STATISTIQUE ET ANALYSE DES DONNÉES

MICHEL ARMATTE

La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher

Statistique et analyse des données, tome 13, nº 2 (1988), p. 65-92
<http://www.numdam.org/item?id=SAD_1988__13_2_65_0>

Statistique et Analyse des Données
1988 - Vol. 13 n°2 - p. 65-92

LA CONSTRUCTION DES NOTIONS D'ESTIMATION ET DE VRAISEMBLANCE CHEZ RONALD A. FISHER

Michel ARMATTE
Université Paris IX-Dauphine

L'étude des articles fondateurs que Ronald A. Fisher consacre dans les années 1920 à l'Estimation permet de suivre pas à pas les phases et arguments de sa construction théorique. Celle-ci associe des matériaux anciens (population et échantillon) et nouveaux (vraisemblance, précision), qui sont redéfinis à la fois par de nouvelles propriétés syntaxiques (efficacité, exhaustivité) et par de nouvelles interprétations sémantiques (l'information) et pragmatiques (la situation expérimentale). La solidité de cette construction est évaluée au travers des controverses nombreuses qui ont entouré sa naissance, sa diffusion, et son dépassement.

The study of the primary papers R.A. Fisher assigns in the twenties to Estimation allows to follow consequently the stages and the arguments of its theoretical construction. This one associate old materials (population, sample) and new materials (likelihood, precision) which are redefined both by new syntactical properties (efficiency, sufficiency) and new semantic properties (information) or pragmatic properties (the experimental situation). The strength of this construction is valued through many controversies which have surrounded its birth, its diffusion and its overtaking.

Bien que l'idée d'estimation apparaisse dès les premiers écrits de la statistique mathématique, chez Jacques Bernoulli ou chez Laplace par exemple, c'est la construction par R.A. Fisher, dans les années 1920, d'une théorie scientifique de l'Estimation, et plus généralement de l'inférence statistique qui va former le noyau du modèle — ou paradigme — dominant, selon lequel la statistique mathématique se constitue comme discipline scientifique autonome et comme méthodologie des sciences expérimentales ou d'observation.

Ce texte « raconte » cette construction et en analyse les phases et les constituants.

Avec quels objectifs et selon quelle méthode?

En quoi ce retour aux sources peut-il intéresser le chercheur, ou l'utilisateur d'outils, qui ont été depuis largement assimilés, popularisés, perfectionnés voire même dépassés par d'autres approches?

Ce sont les premières questions qu'il convient de poser.

1. *Pourquoi faire de l'histoire sociale des sciences*

En effet, contrairement à la culture anglo-saxone dans laquelle existe depuis longtemps une solide tradition de recherche et de débats autour de l'histoire de la statistique mathématique [1], notre propre culture scientifique est a-historique et privilégie une approche formaliste et structurelle; celle-ci

1. En dehors de la référence classique qu'est Todhunter (1865), une première approche peut se faire à l'aide des articles rassemblés par Pearson et Kendall (1970) et Kendall et Plackett (1977), complétée par les ouvrages récents de Mackenzie (1981) et Stigler (1986). Le seul ouvrage français est Benzecri (1982).

Texte reçu le 30.6.88

est une reconstruction des acquis de la discipline selon une logique abstraite où prend le pas sur la logique sociale de leur production. Ce particularisme français est encore accru dans les disciplines mathématiques où le bourbakisme a systématisé cette réorganisation structurelle des concepts.

Or notre culture scientifique tient pour l'essentiel à notre système éducatif qui la met en forme et la reproduit de génération en génération, à travers un appareillage qui filtre la production scientifique pour n'en retenir que les éléments stabilisés et susceptibles de s'arranger selon une stricte cumulativité linéaire; pédagogie oblige. Il en ressort à notre avis toute une série d'effets, ou de méfaits, de la « transposition didactique », qui s'ajoutent à ceux de la production scientifique elle-même, et que l'on peut résumer par le terme systémique de « boîte noire ».

La vérité d'un énoncé théorique, l'efficacité d'un dispositif, sont des propriétés intrinsèques d'une théorie qui fonctionne comme une boîte noire, c'est-à-dire qu'il est devenu très difficile d'en examiner l'intérieur, et encore plus de savoir comment ces propriétés lui ont été conférées.

Par exemple bien peu de statisticiens, et encore moins d'utilisateurs, savent comment s'est construit l'efficacité d'un coefficient de corrélation dans la mesure d'une association entre deux caractères, ce qui ne les empêche pas de s'en emparer, d'en user, et d'en abuser comme d'une force non négligeable, voire décisive, conférée à leur propre argumentation. Il nous semble dommage que utilisateurs et statisticiens aient perdu toute mémoire des acteurs, des matériaux, des enjeux et des stratégies, des investissements de forme (Thevenot, 1985), qui ont construit en même temps le concept et son efficacité.

Or pour peu que l'on ouvre ces boîtes noires, c'est-à-dire que l'on exhume ces matériaux et ces forces qui ont donné naissance à une théorie nouvelle sans préjugé ni restriction à priori sur leur nature, l'on découvre bien vite que l'efficacité ou la vérité de cette théorie sont le résultat d'un *travail de construction* de la part d'un acteur — on l'appelle un savant avant l'ère moderne de la division du travail de recherche — non point isolé mais inscrit dans un champ social d'alliés et de challengers.

Ce que construit ce savant est d'abord, en mathématique, un *enchaînement syntaxique* d'objets et de propriétés qui les relient. C'est par exemple Laplace associant, en 1812, fonctions génératrices et loi « normale » — calcul analytique et calcul des probabilités — au sein de ce qui deviendra le « théorème central limite ». Mais dans cet exemple, comme dans bien d'autres, les éléments associés ne sont pas nouveaux; Laplace n'invente pas les fonctions génératrices, connues de de Moivre et Lagrange, ni la loi « normale » connue depuis de Moivre (1733) comme limite de la loi binominale. Il les dissocie de leur contexte antérieur pour leur donner un autre rôle dans sa construction. Il les réarrange.

Il n'y a pas *découverte*, au sens où cet arrangement, qu'il suffirait de révéler, préexisterait au travail du savant, dans les choses même de la nature.

Il n'y a pas *invention* au sens où les éléments de cet arrangement seraient créés ex-nihilo.

De plus ces éléments ne se limitent pas à des matériaux de nature cognitive, c'est-à-dire des idées et des concepts. Sont enrôlés aussi dans la construction des objets techniques, des forces sociales et politiques, des éléments naturels, des « forces de loi » qui participent à la solidité et l'efficacité de l'édifice, et que seule une volonté d'épuration disciplinaire (aux deux sens du terme) nous fait rejeter dans l'en dehors de la science.

En poursuivant l'exemple de la construction laplacienne, sa syntaxe mathématique n'est pas isolable de sa *sémantique*, c'est-à-dire de la signification des termes mathématiques, par exemple celui d'erreur qui est au centre de la Théorie des Erreurs, laquelle est le fil conducteur de la plupart des travaux mathématiques de l'époque, avec pour principaux champs « d'application » la mécanique céleste et « la figure de la terre ».

Cette sémantique conduit elle-même à une *pragmatique*, c'est-à-dire toute une configuration d'actions politiques sur le milieu social et naturel. Il y a, au delà du cacul astronomique et de la problématique de la figure de la terre, des enjeux propres à une série d'entreprises (La Carte de la France, le système métrique, l'expédition d'Égypte, l'essor de la Marine Marchande) où la science se fait instrument privilégié d'une politique.

Ces usages sémantiques et pragmatiques laissent des *traces :* l'ancrage original de la notion de variable aléatoire dans celle plus restrictive d'erreur persiste dans la terminologie : Fisher, après Pearson, abandonne la terminologie « d'erreur probable » pour celle d'écart-type et de variance, mais continue d'utiliser l'expression « courbe d'erreur » pour la densité d'une variable aléatoire. La prégnance de ces marques originales se lit aussi en creux dans le travail que doivent faire les savants pour casser les associations de certains notions avec des significations ou des connotations antérieures, lorsqu'ils veulent transporter les mêmes concepts dans d'autres champs, dans d'autres réseaux d'associations.

Nous pouvons maintenant répondre à nos questions préliminaires.

Pourquoi faire de l'histoire des sciences et plus particulièrement de la statistique mathématique? Parce que nous ne connaissons bien qu'une forme figée, réifiée, légalisée par la communauté scientifique et didactique, de nos théories scientifiques, et qu'il convient de réouvrir les boîtes noires qui les composent si nous voulons connaître leur mode de production, restituer les enjeux et les rapports de force qui les ont fait naître, soit pour mieux les utiliser et les réintégrer dans nos propres problématiques, soit pour mieux les critiquer ou les dépasser; ceci que nous soyons statisticiens, utilisateurs, ou enfin, concernés plus généralement par la philosophie des sciences. Auquel cas, ce texte peut être lu comme une étude de cas illustrant une certaine méthode socio-historique par un domaine particulier des sciences.

En effet, à la seconde question — comment faire l'histoire de la discipline statistique — nous répondons qu'il ne peut s'agir que d'une histoire sociale, au sens où elle ne se limite pas à l'histoire internaliste des enchaînements d'idées et de concepts, au sens où elle n'exclut pas, à l'inverse, les contenus pour ne retenir que des perturbations qui modifieraient la seule logique interne de la syntaxe mathématique. C'est *une socio-logique* [2] des associations des acteurs, matériaux et forces les plus divers qui font le succès, et parfois la faiblesse d'une théorie, considérée comme système, c'est-à-dire combinaison d'éléments interdépendants, située dans sa concurrence avec d'autres systèmes.

Pour revenir au système ou modèle de l'estimation chez Fisher, il s'agira de repérer dans les textes originaux, dont le noyau est formé par les deux articles de 1922 et 1925, les éléments de sa syntaxe, soit les propriétés d'efficacité et d'exhaustivité associées aux estimateurs du maximum de vraisemblance, mais aussi ceux de la sémantique qui relie cette construction mathématique à une théorie générale de l'inférence où sont redéfinies les notions de population infinie, de probabilité, d'échantillon, d'information et de causalité, sans oublier pour finir la pragmatique du modèle : nous voulons parler du renouvellement du modèle de données empiriques et des problématiques dans les champs d'application de l'Hérédité, de l'Eugénique, de la Génétique et de l'Agronomie auquel Fisher a plus que largement contribué.

L'étude des controverses qui ont à la fois nourri et affaibli cette construction n'est pas un dérivatif à notre enquête. Elle fournit l'occasion d'étudier « à chaud » le processus social de la création scientifique, et la preuve qu'avant de se figer en une boîte noire, la théorie n'est jamais assurée de sa victoire — de sa vérité si l'on préfère —, qu'elle ne doit qu'à une série de conquêtes et de succès dans des épreuves successives.

2. Nous empruntons cette dénomination à David Bloor (1976). Cette approche de la sociologie des sciences est également développée par de nombreux travaux anglo-saxons dont on trouvera une recension dans S. Shapin (1982); elle est prolongée en France par les travaux de Bruno Latour (1984, 1985). Voir aussi l'Année Sociologique (1986) consacrée à la sociologie des sciences et des techniques, et les travaux de Desrosieres et Thevenot sur l'association entre forme statistique et lien politique.

2. *Chronologie*

La notion d'estimation ne date pas bien sûr des travaux de Fisher. Si l'on entend par ce terme la détermination d'un paramètre inconnu caractérisant une épreuve aléatoire à partir de l'observation de quelques-unes de ses réalisations, les tous débuts de l'histoire du calcul des probabilités sont concernés par cette problématique. Nous n'en donnerons que quelques exemples.

Le chapitre de *l'Ars Conjectandi* de Jacques Bernoulli (1713), dans lequel il établit la loi faible des grands nombres peut être lu de façon directe comme un théorème sur la distribution d'une fréquence pour une probabilité donnée; mais bien des indices permettent aussi de le lire comme une tentative de remonter des fréquences observées aux probabilités. Parmi ces indices, d'autres chapitres du même ouvrage introduisent une conception non fréquentiste de la probabilité, traitée en terme de « force de conviction » ou de « raison de croire » et proposent une algèbre — non booléenne — des événements dans laquelle il s'agit de remonter de la chose produite (l'événement), à ce qui la rend probable, au sens de prouvable, soit le nombre et la force de ses arguments. (N. MEUSNIER, 1987).

Mais c'est l'essai fameux du révérend *Bayes*, lu par Price après sa mort (1763), qui donnera le premier un principe permettant de remonter des événements aux causes, jusqu'à constituer un véritable paradigme de l'inférence statistique pendant près d'un siècle et demi. Cet essai contient d'abord (section I) quelques règles du calcul des probabilités, dont le « théorème de Bayes » qui étant donné deux événements successifs (« subséquents ») A et B permet de calculer la probabilité (à posteriori) du premier sachant que le second s'est réalisé, en fonction de sa probabilité à priori :

$$p(A/B) = p(A \text{ et } B) / p(B) = p(B/A) \cdot p(A) / p(B)$$

Cet essai contient ensuite une application de ce théorème à l'estimation du paramètre p d'une épreuve de Bernoulli répétée n fois; la variation continue de p est obtenue par l'artifice non pas d'une composition d'urne, mais du partage aléatoire de la surface d'un billard par le lancer d'une première boule, l'épreuve de Bernoulli consistant à lancer une seconde boule qui s'arrête dans l'une des deux parties du billard définies par le premier lancer. Cet artifice justifie de plus le *postulat* d'une loi à priori uniforme pour p. Bayes obtient alors un jugement à posteriori sur le paramètre p de la forme (dans des notations modernisées) :

$$\text{prob}\,[a < p < b] = \int_a^b p^x (1-p)^y \, dp \,/ \int_0^1 p^x (1-p)^y \, dp$$

où x et $y = n\text{-}x$ sont les nombres de succès et échecs à l'épreuve.

Bayes termine par un « scholium » qui transforme le postulat d'uniformité de la loi à priori en principe général « dans le cas de la probabilité d'événements dont nous ne savons absolument rien avant de faire des tirages ».

Laplace, dans son *Mémoire sur la probabilité des causes* (1774) — bien que n'ayant pas lu Bayes — puis dans la *Théorie analytique des Probabilités* (1920), reprendra et popularisera cette méthode d'estimation dite de « la probabilité inverse » ainsi que le scholium sous la forme d'un principe dit « de raison insuffisante » qui associés, donnent le 6^e principe de Laplace :
soient n causes $C_1, C_2, ..., C_n$, et un événement E, alors :

$$\text{prob}\,(C_i/E) \text{ est proportionnelle à prob } (E/C_i)$$

L'insistance mise ici sur la genèse de cette méthode de la probabilité inverse se justifie par le rôle central — paradigmatique — qu'elle joue tout au long du XIX^e siècle, malgré les commentaires critiques et les attaques en règle qu'elle subit de la part de nombreux probabilistes et statisticiens, en particulier les anglais Mill (1850), Boole (1854) et Venn (1866) et enfin Fisher qui, on le verra plus loin, ne cessera de combattre cette « ineptie ».

Après Bayes, c'est la problématique de la *Théorie des Erreurs* qui fournit le cadre et le prétexte à travers lesquels mathématiciens et astronomes, cherchant le « milieu qu'il faut prendre entre plusieurs observations », testent diverses combinaisons des trois éléments de cette problématique : quel milieu?

Quelle loi de probabilité pour les erreurs? Quelle procédure d'ajustement ou d'estimation?

A défaut d'exhaustivité sur ce thème qui nous éloignerait de notre sujet, citons rapidement quelques travaux qui peuvent être considérés comme précurseurs.

C'est d'abord le premier usage par Lambert (1760) de la méthode du maximum de vraisemblance; puis les trois mémoires contemporains de Lagrange (1773), Laplace (1774) et Daniel Bernoulli (1777), dont le dernier qui suppose une densité des erreurs bornée d'abord elliptique, puis circulaire, et use de la même méthode pour estimer le rayon de son « cercle modérateur ».

C'est ensuite la promotion de la méthode des moindres carrés par Legendre (1805) et Gauss (1809) qui ne connaîtra toute sa gloire que lorsque Gauss (1813) et Laplace (1812) l'auront associée à la loi « normale », justifiée par le « théorème central limite », et à la moyenne arithmétique; la méthode permet d'estimer directement un paramètre, mais aussi le vecteur des coefficients d'un modèle linéaire. Notons que Gauss usera d'abord de la méthode de la probabilité inverse en prenant le critère du maximum de la densité à posteriori, avant que de lui préférer une minimisation d'une fonction de perte dont la forme la plus simple est quadratique.

C'est aussi dans ce contexte que dès 1818, dans un second supplément à sa Théorie Analytique, Laplace cherche à estimer le paramètre y du modèle $a_i = p_i \cdot y + x_i$, où a_i, p_i sont des couples d'observations et x_i des erreurs aléatoires de distribution quelconque mais paire. Il est amené à comparer les variances des deux estimateurs de y que sont la moyenne pondérée (méthode des moindres carrés) et la médiane pondérée (méthode de situation de Boscovich, 1755), et découvre non seulement l'« efficacité » du premier estimateur, mais aussi son « exhaustivité », en ce sens que, pour une distribution normale des erreurs, toute combinaison linéaire des deux estimateurs — dont la loi nécessite leur distribution conjointe — ne peut améliorer le premier de ces estimateurs. Bien sûr, comme le souligne Stigler (1973), le concept n'est pas isolé ni nommé, et nous ajouterons surtout qu'il ne s'intègre pas comme chez Fisher à un modèle d'inférence.

Poursuivant notre chronologie, nous enjamberons ce demi-siècle, qui de 1830 à 1880, voit une prodigieuse avancée de la statistique administrative et de la statistique « morale », en particulier grâce à l'œuvre, en France et Belgique, de Quetelet puis des Bertillon, mais une relative stagnation de la statistique mathématique, surtout en ce qui concerne la théorie inférentielle. Une certaine déconsidération du calcul des probabilités, suivie par un enfermement dans le paradigme de la Théorie des Moyennes conduit les statisticiens à une fixation sur la distribution normale, conçue comme pierre de touche de l'homogénéité des populations, autour de moyennes qui sont autant de centres de gravités sur lesquels peut jouer la « physique sociale » de Quetelet. Les critiques ou les tentatives de dépassement de cette problématique par Cournot, ou Lexis et Dormoy par exemple, n'empêchent pas que toute variabilité soit plus ou moins considérée comme un bruit qui brouille cette mécanique et que seule peut réduire une multiplicité des observations. La loi des grands nombres évite alors de se poser des problèmes d'inférence.

Lorsque la statistique mathématique renaît avec l'école biométrique anglaise, c'est d'abord, au travers des travaux de Galton sur la corrélation (1885-89) sous la forme extrêmement empirique d'une judicieuse analyse de données; celle-ci a cependant presque tout oublié de la mathématique laplacienne, n'en retrouvant que quelques bribes — pour l'essentiel la loi devenue « normale » — et développant au lieu d'une stricte problématique d'estimation, une sorte d'analogie entre les figures de l'analyse empirique et les modèles probabilistes, entre par exemple un coefficient de corrélation empirique évalué par Galton selon une méthode géométrique, et le coefficient du terme produit de la loi binormale.

C'est Karl Pearson qui en quelques années (1892-1906) fera l'essentiel de ce travail de reconstruction mathématique de la statistique. De son œuvre, retenons d'abord, pour ce qui nous préoccupe, la construction d'une méthode d'ajustement de distributions empiriques à un système de courbes paramétrées par les moments ou des fonctions de ceux-ci (dissymétrie et applatissement), dont le principe central consiste à égaliser moments empiriques et moments théoriques en autant d'équations qu'il y a de paramètres à estimer. Méthode dont la simplicité assurera le succès, jusqu'à ce que Fisher n'en fasse une démolition systématique.

En ce qui concerne sa théorie de la régression (1896-98), l'estimateur du « moment-produit » de la corrélation d'une loi binormale, aussi bien que les variances et covariances des coefficients estimés de la régression, sont obtenus par la méthode laplacienne de la probabilité inverse. Pearson — comme Edgeworth d'ailleurs — critique certains aspects du principe de raison insuffisante (sa non-invariance à toute transformation fonctionnelle du paramètre par exemple) mais le justifie par des arguments de régularité statistique, et reprend à son compte la « règle de succession » de Laplace qui en est un corollaire; si l'on a déjà observé p succès au cours de n épreuves bernoulliennes, alors la probabilité d'un $(p+1)$-ième succès à l'épreuve suivante est $(p+1)/(n+2)$.

Il faut dire que cette règle lui sert essentiellement d'argument pour son thème favori de la « persistance des routines de perception », crédo de sa profession de foi phénoménaliste et idéaliste, écrite et répétée dans les éditions successives de la « Grammaire de la Science ». Grande est l'importance de cet ouvrage de philosophie des sciences, qui explicite le sens de bien des choix techniques de Pearson : volonté de s'en tenir aux seuls observables, refus de modèles réalistes, substitution de la contingence à la causalité, importance de la répétition de phénomènes similaires comme seule base aux « lois » scientifiques, d'où prédilection pour les grands échantillons et non-distinction formelle entre échantillon et population.

Lorsque Fisher arrive sur la scène mathématique, les travaux quelquefois dissidents de Yule ou Edgeworth n'ont guère entamé le bastion qu'est devenu le laboratoire de Pearson. Seul, un obscur ingénieur des brasseries, quelques temps stagiaire chez Pearson, et signant ses articles avec le pseudonyme de « Student », ouvre des perspectives de renouvellement que Fisher va saisir. Mais avant d'en venir à cette première intervention de Fisher, qui servira à la fois de préambule à ses travaux sur l'estimation et de premier prétexte à un conflit avec Pearson, il n'apparait pas inutile de donner quelques éléments biographiques.

2. *Éléments biographiques*

Les quelques éléments de la biographie de Fisher, que l'on pourra compléter par l'introduction de Yates et Mather aux « Collected Papers » de Fisher (1971), ou Kendall (1963), ou bien encore le livre de sa fille Joan Fisher-Box (1978), permettront de situer sa contribution à la théorie de l'estimation dans l'ensemble de son œuvre.

Ronald Aymler Fisher (1890-1962) est le benjamin des sept enfants d'un commissaire priseur londonien, et lui-même aura de son mariage en 1917 huit enfants; c'est là sans doute sa contribution au redressement de la natalité dans sa classe sociale, dont le faible taux de reproduction est une motivation essentielle de sa position eugéniste.

Il se montre très vite doué de vives capacités d'abstraction, qu'une très forte myopie semble plutôt favoriser qu'handicaper. Des études de mathématique et de physique l'amènent à une thèse d'optique et un diplôme de « Gonville and Caïus College », et lui permettent d'enseigner dans diverses public schools, tout en s'intéressant de près à la biologie et à l'eugénique, ne serait-ce qu'au travers des écrits de Karl Pearson, qui presque tous associent intimement problématique de l'hérédité, politique eugéniste, et instrumentation mathématique. Fisher associe également ces disciplines dans sa vie professionnelle et militante, mais les dissocie dans des articles spécialisés : ses premières publications concernent successivement ses intuitions d'étudiant sur le maximum de vraisemblance (1912), l'interprétation géométrique des principaux concepts de la statistique (1913), sa première profession de foi eugéniste (1914), et la réintégration des facteurs mendéliens et des notions de dominance et ségrégation dans les modèles biométriques de l'hérédité (1918, 1919).

Face à deux propositions de poste, l'une au « Galton Laboratory » sous la direction de Karl Pearson, l'autre à « Rothamsted Experimental Station » dirigée par John Russel, Fisher opte pour la seconde, qui lui offre davantage de garanties d'indépendance, ainsi qu'un grand nombre de relevés en attente de traitement statistique.

S'ouvre alors pour lui une période extrêmement productive, comptant plus de deux cents articles, dont quelques-uns marquent dès le début un progrès considérable par rapport aux travaux de l'école biométrique. Il s'agit d'abord des deux papiers sur l'estimation (1922 et 1925) qui forment le noyau du corpus que nous analyserons. Il s'agit ensuite, mais parallèlement, des textes dans lesquels Fisher établi les distributions exactes des statistiques d'échantillonnage les plus usitées par la biométrie : moyennes et variances empiriques (1920), coefficient de corrélation linéaire (1915, 1921), coefficients de la régression linéaire (1922), coefficients de corrélation partielle (1924) et multiple (1928). Il s'agit enfin de travaux sur le khi-deux qui corrigent son nombre de degrés de liberté dans une table de contingence, et généralisent son emploi comme critère d'ajustement (1922, 1928).

La nature des données agronomiques expérimentales de Rothamsted et la réhabilitation d'une causalité additive le conduisent à faire de l'analyse de variance le principe de la séparation des sources de variation. Dès la première édition de « Statistical Methods for Research Workers » (1925), le principe de randomisation est introduit, tandis que la notion de plan factoriel émerge dans les années qui suivent, et se trouve intégrée avec les premières analyses de covariance dans la quatrième édition (1932) du même ouvrage.

Parallèlement à ces travaux statistiques, Fisher publie beaucoup sur les thèmes alors fort débattus de la transmission héréditaire des caractères, de la sélection darwinienne, des mécanismes génétiques et des politiques eugénistes qui tirent argument de ces recherches, prenant une part importante dasn l'« Eugenic Society of London » après avoir fondé lui-même dès 1911 celle de Cambridge University. Il publie en 1930 une synthèse de ses travaux : « The genetical Theory of Natural Selection ».

Lorsqu'en 1933, Karl Pearson prend sa retraite, l'institution biométrique éclate en une chaire d'Eugénique, attribuée à Fisher en même temps que la direction de la revue « The annals of Eugenics », et une chaire de Statistiques — la première — attribuée à son fils Egon Pearson en même temps que la direction du laboratoire et de la revue « Biometrika ». La deuxième guerre mondiale, provoquera l'évacuation de London University College, mettra fin à cette cohabitation tout aussi explosive de Fisher avec Pearson fils, et remettra Fisher sur le chemin de Rothamsted. Deux publications importantes marquent cette période agitée : « The Design of Experiment » (1935) et, avec Yates, les « Statistical Tables for Biological Agricultural and Medical Research » (1938).

De 1943 à 1959, Fisher occupera la chaire de Génétique de Cambridge; et tout en complétant certains points de son modèle statistique, ou en peaufinant les synthèses — « Statistical Methods and Scientific Inference » (1956), 13ᵉ édition de « Statistical Methods for Research Workers » (1958) — il publie pour l'essentiel les résultats de ses travaux de génétique (les plus fameux portent sur les groupes sanguins humains et la théorie du facteur rhésus). Son intervention dans la politique sanitaire prit encore une forme controversiale, voire provocatrice, avec ses thèses non orthodoxes sur la cigarette et le cancer (1958, 1959).

Fisher passa les trois dernières années de sa vie en Australie, à l'Université d'Adélaïde, où il avait accepté un poste de chercheur, et où furent édités, post mortem, ses « Collected Papers » (1971). C'est à eux (notés CP), ou bien encore au choix d'articles des « Contributions to Mathematical Statistics » (1950, notés CMS) que renvoie la pagination des extraits que nous citons, dans une traduction française de notre cru.

3. *Distributions exactes et petits échantillons*

Le premier point de rupture par rapport à la statistique de l'école biométrique se cristallisera sur le problème de la distribution du coefficient de corrélation empirique r. Pearson lui avait donné sa formule du « moment-produit » et son statut d'estimateur du coefficient de corrélation d'une loi normale bivariée. Mais bien souvent, la mesure lui suffisait pour des jugements sommaires (une corrélation faible, ou forte, ou de l'ordre de...) sans qu'il éprouve le besoin d'un verdict de significativité; et quand il donnait un « intervalle de confiance », c'était sur la base d'une loi asymptotiquement normale de cet estimateur d'écart-type $(1-\rho^2)/\sqrt{n-1}$, justifiée par la taille importante de ses échantillons.

C'est Gosset, alias « Student », qui se préoccupe le premier en 1908 de la distribution exacte de la moyenne arithmétique d'un petit échantillon d'une variable normale, et dans un second article publié la même année, attire l'attention sur la dissymétrie de la distribution de r pour une corrélation théorique ρ non nulle. C'est la première fois, dans la statistique anglaise tout au moins, que des notations distinctes désignent les paramètres de la population et les statistiques d'échantillonnage associées. Il est remarquable aussi que l'origine de cette préoccupation soit dans la nature différente des données que Student avait à traiter dans la brasserie qui l'employait, et que ce soit le coût de production de ces données et leur enjeu décisionnel qui rendent inopérant sur ce point l'outil statistique forgé par la biométrie :

> *« Le travail de la Brasserie Expérimentale qui concerne des choses telles que le rapport entre l'analyse du malt ou du houblon et le comportement de la bière, et qui prend un jour pour chaque élément d'observation, en limitant ainsi le nombre, demandait une réponse à des questions du genre : si, avec un petit nombre de cas, j'obtiens une certaine valeur de r, quelle est la probabilité qu'il y ait une corrélation supérieure à telle valeur. » (Student, 1915, lettre à Fisher, in Pearson et Kendall)*

Pearson reçoit un premier papier de Fisher qui confirme les observations de Student, mais va plus loin puisqu'il parvient, grâce à des considérations géométriques, à la distribution exacte de r, laquelle ne correspond à aucune des courbes du système paramétrique de Pearson. L'estimateur de ρ n'est plus r, mais $r/(1 + (1-r^2)/2n)$. Et Fisher termine son article par deux tentatives de transformations fonctionnelles de r — soient $r/\sqrt{(1-r^2)}$ et $z = \arg \operatorname{th} r$ — qui visent la construction d'une variable dont la loi soit stable lorsque ρ varie de -1 à $+1$. Seule la première est étudiée à ce premier stade.

Pearson publie ce papier dans son journal Biometrika en 1915, mais, peu satisfait de la valeur opérationnelle des dites transformations, il lance toute l'équipe de son laboratoire dans la tâche gigantesque que constitue la tabulation de la distribution de r pour différentes valeurs de ρ et de n. Les premiers résultats de cette entreprise sont publiés dans Biometrika en 1917 (« A Cooperative Study »), assortis d'une critique vigoureuse de la méthode de Fisher, telle qu'elle apparaît dans le papier de 1912 et la note de 1915, méthode que Pearson assimile à la maximisation de la probabilité inverse, sous l'hypothèse — absurde pour une corrélation — d'une « égale distribution de l'ignorance ».

C'est le début d'un malentendu sur le maximum de vraisemblance dont nous allons étudier les différents aspects, et qui prend la forme d'une controverse ouverte avec Pearson, d'une violence et d'une durée exemplaires. Fisher envoie à Pearson un second papier qui résoud le problème grâce à la transformation en tangente hyperbolique, réexaminée avec succès : elle conduit en effet à une variable z distribuée normalement autour de la valeur

$$\arg \operatorname{th} \rho + \rho/2(n-1) \text{ avec un écart-type de } 1/\sqrt{(n-3)},$$

et ce pour des échantillons assez petits puisque l'approximation est considérée comme suffisamment bonne dès $n = 10$.

Ces résultats rendent caduques, parce qu'inutiles, les laborieuses tabulations de l'équipe de Pearson. Ils sont de plus accompagnés d'une première mise au point de Fisher sur sa méthode sous la forme d'une « Note sur la confusion entre la règle de Bayes et ma méthode de l'évaluation de l'optimum », peu tendre pour les « distingués statisticiens de l'étude coopérative ». La réponse de Pearson est un refus de publier :

« Je ne peux malheureusement pas publier votre papier sous la forme actuelle, sans une réponse à vos critiques, et donc aussi une critique de votre travail de 1912. Je préférerais que vous publiiez ailleurs. Dans les conditions éditoriales et financières actuelles, je suis au regret d'exclure tout ce que je considère, de mon propre jugement, comme erroné, car je ne peux souffrir de controverse. »

Fisher publiera ailleurs (1921), et fournira d'autres contributions sur les distributions exactes de statistiques. D'abord celle du rapport de corrélation (1922 et 1928); puis celle des coefficients d'une régression linéaire (rapportés à leurs écarts-types estimés) selon une loi de Student à $n-p$ degrés de liberté, d'après les indications de ce dernier (1922); tandis que sur ce même problème, Pearson publie encore, quatre ans plus tard, sur le comportement asymptomatique des mêmes coefficients.

La rupture de 1920 entre Fisher et Pearson a donc dès l'origine deux composantes. L'une est le renouvellement de l'outillage statistique que nécessite la prise en considération des *petits échantillons* que Fisher et Student rencontrent dans les essais expérimentaux, et qui est étrangère aussi bien à la pratique de Pearson — « il n'y a que les brasseurs pour s'intéresser à cela » dit-il — qu'à sa philosophie d'une science descriptive des seules régularités que permettent la répétition des « routines de perception ». L'autre est le *désaccord sur la similitude ou non de deux syntaxes mathématiques*, l'une bayesienne, l'autre introduisant le concept de vraisemblance, désaccord entretenu aussi bien par l'entêtement de Pearson que par l'ambiguïté des premières formulations de Fisher. Il faudrait encore ajouter à ces deux motifs de discorde la réintégration par Fisher, dans l'approche biométrique de l'hérédité, des facteurs mendéliens, c'est-à-dire d'une causalité et d'un réalisme qui contreviennent à la conception pearsonienne de la science.

4. *Vraisemblance et probabilité inverse*

La construction de la vraisemblance par Fisher se fait selon une progression qui est la suivante.

Elle est d'abord *la méthode du maximum de vraisemblance*, qui pour exister, doit se distinguer de la méthode de Bayes dont elle est formellement très proche.

Puis elle est construite comme méthode d'estimation surclassant ses concurrentes — la méthode des moments par exemple — grâce à ses *propriétés syntaxiques* et à celles des estimateurs qui lui sont associés.

Enfin la vraisemblance en tant que telle émerge petit à petit comme *concept*, indépendant de la méthode, et occupant une position clé, symétrique à celui de probabilité, dans le modèle plus large de l'inférence inductive.

Certes, il ne s'agit pas d'une progression dont les phases seraient bien isolées dans les textes de Fisher, car il anticipe dès les premiers textes sur son modèle inférentiel général et, vice-versa, il revient sans arrêt, même tardivement, sur cette première phase de dégagement du concept et de résolution des ambiguïtés. Mais c'est une progression qui se lit en tendance chronologique.

C'est en fait le tout premier texte de Fisher (1912) qui introduit le maximum de vraisemblance. Prenant le prétexte de l'ajustement d'une fonction de paramètres $\theta_1,...,\theta_n$ sur une distribution de fréquences observée, il fait une revue des techniques d'estimation de ces θ_i. Le verdict est rapide : la méthode des moindres carrés part d'un bon principe mais conduit le plus souvent à des équations insolubles; la méthode des moments est de nature totalement arbitraire. « Mais nous pouvons résoudre le problème directement »... en choisissant pour les θ, « l'ensemble de valeurs *le plus probable* » à savoir celui qui rend P maximum, P étant la fonction définie par $\log P = \Sigma \log f$, où f est l'ordonnée de la courbe de densité théorique que l'on ajuste. Quelques lignes plus loin, il est dit que la surface de P, dans le cas de deux paramètres, représente « le système de *la probabilité inverse* ».

Fisher payera cher cette terminologie des deux expressions que nous avons soulignées et qui se retrouvent aussi dans la « note » de 1915. C'est l'origine d'une confusion de sa méthode avec celle de Bayes qui est, on l'a vu, un des facteurs de la brouille avec Pearson. Pourtant, dès ce même texte de 1912, Fisher précise que l'intégration de P par rapport aux paramètres est « illégitime » et que « P n'est qu'une probabilité relative, utilisable dans une comparaison point par point, mais non susceptible d'une interprétation en terme de distribution de probabilité sur un domaine, ou en tant qu'estimation d'une probabilité absolue ».

Le texte des « Fondements » (1922) qui est, dira-t-il plus tard, « la première attaque sur une large échelle du problème de l'estimation », baptise P du nom de vraisemblance et consacre, après la mise au point déjà évoquée du texte de 1921, de nouveau quatre pages au distinguo qui doit être fait entre sa méthode et celle de Bayes. Et il reviendra sur ce point dans presque tous les textes sur l'estimation (par exemple 1925, 1930, 1932, 1933, 1934...):

> *(...) « formellement [ma méthode du maximum de vraisemblance] ressemble au calcul du mode d'une distribution de probabilité inverse. Cette ressemblance est tout à fait superficielle : si l'échelle de mesure de la quantité hypothétique est modifiée, le mode changera de position, et peut donc donner la valeur que l'on veut par un changement d'échelle approprié; mais l'optimum, si l'on désigne ainsi la position du maximum de vraisemblance reste inchangé dans une telle transformation. La vraisemblance diffère donc de la probabilité en cela qu'elle ne comporte pas d'élément différentiel, et n'est pas intégrable : elle caractérise un point de l'intervalle de variation, et non un sous-intervalle élémentaire. Il y a une mesure absolue de probabilité puisque l'unité est choisie de telle façon que la somme des probabilités élémentaires soit égale à 1. Mais il n'y a pas une telle mesure absolue de vraisemblance. » (1922, p. 327)*

Et l'exemple est donné de l'estimation d'une proportion p, où les deux méthodes conduisent à maximiser la quantité $Cp^x(1-p)^y$, où C est une constante, et qui est la vraisemblance d'un échantillon donnant lieu à x succès et y échecs, mais, multipliée par dp, devient la probabilité a posteriori que p appartienne à l'intervalle $[p, p+dp]$ selon le théorème de Bayes *et* sous l'hypothèse d'une densité à priori uniforme pour p.

> « Deux concepts radicalement distincts, tous deux d'une grande importance dans l'influence de nos jugements, ont été confondus sous le nom unique de probabilité. Nous proposons d'utiliser le terme de vraisemblance pour désigner l'état de notre information au sujet des paramètres des populations hypothétiques... » (1922, p. 367, conclusion)
>
> « Vraisemblance n'est pas un synonyme de probabilité, et c'est une quantité qui n'obéit pas aux lois du calcul des probabilités. » (1925, p. 707)
>
> « Connaissant la population, nous pouvons exprimer notre connaissance incomplète ou notre attente de l'échantillon en terme de probabilité; connaissant l'échantillon, nous pouvons exprimer notre connaissance incomplète de la population en terme de vraisemblance. » (1930, p. 532)

Pour être plus clair encore, c'est à l'association de plusieurs propositions que Fisher s'en prend sous le nom de méthode de Bayes :

1. Le *théorème de Bayes* qui relie $p(A/B)$ à $p(B/A)$, sa généralisation à une famille finie d'événements A_i dans le cas discret :

$$\text{prob}(A_i/B) = \text{prob}(B/A_i) \cdot \text{prob}(A_i) \; / \; \Sigma_i \text{prob}(B/A_i) \cdot \text{prob}(A_i)$$

et son extension au cas continu par intégration :

$$g(\theta)\,d\theta = f(x,\theta) \cdot h(\theta) \,/\, \int_0^1 f(x,\theta)\, n(\theta) d\theta$$

2. Le transport de ce théorème, des événements « subséquents » de Bayes à la problématique de l'estimation et aux *hypothèses paramétriques*, ce qui suppose qu'un ensemble de valeurs possibles du paramètre est susceptible d'un jugement, a priori ou à posteriori, en terme de mesure de probabilité, bref, qu'on a « le droit » d'écrire prob $(\theta_1 < \theta < \theta_2)$.

3. Le *postulat de répartition uniforme de la loi à priori*.

4. La transformation de ce postulat en scholium par Bayes, puis en *« principe de raison insuffisante »* par Laplace, qui traduit l'ignorance d'une distribution en une connaissance d'une distribution uniforme.

5. La méthode d'estimation qui consiste à choisir la valeur du paramètre qui maximise la distribution a posteriori, pour un échantillon donné, c'est-à-dire le *mode* de cette distribution.

Fisher casse cette association. Certes le théorème des probabilités conditionnelles est hors de cause. Mais il n'est pas question pour lui — dans les années 20 — de probabiliser l'ensemble des valeurs du paramètre : cela est contraire à sa conception fréquentiste de la probabilité :

« Je trouve que le mot de probabilité est mal utilisé dans un tel contexte. La probabilité est un rapport de fréquences et nous ne pouvons rien savoir ici de ces fréquences. » (1922, p. 326)

Si toutefois nous acceptons cette probabilisation de l'ensemble des valeurs du paramètre (par exemple par l'argument d'un tirage dans une superpopulation), il reste que la maximisation du mode de la distribution a posteriori est une méthode dont il redémontre après Boole, Edgeworth et Pearson qu'elle n'est point invariante à une transformation fonctionnelle, c'est-à-dire que le mode de la distribution de p n'est pas le même que celui de p^2 (Edgeworth) ou Arcsin $(2p-1)$ (Fisher, 1922), sauf à considérer que notre ignorance sur p se traduit dans le premier cas par une distribution uniforme de p et dans le second par une distribution uniforme de la fonction de p, ce qui est pour le moins paradoxal.

Dans un texte de 1930 consacré à la probabilité inverse il démonte une fois de plus cette association de deux procédures arbitraires qui composent la méthode de la probabilité inverse :

« Deux éléments totalement arbitraires dans ce processus se sont en fait éliminés mutuellement, la technique non invariante [pour une transformation du paramètre] qui consiste à prendre le mode, et l'hypothèse arbitraire que la densité de la distribution a priori est constante « ... » En fait aucune de ces deux procédures n'a de justification, mais chacune est nécessaire pour éliminer les erreurs introduites par l'autre. » (1930, p. 531)

De fait cette doctrine de la raison insuffisante est une ineptie qui cache l'arbitraire le plus complet. La preuve en est déjà dit-il dans plusieurs textes, que Bayes n'osa point publier son essai de son vivant, « ayant des doutes sur la vérité de son axiome », ce qui est une interprétation fort libre des faits. Mais il accuse principalement Laplace :

(...) « l'introduction par Laplace dans sa définition de la probabilité de l'expression non élucidée « cas également possibles », puisque nous devons savoir quels cas sont également possibles avant de savoir qu'ils le sont, ne peut mener qu'à la « doctrine de la raison insuffisante », à savoir que tous les cas sont également probables (pour nous), à moins que nous ayions des raisons de supposer le contraire, et réduit ainsi toute probabilité à un jugement subjectif. » (1930, p. 528)

Edwards (1972) dans son livre consacré à la vraisemblance, est plus catégorique encore :

« Du point de vue de la méthode du Support [logL], le concept de distribution de probabilité a priori traduisant l'ignorance n'est pas nécessaire, et de fait, apparaît impliquer une contradiction dans les termes. Tout énoncé probabiliste est informatif. Il n'est pas surprenant donc que le concept ne résiste pas à l'examen. »

Concluons ce chapitre anti-bayésien de Fisher par les quatre premières lignes de ce même texte de 1930 :

« Je ne connais qu'un cas en mathématique d'une doctrine qui a été acceptée et développée par les hommes les plus éminents de leur époque, et est encore peut être acceptée par des contemporains, qui en même temps est apparue à une succession d'écrivains de valeur comme fondamentalement fausse et dénuée de fondements. » (1930, p. 528)

La persistance d'une telle erreur est attribuée à « l'illustre autorité de Laplace », au « conservatisme de l'enseignement mathématique » que Fisher lui-même a subi (1936, p. 246), mais surtout au besoin d'inférence rigoureuse, pour lequel la probabilité était inadaptée bien que contrainte et forcée vers cet objectif :

« La probabilité inverse a, je crois, survécu si longtemps en dépit de ses bases insatisfaisantes, parce que ces critiques n'ont, jusqu'à tout récemment, rien proposé pour la remplacer en tant que théorie rationnelle de l'apprentissage par l'expérience. » (1930, p. 531)

« Bien que quelquefois, l'inférence incertaine ait pu être exprimée en terme de probabilité mathématique, il ne s'ensuit pas que la probabilité mathématique soit un concept adéquat pour l'expression rigoureuse de toutes les sortes d'inférences incertaines. » (1935, p. 40)

Le maximum de vraisemblance s'est ainsi construit d'abord *contre* la probabilité inverse, et nous avons montré le travail fourni par Fisher pour l'en démarquer. Mais c'eût été bien sûr insuffisant, s'il ne s'était également construit *pour* ses propres propriétés. Or ces propriétés ne peuvent se justifier qu'à travers un modèle statistique, qui est le cadre de leur validité et de leur interprétation.

5. Le modèle statistique

La présentation de la méthode du maximum de vraisemblance en 1912 ne s'accompagne d'aucune propriété. Elle ne s'inscrit dans aucune théorie générale de l'Estimation. C'est encore un objet mathématique isolé.

C'est le texte des fondements (1922) qui pose clairement, et pour la première fois, ce qu'est l'estimation, la définissant par un double rapport. D'abord un rapport au tout que forme la méthode statistique. Ensuite, un rapport à chacun de ses éléments, à savoir les critères selon lesquels les produits de la méthode — les estimateurs — doivent être évalués.

Les idées de Fisher sur la méthode statistique sont étonnamment claires et réorganisent celle-ci sur la base de deux idées simples :

1. « L'objet de la méthode statistique est la réduction des données. Une masse de données, si importante qu'elle en est inintelligible, doit être remplacée par un relativement petit nombre de quantités qui doivent représenter correctement cette masse, ou, en d'autres mots, doivent contenir la plus grande part possible, si ce n'est la totalité de l'information pertinente contenue dans les données d'origine. » (1922, p. 311)

2. « Cet objectif est accompli par la construction d'une population infinie hypothétique, dont les données observées sont considérées comme un échantillon aléatoire. » (id.)

Enfin, trois phases, et donc trois types de problèmes différents, caractérisent cette coopération de réduction de données :

« 1) Des problèmes de Spécification. Ils apparaissent à travers le choix de la forme mathématique de la population.

2) Des problèmes d'Estimation. Ils impliquent le choix de méthodes de calcul de quantités dérivées de l'échantillon, que nous appellerons statistiques, et qui sont construites pour estimer les valeurs des paramètres de la population hypothétique.

3) Des problèmes de Distribution. Ceux-ci incluent des discussions de la distribution de statistiques d'échantillonnage, ou plus généralement de toutes les fonctions ou quantités dont la distribution est connue. » (1922, p. 313)

De ces définitions de base, qui aujourd'hui nous semblent naturelles parce qu'elles forment toujours l'essentiel du paradigme à l'intérieur duquel nous pensons notre démarche statistique, il faut bien se rendre compte qu'elles opèrent une clarification, voire une rupture par rapport aux pratiques antérieures, dont Fisher dénonce le flou méthodologique. C'est bien une théorie de l'estimation qu'il propose, là où ses prédécesseurs ont accumulés des méthodes pratiques, mais aussi des paradoxes non résolus, faute de raisonner à l'intérieur d'un modèle.

Placer *la réduction de données* au centre de la méthode statistique n'est pas révolutionnaire en soi : tous ceux qui ont calculé des moyennes, des erreurs probables, et des corrélations empiriques avant Fisher faisaient argument de cette réduction.

La théorie de la Moyenne qui est la quintescence de l'astronomie, puis de la statistique morale, est un hymne au pouvoir révélateur d'une telle réduction. Mais l'idée est alors plutôt celle d'une épuration que celle d'une perte d'information : la moyenne, de Laplace à Quetelet, doit permettre de dégager le bon grain de l'ivraie, les causes constantes des causes accidentelles, les lois de la mécanique sociale du bruit dont chaque individu les brouille. À la limite, une telle opération de réduction dégage de l'information pour un statisticien comme Quetelet. Cette idée d'épuration n'est pas non plus absente chez Fisher, face à la redondance d'un échantillon :

« Puisque le nombre de faits indépendants fourni par les données est beaucoup plus grand que le nombre de faits recherchés, une grande part de l'information fournie par un échantillon donné est inutile. C'est l'objet du processus statistique employé dans la réduction de données que d'exclure l'information non pertinente et d'isoler la totalité de l'information pertinente contenue dans les données. »

Cette *notion* d'information, qui deviendra *mesure* puis *concept*, est, chez Fisher, relative à une population hypothétique. Des trois phases du travail statistique données par Fisher, on voit qu'en dehors de l'insistance déjà vue sur les distributions exactes, la plus nouvelle est celle de *Spécification*, et il faut s'arrêter un moment sur cette notion de population hypothétique infinie qui la caractérise. C'est une notion dont le niveau d'abstraction a nécessité plusieurs mises au point de sa part (1925). La population n'est plus l'ensemble ou l'urne dans lequel on fait ou on imagine un tirage, mais la famille de lois de probabilité paramétriques caractérisant la (ou les) variable(s) dont les réalisations forment l'échantillon. Cette population hypothétique résume en sorte tout ce que nos observations possèdent en commun, ou plus précisément ce que nous supposons qu'elles possèdent, c'est-à-dire encore qu'elle explicite l'information a priori dont nous acceptons de tenir compte :

« Il n'y a pas d'erreur à interpréter tout ensemble de mesures indépendantes en terme d'échantillon aléatoire d'une population infinie; car de tels ensembles de nombres sont bien des échantillons aléatoires pris dans la totalité des nombres produits par la même matrice de conditions causales : la population hypothétique que nous étudions est un aspect de la totalité des effets de ces conditions, de quelque nature qu'elles soient. » (1922, p. 313)

Il y a donc une sorte de définition mutuelle des trois concepts de population, échantillon et probabilité.

« Nous n'avons besoin que du concept de population hypothétique infinie, en relation avec celui d'échantillon aléatoire. En dernier recours, l'élucidation d'une idée implique logiquement l'autre. » (1925, p. 700)

Que la population soit infinie est une nécessité absolue car elle permet d'une part de construire la probabilité d'un événement comme limite d'une fréquence relative, et d'autre part elle assure que les n observations de l'échantillon puissent être considérées comme les réalisations de variables indépendantes et de même loi.

« Par exemple, quand nous disons que la probabilité d'obtenir un cinq dans le jet d'un dé est 1/6, nous ne devons pas nous laisser aller à penser que sur 6 lancers de ce dé, un et un seul sera nécessairement un cinq; ni que sur 6 millions de lancers, exactement un million de cinq seront obtenus; mais que sur une population hypothétique d'un nombre infini de lancers du même dé, exactement un sixième sera constituée de cinq. » (1922, p. 312)

Fisher ne s'écartera jamais, dans cette version déductive de la statistique qui va de la population hypothétique à l'échantillon, de cette conception fréquentielle de la probabilité. Sa terminologie même, qui désigne par « distribution de fréquence » ou « courbe de fréquence » ce que nous nommons distribution ou densité de probabilité, ne fait pas ambiguïté pour lui, puisqu'une probabilité n'est qu'une fréquence dans une population infinie.

Une des applications de son modèle statistique est ainsi la réinterprétation de la notion bayesienne d'a priori en terme de super-population : la seule façon à son avis de justifier que l'espace des paramètres soit susceptible d'une mesure de probabilité, est de supposer que la population d'où sont issues les observations — une famille de lois de Bernoulli paramétrée par p, par exemple — est elle-même tirée dans une superpopulation qui exprime une certaine distribution de p. Si cette superpopulation est spécifiée d'une quelconque façon, les bayesiens deviennent autorisés à appliquer la méthode de la probabilité inverse et à user de la probabilité qui dans cette démarche rendue déductive, peut garder son fondement fréquentiel. Mais cela ne fait que repousser un peu plus loin le problème de la spécification.

Après avoir vu comment l'estimation se définit à l'intérieur de l'idée plus générale de modèle statistique, il faut en venir à la batterie de critères qui vont permettre de choisir une statistique comme estimateur d'un paramètre de la population hypothétique.

Dès le texte de Fondements, sont énumérés les trois critères selon lesquels un estimateur peut être évalué :

— *« Une statistique satisfait le critère de convergence [« consistency »] si, quand elle est calculée sur toute la population, elle est égale au paramètre à estimer. »*

— *« La mesure de l'efficacité [« efficiency »] d'une statistique est le rapport de sa précision intrinsèque à celle de la statistique la plus efficace. Il exprime la proportion de l'information utile totale que la statistique utilise. » « Le critère d'efficacité est satisfait par les statistiques, qui calculées sur de grands échantillons, tendent vers une distribution normale avec le plus petit écart-type possible. »*

— *« Une statistique satisfait le critère d'exhaustivité [« sufficiency »] si aucune autre statistique calculée sur le même échantillon ne fournit d'information supplémentaire sur la valeur du paramètre à estimer. »*

Comment sont construites mathématiquement ces propriétés et par quelle syntaxe sont-elles associées à la vraisemblance, c'est ce que nous allons maintenant étudier.

6. *Biais, convergence et critique de la méthode des moments*

La notion de convergence est, dit Fisher, la seule qui ait, jusqu'ici, été prise en compte dans le choix d'un estimateur.

Fisher la définit d'abord comme le fait que « l'application de la statistique à l'ensemble de la population doit être égale au paramètre à estimer (ou une fonction de celui-ci) » (1922, p. 316).

En 1925, la définition est plus précisément celle d'une convergence en probabilité :
« Une statistique est dite estimation convergente d'un paramètre si, lorsqu'elle est calculée sur un échantillon infiniment grand, elle tend à être égale au paramètre. »
Ce qui se ramène à deux conditions :
1 : il existe une limite T_0 vers laquelle tend l'estimateur $T(X)$ du paramètre θ ou de $g(\theta)$
2 : cette limite T_0 doit être égale à θ (ou une fonction $g(\theta)$)
Certains statisticiens (Tassi, 1985) donnent le nom de « consistance » à la première notion de convergence — plus forte puisqu'elle n'est pas asymptotique mais s'applique à la population totale; consistance et convergence étant d'ailleurs équivalentes si T est continue.

La méthode des moments est typiquement, dit Fisher, une méthode fondée sur ce seul critère de convergence, puisqu'elle repose sur la résolution d'un système d'équations dont chacune égalise moment théorique et moment théorique d'un certain ordre, et que l'on sait que les moments empiriques d'un échantillon de variables indépendantes et de même loi convergent vers les moments théoriques correspondants. Mais cela n'empêche pas Fisher de critiquer très sévèrement cette méthode, tout au long de sa carrière, et ce avec de plus en plus de véhémence.

La première critique porte sur le choix arbitraire des moments qui sont choisis : leur nombre doit être égal à celui des paramètres à estimer, mais ils peuvent être centrés ou non centrés et rien n'indique que si il en faut *n*, ce soient les *n* premiers.

Mais surtout les conditions de l'application de la méthode la rendent inutilisable dans de nombreux cas. Ainsi :

« Le système de courbes développé par Pearson a quatre paramètres, et peut être ajusté par le moyen des quatre premiers moments. Pour cela, il faut se restreindre aux courbes dont les quatre premiers moments sont finis. De plus, si la précision de ces quatre moments doit augmenter avec la taille de l'échantillon, c'est-à-dire si leur erreur probable doit être non infinie, ce sont les huit premiers moments qui doivent être finis. Cette restriction nécessite que la classe de distributions pour lesquelles cette condition n'est pas remplie devra être exclue comme « hétérotypique » et que les quatre moments deviennent sans valeur dans cette classe. Il devrait être clair pourtant qu'il n'y a rien d'anormal au sujet de ces prétendues distributions « hétérotypiques » si ce n'est que la méthode des moments ne s'applique pas à elles. » (1922, p. 321)

L'exemple fétiche de Fisher, qui lui permet de ridiculiser la méthode, est celui de la loi de Cauchy, de densité $1/\pi(1+(x-m)^2)$, dont le paramètre de valeur centrale m ne saurait être estimé valablement par le moment empirique correspondant x, moyenne arithmétique de l'échantillon. Car celle-ci ayant la même loi qu'une observation quelconque de l'échantillon, « nous ne nous approchons pas plus de m en prenant la moyenne de 100 valeurs de x, que si nous choisissions l'une quelconque des 100 valeurs » (1922); bref l'estimateur des moments n'est ici même pas convergent.

[C'est seulement] « dans quelques cas spéciaux que le critère de convergence permet seul d'obtenir une solution complète, si par exemple le nombre de classes de fréquences est égal au nombre de paramètres plus un; alors, il existe une seule statistique convergente, et celle-ci est la seule utilisable. Généralement cependant, il y a un grand nombre de statistiques possibles, toutes convergentes, mais d'aucune façon équivalentes. » (1925, p. 703)

Mais bien sûr la critique essentielle de Fisher porte sur l'inefficacité de la méthode des moments dans les cas où elle est applicable. Le papier de 1922 est d'ailleurs encombré de « longues digressions comme celles donnant le traitement exact des corrections de regroupement [*cf.* Shepard] ou celles dans lesquelles l'efficacité de l'ajustement de courbes de Pearson par les moments est examinée » (Note de Fisher sur l'article de 1922). Au bout de plus de vingt pages d'analyse, Fisher conclut que c'est seulement pour une petite région au voisinage de la distribution normale que la méthode des moments est efficace.

La critique technique de la méthode des moments va se transformer petit à petit en conflit ouvert avec Pearson, jusqu'au paroxysme du papier de 1937, écrit au lendemain de la mort de Pearson et, dès son titre, personnellement dirigé contre lui : « professeur Karl Pearson et la méthode des moments ».

C'est une réponse à un dernier papier de Pearson « plein d'amertume et d'attaques véhémentes vis-à-vis d'un écrivain indien, R.S. Koshal, dont l'offense a été de décider après expérience, qu'une courbe de Pearson de type I était mieux ajustée à ses données par la méthode du maximum de vraisemblance que par celle des moments ». Fisher s'autorise à une attaque en règle :

« Pendant ces dernières années, les intimes de Pearson avaient pensé que l'irritation et la controverse pouvaient être dangereux pour sa santé. En conséquence, la discussion de plusieurs points avait été suspendue. » Avec la mort de Pearson, mon obligation d'examiner franchement le statut des méthodes pearsoniennes se réaffirme... » (1937, p. 303)

On peut résumer ainsi les griefs de Fisher dans cet article :

— *« La simple procédure consistant à égaliser les moments n'est pas même due à Pearson, mais remonte à Bessel et Gauss, et fut largement développée par Thiele. »*

— *« Pearson n'a jamais fourni de méthode générale adéquate pour ajuster des données groupées par la méthode des moments. »*

— *« La méthode des moments peut être mieux appliquée qu'il ne le fait lui-même. Par sa méthode, qui modifie à tort les corrections de Sheppard, Pearson obtient des estimateurs qui ne sont même plus convergents. »*

— *« On sait de façon indiscutée, quoiqu'ignorée, que la solution des équations aux moments est elle-même inefficace, c'est-à-dire qu'elle introduit des erreurs d'estimation comparable aux erreurs inévitables de l'échantillonnage. »*

— *« Les biologistes n'ont que peu de raisons de croire que la population sondée suit véritablement une loi correspondant à une courbe de Pearson. »*

— *« Un grand nombre de départements de statistique ont donné une place considérable à cette méthode dans leurs programmes. Il faut rappeler que cette place lui est attribuée aux dépens d'autres sujets qui nécessitent aussi du temps pour être dominés. »*

Et Fisher de citer ses propres méthodes : différences finies, traitement des petits échantillons, analyse de variance, théorie de l'estimation. Il rappelle les multiples mises au point qu'il a déjà faites sur la méthode des moments, en particulier dans son texte des « fondements », auxquelles Pearson en répondit jamais, ou bien encore les incessantes mésinterprétations auxquelles donnèrent lieu sa propre méthode : « Pearson ne s'est jamais donné les moyens de critiquer la théorie de l'estimation, voire même de la comprendre ».

A travers ce ton de règlement de compte, il nous semble qu'il faut voir plus qu'un trait de (mauvais) caractère, dont on pourrait d'ailleurs accuser les deux protagonistes. En rester à cette interprétation psychologique, c'est oublier que cette irritation a sa source dans le processus même de la science, c'est-à-dire la lutte — où tous les coups sont permis — entre deux modèles statistiques, deux configurations de forces à la fois cognitives et sociales, le système des publications que nous analysons n'étant que l'une des scènes où se joue ce conflit. Et Fisher enrage de n'avoir jamais fini de gagner bien qu'ayant le meilleur jeu de cartes. L'on voit ainsi que la force logique des arguments formels de la syntaxe mathématique, n'est qu'une force parmi toutes celles qui construisent l'autorité d'une théorie.

Si nous passons maintenant de la propriété de convergence à celle du biais, c'est pour noter que ce biais, écart espéré entre la vraie valeur du paramètre et son estimateur, n'inquiète pas longtemps Fisher et ne joue qu'un faible rôle dans sa théorie :

« La considération du biais ne nous retiendra guère. Avec des estimateurs convergents, celui-ci doit tendre vers zéro; si nous souhaitons utiliser notre estimateur pour des tests de signification, il est

bon qu'il tende vers zéro plus vite que $1/\sqrt{n}$. Nous pouvons toujours ajuster notre estimateur pour faire que le biais soit absolument nul, mais cela n'est pas nécessaire d'habitude, car estimant un paramètre, il faut se rappeler que nous estimons en même temps son inverse, ou son carré, ou n'importe quelle autre fonction, et un biais nul d'une de ces quantités implique en général un biais de l'ordre de 1/n pour les autres. » *(1935, p. 42)*

7. *Vraisemblance et exhaustivité*

C'est dans un article relativement concis de 1920, aux démonstrations géométriques saisissantes, que Fisher met en évidence le concept d'exhaustivité, bien que le terme n'apparaisse qu'en 1922.

Dans un cheminement très semblable à celui de Laplace que nous avons décrit plus haut (*cf.* aussi Stigler 1973), Fisher compare deux estimateurs de l'écart-type d'une loi normale :

— l'erreur absolue moyenne (mean error) :

$$\sigma_1 = \sqrt{(\pi/2)\cdot\Sigma|x - \bar{x}|/n}$$

— l'erreur quadratique moyenne, ou écart-type (mean square error) :

$$\sigma_2 = \sqrt{(\Sigma(x - \bar{x})^2)/n}$$

Démontrant au passage le théorème de l'indépendance de $\bar{x}$ et de σ_2, il établit séparément les distributions des deux estimateurs et en déduit leur moyenne et écart-type, exactement puis en valeur approchée pour de grands échantillons, soient :

$$E(\sigma_2^2) = \sigma^2(n-1)/n \quad \text{mais} \quad E(\sigma_2) \approx (1 - 3/4n)\sigma \quad \text{et} \quad V(\sigma_2) \approx \sigma^2/2n$$

$$E(\sigma_1) = \sqrt{((n-1)/n)}\sigma \quad \text{et} \quad V(\sigma_1) = \sigma^2(\pi - 2)/2n$$

Si bien que le rapport des variances des deux estimateurs pour de grands échantillons est de $\pi - 2$, ce qu'il interprète ainsi :

« ***Pour obtenir un résultat de précision égale avec la dernière méthode (mean error), il faut augmenter de 14 % le nombre d'observation*** » ***(1920, p. 762).***

Ce résultat est un exemple de ce que Fisher baptisera deux ans plus tard mesure d'efficacité. Mais, sans s'arrêter là, Fisher redémontre, après Gauss, que dans la formule de σ_2, le carré donne le moment empirique de variance minimale par rapport à toute autre puissance, ce qui justifie le choix d'une forme quadratique par d'autres considérations que la commodité. Enfin, il établit, certes pour la taille $n = 4$ seulement, la distribution conjointe, de forme assez complexe, des deux estimateurs, qui « seule, permet de connaître les effets d'un choix portant sur l'une ou l'autre des deux statistiques ».

Or « il existe dans la forme de cette surface de fréquence une distinction qualitative, qui révèle le caractère unique de σ_2. » (...) « Pour une valeur donnée de σ_2, la distribution de σ_1 est indépendante de σ, ... », alors qu'à l'inverse, « pour une valeur donnée de σ_1, la distribution de σ_2 dépend de σ. »

Donc « la valeur de σ_1 ne peut donner aucune information supplémentaire sur σ » que ne donne déjà σ_2, et il en serait de même dit-il de toute autre statistique que σ_2. Ainsi « la totalité de l'information sur σ qu'un échantillon fournit est-elle résumée dans la valeur de σ_2. » (1920)

Fisher conclut cet article sur le fait que « l'unique supériorité de σ_2 est dépendante de la forme de la courbe normale », qui est la loi supposée des observations, et que la loi qui jouerait un rôle analogue pour la statistique σ_1, ne serait rien d'autre que la première loi de Laplace (1774) en $\exp(-|x - m|\sqrt{2}/\sigma)$.

Le texte de 1922 baptise cette propriété exemplaire du nom d'exhaustivité, en donne une définition sous forme de factorisation :

T est un estimateur exhaustif de θ si pour tout autre estimateur T', leur loi conjointe peut se mettre sous la forme d'un produit de deux fonctions dont la seconde est indépendante de θ :

$f(\theta, T, T') = g(\theta, T) \cdot h(T, T')$, généralisée plus tard, par J. Neyman, sous forme de condition nécessaire et suffisante ne faisant pas intervenir d'autre estimateur :

$f(\theta, X) = g(\theta, T) \cdot h(X)$, où X est l'échantillon

Mais surtout, le texte de 1922 associe cette propriété à la méthode du maximum de vraisemblance.

« Pour la solution des problèmes de l'estimation, nous avons besoin d'une méthode qui, pour chaque problème particulier, nous conduise automatiquement à la statistique qui satisfait le critère d'exhaustivité. Une telle méthode est, je crois, fournie par la Méthode du Maximum de Vraisemblance, bien que je ne sois pas satisfait quant à la rigueur mathématique de toute preuve que je puisse fournir à ce sujet. Les lecteurs des pages suivantes sont invités à se faire leur propre opinion sur la possibilité que la méthode du maximum de vraisemblance mène à une statistique exhaustive dans tous les cas. » (1922, p. 323)

Cette affirmation que « toute statistique qui remplit la condition d'exhaustivité doit être une solution obtenue par la méthode de l'optimum » ne permet pas d'affirmer qu'il existe toujours une telle statistique. Et de fait Fisher devra reconnaître qu'il n'en existe pas toujours. Si c'est le cas, alors « cette statistique résume la totalité de l'information sur le paramètre contenue dans l'échantillon. Dans un tel cas, le problème de l'estimation est complètement résolu. »

De plus Fisher montre dès 1922 « qu'une statistique qui satisfait au critère d'exhaustivité, satisfait aussi à celui d'efficacité », dont nous allons maintenant suivre l'émergence.

8. *Vraisemblance et efficacité*

On a vu que le souci de comparer deux estimateurs selon leur variance est ancien et que Fisher propose une mesure de précision relative dès le papier sur l'exhaustivité de 1920. Mais le terme, aussi bien que le lien entre efficacité et vraisemblance datent du texte des « fondements », dans une première version qui est uniquement asymptotique.

C'est en explorant les propriétés de la vraisemblance que Fisher découvre que sa dérivée seconde conduit, au signe près, à l'invariance (l'inverse de la variance) de l'estimateur pour de grands échantillons :

Soit une population définie par la densité $f(x, \theta)$ et T l'estimateur de θ pour un échantillon i.i.d. de taille n :

Si l'on suppose que pour de grands échantillons la loi de T tend vers une loi normale de densité Φ et de variance σ^2, alors Fisher montre (1922, p. 329) que :

$$\mathrm{dLog}L / \mathrm{d}\theta = 0 \qquad (1)$$

est l'équation de vraisemblance, et

$$\mathrm{d}^2 (\mathrm{Log}\,\Phi) / \mathrm{d}\theta^2 = -1/\sigma^2 = E(\mathrm{d}^2 \,\mathrm{Log}\, L / \mathrm{d}\theta^2) = n \cdot E(\mathrm{d}^2 \,\mathrm{Log} f(x, \theta) / \mathrm{d}\theta^2) \qquad (2)$$

où L est la vraisemblance et E l'opérateur d'espérance.

Ainsi, ce concept de vraisemblance fournit-il à la fois l'estimateur, en annulant la dérivée première de L, puis sa variance pour de grands échantillons, en dérivant une seconde fois. L'exemple le plus simple, donné par Fisher, est celui du paramètre p d'une population bernoullienne :

$$L(x, p) = p^k \cdot (1-p)^{n-k} \quad \text{si } k \text{ est le nombre de succès observé}$$

Et

$$\mathrm{dLog}\, L / \mathrm{d}p = k/p - (n-k)/(1-p) = 0 \Rightarrow p = k/n \quad \text{(l'estimateur)}$$

$$\mathrm{d}^2 \,\mathrm{Log}\, L / \mathrm{d}p^2 = -k/p^2 - (n-k)/(1-p)^2,$$

or $E(k) = np$ d'où

$$E(\mathrm{d}^2 \,\mathrm{Log}\, L / \mathrm{d}p^2) = -1/\sigma^2 \Rightarrow \sigma^2 = p(1-p)/n.$$

Fisher écrit en 1922 que la valeur de σ ainsi trouvée est « la plus petite valeur possible de l'écart-type d'une statistique choisie pour estimer un paramètre; elle peut donc être appliquée au calcul de l'efficacité de toute autre statistique ». Mais cette affirmation, bien qu'illustrée par des applications à l'ajustement de courbes du système de Pearson, n'est pas vraiment démontrée, pas plus que ne l'est l'efficacité des estimateurs du maximum de vraisemblance.

Le texte de 1925 « démontre que si une statistique efficace existe, alors elle peut être obtenue par la méthode du maximum de vraisemblance ». (p. 710)

En effet, si T est un estimateur asymptotiquement normal quelconque, et T' l'estimateur du maximum de vraisemblance, Fisher établit que :

$$1 / n\, V(T) = A - n\, A^2\, V(T') \quad (3)$$

où $V(T')$, qui est la variance de T' à l'intérieur des classes définies par T, ne peut être nulle que pour $T = T'$. Dans ce cas $V(T) = V(T') = 1/nA$. Dans tous les autres cas $V(T) > V(T')$.

De fait il faut distinguer ce *critère d'efficacité* (absolue) de la *mesure d'efficacité* (relative) d'une statistique quelconque :

« Le critère d'efficacité nécessite que la valeur fixe vers laquelle tend le produit par n de la variance d'une statistique (de la classe dont nous parlons [estimateurs asymptotiquement normaux]) soit aussi faible que possible. »

« Si nous connaissons la variance de toute statistique efficace, et celle d'une autre statistique à l'étude, l'efficacité de cette dernière peut être calculée par le rapport des deux valeurs. L'efficacité d'une statistique représente la fraction de l'information pertinente disponible qui est effectivement utilisée, dans les grands échantillons, par la statistique en question. » (1925, p. 703).

Toute une série de propriétés corrélationnelles asymptotiques des statistiques efficaces est alors passée en revue. Par exemple :

« Si A est une statistique efficace et B un estimateur non efficace du même paramètre, dont l'efficacité est E, alors pour de grands échantillons, la corrélation entre A et B tend vers la limite $r = \sqrt{E}$. » (1925, p. 705)

L'innovation principale du papier de 1925 se situe toutefois dans la reconnaissance que, pour de petits échantillons, les statistiques efficaces ne peuvent plus être considérées comme équivalentes, et « qu'il faut poursuivre la discrimination parmi le groupe efficace, une discrimination essentielle à l'avance de la théorie des petits échantillons. » (1971)

C'est la notion de *« précision intrinsèque »* d'une distribution, qui va servir à la fois à une extension de la définition de l'efficacité et à l'introduction du concept d'information.

« Si la variance d'une statistique efficace issue d'un grand échantillon de taille n est A/n alors la précision intrinsèque de la distribution est définie par $1/A$. »

C'est ainsi que la précision intrinsèque d'une loi de Cauchy de paramètre m est de 1/2 puisque la variance de l'estimateur du maximum de vraisemblance est $2/n$.

Des formules (2), on en déduit l'expression de cette mesure :

$$1 / A = 1 / n\sigma^2 = -E\,(d^2 \operatorname{Log} f(x, \theta) / d\theta^2) \quad (4)$$

laquelle permet une extension de la définition de l'efficacité pour les petits échantillons :

« L'efficacité est le rapport de la précision intrinsèque de sa loi d'échantillonnage, à la quantité d'information contenue dans les données dont elle a été dérivée » (p. 714), c'est-à-dire encore au produit par n de la précision intrinsèque de la loi de X.

Cette définition a l'avantage de s'appliquer à des échantillons finis, ou à des cas où la distribution asymptotique des estimateurs n'est pas normale, tout en étant compatible avec la définition de l'efficacité donnée dans le cadre asymptotique.

Elle permet si l'on connaît la loi de l'estimateur, d'étudier la vitesse à laquelle l'efficacité de telle statistique tend vers sa valeur limite obtenue pour une taille n infinie, c'est-à-dire encore la vitesse à laquelle la perte d'information tend vers une limite finie.

Fisher montre d'ailleurs que « l'efficacité [ainsi définie] ne peut jamais excéder l'unité » et que ce cas limite d'une perte d'information nulle n'est obtenue qu'à la condition que f puisse se mettre sous la forme :

$$f(\theta, T, T') = g(\theta, T)\, h(T, T')$$

qui est la condition pour que T soit une statistique exhaustive.

Mais « si l'ensemble des échantillons qui, pour une valeur de θ ont la même valeur de $dL / d\theta$, n'ont plus cette même valeur pour d'autres valeurs de θ, il n'existe plus de statistique exhaustive, et il s'ensuivra nécessairement une perte d'information dans l'opération de résumé des données originales par un seul estimateur » (p. 718).

« La quantité d'information perdue peut alors différer pour différentes statistiques efficaces. Mais elle sera minimale pour la solution des équations du maximum de vraisemblance. » (p. 720).

La fin du texte de 1925 est consacrée en outre à l'usage de statistiques auxiliaires qui, combinées avec la statistique optimale, peut annuler asymptotiquement cette perte d'information.

« Ces statistiques auxiliaires, qui en elles-mêmes ne nous disent rien sur la valeur du paramètre, mais nous donnent la qualité d'un estimateur de ce paramètre », ... « ne sont utiles que si différents échantillons de la même taille peuvent fournir différentes quantités d'information. Elles permettent de distinguer alors ceux qui fournissent plus d'information de ceux qui en fournissent moins. » (1935, p. 48)

Ainsi la boucle est-elle bouclée, en ce sens que l'exhaustivité apparaît comme un cas limite de l'efficacité si l'on raisonne en terme d'information, et que la méthode du maximum de vraisemblance conduit toujours à un résultat optimal. Mais c'est bien ce concept d'information sur lequel il nous faut, pour finir, revenir, qui assure l'intégration de tous les éléments sur la théorie de l'estimation chez Fisher.

9. *L'Information de Fisher*

La notion d'information est, on l'a vu, présente dès les premiers textes de Fisher, en fait dès qu'il définit l'objet de la statistique comme étant celui de la réduction des données.

Ce qui n'est d'abord qu'une *notion*, et qui joue déjà un rôle important dans son modèle statistique tel qu'il l'expose en 1922, va devenir une *grandeur mesurable*, puis un *concept* qui intègre les différents éléments de son modèle inférentiel.

La notion est déjà novatrice dans le champ de la statistique. Celle-ci, « science des ensembles nombreux », s'est construite, au XIXe siècle, sur la base d'un vaste mouvement de recueil et de codification de l'information économique et sociale, et la théorie des moyennes fonde son autorité sur la masse et la redondance de cette information diffuse. La même notion de répétition fonde chez Pearson la loi scientifique comme « résumé sténographique des routines de perception » (1892). Mais toute idée de différenciation de l'information véhiculée intrinsèquement par une observation unique selon la loi qui la réalise, suppose une prééminence, qu'il refuse, du modèle théorique sur l'observable.

Alors que la mesure de l'information de Shanon (1949) est davantage une mesure « du choix et de l'incertitude », dans un projet autre de théorie de la communication, « Le concept d'information fisherienne était un concept de précision » dit avec juste raison Lancry (1982); et il en est ainsi de la mesure de l'information (qui n'est d'ailleurs, chez Fisher, ni « secondaire » ni tardive).

La généalogie même de cette mesure d'information est claire. Après avoir, dès 1922 mais dans un exemple, introduit l'expression de « précision intrinsèque d'une courbe d'erreur » (p. 339), Fisher reprend cette notion en 1925. Il la nomme $1/A$, la définit en fonction de la seule loi des observations (*cf.* [4]) et ajoute :

> *« Ce dont nous venons de parler en terme de précision intrinsèque d'une courbe d'erreur peut aussi bien être conçu comme la quantité d'information dans une seule observation appartenant à une telle distribution. » (1925)*

Mais ce rapprochement n'est-il pas abusif, comme l'est l'usage métaphorique du terme information? Il lui sera reproché longtemps, d'associer sa construction mathématique avec un terme aussi chargé de connotations dans le langage commun. Pourtant, n'est-ce pas justement le propre du travail mathématique que de construire, sur la base d'une prénotion dont les significations multiples composent une représentation mentale, un objet univoquement défini par les seules propriétés de la syntaxe mathématique? Fisher (1935) nous donne raison :

> *« Il est évident aussi, qu'en introduisant le concept de quantité d'information, nous ne voulons pas donner un nom arbitraire à une quantité calculable, mais devons nous préparer à justifier le terme employé, en relation avec ce que requiert le sens commun, et dire si le terme est approprié et utile comme instrument de la pensée. Les conséquences mathématiques de l'identification que je propose entre la précision intrinsèque d'une courbe d'erreur et la quantité d'information extraite, peuvent donc être maintenant résumées, afin de juger d'après notre sens commun pré-mathématique, si elles ont bien les propriétés qu'elles doivent avoir. » (1935, p. 47)*

Et d'enchaîner sur les propriétés d'*additivité* de cette mesure, énoncées dès 1925, pour deux, puis n observations.

Il est montré que (1934) :

si $$i(\theta, x) = 1/A = \Sigma(1/f)\,(\mathrm{d}f/\mathrm{d}\theta)^2,$$

où la somme Σ porte sur toutes les observations possibles, mesure la quantité d'information fournie par chacune des observations sur le paramètre θ, pour une densité f donnée, et indépendamment de la méthode d'estimation, alors :

1. « La quantité d'information contenue dans la réunion de deux ensembles indépendants d'observation est la somme des quantités d'information de chaque ensemble pris séparément. »

$$i(\theta, x, y) = i(\theta, x) + i(\theta, y)$$

Cette additivité s'appliquant à tout l'échantillon $X = (x_1, \ldots x_n)$ permet de définir l'information fournie par celui-ci comme $I(\theta, X) = ni(\theta, x) = n/A$.

2. « Si en augmentant le nombre d'observations, l'erreur d'échantillonnage d'un estimateur efficace tend vers la normalité, la quantité d'information est proportionnelle à la précision, constante, de cette distribution limite. » Soit $I(\theta, T) = 1/V$ pour cette classe d'estimateurs.
3. « La quantité d'information fournie par toute statistique ou groupe de statistiques ne peut jamais excéder l'information totale contenue dans les données originales. » Soit en général

$$I(\theta, X) = 1/V.$$

Cette dernière propriété sous forme d'inégalité, que l'on pourrait dire *principe entropique*, a conduit Fisher à développer ce qui n'était alors qu'une analogie forte :

« Comme quantité mathématique, l'information est de façon saisissante semblable à l'entropie dans la théorie mathématique de la thermodynamique. Vous noterez en particulier que les processus réversibles, les changements de notation, les transformations mathématiques, la traduction de données en un autre langage, ou sa réécriture en code ne peuvent être accompagnés d'une perte d'information; mais que les processus irréversibles impliqués dans l'estimation statistique, où les données originales ne peuvent être reconstruites à partir des estimateurs calculés à partir d'elles, peuvent être accompagnés d'une perte, mais jamais d'un gain. » (1935, p. 47)

Peut-être sous l'emprise de cette analogie avec une grandeur physique, ou bien encore sous l'influence de la reconstruction axiomatique de la mesure de probabilité par Kolmogorov dans les mêmes années 30, Fisher a suggéré une reconstruction axiomatique semblable de son concept d'information en prenant comme postulats les propriétés suivantes :

— La mesure vaut zéro pour un échantillon dont la loi est indépendante du paramètre.

— La mesure vaut 1 pour une statistique exhaustive.

— La mesure est bornée supérieurement par sa valeur pour l'échantillon.

— Les mesures sur des observations indépendantes sont additives.

Fisher avoue toutefois « qu'il est personnellement plus incliné à examiner cette quantité telle qu'elle émerge des investigations mathématiques, et à juger de son utilisé selon le libre usage du sens commun, plutôt qu'à l'imposer par une définition formelle. » (1935, p. 47)

Cet attachement à l'usage et la signification d'un concept, plutôt qu'à la rigueur formelle de la construction mathématique, caractérise bien le style, intuitif, incisif, et inductif de sa pensée mathématique, mais aussi la prépondérance chez lui de ce que nous appelons la sémantique de son modèle inférentiel sur le déroulement besogneux de la syntaxe formelle.

Ses textes des années 30 marquent l'intégration et la complétude de sa théorie de l'estimation. A part une mise en correspondance de sa méthode du maximum de vraisemblance avec celle du khi-deux minimum, c'est bien d'une théorie achevée qu'il s'agit, même si d'autres plus tard ajouteront un peu de rigueur à ce paradigme.

Mais de cette théorie de l'estimation, Fisher éprouve le besoin dans les années 30, d'en faire le prototype d'une *théorie générale de l'inférence inductive en univers incertain.* Dans les principaux textes de cette époque, il développe l'idée que cette inférence inductive est une nécessité de l'investigation scientifique, que rigueur et précision ne sont point incompatibles avec le caractère incertain de l'opération, et, enfin, que sa théorie de l'estimation est le noyau de cette théorie générale.

Les deux concepts de quantité d'information et de vraisemblance jouent alors un rôle central et complémentaire dans cette théorie, relayant la mesure de probabilité, réservée aux raisonnements déductifs, et fondée sur la notion de limite de fréquence :

« Ainsi nous avons, en plus de la probabilité de la théorie classique, déjà deux autres caractéristiques quantitatives, appropriées à différentes situations logiques correspondant à différentes sortes d'inférence incertaine. » (1934, p. 4)

Armé de ces deux concepts, Fisher peut batailler contre les systèmes concurrents ou les récupérer en montrant que leurs légitimes préoccupations sont traduites, avec davantage de rigueur et sans distorsion de ce qui fait consensus, par son propre système.

La notion bayesienne de probabilité à priori est, on l'a vu, réinterprétée à travers son concept de super-population. Il note (1932) que la probabilité conjointe de l'échantillon et d'une valeur du paramètre inconnu est alors constituée de deux facteurs assimilables aux contributions de notre connaissance à priori d'une part, à celle de l'échantillon d'autre part; et que si la taille de l'échantillon s'accroît indéfiniment, alors bien sûr le premier facteur s'évanouit et nos conclusions ne dépendent plus de cet à priori. Mais comment interpréter cette convergence entre les deux approches :

« La conclusion qui doit être tirée de cette importance décroissante de notre information a priori n'est pas, trivialement, qu'en introduisant un faux a priori nous ne serons pas trop égaré, mais bien plutôt, et d'une façon fondamentalement différente, que des conclusions ne peuvent être tirées que des seules données. » (1932, p. 259)

Mais cette information a priori n'est pas évacuée du modèle de Fisher; plutôt que de fausser la rigueur du raisonnement par des choix arbitraires, elle est repoussée à la phase de spécification :

[Dans notre approche] « nous sommes dépourvus de toute connaissance des probabilités a priori des différentes valeurs de ces paramètres, ou bien encore, nous ne voulons pas introduire la vague connaissance qui est la nôtre à ce sujet, comme base d'un argument mathématique rigoureux.

La connaissance a priori peut être, et doit être utilisée quand nous en sommes à la spécification des formes de la population que nous avons à considérer. La caractéristique logique principale de cette ligne d'approche, est qu'elle sépare la question de la spécification de la question suivante de l'estimation, laquelle ne peut avoir lieu que lorsqu'une spécification est acceptée. » (1934, p. 7)

De la même façon la tentative de Jeffreys de construire une mesure subjective de probabilité pour traiter ce problème d'inférence incertaine — et dans une moindre mesure celle de Keynes — est mise en pièce par Fisher (1934) car « d'une inconsistance inéradicable », faisant fi de 200 ans d'utilisation de la probabilité en un sens fréquentiste, et mélangeant enfin dans cette construction l'idée de masse d'information en faveur d'une hypothèse et celle de degré selon lequel cette information fournit des arguments à cette hypothèse. Les deux concepts de Fisher distinguent, eux, ces deux approches, et rien n'empêche de considérer que :

« la vraisemblance fournit, dans la situation logique qui prévaut dans les problèmes d'estimation, une mesure de croyance rationnelle analogue, bien que mathématiquement différente, à celle que fournit la probabilité mathématique dans ces problèmes d'inférence déductive pour laquelle la probabilité fut développée. » ... « nous sommes amenés à reconnaître en cette quantité le médium par lequel toute information que nous possédons peut être convoyée de façon appropriée. » (1935, p. 53)

Profondément intégré, prêt à récupérer d'autres approches quand il ne peut totalement interdire leur développement, le système de Fisher est voué par lui à un grand avenir, puisqu'il doit générer ni plus ni moins qu'une nouvelle façon rigoureuse de penser inductivement, qui est comme une alchimie de la vérité, dont le mythe remonte à la théorie des erreurs, mais qui serait enfin mathématiquement dominé :

« L'étude du raisonnement inductif est l'étude de l'embryologie de la connaissance, c'est-à-dire du processus selon lequel la vérité est extraite de son minerai d'origine, dans lequel elle fut amalgamée avec beaucoup d'erreur. » (1935, p. 54)

C'est, pour finir, à la réalité de cette embryologie que nous allons nous intéresser, c'est-à-dire pour reprendre une formule de B. Latour (1984), à la façon dont ce système a résisté aux épreuves qui lui ont été imposées par les autres acteurs, ou par son propre auteur.

10. *Résistance du modèle de Fisher*

Nous utiliserons pour cela, outre les matériaux déjà évoqués, la discussion du papier sur « la logique de l'inférence inductive » lu par Fisher à la Royal Statistical Society, puis publié en 1935. C'est un lieu que ne fréquentait guère Fisher, et qui le confronte sans ménagements aux statisticiens, dont le nombre, la diversité doctrinale et la notoriété vont nous permettre de considérer cette séance comme une sorte de microcosme de la communauté scientifique.

Qu'on en juge par cette liste, dans l'ordre d'intervention, que Fisher, dans sa réponse, caractérise comme un ordre « d'animosité décroissante » : Pr A.L. Bowley, Dr Isserlis, Dr Irwin, Pr Wolf, Dr E.S. Pearson, Pr Greenwood, Dr H. Jeffreys, M.S. Bartlett, Dr J. Neyman.

Ce que font ces distingués statisticiens consiste essentiellement en une démolition de l'immeuble fishérien, et nous pouvons facilement repérer dans leurs arguments l'étage qui en est visé, puis dans la réponse de Fisher, les effets de cet ébranlement.

La première critique se rapporte justement à la syntaxe mathématique de Fisher, jugée « très obscure » (Bowley), inaccessible à une bonne part des lecteurs (Isserlis), et antipédagogique (Greenwood). Ce manque de lisibilité des écrits de Fisher est tour à tour attribué par ses pairs, à son génie transcendant porté sur les court-circuits de la pensée (des « gaps » multiples qui nécessitent de construire soi-même le pont qui les enjambe), à sa roublardise (« il cache la clé », il réécrit dans un autre langage, plus abscons, ce que les autres — Edgeworth ou Pearson — ont trouvé avant lui), à sa modestie (« Fisher a peur d'ennuyer »), ou encore à sa manie d'inventer de nouveaux termes à tout bout de champ.

On reconnaîtra là, outre l'ironie méchante de ces remarques qui n'est pas d'habitude attribuée aux représentants de ce milieu, tout un ensemble de critiques plus ou moins voilées sur la solidité syntaxique de la théorie fishérienne. De fait, la difficulté est réelle pour le mathématicien lecteur, car Fisher multiplie les ellipses, et surtout, il alterne des bribes de démonstrations (ou définitions) formelles et d'autres imagées, empruntées à la sphère des représentations; par exemple, l'efficacité d'une statistique est dans la même phrase un rapport de variance et une fraction d'information disponible, alors que cette dernière notion n'est pas encore définie mathématiquement.

Cet aller et retour permanent entre syntaxe et sémantique déroute et rebute ses lecteurs qui y voient un placage de significations aux connotations discutables :

« L'introduction de ce que les individus grossiers appellent « charabia » porte à la confusion en attachant des significations particulières à des mots bien établis de la langue courante des gens éduqués » dit Greenwood, que « troublent » les nouveaux sens de variance, statistique et information.

« La mesure sur cette base de la quantité de connaissance [d'information] me semble avoir les mêmes dangers que le traitement du coefficient de corrélation ou de son carré en terme de masse de covariation. » (Bowley)

Or de cette prédilection pour *une mathématique qui ne se réduit pas à un jeu syntaxique*, Fisher s'en explique justement à la fin du texte qui est discuté :

« La rigueur, telle qu'elle est comprise en mathématique déductive, n'est pas suffisante. Dans le raisonnement déductif, des conclusions fondées sur un petit nombre de postulats acceptés n'ont besoin que de la rigueur mathématique pour garantir leur vérité. Tous les statisticiens savent que les données sont falsifiées si on n'en utilise qu'une partie. Le raisonnement inductif ne peut pas prétendre à une vérité qui est moins que toute la vérité. Nos conclusions doivent être garanties par la totalité des données, puisqu'à moins que cela nous pouvons être induit en erreur. Ceci bien sûr n'est pas un argument contre l'usage des formes tout à fait précises que peut prendre, si possible, un énoncé. C'est seulement un avertissement à ceux qui peuvent être tentés de penser que le code particulièrement précis des énoncés mathématiques auquel ils ont été exercés au « Collège » est un substitut à l'usage de forces de raisonnement que l'humanité a probablement possédé depuis les temps préhistoriques, et pour lesquelles le processus de codification est encore incomplet, comme le montre l'histoire de la théorie des probabilités. » (1935, p. 54)

Dans cette même discussion, un autre ensemble de critiques porte sur la signification et la portée de son innovation. Il s'agit pour l'essentiel d'un débat sur la vraisemblance. Y a-t-il d'abord nécessité d'un nouveau concept à ce sujet? Est-il vraiment nécessaire de l'opposer à la probabilité (Jeffreys), alors que cette vraisemblance est proportionnelle à la probabilité a posteriori pour les grands échantillons? La notion de vraisemblance n'est-elle pas incluse dans la théorie des probabilités (Neyman)?

Ces questions ne placent plus le débat sur la validité et la lisibilité des écritures de Fisher, mais bien sur ce qu'elles impliquent quant à l'articulation de ces notions par rapport à d'autres préexistantes, sur les bases desquelles la discipline encore jeune est fondée, c'est-à-dire encore *l'insertion du modèle de Fisher dans un modèle plus vaste qui est celui de l'inférence statistique.* Il est même question de savoir — Wolf est le logicien « de service » délégué à ce problème — si l'inférence tout court doit être nécessairement statistique ou mathématique.

Or, une façon de trancher la question de cette insertion sémantique du modèle de Fisher est de le tester sur son efficacité à résoudre des problèmes concrets. Mais quels problèmes? L'irruption de la sphère concrète dans ce débat ne peut se faire directement, puisque ceux qui ont des problèmes relevant de la méthodologie statistique ne sont pas à cette séance de la R.S. pour être les juges en dernière instance de cette efficacité. On assiste donc, ce qui est une phase classique du travail scientifique, à une *traduction* de ces besoins par les statisticiens présents selon des problématiques contradictoires.

Ce sont naturellement E.S. Pearson et Neyman qui sont les mieux placés pour ce niveau de la discussion puisqu'ils construisent dans ces mêmes années 30 un autre modèle d'inférence. Ils ne critiquent pas la validité du modèle de Fisher qu'ils félicitent d'ailleurs, mais critiquent *le domaine* sur lequel s'exerce cette validité.

Egon Pearson par exemple se déclare d'accord avec Fisher pour placer le débat sur le versant de la logique et non celui de la mathématique, mais il lui reproche de surestimer l'importance de sa théorie qui ne saurait couvrir tout le champ de l'inférence. Il propose une catégorisation de la problématique de l'inférence en trois formes qui traduisent mieux la diversité des attentes des utilisateurs, et qui nous sont aujourd'hui familières : estimation ponctuelle, estimation par intervalle, et test d'hypothèse. Or Fisher ne traite que de la première, bien qu'il ait tenté d'aborder la seconde par sa théorie de la probabilité fiduciaire, controversée et concurrencée par la théorie des intervalles de confiance de Neyman et Pearson.

L'intervention, écrite, de Neyman est de même type. Après une analyse élogieuse, quasi kuhnienne, de la révolution que représente le travail de Fisher, et un rappel « des critiques silencieuses mais éloquentes » qui furent le seul accueil fait en dix ans aux quatre éditions de son livre, Neyman fait suivre cette douche écossaise d'un résumé du squelette de la théorie statistique de Fisher qui sert de faire valoir à sa propre conception :

> *« Celui-ci contient le principe de choix parmi les estimateurs, qui est la notion intéressante et importante de quantité d'information. La fonction de vraisemblance semble jouer un rôle secondaire par rapport à cette quantité, dans la production des estimateurs. » ... « Je pense personnellement que la quantité d'information est un concept trop compliqué et trop vague pour servir de principe. Bien sûr, si quelqu'un dit « ne faites pas cela parce que vous allez perdre de l'information » on aura tendance à accorder son comportement à cet avis. Mais une telle action n'est due qu'au pouvoir suggestif des mots « masse d'information perdue ».*
>
> *« Maintenant, qu'est-ce qui peut être considéré comme un principe suffisamment simple et hors de cause dans le travail statistique? Je pense que le concept de base est celui de fréquence des erreurs de jugement. » ... « Le complexe de résultats obtenus dans cette direction peut être considéré comme un système de statistique mathématique alternatif à celui du profeseur Fisher, et entièrement fondé sur la théorie classique des probabilités. »*

Dès lors que peut faire Fisher si ce n'est congratuler Neyman et Pearson pour leurs travaux, qui par ailleurs ne portent pas ombrage aux siens, puisqu'ils en confirment certains résultats.

Pour conclure maintenant sur la résistance du modèle fishérien, il faut dire qu'en arrêtant cette étude au début des années 30, et en la limitant à la seule estimation ponctuelle, nous manquons bien sûr à la tâche d'une évaluation exhaustive.

En ce qui concerne sa syntaxe, il faut rappeler que les propriétés des objets mathématiques mis en place par Fisher ont été par la suite précisées, démontrées avec plus de rigueur, et étendues; il suffit d'évoquer par exemple les noms de Frechet, Darmois, Cramer et Rao attachés à la fameuse inégalité qui exprime la borne de la variance de tout estimateur, le théorème de Koopman et son application au cas particulier de la famille exponentielle, ou encore les travaux sur la robustesse.

La résistance sémantique de la théorie de Fisher ne pourrait s'évaluer que dans sa mise à l'épreuve par les modèles concurrents de l'approche bayesienne, de la phénoménologie pearsonienne, et de la théorie de la décision.

L'école bayesienne contemporaine prouve que le premier de ces modèles concurrents n'a pas succombé aux attaques de Fisher.

La philosophie scientifique de Pearson a été prolongée par les tenants d'une statistique « sans modèle », que l'on retrouve par exemple dans l'école de l'Analyse des Données :

« Karl Pearson a vu son rôle minimisé par la génération de statisticiens élevés dans la doctrine de R.A. Fisher, lequel corrigea maintes inexactitudes de son devancier et donna à la méthode statistique une forme plus cohérente mais dirons-nous moins accueillante aux flots de la nature. On nous permettra de choisir ici le patronage de K. Pearson. » (Benzecri, 1982, p. 46)

Mais cette même École avoue aussi sa dette à l'égard de Fisher en ce qui concerne sa traduction géométrique de la statistique.

Quant au paradigme de la décision qui naît, à la fin de la période que nous avons étudiée, des travaux de Neyman et de Wald, il s'agit effectivement d'un modèle concurrent de celui de l'inférence tel que Fisher la conçoit, mais dont la stratégie est celle de l'assimilation : la théorie de l'estimation, et plus généralement l'inférence statistique sont réinterprétées dans le cadre plus vaste de ce paradigme de la décision. Cette assimilation n'est pas du goût de Fisher, qui oppose, de façon irréductible, la problématique du scientifique à la recherche du modèle vrai à la problématique de l'homme d'action à la recherche d'une règle d'action.

« Fisher aurait maintenu (avec juste raison à mon avis) que l'inférence dans les sciences n'est pas une affaire de décision, et que de toute façon, des critères de choix fondés sur des coûts d'une sorte ou d'une autre, n'étaient pas admissibles. C'est en gros le point de vue anglais contre l'américain. Nous verrons couler beaucoup d'eau sous le pont avant que ce conflit soit résolu. Ne souhaitant pas la controverse, je propose la thèse qu'une telle différence d'attitude est inévitable entre les pays où ce que fait un homme est plus important que ce qu'il pense, et ceux où c'est l'inverse. » (Kendall, 1963, p. 442)

On le voit, le conflit des approches se situe bien ici au niveau de la pragmatique des modèles, c'est-à-dire de leur capacité à résoudre des problèmes, et pour cela, d'abord à leur capacité à les poser en des termes qui sont censés traduire les besoins d'autres acteurs sociaux. Ces besoins peuvent d'ailleurs être segmentés en autant de « marchés » que nécessaire. Il est classique par exemple de distinguer celui de *l'observation*, investi par la biométrie anglaise, celui de *l'expérimentation* auquel la méthode de Fisher est constitutivement associé, et celui de *l'action* pour laquelle l'approche décisionnelle incluant coûts et risques est mieux adaptée. Et cette segmentation me semble plus pertinente que celle qui est faite par Kendall en terme de psychologie nationale.

Une des facettes du conflit de Pearson avec les mendeliens était, surtout après la mort en 1906 de Weldon, sa non-familiarité avec toute forme d'expérimentation biologique. On sait qu'au contraire Fisher a passé une bonne part de sa carrière à la station expérimentale de Rothamsted, et que lui-même élevait, chez lui le plus souvent, toute une quantité de « bestioles » sur lesquelles il testait ses modèles génétiques. Ses publications génétiques sont au moins aussi nombreuses que celles qui sont purement statistiques. Et, ne l'oublions pas, ce fondateur de la Statistique Mathématique ne fut jamais professeur de cette discipline, mais titulaire d'une chaire de Génétique.

Il n'est pas étonnant dans ce cas, que ce que Fisher sut associer dans sa vie, le fut aussi dans son œuvre, que la résistance de sa théorie tienne aussi, ou surtout, à son adéquation à la problématique expérimentale, et que son succès provienne de sa capacité à faire de la méthode statistique une technique de preuve, c'est-à-dire une machine à produire des pièces à conviction pour la science expérimentale. Car l'essentiel disait, avec d'autres, Jensen ([1]), c'est bien de *convaincre*.

([1]) voir texte de DESROSIERES dans le même numéro

Conclusion

Convaincre, nous espérons l'avoir fait. Convaincre de l'intérêt d'un retour aux sources des théories qui forment notre culture statistique contemporaine, c'était notre objectif principal. Nous espérons l'avoir atteint, en restituant non seulement la syntaxe mathématique de la construction d'une théorie de l'estimation par Fisher — ce qui est déjà un objectif pédagogique nécessaire — mais en restituant aussi le tissu d'associations dans lequel est prise cette syntaxe, et qui lui donne son sens, sa valeur pratique et sa durabilité.

Sans doute cette étude est-elle trop limitée, dans sa période et dans son objet, puisqu'elle ne couvre ni toute la production de Fisher, ni même tout ce qui a trait à l'estimation. Nous n'avons pas traité ici de l'estimation par intervalle, et des recherches que Fisher a mené autour de la notion de probabilité fiduciaire comme alternative à la construction par Neyman des intervalles de confiance, avec en son centre, la controverse autour du problème de Behrens.

D'autres matériaux enfin pourraient être convoqués avec bonheur à cette reconstitution, par exemple ceux des travaux plus « appliqués » de Fisher qui illustrent les associations que nous suggérons à plusieurs reprises entre modèle statistique, philosophie des sciences et philosophie politique. Mais on n'en finit jamais avec l'histoire.

BIBLIOGRAPHIE

1. *Articles et Ouvrages de R.A. FISHER*

1912 On an absolute criterium for fitting frequency curves. *Messeng. Math.* 41, 155-90, [CP1].

1915 Frequency distribution of the values of the correlation coefficient in samples from an infinitely large population. *Biometrika* 10, 507-21, [CP4].

1918 The correlation between relatives on the supposition of mendelian inheritance. *Trans. Roy. soc. Edinburgh* 52, 399-433, [CP9].

1919 The causes of human variability. *Eugenic Review*, 10, 213-220, [CP10].

1920 A mathematical examination of methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Mon. Not. Rev. Astr. Soc.*, 80, 758-70, [CP12].

1921 On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4), 1-32, [CP14].

1922 On the mathematical foundation of theoretical statistics. *Phil. Trans. A*, 222, 309-68, [CP18].

1922b The goodness of fit of regression formulae and the distribution of regression coefficients. *J.R. Stat. Soc.*, 85, 597-612, [CP20].

1925 Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, 22, 700-25, [CP42].

1925b *Statistical Methods for Research Workers.* 13e édition 1958, Edinburgh : Oliver and Boyd.

1926 Eugenics : can it solve the problem of decay of civilization? *Eugenic Review*, 18, 128-36, [CP53].

1930 Inverse probability. *Proc. Camb. Phil. Soc.*, 26, 528-35, [CP84].

1930b *The Genetical Theory of Selection.* Oxford : Clarendon Press.

1932 Inverse probability and the use of Likelihood. *Proc. Camb. Phil. Soc.*, 28, 257-61, [CP95].

1934 Probability, likelihood and quantity of information in the logic of uncertain inference. *Proc. Roy. Soc.*, A, 146, 1-8, [CP109].

1934b Two new properties of mathematical likelihood. *Proc. Roy. Soc.*, A, 144, 285-307.

1935 The logic of inductive inference. *J.R. Stat. Soc.*, 98, 39-82, [CMS26] (sans discussion) [CP124] (discussion).

1935b *The Design of Experiment.* Edinburgh : Oliver and Boyd.

1936 Uncertain inference. *Proc. Amer. Acad. Arts. Sci.*, 71, 245-58, [CMS27].

1937 Professor Karl Pearson and the method of moments. *Annals of Eugenic*, 7, 303-18, [CP149].

1950 *Contributions to Mathematical Statistics*, [CMS], Wiley, N.Y. Introduction biographique de Yates et Mather.

1971-1974 *Collected Papers of R.A. Fisher.* Ed. J.H. Benett, Univ. d'Adelaide.

II. *Autres auteurs*

BAYES T. 1764, "An essay toward solving a problem in the doctrine of chances", *Phil. Trans. of the R.S. of London*, 53, p. 370-418. Réédité par Pearson et Kendall, 1970. Édition critique J.P. Clero, Cahiers d'Histoire et de Philosophie des Sciences, 18, Paris 1987.

BENZECRI J.P. 1982, *Histoire et préhistoire de l'Analyse des Données*, Paris : Dunod.

BERNOULLI D. 1778, "The most probable choice between several discrepant observations", trad et commentaires dans Pearson et Kendall 1970.

BERNOULLI J. 1713, *Ars Conjectandi*, trad. N. Meusnier : « Quatrième partie de l'art de conjecturer », Rouen : IREM, 1987.

BLOOR D. 1976, *Sociologie de la logique*, London : Routledge and Regan, trad. Erbnother, Paris. Pandore.

DESROSIÈRES A. 1985, « Histoire de formes : statistiques et sciences sociales avant 1940 », *Rev. Fr. de Sociologie*, vol. XXVI, 2, 277-310.

EDWARDS A.W.F. 1972, *Likelihood*, Cambridge Univ. Press.

EISENHART C. 1981, "Karl Pearson", in GILLEPSIE, Dictionary of Scientific Biography.

FISHER-BOX Joan 1978, *R.A. Fisher, the life of a scientist*, N.Y. : John Wiley and Sons.

GOSSET W.S. ("STUDENT") 1908, "The probable error of a correlation coefficient, *Biometrika*, 6, 302.

HACKING I. *The Emergence of Probability. Logic of Statistical Inference*, Cambridge Univ. Press.

KENDALL M.G. 1963, "Ronald Aymler Fisher, 1890-1962", *Biometrika*, 50, 1-15. Repris dans Pearson E.S. et Kendall M.G. 1970.

KENDALL M.G. et PLACKETT R. 1977, *Studies in the history of statistics and probability*, vol. II, London : Griffin.

LANCRY P.J. 1982, *Théorie de l'information en économie*, Paris : Economica.

LAPLACE P.S. 1774, «Mémoire sur la probabilité des causes par les événements », *Œuvres Complètes*, vol. VIII, 27-65

LAPLACE P.S. 1812, *Théorie Analytique des probabilités*, 3^e^ édition avec suppléments 1920, Œuvres complètes, VII.

LATOUR B. 1984, *Les microbes : guerre et paix*, suivi de *Irreductions*, Paris : A.M. Metailié.

LATOUR B. 1985 « Les vues de l'esprit », *Culture technique*, 14, 4-29.

MACKENZIE D. 1981, *Statistics in Britain 1865-1930. The Social construction of scientific knowledge*, Edinburgh Univ. Press.

MEUSNIER N. 1987, « Jacques BERNOULLI et l'Ars Conjectandi, Documents pour l'étude de l'émergence d'une mathématisation de la stochastique », Université de Rouen - IREM.

PEARSON E.S. 1968, "Some early correspondance between W.S. Gosset, R.A. Fisher and Karl Pearson, with notes and comments", *Biometrika*, 55, 445-57. In Pearson et Kendall 1970.

PEARSON E.S. et KENDALL M.G. 1970, *Studies in the History of Statistics and Probability*. Vol. I, London : Griffin.

PEARSON K. 1892, *La Grammaire de la Science*, Traduction sur la 3^e^ édition (1911) par L. March, Paris : Alcan.

PEARSON K. 1895, « Contribution to the Mathematical Theory of Evolution, II. Skew variation in homogeneous material », *Phil. Trans. of the R.S.*, 186A, 343-314.

PEARSON K. et FILON L.N.G. 1898, "Mathematical Contributions to the Theory of Evolution IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation", *Phil. trans. of the R.S.*, 191A, 229-311.

PEARSON K., SOPPER H.E., YOUNG A.W., CAVE B.M., et LEE A. 1917, "On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R.A. Fisher. A Cooperative Study, *Biometrika*, 4, 328-413.

SHAPIN S. 1982, « L'histoire des Sciences est-elle possible », *History of Science*, 20, 157-211, trad. Erbnother, Paris : Pandore, 1975.

STIGLER S.M. 1973, « Laplace Fisher and the discovery of the concept of sufficiency », *Biometrika*, 60, 439-45, rééd. Kendall et Plackett 1977.

STIGLER S.M. 1986, *The History of Statistics : the measurement of uncertainty before 1900*, Harvard Univ. Press.

TASSI P. 1985, *Méthode statistique*, Paris : Economica.

THÉVENOT L. 1985, « Les investissements de forme » in L. Thevenot (éd.) : *Conventions économiques*, Paris : PUF.

TODHUNTER I. 1865, *A History of the Mathematical Theory of Probability*. Cambridge et London : MacMillan.