce qui en tient lieu à l'époque — désireux de connaître des éléments de leur puis-

sance : population, potentiel militaire, richesses... L'idée de recensement, ou de liste d'inventaire, apparaît donc de façon tout à fait naturelle dans l'histoire, impliquant en outre une impression de précision de la plus haute qualité.

Les premiers recensements semblent remonter à la civilisation sumérienne, de 5 000 à 2 000 ans avant notre ère, pour laquelle on possède des listes d'hommes et de biens inscrits sur des tablettes d'argile. Le relevé des personnes et des biens a lieu régulièrement en Mésopotamie 3 000 ans avant Jésus-Christ. L'Egypte paraît avoir été la première nation à organiser des recensements systématiques de population, au moins depuis l'an 2900 avant J.-C., mais aussi à institutionnaliser des recensements à finalité fiscale (2700 à 2500 ans avant J.-C.). Elle semble même avoir la primeur du principe de la déclaration obligatoire : en effet, sous le pharaon Amasis II, au VIe siècle avant notre ère, tout individu était tenu de déclarer ses sources de revenu et son activité. Tout manquement à cette règle était puni de mort.

Il serait vain et fort long de dresser la liste des premières tentatives de dénombrement exhaustif. L'enseignement essentiel en est que les civilisations qui les engendrèrent — Japon, Rome, Inca, Inde — possédaient un système administratif fort, autoentretenu par l'amélioration de la connaissance quantitative de l'Etat.

II. — Les premières tentatives d'extrapolation

A partir du XIIIe siècle, les données deviennent plus nombreuses grâce à la prolifération des rôles fiscaux. Le plus célèbre est, en France, « l'état des paroisses et des feux des baillages et sénéchaussées de France » dressé en 1328. Le « feu », référence à un foyer, une famille ou une habitation, sera un élément essentiel

pour estimer une population bien que devenant vite une donnée sociale plutôt que démographique [cf. l'ouvrage de J. et M. Dupaquier, *Histoire de la démographie*, Paris, Librairie Académique Perrin, 1984].

Le XIVe siècle voit le début des enregistrements des actes d'état civil (commune de Givry, 1334). En août 1539, François Ier rend obligatoire la tenue de registres de naissances, par l'article 51 de l'édit de Villers-Cotterêts ; sous Henri III, l'édit de Blois (mai 1579) étend cette obligation aux mariages et aux décès.

Cependant, durant toute cette période où s'affirme la statistique administrative, les recensements sont très rares (recensement de Paris en 1590 — 200 000 habitants —, du Berry en 1565, du Bourbonnais en 1569).

Les progrès fondamentaux de la statistique vont apparaître lors de la seconde moitié du XVIIe siècle, avec le besoin que ressentent les monarques et leurs conseillers de connaître et expliquer les phénomènes économiques et sociaux. Cette époque coïncide avec le développement de l'*école d'arithmétique politique anglaise*, précurseur de la statistique inférentielle. Cette école, dont les fondateurs sont John Graunt (1620-1674) et William Petty (1623-1687), est guidée par le souci de la quantification et la recherche de constantes de comportement permettant des estimations et des prévisions : nombre d'enfants par femme, temps entre deux naissances pour une même mère, nombre d'habitants par maison, nombre d'individus par feu, proportion de décès, etc. Les techniques de *multiplicateur* de l'arithmétique politique vont être utilisées au détriment des recensements, et favoriser l'appartion d'enquêtes partielles. Parallèlement, le remplacement d'une connaissance exhaustive par une extrapolation fondée sur l'examen d'une partie de la population est une attitude qui commence à trouver des éléments de justification par l'apparition du calcul des probabilités (cf. chapitre III).

En France, deux noms dominent cette époque : Colbert et Vauban. En 1664, Jean-Baptiste Colbert ordonne aux intendants de procéder à une vaste enquête sur l'état des provinces ; celle-ci est précédée d'une enquête-pilote commandée à son frère dans les provinces de l'Est. Le marquis de Vauban s'intéresse encore de plus près à la connaissance du chiffre, que ce soit par des recensements ou des enquêtes. Auteur de *Méthode généralle et facille pour faire le dénombrement des peuples*, publiée en 1686, il préconise l'utilisation d'échantillons de terres arables dans chaque province pour estimer au mieux les capacités agricoles ; ces procédures sont reprises dans les célèbres *mémoires dressées par les Intendants du Royaume*. Certains esprits commencent même à demander un organisme centralisateur de la collecte chiffrée. L'un des premiers, sinon le premier, est Charles Castel, dit l'abbé de Saint-Pierre (1658-1743), qui réclame, dès le début du XVIIIe siècle, la création d'un *Bureau pour recueillir les divers dénombrements*.

En Angleterre, les méthodes de *l'arithmétique politique* poursuivent leurs progrès. A titre d'exemple, détaillons l'estimation de population de W. Petty, décrite dans l'article de Jacqueline Hecht dans *Pour une histoire de la Satistique*, INSEE, Economica, 1987 :

« Il évalue tout d'abord le nombre des maisons à Londres à 88 000 pour 1686. Etablissant une moyenne entre la proportion des feux par maison à Dublin et à Bristol, il évalue le nombre de feux londonniens à 105 000, à peu près équivalent au chiffre officiel du Bureau des Feux (105 315). En supposant que chaque famille compte en moyenne six membres, et que 10 % des maisons abritent deux familles chacune, les autres n'en ayant qu'une, il conclut que les 105 000 maisons londoniennes abritent 695 000 habitants. »

Ces techniques d'extrapolation ou du *multiplicateur* connaissent vite un énorme succès, contribuant à mettre un frein aux recensements jugés par certains onéreux, peu précis et « monstrueux » (Saint-Simon).

Elles servent de base aux calculs faits par d'Expilly, Messance (*Recherches sur la population des généralités d'Auvergne, de Lyon et de Rouen* , publié en 1766) et le baron Moheau (*Recherches et considérations sur la population de la France*, 1778). L'abbé Jean-Joseph d'Expilly (1719-1793) utilise une extrapolation fondée sur les feux. Possédant leur nombre total, il le multiplie par une autre estimation du nombre d'habitants par feu. Il introduit en outre une « stratification campagnes-villes », et adopte un multiplicateur 4,5 pour les premières et 5 pour les secondes. Messance et Moheau préconisent des enquêtes partielles pour évaluer la population : sélectionnant un grand nombre de paroisses ou de petites villes, ils dénombrent les habitants, établissant sur la base des cinq ou dix dernières années de l'Etat Civil une année moyenne des naissances, et calculant le nombre de naissances par habitant, taux qui sera appelée multiplicateur des naissances. Sous l'hypothèse que celui-ci est proche de celui qui serait observé sur la France entière, il suffit de le multiplier par le total des naissances d'une année pour avoir une idée de l'ensemble de la population.

Prenons quatre exemples de ces techniques d'extrapolation. En 1783, Pierre-Simon de Laplace expose à l'Académie des Sciences une procédure de détermination du multiplicateur consistant à réaliser une enquête portant sur un million d'habitants répartis sur plusieurs régions. Il propose de calculer l'incertitude de l'estimateur — *l'erreur à craindre* — en supposant que les régions sont tirées au hasard. En l'an XI (1802), Jean-Antoine Chaptal, alors ministre de l'Intérieur, fit « sonder » 50 000 habitants dans 30 départements et obtient une valeur de 28,35. Antoine Lavoisier estime par une méthode d'échantillonnage la superficie cultivée et le cheptel en chevaux, bœufs, moutons et porcs, population sur la base du nombre de naissances et de celui des habitants par maison. Ses

l'institut international de Statistique, 9, livraison 2, p. 176-183). Il présente un mémoire écrit intitulé *Observations et expériences concernant des dénombrements représentatifs* dans lequel il dresse un bilan satisfaisant de deux expériences norvégiennes d'enquêtes partielles, dont l'une sur la distribution des revenus :

« ... Après avoir fait un choix de 127 communes rurales et de 23 villes représentatives, nous avons d'abord pris les bulletins individuels recueillis par le recensement général des hommes âgés de 17, 22, 27, 32, 37, etc., jusqu'à 97 ans, et demeurant dans les communes en question. De cette façon nous aurions eu 1/5 de la population mâle adulte des communes représentatives ; mais comme nous désirions réduire encore davantage les chiffres, nous avons employé seulement les bulletins des individus commençant par les lettres A, B, C, L, M, ou N. Aux principaux renseignements donnés par le recensement (âge, état civil, profession, etc.) nous avons ajouté à l'aide d'une investigation particulière l'indication des fortunes et des revenus des individus respectifs. De cette manière nous avons obtenu un dénombrement représentatif spécial comprenant en moyenne 3,3 pour cent de la population mâle des villes et 1,6 pour cent de celle des campagnes... ».

Kiær ne définit pas *a priori* la notion de représentativité, mais la contrôle a posteriori, par comparaison de la structure par activité de l'échantillon obtenu avec celle issue du recensement. Il pose la question :

« ... De quelle manière un dénombrement représentatif doit-il être exécuté pour qu'il puisse présenter une miniature aussi correcte que possible de la société entière ?... »

Bien que Kiær ait pris toutes les précautions oratoires nécessaires (« ... il va sans dire que les dénombrements partiels représentatifs n'ont pas pour but de remplacer les recensements généraux... »), les réactions en Assemblée générale de l'IIS du 30 août sont extrêmement violentes.

L'un des rares à ne pas désavouer A. Kiær est le Français Emile Levasseur :

« Entre les statistiques générales [...] et les monographies [...], n'y a-t-il pas la place pour un troisième procédé, l'exploration statistique, [...] à l'aide des moyens de la statistique appliqués non à la totalité [...] mais à un nombre déterminé et restreint ? »

Malgré toutes ces critiques, l'idée de « une partie pour le tout » est lancée à un public de statisticiens.

En 1897, à Saint-Pétersbourg, Kier confirme que la représentativité est décidée après contrôle avec les résultats d'un recensement, et conteste les monographies « qui ne donnent pas une vraie miniature de la totalité ». Il reçoit, cette même année, l'appui de la Conférence des Statisticiens scandinaves, qui adopte une motion proclamant l'intérêt des méthodes de sondages représentatifs.

Lors du congrès de Budapest (1901), Kiær précise :

« ... la méthode représentative demande un grand nombre d'unités d'observation réparties de telle manière que [celles] de différents caractères soient représentées autant que possible dans les mêmes rapports qu'elles existent dans l'ensemble... »

Outre cette anticipation de la méthode des quotas, il entrevoit les procédures de rééchantillonnage.

Dans la discussion qui suit, L. von Bortkiewicz, professeur à l'Université de Berlin, suggère de recourir au calcul des probabilités pour tester statistiquement l'écart existant entre les répartitions de l'échantillon et de la population totale sur les variables-clés. L'intervention de von Bortkiewicz marque la première introduction sérieuse des probabilités dans les études de l'IIS consacrées à l'échantillonnage représentatif, et la première recommandation de l'utilisation d'un test de significativité.

Kiær reçoit en outre l'appui de Carroll Wright, du ministère du Travail des États-Unis, qui dans une lettre adressée au rapporteur du Comité de l'IIS chargé d'étudier la méthode représentative, confirme la bonne opinion de cette technique, fréquemment utilisée avec succès, dans la statistique officielle américaine.

Après 1903, l'ordre du jour de l'IIS ne mentionne plus la question du sondage représentatif jusqu'en 1925, principalement en raison de l'absence de la personnalité de A. Kiær, qui décédera dans l'intervalle. Durant ces 22 ans, l'introduction du raisonne-

ment probabiliste dans les méthodes partielles se poursuit.

Trois noms sont à mentionner. En 1911, von Bortkiewicz prévoit la césure entre statistique privée et statistique officielle, la première tendant à être partielle et sélective, la seconde étant générale (au sens du recensement) et complète ; la statistique privée courra alors le risque d'être non représentative. Il étend également la notion de représentativité aux tests sur les lois jointes et non seulement sur les marges.

Durant la même période, des contributions essentielles sont apportées par l'Anglais Arthur Bowley (1869-1957), tout d'abord par sa communication de 1906 à la Royal Statistical Society, puis par un livre publié en 1920, et par sa participation à la commission chargée par l'IIS, en mai 1924, d'étudier l'application de la méthode représentative. Bowley développe l'échantillonnage aléatoire, la stratification, établit l'équation d'analyse de la variance en univers stratifié, les formules de la variance d'un total dans les cas simple et stratifié. Ses travaux sont d'autant plus innovateurs que les tenants de la méthode représentative utilisent plutôt, pour sélectionner les unités observées, un « choix judicieux » — le *purposive selection*. Bowley a une claire vision des défauts de cette méthode, tout simplement parce que dans le sondage par choix judicieux, l'origine de l'aléatoire nécessaire pour appliquer la théorie des probabilités n'est pas spécifiée.

Troisième grand nom de la période inter-congrès, A. I. Tchuprow qui définit l'allocation optimale dans un échantillonnage stratifié, ce qui est en contradiction avec l'idée de représentativité courante où l'échantillon est un « modèle réduit » de la population entière.

Le point final (?) de la polémique sur la méthode représentative peut être situé lors du congrès de l'IIS de Rome en 1925, lors de la présentation faite par le rapporteur Adolphe Jensen, chef du Département de la

Statistique du Danemark, des conclusions de la Commission *ad hoc* nommée en 1924 et déjà mentionnée (outre A. Jensen, cette commission est composée de Arthur Bowley, Corrado Gini, Lucien March, Verrijn Stuart et Frantz Zizek). Jensen développe deux communications intitulées *Rapport sur la méthode représentative en statistique* et *La méthode représentative en pratique* et Bowley y expose *Measurement of the precision attained in sampling* (*Bulletin de l'Institut international de Statistique*, 22, livraison 1, p. 359-439). La polémique initiale ayant disparu, la résolution adaptée par l'IIS distingue l'échantillonnage aléatoire (choix au hasard, *random sampling*) et l'échantillonnage raisonné (choix judicieux, *purposive selection*).

Après 1925, le point de discussion n'est plus sur « faut-il échantillonner ou non » mais sur « comment tirer l'échantillon ? ». Le débat va vite concerner les tenants du *random sampling* (la « prise au hasard » de l'IIS) et ceux de la *purposive selection* (le choix judicieux).

V. — Un cheminement parallèle : le début des sondages d'opinion aux Etats-Unis

Les sondages d'opinion sont nés du désir d'obtenir des informations, si possible quantitatives, sur l'état d'esprit de la population. A cet effet, Daniel Defœ en Angleterre et le comte de Lafayette en France mettent en place, dès le début du XIXe siècle, des réseaux de correspondants répartis géographiquement.

Le vrai berceau des sondages d'opinion se trouve cependant aux Etats-Unis, à l'occasion des couvertures de presse des élections présidentielles. Dès 1824, le *Harrisbourg Pennsylvanian* et le *Raleigh Star* font des enquêtes préélectorales par consultation individuelle d'électeurs (les « votes de paille »). Par la suite,

d'autres journaux vont reprendre cette habitude très appréciée du public : *New York Herald, Chicago American, Columbus Dispatch,* et *The Literary Digest.*

Ces échantillons ne reposent sur aucun critère de représentativité (le mot n'apparaîtra d'ailleurs que 70 ans plus tard). Leurs propriétés sont dues essentiellement à des tailles fort élevées : 30 000 pour le *New York Herald* en 1905, plus de 2 000 000 pour le *Literary Digest* en 1936.

Une date cruciale pour l'histoire de l'échantillonnage est le mardi 3 novembre 1936, jour de la publication des résultats de l'élection présidentielle. Alors que le *Literary Digest* a prédit la victoire de Landon, en utilisant la technique des votes de paille qui lui a fort bien réussi depuis 1916, c'est F. D. Roosevelt qui est élu. Celui-ci avait été donné favori par trois « sondages » réalisés indépendamment par Archibald Crossley, Elmo Roper et George Gallup (ce dernier a créé son propre institut en 1935). Ces trois études ont été réalisées selon la méthode par choix judicieux : tirage des personnes interrogées avec contrôles par quotas, une « Amérique en microcosme » d'après le mot de E. Roper.

En dépit de sa grande taille, l'échantillon du *Literary Digest* s'était révélé biaisé ; la technique des votes de paille disparaît alors. Dès 1937, de nombreuses revues américaines créent une rubrique d'opinion, nourrie par des consultations réalisées par des organismes privés utilisant les sondages représentatifs.

L'année 1938 voit l'apparition des premiers instituts d'études de l'opinion en Grande-Bretagne et en France ; Jean Stœtzel, professeur de sociologie à la Sorbonne, fonde l'IFOP (Institut français d'Opinion publique). De nombreux autres pays vont suivre ce mouvement et en 1947 se tient, à Paris, le premier congrès international en ce domaine.

VI. — La situation en Russie pré-révolutionnaire

Au XIXe siècle, en même temps qu'apparaissent et se généralisent, aux Etats-Unis, les sondages d'opinion, l'Empire russe connaît un développement des méthodes fondées sur une vision partielle de la population.

La cadre est institutionnel : en 1864, le pouvoir central crée les zemstvos, gouvernements locaux au sein desquels existent des départements statistiques chargés de recueillir, en particulier, les informations sur l'état de l'agriculture pour aider à définir les orientations de la politique économique rurale. Le père spirituel de la « statistique zemstvo » est Alexander Ivanovitch Tchuprov (1842-1908), professeur d'économie politique et de statistique à l'Université de Moscou, membre de l'IIS dès sa création en 1885, fondateur de la Société juridique de Moscou.

L'existence d'instituts statistiques dans les zemstvos favorise l'apparition de nombreux statisticiens mathématiciens qui vont contribuer fortement au développement des méthodes : outre A. I. Tchuprov, il faut mentionner son fils Alexander Alexandrovitch Tchuprov (1874-1926), professeur au département Economie de l'Institut polytechnique de Saint-Pétersbourg, A. V. Peshekhonov (1867-1933), V. G. Groman, qui deviendra après la révolution d'Octobre l'un des grands responsables de l'Office central statistique et du Gosplan, et A. G. Kovalevsky.

Il semble que, très tôt, les techniques d'échantillonnage aient été utilisées par les statisticiens russes, dans l'impossibilité d'observer exhaustivement les territoires de leur compétence. Dès 1894, à Moscou, A. I. Tchuprov présente une communication sur la « méthode monographique » consistant en la sélection d'éléments typiques déterminés à partir d'une connaissance exhaustive d'une population. En 1897, au congrès de l'IIS à Saint-Pétersbourg, il intervient sur

les méthodes représentatives que le Norvégien A. Kiær a exposées à Berne en 1895. Si A. I. Tchuprov jouera ensuite un grand rôle pour relayer et populariser les idées de Kiær auprès des statisticiens russes, au sens où il peut faire état de l'utilisation de ces techniques dans les pays de l'Europe de l'Ouest, il est admis que les sondages — au sens large du terme — sont néanmoins communément utilisés dans la statistique russe indépendamment de Kiær : en 1896, Peshekhonov procède à un échantillonnage au 1/10 de zones géographiques, une sélection au 1/5 est tirée pour le district de Viatka en 1906, etc.

Une avancée méthodologique essentielle apparaît en 1910, parallèlement aux travaux de Bowley en Angleterre : A. A. Tchuprov étudie l'échantillonnage aléatoire, développe son utilisation pour mesurer la précision des estimateurs, et fait référence au tirage par grappes et à l'échantillonnage stratifié, avec et sans remise. Influencés par A. Markov, conscient de l'importance de l'utilisation des probabilités en statistique, Tchuprov et ses élèves vont obtenir de nombreux résultats qui ne seront découverts, ou popularisés, que bien plus tardivement en Europe de l'Ouest, notamment par J. Neyman.

Le meilleur exemple du niveau atteint en Russie par les sondages, en pratique et en théorie, est fourni par le contenu d'un ouvrage publié en 1924 par A. G. Kovalevsky. Au-delà d'une présentation mathématique du tirage systématique, Kovalevsky donne un traitement rigoureux du sondage par stratification, et établit l'allocation optimale par strate — ce qui est contraire à l'interprétation de la notion de représentativité admise à l'époque. Ce résultat ne sera découvert par Neyman que dix ans plus tard.

Les difficultés dues à l'éloignement, à la langue et, bien sûr, aux événements liés à la révolution d'Octobre ont longtemps empêché de bien saisir l'importance de l'école russe dans la « statistique partielle ». Il semble

dorénavant bien établi que, dès l'origine, le développement de la méthodologie et de l'utilisation des sondages en Russie tsariste ait été considérable, et que nombre de statisticiens russes du début peuvent être considérés comme les précurseurs et les initiateurs de la méthode, au même titre que Kiær, Bowley et Neyman.

VII. — Les débuts du développement contemporain

Après 1925, la discussion sur la méthode représentative ne porte plus sur son principe, mais sur ses modalités d'application. Le Congrès de l'IIS ayant distingué méthode aléatoire et méthode raisonnée, les travaux approfondissent cette double approche. Dans la lignée des résultats de Bowley et Tchuprov, Jerzy Neyman (1894-1981) peut être considéré comme l'un des fondateurs de la théorie statistique des sondages, entre autres par son exposé du 19 juin 1934 devant la Royal Statistical Society (publié dans le *Journal of Royal Statistical Society* de cette même année).

L'apport de Neyman va être considérable car, le premier, il va prendre parti entre échantillonnage aléatoire et échantillonnage raisonné qui sont placés sur le même plan par Bowley et la recommandation de l'IIS. Pour cela, il s'appuie sur les travaux de C. Gini et L. Galvani qui, en novembre 1926 utilisent la méthode par choix raisonné pour tirer un échantillon à 15 % du recensement général effectué en Italie en 1921, échantillon que les deux statisticiens italiens souhaitent « représentatif du pays dans ses composantes démographique, sociale, économique et géographique ». Partant des résultats sur la théorie de l'estimation de Ronald Fisher, Neyman exhibe les hypothèses à partir desquelles les estimateurs issus de la méthode par choix raisonné ont de bonnes propriétés statistiques de convergence, hypothèses qui, en pratique, ne sont

jamais toutes satisfaites simultanément. Il montre donc, *a contrario*, que les deux techniques d'échantillonnage ne sont pas équivalentes,et qu'il convient de recommander la méthode aléatoire puisque « ... la confiance générale placée dans la sélection judicieuse devra être quelque peu réduite... »

Au-delà de cette question — fondamentale — de choix entre méthodes d'échantillonnage, Neyman contribue à de nombreux développements : à l'échantillonnage aléatoire simple, avec ou sans remise, et à l'échantillonnage stratifié, déjà envisagés par Bowley, il ajoute l'échantillon par grappes, technique qui a sa préférence ; il fait grandement avancer la théorie de l'estimation en sondages ; il étudie l'allocation optimale d'un échantillon par strates, ainsi que le choix de l'allocation minimisant la variance de l'estimateur, à plan d'échantillonnage fixé. Ce dernier point est important, puisqu'il remet en cause le principe de représentativité communément admis consistant en une homothétie de la population. Dans l'histoire du concept de représentativité, l'apport de Neyman sur ce point est essentiel.

Etablissant la supériorité des techniques de sondage aléatoire en univers stratifié, Neyman oriente une grande partie de la recherche dans cette voie. Ironie des temps, les critiques — justifiées — que Neyman adresse à la méthode représentative à choix judicieux coïncident avec l'émergence de la vogue des sondages d'opinion effectués selon cette même technique, ou plutôt l'une de ses diverses variantes, la méthode des quotas.

En 1938, Neyman mène à son terme l'échantillonnage à deux degrés, et l'utilisation conjointe des fonctions de coût et de la variance pour définir les allocations optimales.

Toutes les bases de la statistique des sondages sont alors posées. Deux grandes voies de développement apparaissent : l'une est liée aux problèmes de la

pratique (modes de tirages, estimation, variabilité, non-réponses, etc.), l'autre fait référence au modèle probabiliste sous-jacent à la théorie des sondages. Les avancées de la première catégorie ont permis la construction de la méthodologie fondamentale des sondages. Citons-en quelques grandes étapes sans aucune prétention d'exhaustivité : Kiær (1897), Tchuprov (1923) et Neyman (1934) ont évoqué à des titres divers les tirages à probabilités inégales. Morris Hansen et William Hurwitz, en 1943, établissent les résultats d'estimation et de précision des estimateurs lorsque les unités élémentaires sont tirées avec des probabilités inégales, pour des plans de sondage avec remise. La détermination des probabilités d'inclusion optimales est due aux mêmes auteurs, en 1949. D. Horvitz et T. Thompson généraliseront, en 1952, les résultats précédents aux tirages sans remise, avec probabilités inégales. Les conditions de positivité des estimateurs de la variance dues à F. Yates et P. Grundy sont publiées en 1953.

Parallèlement, se développent les procédures d'estimation des paramètres d'intérêt lorsqu'une information supplémentaire existe. Les estimateurs par le quotient, par la différence et par la régression apparaissent en filigrane chez Neyman (1934), mais sont approfondis par W. Cochran (1942), M. Hansen, W. Hurwitz et W. Madow (1953).

La pratique laisse une place importante aux problèmes de non-réponses, totales ou partielles, auxquels des éléments de solutions commencent à être apportés par Hansen et Hurwitz (1946) et Tore Dalenius (1957), et aux enquêtes à passages répétés. On peut considérer que Raymond Jessen (1942) apporte, une solution à ces questions spécifiques.

L'étude de la variabilité des estimateurs est également un vaste champ d'investigation, essentiellement pour des statistiques non linéaires. Aux méthodes de linéarisation usuelles en statistique classique vont se

superposer les procédures de réplication d'échantillon : P. Mahalanobis (1946), W. Deming (1960), les techniques de demi-échantillons équilibrés (P. Mac - Carthy, 1966), qui anticipent les applications du jackknife et du bootstrap en pleine actualité.

CHAPITRE V

L'ÉMERGENCE DE LA STATISTIQUE MATHÉMATIQUE : ESTIMATION ET TESTS

I. — Introduction

En règle générale, une discipline scientifique ne naît pas dans l'absolu, mais émerge dans un contexte bien particulier décrit, entre autres éléments, par le niveau mathématique du moment et l'état des besoins des praticiens. La statistique mathématique n'échappe pas à cette logique et, s'il est généralement convenu que les années 1885-1925 marquent la naissance de la statistique en tant que théorie cohérente, c'est que l'environnement scientifique des dernières années du XIXe siècle est très fourni. En effet, non seulement existent déjà des outils statistiques puissants, tels que les moindres carrés, la régression, la corrélation, mais les théories mathématiques sur lesquelles va s'appuyer la statistique sont en plein essor : la théorie des erreurs est déjà bien développée, la théorie de l'intégration au sens de Riemann est en place, les travaux de Cantor, Péano et Borel commencent à être diffusés.

Parallèlement, les domaines où le besoin de prise en compte de l'aléatoire se fait sentir sont de plus en plus nombreux, et ne se limitent plus à l'astronomie ou la géodésie : le champ d'expérience s'étend à l'agronomie, la biométrie, la psychométrie, l'économie.

Enfin, l'apparition des premiers enseignements de statistique pour des mathématiciens va favoriser l'éclosion d'une génération de « statisticiens mathéma-

ticiens » dont les connaissances théoriques en analyse, en algèbre et en géométrie vont permettre l'émergence de la statistique mathématique comme discipline scientifique à part entière.

II. — Les premières méthodes d'estimation

Il serait bien sûr complétement faux d'écrire que le problème de l'estimation naît avec la théorie développée par R. A. Fisher. Dès les débuts du calcul des probabilités, on voit poindre l'idée qui consiste à donner une valeur numérique à un paramètre associé à une expérience aléatoire à partir d'une séquence de réalisations de cette expérience. La loi faible des grands nombres de Jacques Bernoulli [*Ars Conjectandi*, 1713] en est un exemple.

Quelques années plus tard, l'œuvre de Thomas Bayes publiée en 1763, après sa mort — parce qu'il n'était pas sûr de ses conclusions, *dixit* Fisher —, engendre, par l'utilisation de la probabilité conditionnelle, une méthode d'estimation qui consiste à maximiser en le paramètre cette probabilité *a posteriori* ; en usant de notations « modernes », avec $f(x \mid \theta)$ la loi *a priori* de la v.a. X pour une valeur du paramètre θ, $g(\theta)$ la loi du paramètre θ sur la population de référence, la densité marginale de l'échantillon $(x_1, ..., x_n)$ est :

$$f(x_1, ..., x_n) = \int_\theta \prod_{i=1}^{n} f(x_i \mid \theta)\, g(\theta)\, d\theta.$$

Conditionnellement à $(x_1, ..., x_n)$, la loi *a posteriori* de θ est donc :

$$f(\theta \mid x_1, ..., x_n) = \frac{g(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)}{\int_\theta g(\theta) \prod_{i=1}^{n} f(x_i \mid \theta)\, d\theta}$$

Pierre Simon de Laplace, successivement dans *Mémoire sur la probabilité des causes* (1774) puis dans *Théorie analytique des Probabilités* (1812), développe la procédure consistant à maximiser, en $\theta, f(\theta \mid x_1, ..., x_n)$.

Pour être exact, Bayes postule seulement une loi uniforme sur θ, reflétant l'ignorance totale sur le paramètre qu'il érige en principe général et que Laplace dénommera le « principe de la raison insuffisante ». Cette technique porte le nom d'*estimation par la probabilité inverse* .

Mentionnons également l'apport de la méthode des moindres carrés (Legendre, Gauss, Laplace), déjà vue par ailleurs (chapitres II et III), puisque la méthode permet d'estimer le vecteur des cœfficients d'un modèle linéaire.

En 1818, dans la dernière partie du second supplément à la *Théorie analytique des Probabilités*, Laplace s'intéresse à l'estimation de y dans le modèle linéaire

$$a_i = yp_i + x_i \qquad (i = 1 \text{ à } n)$$

où (a_i, p_i) sont connus, y est un paramètre à estimer et x_i la réalisation d'un aléa pair (notations originales de Laplace). Dans ce supplément, Laplace fait référence à une méthode introduite par Boscovich en 1757, appelée « méthode de situation », pour laquelle il a développé dès 1799 dans *Mécanique céleste* un algorithme de résolution pour le problème :

$$\text{Min} \sum_{i=1}^{n} |yp_i - a_i| .$$

Si tous les p_i sont positifs et les indices tels que

$$\frac{a_1}{p_1} \geq \frac{a_2}{p_2} \geq ... \geq \frac{a_n}{p_n},$$

il définit l'entier k tel que

$$\sum_{i=1}^{k-1} p_i < \sum_{j=k}^{n} p_j \text{ et } \sum_{i=1}^{k} p_i > \sum_{j=k+1}^{n} p_j ;$$

alors la valeur optimale de y est $y^* = \frac{a_k}{p_k}$, estimateur se ramenant à la médiane si toutes les variables exogènes p_i sont constantes (prises égales à 1). Laplace détermine ensuite la loi de $y^* - y$, et montre que, lorsque n tend vers l'infini, $y^* - y$ converge en loi vers la loi normale N $(0, \sigma)$ avec :

$$\sigma = \frac{1}{2f(0) \sum_{i=1}^{n} p_i^2} .$$

où f est la densité de l'aléa x_i.

Il compare ensuite l'estimation $\hat{y}$ des moindres carrés avec y^*, et montre que $\hat{y}$ est préférable à y^* si :

$$4 f^2(0) \int_{\mathbf{R}} x^2 f(x) dx > 1$$

et donc en particulier si les x_i sont « normaux ».

La place détermine alors la loi jointe du couple $(y^* - y, \hat{y} - y)$ en inversant la fonction caractéristique. Pour n grand, il établit que cette loi peut être approximée par une loi normale bivariée et utilise ce résultat pour déterminer la valeur de la constante λ minimisant, V_{as} désignant la variance asymptotique :

$$\min_{\lambda} V_{as} [\hat{y} - \lambda (\hat{y} - y^*)].$$

Les travaux de Laplace ont presque cent ans d'avance sur ceux de Fisher ; il n'est pas loin du concept d'exhaustivité sur lequel va reposer toute la théorie fishérienne de l'estimation. Certains statisticiens considèrent, peut-être à juste titre, que Pierre Simon de Laplace est le premier théoricien de la statistique moderne.

A la fin du XIXe siècle, Karl Pearson, travaillant sur les familles de densités de probabilités, introduit la méthode des moments, technique d'estimation consistant à résoudre un système d'équations égalant moments théoriques et moments empiriques, le nombre d'équations à prendre en compte étant égal à la dimension du paramètre à estimer. Les fondements de la méthode des moments sont asymptotiques et traduisent la convergence des moments empiriques calculés sur un grand échantillon vers les moments théoriques associés. Dès 1894, K. Pearson utilise cette technique pour estimer les paramètres π, m_1, σ_1, m_2, σ_2 d'un mélange de lois normales $\pi\ P_1 + (1 - \pi)\ P_2$, avec $P_1 = N\ (m_1, \sigma_1)$ et $P_2 = N\ (m_2, \sigma_2)$, obtenant la célèbre équation du neuvième degré dite « nonique de Pearson ».

III. — La théorie de l'estimation au sens de Fisher

L'étape suivante de la théorie de l'estimation débute en avril 1912 avec la première publication de R. A. Fisher utilisant le maximum de vraisemblance (bien que ne portant pas encore ce nom), méthode développée et popularisée par deux articles datant de 1922 et 1925 (« On the mathematical foundations of theoritical statistics », *Phil. Trans. Roy. Soc.*, A, 222, p. 390-368, et « Theory of Statistical estimation », *Proc. Comb. Phil. Soc.*, 22, p. 700-725). Au delà de la technique de l'estimation elle-même, l'important est que, pour la première fois, un modèle d'inférence statistique est proposé, contenant trois principes fondamentaux : la vraisemblance, l'exhaustivité, l'information.

Fisher possède une vision globale des problèmes statistiques, ainsi qu'en témoignent les extraits ci-après de « On the mathematical foundations of theorical statistics » (1922) :

« L'objet de la méthode statistique est la réduction des données. Une masse de données doit être remplacée par un petit nombre de quantités représentant correctement cette masse, et contenant autant que possible la totalité de l'information pertinente contenue dans les données d'origine. Cet objectif est accompli par la construction d'une population infinie hypothétique. La statistique comporte des problèmes de spécification apparaissant à travers le choix de la forme mathématique de la population, des problèmes d'estimation, impliquant le choix de méthodes de calcul de quantités dérivées de l'échantillon, que nous appellerons statistiques, construites pour estimer les valeurs des paramètres de la population hypothétique, et enfin des problèmes de distribution. »

L'approche de Fisher repose sur le concept de vraisemblance, introduit en 1912 : si f est la densité théorique d'une variable observable X dépendant de p paramètres $\theta_1, \ldots, \theta_p$, « il faut choisir pour les θ_i la valeur qui rend P maximum où P est la fonction définie par :

$$\text{Log P} = \sum_{i=1}^{n} \text{Log} f(x_i, \theta_1, \ldots, \theta_p)$$

Fisher commet néanmoins la maladresse d'écrire que « dans le cas de deux paramètres, la surface de P représente la probabilité inverse », donnant par là l'impression qu'il n'a fait que retrouver la méthode de Bayes vue au début du § II. Il s'agit là de l'une des causes de la brouille entre Pearson et Fisher, le premier ayant beau jeu de proclamer l'absurdité d'une technique de maximisation d'une probabilité inverse calculée sous l'hypothèse souvent aberrante de loi uniforme sur les paramètres. Toute sa vie, Fisher se défendra contre l'assimilation de la vraisemblance à une probabilité, puisqu'il n'y a pas d'élément différentiel :

« La vraisemblance caractérise un point et non un intervalle élémentaire ; vraisemblance n'est pas synonyme de probabilité. »

Au-delà de la maximisation de la vraisemblance, Fisher s'intéresse aux qualités que l'on peut attendre d'un estimateur. La notion de convergence en proba-

bilité apparaît en 1925, encore que, dès 1922, est introduit le concept de consistence : une statistique calculée sur l'ensemble de la population de référence doit être égale au paramètre à estimer. L'estimation sans biais est également définie, mais peu utilisée en raison de la convergence.

En 1920, Fisher publie un important article dans lequel interviennent les notions d'exhaustivité, d'efficacité et d'information (« A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean-square error », *Mon. Nat. Roy. Ast. Soc.*, 80, p. 758-770). L'article repose sur l'estimation de l'écart type σ à partir d'un échantillon $(x_1, ..., x_n)$ de la loi normale $N(m, \sigma)$. Comparant :

$$\sigma_1 = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \text{ et } \sigma_2 = \sqrt{\frac{1}{n} \Sigma (x_i - \bar{x})^2},$$

Fisher montre l'indépendance de $\bar{x}$ et σ_2, calcule les écarts types à distance finie et asymptotiques de σ_1 et σ_2, définit l'efficacité relative de σ_1 par rapport à σ_2, et établit que, conditionnellement à σ_2, la loi de σ_1 ne dépend pas de σ ; il conclut que la connaissance de σ_1 n'apporte aucune information supplémentaire sur σ :

« Toute l'information sur σ contenue dans l'échantillon est résumée dans la valeur de σ_2. »

Cette propriété ne sera appelée exhaustivité qu'en 1922, année où Fisher énonce le critère de factorisation et lie exhaustivité et vraisemblance.

Bien que la comparaison des variances de deux estimateurs, l'efficacité relative, ait déjà été utilisée par Laplace (cf. § II), Fisher montre que :

$$-1 / E\left(\frac{\partial^2 \text{Log P}}{\partial \theta^2}\right)$$

est la plus petite valeur possible pour la variance d'un estimateur, et définit l'efficacité absolue (« précision intrinsèque », 1925).

Le concept d'information, cité par analogie avec le langage courant dans le texte de 1920, est développé de 1925 à 1934 : définition par $-\mathrm{E}\left(\frac{\partial^2 \mathrm{Log\ P}}{\partial\theta^2}\right)$, additivité en cas d'échantillon indépendant, liaison avec l'exhaustivité. Enfin, Fisher définit les fonctions pivotales en 1930.

Ainsi, dès 1930, la statistique dispose d'un modèle général d'estimation et d'inférence, dont la cohérence a été établie en vingt ans à peine. Certes, le modèle fishérien a été critiqué par la suite, en particulier en raison du rôle central joué par les statistiques exhaustives dont l'existence n'est pas toujours assurée, mais nul ne peut nier son importance dans l'histoire de la statistique.

IV. — L'estimation par intervalle

Une forme différente du problème de l'estimation apparaît dans les années trente, l'estimation « par intervalle ». Le premier article dans lequel cette technique est mentionnée est dû à un agronome polonais, W. Pytkowski (« The dependence of the income in small farms upon their area, the outlay and the capital invested in cows », *Biblioteka Pulanska* , 1932). Néanmoins, le fondateur de la théorie des régions de confiance est Jerzy Neyman, présentée en annexe de l'article sur la théorie des sondages publié en 1934 (cf. chapitre IV), et définitivement établie par les deux articles de référence qui sont « Outline of a theory of statistical estimation based on the classical theory of probability », *Phil. Trans. Roy. Soc.*, 1937, A, 236, p. 333-380, et « L'estimation statistique traitée comme un problème classique de probabilité », *Actualités sci. et ind.*, 1938, p. 25-57.

V. — La théorie des tests

Le concept de test d'hypothèse et la théorie sous-jacente apparaissent bien plus tardivement que la théorie de l'estimation. Cependant, le critère d'adéquation de $\varkappa^2$ de Pearson (1900) relève du problème du test d'ajustement à une loi de probabilité parfaitement connue ; les résultats sur le degré de liberté du $\varkappa^2$ à utiliser lorsque des paramètres de la loi testée doivent être estimés n'apparaitront que bien après, par G. U. Yule, ancien assitant de K. Pearson (1922), et R. Fisher : « The conditions under which $\varkappa^2$ measures the discrepancy between observation and hypothesis », *J. Roy. Stat. Soc.*, 1924, p. 442-450.

Les fondateurs de la théorie des tests d'hypothèses sont Jerzy Neyman et Egon Pearson, fils de Karl. Néanmoins, J. Neyman mentionne qu'à sa connaissance, la première approche d'un test d'hypothèse est due à Laplace (encore !) qui, en 1812, se demande si les comètes sont des éléments réguliers du système solaire ou au contraire des intrus venant d'un autre système. Travaillant sous cette dernière hypothèse, il la quantifie en écrivant qu'alors l'angle fourni par la trajectoire d'une comète et l'écliptique (le plan dans lequel la terre tourne autour du soleil) doit être uniformément distribué entre 0 et 90°.

La construction de la théorie des tests de Neyman et Pearson a lieu entre 1926 et 1933, les références fondamentales étant « On the use and interpretation of certain test criteria for purposes of statistical inference », *Biometrika* , 1928, p. 175-240 et 263-337, et « On the problem of the most efficient tests of statistical hypotheses », *Phil. Trans., Roy. Soc.* A., 1933, p. 289-337. Au delà des principes et des techniques développés par Neyman et Pearson, leur apport essentiel est d'élaborer une conception statistique différente du modèle de vraisemblance de Fisher par l'intermédiaire d'un modèle décisionnel qui servira

de base aux travaux d'Abraham Wald (*Statistical Decision Functions* , 1950). Ils présentent en effet leur optique comme un problème de choix entre deux décisions possibles : accepter ou refuser l'hypothèse privilégiée.

Outre la terminologie générale des tests, toujours employée, le principe de référence apparaît dès 1928 : le rapport des vraisemblances sous l'hypothèse nulle et l'hypothèse alternative (« le λ — principe »). Il semble d'ailleurs que l'idée soit de E. Pearson, tout acquis à la notion de vraisemblance de Fisher, et que Neyman ait été d'abord fort réticent, faisant allusion aux probabilités inverses de Bayes. Les réflexions de Neyman et Pearson se poursuivent (à distance) ; leurs travaux de 1930 et 1931 (publiés en 1933) portent sur la protection de l'hypothèse nulle contre un abandon hâtif, le risque β de seconde espèce, la recherche de la *best critical region* (appellation de 1930) qui devient en 1931 le « uniformly most powerful (UMP) test » (test uniformément le plus puissant).

Après 1933, Neyman s'intéresse aux cas où un test UMP n'existe pas ; il recherche alors des compromis, des « bonnes régions critiques », développe la théorie des tests sans biais, et établit la liaison entre les tests et la méthode de l'estimation par intervalle de confiance.

Comment ne pas conclure en constatant que la collaboration entre Neyman et E. Pearson ne fut jamais aussi forte qu'entre 1928 et 1933, lorsqu'ils étaient l'un en Pologne, l'autre en Angleterre ; quand, en 1934, Neyman arrive au Département de Statistique de l'University College de Londres, dirigé par Pearson, leurs publications communes s'arrêtent.

CHAPITRE VI

STATISTIQUE NON PARAMÉTRIQUE ET ROBUSTESSE

Au-delà des développements récents engendrés dans chacun des domaines divers mentionnés, tels que l'Analyse des Données, les modèles stochastiques de type ARIMA en séries chronologiques, les modèles de superpopulation en théorie des sondages, il est communément admis que, au niveau des méthodes, la statistique non paramétrique et les procédures robustes sont les dernières nées. Cet ouvrage nous paraîtrait incomplet s'il ne resituait pas ces techniques dans une perspective historique.

I. — Les premières réflexions sur les valeurs aberrantes

Le problème de l'existence de valeurs aberrantes (*spurious observations* ou *outliers* en anglais) porte en lui le germe des méthodes robustes. Il est d'autant plus important qu'il est rencontré dans de nombreuses disciplines, de l'astronomie à la chimie et la physique. Après Boscovich qui, dès 1770, s'interroge sur l'opportunité de calculer la moyenne d'un échantillon après suppression de la plus petite et de la plus grande observation, il semble que le premier à avoir préconisé le rejet des valeurs aberrantes ait été Legendre, en 1805. Cette attitude est loin d'être suivie par le monde scientifique de l'époque, plutôt enclin à conserver

l'entièreté des données : le célèbre astronome prussien Bessel, en 1838, défend le fait que toutes les observations doivent être utilisées, avec le même poids.

En 1852, l'astronome et mathématicien de Harvard Benjamin Pierce énonce le premier critère probabiliste de rejet des *outliers*, immédiatement combattu par un autre astronome, G. B. Airy.

Le débat sur rejet ou non dure près de vingt ans, essentiellement entretenu par les astronomes anglais et américains : W. Chauvenet (1863), J. W. Glaisher (1872), E. J. Stone (1873). Bien que les tenants de l'intégralité de l'échantillon soient encore nombreux, le nombre de critères de détection des points aberrants croît sans cesse ; néanmoins, l'essentiel de la réflexion statistique consiste alors à utiliser la moyenne sur les données restantes.

Pourtant, des procédures d'estimation plus élaborées existent, parfois depuis longtemps. Ainsi, en 1763, l'astronome (encore !) anglais James Short utilise la moyenne de trois moyennes : celle, m_1, de l'échantillon, celle des observations situées à moins d'une seconde de m_1, celle des observations situées à moins d'une demi-seconde de m_1.

Des estimateurs pondérant les données de façon différente, et éventuellement par zéro les valeurs extrêmes, sont développés par A. de Morgan (1847), Glaisher (1873), W. Jevons (1874), Stone (1875) et F. Edgeworth (1884, 1885).

La réflexion sur les critères de rejet se poursuit également. En 1884, T. Wright propose de rejeter toute observation X_i d'un échantillon $(X_1, ..., X_n)$ dont l'écart $X_i - X$ dépasse 5 fois « l'erreur probable », c'est-à-dire, pour une loi normale, la moitié de l'intervalle interquartile $0{,}6745\ \sigma$.

En 1899, W. Sheppard introduit la première règle de rejet pour des données multidimensionnelles. D'autres règles seront introduites ensuite par Student, en 1927, par W. Thompson en 1935, et, plus près de

nous et toujours d'actualité, par F. Grubbs (1950, 1969) et W. Dixon (1950, 1951).

II. — Les prémisses de la robustesse et de la statistique d'ordre

Nous avons déjà noté que la deuxième moitié du XIXe siècle a vu apparaître des estimateurs pondérés ; néanmoins les ancêtres des L-estimateurs, fonctions linéaires de la statistique d'ordre, sont employés bien avant. Comme détaillé au chapitre V, en 1818, dans sa *Théorie analytique des Probabilités*, Laplace estime le paramètre y dans le modèle linéaire :

$$a_i = y\, p_i + x_i \quad (i = 1 \text{ à } n)$$

où les aléas x_i sont de loi symétrique, en minimisant $\sum_{i=1}^{n} |x_i|$. En 1816, voulant estimer l'erreur probable dans une loi N (m, σ), Gauss suggère de fonder l'estimateur sur $\sum_{i=1}^{n} |X_i - \overline{X}|$, intuition dont l'usage rigoureux fut établi par Dirichlet.

En 1875, F. Galton développe l'emploi de la médiane ; Edgeworth l'utilise pour des distributions de probabilité à queues plus épaisses que celles de la loi normale (1887, 1888). En 1889, Galton suggère d'estimer m et σ, dans une loi N (m, σ), par :

$$\hat{m} = \frac{u_p\, X_{n,q} - u_q\, X_{n,p}}{u_p - u_q} ; \hat{\sigma} = \frac{X_{n,p} - X_{n,q}}{u_p - u_q}$$

où u_p et u_q sont les fractiles vrais d'ordre p et q de la loi N (0, 1) et $X_{n,p}$ (resp. $X_{n,q}$) le fractile d'ordre p (resp. q) calculé sur l'échantillon.

Dix ans après, Sheppard démontre la normalité jointe de $\hat{m}$ et $\hat{\sigma}$, et donne une classe générale d'estimateurs de m et σ fondée sur un nombre fini quelconque de fractiles. Néanmoins, tous ces dévelop-

pements tournent autour de la loi normale, et ne correspondent pas toujours avec les distributions à queue épaisse, génératrices potentielles de points aberrants.

Les premiers modèles adaptés à cette situation sont dûs à Simon Newcomb, probablement le plus célèbre astronome américain du XIXe siècle. En 1882, Newcomb étudie les mélanges de lois normales de variances différentes, modèle lui permettant de ne pas éliminer systématiquement les observations aberrantes. En 1886, il donne un estimateur de tendance centrale plus « robuste » que la moyenne en utilisant la théorie de la décision bayésienne. Dans le même article, Newcomb introduit un concept très proche de la sensibilité de Tukey ou de la fonction d'influence de F. Hampel ; de même, en 1912, une publication posthume fournit un estimateur pondéré qui n'est pas loin d'être un M-estimateur comme introduit par P. Huber.

Pour conclure cette partie, il convient de mentionner que le qualificatif de robuste utilisé dans les lignes qui précèdent l'a été par anticipation. Il est employé pour la première fois par G. Box en 1953, dans un article intitulé « Non-normality and tests on variances », *Biometrika*. Même K. Pearson ne l'emploiera pas, alors que, dès 1931, il note la haute sensibilité des procédures usuelles de test d'égalité des variances au non-respect de l'hypothèse de loi normale.

III. — Les origines des L et M-estimateurs

Le paragraphe précédent contient de nombreux exemples de statistiques appartenant à la classe des L-estimateurs, c'est-à-dire des fonctions linéaires de la statistique d'ordre. Les développements théoriques des L-estimateurs peuvent être considérés comme fondés par deux articles de Percy Daniell, mathématicien

anglais en poste à l'Université de Sheffield (« Observations weighted according to order », *Amer. Journ. Math.*, 1920, 42, p. 222-236, et « Integral products and probability », *Amer. Journ. Math.*, 1921, 43, p. 143-162). Daniell y établit les moments de toute fonction de la statistique d'ordre. Il définit également, pour un échantillon ordonné en sens croissant $X_{(1)}, \ldots, X_{(n)}$ d'une v.a. X de densité f, la statistique :

$$L = \frac{1}{n} \sum_{i=1}^{n} h\left(\frac{i}{n+1}\right) X_{(i)}$$

et détermine la variance asymptotique de L. Daniell énonce enfin les critères de choix optimal de h, et étudie en détail les propriétés de la moyenne tronquée (*trimmed mean*). Il est curieux de constater qu'il faudra attendre trente ans pour que les résultats obtenus par Daniell — par ailleurs très connu pour ses travaux en théorie de l'intégration — soient parfaitement perçus par le monde des statisticiens.

En 1972, Peter Huber introduit la classe des M-estimateurs associée à une fonction objectif $\varrho(.\,, \theta)$ comme la solution en θ de l'équation : $\sum_{i=1}^{n} \Psi(X_i, \theta) = 0$, où Ψ est la dérivée de $\rho(.\,, \theta)$ par rapport à θ. Pour un paramètre de position, on obtient $\sum_{i=1}^{n} \Psi(X_i - \theta) = 0$. Or, en 1844, dans un article intitulé « On the method of least squares », *Trans. Comb. Phil. Soc.*, 8, p. 204-219, R. Ellis écrit :

« On ne peut donner aucune raison satisfaisante au fait que la règle de la moyenne arithmétique solution de $\sum_{i=1}^{n} (X_i - \theta) = 0$, soit privilégiée par rapport à d'autres estimateurs inclus dans l'équation générale $\sum_{i=1}^{n} f(X_i - \theta) = 0$ où f est une fonction paire. »

IV. — Les tests non paramétriques

Les méthodes non paramétriques ne sauraient posséder des fondements aussi lointains que ceux de la robustesse, d'une part parce que les problèmes concrets, tels que ceux des valeurs aberrantes, sont fort anciens et, d'autre part, parce que la base théorique à partir de laquelle les principes non paramétriques se développent est constituée par le modèle de la statistique inférentielle classique.

Le premier problème relevant des méthodes non paramétriques semble être l'adéquation d'un échantillon à une loi de probabilité donnée. Le célèbre test du $\varkappa^2$ de Karl Pearson est publié en 1900. Bien plus tard, H. Cramer (1928) et R. von Mises (1931) construisent des statistiques de test fondées sur $[F_n(x) - F_0(x)]^2$, carré de la différence entre la fonction de répartition empirique et celle de la loi testée. En 1933, A. Kolmogorov introduit le critère sup $|F_n(x) - F_0(x)|$ dont il détermine la loi asymptotique ; en 1939, N. Smirnov utilise la statistique de Kolmogorov pour un test unilatéral, et propose un test de comparaison des lois de deux échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_n)$ basé sur le critère sup $|F_n(x) - G_n(x)|$, où F_n (resp. G_n) est la fonction de répartition empirique de X_i (resp. Y_i). Il faudra attendre le début des années soixante pour étendre ce test à k échantillons ($k > 2$).

Un autre domaine précurseur est la mesure des liaisons et donc les tests d'indépendance. Dans le contexte de la psychométrie, très en vogue au début du siècle, naît une première mesure d'association, le cœfficient de corrélation des rangs de C. Spearman (« The proof and measurement of association between two things », *Amer. Journ. Psychol.*, 15, 1904, p. 72-101). Comme son nom l'indique, le cœfficient de Spearman consiste à ordonner indépendamment les deux échantillons $(X_1, ..., X_n)$ et $(Y_1, ..., Y_n)$ d'une

variable bidimensionnelle (X, Y) et à calculer la corrélation sur les rangs ainsi obtenus.

Dans la même optique, une autre mesure d'indépendance est due à M. G. Kendall, en 1938 (« A new measure of rank correlation », *Biometrika*, 30, p. 81-93). A partir des lois exactes ou asymptotiques de ces deux cœfficients sont bâties des procédures de tests qu'il est tentant de vouloir comparer aux techniques usuelles dans le cas de la loi normale. Ces comparaisons vont être fondées sur les travaux de E. Pitman qui, publiés de 1937 à 1948, donnent naissance au concept d' « efficacité relative asymptotique ». L'efficacité du test de Spearman par rapport au test classique sur le cœfficient de corrélation linéaire dans la loi normale bivariée est calculée en 1936 par H. Hotelling et M. Pabst.

C'est pour savoir si deux échantillons proviennent d'une même loi que les outils non paramétriques vont se révéler performants, par l'apparition des tests de rangs. Bien qu'il n'ait pas été le premier à proposer une telle approche, on peut considérer F. Wilcoxon comme le père des tests de rang (« Individual comparisons by ranking methods », *Biometrika*, 1, 1945, p. 80-83), suivi par H. Mann et D. Whitney qui étudient, en 1947, un test pour savoir si une variable est stochastiquement plus grande qu'une autre. Pitman montre, en 1948, que l'efficacité du test de Wilcoxon par rapport au test de Student est très acceptable dans le cas gaussien : 95,5 %. En 1956, J. Hodges et E. Lehmann montreront que cette efficacité dépasse toujours 86,4 % quelle que soit la loi considérée. De nombreux autres tests de rang existent pour le « problème des deux échantillons » — Mood (1954), Hœffding (1951), Freund et Ansari (1957), Ansari et Bradley (1960)... —, mais ceux-ci sont trop proches pour figurer dans cette perspective historique. Nous citerons néanmoins les travaux de W. Kruskal et W. Wallis (1952,1953) qui étendent les résultats de Wilcoxon à la comparaison de plusieurs

échantillons ; l'efficacité du test de Kruskal-Wallis est établie par F. Andrews en 1954.

Simultanément au développement de l'utilisation des statistiques de rang, les résultats probabilistes donnant les lois asymptotiques des rangs sont établis. Kolmogorov (1933) puis Hotelling et Pabst (1936) trouvent les lois limites des statistiques de Kolmogorov et Pearson sous l'hypothèse nulle et sous quelques hypothèses alternatives précises. Les deux publications théoriques fondamentales sont celle de von Mises (« On the asymptotic distribution of differentiable statistical functions », *Ann. Math. Statist.*, 1947, p. 309-348) et celle de Hœffding qui introduit la classe des U-statistiques (« A class of statistic with asymptotically normal distribution », *Ann. Math. Statist.*, 1948, p. 293-325).

Enfin, bien avant Wilcoxon, M. Friedman (1937) emploie les rangs pour se libérer de l'hypothèse de normalité dans le modèle de l'analyse de la variance à deux facteurs.

V. — Compléments sur le non paramétrique

La construction d'intervalles de confiance à partir des statistiques de rang apparaît en 1947, fondée sur les résultats de Wilcoxon. Néanmoins, les développements de l'estimation ponctuelle et par intervalle sont récents, dûs à Hodges et Lehmann, dans les années soixante.

Dans un autre ordre d'idées, la recherche s'est orientée vers la comparaison des diverses méthodes non paramétriques sur la base de critères d'optimalité. Le principe naturel est celui du minimax : minimum du maximum de la variance pour les problèmes d'estimation, minimum du maximum de la probabilité d'accepter à tort l'hypothèse nulle pour les problèmes de test, le maximum étant pris, dans chaque cas, sur une classe spécifiée de lois.

Les premiers travaux d'optimalité minimax semblent dûs à Hœffding (1951) qui montre que le test des signes est minimax.

Enfin, comment ne pas mentionner la classe des tests de permutation, par ailleurs en pleine actualité ? Ils ont été introduits par Fisher en 1935, puis étudiés par Pitman, Scheffe (1943), Wald et Wolfowitz (1944) entre autres. En 1949, Lehmann et Stein démontreront l'optimalité de tels tests au sein de la classe des procédures *distribution-free*.

CHAPITRE VII

L'ANALYSE DES SÉRIES CHRONOLOGIQUES

L'analyse des séries chronologiques occupe dans le domaine de la statistique une place importante et originale. Sa caractéristique essentielle réside en une dépendance des phénomènes étudiés vis-à-vis du temps, concept essentiel tant au niveau scientifique que philosophique. Il est peu de disciplines qui ne soient confrontées à l'étude de variables évoluant dans le temps et qu'on désire décrire, expliquer, contrôler ou prévoir. L'économie, la gestion des entreprises, la météorologie, la démographie, les sciences expérimentales... constituent quelques exemples de domaines privilégiés.

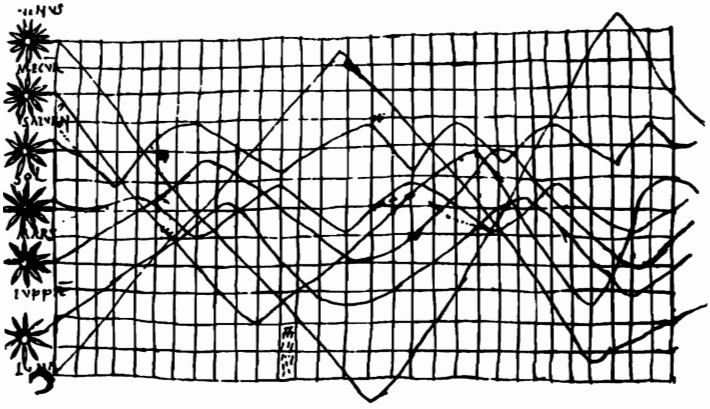
La diversité des méthodes d'analyse est non seulement fonction d'une évolution historique du problème mais aussi du fait qu'il n'existe pas nécessairement *une seule* bonne façon de procéder à l'étude d'une série chronologique. Nous aurions pu aborder les modèles non linéaires, la robustesse, le contrôle, les modèles d'intervention et de fonctions de transfert, les modèles multivariés et leur comparaison avec les modèles économétriques.

I. — Approche graphique

Une *série chronologique* (encore appelée *série temporelle* ou *chronique*) est un ensemble d'observations

qui se distinguent par le rôle spécifique que joue l'ordre dans lequel elles ont été recueillies. Cette particularité est apparue très tôt, dès que la représentation graphique a pu être utilisée : on en trouve des exemples avec les problèmes de notation musicale proportionnelle au Moyen Age.

En 1936, dans un article traitant de l'image au Moyen Age et intitulé « A note on a 10th century graph », *Osiris,* vol. 1, p. 260-262, H. G. Funkhauser reproduit un diagramme présentant l'inclinaison des orbites des planètes en fonction du temps ; ce graphique (que nous reproduisons ci-après) fait partie d'un manuscrit du xe ou xie siècle et a pour but d'illustrer un commentaire du *Songe de Scipion* de Cicéron (*De Republica*, 6, 14), extrait des *Saturnales* de Macrobius (395). Il est considéré par M. G. Kendall (1973) comme le plus ancien diagramme temporel connu dans le monde occidental.



Il est assez normal de trouver dans l'astronomie la source de l'histoire des séries temporelles. L'observation des astres était fort répandue dans l'antiquité. La Grèce, notamment, a légué en ce domaine de précieux

témoignages, même si le fatal incendie de la bibliothèque d'Alexandrie, en 640, détruisit de nombreux documents essentiels.

Le témoignage mentionné ci-dessus est cependant isolé. Il faut en effet attendre le XVIIIe siècle pour voir apparaître de façon plus systématique des représentations graphiques de séries temporelles dans la littérature.

On peut tout d'abord citer le rôle du Suisse Johan Heinrich Lambert (1728-1771), d'origine alsacienne, dont nous avons déjà parlé dans le chapitre II.

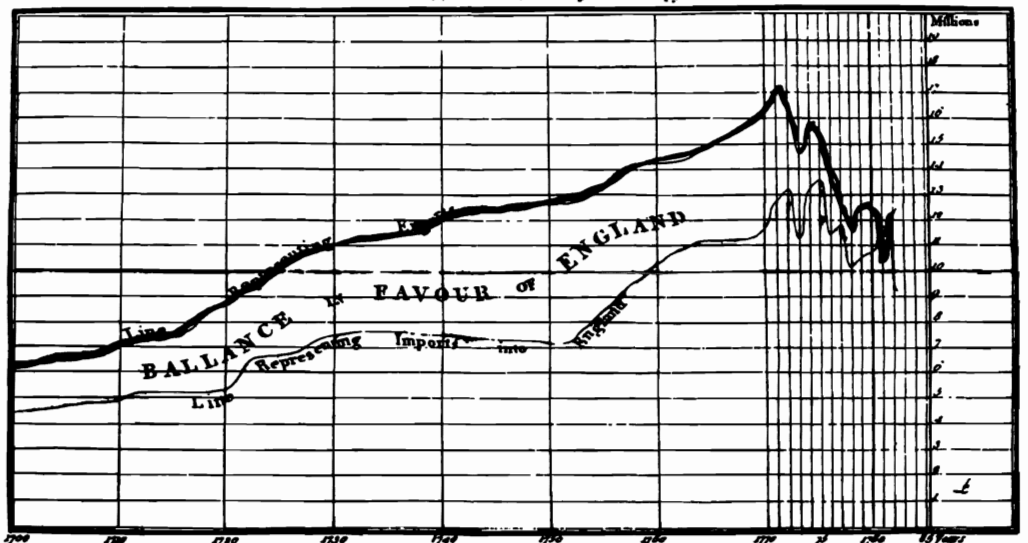
Il faut aussi insister sur l'importance du travail de l'économiste politique anglais William Playfair (1759-1823). C'est lui qui publia les premières représentations graphiques de séries chronologiques économiques dans un ouvrage intitulé *The commercial and political atlas* (London, 1786), dont le graphique ci-contre est extrait. Il traite des importations et exportations en Angleterre de 1700 à 1782.

II. — Analyse fréquentielle

Le développement des connaissances mathématiques aux XVIIIe et XIXe siècles a permis de dépasser la visualisation graphique. C'est par l'approche fréquentielle que se développèrent les premiers travaux.

L'astronomie, nous l'avons dit, a engendré des méthodes dont se sont inspirés ceux qui avaient à traiter de séries chronologiques. Il en fut de même pour la physique. Ces sciences se sont penchées notamment sur l'étude des phénomènes cycliques et la recherche des composantes périodiques des phénomènes évoluant dans le temps. Les écrits de J. L. Lagrange (1736-1813) intitulés *Recherches sur la manière de former des tables des planètes* et *Sur les interpolations*, publiés en 1772, sont cités comme une origine très probable des travaux dans ce domaine [Davis, 1941]. Ce sont cependant les résultats obtenus par J.-B. Fourier (1768-1830) qui constituent généra-

CHART of all the IMPORTS and EXPORTS to and from ENGLAND
From the Year 1700 to 1782 by W. Playfair



The Divisions at the Bottom, express YEARS, & those on the Right hand MILLIONS of POUNDS

J. Anske Sculp.t

Publish'd as the Act directs 20.th Aug.t 1785

lement la référence essentielle à la naissance de l'analyse fréquentielle (encore appelée analyse harmonique). Ce mathématicien célèbre a établi, dans sa *Théorie analytique de la chaleur* publiée en 1807, qu'une fonction $x(t)$ pouvait être décomposée en une somme de termes périodiques exprimés par des fonctions sinusoïdales et cosinusoïdales, reprenant de façon approfondie une idée émise par L. Euler (1707-1783). Ces travaux, ainsi que ceux de G. C. Stokes (« Note on searching for periodicities », *Proc. Roy. Soc.*, 1879), ont permis à A. Schuster de construire au début du xxe siècle le périodogramme qui l'a rendu célèbre, par l'intermédiaire de deux articles : « On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena », *Terrestrial Magnetism*, vol. 3, p. 13-41, 1898, et « On the periodicities of sunspots », *Philos. Trans. Roy. Soc., Serie A.*, vol. 206, p. 69-100, 1906. L'intérêt de Schuster s'est porté sur l'étude des taches solaires qui constitue un « classique » dans la littérature consacrée aux séries chronologiques. Le comportement cyclique de ces taches et leur amplitude variable représentaient à l'époque un sujet de discussion très important. Les données utilisées portaient sur la période 1750-1900. Un autre exemple tout aussi connu est le périodogramme construit par W. Beveridge, en 1922, afin d'étudier le comportement du prix du blé en Europe occidentale sur une période d'environ trois cents ans (1545-1845) : « Wheat prices and rainfall in Western Europe », *J. Roy. Stat. Soc.*, vol. 85, p. 412-459 (followed by a discussion : 460-478), 1922.

Il est intéressant de noter que la démarche de Schuster repose sur la façon de présenter une série, qui fut introduite par Buys-Ballot en 1847. Cette façon d'opérer était cependant limitée par le caractère déterministe de la démarche, comme cela fut établi ultérieurement. Le principe de travailler dans le

domaine fréquentiel en utilisant un spectre échantillon était cependant une idée très fructueuse, comme le soulignent tous les travaux qui en ont résulté.

La considération d'un périodogramme dans un contexte aléatoire est due à E. Slutsky : « Sur l'extension de la théorie des périodogrammes aux suites des quantités dépendantes », *Comptes rendus de l'Académie des Sciences de Paris*, vol. 189, p. 722-733, 1929. Ce dernier a détaillé la plupart des propriétés statistiques de cet outil. Par ailleurs, en 1930, N. Wiener proposa une procédure générale d'analyse harmonique des séries chronologiques en abordant l'étude des processus vectoriels : « Generalized harmonic analysis », *Acta. Math.*, vol. 55, p. 117-258. Cette approche allait déboucher sur l'étude des relations entre séries qui, dans le domaine fréquentiel, conduisit à l'élaboration de *périodogrammes croisés*, dont le nom fut introduit par Whittle en 1953 et, de manière plus générale, à des *périodogrammes d'ordre k* (Brillinger et Rosenblatt, 1967).

Le spectre échantillon définissant le périodogramme peut être vu comme la transformée de Fourier de l'autocorrélation (voir ci-dessous). Il fut dès lors à la base d'une série de développements placés sous le titre général d'*analyse spectrale*. On ne peut citer ici tous ceux qui ont contribué à son épanouissement. Le lecteur intéressé pourra, par exemple, consulter [Priestley, 1981].

III. — Décomposition d'une série chronologique

L'analyse harmonique avait à l'origine pour but de mettre en évidence une ou plusieurs « composante(s) » cycliques d'une série chronologique. Il existe cependant d'autre éléments qui interviennent dans l'évolution d'un phénomène. C'est ainsi que W. M. Persons a proposé, en 1919, une approche globale de décompo-

sition d'une série chronologique en termes de composantes tendancielle, cyclique, saisonnière et accidentelle. Une littérature importante a été consacrée à la détermination et (ou) à l'élimination de l'une ou l'autre d'entre elles. Dans cette voie, la *tendance* et la *composante saisonnière* ont été l'objet d'une attention toute particulière.

La détermination d'une tendance a été étroitement associée à des méthodes d'ajustement analytique de points observés par des courbes dont la plus usitée fut (et est toujours) la droite. C'est pourquoi l'analyse de cette composante a souvent été rattachée au problème de la régression (cf. chapitre II). Ainsi, avant la seconde guerre mondiale, la *Cowles Commission* fit des ajustements linéaires de séries de prix (relevés entre 1831 et 1930) à des fins de prévision. Dans la lignée, on trouve des études portant sur des ajustements qui, par transformations élémentaires des variables, se ramènent au cas linéaire.

Par ailleurs, on vit se développer des recherches de tendance non linéaire. Un domaine particulièrement attractif concerne les *courbes de croissance* dont les plus célèbres sont l'*exponentielle modifiée*, la *courbe de Gompertz* et la *logistique*.

Benjamin Gompertz établit sa courbe en 1825 afin de graduer des tables de mortalité (*Philosoph. Trans. Roy. Soc., Serie A*, vol. 115, p. 513-580). La courbe logistique, quant à elle, a été introduite en 1844 par P. F. Verhulst dans les *Mémoires de l'Académie royale de Bruxelles*, vol. 18, p. 1. Son usage fut important, en particulier dans les études touchant à la démographie et à l'épidémiologie.

Une autre approche de la tendance concerne l'usage des *moyennes mobiles* consistant à remplacer chaque observation par la moyenne arithmétique de valeurs qui l'entourent. Si $\{y_1, y_2, ..., y_n\}$ est une série d'observations prélevées à des dates équidistantes, on définit une valeur lissée

$$y_t^* = \sum_{j=-p}^{q} w_j y_{t+j}$$

où p et q sont des entiers non négatifs et $\{w_j\}$ un ensemble de coefficients de pondération dont la somme vaut 1. Cet outil fut fort utilisé pour aborder le problème de la tendance dans la mesure où cette technique permet « d'éliminer » des variations accidentelles ou périodiques.

L'usage des moyennes mobiles a été développé dans deux directions. La première concerne le problème de la graduation en actuariat. La seconde traite de l'étude des séries économiques. De nombreuses expressions ont été développées dans des contextes très précis et portent le nom de ceux qui les ont introduites (Spencer, Henderson...). Le recours à des moyennes mobiles s'est cependant particulièrement développé dans le cadre des méthodes de *désaisonnalisation*. Plus précisément, la séparation des fluctuations saisonnières des autres composantes a été étudiée à l'origine par M. T. Copeland en 1915 (« Statistical indices of business conditions », *Quart. J. of Economics*, vol. 29, p. 522-562) et W. M. Persons en 1919 (« An index of general business conditions », *Review of Economics and Statistics*, vol. 1, p. 111-205), et reprise par F. R. Macauley en 1930 (« The smoothing of time series », *Nat. Bur. of Econ. Res.*, p. 121-136). Ce dernier introduisit la méthode des rapports à la moyenne mobile qui connaît encore aujourd'hui un succès relativement important. Cette procédure est à la base d'une technique de désaisonnalisation utilisée par de nombreux offices statistiques gouvernementaux : la méthode Census dont la première version date de 1954. Elle fut modifiée à de nombreuses reprises et est connue actuellement sous la version X-11.

Il faut noter que ces différentes approches se sont développées dans un contexte déterministe. Elles ont bien sûr évolué en y intégrant généralement un

caractère stochastique par l'adjonction de perturbations aléatoires.

IV. — Approche temporelle

L'approche temporelle est liée à l'usage du corrélogramme et, plus généralement au concept de corrélation sérielle. On attribue généralement les premières références à cette notion à un article du physicien anglais J. H. Poynting qui, en 1884, s'interrogeait sur les relations existant entre d'une part les mouvements du prix du blé en Angleterre, en France et au Bengale et, d'autre part, de l'importation du coton et de la soie en Grande-Bretagne (*J. Roy. Stat. Soc.*, vol. 47, p. 34-64). Poynting n'a pas introduit explicitement le coefficient de corrélation sérielle mais il a attiré l'attention sur la nécessité d'utiliser ce concept. Le premier usage de ce coefficient est à porter au crédit, semble-t-il, de R. H. Hooker qui, en 1901, étudia l'association existant entre le taux de mariage en Angleterre et l'indice du commerce (« Correlation of the marriage-rate with trade », *J. Roy. Stat. Soc.*, 64, p. 485-492). Si $\{x_1, x_2, ..., x_n\}$ et $\{y_1, y_2, ..., y_n\}$ sont deux séries de valeurs relevées aux mêmes dates d'observation (supposées équidistantes et représentées par les n premiers nombres naturels), ce coefficient peut être défini par $r_k = c_k/c_0$ où la covariance (sérielle) c_k est fournie par l'expression

$$c_k = \frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x})\ (y_{t+k} - \bar{y}),$$

dans laquelle $\bar{x}$ et $\bar{y}$ sont les moyennes des valeurs $\{x_t\}$ et $\{y_{t+k}\}$ utilisées dans cette expression. Lorsque les séries $\{x_t\}$ et $\{y_t\}$ sont confondues, on parle de coefficient d'autocorrélation. L'utilisation de ce dernier et son histoire sont fortement liés à la méthode des différences dont les travaux les plus importants sont

dus à O. Anderson [1929]. Cette méthode reposait sur l'hypothèse que tout élément d'une série chronologique pouvait se décomposer en deux parties ; la première était structurelle, la seconde accidentelle :

$$y_t = f_t + e_t.$$

En calculant les différences successives $\nabla y_t = y_t - y_{t-1}$, $\nabla^2 y_t = \nabla (\nabla y_t)$, ... on a constaté que la variance des séries correspondantes avait tendance à décroître jusqu'à une limite représentant la variance de la partie accidentelle. C'est ainsi que Gosset (*alias* Student) écrivit dans un article publié en 1914 (« The elimination fo spurious correlation due to position in time or space », *Biometrika*, 10, p. 179-180) :

« ... if we wish to eliminate variability due to position in time or space and to determine whether there is any correlation between the residual variations, all that has to be done is to correlate the 1st, 2nd, 3rd, ..., *n*th differences between successive values of the other variable. When the correlation between the two n th differences is equal to that between the two $(n + 1)$th differences, this value gives the correlation required. »

On ne peut cependant parler de la corrélation sérielle sans mentionner les résultats obtenus par Ernst Abbe dans ce domaine. Ce spécialiste de l'optique, né en 1840 et décédé en 1905, publia quelques recherches d'ordre statistique. Si le résultat le plus important concerne la découverte de la variable $\varkappa^2$ par l'étude de $\Delta = \sum_{i=1}^{n} Z_i^2$, *où* Z_i $(i = 1, 2, ..., n)$ sont n variables N (0, 1) indépendantes, Abbe étudia également la variable :

$$\theta = \sum_{j=1}^{n} (Z_j - Z_{j+1})^2$$

où $Z_{n+1} = Z_1$ et, ensuite le rapport θ / Δ, obtenant la loi du coefficient de corrélation sérielle circulaire dont les propriétés ont été retrouveés par R. L. Anderson en 1942.

V. — Modèles stochastiques

L'approche temporelle, comme les travaux dans le domaine fréquentiel, se sont développés à l'origine dans un contexte déterministe. La prise en compte d'un point de vue aléatoire allait permettre à l'analyse des séries chronologiques de prendre un nouvel essor par la mise au point d'instruments originaux, et de toucher des domaines d'application autres que l'astronomie, l'économie ou la démographie. L'article de référence en la matière est dû à G. U. Yule et date de 1927 : « On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers », *Royal Society of London, Philosophical Transactions*, Series A 226, p. 267-298. L'idée de l'auteur était de décrire la série $\{y_t\}$ du nombre de taches solaires (mentionné plus haut) par *un modèle autorégressif d'ordre 2* qui s'exprimait, en ne considérant que les écarts à la moyenne, sous la forme :

$$Y_t = \alpha\, Y_{t-1} + \beta\, Y_{t-2} + \varepsilon_t$$

où α et β sont des paramètres réels et $\{\varepsilon_t\}$ une suite de variables aléatoires non corrélées et identiquement distribuées, appelée *bruit blanc*.

A titre purement anecdotique, on rapporte que les travaux de Yule sont la conséquence d'une réflexion née de l'observation du mouvement oscillatoire d'un pendule bombardé de petits pois lancés par un enfant. Yule en déduit que l'impact des chocs passés est « intégré » dans le mouvement ultérieur du pendule. Cette idée est à la base de l'approche temporelle dans l'étude des processus aléatoires, dont l'analyse des séries chronologiques va, dès lors, largement bénéficier.

Les travaux de Yule allaient trouver une aide précieuse dans le théorème ergodique de Birkhoff (1931) et Khintchin (1932) ainsi que dans le théorème de Wold qui, en 1938, établit que la partie non

déterministe X_t d'un processus stationnaire pouvait être représenté comme une « moyenne mobile » de bruits blancs. Ces travaux sont à l'origine des modèles autorégressifs (AR), moyenne mobile (MA) et mixtes (ARMA) qui ont conduit G. E. P. Box et G. M. Jenkins à proposer, en 1970, une méthodologie de modélisation des séries chronologiques à des fins de description, de prévision ou de contrôle (*Time Series Analysis : forecasting and control*, San Francisco, Holden Day). Il existe de nombreuses références en ce domaine (dont celle des inventeurs de cette méthode). Pour la littérature d'expression française, nous renvoyons le lecteur à la bibliographie.

VI. — La prévision

Le problème de la prévision constitue indéniablement un des objectifs les plus attractifs de l'analyse des séries chronologiques. Il se pose comme suit dans le cas univarié : sur la base d'une série $(y_1, ..., y_T)$, on désire obtenir une estimation $\hat{y}_T(h)$ de la valeur de y_{T+h}, dite prévision faite à l'instant T pour un horizon h. Ce problème a reçu une attention particulière en dehors de la modélisation ARMA. C'est ainsi que l'usage des modèles linéaires a conduit à des méthodes de prévision par régression selon une procédure d'extrapolation accompagnée de zone de confiance pour la valeur prédite. D'un point de vue historique, il est intéressant de mentionner aussi la classe des méthodes de *lissage exponentiel* dont le cas le plus simple est décrit par la formule de récurrence

$$\hat{y}_T(1) = \alpha y_T + (1 - \alpha)\hat{y}_{T-1}(1)$$

où α est un paramètre compris entre 0 et 1, choisi généralement de façon à minimiser la somme des carrés des erreurs de prévision. Ces méthodes ont été principalement abordées par Brown en 1962 (*Smoothing, forecasting and prediction*, Englewood Cliffs,

New Jersey, Prentice Hall). Elles ont été généralisées par Holt et Winters.

Une dernière référence en ce domaine concerne une procédure qui trouve son origine dans les travaux du danois Thoward Nicolai Thiele (1838-1910). Ce dernier, désirant traiter de séries chronologiques dans un problème de géodésie, proposa un modèle composé d'un terme de régression, d'une composante du type bruit blanc et d'une autre représentant un mouvement brownien. La recherche de valeurs prédites pour ce dernier le conduisit à proposer une méthode récursive de calcul qui annonçait celle qui porte actuellement l'appellation de *filtre de Kalman*.

CHAPITRE VIII

L'ANALYSE DES DONNÉES

I. — Introduction

« Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (...) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques. »

C'est ainsi que Jean-Paul Benzécri commence le tome 2 de son ouvrage fondamental *L'Analyse des Données* (1973), tome consacré à l'analyse des correspondances. Mots forts traduisant un conflit d'écoles qui s'est exaspéré beaucoup plus en France que dans le monde anglo-saxon. Depuis, les passions se sont apaisées, chacun ou presque reconnaissant les mérites des deux approches somme toute complémentaires.

L'analyse des données n'a pas été, loin s'en faut, le seul domaine de débat en statistique. Ainsi, K. Pearson et R. Fisher, puis R. Fisher et J. Neyman se sont rarement ménagés (cf. chapitres V et IX) ; que dire également du conflit entre bayésiens et non-bayésiens, ou de la querelle opposant écoles anglaise et américaine pour l'insertion de la statistique dans la théorie de la décision ? Toute confrontation d'idées est positive, et l'analyse des données a répondu à un besoin. Avec son approche plus inductive que déductive, cherchant *a posteriori* une structure dans l'amas des données, elle réhabilite l'individu-observation. Par sa nature descriptive, elle fait appel aux représentations graphiques et à la géométrie. N'oublions pas, d'ail-

leurs, que le sens géométrique domine l'œuvre de Fisher, et, plus avant, celle de Laplace, Gauss et Legendre.

Comme dans les autres chapitres, il est hors de question de prétendre à l'exhaustivité. Aussi, nous insisterons plus sur le cheminement des idées et des résultats dans différentes branches de l'analyse des données que sur les apports récents. A cet égard, le début des années 1970 sera un seuil rarement franchi.

II. — La psychométrie

A partir de la fin du XIXe siècle, la loi normale a pris une forme multidimensionnelle avec, dans un premier temps, F. Galton et K. Pearson en dimension 2, puis au premier quart du XXe siècle, dans le cas général avec R. Fisher. Les premiers en font grand usage pour l'étude de comportements humains : anthropométrie, génétique, biométrie. Souvenons-nous, en effet, que F. Galton est le cousin de Charles Darwin.

Cette loi normale va trouver un nouveau champ d'application avec la naissance de la psychométrie ou analyse factorielle, école essentiellement américaine qui a pour objet la mesure statistique de l'intellect. Les études font d'abord référence à des grandeurs physiques. Dès 1764, Lambert avait remarqué que le seuil de perception était proportionnel à la grandeur de l'objet. Très vite, on passe aux notes élaborées à partir de tests. Si l'on peut citer Cattel, élève de Galton, mesurant en 1890 des aptitudes mentales pour prédire la réussite scolaire, c'est généralement à Charles Spearman que l'on attribue le premier article d'analyse factorielle, dont le titre peut surprendre par son ambition : « General intelligence objectively determined and measured », *Amer. Jour. Psychol.*, 1904. Plus tard, vers 1930, Louis Leon Thurstone reprendra les travaux de Spearman pour les développer et y introduire le calcul matriciel naissant.

Thurstone postule un modèle linéaire à partir des notes centrées réduites ; dans ce modèle, une note est, à l'erreur gaussienne près, une combinaison linéaire de facteurs communs (un seul chez Spearman) et d'un facteur dit spécifique. On conçoit qu'un tel modèle oblige Thurstone à distinguer partie commune et partie spécifique, ce qui l'amène à traiter la matrice de corrélation des notes avec diagonale modifiée par la « *communality* ». C'est à cette époque que Harold Hotelling, pour un tout autre problème, met au point une méthode de diagonalisation par itérations successives, méthode qui sera fort utile pour l'analyse factorielle.

Bien d'autres psychométriciens suivront Thurstone, discutant ici la « *communality* », là les axes obliques. Il convient de s'attarder quelque peu sur les travaux d'un grand psychométricien décédé en 1988, Louis Guttman. Travaillant sur des données qualitatives et les réponses 1 ou 0 des individus aux modalités, Guttman arrive aux formules de l'analyse des correspondances (« The quantification of a class of attributes », P. Horst éd., *SSRC*, 1941, New York). Cependant, ne possédant pas de moyens de calculs, Guttman s'ingénie à introduire des modèles. Présupposant un ordre sur les données, il parvient à son célèbre *scalogramme* où, par des permutations successives, doivent apparaître des parallélogrammes sur la structure diagonale du tableau, générateurs de facteurs suivant un effet polynomial connu sous le nom d' « *effet Guttman* ». Bien d'autres modèles suivront, ainsi que d'autres psychométriciens dont Cyril Burt, continuateur de l'école anglaise, et W. S. Torgerson avec son ouvrage *Theory and Methods of Scaling* (Wiley, 1958).

III. — L'analyse multivariée

Le début de ce siècle voit naître un nombre considérable de résultats relatifs au modèle multinormal. Ils furent l'œuvre de Sir Ronald Fisher bien sûr, mais aussi de Harold Hotelling et de l'école indienne centrée autour de P. C. Mahalanobis avec R. C. Bose, auteur d'un remarquable historique *Early history of multivariate statistical analysis* (in *Multivariate Analysis IV*, Krishnaiah éd., North Holland, 1977, p. 3-22) et S. N. Roy. Parmi ces résultats, trois s'avèrent cruciaux à l'égard de l'analyse des données.

L'analyse en composantes principales (ACP) apparaît en 1901 avec K. Pearson et est située dans un cadre multivarié par Hotelling en 1933, qui fut, à l'en croire, précédé par Thurstone en 1931. Hotelling résoud le problème suivant : étant donné n variables normales $x_1, \ldots, x_n$ centrées réduites, trouver la combinaison linéaire normée $z = a_1 x_1 + \ldots + a_n x_n$ ayant une variance maximum. Puis il itère le procédé pour une combinaison linéaire indépendante de z de variance maximum. La loi asymptotique des cœfficients sur échantillonnage est établie par Girshick en 1939.

Dès 1913 (« Correlation of sums and differences », *Br. Jour. Psychol.*, 5, p. 417-423), Spearman définit le problème de la corrélation canonique pour deux ensembles de lois normales $x_1, \ldots, x_n$ et $y_1, \ldots, y_p$: trouver deux combinaisons linéaires $z = a_1 x_1 + \ldots + a_n x_n$ et $t = b_1 y_1 + \ldots + b_p y_p$ ayant une corrélation maximum. Mais la solution date de 1936 avec Hotelling qui donne en plus une interprétation géométrique ainsi que la loi conjointe de z et t pour $n = p = 2$. Ces résultats seront étendus par Girshick en 1939.

Enfin, dans l'analyse discriminante, Fisher (1936) recherche une combinaison linéaire $z = a_1 x_1 + \ldots + a_n x_n$ mesurée sur deux populations qui, dans un but d'affectation, maximise le quotient de la variance interclasse par la variance intraclasse.

Ces résultats sont précédés, accompagnés et suivis de bien d'autres. A défaut de ne pouvoir tout citer, il faut néanmoins mentionner les apports de Fisher. Nous lui devons l'utilisation du modèle linéaire qui, de la régression, conduira à l'analyse de variance et aux plans d'expérience. Il établit correctement de nombreuses propriétés : loi du cœfficient de corrélation, démonstration rigoureuse de la loi de la statistique de Student généralisée au cas multivarié par le T^2 de Hotelling (1931), étude de la corrélation multiple (1928), aidé en cela par Wishart (1928) qui obtient la loi de la matrice de covariance empirique et par Wilks (1932), auteur de la variance généralisée. On lui doit également le test de comparaison des variances (1926) ainsi que le lien entre l'analyse discriminante et le D^2 de Mahalanobis (1928) dont la loi sera établie par Bose en 1936. Enfin, si Roy et Hsu obtiennent en 1939 la loi des valeurs propres de la matrice de covariance empirique, leurs premières équations sont établies par Fisher la même année.

IV. — L'analyse multidimensionnelle

Il convient ici de distinguer ce que les anglo-saxons nomment *multivariate analysis* (voir § III) et *multidimensional scaling* (§ IV). L'analyse multivariée fait entièrement référence au modèle probabiliste et plus précisément multinormal. Dans ce modèle, les psychométriciens ont largement puisé ou apporté, ce que confirme Cyril Burt dans des propos tenus lors d'un colloque CNRS en 1955 et rapportés par Benzécri dans son livre *Histoire et préhistoire de l'analyse des données* :

« Les analogies entre l'analyse factorielle et l'analyse de la variance sont évidentes. »

Le *multidimensional scaling* est plus proche des principes de l'analyse des données évoqués dans

l'introduction. Oubliant (dans un premier temps ?) le modèle probabiliste, on construit sur la base des observations une approche plus géométrique et descriptive. Nombre de méthodes de l'analyse multivariée, comme les trois citées au paragraphe précédent, peuvent être présentées dans cette optique. Ainsi l'ACP peut être formulée à partir de l'échantillon comme la recherche de l'axe ou du sous-espace de plus grand allongement du nuage des observations dans l'espace des variables. C'est selon cette optique géométrique qu'elle fut introduite par K. Pearson en 1901, reprise par Frisch en 1929 avant que Hotelling ne lui donne un aspect multivarié. Cette approche multidimensionnelle de l'ACP est confirmée par M. G. Kendall dans *A course in multivariate analysis*, Londres, Griffin, 1957 :

« Dans l'analyse en composantes, on part des observations et l'on cherche des composantes on va des données vers un modèle hypothétique. En analyse factorielle, c'est l'inverse. »

Depuis, revenue à ses origines, l'ACP a fait l'objet de nombreuses études. Nommons C. R. Rao, J. Gower héritier de Rothamsted Station, et L. R. Tucker jetant entre 1963 et 1966 les bases d'une extension à un cube de données.

Dès lors que l'on oublie le modèle multinormal, l'analyse canonique peut être appliquée au cas discontinu. Deux variables qualitatives sont considérées comme deux ensembles de variables, chacun étant défini par les modalités codées 0 ou 1. Les combinaisons linéaires données par les corrélations canoniques et calculées par Maung en 1941 ne sont que les facteurs de l'analyse des correspondances. Dans cette voie, plus que la corrélation, ce sont les combinaisons linéaires, sources de réduction de dimension, qui importent. Depuis, l'analyse canonique a été étendue à plusieurs ensembles de variables (P. Horst en 1961, J. D. Carroll en 1968, J. R. Kettenring en 1971). Le

modèle multinormal peut être également abandonné en analyse discriminante. R. Fisher ne s'en privera pas dès 1940 pour appliquer celle-ci à des variables qualitatives. Il lui faudra alors coder ces variables et les équations obtenues seront reprises par Maung en 1941 pour établir les corrélations canoniques.

V. — L'analyse des correspondances

L'analyse (factorielle) des correspondances, l'AFC, a été découverte et redécouverte par de nombreux auteurs durant plusieurs décennies, au gré des investigations et des techniques utilisées. Elle est ainsi appelée méthode des poids de Guttman, ACP sur données qualitatives, théorie de la quantification d'Hayashi, régression linéaire simultanée, biplot, dual scaling, reciprocal averages, additive scoring, canonical scoring, etc.

Les approches de Guttman, Fisher et Maung ont déjà été mentionnées. Attardons-nous sur celle de J.-P. Benzécri en puisant dans son livre *Histoire de préhistoire de l'analyse des données*. L'auteur fait remonter à 1962 le nom d'analyse des correspondances et ses premiers travaux. Partant d'un tableau de contingence (f_{ij}) et des marges $f_{i.}$ et $f_{.j}$, Benzécri compare deux modalités lignes i et i' à travers les distributions conditionnelles $f_{ij}/f_{i.}$ et $f_{i'j}/f_{i'.}$, et propose une distance quadratique :

$$d^2(i, i') = \sum_j (1/f_{.j}) [f_{ij}/f_{i.} - f_{i'j}/f_{i'.}]^2 .$$

Il nomme cette distance *distance distributionnelle*, nom qui deviendra vite, pour des raisons évidentes, *distance du* χ^2. On est alors dans le schéma classique de l'ACP du nuage des lignes placé dans l'espace des colonnes pour réduire la dimensionnalité. Doublant cette analyse d'une analyse duale sur les colonnes, Benzécri et Brigitte Escofier traitent différents exem-

ples, notamment de linguistique, sans qu'aucun lien n'existe entre l'analyse des lignes et celle des colonnes. Ce n'est qu'ultérieurement, avec l'usage de l'ordinateur, que ces liens, non évidents dans cette approche, apparaîtront numériquement et seront démontrés par B. Escofier. Ils feront l'objet de sa thèse soutenue en 1965 avec notamment la célèbre *formule barycentrique* ou *formule de transition* :

$$\alpha_{ih} = (1/\sqrt{\lambda_h}) \sum_j (f_{ij}/f_{i.})\ \beta_{jh}\,.$$

Sur le h-ième axe principal, la composante principale α_{ih} est, à un cœfficient près, la moyenne pondérée des composantes des colonnes et réciproquement. Ce sont les *reciprocal averages* établies par Richardson et Kuder en 1933, Horst en 1935, Hirschfeld en 1935. A partir de cœfficients β quelconques, on obtient des cœfficients α par des moyennes pondérées, puis de nouveaux β. En itérant le processus, il y a convergence à un facteur de normalisation près vers un vecteur propre d'une matrice de type AA'. C'est par une démarche semblable que Richardson, Kuder, Horst et Hirschfeld découvrent la méthode des reciprocal averages. Cette démarche est également proche de celle d'Eckert et Young qui établissent en 1936, non sans erreur, la décomposition d'une matrice rectangulaire à laquelle les tenants de l'analyse factorielle de l'époque et de nos jours font référence. Cette décomposition n'est autre que celle en valeurs singulières établie par Schmidt en 1907. Néanmoins, les équations des composantes viennent ainsi naturellement mais elles restent injustifiées quant aux critères d'optimalité de l'ACP.

Tout ceci montre à l'évidence combien de personnes ont travaillé indépendamment et selon des optiques différentes. Encore faudrait-il parler d'Hayashi (1952) étudiant et étendant la quantification de Guttman, ou Lancaster (1958)... La paternité de l'AFC

est donc bien difficile à établir tant les démarches ont été incomplètes et parallèles. Ainsi Benzécri écrit, en faisant référence aux travaux de Guttman, Fisher et Maung, Hayashi :

« La priorité de ces auteurs est certaine : la seule originalité que puissent revendiquer les chercheurs français est d'avoir conjugué, avec une méthode découverte indépendamment par plusieurs auteurs, des idées et des problèmes multiples dont la synthèse n'était pas faite. »

VI. — L'approche non métrique

Toutes les méthodes d'analyse des données évoquées dans les paragraphes précédents reposent sur la géométrie euclidienne. Qu'il soit celui des variables, celui des modalités des variables qualitatives ou des individus, l'espace vectoriel est toujours sous-jacent et normer des variables ou introduire des poids revient géométriquement à choisir un produit scalaire. Si Benzécri, dans son approche de l'analyse des correspondances, choisit en premier lieu une distance plutôt qu'un espace, il prend garde à la définir par une forme quadratique adaptée à la géométrie euclidienne.

Cependant, pour certains champs d'application et certains types de données, des spécialistes ont proposé des indices non euclidiens et même non métriques. C'est, en particulier, le cas des données sur signes de présence/absence où un grand nombre d'indices a vu le jour. Dans une étude comparative, Zdenek Hubalek en recense 43 (« Cœfficients of association and similarity based on binary — presence/absence — data : an evaluation », *Biol. Rev.*, 1982, 57, p. 669-689). Ils ont été introduits dans des domaines aussi divers que l'écologie (Jaccard, 1901), la taxinomie (Sneath, 1957), la psychologie, la biologie et certains apparurent indépendamment, tel l'indice d'Ochiai (1957) qui n'est autre que l'indice de Driver et Krœber créé en 1932 pour l'ethnologie.

Dans l'approche géométrique de l'analyse des données, il est important alors de savoir si l'on peut construire une figure euclidienne respectant les distances choisies, ou, en termes plus mathématiques, si un espace métrique fini est isométriquement plongeable dans un espace euclidien. On peut attribuer les premiers travaux dans cette voie à Minkovski, Dirichlet, Hermite et Gauss, mais c'est avec M. Fréchet (1935), I. J. Schœnberg (1935, 1937), G. Young et A. Householder (1938) que sont établies les conditions d'immersion et les transformations des distances euclidiennes. Ces résultats seront étendus aux espaces L en 1966 par J. Bretagnolle, D. Dacunha-Castelle et J. L. Krivine. La première approximation euclidienne, dite de constante additive, est obtenue par J.C. Lingœs en 1971, offrant ainsi une représentation euclidienne multidimensionnelle qui ouvre la voie aux techniques de l'ACP.

L'analyse des proximités est aussi par essence de nature non métrique. Seules y interviennent les inégalités entre les diverses distances. On doit à J. B. Kruskal (1964) l'approximation au sens des moindres carrés respectant un système donné d'inégalités et à R. N. Shepard (1962) un algorithme d'obtention d'une figure de faible dimension satisfaisant un tel système. D'autres algorithmes ont suivi où sont souvent associées contraintes de dimensionnalité et contraintes d'inégalité. Ce sont souvent des algorithmes alternés où, pour la dimension, on minimise par des méthodes de gradient un critère d'adéquation. Parmi eux, citons les algorithmes de J. B. Kruskal (1964), V. E. Mac Gee (1966), L. Guttman et J.C. Lingœs (1967).

VII. — La classification hiérarchique

Le terme classification recouvre bien des aspects : partitionnement d'un ensemble d'objets avec ou non des classes empiétantes, affectation d'objets à des

groupes pré-définis, sériation ou classement, taxinomie et décomposition d'une population en arbre. Dans l'impossibilité d'en tracer l'histoire en quelques lignes, nous nous concentrerons sur la représentation sous forme d'arbre qui semble la plus proche des méthodes évoquées dans les paragraphes précédents. En outre, n'est-ce pas déjà sous forme d'arbre que les naturalistes du XVIIIe siècle ont cherché à classer les espèces ?

L'arbre le plus usuel est l'arbre hiérarchique ou *dendogramme*, associé à des partitions emboîtées. Une procédure classique de construction est l'algorithme ascendant où les classes sont progressivement regroupées. Son principe est déjà présent chez M. Adanson en 1757 (*Histoire naturelle du Sénégal*, Paris, Bauche). Partant d'un indice de distance entre objets, on définit un indice d'agrégation entre classes. Dès lors, il suffit de regrouper les deux objets les plus proches puis d'itérer la procédure avec les classes. Il y a donc autant d'algorithmes que d'indices d'agrégation choisis. Parmi ceux-ci, il faut mentionner le simple lien (distance minimum entre objets), le lien complet (distance maximum) introduit par Sorenson (1948), l'indice de la moyenne dû à Sokal et Michener (1958), l'indice de Ward fondé sur la variance (1963). Lance et Williams en donnent une forme synthétique en 1967.

Dans le dendogramme, la distance entre objets est définie par la hauteur de la plus petite classe les contenant. Celle-ci vérifie une inégalité particulière dite ultramétrique. On attribue généralement à Johnson (1967) le théorème d'équivalence entre ultramétriques et hiérarchies. Pourtant dès 1965, J.-P. Benzécri l'enseigne à l'Université de Rennes, définissant en outre une hiérarchie en termes ensemblistes ; en 1966, J. D. Apresjan propose une hiérarchie particulière et, la même année, N. Jardine et R. Sibson travaillent sur ce sujet. Au cours de la même période (1967), Johnson, Jardine et Sibson étudient les ultra-

métriques majorant ou minorant une distance donnée, trouvent des éléments minimaux dans le premier cas et un élément maximum, la sous-dominante, dans le second. Ils montrent que ces approximations sont construites par les algorithmes du lien complet et du simple lien. Toujours à la même époque, Jardine et Sibson établissent que la sous-dominante est l'ultramétrique solution d'un problème min-max dans le graphe complet du tableau de distance original. Enfin, J. Gower et G. Ross (1969) montrent la liaison existant avec l'arbre de longueur minimum. En quelques années, toutes les bases de la classification hiérarchique étaient jetées. L'extension à des arbres non hiérarchiques suivra après 1970.

Chapitre IX

QUELQUES GRANDS NOMS

Comme nous l'avons précisé dans le premier chapitre, nous avons voulu accorder une attention particulière à quelques personnages qui, pour des raisons multiples, ont influencé la méthodologie statistique. On aurait pu en citer d'autres et nous ne prétendons pas avoir retenu les seuls grands noms possibles. Choisir n'est pas exclure ! Pour des raisons de facilité, nous avons préféré les présenter par ordre alphabétique.

I. — Ronald Aylmer Fisher

Ronald Aylmer Fisher est né le 17 février 1890, à East Finchley (Londres), le plus jeune d'une famille de huit enfants. Dès sa plus petite enfance, il se révèle très brillant et précoce dans le raisonnement mathématique — en particulier en géométrie — aidé peut-être en cela par une très mauvaise vue, dont il souffrira toute sa vie. En 1909, il obtient une bourse d'études par le Gonville and Caius College de Cambridge, dont il sort diplômé en 1913 après avoir étudié les mathématiques, la loi des erreurs, la mécanique et la théorie des quanta. C'est pendant son séjour à Cambridge qu'il se passionne également pour la biométrie, la génétique et l'eugénisme, suite à la vulgarisation des travaux de Mendel (1865) ainsi qu'aux articles de Galton sur l'hérédité et de K. Pearson sur les mathématiques de l'évolution.

Après avoir publié, en 1912, son premier article, intitulé « On an absolute criterion for fitting frequency curves » dans lequel il utilise — sans la nommer ainsi — la méthode du maximum de vraisemblance, Fisher trouve un emploi statistique chez Mercantile and General Investment Company, à Londres, qu'il occupera jusqu'en 1915, date de la parution, dans *Biometrika*, de « Frequency distribution of the values of the correlation cœfficient in samples from an indefinitely large population ». Ce papier le rend célèbre dans le monde statistique, et lui permet de contacter Gosset, avec lequel il correspondra régulièrement jusqu'à la mort de celui-ci en 1937.

Après quatre années passées à enseigner, car dispensé du service militaire à cause de sa vue, un poste lui est proposé à la Station expérimentale de Rothamsted. Il l'accepte en octobre 1919, de préférence à la charge que K. Pearson souhaitait lui voir prendre au Laboratoire Galton de l'University College, non seulement pour l'intérêt qu'il perçoit à mener des recherches en agriculture et en agronomie mais aussi en raison d'une animosité envers Pearson. Il y reste jusqu'en 1933, année où Karl Pearson part à la retraite. Le département de ce dernier à l'University College est alors scindé en un département d'Eugénisme, que Fisher (élu en 1929 « Fellow de la Royal Society ») prend en charge, et en un département de Statistique appliquée qui sera confié à Egon Pearson en 1935. Son dernier poste sera, à partir de 1943, la chaire de Génétique à Cambridge, jusqu'à sa retraite en 1957. Il s'installe alors, à la surprise générale, à Adélaïde en Australie, où il meurt le 29 juillet 1962.

La contribution de Ronald Fisher à la science statistique est énorme, non seulement par ses 395 publications, mais surtout parce qu'il a fondé les concepts de référence de la statistique mathématique moderne. Il serait vain de détailler l'ensemble de ses

travaux; nous nous contenterons d'examiner la période de son séjour à Rothamsted.

Nous avons vu que, dès 1912, Fisher utilise le maximum de vraisemblance. La même année, il écrit à Gosset pour lui proposer une démonstration géométrique de la « loi de Student », en représentant l'échantillon dans $\mathbf{R}^n$, et pour introduire la notion de degré de liberté. L'article de 1915 porte en lui les germes de sa brouille avec Pearson. En effet, ce dernier publie en 1917 une critique des idées de Fisher sur le maximum de vraisemblance sans l'en avoir informé (cf. § III du chap. V). Pearson réfute également la proposition de Fisher de travailler non pas sur le cœfficient de corrélation r mais sur la variable $z = (1/2)\ \mathrm{Log}\ [(1 + r) / (1 - r)]$, transformation née du génie géométrique de Fisher. En 1918, Fisher publie « The correlation between relatives on the supposition of Mendelian inheritance », article refusé en 1916 par la Royal Society et dans lequel est présentée l'analyse de la variance. Attribuant ce refus à l'intervention de Pearson, la rupture est consommée essentiellement, semble-t-il, du fait de Fisher qui, tout au long de sa vie, ne supportera pas la critique.

Les années passées à Rothamsted voient Fisher définir et utiliser les principes de la randomisation, des plans d'expérience, des carrés latins, qui seront repris dans un ouvrage édité en 1935 et intitulé *The design of experiments*. En 1925 puis 1928 (« Applications of Student's distribution », « The general sampling distribution of the multiple correlation cœfficient ») il énonce les lois de probabilité des statistiques définies dans le cadre du modèle linéaire général avec aléa gaussien. Dans la même période, Fisher développe la théorie de l'estimation (1922 : « On the mathematical foundations of theoretical statistics » ; 1925 : « Theory of statistical estimation »). Il établit clairement la distinction entre les statistiques, estimateurs calculés sur un échantillon, et les paramètres à estimer définis

sur la population, ainsi que la théorie générale de l'estimation selon le principe du maximum de vraisemblance ; travaillant sur l'écart type d'une loi N (m, σ), il définit la propriété d'*exhaustivité*. Il introduit une quantité d'information contenue dans un échantillon, ainsi que les concepts de convergence et d'efficacité. Dès 1922, il établit l'efficacité asymptotique des estimateurs du maximum de vraisemblance. L'ensemble de ses travaux sur la théorie de l'estimation est publié en 1925 dans *Statistical Methods for Research Workers*, que l'on peut appeler sans risque d'être contredit, le best-seller de la statistique : 14 éditions, traductions en 6 langues.

Au cours de son séjour à Rothamsted, Fisher s'intéresse à l'analyse harmonique, aux méthodes fiduciaires, à l'analyse multivariée, à l'analyse de la variance et de la covariance, aux fonctions pivotales et aux lois des valeurs extrêmes (article écrit en 1928 avec L. Tippett : « Limiting forms of the frequency distribution of the largest or smallest member of a sample ») ; en 1929, reprenant le vieux problème du calcul des « moments des moments », il conçoit — au cours d'un voyage en train — le concept de k-statistique, fonction de l'échantillon $k_p = k\,(x_1, ..., x_n)$ telle que $E\,(k_p) = K_p$, cumulant d'ordre p sur la population.

L'année 1930 voit paraître un livre sur la passion initiale de Fisher, *The Genetical Theory of Natural Selection*.

La même année débute la querelle entre Fisher et Jerzy Neyman, qui n'est pas d'accord avec la statistique fiduciaire de Fisher et n'a pas la même conception philosophique de la statistique que lui. Ce dernier reproche alors à Neyman de lui avoir volé ses idées pour établir sa théorie de l'estimation par intervalle de confiance, et de critiquer l'existence d'un biais dans la statistique z en l'utilisant sous une hypothèse nulle différente de celle sous laquelle Fisher l'a conçue. Sans

porter de jugement, on ne peut que constater *a posteriori* que l'harmonie ne régna jamais entre Fisher, Neyman et K. Pearson.

Le départ de Fisher de Rothamsted ne marque pas la fin de ses travaux. Aux 97 articles publiés entre 1919 et 1933 il convient d'ajouter les 91 publications entre 1934 et 1943, et les 73 papiers écrits lors de ses fonctions à Cambridge ; il publie, sous forme de livres, *Statistical Tables for biological, agricultural and medical research* (1938, avec F. Yates), *The theory of imbreeding* (1949), *Contributions to Mathematical Statistics* (1950), et *Statistical methods and statistical inference* (1956).

L'abondante productivité scientifique de R. Fisher jusqu'à sa mort n'a jamais gêné sa participation aux instances statistiques : président de Gonville et Caius College de Cambridge, président de la Royal Statistical Society, président de la Société de Biométrie, président de l'Institut international de Statistique (IIS). A tout point de vue, Fisher a été l'homme qui a fait de la statistique une science moderne.

II. — Francis Galton

Francis Galton naît en 1822 près de Birmingham dans une famille lui assurant un environnement intellectuel extrêmement favorable : son grand-père est membre de la Royal Society, et son cousin n'est autre que Charles Darwin. Cet environnement ne lui permettra cependant pas de mener à leur terme des études de médecine, ni d'obtenir un diplôme de mathématiques à Cambridge.

S'il a laissé un nom célèbre dans la statistique, Galton se révèle très vite un touche-à-tout génial et polyvalent ; ainsi, l'édition 1930 du *Larousse du xxe siècle* le définit comme un « voyageur et savant ». Ses premières publications datent de 1850, et concernent la géographie qui sera la grande passion de sa

vie : il explore le Nil blanc en Egypte (1846) et le Sud-Ouest africain (1850). Il s'intéresse également à la météorologie, et l'histoire retiendra qu'il est le créateur du mot « anticyclone » et de l'approche théorique de ce phénomène. Elu membre, en 1860, de la Royal Society pour la renommée de ses travaux en géographie, il fera également partie pendant presque toute sa vie du Conseil de la Royal Geographic Society. Pour parfaire le profil du personnage, il faut savoir qu'il est aussi l'inventeur du sac de couchage.

A partir de 1865, il commence à s'intéresser aux méthodes statistiques et à leur application à toutes sortes de domaines : génétique, anthropométrie, éducation, psychologie. Il essaie de décrire au mieux, et si possible de quantifier, le comportement humain et son évolution. A ce titre, Galton est considéré comme l'un des pères de la biométrie. Son champ d'études du comportement humain est vaste, allant de la recherche de la liaison existant entre la taille des parents et celle des enfants à la question de savoir si les femmes de Londres sont les plus belles d'Angleterre — et donc du monde.

Les travaux de Galton sont fondés sur la mesure quantitative faite à partir de la loi normale. Sa contribution essentielle à la statistique est probablement le concept de corrélation et sa mesure par le cœfficient de corrélation. Il utilise pour la première fois le terme corrélation dans son livre *Hereditary Genius* (1869), sans toutefois en donner une définition précise, en notant que la moralité d'un individu et son instabilité morale sont non corrélées. En 1877, étudiant la relation existant entre la taille des enfants, prise en écart par rapport à la moyenne de leur génération, et celle des parents — prise également en écart -, il constate que cet écart à la moyenne diminue, c'est-à-dire *régresse*. Il introduit donc une mesure numérique de la *régression* - ou de la réversion — qui n'est autre que notre cœfficient actuel de

régression. Avec l'aide d'un ami mathématicien de Cambridge, Hamilton Dixon, il exprime cette liaison en terme de loi normale bivariée avec une ellipse dans la partie exponentielle. Il détermine les axes de l'ellipse, se montrant en cela un précurseur de l'analyse en composantes principales.

En 1877 également, constatant que les coefficients de la régression de y en x et de x en y sont les mêmes si les variables sont normalisées en termes de variabilité ($\sigma_x = \sigma_y = 1$), il appelle alors ce paramètre coefficient de co-relation (ou corrélation). Comme son calcul se place dans le cadre de la recherche d'une relation de régression, il l'appelle r, initiale du mot régression. Notons cependant que Galton ne calcule pas r en normant ses variables par leur écart type, mais par l'écart interquartile.

Galton jouera également un rôle important dans le développement de la médiane et des fractiles, en relation avec les théories de l'eugénisme qui visent à améliorer — et mesurer cette amélioration — de la race humaine.

L'apport de Galton dépasse la méthodologie statistique, et touche également l'organisation. Très lié, à la fin de sa vie, à Karl Pearson, alors professeur de Mathématiques Appliquées à l'University College de Londres, il finance la création du journal *Biometrika* (1901). Il crée également un « Eugenics Record Office » qui fusionnera plus tard avec le Laboratoire de Biométrie de Karl Pearson pour devenir le Laboratoire Galton. Enfin, à sa mort, il laisse un legs de 45 000 livres sterling pour que soit créée la « Galton Chair of National Eugenics » dont le premier titulaire sera Karl Pearson.

Génie polyvalent, parfois inquiétant dans sa recherche systématique de la sélection de l'élite, Galton a toujours essayé de vulgariser ses travaux et de représenter physiquement les relations aléatoires qu'il étudie, par la construction, avec plus ou moins de succès,

de diverses machines. Lorsqu'il meurt en 1911, à Haslemere, il laisse plus de 300 publications, dont 17 ouvrages aussi brillants qu'éclectiques.

III. — Corrado Gini

Corrado Gini est né en 1884 à Motta di Livenza, dans la province de Trévise. Il complète sa fonction juridique (effectuée à l'Université de Bologne) en suivant des enseignements de mathématiques, ce qui lui permet de développer sa réflexion dans des domaines très diversifiés. Sa soif de toucher à des aspects multiples de l'existence humaine se traduit par la diversité des cours qu'il dispense dans les universités de Cagliari, Padoue et Rome. Il sera en effet professeur de biométrie, démographie, droit constitutionnel, économie politique, sociologie et statistique. Dans ce dernier domaine, qui nous intéresse particulièrement, il faut souligner que Gini a une influence importante et représentera longtemps l'école statistique italienne. Bien qu'on lui attribue un intérêt essentiel vers la statistique descriptive — à laquelle il apporte une contribution certaine —, Gini commence sa carrière en s'intéressant à la mesure d'une probabilité tout particulièrement dans l'étude de la proportion des sexes masculin et féminin à la naissance. Son premier ouvrage date de 1908 et est intitulé *Il sesso dal punto di vista statistica : le leggi della produzione dei sessi.* Il s'intéresse dans la suite à l'étude des distributions de probabilité et tout particulièrement à la distance entre deux distributions ; il développe aussi sa recherche dans l'étude des mesures d'association et de la « transvariation ».

Dans le domaine de la statistique descriptive, il introduit notamment le concept de *différence moyenne* qui, pour n valeurs $x_1, x_2, \ldots, x_n$, est définie par

$$g = \frac{1}{n\,(n-1)} \sum_{ij} |x_i - x_j| .$$

Même si cette mesure de variabilité avait déjà été étudié antérieurement (cf. par exemple l'article de Jordan paru en 1869 dans *Astronomische Nachrichten*, vol. 74, p. 209-226), Gini la considère de façon telle qu'elle lui sera principalement attribuée.

Dans le domaine de l'économie, Gini joue un rôle important dans la mesure des inégalités de revenu. Parmi les nombreuses approches de ce problème, le coefficient qu'il met au point est certainement le plus connu. Son interprétation graphique via la courbe de Lorenz a favorisé cette situation. Les résultats de Gini dans ce domaine datent des années 1910-1914 (cf. par exemple « Sulla misura della concentrazione e della variabilita dei caratteri », *Atti del Istituto Veneto di Scienze Lettere e Arti*, 1914). Par ailleurs, il contribue aussi à la théorie des indices de prix, notamment dans « Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues », *Metron*, vol. 4, p. 3-162, 1924.

Nous n'évoquerons pas ses travaux en démographie (fertilité différentielle, migration, études de populations...) et en sociologie (santé publique...). Notons encore que sa direction de l'Office central de Statistique et son œuvre d'éditeur dans les revues *Metron* et *Genus* ont contribué à faire de Gini l'homme influent que l'on connaît. Il décède en 1965.

IV. — William Sealy Gosset

Les étudiants qui utilisent à longueur d'année la statistique de Student ignorent bien souvent qu'il s'agit là du pseudonyme employé par William Sealy Gosset pour ses nombreuses publications. Gosset est né en 1876 à Canterbury ; il commence son éducation à Winchester, et étudie avec succès la chimie et les mathématiques au New College d'Oxford.

Contrairement à la grande majorité des statisticiens qui ont contribué au développement de leur science, Gosset n'opte pas pour une carrière de type universitaire. En 1899, en effet, il prend un emploi de brasseur au sein de la célèbre brasserie Guinness, en poste à la St James' Gate Brewery de Dublin où il passera l'essentiel de sa vie. Gosset profite de la politique, mise en place par la société Guinness, de recrutement de jeunes scientifiques d'Oxford ou de Cambridge afin d'exploiter la grande masse des données expérimentales disponibles.

Il s'aperçoit très vite que ce que nous appelons maintenant la théorie aymptotique ne s'applique pas : les conditions des expérimentations varient tellement en fonction de la température, l'origine des produits — malt, houblon —, que les données cohérentes entre elles sont peu nombreuses. Tout naturellement, ses connaissances mathématiques vont très vite le conduire à réfléchir sur la théorie des erreurs. Le témoignage d'un de ses collègues confirme que dès 1903 il calcule des probabilités d'erreur ; en 1904, il rédige chez Guinness une note interne — non publiée — intitulée « The application of the law of error to the work of the Brewery ». Dans cette note, la loi des erreurs est gaussienne, écrite sous la forme popularisée à l'époque par G. B. Airy dans « Theory of Errors of Observations » :

$$f(x) = (1/c\sqrt{\Pi}) \exp(-x^2/c^2).$$

Il présente à ses collègues, dans un langage clair et accessible à des non-spécialistes, les résultats d'estimation de c dus à Airy ou Merriman, et émet des doutes sur la loi de Airy ($V(X \pm Y) = V(X) + V(Y)$) lorsqu'il existe une liaison entre X et Y.

En 1905, Gosset rencontre Karl Pearson, avec lequel il établira des liens d'amitié ainsi qu'avec son fils Egon ; il correspondra fréquemment avec Karl puis Egon, leur soumettant et proposant des idées. Utilisant

une pratique courante chez Guinness, il séjourne, au cours de l'année scolaire 1906-1907, au Laboratoire de Biométrie de l'University College de Londres, dirigé par K. Pearson. C'est pendant cette période sabbatique qu'il élabore les bases des travaux qui vont contribuer à sa célébrité.

Dans son premier article, publié en 1907 dans *Biometrika* et intitulé « On the error of counting with a hæmacytometer », il établit la convergence de la loi binomiale vers la loi de Poisson. Ce résultat, déjà établi depuis fort longtemps en particulier par von Borkiewicz sur les suicides de femmes en Allemagne, rend Gosset populaire car il répond à une question que se posent alors de nombreux biologistes.

Les deux publications fondamentales apparaissent en 1908, toujours dans *Biometrika* : « The probable error of a mean », « Probable error of a correlation coefficient ».

Partant d'un échantillon $X_1, \ldots, X_n$ de n observations d'une variable normale de loi $N(m, \sigma)$, « Student » obtient les moments de $S^2 = \Sigma (X_i - \overline{X})^2/n$, et montre que ces moments correspondent exactement à ceux de la loi de Pearson de type III. Il montre ensuite que le coefficient de corrélation entre $\overline{X}^2$ et S^2 est nul, et, faisant l'hypothèse qu'il en est de même pour $\overline{X}$ et S, en déduit la loi de $z = \overline{X}/S$.

Il établit les propriétés de cette densité, donne la table de sa fonction de répartition pour $n = 4$ à 10, et étudie son approximation par une loi normale $N(0,1/\sqrt{n-3})$.

Il serait vain de détailler chacun des 19 autres articles publiés par Gosset jusqu'à sa mort, en 1937. Chacun d'eux apporte un éclairage nouveau sur tel ou tel point des techniques statistiques. Citons cependant, en 1909, « The distribution of the means of samples which are not drawn at random », où il établit la célèbre formule $V(\overline{X}) = (\sigma^2/n)\,[1 + (n-1)\,\rho]$ pour une moyenne de variables corrélées entre elles, et « An

experimental determination of the probable error of Dr Spearman's correlation coefficients ».

Enfin, au-delà de ses articles, Gosset eut une grande influence par les diverses correspondances qu'il adressa tout au long de sa vie à ses amis. Ainsi, Egon Pearson reconnaissait-il qu'une lettre de Gosset, écrite en 1926, avait eu une considérable influence sur ses travaux et ceux de Neyman sur la théorie des tests, jetant les bases de la comparaison des vraisemblances des hypothèses H_0 et H_1, et lui donnant même l'idée du mot « puissance ».

Comment ne pas conclure en affirmant que pour Gosset, l'ensemble de ces recherches avait pour objectif principal l'amélioration de la production de Guinness et, plus généralement, le développement de l'expérimentation statistique en agriculture et en agronomie. Précurseur des statisticiens industriels et des statisticiens d'entreprise, Gosset a toujours été considéré par les statisticiens comme un conseiller scientifique de sa société, et par les brasseurs comme l'un d'entre eux consacrant son temps libre à la statistique.

Il achèvera sa carrière chez Guinness en prenant en charge, à partir de 1935, la nouvelle brasserie de Londres, où il décédera en 1937.

V. — Jerzy Neyman

Jerzy Neyman est né le 16 avril 1894 à Bendery, en Russie, au sein d'une famille d'origine polonaise déportée après l'insurrection de 1863. Il va à l'école à Simferopol puis à Kharkov, après la mort de son père en 1906. Six ans plus tard, il entre à l'Université de Kharkov pour étudier les mathématiques et la physique. C'est là que Neyman s'intéresse aux travaux de Pearson, tout en se passionnant pour les développements de la théorie de la mesure de Lebesgue, domaine dans lequel il publie ses premiers articles.

Lors de la révolution russe, de 1917 à 1921, les conditions de vie sont difficiles pour Neyman, qui est arrêté quelques semaines, puis relâché car son savoir est nécessaire pour le bon fonctionnement de l'Université.

En 1921 — il a donc 27 ans —, Neyman arrive en Pologne à l'occasion d'un échange de prisonniers. Il obtient un poste de statisticien à l'Institut national d'Agriculture de Bydgosz, soutient sa thèse de doctorat en 1923. En 1924, une bourse d'étude lui permet un premier séjour d'un an à l'University College, chez Karl Pearson. Il y rencontre Fisher, Gosset, Karl et Egon Pearson avec lesquels il entame une forte collaboration, et fait apprécier ses travaux. En 1925, il est à Paris et assiste aux cours de Lebesgue, Hadamard et Borel.

De retour en Pologne, il prend la direction du département statistique de l'Institut Nencki de Biologie expérimentale jusqu'en 1934, année où Egon Pearson lui propose un poste de professeur associé dans son département de Statistique appliquée, à l'University College. Neyman y demeure quatre ans, et part en août 1938 pour l'Université de Californie, à Berkeley. Certes, la fonction est intéressante puisqu'on lui propose de créer un Département Statistique, mais il semble également que Neyman ait voulu fuir la Seconde Guerre mondiale qu'il sentait fort proche.

Il restera à Berkeley jusqu'à la fin de ses jours, ayant parfaitement réussi dans sa mission d'animation du Département Statistique, dans ses fonctions de directeur et celles de professeur. Parallèlement, il développe l'Institute of Mathematical Statistics, et crée la Bernoulli Society, section de l'Institut international de Statistique. Il meurt le 5 août 1981 à Berkeley.

Neyman fut un esprit extrêment fertile à toutes les époques de sa vie, s'intéressant aux probabilités, à la physique statistique, aux processus stochastiques. Cependant, sur le plan statistique, la période 1928-

1938 fut celle où s'élaborèrent les recherches qui ont laissé le plus d'éléments contribuant à la statistique mathématique au niveau des concepts. Comment ne pas mentionner l'article publié en 1928 dans *Biometrika* où E. Pearson et Neyman établissent les bases de la théorie des tests telle que nous la connaissons encore aujourd'hui avec les notions d'hypothèse nulle à tester contre toutes les autres hypothèses (idée combattue par R. Fisher). Le célèbre duo poursuit ses recherches, et ses publications de 1933 et 1934 établissent les principes de la maximisation de la puissance tout en contrôlant le niveau et énoncent le concept de test sans biais.

En 1934, Neyman crée la théorie « moderne » des sondages dans une conférence devant la Royal Statistical Society, publiée sous le titre « On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection » dans le *Journal of the Royal Statistical Society*. Nous ne nous étendrons pas plus sur l'apport de Neyman à la théorie des sondages aléatoires, développé au chapitre IV. Il est important de noter que cet article contient les germes de l'estimation par intervalle de confiance, que Neyman expose complètement en 1937 en liaison avec la théorie des tests.

Les intervalles de confiance ne rapprochent pas Fisher de Neyman, le premier ayant développé des méthodes fiduciaires dont les résultats ne concordent pas toujours avec ceux de Neyman, comme pour la différence de deux moyennes de lois normales de variances différentes (problème de Behrens-Fisher). La rupture entre ces deux génies de la statistique est alors consommée, et il est juste de reconnaître que Neyman ne répondra presque jamais aux critiques et aux attaques de Fisher, sauf en 1961 où il célèbre avec humour « les noces d'argent de sa dispute avec Fisher ».

Comment ne pas mentionner également les travaux

réalisés pendant la guerre, à Berkeley, publiés en 1949, sur les statistiques BAN (*best asymptotically normal*), leur utilisation en estimation et en test, et la même année, les études sur les tests localement optimaux en présence de paramètres nuisibles, qu'il appelle C (α) en l'honneur de Cramer ?

Mathématicien probabiliste à l'origine, Neyman reste l'exemple de l'esprit créateur, ouvert et curieux à toutes les disciplines et à toutes les cultures.

VI. — Karl et Egon Pearson

Né en 1857 à Londres, Karl Pearson a marqué la statistique d'une empreinte durable. Sa formation de mathématicien au King's College de Cambridge lui permet d'occuper la chaire de Mathématiques appliquées au University College de Londres, en 1884. Mais l'intellectuel qu'est Karl Pearson avait d'autres cordes à son arc. La physique, l'histoire, la philosophie sont autant de disciplines qui le passionnent. Cet éclectisme est dû en partie à l'influence de son père, William Pearson, avocat et membre du Conseil royal, qui fut lui-même un homme de grande culture.

Après avoir obtenu son « degree with mathematical honors » en 1879, il se tourne vers le droit et s'inscrit au Barreau en 1881. Attiré par l'Allemagne, il séjourne à Heidelberg et Berlin, s'intéressant aussi bien à la physique qu'à la méthaphysique, à l'histoire des religions qu'au droit romain. Il poursuit ses activités diverses dès son retour en Angleterre.

Sa première œuvre, largement autobiographique, date de 1880 : *The New Werther*, écrite sous forme de lettres d'un jeune homme, Arthur, à sa fiancée. Ce petit livre annonce les œuvres qui suivront : *The Trinity, A Nineteenth Century Passion Play* (1882), *Die Fronica* (1887), *The Ethic of Freethought* (1886), *The Grammar of Science* (1892).

Dans un livre consacré à l'histoire de la statistique,

on ne peut passer sous silence un écrit, conséquence d'un des examens présentés par Pearson durant ses études.

Parmi ses professeurs célèbres qui ont pour noms Arthur Cayley (algèbre linéaire, opérations sur les matrices), James Clark Maxwell (théorie électromagnétique de la lumière), Georges Gabriel Stokes (fluorescence)... se trouve Isaac Todhunter, auteur d'une *Histoire de la Théorie mathématique des Probabilités*. La question d'examen posée par ce dernier traitait d'une démonstration liée au manuscrit incomplet de Todhunter portant sur l'*Histoire de la Théorie de l'Elasticité*. C'est ainsi que Pearson, après le décès de son professeur, fut invité à achever et publier l'œuvre de ce dernier, établissant ainsi une collaboration qui s'avérera fidèle et fructueuse avec Cambridge University Press.

A l'âge de 33 ans, la carrière de Pearson va se tourner vers la statistique. Deux événements sont à l'origine de cette situation. Le premier est la publication en 1889 de l'ouvrage de Galton intitulé *Natural Inheritance* (cf. § II). Le second est la nomination de W. F. R. Weldon à la chaire de zoologie à l'University College. Pearson se lie d'amitié avec ce dernier et trouve dans les domaines d'intérêt de Weldon de multiples sources à sa réflexion statistique. C'est ainsi que Weldon recourt à des méthodes statistiques élémentaires, aidé par Galton, pour aborder la théorie de la sélection naturelle de Darwin et fournit à Pearson des données lui permettant d'obtenir des résultats théoriques. L'admiration que Pearson porte à Galton s'est traduite par une biographie en trois volumes intitulée *The Life, Letters and Labours of Francis Galton* (publiés entre 1914 et 1930).

Il est difficile de dresser une liste exhaustive des contributions de Karl Pearson à la statistique. Citons-en quelques-unes en respectant la chronologie de leur publication.

a) L'étude des courbes de densité donne lieu à des articles importants : « On the dissection of asymmetrical frequency curves », *Phil. Trans, Ser. A,* vol. 185, 1894, et « Skew variation in homogeneous material », *Phil. Trans, Ser. A,* vol. 186, 1895. Ces deux articles ont été rassemblés sous un intitulé commun : *Contributions to the mathematical theory of evolution.* L'objectif de Pearson est de construire un système de courbes de densité basé sur une seule équation aux différences avec des paramètres obtenus par la méthode des moments.

b) La corrélation (cf. chapitre II) doit à Pearson l'étude de son cœfficient *r* et de ses généralisations. Son exposé de base est contenu dans l'article intitulé « Regression, heredity and panmixia », *Phil. Trans, Ser. A,* vol. 187, 1896. Pearson s'est appuyé sur les travaux de Galton pour construire les formules que nous utilisons couramment. Sa théorie générale de la corrélation portant sur trois variables est la base de très nombreux travaux soit personnels, soit rédigés par ses élèves et collaborateurs comme G. U. Yule et W. F. Sheppard.

c) L'année 1900 lui permet de mettre au point un critère de la qualité d'un ajustement qui va influencer toute la statistique, celui du *chi-carré* (encore dénommé chi-deux). L'article de référence en la matière est intitulé « On the 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 from random sampling », *Phil. Mag., Ser. 5,* vol. 50, 1900. Son souci de mesurer le degré d'adéquation d'un modèle théorique à des observations réelles l'a conduit à utiliser des données variées. Huit exemples sont en effet traités dans cet article. Certains concernent des observations du domaine des Sciences naturelles. D'autres traitent de données artificielles liées à des lancements de dés. Notons d'ailleurs que Pearson a

recouru de façon importante à des lancements de pièces de monnaies ou de dés effectués par lui-même, ses amis (dont Weldon) et ses élèves. Il est à noter que ces travaux reposent aussi sur une correspondance qu'il a entretenue avec Edgeworth.

d) Nous serions incomplets si nous ne citions pas les travaux de Pearson consacrés à la *dispersion* . C'est lui qui introduit le terme *standard deviation* et la notation σ dans son article de 1894 cité plus haut. Il a aussi établi l'habitude de mesurer la « variance échantillonnée » de la plupart des statistiques.

Deux autres activités de Pearson sont encore à souligner. La première concerne son rôle d'éditeur de *revues*. Pearson assume cette fonction pour les *Annals of Eugenics* (publiés depuis 1925 par le Eugenic Laboratory), les *Drapers' Company Research Memoirs* (publiés à partir de 1904) et surtout pour *Biometrika* (datant de 1901). Cette dernière revue a été fondée par Weldon, Pearson et Davenport « in consultation with F. Galton ». A la mort de Karl Pearson, son fils, Egon, lui succédera en tant qu'éditeur de cette publication. La seconde activité concerne le souci de construire des *tables statistiques* dans le but de fournir aux utilisateurs de cette discipline un outil de travail. Beaucoup de ces tables furent publiées dans Biometrika. Les premières d'entres elles ont été rassemblées en 1914 dans un ouvrage en deux volumes intitulé *Tables for Statisticians and Biometricians* et publié par le laboratoire de Pearson à l'University College.

On ne peut évoquer la vie de Karl Pearson sans préciser que de nombreuses controverses ont émaillé son existence. Ses prises de position concernant le socialisme, l'émancipation de la femme et la libre pensée ont marqué sa jeunesse. Son souci, en compagnie de Weldon, d'appliquer une méthodologie statistique à l'étude de l'évolution a engendré de vives oppositions qui les ont conduits à créer *Biometrika*. Il est symptomatique que Pearson n'ait jamais été

nommé « Fellow of the Royal Statistical Society ». Et que dire des controverses entre R. A. Fisher et lui, après la première guerre mondiale. Il n'empêche que des chercheurs de tous pays fréquentèrent son laboratoire, tissant ainsi une toile dont l'importance encore actuelle souligne son rôle magistral.

En 1933, trois ans avant sa mort, son département est divisé en deux : le premier géré par Fisher, le second par son fils Egon.

Egon Sharpe Pearson n'est pas seulement le fils de son père. Né à Hampstead, près de Londres, en 1895, il entame des études de mathématiques au Trinity College de Cambridge en 1914. Des problèmes de santé et la guerre mondiale perturbent quelque peu ses études. En 1919, ayant obtenu son *first degree*, il se tourne vers l'astronomie. Bien qu'occupant une place dans le département de statistique appliquée de l'University College (département fondé par Karl Pearson en 1911), il développe sa réflexion sans être trop influencé par son père. Bien plus, à la recherche d'orientations nouvelles, il se lie en 1925 avec Jerzy Neyman, alors que ce dernier effectue une visite dans son département. Sa correspondance avec Gosset est aussi la source de travaux importants. Cette double influence est particulièrement fertile jusqu'en 1938, date du départ de Neyman pour les Etats-Unis, un an après la mort de Gosset. Ce travail en commun permet de jeter les bases de l'approche, dite de « Neyman-Pearson », des tests d'hypothèses ; il s'étend de 1926 à 1933, date de la parution de l'article intitulé « On the problem of the most efficient tests of statistical hypotheses », *Philos. Trans. Roy. Soc. London, Serie A*, vol. 231, p. 289-337.

Par ailleurs, sous l'influence de Gosset, il développe des résultats intéressants dans le domaine de la *robustesse* , aidé en cela par les tables de nombres au hasard produites par Tipett entre 1925 et 1927, qui

lui permettent de recourir à ce qu'on appellera plus tard simulation.

Egon Pearson joue aussi un rôle important dans la recherche appliquée à l'industrie et à l'agriculture. Son action à la Royal Statistical Society (traduite dans les sections B et C du journal de cette association) et sa participation au British Standards Institution ont été déterminantes.

Après la mort de son père en 1936, il consacre une partie importante de son temps à réviser les tables statistiques de ce dernier, tout en continuant à travailler dans la voie tracée par Neyman et lui-même.

Après la seconde guerre mondiale, il participe au développement de l'usage du contrôle de qualité et devient l'un des fondateurs de la Société de Recherche opérationnelle britannique en 1948.

Après avoir quitté la direction du département de statistique de l'University College en 1960, il se consacre principalement à l'élaboration d'une version fructueusement commentée des conférences que son père consacra à l'*Histoire de la Statistique aux xviie et xviiie siècles*, entre 1921 et 1933. Cet ouvrage est publié chez Macmillan, en 1979, un an avant son décès à Midhurst, dans le Sussex.

VII. — **Adolphe Quételet**

Lambert-Adolphe-Jacques Quételet est né à Gand en 1796. Fils d'un Picard installé depuis peu dans cette ville, il doit rapidement chercher des moyens d'existence car son père meurt en 1803. C'est ainsi qu'en 1813 il est nommé professeur « de dessin, de mathématiques et de grammaire » dans un collège privé d'Audenarde, avant de retourner en 1815 enseigner les mathématiques dans son lycée de Gand qui, d' « impérial » qu'il était durant ses études, vient juste de devenir « royal ». Il rêvait d'être un poète : il se retrouve mathématicien (il fallait vivre !). Sa thèse de

doctorat, intitulée *De quibusdam locis geometricis necynon de curva focali*, présentée en 1819 à l'Université de Gand, est la première thèse de cette institution, fondée quelques années plus tôt. Attiré à Bruxelles, il est nommé professeur de mathématiques élémentaires à l'Athénée de cette ville.

En 1823, il a l'idée de fonder un observatoire et est envoyé à Paris pour s'initier à la pratique de l'astronomie. C'est à l'Observatoire de Paris qu'il rencontre Fourier, Laplace et Poisson. C'est ainsi qu'il a ses premiers contacts avec la statistique.

Chargé de visiter les centres d'astronomie italiens, il vit la Révolution belge de loin, tremblant pour la construction de « son observatoire ».

1832 est une année importante pour Quételet. Il s'installe à l'Observatoire de Bruxelles et est chargé par le gouvernement belge d'assister à la réunion de l'Association britannique pour l'Avancement des Sciences. Sa communication sur les étoiles filantes et le magnétisme est certes intéressante, mais sa suggestion de créer une section spéciale de statistique, après des entretiens avec Malthus, Jones et Babbage aura encore plus d'impact. C'est ainsi qu'est créée en 1834 la Statistical Society de Londres qui commence à publier en 1837 les *Transactions of the Statistical Society of London.*

En 1834, il décline l'invitation qui lui est faite par l'Université Libre de Bruxelles de lui confier une chaire, préférant rester fidèle à son observatoire. La même année, il est élu secrétaire perpétuel de l'Académie royale de Belgique. Ses recherches dans le domaine de l'astronomie et de la climatologie vont occuper une place importante de la vie de Quételet dans les années qui suivent. Mais, parallèlement à ces activités, il va se tourner vers la statistique. Invité à collaborer à la publication du recensement de la population belge en 1829, il acquiert rapidement une compétence en la matière. C'est ainsi qu'en 1841, il

devient le premier président de la Commission centrale de Statistique. En 1851, il lance l'idée

« d'inviter à se réunir en un congrès de statistique universelle, à Bruxelles, les savants des différentes parties du monde qui s'occupent de statistique, afin d'encourager et de développer les travaux qui se rapportent à cette science et, s'il est possible, de les coordonner par l'adoption de bases uniformes ».

Le premier congrès international de statistique se tient donc à Bruxelles en 1853. Quételet, désigné comme président du bureau provisoire, prononce le discours d'ouverture et définit des objectifs qui marqueront les congrès qui suivront pendant de très nombreuses années. Deux cents trente-sept savants provenant de vingt-six Etats différents participent à cette réunion.

En 1855, Quételet est victime d'une crise d'apoplexie dont les conséquences auraient pu lui être funestes. Sa volonté de surmonter cet accident ne suffit pas à en vaincre les séquelles. Après cette date, les nombreuses publications de Quételet seront surtout des compilations de travaux antérieurs. En 1856, il abandonne les travaux de l'Observatoire concentrant son énergie sur ses activités de président de la Commission centrale de Statistique et de secrétaire perpétuel de l'Académie, qu'il était devenu dès 1834.

Ses participations aux différents congrès de statistique ultérieurs sont l'occasion pour lui de défendre le principe d'une *statistique scientifique* reposant sur le calcul des probabilités. Ce principe conduira le congrès de Florence de 1867 à créer une section spéciale chargée de s'occuper de ce problème.

Comme nous l'avons déjà souligné, Quételet se consacre surtout, après 1855, à synthétiser ses recherches antérieures, tout en y apportant des compléments significatifs. C'est ainsi qu'il publie une *Histoire des sciences mathématiques et physiques chez les Belges* (1864), et *Météorologie de la Belgique, comparée à celle du Globe* (1867).

Mais c'est dans le domaine de la statistique que ses écrits ont le plus marqué ses contemporains. En 1869, en effet, paraît une seconde édition d'un ouvrage publié en 1835, qui constitue en fait une réédition de la plupart des œuvres statistiques de Quetelet. La *Physique sociale* en est le titre. A travers ce livre, Quételet veut développer une étude de l'homme, basée sur le calcul des probabilités. Le fait que cette œuvre soit une compilation relativement mal ordonnée de chapitres rédigés à des époques différentes a suscité des critiques à son égard. Il est cependant essentiel de replacer chaque chapitre dans son contexte historique. C'est en se référant à ses écrits antérieurs à 1831 (études statistiques sur les naissances, les décès, les crimes...) et à ceux qui ont constitué *L'homme et le développement de ses facultés*, publié en 1845 et dans lequel il élabore la *mécanique sociale* comme l'étude des lois qui régissent l'homme du point de vue physique, intellectuel et moral (qui ne concerne pas chaque individu, mais qui traite du développement d'un « homme moyen »), que l'on peut mieux comprendre l'apport de Quételet.

Dans son œuvre, Quetelet a présenté de nombreux tableaux portant sur trois et quatre variables, dépassant ainsi la simple analyse bivariée. Il faut aussi souligner que les lois auxquelles il s'est attaché concernent autant l'étude de distributions que celle de tendances. Citons en particulier, dans ce dernier cas, le concept de *penchant* (au crime) si cher à Quételet.

Jusqu'à sa mort, qui survient en 1874, il est salué comme le premier « qui a contribué à faire de la statistique une science morale ». Quételet a su donner à la statistique du milieu du XIXe siècle une impulsion dont l'influence ultérieure est indéniable.

ÉLÉMENTS DE BIBLIOGRAPHIE

OUVRAGES GÉNÉRAUX

Benzécri J.-P. (1982), *Histoire et Préhistoire de l'analyse des données*, Dunod.

INSEE (1987), *Pour une histoire de la statistique*, Paris, Economica.

Kendall M. G. and Plackett R. L. (1969), *Studies in the history of statistics and probability*, vol. 2, London, Charles Griffin.

Kotz S. and Johnson N. L., Eds (1982 à 1988), *Encyclopedia of statistical science*, 9 vol., New York, J. Wiley.

Kruskal W. H. and Tanur J. M., Eds (1978), *International encyclopedia of Statistics*, 2 vol., New York, Macmillan and Free Press.

McMullen and Pearson E. S. (1939), William Sealy Gosset (1876-1937), *Biometrika*, 30, p. 205-250.

Owen D. B., Eds (1976), *On the history of statistics and probability*, « Statistics : Textbooks and Monographs », 17, New York, M. Dekker.

Pearson E. S., Ed. (1929), *The History of statistics in the XVIIth and XVIIIth centuries*, New York, Macmillan.

Pearson E. S. and Kendall M. G., Eds (1969), *Studies in the history of statistics and probability*, vol. 1, London, Charles Griffin.

Stigler S. M. (1986), *The history of statistics ; the measurement of uncertainty before 1900*, Cambridge, Harvard Univ. Press.

Westergaard H. (1932), *Contributions to the history of statistics*, London, King.

CHAPITRE 2

Beniger J. R. and Robyn D. L. (1978), Quantitative graphics in statistics : a brief history, *The American Statistician*, vol. 32, p. 1-11.

Bertin J. (1977), *La graphique et le traitement graphique de l'information*, Paris, Flammarion, « Nouvelle Bibliothèque Scientifique ».

Eisenhart D. (1961), Boscovich and the combination of observations, in *Roger Joseph Boscovich*, ed. by L. L. Whyte, London, Allen and Unwin, p. 200-212.

Fienberg S. E. (1979), Graphical methods in statistics, *The American Statistician*, vol. 33, p. 165-178.

Fisher I. (1922), *The making of index numbers : a study of their varieties, tests and reliability*, Boston, Houghton Miffin.

Harter H. L. (1974), The Method of least squares and some alternatives-part I, *Int. Stat. Rev.*, vol. 42, p. 147-174.

Plackett R. L. (1958), The principle of the arithmetic mean, *Biometrika*, vol. 45, p. 130-135.

Tufte E. R. (1983), *The visual display of quantitative information*, Cheshire, Graphic Press.

CHAPITRE 3

David F. N. (1955), Dicing and gaming (a note on the history of probability), *Biometrika*, vol. 42, p. 1-15.
Kendall M. G. (1956), The beginnings of a probability calculus, *Biometrika*, vol. 43, p. 1-14.
Todhunter I. (1965), *A History of the mathematical theory of probability from the time of Pascal to that of Laplace*, Cambridge, Cambridge University Press.

CHAPITRE 4

Cassel C., Sarndal S. and Wretman J. (1979), *Foundations of inference in survey sampling*, New York, J. Wiley.
Droesbeke J.-J., Fichet B. et Tassi Ph., Eds (1987), *Les sondages*, Paris, Economica.
Hansen M. and Madow W. (1974), Some important events in the historical development of sample survey, in *On the history of Statistics and Probability*, éd. D. Owen, New York, Marcel Dekker.
Hansen M., Dalenius T. and Tepping B. (1985), The development of sample surveys of finite populations, A celebrations of statistics, *The ISI Centenary Volume*, A. Atkinson and S. Fienberg éd., Springer Verlag.
Kruskal W. and Mosteller F., Representative Sampling, *Inst. Statist. Rev.*, 47, 1979, p. 13-24, p. 111-127, p. 245-265 ; 48, 1980, p. 169-195.
Zarcovich S. (1966), *Sondages et recensements*, Rome, FAO.

CHAPITRE 5

Fisher-Box J. (1978), *R. A. Fisher, the life of a scientist*, New York, Wiley
Cramer H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.
Fisher R. A. (1925), *Statistical Methods for Research Workers*, 1re éd., Edinburgh, Oliver and Boyd.
Pearson E. S. (1966), *Selected Papers*, Cambridge University Press.
Reid C. (1982), *Neyman from life*, New York, Springer Verlag.

CHAPITRE 6

Huber R. (1981), *Robust Statistics*, New York, Wiley.
Hajek J. and Sidak Z., *Theory of rank tests*, New York, Academic Press.
Lecoutre J. P., Tassi P. (1987), *Statistique non paramétrique et robustesse*, Paris, Economica.

CHAPITRE 7

Anderson O. (1929), *Die Korrelazionrechnung in der Konjoncturforschung*, Bonn, Schroeder.
Coutrot B. et Droesbeke J.-J. (1990), *Les méthodes de prévision*, 2e éd., Paris, Presses Universitaires de France, coll. « Que sais-je ? », n° 2157.
Davis H. T. (1941), *The analysis of economic time series*, Bloomington, Indiana, The Principia Press.
Droesbeke J.-J., Fichet B. et Tassi Ph., Eds (1989), *Séries chronologiques. Théorie et pratique des modèles ARIMA*, Paris, Economica.
Gourieroux Ch. et Monfort A. (1983), *Cours de séries temporelles*, Paris, Economica.
Kendall M. G. (1973), *Time Series*, London, Griffin.
Priestley M. B. (1981), *Spectral analysis and time series*, New York, Academic Press.

CHAPITRE 8

Bouroche J.-M. et Saporta G. (1980), *L'analyse des données*, Paris, Presses Universitaires de France, coll. « Que sais-je ? », n° 1854.
Gif A. (1981), *Non-linear multivariate analysis*, Université de Leiden.
Jardine N. et Sibson R. (1971), *Mathematical Taxonomy*, New York, Wiley.
Sokal R. et Sneath (1963), *Principles of numerical taxonomy*, San Francisco, Freeman.

CHAPITRE 9

Gini C. (1955), *Memorie di metodologia statistica*, vol. 1 : *Variabilita e Concentrazione*, Roma, Libreria Eredi Virgilia Veschi.
Gini C. (1960), *Memorie di metodologia statistica*, vol. 2 : *Transvariazione*, Roma, Libreria Goliardica.
Lottin J. (1912), *Quételet : statisticien et sociologue*, Paris, Félix Alcan.
Mailly Ed. (1875), Essai sur la vie et les ouvrages de Lambert-Adolphe-Jacques-Quételet, *Annaire de l'Académie royale de Belgique*, 41, p. 109-297.
Neyman J. and Pearson E. S. (1967), *Joint statistical papers*, Cambridge University Press.
Pearson K. (1914, 1924, 1930), *The life, letters and labours of Francis Galton*, 3 vol., Cambridge University Press.
Pearson E. S. (1938), *Karl Pearson : An appreciation of some aspects of his life and work*, Cambridge, Cambridge Univ. Press.
Pearson E. S. (1966), *Selected papers*, Cambridge University Press.

TABLE DES MATIÈRES

CHAPITRE PREMIER. — L'histoire ? Une (certaine) histoire 3

— II. — La statistique descriptive 5

— III. — La probabilité 22

— IV. — Les sondages 39

— V. — L'émergence de l'inférence statistique 57

— VI. — Statistique non paramétrique et robustesse 67

— VII. — L'analyse des séries chronologiques 76

— VIII. — L'analyse des données 89

— IX. — Quelques grands noms 101

BIBLIOGRAPHIE 125

Imprimé en France
Imprimerie des Presses Universitaires de France
73, avenue Ronsard, 41100 Vendôme
Août 1990 — N° 35 811